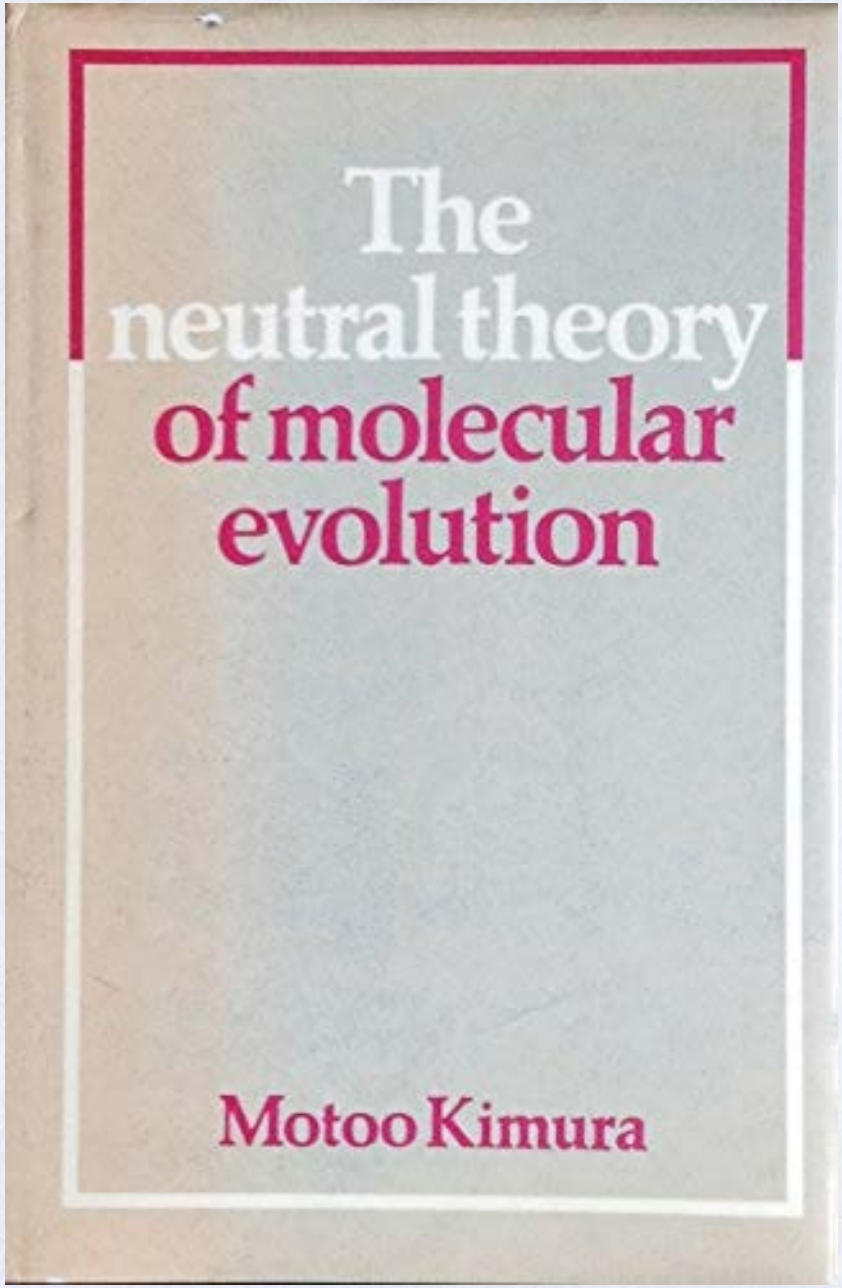
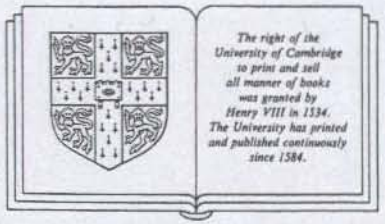


THE NEUTRAL THEORY
OF MOLECULAR EVOLUTION

MOTOO KIMURA
National Institute of Genetics, Japan

1983



CAMBRIDGE UNIVERSITY PRESS
Cambridge
London New York New Rochelle
Melbourne Sydney

Contents

Preface

This book represents my attempt to convince the scientific world that the main cause of evolutionary change at the molecular level – changes in the genetic material itself – is random fixation of selectively neutral or nearly neutral mutants rather than positive Darwinian selection. This thesis, which I here call the neutral theory of molecular evolution, has caused a great deal of controversy since I proposed it in 1968 to explain some then new findings in evolution and variation at the molecular level. The controversy is not surprising, since evolutionary biology has been dominated for more than half a century by the Darwinian view that organisms become progressively adapted to their environments by accumulating beneficial mutants, and evolutionists naturally expected this principle to extend to the molecular level. The neutral theory is not antagonistic to the cherished view that evolution of form and function is guided by Darwinian selection, but it brings out another facet of the evolutionary process by emphasizing the much greater role of mutation pressure and random drift at the molecular level.

The neutral theory has two roots. One is the stochastic theory of population genetics whose foundation traces back to the pathbreaking work of R.A. Fisher, J.B.S. Haldane and Sewall Wright early in the 1930s, and is mathematical in nature. The other is molecular genetics, which has revolutionized our concept of life and whose impact we are still feeling. In particular, molecular advances have permitted the study of evolution at the most basic level, DNA itself – something not imagined two decades ago. The study of intraspecific genetic variability has been similarly revolutionized by new molecular approaches.

Many of the arguments employed to support or refute the neutral theory are quantitative in nature, and are often meticulous and difficult; a careful

and detailed explanation is necessary. This is one reason why the book has become larger than originally intended. In addition, the past few years have produced an explosion of new molecular knowledge bearing on the theory. Naturally, I have tried to incorporate as much of this as I can.

It brings me satisfaction to observe that the neutral theory has survived over a decade of severe tests of many kinds; I believe the probability of its future survival is high, although refinements and modifications will no doubt be made. It is often said that science develops through cooperation of many people throughout the world, and the subject treated in this book is no exception. Readers will see that scientists from various nations have participated in the enterprise. Their work, even when done to refute the theory, have pointed the way to revisions to make it more realistic. The neutral theory of evolution has had an evolution of its own through the interaction between theoretical analyses and experiments or observations, as in the typical scientific paradigm.

The writing of this book has cost me three years of hard work, but I feel some excitement in the realization that in 1982 we commemorate the hundredth anniversary of the death of Charles Darwin. His theory of evolution by natural selection has been the great unifying principle in biology. As new molecular knowledge unfolds we have a new uncharted territory awaiting exploration. I hope that the neutral theory represents a step forward in our search for molecular understanding of evolution and variability.

Drafts of parts of this book have been read by Nigel Calder, Daniel Hartl, Thomas Jukes, Takashi Miyata, Terumi Mukai, Masatoshi Nei, Tomoko Ohta, William Provine, Jacques Ruffié and Naoyuki Takahata. Their comments were much appreciated. I am particularly grateful to James Crow for numerous criticisms and for helping me over various semantic hurdles. Kenichi Aoki also went carefully through the entire manuscript and made suggestions for improved presentation. I would like to take this opportunity to express my indebtedness to James Crow for his guidance and help and to Tomoko Ohta for her cooperation and constructive criticism. Without their help I would never have been able to develop the theory presented here.

This book is dedicated to the late Professor Taku Komai, who was the pioneer evolutionary geneticist in Japan and to whom I am deeply indebted, particularly for his help and encouragement when I first started as a mathematical geneticist. I should like to think that, were he still alive, he would greet this book with approval and satisfaction.

National Institute of Genetics, Mishima

Introduction

The neutral theory asserts that the great majority of evolutionary changes at the molecular level, as revealed by comparative studies of protein and DNA sequences, are caused not by Darwinian selection but by random drift of selectively neutral or nearly neutral mutants. The theory does not deny the role of natural selection in determining the course of adaptive evolution, but it assumes that only a minute fraction of DNA changes in evolution are adaptive in nature, while the great majority of phenotypically silent molecular substitutions exert no significant influence on survival and reproduction and drift randomly through the species.

The neutral theory also asserts that most of the intraspecific variability at the molecular level, such as is manifested by protein polymorphism, is essentially neutral, so that most polymorphic alleles are maintained in the species by mutational input and random extinction. In other words, the neutral theory regards protein and DNA polymorphisms as a transient phase of molecular evolution and rejects the notion that the majority of such polymorphisms are adaptive and maintained in the species by some form of balancing selection.

The word 'neutral' is not used in a strict, literal sense. The emphasis is not on neutrality *per se*, but on mutation and random drift as the main explanatory factors. The mutant genes that are important in molecular evolution and polymorphism are assumed to be nearly enough neutral for chance to play the major role. As the theory has developed, more attention has been given to selective molecular constraints, selection acting in indirect ways (as in tRNA abundance-codon usage relationship), and very weak negative selection acting on genes drifting to fixation. The theory does not, then, assume that selection plays no role; however, it does deny that any appreciable fraction of molecular change is due to positive selection or

that molecular polymorphisms are determined by balanced selective forces.

One possibility would be to rename the theory the 'mutation-random drift theory', but the term 'neutral theory' is already widely used and I think it better not to change horses in midstream. I want the reader to realize that 'neutral theory' is shorthand for 'the theory that at the molecular level evolutionary changes and polymorphisms are mainly due to mutations that are nearly enough neutral with respect to natural selection that their behavior and fate are mainly determined by mutation and random drift'. I also must emphasize that the theory does not deny the occurrence of deleterious mutations. On the contrary, selective constraints imposed by negative selection are a very important part of the neutralist explanation of some important features of molecular evolution, as I shall detail in chapter 7.

Classical evolution theory has demonstrated beyond any doubt that the basic mechanism for adaptive evolution is natural selection acting on variations produced by changes in chromosomes and genes. Such considerations as population size and structure, availability of ecological opportunities, change of environment, life-cycle 'strategies', interaction with other species, and in some situations kin or possibly group selection play a large role in our understanding of the process. This field has been greatly enriched by new molecular understanding, which has revealed totally new and unexpected additional possibilities and constraints. One now has to consider the way the primary sequence of amino acids is converted into a three-dimensional structure, which amino acids are hydrophilic and which hydrophobic, which are on the surface and which are buried inside, which are associated with essential functions and which are not, and other detailed aspects of our increasingly deep knowledge of proteins. There are DNA constraints caused by secondary folding and pairing of RNAs, by the matching of codon usage with tRNA abundance and vice versa, and by the processing of RNA to remove intervening sequences. There is the whole new question of the evolution of the genetic code, made more pertinent by the discovery that mitochondria have a somewhat different and variable code dictionary. There is new understanding of chromosome evolution, made possible by new staining and labeling techniques. There is the likelihood of discovering many more pseudogenes – seemingly functionless analogs of known genes, which have been revealed by cloning and rapid DNA sequencing methods. We have to consider 'selfish DNA', transposons, and other possible mechanisms by which the genome increases and decreases in size and the role of 'junk DNA'. So the study of adaptive evolution remains the exciting subject that

it has been since Darwin, but greatly enriched by new and much deeper levels of understanding brought about by the molecular revolution.

But, in my view, the most surprising possibility arising from molecular studies is that the great preponderance of nucleotide changes over time and of nucleotide variability in populations at any one time are selectively neutral or nearly neutral, so that increases and decreases in the mutant frequencies are due mainly to chance. It is remarkable, I think, that their behavior is calculable from the theory of stochastic processes, a theory which until recently has been regarded as too academic to have actual biological applications. Furthermore, as I shall explain in chapter 7 and elsewhere in this book, many of the newly discovered phenomena enumerated above lend support to the neutral theory, a most dramatic example being the rapid evolutionary change of pseudogenes.

Superimposed on this random change of nucleotide frequencies are the directional, adaptive changes brought about by natural selection in an ever-changing environment and the systematic process of removal of deleterious mutations. These are subjects of traditional, and continuing interest to evolutionists.

The whole picture, however, must include *all* the changes that occur in the DNA, however slight in their phenotypic effects. Just as the mathematical treatment of random processes in physics has contributed to a deeper understanding, the mathematical treatment of random genetic processes can illuminate some previously dark secrets of evolution. Furthermore, we should not overlook the possibility that some of the 'neutral' alleles may become advantageous under an appropriate environmental condition or a different genetic background; thus, neutral mutants have a latent potential for selection. This means that polymorphic molecular mutants, even if selectively neutral under prevailing conditions of a species, can be the raw material for future adaptive evolution. To regard random fixation of neutral mutants as 'evolutionary noise' is inappropriate and misleading. Also, there is the possibility (as I shall show in section 6.7) that extensive neutral evolution occurs under stabilizing phenotypic selection, if a large number of segregating loci (or sites) are involved in a quantitative character. In this case, genes that are substituted by random drift and those that are responsible for phenotypic variability belong to the same class.

An introductory caution is needed. *Webster's Third International Dictionary* defines evolution as 'the process by which through a series of changes or steps any living organism or group of organisms has acquired the morphological and physiological characters which distinguish it'. This definition was entirely appropriate for Darwin's time and for the first half of

the twentieth century. (Note that Darwin used the term 'descent with modification' and he stated that species change by 'preservation and accumulation of successive slight favourable variations'.) But, with all the changes that molecular biology has revealed – none of which is visible to the naked eye – a much broader definition is needed. In this book, as in earlier discussions of the neutral theory, I include in the word evolution *all* changes, large and small, visible and invisible, adaptive and nonadaptive. In some cases, evolution may occur even by random fixation of very slightly deleterious mutants, whose selection coefficients are comparable to or only slightly larger than the mutation rates.

The first two chapters of this book will be devoted to the historical development of theories on the mechanism of evolution. This will set the neutral theory of molecular evolution in proper historical perspective. In the latter half of chapter 2, I shall document how the neutral theory came to be proposed. Chapter 3 will serve as a systematic introduction to the theory. Readers who want to learn quickly about the theory are invited to read this chapter, skipping over the first two chapters. However, a quick reading of the latter half of chapter 2 will also help to deepen understanding of the theory. In chapter 4, the main features of evolution at the phenotypic level, as exemplified by the history of vertebrates, will be presented, and these will be contrasted with the characteristics of molecular evolution, with special reference to evolutionary rates. A more general discussion of the main features of molecular evolution will follow in chapter 5.

Now, in order to make proper appraisal of the neutral theory, a correct understanding of natural selection is essential. So, the definition, types and action of natural selection will be reviewed in chapter 6. This chapter will serve as an introduction to classical population genetics theory. Most material in this chapter, except that in the last section, is concerned with deterministic theory. I have encountered with dismay a number of occasions in which natural selection is invoked as a panacea to explain virtually any aspect of evolution and variation. It is easy to invent a selectionist explanation for almost any specific observation; proving it is another story. Such facile explanatory excesses can be avoided by being more quantitative. This is one reason why chapter 6 was written. In chapter 7, I shall show how the neutralist paradigm can cope with diverse observational facts coming from molecular evolutionary studies. In fact, data from the new molecular revolution have recently added supporting evidence for the neutral theory, and this will be discussed extensively in this chapter. I am convinced that no other existing theory can give a better and more consistent explanation of these facts. This chapter together with the

next chapter constitute the core of this book. In chapter 8, I shall present a rather extensive account of the stochastic theory of population genetics and various models that have been devised to treat evolution and variation at the molecular level. For most readers, this chapter, particularly from section 8.3 onward, may be difficult reading. This is inevitable because of the intricate nature of the subject, although I have done my best to present the material as clearly as possible. Readers who are mainly interested in the biological aspects (rather than mathematical properties of various models) may proceed to the next chapter, after finishing the first two sections (sections 8.1 and 8.2). Chapter 9 treats the problem: what is the mechanism by which the genetic variability at the molecular level is maintained? This problem has been regarded by some as the most important problem currently facing population genetics. Here again, I shall show that the neutral theory has not only withstood various tests but has proved useful in explaining levels of heterozygosity under various circumstances (such as haploidy vs. diploidy, monomeric vs. dimeric or tetrameric enzyme loci, etc.). Recently, the neutral theory has gained strength in treating the problem of intraspecific variability at the molecular level, and it is quite possible that this problem has essentially been solved by the theory.

In the last chapter (chapter 10), I shall summarize the whole work and give a few concluding remarks.

I am convinced that the neutral theory, as explained in this book, is correct in its essential details, although it will doubtless be refined in the future as it has been in the past. The final arbiter is time; but new data are appearing so rapidly that the time may be quite short.

1

From Lamarck to population genetics

At the beginning of the nineteenth century it was almost universally believed that all living beings on earth were immutable divine creations. In bold contrast, the French naturalist, Jean Baptiste Lamarck argued that life could change continuously on a grand scale. Moreover, he provided for the first time an appealing, easily understood process by which such changes could occur. Lamarck published his book entitled *Philosophie Zoologique* in 1809 propounding this idea when he was already sixty-five years old. Apparently the book attracted little attention when it appeared. It is said that, in his old age, he was neglected by his contemporaries, became blind, and died miserably. His fame came only long after his death when evolution became a heated subject following the publication of Darwin's *The Origin of Species*. Some of Darwin's opponents drew from Lamarck a theory called neo-Lamarckism, emphasizing the direct effect of environment as the prime factor for evolution. (For an authoritative account of the nineteenth-century approaches to evolution, readers may refer to Simpson, 1964.)

Here our main interest in Lamarck is that he proposed, probably for the first time in biology, a general theory to explain how evolution occurred. As is well known, he assumed that the effect of use and disuse, which in animals is induced by their living conditions, is inherited by offspring and this causes a perfecting tendency in evolution. For example, the ancestors of the giraffe tried to reach leaves on high twigs by continuously stretching their legs and necks. In each generation, as this effort was continued, it caused these parts to be a little longer, and this tendency was passed on to their offspring. Through the accumulation of small increments of elongation, the present giraffe finally evolved. Lamarck's theory rests on the assumption that acquired characters caused by use or disuse are inherited.

It is now clear that inheritance of such acquired characters does not

occur, and therefore the Lamarckian theory of evolution is wrong. As remarked by Simpson (1964) it is a pity that such a charming theory is not true.

It was to August Weismann's credit that he openly challenged the Lamarckian hypothesis. Through a series of writings he showed that the hypothesis is both unnecessary and improbable and that the supposed evidence for it is weak. (For the role of Weismann in the history of genetics, readers are invited to consult Sturtevant, 1965.) It is a well-known story that he cut off the tails of mice for twenty-two successive generations but found no decrease in the tail length at the end of that time. The vigorous development of Mendelian genetics in this century, culminating in modern molecular genetics, has shown beyond doubt that there is no evidence for inheritance of acquired characters. Yet there are evolutionists even now who cling to this hypothesis, claiming that the modern theory of evolution based on Mendelian genetics is unsatisfactory. Being almost irresistible it is regularly revived and the discovery of reverse transcriptase has invited a new round of Lamarckian speculations (e.g. Steele, 1979).

Weismann led his school of 'neo-Darwinians' in the fight against neo-Lamarckians; his theoretical writings caused one of the most heated controversies in the early post-Darwinian period, making evolutionary studies in the later part of the nineteenth century highly turbulent. Weismann was also an avid selectionist, advocating the selection theory more forcefully than Darwin; it is said that he 'out-Darwined Darwin' (Simpson, 1949). In fact, Weismann rejected all of Darwin's theory of evolution except natural selection.

Scientific studies of evolution really started with Charles Darwin. He published his book *The Origin of Species* when he was fifty years old (Darwin, 1859), half a century after Lamarck's *Philosophie Zoologique*. With his masterly writing and wide ranging examples, Darwin not only persuaded the world that evolution has actually occurred, but also he showed through his theory of natural selection why adaptive evolution is an inevitable process. *The Origin of Species* has had immeasurable influence not only on biology but also on human thought in general. We cherish Darwin for we owe to him our enlightened view of the nature of living things, including ourselves; our civilization would be pitifully immature without the intellectual revolution led by Darwin, even if we were equally well off economically without it. H.J. Muller (1960), in celebrating the hundredth anniversary of the publication of *The Origin of Species*, remarked that it can justly be considered as the greatest book ever written by one person.

Extrapolating from artificial selection which proved to be most efficient in producing domestic races of plants and animals useful to man, Darwin reasoned that the principle of selection applies equally in nature. Because more individuals of each species are born than can possibly survive, a struggle for existence follows, and any variation, however slight, if in any manner profitable to the individual, will have a better chance of surviving. Then, under 'the strong principle of inheritance', such a variation tends to be propagated. He writes:

> Slow though the process of selection may be, if feeble man can do much by his powers of artificial selection, I can see no limit to the amount of change, to the beauty and infinite complexity of the coadaptations between all organic beings, one with another and with their physical conditions of life, which may be effected in the long course of time by nature's power of selection.

Darwin emphasized the importance of accumulation of small beneficial variations, thereby causing a gradual and continuous process of adaptive evolution.

When Darwin formulated his theory, the mechanism of inheritance and the nature of heritable variations were unknown, and this prevented him from being fully confident of the role of natural selection. In fact, through successive editions of *The Origin*, he gradually weakened his claim that natural selection is the main cause of evolution. He conceded that inheritance of acquired characters also plays an important role in evolution. It is all too easy to forget that a storm of opposition and criticism once raged against the Darwinian view, for it is now so well established as to be regarded as almost sacrosanct.

With the rise of Mendelian genetics in this century the way was open to elucidate the mechanism of inheritance and the nature of heritable variations, which Darwin vainly struggled to understand. The dawn of the Mendelian era, however, was stormy. Soon, a bitter conflict arose between the biometricians championed by Karl Pearson and W.F.R. Weldon, and the Mendelians led by William Bateson (see Provine, 1971). Actually, the conflict between these two groups had started before the rediscovery of Mendel's law in 1900. Weldon, who was a biologist, came to the belief, stimulated by Francis Galton, that evolution can best be studied by the statistical method. He made many measurements on animal and plant characters with the aim of estimating the evolutionary rate and intensity of natural selection. Through personal contact with Weldon, Karl Pearson, who was an outstanding applied mathematician, became interested in the

problems of evolution. Although the theory of heredity he formulated was wrong, the statistical methods which Pearson developed, such as the χ^2 method, turned out to be of enormous value for later studies of evolution and variation, as pointed out by Haldane (1957a). Both Weldon and Pearson believed, following Darwin, that evolution occurs gradually by natural selection operating on small differences.

On the other hand, William Bateson, through his studies of variation in plants and animals, came to the definite conclusion that evolution could not occur by natural selection acting on continuous variations, contrary to Darwin's view. Bateson emphasized instead the importance of discontinuous variation.

With the rediscovery of Mendel's laws, the conflict between Mendelians and biometricians was exacerbated. While Bateson was impressed by the importance of Mendel's laws, both Weldon and Pearson vigorously attacked Mendelism. It is said that from the strenuous effort to disprove Mendelism by searching for exceptions in the huge volumes of stud books of race horses, Weldon became exhausted, contracted pneumonia, and died in the prime of his life (see Pearson, 1906). As one who has been honored by the Weldon Memorial Prize, I feel a particular interest in Weldon's life, and would like to add that he made an important discovery concerning the action of natural selection. He studied (Weldon, 1901) snail shells and measured the numbers of turns of the spiral in a given length of axis. By comparing this number in young shells and in the corresponding part of the adult shells, he found that young snails that had too many or too few turns showed a higher mortality than those whose shells were near the average. This is one of the first reports of 'centripetal selection', as pointed out by Haldane (1959).

After Weldon's death, biometricians were in retreat and the victory of Mendelians was soon evident, being supported by overwhelming facts. Many Mendelians at that time doubted that natural selection acting on small continuous variations could be effective in producing evolutionary change as envisaged by Darwin. Rather, they adopted the mutation theory of Hugo de Vries, claiming that a new species arises by mutational leaps rather than by gradual natural selection. The mutation theory, proposed at the beginning of this century, became very popular among biologists and found many adherents at that time.

It is now known that the 'mutations' observed by de Vries in the evening primrose *Oenothera lamarckiana* were probably due to this plant being heterozygous for special multiple chromosome rearrangements. As remarked by Sturtevant (1965), it is ironic that few of those mutations would

now be called mutations. Nevertheless, de Vries' theory, by calling wide attention among scientists to the possibility of mutation as the real cause of genetic variation, opened the way to its eventual confirmation and elucidation by H.J. Muller, to whose great contribution to our understanding of the mechanism of evolution we have to return later.

The first decade of this century witnessed active experiments to settle the issue whether natural selection on continuous characters is effective (as assumed by Darwin) or not. Among such experiments, those made by Wilhelm Johannsen are probably best known. He proposed the pure line theory, showing that selection is ineffective within a pure line. The exciting but confusing atmosphere of this period is well described by Provine (1971).

Gradually, however, it became understood that Mendelism and Darwinism are mutually compatible. Such a change of atmosphere was due to the vigorous development of genetics, of which *Drosophila* genetics was particularly important, revealing that mutational changes can be very small. Eventually, population genetics was developed through the effort to bring about a synthesis of Darwinism and Mendelism by the method of biometry.

The contributions of Hardy (1908) and Weinberg (1908) serve as a convenient starting point to discuss the history of population genetics. They showed that under random mating and Mendelian inheritance, genotypic frequencies at an autosomal locus remain unchanged from generation to generation. Furthermore, if a pair of alleles A and a occur with relative frequencies of p and q (where $p+q=1$), then the equilibrium genotypic frequencies are

$$p^2AA : 2pqAa : q^2\text{aa}.$$

The finding that genotypic frequencies remain unchanged was significant, for some biometricians at that time mistakenly criticized Mendel's laws by saying that if they were true, a dominant trait (such as brachydactyly in man) would increase in frequency until it reaches 75%. Hardy's paper helped to dispel this misconception. Now that Mendelian inheritance is well established there is no need for us to worry about such a matter.

A part of the findings by Hardy and Weinberg which we can still appreciate is the rule that genotypic (or zygotic) frequencies in a random mating population can be computed by simply multiplying relevant gene frequencies. It is, however, no more than a useful rule. Their findings serve as a convenient starting point for the teaching of population genetics. However, I find it surprising that the Hardy–Weinberg principle is customarily presented with an exaggerated importance attached to its

significance. This tendency was probably started by Dobzhansky who presented 'the Hardy–Weinberg law' in the third edition (1951) of his influential book *Genetics and the Origin of Species* saying that this law is the foundation of population genetics and of modern evolution theory. In my opinion, the pedantic, and intellectually sterile way that this law is often presented is almost detrimental to students, so that some reform is needed in the teaching of elementary population genetics. In many textbooks, after solemnly announcing that this is the most important law in population genetics, it is introduced by saying that (i) with no disturbing influences (such as mutation, selection, migration, random sampling of gametes etc.), the genotypic frequencies remain unchanged under random mating, and, (ii) the genotypic frequencies are given by p^2AA, $2pqAa$ and q^2aa. Furthermore, the first point (i) is emphasized as the more essential part of the law, which, I think, is anachronistic.

The fact that gene frequencies do not change without disturbing factors is obvious without any computation as soon as we note that homologous genes on the chromosomes segregate regularly at meiosis and that genes are self-reproducing entities. We tend to forget that formation of genotypes (zygotes) is simply a grouping by two of homologous genes. Higher organisms including ourselves happen to have a well-developed diploid phase; probably no one would pay any attention to the Hardy–Weinberg law if we humans were haploid. The Hardy–Weinberg law would be a totally useless thing if its validity were restricted to the condition (i). On the other hand, as a rule of thumb, it is quite useful; genotypic frequencies can be predicted with very good approximation by simply multiplying relevant gene frequencies *even when gene frequencies are changing*, generation after generation, under natural selection, if we enumerate genes and genotypes immediately after fertilization.

During the two decades that followed the contributions of Hardy and Weinberg, the population consequences of Mendelian inheritance were worked out by R.A. Fisher, J.B.S. Haldane and Sewall Wright. By the early 1930s the mathematical theory of classical population genetics was essentially complete.

Among these three founders of theoretical population genetics, Fisher appears to have had the strongest influence on the formation of the orthodox view that the rate and the direction of evolution are almost exclusively determined by natural selection, with mutation, migration and random drift playing subsidiary roles.

Incidentally, the term neo-Darwinism has often been used to represent the orthodox evolutionary genetics because it is in the tradition of

Weismann, rejecting inheritance of acquired characters and attaching paramount importance to natural selection. In the United States, a more impressive name 'Synthetic Theory of Evolution' has been widely used, perhaps to emphasize that various factors are taken into consideration.

In the field of theoretical population genetics, Fisher achieved a number of breakthroughs. In 1918 he published a profound paper treating correlation between relatives by statistical methods based on the supposition of Mendelian inheritance (Fisher, 1918). Like many of his other papers, the subject was worked out so thoroughly with his impressive mathematical power that subsequent studies by a number of mathematical geneticists have been able to add relatively little. In fact, after more than half a century, it still remains one of the most profound papers on biometrical genetics. In 1922, Fisher published a paper with the rather odd title, 'On the dominance ratio'. Although the mathematical treatments in it were incomplete and even contained significant errors, the paper was full of originality. With this paper, Fisher initiated the use of stochastic methods in population genetics, in which random fluctuation of gene frequencies from generation to generation is incorporated. Fisher considered random sampling of gametes as the factor causing random fluctuation in gene frequency and, assuming no selection, he investigated its effect (which he called 'Hagedoorn effect') on the decrease of genetic variability of a species. Using a gene frequency transformed to stabilize the variance, he derived a partial differential equation that is commonly used to describe the conduction of heat. He then solved it and found that the variance decreased at the rate $1/(4N)$ per generation, where N is the number of breeding individuals. From this he concluded that the decrease of variability by this cause is so inconceivably slow that it is of little consequence in evolution. As compared with this, he considered that a most trifling amount of selection has an important effect. Unfortunately, Fisher's treatment contained a few errors. In particular, the correct rate of decay of variance was $1/(2N)$, as shown by Wright using an entirely different method. Later, Fisher (1930a) presented a more complete treatment of the problem in which the errors were amended and the results were greatly extended. But he did not find any reason to change his earlier conclusion on the role of random sampling of gametes; population sizes of most species are so enormous that random sampling is a totally unimportant factor in evolution. Fisher's 1922 paper also contains a discovery which made a profound influence on the later thinking of population and evolutionary geneticists. He showed that if selection favours the heterozygote, there is a condition of stable equilibrium, and both alleles will continue in the stock. In other words,

overdominant alleles are actively maintained in the population by natural selection.

Fisher's 1930 paper is an elaboration of the stochastic problems treated in his 1922 paper. He developed an ingenious functional equation method to treat the fate of rare mutant alleles in the population. Although the method is now familiar among mathematicians in the study of 'branching processes', the level of sophistication relative to that time is most impressive. His penetration into the problem and mastery of his mathematical analyses makes me often feel that Fisher's mathematical results cannot easily be surpassed. Our contemporary advantage mainly comes from the unexpected robustness and versatility of the diffusion equation method (now called 'diffusion models', see Kimura, 1964); without our current knowledge Fisher must have felt insecure in using these to study the behavior of rare mutant alleles. At any rate, he arrived at the conclusion that chance in the form of random gene frequency drift is unimportant in evolution. For me this appears to be almost an anticlimax, after seeing his most impressive mathematical treatments of the stochastic behavior of mutants.

His whole views on evolution at that time were summarized in his book *The Genetical Theory of Natural Selection* (Fisher 1930b). To bring out the synthesis of Darwinism, Mendelism and biometry, no other single book contributed more than this one. Since its publication, it became a bible for many biologists who tried to understand evolution in terms of natural selection. Like the Bible, it brought revelation and at the same time some biassed thinking. Fisher must be largely responsible for the panselectionism later prevalent in England.

One of his pet theories is his explanation of how dominance evolved ('theory of dominance'). Fisher claimed that the dominance of the wild type gene over mutant alleles is caused by long continued selection of modifiers which make mutant heterozygotes resemble the wild type homozygotes (see Fisher, 1930b). Fisher assumed that mutant heterozygotes were originally intermediate in fitness between two homozygotes (i.e. no dominance at the start). Since mutant heterozygotes are in general rare, the selection for such modifiers is extremely weak, of the order of mutation rate, but Fisher claimed that in the long course of evolution a very weak force will eventually be effective. He quantified the efficiency of selection involved by deriving a formula which, according to him, represents the rate of progress in the modification of the heterozygote, compared to the rate of progress which would be effected by selection of the same intensity, acting upon a population entirely composed of heterozygotes. His formula, using terminologies in the present book, is

$$2vw_{12}/[(1-w_{12})(2-w_{12})] \tag{1.1}$$

where v is the mutation rate and w_{12} is the fitness of heterozygotes relative to the wild homozygotes. Fisher's theory of dominance was criticized by Wright (see for example, Wright, 1934. For details, see chapter 15, vol. 3 (1977) of his treatise). Wright pointed out that such weak selection would be easily upset by other forces of selection that might be induced by the modifiers' effects on the much more abundant wild type homozygotes, and also it would be overwhelmed by mutation pressure and random drift. Haldane (1930) was also critical, arguing that it is much more likely that selection occurs among wild type alleles, favoring a more active and therefore more dominant wild type.

It now appears that Fisher's theory of dominance is either untenable or at best holds only for special cases. Recent studies of lethal and slightly deleterious mutants in *Drosophila* clearly show (see Mukai and Yamaguchi, 1974) that slightly deleterious mutants have in general a much higher degree of dominance ($0.3 \sim 0.4$) than lethal and semilethal ones (degree of dominance $0.01 \sim 0.02$). These studies concern only viability, but the same conclusion also holds for total fitness. Mutants with mild homozygous effects have a much greater relative effect on heterozygous fitness than do drastic mutants (Simmons and Crow, 1977; Crow, 1979). It is clear from equation (1.1) and was emphasized by Fisher that, as the heterozygote is closer to the normal homozygote in fitness, the rate of dominance modification increases rapidly. We should therefore expect that mildly deleterious mutants should be more nearly completely recessive than lethals. However, the opposite is true. A similar criticism was made by Charlesworth (1979) (see also Charlesworth and Charlesworth, 1979). As compared with Fisher's theory, alternative hypotheses due to Wright (see Wright, 1977) and especially to Muller (1933) are more plausible: the wild type allele usually produces a useful substance which recessive mutants lack the ability to produce or produce in lesser amount. Dominance can be understood in terms of the dose–response curve. The wild type genes are selected in the course of evolution for ability to produce useful substances with a sufficient margin of safety to withstand external and internal disturbances; dominance follows automatically.

Throughout his writings, Fisher consistently minimized the role played by random drift in evolution, and this must have discouraged young English geneticists from working on this topic. What turned out, however, is that in theoretical population genetics, the main progress in recent years

has been the treatment of gene frequency changes as stochastic processes (See Li, 1977a). The lack of interest shown by English geneticists in this field is probably due to Fisher's influence. The situation is rather different in Scotland where Alan Robertson of the Institute of Animal Genetics has done outstanding work on small populations, influenced by Wright.

Haldane's contributions to population genetics are not as original as Fisher's, but they are more diverse, more open minded, and often biologically more appropriate. Starting in 1924, he published a series of papers entitled 'A mathematical theory of natural and artificial selection'. With these papers he opened up the systematic mathematical attack on the process of change in gene frequencies by natural selection. The work is based on his conviction that a satisfactory theory of natural selection must be quantitative (as stated in his first paper, part I of the series; Haldane, 1924), and that only through quantitative investigation can the adequacy of the genetic theory of natural selection be tested. He considered various kinds of selection, such as zygotic, gametic, familial and X-linked. A typical situation considered is as follows. In a very large random mating population, the dominant types (*AA*, *Aa*) are favored, and, as compared with them, $1 - k$ recessive (*aa*) individual survive to breed, where k is called the coefficient of selection in favor of *A*. Then, Haldane derived a finite difference equation which transforms the gene ratio (i.e. ratio of the frequency of *A* and *a*) from one generation to next. The equation obtained was nonlinear, and although he could not obtain the exact solution, he worked out approximate solutions sufficient for practical purposes. He then constructed a table from which one can readily obtain the relation between selection intensity and the time change of genotypic frequencies. For example, if $k = 0.001$ it takes about 16 483 generations for *A* to increase from 1% to 99%. If k is ten times larger, the number of generations required becomes one-tenth as long. Haldane showed that in general the number of generations required for a given change is inversely proportional to the intensity of selection. He applied this theory to a case of industrial melanism in moths that occurred in Manchester and found that the selection coefficient involved was at least 0.332 and probably as large as 0.5. Some thirty years later, thanks to the work of Kettlewell (1955) and others, it was shown that selection intensities of this magnitude commonly occur in these moths, especially from predation by birds.

Such treatment of the change of gene frequencies, as initiated by Haldane, can be called deterministic, since no random elements, particularly, random sampling of gametes, are taken into account. Even if more sophisticated

methods have been developed since, the deterministic treatment is still widely used, and usually is found satisfactory when the population is large. Furthermore, because of its simplicity, this approach is still the most useful and often the only manageable one for many problems. In the subsequent papers of the series, he investigated the effects of various factors on the change of gene frequencies, such as partial inbreeding, partial assortative mating, incomplete dominance (including over-dominance) for autosomal as well as sex-linked loci, multifactorial inheritance, linkage (but without selection), polyploidy, generation overlap, and isolation. These papers contain results which are still useful but often overlooked. Among the series, two papers, parts V and parts VIII, deserve particular mention. In part V, Haldane (1927) took up a stochastic problem and investigated the probability of fixation of mutant genes, using the generating function method suggested by Fisher (1922). He showed for the first time that a dominant mutant gene having a small selective advantage k in a large random mating population has a probability of about $2k$ of ultimately becoming established in the population. Another outstanding paper is part VIII entitled 'Metastable populations' (Haldane, 1931) in which he investigated the situation where mutant genes are disadvantageous singly but become advantageous in combination. Assuming no linkage complications, he treated the two-locus case with extreme elegance, determining trajectories of points representing the genetic composition of a population in a two-dimensional co-ordinate system. He discussed transformation of the *AABB* population into the *aabb* population, and wrote that a species which is liable to transformation of this kind may be called 'metastable'. He suggested that the process of species formation may be a rupture of the metastable equilibrium, and that such rupture will be especially likely in small isolated communities. It is interesting to note the similarity of this work to the theory of Wright, which was propounded independently and which describes evolution as a trial and error process in terms of a multidimensional adaptive surface (Wright's 'shifting balance theory' to be discussed subsequently).

Haldane's work up to this stage is summarized in his book *The Causes of Evolution* published in 1932. This book is much easier to read than Fisher's 1930 book, and it must have contributed greatly to making the biological world realize that Mendelism and Darwinism are mutually compatible. Haldane took many opportunities to write on the use of mathematics in biology, particularly in relation to the genetical theory of evolution. He wrote, in one of his papers (Haldane, 1938):

> At present one may say that the mathematical theory of evolution is in a somewhat unfortunate position, too mathematical to interest most biologists, and not sufficiently mathematical to interest most mathematicians. Nevertheless, it is reasonable to suppose that in the next half century it will be developed into a respectable branch of Applied Mathematics.

His prediction is now fully realized, and the use of mathematical models is now becoming more and more popular in biology. His confidence in the ultimate value of making evolutionary studies more quantitative is well illustrated in his spirited article entitled 'A defense of beanbag genetics' (Haldane, 1964). He wrote this shortly before his death, in rebuttal to Ernst Mayr's (1963) criticism that population genetic theories regard a population as a bag full of colored beans and that such a simplified way of thinking is misleading.

As compared with Fisher and Haldane, Wright's view of evolution is distinguished by his emphasis on the importance of random genetic drift and the prevalence of non-additive gene interaction systems. Following his earlier work on systems of mating (Wright, 1921) in which he developed a new and powerful treatment of inbreeding and assortative mating using his method of path coefficients, he published in 1931 a paper entitled 'Evolution in Mendelian populations'. Together with Fisher's (1930b) book *Genetical Theory of Natural Selection* and Haldane's (1932) *The Causes of Evolution*, this paper represents a culmination of classical population genetics, whereby the synthesis of Darwinism and Mendelism was fully achieved.

Since 1932, Wright developed a theory of evolution that he later called 'the shifting balance theory' (Wright, 1932, 1970). Examining various modes of transformation under mutation pressure (as may occur in the degeneration of useless organs), mass selection, pure sampling drift and migration, Wright came to the conclusion that a large, subdivided population structure is most favorable for rapid evolutionary progress through the shifting balance process. According to him (Wright, 1977), three phases are involved in the process: (i) Random drift – extensive gene frequency drift occurs in each deme (local population) due to accidents of sampling or to fluctuations in the coefficients measuring various evolutionary pressures. (ii) Mass selection – by chance a deme may cross one of the innumerable 'two-factor saddles' in the surface of fitness values leading to rapid genetic change in this deme by mass selection. (iii) Interdeme selection – a deme which came to a new fitness peak superior to surrounding demes will expand through interdemic selection. This process may spread through concentric circles originating from different centres, and two such circles may overlap and give rise to a new, still superior centre for further progress. Thus, a virtually infinite field of interaction systems may be explored with only a small number of novel mutations.

Wright's theory has been widely misunderstood as proposing random drift as a significant alternative to natural selection in the phenotypic evolution of species. That this is not what he originally intended has been emphasized by Wright himself (see p. 453, vol. 3 (1977) of Wright's treatise).

For Wright, the shifting balance process involving joint action of all evolutionary factors is the major basis for evolution. During the decade or two that followed his 1931 paper, Wright's theory of evolution attracted much attention. Its popularity was greatly enhanced by its detailed exposition in Dobzhansky's book *Genetics and the Origin of Species* (1937) which was widely read by biologists. I still remember my fascination when I first read about Wright's work in that book when I was an undergraduate student.

Wright's theory was criticized by Fisher and his school in England, and a controversy arose (Fisher and Ford, 1950; Wright, 1951a; Fisher, 1953; Ford, 1965) which continues to this day. The issue, obviously, is whether random drift has any important role to play in evolution. Fisher was confident that the number of individuals making up a species is generally so large that the chance effect due to random sampling of gametes is negligible. He also thought that for most mutant alleles the product of the population size and the selection coefficient is unlikely to be restricted to the neighborhood of zero in the course of evolution so that selectively neutral mutants must be extremely rare. On the other hand, as mentioned above, for Wright random drift in local populations is important in progressive evolution. He considered that a species consisting of one large panmictic population will soon reach an adaptive plateau by mass selection and will end up in evolutionary stagnation. Fisher (1953) thought that such an event is unlikely, for the population would contain myriads of genotypes, and if any of the extreme (homozygous) forms among them has, in some particular locality, a selective advantage, 'there is nothing to prevent a shift in the gene ratios yielding higher adaptation'.

Wright's theory has also been criticized by Mayr (1963 see p. 520) who considers that in all widespread, successful species there may be sufficient migration between demes to make gene frequencies of all local populations very similar. In my opinion, no compelling evidence for the shifting balance theory is available, although it is an attractive hypothesis. I also think it

somewhat surprising that, despite his passion for the shifting balance process, Wright has not made much quantitative study of the process, nor made serious attempts to test it against critical observations.

Although Wright's theory of evolution is controversial, he has made many fundamental contributions, since his 1931 paper, to our understanding of the stochastic behavior of mutant alleles in finite populations. He has worked out gene frequency distributions under irreversible as well as reversible mutation (Wright, 1937, 1938a, 1942, 1945; for review see vol. 2, 1969 of Wright's treatise). Ensuing development of the subject in terms of diffusion models (see Kimura, 1964) owes much to his work.

2

Overdevelopment of the synthetic theory and the proposal of the neutral theory

2.1 Formation of the synthetic theory as the orthodox view

In the previous chapter, a history of the theories on the mechanism of evolution was presented – a history leading to the establishment of classical population genetics by Fisher, Haldane and Wright early in the 1930s. In discussing the synthesis of Darwinism with Mendelian genetics, there is one more name that we should not forget. This is H.J. Muller who by the early 1920s had elucidated the fundamental nature of gene mutation and its bearing on evolution by natural selection.

His dramatic demonstration of the mutagenic action of X-rays (Muller, 1927) for which he was later awarded the Nobel Prize is well known. But less well known are his fundamental contributions to our understanding of the mechanism of evolution. He propounded forcefully the view that the basis of adaptive evolution by natural selection lies in the remarkable property of the gene that not only is each gene self-reproducing, but also mutated forms of a gene are again self-reproducing (Muller, 1922, see also p. 175 of Muller, 1962). Natural selection, then, can be defined as the differential multiplication of mutant types. He went so far as to claim (Muller, 1929, see p. 188 of Muller, 1962) that the gene is the very basis of life. This was at a time when many biologists still regarded the gene as a hypothetical entity which was assumed to explain the results of crossing. This 'gene primacy' view was later much extended by him (Muller, 1966). According to Muller, '*the criterion for life is the potentiality of evolution by* Darwinian *natural selection*'. He also established the principle that gene mutations, being blind changes, would usually be harmful, and that the more drastic ones would tend to be the more harmful. Nowadays, we seldom appreciate Muller's insight, for these views are shared by us as common sense. Another principle to which he called attention, as a corollary of the above principle, is that

mutant genes tend to have less dominance than their 'normal' progenitors. He applied these principles not only to evolution but also to the problems of human welfare; his paper entitled 'Our load of mutations' (Muller, 1950) is a classic in human population genetics. One remarkable foresight of Muller was that most mutant genes have a certain degree of dominance and that it is probably greater in the case of the less detrimental mutant genes than the more detrimental ones. Since then this has been amply confirmed by the work of Crow, Mukai and their collaborators (for review, see Simmons and Crow, 1977; Crow, 1979).

In addition to bringing in these conceptual revolutions, Muller created elegant, special purpose stocks, including the famous '*ClB*' chromosome, which made *Drosophila* the unrivalled experimental organism for genetical studies, to be surpassed only much later by *E. coli*. Studies of lethal and detrimental mutant genes in natural populations of *Drosophila* would certainly have been impossible without Muller's technical breakthrough. It is clear that Muller, more than anybody else, contributed to the establishment of the very foundation for studies of evolution by genetic methods.

Upon the foundation built by Fisher, Haldane and Wright, as well as by Muller, various studies, including investigation of natural populations by Th. Dobzhansky and his school, paleontological studies by G.G. Simpson, 'ecological genetics' by E.B. Ford and his followers, and the speciation theories of Ernst Mayr were added one by one until the synthetic theory grew into an enormous and impressive edifice.

Dobzhansky's main contribution to the science of population genetics, other than his influential writings, is his finding with A.H. Sturtevant that chromosome polymorphisms involving inversion (i.e. reversed gene order in a chromosome region) are abundant in some species of *Drosophila* (see Dobzhansky, 1951, 3rd edition, for review). He carried out extensive and successful investigation of inversion polymorphisms particularly in natural populations of *D. pseudoobscura*. In addition to observations in natural populations, he did cage experiments to study the frequency changes of chromosomes having different inversions, in which Wright collaborated to estimate fitnesses of various genotypes (Wright and Dobzhansky, 1946). The statistical analyses showed that the inversion heterozygotes have higher fitness than inversion homozygotes. The data were also analysed under the hypothesis of frequency-dependent selection, which was equally consistent with the data. Dobzhansky and his associates then attempted to make a more direct test of heterosis of polymorphic inversions, for which they obtained supporting evidence. Although the genic contents that differentiated these inversions were unknown, such prevalence of inversion

polymorphisms and experimental evidence suggesting marked heterosis of these inversions led Dobzhansky to hold the view that overdominance or heterozygote advantage at individual gene loci is prevalent in natural populations.

The enthusiasm toward the omnipresence of heterozygote advantage was spurred by Lerner's (1954) proposal of 'genetic homeostasis'. According to him, Mendelian populations (i.e. reproductive communities) are possessed of self-equilibrating properties tending to retain a genetic composition that produces a maximum mean fitness. He claims that the most likely mechanism for this is heterozygote advantage, and for normal development of the individual, an obligate level of heterozygosity is needed. He went so far as to suggest (Lerner, 1954, p. 75) that the pre-Mendelian notion of 'inbreeding *per se*' being harmful is partly justified. In other words 'not only gene contents but homozygosity as such must be considered to play a rôle in inbreeding degeneration'. Lerner questioned Muller's (1950) concept of genetic load on the ground that 'balance in populations based on the selective advantages of heterozygotes' is a phenomenon extending beyond isolated cases of polymorphism. In addition to overdominance, Lerner emphasized the importance of epistatic interaction in fitness, being influenced by Wright's concept of evolution as an irregularly shifting state of balance.

This line of thought attaching paramount importance to heterozygosity as the 'adaptive norm' and epistasis as the cause of 'coadapted gene complexes' was further elaborated by Dobzhansky (1955) who presented the leading paper at the twentieth Cold Spring Harbor Symposium dedicated to the topic of population genetics. He set up two contrasting working hypotheses called the 'classical' and the 'balance' hypotheses. According to his formulation, the classical hypothesis assumes that evolutionary changes consist of gradual substitution and eventual fixation of the more favorable allele in place of the less favorable one. Under this hypothesis, he asserts that most individuals in a population should be homozygous for most genes, and heterozygous loci will be a minority. The chief sources of heterozygosity, rare as it is, are four under the classical hypothesis: (1) deleterious mutants which are eliminated by natural selection in a certain number of generations; (2) adaptively neutral mutants; (3) adaptive polymorphism maintained by the diversity of the environments; (4) the rare good mutants in the process of spreading through the population. On the other hand, the balance hypothesis assumes that the adaptive norm is an array of genotypes heterozygous for a number of alleles. Homozygotes for these alleles occur in normal outbred

populations only in a minority of individuals, and they are inferior to heterozygotes in fitness. Then, there will be a selection pressure favoring the development of series of multiple alleles at many loci. Dobzhansky's description of the balance hypothesis is very similar to Lerner's claim of genetic homeostasis. It is abundantly clear from his writings that Dobzhansky placed himself as the chief protagonist of the balance hypothesis and put Muller in the opposing camp of classicists.

To support his case, Dobzhansky (1955) presented results of various experiments, among which the most remarkable was the production through recombination of what he called synthetic lethals. In one experiment, he reported that 10 second chromosomes were taken from a population of *D. pseudoobscura*. These chromosomes were normal to subvital when homozygous. Then, females heterozygous for all possible combinations of these 10 chromosomes were bred to produce offspring, and 450 chromosomes from them were tested for homozygous viability. It turned out that among these 450 chromosomes, 19 were lethal and 57 were semilethal in double dose. It is evident, if the experiment is valid, that a mere 10 chromosomes can give rise to a great abundance of genetic variability. Later investigations by other geneticists, however, could not confirm the existence of many synthetic lethals (Hildreth, 1956; Spiess and Allen, 1961; Temin *et al.*, 1969). Dobzhansky also reported that his group found that some lethals were heterotic, contrary to Muller's claim. No supporting evidence for it, however, was obtained in the large-scale study of Mukai and Yamaguchi (1974). They also showed that contrary to Dobzhansky's claim, 'super-vital' chromosomes do not occur, at least in *Drosophila melanogaster*. Furthermore, they showed that lack of correlation between the fitness of a heterozygote and the sum of fitnesses of the constituent chromosomes when homozygous, as claimed by Dobzhansky, is due to mixing of lethals and mildly detrimentals which have very different degrees of dominance (Mukai and Yamazaki, 1964; Wills, 1966). Also lethal frequencies in natural populations are too low to permit any appreciable overdominants, or even complete recessives, as first noted by Sturtevant (Crow and Temin, 1964; Crow, 1964).

Despite its shaky evidence, this paper of Dobzhansky had a tremendous influence on subsequent opinions among population geneticists, particularly in the United States.

In the latter half of 1960s when abundant enzyme polymorphisms started to be uncovered in many organisms by electrophoretic methods, some authors wrote as if Dobzhansky predicted the existence of such extensive polymorphisms by his balance hypothesis. The truth is, however, that most

of the experimental evidence on which Dobzhansky thought to support his balance hypothesis turned out to be invalid (see for example, Mukai and Yamaguchi, 1974).

Muller (1958) strongly opposed the claim that overdominance plays a predominant role in maintaining genetic variability, let alone the claim, as made by Lerner (1954), that heterozygosity *per se* tends to be beneficial. He was particularly afraid that such a view might be used to minimize the harmful genetic effects of radiation. Yet it is wrong to assert that Muller held the view that species are genetically homogeneous. In a paper entitled 'Redintegration on the symposium on genetics, paleontology, and evolution', Muller (1949) said:

> Moreover, it must be remembered that in most natural crossbreeding populations there has occurred a great accumulation of mutations that have arisen through the course of hundreds of preceding generations and that have not yet become established but with regard to which the population is still heterogeneous. Thus there may be a very considerable amount of hereditary variability of that character in the population, allowing its plasticity in response to selection in virtually any direction, and this is again despite the fact that mutation in any individual gene is so exceedingly rare.

The thesis that balancing selection rather than recurrent mutation is the main cause of genetic variability within species is also implicit in the concept of genetic polymorphism first proposed by E.B. Ford in 1940 (quoted from Ford 1965). He defines genetic polymorphism as the occurrence together in the same habitat of two or more discontinuous forms of a species, in such proportions that the rarest of them cannot be maintained simply by recurrent mutation. In my opinion, this was a peculiar definition, for it tacitly presupposes the mechanism involved; mutation-selection balance was automatically excluded from it. This definition has since been widely quoted in textbooks of genetics. Ford claims that the genetic polymorphism is most frequently maintained by heterozygous advantage (see Ford, 1965, p. 63). Actually, Ford's argument for the prevalence of overdominance is very simple, as clearly stated in another of his books *Ecological Genetics* (Ford, 1964). He says that 'the heterozygous advantage tends to be *evolved* in polymorphisms' and that the heterozygote will have 'nothing but advantage, and be superior to the homozygotes which will have both advantage and disadvantage'. At the same time, a high intensity of selection involved was suggested, in contrast

to the traditional view that adaptive evolution consists of gradual accumulation of small mutational steps. Ford's opinion found many adherents, particularly in England.

It seems to me that they overlooked the fact that only those cases with very large selection coefficients could be detected in nature by the methods of ecological genetics. In fact, the difficulty of detecting a small selection intensity can be seen in the experimental work of Mukai and Yamaguchi (1974) who investigated the relative viabilities of lethal chromosomes in heterozygous condition in *Drosophila melanogaster*. They extracted 691 second chromosomes (274 lethal-carrying and 417 lethal-free chromosomes) from the Raleigh, N.C., population. Using these chromosomes, they made 688 crosses which produced 241 lethal-free heterozygotes ($+/+'$), 350 single-lethal heterozygotes ($+/l$) and 97 double-lethal heterozygotes (l/l'). In their experiment, the average number of flies counted per cross was roughly 1400, bringing the grand total for all the crosses to 968 820 flies (Mukai, personal communication). Thus, to establish the result that the recessive lethal mutants lower the viability of heterozygotes by about 1.2% on the average, they counted nearly one million flies under controlled conditions.

Along with the emphasis on overdominance and epistatic interaction in fitness, a new trend developed. One of the leaders in this development was Ernst Mayr. According to him (Mayr, 1955), classical population genetics, which describes in mathematical terms the frequencies of genes, is based on the oversimplification of assuming a fixed, absolute selective value at a single locus, and this is no longer tenable. Rather, we must regard the fitness of a gene as relative; it depends on such factors as other genes with which it combines, past evolutionary history of the population, degree of inbreeding, population size, and so on. He pointed out that such a new development was gathering momentum for some years, and is, so to speak, 'a theory of relativity in the field of population genetics'.

Similarly, Waddington (1957) in his book entitled *The Strategy of the Genes* criticized the work of Fisher, Wright and Haldane saying that the mathematical theory they developed lacks two things that one normally expects from a mathematical theory. First, it has not led to any noteworthy quantitative statements about evolution, since the formulae involve parameters of selective advantage, effective population size, mutation rate etc., which are not known accurately. Secondly, it has not revealed new relations and processes that can explain phenomena which were previously obscure. In this book, Waddington introduced a horde of neologisms such as metaselection in analogy with metaphysics, homeostatic and quasi-homeostatic selection, canalising selection in contrast to normalizing selection, persistent drift vs. intermittent drift, spurious and true fitness cross-section, creode etc., which for the most part are not useful for discussing evolutionary proeesses in natural populations.

In the year 1959, the hundredth anniversary of the publication of Darwin's *The Origin of Species* was celebrated throughout the world. That year's Cold Spring Harbor Symposium was held under the title 'Genetics and Twentieth Century Darwinism'. In the leading paper entitled 'Where are we?', Mayr (1959) again speaks of the newer population genetics as the genetic 'theory of relativity'. Referring to the work of Fisher, Wright and Haldane, he remarks: 'but what, precisely, has been the contribution of this mathematical school to the evolutionary theory, if I may be permitted to ask such a provocative question?' His own newer population genetics, however, was entirely verbal and lacked any quantitative treatment, quite in contrast to the theory of relativity in physics.

Despite these various attempts to glorify the synthetic theory of evolution, actually very little substantial progress was made at this time. On the whole, it was a time of stagnation, rhetorical arguments dominating over meticulous scientific reasoning.

By the early 1960s, consensus seems to have been reached that every biological character can be interpreted in the light of adaptive evolution by natural selection, and that almost no mutant genes are selectively neutral. For example, in his book *Animal Species and Evolution*, Mayr (1963, see chapter 8) claims that he considers it exceedingly unlikely that any gene will remain selectively neutral for any length of time. He also says that selective neutrality can be excluded almost automatically whenever polymorphism or character clines are found in natural populations. He even suggests that it would clarify evolutionary discussions if authors would refrain from invoking 'genetic drift' as a cause of evolution.

Also, Ford (1964), in his book *Ecological Genetics*, claims that not only are neutral genes very rare, but also they are unlikely to attain appreciable frequencies, because their neutrality will be upset by changes in the environment and in the genetic outfit of the organism. The rationale of Ford's claim was drawn from Fisher (1930b) who had always been an extreme selectionist. In fact, in one of his papers, he (Fisher 1936) states as follows. 'Evolution is progressive adaptation and consists of nothing else. The production of differences recognizable by systematists is a secondary by-product, produced incidentally in the process of becoming better adapted.'

If we call the view that invokes positive selection to explain even the most

minute biological differences as 'panselectionism', this was a very successful paradigm at that time (and is still held by many biologists even now). I can attest that the above represents the then prevailing tide of thought, for I myself was greatly influenced by it.

Of course, panselectionism is not new. The history of evolutionary writings, especially around the turn of the century, is full of far fetched, imaginative ideas for selective advantages of specific traits. For example, one author attributed the pink color of flamingos to protective coloration against the setting sun. But the panselectionism of the synthetic theory is much more sophisticated. In the 1950s all observations and experiments appeared to support it.

Another prevailing opinion at that time was that the rate and direction of evolution is essentially determined by positive selection, with mutation playing only a minor and subsidiary role. It was said that even if mutations were stopped completely, outbreeding species usually have enough genetic variability that can be generated by recombination so that evolution can proceed for very long time. This opinion appears to have originated from Fisher (see page 96 of Fisher, 1930b). Even Muller (1949) states that he defends the thesis that evolutionary rates are chiefly determined by selective factors rather than by mutation rates. I note, however, that Fisher and Muller were thinking of adaptive evolution.

Looking back, I think that it is a curious human nature that if a certain doctrine is constantly being spoken of favorably by the majority, endorsed by top authorities in their books and taught in classes, then a belief is gradually built up in one's mind, eventually becoming the guiding principle and the basis of value judgement. At any rate, this was the time when the panselectionist or 'neo-Darwinian' position was most secure in the history of biology: the heyday of the traditional 'synthetic theory' of evolution.

Led by the *Zeitgeist*, a great deal was said about how gene pools of the species are organized and how they change in evolution. However, these were inferences based on observations at the phenotypic level, and in reality, there was no way of actually knowing what is going on in evolution at the level of the internal structure of the gene. As mentioned already, much importance was claimed for epistatic and heterotic gene interactions in fitness. Such terms as *integrated coadapted complex*, *genetic revolution*, *cohesion of the gene pool*, as well as *genetic homeostasis* were introduced and accepted by some. But, in my opinion, they were more rhetorical than scientific.

Throughout this period, but largely independent of fashion, the mathematical theory of population genetics grew gradually to be quite sophisticated, a rather unusual occurrence in biology. Particularly noteworthy is the theoretical framework called diffusion models (Kimura, 1964) which makes use of partial differential equations known as diffusion equations in mathematics. It enables us to treat the behavior of mutant alleles by incorporating random changes due to random sampling of gametes in reproduction as well as deterministic changes due to mutation and selection. This powerful method is an outgrowth of the great work of Fisher and Wright. My own life as a scientist has been largely devoted to the development of diffusion models stimulated by the pioneering work of Wright. Although it involves approximation, the diffusion equation method gives answers to many important but difficult problems which are inaccessible to other methods. For example, this method gives solutions to such problems as 'what is the probability that a single mutant with arbitrary dominance that appears in a finite sized population, and has a certain selective advantage, will eventually spread through the whole population?' (Kimura, 1957, 1962). This probability is usually called the *probability of gene fixation* in population genetics.

Despite such a development, its applicability to evolution had been quite limited until the era of molecular biology was ushered in, because population genetics theory is built on the concept of gene frequencies, that is, the proportion of various allelic genes in the population. Conventional studies of evolution are conducted at the phenotypic level, and there was no direct way of unambiguously connecting the two.

Such limitations were removed with the advent of molecular genetics. Two developments ensued. First, it became possible, through comparative studies of amino acid sequences of proteins among related organisms, such as comparisons of hemoglobins among vertebrates, and the use of paleontological data, to estimate the evolutionary rates of amino acid substitutions (Zuckerkandl and Pauling, 1965). This enabled us to estimate the evolutionary rates of nucleotide substitutions inside the genes. Secondly, the development of electrophoretic techniques enabled us to detect enzyme variability among individuals rapidly, and these studies have disclosed a wealth of polymorphic variants at the enzyme level in many organisms (Harris, 1966, Lewontin and Hubby, 1966). This led us to estimate genetic variability within species on a much finer scale than with conventional gene markers.

So, at last, the time had arrived to apply the mathematical theory of population genetics to determine how genes evolve. One should have

expected then that the principle of Darwinian positive selection or the survival of the fittest would be clearly shown to prevail at this most fundamental level.

For the benefit of readers who are not familiar with molecular genetics, I shall briefly summarize here the nature of genes and mutations at the molecular level. A gene, or more precisely a cistron, may be regarded as a linear message written with four kinds of DNA bases usually denoted by the letters A, T, G and C, where A = adenine, T = thymine, G = guanine and C = cytosine. When the gene acts, this message is transcribed into another molecule called messenger RNA, using four letters similar to those in DNA sequences but having U (= uracil) instead of T. Then, the information in the messenger RNA is translated into a polypeptide which is made by stringing together amino acids, of which there are twenty different kinds. Finally, the

Table 2.1. *Standard RNA code table*

1	2				3
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
U	Phe	Ser	Tyr	Cys	C
U	Leu	Ser	Term.	Term.	A
U	Leu	Ser	Term.	Trp	G
C	Leu	Pro	His	Arg	U
C	Leu	Pro	His	Arg	C
C	Leu	Pro	Gln	Arg	A
C	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
A	Ile	Thr	Asn	Ser	C
A	Ile	Thr	Lys	Arg	A
A	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
G	Val	Ala	Asp	Gly	C
G	Val	Ala	Glu	Gly	A
G	Val	Ala	Glu	Gly	G

Ala, Alanine (A); Arg, Arginine (R); Asn, Asparagine (N); Asp, Aspartic acid (D); Cys, Cysteine (C); Gln, Glutamine (Q); Glu, Glutamic acid (E); Gly, Glycine (G); His, Histidine (H); Ile, Isoleucine (I); Leu, Leucine (L); Lys, Lysine (K); Met, Methionine (M); Phe, Phenylalanine (F); Pro, Proline (P); Ser, Serine (S); Thr, Threonine (T); Trp, Tryptophan (W); Tyr, Tyrosine (Y); Val, Valine (V); Term., Chain terminating codon. Letters in parentheses are one letter expressions used by Dayhoff.

polypeptide chain is folded to become a functional protein. It is remarkable that protein molecules, which really make our lives possible and which include enzymes (such as cytochrome c, hormones (such as insulin) and structural materials (such as keratin), are all made in this way. The genetic message is composed in such a way that a set of three consecutive letters or 'triplet' forms a code word or 'codon' for an amino acid. With four possible letters at each position of a triplet, there are 4^3 or 64 codons. Of these, 61 are used to code for the 20 amino acids, while the remaining three codons serve as punctuation marks ('chain termination'). The entire 64 code words have been deciphered (see Table 2.1). Gene mutations are changes in the DNA message, and they can be classified into two groups; (1) replacements of one nucleotide base for another, and (2) structural changes consisting of deletions and insertions of one or more nucleotide bases as well as transpositions and inversions of larger DNA segments. It is known that nucleotide substitutions are found more frequently than structural changes when homologous genes are compared among related organisms. In other words, nucleotide substitutions are the most prevalent evolutionary changes at the molecular level.

2.2 Proposal of the neutral mutation-random drift hypothesis

The picture of evolutionary change that actually emerged from molecular studies, however, seemed to me quite incompatible with the expectations of neo-Darwinism. Among the salient features that were soon disclosed were the approximate uniformity of the rate of amino acid substitutions per year for each protein among diverse lineages, an apparent randomness in the pattern of substitutions, and a very high overall rate amounting to at least one mutant nucleotide substitution per genome every two years when extrapolated to the haploid amount of DNA in mammals.

This rate was obtained by using a conservative estimate that one amino acid substitution occurs in 28×10^6 years for a polypeptide chain consisting of 100 amino acids and also by assuming that the mammalian genome comprises some 4×10^9 nucleotide pairs. I then realized that this estimate of the rate of mutant substitution per genome per generation is at least several hundred times higher for mammals than the famous estimate previously obtained by Haldane (1957b). His estimate was based on his concept called 'cost of natural selection' which I later called 'substitutional load' (Kimura, 1960a). Haldane considered the situation in which an original gene becomes disadvantageous due to change of environment, while a mutant allele that was originally less fit becomes advantageous. He then computed the amount of selective death or 'cost' needed to carry out

gene substitution in the species. He showed that the cost D required to carry out one gene substitution by natural selection is independent of the selection coefficient but depends on the initial frequency of the substituted mutant gene. He obtained an elegant result; $D = -2\log_e p$ for a semi-dominant mutation, where D is the fraction of cumulative selective deaths and p is the initial mutant frequency. For example, if the initial frequency of the advantageous mutant is one in a million ($p = 10^{-6}$), $D = 27.6$. If the mutant is completely recessive, D becomes much larger. Haldane took $D = 30$ as a representative value in the actual course of evolution. If selection is taking place slowly at a number of loci with an average rate of one mutant substitution every n generations, the fitness of the species will fall below the optimum by a factor of $30/n$, namely, selection intensity $I = 30/n$. He conjectured that $I = 0.1$ is a reasonable figure for horotelic evolution (standard rate evolution) and suggested that mutants are substituted at the rate of about 1/300. In other words, one gene substitution occurs every 300 generations. He also considered that this accords with the observed slowness of evolution at the phenotypic level.

Haldane's treatment is based on a deterministic model that disregards the effect of random sampling of gametes in finite populations. To remove this limitation, I used the diffusion equation method and obtained the cost or the substitutional load in a finite population. For semidominant mutants, the load for one gene substitution is approximately $L(p) = -2\log_e p + 2$, where we can put $p = 1/(2N)$ for a molecular mutant (see Kimura and Maruyama, 1969 for details). This approximate formula is valid if the mutant allele has a definite selective advantage over the pre-existing allele. This formula shows that as compared with Haldane's formula the cost is larger by about 2, although this difference is usually relatively small.

Applying this and similar formulae under the assumption that the majority of mutant substitutions at the molecular level are carried out by positive natural selection, I found that the substitutional load in each generation is so large that no mammalian species could tolerate it. For example, for a species consisting of a half million individuals, $L(p) = 29.6$ with $p = 1/(2N) = 10^{-6}$. So, even equating one year with one generation, the load per generation is roughly 30. This means that to maintain the same population number and still carry out mutant substitutions at the rate of one substitution every two years (i.e. the average substitution load of 15), each parent must leave $e^{15} \approx 3.27 \times 10^6$ offspring for one of the offspring to survive and reproduce.

This was the main argument used when I presented the neutral mutation-random drift hypothesis of molecular evolution (1968a). An additional

reason not mentioned but realized at that time is that an unusually high rate of production of advantageous mutants is required to explain the high rate of molecular evolution by natural selection, especially assuming mutants with only a slight selective advantage, s_1, say 0.001 or less. To see this, let us consider a mammalian species having a large body size and a generation span of six years. For such a species, the effective population number (but not necessarily the actual number) may be 10^5 or less. Using the formula for the rate of evolution in terms of mutant substitutions (see equation 3.22 in chapter 3), I argued as follows: since we have $v = k/(4N_e s_1)$, where v is the mutation rate per gamete, k is the rate of mutant substitution per generation and N_e is the effective population size, if we put $k = 3$ (equivalent to a substitution rate of 0.5 per year), $2N_e = 10^5$ and $s_1 = 10^{-3}$, we get $v = 0.015$ or 1.5% per gamete per generation. This is a very high rate of production for advantageous mutations since it is the same order of magnitude as the total rate for lethal and semilethal mutations per gamete. Clearly, advantageous mutations should be much less frequent than deleterious mutations. If k is larger, or N_e or s_1 is smaller, v becomes still higher. The assumption that advantageous mutations are being produced at such a high and constant rate throughout the long history of evolution seemed to me to contradict a basic principle of adaptive evolution by natural selection. Because, one should expect that each substitution of an advantageous mutation must generally decrease the probability of subsequent mutations being advantageous unless the environment changes drastically.

Turning now to the problem of extant genetic variability within species, the new observations were also remarkable. As mentioned in the previous section, application of electrophoretic methods suddenly disclosed, starting in 1966, a wealth of genetic variability showing high incidence of molecular polymorphism at a large fraction of loci in various organisms. According to Lewontin and Hubby (1966), the average per locus heterozygosity is about 12% and the proportion of polymorphism is 30% for 18 randomly chosen loci in *D. pseudoobscura*. A similar result was reported by Harris (1966) in man. In these organisms, then, there must be thousands of loci where variants are segregating at high frequencies, and each individual may be heterozygous at hundreds or even thousands of loci. However, these polymorphisms were accompanied by no visible phenotypic effects and no obvious correlation with environmental conditions. If these were generally maintained by overdominance at individual loci, it seemed to me that the total amount of segregation load (Crow, 1958) or selective deaths involved must be intolerably large for higher animals.

The segregation load is created because in each generation inferior homozygotes are produced by segregation. Consider a locus with a pair of overdominant alleles A_1, and A_2, and let the relative fitnesses of the three genotypes A_1A_1, A_1A_2 and A_2A_2 be respectively $1-s_1$, 1 and $1-s_2$ ($0<s_1\leqq 1$, $0<s_2\leqq 1$). Then, the average fitness of the equilibrium population is less by

$$L_s = s_1 s_2/(s_1+s_2) \tag{2.1}$$

than the fitness of the heterozygote. If $s_1=s_2=s$, then $L_s=s/2$. For example, if the selection coefficient against the homozygotes is 1‰ the segregation load is $L_s=0.005$. Suppose that n overdominant loci are segregating independently and that their contribution to the fitness is multiplicative; then the fraction of selective elimination per individual due to n segregating loci is $1-e^{-sn/2}$. As a numerical example, let $s=0.01$ and $n=2000$, then this becomes $1-e^{-10}\approx 0.9999546$. This means that if 2000 overdominant loci are segregating, each with 1% heterozygote advantage, and if the selection is carried out by premature death of less fit homozygotes, each individual must produce on the average roughly 22 000 young in order to maintain the population number constant from generation to generation. It is evident that no mammalian species can afford such reproductive waste.

Considering these fascinating but puzzling observations, it occurred to me during 1967 that a majority of nucleotide substitutions in the course of evolution must be the result of random fixation of selectively neutral or nearly neutral mutants rather than positive Darwinian selection, and many of the enzyme polymorphisms are selectively neutral and maintained by the balance between mutational input and random extinction. I presented this thesis at the Genetic Club Meeting in Fukuoka in November 1967, and also submitted a short paper to *Nature* in December, which was published in February of the next year (Kimura, 1968a).

Then, in 1969 very strong support for my thesis, particularly with respect to molecular evolution, came with the publication of the paper entitled 'Non-Darwinian evolution' written jointly by Jack Lester King and Thomas H. Jukes. These authors arrived at the same idea as mine independently and corroborated their claim with a wealth of cogent data from molecular biology. The only difference is that whereas King and Jukes (1969) paid little attention to protein polymorphisms, my neutral theory is equally concerned with the maintenance of protein polymorphisms (Kimura, 1968a, b).

The possibility that alleles responsible for enzyme polymorphisms are not strongly selected and may be maintained by mutation pressure or by

very slight heterozygote advantage was also suggested by Crow (1968). A year earlier, a similar view that these polymorphisms are almost neutral in the contemporary populations was briefly mentioned by Robertson (1967) in relation to Harris' (1966) observation of human polymorphisms.

Still earlier, and indeed before the wealth of enzyme polymorphisms was discovered, we (Kimura and Crow, 1964) investigated theoretically the number of isoalleles that can be maintained in a finite population. In this paper we considered three possibilities: (1) a system of selectively neutral isoalleles whose frequency in the population is determined by the mutation rate and random drift; (2) a system of mutually heterotic alleles; (3) a mixture of heterotic and harmful mutants. We came to the conclusion, from the consideration of the segregation load in a finite population that the number of segregating loci due to overdominance for fitness is likely to be a minority (contrary to the claim of Wallace (1958)). Therefore, we thought that the typical *Drosophila* is likely to be homozygous for the majority of its selected genes. However, this by no means rules out the possibility of high heterozygosity with respect to selectively neutral or nearly neutral isoalleles. In fact, we wrote in that paper as follows: 'In very large populations, the possibility of many very nearly neutral, highly mutable multiple isoalleles cannot be ruled out, although there is no experimental evidence for the existence of such systems.' The experimental evidence was not to come until two years later.

As mentioned already, a straightforward interpretation of the then newly discovered, abundant protein polymorphisms in terms of the neutral isoallele model was given by Kimura (1968a, b). A few years later, the theory was elaborated with more evidence and with more confidence by Kimura and Ohta (1971a). In this paper, we claimed that evolutionary substitutions of amino acid and extant protein polymorphisms are not two independent phenomena but merely two different aspects of a single phenomenon caused by random frequency drift of neutral or nearly neutral alleles in finite populations. In short, protein polymorphism is a phase of molecular evolution.

The proposal of the neutral theory or non-Darwinian evolution led to a controversy called 'neutralist-selectionist controversy'. The pros and cons of the neutral theory have been debated in various scientific meetings and publications (see for example, *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* volume V: *Darwinian, Neo-Darwinian, and Non-Darwinian Evolution*, ed. L.Le Cam, J. Neyman, and E.L. Scott (1972), University of California Press; *Trends in Biochemical Sciences* 1: N152–N154).

The arguments for neutral and selective mechanisms have been reviewed

in the papers of Crow (1972, 1981) and Harris (1976), and in the books of Calder (1973), Lewontin (1974) and Ruffié (1976, 1982). Particularly, in Lewontin's book, a very detailed and critical appraisal both of selectionist and neutralist stands is made, and it has been widely quoted. Lewontin claims that the name 'neutral mutation theory' is inappropriate and he suggests that a more appropriate term is 'neoclassical theory', referring to classical vs. balance theories as coined by Dobzhansky (1955). I think that Dobzhansky's classical theory is a strawman, essentially a caricature of Muller's views and therefore a term derived from it is inappropriate. Furthermore, the picture of the neutral theory presented in Lewontin's book is, in my view, somewhat biassed and I intend to correct such a distorted picture in this book.

My claim that the observed rate of molecular evolution is too great to be explained by positive selection in the light of Haldane's cost principle was immediately met with strong criticism by Maynard Smith (1968). He showed that if we assume a 'threshold model' in which a certain fraction, such as 50%, of individuals having more than a threshold number of favorable alleles survive for reproduction, then a much larger number of gene substitutions can be carried out without a heavy genetic load, as compared with the case in which selection acts independently on different loci. For example, when 50% of the population is selected each generation, the number of loci at which selection can act simultaneously with a selection coefficient 0.01 is 25 500 under threshold selection, whereas the number of loci is only 138 under the multiplicative fitness model (both assuming that the population is haploid and that the frequency of the more favorable allele at each locus is 0.5). He argues that the assumption of multiplicative fitnesses between loci which underlies the Haldane cost principle is only valid under exceptional circumstances, and hence Haldane's conclusion that it will typically take 300 generations per gene substitution is unjustified. Maynard Smith went on to state that the rate of evolution could be greater than this by one or more orders of magnitude, and that my conclusion that a large proportion of gene substitutions at the molecular level are selectively neutral and have occurred by drift, although it may be true, does not follow necessarily from the cost of selection argument.

I think that there are several unrealistic features in Maynard Smith's model. First, he assumes that the fitness of individuals is solely determined by genes, i.e. 100% heritability. There is the possibility, however, that the broad-sense heritability of fitness is very low. In fact, Mukai, Schaffer and Cockerham (1972) showed in *Drosophila melanogaster* that the fitness of

individuals as a quantitative character has an extremely low, broad-sense heritability (H^2) of about 0.004, so that threshold or truncation selection cannot work effectively without entailing very severe culling, that is, heavy cost. Secondly, Maynard Smith assumes tacitly that an advantageous mutant exists ready-made at every one of a vast number of loci at any time without waiting for environmental changes. I think that such an assumption is highly unrealistic. For one thing, if it were true, evolution would be vastly more rapid in a species with a very large population size than one with a small population size. No such pattern is discernible at the molecular level. The rate of molecular evolution is roughly constant per year for individual proteins, such as the hemoglobin α chain, among diverse lineages (see chapter 4). I think that it is much more realistic to assume, as Haldane (1957b) did, that advantageous mutants are derived from a reservoir of previously deleterious mutants through environmental changes. A similar model as that of Maynard Smith was presented by Sved (1968) who however realized that the limiting factor in this type of formulation is the availability of favorable mutations. Maynard Smith's calculations were also criticized as unrealistic by O'Donald (1969) from a different viewpoint.

The suggestion that a threshold or truncation selection model can also be invoked to avoid the heavy segregation load even if polymorphisms are maintained at a very large number of loci by overdominance had previously been made by King (1967), Sved *et al.* (1967) and Milkman (1967). The models of these authors are essentially similar, so I shall explain that of King. In his model, it is assumed that each homozygous locus contributes one unit to a measure called 'survival factor parameter'. Environmental factors also contribute to this parameter. It is then assumed that individuals with the parameter values greater than a certain threshold value fail to survive. In other words, it is assumed that truncation selection is practiced in nature as in animal breeding. Using this model, King showed that a large number of heterotic loci each with a significant selective advantage (say, 0.01) can be maintained under a modest selective elimination such as 50%. One of the difficulties with this model is that, as pointed out by Lewontin (1974), an unrealistically large amount of fitness reduction is expected to occur upon inbreeding. Let us take, as an example, one of the cases tabulated by King (1967). Assuming that the average survival rate is 50% and that the proportion of variance due to overdominance is 0.033, it can be shown that 1000 loci each with heterozygote frequency 0.3 and with 1% selective advantage of the heterozygotes over either homozygotes can be maintained. If we calculate what would happen if the genome is made completely homozygous (i.e. 100% inbreeding), it turns out that survival of

the resulting homozygotes is less than one in ten thousand (10^{-4}). This is contradicted by experimental studies in *Drosophila* (Temin *et al.*, 1969).

Although these *Drosophila* studies measured viability only, there is experimental evidence that fitness differences are about five times as great as those for viability alone, still inconsistent with the enormous inbreeding decline that King's numbers would predict.

Nevertheless, these models assuming truncation selection have been hailed by some selectionists as the alternative solution to the neutral theory (and in fact, a more desirable one) for understanding molecular evolution and polymorphism. Not much attempt, however, has been made to confirm that this type of selection actually occurs and is really responsible for gene substitution and maintenance of genetic variability in nature.

I believe that during the decade since the origination of the neutral theory, supporting evidence for it increased, particularly by the recent finding that nucleotide changes that cause no amino acid changes (called *synonymous* or *silent* substitutions) occur at much higher rates in evolution than those which lead to amino acid changes (Salser, 1978; Nichols and Yanofsky, 1979).

Since natural selection acts through phenotypes of the organisms for which the structure and function of proteins play a decisive role, one should expect that the mutations that do not cause amino acid changes in proteins, other things being equal, are much less subject to natural selection than those which cause amino acid changes. Yet, abundant data have emerged in the last few years showing that synonymous changes, and also nucleotide changes in the non-coding regions (that do not participate in protein formation) in DNA are the most prevalent evolutionary changes at the molecular level (Kimura, 1977; Jukes, 1978b).

In addition, it has been discovered very recently that some genes of eukaryotes (higher organisms with a true nucleus) contain regions called *intervening sequences* or *introns* which are not included in the mature messenger RNA and therefore do not participate in protein formation (see Crick, 1979 for review). Already, there is evidence that evolutionary nucleotide substitutions in introns are very rapid (van den Berg *et al.*, 1978). As I shall examine in more detail later, *the rule is that those molecular changes that are less likely to be subject to natural selection occur more rapidly in evolution.*

I think that King and Jukes had remarkable insight when they wrote in 1969: 'if DNA divergence in evolution includes the random fixation of neutral mutations, then the third-position nucleotides should change more rapidly', reasoning that most of the nucleotide changes at the third position

of the codon are synonymous and therefore are more likely to be neutral.

In this context, I wrote in one of my papers on the neutral theory (Kimura, 1968b): 'The recent findings of "degeneracy" of DNA code, that is, existence of two or more base-triplets coding for the same amino acid, seem to suggest that neutral mutations may not be as rare as previously considered'. In the same paper, however, I also added a note of caution: 'It is important to note that probably not all synonymous mutations are neutral, even if most of them are nearly so'. Perhaps, this was a pertinent statement in the light of the recent finding that synonymous codons are often used in 'non-random' or unequal fashion (see section 7.5 for detailed discussion on this phenomenon).

3

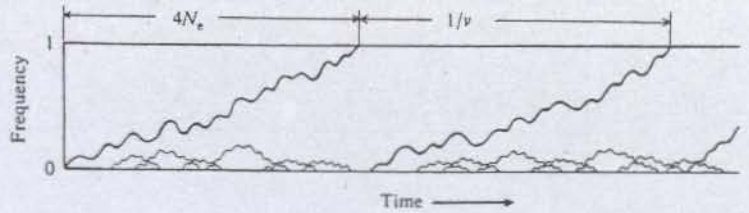
The neutral mutation-random drift hypothesis as an evolutionary paradigm

3.1 Chance acting on selectively equivalent alleles

The neutral mutation-random drift hypothesis (or the neutral theory for short) holds that at the molecular level most evolutionary change and most of the variability within species are not caused by Darwinian selection but by random drift of mutant alleles that are selectively neutral or nearly neutral. The essential part of the neutral theory is not so much that molecular mutants are selectively neutral in the strict sense as that their fate is largely determined by random genetic drift. In other words, the selection intensity involved in the process is so weak that mutation pressure and random drift prevail in molecular evolution.

In order to see more fully what the neutral theory purports, and how it differs from the traditional synthetic theory, let us consider an evolutionary process in which mutant genes are substituted one after another within the species. Each such substitution is made up of a sequence of events in which a rare mutant form which appeared, usually singly represented, in the population, finally spreads through the whole population reaching fixation, i.e. frequency of 100%. Fig. 3.1 illustrates the course of change in the

Fig. 3.1. Behavior of mutant genes following their appearance in a finite population. Courses of change in the frequencies of mutants destined to fixation are depicted by thick paths. N_e stands for the effective population size and v is the mutation rate.



frequencies of mutant alleles following their appearance in a finite population.

Although a large number of mutants arise in each generation in any reasonably large population, the majority are lost by chance within a small number of generations (Fisher, 1930b; Kimura and Ohta, 1969b). It is often not realized that this is true not only for deleterious and selectively neutral mutants but also for advantageous mutants unless the advantage is very large. For example, if a mutant has a 1% selective advantage, the chance is only about 2% that it will eventually spread through the whole population (Haldane, 1927; Fisher, 1930b). In the remaining 98% of the cases, it will be lost by chance from the population without being used in evolution. Thus, there is a vast difference between the total number of advantageous mutants that have ever occurred in any species in the course of evolution and the number that have actually been incorporated (fixed) into the species. Actually, it is a lucky minority that manage to increase their frequencies and spread in the species, finally reaching the state of fixation. Each such event usually takes a large number of generations. For a selectively neutral mutant, it takes on the average $4N_e$ generations until fixation if we exclude the cases in which it is lost (Kimura and Ohta, 1969a). In Fig. 3.1, the courses of change in the frequencies of mutants destined for fixation are depicted by thick paths.

According to the traditional interpretation, mutant genes must have some selective advantage for them to spread through the species. Occasionally, however, a selectively neutral mutant may be carried along by 'hitchhiking' (i.e. by being linked to a favorable gene) to reach a high frequency, but this is a secondary phenomenon. On the other hand, the neutral theory is based on the well-known fact in population genetics that mutants do not need a selective advantage for some of them to spread through the population. If mutants are selectively equivalent to the pre-existing forms from which they are derived, their fate is left to chance and their frequencies increase or decrease fortuitously as time goes on. Such a fluctuation arises not only because of the small initial number of mutants, but also because of the finite size of populations; in each generation a relatively small number of gametes are sampled out of a vast number of male and female gametes produced, to create the individuals of the next generation. Although the overwhelming majority of such neutral mutants are lost by chance, a tiny minority will eventually become fixed in the population. If neutral mutations are common at the molecular level, and if such a random process goes on incessantly for a very long time, say tens of millions of generations, the genetic composition of the population will

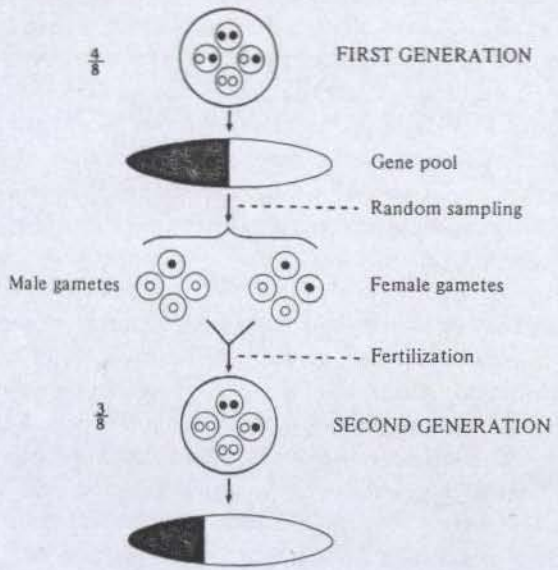
change tremendously. Even though a mutant gene may have only one chance in a million of spreading through the population, in a population large enough to produce 10 mutants at a locus each year there will be 10 mutant substitutions at this locus each million years. This random process is known as 'random genetic drift' (or 'random drift' in short) in population genetics.

Here, I would like to point out that we must keep clear the distinction between *gene mutation* at the individual level and the *mutant (gene) substitution* at the population level. Not infrequently, in the literature of molecular evolution, the differences of amino acids between homologous proteins are referred to simply as 'randomly distributed point mutations', but such an expression tends to obscure the basic distinction between these two concepts; the amino acid differences between species reflects the results of mutant substitutions rather than simply gene mutations.

3.2 Random genetic drift due to finite population size

The concept of random genetic drift is basic for the neutral theory, so I shall explain this in some detail. By random genetic drift I mean

Fig. 3.2. Illustration of the process of random change of gene frequencies during one generation in a population consisting of four breeding individuals. This corresponds to a population of four monoecious individuals in which random union of gametes takes place in reproduction.

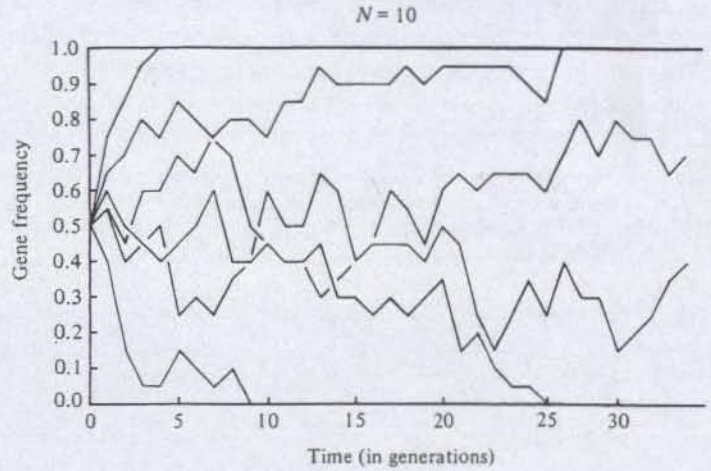


random fluctuation of gene frequencies in a population caused by random sampling of gametes in reproduction. In any sexually reproducing species, the total number of individuals is not only finite, but also can be regarded as a random sample chosen from a much larger collection of male and female gametes (or 'gene pool') produced by the parental generation. The amount of fluctuation in gene frequencies (that is, proportion of various alleles) is expected to be larger, the smaller the population.

Fig. 3.2 illustrates how gene frequencies change from one generation to the next assuming a model population consisting of four individuals. In this figure, a pair of alleles A_1 and A_2 are expressed by black and white balls, and the frequency of the 'black' allele changes from 4/8 (i.e. 50%) to 3/8 (i.e. 37.5%) in one generation by chance. More generally, if the frequency of A_1 in the population happens to be x at a given moment, then the change of the frequency of A_1 in one generation has mean zero and variance (that is, the squared standard deviation) $x(1-x)/(2N)$, where N is the population size in the next generation. The process of random drift may also be illustrated as in Fig. 3.3 where a few courses of change ('sample paths') in gene frequency are shown starting from 50%.

There is another method of representation, suitable for mathematical treatment of the process of random drift. The process is described in terms

Fig. 3.3. A few sample paths representing processes of random gene frequency drift starting from 50%. These are chosen from sample paths actually obtained in Monte Carlo experiments by a computer, simulating a population comprising ten breeding individuals ($N = 10$) and starting from the gene frequency 0.5.



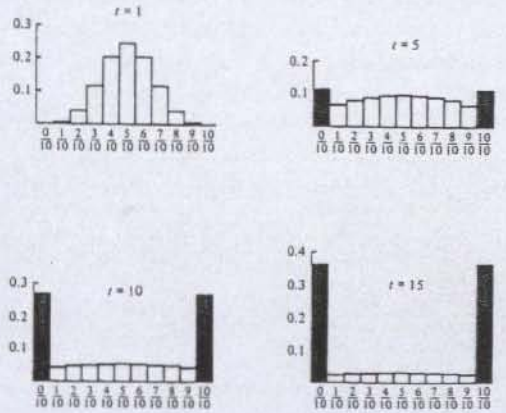
of the probability distribution of gene frequencies. We consider an infinite collection of populations each consisting of N breeding individuals in which a pair of alleles A_1 and A_2 are segregating. We shall denote the frequency of A_1 by x, so that the frequency of A_2 is $1-x$. Let us assume that at generation zero ($t=0$) all the populations in this ensemble contain A_1 in the same frequency (say p), and then investigate how the proportions of populations having A_1 in various frequencies (that is, the distribution of x) change in subsequent generations.

In Fig. 3.4. the course of change is illustrated assuming $N=5$ and $p=0.5$. The abscissa (x) represents the frequency of A_1 and the height of the column the proportion of the populations having this A_1 frequency. The two terminal frequency classes ($x=0$ and $x=1$) are shown by black columns.

As may be seen from this figure, the frequency distribution of unfixed classes becomes more and more spread out as time goes on, and eventually A_1 becomes either fixed in the population or lost from it (fixation of A_2). After 10 generations, the distribution is nearly flat and decreases at the rate 0.1 per generation which is equal to $1/(2N)$.

When the population size N is larger, say one hundred or more, histograms representing the gene frequency distribution can be approxi-

Fig. 3.4. Histograms showing the course of change of the gene frequency distribution with time (generations 1, 5, 10 and 15) through random sampling drift. In this illustrated example, the population consists of five monoecious individuals breeding at random ($N=5$), and the frequency of A_1 starts from 50% at $t=0$ (i.e. $p=0.5$). The abscissa represents the frequency of A_1 in the population. The black column at the left represents the proportion of populations in which A_1 is lost. Similarly, the black column at the right represents the proportion of populations in which A_1 is fixed.



mated without much error by a smooth frequency curve. Also, the time parameter t can be treated as continuous ($t \geqq 0$), although we can still measure time with one generation as the unit.

Let $\phi(p, x; t)$ be the probability density that the frequency of allele A_1 in the population becomes x in the t-th generation, given that it is p at $t=0$. Then it can be shown that $\phi(p, x; t)$ satisfies the diffusion equation

$$\frac{\partial \phi}{\partial t} = \frac{1}{4N}\frac{\partial^2}{\partial x^2}[x(1-x)\phi], \qquad (0 < x < 1), \tag{3.1}$$

where $\phi = \phi(p, x; t)$. For the rationale of this equation readers may refer to Kimura (1964) and Crow and Kimura (1970; see pp. 382–3). The exact solution of this equation which satisfies the condition that the initial gene frequency is p was obtained by Kimura (1955), and is expressed as follows;

$$\phi(p, x; t) = \sum_{i=1}^{\infty} p(1-p)i(i+1)(2i+1)F(1-i, i+2, 2, p) \times F(1-i, i+2, 2, x)e^{-i(i+1)t/(4N)}, \tag{3.2}$$

where $F(\cdot, \cdot, \cdot, \cdot)$ stands for the hypergeometric function.

Fig. 3.5. The course of change in the frequency distribution of allele A_1 due to random genetic drift. In Figure (*a*), the initial frequency (p) of A_1 is 0.5, while in Figure (*b*) it is 0.1. In both figures, t stands for time and N stands for the effective population number. The abscissa is the frequency (x) of A_1 and the ordinate is the probability density (ϕ). (From Kimura, 1955).

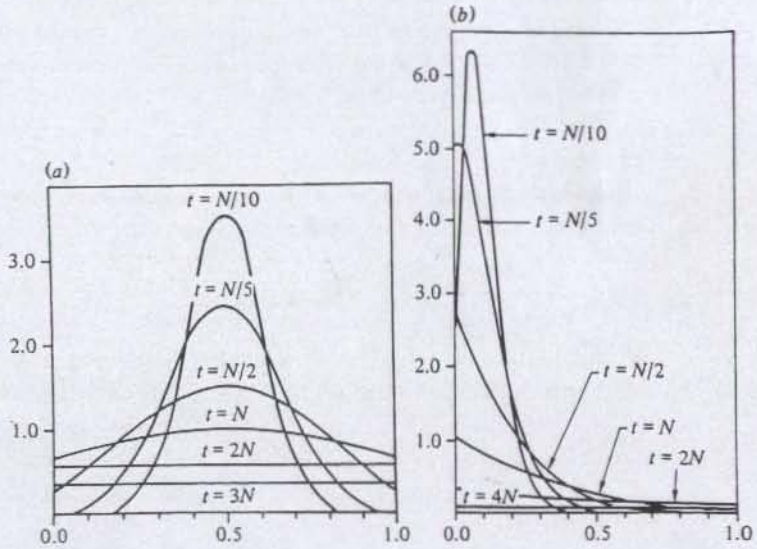


Fig. 3.5 illustrates the course of change of the gene frequency distribution of unfixed classes for the two following cases; in Fig. 3.5*a* the initial frequency (p) is 0.5 and in Fig. 3.5*b* it is 0.1. It may be seen from Fig. 3.5*a* that, if $p = 0.5$, the distribution curve becomes nearly flat and that genes are unfixed in roughly 50% of the cases after $2N$ generations. On the other hand, if $p = 0.1$ (Fig. 3.5*b*), it takes $4N$ or $5N$ generations before the distribution curve becomes practically flat. In general, it can be shown that, after a large number of generations, the distribution of unfixed classes ($0 < x < 1$) becomes flat and decays at the rate $1/(2N)$ per generation.

It might be thought that there is little interest in the flat distribution of unfixed classes, since by the time the distribution becomes flat most of the genes are already fixed or lost. Why, then, did Fisher and Wright emphasize the flat shape of the distribution. The reason is that, instead of thinking of a group of populations all starting with the same frequency, they were considering a group of populations starting from a wide variety of initial frequencies.

3.3 Effective population size

Actual populations are expected to be much more complicated in breeding structure than the model populations considered in the previous section. Therefore, it is desirable if such a complicated situation can be reduced to an equivalent simple case for which the mathematical treatment is much easier. The concept of effective population number (or size) was introduced by Wright (1931) to meet this need, and it has become one of the fundamental concepts in treating the problem of random genetic drift.

Here I shall quickly summarize some of the important results obtained on this topic (for more details, see Kimura and Ohta, 1971b, chapter 3). In the following, N_e denotes the effective size and N denotes the actual size of the population.

(i) If a population consists of N_m breeding males and N_f breeding females, the effective size of the population is

$$N_e = \frac{4N_m N_f}{N_m + N_f} \tag{3.3}$$

(Wright, 1931). This shows that if the number of breeding males and females differs greatly, the effective size N_e depends mainly on the less numerous sex. In the extreme case in which the number of females is indefinitely greater than that of males, $N_e = 4N_m$. Thus, if a herd is headed by one male, the effective size of the herd is only four even when the number of females is indefinitely large.

(ii) If N parents furnish varing number (k) of gametes to the next generation, and if the population is stationary in size ($\bar{k} = 2$) and mating is at random,

$$N_e = \frac{4N - 2}{\sigma_k^2 + 2}, \tag{3.4}$$

where σ_k^2 is the variance of k (Wright, 1938b). Note that if the progeny number follows the Poisson distribution so that $\sigma_k^2 = \bar{k} = 2$, the effective size N_e is approximately equal to the actual number N. If the variance is larger than the mean, as is likely in most natural populations, the effective number is less than the actual number. On the other hand, if the parents leave exactly the same number of offspring so that $\sigma_k^2 = 0$, we get $N_e = 2N - 1$, namely, the effective size is about twice the actual size. Such a situation is unlikely to occur in nature but may be realized in experimental populations under human control.

(iii) If the number of individuals changes cyclically with a relatively short period of n generations, the effective size is

$$N_e = n \Big/ \sum_{i=1}^{n} (1/N_i). \tag{3.5}$$

In other words, N_e is equal to the harmonic mean (i.e. reciprocal of the mean of the reciprocals) of the number of individuals over one cycle (Wright, 1938). Thus the effective size is controlled largely by the phase of small numbers. For example, if an insect population expands in five generations from 10 to 10^6 in a geometric series and then returns to the initial size through the following five generations, $N_e \approx 54$. Such a periodic reduction in the number of breeding individuals can play an important role in making the effective size much smaller than the average actual population size. For the effect of a population bottleneck on genetic variability, readers should consult Nei *et al.* (1975). The effective size also decreases if the difference in fecundity among individuals has a high heritability (Robertson, 1961; Nei and Murata, 1966).

(iv) If generations overlap, Crow and Kimura (1972) showed that for a stationary population

$$N_e = N_0 \tau \bar{l} \tag{3.6}$$

approximately, where N_0 is the number born each year, τ is the average age of reproduction and

$$\bar{l} = \int_0^\infty l_y^2 b_y dy, \tag{3.6a}$$

where l_y is the probability of surviving from birth to age y, and $b_y\,dy$ is the

expected number of births to an individual in the age interval y to $y + dy$. When the birth rate is uniform and the death rate is low during the period of reproduction, the above formula (3.6) becomes essentially the same as the formula by Nei and Imaizumi (1966) which is

$$N_e = N_m \tau, \tag{3.7}$$

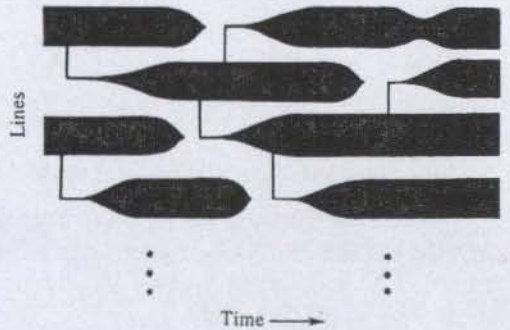
where N_m is the number born per year who survive until the mean age of reproduction and τ is the mean age of reproduction. An important advance in this subject was made by Hill (1972). He showed that the effective size of a random mating population with overlapping generations is equal to the effective size of a population with discrete generations which has the same number of individuals entering the population each generation and the same variance of lifetime progeny production. His formula for monoecious diploids is

$$N_e = (4N - 2)L/(\sigma_n^2 + 2), \tag{3.8}$$

where N is the number born every year, L is the generation interval in years, and σ_n^2 is the variance of lifetime family size per parent, or, more strictly, the variance per individual of the number of gametes contributed to the next generation. Note that many newborn individuals may have no progeny. (For a simplified derivation, see Hill, 1979.)

(v) If a species consists of a large number of more or less isolated subpopulations, and if extinction and subsequent replacement occur frequently among subpopulations, the effective size of the species is greatly reduced as compared with the situation in which a whole species forms a random mating or panmictic unit. This property is particularly pertinent when we consider genetic variability of a bacterial species (see chapter 9). It

Fig. 3.6. Diagram illustrating the population structure of an asexually reproducing haploid species consisting of a large number of lines which are subject to frequent extinction and subsequent replacement.



can be shown (see, Maruyama and Kimura, 1980) that if a haploid species consists of n lines which reproduce asexually and each of which is subject to random extinction and subsequent replacement (see Fig. 3.6), then the effective population size of the species is

$$N_e = \bar{N} + n/(2\lambda) + n\bar{N}v/\lambda, \tag{3.9}$$

where $\bar{N}$ is the harmonic mean of the effective population size per line per unit time, n is the number of lines (assumed to be large), λ is the rate of line extinction and v is the mutation rate (assuming the infinite neutral allele model and measuring these rates using the same unit of time). The amount of genetic variability within a haploid species may be measured in terms of 'virtual heterozygosity' which is defined as one minus the sum of squares of allelic frequencies. At equilibrium in which mutational production of new alleles and random extinction of pre-existing alleles balance each other, the virtual heterozygosity is given by $2N_e v/(1 + 2N_e v)$. Note that the effective size of the species N_e can be several orders of magnitude less than its apparent size which is equal to $n\bar{N}$, where $\bar{N}$ is the arithmetic mean of the effective population size per line per unit time. In a diploid species consisting of n colonies (subpopulations), if migration takes place between colonies at the rate m per generation (assuming the island model of population structure) in addition to extinction and recolonization of colonies, it can be shown that the effective size of the species is

$$N_e = \bar{N} + n/[4(v + \lambda + m)] + n\bar{N}(v + m)/(v + \lambda + m), \tag{3.10}$$

where v is the mutation rate and λ is the colony extinction rate per generation. If the rate of colony extinction (λ) is much larger than the migration rate (m) of individuals, the effective size of the species is much smaller than $n\bar{N}$, the product of the number of colonies and the effective size of an individual colony. The average heterozygosity at equilibrium is $4N_e v/(1 + 4N_e v)$, where N_e is given by (3.10).

Finally, I would like to call the readers' attention to the fact that the effective population size applicable to neutral evolution is that of the entire species (or subspecies if this forms an independent reproductive unit), but not the effective size of a local population which forms a part of it. This is particularly important when we consider the problem of subdivided population structure in general.

3.4 Neutral and nearly neutral mutations

For the behavior of mutant alleles to be mainly controlled by random drift, it is not necessary that they be strictly neutral, that is, completely equivalent with respect to fitness. What is required is that the

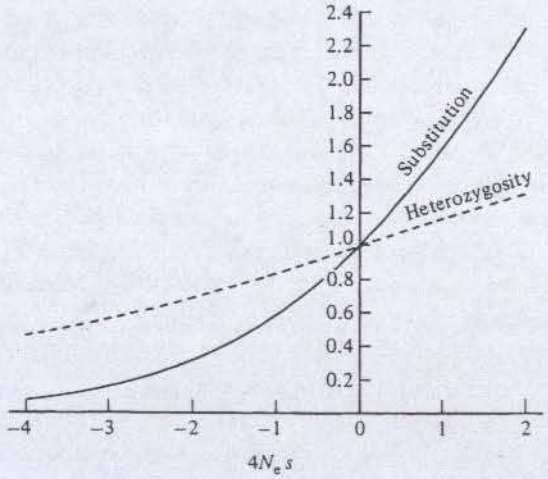
magnitude of their selective advantage or disadvantage, as measured by the selection coefficient, s, does not exceed, roughly speaking, the reciprocal of twice the effective population size; $|s| < 1/(2N_e)$. In my 1968 paper on the neutral and nearly neutral isoalleles (Kimura, 1968b), I said that a mutant gene may be defined as 'almost neutral' if $|2N_e s|$ is much smaller than unity, that is, $|s| \ll 1/(2N_e)$. Li (1978) claims that this definition is too strict and suggests that a more reasonable definition of almost neutral is $|s| \leqq 1/N_e$.

To see the behavior of alleles in the neighborhood of neutrality, let us consider the probability that a mutant allele eventually becomes fixed in the population. Let $u(p)$ be the probability that a mutant allele whose initial frequency is p will ultimately be fixed in the population. We consider the simple case of genic selection in which the mutant allele has selective advantage s in single dose and $2s$ in double dose, that is, the case of semidominance (or 'no dominance'). Then, it can be shown (Kimura, 1957, 1962) that

$$u(p) = (1 - e^{-Sp})/(1 - e^{-S}), \tag{3.11}$$

where $S = 4N_e s$. The probability of ultimate fixation of an individual mutant, denoted by u, can then be obtained by putting $p = 1/(2N)$. If $|s|$ is

Fig. 3.7. Relative contribution of a weakly selected mutant to gene substitution and heterozygosity. The solid curve represents u/u_0 as a function of $4N_e s$, where N_e is the effective population size and s is the selection coefficient of the mutant. The broken curve represents $H_T/H_{T,0}$.



small, we have

$$u = \frac{S}{2N(1 - e^{-S})} \tag{3.12}$$

as a good approximation. In these formulae, the selection coefficient (s or $S/4N_e$) may be positive or negative depending on whether the mutant is advantageous or disadvantageous. When the mutant is strictly neutral ($s \to 0$), the formula reduces to

$$u = 1/(2N). \tag{3.13}$$

If we denote this fixation probability as u_0, and if we consider the ratio u/u_0, then we get

$$u/u_0 = S/(1 - e^{-S}). \tag{3.14}$$

This represents the contribution of a mutant having selective advantage s to gene substitution, relative to that of a strictly neutral mutant.

In Fig. 3.7, u/u_0 is plotted as a solid curve. For example, if $s = -1/(4N_e)$, this ratio is about 0.58.

Next, let us consider the relative contribution of a weakly selected mutant to population heterozygosity. Again assuming semidominance in fitness, let $H(p)$ be the sum of the fraction of heterozygotes involving a mutant allele over all generations until either fixation or loss, starting from the frequency p. Then, it can be shown (Kimura, 1969b; this corresponds to equation 15′ therein with $\nu_m = 1$) that

$$H(p) = 2[u(p) - p]/s, \tag{3.15}$$

where $u(p)$ is the probability of ultimate fixation as given in (3.11). For a single mutant, $p = 1/(2N)$, and the corresponding sum of heterozygotes is

$$H_T = \frac{S - 1 + e^{-S}}{Ns(1 - e^{-S})}, \tag{3.16}$$

where $S = 4N_e s$ as before. To simplify our treatment, let us assume that the actual and the effective population sizes are equal, that is, $N_e = N$. Then, we get

$$H_T = 4(S - 1 + e^{-S})/[S(1 - e^{-S})]. \tag{3.17}$$

For a strictly neutral mutant ($S \to 0$), this reduces to $H_T = 2$ which we denote by $H_{T,0} = 2$. Thus, the ratio

$$H_T/H_{T,0} = 2(S - 1 + e^{-S})/[S(1 - e^{-S})], \tag{3.18}$$

where $S = 4N_e s$, represents the relative contribution to heterozygosity of a mutant having selective advantage s, taking a strictly neutral mutant as the standard.

In Fig. 3.7, this ratio is plotted as a broken curve. It is clear that contribution to heterozygosity changes much more slowly than the contribution to gene substitution as s deviates from zero (i.e. strict neutrality). For example, if $s = -1/N_e$ or $S = -4$, we have $H_T/H_{T,0} \approx 0.46$, whereas $u/u_0 \approx 0.075$. This is easily understandable since deleterious alleles (unless their effect on fitness is large) behave much like strictly neutral alleles when their frequencies are low, being strongly influenced by random sampling drift. Thus deleterious alleles can contribute significantly to heterozygosity. On the other hand, they are effectively selected against at high frequencies, being prevented from becoming fixed, and therefore cannot contribute significantly to gene substitution.

3.5 Population dynamics of mutant substitution

I shall now treat quantitatively the process of accumulation of new mutants within the species in the course of evolution. Let us fix our attention on the process by which mutant genes are substituted one after another in the population as depicted by thick paths in Fig. 3.1. We first consider the entire set of nucleotides making up the genome (haploid chromosome set). For a human, this number is very large, of the order of 3.5 billion (3.5×10^9). Since the mutation rate per nucleotide site is low (say, 10^{-8} or less), we may assume that whenever a mutant appears, it occurs at a new site in which mutant forms are not already segregating in the population. This assumption (Kimura, 1969b) is now known as the 'infinite site model' in population genetics (see Kimura, 1971).

Let v be the mutation rate per gamete per generation. Since each individual has two sets of chromosomes, there are $2N$ chromosome sets in a population of N individuals, and therefore $2Nv$ new, distinct mutants will be introduced into the population in each generation (for a haploid population, the corresponding number is Nv). Let u be the probability of a single mutant ultimately reaching fixation in the population. For a semidominant mutant, this is given by equation (3.12). When the process is viewed over a very long time, the rate per generation of mutant substitutions in the population is equal to the number of new mutations produced in each generation multiplied by their ultimate fixation probability. Thus, we have

$$k = 2Nvu. \tag{3.19}$$

Note that $2Nv$ new mutants appear in each generation of which a fraction u eventually reach fixation. Thus k represents *the rate of evolution in terms of mutant substitutions.*

Although we considered the genome as a whole in the above consideration, the results can also be applied with good approximation to a single gene locus consisting of several hundreds or more of nucleotide sites. So, in the following, we restrict our consideration to a gene locus or cistron, and express the evolutionary rate taking the nucleotide or codon as the unit. Also, we often find it convenient to consider the protein coded by a gene rather than the gene itself.

In considering mutations at the molecular level, two very important points that we must keep in mind are: (i) the number of possible allelic states at any gene locus (cistron) is so large as to be practically infinite, and (ii) back mutation in the strict sense is so rare as to be practically negligible in any short interval of time. As an example, let us take the cistron coding for the α chain of the mammalian hemoglobins. This polypeptide consists of 141 amino acids, and so this corresponds to 423 nucleotide sites (actually, the gene coding for this polypeptide contains intervening sequences or 'introns', but we neglect these parts). This allows 4^{423} or some 10^{254} allelic states through base replacements alone, because each nucleotide site may be occupied by any one of the four kinds of nucleotide bases. Thus, for any one of these alleles, there are 3×423 or 1269 other alleles that can be reached by a single step base replacement. The probability of returning to the original allele from any one of the derived alleles by further single base replacement is only 1 in 1269, assuming that all base replacements occur with equal probability.

Incidentally, these considerations justify a particular model of allelic mutations (called 'infinite allele model') which assumes that whenever mutation occurs it leads to a new, non-preexisting allele (Kimura and Crow, 1964). Although this model is suitable for treating the level of extant genetic variability in the species (see chapter 8), it is not suitable to treat the process of mutant substitutions we are considering here. At a highly polymorphic locus, fixation of a single allele is unlikely to occur and what we see is the shifting of allelic frequencies with time. So, unless we identify mutations at individual nucleotide sites, no clearcut interpretation of the rate of mutant substitutions is possible. This is why treatment of mutant substitutions by using the infinite allele model, as was done by Guess and Ewens (1972), is unsatisfactory for practical biological use.

Returning to equation (3.19) for the rate of evolution, let us consider two contrasting cases, the case of neutral mutations and that of definitely advantageous mutations. First, let us assume that mutant alleles are selectively neutral. In this case, as shown in (3.13), $u = 1/(2N)$. Substituting this in (3.19), we get

$$k = v. \tag{3.20}$$

In other words, the rate of evolution in terms of mutant substitutions per generation is equal to the mutation rate per gamete independent of population size (Kimura, 1968a). This remarkable property is only valid for neutral alleles. However, as shown in the previous section, if a mutant has selection coefficient much smaller in absolute value than $1/(2N_e)$, it behaves like a neutral mutant so that this equation holds approximately.

Next, let us assume that the mutant allele has a definite selective advantage such that $4N_e s \gg 1$. Then, from (3.12) we have

$$u = 2sN_e/N \tag{3.21}$$

with good approximation (Kimura, 1964). Substituting this in (3.19) we obtain

$$k = 4N_e sv. \tag{3.22}$$

This means that the rate of evolution depends on the effective population size (N_e) and selective advantage (s), as well as on the rate (v) at which advantageous mutants are produced each generation. One should expect in this case that the rate of evolution would depend strongly on the environment, being high for a species offered a new ecologic opportunity but low for those kept in a stable environment. I think it highly unlikely that the product $N_e sv$ remains the same for diverse lineages of vertebrate evolution, irrespective of whether the evolution at the phenotypic level is very rapid (as in the line leading to man) or practically stopped (as in the line leading to carp). As we shall see in the next chapter, the observed rate of evolution for a given molecule, say hemoglobin, is constant per year among diverse lineages, and equation (3.22) appears to be incompatible with such an observation. Equation (3.22) would also be incompatible with the observation if observed rates were constant per generation rather than per year. The observed constancy is much more compatible with the expectation of the neutral theory (see equation 3.20), although doubts have been expressed by various authors as to why the neutral mutation rate per year (rather than per generation) is equal between organisms having very different generation lengths. This problem will be discussed later (see section 8.7) when we treat the model of 'effectively neutral mutations' (Kimura, 1979a). Note that the problem does not arise when comparisons are made among organisms with similar generation lengths but living under very different conditions.

In considering the population dynamics of mutant substitutions, we need to know, in addition to the probability of gene fixation, the average length of time involved for each substitution. A general theory on this subject has been worked out by Kimura and Ohta (1969a) using the diffusion equation

method. The theory gives the average number of generations until fixation (excluding the cases of eventual loss), assuming that the initial frequency of the mutant allele is p. In the special case of a selectively neutral mutant, putting $p = 1/(2N)$, the average number of generations until fixation becomes approximately

$$\bar{t}_1 = 4N_e. \tag{3.23}$$

In other words, it takes, on the average, four times the effective population size for a selectively neutral mutant to reach fixation by random frequency drift. In addition, it was shown that the number of generations until fixation

Fig. 3.8. Probability distribution of the length of time until fixation of a selectively neutral mutant. The abscissa represents the time measured with $4N_e$ generations as the unit, and the ordinate the probability density (from Kimura, 1970).

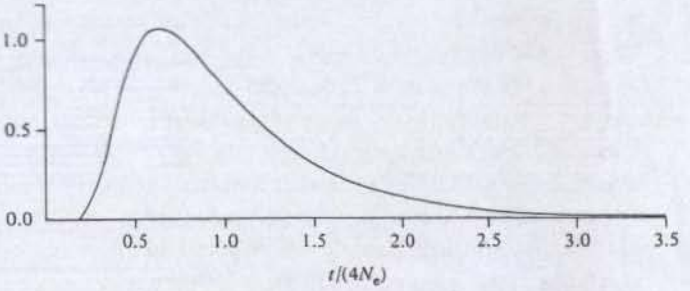
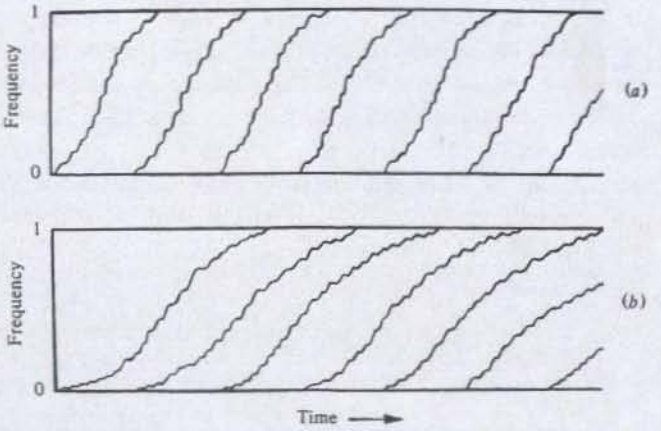


Fig. 3.9. Processes of mutant substitutions in the population. Individual mutant alleles increase much more rapidly within the population in case (*a*) than in case (*b*), but the rate of evolution in terms of mutant substitutions is the same in both cases.



has a standard deviation of about $2.15N_e$ or roughly half the mean (Kimura and Ohta, 1969b; Narain, 1970). Furthermore, in the case of neutral mutants, the probability distribution of the length of time until fixation has been obtained (Kimura, 1970). This distribution is illustrated in Fig. 3.8. From the figure, it may be seen that fixation before $0.8N_e$ generations is quite unlikely to occur. For selectively advantageous mutants, the average length of time until fixation is shorter, while if they are overdominant, the time is prolonged.

Generally speaking, it takes a very long time for each mutant to reach fixation in the population, and therefore, we must consider a still longer time when we define the rate of evolution as expressed by equation (3.19). Note that the rate of evolution thus defined is independent of the rate at which individual mutant alleles increase or decrease within the population. What matters is the average interval between two consecutive fixations. Thus, the two cases depicted in Fig. 3.9 have the same k values even if individual mutant alleles increase much more rapidly within the population in case (*a*) than case (*b*). For neutral mutants, the average time between two consecutive fixations is $1/v$ generations, where v is the mutation rate.

3.6 On some misunderstandings and criticisms

Since a number of criticisms and comments have been made regarding my neutral theory, often based on misunderstandings, I would like to take this opportunity to discuss some of them. The neutral theory by no means claims that the genes involved are functionless as mistakenly suggested by Zuckerkandl (1978). They may or may not be, but what the neutral theory assumes is that the mutant forms of each gene participating in molecular evolution are selectively nearly equivalent, that is, they can do the job equally well in terms of survival and reproduction of the individual.

Sometimes, neutral changes are referred to as evolutionary 'noise' by anti-neutralists. I think, however, this is a misnomer. Just as synonyms are not 'noise' in language, it is not proper to regard the substitution of neutral alleles simply as noise or loss of genetic information. Thoday (1975) classifies 'neutral' allele pairs into four kinds; strictly neutral, quasi-neutral, conditionally neutral and pseudo-neutral. He says that strictly neutral allele pairs clearly make no contribution to fitness, and if they exist, they are mere evolutionary noise. It seems to me to be more appropriate to say that strictly neutral alleles are absolutely noiseless.

If neutral variations are represented by amino acid changes in a protein, this means that such changes are equally acceptable for the working of the protein in the body. Furthermore, this equality need not be perfect for the neutral theory to apply. All one needs is that the resulting difference in fitness is very small as explained in section 3.4 previously. Lewontin (1974), in his book (see p. 197) asserts that if amino acid substitutions are completely indifferent to the action of natural selection, they are 'genetic junk'. Apparently, he confuses equal function with no function. I think that only genetically inert parts deserve to be called genetic junk.

As to the possibility that selectively equivalent mutations occur at the molecular level, I would like to emphasize the fact that in organisms, particularly in higher forms, physiological homeostasis is a well-developed buffer against internal and external environmental disturbances. So, fluctuation of environmental conditions by no means automatically implies fluctuation of Darwinian fitness of mutants. This important point is often overlooked by mathematical population geneticists working on problems of molecular evolution. The term 'fluctuation of environment' is routinely used among them to denote 'fluctuation of selection coefficients' without inquiring what biological basis there is for the relative fitness of alleles at thousands of loci to fluctuate conspicuously with certain duration (see for example, Matsuda *et al.*, 1978).

Some criticisms of the neutral theory come from a wrong definition of natural selection. Here I must emphasize that the term 'natural selection' should be used strictly in the Darwinian sense; it acts through differential survival and reproduction of the individual. Very often, the existence of detectable functional difference between two molecular forms is taken as evidence for the existence of natural selection. To prove that natural selection is involved, the survival rates and the fecundity must be investigated. This is often overlooked by biochemists when they discuss natural selection. It is unfortunate that term 'selection' is sometimes used in quite a meaningless way. For example, Zuckerkandl (1978) states that 'Positive selection does not *per se* confer biological significance upon a mutant.' I also think that some of the misunderstanding could be avoided by making clear distinction between positive and negative selection. The latter, which is concerned with the elimination of deleterious mutants, has little to do with gene substitution unless the mutant effect is extremely small – so small that the deleterious effect is overriden by random genetic drift (see section 3.4). It is known, since the great work of Muller in the early days of Drosophila genetics, that negative selection is the most common form of natural selection (see Muller, 1962; also Haldane, 1954). 'Stabilizing' or 'centripetal' selection at the phenotypic level reflects negative selection, whereas 'directional' selection usually involves positive selection (this topic will be treated in more detail and in quantitative terms

in chapter 6). The existence of negative selection does not contradict the neutral theory.

Also, the distinction between gene mutations at the individual level and substitution of mutant forms at the population level must be made clear. Only the substitutions of the mutant forms in the population are directly related to molecular evolution. In too many discussions, such distinctions are ignored. I would like to emphasize that for advantageous mutants, the rate of evolutionary mutant substitutions is greatly influenced by the population size and the degree of selective advantage in addition to the mutation rate, as I showed in the previous section. Yet, some biochemists who discuss molecular evolution pay attention only to the mutation rate.

Recently, some authors, such as Zuckerkandl (1976) and Gillespie (1977) claim that they produced 'selectionist' theories whose consequences are indistinguishable from the neutral theory. I believe, these claims are all based on mistaken theories. Zuckerkandl (1976) considers different component factors of fitness for an amino acid replacement, and assumes that most alleles are genetically 'sufficient' as long as they are present alone, but compete with each other when other alleles coexist. According to him, after a more advantageous allele becomes fixed by selection, its advantage is progressively reduced by subsequent fixations of alleles, making room for additional new alleles to become fixed. Zuckerkandl's model is elusive and difficult to quantify, but he seems to think that if the mutation rate is constant, the rate of mutant substitutions by natural selection under constant environment is constant. Clearly he overlooks the effect of population size on the rate of evolution through accumulation of advantageous mutations. In a large population, the number of advantageous mutations that appear each generation in the population is much larger than in a small population. Also, various mutations will not necessarily appear in such a sequence as he postulated.

In his 'random environment model', Gillespie (1977) assumes that selection coefficients fluctuate with a certain autocorrelation, but in his treatment he completely ignores the contribution of mutations. As shown by Takahata and Kimura (1979), if mutations are taken into account, Gillespie's conclusions on heterozygosity must be modified drastically. At any rate, all these pseudo-neutral theories contain a peculiar feature; they claim that natural selection acts in such a way that its consequences are indistinguishable from no selection. It may be possible to construct a selection model that mimics one consequence of a neutral model, but it seems most unlikely that a model can be constructed that mimics *all* consequences of neutrality. Consider the difficulty that mathematicians

have had in constructing deterministic processes that mimic random number sequences, or the difficulty that an unscrupulous experimenter would have in generating large amounts of fictitious data that no combination of statistical tests could detect.

Some authors, such as Gatlin (1976), criticize the neutral theory on the ground that the arrangement of nucleotide bases in the genome deviates from complete randomness. As pointed out by Kimura and Ohta (1977), functional constraints are consistent with neutral substitutions within a class of mutants. For example, if a group of amino acids are constrained to be hydrophilic, there can be random changes within the codons producing such amino acids, but the pattern of nucleotides will be non-random. There is, of course, negative selection against hydrophobic mutants in this region, but, as mentioned before, negative selection does not contradict the neutral theory. The neutral theory does not assert that *all* amino acids are equivalent at a certain site, only that the majority of evolutionary changes concern those mutants that are equivalent. In addition to functional constraint, there is the possibility that mutation rates are not equal among the four nucleotide bases. Furthermore, as shown by Koch (1971), there is also a possibility that base-pair substitution mutation rates are influenced by neighboring base pairs.

A well-known example of 'non-random' association of adjacent nucleotide bases occurs in vertebrates where the dinucleotide sequence CpG is less frequent than would be expected from the DNA base composition (see for example Jukes, 1978a), and this has been attributed to natural selection (Subak-Sharpe *et al.*, 1966). Recently, however, it has been shown that unequal mutation rates are mainly responsible for this phenomenon. According to Bird (1980), the observed shortage of the dinucleotide sequence CpG in the vertebrate DNA is related to DNA methylation; namely, mCpG is the major methylated sequence in animals and mCpG tends to mutate to TpG, thereby producing TpG and CpA after replication. This means that loss of two CpGs leads to the gain of one TpG and one CpA. Bird (1980) has shown that there is a strong correlation between CpG deficiency and level of DNA methylation among animals. The vertebrates have the highest levels of DNA methylation and at the same time show the most extreme CpG deficiency, while the insects have very poorly methylated DNA and display no significant CpG deficiency. Partially methylated echinoderm DNA lies between these two groups in CpG deficiency. He has also shown, using extensive data coming from various animal groups, that there is a remarkable correlation between CpG deficiency and TpG + CpA excess. These results support the suggestion of Salser (1978) that methylated

CpG is a mutational 'hot spot' but not the claim of Subak-Sharpe *et al.* (1966) that the CpG shortage in the vertebrate genome is due to natural selection.

There is thus no evidence that the non-randomness Gatlin observed was brought about by adaptive mutant substitutions rather than by functional constraints and negative selection. Similarly, and more recently, the observed inequality (or 'non-randomness') of synonymous codon usage has been mentioned as evidence against the neutral theory, but such an observation can also be explained by general constraints and negative selection. We shall consider the problem of non-random synonymous codon usage more in detail later (see section 7.5).

Some criticism and skepticism regarding the neutral theory of molecular evolution come from the misunderstanding that the neutral theory holds that *all* the mutations at the molecular level are strictly neutral. This was never my intent, and is certainly not what the theory purports. The existence of a great deal of negative selection and some positive selection is not incompatible with the neutral theory, which holds that *most* evolutionary *changes* are caused by random genetic drift of mutant alleles that are selectively nearly equivalent.

4

Molecular evolutionary rates contrasted with phenotypic evolutionary rates

4.1 Some features of phenotypic evolution

Pattern of evolution as exemplified by the history of vertebrates

Before we treat evolutionary rates at the molecular level, I shall briefly mention some features of phenotypic evolution as revealed by fossil records. This will help to show that the pattern of molecular evolution is quite different from that of phenotypic evolution.

Through paleontological studies in the past hundred or so years, much valuable information has been obtained on the evolutionary history since the Cambrian period (600–500 million years ago) when abundant fossil records became available. Particularly noteworthy is the elucidation of the history of vertebrates, a group to which we ourselves belong. (For authoritative accounts of this subject, see Romer, 1967, 1968 and McFarland *et al.*, 1979.) The history of vertebrates is also relevant when we discuss the evolutionary rate of globins in the next section. So I shall summarize some relevant aspects of vertebrate history.

The oldest known true vertebrates were jawless fish, the agnatha, whose fossils are found in the Ordovician period (500–440 million years ago; for geological periods, see Fig. 4.2). Bony scales of such animals are known from late in the Cambrian period. They ranged in size from several to fifty centimeters long, and had curious appearances, with unpaired fins. They probably lived in seawater, and sucked in water and its food contents with powerful muscular pulses; they filtered water through gills on each side behind the mouth. Their fossils are abundant in the Silurian period (440–400 million years ago). Their bodies are covered by armor plates, hence they are named the ostracoderms.

It is generally supposed that this armor developed as protection against

the then dominant predaceous arthropod called eurypterid (sea scorpion). The bony armor plate led to the development of internal bone, first around in the head region, and later, as internal, solid masses of bone, as seen in higher fishes. Considering the fact that without a bony skeleton the more advanced classes of vertebrates would have been impossible, Romer (1967) remarks that we must thank the eurypterid enemies of our early ancestors.

A new era in fish evolution was opened, when, late in the Silurian period, they developed jaws through the enlargement of a pair of bars which formerly supported gill slits. During the Devonian period (400–350 million years ago), much expansion and radiation of jawed fishes occurred. At the same time, jawless ostracoderms dwindled, and by the end of the Devonian, they became extinct; only a few of their descendants survive to this day in primitive and specialized forms as lampreys and hagfishes.

The next major step in vertebrate evolution was the conquest of the land, started by amphibians and completed by reptiles. Geological evidence suggests that during the Devonian period there were many shallow lagoons and that seasonal droughts were common. Under such conditions, fishes that could crawl out of a dried pool with the help of their fins and move to other pools with water must have had a better chance of survival. Thus, fins eventually developed into legs. Frequent droughts also put a premium on those who could use swim bladders to breathe air. A group of 'lobe fin' fishes called crossopterygians gave rise to the ancestors of land vertebrates. Although crossopterygians flourished during the Devonian period, they rapidly declined in numbers by the end of the Paleozoic. This was probably caused by loss in competition with amphibians who descended from them and who shared a similar environment.

It was for a long time believed that crossopterygians had become completely extinct. However, a few decades ago, a relict survivor belonging to fishes termed coelacanths was caught deep in the Indian Ocean and was named *Latimeria* – sensational news in science. This is probably the most famous example of a living fossil.

The first four-footed vertebrates (tetrapods) appeared in the Devonian period. They were amphibians who had to go back to water for reproduction where eggs were laid and developed as frogs do today. Among early amphibians, there were transitional forms, such as *Ichtyostega*, which had four short legs but retained a tail fin like a fish. From amphibians evolved reptiles that lay an amniote type egg protected by a hard shell. This new 'invention' made full terrestrial life possible. It is likely that this evolutionary change also arose as an immediate adaptation to seasonal drought; eggs in a pond must have died when it dried up. So, the laying of eggs on dry land probably occurred before the adults abandoned aquatic life completely.

By the Carboniferous period (350–270 million years ago) reptiles appeared. In order for animals to move from water to land, plants must have preceded them. The extensive development of vascular land plants occurred in the Devonian period, and in the following Carboniferous they flourished, forming great forests on the earth. This allowed a boom of insects toward the end of Carboniferous period, furnishing some of the foods for the great radiation of reptiles that occurred in the Permian period and the subsequent Mesozoic era.

The continents of the earth then were quite different from those which we know now. As first proposed by the German geophysicist Alfred Wegener more than half a century ago, all the land surfaces were once amassed as a single super continent called Pangea. Later this gradually split and drifted apart, and eventually the present continents were formed. Wegener's hypothesis of 'continental drift', which had long been treated with scorn and disbelief was shown to be essentially correct by recent studies of earth sciences.

Toward the end of the Paleozoic and by roughly 230 million years ago, Pangea was formed. Before the end of the Triassic period (225–180 million year ago) the dinosaurs became 'the undisputed masters of Pangea' as remarked by Calder (1973). A spectacular radiation of reptiles occurred during the Mesozoic era (225–70 million years ago). They include not only dinosaurs but also flying reptiles, marine reptiles and the ancestors of birds. Although most of the ruling reptiles (Archosauria) became extinct at the end of the Mesozoic, the birds were very successful and became a distinct class of vertebrates that survive today. Their elegant but energy-expensive life in the air protects them from predator enemies on the ground. Thus, whenever such enemies were absent, the birds have tended to abandon flight to become ground-dwellers.

During the Permian (270–225 million years ago), mammal-like reptiles called the therapsids appeared, and they dominated the land until the Triassic (225–180 million years ago). They then dwindled in number, and during much of the Mesozoic, they gave way to dinosaurs. It appears that efficient bipedal running, which developed in the ruling reptiles with highly predaceous habits, was more successful in competition over the improved quadrupedal gait of therapsids.

The first mammals appeared at the end of the Triassic, about 200 million years ago. Scant fossil evidence suggests that our Mesozoic, mammal ancestors were small, inconspicuous creatures like the mouse or rat, both in

size and habit. Presumably they were mostly nocturnal. The egg-laying mammals of Australia which are known as monotremes consisting of platypus (*Ornithorhynchus*) and spiny anteater (Echidna) represent a side-line which split off from the main mammalian line at very early stage (see Fig. 4.2). They owe their present survival to geographical isolation in an area where competition or predation by more advanced mammals has been absent.

For the main line of mammals, the Mesozoic was a time of trial, but a significant one, for the outstanding characteristics of the mammals, particularly intelligence and temperature-independent activity, must have been much improved by the struggle for survival under the reptilian tyranny. Romer (1967) remarked that 'As mammals, we owe a debt of gratitude to the dinosaurs for their unintended aid.'

Mammals are characterized by nursing the young and caring for them, so that the time needed for the development of brain is secured, and the completed brain is used efficiently during the life of the individual. No doubt the development of a better brain (the hardware for information processing) is the real cause of mammalian success. It is possible that discussion of K vs r selection (MacArthur and Wilson, 1967) as 'adaptive strategy' which is now so fashionable in mathematical ecology has little relevance when we consider the trend of mammalian adaptive evolution. Producing a small number of offspring and caring for them intensely to ensure their survival against the vagaries of environment is, I believe, the real trend dominating mammalian evolution in the long run.

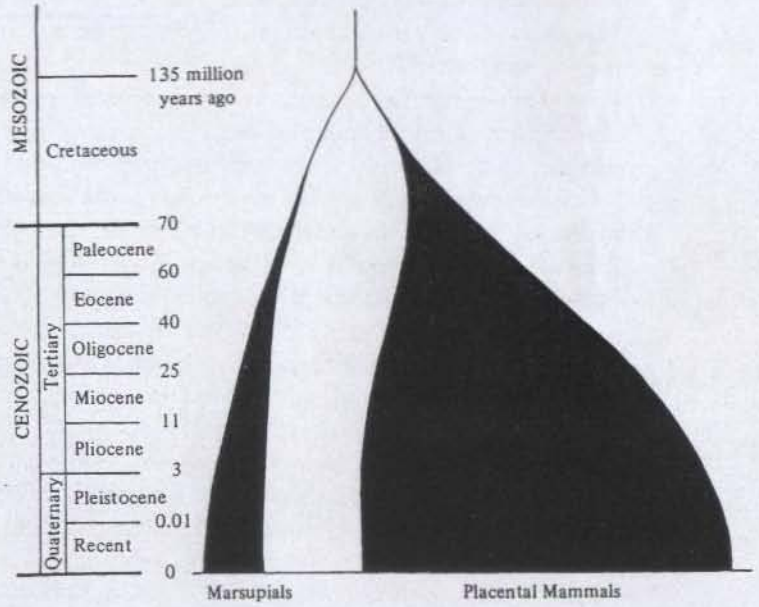
Toward the end of the Mesozoic all the dinosaurs, which had flourished until then, became extinct. The real cause of such large scale extinction of the great reptiles on land, in water and in air still remains as a major enigma in paleontology.

Recently, a bold hypothesis was proposed by Alvarez *et al.* (1980) to explain this Cretaceous-Tertiary (C-T) extinction some 65 million years ago, based on their observation that the concentration of iridium abruptly increased in a layer of clay deposited at precisely this period of time. According to their hypothesis, a large (estimated to be 10 ± 4 kilometers in diameter) asteroid hit the earth then, and a huge amount of dust was hurled by the impact into the stratosphere. The dust stayed in the air several years and was distributed worldwide. The resulting darkness suppressed photosynthesis, attacking food chains at their origin. The large herbivorous and carnivorous animals, including dinosaurs, became extinct. It appears that this dramatic hypothesis has since been strengthened by additional observations (see Kerr, 1980).

The mammals were not very diverse at the beginning, and their spectacular radiation starts with the Cenozoic era, filling various ecological niches previously occupied by the extinct reptiles (see Fig. 4.1 for subdivisions of the Cenozoic era). Of the two main groups of mammals, that is, marsupials (Metatheria) and placentals (Eutheria), the former have had only limited success. A truly amazing spread of mammals was achieved by the Eutheria, making the Cenozoic era the 'Age of Mammals' (for detailed account of this subject, see Kurtén, 1971). This is particularly evident in the Paleocene epoch, the first 10 million years of the Cenozoic era. According to Kurtén (1971), six, ten and fourteen new mammalian families arose in the early, middle and late parts of the Paleocene, of which only nine out of the last mentioned families became extinct by the end of the Paleocene; adaptive radiation occurred at an unprecedented pace.

In the plant kingdom, during the Cretaceous period, but well before dinosaurs became completely extinct, explosive evolution occurred in the flowering plants (angiosperms), producing various families which flourish today. Along with the mammals, flowering plants characterize the Cenozoic era.

Fig. 4.1. Mammalian expansion during the Cenozoic era. This figure is constructed from data given in Romer (1968) and Kurtén (1971).



Evolution of primates is obviously of great interest to us, for, as an order, they include man himself (They also include lemurs, monkeys and apes). Primates are predominantly tree-dwellers, and their common ancestor is believed to have been an animal similar to the tree shrews (*Tupaia*) which live in the East Indies today. Many features which characterize the primates are the products of life in trees. Prominent among them are hands with an opposable thumb which have good grasping power, eyes forwardly turned and endowed with well-developed sight, and an ability to gauge distance with great precision. The brain is highly developed in order to command these faculties. On the other hand, the sense of smell is much reduced. Fossil evidence from Ethiopia suggests that *Australopithecus*, which lived some four million years ago, represents a direct human ancestor.

Human evolution was definitely under way when life in the trees was abandoned, and an upright posture with bipedal gait was established by some three million years ago. It is suggested that in the late Tertiary times there was a process of reduction of forest area and replacement by grassy savannas. This forced the ancestors of man to become ground dwellers. Man didn't leave the trees; the trees left him.

As a creature, man was not a particularly strong one, and after he left the protection of the trees, he could only survive by cunning which was made possible by brain development. Some 2.0 million years ago, an advanced hominid with a brain size of 750 cm^3 was already making stone tools (Washburn, 1978; dating adjusted). This brain size is about half that of modern man (1400 cm^3). It is possible, as pointed out by Muller (1959), that there was positive feedback between the practice of using tools and improvement of genetic ability for better use of them. This helped natural selection to act efficiently, leading to rapid improvement of intelligence.

Despite a wealth of fossils that have recently been unearthed, the time of divergence between man and the African apes (chimpanzee and gorilla) is uncertain, and it has been much debated. The traditional paleontological studies based on comparison of anatomical structures (such as teeth) have placed the point of divergence somewhere between 20 and 30 million years ago. In particular, fossils designated *Ramapithecus* and *Sivapithecus*, dating back to 14 million years ago, seemed to require an early divergence (Simons, 1981). On the other hand through immunological studies on the albumin of various primate species, Sarich and Wilson (1967) obtained a result suggesting that the divergence occurred only about 5 million years ago. Their claim is based on the empirical rule that logarithmic immunological distance (log ID) serves as a 'molecular clock', that is, it increases at a constant rate with chronological time. Although they did not know the

reason, they took it for granted that the molecular clock is accurate for all cases including human and ape lines. More recently, King and Wilson (1975) compared forty-four enzyme loci between man and chimpanzee, and studied their genetic difference in terms of Nei's (1972) measure of distance. The new result also showed that man and chimpanzee are genetically very close, again suggesting a divergence time of the order of 5 million years ago. It can be counted something of a victory for molecular methods of dating that *Ramapithecus* and *Sivapithecus*, formerly thought to be ancestral to man, have now been related to the orang-utan (Pilbeam, 1982). There is no longer any gross contradiction between the molecular data and the appearance of hominids about 4 million years ago, in the fossil record.

The problem of the divergence time between human and ape lines was approached from a fresh angle by Kortlandt (1972). He suggested that the appearance of Rift Valleys in Africa and subsequent flooding by rivers and lakes caused isolation between the ancestor of the gorilla and chimpanzee in West Africa on the one hand, and the human ancestor in East Africa on the other. He dated the split some 14 million years ago, but it is possible that the mechanism he proposed operated ten million years later, because geological activity has been continuous in the Rift Valleys throughout the Pliocene and Pleistocene epochs. Divergence of the orang-utan (now living in forests in Asia) and the African apes, perhaps 10 million years ago, may be similarly related to the rifting that opened the Red Sea.

The history of vertebrate evolution summarized above clearly shows that ecologic opportunities play an essential role in rapid phenotypic evolution, and that progressive evolution is almost always brought about as a result of organisms' response to environmental challenge. Production of man from a primitive jawless fish in half a billion years is a remarkable example of progressive evolution, but we should not forget that degeneration and extinction are much more common in evolution. Haldane (1958) calls attention to the fact that probably for every case of progressive evolution in the sense of descendants being more complex in structure and behavior than their ancestors, there have been ten of regressive evolution. The main reason that evolution on the whole appears to be progressive is simply because a species that acquired a new capacity was more likely to give rise to various descendant species than one which lost some capacities.

As a feature of phenotypic evolution, one can not fail to mention its opportunistic nature. This is most clearly shown by the phenomenon known as convergence. A striking example of this is given by the adaptive radiation of marsupials in Australia as compared with the radiation of placental mammals. In the former, there are a marsupial anteater, mole-like

marsupials (the pouched 'mole') a dog-like one (the Tasmanian 'wolf'), squirrel-like ones including marsupial 'flying squirrels', and so on (see Simpson, 1949). A similar parallelism exists between North and South American mammals (including extinct forms). These examples show that similar ecologic opportunities were seized by animals of two different but related groups, producing forms with similar ways of life. That such similarities, however, are superficial has been revealed by detailed anatomical studies. It is evident that the underlying genetic structures of these two groups are also different.

Opportunism in evolution is an eloquent testimony that evolution at the level of form and function is largely determined by Darwinian natural selection that brings about adaptation of organisms to their environments. Positive natural selection cares little how such form and function are brought about genetically; it only cares whether the resultant phenotypes are good for the organism's survival and reproduction. This means that consideration of 'adaptive strategies' may be relevant for treating phenotypic evolution, but not necessarily for the genetic changes that underlie phenotypic evolution, and *a fortiori* not for molecular evolutionary changes.

Rates of phenotypic evolution

We continue to treat evolutionary change at the phenotypic level, but concentrate our attention on the problem of its rate (speed). The existence of striking differences in evolutionary rates among animal groups was discussed by Darwin (1859) in his book *The Origin of Species*. He points out (see chapter 10 of *The Origin*) that species of different genera and classes have not changed at the same rate, or in the same degree. For example, the Silurian *Lingula* differs but little from the living species of this genus, whereas most of the other Silurian molluscs and all the Crustaceans have changed greatly. He goes on to suggest that 'productions of the land seem to change at a quicker rate than those of the sea'. He noted also that there is reason to believe that organisms considered high in the scale of nature generally change more quickly than those that are low, though there are exceptions to this rule.

More recently, Rensch (1960), by investigating extensively the age of species and higher categories of various animals, concludes that evolution is slower in the more primitive marine groups than in higher terrestrial animals. As to the cause of such findings, he attributes this to the influence of environment, namely, marine animals are much less subject to periodic changes of important ecological factors, such as temperature, moisture, and

so on, than terrestrial ones. In addition, in most marine animals, population size is less likely to fluctuate, and effect of natural selection tends to remain constant.

Rensch notes that the appearance of a completely new structural type is often succeeded by a period of 'explosive' evolution (tachytely), and then, phases of slow change (bradytely) may preserve certain animal types for a very long period. He points out an interesting fact that the number of generations per unit time is not generally related to the rate of evolution. For example, in elephants, the rate of evolution was relatively fast, while in rodents with small body size, even though the number of generations per unit time was 40 to 100 times greater than that of elephants, evolution in these rodents is slow. (He attributes this slowness to a greater constancy of the environment and the larger size of the reproductive communities in the rodents.) As a similar example, we can add that human evolution at the phenotypic level has been very rapid despite the fact that his generation span is exceptionally long as compared with other animals.

Extensive and careful analyses of the rates of evolution based on fossil records has been made by Simpson (1944, 1953). He recognizes two types of rates, morphological and taxonomic rates. The morphological rate is concerned with how a morphological character or a character complex changes along a lineage ('phylum'). In this respect, it is a 'phylogenetic rate'. For example, we inquire how rapidly the lengths of teeth changed along the line leading to the present horse from its ancestor *Eohippus* that lived in the lower Eocene.

On the other hand, the taxonomic rate is concerned with the duration and abundance of a certain taxonomic unit. It may be phyletic in the sense that we inquire how rapidly a new genus appears and replaces the old one along a single lineage. For example, in the line leading to the present horse (*Equus*) starting from *Hyracotherium*, there are eight successive genera (we exclude *Equus* since its span is incomplete). The time covered is roughly 60 million years. This means that the average phyletic taxonomic rate is 0.13 genera per million years (or reciprocally, 7.5 million years per genus). This is within the range of 'standard' rates which Simpson called 'horoteric'. Another example is the evolution of Triassic and earlier ammonites. In this case, the rate is 0.05 genera per million years, much slower than in the horse evolution. It is possible, as pointed out by Simpson (1953), that a taxonomic difference based on the judgement of experienced workers is a more suitable indication of the totality of evolutionary changes than any collective measure of difference of morphological characters.

The taxonomic rate may also be concerned with changes in the numbers

or the frequencies of phyla (or other taxonomic groups). The rate of such changes may be called the taxonomic frequency rate. Although this rate is not directly related to the phenotypic evolutionary rate in the ordinary sense, I shall quote an example from Simpson (1953). He presents time-frequency curves involving the frequencies of the first and last appearances as well as total frequencies of ungulate (i.e. hoofed animal) families from latest Cretaceous until Pleistocene. From the curves, the rate of change in total frequency, the rate of origin (per million years) and the rate of extinction of families can be computed. His analysis shows that origination is highest at about the middle (i.e. around Eocene) of this period, while extinction is much more pronounced in the latter half of the period.

Returning to evolutionary change along individual lineages, there are some examples of very slow evolution. Particularly noteworthy are 'living fossils' (called 'immortals' by Simpson, 1949) that survive today and that show little change since their first appearance in the very remote past. Outstanding examples are: the sea shell *Lingula* which looks unchanged since its Ordovician ancestor more than 400 million years ago, the lizard-like reptile *Sphenodon* which has survived almost unchanged since the Jurassic, the opossum (unchanged since Cretaceous), and crocodiles (since early Cretaceous). These animals must have been well adapted to environments that have not changed for very long time. Negative selection has worked so efficiently on them to weed out mutations that, at the phenotypic level, the status quo has been maintained. On the other hand, there is no reason to believe that this is caused by reduced mutation rates. The fact that side branches (that later became extinct) arose once in a while, for instance, from opossums arose the carnivorous marsupials of South America, is consistent with this interpretation.

According to Simpson (1949), some general features common to many of the 'immortals' are as follows. Most of them are relatively small animals. They have rather large local populations, have some latitude in adaptation, and live in conditions that are relatively constant. Most of them were progressive animals when they arose and had then evolved rapidly into new adaptive types by first seizing opportunities as they arose; they adapted so well to their environmental conditions that later forms could not displace them.

Finally, I would like to mention the work of Haldane (1949) on quantitative measurement of the rate of evolutionary change ('morphological rates' in the terminology of Simpson, 1953). He claims that if we wish to compare the rates of evolutionary changes of different organs, or of the same organ in different genera, the most appropriate measure is in terms of

the percentage change. He suggests that if the mean length of a certain structure (say, a tooth length) has increased from x_1 cm to x_2 cm in the time t, then the rate of change per unit time can be estimated by

$$(\log_e x_2 - \log_e x_1)/t. \tag{4.1}$$

Then, he found, using the data of Simpson (1944), that in the evolution of the horse the height of certain teeth increased on the average at the rate of about 4×10^{-8} per year, or 4% per million years. He proposed the term *darwin* as a unit of evolutionary rate, representing a change in measurement at the rate of 10^{-6} per year, or equivalently, increase or decrease of size by a factor of e (the base of natural logarithm, i.e. 2.718 . . .) per million years. In these terms the horse rate is about 40 millidarwins. Using the same method, Haldane found that the body length of six suborders of dinosaurs increased during the Mesozoic era at the rate of about 2.6×10^{-8} or 26 millidarwins which is slightly less than the horse rate. If we apply Haldane's method to the increase of brain size in hominid evolution (as discussed in the previous subsection), we find that the rate is near 1/5 darwin, since the cranial capacity doubled in about 3 million years. This is an example of very rapid evolution at the phenotypic level. Most examples of rates faster than one darwin involve human intervention, such as happened to cause the very rapid changes in domestic animals and plants.

Since Simpson and Haldane discussed these issues, a great debate has developed among paleontologists about whether phenotypic evolution is gradual and essentially continuous, or proceeds in a punctuated fashion, with sudden 'speciation events' being followed by evolutionary stasis. (For a review of different opinions, see Schopf, 1977). If the punctuational mode of evolution occurs, then the rates of change must be much faster than Haldane's figures suggest, during speciation; thereafter, every species becomes a living fossil, in its resistance to change, until it is superseded. But at the time of writing the arguments about the tempo of change are still raging, and it is too early for a non-paleontologist to judge the likely conclusion.

4.2 Rate of evolution at the molecular level

Comparative studies of protein sequences

It had long been a dream of population geneticists to find out at what rate genes are substituted in the species in the course of evolution. Such a rate should characterize the speed of evolution more unambiguously than any other measure of evolutionary rates based on phenotypes.

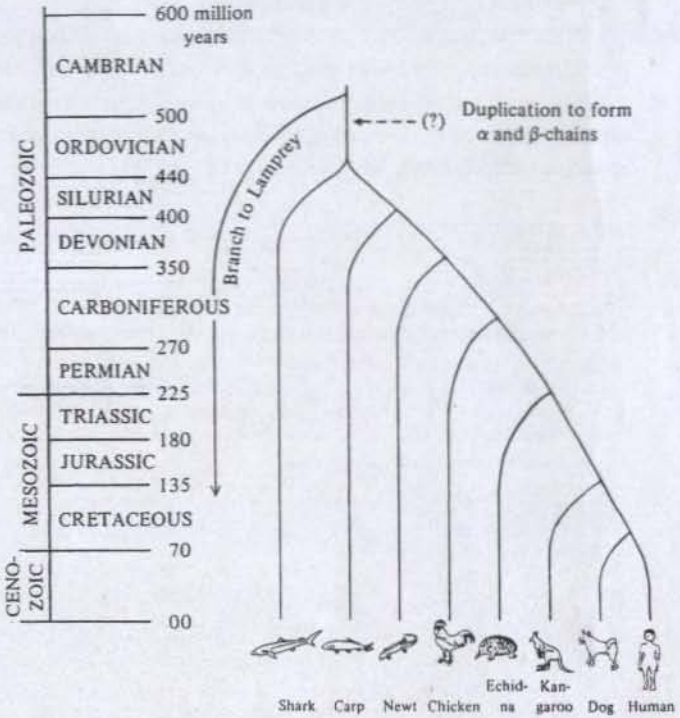
Simpson (1953) states that 'genetic rate' is the ideal measurement of the speed of evolution, although this was not then feasible. Later, this became possible by the development of molecular genetics which can treat genes in molecular terms. Specifically, comparative studies of amino acid sequence of proteins enabled us for the first time to obtain reliable estimates of the rates at which new mutant genes are incorporated into the species in evolution. As pointed out by Crow (1969), we can bypass classical genetic analysis by directly comparing proteins coded by genes between, say, mice and whales, even if they can not be crossed with each other.

During the past decade and a half, a large body of data have accumulated as shown by extensive compilation of protein sequences by Dayhoff (1969, 1972, 1976, 1978). Among various proteins, globins seem to have been studied most extensively (as well as intensively), although studies of cytochrome c sequences are nearly comparable. Particularly, globin sequences among vertebrates are well studied, and they serve as an excellent model system to discuss the evolutionary rate at the molecular level. Here we shall mainly be concerned with hemoglobin and myoglobin. Hemoglobin is the oxygen-carrier in the blood, carrying oxygen from the lungs to tissues, and also bringing the carbon dioxide back to the lungs. On the other hand, myoglobin has the function of storing oxygen in the tissue, and it exists as a monomer. These globins are essential for a vertebrate's life. In sharks and higher vertebrates, the hemoglobin molecule is a tetramer consisting of two identical α-chains and two identical β-chains. In the sea lamprey (*Petromyzon marinus*), however, hemoglobin in the blood exists predominantly as monomers (single chains), although six different genes seem to exist for it (Hendrickson and Love, 1971). The core of these molecules is the part known as the heme (containing an iron atom) to which the oxygen molecule attaches. For more details on the structure and action of proteins, readers are invited to consult an excellent introductory book by Dickerson and Geis (1969).

Let us compare the hemoglobin α-chain among vertebrates. In mammals, this chain consists of 141 amino acid sites, and there is no ambiguity in detecting homology among them. For example, if we compare the human and Japanese monkey, we find that their α chains exactly match each other except at four amino acid sites (Matsuda, 1978). Thus, the percentage amino acid difference between the two is only 2.8. Since each site of a protein sequence can be occupied by any one of 20 amino acids, the probability of two sequences matching each other even for a stretch of 10 consecutive amino acid sites by chance is very small (note that $1/20^{10} \approx 10^{-13}$). Of these four amino acid differences between man and Japanese monkey, three can

be interpreted by the code table as each being derived by a single mutational change (base substitution), while the remaining one requires at least two mutational steps. It is evident that the two α-chain genes have descended from a common ancestral gene through five or so base substitutions. If, on the other hand, we compare human with dog, their α-chains differ by 23 amino acid sites which amounts to 16.3% of the total. For still more remote

Fig. 4.2. Percentage amino acid differences when the α hemoglobin chains are compared among eight vertebrates together with their phylogenetic relationship and the times of divergence.



Shark	59.4	61.4	59.7	60.4	55.4	56.8	53.2
Carp		53.2	51.4	53.6	50.7	47.9	48.6
Newt			44.7	50.4	47.5	46.1	44.0
Chicken				34.0	29.1	31.2	24.8
Echidna					34.8	29.8	26.2
Kangaroo						23.4	19.1
Dog							16.3
Human							

α-chain

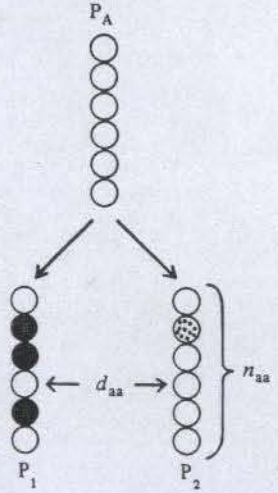
Percent Amino Acid Differences

comparisons, differences become larger. For example, human and carp α chain differ from each other at roughly 50% of the amino acid sites. In addition, these two chains are differentiated by insertions or deletions that amount to three amino acid sites. Fig. 4.2 shows the percentage amino acid differences in the pairwise comparisons among the α-chains of eight vertebrates, together with their phylogenetic relationship and the times of divergence. The table in this figure was constructed based on Alignment 32 of Dayhoff (1978), but it differs slightly from her Matrix 34; we excluded insertions and deletions when they exist between sequences, and we concentrated on replacements among the 20 kinds of amino acids.

It may be seen from Fig. 4.2 that there is a strong parallelism between the divergence time and the percentage amino acid differences. The parallelism becomes even more evident if we use, instead of mere percentage differences, the estimated number of amino acid substitutions that have actually occurred in the course of evolution. The estimates can be obtained as follows. To facilitate its later use, I shall explain the procedure in general terms.

Let n_{aa} be the total number of amino acid sites for which two homologous

Fig. 4.3. Diagram illustrating two homologous polypeptide chains P_1 and P_2 which descended from an ancestral chain P_A by accumulating mutations (solid and dotted circles); n_{aa} is the total number of amino acid sites, and d_{aa} is the number of amino acid differences. In this illustrated example, $n_{aa} = 6$ and $d_{aa} = 3$. Note that although four mutations have accumulated, differences can be detected only at three sites (we exclude cases in which two mutations lead to states which are not detectably different).



protein sequences are compared, and suppose that they differ from each other at d_{aa} sites (see Fig. 4.3). The fraction of amino acid differences which we denote by p_d is given by d_{aa}/n_{aa}. What we are aiming at is to estimate the number of amino acid substitutions that have actually occurred in evolution. If two mutations (amino acid replacements) occur at the homologous sites in two lineages, one leading to P_1 and another leading to P_2, or two mutations occur at the same site in one of the lines, then this is detected merely as one amino acid difference between P_1 and P_2. There may be cases in which the first mutation leads to a difference in amino acid but the second mutation cancels the difference, but we neglect such cases because such events are relatively rare (less than 5% of the total). To correct for superimposed mutations, a simple but very useful procedure is to assume that amino acid substitutions follow the Poisson law in statistics (Zuckerkandl and Pauling, 1965; Margoliash and Smith, 1965).

Let K_{aa} be the average number, per site, of amino acid substitutions between the two polypeptides compared. We assume that the probabilities of 0, 1, 2, ... amino acid substitutions occurring at a particular site are given by the Poisson series

$$e^{-K_{aa}} + K_{aa}e^{-K_{aa}} + \frac{K_{aa}^2}{2!}e^{-K_{aa}} + \ldots . \tag{4.2}$$

This assumption is based on the consideration that a substitution at each site is a very rare event for any given period, but extending over an enormous period, the probability of an occurrence is appreciable.

In particular, the probability that no substitutions occur at a site is $e^{-K_{aa}}$. By equating this to the fraction of amino acid sites for which two polypeptides are identical, that is,

$$e^{-K_{aa}} = 1 - p_d, \tag{4.3}$$

we obtain

$$K_{aa} = -\log_e(1 - p_d) \approx -2.3 \log_{10}(1 - p_d), \tag{4.4}$$

where $\log_e$ denotes the natural logarithm, and $p_d = d_{aa}/n_{aa}$ is the fraction of amino acid differences. Note that the table in Fig. 4.2 lists values of $100\,p_d$. The standard error of K_{aa} is given by

$$\sigma_K = \sqrt{\frac{p_d}{(1 - p_d)n_{aa}}} \tag{4.5}$$

(Kimura, 1969a). Strictly speaking, equation (4.3) is valid only when the rates of substitution are the same among different sites, but it is valid as a first approximation even if the rates differ, provided the average substitution rate is low (see Nei and Chakraborty, 1976).

The rate of evolutionary amino acid substitution per site per year is then given by

$$k_{aa} = K_{aa}/(2T), \tag{4.6}$$

where T is the number of years that have elapsed since the evolutionary divergence of the two chains from their common ancestor. Note that the factor 2 in the denominator corresponds to two branches in a phylogenetic tree involved.

For example, if we compare the α chain of man with that of the shark, we find that they have the same amino acid at 65 sites and a different amino acid at 74 sites. In addition, these two chains are differentiated by insertions or deletions (often called 'gaps' in sequence comparison) that amount to 11 amino acid sites. Thus $n_{aa} = 65 + 74 = 139$ and $d_{aa} = 74$ so that $p_d = 74/139 = 0.532$ and $K_{aa} = -\log_e(1 - p_d) = 0.760$ with $\sigma_K = 0.091$. Similarly, for the comparison of human and carp α chains, we find that $d_{aa} = 68$ and $n_{aa} = 140$ if we exclude 3 gaps. Then, we obtain $K_{aa} = 0.666$ and $\sigma_K = 0.082$.

In Fig. 4.4, the estimated number of amino acid substitutions per site (K_{aa}) is plotted against the time of divergence T as solid dots. For example, the point marked 1 represents the average value of K_{aa} obtained by comparing the α chain of the shark with those of the other seven animals. Similarly, the point marked as 2 represents the average K_{aa} obtained by comparing the carp α chain with those of the newt, chicken etc. From this

Fig. 4.4. Relationship between K_{aa}, the number of amino acid substitutions (ordinate) and the time of divergence in millions of years (abscissa).

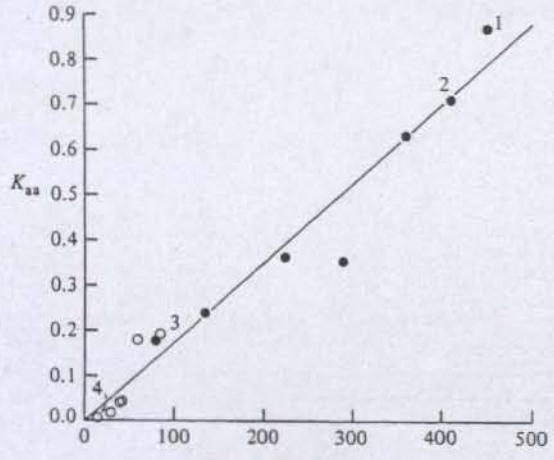
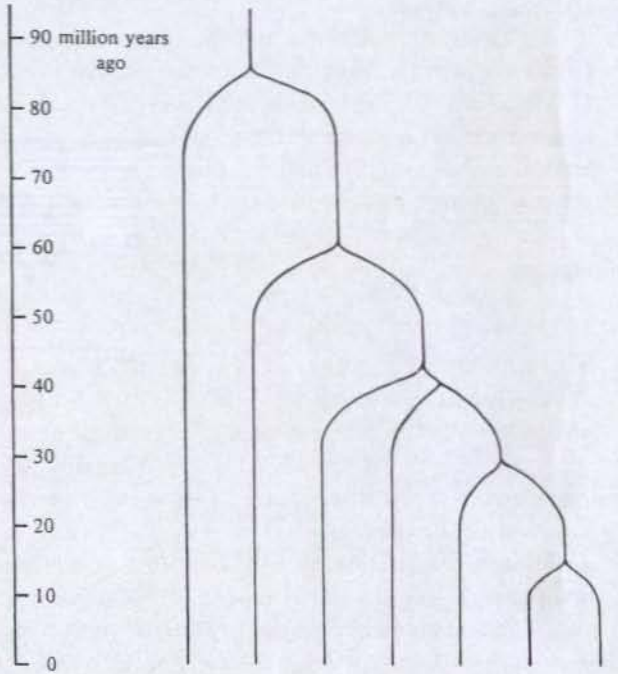


figure it is evident that these seven points are reasonably close to being on a straight line. The regression of K_{aa} on T (solid line in Fig. 4.4) has the slope 0.00179, which means $k_{aa} = K_{aa}/(2T) \approx 0.9 \times 10^{-9}$ per amino acid site per year. In other words, the rates of amino acid substitutions are roughly the same among diverse lineages of vertebrate evolution, and the representative

Fig. 4.5. A phylogenetic tree of primates (including the tree shrew) together with a table showing the number of amino acid differences when their hemoglobin β-chains are compared. Note that the β-chains in these animals consist of 146 amino acids. The numbers involving the tree shrews are based on the amino acid sequence of the β-chain determined by Maita *et al.* (1977).



β-chain — Number of amino acid differences

Tree Shrew	26	25	24	25	27	26
Brown Lemur		25	23	25	22	25
Spider Monkey			8	4	7	6
Japanese Monkey				5	6	7
Gibbon					3	2
Gorilla						1

value 0.9×10^{-9} is not very different from 10^{-9}. The only significant deviation observed is for the chicken vs. mammals comparison suggesting that the α-chain evolved more slowly in birds unless the divergence time is in error.

Fig. 4.4 also contains observed values of K_{aa} similarly computed for the hemoglobin β-chains of the primates (including the tree shrew) using data given in Fig. 4.5. They are plotted as open circles. Among them the circle marked as 3 represents the comparison of the tree shrew with the rest of the primates, while the one marked as 4 represents the human-gorilla comparison. They show that the primate β-chains evolved at the rates not very different from the value obtained for the vertebrate α, i.e. roughly at 10^{-9}/a.a. site/year.

One apparent anomaly is that the rate of substitution in the human-gorilla line is much lower than in other lines such as in the tree shrew line. Let us inquire, then, what is the probability that the number of substitutions happens to be one or less during 15 million years of separation (as in man and the gorilla), assuming that the intrinsic substitution rate is 10^{-9} per site per year for the β chain consisting of 146 amino acid sites. This probability is given by the first two terms in the Poisson series with the mean 146 K_{aa}, that is,

$$P = e^{-X}(1 + X), \tag{4.7}$$

where $X = 146\ K_{aa}$. Note that the sum of Poisson variables is again a Poisson variable. Putting $K_{aa} = 2Tk_{aa} = 2 \times 1.5 \times 10^{7} \times 10^{-9} = 0.03$, we obtain $P = 0.067$. In other words, it is expected to occur by chance every 15 trials. So, from this observation alone, we can not definitely say that the evolutionary rate is significantly lower in the human-gorilla line than in other vertebrate lines.

Similar estimates are obtained both for the α and β hemoglobin chains, when comparisons are made between six mammals, human, bovine, horse, dog, rabbit and mouse: $k_{aa} = (1.12 \pm 0.09) \times 10^{-9}$ for the α-chain and $k_{aa} = (1.28 \pm 0.14) \times 10^{-9}$ for the β-chain. It is assumed here that these animals diverged from each other 80 million years ago. In addition, from the comparison of myoglobin (consisting of 153 amino acids) between six mammals, human, harbor seal, badger, horse, bovine and sperm whale (see Table 4.4), we obtain, as the rate of substitution per site per year, $k_{aa} = (1.01 \pm 0.09) \times 10^{-9}$ which is very similar to the above estimates.

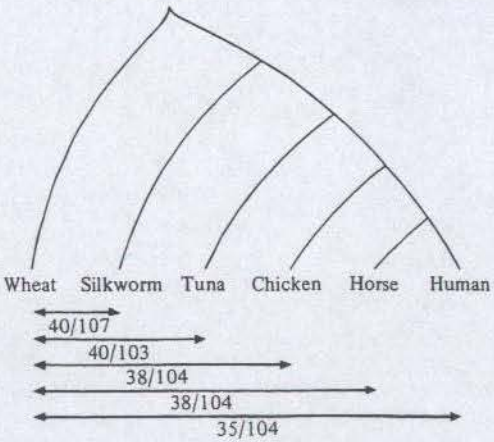
We can hypothesize then that globins evolved throughout the course of vertebrate evolution at roughly a constant (uniform) rate of 10^{-9} per site per year. As I shall discuss in more detail in the next subsection, constancy is

in terms of rate per year; generally speaking, generation time has little effect on the rate of amino acid substitutions.

Approximate constancy of evolutionary rate as well as the fortuitous nature of amino acid substitution in evolution are also apparent in cytochrome c. This molecule performs the important function of carrying electrons in the respiratory chain, and it is distributed more widely among eukaryotes (i.e. organisms with a true nucleus) than hemoglobin. Fig. 4.6 shows the fraction of amino acid differences of cytochrome c of five animals as compared with that of wheat. Note that, when compared to wheat, the various animals have differentiated to about the same extent. In this case, however, the rate of amino acid substitution per year is about one third that of hemoglobin.

Table 4.1 lists values of k_{aa} for several proteins. The highest rate is represented by fibrinopeptides ($k_{aa} = 8.3 \times 10^{-9}$ in this table), while the lowest rate is represented by histone H4 ($k_{aa} = 0.01 \times 10^{-9}$). At the moment, we do not know for sure how representative these figures are for all cistrons (genes) in the mammalian genome. It is possible that the frequency distribution of these rates does not follow a normal distribution, so the median is more suitable as a representative value than the arithmetic mean. It turns out, however, that the median of the evolutionary rates of twenty-one proteins listed in Table 6.1 of McLaughlin and Dayhoff (1972) is 1.3×10^{-9} per amino acid site per year (Kimura, 1974). Similarly, from Table 1 of Dayhoff (1978) who lists evolutionary rates for sixty proteins, we

Fig. 4.6. A phylogenetic tree and the fraction of amino acid differences in cytochrome c. In counting the number of amino acid differences, unmatching parts ('gaps') due to insertions or deletions were excluded.



find that the median rate is 0.74×10^{-9}/a.a./yr. This may be compared with 1.6×10^{-9}/a.a./yr which was obtained by King and Jukes (1969) as the average rate of seven proteins. On the whole we can say that, roughly speaking, 10^{-9} per site per year is the standard rate of molecular evolution in terms of amino acid substitution, and that vertebrate globins are representative. I suggested (Kimura, 1969a) that we use 10^{-9}/a.a./yr as a unit of evolutionary rate, calling it the *pauling*. In terms of this, the hemoglobin rate is very near to one *pauling*, while the cytochrome c rate is about 30 *centipaulings*. Unfortunately, this term has not gained popularity among students of molecular evolution, as is also the case for the term *darwin* among students of phenotypic evolution. As compared with more popular designations such as 'PAMs per 100 million years' meaning 'accepted point mutations per 100 residues per 100 million years' (see p. 3 of Dayhoff, 1978), or the 'Unit Evolutionary Period', or UEP in short, meaning 'the time in millions of years for one per cent change in amino acid sequence to show up between two divergent lines' (Dickerson, 1971), expressions in *pauling* units are much more simple and straightforward, and I hope that the unit will be used more widely. Perhaps, a less personality-related designation such as

Table 4.1. *Evolutionary rates in terms of amino acid substitutions. These rates are based mostly on data from mammalian order. They are expressed per amino acid site per year taking* 10^{-9} *as the unit ('pauling')*

Proteins	$k_{aa} \times 10^9$/yr
Fibrinopeptides	8.3
Pancreatic ribonuclease	2.1
Lysozyme	2.0
Hemoglobin α	1.2
Myoglobin	0.89
Insulin	0.44
Cytochrome c	0.3
Histone H4	0.01

These values are quoted from Dayhoff (1978), except for the lysozyme rate which is derived from Wilson *et al.* (1977) using the relation k_{aa} (in paulings) = 5/UEP (in millions of year), the fibrinopeptide rate which is from Ohta and Kimura (1971a) who treat rapidly evolving parts of fibrinopeptides A and B, and the cytochrome c rate which is from McLaughlin and Dayhoff (1972).

'Molecular Evolutionary Unit', or MEU for short (rather than *pauling*) might be preferable.

The relationship, as given by equation (4.4) between the observed number of amino acid differences in two homologous sequences and the estimated number of amino acid substitutions that might have occurred in the course of evolution, should be a good approximation, as long as the fraction of amino acid differences is not very large (say, less than 50%). A more accurate relationship between these two quantities was given by Dayhoff (see Table 23 on p. 351 of Dayhoff, 1978) in the form of a conversion table. She obtained it by multiplying a 'mutation probability matrix' for various evolutionary distances. It takes into account the fact that some amino acid pairs tend to replace each other more frequently in evolution than expected on the random basis. As compared with Dayhoff's relationship, equation (4.4) underestimates the number of amino acid substitutions, especially when the evolutionary distance is remote.

The following empirical formula which is a simple modification of equation (4.4), however, brings the relationship between K_{aa} and p_d quite close to Dayhoff's table.

$$K_{aa} = -\log_e\left(1 - p_d - \frac{1}{5}p_d^2\right) \tag{4.8}$$

As seen from Table 4.2, this formula is valid for p_d up to about 0.7. In these estimations, synonymous nucleotide substitutions that do not lead to amino acid changes are not taken into account. If nucleotide substitutions

Table 4.2. *Relationship between the fraction of observed amino acid differences* (p_d) *and the estimated number of amino acid substitutions* (K_{aa})

	K_{aa}		
p_d	Dayhoff[a]	Equation (4.4)	Equation (4.8)
0.01	0.01	0.01	0.01
0.05	0.05	0.05	0.05
0.10	0.11	0.11	0.11
0.20	0.23	0.22	0.23
0.30	0.38	0.36	0.38
0.40	0.56	0.51	0.57
0.50	0.80	0.69	0.80
0.60	1.12	0.92	1.11
0.70	1.59	1.20	1.60
0.80	2.46	1.61	2.63

[a] From Table 23 of Dayhoff (1978).

were equally frequent at all three positions in a codon, synonymous nucleotide substitutions should amount to roughly 25% (Kimura, 1968b), but in reality synonymous substitutions are much more frequent as will be explained later.

In addition, there are a few other methods, such as Fitch's (1971c) minimum phyletic distance method, which were devised for the purpose of estimating the number of nucleotide substitutions from amino acid sequence data. As shown by Nei and Chakraborty (1976), however, there is a high correlation among these estimates, and in practice any of them may be used for constructing evolutionary trees or relating nucleotide substitutions to evolutionary time.

Constancy of molecular evolutionary rates

As compared with the evolution at the phenotypic level, molecular evolution has the remarkable feature that the rate of evolution is approximately constant (i.e. the same) throughout diverse lineages for a given protein. We have already noted this remarkable phenomenon (see previous subsection), but since this has much bearing on the neutral theory, I shall examine this problem in more detail in this subsection.

Table 4.3 lists observed numbers of amino acid differences between six mammals (human, mouse, rabbit, dog, horse, bovine) when their hemoglobin α-chains are compared. It is likely that these six mammals diverged from each other late in the Mesozoic some 80 million years ago, and that they have accumulated mutations in their genes coding for the α hemoglobin independently since their divergence. We shall inquire whether the intrinsic rates of amino acid substitutions among the six lineages are

Table 4.3. *Observed numbers of amino acid differences between six mammals when their hemoglobin α-chains are compared. The ratio of the observed to the expected variances turns out to be as follows:* $R = 1.26$

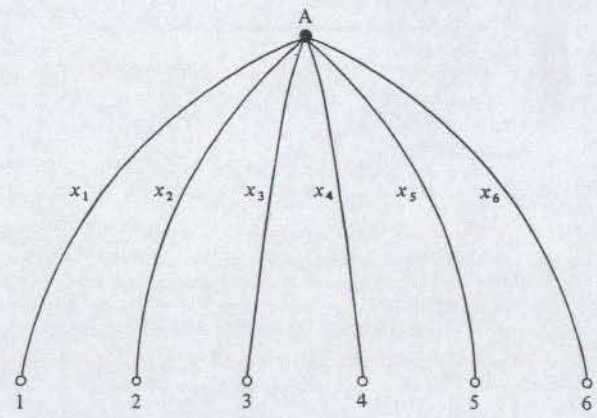
Human	Mouse	Rabbit	Dog	Horse	Bovine
Human	18	25	23	18	17
Mouse		27	25	24	19
Rabbit			28	25	25
Dog				27	28
Horse					18
Bovine					

equal and whether variation of the observed numbers of substitutions as shown in Table 4.3 lie within the limits of normal statistical fluctuation.

Let $x_1, x_2, \ldots, x_n$ be the number of amino acid substitutions in n lineages since divergence from their common ancestor A (see Fig. 4.7, where $n = 6$). We denote the mean and variance of the x's by $\bar{x}$ and V_x. Further, let $\bar{D}$ and V_D be the mean and variance of the number of amino acid differences when n species are compared and therefore $n(n-1)/2$ comparisons are made. It can then be shown that $\bar{x} = \bar{D}/2$ and $V_x = (n+1)V_D/[2(n-1)]$. From Table 4.3, the mean and the variance of the 15 observed values (after Poisson correction using equation (4.4), and noting $n_{aa} = 141$) turn out to be as follows: $\bar{D} = 25.34$ and $V_D = 22.81$. Therefore, we obtain $\bar{x} = 12.67$ and $V_x = 15.97$.

In order to see whether the variation of the number of amino acid substitutions among lineages is larger than what is expected by pure chance, an appropriate quantity is $R = V_x/\sigma_x^2$, where σ_x^2 is the theoretically expected variance. Thus, R is the ratio of the observed and the expected variances. Since, the x_i's follow the Poisson distribution, and since for this distribution the variance is equal to the mean, we may put $\sigma_x^2 = \bar{x}$, giving, in this case, $R = 1.26$. The statistic $(n-1)R$ should follow the χ^2 (chi-square) distribution with $n-1$ degrees of freedom. In the present case $n = 6$ and $(n-1)\,R = 6.3$. From a table of the χ^2 distribution we find that deviation of R from unity is not statistically significant. Clearly, we cannot reject the hypothesis that the intrinsic rates of amino acid substitutions in these six

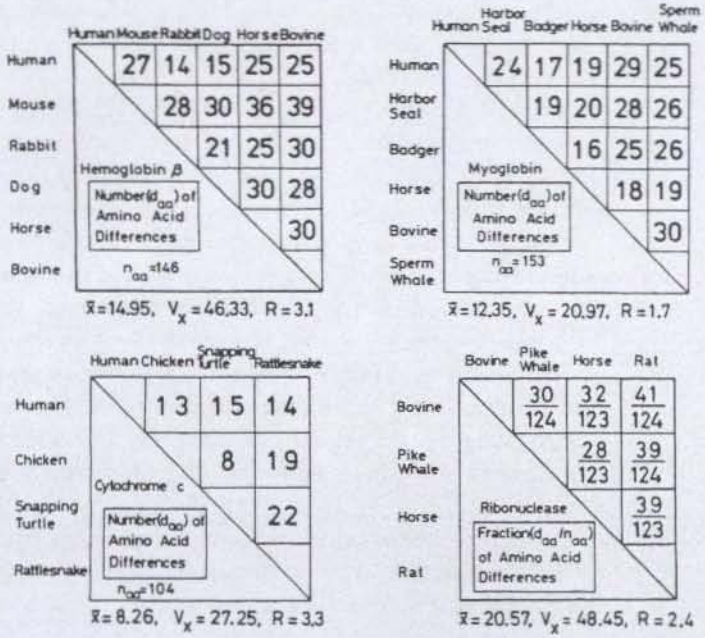
Fig. 4.7. Divergence of six mammals (see Table 4.3) from their common ancestor A. Letters x_1, x_2, etc. denote the number of amino acid substitutions in the individual lineages. For the data in Table 4.3, the mean and variance of x_i's are $\bar{x} = 12.67$ and $V_x = 15.97$.



lineages are the same. Note that if the point of divergence of these six lineages is not exactly the same, R tends to overestimate the true variance ratio, so that the present test for constancy is conservative.

Similar tests can be made for hemoglobin β, myoglobin, cytochrome c and pancreatic ribonuclease as shown in Table 4.4. For each of these molecules, the animals involved are considered to have diverged from a common ancestor at about the same time and therefore they have evolved separately for the same length of time. It turns out that $R = 3.1$ for hemoglobin β, $R = 1.7$ for myoglobin, $R = 3.3$ for cytochrome c and $R = 2.4$ for ribonuclease. Of these four cases, the deviation from unity is significant

Table 4.4. *Test for the uniformity of evolutionary rates among lineages within molecules (similar to the one explained in the text for the hemoglobin α) applied to hemoglobin β, myoglobin, cytochrome c and pancreatic ribonuclease. Data are taken from Dayhoff* (1978). *Note that the lengths of polypeptide chains are $n_{aa} = 146$ for hemoglobin β, $n_{aa} = 153$ for myoglobin, and $n_{aa} = 104$ for cytochrome c. For ribonuclease, $n_{aa} = 123$ or 124 depending on the comparison, because gaps (i.e. insertions or deletions) occur in sequence comparisons, and these gaps are excluded in computing fractions of amino acid differences*



for hemoglobin β (at about 1% level) and cytochrome c (3% level), but is not significant for the remaining two cases. These results suggest that although the strict constancy may not hold, yet a rough constancy of the evolutionary rate for each molecule among various lineages is a rule rather than an exception. Note that the rates of evolution may be different between different molecules. In fact, in the present case, the rate of evolution of pancreatic ribonucleases among mammals ($k_{aa} = 2.07 \times 10^{-9}$) turns out to be about eight times larger than the evolutionary rate of cytochrome c among vertebrates ($k_{aa} = 0.26 \times 10^{-9}$). The average value of R for the above five molecules (including hemoglobin α) is $\bar{R} = 2.6$. This is consistent with the result of our previous analysis (Ohta and Kimura, 1971b) in which we showed that the observed variances of the evolutionary rates for hemoglobins and cytochrome c among mammalian lines are roughly 1.5 to 2.5 times as large as the variance theoretically expected, if the variation is due to pure chance.

Let us investigate if such an approximate constancy also holds for more remote comparisons. We choose to examine the divergence between the α and β chains of hemoglobin. It is clear that these two chains originated by gene duplication which occurred in the remote past before the teleost appeared. In the second and third columns of Table 4.5, we compare the divergence between the α and β hemoglobin chains of man and the divergence between the α chain of the carp and the β chain of man. This

Table 4.5. *Comparison of amino acid differences between α and β hemoglobins*

Type of change[a]	Amino acid differences: Human α vs. Human β	Carp α vs. Human β	Platypus α vs. Platypus β	Shark α vs. Shark β
0	62	61	62	50
1	55	49	49	56
2	21	29	28	32
3	0	0	0	1
Gap	9	10	9	11
Total	147	149	148	150

[a] Numbers 0, 1, 2 and 3 respectively denote the type of amino acid differences that can be interpreted from the genetic code table as due to a minimum of 0, 1, 2 and 3 nucleotide substitutions (see Table 4.6) in these four sets of comparisons involving the α and β hemoglobin chains. The number of gaps (expressed as equivalents of the number of amino acids) is also listed for each comparison.

table is constructed based on Alignment 32 of Dayhoff (1978) with the help of Table 4.6 where the minimum number of nucleotide substitutions required to convert one amino acid to another is listed for 20 × 20 amino acid combinations.

Since the human and carp α chains differ from each other at about 50% of the amino acid sites (see the table at the bottom of Fig. 4.2), the data suggest that the two genes coding for the α and β chains have diverged independently from each other and to the same extent in the two lineages since their origin by gene duplication. It is remarkable that mutant substitutions have occurred at practically the same average rate in two separate lineages that have evolved independently for nearly half a billion years. For these comparisons, the rate of amino acid substitutions turns out to be about 0.81×10^{-9} if we use equation (4.4) (Poisson correction) or

Table 4.6. *The minimum number of nucleotide base substitutions ('mutational steps') required to convert one amino acid to another. For example, to change from threonine to valine, at least two changes, A → G in the first position and C → U in the second position of the codon, are required*

	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Ala A	0	2	1	1	2	1	2	2	2	2	2	2	1	2	2	1	1	1	2	2
Cys C	2	0	2	3	1	1	2	2	3	2	3	2	2	3	1	1	2	2	1	1
Asp D	1	2	0	1	2	1	1	2	2	2	3	1	2	2	2	2	2	1	3	1
Glu E	1	3	1	0	3	1	2	2	1	2	2	2	2	1	2	2	2	1	2	2
Phe F	2	1	2	3	0	2	2	1	3	1	2	2	2	3	2	1	2	1	2	1
Gly G	1	1	1	1	2	0	2	2	2	2	2	2	2	2	1	1	2	1	1	2
His H	2	2	1	2	2	2	0	2	2	1	3	1	1	1	1	2	2	2	3	1
Ile I	2	2	2	2	1	2	2	0	1	1	1	1	2	2	1	1	1	1	3	2
Lys K	2	3	2	1	3	2	2	1	0	2	1	1	2	1	1	2	1	2	2	2
Leu L	2	2	2	2	1	2	1	1	2	0	1	2	1	1	1	1	2	1	1	2
Met M	2	3	3	2	2	2	3	1	1	1	0	2	2	2	1	2	1	1	2	3
Asn N	2	2	1	2	2	2	1	1	1	2	2	0	2	2	2	1	1	2	3	1
Pro P	1	2	2	2	2	2	1	2	2	1	2	2	0	1	1	1	1	2	2	2
Gln Q	2	3	2	1	3	2	1	2	1	1	2	2	1	0	1	2	2	2	2	2
Arg R	2	1	2	2	2	1	1	1	1	1	1	2	1	1	0	1	1	2	1	2
Ser S	1	1	2	2	1	1	2	1	2	1	2	1	1	2	1	0	1	2	1	1
Thr T	1	2	2	2	2	2	2	1	1	2	1	1	1	2	1	1	0	2	2	2
Val V	1	2	1	1	1	1	2	1	2	1	1	2	2	2	2	2	2	0	2	2
Trp W	2	1	3	2	2	1	3	3	2	1	2	3	2	2	1	1	2	2	0	2
Tyr Y	2	1	1	2	1	2	1	2	2	2	3	1	2	2	2	1	2	2	2	0

0.96×10^{-9} if we apply equation (4.8) (Dayhoff correction) assuming that the α–β duplication occurred about 5×10^{8} years ago. Since the comparisons are remote, it is possible that the Dayhoff correction is more appropriate than Poisson correction for them. At any rate, these estimates of amino acid substitutions roughly agree with 0.9×10^{-9}, the rate which we obtained in the previous subsection by comparing the α hemoglobin sequences among vertebrates. Similar comparisons of the α and β chains can be made for other animals, such as α-2 vs. β in chicken and α vs. β in the platypus, giving very similar results.

These observations show that the rate of molecular evolution in terms of mutant substitutions is constant per year, and, surprisingly, this does not seem to depend on such factors as generation time, living conditions, and the population size. As pointed out in the previous chapter, for selectively neutral mutants, the rate of evolution in terms of mutant substitutions is equal to the mutation rate per gamete, and it is independent of the population size (see equation 3.20). Thus the observed constancy can be explained most easily by assuming that the majority of amino acid substitutions is caused by random fixation of selectively neutral mutations, and that the rate of occurrence per gamete of neutral mutations is constant per year. In this connection, I pointed out earlier (Kimura, 1969a) that if hemoglobins and other molecules of 'living fossils' are shown to have undergone as many DNA base (and therefore amino acid) substitutions as corresponding genes (proteins) in more rapidly evolving species, this supports my neutral theory.

Recently, the amino acid sequences of the β chain and the α chain of the principal hemoglobin of the Port Jackson shark have been determined (Fisher *et al.*, 1977). According to Romer (1968), this shark is a relict survivor of a type of ancestral shark which had numerous representatives in the late Paleozoic days, notably in the Carboniferous period (270–350 million years ago). So, this shark is well entitled to be called a living fossil. The right-most column of Table 4.5 lists the result of comparison between the α and β chains of the Port Jackson shark. If we compare this with the corresponding result for the human, it is clear that the genes coding for the α and β chains in the shark have diverged to roughly the same extent (or slightly more) as have the corresponding two genes in man, by accumulating random mutations since the origin of the α and β chains by duplication.

Another example of a living fossil is the opossum. To test my prediction on the rate of molecular evolution in living fossils, Stenzel (1974) sequenced the α chain of this animal. From his data (see Table 2 of Stenzel, 1974), it

turns out that the α chain of the opossum evolved at the rate $k_{aa} = 1.7 \times 10^{-9}$ per year since it diverged from the placental mammals some 135 million years ago. Thus, Stenzel (1974) comments that 'this result exceeds the expectation of the neutral mutation theory'. I would like to point out that the primary structure of the opossum hemoglobin α as sequenced by Stenzel has quite an unusual feature. It is known (Jukes, 1971) that, in vertebrates, amino acid positions 58 and 87 of the α hemoglobin are invariably occupied by *His* (which is bound to heme), yet in the opossum, *Gln* rather than *His* occurs at position 58. This could be called a pathological hemoglobin in the ordinary standard. It is possible that the opossum α hemoglobin lost some selective constraints due to unknown causes (such as gene duplication) in the past history of its evolution, with the result that it started to accumulate mutations at an unusually high rate. If so, this is similar to the rapid amino acid substitutions found in guinea pig insulin associated with the loss of the usually invariant histidine B10 (to be discussed in the next chapter). The myoglobin sequence of the opposum is also known (see Alignment 33 in Dayhoff, 1976), and for this molecule, there is nothing unusual in its sequence. The rate of evolution turns out to be $k_{aa} = 0.6 \times 10^{-9}$ per year. As compared with the standard value, 0.9×10^{-9} for this molecule, the rate of myoglobin evolution in the opossum is slightly low, but not by very much.

Constancy of the evolutionary rate is one of the most controversial subjects in molecular evolution, particularly when this is interpreted in the light of the neutral theory. In fact, the claim of constancy as well as its neutralist interpretation have been criticized by many authors. Conspicuous among them is Lewontin who declares in his book *The Genetic Basis of Evolutionary Change* (1974; see p. 229) that 'the claimed "constancy" is simply a confusion between an average and a constant'. He goes on to suggest that it is like maintaining that the temperature never changes in Chicago because the total number of degree days measured there was the same in the last fifty years of the nineteenth century as it was in the first fifty years of the twentieth. Earlier, Stebbins and Lewontin (1972) said that 'the so-called "constancy" of rates over millions of years is nothing but the law of large numbers'. This type of criticism has become popular (see, for example, Powell, 1975; Spiess, 1977), and was even quoted in a textbook on evolution by Grant (1977) who dismisses my claim of the constancy of molecular evolution as 'elementary fallacy'.

I would like to point out that Lewontin's criticisms only come from his misunderstanding of the real nature of molecular evolution. Take for example, the hemoglobin α chain. As we have seen already, one amino acid

substitution occurs on the average roughly every seven million years along a lineage. It is clear that in order to get a meaningful estimate of the intrinsic evolutionary rate we must observe at least several million years, preferably much longer. The situation is somewhat similar to an accident statistics in actuarial science. In such a situation, 'the law of small numbers' (the Poisson law) rather than the law of large numbers is appropriate. What is really important, however, is that comparisons involve two or more distinct lineages, such as shown in Fig. 4.2, 4.4 and 4.5 or Tables 4.3 and 4.4, and we are trying to estimate the intrinsic rates of evolution in different lineages to make comparisons. Just as death rates do not become equal between man and an insect by merely averaging over a long period or a large number of individuals, there is no reason to expect that two averages converge toward each other unless their intrinsic parameters are the same. My point is that the intrinsic evolutionary rates are essentially determined by the structure and function of the molecules but not by environmental conditions.

In discussing the present problem, we must make clear the distinction between the observed evolutionary rate and the intrinsic parameter underlying it. Let us consider a gene coding for the α hemoglobin and see how mutant substitutions causing amino acid changes proceed in the course of evolution in a mammalian species. If the species concerned consists of half a million individuals per generation ($N = 0.5 \times 10^6$) and if the generation span is two years, then the total number of mutations that will occur at this locus in the course of 7 million years in the species is 3.5 million, if we assume the mutation rate per generation per gamete of 10^{-6}. Note that, of these 3.5 million mutations, only one succeeds in becoming fixed in the species. Admittedly, there are number of recurrent mutations, and also not all mutations are selectively neutral, but these are not important to the issue.

As soon as we realize this, and regard the process of amino acid substitutions as stochastic, it should not be difficult to understand that the intrinsic parameter of the rate of substitution and the observed frequency of substitutions are two different things. The number of observed substitutions during a certain period, say an interval of 10 million years, must fluctuate greatly as we go down the long history of the evolution of mammals, but this does not necessarily mean that the intrinsic rate of evolution itself fluctuates. In some period, no substitution may occur, and in another period of the same length two or more substitutions may occur by chance. This may be compared with a sequence of trials each of which consists of throwing 20 coins at random. The event of all 20 coins showing face up simultaneously occurs roughly once every one million trials,

although the intrinsic probability of occurrence of such an event remains the same all the time.

That the molecular evolutionary rate is constant (or nearly constant) in physical time (years) was referred to as 'existence of molecular evolutionary clock' by Zuckerkandl and Pauling (1965). Since then the term molecular clock has been very popular in the field of molecular evolution.

Enthusiastic support for the concept was expressed by Wilson *et al.* (1977) who wrote that 'the discovery of the evolutionary clock stands out as the most significant result of research in molecular evolution'. On the other hand, these authors do not consider the neutralist-selectionist controversy very important as the basis of the clock. Referring to the controversy, they say that although some workers regard the resolution of the problem as the major one confronting molecular evolutionists, 'we disagree'. They are interested in the clock itself and in using it to establish evolutionary mileposts, not its mechanics. By the way, Fitch and Langley (1976) write that 'the clock is erratic from moment to moment but satisfactory if enough substantial ticks are counted'. It seems to me that they share the same fallacy as Lewontin in claiming that 'two mean value estimates can be equally good' if the sample is large.

The constancy of the evolutionary rate in terms of amino acid substitutions is a scientific hypothesis, and to agree or disagree one has to resort to observational facts and their statistical analysis rather than to rhetorical arguments. This was why we investigated (Ohta and Kimura, 1971b) the variance of evolutionary rates among lines. We obtained a result suggesting that the observed variance is slightly larger than the variance theoretically expected, if the variation is due to pure chance. More elaborate statistical analyses were carried out by Langley and Fitch (1974). Using a maximum likelihood procedure, and incorporating the assumption of minimum evolution, they analysed the variation of evolutionary rates among the branches of a phylogenetic tree involving simultaneously the evolution of the α and β hemoglobins, cytochrome c, and fibrinopeptide A. They found that the variation of evolutionary rates among branches ('among legs over proteins') is significantly higher than expected by pure chance, with a χ^2 value about twice its degrees of freedom. Since the expected value of χ^2 is equal to its degrees of freedom, their results mean that variation of evolutionary rate in terms of mutant substitutions among lines is about twice as large as that expected from chance. Then they use this result to emphasize non-constancy of the molecular evolutionary rate and conclude that 'a large proportion of the observed substitutions were selectively fixed and can be expected to reflect environmental changes.

There is no reason to expect relative rates of selective substitution to be constant in time or among lines of descent.' In other words, they reject neutral mutations as the major cause of substitutions. At the same time, however, they showed that when the estimated numbers of mutant substitutions are plotted against the corresponding geological dates, they are beautifully on a straight line (similar to our Fig. 4.4). This suggests uniformity of the evolutionary rates. The only exceptions are three observed points for primates. In my opinion, emphasizing local fluctuations as evidence against the neutral theory, while neglecting to inquire why the overall rate is intrinsically so regular or constant is picayunish. It is a classic case of 'not seeing the forest for the trees'. The remarkable constancy of evolutionary rates per year at the molecular level has also been shown by Wilson and his group using an immunological method involving albumin over a wide range of vertebrate species (see Wilson *et al.*, 1977).

An exaggerated claim of non-constancy was made by Goodman *et al.* (1974, 1975) in their study of globin evolution. They maintain that the mutant substitutions occurred at a very high rate in the early stage of globin evolution particularly soon after duplication to form myoglobin, α and β hemoglobins, and that this was followed by a markedly decreased rate from the amniote ancestor to the present for the last 300 million years. According to Goodman *et al.* (1975), from 500 million to 400 million years ago the genes descending from 'the basal vertebrate through the haemoglobin-myoglobin and β-α ancestors to the teleost-tetrapod α ancestor evolved at the rate of 109 NR%' where NR% means the number of nucleotide replacements per 100 codons per 10^8 years. This corresponds to $k_{aa} = 10.9 \times 10^{-9}$ per amino acid per year in the terminology of this book, and it is some 10 times higher than our estimate of 10^{-9} per amino acid site per year. On the other hand, for the last 300 million years, the rate estimated by them is only 15NR% (i.e. $k_{aa} = 1.5 \times 10^{-9}$) which is not very different from $k_{aa} = 1.12 \times 10^{-9}$ that we obtained by comparing α chains among several mammals. Together with their specious explanation that when new functions emerged in duplicated globin genes rapid mutant substitutions occurred by *positive* Darwinian selection, Goodman *et al.*'s (1975) claim of non-constancy has been quoted widely as a piece of evidence against the neutral theory (see, for example, Vogel and Motulsky, 1979).

If we examine carefully the data used by Goodman *et al.* (1974, 1975), we find that their estimate of the high evolutionary rate (109 NR%) essentially comes from their assumption that the gene duplication, leading to lamprey globin on the one hand and the ancestral globin on the other which

produced myoglobin as well as α and β globins, occurred 500 million years ago (let us call this point *a* on the phylogenetic tree; see Fig. 1 of Goodman *et al.*, 1975). They also assume that this was followed by myoglobin vs. β–α hemoglobin divergence about 30 million years later (call this point *b*). In addition, they assume that the duplication which produced the α and β hemoglobins occurred about 425 million years ago (call this point *c*) and that the ancestor of carp α and chicken α diverged 400 million years ago (call this point *d*). They allocate 50 mutations between point *a* and *b*, 58 mutations between *b* and *c*, and 51 mutations between *c* and *d*. In other words, for the time span of 100 million years (i.e. between points *a* and *b*), they assume the establishment of two gene duplications and allocate a total of 159 mutations (among about 145 amino acid sites).

Their choice of gene-duplication points in time entails some problems, of which the most serious is their assumption that the duplication responsible for the myoglobin vs. hemoglobin divergence occurred about 470 million years ago (point *b*) after the ancestor of jawed fishes diverged from the ancestor of lamprey. In other words, they assume that the history of myoglobin is younger than that of the jawless fish (agnatha).

It is much more likely that the gene duplication that had led to the lamprey (*Petromyzon marinus*) globin whose sequence was used in their analysis and the ancestral molecule from which myoglobin and α–β globins had descended is much older than they postulate. In other words, it is likely that the duplicated globin genes already existed in the ancestral agnatha that appeared 500 million years ago (point *a*). According to Hendrickson and Love (1971) the lamprey globin that has been sequenced is only one of six hemoglobins present in *P. marinus* blood, namely, hemoglobin V, which is the most prevalent of the six hemoglobins. This suggests that gene duplication occurs rather frequently in evolution. Much more important is the recent finding by Romero-Herrera *et al.* (1979) showing that true myoglobin exists in the red muscle of the heart in the lamprey. Although the amino acid sequence is yet to be determined, this finding completely invalidates Goodman *et al.*'s assumption that myoglobin originated only some 470 million years ago *after* the jawed-fish diverged from the agnatha, and that therefore myoglobin is newer than Hb V of the lamprey. In this respect, Hunt *et al.* (1978) is much more judicious in placing the lamprey globin branch on the opposite side of the phylogenetic tree from the other vertebrate globins, stating that the gene duplication responsible for myoglobin and lamprey globin must have occurred at a very ancient time. (In Hunt *et al.*'s figure this duplication point is depicted as being one billion or so years old). They also place the divergence point of α and β globins

approximately 500 million years ago, which is a much more appropriate choice than that of Goodman *et al.* who place it about 425 million years ago.

Goodman (1978) as well as Goodman *et al.* (1974, 1975) also criticize my claim of the constancy by saying that I failed to consider multiple mutations in my estimation procedure, writing as if their method is much superior to the method I use. So, let us examine this aspect of their work. Their method consists of first estimating the number of mutant substitutions using the principle of 'maximum parsimony' and then modifying the estimate, sometimes in enormous amount by their 'augmentation procedure'. Take, for example, evolutionary change of the α hemoglobin along the two lineages, one leading to the carp and the other leading to man from their common ancestor some 400 million years back. According to Goodman *et al.* (1974) the number of mutations accumulated along the lineage leading to the carp amounts to 64, while the corresponding number accumulated along the lineage to man amounts to 56, both being estimated by the maximum parsimony method as 'A-solution' values (see their Figure 2). These two numbers certainly do not show that the intrinsic rates of mutant substitutions are significantly different in these two lineages. Then, they modify these values by augmenting 48 mutations to the lineage leading to the carp and 33 mutations to the lineage leading to man. So the final estimates of mutant substitutions turn out to be 112 and 89 respectively for these two lineages (see Figure 6 of Goodman *et al.*, 1974). Although no statistical errors are available regarding these estimates, presumably, they are very large, and the difference between these two values can not be statistically significant. This can be seen clearly by the fact that in another paper published next year (Goodman *et al.*, 1975), they allocate respectively 109 and 105 mutations to these two lineages (see their Figure 1) instead of 112 and 89 as mentioned above. That their augmentation procedure itself contains serious statistical problems has been pointed out by Tateno and Nei (1978) and also by Fitch (1980). In addition, it is likely that the maximum parsimony method used by them is inherently error-ridden, for there is no guarantee that evolution proceeds through the shortest path.

According to Goodman *et al.* (1974), their maximum parsimony method can often distinguish between different codons for the same amino acid and this discriminating power increases its utility in reconstructing phylogeny (see page 4 of Goodman *et al.*, 1974). They have published extensive globin codon sequences (in terms of the RNA code) which they determined by this method. So, it is of interest to check if their determinations are valid now that the actual nucleotide sequence has become available for a few globin

genes, thanks to the new technology of amplifying gene copies in a bacterial plasmid followed by rapid DNA sequencing using the Maxam-Gilbert method. Recently, the sequence of rabbit α-globin messenger RNA (mRNA) has been determined by Heindell *et al.* (1978) using this procedure. This allows us to examine the corresponding maximum parsimony codon sequence determined by Goodman *et al.* (1974).

It turns out that, of 51 codons predicted by Goodman *et al.* for rabbit α-globin (see Table 4 of their 1974 paper), 44 are simply wrong and only the remaining 7 codons are correct (Kimura, 1981a). In other words, only 14% of the maximum parsimony codons have turned out to be correct. This poor performance is surprising because most of the indeterminancy is at the third position of the codon and even random assignment of one of four nucleotide bases to this position can achieve 25% fit on the average. A similar examination can be made for the maximum parsimony β-globin codon sequence of man (see Table 5 of Goodman *et al.*, 1974), since the entire sequence of human β-globin mRNA has been determined by Marotta *et al.* (1977). In this case, it turns out that about half of the maximum parsimony codons are wrong, the worst error being the assignment of AGU for the serine at the amino acid position 44 where the actual codon is UCC; none of the three letters is correct.

It might be argued that the bases given in the third position of the codons in the numerous tables of Goodman's 1974 paper were not intended, after all, to be correct. As explained somewhat ambiguously in the footnote to his Table 1, curiously U in the third position stands for either U or C, G stands for either A or G, and sometimes these may stand for any of the four bases. This being the case, the above criticisms of the maximum parsimony codons lose some force, although this does not mean that Goodman's maximum parsimony operation is a valid one. In fact, the fallacy of maximum parsimony codons has been revealed in a dramatic way by a recent study of Holmquist (1979). Previously, Tateno and Nei (1978) used a computer to perform Monte Carlo experiments of molecular evolution, simulating the process of divergence of nucleotide sequences by randomly accumulating mutational changes (each hypothetical sequence comprised 300 nucleotides, corresponding to 100 codons, the first half of which were assumed to be variable and the remaining half constant). Starting from a common ancestor A, a phylogenetic tree involving twenty-one contemporary sequences was produced. What is important is that, in these simulation experiments, unlike the evolutionary processes in nature, the exact topology, the numbers of accumulated nucleotide substitutions in each branch of the phylogenetic tree, and also, all the ancestral sequences at the

branch points are known. Holmquist obtained the data for these simulation experiments from Nei and Tateno, and sent the twenty-one contemporary sequences (after being translated into protein sequences) to Goodman asking him to obtain the maximum parsimony solution including both the topology and reconstructed ancestral amino acid sequences (this was called the 'inferred topology' solution). Holmquist also asked Goodman to provide him with the ancestral amino acid sequence reconstructions, given the correct topology (the solution thus obtained was designated the 'known topology' solution). These amino acid sequences obtained by Goodman were then compared with the correct ancestral sequences. It turned out that there were numerous errors in the amino acid sequences inferred by the method of maximum parsimony. Particularly noteworthy was the high rate at which errors accumulated as the distances of nodal sequences from the contemporary sequences increased. For nodes 200 or more replacements distant, the error was more than 80%. Actually, the errors are so numerous in remote ancestral reconstructions, even when the correct topology is supplied, that the reconstructed sequences are totally worthless (they even contain chain terminating codons!). It is clear now that any discussion which involves early stages of globin evolution based on maximum parsimony reconstruction is meaningless (Kimura, 1981b).

All these results show clearly that Goodman *et al.*'s (1974, 1975) work is fraught with errors and uncertainty, and contrary to their claim, there is no evidence that the evolutionary amino acid substitutions were very rapid in the early stage of vertebrate globin evolution. Their claim that the high evolutionary rate is due to positive Darwinian evolution, although plausible sounding, has no support from molecular data. It is much more likely, as we shall discuss in more detail later (chapter 5), that the high rate is usually caused by removal of pre-existing functional constraint, allowing a large fraction of previously harmful mutations to become selectively neutral so that they become fixed by random drift. In other words, when an exceptionally high rate is encountered in molecular evolution, we should suspect loss of constraint rather than acquirement of new function.

These arguments of course do not mean that no adaptive mutant substitutions occurred in the course of globin evolution whereby the functional hemoglobin $\alpha_2\beta_2$ tetramer emerged from monomeric hemoglobin (somewhat similar to Hb V of the lamprey) by gene duplication and later differentiation. Indeed a long course of selective tuning must have been involved. What I want to claim is that superimposed on such adaptive change, a much larger number of selectively neutral or nearly neutral substitutions have occurred by random drift in hemoglobin evolution.

Comparative studies of nucleotide sequences

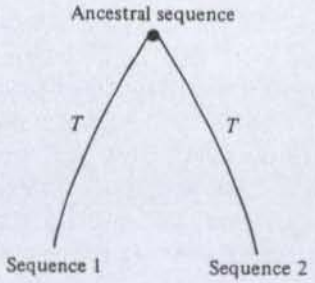
Although comparative studies of protein sequences (as treated in the previous subsection) have been a major tool for the studies of molecular evolution during the last 15 years or so, we are now experiencing a new change in the type of data being obtained, with an outburst of reports on nucleotide sequences. This is a result of the revolutionary technology of amplifying gene copies in bacterial plasmids and then rapidly sequencing the DNA by the Maxam-Gilbert method (Maxam and Gilbert, 1977). In fact, during the last few years, data on nucleotide sequences of various parts of the genome in diverse organisms have started to accumulate at an accelerated, almost explosive rate. Before long, comparative studies of amino acid sequences will be superseded by those of nucleotide sequences.

Already, it has become increasingly evident that the preponderance of synonymous and other silent base substitutions is a general but remarkable feature of molecular evolution and that this is consistent with the neutral theory (Kimura, 1977, 1979b, 1980b, 1981c; Jukes, 1978, 1980a, b; Jukes and King, 1979; Miyata and Yasunaga, 1980; Miyata, *et al.*, 1980b; Nichols and Yanofsky, 1979).

In estimating the evolutionary distances between homologous sequences in terms of the number of base substitutions, corrections for multiple and revertant changes at homologous sites are essential. This is because there are only four kinds of bases in nucleotide sequences and even two random sequences show 25% match on the average at individual sites. So we need suitable statistical methods which incorporate such corrections in the estimation of evolutionary distances (and, also, the evolutionary rate when the divergence time is known).

Let us consider a pair of homologous sites in two nucleotide sequences compared (denoted by sequences 1 and 2). In what follows we use RNA (rather than DNA) codes so that the four bases are expressed by letters U, C, A and G. This facilitates translating back and forth between codon (triplet) and amino acid, when needed, using a standard code table. We investigate how the homologous sites differentiate from each other in the course of evolution starting from a common ancestor T years back (Fig. 4.8). At individual sites, bases are substituted one after another in the course of evolution. To formulate this process, we assume a model of evolutionary base substitutions as shown in Fig. 4.9. In this figure, α is the rate of transition type substitutions, and 2β is that of transversion type substitutions, so that the total rate of substitutions per site per year is $k = \alpha + 2\beta$. Note that α and β refer to evolutionary rates of mutant substitutions in the species rather than the ordinary mutation rates at the level of individuals.

Fig. 4.8. Divergence of two homologous sequences from a common ancestral sequence T years back.



The total number of base substitutions which separate two sequences and therefore involve two branches each with length T (see Fig. 4.8) is given by $2Tk$ which we denote by K.

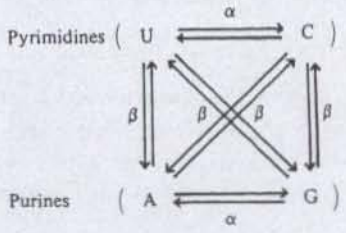
$$K = 2Tk = 2\alpha T + 4\beta T. \tag{4.9}$$

In comparing homologous sites of sequences 1 and 2, we note that there are twelve combinations of different bases as shown in Table 4.7. Let P be the

Table 4.7. *Various types of different base pairs at homologous nucleotide sites in two sequences compared*

	Type of difference											
	Transition type				Transversion type							
Sequence 1	U	C	A	G	U	A	C	G	U	G	C	A
Sequence 2	C	U	G	A	A	U	G	C	G	U	A	C
(Frequency)	(P)				(Q)							

Fig. 4.9. Model of evolutionary base substitutions. In this model, the rate of transition type substitution (α) need not be equal to the transversion type substitutions (β).



probability (relative frequency) that, at time T, homologous sites are occupied by UC, CU, AG or GA. In other words, P is the probability of homologous sites showing the transition type base differences. Similarly, let Q be the probability of homologous sites showing the transversion type differences (Table 4.7). Then, it can be shown that P and Q satisfy the following set of differential equations (for derivation, see Kimura, 1980b).

$$\frac{dP}{dT} = 2\alpha - 4(\alpha+\beta)P - 2(\alpha-\beta)Q$$
$$\frac{dQ}{dT} = 4\beta - 8\beta Q \tag{4.10}$$

The solution of this set of equations which satisfies the condition $P = Q = 0$, at $T = 0$, that is, the two sequences are identical at the start, is as follows.

$$\left.\begin{aligned} P &= \tfrac{1}{4} - \tfrac{1}{2}e^{-4(\alpha+\beta)T} + \tfrac{1}{4}e^{-8\beta T} \\ Q &= \tfrac{1}{2} - \tfrac{1}{2}e^{-8\beta T} \end{aligned}\right\} \tag{4.11}$$

From these equations we get

$$4(\alpha+\beta)T = -\log_e(1-2P-Q) \tag{4.12}$$

and

$$8\beta T = -\log_e(1-2Q). \tag{4.13}$$

Since the evolutionary distance in terms of the number of base substitutions between the two sequences is given by equation (4.9), we obtain

$$K = -\tfrac{1}{2}\log_e[(1-2P-Q)\sqrt{1-2Q}] \tag{4.14}$$

(Kimura, 1980b). This equation may be used to estimate the evolutionary distance between two sequences in terms of the number of base substitutions per site that have occurred in the course of evolution extending over T years. In this equation, $P = n_1/n$ and $Q = n_2/n$, where n_1 and n_2 are respectively the numbers of sites at which two sequences differ from each other with respect to transition type and transversion type substitutions, and n is the total number of sites compared.

In the special case of $\alpha = \beta$, equations (4.11) reduce to $P = Q/2 = [1-\exp(-8\alpha T)]/4$, so that (4.14) reduces to

$$K = -\tfrac{3}{4}\log_e(1-\tfrac{4}{3}\lambda), \tag{4.15}$$

where $\lambda = P + Q = 3Q/2$ is the fraction of sites for which two sequences differ from each other. This formula is well-known (see Kimura and Ohta, 1972), and it was first obtained by Jukes and Cantor (1969). However, in actual situations, particularly when the third codon positions are compared, P is often larger than Q, and therefore the assumption of $\alpha = \beta$ or $P = Q/2$ is not always realistic. This is one reason why formula (4.14) is better than (4.15). Equation (4.14) also has a desirable property that as P and Q get small, it converges to $K = P + Q$ independent of α and β. The base substitution rate per year is then given by $k_{nuc} = K/(2T)$ if the divergence time T is known, where the subscript nuc means that the estimate refers to the rate per nucleotide site.

Since a large fraction of substitutions at the third codon positions are synonymous, that is, they do not cause amino acid changes, it would be of interest to estimate the synonymous component of the substitution rate at this position. From the standard RNA code table we note that for a given pair of bases in the first and the second codon positions, roughly speaking there are two situations; either the third position is completely synonymous (four-fold degeneracy) or synonymy is restricted within purines or pyrimidines (two-fold degeneracy). These two situations occur roughly in equal numbers. Thus, the synonymous component of substitutions at the third position which we denote by k_S' may be estimated by $0.5(\alpha+2\beta)+0.5\alpha$ or $\alpha+\beta$. Let $K_S' = 2Tk_S' = 2(\alpha+\beta)T$ be the synonymous component of the distance, then, we get

$$K_S' = -\tfrac{1}{2}\log_e(1-2P-Q), \tag{4.16}$$

if we apply equation (4.12).

It is desirable to have formulae for the error variance (due to sampling) of the estimated values of K and K_S'. Let n be the number of nucleotide sites for which the two sequences are compared with each other. Then the large sample standard errors of K and K_S' are

$$\sigma_K = \frac{1}{\sqrt{n}}\sqrt{[(a^2P + b^2Q) - (aP+bQ)^2]} \tag{4.17}$$

and

$$\sigma_{K'S} = \frac{\sqrt{[4P+Q-(2P+Q)^2]}}{2(1-2P-Q)\sqrt{n}}, \tag{4.18}$$

where $a = 1/(1-2P-Q)$ and $b = (1/2)[1/(1-2P-Q) + 1/(1-2Q)]$.

As an example, let us compare the nucleotide sequence of the rabbit β globin (Efstratiadis *et al.*, 1977) with that of chicken β globin (Richards *et al.*, 1979). There are 438 nucleotide sites that can be compared, corresponding to 146 amino acid sites (codons). Among these sites, we find that there are 58 sites for which these two sequences show transition type differences, and 63 sites with transversion type differences. Thus, $P = 0.132$,

$Q = 0.144$ and we obtain $K = 0.348$. Mammals and birds probably diverged during the Carboniferous period (see Romer, 1968), so we tentatively take $T = 300 \times 10^6$ years. The evolutionary rate per site is then $k_{nuc} = K/(2T) = 0.58 \times 10^{-9}$ per year. This is the overall rate per site, but it is much more interesting to estimate separately the evolutionary rates for the three codon positions. For the first position, there are 146 nucleotide sites compared, and we find $P = 15/146$ and $Q = 21/146$, giving $K_1 = 0.300$, where subscript 1 denotes that it refers to the first codon position. Similarly, for the second position, we find $P = 7/146$ and $Q = 18/146$ so that $K_2 = 0.195$. Finally, for the third position, $P = 36/146$ and $Q = 24/146$, and we get $K_3 = 0.635$, which is much higher than the corresponding estimates for the first and second positions. We can also estimate the synonymous component of the evolutionary distance per third codon position. From equation (4.16), this turns out to be $K_S' = 0.535$.

In Table 4.8, results of similar calculations are listed for various comparisons involving the human β (Marotta *et al.*, 1977), the mouse β (Konkel *et al.*, 1978) and rabbit α globin sequences (Heindell *et al.*, 1978), in addition to the chicken and rabbit β-globin sequences. Except for the last two comparisons involving the abnormal, globin-like α-3 gene (Nishioka *et al.*, 1980), the table suggests that the relationship $K_2 < K_1 < K_3$ holds generally; the evolutionary mutant substitutions are most rapid at the third position, and this is followed by the first position, and then the second position, where the substitutions are the slowest. This can readily be interpreted by the neutral theory as follows. Among the three codon positions, base substitutions at the second positions tend to produce more drastic changes in the physico-chemical properties of amino acids than those at the first positions. Take for example the codon for Pro. The code table shows (see Table 2.1) that it is CC*N*, where *N* is any one of four bases. Substitutions of U, A and G, for base C at the first position, lead respectively to Ser, Thr and Ala. In terms of Miyata's distance (based on polarity and volume differences between an amino acid pair; see Miyata *et al.*, 1979), they are 0.56, 0.87 and 0.06 units apart from Pro, with the average distance of about 0.5. On the other hand, the corresponding average distance resulting from substitutions for C at the second position of the codon turns out to be about 2.5.

This means that mutational changes at the first position have a higher chance of not being harmful (i.e. selectively neutral or equivalent) than those at the second position, and therefore, have a higher chance of being fixed in the species by random drift (Kimura and Ohta, 1974). This type of reasoning applies more forcibly to the third position (as compared with the

Table 4.8. *Evolutionary distances in terms of the number of base substitutions estimated for several comparisons of globin sequences, K_1, K_2, and K_3 denote the number of base substitutions at the first, second and third positions of codons, while K_S' stands for the estimated number of substitutions due to synonymous changes at the third position. Estimated values of these parameters together with their standard errors are listed*

Comparison	Evolutionary distances per nucleotide site			
	K_1	K_2	K_3	K_S'
Chicken β vs. Rabbit β	0.30 ± 0.05	0.19 ± 0.04	0.64 ± 0.11	0.53 ± 0.10
Human β vs. Mouse β	0.17 ± 0.04	0.13 ± 0.03	0.34 ± 0.06	0.28 ± 0.05
Human β vs. Rabbit β	0.06 ± 0.02	0.06 ± 0.02	0.28 ± 0.06	0.25 ± 0.05
Rabbit β vs. Mouse β	0.16 ± 0.04	0.13 ± 0.03	0.43 ± 0.07	0.36 ± 0.07
Rabbit α vs. Rabbit β	0.54 ± 0.09	0.44 ± 0.07	0.90 ± 0.15	0.69 ± 0.13
Rabbit α vs. Mouse α-1	0.12 ± 0.03	0.11 ± 0.03	0.54 ± 0.09	0.47 ± 0.09
Rabbit α vs. Mouse α-3	0.27 ± 0.06	0.28 ± 0.06	0.69 ± 0.13	0.56 ± 0.12
Mouse α-1 vs. Mouse α-3	0.16 ± 0.04	0.20 ± 0.05	0.30 ± 0.06	0.22 ± 0.05

first and the second positions), since a majority of mutational changes at this position do not cause amino acid changes.

As an index for the relative frequencies of synonymous and amino acid-altering substitutions, we may take the ratio $K_S' : \frac{1}{2}(K_1 + K_2)$. This is 4.17 for the human β-rabbit β comparison ($T = 8 \times 10^7$ years), 2.16 for the chicken β-rabbit β comparison ($T = 3 \times 10^8$ years) but only 1.41 for the rabbit α-rabbit β comparison ($T = 5 \times 10^8$ years). Perier *et al.* (1980) claims, based on similar analysis of preproinsulin and globin evolution, that initially the silent sites accumulate changes some seven times more rapidly than replacement sites (i.e. amino acid altering changes), but that accumulation saturates at 85–100 million years. Such a 'saturation' phenomenon of silent substitutions is probably an artefact (they are more apparent than real), coming from difficulty of detecting all the synonymous substitutions, as more and more of them accumulate in the codons. One reason why I suspect this is that the higher rates are obtained from more recent comparisons (such as between mammals), and that because the evolution proceeds from the remote past toward the more recent past, but not in the reverse direction, the 'saturation' phenomenon can only be superficial. If evolutionary change really decreased, the more recent comparisons should show slower changes, but what is observed is the other way round. Furthermore, as compared with the first and the second positions, the third position shows marked deviation of the base composition from equality (e.g. G36%, C30%, U27%, A7% in human β) and this complicates the problem of estimation for a very large value of T. For additional discussion on the estimation of synonymous substitution at position 3 of codons readers may refer to Kimura (1981c).

Much more fascinating are the last two comparisons in Table 4.8 which involve the globin-like α-3 gene recently sequenced in the mouse (Nishioka *et al.*, 1980). This gene completely lacks two intervening sequences normally present in all the α and β globin genes, and it does not encode globin. In other words, it is inactive in the production of stable mRNA. However, from a comparison of this sequence with the normal, adult mouse α globin (α-1) and the rabbit α globin nucleotide sequences, it is evident that this gene evolved from a normal ancestral α globin gene through duplication and subsequent loss of its intervening sequences. This must have occurred after the mouse and the rabbit diverged from their common ancestor some 80 million years ago. This α globin-like gene acquired in its coding region a number of insertions and deletions of nucleotides. In making sequence comparisons, therefore, I chose only those codons of the α-3 gene which do not contain such changes and which are either identical

with or differ only through base substitutions from the corresponding (homologous) codons of the mouse α-1 and the rabbit α genes.

As seen from the last three lines of Table 4.8, this 'dead gene' (α-3) evolved at a much faster rate than its normal counterpart (mouse α-1 gene), particularly with respect to the first and the second codon positions. This is easy to understand from the neutral theory. Under a normal situation, each gene is subject to a selective constraint coming from the requirement that the protein which it produces must function normally. Evolutionary changes are restricted within such a set of base substitutions. However, once a gene is freed from this constraint, as is the case for this globin-like α-3 gene, practically all the base substitutions in it become indifferent to Darwinian fitness, and the rate of base substitutions should approach the upper limit set by the mutation rate (this holds only if the neutral theory is valid, but not if the majority of base substitutions are driven by positive selection). If the rates of synonymous substitutions are not very far from this limit (Kimura, 1977) we may expect that the rates of evolution of a 'dead gene' are roughly equal to those of synonymous substitutions. Recently Miyata and Yasunaga (1981) computed the evolutionary rate of a mouse pseudo alpha globin gene ($\psi\alpha$ 30.5) which was sequenced by Vanin *et al.* (1980) and which appears to be essentially equivalent to the α-3 gene. They also obtained a result supporting this prediction.

In addition to coding regions, it would be interesting to estimate base substitution rates in non-coding regions such as introns. I use data presented by van Ooyen *et al.* (1979) who investigated the similarity between the nucleotide sequences of rabbit and mouse β-globin genes. They list (see their Table 2) separately the numbers of 'transition' and 'transversion' type differences between the homologous parts of these sequences. For the small introns, excluding five gaps that amount to six nucleotides, $P = 27/113, Q = 18/113$ and $n = 113$. Using equations (4.14) and (4.17), we get $K = 0.60 \pm 0.12$. This is not significantly different from the substitution rate at the third position $K_3 = 0.43 \pm 0.07$. The large introns of rabbit and mouse β globin genes differ considerably in length, being separated from each other by fourteen gaps (determined by optimization of alignment of the sequences) which amount to 109 nucleotides. Excluding these parts, $P = 113/557$ and $Q = 179/557$, from which we get $K = 0.90 \pm 0.07$. This value is significantly larger than K_3. It is likely, as pointed out by van Ooyen *et al.*, that insertions and deletions occur rather frequently in this part in addition to point mutations, and that they inflate the estimated value of the 'nucleotide substitution rate', since a majority of these changes may also be selectively neutral and subject to random fixation by genetic drift.

5

Some features of molecular evolution

There are several features which distinguish molecular evolution from phenotypic evolution. These features may be summarized in the form of five principles (Kimura and Ohta, 1974) as follows; the first four principles are empirical, while the last one is theoretical and helps us to understand the nature of 'selectively neutral mutations'.

(i) *For each protein, the rate of evolution in terms of amino acid substitutions is approximately constant per year per site for various lines, as long as the function and tertiary structure of the molecule remain essentially unaltered.*

This is the rate-constancy hypothesis which Zuckerkandl and Pauling (1965) called the 'molecular evolutionary clock'. This is one of the most controversial subjects in molecular evolution. Particularly, it has been debated in relation to the date of human–ape divergence (see Wilson *et al.*, 1977), and to evidence for and against the neutral theory. I have already discussed extensively the problems relating to the constancy of molecular evolutionary rates in chapter 4, so I shall not repeat the same material here. Suffice it to say that the evidence is good in globins among mammals, although the observed variance of evolutionary rates among lines is slightly larger than what is expected by chance. There is also rough constancy of the evolutionary rates in cytochrome c and fibrinopeptides (Dickerson, 1971) although each has a different rate from that of globins (see also Table 4.1). Recently, Dayhoff (1978) has presented an extensive list of 'rates of mutation acceptance', that is, evolutionary rates in terms of mutant substitutions, in diverse protein families from which a reasonable estimate can be made. She states that although there is a factor of almost 400 between the fastest changing families and the slowest histone, the rate of

change of proteins within a family seldom varies by more than a factor of two or three, particularly when the proteins fill the same functional niche in different organisms. One could perhaps say that the general success of constructing consistent phylogenetic trees through comparative studies of protein and nucleotide sequences speaks for the validity of the rate-constancy hypothesis.

This does not mean that the hypothesis is absolutely valid. In fact, some cases are reported for which the hypothesis does not appear to hold very well. For example, according to Matsubara *et al.* (1978) the molecular evolution of chloroplast-type ferredoxin among plants occurred rather rapidly in angiosperms but rather slowly in blue-green algae. However, they are cautious enough to point out that the time of divergence of angiospermae is still obscure and that these observations might suggest the polyphyletic origin of angiospermae. Jukes and Holmquist (1972) report that rattlesnake cytochrome c has evolved three or four times as rapidly as turtle cytochrome c. A more systematic attempt to examine rate constancy has recently been made by Sneath (1980) who called attention to the importance of distinguishing the rate of observed occurrences from the 'implied underlying rate'. He examined the coefficients of variation of underlying rates using data from cytochrome c, globin, fibrinopeptide, pancreatic ribonuclease and snake toxin, and obtained results suggesting that underlying rates are remarkably constant except for cytochrome c, although he admits that there are uncertainties in measures of evolutionary rate.

Wilson *et al.* (1977) attempted to test whether the generation-time hypothesis is valid, that is, if the evolutionary rate is primarily dependent on the number of generations since divergence and only secondarily dependent on absolute time. Using a dozen molecules, they compared the evolutionary rate in mammals which have a short generation time with the rate in mammals having a long generation time. It turned out that none of the observed values falls exclusively within the range predicted by the generation-time hypothesis. On the other hand, observed values are consistent with the assumption that the rate depends on the absolute time (years). Whether the rate can be extrapolated validly from higher animals down to bacteria based on the rate-constancy hypothesis is debatable and even doubted, but the apparent success of constructing a consistent phylogenetic tree encompassing all the kingdoms of living organisms using 5S ribosomal RNA (Hori and Osawa, 1979) suggests that even here the hypothesis is not totally invalid. The observation that the rate of evolution in terms of mutant substitutions for a given molecule (say hemoglobin α in

mammals) is constant per year, completely independent of living conditions and the organism's way of life, and apparently independent of generation span can be explained much more easily by the neutral theory than by the selectionist theory. It is not necessary that the evolutionary rate be absolutely constant (without exception) for the neutral theory to be valid, although exceptional cases have to be explained based on the neutral theory if the theory is valid.

As shown in chapter 4, under the neutral theory we have $k = v$ (equation 3.20), that is, the rate of evolution is equal to the mutation rate for neutral alleles. Taking into account the fact that a certain fraction of mutations may be deleterious, this equation can also be expressed as

$$k = v_0 = v_T f_0, \tag{5.1}$$

where v_T is the total mutation rate and f_0 is the fraction of neutral mutations. Note that the existence of deleterious mutations does not contradict the neutral theory, since they make no contribution to adaptive evolutionary change. We then assume, to explain the observation, that the rate of neutral mutation is constant per year for a given molecule. Definitely advantageous mutants may also occur but we assume that they are too rare to be of any significance in the ordinary observation of the rate of molecular evolution.

On the other hand, if the mutant substitutions are driven by positive natural selection, we have $k = 4N_e sv$ (equation 3.22), that is, the rate is equal to four times the product of effective population size (N_e), the mutation rate (v) for definitely advantageous alleles and the selection coefficient (s) representing their advantage. If we take into account the possibility that deleterious mutants occur in addition, the selectionist equation which corresponds to (5.1) becomes

$$k = 4N_e \bar{s}_A v_T f_A, \tag{5.2}$$

where v_T is the total mutation rate, f_A is the fraction of definitely advantageous mutants and $\bar{s}_A$ is the average selection coefficient of such mutants. Note that $N_e \bar{s}_A$ must be much larger than unity; otherwise they behave more like selectively neutral mutants than selectively advantageous mutants. Here we neglect selectively neutral or nearly neutral mutations, for they do not contribute significantly to evolutionary mutant substitutions from the selectionist standpoint. Then, in order to explain the observed constancy of the evolutionary rate, we must assume that the product $N_e \bar{s}_A v_T f_A$ is kept constant among diverse lineages for a given protein. In other words, a highly complicated and arbitrary set of assumptions must be invoked regarding mutation, gene interaction and

ecological conditions, as well as population size, in order to explain the approximate constancy per year from the selectionist (or the neo-Darwinian) viewpoint. On the other hand the neutral hypothesis requires only one *ad hoc* assumption, namely that f_0 increases with the length of the reproductive generation time. This is not unreasonable when we realize that larger organisms with longer life cycles tend to have smaller populations; hence, a larger fraction of mutants would be expected to be selectively neutral. We shall discuss this problem more in detail later (see, section 8.7). Note that this problem does not arise when we compare two organisms having the same generation span but living in totally different environments.

One important point which is often overlooked is the effect of population size; it is rather common to find the statement (see for example Perier *et al.*, 1980, p. 561) that the constancy in the evolutionary rate of mutant fixation can be explained by constancy in the 'rate of selection'. Such a statement is clearly not valid. If evolution is the result of the gradual accumulation of rare, advantageous mutations, such mutants should appear in vastly greater numbers in a large population than in a small one, so that evolution is much more rapid in the former.

Unlike phenotypic evolution for which no detailed knowledge about underlying genetic change is available, we can go much deeper in the case of molecular evolution, to analyse the process of genetic change involved.

It is well known that the majority of evolutionary amino acid substitutions and also the hemoglobin variants found in human populations can be interpreted as due to single base substitutions (see for example, Vogel and Motulsky, 1979). In globins, as we have seen already, amino acid substitutions occur at the rate of about 10^{-9} per amino acid site per year. This means that, roughly speaking, in the α chain consisting of 141 amino acids, one substitution occurs on the average every 7 million years. Consider a mammalian species with an average generation span (g) of two years and comprising half a million individuals each generation (its effective size may be much smaller). Let us suppose that the mutation rate (v) per generation of α globin is 10^{-6}. We restrict our consideration only to those mutations that cause amino acid changes. During 7 million years, the total number of α globin mutations which occur in the species is then $2Nv \times 7 \times 10^6/g = 2 \times (0.5 \times 10^6) \times 10^{-6} \times 7 \times 10^6/2 = 3.5 \times 10^6$. Thus, 3.5 million mutations occur, among which only one becomes fixed. Not all of them are unique.

Let us inquire how often the same α-globin mutations recurred during this period. For this purpose we consider the total number of alleles at this

locus that can be generated through single base substitutions. Although each amino acid site (codon) can be occupied by one of 20 different amino acids, the number generated per codon through single base substitutions is somewhat less than 19. In fact, the appropriate number of different amino acids is about 6.5. This is shown as follows. Consider base substitutions separately at individual nucleotide sites within a codon. At the first and also at the second positions, a base substitution leads to a different amino acid in the majority of cases, so with respect to these two positions, the possible number of different amino acids produced is $3+3$ or 6. At the third position, roughly speaking two cases occur in equal frequencies. When the codon is four-fold degenerate (as in Val), no change occurs, but when it is two-fold degenerate a base substitution can produce a different amino acid (such as Asp vs. Glu when the codon is GAN). On the average, therefore, the number of different amino acids which are generated at the third position is 0.5. This gives, as the total number of different amino acids generated per codon through single base substitutions, $3+3+0.5$ or 6.5. For the α-globin as a whole, the approximate number of possible alleles that can be generated through single base substitutions is 6.5×141 or 916.5.

Taking 1000 as the possible number of α-globin alleles in order to continue our rough calculation and dividing 3.5 million mutations by 1000, we find that the same mutant allele occurred 3500 times during 7 million years. If any one of these 1000 alleles had had a significant selective advantage, it would have been fixed in the species without the necessity for such repeated occurrence. As explained in chapter 3, if a mutant has a selective advantage s, the probability of its fixation is $u = 2sN_e/N$, where we assume $4N_e s \gg 1$ (see equation 3.19). Suppose that the effective size is half of the actual size ($N_e = N/2$), in which case we have $u = s$. Then the probability that all the mutations were lost (i.e. unsuccessful) until the 3500th one with respect to this allele is e^{-3499s}. This probability must be fairly high, for such a pattern is commonly observed in globin evolution.

So, let us take this probability as 0.5, so that $e^{-3499s} = 0.5$. This gives $s = 2 \times 10^{-4}$. If the number of different advantageous alleles is m, then this value of s has to be divided by m. A similar estimate of $\bar{s}_A$ can be obtained if we equate the reciprocal of k or $4N_e v_T f_A \bar{s}_A$ to 7×10^6 assuming $4N_e = 10^6$, $v_T = 10^{-6}$ and $f_A = m/1000$. Such a small selective advantage, even if it exists, will never be measurable in practice. The story is much more simple and straightforward from the standpoint of the neutral theory; the mutation rate for neutral alleles per generation is 141×10^{-9} or $v_0 = 1.4 \times 10^{-7}$ which is roughly 1/7 of the total mutation rate assumed, i.e. $v_T = 10^{-6}$, the rest of the mutations being deleterious.

Before we proceed, I would like to discuss briefly the claim that hemoglobins from different mammals are not physiologically equivalent and that differences reflect adaptations of animals to their environments. For example, Tomita and Riggs (1971) report that mouse hemoglobin has the largest Bohr effect and elephant hemoglobin the smallest, with human hemoglobin occupying an intermediate position (although its difference from mouse hemoglobin appears to be rather small). Here, Bohr-effect means that when partial pressure of carbon dioxide increases or pH is lowered, the hemoglobin molecules tend to release oxygen even if oxygen partial pressure remains the same. It is expected that the smaller the mammal the larger the metabolic rate. Therefore, they claim that the observed relationship between the body weight and the Bohr effect must be adaptive and that this does not support the neutral theory. I cannot deny their claim, but I think it probable that such adaptive substitutions, if real, constitute a small fraction of the total amino acid substitutions in evolution. There must be at least a few thousand living species of mammals. It is rather difficult to believe that each of the globin molecules in a different species has special qualities that fit it exactly to the needs of the species, and that these needs must therefore be different from the exact needs of all other species for oxygen transport in the blood.

(ii) *Functionally less important molecules or parts of molecules evolve (in terms of mutant substitutions) faster than more important ones.*

(iii) *Those mutant substitutions that are less disruptive to the existing structure and function of a molecule (conservative substitutions) occur more frequently in evolution than more disruptive ones.*

These two principles have very important bearings on the neutral theory of molecular evolution, so I shall elaborate them further in chapter 7. Here I only point out that they have very simple interpretations from the standpoint of the neutral theory: the probability of a mutation not being harmful and therefore selectively neutral is larger if the mutation occurs in a functionally less important molecule or a part of a molecule, and thus has a higher chance of being fixed in the population by random genetic drift. This probability is also higher if the mutation is 'conservative', that is, if it has a less drastic effect. One of the most extreme situations is a synonymous substitution which occurs predominantly at the third codon position.

(iv) *Gene duplication must always precede the emergence of a gene having a new function.*

The importance of gene duplication in evolution was noted earlier by the great *Drosophila* workers of the Morgan school. The crucial point pertinent here is that the existence of two copies of the same gene enables one of the copies to accumulate mutations and to eventually emerge as a new gene, while the other copy retains the old function required by the species for survival through the transitional period. Shielded by the normal counterpart in the corresponding site of the duplicated DNA segment, mutations that would have been rejected before duplication can now accumulate, and through their accumulation, the stage is set for the emergence of a new gene.

The creative role which gene duplication plays in evolution has been much clarified by Ohno (1970) in his stimulating book, in which he considers new evidence based on modern molecular, cytological and paleontological researches. Previously, he proposed the bold hypothesis that our vertebrate ancestors during the fish stages went through probably two rounds of tetraploidization before coming on land to live (Ohno *et al.*, 1968). This occurred before a rigid chromosomal sex-determining mechanism was established. Together with his more recent paper (Ohno, 1973), Ohno has made an important contribution to the modern evolutionary theory by bringing to light the remarkably conservative nature of mutant substitutions in evolution. Strong similarity of the X-chromosome among mammals ('Ohno's Law') can readily be explained by noting that the X-chromosome exists in single dose in males and also effectively in single dose in females due to random X-chromosome inactivation, and is therefore exposed more to natural selection than autosomes.

During the last few years, observations showing the prevalence of duplicated or repeated structures within the genome of higher organisms have accumulated rapidly. Previously, we knew that the α and β chains of hemoglobin were the products of a gene duplication. These were descended from a single chain similar to the one found in the blood of the lamprey. Before this event, hemoglobin and myoglobin diverged by gene duplication, and after this event, the γ and δ chains were derived also by gene duplication from the β chain in the ancestor of humans (γ is a component of Hb A_2 and δ is that of fetal hemoglobin HbF). As far as the appearance of a new function is concerned, the last major important adaptive change is the one from a monomer to a tetramer ($\alpha_2\beta_2$) which occurred about 500 million years ago. This change enabled hemoglobin to give up its oxygen readily when pH is lowered in the tissue, and to combine with oxygen readily when the pH is

raised in the lungs. This new function gave animals the ability to escape from predators or to capture prey more rapidly than the possessors of a monomeric hemoglobin. It has been known that a normal human genome contains at least one each of β, δ, ε and ζ hemoglobin genes, and that the genes for α and γ genes are duplicated (for a useful review of the genetics of hemoglobins, see Vogel and Motulsky, 1979). The incidence of triplicated α-globin loci in humans is reported by Goossens *et al.* (1980). According to them, individuals heterozygous for the triple α loci ($\alpha\alpha/\alpha\alpha\alpha$) show no consistent abnormal clinical manifestations. The triplicated α-globin loci must have been produced by unequal crossing-over. In the BALB/c mouse, the entire β-globin complex consists of seven linked genes (Jahn *et al.*, 1980).

Gene duplication also produced non-functional 'pseudogenes'. According to a recent report by Nishioka *et al.* (1980), the mouse has at least two non-functional α-genes, α-3 and α-4; α-3 lacks the two introns that normally exist in functional globin genes, and α-4 contains mutations analogous to those in certain human pathological hemoglobin M. The α-3 sequence as determined by Nishioka *et al.* (1980) shows that it contains a number of insertions and deletions as compared with the normal α sequence and it cannot encode globin. A human α-psuedogene was recently described (quoted from Proudfoot, 1980).

Another interesting example showing that a gene with a new function emerged by duplication comes from the comparison of α-lactalbumin and lysozyme. Lysozyme is an enzyme that attacks and lyses the bacterial cell wall. It is found widely in diverse organisms. On the other hand, α-lactalbumin is not by itself an enzyme but is a component of the lactose synthetase system which is found only in mammary glands during lactation. According to Dickerson and Geis (1969), the amino acid sequences of chicken (hen egg white) lysozyme, human lysozyme and bovine α-lactalbumin show enough similarity among them that they are clearly descended from a common ancestor. Also, these proteins appear to have essentially the same polypeptide chain folding, supporting the case for the common origin. The problem is when α-lactalbumin diverged from vertebrate lysozyme. In terms of observed amino acid differences, both chicken and human lysozymes are almost equally distant from bovine α-lactalbumin ($d_{aa} = 79$ and 82), but these two lysozymes are much nearer to each other ($d_{aa} = 57$) as compared with their distances from α-lactalbumin. On the other hand, it is natural to assume that the α-lactalbumin diverged from lysozyme in the incipient mammalian line when the milk-producing system was being developed. If the rate-constancy hypothesis is valid, one should expect that the distance between bovine α-lactalbumin and human

lysozyme (both being in the mammalian line) is much nearer than the distance between bovine α-lactalbumin and chicken lysozyme (bovine–chicken divergence being much older than the origin of the mammal), contrary to observations. To resolve such a contradiction, Dickerson and Geis (1969) claim that the fallacy lies in the rate-constancy hypothesis, and that evolutionary alteration in α-lactalbumin was much more rapid than that of lysozyme. Observed amino acid differences would suggest that the former is at least twice as rapid. According to them, when a duplicated gene is being altered to perform a new function, selection pressure becomes unusually severe so that the sequence will change unusually rapidly. This explanation has since been accepted widely, being treated as if it were an established fact. Subsequent study by White *et al.* (1977), however, cast doubt on its validity. These authors made more extensive comparisons involving α-lactalbumins of man, guinea pig, and cow, and also comparisons involving lysozymes of man, rat, and baboon. They found that evolutionary rates in these two proteins are about the same in mammals, and they proposed an alternative interpretation that the duplication event responsible for the origin of lactalbumin from lysozyme was more ancient than the bird–mammal divergence. If so, the rate-constancy hypothesis is intact.

In my opinion, the hypothesis that evolutionary mutant substitutions occurred at a very high rate immediately after gene duplication cannot be ruled out in some cases, but it may be explained more naturally by assuming random fixation of slightly deleterious mutants than by invoking an unusually high rate of adaptive mutant substitutions. The above examples clearly show that duplication can produce a new gene with a useful function, but also it tends to lead to the degeneration of one of the duplicated copies. In fact, it is likely that duplications produce degenerated, non-functional genes much more often than useful new genes as pointed out by Nei (1969) and Ohta and Kimura (1971a). This is because many mutations (including deletions and insertion of nucleotides), which would have been definitely deleterious before duplication, become not harmful (neutral) or only very slightly deleterious after duplication, thus enabling them to spread in the population by random drift (for a mathematical treatment of this problem, see Kimura and King, 1979). What is important is that, as first emphasized by Muller and Bridges (see Carlson, 1981), duplications enable genes to make evolutionary experiments which have been forbidden before, liberating them from incessant natural selection whose overwhelming activity is 'stabilizing' or keeping the status quo. In other words, gene duplications create conditions that enable random drift

to operate much more prominently on mutants than was possible before duplication. This allows fixation of mutants that are slightly deleterious for contemporary conditions but which may have other useful effects for adaptation to a new environment. Thus, if we take into consideration the role of gene duplication, the neutral theory has a bearing on the problem of progressive evolution. An apparently paradoxical statement made by Ho and Saunders (1979) that 'a relative lack of natural selection may be the prerequisite to major evolutionary advance' could be understood in the light of the neutralist paradigm.

The haploid chromosome set of a mammal contains about 500 times as much DNA as the genome of *E. coli*, and this must have been a result of repeated duplication in the course of evolution. That gene duplications must have played a spectacular role in progressive evolution from prokaryotes to mammals may also be seen from Dayhoff's (1978) conjecture that the approximately 50 000 proteins which are estimated to exist in humans can be grouped into about 500 superfamilies, each containing about 100 sequences.

We should not forget that underneath all these problems, there is the important problem of how the duplication itself becomes fixed in the population in the first place. One popular argument (see for example, Ohno, 1970, p. 65) is that if a locus contains a pair of overdominant alleles, say A_1 and A_2, then production of a chromosome which has two loci A_1 and A_2 as a result of duplication can attain permanent heterozygous advantage without producing unfit homozygotes. Note that in this case, crossing-over between homologous chromosomes is required. Spofford (1969) presented a population genetical treatment to show that such a heterotic duplication can be incorporated into a species.

However, it is possible, and indeed much more likely that intra-chromosomal duplication leads to a chromosome segment of the form A_1A_1. If such a duplication is selectively neutral or only very slightly deleterious, it becomes fixed in the population by random genetic drift. (Further increase to become triplicated etc. can be brought about by unequal crossing-over between sister chromatids or out of register synapsis and crossing-over between homologous chromatids, as in the case of the multigene family which I shall discuss subsequently.) Like ordinary point mutations, gene duplications are constantly occurring at low frequencies in the population, and the 'mutation pressure' for intra-chromosomal duplication in conjunction with random drift, during the long course of evolution, must be the main cause for the widespread occurrence of repeated structures in the genome of higher organisms. A somewhat similar

discussion along this line together with numerical examples are contained in Mayo (1970).

Tetraploidization differs from intra-chromosomal duplication in that all loci are duplicated in one step. This form of producing duplicated loci, although common in plants, is rather rare in animals; but in fishes belonging to the suborder *Salmonidea*, such as trout and salmon, there are autotetraploid species which have progressed toward diploidization in various degrees (Ohno 1970). In this connection, the observations made by Allendorf (1978) and Ferris and Whitt (1977) are particularly relevant to the neutralist-selectionist controversy. In both salmonid and catostomid fish, there are a number of duplicated loci in which one of the gene copies has lost function after tetraploidization (probably due to fixation of a non-functional 'null' allele which was definitely deleterious before duplication but no longer so after duplication) and such loci amount to approximately 50% (this is estimated to have occurred during the span of 50 million years). Allendorf (1978) called attention to an interesting relationship which may be expected between the level of heterozygosity and the frequency of loss of duplicated gene expression. He points out that if the heterozygosity in enzyme loci is maintained by balancing selection such as heterozygote advantage (overdominance), duplicated loci tend to be maintained in a state of 'fixed heterozygosity' as in Spofford's (1969) model for intra-chromosomal duplication. This tendency should be higher if the over-dominance is stronger, and, hence if the average heterozygosity is higher. In other words, the higher the heterozygosity, the lower the chance that loss of duplicated gene expression occurs after tetraploidization. On the other hand, if alleles for enzyme polymorphisms are selectively neutral or nearly neutral and maintained in the population by the balance between mutational input and random extinction by sampling drift, then more heterozygosity will result if the mutation rate for neutral alleles is higher. It is expected that the effective mutation rates for null alleles and those for neutral alleles go in parallel. Also, if a molecule is subject to less selective constraint, a larger fraction of functional mutations are expected to be neutral (not harmful), and at the same time, non-functional null mutants tend to be less harmful, leading to the same type of expectation. Then, the loss of duplicated gene expression should occur more frequently after tetraploidization if the level of heterozygosity is higher.

Keeping these opposite predictions in mind, Allendorf (1978) examined data available in both salmonids and catostomids. It turned out that among twelve proteins in the rainbow trout, the average heterozygosity ($\bar{H}$) is 6% for those which retain duplicate gene expression but 15% for those which

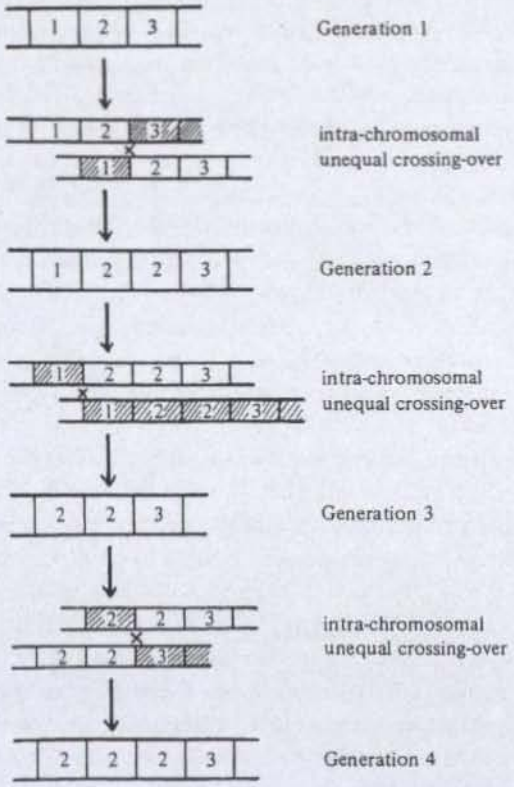
lost it. In other words, higher heterozygosity and liability to loss of duplicated gene expression go together, in agreement with the prediction of the neutral theory.

Very similar results were also obtained using data on 26 species of catostomid fish for 12 enzymes reported by Ferris and Whitt (1977). Allendorf concludes that these results support the neutral theory. Pursuing this problem further, Takahata and Maruyama (1979) investigated mathematically the process by which null alleles become fixed at independent duplicated loci in the course of evolution. Solving the two-dimensional Kolmogorov backward equation numerically, they obtained the rate at which one of the active genes is lost and the amount of heterozygosity at specified time, assuming various selection modes including the cases in which null alleles are completely recessive (only deleterious when doubly homozygous), partially recessive and overdominant. It turned out that in the case of complete recessivity, the time taken for fixation of a null allele at one of the duplicated loci to occur with probability 50% should be much shorter than 50 million years. Also, there should be a large amount of heterozygosity among related species at those loci where loss of function is still in progress. This too does not fit observations. In the case of overdominance, fixation is retarded but the level of heterozygosity remains very high, contrary to observations. The only realistic assumption which is consistent with observations is that null alleles are very slightly deleterious in all heterozygous combinations in which $N_e s$ (the product of effective size and selection coefficient against heterozygotes carrying at least one null allele) is around unity. What is important is that null alleles, despite very slightly deleterious effects, can be fixed in the species by random drift after duplication.

I cannot leave our discussion on the evolutionary role of gene duplication without mentioning the newly proposed concept of 'multigene family' (Hood *et al.*, 1975). It represents a group of tandemly repeated genes which is characterized by the four properties; multiplicity, close linkage, sequence homology, and related or overlapping phenotypic function. Examples are immunoglobulin genes, ribosomal RNA (rRNA) genes, and histone genes. These genes are repeated many times in the genome, and they exhibit two interesting evolutionary features, namely, rapid change in family size and coincidental evolution. The last mentioned phenomenon, that is, coincidental evolution, is also called horizontal evolution. I shall explain this fascinating but puzzling phenomenon taking the spacer regions associated with ribosomal RNA genes in eukaryotes as an example (for more details, see Hood *et al.*, 1975). In the African toad *Xenopus laevis*, the gene for 18S–

28S rRNA is repeated about 450 times at a single chromosomal site (the nucleolar organizer), each repeating unit consisting of 18S, 28S and spacer regions. Essentially the same number of repeats with the same structure are found in the related species *Xenopus mulleri*. However, when nucleotide sequences are compared between these two species, the 18S and 28S regions are shown to be identical, but the spacer regions have diverged significantly. Since the spacer has no known function, it is reasonable to assume, from the standpoint of the neutral theory, that it is subject to less selective constraint and that therefore mutations have accumulated more rapidly by random drift. What is puzzling, however, is that, within each species, sequences of spacer regions are essentially identical. How is such intraspecific homogeneity of the spacer regions maintained while they are evolving rapidly?

Fig. 5.1. Diagram illustrating the process of coincidental evolution through intra-chromosomal unequal crossing-over, leading to fixation of gene 2 by chance in this part of the chromosome.



An ingenious explanation has been put forward by Smith (1974) and Black and Gibson (1974) assuming homologous but unequal crossing over (see also Tartof, 1974). Figure 5.1 illustrates the process by which a multigene family within a chromosome becomes homogeneous under repeated 'intra-chromosomal' (i.e. between-sister-chromatid) unequal crossing-over. As pointed out by Hood *et al.* (1975) as well as by Smith, the process is analogous to random fixation of one of the alleles in a finite population, so that the term 'crossover fixation' has been proposed to designate it. The average time until one of the variant copies of the family becomes fixed in a chromosome lineage is called the crossover fixation time. Smith, and Black and Gibson performed simulation experiments to obtain some quantitative estimates regarding the number of crossovers that are needed to attain a certain level of homogeneity among the family members. Note that intra-chromosomal unequal crossing-over enables a mutation which occurred in one of the members of a multigene family to spread through the family by chance (horizontal spreading of a mutation). This causes nearby members to be more similar genetically than those far apart on the chromosome (for a mathematical treatment, see Kimura and Ohta, 1979). In conjunction with random genetic drift which enables one of the copies of homologous multigene families to fix randomly in the population, coincidental evolution takes place. The mathematical theory of coincidental evolution based on population genetics has been greatly advanced by Ohta (1976a, 1978a, b, 1979, 1980a). The various researches that she has done on evolution and variation of multigene families have recently been presented in an organized form in her excellent monograph (Ohta 1980b). In this monograph, the evolutionary process of a multigene family is treated as a double diffusion process of mutants: diffusion of mutants on one chromosome by unequal crossing-over and diffusion of such chromosomes in the population by random genetic drift. The identity coefficients (probability of identity of genes) within a chromosome and between chromosomes have been studied. Of particular interest is her application of the theory to the problem of amino acid diversity of immunoglobulins, a very well studied example of a multigene family. She found that the framework regions inside the variable region of immunoglobulins evolve slightly faster (in terms of amino acid substitutions) than hemoglobins, whereas hypervariable regions that comprise the antigen binding sites evolve at about the same rate as fibrinopeptides. Furthermore, the high variability at the hypervariable sites within the species is at about the level theoretically expected from the high evolutionary rates at these sites. It is interesting to note that the diffusion equation method and the neutralist

paradigm are useful to elucidate the nature of the multigene family in quantitative terms.

Although the multigene family, in its characteristic form, is represented by large-scale repetitive structures (such as 5S ribosomal gene which is repeated some 24000 times in *X. laevis* and 9000 times in *X. mulleri*), small multigene families are also of much interest from the standpoint of evolution and population genetics. Zimmer *et al.* report (1980) that in five species of apes (chimpanzee, pygmy chimpanzee, gorilla, orangutan and gibbon), the adult hemoglobin α genes exist in a closely linked duplicate state as in humans, and that rapid cycles of gene duplication and loss seem to have been going on in this gene region. That coincidental evolution (these authors try to rename it 'concerted evolution') is actively occurring seems to be evident from the fact that α chains within species are almost identical. These authors estimate that α polypeptides within these species are about ten times more alike than is expected for 10 million years of divergent evolution (which the authors believe to have occurred since the common ancestors of these apes). They also found, based on their analysis using five restriction enzymes, that, in the chimpanzee, a polymorphism exists with respect to the number of α genes: of the five chimpanzee individuals studied, the three-gene state, rather than the two-gene state was predominant. This contrasts with the corresponding polymorphism in humans, where, according to Goossens *et al.* (1980), individuals having triple α loci and also those having a single α locus are a minority, the majority having duplicated α genes. A mathematical theory of coincidental evolution in small multigene families has been worked out by Ohta (1981a) using two models. Applying her theory to the above observations, she estimated that the rate of unequal crossing-over is probably about 10^{-4} per generation and the selection coefficent against chromosomes with three genes or one gene would be roughly 10^{-2}. She also pointed out that non-functional genes could originate by unequal crossing-over and if such genes are selectively neutral, they will accumulate in the population by random drift.

More recently, in addition to unequal crossing-over, gene conversion has been suggested to play an important role in the concerted or coincidental evolution of repeated gene families (Baltimore, 1981). Roughly speaking, one gene conversion results in duplication of a gene copy and, at the same time, causes loss of another. Thus it would be equivalent to one cycle of unequal crossovers of duplication and deletion (Birky and Skavaril, 1976; Ohta, 1977a). The main difference of the effect of conversion from that of unequal crossing-over is that it is not accompanied by contraction or expansion of a multigene family, that is, change of the gene number per family.

When directionality (or 'polarity') exists in gene conversion, the process of divergence of multigene families between species will be much accelerated, Dover (1982) named this and random conversion processes 'molecular drive', and he emphasized that this phenomenon was essential for the evolution of certain coding and non-coding families in *Drosophila* that are dispersed on several chromosomes. Spreading of mutant genes by directional gene conversion within a single genome has been studied mathematically by Nagylaki and Petes (1982), but the theory for treating molecular drive at the level of population is not yet available.

(*v*) *Selective elimination of definitely deleterious mutants and random fixation of selectively neutral or very slightly deleterious mutants occur far more frequently in evolution than positive Darwinian selection of definitely advantageous mutants.*

This is an extended form of the original neutral theory, and is based on the thesis put forward by Ohta (1973a, 1974) which argues that very slightly deleterious mutations as well as selectively neutral mutations play an important role in molecular evolution. Adaptive changes due to positive Darwinian selection no doubt occur at the molecular level (and of course at the phenotypic level), but I believe that definitely advantageous mutant substitutions are a minority when compared with a relatively large number of 'non-Darwinian' type mutant substitutions, that is, random fixations of mutant alleles in the population through the process of random drift of gene frequency. This leads us to an important principle for the neutral theory stating that '*the neutral mutants' are not the limit of selectively advantageous mutants but the limit of deleterious mutants when the effect of mutation on fitness becomes indefinitely small.* This means that mutational pressure causes evolutionary change whenever the negative-selection barrier is lifted.

As an application of this principle, let us consider the high evolutionary rate of guinea pig insulin. Generally speaking, insulin (A and B chains) has a rather low evolutionary rate of $(0.3 \sim 0.4) \times 10^{-9}$/amino acid/per year. However, in the guinea pig, and also, in its close relative the coypu (both belonging to a group called 'hystricomorph rodents'), it is known that very rapid evolutionary change has occurred with a substitution rate almost ten times as high as the above value (see King and Jukes, 1969, and for more data, see Jukes, 1979). From the neo-Darwinian point of view, one might naturally consider such a rapid evolutionary change to be the result of adaptive change by natural selection. In fact, even King and Jukes (1969) in their paper on 'Non-Darwinian evolution' invoked 'positive natural

selection' to explain the rapid change. However, it is much more likely, as pointed out by Kimura and Ohta (1974), that guinea pig insulin lost its original selective constraint in the process of speciation. This allowed the accumulation of mutations which previously would have been rejected. This inference is supported by a report of Blundell *et al.* (1971) who studied the three-dimensional structure of insulin molecules. According to them, in the guinea pig, zinc is absent from the insulin-producing cells (coinciding with the loss of usually invariant histidine B10). This suggests a drastic change in the tertiary structure. It is assumed then that, with the loss of the zinc constraint, mutations in guinea pig insulin started to accumulate at a very high rate. Our claim is consistent also with the observation (for data, see for example, Jukes, 1979) that the rate of evolutionary amino acid substitution in guinea pig and coypu is roughly equal to that of C-peptide of proinsulin, which has a very high evolutionary rate in all cases and which is considered to have much less functional constraint than insulin. Furthermore the observation that guinea pig insulin has a very low biological activity, about 2% of that of bovine, as stated by Blundell and Wood (1975) seems to support the neutralist interpretation. That all hystricomorph rodent insulins are much less biologically active than bovine insulin was recently reported by Horuk *et al.* (1979). Blundell and Wood (1975), however, propose the hypothesis that guinea pig insulin is the result of adaptation to a local shortage of zinc in guinea pig ancestors.

Another interesting case which can be explained by random fixation of very slightly deleterious but almost neutral mutations is the loss of ascorbic acid synthesizing ability in man and a few other species. It was suggested by Jukes and King (1975) that this loss had occurred by a 'neutral evolutionary change'. According to them, the ability to synthesize ascorbic acid (vitamin C) is in general a characteristic of terrestrial vertebrates. However, besides humans, the ability is not present in monkeys, guinea pigs, fruit-eating bats and some passerine birds. These animals consume food rich in ascorbic acid, and it is assumed that the mutants leading to loss of the synthesizing ability were not significantly harmful and became fixed in the species through random gene frequency drift under continued mutation pressure. The alternative hypothesis is that the loss was adaptive, and occurred by positive natural selection. Jukes and King's hypothesis, however, appears to be more plausible since many herbivorous vertebrate species, which consume food high in ascorbic acid, have retained the ability to synthesize it. The observation that the loss occurred in species that are widely scattered in phylogeny also supports their hypothesis. For a mathematical treatment of the average length of time until fixation of a neutral or very slightly deleterious allele in a finite population by random drift under continued mutation pressure, readers may refer to Kimura (1980a).

When we consider the action of natural selection at the molecular level, we must keep in mind that higher order (i.e. secondary, tertiary and quaternary) structures rather than the primary structure (i.e. amino acid sequence) are subject to selective constraint, usually in the form of negative selection, that is, elimination of functionally deleterious changes. The existence of selective constraint, often inferred from non-randomness in the distribution of amino acids or nucleotide bases, does not contradict the neutral theory. That random drift can proceed underneath phenotypic constancy may not be obvious from common sense, but this subtle process can be treated mathematically (see section 6.8).

If mutation pressure is mainly responsible for evolutionary change, the upper limit of evolutionary rate is set by the total mutation rate. In this connection, it is interesting to note that the high evolutionary rate of guinea pig insulin is not very different from the fibrinopeptide rate, and this rate in turn is roughly equal, when expressed in terms of nucleotide substitutions, to the rate of nucleotide substitution in the DNA of the mammalian genome, most of which does not function as genes (see Ohta and Kimura, 1971a). We note also that accumulation of very slightly deleterious mutants by random drift is essentially equivalent to the deterioration of the environment, and definitely adaptive gene substitutions must occur from time to time to save the species from extinction. I have to emphasize again that the above argument does not mean that adaptive changes do not occur at the level of information macromolecules. On the contrary, the marvelous function of molecular machineries on which life depends must be the products of positive Darwinian selection. However, most of the molecules that we are studying now, such as hemoglobins, cytochrome c, transfer RNAs, etc., must have had their essential designs perfected very far in the past; the changes thereafter have been mostly variations on the theme. This is clearly shown, for example, by the fact that all the globin molecules, including hemoglobins of an insect and a marine annelid worm have exactly the same tertiary structure characterized by the 'myoglobin fold' (see Dickerson and Geis, 1969), despite an extensive change in the primary structure. Sometimes, extensive adaptive interpretations of individual amino acid substitutions are given (such as Goodman *et al.*, 1974 for globin evolution), but we need to take these adaptive stories with a pinch of salt.

In general, adaptive mutant substitutions should represent an organism's response to environmental challenge. Thus, there are no such things as universally advantageous mutants; 'advantageous mutants' are a very

special, small class of mutants derived from a very large set of almost universally deleterious mutations. If a paradoxical expression is allowed, we could perhaps say that advantageous mutations are a special form of deleterious mutations. The ways in which mutations become advantageous are so opportunistic that it is not easy to find simple rules to describe them. In this context, a theory due to Ohta (1972) may prove to be important. According to this theory, it is almost impossible for a mutant to find itself advantageous under all conditions required by the environment. On the other hand, if the environment is uniform a mutant will have a better chance to be advantageous. Thus, the probability of a mutant being advantageous to the species as a whole increases in a more uniform environment. Since smaller populations are likely to be distributed over less varied environments, it is expected that rapid evolution at the phenotypic level occurs in species with small population sizes, although their probability of extinction is large.

Although extensive discussion on speciation is outside the scope of this book, random fixation of neutral or slightly deleterious alleles has obvious relevance to the problem of speciation. The central issue here is how reproductive isolation develops. It has been pointed out by several authors (see, for example, Dobzhansky, 1970; Nei, 1975) that for the origin of reproductive isolation, mutational changes at two or more loci are generally required. Roughly speaking, mutant alleles at different loci which are individually harmless (i.e. neutral) but detrimental in combination must be involved.

One simple model suggested by Dobzhansky (1951, 1970) is as follows. Consider a pair of alleles A and A' in the first locus and B and B' in the second. Suppose that a population has the genetic constitution $AABB$ at the start and, then, this population is split into two geographically isolated subpopulations; in one of them, the mutant allele A' becomes fixed by random drift so that a local race $A'A'BB$ is formed, while in the other subpopulation, allele B' becomes fixed, giving rise to a race $AAB'B'$. Under the assumption that A and A' are selectively equivalent (in the absence of B'), random fixation of A' may occur in the first subpopulation unless the frequency of B' happens to be high beforehand. A similar situation may hold for allele B' in the second subpopulation. Note that the cross $A'A'BB \times AAB'B'$, if it occurs, leads to disadvantageous offspring $A'AB'B$ due to epistatic interaction between A' and B', thus contributing to the establishment of reproductive isolation. Later, when the geographical barrier between these two populations is removed, they may be able to evolve as sympatric species.

6

Definition, types and action of natural selection

6.1 Meaning of natural selection

For a proper appraisal of the neutral theory, it is essential to have a correct understanding of the role which natural selection plays in changing and maintaining gene frequencies in natural populations. Therefore, I shall briefly discuss the definition, types and actions of natural selection.

In the Introduction of *The Origin of Species*, Darwin (1859) states:

> As many more individuals of each species are born than can possibly survive; and as, consequently, there is a frequently recurring struggle for existence, it follows that any being, if it vary however slightly in any manner profitable to itself, under the complex and sometimes varying conditions of life, will have a better chance of surviving, and thus be *naturally selected*. From the strong principle of inheritance, any selected variety will tend to propagate its new and modified form.

Although much progress has been made in biology since Darwin's time, his theory of natural selection still remains as the only scientifically acceptable theory to explain why organisms are so well adapted to their environments. A large part of population genetics may be regarded as an attempt to treat the action of natural selection and its consequences in quantitative terms (based on genetics and using the method of biometry).

From the standpoint of genetics, one can define natural selection as the differential multiplication of variant (or mutant) types (Muller, 1960). In other words, we can say that natural selection occurs when there are differences in fitness between genotypes. Here, the term fitness represents the organism's ability to leave offspring to the next generation. For each genotype, fitness is measured by the number of offspring per individual.

Needless to say, a large number of young may die prematurely (as in insects and fishes), so the number of offspring that reach maturity rather than simply those born is relevant. What is important is that we count the number of offspring at the same stage of development as we count the parents. In human populations, fitness may conveniently be measured by the number of daughters per mother. For details on the nature and measurement of fitness, see Crow and Kimura (1970).

Natural selection, however, acts through survival and reproduction of individuals based on phenotypes rather than directly on genotypes. Furthermore, the way in which natural selection acts in individual situations is likely to be so diverse and intricate that the full details are often beyond our comprehension. To overcome such a difficulty and to get a unified view of the working of natural selection, a few fundamental concepts and a great deal of terminology have been proposed.

One simple but useful way of classifying natural selection is to divide it into two distinct types, positive and negative selection. We say that positive selection is at work if a mutant form of a gene (simply called 'mutant') arises in the population and if it enhances the average fitness of individuals carrying it, thereby helping itself to spread through the population. This is the type of natural selection which was assumed by Darwin as the main cause of adaptive evolution in the biological world, and it can justly be called Darwinian selection. The selected gene need not be a new mutant. It could be a previously deleterious gene that has been rendered beneficial by a change in the environment.

On the other hand, if a mutant reduces the fitness of individuals due to its deleterious effects on viability or fertility, then it tends to be eliminated from the population. This type of selection may be called negative selection. The term 'purifying selection' is used by Lewontin (1974) to represent this selection.

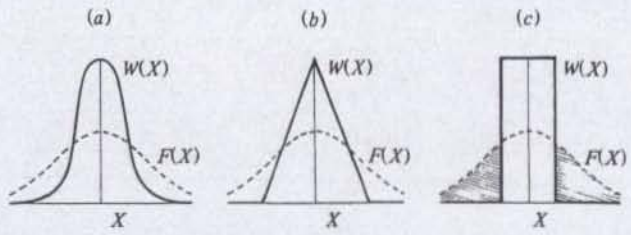
Despite its biological importance, positive selection is seldom observed at work in nature. A few well-known, and constantly cited examples are industrial melanism in month (Kettlewell, 1955, 1956, 1958), DDT resistance in insects and antibiotic resistance in bacteria. As compared with these, examples of negative selection are abundant; it is popular to associate unfavourable effects and deformities with mutations, as so many textbook examples of mutants are of this nature. Furthermore, intensive studies of recessive lethals and detrimental mutants in *Drosophila* populations have shown beyond doubt that the majority of these mutant genes are unconditionally deleterious both in homozygous and heterozygous states (Mukai and Yamaguchi, 1974; Mukai *et al.*, 1972).

Natural selection acts primarily on phenotypes and only secondarily on genotypes (through its effect on phenotypes). The relationship between phenotypes and genotypes, moreover, if often not straightforward, particularly when a phenotypic character is determined by many genes as in quantitative characters (such as height in man and weight in mouse) for which environmental effects also play important roles. It is convenient, therefore, to consider selection at two levels, phenotypic and genotypic.

6.2 Phenotypic selection on quantitative characters

We shall consider phenotypic selection first. Following Mather (1953), selection on quantitative characters is customarily classified into three types, *stabilizing*, *directional* and *disruptive*. Let us assume that the phenotypic measurement (X) follows a certain bell shaped distribution $F(X)$ (such as the normal distribution). Selection is called *stabilizing* if individuals near the mean value of the character have the highest fitness (optimum), and those that deviate in a positive or negative direction are selected against. In other words, it weeds out extremes. A few examples of stabilizing selection are illustrated in Fig. 6.1, where the fitness function $W(X)$ is shown by a solid curve and frequency function $F(X)$ by a broken curve. The importance of stabilizing selection lies in the fact that this is by far the commonest type of natural selection. The best known example in man is the relation between the birth weight and neo-natal mortality studied by Karn and Penrose (1951). They showed that babies whose weights are near the mean have the highest survival rate and those who deviate from this in either direction tend to have higher mortality. For detailed accounts on natural selection in human populations, readers are invited to consult Cavalli-Sforza and Bodmer (1971). For additional

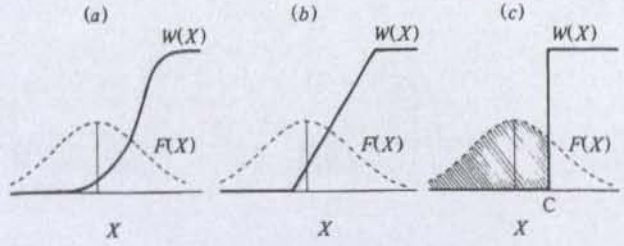
Fig. 6.1. Three examples of fitness function which lead to stabilizing selection. In this type of selection, individuals with a phenotypic value (X) near the mean are favored and those with more extreme values are selected against. In each example, the fitness function $W(X)$ is shown with a solid curve and the frequency function $F(X)$ by broken curve. In example (*c*), individuals corresponding to the shaded area are completely eliminated.



examples of stabilizing selection see, for example, Mather (1973) and Wallace (1968). To represent this type of selection, Haldane (1959) prefered the term *centripetal selection*, originally due to Simpson (1944), because the term stabilizing selection is sometimes used to represent another type of selection, that is, selection for keeping the development of an individual along the normal course in the face of environmental disturbances. This latter type of selection has also been called *canalising selection* by Waddington (1957). On the other hand, the term *normalizing selection* is often used in the same sense as centripetal selection. Stabilizing selection has the characteristic that, after selection, the variance of the character is reduced but no appreciable change occurs in the mean.

Next, let us consider *directional selection*. If the optimum phenotypic value is far removed from the mean, selection tends to shift the phenotypic value toward the optimum. This type of selection is called directional (or directed) selection. It is also called linear selection by Simpson (1944) and Spiess (1977). Darwin postulated this type of selection by extrapolating from artificial selection in order to explain evolutionary change in nature. Now, we know that directional selection occurs much more rarely than stabilizing selection. It must mainly be restricted to the situation when a species is faced with a new environmental challenge. The best authenticated case is industrial melanism in Lepidoptera. Artificial selection, such as selective breeding for milk yield, still gives us much useful information in considering the process of directional selection. Fig. 6.2 depicts three examples of fitness functions (shown by solid curves), which lead to directional selection. In this figure, (*c*) represents 'truncation selection' which is familiar in animal breeding. The point C in the abscissa is the truncation point; only those individuals whose phenotypic value is greater than C are saved for breeding. In this figure, the shaded area represents the fraction of individuals that fail to leave offspring. In nature, however, such a

Fig. 6.2. Three examples of fitness functions which lead to directional selection. Solid lines represent fitness functions and broken curves represent frequency functions. In example (*c*) individuals corresponding to the shaded area are completely eliminated.



clear cut truncation is unlikely to occur. A more likely situation is represented by (*a*) or (*b*). Truncation selection is known to be the most efficient form of directional selection. What is surprising is that when this is modified so that the fitness increases more or less linearly over a range of one or two standard deviations of the value of the selected character, although the efficiency is reduced, the reduction is slight, as shown by Crow and Kimura (1979).

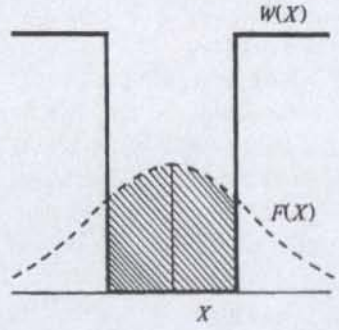
The third type of phenotypic selection, that is disruptive selection, occurs when two or more different optima exist among individuals in one population. An example of this type of selection is illustrated in Fig. 6.3, where the two extremes of the distribution are favoured in selection, but the intermediates are eliminated. This may occur if the habitat of a species contains two distinct niches and each of the two extremes is fitted to one of the two niches. This is an opposite situation to centripetal selection; therefore, it may be called centrifugal selection. According to Haldane (1959), centrifugal selection occurs, but rarely except after the crossing of two species.

6.3 Genotypic selection

A straightforward quantitative approach to the effect of natural selection on genetic change in natural populations is to bypass phenotypic selection and assign fitness values directly to each genotype, and then compute the change of gene frequencies.

The situation is particularly simple if a pair of alleles, say, A_1 and A_2, are segregating in a very large haploid population with discrete, non-overlapping generations. Let W_1 and W_2 be the fitnesses of A_1 and A_2, and

Fig. 6.3. An example of disruptive (centrifugal) selection. Individuals at the two tails of distribution are favored, and those with intermediate phenotypic values (corresponding to individuals in the shaded area) are completely eliminated.



let p be the frequency (more precisely, the relative frequency) of A_1, so that $q = 1 - p$ is the frequency of A_2. After one generation of selection, the frequency of A_1 changes from p to p' according to the relation

$$p' = pW_1/\bar{W}, \tag{6.1}$$

where $\bar{W} = pW_1 + qW_2$ is the mean fitness of the population. The change from one generation to the next $(p' - p)$, as usually denoted by Δp is

$$\Delta p = (W_1 - W_2)pq/\bar{W}. \tag{6.2}$$

It is often convenient to consider the selection coefficient: let $1 + s$ and 1 be the relative fitnesses of A_1 and A_2 so that $1 + s = W_1/W_2$. Then, we get

$$\Delta p = sp(1 - p)/\bar{w}, \tag{6.3}$$

where $\bar{w} = 1 + sp$, and s is the selection coefficient for A_1. If the absolute value of s is small, the denominator $\bar{w}$ is very near unity, so that it may be neglected in equation (6.3). Furthermore, it can be shown (Kimura, 1978) that if the population number regulating mechanism is taken into account, and if one of the alleles (A_1) is advantageous over the other ($s > 0$) so that it increases from a low frequency to a high frequency, then in the process, the approximation formula $\Delta p = sp(1 - p)$ is valid under wider circumstances than the above treatment suggests. Generally speaking, unless the selection is very intense, the process of deterministic change in gene frequency in a haploid population can be described with sufficient accuracy by the differential equation

$$\frac{dp}{dt} = sp(1 - p), \tag{6.4}$$

where t stands for time. In this equation, t is a continuous variable, but it may be measured taking one generation as the unit.

It is often convenient to apply the logit transformation to the gene frequency p, so that

$$z = \log_e \frac{p}{1 - p}. \tag{6.5}$$

Note that the new variable z is the natural logarithm of the gene ratio. It changes continuously from $-\infty$ to $+\infty$ as p changes from 0 to 1. Then, equation (6.4) reduces to

$$\frac{dz}{dt} = s(t), \tag{6.6}$$

where $s(t)$ is the value of s at time t. This can readily be solved, and z is expressed as a function of time t.

$$z_t = z_0 + \int_0^t s(\tau)\,d\tau. \tag{6.7}$$

Note that if $s(\tau) = s$, a constant, then $z_t - z_0 = st$ or $t = (z_t - z_0)/s$. Note also that the frequency of allele A_1 can be obtained by using the relation,

$$p_t = e^{z_t}/(e^{z_t} + 1). \tag{6.8}$$

As an example, let us assume that A_1 has a constant selective advantage of 1% over A_2 so that $s(t) = 0.01$, and ask the question; how long does it take for the gene frequency to change from 0.5% to 99.5%? Since, this corresponds to change of z from -5.293 to $+5.293$, the answer is 10.586/0.01 or 1058.6 generations. Such a unidirectional change in gene frequency may also be called 'directional selection', using the same terminology as used to describe an analogous situation in phenotypic selection.

Equation (6.6) can also be used to describe the process of 'genic selection' in a random mating diploid population. This is a type of selection in which there is no dominance in fitness, namely, the heterozygote has a fitness exactly intermediate between those of the two homozygotes. In this case, we may denote the relative fitnesses of A_1A_1, A_1A_2 and A_2A_2 as $1 + s$, 1 and $1 - s$. Then, assuming that $|s|$ is small, this leads to the same equation as (6.6) as a good approximation.

Generally speaking, however, the situation is much more complicated in a diploid population even under random mating. I shall briefly summarize a few important cases to facilitate discussion in the remaining part of this book. For more details, and also for treatments of various cases, readers may consult Crow and Kimura (1970) and Wright (1969).

Consider a random mating, diploid population, and assume that a pair of alleles A_1 and A_2 are segregating. We denote the relative fitnesses to A_1A_1, A_1A_2 and A_2A_2 as w_1, 1 and w_2, (We use the lower case letter w to denote the relative fitness. The capital letter W is reserved to represent the absolute fitness.) If A_1 is unconditionally advantageous over A_2 so that $w_1 \geqq 1$ and $w_2 \leqq 1$ (but excluding the neutral case of $w_1 = w_2 = 1$), then A_1 increases at the expense of A_2. Unless countered by mutation, A_2 will eventually be lost from the population. Similarly, if A_2 is advantageous over A_1, then A_2 increases at the expense of A_1. These two cases represent directional selection at the genotypic level.

We shall next consider some cases in which Δp due to selection changes sign in the interval (0, 1) so that an internal equilibrium point exists. One of the best known examples is overdominance, where the heterozygote has higher fitness than either homozygote ($0 \leqq w_1 < 1$ and $0 \leqq w_2 < 1$). Let $w_1 = 1 - s_1$ and $w_2 = 1 - s_2$; then s_1 is the selection coefficient against A_1A_1 and s_2 is that against A_2A_2 ($s_1 > 0$, $s_2 > 0$). In terms of these selection

coefficients, the change of the frequency of A_1 in one generation is

$$\Delta p = \{s_2 - (s_1 + s_2)p\}p(1-p)/\bar{w}, \tag{6.9}$$

where

$$\bar{w} = 1 - s_1 p^2 - s_2(1-p)^2. \tag{6.10}$$

The internal equilibrium value of p which will be denoted by $\hat{p}$ and which satisfies $\Delta p = 0$ is obtained by putting $s_2 - (s_1 + s_2)p = 0$. This yields

$$\hat{p} = s_2/(s_1 + s_2). \tag{6.11}$$

In terms of this equilibrium frequency, equation (6.9) is expressed as

$$\Delta p = -(s_1 + s_2)(p - \hat{p})p(1-p)/\bar{w}. \tag{6.12}$$

From this, it is easy to see that if p is larger than $\hat{p}$, Δp is negative, but if p is less than $\hat{p}$ then Δp is positive. This means that if the gene frequency departs from its equilibrium value, then it tends to return to that point. In other words, the equilibrium is stable. Ever since this was demonstrated by Fisher (1922), overdominance has been considered by many geneticists as the main factor maintaining genetic variability within species. However, well established examples of overdominance, or a single-locus heterosis in fitness, are scarce. So far, sickle cell anemia in man is probably the best example, and this is constantly being cited.

The hemoglobin mutant Hb^S responsible for sickle cell anemia differs from the normal allele Hb^A by having valine at position 6 of the β-chain instead of glutamic acid in Hb^A (Ingram, 1963). The Hb^S allele has a selective disadvantage by causing severe anemia in the homozygote, but gains advantage by producing resistance to malaria in $Hb^S Hb^A$ heterozygotes (Allison, 1955). By applying terminologies defined for phenotypic selection, we may call selection which maintains overdominance equilibrium stabilizing selection at the genotypic level. Note that $\bar{w}$ as a function of p takes its maximum value at $\hat{p}$ given by (6.11). So the overdominance equilibrium represents a local maximum of the fitness function.

In the opposite situation in which the two homozygotes show more advantage than the heterozygote ($w_1 > 1$ and $w_2 > 1$). the equilibrium frequency is again given by (6.11), but it is unstable. This can immediately be

Table 6.1. *A simple model of frequency-dependent selection*

Genotype	A_1A_1	A_1A_2	A_2A_2
Fitness	$1 + a - bp$	1	$1 - a + bp$
Frequency	p^2	$2p(1-p)$	$(1-p)^2$

seen by noting that in equation (6.12), if p is higher than $\hat{p}$, Δp is positive, whereas if p is less than $\hat{p}$, Δp is negative. In either case, deviation from equilibrium will be enlarged in each succeeding generation. Thus, if the frequency of an allele is lower than its equilibrium value, the allele will eventually be lost from the population, whereas if it is higher, it will eventually be fixed. The term, disruptive (or centrifugal) selection may be applied to this case of heterozygote disadvantage (i.e. negative overdominance).

A type of frequency-dependent selection in which each allele becomes advantageous when rare ('minority advantage') has long been considered an important mechanism for maintaining genetic variability within species (Wright, 1949; Haldane, 1954; Clarke and O'Donald, 1964). This type of selection will arise, among other ways, if there is division of labor among different genotypes.

As a simple model of frequency-dependent selection, let us assume, following Wright and Dobzhansky (1946), that the relative fitnesses of A_1A_1, A_1A_2 and A_2A_2 are expressed as $1 + a - bp$, 1 and $1 - a + bp$, where a and b are positive constants ($b > a > 0$), and p is the frequency of A_1 (see Table 6.1). Under random mating, the change of gene frequency per generation by selection is

$$\Delta p = p(1-p)(a - bp)/\bar{w}, \tag{6.13}$$

where

$$\bar{w} = 1 - (a - bp)(1 - 2p). \tag{6.13a}$$

In this model, A_1 tends to increase if its frequency (p) is less than a/b, but tends to decrease if it is higher than a/b. Clearly $\hat{p} = a/b$ is the stable equilibrium point. Furthermore, at this equilibrium state, all three genotypes have equal fitnesses. Note that this is not an intrinsic property of frequency-dependent selection in general. It holds only when the heterozygote is exactly intermediate between the two homozygotes in fitness. In fact this property is lost in a slightly more general form of frequency-dependent selection which I shall treat later (see Table 6.2).

6.4 Other terminologies and concepts for natural selection

The term 'balancing selection' was proposed by Dobzhansky (1965) to represent a class of selection which leads to genetic heterogeneity or polymorphism. It includes overdominance and frequency-dependent selection with minority advantage which we discussed above. Dobzhansky also speaks of 'diversifying selection' as a form of balancing selection, and says that some balanced polymorphisms are maintained by such selection.

According to him, this term is synonymous with disruptive selection (see Dobzhansky, 1970, p. 127), and occurs when different genotypes are favored in different ecological niches. I have to remark, however, that this type of selection does not necessarily lead to balanced polymorphisms, and often special conditions have to be satisfied for polymorphisms to be maintained (see, for example, Maynard Smith and Hoekstra, 1980).

Temporal change in the intensity and the direction of natural selection either accelerates fixation or retards it (maintains polymorphism) depending on the situation. An elegant model pertinent to this topic was investigated by Haldane and Jayakar (1963). Let A and a be a pair of alleles at an autosomal locus and assume that allele a is completely recessive with respect to fitness. Assuming random mating, they showed that for biologically realistic circumstances the condition for polymorphism to be maintained is that the arithmetic mean of the fitness of the recessive relative to the dominant is larger than unity, and that its geometric mean is less than unity. For example, if the recessive is normally 5% to 10% fitter than the dominant, but an epidemic disease kills off all the recessives every twenty generations, this condition may be fulfilled. During each epidemic the frequency of the recessive gene decreases suddenly, and then it steadily increases until the next epidemic attacks. For more recent studies on this type of problem, readers are invited to consult Karlin and Lieberman (1974). Incidentally, it is advisable not to include the cases in which selection coefficients fluctuate with time (such that different alleles or genotypes are favored in alternate generations) in the category of disruptive selection.

In addition, there are a number of terminologies which have been proposed to focus our attention on various aspects of the action of natural selection. Wallace (1968) uses the terms 'hard' and 'soft' selection in describing the way in which the proportion of breeding individuals is determined. This may be explained most simply by using the case of truncation selection as depicted in the rightmost diagram (*c*) of Fig. 6.2, where individuals with a phenotypic value (X) larger than the 'cut-off point' C survive to leave offspring, the rest being eliminated. Under hard selection, this point is fixed, and therefore the proportion of the population to survive and reproduce fluctuates as the location of the frequency distribution (broken curve) fluctuates; if the distribution moves to the left, the surviving proportion decreases, whereas if the distribution moves to the right, this proportion increases. On the other hand, under soft selection, the cut-off point can move as the location of the distribution curve fluctuates in such a way that the surviving fraction is kept constant. Therefore, under soft selection, the average fitness of the population remains the same even if the genetic constitution of the population changes drastically. This type of selection does not apply to recessive lethals that unconditionally kill homozygotes. However, if we consider the fact that weak competitors could still survive if strong competitors are absent, soft selection may not be uncommon in nature.

MacArthur and Wilson (1967), in their well-known work on island biogeography, consider two types of natural selection which they termed r selection and K selection. Let N be the population number and suppose that its rate of change follows the equation

$$dN/dt = rN(K-N)/K, \tag{6.14}$$

where r is the intrinsic rate of population increase and K is the carrying capacity. If N is much smaller than K, the population increases exponentially at the rate r, but as N approaches K, the rate of growth diminishes until a steady state is reached, where $N = K$. Equation (6.14) is often called the logistic equation. In an uncrowded environment, the fitness is largely determined by r so that types which harvest the most food (thus having the largest rate of increase), even if they are wasteful, will be favored. This situation is called r selection. Evolution here favors 'productivity'. In a crowded environment, on the other hand, the carrying capacity may be more important. Genotypes which can replace themselves at the lowest food level will win. This situation is called K selection. Evolution here favors 'efficiency' of conversion of food into offspring. The concept of K vs. r selection has been extremely popular among ecologists. For example, Roughgarden (1979), assuming a 'tradeoff' between r and K, discusses the behavior of a population consisting of a pair of alleles with different r and K values. On the whole, however, the concept seems to me to have been overused without much evidence.

Although various types of selection can be envisaged and their roles have been argued, it is not easy to estimate the intensity of selection involved. One important point is to clarify quantitatively how much selection is allocated to individual loci when the overall phenotypic selection intensity is given. This problem will be taken up in the next section.

6.5 Genetic load

As a measure of genotypic selection intensity, genetic load is defined by Crow (1958) as the fraction by which the population fitness is reduced in comparison to the optimum (i.e. the most fit) genotype. In symbols, this is expressed as

$$L = \frac{w_{op} - \bar{w}}{w_{op}}, \tag{6.15}$$

where $\bar{w}$ is the mean fitness of the population and w_{op} is that of the optimum

genotype. In the following, I shall summarize some important results obtained from the studies of the genetic load [for more details, readers may refer to Crow and Kimura (1970, p. 297), and Kimura and Ohta (1971b, p. 44)].

The mutational load (L_{mut})

This is the load created by elimination of recurrent harmful mutations. Consider an autosomal locus in a very large, random mating diploid population. It was first shown by Haldane (1937) that the decrease of fitness due to completely recessive deleterious mutations is equal to the mutation rate v (per gamete). This means $L_{mut} = v$ in this case. For semidominant mutations, $L_{mut} = 2v$, that is, the load is equal to the mutation rate per zygote. What is important in these cases is that the load is almost independent of the degree of the harmfulness of the mutants. Essentially the same conclusion was derived independently by Muller (1950) using the concept of 'genetic death'. More generally, if fitnesses of mutant heterozygotes and homozygotes are $1 - hs$ and $1 - s$ relative to the normal homozygote (whose fitness is assumed to be unity.), L_{mut} changes from v to $2v$ as h changes from 0 to 1. It can be shown, however, that L_{mut} is very close to $2v$ if sh^2 is much larger than $2v(|1 - 2h|)$. Since for lethal and semilethal genes, the mutation rate and degree of dominance are roughly $v = 10^{-5}$ and $h \approx 0.02$, the formula

$$L_{mut} = 2v \tag{6.16}$$

holds as a good approximation to describe the mutational load in natural populations. In other words, the mutational load is equal to the mutation rate per zygote. This rule was called the Haldane–Muller principle by Crow (1957).

More generally, we can treat the mutational load as a function of s, h and v (Kimura, 1961). Also, it can be shown that the mutational load is unaltered even if the population-regulating mechanism is taken into account (Kimura and Ohta, 1971b, p. 48). However, there are a few circumstances under which the Haldane–Muller principle needs some modification. Of particular importance is the case of 'reinforcing type' epistasis between loci. This is the case in which the simultaneous existence of two or more mutant genes within one individual causes a more deleterious effect than the sum of the deleterious effects of the individual mutants. It was shown by Kimura and Maruyama (1966) that if the deleterious effect is proportional to the square of the number of mutant genes, the load under random mating becomes roughly half as large as in the case of

no epistasis. One should expect that with stronger reinforcing epistasis, the load would become smaller. More recently, Crow and Kimura (1979) developed a theory which enables us to compute the efficiency of truncation selection, and they applied the theory to viability mutants in *Drosophila*, showing that the actual load involved can indeed be smaller. Previously, it was shown by Mukai (1964) that mutants causing a small decrease in homozygous viability (which he called 'viability polygenes') occur at a much higher rate than lethals and semilethals, amounting to about 0.4 per gamete per generation. Each of these mutants reduces viability by about $2 \sim 3\%$ on the average in the homozygous state. They have a much higher degree of dominance than lethals (with $h \approx 0.4$ for newly arisen mutants) and are approximately additive in effect between loci. If we apply the Haldane–Muller principle, the mutational load due to viability polygenes alone should amount to 0.8. Under truncation selection (or reasonable departure from it as shown by (*a*) and (*b*) of Fig. 6.2), however, the load can be significantly reduced. Let us choose as the underlying phenotypic value X the pleiotropic heterozygous effects of the viability mutants. This choice is reasonable since these mutants exist mostly in heterozygous condition (because of low frequencies of individual mutants), and their effects on fitness include not only viability but also fertility. Here X is the underlying scale on which truncation selection acts, and it corresponds to 'fitness potential' of Milkman (1978). The selection coefficient (s) against the mutant and the standard deviation (σ) of the fitness potential can be estimated experimentally. They are approximately $s = -0.02$ and $\sigma = 0.02 \sim 0.04$. Let A be the average effect of a single mutant on fitness potential, and suppose that 10% truncation selection ($\bar{w} = 0.9$) occurs. Then it can be shown that $A = -0.002 \sim -0.005$. This means that with truncation selection, mutants with an effect on fitness of $0.1 \sim 0.4\%$ are eliminated from the population as efficiently as if they had an effect of 2.0% but were eliminated by additive selection. Truncation selection eliminates the mutants in groups so that the load becomes much smaller than expected from the Haldane–Muller principle. Although truncation selection in the strict sense is unlikely to occur in nature, selection of types (*a*) and (*b*) of Fig. 6.2, which we may call *quasi-truncation selection*, are sufficiently realistic, and have an efficiency not very far from truncation selection.

The Haldane–Muller principle also requires modification in a small population in which mutant frequencies tend to drift away from their equilibrium values. This problem was investigated by Kimura, Maruyama and Crow (1963) who showed that in very small populations the load is considerably larger than in a large population.

The segregational load

This load arises when the optimum genotype is not fixable. The most important example is overdominance. Consider a random mating population and suppose that a pair of alleles A_1 and A_2 are segregating with respective frequencies p and $q(=1-p)$. Let $1-s_1$, 1 and $1-s_2$ be the relative fitnesses of the three genotypes A_1A_1, A_1A_2 and A_2A_2, respectively, so that s_1 and s_2 are the selection coefficients against two homozygotes, A_1A_1 and A_2A_2. As shown already (see equations 6.9–6.12), the equilibrium frequency $p=s_2/(s_1+s_2)$ is stable if $s_1>0$ and $s_2>0$. Then, substituting this in equation 5.10a and noting $w_{op}=1$, we obtain for the segregational load

$$L_s = s_1 s_2/(s_1+s_2). \tag{6.17}$$

Since the selection coefficients s_1 and s_2 are ordinarily expected to be much larger than the mutation rates, the segregational load due to overdominance is also expected to be much larger per locus than the corresponding mutational load. When multiple overdominant alleles A_1, $A_2,\ldots,A_n$ are segregating, and if all the heterozygotes have the same fitness but homozygote A_iA_i has the relative fitness $1-s_i$ $(i=1,2,\ldots,n)$ as compared with the fitness of the heterozygotes, then the segregational load becomes $L_s=\bar{s}/n$, where $\bar{s}$ is the harmonic mean of the n selection coefficients (see Crow and Kimura, 1970, p. 305: Kimura and Ohta, 1971b, p. 56). Then, if n is large, the load can be small. However, in a finite population, this number is limited by random drift even if an infinite sequence of mutually heterotic alleles may be produced by mutation. Kimura and Crow (1964) investigated the effective number of overdominant alleles under a given mutation rate and effective population size, using the diffusion equation method and assuming that whenever mutation occurs, it leads to a new, not pre-existing overdominant allele. They arrived at the conclusion that overdominance is not very efficient as a mechanism for keeping a large number of alleles per locus in a finite population, and that the load per locus due to overdominance becomes sufficiently large that it is unlikely that a very large number of segregating loci can be maintained each with a significant overdominance effect (assuming independent multiplicative effect to individual loci on total fitness). As I discussed in chapter 2, to avoid the difficulty of the heavy segregational load, models assuming truncation selection were proposed by King (1967), Sved, Reed and Bodmer (1967) and Milkman (1967). Essentially the same type of models but with some refinement were again invoked by Sved (1976), Milkman (1978) and Wills (1978) to explain the maintenance of protein polymorphisms by

overdominance. They assume that the fitness character is proportional to the number of heterozygous loci, and argue that truncation selection or a reasonable deviation from it, in conjunction with 'soft selection' can avoid the difficulty of the heavy load and still maintain polymorphism at thousands of loci. In particular, Wills (1978) uses the term 'rank-order selection' to designate the type of natural selection involved. He assumes that the fitness potential is directly proportional to the average heterozygosity, and that fitness itself is determined by the relative ranking of the organism's fitness potential and by the severity of the environment. He claims that rank-order selection is capable of maintaining all genetic polymorphisms. According to his treatment, the number of polymorphisms that can be retained at a particular selective equilibrium increases as the square of the population size. It is highly unrealistic, I think, to assume, as Wills does, that rank-order selection based solely (or predominantly) on the number of heterozygous sites in the genome, even including the length polymorphism in the non-transcribed spacer regions of *Xenopus* rRNA genes (Reeder *et al.*, 1976), is constantly at work to shape the genetic structure of natural populations. One needs unconditional faith in the benefit of heterozygosity *per se* to maintain such a view. The paucity of overdominance variance as compared with additive genetic variance in viability in various populations of *Drosophila melanogaster* (Mukai, 1977), among other things, speaks strongly against such a view (see also section 9.2). Also, the theoretical results suggesting that the level of polymorphisms increases as the square of the population size simply disagree with observations (see, for example, Soulé, 1976).

Genetic load under frequency-dependent selection

It is often claimed (see for example, Thoday, 1976) that frequency-dependent selection with minority advantage is not accompanied by any genetic load. To see if such a claim is generally true, let us consider a simple model of frequency-dependent selection as shown in Table 6.2, where s_1, s_2, c_1 and c_2 are positive constants, and p and q are the frequencies of A_1 and

Table 6.2. *A model of frequency-dependent selection*

Genotype	A_1A_1	A_1A_2	A_2A_2
Fitness	$1-s_1(p-c_1)$	1	$1-s_2(q-c_2)$
Frequency	p^2	$2pq$	q^2

A_2. This model (Kimura, 1960b) is based on the consideration that there are two environments E_1 and E_2 in the habitat of a species and that A_1A_1 individuals are well adapted to E_1 but not to E_2, and, similarly A_2A_2 individuals are well adapted to E_2 but not to E_1. It is also assumed that the heterozygous (A_1A_2) individuals are uniformly adapted to both environments. Under random mating, the mean fitness of the population is

$$\bar{w} = 1 - s_1p^2(p - c_1) - s_2q^2(q - c_2), \tag{6.18}$$

where $q = 1 - p$, and the gene frequency equilibrium is attained if the average fitnesses of A_1 and A_2 are equal so that

$$s_1p(p - c_1) = s_2q(q - c_2). \tag{6.19}$$

To simplify our analyses, we shall assume in what follows that the two coefficients s_1 and s_2 are equal ($s_1 = s_2$), and they will be denoted by s. From equation (6.19), the equilibrium frequency can readily be obtained,

$$\hat{p} = (1 - c_2)/(2 - c_1 - c_2). \tag{6.20}$$

In what follows, the hat (^) on various symbols means that they represent equilibrium values. The change of the frequency of A_1 in one generation is

$$\Delta p = spq[(1 - c_2) - (2 - c_1 - c_2)p]/\bar{w}, \tag{6.21}$$

where

$$\bar{w} = 1 - sp^2(p - c_1) - sq^2(q - c_2). \tag{6.22}$$

From equation (6.21), it is clear that for the equilibrium to be stable, the condition $2 - c_1 - c_2 > 0$ must be satisfied. In addition, $0 < \hat{p} < 1$ or $0 < 1 - \hat{p} < 1$ for a non-trivial equilibrium to exist. Combining these conditions, we obtain, as the necessary and sufficient condition for the existence of non-trivial stable equilibrium,

$$\left.\begin{array}{l} 0 < c_1 < 1 \\ \text{and} \\ 0 < c_2 < 1. \end{array}\right\} \tag{6.23}$$

The relative fitnesses of A_1A_1 and A_2A_2 at equilibrium are then

$$\hat{w}_{11} = 1 - s(1 - c_1)(1 - c_1 - c_2)/(2 - c_1 - c_2) \tag{6.24}$$

and

$$\hat{w}_{22} = 1 - s(1 - c_2)(1 - c_1 - c_2)/(2 - c_1 - c_2). \tag{6.25}$$

The mean fitness of the population at equilibrium is

$$\hat{\bar{w}} = 1 - s(1 - c_1)(1 - c_2)(1 - c_1 - c_2)/(2 - c_1 - c_2)^2. \tag{6.26}$$

We now consider the following three cases. First, if $c_1 + c_2 = 1$, then all the genotypes have the same fitness at equilibrium ($\hat{w}_{11} = 1 = \hat{w}_{22}$) so that there is no load at equilibrium. This corresponds to the case considered in the previous subsection (see equations 6.13 and 6.13a). Secondly, if $0 < c_1 + c_2 < 1$, then $\hat{w}_{11} < 1$ and $\hat{w}_{22} < 1$, so that the heterozygote has the highest fitness. This means that A_1 and A_2 are overdominant at equilibrium. Then the load is

$$L = s(1 - c_1)(1 - c_2)(1 - c_1 - c_2)/(2 - c_1 - c_2)^2. \tag{6.27}$$

Thirdly, if $1 < c_1 + c_2 < 2$, then $\hat{w}_{11} > 1$ and $\hat{w}_{22} > 2$ so that either homozygote is more fit than the heterozygote. Therefore, A_1 and A_2 show negative overdominance at equilibrium. What is remarkable in this case is that, unlike the case of negative overdominance with constant fitnesses, the equilibrium is stable. Now, let us suppose that $c_1 > c_2$ so that A_2A_2 is the most fit among the three genotypes. Then the load is

$$L = s[(1 - c_2)^2(c_1 + c_2 - 1)/(2 - c_1 - c_2)^2]/\hat{w}_{22}. \tag{6.28}$$

An additional point of interest, as compared with the case of constant fitnesses, is that the gene frequency at stable equilibrium and the gene frequency which gives the maximum $\bar{w}$ are in general different. In fact, the frequency of A_1 which maximizes $\bar{w}$ (see equation 6.22) is

$$\tilde{p} = (3 - 2c_2)/(6 - 2c_1 - 2c_2), \tag{6.29}$$

which is different from $\hat{p}$ given by (6.20).

Drift and dysmetric load

In the above model of frequency-dependent selection (Table 6.2), we have seen that in the special case of $s_1 = s_2$ and $c_1 + c_2 = 1$, all three genotypes have the same fitness at stable equilibrium, and therefore there is no genetic load at the equilibrium. In this case, the model reduces to the one given in Table 6.1. Note that in this special case, the heterozygote is assumed always to have a fitness exactly midway between the two homozygotes (case of 'no dominance'). Even in such a simplified situation, however, it is unrealistic to assume that the load is non-existent: in nature, it is likely that the gene frequencies deviate from their stable equilibrium point due to random drift and this creates a load.

Generally, we can define the drift load by

$$L_{\text{drift}} = [\hat{\bar{w}} - E(\bar{w})]/\hat{\bar{w}}, \tag{6.30}$$

where $\hat{\bar{w}}$ is the mean fitness in an infinitely large population at equilibrium and $E(\bar{w})$ is the expected value of the mean fitness in a finite population. It has been shown by Kimura and Ohta (1970) that in the above model of frequency-dependent selection (Table 6.1), the drift load in a finite popu-

lation of the effective size N_e is

$$L_{drift} = 1/(2N_e). \tag{6.31}$$

What is remarkable is that this load is independent of parameters a and b, and depends only on the effective population size. A similar property of drift load was discovered earlier by Robertson (1970) who showed that, if a pair of overdominant alleles are segregating in a population of effective size N_e, then $L_{drift} = 1/(4N_e)$ independent of the selection coefficients.

One important property of frequency-dependent selection, as we have shown already, is that the gene frequency at stable equilibrium ($\hat{p}$) is not generally equal to the frequency ($\tilde{p}$) which gives the maximum mean fitness $\bar{w}$ (compare equations 6.20 and 6.29). We shall consider this problem a little further using the model given in Table 6.1. In this model, the gene frequency at equilibrium is $\hat{p} = a/b$ and the corresponding mean fitness is $\hat{\bar{w}} = 1$. On the other hand, the gene frequency which maximizes $\bar{w}$ is $\tilde{p} = 1/4 + a/(2b)$, and the corresponding maximum fitness is $\bar{w}_{max} = 1 + (2a - b)^2/(8b)$. This means that natural selection keeps the population fitness short of its possible maximum value. Thus, we can regard this deficiency of population fitness as a load. Apparently, J.B.S. Haldane was the first who noticed this fact. In a personal communication to J.F. Crow (1959, unpublished), Haldane proposed to call this type of load 'dysmetric load'. So, let us formally define the dysmetric load by

$$L_{dys} = (\bar{w}_{max} - \hat{\bar{w}})/\bar{w}_{max}. \tag{6.32}$$

In the above model of frequency-dependent selection, we have $L_{dys} = (2a - b)^2/[8b + (2a - b)^2]$. Wright and Dobzhansky (1946), using the same model to analyse the polymorphism involving ST and CH chromosomes in *Drosophila pseudoobscura* obtained the estimates $a = 0.902$ and $b = 1.288$. Thus, $L_{dys} = 0.0252$ or about 2.5% for this polymorphism.

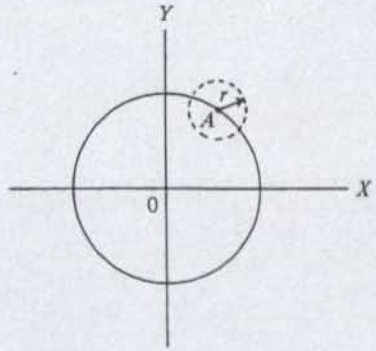
Substitutional load

In the process of substituting one allele for another by positive natural selection (due to a change of environment), elimination of the less fit allele is involved and therefore a genetic load is created. The concept of substitutional load originated with Haldane (1957b) who called it 'the cost of natural selection'. This load has a remarkable property that it is almost independent of the selection coefficient but is largely determined by the initial frequency of the mutant allele involved. Since I have already discussed this topic extensively in chapter 2 in relation to my proposal of the neutral theory (see, section 2.2), I shall not repeat it here, except to mention that the concept is still a controversial one, having been criticized by a number of authors. One popular criticism is that the substitution of a more advantageous allele for a less advantageous one cannot be considered a load, since the fitness of the species is thereby increased. In my opinion, this type of criticism overlooks the important biological fact that for each species the environment, both physical and biotic, is constantly deteriorating, while advantageous mutants are always very rare at the start. There is some possibility, however, that under a certain circumstance the amount of load or cost may not be as large as the Haldane cost principle implies. Nei discusses this problem from the standpoint of the fertility excess required for gene substitution (Nei, 1975, pp. 61–66). Various topics relating to the cost or substitutional load are treated in Crow (1970), Crow and Kimura (1970, pp. 244–52), Felsenstein (1971) and Kimura and Ohta (1971b, pp. 72–88).

6.6 Fisher's model of adaptive process

In order that an organism, or a part, may become well adapted to a given environment, simultaneous adjustment of various component traits is required. Fisher (1930b) explains such a process using, as an analogy, the adjustment of a microscope for distinct vision. This is achieved by slightly changing various parts independently, such as moving each of the lenses a little, either longitudinally or transversely, or twisting through an angle, altering the refractive index, and so on. In more abstract terms, he suggests that the way in which adaptation proceeds may be represented geometrically. In a space of two dimensions, Fisher's model may be illustrated as in Fig. 6.4. In this figure, the degree of adaptation or conformity is measured by the distance of a point A (representing the present state) to a fixed point 0 corresponding to the most fit or optimum state. Any point inside the circle

Fig. 6.4. Illustration of Fisher's model of adaptive process in a space of two dimensions.

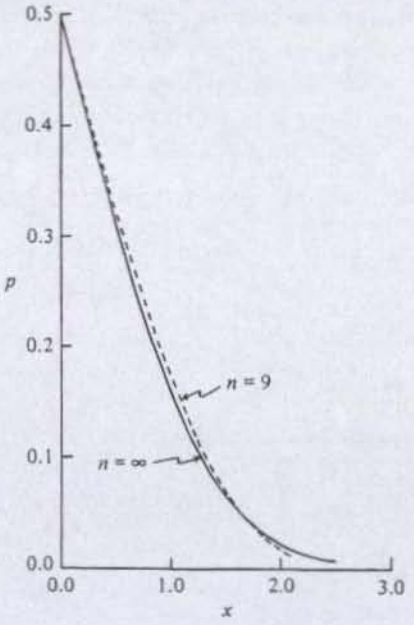


passing through A and centered at 0 represents a state with superior adaptation. If A is shifted through a fixed distance, r, in any direction (by mutation or recombination), the translation will improve the adaptation if it carries the point inside this circle, but will impair it if the new position is outside. If r is very small, we can easily see that the chances of these two events are approximately equal. This means that the chance of improvement in adaptation approaches the limiting value 1/2 as r goes to zero. On the other hand, if r is greater than the diameter (d) of the circle, the chance of improvement is zero. A similar argument can be made in a space of three dimensions, and in this case, according to Fisher, the probability of improvement is $(1-r/d)/2$ for $0<r\leqq d$. It is desirable to obtain the probability of improvement when the number of dimensions is large, for adaptation usually involves conformity in a large number of respects. According to Fisher, when the number of dimension (n) is large, the probability of improvement is

$$p=\frac{1}{\sqrt{(2\pi)}}\int_x^\infty e^{-\lambda^2/2}\,d\lambda, \tag{6.33}$$

where $x=r\sqrt{n/d}$.

Fig. 6.5. Relationship between the probability of improvement p (ordinate) and the magnitude of change x (abscissa) where $x=r\sqrt{n/d}$ and n is the number of dimensions.



I have worked out the same problem using multi-dimensional geometry, and obtained the result that, for $0\leqq r/d\leqq 1$, the probability of improvement is

$$p=\frac{1}{2}I_z(\nu/2,1/2), \tag{6.34}$$

where $z=1-(r/d)^2$ and $\nu=n-1$. In this formula, the function $I_z(\cdot,\cdot)$ is given by the ratio.

$$I_z(a,b)=B_z(a,b)/B(a,b), \tag{6.35}$$

where

$$B_z(a,b)=\int_0^z x^{a-1}(1-x)^{b-1}\,dx \tag{6.36}$$

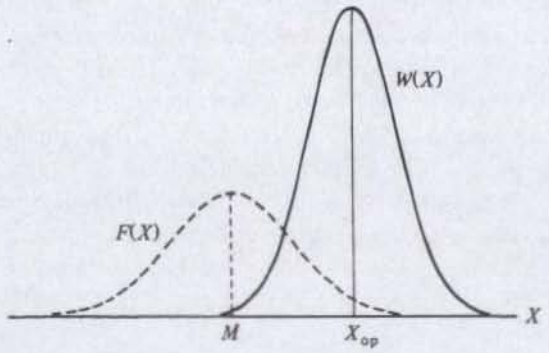
is the incomplete beta function of the first kind, and $B(a,b)$ is Euler's beta function. It can then be shown that if $\nu=n-1$ is large, (6.34) can be approximated by (6.33). Useful formulae relating $I_z(\cdot,\cdot)$ are found in Abramowitz and Stegun (1964). Also, the incomplete beta function is tabulated extensively by Pearson (1934). In Fig. 6.5, the relationship between the probability of improvement (p) and the magnitude of change (x) is illustrated by the solid curve for $n=\infty$ and broken curve for $n=9$. Note that Fisher's formula (equation 6.33) gives a good approximation already for a space of nine dimensions.

6.7 Relation between phenotypic selection and genotypic selection

A general formulation

In order to assess how extensively random drift operates in evolution, it is essential for us to know how much selection is induced

Fig. 6.6. An example of a set of frequency and fitness functions. The frequency function (broken curve) has a mean M and variance σ^2. The fitness function (solid curve) takes its maximum value at $X=X_{op}$.



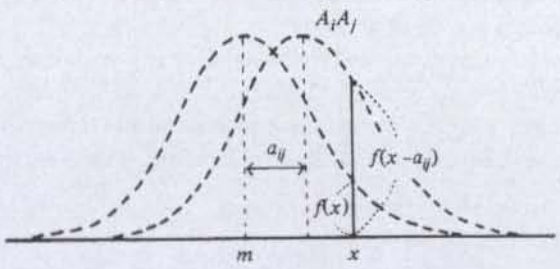
among alleles at individual loci under a given intensity of overall phenotypic selection. Let us consider a quantitative character, such as height, weight, concentration of some substance, or a more abstract quantity which represents Darwinian fitness in an important way. We assume that the character is determined by a large number of gene loci each with a very small phenotypic effect in addition to being subjected to environmental effects (the following results also hold if we consider smaller genetic units such as nucleotide sites rather than conventional gene loci). To simplify the treatment, we also assume that the genes are additive with respect to the character. We use the method of Kimura and Crow (1978) and denote by X the measured phenotypic character with the mean M and the variance σ^2, and by X_{op} the optimum phenotypic value. Let $F(X)$ and $W(X)$ be the relative frequency and the fitness of individuals with character value X (see Fig. 6.6).

Consider a particular locus in which a pair of alleles A_1 and A_2 are segregating with respective frequencies $1-p$ and p. We assume a random mating diploid population, and let X_{ij} be the average phenotypic value of A_iA_j individuals, where $i=1$ or 2, and $j=1$ or 2 ($X_{12}=X_{21}$).

It is often convenient to measure various quantities related to the character value in units of the phenotypic standard deviation (σ). For this purpose, lower case letters will be used. Also, unless otherwise stated, we shall take X_{op} as the origin so that $x=(X-X_{op})/\sigma$. The corresponding frequency and fitness functions will be denoted by $f(x)$ and $w(x)$. We also let $m=(M-X_{op})/\sigma$ and $a_{ij}=(X_{ij}-M)/\sigma$. Note that a_{ij} is the deviation of A_jA_j from the population mean in σ units.

We assume that the background distribution of the character is the same among different genotypes at this locus, and that this is given by $f(x)$ with good approximation, because individual gene effects are assumed to be

Fig. 6.7. Diagram illustrating the meaning of equation (6.37) where a_{ij} is the deviation of the average phenotypic value of A_iA_j individuals from the population mean in σ units.



extremely small. Let w_{ij} be the relative fitness of A_iA_j, then

$$w_{ij}=\int_{-\infty}^{\infty} w(x)f(x-a_{ij})\mathrm{d}x, \tag{6.37}$$

as illustrated in Fig. 6.7. Assuming that a_{ij} is small, we expand $f(x-a_{ij})$ in Taylor series,

$$f(x-a_{ij})=f(x)-a_{ij}f'(x)+(a_{ij}^2/2)f''(x)-\ldots, \tag{6.38}$$

as was done in Kimura and Crow (1978), but here we retain the second-order term, so that, we get, from equation (6.37),

$$w_{ij}=b_0-a_{ij}b_1+a_{ij}^2b_2/2, \tag{6.39}$$

where

$$b_0=\int_{-\infty}^{\infty} w(x)f(x)\,\mathrm{d}x, \tag{6.39a}$$

$$b_1=\int_{-\infty}^{\infty} w(x)f'(x)\mathrm{d}x, \tag{6.39b}$$

and

$$b_2=\int_{-\infty}^{\infty} w(x)f''(x)\mathrm{d}x. \tag{6.39c}$$

Here, the prime denotes differentiation.

Let us denote by a the effect on the character value x of substituting A_2 for A_1. Under random mating the frequencies of the three genotypes, A_1A_1, A_1A_2 and A_2A_2, are respectively $(1-p)^2$, $2(1-p)p$ and p^2, where p is the frequency of A_2. Assuming additive gene effects on x, it is not difficult to show that $a_{11}=-2ap$, $a_{12}=a(1-2p)$ and $a_{22}=2a(1-p)$. Then, using equation (6.39), we can compute the mean fitness of the population, $\bar{w}=w_{11}(1-p)^2+w_{12}2(1-p)p+w_{22}p^2$, and the mean fitness of A_2, $w_2=w_{21}(1-p)+w_{22}p$, and they turn out to be as follows.

$$\bar{w}=b_0+a^2p(1-p)b_2 \tag{6.40}$$

and

$$\bar{w}_2=b_0-a(1-p)b_1+a^2(1-p)b_2/2. \tag{6.41}$$

The change of the frequency of A_2 in one generation is given by the formula

$$\Delta p=p(w_2-\bar{w})/\bar{w}, \tag{6.42}$$

(see, Crow and Kimura 1970, p. 180). Therefore, substituting equations (6.40) and (6.41) into this formula, and neglecting terms involving a^3

and higher order terms, we obtain

$$\Delta p = \frac{p(1-p)}{b_0}[-ab_1 + a^2 b_2(\tfrac{1}{2} - p)]. \tag{6.43}$$

This agrees with Bulmer (1971a). Equating this with the standard formula for the gene frequency change under genic selection, i.e. $\Delta p = sp(1-p)$ (see equations 6.3 and 6.4), we obtain

$$s = -ab_1/b_0 + a^2 b_2(\tfrac{1}{2} - p)/b_0, \tag{6.44}$$

as the selection coefficient which represents the advantage of A_2 over A_1. In this formula, a is the effect of substituting A_2 for A_1 on the character value x.

Application to stabilizing selection

Since stabilizing selection is known to be the most prevalent type of natural selection, we consider the effect of such selection in the following treatments. More specifically, we assume that the frequency function and the fitness functions are given by normal distributions,

$$f(x) = \frac{1}{\sqrt{(2\pi)}}\exp[-(x-m)^2/2] \tag{6.45}$$

and

$$w(x) = \exp(-kx^2). \tag{6.46}$$

Then equation (6.44) reduces to

$$s = -m\lambda a + (\lambda^2 m^2 - \lambda)(\tfrac{1}{2} - p)a^2, \tag{6.47}$$

where $\lambda = 2k/(1+2k)$. An equivalent result was obtained earlier by Bulmer (1972). Note that if we use the original scale (X), and express the density function of the frequency distribution of the character by $F(X) = [1/\sqrt{(2\pi\sigma)}]\exp[-(X-M)^2/(2\sigma^2)]$, the fitness function by $W(X) = \exp[-K(X-X_{op})^2]$, and the effect of allele substitution by A, then the parameters in equation (6.47) are $m = (M - X_{op})/\sigma$, $\lambda = 2K\sigma^2/(1+2K\sigma^2)$, and $a = A/\sigma$.

From equation (6.47), it may be seen that if $|m| \gg |a|$, that is, if the deviation of the mean from the optimum is much larger than the effect of the allele substitution, we have $s/a \approx -\lambda m$ so that the situation is similar to truncation selection (see Milkman, 1978; Crow and Kimura, 1979), and natural selection acts very efficiently to change the mean toward the optimum. During this short period of directional selection, extensive shift of gene frequencies is expected to occur at many loci, but this process itself will seldom cause gene substitutions.

If, on the other hand, the mean is at the optimum ($m = 0$), we have $s = -\lambda(\tfrac{1}{2} - p)a^2$ from equation (6.47). In this case, A_2 is disadvantageous if $p < \frac{1}{2}$, but advantageous if $p > \frac{1}{2}$. This selection is frequency dependent, and alleles behave as if negatively overdominant. The change of gene frequency is then given by $\Delta p = \lambda a^2 p(1-p)(p - \frac{1}{2})$ in agreement with Robertson (1956) and Wright (1935). What is pertinent to our evolutionary consideration is that, here, every new mutation is deleterious, because $\Delta p < 0$ if p is small. Furthermore, if a large number of loci are segregating each with a very small effect, a is small so that a^2 is an exceedingly small quantity. This applies with particular force if we consider individual nucleotide sites rather than the conventional gene loci, for it has been estimated (Kimura, 1974) that the average individual in a large mammalian population is likely to be heterozygous at a million or so nucleotide sites. This result is consistent with Ohta's claim (Ohta, 1973, 1974) that the majority of mutants at the molecular level are nearly neutral but very slightly deleterious. Furthermore, it can be shown (Kimura, 1981b) that negatively overdominant alleles are far more likely to be fixed by random genetic drift than unconditionally deleterious alleles having the same magnitude of selection coefficients. It is interesting and rather unexpected that 'balancing selection' at the phenotypic level leads to the unbalance of the polymorphic state of the genes involved. In this sense the selectionist's claim of balancing selection at the molecular level is wrong in its foundation.

Selection intensity

The intensity of natural selection under stabilizing selection can be measured in terms of the load, L, that is, by the fraction of individuals which are eliminated in each generation by natural selection due to their phenotypic values deviating from the optimum. For the frequency and fitness functions given as equations (6.45) and (6.46), we have

$$L = 1 - \sqrt{(1-\lambda)}\exp(-\lambda m^2/2). \tag{6.48}$$

For $m = 0$, this reduces to

$$L = 1 - \sqrt{(1-\lambda)} \tag{6.49}$$

In the special case in which the fitness function has the same variance as the frequency function, i.e. when $K = 1/(2\sigma^2)$ or $k = \frac{1}{2}$, we get $L \approx 0.293$ or roughly 30% elimination.

In general, L is likely to be small for any single character in mammals. For example, according to Haldane (1959), the intensity of selection acting on birth-weight of babies through their neonatal mortality is $L = 0.027$. If L is small, equation (6.49) may be replaced, with good approximation, by

$$L = \lambda/2. \tag{6.50}$$

For $\lambda \leqq 0.1$, the error involved with this approximation is less than about

2.5%. Even for $\lambda = 0.3$, the error is about 8%. Thus, equation (6.50) may be used for all practical purposes.

Connecting total phenotypic selection with selection at an individual locus

During its lifetime, an individual is subject to natural selection on a large number of quantitative characters, many of which are correlated. Let us assume, to simplify the treatment, that we can choose a certain number, say n_c, of independent characters, which collectively represent, with a first approximation, the total pattern of selection. This may be called the total representative phenotype. Various parameters pertaining to the i-th character will be expressed by subscript $i (i = 1, 2, \ldots, n_c)$.

Since the total selection intensity is limited, the selection intensity, as measured by L_i to which each component character is subject is expected to be small, if the number of characters involved is large. Let L_T be the total selection intensity, then

$$(1 - L_T) = \prod_i (1 - L_i) \tag{6.51}$$

so that

$$L_T \approx 1 - \exp\left(-\sum_i L_i \right). \tag{6.52}$$

To simplify the treatment still further, let us suppose that L_i are all equal among component characters, then $L_i \approx -(1/n_c)\log_e(1 - L_T)$. The selection coefficient per locus (or site) is $-\lambda_i(\frac{1}{2} - p_i)a_i^2$, where $\lambda_i = 2k_i/(1 + 2k_i)$. Noting $L_i \approx \lambda_i/2$, and dropping the subscript i, we get

$$s = [\log_e(1 - L_T)](1 - 2p)a^2/n_c, \tag{6.53}$$

approximately. In the right hand side of this formula, we note that $a^2/n_c = A^2/(n_c\sigma^2)$, where $n_c\sigma^2$ is the variance of the total phenotype.

It may be interesting to inquire how much selection is induced per nucleotide site, if every site contributes equally to the total phenotype. Let $\bar{h}_e$ be the average heterozygosity per site, and if we denote by n_{nuc} the total number of nucleotide sites concerned, then $n_{nuc}A^2\bar{h}_e = n_c\sigma^2\rho^2$, where A is the effect of substituting one nucleotide on a component phenotype and ρ^2 is broad sense heritability. Thus, the coefficient for stabilizing selection, s_s, as defined by the relation $s = -s_s(1 - 2p)$, turns out to be as follows:

$$s_s = -[\log_e(1 - L_T)]\rho^2/(n_{nuc}\bar{h}_e). \tag{6.54}$$

This represents the selection intensity involved in nucleotide substitution under stabilizing selection (assuming $m = 0$). To obtain an estimate of s_s, let

us assume that the average heterozygosity per enzyme locus with respect to electrophoretically detectable alleles is 0.1 and assume (rather conservatively) that in addition there is twice as much heterozygosity with respect to silent alleles. Then, if the average number of nucleotide sites that make up a locus is 10^3, we get $\bar{h}_e = 3 \times 10^{-4}$. Extrapolating this to the total genome of a mammal with 3.5×10^9 nucleotide sites, the average number of heterozygous nucleotide sites per individual is $n_{nuc}\bar{h}_e = 1.05 \times 10^6$. The total selection intensity cannot be very high for mammals with large body size. For the human species, Deevey (1960) estimates that throughout its main course of evolution it had a reproductive pattern such that about half the children ever born have lived to become sexually mature. Thus, as typical values of the load for a mammal (such as the human species), we suppose that $L_T = 0.5$. As to broad sense heritability we tentatively assume that $\rho^2 = 0.5$. Then, we obtain $s_s = 3.3 \times 10^{-7}$, which is a very small selection coefficient for stabilizing selection and shows that the majority of mutations at the molecular level are nearly neutral but very slightly deleterious. This allows extensive random drift to operate on them. This is consistent with Ohta's (1973, 1974) hypothesis of very slightly deleterious mutations. However, the fitness of the species does not drift downward with this model the way it does according to Ohta's hypothesis. Also, in this view, those genes which are substituted by random drift and those that are responsible for the phenotypic variability of quantitative traits belong to the same class. It is possible that many, and even most, of the mutants affecting a quantitative trait are regulatory rather than structural. DNA outside the coding region may be more important from this standpoint than translated DNA. The present analysis agrees with Lande (1980) who suggests that many polygenic changes can accumulate by random drift because they have little or not net phenotypic effect. Needless to say, some sites produce much larger phenotypic effects than others and therefore are subject to stronger selection. On the other hand, a certain fraction of sites (presumably a large fraction) produce no phenotypic effects at all, and therefore are completely neutral with respect to natural selection.

6.8 Behaviour of mutant alleles in a finite population under stabilizing selection

Probability of fixation of a mutant allele under stabilizing selection

Before we proceed to the next chapter treating problems on selective constraint and the rate of evolution, it is desirable to investigate

how extensively random drift operates under stabilizing selection, the most prevalent mode of natural selection. This will help us understand particularly how 'non-random codon usage' (see section 7.5) can be explained in the framework of the neutral theory based on the stochastic theory of population genetics.

We have shown in section 3.5, that the rate of evolution in terms of mutant substitutions is given by $k = 2Nvu$ (see equation 3.19), where N is the population size, v is the mutation rate and u is the probability of a single mutant ultimately reaching fixation in the population. We have also shown, in the previous section (section 6.7) of this chapter, that every mutation is slightly deleterious under stabilizing selection. Thus, in order to see the extent to which random drift operates under stabilizing selection, it is necessary to investigate the probability (u) of fixation of a mutant allele under this mode of selection.

We consider a quantitative character determined by a large number of loci each with a very small effect. Let us assume that a pair of alleles A_1 and A_2 are segregating at a given locus, with relative frequencies, $1-p$ and p, in a random mating, diploid population. In the following, we assume that A_2 is a mutant allele produced from A_1. We denote by N the number of individuals composing the population, and by N_e the effective size (for the meaning of N_e, see section 3.3; roughly speaking, N_e is equal to the number of breeding individuals in one generation). This number is likely to be considerably smaller in most cases than the census number N.

As shown in the previous section, if the frequency function and the fitness function are given by normal distributions (with equations 6.45 and 6.46), and if a is the effect of substituting A_2 for A_1, then (see equation 6.47), we have

$$s = \beta_1 - \beta_2(1-2p), \tag{6.55}$$

where $\beta_1 = -\lambda ma$ and $\beta_2 = \lambda(1-\lambda m^2)a^2/2$, in which $\lambda = 2k/(1+2k)$.

Disregarding mutational change for a moment, let us denote the frequency of A_2 by x (this should not be confused with x in the previous section, where x is used to represent the character value in σ units). Then, the mean and the variance in the change of x during one generation are

$$M_{\delta x} = [\beta_1 - \beta_2(1-2x)]x(1-x) \tag{6.56}$$

and

$$V_{\delta x} = x(1-x)/(2N_e). \tag{6.57}$$

Let $u(p)$ be the probability that A_2 eventually becomes fixed in the population (i.e. reaches 100% in frequency), given that its initial frequency is p. Then, $u(p)$ can be expressed in terms of $M_{\delta x}$ and $V_{\delta x}$ by using a general

formula for the probability of fixation obtained by Kimura (1962), namely, by

$$u(p) = \int_0^p G(x)\mathrm{d}x \Big/ \int_0^1 G(x)\mathrm{d}x, \tag{6.58}$$

where

$$G(x) = \exp\left\{-2\int \frac{M_{\delta x}}{V_{\delta x}}\mathrm{d}x\right\}. \tag{6.59}$$

(see p. 424 in Crow and Kimura, 1970, or p. 173 of Kimura and Ohta, 1971b).

We are particularly interested in the probability of fixation of A_2 when it is initially singly represented in the population. This probability, which is denoted by u, is given by $u(p)$ with $p = 1/(2N)$. We then obtain

$$u = \frac{1}{2N\int_0^1 \exp[-B_1x + B_2x(1-x)]\mathrm{d}x}, \tag{6.60}$$

where $B_1 = 4N_e\beta_1$ and $B_2 = 4N_e\beta_2$.

In the above treatment, we have assumed that m, the deviation of the mean from the optimum [see Fig. 6.6, and note $m = (M - X_{op})/\sigma$], remains unchanged throughout the process. This assumption appears to be unrealistic, because, if $m \neq 0$, one should expect that $|m|$ tends to be reduced by the directional component of selection much more rapidly than the mutant gene drifts toward fixation. There is an important possibility, however, that this change is opposed by mutational pressure so that m remains constant under continued stabilizing selection, although $|m|$ at equilibrium is likely to be small. This occurs when the optimum and the mutational equilibrium point do not coincide. We shall assume such a case when we discuss the problem of 'non-random' synonymous codon usage in the next chapter.

We have already seen that the selection which acts on the mutant allele is frequency dependent and that alleles behave as if negatively overdominant. This suggest that mutants under stabilizing selection are far more likely to be fixed by random drift than unconditionally deleterious mutants having a comparable selection coefficient. To show that this is indeed the case, I list in Table 6.3 some examples of the probabilities of fixation (u) for these two cases. In the case of stabilizing selection, we let $m = 0$ and denote $\lambda a^2/2$ by s_s (meaning selection coefficient for stabilizing selection), so that $\Delta p = -s_s p(1-p)(1-2p)$. For the unconditionally deleterious case, we denote the selection coefficient against A_2 by $-s'(s' > 0)$, so that the probability of

fixation is given by

$$u = S'/[2N(e^{S'} - 1)], \tag{6.61}$$

where $S' = 4N_e s'$ (see p. 426 of Crow and Kimura, 1970). In both cases, u is tabulated (Table 6.3) taking the probability of fixation of the completely neutral case as the unit, i.e. it is expressed as a multiple of $u_0 = 1/(2N)$. It is clear from this table that there is an enormous difference in fixation probability between the two cases, and that under stabilizing selection, extensive neutral evolution is possible even when $4N_e s_s$ is 8 or more. For $B = 4N_e s_s > 8$, it can be shown that $u/u_0 \approx \sqrt{(B/\pi)}\exp(-B/4)$.

Gene frequency distribution

Let us now incorporate mutational pressure and investigate the probability distribution of allelic frequencies at statistical equilibrium attained under stabilizing selection in a finite population. We denote by $\phi(x)$ the probability density such that $\phi(x)dx$ represents the probability that the frequency of A_2 in the population lies in the range $x \sim x + dx$, where $0 < x < 1$. We assume reversible mutations between the two alleles, and let v_1 be the mutation rate from A_1 to A_2, and let v_2 be the rate in the reverse direction. Then the mean and the variance in the change of A_2 in one generation are

$$M_{\delta x} = [\beta_1 - \beta_2(1 - 2x)]x(1 - x) - v_2 x + v_1(1 - x) \tag{6.62}$$

and

$$V_{\delta x} = x(1 - x)/(2N_e). \tag{6.63}$$

We use Wright's (1938) formula for the steady state gene frequency

Table 6.3. *Relative probability of fixation* (u/u_0) *of a negatively overdominant mutant and that of an unconditionally deleterious mutant, where* u_0 *is the probability of fixation of a completely neutral mutant. In this table,* $\bar{S}$ *stands for* $4N_e s_s$ *for the negatively overdominant case and* $4N_e s'$ *for the unconditionally deleterious case.*

$\bar{S}$	Negatively overdominant	Unconditionally deleterious
0	1.00	1.00
1.0	0.84	0.58
8.0	0.23	2.7×10^{-3}
16.0	0.042	1.8×10^{-6}
20.0	0.017	4.1×10^{-8}
30.0	0.0017	2.8×10^{-12}

distribution, that is,

$$\phi(x) = \frac{C}{V_{\delta x}}\exp\left\{2\int\frac{M_{\delta x}}{V_{\delta x}}dx\right\}, \tag{6.64}$$

where the constant C is determined so that

$$\int_0^1 \phi(x)\,dx = 1. \tag{6.64a}$$

(see pp. 434–435 of Crow and Kimura, 1970). Then we obtain

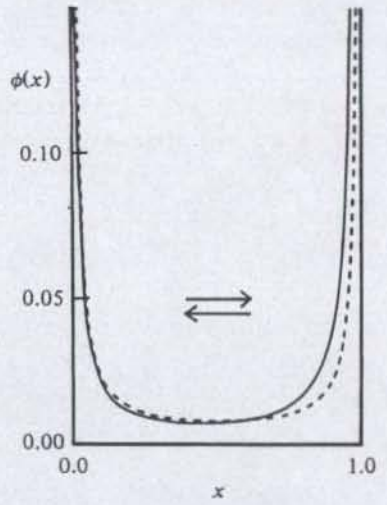
$$\phi(x) = C\, e^{B_1 x - B_2 x(1-x)} x^{V_1 - 1}(1 - x)^{V_2 - 1}, \tag{6.65}$$

where $B_1 = 4N_e\beta_1$, $B_2 = 4N_e\beta_2$, $V_1 = 4N_e v_1$ and $V_2 = 4N_e v_2$.

The probability of A_2 being temporarily fixed in the population may be obtained by integrating $\phi(x)$ from $1 - 1/(2N)$ to 1, and we obtain, with sufficient accuracy,

$$f_2 = Ce^{B_1}/[V_2(2N)^{V_2}]. \tag{6.66}$$

Fig. 6.8. Probability distribution (ϕ) of the frequency (x) of allele A_2. The solid curve represents the distribution under stabilizing selection with $V_1 = V_2 = 4 \times 10^{-3}$, $B_1 = 0.847$ and $B_2 = 2$ (see equation 6.65). The distribution is restricted to the cases in which both A_1 and A_2 coexist in the population ($0 < x < 1$). Actually, for the set of parameters given above, the allele A_2 is either fixed in the population ($x = 1$) or lost from it ($x = 0$) in the great majority of cases. The symmetrical broken line represents the completely neutral case ($B_1 = B_2 = 0$) under the same mutation rate and population size ($V_1 = V_2 = 4 \times 10^{-3}$). Horizontal arrows indicate that during the long course of evolution, turnover from one fixed state (say, A_1 fixed) to another (A_2 fixed) occurs via polymorphic states.



Similarly, the probability of A_1 being temporarily fixed in the population (i.e. A_2 lost) is

$$f_1 = C/[V_1(2N)^{V_1}]. \tag{6.67}$$

Then, the ratio of f_2 to f_1 is

$$f_2/f_1 = e^{B_1}(V_1/V_2)(2N)^{V_1 - V_2}. \tag{6.68}$$

We shall be mainly concerned, in this book, with the situation in which both V_1 and V_2 are much smaller than unity, and alleles are fixed most of the time. This situation is particularly pertinent when we consider an individual nucleotide site rather than a conventional gene locus, because the mutation rate per site must be of the order of 10^{-8} rather than 10^{-5}.

As an example, let us consider a site for the third position of a codon. We group the four bases into two classes, denoting bases A and U as A_1, and C and G as A_2. Then, there are two alleles at this site and the distribution of the frequency of A_2 is given by equation (6.65). Let us assume that the mutation rates are equal in both directions with $v_1 = v_2 = 10^{-8}$ per generation, and that the effective population size is a hundred thousand ($N_e = 10^5$) so that $V_1 = V_2 = 4 \times 10^{-3}$. Under such conditions, the alleles are predominantly in the fixed state. Let us further assume that fixation of A_2 occurs more frequently than A_1 with the ratio $f_2/f_1 \approx 7/3$ as in hemoglobin genes. This leads to $B_1 = \log_e(7/3) = 0.847$ from equation (6.68). As to the effect of a single mutant substitution, we assume a small value, $a = 10^{-2}$ so that $a^2 = 10^{-4}$. This corresponds to 3×10^4 protein-coding genes with an average heterozygosity of 0.17 per locus and a broad sense heritability (ρ^2) of 50%. As to the intensity of stabilizing selection for optimizing the codon usage (see section 7.5) with respect to change at the third codon positions (which are mostly synonymous), its intensity, if any, must be rather small. So, we assume $L = 0.05$ (1/10 of the total load) so that $\lambda = 0.1$. Noting that $\lambda m^2 \ll 1$, we get $B_2 = 4N_e\beta_2 \approx 4N_e\lambda a^2/2 = 2$. Fig. 6.8 illustrates the frequency distribution of the allele A_2 for unfixed classes ($0 < x < 1$), assuming these parameters.

7

Molecular structure, selective constraint and the rate of evolution

7.1 Conservative nature of mutant substitutions

It has been known, since the early work of Zuckerkandl and Pauling (1965), that in protein evolution, amino acids which are similar in structure and chemical properties are substituted more frequently than dissimilar ones. They considered that such 'conservative' substitutions are likely to cause only a small change in the function of the molecule, and are therefore more likely to be 'adopted by natural selection'. They noted, however, that 'apparently chemists and protein molecules do not share the same opinion' regarding what are the most prominent properties of a residue.

So, they assumed that the best criterion for conservatism of a substitution is a high frequency of occurrence during protein evolution. One shortcoming of such a criterion is that arguments easily become circular.

It is desirable, therefore, to use a more objective measure of amino acid similarity, and then investigate whether there is a positive correlation between such similarity and the frequency of evolutionary substitution. Another point about which we have to be careful in such an analysis is that the genetic code has the property, already noted by Zuckerkandl and Pauling (1965), that a single base substitution often leads to a similar amino acid or to a synonymous codon. In other words, the code itself predisposes to conservative mutational changes. Epstein (1967) seems to be the first to have investigated the conservative nature of substitutions by using an objective index to represent the difference between a pair of amino acids. His index (called the coefficient of difference) is based on polarity and size differences, and he compared its distribution for amino acid substitutions in homologous proteins (subtilisin, α and β hemoglobins, cytochrome c, ribonuclease etc.). The results of his study showed that for all the observed

comparisons of homologous proteins, the values of the coefficient of difference were considerably lower than could be accounted for by 'random mutations', even if the conservative nature of the genetic code is taken into account.

Later, Clarke (1970) made a similar study employing Sneath's (1966) index (D) of chemical dissimilarity between amino acids and using Dayhoff's data on the relative frequencies of amino acid substitutions (see Dayhoff and Eck 1968). Taking into account the differences in frequencies and in probabilities of mutational changes (i.e. differences in the number of synonymous codons among amino acids), he found a statistically significant negative correlation between the frequency of substitutions (R) and the chemical dissimilarity (D). This indicates that amino acid substitutions which are accompanied by small chemical changes occur much more frequently in evolution than those accompanied by large ones. He also found that the average substitution rate is only about one-ninth the maximum substitution rate, which occurs at $D = 0$. He claimed that these observations invalidate several 'less critical' studies of protein evolution (in which he includes King and Jukes, 1969 and possibly Kimura, 1968a, 1969a). According to him, 85–90% of mutations causing amino acid changes are disadvantageous, but this is based only on one kind of selective constraint. In reality, he notes, there are many others, such as limitations imposed by the kind of protein, by the position of the amino acid in the protein, and by the type of organism in which the protein is found, and, when all such factors are taken into account, "the average proportion of 'neutral' mutations will be nearer 0.1 per cent than 10 per cent". Jukes and King (1971), in their rebuttal to Clarke, questioned the validity of his statement and also pointed out that if selectively neutral and nearly neutral changes are common in evolution, one should expect exchanges between chemically similar amino acids to occur more frequently than those between dissimilar ones (in agreement with observations). On the other hand, they said, for positive Darwinian evolution to occur, there must be a difference in the function of the changed protein, and one should expect to find radical changes favored over conservative ones. Clarke (1971) in turn replied, referring to Fisher (1930b) who showed that a smaller mutational change has a higher chance of being advantageous (see, section 6.6 in the previous chapter), saying that Fisher demonstrated that selection is very unlikely to favor extreme changes. It seems to me that Clarke overlooked the point that mutations, if their effects are very small, are likely to have very small selective advantages and therefore have correspondingly small probabilities of fixation, i.e. lower chances of spreading through the species.

As I shall explain later, this means that Fisher's argument can not be used to explain in selectionist terms the observation that the smaller the dissimilarity between the amino acids, the higher the rate of their substitutions in evolution.

From the standpoint of the neutral theory, the nature of conservative substitutions can readily be explained by noting that the smaller the difference between two amino acids, the higher the probability that they are selectively equivalent rather than the mutant being harmful. (The neutral theory in its strict form assumes that none is beneficial.) Therefore, the probability of an interchange being selectively neutral is higher between more similar amino acids, and therefore evolutionary substitutions of such amino acids should occur more frequently by random genetic drift. In terms of equation (5.1), the fraction f_0 is larger for mutations leading to a more similar amino acid.

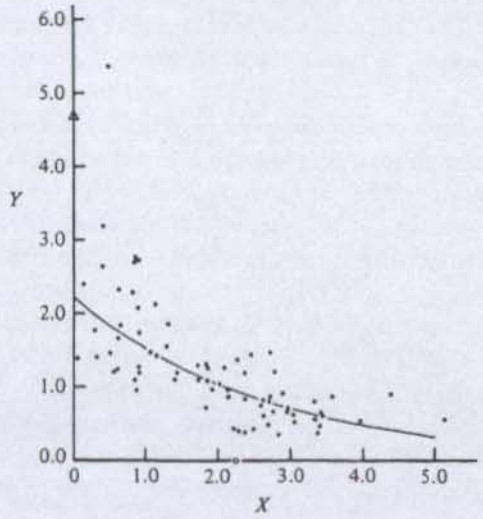
Sometimes the neutral theory is misunderstood to mean that substitutions of various types of amino acids should occur equally frequently (for which the term 'randomness' is commonly applied), and the observed pattern of inequality ('non-randomness') is claimed to be inconsistent with the neutral theory. For example, Grantham (1974) claims that the controversy with respect to the mechanism of molecular evolution is essentially the controversy between 'randomness and physicochemical determinism', and he identifies the neutralist with 'randomness proponent'. To investigate the conservative nature of amino acid substitutions, Grantham (1974) introduced a formula to represent the physico-chemical difference between amino acids similar to what has been done by Epstein (1967) and Clarke (1970). His formula for D_{ij}, giving the distance between two amino acids i and j, is based on three properties, polarity, volume and composition, where composition is defined as the ratio of atomic weights of noncarbon elements in end groups or rings to that of carbons in the side chain. He then investigated the correlation between D_{ij} and the relative substitution frequency (RSF) as given by McLachlan (see Table A1 in Appendix of McLachlan, 1972). It turned out that linear regression of log RSF on D values gave a correlation coefficient of -0.72 which represents a much stronger negative correlation than those obtained by using the Sneath difference and Epstein formula. His conclusion, however, agrees with the previous work in showing that fixation of mutations between dissimilar amino acids is generally rare.

More recently, Miyata *et al.* (1979) proposed a simple distance measure based only on polarity (p) and volume (v), each weighted by the reciprocal of their standard deviations, σ_p and σ_v. According to them, if we treat only

amino acid substitutions that can be attributed to single step mutations, their distance measure gives a high negative correlation with the observed substitution rate, showing clearly the conservative nature of substitutions in evolution. They also examined data on abnormal human hemoglobins, and found a positive correlation between the frequency and amino acid difference. This is understandable since in the case of abnormal hemoglobin variants, amino acid substitutions causing more severe pathological effects are more likely to be detected and described.

The conservative nature of amino acid substitutions has also been investigated by Bogardt *et al.* (1980) for the mammalian myoglobins. Using a 'structural discriminate function' (D_s) based on four properties of amino acids, volume, polarity, isoelectric point and hydrophobicity, and making a residue-by-residue comparison of the known mammalian myoglobins, they found that amino acid substitutions which are compatible with the retention of the original conformation of the protein appear to be favored in evolution. In particular, they found that the coefficient denoted Δ_m, which is essentially the average D_s value per amino acid site, is larger for the external

Fig. 7.1. Relationship between physico-chemical difference (X) and relative substitution frequency (Y) of amino acids in protein evolution. Observed values are plotted with solid dots, and the exponential function ($Y = 2.22e^{-0.376X}$) fitted to the observations is shown by the curve. The triangle on the Y-axis represents the expected frequency of synonymous substitutions estimated by using values obtained through comparative studies of the nucleotide sequences of hemoglobins. For details, see text.



residues than for the interior ones, suggesting that exterior residues allow less conservative (i.e. more radical) substitutions. This is easily understandable because in the functionally less important (exterior) part of the molecule, more diverse amino acids can be tolerated without impairing its working (i.e. can become selectively equivalent or neutral). They consider, however, that the majority of 'neutral' amino acid substitutions are the result of 'selective fixation of acceptable alternative stabilizing mutations'. As I have shown in the previous chapter (see section 6.8), extensive neutral evolution can occur under stabilizing selection, but evolutionary mutant substitutions involved cannot be called 'selective fixation', because fixation of mutants is caused by random drift rather than by selection. In fact, the selection involved is extremely weak negative selection so that random drift can easily override it.

In order to examine in more detail the conservative nature of amino acid substitutions from the standpoint of population genetics, I constructed a graph (Fig. 7.1) showing the relationship between the physico-chemical difference and the relative frequency of amino acid substitutions. In this figure, the abscissa (X) stands for Miyata *et al.*'s (1979) distance and the ordinate (Y) represents the relative frequency of amino acid substitutions as listed in Table A1 of McLachlan (1972). This is based on comparison of 17 homologous families of proteins, including 13 hemoglobins, 8 cytochromes c, 14 antibody light-chain variable regions, etc. It appears to be the most extensive compilation of amino acid substitution frequencies available so far. Note that the substitution frequencies here are normalized by taking into account the fact that different amino acids have different frequencies of occurrence. In Fig. 7.1, observed values are shown by dots and an exponential function $Y(X) = Ae^{BX}$ fitted to these observations is shown by a solid curve. Although there are 190 combinations (amino acid pairs) between 20 amino acids, I have chosen only 75 pairs which involve single mutational steps, that is, pairs which can be mutually converted on the code table through single base substitutions. Also, in fitting the exponential function $Y = Ae^{BX}$ (or, equivalently, the linear function $\ln Y = \ln A + BX$) to the data, I excluded one pair (Cys↔Phe; the open circle in Fig. 7.1) for which no observation is recorded in McLachlan's table A1. The linear regression of $\ln Y$ on X turned out to give $B = -0.376 \pm 0.04$ for the slope and $\ln A = 0.797 \pm 0.09$ for the intercept. The correlation coefficient is -0.742 and the mean value of X is $\bar{X} = 1.93$. The relative frequency which corresponds to the average distance between members of a substituted pair of amino acids is $Y(\bar{X}) = 1.07$ and this is less than half of $Y(0) = 2.22$.

It is interesting, particularly from the standpoint of the neutral theory, to

inquire if the relative frequency extrapolated at $X = 0$ is consistent with what would be expected from synonymous substitutions (which do not cause amino acid changes). For this purpose, we first estimate the ratio, which we shall denote by $R_{S/A}$, of synonymous to amino acid altering substitutions per site per equivalent mutational input using data obtained through comparative studies of nucleotide sequences. As explained in the last subsection of chapter 4, the estimated number of substitutions due to synonymous changes at the third position of the codon is given by K'_S. We note that about $(1 + 1/3)/2$ or $2/3$ of random nucleotide substitutions at the third position of the codon are synonymous, because, for a given combination of bases in the first and second positions of the codon, a change at the third position is completely synonymous in about half the cases, but only partly synonymous (purine↔purine or pyrimidine↔pyrimidine) in the remaining half. On the other hand, a nucleotide substitution either at the first or the second positions of the codon almost always leads to an amino acid change. Therefore, the expected value of $R_{S/A}$ may be estimated by the ratio of $K'_S/(2/3)$ to $(K_1 + K_2)/2$, namely, by $3K'_S/(K_1 + K_2)$. We use values listed in Table 4.8 on hemoglobins (which show standard amino acid substitution rates) but exclude comparisons involving the abnormal $\alpha - 3$ (a 'dead gene') and the α vs. β comparison (which is too remote to estimate K'_S accurately). The estimated value of $3K'_S/(K_1 + K_2)$ turned out to be 4.43 ± 1.64. If we multiply this by $Y(\bar{X})$, it should give a rough estimate for the synonymous substitution rate to be plotted on the Y-axis in Fig. 7.1. This gave 4.7 ± 1.8, indicated by the open triangle in Fig. 7.1. Although this is considerably larger than $Y(0) = 2.22$, considering a large statistical error of this estimation procedure, we may conclude that the data on synonymous substitution are consistent with the protein data in regard to the relationship between the rate of evolution and the magnitude of change, namely, the smaller the mutational change, the higher the evolutionary rate in terms of mutant substitutions.

From the standpoint of the neutral theory, this can readily be explained by using equation (5.1), i.e. $k = v_T f_0$; the probability (f_0) of a mutational change being selectively neutral becomes larger if the amino acid replacement involved is associated with a smaller physico-chemical difference, and therefore k becomes larger.

Let us now investigate what relation should be expected between k and x, where x is the magnitude of a mutational change, if we assume that mutant substitutions are caused exclusively by positive Darwinian selection. Under this assumption, k is proportional to the product $P_a(x)u_a(x)$, where $P_a(x)$ is

the probability that a mutation with the effect x happens to be selectively advantageous and $u_a(x)$ is the probability of such a mutation ultimately becoming fixed in the population.

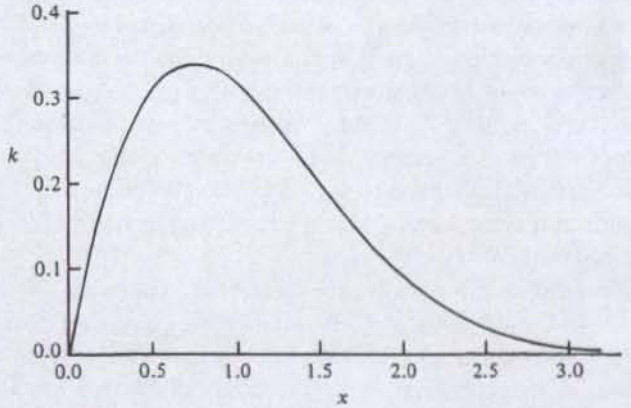
If we assume that the selective advantage s is proportional to x, then $u_a(x) \propto 2x$ because $u \approx 2sN_e/N$ as shown in equation (3.21). As to $P_a(x)$, we follow Fisher (1930b) and assume that it is given by equation (6.33), namely, $P_a(x)$ decreases as x increases (see Fig. 6.5). Then

$$k \propto 2xP_a(x) \tag{7.1}$$

and this is illustrated in Fig. 7.2. Unlike the observation illustrated in Fig. 7.1, this curve is bell-shaped, with the maximum evolutionary rate occurring for an intermediate mutational effect, contrary to the claim of Clarke (1971). Note that even if mutations having smaller effects have a higher chance of becoming advantageous as Fisher (1930b) showed, the selective advantage of such mutations must be correspondingly small. In the case of ordinary phenotypic characters ('quantitative characters'), there will be an additional factor to be considered, namely, mutations having small phenotypic effects may occur more frequently than those having large effects. In the case of molecular evolution, this last factor can be eliminated, as we have done above, by taking into account the difference in the frequencies of amino acids and/or difference in the numbers of synonymous codons among different amino acids.

It might be argued that the selective advantage (s) may not vanish even

Fig. 7.2. Relationship expected between the magnitude of mutational change x (in an arbitrary unit) and the relative evolutionary rate k (in terms of mutant substitutions), if the mutant substitutions are caused exclusively by positive Darwinian selection.



when the mutational effect (x) reduces to zero; rather, it converges to a certain positive value s_0 which is much larger than $1/(2N_e)$. If this is the case, $P_a(x)$ should be close to 1/2 for $x \approx 0$, that is, the synonymous mutations should have definite selective advantage in 50% of cases. Then, $f_A = 0.5$ and $\bar{s}_A = s_0$ in equation (5.2), and we have $k = 2N_e s_0 v_T$ for synonymous substitutions. This means that k is much larger than the total favorable mutation rate v_T. This is in sharp contrast to the neutralist prediction that, in general, k is smaller than v_T even if k may be reasonably close to v_T. As we shall show in section 7.3 of this chapter, synonymous substitutions are characterized by the remarkable property that not only their rates are high but also they are roughly equal to each other in different genes even when their amino acid-altering substitution rates are very different. This is much more consistent with the neutral theory than the selectionist interpretation.

7.2 Molecular structure and selective constraint

Along with the conservative nature of mutant substitutions in molecular evolution (as was discussed in the previous section), it has become increasingly evident that functionally less important molecules or parts of a molecule evolve (in terms of mutant substitutions) faster than more important ones. Among the proteins so far investigated, the highest rate is represented by fibrinopeptides (see Table 4.1). They have an estimated evolutionary rate of 8.3×10^{-9} per amino acid site per year, although a lower estimate 4.5×10^{-9} has also been obtained (Barnard *et al.*, 1972). On the other hand, the lowest evolutionary rate is represented by histone H4 with a rate 0.008×10^{-9} per amino acid site per year. This rate may be obtained immediately by noting the well-known fact that histone H4 (consisting of about 100 amino acids) of pea plants differs from that of calf thymus by only two amino acid replacements (see Isenberg, 1979), and by assuming that the plants and animals diverged some 1.2 billion years ago.

It is interesting to note that fibrinopeptides, the most rapidly evolving molecules, have little known function after they become separated from fibrinogen in the blood clot. Needless to say, fibrinopeptides are not completely inert but have a demonstrable physiological action. However, this action is attributable to a small portion of the peptide (Laki, 1965) so that the rest of the molecule is free to change. We shall discuss this problem again later.

The relationship between the functional importance (or more strictly, functional constraint) and the evolutionary rate has been beautifully explained by Dickerson (1971) as follows. In fibrinopeptides, virtually any amino acid change (mutant substitution) that permits the peptides to be

removed is 'acceptable' to the species. Thus the rate of evolutionary substitution of amino acids may be very near the actual mutation rate. Hemoglobins have the definite function of carrying oxygen so that structural requirements are more restrictive than for fibrinopeptides, with the result that the evolutionary rate is lower. Cytochrome c interacts with cytochrome oxidase and reductase, both of which are much larger than it, and there is more functional constraint in cytochrome c than in hemoglobins. Thus it has a lower evolutionary rate than hemoglobins. Histone H4 binds to DNA in the nucleus, and it is quite probable that a protein so close to the genetic information storage system is highly specified with little evolutionary change permitted over a billion years. From the standpoint of the neutral theory, these observations can be interpreted as follows. The probability of a mutational change being neutral depends strongly on molecular constraint. If a molecule or a part of a molecule is functionally less important, then the probability of a mutational change in it being selectively neutral (i.e. selectively equivalent) is higher. In other words, the same job can be done equally well by a variety of amino acids. Thus, the rate of evolution in terms of mutant substitutions by random drift becomes high.

Also, if a molecule is exposed less to natural selection, the probability of a mutation in it being selectively neutral is higher. The observation that the δ chain of hemoglobin $A_2(\alpha_2\delta_2)$, which forms the minor component of adult hemoglobin, shows a higher evolutionary rate and a higher level of polymorphism than the β chain which forms the major component $A(\alpha_2\beta_2)$ fits to this picture (Boyer *et al.*, 1969).

The evolutionary rates differ not only for different molecules but also for different parts of one molecule. As an example, let us take hemoglobin, one of the proteins whose structure is best known. There are several helical and non-helical regions, and as with other globular proteins, the inside of the molecule is filled with hydrophobic amino acids and is relatively conservative in evolution. Particularly, the heme pocket is known to play an extremely important role in the normal functioning of the molecule, and replacements of amino acids in this part tend to cause abnormalities such as unstable hemoglobin (Perutz and Lehmann, 1968). On the other hand, amino acids at the surface usually play no special role in the function, and substitutions of mutant amino acids in this part often show no clinical symptoms in heterozygous condition in man.

In order to compare evolutionary rates in these two parts, we analysed (Kimura and Ohta, 1973) sequence data for the α and β chains mainly among mammals. The results showed that in both chains the evolutionary rates at the surface are almost ten times higher than in the heme pocket

(Table 7.1). In addition, it is known (Jukes, 1971) that two histidines binding to the heme in the globin molecule are almost completely invariant. In other words, almost no mutant substitutions seem to have occurred in these two histidine positions in the entire history of vertebrate evolution extending some 500 million years and including a great many lines. The opossum α globin which was sequenced by Stenzel (1974) and which was discussed in chapter 4 must be a rare exception. The Perutz model of hemoglobins (see Perutz and Lehmann, 1968) helps us greatly to interpret such observations in terms of the structure and function of these molecules. More generally, according to the oil drop model of globular proteins (see Dickerson and Geis, 1969), the inside of a molecule is filled with non-polar (hydrophobic) amino acids, while the surface parts are occupied by polar (hydrophilic) amino acids. The functionally vital 'active center' is located inside a crevice, and the rate of evolutionary substitutions of amino acids in this part is expected to be very low due to stringent constraint. On the other hand, the surface parts are usually not very critical in maintaining the function or the tertiary structure, and the evolutionary rates in these parts are expected to be much higher, due to the possibility of a higher fraction of mutations being selectively neutral.

It might be argued that the difference of the evolutionary rate in different parts of the molecule is due to a difference in the mutation rate rather than a difference in selective constraint. That this is not the case is evident from an extensive study of hemoglobin variants in human populations. Vogel (1969) analyzed the distribution of amino acid replacements in the α and β chains using reported hemoglobin variants in human populations. He obtained a result suggesting that there are no codons within the α and β cistrons with an increased tendency to mutate. It appears as if mutations occur in the functionally rigid heme pocket just as frequently as in the functionally less

Table 7.1. *Comparison of evolutionary rates between the surface and the heme pocket in the α and β hemoglobin chains. The listed values represent the rates of amino acid substitution in units of* 10^{-9}*/amino acid/year ('paulings').*

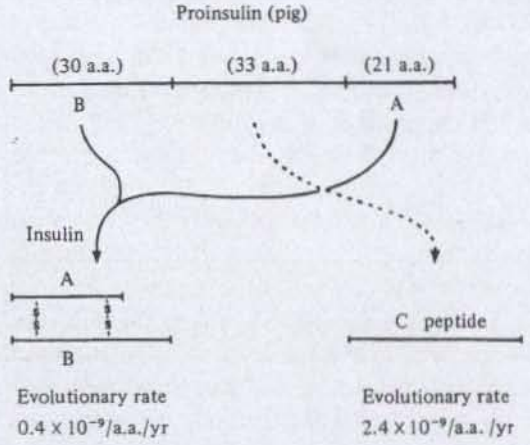
Region	Hemoglobin α	Hemoglobin β
Surface	1.35	2.73
Heme pocket	0.165	0.236

After Kimura and Ohta (1973).

rigid surface parts. Particularly noteworthy in this context is the fact that hemoglobin M variants are reported from various parts of the world. These variants represent the replacement of tyrosine for one of the two histidines that are linked to the heme, and they cause hemoglobin M disease. These variants are rare, but they are by no means excessively rare compared with other hemoglobin variants. Even within Japan, all the four types of hemoglobin M variants have been found (for details, see Yanase 1971). They are callled M Osaka (α58, His → Tyr), M Iwate (α87, His → Tyr), M Kurume (β63, His → Tyr) and M Akita (β92, His → Tyr).

The reason that these two histidine sites are invariant in evolution is that mutations at these two sites are highly deleterious or practically lethal in homozygous conditions, even if heterozygotes survive. It is known that about 30% of the hemoglobin molecules of the blood of M Iwate and other Hb M heterozygotes are abnormal (Met Hb); the rest are normal (Shibata *et al.*, 1966). Thus, heterozygotes would be expected to be selectively disadvantageous and these mutants have almost never been able to spread through vertebrate species in their course of evolution. Similarly, according to Hayashi and Stamatoyannopoulos (1972), the penultimate tyrosine at position HC2 in vertebrate hemoglobins is evolutionarily invariant. They studied chemical properties of the variant Hb Bethesda and Hb Rainier, in which this tyrosine is replaced respectively by histidine and by cysteine in the β chain. They found that, as expected from the Perutz model, these substitutions are associated with extreme disorganization of the normal hemoglobin oxygenation function. Thus, mutations at these amino acid

Fig. 7.3. Comparison between the evolutionary rate of insulin (A + B peptides) and that of the middle segment (C peptide) of proinsulin.



sites, despite repeated occurrence in the course of vertebrate evolution, have not been able to spread through the whole population in any species.

Another interesting example, showing that a functionally less important part of a molecule evolves faster than a more important part, is the middle segment (C) of the proinsulin molecule. This part is removed when the active insulin (consisting of A and B peptides) is formed. It is known that this part evolves at about the rate 2.4×10^{-9} per amino acid site per year which is several times as fast as that of insulin. Evolutionary rates given in Fig. 7.3 were computed by using data given in Jukes (1979); the average difference between mammals except the guinea pig listed in his Table 2 is 6% for the A + B peptides, whereas it is 32% for the C peptides. No doubt, many similar examples will be forthcoming, since the tailoring of a precursor molecule to produce a functionally active form appears to be a common phenomenon.

As an additional example showing rapid evolution of a functionally less important part, I would like to cite the report of Barnard *et al.* (1972). According to them, sequences 15–24 of pancreatic ribonuclease evolves at a very high rate comparable to rapidly evolving parts of fibrinopeptide, and this 'hypervariability' can be correlated with a lack of any contribution of this part either to the enzymatic activity or to the maintenance of structure required for the activity. These authors compare (in their Table 3) the amino acid composition in the hypervariable segment of fibrinopeptide A, residues 15–24 of ribonuclease, and proinsulin C peptide with the amino acid composition of a 'random protein'. Since they do not agree, the authors claim that there is a selection pressure acting on all regions of a protein, and that truly neutral mutations do not generally become fixed in evolution. I would like to point out that the amino acid composition of a given protein for a limited group of species can deviate from that of the ideal random protein simply by a historical accident. In addition, selective constraint (negative selection) may still be acting to some extent even in hypervariable regions, but the existence of selective constraint itself does not contradict the neutral theory as I have already emphasized in previous places in this book.

In order to make a case for adaptive substitutions at the molecular level, Stebbins (in Stebbins and Lewontin, 1972) set out to analyse the data on fibrinopeptides. He made use of the sequence data on artiodactyl fibrinopeptides A and B presented by Mross and Doolittle (1967). For each amino acid site, he investigated the frequencies of substitutions together with Sneath's (1966) correlation coefficients between the substituted amino acids. Stebbins classified substitutions into two types, conservative and

radical; if Sneath's correlation coefficient between a pair of residues is higher than 0.6, such a substitution is called conservative, while if it is lower than 0.6, the substitution is called radical. According to him, the results of this analysis were 'unequivocal' in showing that substitutions are unequally distributed on the molecule, and that radical substitutions are concentrated in those sites having the largest number of substitutions, while the conservative substitutions are scattered over sites where substitutions are rare. Taking into account a report that the residue at position 13 of fibrinopeptide A influences the rate at which thrombin splits the molecule, Stebbins and Lewontin (1972) postulated that radical substitutions recorded in the highly variable parts (positions 12 to 19 in fibrinopeptide A and positions 18 to 21 in fibrinopeptide B) took place during artiodactyl evolution in association with an adaptive shift. These authors believe that these radical substitutions are associated with alterations in the accessory adaptive properties of the fibrinogen molecule and its derivatives, and that such substitutions have aided very different artiodactyls in acquiring various physiological and ecological properties.

Despite its plausibility, Stebbin's analysis of fibrinopeptides contains some problems. Consider fibrinopeptide A first. It is known that the enzyme thrombin cuts the fibrinopeptides from the fibrinogen. The point at which it cuts the A chain is the first arginine residue that is encountered from the amino terminus. If we compare fibrinopeptide A sequences, some of their amino acid residues are invariant or nearly so, such as glycine at position 4 (i.e. 4 residues before the arginine which is also invariant). Valine at position 2 is highly conserved; only one substitution (Val→ Ala) is observed in the data from artiodactyls. Since the Sneath correlation coefficient (ϕ) for Val-Ala pair is 0.675, this substitution is conservative. Also, glycine at position 5 is highly conserved; only one substitution (Gly→ Thr) is found in the data, and it is radical, for the correlation involved is $\phi = 0.396$.

As compared with this region, positions 12–14 are highly variable. At position 13, for example, the occurrence of four substitutions is evident in the artiodactyl data; Ala-Val ($\phi = 0.675$), Pro-Ala ($\phi = 0.553$), Val-Pro ($\phi = 0.473$) and Asp-Ala ($\phi = 0.156$). According to Stebbin's criterion, only the first substitution is conservative ($\phi > 0.6$), the rest being radical. Thus, substitutions at this position may be called predominantly (i.e. 75%) 'radical'. In general, however, if we adopt this criterion, most substitutions turn out to be radical. In fact, among 190 correlation coefficients between pairs of amino acids as listed in Table 2 of Sneath (1966), only 16 are larger than 0.6, so that 174/190 or 92% of them are radical. The average of 190 correlation coefficients is $\bar{\phi} = 0.376$. It may be more appropriate, therefore,

to choose 0.4 instead of 0.6 as the borderline in judging whether a substitution is conservative or radical. So, if we call substitutions having correlation coefficients larger than 0.4 conservative, then, the observed substitutions at position 13 are predominantly (i.e. 75%) 'conservative', with only one radical substitution (i.e. Asp-Ala; $\phi = 0.156$). In addition, a comparison of various A sequences clearly shows that insertion or deletion must have occurred frequently in this highly variable region, and exact alignment appears to be very difficult; there are too many gaps, and no one really knows where the gaps belong. The situation is much worse in fibrinopeptide B. In general, among animals, the fibrinopeptides A vary in length from 14 to 19 residues, and the fibrinopeptides B vary in length from 9 to 21 residues (see D-90 of Dayhoff, 1972). The results of the analysis by Stebbins on the nature of amino acid substitutions using fibrinopeptides are therefore not 'unequivocal', contrary to his assertion. These and other statistical considerations, I believe, make Stebbin's analysis dubious, to say the least. Furthermore, the demonstration that the rates of amino acid substitutions are not 'randomly distributed' (i.e. not equal) over the peptide chains, although it shows that negative selection is acting (which is not incompatible with the neutral theory), is not sufficient to show that positive Darwinian selection is the main agent for gene substitution.

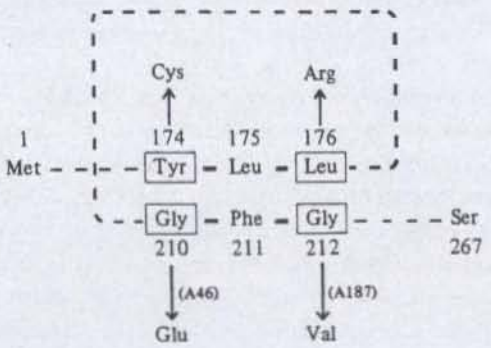
As we have seen through several examples, the contrast between the neutralists and selectionists is most clearly seen in their interpretations with respect to a rapidly evolving molecule (such as fibrinopeptides) or a part of one molecule (such as C peptide of proinsulin). According to the neutral theory, they are not functionally important and therefore a large fraction of mutations are neutral, with the result that mutations accumulate rapidly by random drift. On the other hand, according to the selectionist interpretation, a rapidly evolving molecule or a part of one molecule has some perhaps unknown function, and undergoes rapid adaptive improvements by accumulating many slightly advantageous mutations by positive Darwinian selection. Which of these two interpretations is more appropriate will be decided as more data accumulate.

Here a cautionary remark is in order. In discussing the evolutionary rate, we should not confuse the rate of evolution in terms of mutant substitutions with the rate of change in the frequency of an individual mutant allele. The former (denoted by k) is measured by counting the number of amino acid (or nucleotide) substitutions during a long span, say, tens of millions of years, while the latter (denoted by Δp) is measured by recording the change during a short period, say, a century or less (as in industrial melanism). It is a pity that these two are sometimes confused even by a few population geneticists

when they criticize the neutral theory. Note that for a neutral allele, the amount of change in gene frequency per generation (Δp) can be extremely small especially in a large population, its mean being zero and standard deviation being of the order of $1/\sqrt{N_e}$. The evolutionary rate in terms of mutant substitutions (k), however, is invariant with respect to the population size N_e (see equation 3.20). On the other hand, for a positively selected allele, k increases enormously as the population size increases (see equation 3.22), but Δp stays essentially unaltered.

For our deeper understanding of the mechanism of molecular evolution, it is essential to search for the relationship between the pattern of mutant substitutions on the one hand and the tertiary structure and function of the molecule on the other. We are still unable to describe in concrete physicochemical terms how molecular constraints become shifted as new mutants are substituted one by one in the course of evolution. This is particularly true for proteins, although Fitch's concept of concomitantly variable codons or 'covarions' may serve as a useful guide. According to the statistical analysis of Fitch and Markowitz (1970), only 10% of codons in cytochrome c can accept mutations at any moment in the course of evolution, but the positions of covarions inside the molecule change with time. Also, it was found (Fitch, 1971a) that the proportion of covarions is about 30% in hemoglobin α, but nearly 100% in fibrinopeptide A. What is really interesting is his finding that if the rate of amino acid substitutions is calculated on the basis of covarions, cytochrome c, hemoglobin α and fibrinopeptide A are all evolving at about the same rate. From the standpoint of the neutral theory, covarions can be interpreted in a natural

Fig. 7.4. Diagram illustrating possible relationship between regions in reversion mutation studies of tryptophan synthetase A-protein. Data from Yanofsky *et al.* (1964). Copyright 1964 by the American Association for the Advancement of Science.



way by assuming that they represent amino acid sites which can accept mutations by random genetic drift without impairing the structure and function of the molecule. Fitch (1971b) has also been able to estimate the persistence of a covarion, that is, the probability that a codon which belongs to a covarion group remains so after a mutation becomes fixed at a different codon. His estimate of the persistence turns out to be less than 0.25 suggesting fairly rapid turnover of covarions within a molecule.

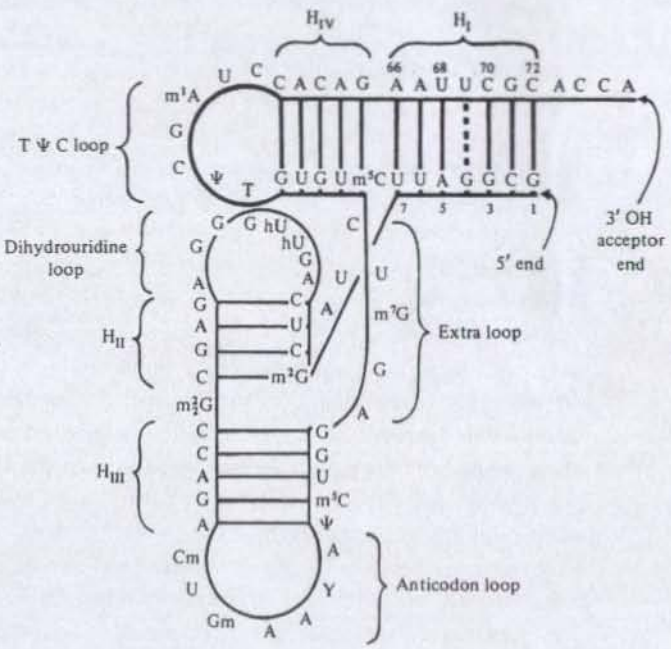
Probably, what determines the pattern of interaction between amino acids in evolution is their physical proximity or direct contact within the folded protein. In this connection, studies of mutants of the tryptophan synthetase A-protein in *E. coli* by Yanofsky and his associates (Yanofsky *et al.*, 1964) are most instructive (see also, Yanofsky, 1967, and Watson, 1970). This protein consists of 267 amino acids. Position 210 in the Wild type protein is occupied by glycine (see Fig. 7.4), but mutation studies show that substitution of serine for glycine leaves the enzyme fully active. On the other hand, mutant protein A46 which has glutamic acid at position 210 is nonfunctional. Extensive reversion studies showed that if a further change Tyr → Cys occurs at position 174 (which is 36 amino acids away), then this gives a functionally active enzyme (see Fig. 7.4). However, a single mutation Tyr → Cys at position 174 (but with Gly at 210 as in the wild type) produces an inactive protein. Similarly, a nonfunctional change Gly → Val (mutant A187) at position 212 and the recovery of function by a second change Leu → Arg at position 176 (again, 36 amino acids away from the primary mutational site) were found. These observations led Yanofsky *et al.* (1964) to propose a model of spatial relationships of these sites in the folded protein as illustrated in Fig. 7.4.

The pattern by which selective constraint shifts as amino acids are substituted one after another in the actual course of evolution must in general be more subtle than illustrated above. It is likely, however, that a substitution of an amino acid at one position often influences the set of sites where selectively neutral changes are allowed to occur in other parts of the molecule.

In the following, I shall try to show that shifting of selective constraints can be explained in more concrete molecular terms in transfer RNA (tRNA), particularly regarding base substitutions at its helical regions. As is well-known, the secondary structure of all the tRNA molecules can be arranged in a clover leaf configuration, although the molecules are folded in a more complex fashion to form the tertiary structure (Kim *et al.*, 1973, 1974; Klug *et al.*, 1974). In Fig. 7.5, yeast phenylalanine tRNA is illustrated by taking into account the fact that its three-dimensional structure is L-

shaped (Kim *et al.*, 1973, 1974). In general, the tRNA molecule contains four helical regions, as denoted by H_I, H_{II}, H_{III} and H_{IV} in Fig. 7.5. They correspond respectively to the amino acid acceptor, DHU (dihydrouridine), anticodon and TΨC stems. In these regions, bases are mostly paired according to Watson–Crick rules, although there are one or two mispairs, especially in the first helical region. It was suggested (Jukes, 1969; Holmquist *et al.*, 1973) that in these regions the maintenance of the double helical structure is the only requirement. This means that individual bases may change freely in the course of evolution as long as the base substitutions are coupled in such a way as to maintain Watson–Crick pairing at all but a few sites in these regions. According to Holmquist *et al.* (1973), among the mispairings in the helical regions, G-U (or U-G) pairs that do not interfere with helicity occur much more frequently than other forms of mispairings. In fact, G-U mispairs amount to more than 60% of all the 'non-Watson–Crick pairs'. It is interesting to note also that in these regions G-C pairs that have three hydrogen bonds occur much more frequently (69%) than A-U pairs that have two hydrogen bonds. This

Fig. 7.5. Diagram illustrating the structure of the yeast phenylalanine tRNA based on data given by Kim *et al.* (1974). Copyright 1974 by the American Association for the Advancement of Science.



suggests the importance of maintaining a tight double stranded structure. On the other hand, in the non-helical regions consisting of various loops, we find far fewer G-C pairs (39% G + C), suggesting the opposite requirement of maintaining an open, unpaired structure necessary for molecular recognition. The majority of the 'invariant sites' also occur in these loops.

Let us now examine in more detail the helical regions, particularly the amino acid acceptor stem (H_1 or a.a. stem) consisting of seven base pairs. Fig. 7.6 illustrates a few examples having 0, 1 and 2 mispairings. Also, in Table 7.2, the frequencies of various types of mispairs are listed. The average number of mispairings per a.a. stem is about 0.59, of which the most

Table 7.2. *Frequencies of various types of mispairings in the amino acid acceptor stem. Observed frequencies were computed by using 177 tRNA sequences compiled by Gauss and Sprinzl (1981), but excluding one* E. coli *temperature-sensitive tRNA. In this table 'Non-GU' includes all mispairings other than G-U pair (such as G-A and U-U). Expected frequencies were computed under the assumption that frequencies of mispairs among seven paired sites are statistically independent.*

Type of mispairs	Frequency (%)	
	Observed	Expected
None	55.1	54.4
1 G-U	25.0	28.6
2 G-U	9.7	7.1
3 G-U	0.6	1.0
1 Non-GU	6.8	5.8
1 G-U + 1 Non-GU	2.8	3.1
Total	100.0	100.0

Fig. 7.6. Examples showing 0, 1 and 2 nucleotide mispairings in the amino acid acceptor stem of the tRNA molecule.

Phenylalanine *E. coli*	Phenylalanine Yeast	Phenylalanine Wheat	Lysine *Drosophila*	Alanine Yeast
1 G – C 72	G – C	G – C	G – C	G – C
C – G	C – G	C – G	C – G	G – C
3 C – G 70	G – C	G – C	C – G	G ··· U
C – G	G ··· U	G – C	C – G	C – G
5 G – C 68	A – U	G A	G ··· U	G – C
G – C	U – A	G – C	G ··· U	U U
7 A – U 66	U – A	A – U	C – G	G – C

predominant (0.49) is G-U pair, while the remaining minority (0.10) is represented by various other types of mispairs such as G-A, A-C and U-U. Furthermore the incidence of a non-G-U type mispairing appears to be largely independent of whether one G-U pair exists in another part of this stem or not. The observed frequencies of various types of mispairs agree reasonably well with frequencies computed under the assumption that occurrence of the mispairs among the seven paired sites are statistically independent.

We can visualize the process of evolutionary change in this part of the molecule as an alternating sequence of opening and closing of Watson–Crick pairs. As pointed out by Ohta (1973), the first opening is likely to be very slightly deleterious, but a mutation causing such a change can spread in the population by random drift. It is possible that one mispairing, especially when it is G-U, has such a small deleterious effect that it is practically neutral with respect to natural selection. Thus, G-U serves as a most natural transitional state going from G-C to A-U or vice versa as pointed out by Jukes (1969). On the other hand, non-G-U type mispairings cause a definite weakening of the bonding, and although one such mispairing may sometimes be allowed, mutations leading to more than one such mispairing are eliminated from the population because of marked deleterious effects; only when the existing mispairing is closed by a mutant substitution, is the molecule ready to accept a new mutation. Even G-U pairs, when more than two accumulate in this region, appear to show deleterious effects. Thus, although there are seven base pairs (fourteen nucleotide sites) in this region, the effective number of variable sites at any moment in the course of evolution is smaller.

Ozeki and his associates used nonsense suppressor mutants of *E. coli* to show that the proper functioning of this region of tRNA depends upon the formation of Watson-Crick pairs (or sometimes G-U pair) but not upon the nature of individual bases themselves. For example, in su$^+$2 ($tRNA_2^{Gln}$) which inserts glutamine at the 'amber' codon (UAG), base pairs in the amino acid acceptor stem are as follows.

7 6 5 4 3 2 1
A U G G G G U
· · · · · · ·
U A C C C C A

It was found that if A at position 7 changes to C (a mutant called C7), the molecule becomes defective. Also, the mutant called G65 in which U (opposite to A at position 7) changes to G is defective. However, normal function is restored in the double mutant called C7-G65 which has a C-G

pair instead of the original A-U. Many interesting examples of this sort are reviewed in Ozeki *et al.* (1980).

7.3 Synonymous and other silent substitutions

We have already discussed, in chapter 4, the high evolutionary rates of synonymous nucleotide substitutions (that do not cause amino acid changes). We have also noted that since synonymous substitutions are much less likely to be subject to natural selection than amino acid altering substitutions, the observed high rates support the neutral theory.

Fig. 7.7 illustrates the preponderance of synonymous substitutions as revealed by comparison of the coding region of the human β-globin (Lawn *et al.*, 1980) with that of the rabbit β-globin (van Ooyen *et al.*, 1979) gene. In this figure, only those codon pairs which contain at least one base difference are shown. Small boxes indicate places where bases are different at homologous sites. It is evident that synonymous base differences outnumber the amino acid differences by a factor of about 1.7. Note that β-globin in mammals consists of 146 amino acids and therefore corresponds to 438 nucleotide sites.

In this section, we further investigate the evolutionary rates of synonymous and other silent substitutions, and in the next section, the rate of nucleotide substitutions in pseudogenes (that have lost their functions as genes). To estimate the rate of base substitutions, let us assume a model as illustrated in Fig. 7.8. This model is called the 'three substitution-type (3ST) model' (Kimura, 1981c) and is similar to the corresponding model used in chapter 4 (see Fig. 4.9). The only difference is that two parameters β and γ rather than a single parameter (β) are assigned for transversion type substitutions and therefore the model is slightly more general. We mostly follow the notations used in chapter 4 (such as using the RNA codes and expressing the four bases by the letters U, C, A and G). In the present model the rate of base substitutions per site per unit time (year) is

$$k = \alpha + \beta + \gamma. \tag{7.2}$$

Note that α, β and γ here refer to the evolutionary rates by which bases are substituted in the species in the course of evolution rather than ordinary mutation rates at the level of individuals. The total number of base substitutions per site which separate the two sequences and therefore involve two branches each with length T (T being the time since the two diverged from a common ancestor) is given by $2Tk$ which we denote by K.

$$K = 2Tk = 2(\alpha + \beta + \gamma)T. \tag{7.3}$$

When we compare two homologous sequences, say, sequences 1 and 2, there are twelve possible types of base differences as shown in Table 7.3. We denote by P the probability (frequency) of homologous sites showing transition-type base differences. In other words, P is the probability that two homologous sites are occupied by UC, CU, AG or GA. Similarly, we denote by Q the probability of homologous sites being occupied by UA, AU, CG or GC and by R the probability of UG, GU, CA or AC. Thus,

Fig. 7.7. Diagram illustrating the preponderance of synonymous changes over amino acid-altering changes, as revealed by comparison of the human and rabbit β-globin genes. The encircled numbers indicate positions where amino acids differ between the two sequences, while small boxes indicate places where bases are different. Only those codon positions are shown where at least one base difference exists.

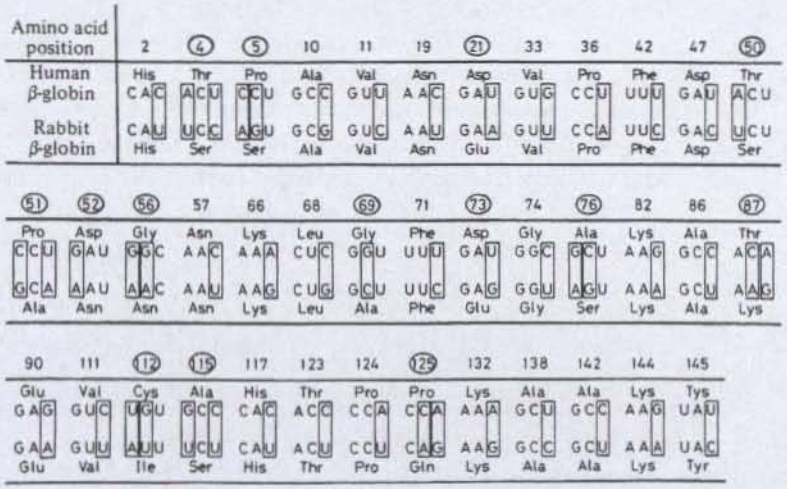
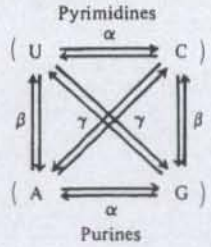


Fig. 7.8. Three substitution-type (3ST) model of evolutionary base substitutions. In this model α is the rate of 'transition' type substitutions, while β and γ are those of 'transversion' type substitutions.



$Q + R$ represents the probability of homologous sites showing transversion-type differences.

Then, the evolutionary distance in terms of the number of base substitutions between the two sequences may be estimated by

$$K = -(1/4)\log_e[(1-2P-2Q)(1-2P-2R)(1-2Q-2R)]. \tag{7.4}$$

This is the sum of the three components which are

$$2\alpha T = -(1/4)\log_e[(1-2P-2Q)(1-2P-2R)/(1-2Q-2R)], \tag{7.4a}$$

$$2\beta T = -(1/4)\log_e[(1-2P-2Q)(1-2Q-2R)/(1-2P-2R)] \tag{7.4b}$$

and

$$2\gamma T = -(1/4)\log_e[(1-2P-2R)(1-2Q-2R)/(1-2P-2Q)]. \tag{7.4c}$$

In the special case in which the two types of 'transversion' substitutions occur equally frequently so that $\gamma = \beta$, this formula reduces to

$$K = -(1/2)\log_e[(1-2P-Q')\sqrt{(1-2Q')}], \tag{7.5}$$

where $Q' = Q + R$ is the total fraction of 'transversion' substitutions. This agrees with equation (4.14) in chapter 4. If the divergence time T is known, the rate of base substitutions per site per year can be estimated by

$$k = K/(2T). \tag{7.6}$$

It is known that a large fraction of base substitutions at the third position of codons are synonymous, so it is of interest to estimate the synonymous

Table 7.3. *Various types of different base pairs at homologous nucleotide sites when two homologous sequences* (1 *and* 2) *are compared, and designation of their frequencies* (*probabilities*)

	Transition type				Transversion type							
Sequence 1	U	C	A	G	U	A	C	G	U	G	C	A
Sequence 2	C	U	G	A	A	U	G	C	G	U	A	C
Frequency	P				Q				R			

Table 7.4. *Evolutionary distances per site (together with standard errors) as estimated by 3ST method.* K_i, $(i = 1, 2, 3)$, *denotes the number of base substitutions at position i of codons that separates the two sequences compared, and* K_S' *denotes the synonymous component at position 3*

	Evolutionary distance per nucleotide site			
Comparison	K_1	K_2	K_3	K_S'
Human vs. rat pregrowth hormones	0.26 ± 0.04	0.18 ± 0.03	0.53 ± 0.07	0.44 ± 0.07
Human vs. rat I preproinsulins:				
A + B chains (insulin)	0.04 ± 0.03	0.00[a]	0.46 ± 0.12	0.38 ± 0.12
C peptide	0.18 ± 0.06	0.27 ± 0.10	0.95 ± 0.46	0.77 ± 0.51
Human vs. rabbit β globins	0.06 ± 0.02	0.06 ± 0.02	0.28 ± 0.05	0.25 ± 0.05
Rabbit vs. mouse β globins	0.16 ± 0.03	0.13 ± 0.03	0.43 ± 0.07	0.36 ± 0.07
Rabbit vs. mouse α-globins[b]	0.12 ± 0.03	0.12 ± 0.03	0.54 ± 0.09	0.47 ± 0.09
S. purpuratus vs. *P. miliaris*:				
Histone H2B	0.09 ± 0.03	0.02 ± 0.01	0.48 ± 0.10	0.43 ± 0.10
Histone H3	0.008 ± 0.008	0.008 ± 0.008	0.47 ± 0.08	0.41 ± 0.08

[a] No observed changes among 51 codons.
[b] Mouse α-1 gene of Nishioka *et al.* (1980).

component of the number of base substitutions at position 3. This is done by the formula

$$K_S' = -(1/4)\log_e[(1-2P-2Q)(1-2P-2R)]. \quad (7.7)$$

In the special case $\gamma = \beta$, this reduces to equation (4.16) of chapter 4. In addition, formulae for the error variance (due to sampling) of these estimates can be derived. For details, readers may refer to Kimura (1981c).

Table 7.4 lists the results of applying these formulae to various comparisons. In the first line of the table, the nucleotide sequence of human pregrowth hormone (presomatotropin) (data from Martial *et al.*, 1979) is compared with that of rat pregrowth hormone (Seeburg *et al.*, 1977). The K values are 0.26, 0.18 and 0.53 respectively for the first, second and third codon positions, and 0.44 for the synonymous component at the third position. Since the human and the rat probably diverged late in the Mesozoic some 80 million years ago, we may take $T = 8 \times 10^7$ for the divergence time. Then the evolutionary rates per year at these three codon positions are $k_1 = 1.6 \times 10^{-9}$, $k_2 = 1.1 \times 10^{-9}$ and $k_3 = 3.3 \times 10^{-9}$, and $k_S' = 2.8 \times 10^{-9}$ for the synonymous component. As far as amino acid altering substitutions are concerned, pregrowth hormone appears to have an evolutionary rate slightly higher than hemoglobins.

Comparison of human preproinsulin (Bell *et al.*, 1980; Sures *et al.*, 1980) with rat preproinsulin gene I (Cordell *et al.*, 1979; Lomedico *et al.*, 1979) presented in the second and third lines in Table 7.4 is interesting in showing large differences in the evolutionary rates between A + B chains (insulin) and C peptide in the first two codon positions. In other words, with respect to amino acid substitutions, insulin evolves only some 1/10 as rapidly as proinsulin C peptide. Yet with respect to the synonymous component (K_S'), the evolutionary rate of insulin is not significantly lower than that of the C peptide. Furthermore, K_S' value of preproinsulin is roughly equal to that of pregrowth hormone (both regarding human-rat divergence). The K values listed in the fourth, fifth and sixth lines in the table were computed using data on human β globin (Marotta *et al.*, 1977), rabbit β globin (Efstratiadis *et al.*, 1977), mouse β globin (Konkel *et al.*, 1978), rabbit α globin (Heindel *et al.*, 1978) and mouse α-1 globin genes (Nishioka *et al.*, 1980).

If we compare these values with the corresponding values listed in Table 4.8 (see lines 3, 4 and 6) we note that, as far as hemoglobins α and β for these pairs of animals are concerned, the present model gives almost exactly the same estimates as those obtained by using the simpler model (see Fig. 4.9) assumed in chapter 4. The divergence time of the rabbit and the mouse, as well as that of the human and the rabbit may be assumed to be eighty million years ($T = 8 \times 10^7$).

Here, divergence of introns will be of interest. I have compared the small introns (intron 1) of the human and rabbit. These two homologous introns can be aligned in a satisfactory way by inserting three gaps (amounting to four nucleotides) in the rabbit intron. Excluding these gaps, and also the first two and the last two nucleotide positions (which must be invariant in accordance with the GT/AG rule for splicing; see Nishioka *et al.*, 1980), we get $P = 16/122$, $Q = 8/122$ and $R = 11/122$. Applying equation (7.4) and the corresponding formula for the standard error (see Kimura, 1981c), we get the distance per site $K = 0.36 \pm 0.053$. This may be compared with the corresponding estimate for the synonymous component, $K_S' = 0.25$ (see line 4 in Table 7.4). Noting that 2/3 of substitutions at the position 3 are synonymous, we may estimate a corrected synonymous distance by $K_S = K_S'/(2/3)$ which yields $K_S = 0.38$. Thus, the rate of mutant substitutions in the small intron is approximately equal to that of synonymous substitutions.

The last two lines of Table 7.4 show evolutionary distances between two sea urchin species *S. purpuratus* and *P. miliaris* for histone H2B and H3 sequences, as computed using data reported by Sures *et al.* (1978) and Schaffner *et al.* (1978). These two species belong to different families which diverged somewhere between 6×10^7 and 16×10^7 years ago (Kedes, 1979). The values of K_1 and K_2 obtained show that these histones have a very low evolutionary rate in terms of amino acid replacements, and furthermore that the evolutionary rate of H3 is only 1/7 of the rate of H2B. However, with respect to the synonymous component of the substitution rates at the third codon position, these histones are very similar. Furthermore, their synonymous substitution rates are not very different from the corresponding rates of pregrowth hormone, insulin and the α and β globins. If we tentatively assume $T = 10^8$ years for the divergence time of these sea urchin species, we obtain $k_S' = 2.2 \times 10^{-9}$ for histone H2B and $k_S' = 2.1 \times 10^{-9}$ for histone H3.

Fig. 7.9. A part of comparison between the histone H4 mRNA sequences of *S. purpuratus* and *L. pictus* together with the corresponding amino acid sequence of cow histone H4. This diagram is constructed using data from Grunstein *et al.* (1976) and Grunstein and Grunstein (1978).

L. pictus	Messenger RNA	GAU	AAC	AUC	CAA	GGA	AUA	ACU	AAA	CCG	GCA
S. purpuratus	Messenger RNA	GAU	AAC	AUC	CAA	GGC	AUC	ACC	AAG	CCU	GCA
Histone H4	Amino acid sequence	Asp	Asn	Ile	Gln	Gly	Ile	Thr	Lys	Pro	Ala
		24	25	26	27	28	29	30	31	32	33

Even more conservative than histone H3 and probably the most conservative of all proteins is histone H4. As mentioned already (see section 7.2) the rate of amino acid substitution of histone H4 is estimated to be $k_{aa} = 0.008 \times 10^{-9}$ per amino acid site per year. A remarkable fact is that even in such a highly conserved protein, a large number of synonymous changes have occurred in the course of evolution as shown in Fig. 7.9, where a part of sequence comparison between histone H4 messenger RNAs of two sea urchin species *Lytechinus pictus* and *Strongylocentrotus purpuratus* is presented. Using the data reported by Grunstein *et al.* (1976), I estimated the evolutionary distance of these two species at the third codon position of histone H4 and obtained the result $K_3 = 0.44 \pm 0.163$ (Kimura, 1977). According to Grunstein *et al.* (1976), these sea urchins shared a common ancestor approximately 6×10^7 years ago. Therefore, if we assume $T = 6 \times 10^7$, we obtain $k_3 = (3.7 \pm 1.4) \times 10^{-9}$. This value is again comparable to the substitution rate at the third codon position of various other proteins.

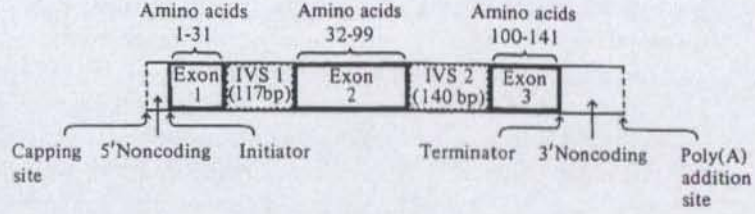
As the final example of a highly conserved protein, I would like to take tubulin, which is the subunit protein of microtubules. Recently, the nucleotide sequence of the C terminal region of rat α-tubulin has been determined (Ginzburg *et al.*, 1981), and a comparison of this with the corresponding region of chick α-tubulin (Valenzuela *et al.*, 1981) has become possible. I use the data presented in Fig. 2 of Ginzburg *et al.* (1981). Excluding the termination codon, 145 codon positions can be compared. In this region, we find only one amino acid difference between these two α-tubulins. This gives $K_{aa} = 0.0069$ as an estimate for the number of amino acid substitutions. On the other hand, 51 codons are involved in synonymous changes, of which one involves the first and also the third codon positions (AGA vs. CGC, both coding for Arg). At the third codon position, the frequencies of the three types of differences are $P = 33/145$, $Q = 6/145$ and $R = 12/145$. Then we get $K_3 = 0.51 \pm 0.08$ as the estimate of the number of base substitutions that separate the rat and chick α tubulins at the third codon positions. Also, we get $K'_S = 0.44 \pm 0.08$ as the synonymous component. It is interesting to note that these values are not very different from the corresponding values ($K_3 = 0.64 \pm 0.11$ and $K'_S = 0.53 \pm 0.10$) obtained by comparing the chicken β-globin with the rabbit β-globin genes (see Table 4.8), despite an enormous difference in the amino acid-altering substitution rates between β-hemoglobin and α-tubulin.

The fact that the rates of synonymous substitutions are not only higher than those of amino acid-altering substitutions but also are approximately equal among different genes has also been revealed by extensive studies of Miyata *et al.* (1980b), who compared some 50 pairs of homologous

nucleotide sequences of different genes. They obtained 5.1×10^{-9} per nucleotide site per year as the estimate for the average evolutionary rate of synonymous substitutions. A rough estimate which I have obtained from Table 7.4 for the synonymous component of the substitution rate at position 3 of the codons is $k'_S = 2.6 \times 10^{-9}$ per site per year. If we assume that the substitutions of the four kinds of bases occur in equal frequencies, then about 2/3 of nucleotide substitutions turn out to be synonymous at the third codon position. Therefore, the evolutionary rate of synonymous substitutions may be estimated by dividing k'_S by 2/3 which yields $k_S = 3.9 \times 10^{-9}$ per year. This estimate is not very different from Miyata *et al.*'s estimate of 5.1×10^{-9}.

Miyata *et al.* (1980b) also investigated the evolutionary rate in the noncoding regions. They found that the intervening sequences and the 5′ portion of the 3′ noncoding region show considerable sequence divergence, the extent of which is almost comparable to that in the synonymous codon sites (for terminologies on various parts of the eukaryotic gene, see Fig. 7.10). On the other hand, they found that the other blocks consisting of the 5′ noncoding region and the 3′ portion of the 3′ noncoding region are strongly conserved, showing approximately half of the divergence of the synonymous sites. This is understandable since these regions contain various signals for transcriptional control and modification of mRNA which impose strong constraints against free change by mutant substitutions. Such signals include TATA ('tar-tar' box) for the initiation of transcription, a signal for capping, the AATAAA signal for the poly (A) addition, and several others (see, for example, Leder *et al.*, 1980; Vanin *et al.*, 1980).

Fig. 7.10. Structure of human α-globin (α2) gene. This diagram has been drawn based on data from Proudfoot and Maniatis (1980). The structure of the β-globin gene is similar, the main difference being that the second intron is larger (573 base pairs in rabbit β-globin; see van Ooyen *et al.*, 1979). In this figure, the symbol IVS stands for the intron.



All these observations can be interpreted readily by the neutral theory which claims that the majority of mutant substitutions in the species are caused by random fixation of selectively neutral (i.e. selectively equivalent, but not necessarily functionally identical) mutants, rather than by positive Darwinian selection. Although favorable mutations no doubt occur, the theory assumes that they are rare enough to be neglected. The neutral theory predicts that the probability of a mutation being selectively neutral (that is, not harmful) is larger, the less the mutation disrupts the existing structure and function of the molecule. At the limit in which all the mutations are selectively neutral, the rate of evolution per site (k) becomes equal to the total mutation rate (v) per site. It is likely that synonymous mutations are not very far from this limit (Kimura, 1977), and therefore, the evolutionary rates of synonymous substitutions per site are nearly equal for different molecules. Of course, such an equality is only approximate, because, as we shall discuss later (see section 7.5), there might still be a selective constraint as reflected by 'non-random' synonymous codon usage. But, the amount of selective constraint involved must be rather small. Rapid evolutionary change in introns can also be explained in a similar way (see chapter 4).

An opposing view has been proposed by Perier *et al.* (1980). They claim that the driving force for fixation is positive natural selection operating on some fraction of amino acid altering ('replacement') changes, and, that such selective fixation carries along with it neutral alterations (including changes at silent sites), that have accumulated in those regions of the DNA. In other words, they invoke the 'hitch-hiking' effect to explain fixation of synonymous changes.

I have pointed out (Kimura, 1981c) that, unless we ignore well-established principles of population genetics, such an explanation cannot account for actual observations. In fact, such 'hitch-hiking' cannot bring about substitutions of neutral mutants at a very high rate when the selected changes occur at a very low rate. Take, for example, the histone H4 gene. The rate of replacement is exceedingly low, yet, synonymous base substitutions occur at a rate comparable to that of replacements in fibrinopeptides, one of the most rapidly evolving molecules.

We can treat the problem more in detail. Since the effect of hitch-hiking on fixing linked alleles extends over only short distances around a selectively driven gene, we consider a small segment of DNA, such as a gene locus, within which crossing-over is so infrequent as to be negligible. Let us suppose that a single new, advantageous mutant allele has appeared at this locus. In order that this selected mutant can bring other unselected (neutral) mutants to fixation by hitch-hiking, the gene copy in which this advantageous mutant appeared must also contain at the same time a number of neutral mutants. Furthermore, in order to make the rate of substitution of neutral mutants per site m times higher than that of selectively driven mutations (in this case amino-acid altering changes), each gene copy in the population must contain on the average m neutral mutants, irrespective of whether an advantageous mutation happened to occur in it or not. This factor m must be very large, probably 1000 or more in histone H4. On the other hand, if each gene copy contains a large number of neutral mutants, the corresponding (homologous) genes in different individuals have so many base differences from each other that there is no such thing as a species-specific nucleotide sequence of a particular gene. In other words, every individual in the species would have a quite different homologous sequence. This is contrary to observations.

Furthermore, hitch-hiking can not explain the observation that when genes of different proteins are studied, the evolutionary rates of synonymous substitutions are not only high among them, but also roughly equal to each other, even when their amino acid altering substitution rates differ a great deal, as we have discussed extensively in this section.

Jukes (1980a) who reviewed the occurrence of synonymous substitutions in various organisms points out that half of the nucleotide substitutions during the divergence of genes in animals, bacteria and viruses are silent changes. He supports the view that these changes result from random drift acting on neutral and nearly neutral changes (see also Jukes, 1980b).

The most important role of DNA is self-perpetuation, but it has the additional properties of accumulating mutations and occasional duplications.

The gene primacy view of life which had long been held by Muller (see Muller, 1967) has been extended and elaborated with respect to behavioral characters under the name of sociobiology (Wilson, 1975). One aspect of the thinking of sociobiologists is exemplified by the title of the book *The Selfish Gene* by Dawkins (1976).

Paraphrasing this, Orgel and Crick (1980) proposed the concept of selfish DNA (see also Doolittle and Sapienza, 1980). According to these authors, selfish DNA is a piece of DNA that has little or no phenotypic effect, yet spreads in the species because of its rapid replication within the genome. The only function of such DNA is survival within genomes, so that it may be regarded as 'the ultimate parasite'. Highly and moderately repetitive nucleotide sequences in eukaryotes are considered to belong to this class of DNA. Transposable elements such as *copia* in *Drosophila* may also belong to this class.

The concept of selfish DNA has attracted a great deal of attention and

has been much discussed (see, for example, Orgel *et al.*, 1980). One problem which is often overlooked in such discussions is that extra replication alone within a genome is not sufficient for a selfish element to spread through the species. Transposition to another genome (chromosome set) within a diploid individual, and spread to other individuals by sexual processes must operate for the selfish element to spread through the species.

This means that random genetic drift must play a very important role in determining the fate of selfish DNA in the population, because, by definition, selfish DNA makes no specific contribution to organismal phenotype, and is therefore, selectively neutral in the Darwinian sense. It is desirable therefore to develop a population genetical theory to treat the evolution of selfish DNA. A start along this line has been made by Ohta and Kimura (1981), and Ohta (1981b). In these papers, two classes of selfish DNA have been distinguished, namely, tandemly repetitive sequences and dispersed repeated families.

In reality, however, these two classes may not be completely independent, as suggested by the recent paper by Childs *et al.* (1981) reporting a novel class of genetic elements called orphons. These are dispersed, solitary elements which are considered to have been displaced from tandemly repeated multigene families to new locations (somewhat like orphans that have lost their families). The authors (Childs *et al.*, 1981) analysed in detail histone gene orphons from the sea urchin *L. pictus*, and found that there are more than 50 gene orphons in each urchin genome.

In addition, they found that orphon loci are polymorphic in the population. Orphons appear to have been separated from parental families relatively recently in evolutionary times, and are endowed with properties which differentiate them from ordinary transposons. The population genetical implications of such elements is yet to be worked out.

7.4 Rapid evolutionary change of pseudogenes

A pseudogene is a region of DNA which shows definite homology with a known functional gene but has lost ability to produce a functional product due to mutational changes (Proudfoot, 1980). It is sometimes called a 'dead gene', although some have objected to this expression (see for example, Vanin *et al.*, 1980) saying that a pseudogene may have an unknown but important function such as controlling the expression of normal genes.

In my opinion, it is much more appropriate to regard pseudogenes as indeed dead; they have been liberated from the constraint of negative selection after they were created by gene duplication followed by loss of their function by mutation. They have then been accumulating various mutational changes at the maximum speed allowable under mutational pressure and random drift. If this view is correct, their evolutionary change must conform to the quantitative prediction of the neutral theory.

Comparison of pseudogenes with their normal counterparts reveals many remarkable features. Not only do base substitutions occur at a very high rate but also the deletion and addition of nucleotides occur quite frequently. Unlike the 'conservative' mode of change that characterizes the evolution of many normal genes, base substitutions at the first and the second codon positions in the pseudogenes occur just as frequently as those at the third position. We have already discussed briefly these features with respect to the globin-like α-3 gene (Nishioka *et al.*, 1980) toward the end of chapter 4.

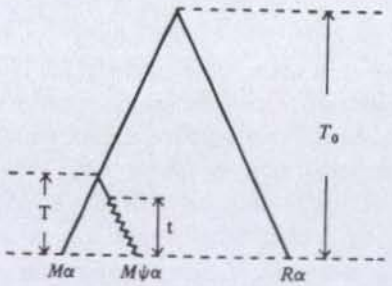
In addition, various signals in the non-coding region, which are essential for transcriptional control (such as the Hogness or 'tar-tar' box) and other post-transcriptional modifications often change to non-functional forms. For example, in the mouse, there are two pseudo α-globin genes denoted by α-3 and α-4 (Leder *et al.*, 1980). In one of them (α-3), the promoter site ('tar-tar' box) has changed from the normal TATA to an aberrant TAGA. From the result of experiments by Wasylyk *et al.* (1980), we can infer that the latter must be defective in carrying out the important function of initiation of transcription. They showed that by a single-base change in the third nucleotide of the T-A-T-A box of the conalbumin gene from T to G, the efficiency of specific transcription of that gene is drastically decreased in an *in vitro* system. Also, it has been shown by Tsujimoto *et al.* (1981) that, in the silk fibroin gene from *Bombyx mori*, the deletion of the second TA in the TATAAAA sequence at position $-30 \sim -24$ results in a severe reduction of the fibroin gene transcription. In another pseudogene (α-4), the poly (A) addition signal has changed from the normal AATAA to an abnormal AATAC.

Since we shall have more to say on α-3 later, I mention here a rather surprising set of changes found in the coding region of α-4. According to Leder *et al.* (1980), α-4 contains point mutations that would substitute a tyrosine for a histidine in amino acid position 58. This corresponds to hemoglobin M Boston that causes methemoglobinemia in man. It also has undergone changes that shift the normal termination codon out of phase so as to continue the globin reading for an additional 40 amino acids, analogous to hemoglobin Constant Spring in man. Since no such abnormal hemoglobins are found in the normal mouse, it is evident that α-4 is not expressed. These changes, together with other changes mentioned already

suggest that pseudogenes are in a headlong process of disintegration to become what Ohno (1972) called 'junk' DNA.

Although some casual calculations were reported earlier (Kimura, 1980b; Proudfoot and Maniatis, 1980) showing that pseudogenes evolved faster (in terms of base substitutions) than their normal counterparts, it was Miyata and Yasunaga (1981) who first made a careful analysis of evolutionary rate of a pseudogene. They made use of the data reported by Vanin *et al.* (1980) on the nucleotide sequence of a mouse α-globin-related pseudogene which is called ψα30.5 and which is essentially equivalent to the mouse α-3 gene of Nishioka *et al.* (1980). This gene has the remarkable feature that it completely lacks the two introns present in all the functional α- and β-globin genes. In other words, it resembles α-globin processed messenger RNA. Miyata and Yasunaga (1981) compared the sequence of this pseudogene and the productive α-globins from the mouse and the rabbit. They found that for every functional or structural block that can be compared, the mouse ψα30.5 gene (abbreviated as Mψα) is always much nearer to the normal adult α-globin gene of the mouse (Mα) than that of the rabbit (Rα). The only exception for this is the 3′ portion of the 3′ noncoding region, and this may be attributed to chance deviations and/or structural changes due to unequal crossing-over at this region in Mψα (for the structure of the α-globin gene, see Fig. 7.10). The result suggests that the pseudogene was created from a normal α-globin gene by gene duplication in the mouse genome after the mouse line diverged from the rabbit line. (According to the recent report by Lueders *et al.*, 1982, the pseudogenes αψ3 and αψ4, i.e. α-3 and α-4 mentioned above, of the mouse are located on

Fig. 7.11. Phylogenetic relationships assumed by Miyata and Yasunaga (1981) to analyse the evolutionary history of the mouse pseudo α-globin gene (Mψα). Letters Mα and Rα stand for the normal adult α-globin gene of the mouse and that of the rabbit. The zigzag part of a branch indicates the period during which the duplicated gene has accumulated mutations at a very high rate after loss of function.



chromosomes 15 and 17, respectively, whereas the three functional α-globin genes are clustered on chromosome 11. Furthermore, retrovirus-like elements exist on both sides of the αψ3 gene, suggesting that retroviral functions could have been instrumental in the formation of this pseudogene.)

In order to reconstruct the evolutionary history of this pseudo α-gene, they assumed a molecular phylogenetic tree involving the three genes as illustrated in Fig. 7.11.

In carrying out their analysis, Miyata and Yasunaga (1981) carefully noted that the third exon of the mouse pseudo α-globin gene differs from the other exons in that it is highly conserved. This suggests the possibility that this part of Mψα was derived from Mα or one of closely related α-globin genes by double unequal crossing-over (or gene conversion) at a relatively recent time, so they excluded this part in their analysis. They assumed that the gene duplication that led to Mψα occurred in the mouse line T years ago and that the duplicated gene evolved at the same rate as the normal α globin gene until t years ago. Let us designate the amino acid altering substitution rate by k_A and the synonymous substitution rate by k_S. After that time, they assumed that Mψα became inactive and all the sites in it have evolved at the same rate, which we denote by k_0.

Assuming that the divergence time of the mouse and the rabbit (T_0 in Fig. 7.11) is 75 million years ago, they obtained $T = 24$ million years for the duplication and $t = 17$ million years for the time since the loss of function. They also obtained $k_A = 0.82 \times 10^{-9}$, $k_S = 6.6 \times 10^{-9}$ and $k_0 = 12.6 \times 10^{-9}$ per site per year, so that $k_0/k_S = 1.91$. This means that the rate of mutant substitutions in Mψα is 1.9 times as high as the synonymous substitution rate of the productive α genes. This suggests that, in the normal functioning gene, there are still some selective constraints so that not all synonymous mutations are selectively neutral. Miyata and Hayashida (1981) consider that the evolutionary rate of the pseudogene may correspond to the maximum rate predicted by the neutral theory (Kimura, 1977), namely, this corresponds to the case of $f_0 = 1$ in equation (5.1). Whether the rate of evolution in terms of mutant substitutions in this case is really equal to the rate of point mutations per gamete (both being measured in years) is an extremely important problem for testing the validity of the neutral theory. This is one of the problems left for the future.

A similar analysis was carried out by Takahata and Kimura (1981) using a model of mutation which is more general than the 3ST model illustrated in Fig. 7.8. We followed Miyata and Yasunaga (1981) in excluding the third exon in the analysis. The main difference is that we assumed that Mψα lost

its function immediately after its birth by duplication (corresponding to $t = T$ in Fig. 7.11). We calculated the evolutionary rate separately for the three codon positions, denoting by k_i the rate at position $i (i = 1, 2$ or $3)$ for the normal gene. We assumed that duplication occurred T_d years ago, and immediately thereafter, the duplicated gene became 'dead' and started to evolve at the rate k_i' instead of k_i. Then the following ratios were obtained for the first, second and the third codon positions: $k_1'/k_1 = 11.5$, $k_2'/k_2 = 13.9$ and $k_3'/k_3 = 0.9$. Also, the ratio T_d/T_0 for the three codon positions turned out to be 0.26, 0.42 and 0.43, with an average of 0.37. Thus, if we assume $T_0 = 8.0 \times 10^7$ years for the divergence time of the mouse and the rabbit, we get $T_d = 3.0 \times 10^7$ years, namely, this dead gene was created about 30 million years ago by duplication. Although these values are tentative, it is likely that base substitution rates increased some ten times in the first two codon positions after the gene lost its function with the result that selective constraint disappeared entirely.

An elegant statistical analysis of the evolutionary rates of pseudogenes has been presented by Li *et al.* (1981). They used sequence data on human globin pseudogene $\psi\alpha 1$ (Proudfoot and Maniatis, 1980), rabbit globin pseudogene $\psi\beta 2$ (Lacy and Maniatis, 1980) and mouse $\psi\alpha 3$ of Nishioka *et al.* (1980). By comparing each of these with their normal counterparts and more remotely related globin sequences similarly to what was done by Miyata and Yasunaga (see Fig. 7.11), Li *et al.* estimated the rates of nucleotide substitutions at the first, second and third codon positions (i.e. k_1, k_2 and k_3 in our terminology), the time since duplication (corresponding to T in Fig. 7.11), and the time since loss of function (t in Fig. 7.11). They also estimated the rate of substitution after a gene became nonfunctional, assumed to be equal for all the three codon positions. Let us denote this rate by k_0. They applied a least-squares method to obtain estimates for t and k_0.

The estimated values when averaged for the three cases (set of trees) turned out to be as follows. The average substitution rates at the first, second and third codon positions of codons in the normal globin genes are $k_1 = 0.71 \times 10^{-9}$, $k_2 = 0.62 \times 10^{-9}$ and $k_3 = 2.64 \times 10^{-9}$, and the substitution rate per site of the pseudogenes during their nonfunctional periods is $k_0 = 4.6 \times 10^{-9}$ (per years). They noted that 4.6×10^{-9} is one of the highest rates of nucleotide substitution so far estimated.

Although Li *et al.*'s estimate for k_0 is less than half as large as the corresponding estimate by Miyata and Yasunaga, the ratio $k_0/k_3 = 1.74$ does not differ much from a similar ratio $k_0/k_s = 1.91$ obtained by Miyata and Yasunaga. These results suggest that the synonymous substitution rate,

even if high, is not the real maximum rate predicted by the neutral theory, and therefore the synonymous mutations are subject to negative selection even if the intensity of selection involved must be extremely weak. We shall discuss this problem in the next section in relation to the 'non-random' usage of synonymous codons. To sum up, various observations on the evolution of pseudogenes can be interpreted most naturally by the neutral theory of molecular evolution.

7.5 'Non-random' usage of synonymous codons

It has now been well established that synonymous base substitutions (i.e. base substitutions that do not cause amino acid changes) occur much more frequently in evolution than amino acid altering substitutions, sometimes in a dramatic fashion as in the histone H4 gene (see section 7.3 above). This fits the expectation of the neutral theory quite well, for, as compared with amino acid substitutions, synonymous substitutions are very much less likely to change organismal phenotypes, and therefore, other things being equal, must be much less subject to Darwinian natural selection.

On the other hand, it has become increasingly evident that synonymous or 'degenerate' codons are used quite unequally or in 'non-random' fashion in many genes of various organisms (see, for example, Grantham *et al.*, 1980a and Grantham, 1980). It appears that non-random codon usage is the rule rather than an exception, and this is often mentioned as evidence against the neutral theory. The situation is analogous to what was claimed, several years back, against the neutral theory based on observations on the conservative nature of amino acid substitutions (see section 7.1 in this chapter). Note that the existence of selective constraint by no means contradicts the neutral theory.

In this section, I shall try to show that non-random usage of synonymous codons can also be explained satisfactorily in the framework of the neutral theory, although the problem involved is more intricate and, therefore, a more careful analysis is required.

One of the best examples showing non-random usage of synonymous codons is afforded by codons coding for the amino acid leucine (Leu). As seen from the standard code table, there are six codons that code for Leu; they are UUA, UUG, CUU, CUC, CUA and CUG. Of these six, the codon CUG is used predominantly by genes in bacteria (Grantham *et al.*, 1981). This may be seen in Fig. 7.12 where the relative frequencies of usage (expressed per thousand) of 61 codons (excluding the initiator and

terminator codons) among 25 genes of *E. coli* are listed under the heading ECO. In fact, the frequency of usage of CUG amounts to about 60% of all the six codons coding for leucine.

A quite different pattern of codon usage is apparent in yeast

Fig. 7.12. Frequencies of codon usage (expressed per thousand) in three organisms, a bacterium, yeast and man (excluding initiator and terminator codons). The columns headed by ECO list observed codon frequencies per mille for 25 genes of *E. coli* (data from Fig. 3(b) of Grantham *et al.*, 1981). The columns headed by SAC list codon frequencies (per mille), computed as the average for 3 genes, coding for ADH1, Iso-1-cytochrome c and glyceraldehyde-3-PO_4 dehydrogenase of *Saccharomyces cerevisiae* using data compiled by Grantham, Gautier and Gouy (1980a). The columns headed by HOM list similar frequencies for 4 human genes, alpha-chorionic gonadotropin, beta-globin, pre-growth hormone and chorionic somatomammotropin, computed by using data compiled by the same authors.

	ECO	SAC	HOM		ECO	SAC	HOM		ECO	SAC	HOM		ECO	SAC	HOM
Phe				Ser				Tyr				Cys			
	UUU	18	6 22		UCU	17	36 11		UAU	14	6 24		UGU	5	14 12
	UUC	18	30 38		UCC	13	21 28		UAC	12	29 18		UGC	6	3 26
Leu					UCA	8	3 9	Term				Term			
	UUA	7	3 1						UAA	–	– –		UGA	–	– –
	UUG	7	50 4		UCG	9	3 10		UAG	–	– –	Trp			
													UGG	12	6 8
Leu				Pro				His				Arg			
	CUU	8	3 5		CCU	5	3 10		CAU	16	6 12		CGU	30	0 1
	CUC	8	0 30		CCC	4	0 12		CAC	9	17 24		CGC	21	0 11
	CUA	2	6 14		CCA	7	30 16	Gln					CGA	3	0 0
	CUG	47	0 64		CCG	19	0 5		CAA	13	14 10		CGG	4	0 5
									CAG	29	0 38				
Ile				Thr				Asn				Ser			
	AUU	24	29 5		ACU	21	27 11		AAU	10	6 4		AGU	7	0 10
	AUC	32	28 17		ACC	22	35 20		AAC	25	38 38		AGC	9	0 12
	AUA	5	0 2		ACA	5	6 17	Lys				Arg			
Met									AAA	46	20 17		AGA	5	34 4
	AUG	22	18 20		ACG	10	0 12		AAG	18	79 41		AGG	2	0 20
Val				Ala				Asp				Gly			
	GUU	28	44 6		GCU	37	54 20		GAU	25	15 20		GGU	33	84 12
	GUC	9	43 18		GCC	20	33 26		GAC	27	39 32		GGC	29	9 33
	GUA	21	0 4		GCA	31	0 9	Glu					GGA	4	0 0
	GUG	17	6 40		GCG	23	0 2		GAA	37	47 21		GGG	6	6 12
									GAG	18	8 34				

(*Saccharomyces*) for which the relative frequencies are listed in the next column headed by the word SAC. For example, among the six codons coding for leucine, UUG is by far the most frequently used codon (amounting to about 80%). In the last column headed by HOM, frequencies of codon usage are listed for some human genes. It appears that, as far as the pattern of codon usage is concerned, man is nearer to bacteria than to yeast.

The most thorough analysis of the pattern of codon usage which has been made so far is that of Grantham and his associates (Grantham *et al.*, 1980a, b, 1981). They compiled extensive tables of codon usage from reported messenger RNA sequences of various organisms. The number of such sequences amounted to 90 in January, 1980, and 119 in June, 1980, and more recently, 161 in 1981.

They applied 'correspondence analysis', a multivariate statistical method, to the distances between messenger RNAs based on differences in the usage of 61 codons; each messenger was treated as a point in a multidimensional space. Then the data were projected on to a plane whose horizontal and vertical axes correspond to the first and second most important factors. Grouping was done by an automatic classification. The results of analysis of 119 mRNAs (Grantham *et al.*, 1980b) showed that most genes in a genome have the same 'coding strategy' with respect to choices among synonymous codons. Thus, mammalian, bacterial, virus, mitochondrial and yeast plus slime mold genes fall in different classes. In other words, there is a consistent choice of degenerate bases, and this confirms 'the genome hypothesis' of Grantham *et al.* (1980b) which states that each gene in a genome tends to conform to its species' usage of the codon catalog.

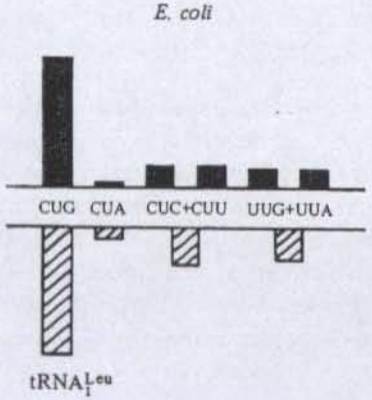
On the other hand, application of correspondence analysis to amino acid frequencies of 119 proteins show that no grouping of proteins by genome type is evident; viral, bacterial, mammalian and other proteins lie in the same class. This means that mRNA sequences contain information other than that necessary for coding proteins, the other information being mainly concerned with the choice of degenerate bases.

Grantham (1980) claims, based on these observations, that messenger RNA is an evolutionary structure in its own right, and that the work by his group reveals protein-independent molecular evolution of a 'non-neutral character'. In my opinion, however, what he calls 'non-neutral' is more an expression of selective constraint rather than that of the adaptive mutant substitutions that Grantham seems to have in mind.

Then, the problem is what is the main cause of selective constraint that brings about such a choice of degenerate codons characteristic of each

genome type. An important step forward for understanding this cause has recently been made by Ikemura (1980, 1981a, b). Using two-dimensional polyacrylamide gel electrophoresis, he separated the 26 known tRNA of *E. coli* and measured their relative abundance in terms of molecular numbers in the cells. He then investigated the relationship between the tRNA abundance and the frequency of usage of the corresponding codons in *E. coli* genes, such as tryptophan synthetase A protein (*trp* A), lac operon repressor (*lac* I) and ribosomal protein (r-protein) genes. A very strong correlation was found between the tRNA abundance and choice of codons among synonymous codons and also for codons corresponding to different amino acids. Of particular interest, in the present context, is the finding that among synonymous codons for an amino acid the most frequently used codon invariably corresponds to the most abundant isoaccepting tRNA species. Fig. 7.13 illustrates this relationship for the six synonymous codons coding for leucine. The solid columns above represent the relative frequencies of codon usage and the hatched columns below represent the abundance of the cognate tRNA species. It is clear that the agreement

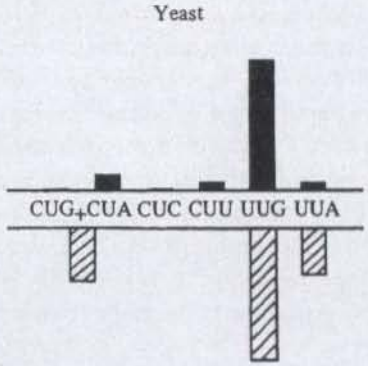
Fig. 7.13. Diagram illustrating the relationship between the relative frequencies of usage of codons for leucine (solid columns above) and the relative abundance of the corresponding cognate tRNA species (hatched colums below) in *E. coli*. The frequencies of usage are based on the data given in Fig. 7.12, while the data on the abundance of tRNA's are from Ikemura (1981a, b). The plus sign between codons CUC and CUU means that these two codons are recognized by a single tRNA species (i.e. $tRNA_2^{Leu}$), whose relative abundance (taking that of $tRNA_1^{Leu}$ as unity) is given by the hatched column below. There is a similar meaning for the plus sign between UUG and UUA. The relative amount of tRNA corresponding to CUA, which is a minor species, is 0.1 (Ikemura, personal communication).



between these two is excellent. Note that codons CUC and CUU are recognized by a single tRNA species ($tRNA_2^{Leu}$), and, so also are UUG and UUA.

Ikemura's proposal that codon usage is mainly constrained by the availability of the corresponding tRNA species is much strengthened by his recent finding that such a correlation also exists in yeast which has a very different pattern of codon usage from that of *E. coli*. For example, among six synonymous codons coding for leucine, the most frequently used one is UUG in yeast. Ikemura found (personal communication, July, 1981) that, in this case, the most abundant cognate tRNA species is exactly the one which recognizes UUG (see Fig. 7.14). Incidentally, these findings cast doubt on the validity of the suggestions made by Modiano *et al.* (1981) that UUA and UUG are 'pretermination codons' (that can mutate to termination codons by a single nucleotide substitution) and that, since four other nonpretermination codons are available for leucine, codons UUA and UUG are never used in the normal α- and β-globin genes. They claim that usage of pretermination codons is avoided whenever possible as an evolutionary strategy for reducing the rate of mutation with drastic effects. It seems to me that the selective advantage coming from such a strategy is too small (presumably the order of the mutation rate) to be effective in the actual course of evolution.

Fig. 7.14. Diagram illustrating the relationship between the frequencies of usage of synonymous codons for leucine and the abundance of the corresponding cognate tRNA species in yeast. As in Fig. 7.13, the solid columns represent the relative frequencies of usage and the hatched columns the relative tRNA abundance. The frequencies of usage are based on the data given in Fig. 7.12, while the data on the relative abundance of tRNAs are from Ikemura (personal communication).



Previously, Fitch (1980) examined the mRNA sequences of β-hemoglobin for human, mouse and rabbit, and observed that there is a significant bias against the use of codons differing from terminating codons by only one base. Rarity of usage of UUA and UUG is also evident in other human genes (see Fig. 7.12), and, it is likely, if Ikemura's hypothesis is valid, that this is due to the paucity of the tRNA species that recognize these two codons rather than to these being 'pretermination' codons.

Let us assume that, in general, the choice of synonymous codon is largely constrained by tRNA availability, and that this is related to translational efficiency as suggested by Post *et al.* (1979) and Ikemura (1981b). This leads to a hypothesis that preferential codon usage represents the optimum state in which the population of synonymous codons matches that of cognate tRNA available in the cell. This will help to carry out more efficient cell function leading to higher Darwinian fitness. Then, the concept of stabilizing selection can be applied to treat the problem (Kimura, 1981d). This approach must be compatible also with 'the genome hypothesis' of Grantham *et al.* (1980b) which states that a surprising consistency in the choice of degenerate bases exists among genes of the same or similar genomes, and that 'the genome and not the individual gene is the unit of selection'.

In the following treatment, we assume that the pattern of codon usage changes in the course of evolution in such a way that they match those of cognate tRNAs. There is however, the reverse possibility of cognate tRNAs (as determined by the numbers of their genes) being adjusted to match the relative frequencies of synonymous codons. We do not know at present which one of these two may be regarded as the primary cause. It is quite likely that the influence is reciprocal, each regulating the other's change, and they shift very slowly in unison. In this section, however, I follow Ikemura's (1981a, b) opinion and emphasize the aspect of the codon usage being determined by the frequency of tRNAs.

We make use of the results obtained toward the end of the previous chapter (see section 6.8) where the theories on the probability of fixation of a mutant allele and the distribution of allelic frequencies are presented (assuming two allelic states). Let us consider a particular site corresponding to the third codon position, and, to simplify the treatment, let us group the four bases into two classes, designating bases U (uracil) and A (adenine) as A_1 and G (guanine) and C (cytosine) as A_2. We denote the frequency of allele A_2 by x so that the frequency of A_1 is $1-x$. If v_1 is the mutation rate from A_1 to A_2 and if v_2 is the rate in the reverse direction, then equation

(6.65) may be used to describe the steady state distribution of x under mutation and random drift.

Let Q_{op} be the optimum proportion of A_2 (G or C) at the third codon position, and assume that mutation rates are equal between A_1 and A_2, i.e. $V_1 = V_2$ in equation (6.65). The validity of this assumption (i.e. equality of mutation rates), will be discussed later. Then the mean of x (i.e. $\bar{x}$) does not coincide with Q_{op} unless $Q_{op} = 0.5$. So, we assume that stabilizing selection is at work to hold $\bar{x}$ near Q_{op}. At individual sites, however, A_2 is either fixed or lost most of the time (see Fig. 6.8). Let f_2 be the probability that A_2 is fixed in the population at a given site. Similarly, let f_1 be the probability that A_1 is fixed (A_2 is lost). Then, from equation (6.68), we have $f_2/f_1 = \exp(B_1)$, where $B_1 = 4N_e\beta_1 = -4N_e\lambda ma$. Thus, we can estimate B_1 by the relation $B_1 = \ln(f_2/f_1)$, and we obtain $B_1 = 0.85$ for $f_2/f_1 = 0.7/0.3$. Since $-\lambda ma = B_1/(4N_e)$, and since N_e in most mammalian species must be at least 10^4, the intensity of selection which acts at an individual site to produce 'non-random' codon usage is an exceedingly weak one, leaving plenty of room for random drift to operate. This is consistent with Latter's (1975) claim that mutations responsible for enzyme polymorphisms are very slightly deleterious with 'Ns' values in the range 1–3.

One important question that remains is to what extent the rate of evolution in terms of mutant substitutions is influenced by such selection. Since the relative evolutionary rate (in terms of mutant substitutions) under stabilizing selection as compared with the strictly neutral case is given by u/u_0 with $u_0 = 1/(2N)$, we have, from equation (6.60),

$$u/u_0 = 1 \Big/ \int_0^1 \exp[-B_1 x + B_2 x(1-x)]\,dx, \tag{7.8}$$

where $B_1 = -4N_e\lambda ma$ and $B_2 = 2N_e\lambda(1-\lambda m^2)a^2$. If we assume that $2N_e\lambda a^2$ is negligibly small so that $B_2 \approx 0$, then we get

$$u/u_0 \approx 2f_1 f_2 \ln(f_2/f_1)/(f_2 - f_1). \tag{7.9}$$

For $f_2/f_1 = 0.7/0.3$, as we observe at the third codon position in globin and other mammalian mRNAs, we get $u/u_0 = 0.89$. In other words, the evolution is retarded by about 10% from what is expected under complete selective neutrality. Under the more extreme situtation of $f_2/f_1 = 0.9/0.1$, we get $u/u_0 \approx 0.49$ which means nearly 50% retardation. If B_2 is not negligible, however, the amount of retardation becomes larger. Assuming that $f_2/f_1 = 0.7/0.3$, we get $u/u_0 = 0.63$ if $B_2 = 2$. This corresponds to the case for which the gene frequency distribution is illustrated in Fig. 6.8 by a

solid curve. If $B_2 = 4$, we get $u/u_0 = 0.44$, and if $B_2 = 10$ we get 0.14. Probably, $B_2 < 2$ is more realistic than such larger values because a is likely to be small and therefore a^2 must be an exceedingly small quantity.

In actual situations, however, there are other complications. First, the number of possible 'alleles' (bases) at each nucleotide site is four rather than two, and inequality of the frequencies of U and A, and also of C and G might further retard the evolutionary rate, although as far as the ratio u/u_0 is concerned, its effect may not be very large. Secondly, the amount of the selective constraint may differ from gene to gene depending on production levels of individual genes. According to Ikemura (19801a, b) the amount of change in fitness (Δw) due to the change of production rate (Δr) caused by a synonymous alteration is given by the equation

$$\Delta w = Pn \cdot \Delta r, \tag{7.10}$$

where Pn is a parameter which is closely related to the production level of a gene (Ikemura, 1981b). This implies that, for a given change (Δr) introduced by a synonymous alteration, the extent of fitness change (Δw) is proportional to Pn. In other words, the effect on fitness is larger for genes of more abundant proteins (such as r-protein genes). Ikemura reasoned that the proportion of mutations that can be regarded as neutral or nearly neutral should be larger for genes with low Pn values than for genes with high Pn values.

Ikemura's equation (7.10) could be paraphrased in more general terms to fit the neutral theory as follows;

$$\bar{s}' \propto Pn \cdot |a|, \tag{7.11}$$

where $\bar{s}'$ is the average selection coefficient against the mutant and $|a|$ is the effect of substituting the mutant for the pre-existing allele.

As to the difference in the abundance of different tRNA species, Ikemura (1981a) has shown that, in *E. coli*, abundance is proportional to the number of genes concerned; an abundant tRNA species is encoded by several gene copies and a minor tRNA usually by a single copy.

On the other hand, the validity of our assumption that mutation rates (not necessarily the rates of evolutionary base substitutions) are equal among the four bases (and therefore between A_1 and A_2) seems to be well supported from data obtained through comparison of base compositions of pseudogenes with those of the normal counterparts. In their study of the evolutionary rates of pseudogenes, Miyata and Hayashida (1981) made the interesting observation that U-, C-, A- and G-ending codons occur much more uniformly in globin pseudogenes than in their functional counterparts. It is expected, however, that even in pseudogenes there still remains

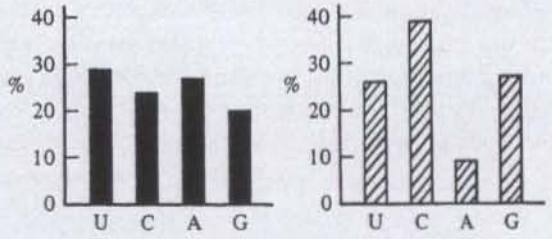
the effect of base composition at the time when they were functional. So, in order to eliminate such an effect as much as possible, I have only scored the base frequencies at the third codon positions for which globin pseudogenes and their normal counterparts differ from each other.

If pseudogenes are free from selective constraint and their base substitutions have occurred at high rates, the base composition among the set of the third codon positions chosen from the pseudogenes should reflect better the pattern of mutational changes unaffected by selection. Table 7.5 lists the frequencies of the four bases thus scored at the third codon positions of four globin pseudogenes. These frequencies were obtained by comparing mouse α-3 with mouse α-1 (data from Nishioka *et al.*, 1980), rabbit $\psi\beta 2$ with rabbit $\beta 1$ (data from Lacy and Maniatis, 1980), human $\psi\alpha 1$ with human α2 (data from Proudfoot and Maniatis, 1980), and mouse βh3

Table 7.5. *The observed numbers of four nucleotide bases at the third codon position in globin pseudogenes. Base frequencies were scored only in those third positions where the pseudogene differs from one of its functional counterparts in the same species. For details see text*

	Observed numbers				
Pseudogene	U	C	A	G	Total
Mouse α-3	11	6	6	6	29
Rabbit $\psi\beta 2$	9	11	9	5	34
Human $\psi\alpha 1$	17	7	10	10	44
Mouse βh3	1	7	11	6	25
Total	38	31	36	27	132
%	28.8	23.5	27.3	20.4	100

Fig. 7.15. The solid columns at the left represent the relative frequencies of the four bases at the third codon positions in the globin pseudogenes, while the shaded columns at the right show those in the functional globin genes. For details, see text.



with mouse β_{maj} (data from Miyata and Hayashida, 1981). It is clear from this table that the four bases occur in roughly equal frequencies (supporting the assumption $V_1 = V_2$). On the other hand, the base composition at the third codon positions in the normal counterparts of these pseudogenes deviates significantly from equality. In Fig. 7.15, base frequencies (in %) thus obtained are compared between globin pseudogenes (solid columns at the left) and their functional counterparts (shaded columns at the right). It is interesting to note that the latter are very similar not only to the base composition at the third codon positions of the globin sequences in general, but also, to those of the third positions among four-fold degenerate ('quartet') codons in mammalian genes as illustrated by Grantham (1980) in his Fig. 3. Miyata and Hayashida (1981) seem to be correct in claiming that the evolutionary rate of pseudogenes corresponds to a limiting rate as predicted by Kimura (1977) based on the neutral theory, and with the assumption that there are some selective constraints against synonymous changes.

In their earlier attempts to estimate the proportion of all point mutations which are evolutionarily unacceptable (using data on human and rabbit β-hemoglobin mRNA sequences), Boyer *et al.* (1978) obtained, among other things, the following conclusions. (1) At least 95% of amino acid-altering substitutions have been unacceptable by primate β-hemoglobin genes. (2) Roughly 70% of synonymous β-Hb mutations are adaptively undesirable. (3) Only about 11% of all kinds of point mutations are acceptable and the mutation rate is at least 10^{-8} per site per year. (4) Non-randomness in the codon choice arises due to selection exerted on pretranslational events. (5) Selective constraint on nucleotide doublets influences not only non-synonymous changes in evolution but also contributes to the genesis of genetic polymorphisms.

Although their conclusions may need some revisions in the light of the more recent studies which I have described already, Boyer *et al.* (1978) are justified in their claim that not all synonymous mutations are strictly neutral and, indeed, a significant fraction of them are subject to negative selection (for the problem of CpG doublet deficiency, readers may refer to Bird's (1980) observation discussed toward the end of chapter 3).

Probably the most significant fact revealed from the various studies summarized in this section is that evolution (in terms of base substitutions) is *slowed down rather than accelerated* by non-random choice among synonymous codons. It supports the neutralist interpretation that the selection involved is negative (coming from stabilizing selection) rather than positive Darwinian selection usually envisaged by the selectionists.

Although much remains to be clarified, I find it pleasing that such diverse observations as the rapid change of pseudogenes and non-random usage of synonymous codons can be understood in a unified and quantitative way under the framework of the neutral theory by incorporating the concept of stabilizing selection. Here again, the paramount importance of random genetic drift in molecular evolution is evident.

8

Population genetics at the molecular level

8.1 Why a stochastic treatment is required

This chapter is intended to serve as a preliminary to the next chapter where the mechanism for the maintenance of genetic variability at the molecular level will be discussed. Since the neutral theory treats intraspecific variability as one aspect of the stochastic theory of molecular evolution, it is essential to have a good understanding of the stochastic theory, particularly as it is applied to molecular mutants in finite populations. Actually, I have already presented some accounts of the theory, notably in chapter 3, and also in section 6.8 of chapter 6. In this chapter I shall present more systematically the relevant aspects of the theory treating gene frequency changes as stochastic processes, where the term stochastic process means the mathematical formulation of chance events proceeding in time.

The necessity of stochastic treatment stems from the uniqueness of mutations and the finiteness of populations; it comes from the fact that individual molecular mutants are practically unique because of their exceedingly low incidence, so that in a finite population their fates are largely controlled by the random sampling of gametes in reproduction (see chapter 3).

To see these points in more concrete terms, let us consider a hypothetical mammalian population whose generation length is exactly one year, and which comprises a hundred thousand individuals each year. To simplify the story, let us also assume, for the moment, an ideal situation in which the effective size of the population is equal to the actual size (i.e. $N_e = N = 10^5$). We consider a gene locus that codes for an imaginary protein consisting of 333 amino acids (somewhat similar in size to alcohol dehydrogenase). This protein corresponds to 999 nucleotide sites in the coding region of the gene.

Let us also assume that the mutation rate per site per generation, denoted by v_{nuc}, is 5×10^{-9} and that the mutation rates among the four nucleotide bases (A, C, G and T) are all equal so that the mutational change from one specific base to another (say from C to A) occurs at the rate $(5/3) \times 10^{-9}$. This means that the expected number of new mutations per site per generation in the whole population is $2Nv_{nuc}$ or 0.001. With such a low rate, a mutant gene, when it appears at a given site is likely to be represented only once in the population. Also, with this rate, mutations occur at a given site on the average at intervals of 1000 generations in the population. Furthermore, it can be shown that the great majority of mutants are lost from the population by chance within 20 or so generations.

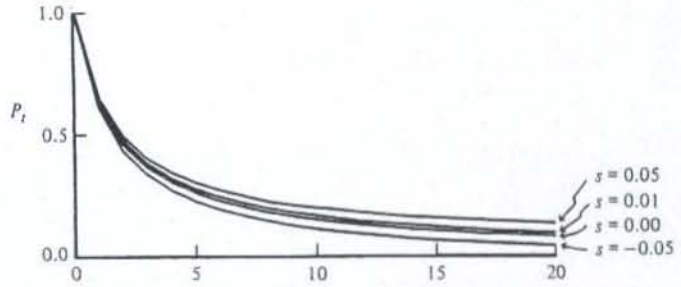
The mutation rate, $v_{nuc} = 5 \times 10^{-9}$, was obtained from the estimated rate of evolution of globin pseudogenes, which on the neutral hypothesis is equal to the mutation rate (see section 7.4). There is independent evidence, not requiring the neutral assumption, from the frequency and distribution of rare human hemoglobin variants in the Japanese population. Using these data, Kimura and Ohta (1973a) estimated the mutation rate per amino acid as $v_{aa} = 4.5 \times 10^{-9}$, which corresponds to $v_{nuc} = 1.5 \times 10^{-9}$. The agreement is as good as can be expected, given the large uncertainty of the estimates, and I shall take 5×10^{-9} as a representative value.

In Fig. 8.1, the survival probability of a newly arisen mutant, as computed by the recurrence formula

$$P_t = 1 - \exp\{-(1+s)P_{t-1}\} \tag{8.1}$$

is illustrated for a few cases having different selection coefficients. In this formula, P_t stands for the survival probability of a mutant after t

Fig. 8.1. Probability of survival (P_t) of a mutant as a function of time t (in generations) after its occurrence. The selection coefficient of the mutant is denoted by s. The course of change in the survival probability was computed by using equation (8.1).



generations ($t = 0, 1, 2, \ldots$), assuming $P_0 = 1$. This problem was first solved by Fisher (1930a, b). For an explanation, readers may refer to Crow and Kimura (1970, p. 421). What is remarkable, and still sometimes overlooked, despite the great influence of Haldane and Fisher, is that for individual mutants their usual fate is extinction and that this applies even to mutants with a definite selective advantage $4Ns \gg 1$ unless the advantage is very large.

In the following I shall consider the behavior of neutral mutants in some detail, because of their obvious importance for this book. Furthermore, the behavior of very weakly selected mutants is similar in many respects to that of neutral mutants. The average length of time until fixation of a selectively neutral mutant, excluding the cases of eventual loss, is $4N_e$ generations (Kimura and Ohta, 1969a), which means 400 000 generations (or years) in our example. On the other hand, the average number of generations until loss of a mutant, excluding the cases of eventual fixation, is

$$\bar{t}_0 = 2(N_e/N)\log_e(2N) \tag{8.2}$$

generations. This turns out to be 24.4 generations for the present example. Thus an enormous difference exists between the average time until fixation and that until extinction. The latter, however, has a standard deviation

$$\sigma(t_0) \approx 4N_e/\sqrt{N}, \tag{8.3}$$

which turns out to be approximately 1265 and which is much larger than its mean. On the other hand, the time until fixation has a standard deviation of about $(2.15)N_e$ generations, roughly half the mean (Kimura and Ohta, 1969b).

Although the overwhelming majority of mutant genes are lost from the population, a lucky minority spreads through the whole population. For a neutral mutant, this probability of fixation is equal to $1/(2N)$ which is 1 in 200 000 in our example.

During the long course of evolution, turnover of allelic states will occur; the predominant nucleotide base that occupies a given site changes from one form to another (say, from C to A). In our present example, the average interval that separates two consecutively fixed states is roughly $1/v_{nuc}$ or 2×10^8 generations.

To understand the process involved, a consideration of the steady-state gene frequency distribution will be of help. Since the possible number of different bases at a given nucleotide site is four, three frequency parameters are required to represent allelic frequencies, and four-dimensional space is needed to illustrate the distribution. However, if mutation rates are equal among the four bases and if they are selectively neutral (i.e. equivalent), we

may simplify our treatment by fixing our attention on one of the bases, say A, while grouping the remaining three bases as one class (calling this A'). Let x be the frequency of base A at a given site in the population. Then, the steady state distribution of x is given by Wright's formula

$$\phi(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(1-x)^{\alpha-1}x^{\beta-1}, \tag{8.4}$$

where in this case $\alpha = 4N_e v_{nuc}$ and $\beta = \alpha/3$, and $\Gamma(\cdot)$ stands for the Gamma function (see Kimura, 1968b; this is a special case of equation 3.5 therein for $K = 4$): $\phi(x)\mathrm{d}x$ can be interpreted as the probability of the frequency of A lying within the short interval between x and $x + \mathrm{d}x$. This distribution allows another interpretation which is more appropriate in the present context; if we trace the allelic composition at this site in the long course of evolution, the relative number of times in which the allele A happens to have frequency x is proportional to $\phi(x)$. The diagram at the left side of Fig. 8.2 depicts this distribution for $N_e = N = 10^5$ and $v_{nuc} = 5 \times 10^{-9}$ ($\alpha = 0.002$). It is evident that the distribution is markedly U-shaped.

The probability of A being temporarily fixed in the population may be obtained by integrating $\phi(x)$ from $x = 1 - 1/(2N)$ to $x = 1$, and we obtain

$$f_1 = C_1(2N)^{-\alpha}, \tag{8.5}$$

where $C_1 = \Gamma(\alpha+\beta)/[\Gamma(\alpha+1)\Gamma(\beta)]$ (Kimura, 1968b). In our present example, this turns out to be 0.24397. The probability of complete fixation of one or another of the four bases is $4f_1$ or 0.97588, so that the probability that the population contains two or more different bases simultaneously is $(1 - 4f_1)$ or 0.02412. These calculations give the probabilities of the fixed state or monomorphism (and also those of a segregating state or polymorphism) in the strict sense of the words. In practice, however, when we define monomorphism, it is more convenient to include cases in which rare variants exist in the population, because they are unlikely to be detected with a limited sample size.

So let us call a population 'monomorphic' if the total frequency of 'variant' alleles (or less frequent bases) is q or less. It is customary to take $q = 0.01$. Under this definition, the probability that the population is monomorphic (at the level of q) with respect to nucleotide base A is

$$P_{mono}(\mathrm{A}) = \int_{1-q}^{1} \phi(x)\mathrm{d}x \approx C_1 q^{\alpha}. \tag{8.6}$$

In our present example where $\alpha = 0.002$, $C_1 \approx 0.25$ and $q = 0.01$, we get $P_{mono}(\mathrm{A}) \approx 0.2477$. This means that the probability of polymorphism, as computed by $1 - 4P_{mono}(\mathrm{A})$ is 0.009168.

The average homozygosity at this site is

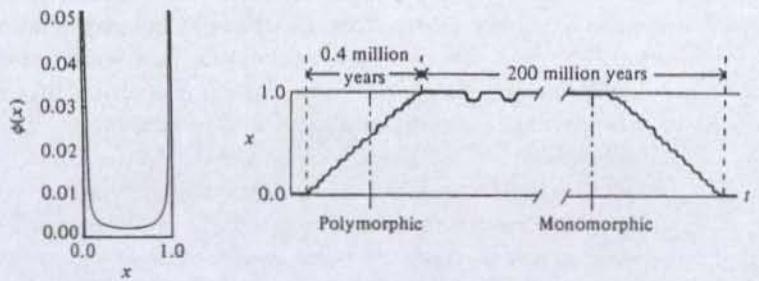
$$\bar{h}_o = 4\int_0^1 x^2\phi(x)dx = (\alpha+3)/(4\alpha+3), \tag{8.7}$$

which, in our example, turns out to be 0.998005. Then, the average heterozygosity ($\bar{h}_e$) at this site, as given by $1-\bar{h}_o$, amounts to 0.001995.

These calculations show that, at individual nucleotide sites, a polymorphic state is very rare. This ensures the existence of species-specific nucleotide sequences. In our hypothetical mammal, a change from one fixed state to another occurs at an average interval of 200 million years (generations), and, in between, the transitional (i.e. polymorphic) states occupy an average length of 400 000 years, which is a fraction amounting to only 1/500. The right hand side diagram of Fig. 8.2 illustrates the course of such changes.

Let us now consider evolutionary changes of the gene as a whole. In our example, the coding region of the gene comprises 999 nucleotide sites corresponding to 333 codons, and therefore the total mutation rate amounts to about 5×10^{-6} per gene per generation. Among the base changes in this region, those occurring at the first and second codon positions usually lead to amino acid changes, while only some 1/3 of those at the third positions lead to amino acid changes. Thus the total mutation rate for amino acid altering changes amounts to approximately 3.9×10^{-6} which we shall denote by $v_{T(A)}$. Let us assume (following Shaw, 1970) that on

Fig. 8.2. Diagrams illustrating the process of evolutionary base substitutions at a given nucleotide site in a hypothetical mammal consisting of 10^5 breeding individuals and having a generation span of one year. It is assumed that the mutation rate per site per generation is $v_{mut}=5\times10^{-9}$ and that mutation rates between any two of the four bases (A, C, G, T) are all equal. The diagram on the left-hand side shows the probability distribution of the frequency (x) of base A, while the diagram on the right-hand side depicts a typical course of change in the frequency of A with time (t). For details, see text.



the average 1/3 of the amino acid replacements can be detected by standard methods of electrophoresis. Then, the rate for electrophoretically detectable mutations is 1.3×10^{-6} per generation which we shall denote by $v_{T(E)}$. A large fraction of such mutational changes are likely to be deleterious, so let us assume that on the average only 1/10 of amino acid altering mutations are selectively neutral, while the remaining 9/10 are definitely deleterious contributing neither to molecular evolution nor polymorphism. Under these simplifying but not totally unrealistic assumptions, the mutation rate for electrophoretically detectable selectively neutral mutations is 1.3×10^{-7} per gene per generation. This rate will be denoted by $v_{0(E)}$.

Then, how much genetic variability is expected at this locus in the form of protein polymorphism (as detected by the standard electrophoretic method)? To answer this question, we use the 'model of infinite alleles' which we shall explain in more detail in the next section. This model is based on the consideration that a gene consists of a large number of nucleotide sites each of which may be occupied by one of four kinds of bases and therefore, in combinatorial terms, the possible number of allelic states is astronomical. In the context of the present calculation, this amounts to the assumption that whenever an amino acid change occurs in the protein, it represents a new type, not pre-existing in the history of the population. Under this model, the expected heterozygosity at this locus is

$$\bar{H}_e = \frac{4N_e v_{0(E)}}{4N_e v_{0(E)}+1} \tag{8.8}$$

(Kimura and Crow, 1964). For $N_e=10^5$ and $v_{0(E)}=1.3\times10^{-7}$, this gives $\bar{H}_e\approx0.049$ or 4.9% heterozygosity. Among vertebrate species so far investigated by various authors, the distribution of the average heterozygosity has a mode around 5% (see chapter 9), so this is a realistic value. The infinite allele model also allows us to compute the probability of protein polymorphism by the formula

$$P_{poly}=1-q^V \tag{8.9}$$

(Kimura and Ohta, 1971b), where $V=4N_e v_{0(E)}$. For $q=0.01$, this turns out to be 0.213; namely, if a large number of equivalent loci in the genome of this mammal could be studied, the expected fraction of polymorphic loci is 21.3%. This is again a realistic value. According to Nevo (1978) the mean proportion of polymorphic loci per population (P) and the average heterozygosity per locus per individual (H) are respectively $P=0.147\pm0.098$ and $H=0.0359\pm0.0245$ for 46 mammalian species.

The next question is: what is the rate of evolution in terms of amino acid substitutions in our hypothetical mammal? This rate is the one usually

estimated by comparative studies of homologous protein sequences and it is expressed taking one year as the unit of time. In this book we have denoted this by k_{aa} (see chapter 4). Theoretically, this rate can be obtained by noting that it is equal to the rate at which amino acid altering mutations that are selectively neutral occur per gamete per year (see equation 3.20). In our example, we have assumed that only 1/10 of amino acid altering mutations are selectively neutral and that $v_{nuc} = 5 \times 10^{-9}$. Therefore, the relevant mutation rate per codon (which comprises three nucleotide positions) is $v_{0(A)} = (1 + 1 + 1/3)v_{nuc} \times (1/10)$ per year, which turns out to be 1.17×10^{-9}. Thus, for our hypothetical mammal, the rate of molecular evolution is $k_{aa} = 1.17 \times 10^{-9}$ per amino acid site per year.

We have already seen in chapter 4 that the median of the evolutionary rates of 21 proteins (estimated mostly for mammals) is 1.3×10^{-9} per amino acid site per year, so the present k_{aa} value is quite realistic. Note that this rate is independent of the effective population size, because it represents the selectively neutral substitutions.

Finally, I would like to point out that although we made a simplifying assumption that the effective and actual population sizes are equal, this is a highly unrealistic assumption. In natural populations the relevant effective population size (N_e) for discussing evolution and variation at the molecular level must usually be much smaller than the apparent population size (N). If there is a great deal of mortality between censusing and reproduction or if reproduction is restricted to just a few (as for males in highly polygynous species) the effective number may be orders of magnitude less than the census number. As shown in chapter 3, such factors as inequality of the numbers of breeding males and females, fluctuation of the population size from generation to generation, inheritance of fertility, local extinction of colonies followed by recolonization etc. all contribute to make N_e smaller than N. Especially, if the population size goes through a sequence of bottlenecks in the course of evolution and if recovery from reduced population size after each bottleneck is slow, the average heterozygosity will be much reduced (Nei *et al.*, 1975), contributing greatly to a decreased value of N_e. This means that we tend to overestimate the average heterozygosity if we apply equation (8.8) taking N_e as the value of the population size of a species when it is in a flourishing condition. In nature, even if some species are widely distributed and cover an enormous area comprising an immense number of individuals, it is rather unlikely that they have always been so during the previous millions of years and will remain so in the coming millions of years. Sooner or later such a state will be disrupted by the process of speciation. For an average mammal, such as we have in

mind, it is reasonable to assume that, unless it becomes extinct, it changes into a new species at intervals of half million years or so, and that in each stage of speciation, a significant reduction of population size occurs. Probably no species on the earth can maintain an immense population size for millions of generations. This must be the main reason why the observed heterozygosity per enzyme locus in various organisms seldom exceeds 30% even in very large populations such as insects. In mammals, according to Nevo's (1978) Fig. 2, the average heterozygosity (H) among 48 mammalian species is distributed almost uniformly between 0% and 7%, although, in two exceptional cases, the H values are around 10%.

In our hypothetical mammal, we showed that, at an individual nucleotide site, the turnover of bases occurred at an average interval of 200 million years. During this period, then, it must have gone through 400 episodes of speciation. Even the duration of a neutral polymorphic state, which is $4N_e \pm 2.15N_e$ or $400\,000 \pm 215\,000$ generations, can sometimes be longer than the history of a species.

Thus, although the change of gene frequencies is random, the process of change is extremely slow by ordinary standards. Suppose that the frequency of a polymorphic allele at a given locus happens to be p, then the change in one generation of this frequency in the whole species has a mean of zero and variance

$$\sigma^2_{\delta p} = p(1 - p)/(2N_e). \tag{8.10}$$

Therefore, a probable amount of change (i.e. $\sigma_{\delta p}$) in one generation is less than 0.12%. Even with observations extending over 10 generations for a polymorphic allele whose frequency is around 50%, the probable amount of change is only about 0.35%.

If the species consists of a number of local colonies or 'demes', the changes in individual demes are expected to be larger, but as a whole they tend to average out to the above value. If, by chance or by some unknown cause, the gene frequency of one deme is greatly disturbed, it will sooner or later be brought back, through migration from neighboring demes, to a frequency not very different from the average frequency of the species as a whole. This may give the impression that the polymorphism is balanced. Actually, the polymorphism is transient if it is selectively neutral, but for our ephemeral existence, it is almost as if permanent.

Incidentally, a rapid decrease of the population size during the last few generations (such as is found in some game animals due to overhunting) is unlikely to cause a significant reduction in heterozygosity, except for extreme situations. To see this point, let us suppose that in our hypothetical

mammal the effective population size has been halved in each generation over the last ten generations. During this process, the population size has decreased from $N_e = 10^5$ to $N_e = 97.7$, roughly a thousand fold reduction. Yet, the heterozygosity is expected to have decreased from $\bar{H}_e = 4.90\%$ to $\bar{H}_e = 4.85\%$, a very insignificant decrease indeed.

8.2 Some models of allelic mutations

It had long been customary in population genetics to assume a model of mutations in which there are two possible alleles at a locus and mutations occur reversibly between the two, as may be expressed in the form;

$$A \rightleftarrows a.$$

Such a model, although simple and still useful for some cases, cannot in general be used to estimate genic variability at the molecular level, because the total number of allelic forms that are possible at any gene locus is enormous. For example, at a locus comprising 1000 nucleotide sites, the number of alleles that can be produced by base replacements alone (excluding structural changes) is 4^{1000} or roughly 10^{602}. For any one of these, there are 3×1000 or 3000 alleles that can be reached by a single-step base replacement, so the chance of returning to the original allele from any of them by a further single base change is only 1 in 3000 (assuming equal probability for all mutational changes). The total number of possible alleles is almost infinite as compared with the total number of alleles contained in a population at any time. These considerations were used by Kimura and Crow (1964) to justify the model which was later called the model of infinite alleles (Kimura, 1971). In this model, it is assumed that the possible number of isoallelic states at a locus is so large that each new mutant represents an allelic state not pre-existing in the population. Schematically, the model may be expressed as follows.

$$A_1 \rightarrow A_2 \rightarrow A_3 \rightarrow A_4 \rightarrow A_5 \rightarrow \ldots$$

If the alleles produced by mutation are selectively neutral (i.e. selectively equivalent), the level of heterozygosity that may be attained at statistical equilibrium where mutational production of new alleles is counterbalanced by random extinction of alleles can be calculated as follows.

Let v be the mutation rate per locus per generation, and let f_t be the probability that two homologous genes in an individual are identical in allelic state in the t-th generation ($t = 0, 1, 2, \ldots$). In other words, f_t is the probability of an individual being homozygous at a given locus in the t-th generation. We assume a random mating, monoecious population of

effective size N_e in which self-fertilization occurs with probability $1/N_e$. Then we have

$$f_{t+1} = (1-v)^2\left[\frac{1}{N_e}\left(\frac{1+f_t}{2}\right) + \left(1-\frac{1}{N_e}\right)f_t\right]. \tag{8.11}$$

This may be obtained by noting that two homologous genes within an individual in the next (i.e. $t+1$-th) generation are either derived from the same individual or from different individuals in the present (i.e. t-th) generation with probabilities $1/N_e$ and $(1-1/N_e)$ respectively: if they are derived from the same individual, they are either derived from the same gene or from two homologous genes each with probability 1/2. In this case, if they are derived from the same gene, we have $f = 1$, but if they are derived from homologous genes, f remains unchanged. Thus, we get the first term in the braces on the right-hand side in the above equation. On the other hand, if they are derived from two different individuals, f also remains unchanged. Note that under continued random mating among monoecious individuals, the probability of identity in allelic state is the same whether we take two homologous genes from the same individual or from two different individuals. Thus we obtain the second term in the braces. Finally, the term $(1-v)^2$ comes from the fact that only when neither of the two genes in the uniting gametes has mutated, do they contribute to the homozygosity in the next generation. Although we assumed a population of monoecious organisms in this derivation, the situation is essentially similar for a population with separate sexes unless N_e is extremely small (see Crow and Kimura, 1970, p. 102).

If we assume that v is very small but N_e is very large ($v \ll 1 \ll N_e$), equation (8.11) may be replaced by the differential equation

$$\frac{df_t}{dt} = \frac{1}{2N_e} - \left(2v + \frac{1}{2N_e}\right)f_t. \tag{8.12}$$

From this we obtain

$$f_t = f_\infty + (f_\infty - f_0)\exp\left[-\left(2v + \frac{1}{2N_e}\right)t\right], \tag{8.13}$$

where

$$f_\infty = \frac{1}{4N_e v + 1} \tag{8.14}$$

is the expected homozygosity at equilibrium.

In actual populations, f can be estimated by the sum of squares of allelic frequencies. The reciprocal of this has been called the 'effective number of alleles' by Kimura and Crow (1964). The average heterozygosity per locus

at equilibrium is given by

$$\bar{H}_e = 1 - f_\infty = \frac{4N_e v}{4N_e v + 1} . \tag{8.15}$$

Equation (8.13) shows that if the population is not in an equilibrium state, it approaches this state at the rate $(1 + 4N_e v)/(2N_e)$ per generation. This means that if the effective population size is very large, an enormous length of time is required (of the order of $4N_e$ generations unless $4N_e v$ is large) for the population to reach a new equilibrium, for example, after the population number has suddenly increased many-fold.

The above formula for $\bar{H}_e$ refers to the average heterozygosity of the equilibrium population. If we consider an infinite collection of equivalent populations undergoing independent evolutionary change by mutation and random drift (with the same values of v and N_e), the amount of heterozygosity in each population is a random variable. In other words, one minus the sum of the squares of the allelic frequencies,

$$H_e = 1 - \sum_i x_i^2$$

(where x_i is the frequency of i-th allele), may differ randomly from population to population, and $\bar{H}_e$ in equation (8.15) represents the average value of H_e over all populations at equilibrium. Then, the observed heterozygosity of an arbitrarily chosen population is likely to deviate from the mean value.

The problem of obtaining the variance of the heterozygosity (σ_H^2) is much more difficult than obtaining the mean, but this has been solved by Stewart (1976). His formula is

$$\sigma_H^2 = \frac{2\theta}{(1+\theta)^2(2+\theta)(3+\theta)}, \tag{8.16}$$

where $\theta = 4N_e v$. Actually, Stewart (1976) derived the relevant formula for a model assuming an arbitrary number, K, of allelic states, and the formula above corresponds to the special case $K = \infty$. I shall discuss some properties of the K-allele model later.

It is also important to investigate the numbers and relative frequencies of alleles in the population. Under the infinite allele model, turnover of alleles occurs continuously in the population, old alleles being replaced by new ones as time goes on, and in this sense no steady state exists: alleles are perpetually in flux. However, if we fix our attention on those alleles actually contained in the population at a given moment, and inquire as to the relative abundance of alleles having different number of representatives, evidently there is a definite pattern of relative abundance. This allows us to

define a distribution of allelic frequencies at steady state which we denote by $\Phi(x)$. This distribution has the meaning such that $\Phi(x)\,dx$ represents the expected number of alleles whose frequency is in the range x to $x + dx$. For neutral mutations, it has been shown by Kimura and Crow (1964) that

$$\Phi(x) = 4N_e v(1-x)^{4N_e v - 1} x^{-1}. \tag{8.17}$$

Note that this type of frequency distribution differs from the probability distribution of a given allele at a locus among equivalent populations, such as equation (8.4) in the previous section. The distribution $\Phi(x)$ was called 'frequency spectrum' by Ewens (1972). In the following, I shall mainly use the letter M to stand for $4N_e v$, although letter θ may also be used from time to time (in either case, I shall endeavor to keep the meaning of various letters straight). In terms of this distribution $\Phi(x)$, the statement that the frequencies of various alleles at a locus in a population adds up to unity may be expressed as

$$\int_0^1 x\Phi(x)dx = 1.$$

This can be confirmed by substituting $M(1-x)^{M-1}x^{-1}$ for $\Phi(x)$ and integrating the left-hand side. Similarly, the average value of the sum of the squares of the allelic frequencies or the average homozygosity is

$$\bar{f} = \bar{H}_0 = \int_0^1 x^2\Phi(x)dx = \frac{1}{M+1}, \tag{8.18}$$

where $M = 4N_e v$. This agrees with equation (8.14).

The expected number of alleles whose frequencies are in the interval (x_1, x_2) may be obtained from

$$\bar{n}_a(x_1, x_2) = \int_{x_1}^{x_2} \Phi(x)dx. \tag{8.19}$$

If allelic frequencies are scored at L equivalent loci, this value has to be multiplied by L. One important property of this distribution is that although it contains two parameters N_e and v, they enter the formula only as a product $N_e v$ so that the distribution $\Phi(x)$ is determined completely by a single parameter $M = 4N_e v$. Furthermore, as shown in equation (8.15) the average heterozygosity ($\bar{H}_e$) is equal to $M/(M+1)$ so that if we know this we can estimate M by the relation

$$M = \bar{H}_e/(1 - \bar{H}_e). \tag{8.20}$$

Then we can compute the expected numbers of alleles in various frequency classes using equation (8.19) and compare them with the corresponding observed numbers. As an example, I shall use the data from Ward and

M. According to these authors, the distribution of the evolutionary rate of amino acid substitutions and that of the molecular weight suggest that an appropriate value of α is about 1 (see also Fuerst *et al.*, 1977). The relationship between the average heterozygosity ($\bar{H}_e$) and $\bar{M}$ is more complicated in this model, but Nei (1975) derived a useful formula

$$\bar{H}_e = \frac{\bar{M}}{1+\bar{M}} - \frac{V_M}{(1+\bar{M})^3} \tag{8.23}$$

which gives a good approximation to $\bar{H}_e$ except when α is small and $\bar{M}$ is large. This shows that variation of the mutation rate among loci tends to decrease the average heterozygosity. In the case of the plaice, $\bar{H}_e = 0.102$, and we get $\bar{M} = 0.128$. The dotted columns in Fig. 8.3 represent the expected numbers of alleles obtained by using equation (8.22) assuming $\alpha = 1$ and $\bar{M} = 0.128$. It is interesting to note that the expected values thus obtained do not differ very much from those derived by equation (8.17) (compare the dotted columns with the dashed columns). Therefore, the excess of rare alleles stands out even if the variation of mutation rate among loci is taken into account.

Let us examine in more detail the occurrence of rare alleles whose frequencies are less than q, where we shall usually take $q = 0.01$ when we apply the theory to actual observations. It can be shown mathematically (see section 8.4 below) that, in the neighborhood of $x = 0$, the behavior of mutants in general, including those having a mild selective advantage or disadvantage, is essentially the same as that of selectively neutral mutants. Thus the formula

$$\bar{n}_a(x<q) = \int_{1/(2\bar{n})}^{q} \Phi(x)\mathrm{d}x \approx M\log_e(2\bar{n}q) \tag{8.24}$$

is valid, as pointed out by Nei (1977), if $4N_e|s|q$ is small, where $|s|$ is the absolute value of the selection coefficient of a mutant allele. In this formula M stands for $4N_e v$; however v here represents the mutation rate for all the alleles for which $4N_e|s|q \ll 1$ rather than just the selectively neutral alleles. In fact, Nei (1977) proposed the use of equation (8.24) for estimating the mutation rate for protein loci. In the following discussion we shall denote M in this equation by M_q in order to distinguish it from M obtained by equation (8.20). If rare variants are scored at more than one locus with a large sample for each locus (the mean size being $\bar{n}$ per locus), and if N_e is known, we can estimate the total mutation rate per locus by

$$v_T = M_q/(4N_e), \tag{8.25}$$

where

$$M_q = \bar{n}_a(x<q)/\log_e(2\bar{n}q), \tag{8.26}$$

in which an appropriate value for q is 0.01.

Since the sample size has to be large in order to estimate M_q from observations on rare variants, I have chosen from the data by Ward and Beardmore (1977; see their Tables 2 and 3), 11 loci for which the sample size per locus is larger than 1000. The average sample size per locus turned out to be $\bar{n} = 1956.1$. Of these 11 loci, 8 are polymorphic and 3 are monomorphic. The average heterozygosity of these 11 loci is 0.147 which is not very different from the average heterozygosity of 46 loci, i.e. 0.102. Among these loci, 30 alleles are found whose frequencies in the sample are less than 0.01. Therefore, $\bar{n}_a(x<0.01) = 30/11 = 2.73$ per locus. Then, applying equation (8.26), where we assume $q = 0.01$, we get $M_q = 2.73/\log_e(2 \times 19.56) = 0.74$. If we denote the total mutation rate per locus by v_T, then $v_T = 0.74/(4N_e)$. The real value of N_e is not known but if it is 10^6, we have $v_T = 1.6 \times 10^{-7}$.

One important point here is that, even when the actual value of N_e is not known, we can estimate the proportion (which we shall denote by P_{neut}) of mutations that are selectively neutral (at the time of their occurrence) by the following method. If the neutral theory is correct and if the mutant alleles responsible for the protein polymorphism are selectively neutral, then the value of M, i.e. 0.114 which we have obtained by substituting the observed average heterozygosity $\bar{H}_e = 0.102$ into equation (8.20), should represent $4N_e v_0$, where v_0 is the mutation rate for selective neutral alleles that are electrophoretically detectable. Therefore, we have

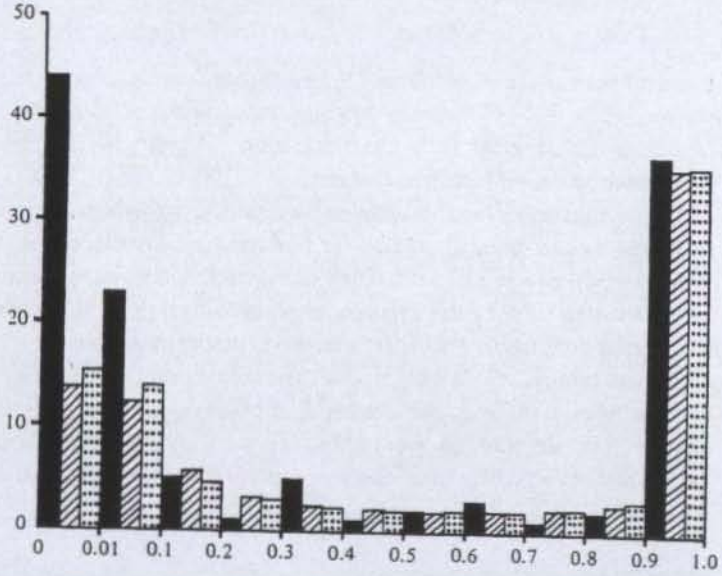
$$P_{neut} = v_0/v_T = M/M_q \tag{8.27}$$

Substituting the values of M and M_q, as estimated above, into this equation, we get $P_{neut} \approx 0.15$. This means that one mutation out of 6.5 on the average is selectively neutral while the remaining 5.5 are too deleterious to contribute to protein polymorphism.

The occurrence of rare variants is also reported in the Japanese macaque (*Macaca fuscata fuscata*) studied by Nozawa and his associates (see for example, Nozawa *et al.*, 1982). They surveyed 32 independent protein loci and obtained 1.3% as the average heterozygosity, which is a rather low value even for mammals. Their extensive studies made so far yield the following data (K. Nozawa, personal communication, 1981). The average number of rare variants per locus is $\bar{n}_a(x<0.01) = 23/32 = 0.719$, and the average sample size is $\bar{n} = 1609.9$. Thus we get $M_q = \bar{n}_a(x<0.01)/\log_e(2\bar{n} \times 0.01) = 0.207$. The observed average heterozygosity per locus is

Beardmore (1977) on protein variation in the plaice, *Pleuronectes platessa*, a marine flatfish. They screened electrophoretically detectable variations at 46 protein loci (39 enzymes and 7 non-enzyme proteins) taking very large samples from the Bristol Channel population. This probably represents the most comprehensive investigation of protein variation in fishes. The mean heterozygosity per individual per locus turned out to be 0.102 ± 0.026. The numbers in the sample differ from locus to locus: there are more than 2000 in 8 loci, between 2000 and 500 in 9 loci, between 500 and 100 in 16 loci, but less than 100 in the remaining 13 loci. Using equation (8.20), we get $M = 0.114$ for $\bar{H}_e = 0.102$. Then, the expected number of alleles in various frequency classes, $\bar{n}_a(x_1, x_2)$, can be computed by integrating $L\Phi(x)$ or $46M(1-x)^{M-1}x^{-1}$ from $x = x_1$ to $x = x_2$. In Fig. 8.3, the expected numbers thus computed are shown by hatched columns. The corresponding observed numbers are represented by solid columns. (As to the dotted columns, I shall explain them later.)

Fig. 8.3. Histogram showing observed and expected distributions of allelic frequencies. The solid columns represent observed numbers of alleles at 46 protein loci in the plaice, *Pleuronectes platessa* (data from Ward and Beardmore, 1977). The hatched columns represent the expected numbers of alleles based on the infinite allele model of Kimura and Crow (1964) while the dotted columns represent the expected numbers computed by using Nei *et al.*'s (1976a) model incorporating possible variation of the mutation rates among loci. Abscissa: allele frequency. Ordinate: number of alleles in various frequency classes.



In comparing the observed distribution of allelic frequencies with the theoretical one, we must keep in mind that the distribution $\Phi(x)$ refers only to those alleles actually contained in the population; although there are an infinite number of possible alleles, only a limited number of them are present at any moment in a population, and we do not include the countless number of missing alleles. This means that each monomorphic locus makes only one contribution to an allelic class containing $x = 1.0$. Therefore, I included monomorphic loci in the right-most solid column which represents the number of alleles in the frequency range [0.9, 1.0]. On the other hand the left-most class which represents the number of alleles in the frequency range (0, 0.01) contains many rare variants, and they tend to come from loci for which large samples were scored. The expected number $\bar{n}_a(x < 0.01)$ in this class was computed by taking, as the lower limit of integration, the reciprocal of twice the average number of the sample size per locus, namely, 46 times the value of

$$\bar{n}_a(x < 0.01) = \int_{1/(2\bar{n})}^{0.01} \Phi(x)\,dx = M \int_{1/(2\bar{n})}^{0.01} (1-x)^{M-1}x^{-1}dx, \quad (8.21)$$

where $\bar{n}$ is the average sample size, which is 678.9 in the plaice studied by Ward and Beardmore (1977).

It is clear from the figure that the expected number of alleles in this class is much smaller than that observed. A similar tendency is noticeable for the next frequency class (0.01, 0.1). Except for these two classes, the agreement between observed and expected numbers is satisfactory. Similar observations showing an excess of rare alleles were presented by Ohta (1975, 1976b) for *Drosophila* and human populations. In the above analysis of the plaice data, we have treated the 46 loci as if they are equivalent in having the same $M(= 4N_e v)$ values. However, it is much more likely, and indeed realistic to assume that mutation rates for neutral alleles differ among loci due to differences of the gene size as well as differences in the functional constraints of proteins for which they code. To take such variation into account, Nei *et al.* (1976a) proposed an infinite allele model in which is incorporated the assumption that the mutation rate (more precisely the mutation rate for neutral alleles) among loci follows a gamma distribution. They showed that if M or $4N_e v$ follows a gamma distribution with mean $\bar{M}$ and variance V_M, then the formula for the distribution of allelic frequencies corresponding to equation (8.17) becomes

$$\Phi(x) = \frac{\bar{M}x^{-1}(1-x)^{-1}}{[1-(\bar{M}/\alpha)\log_e(1-x)]^{\alpha+1}}, \quad (8.22)$$

where $\alpha = \bar{M}^2/V_M$ is the reciprocal of the squared coefficient of variation of

M. According to these authors, the distribution of the evolutionary rate of amino acid substitutions and that of the molecular weight suggest that an appropriate value of α is about 1 (see also Fuerst *et al.*, 1977). The relationship between the average heterozygosity ($\bar{H}_e$) and $\bar{M}$ is more complicated in this model, but Nei (1975) derived a useful formula

$$\bar{H}_e = \frac{\bar{M}}{1+\bar{M}} - \frac{V_M}{(1+\bar{M})^3} \tag{8.23}$$

which gives a good approximation to $\bar{H}_e$ except when α is small and $\bar{M}$ is large. This shows that variation of the mutation rate among loci tends to decrease the average heterozygosity. In the case of the plaice, $\bar{H}_e = 0.102$, and we get $\bar{M} = 0.128$. The dotted columns in Fig. 8.3 represent the expected numbers of alleles obtained by using equation (8.22) assuming $\alpha = 1$ and $\bar{M} = 0.128$. It is interesting to note that the expected values thus obtained do not differ very much from those derived by equation (8.17) (compare the dotted columns with the dashed columns). Therefore, the excess of rare alleles stands out even if the variation of mutation rate among loci is taken into account.

Let us examine in more detail the occurrence of rare alleles whose frequencies are less than q, where we shall usually take $q = 0.01$ when we apply the theory to actual observations. It can be shown mathematically (see section 8.4 below) that, in the neighborhood of $x = 0$, the behavior of mutants in general, including those having a mild selective advantage or disadvantage, is essentially the same as that of selectively neutral mutants. Thus the formula

$$\bar{n}_a(x<q) = \int_{1/(2\bar{n})}^{q} \Phi(x)\mathrm{d}x \approx M\log_e(2\bar{n}q) \tag{8.24}$$

is valid, as pointed out by Nei (1977), if $4N_e|s|q$ is small, where $|s|$ is the absolute value of the selection coefficient of a mutant allele. In this formula M stands for $4N_e v$; however v here represents the mutation rate for all the alleles for which $4N_e|s|q \ll 1$ rather than just the selectively neutral alleles. In fact, Nei (1977) proposed the use of equation (8.24) for estimating the mutation rate for protein loci. In the following discussion we shall denote M in this equation by M_q in order to distinguish it from M obtained by equation (8.20). If rare variants are scored at more than one locus with a large sample for each locus (the mean size being $\bar{n}$ per locus), and if N_e is known, we can estimate the total mutation rate per locus by

$$v_T = M_q/(4N_e), \tag{8.25}$$

where

$$M_q = \bar{n}_a(x<q)/\log_e(2\bar{n}q), \tag{8.26}$$

in which an appropriate value for q is 0.01.

Since the sample size has to be large in order to estimate M_q from observations on rare variants, I have chosen from the data by Ward and Beardmore (1977; see their Tables 2 and 3), 11 loci for which the sample size per locus is larger than 1000. The average sample size per locus turned out to be $\bar{n} = 1956.1$. Of these 11 loci, 8 are polymorphic and 3 are monomorphic. The average heterozygosity of these 11 loci is 0.147 which is not very different from the average heterozygosity of 46 loci, i.e. 0.102. Among these loci, 30 alleles are found whose frequencies in the sample are less than 0.01. Therefore, $\bar{n}_a(x<0.01) = 30/11 = 2.73$ per locus. Then, applying equation (8.26), where we assume $q = 0.01$, we get $M_q = 2.73/\log_e(2 \times 19.56) = 0.74$. If we denote the total mutation rate per locus by v_T, then $v_T = 0.74/(4N_e)$. The real value of N_e is not known but if it is 10^6, we have $v_T = 1.6 \times 10^{-7}$.

One important point here is that, even when the actual value of N_e is not known, we can estimate the proportion (which we shall denote by P_{neut}) of mutations that are selectively neutral (at the time of their occurrence) by the following method. If the neutral theory is correct and if the mutant alleles responsible for the protein polymorphism are selectively neutral, then the value of M, i.e. 0.114 which we have obtained by substituting the observed average heterozygosity $\bar{H}_e = 0.102$ into equation (8.20), should represent $4N_e v_0$, where v_0 is the mutation rate for selective neutral alleles that are electrophoretically detectable. Therefore, we have

$$P_{neut} = v_0/v_T = M/M_q \tag{8.27}$$

Substituting the values of M and M_q, as estimated above, into this equation, we get $P_{neut} \approx 0.15$. This means that one mutation out of 6.5 on the average is selectively neutral while the remaining 5.5 are too deleterious to contribute to protein polymorphism.

The occurrence of rare variants is also reported in the Japanese macaque (*Macaca fuscata fuscata*) studied by Nozawa and his associates (see for example, Nozawa *et al.*, 1982). They surveyed 32 independent protein loci and obtained 1.3% as the average heterozygosity, which is a rather low value even for mammals. Their extensive studies made so far yield the following data (K. Nozawa, personal communication, 1981). The average number of rare variants per locus is $\bar{n}_a(x<0.01) = 23/32 = 0.719$, and the average sample size is $\bar{n} = 1609.9$. Thus we get $M_q = \bar{n}_a(x<0.01)/\log_e(2\bar{n} \times 0.01) = 0.207$. The observed average heterozygosity per locus is

$\bar{H}_e = 0.013 \pm 0.0014$, from which we get $M = \bar{H}_e/(1-\bar{H}_e) = 0.0132$. Using these values, we obtain $P_{neut} = M/M_q \approx 0.064$. This means that, roughly speaking, one mutation out of 16 is selectively neutral in Japanese monkeys. This is less than half as large as the corresponding value obtained for the plaice.

The census number of the total population of the Japanese macaque is estimated to be $20\,000 \sim 70\,000$. It is also estimated that the effective population size is about one-third of its census number (cited from Nozawa *et al.*, 1975). Following Nei (1977), if we assume $N_e = 2 \times 10^4$, we obtain $v_T = M_q/(4N_e) = 2.6 \times 10^{-6}$. On the other hand, the mutation rate for neutral alleles is $v_0 = M/(4N_e) = 1.65 \times 10^{-7}$ per generation.

The two estimates for P_{neut}, i.e. 0.15 for the plaice and 0.064 for the Japanese macaque, considering the possibly large statistical errors involved, suggest that 1/10 is a realistic value for the probability of amino acid-altering mutations being selectively neutral. Note that this is the value which we assumed for our hypothetical mammal discussed in the previous section.

Let us consider more generally the number of alleles contained in the population and also in the sample, still assuming the model of infinite neutral alleles. The expected value of the number of different alleles in the population is

$$\bar{n}_a = \int_{1/(2N)}^{1} \Phi(x)\mathrm{d}x = 4N_e v \int_{1/(2N)}^{1} (1-x)^{4N_e v - 1} x^{-1}\mathrm{d}x. \tag{8.28}$$

Note that the lower limit of integration is $1/(2N)$ rather than zero, and that $\bar{n}_a$ depends much on N even when $4N_e v$ is constant; $\bar{n}_a$ becomes larger as N increases. This is understandable because the larger the population the higher the chance of different rare variants existing somewhere in the population. In fact, $\bar{n}_a$ becomes indefinitely large as $N \to \infty$ while keeping $4N_e v =$ constant. This may be compared with the effective number of alleles which is denoted by n_e and which is defined as the reciprocal of the sum of squares of allelic frequencies. For neutral mutants the expected value of this is

$$\bar{n}_e = 1 \Big/ \int_0^1 x^2 \Phi(x)\mathrm{d}x = 4N_e v + 1. \tag{8.29}$$

In general, the effective number of alleles (n_e) is smaller than the actual number of different alleles (n_a) in the population. Only when various alleles occur in equal frequencies, do we have $n_e = n_a$. One characteristic of n_e is that it depends on $4N_e v$ but is largely independent of the sample size n unless n is very small. As an example showing the difference between these

two measures of the number of alleles, let us suppose that four alleles, A_1, A_2, A_3 and A_4, occur with respective frequencies 0.5, 0.3, 0.15 and 0.05. Then the sum of squares of allelic frequencies is $f = (0.5)^2 + (0.3)^2 + (0.15)^2 + (0.05)^2 = 0.365$ so that $n_e = 1/f = 2.74$, which is smaller than the actual number $n_a = 4$. Clearly, rare alleles contribute little to n_e.

Next, we shall consider the number of alleles in a sample. Suppose we draw a sample of size n from the population. Then what is the expected number of alleles in the sample? Consider a particular allele whose frequency in the population happens to be x. The probability that this allele is represented at least once in the sample is $1-(1-x)^{2n}$. The expected number of alleles whose frequencies in the population are in the range $x \sim x + \mathrm{d}x$ is $\Phi(x)\mathrm{d}x$. Therefore, the expected number of different alleles in the sample of size n is

$$\bar{n}_a = \int_0^1 [1-(1-x)^{2n}]\Phi(x)\mathrm{d}x. \tag{8.30}$$

It was shown by Ewens (1972) that if all the alleles are selectively neutral so that $\Phi(x)$ is given by (8.17), the relationship between the expected number of alleles in the sample and the population parameter $\theta = 4N_e v$ is

$$\bar{n}_a = \frac{\theta}{\theta} + \frac{\theta}{\theta+1} + \ldots + \frac{\theta}{\theta+2n-1} \tag{8.31}$$

He showed that if the population parameter θ is estimated by this formula by substituting for $\bar{n}_a$ the actual number of different alleles in the sample (n_a'), it has some desirable statistical properties. First, the estimator, say $\hat{\theta}$, thus obtained is indeed the maximum likelihood estimator of θ, namely,

$$n_a' = \frac{\hat{\theta}}{\hat{\theta}} + \frac{\hat{\theta}}{\hat{\theta}+1} + \ldots + \frac{\hat{\theta}}{\hat{\theta}+2n-1}. \tag{8.31a}$$

Secondly, this estimator is sufficient, that is, all information regarding θ is contained in n_a'; namely, if n_a' is given, the joint probability distribution of the number of representatives of various alleles in the sample is independent of θ. He claims that, as compared with this, the common method of estimating θ by equating $1/(1+\theta)$ to the expected homozygosity (see equation 8.18) is much less satisfactory. According to Ewens (1979), θ estimated by $1/f - 1$, where f is the sum of squares of allelic frequencies in the sample, is strongly biassed and is accompanied by a much larger mean square error than $\hat{\theta}$, and therefore undesirable. Ewens (1972) also proposed approximate statistical tests to check whether the observed frequencies of polymorphic alleles at individual loci conform to the neutral theory. Although Ewens' sampling theory is statistically elegant, in my opinion, its

applicability to actual situations is quite limited. As I have shown already, there are probably some ten times as many slightly deleterious alleles as selectively neutral alleles at the time of occurrence, so that, if the sample size is large (which is essential for any single locus test such as Ewens' to be meaningful), slightly deleterious alleles will inflate the n_a' value and this will greatly reduce the validity of Ewens' tests. This point has already been made by Nei (1975) and Ohta (1977b), and more recently by Li (1979). We shall discuss more extensively in chapter 9 various statistical tests which have been proposed to check the neutral theory through the use of observations on protein polymorphisms.

In the above treatments, we have used a model assuming an infinite number of allelic states. However, it is sometimes more convenient to use a model in which the number of possible allelic states is K rather than infinity. Particularly, the symmetric K allele model in which mutations occur in equal frequencies in all directions among K alleles, $A_1, A_2, \ldots, A_K$, is useful (Kimura, 1968b).

Let v be the mutation rate per gene per generation. We shall denote by v_1 the mutation rate by which each allele mutates to one of the remaining $(K-1)$ alleles so that $v_1 = v/(K-1)$. Then, it can be shown (Kimura, 1968b) that under random mating, the expected homozygosity at equilibrium (still assuming selectively neutral mutations) is

$$\bar{H}_0 = E\left(\sum_{i=0}^{K} x_i^2\right) = \frac{1+(2N_e-1)(2v_1-v_1^2K)}{2N_e-(2N_e-1)(1-v_1K)^2}, \tag{8.32}$$

where x_i is the frequency of the i-th allele A_i. In this formula, if we let $K \to \infty$ while holding $(K-1)v_1 = v = \text{constant}$, then $v_1 \to 0$ and $v_1^2 K \to 0$, and therefore we get

$$\bar{H}_0 = \frac{1}{2N_e-(2N_e-1)(1-v)^2}. \tag{8.33}$$

Since the mutation rate v is much smaller than unity, the denominator of this formula reduces to $1+4N_ev$ with good approximation, thus leading to a formula equivalent to (8.14). Also, from equation (8.32), if we assume $N_e \gg 1 \gg v$, we get a useful formula

$$\bar{H}_0 = \frac{1+4N_ev/(K-1)}{1+4N_evK/(K-1)}, \tag{8.33a}$$

which agrees with (8.14) when $K = \infty$. The average heterozygosity may be obtained by $\bar{H}_e = 1 - \bar{H}_0$.

The actual value of homozygosity (or sum of squares of allelic frequencies) in a particular population is likely to deviate from the average value $\bar{H}_0$ due to random frequency drift. The variance of such deviations at a steady state in which mutation and random sampling balance each other has been obtained by Stewart (1976). His formula is

$$\text{Var}\,(H) = \frac{2\theta(1+\theta/L)}{(1+\theta+\theta/L)^2(2+\theta+\theta/L)(3+\theta+\theta/L)}, \tag{8.34}$$

Where $L = K-1$ and $\theta = 4N_ev$. Since $H_e = 1 - H_0$ this formula applies both to homozygosity and heterozygosity. In this formula, if we let $L \to \infty$, we obtain equation (8.16) which gives the variance of homozygosity (or heterozygosity) for the infinite allele model. An interesting point made by Stewart in relation to the nonrecurrent case (i.e. $L \to \infty$) is that the ratio of standard deviation σ_H to the mean $\bar{H}_e$ is about 0.41 when $\theta = 1.0$, but increases to 1.75 as θ decreases to 0.1 and to 5.75 when $\theta = 0.01$. This means that if $4N_ev$ is small, the standard deviation of heterozygosity (H_e) among equivalent loci may be several times greater than its average value.

In our hypothetical mammal which we discussed in the previous section, the average heterozygosity for electrophoretically detectable alleles is 4.9%, for which the ratio $\sigma_H/\bar{H}_e$ turns out to be about 2.5. It is important to note that even if the intrinsic parameter $\theta = 4N_ev$ remains constant at a given locus, the actual level of heterozygosity will deviate a great deal from the expected value ($\bar{H}_e$) during the long course of evolution. Thus, the familiar argument by the selectionists that the level of heterozygosity is determined as an optimum strategy for the species to adapt to its environment will lose much of its force.

In addition to selectively neutral alleles, it is natural to inquire how much variability can be maintained in a finite population when mutations are heterotic, that is, overdominant. Kimura and Crow (1964) investigated an infinite allele model with overdominant mutations. In this model it is assumed that all the heterozygotes have equal fitnesses, and that all the homozygotes have equal fitnesses which are lower by s as compared with the heterozygotes. In other words, s is the selection coefficient against homozygotes. It is expected that many more alleles are maintained within a population if all the alleles produced by mutation are overdominant than if they are neutral. Kimura and Crow developed an approximate method to treat this problem. For example, in a population of $N_e = 10^5$, if the mutation rate is $v = 10^{-5}$ and the selection coefficient is $s = 10^{-3}$, then we get $n_e = 13.5$ as the effective number of alleles. The average heterozygosity is $\bar{H}_e = 0.926$. This may be compared with the selectively neutral case having the same mutation rate 10^{-5} and the same effective population number 10^5, for which $n_e = 4N_ev + 1 = 5.0$. This gives an average heterozygosity of

$\bar{H}_e = 0.80$. In this case, overdominance plays a relatively small role in increasing the average heterozygosity, since high heterozygosity is already attained by mutation alone.

If the mutation rate is lower, such as $v = 10^{-6}$, in a population of $N_e = 10^5$, overdominance is more effective. For example, if $s = 10^{-3}$, the effective number of alleles maintained is roughly 8 as compared with 1.4 for the corresponding case of $s = 0$.

Although it is unrealistic to assume overdominance for many loci in the genome, the possibility can not be excluded that in a small fraction of the loci, variant alleles may be maintained by overdominance. An interesting case has been reported by Kerr (1967). This is the sex-determining alleles found in bee populations in Brazil. In this case, alleles are haplo-viable but homozygous lethal so that they are overdominant with $s = 1$ in females while the males may be regarded as gametes. In fact, Kerr applied the theory of Kimura and Crow (1964) to analyse his data on the number of sex-determining alleles actually found in local populations of bees.

The infinite allele model with heterotic mutations has also been studied by Wright (1966). More recently Watterson (1977a) developed a statistical method for testing departures from selective neutrality in the direction of heterozygote advantage or disadvantage based on sample homozygosity (see also Watterson, 1978). Like Ewens' (1972) test, Watterson's test is based on a sample from a single locus. As compared with the tests which make use of a large number of loci simultaneously (Nei and Roychoudhury, 1974; Nei 1975; Ohta, 1975, 1976b) it contains some drawbacks, including low statistical power, as pointed out by Li (1979).

There are additional models of molecular population genetics (e.g. the stepwise mutation model and the infinite site model) which we need to take up, but I shall relegate them to later sections. In the next two sections, I shall present some important concepts in the stochastic theory of population genetics and their mathematical treatment.

8.3 The age of a neutral mutant in a finite population

To understand the pattern of genic variability within a population, the concept of the age of a mutant allele is very useful. So, I shall consider, in this section, the problem of estimating the age of a mutant. In other words, we ask the question; how many generations a mutant allele with a given frequency x has persisted in the population since it appeared by mutation? This problem was first treated mathematically by Kimura and Ohta (1973b) and solutions for some simple situations were given by them. This was followed by more general treatments by Maruyama (1974), and

Maruyama and Kimura (1975). Since then, the problem has attracted much attention by mathematical geneticists and a number of papers have been published (see, for review, Watterson, 1977b and Ewens, 1979). Here, I shall consider only some simple cases pertinent to this book.

If a neutral allele with current frequency x has increased from a very low frequency since its origin by mutation, it can be shown (Kimura and Ohta, 1973b) that the average age, in terms of generations, is given by

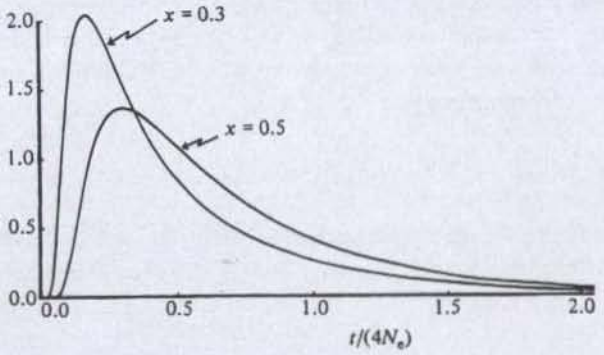
$$\bar{a}(x) = -4N_e\left(\frac{x}{1-x}\right)\log_e x. \tag{8.35}$$

In deriving this formula, it was assumed that the mutant allele has not itself mutated since its origin. Using the terminology of Maruyama (1974), this gives the 'age before fixation'. As an example, let us suppose that the frequency of an allele is 10%. Then we have $\bar{a}(0.1) \approx 1.02N_e$, namely, the average age in generations is roughly equal to the effective size of the population. Also, for $x = 0.3$ and 0.5, we have $\bar{a}(0.3) \approx 2.06N_e$ and $\bar{a}(0.5) \approx 2.77N_e$. These examples suggest that polymorphic alleles, if they are selectively neutral, tend to be very old, sometimes older than the age of the species. Under the same assumptions of no selection and no further mutation, the mean square age (before fixation) is given by

$$\overline{a^2(x)} = 32N_e^2\left(\frac{x}{1-x}\log_e x - \int_0^x \frac{\log_e z}{1-z}dz\right) \tag{8.36}$$

(Kimura and Ohta, 1973b). For $x = 0.1$, this is roughly equal to $2.8N_e^2$,

Fig. 8.4. The distribution of the age (before fixation) is illustrated for alleles whose current frequencies in the population are 30% and 50% respectively. Abscissa: age measured with $4N_e$ generations as the unit; ordinate: probability density. Note that because of the reversibility of the stochastic process involved, these give the distributions of time until extinction of an allele whose current frequencies are $x = 0.3$ and 0.5.



giving a variance of about $1.8N_e^2$. This means that although the age (before fixation) of a mutant allele whose current frequency happens to be 10% is about N_e on the average, it is accompanied by a large uncertainty with standard deviation of roughly $1.4N_e$ generations which is larger than the mean.

To show the highly skewed nature of the distribution of the age the distributions for $x = 0.3$ and 0.5 are given in Fig. 8.4. These curves were drawn based on the following rationale. The probability density of the age (denoted by t) may be obtained by

$$\lim_{p \to 0}\left[\phi(p,x;t)\Big/\int_0^\infty \phi(p,x;t)\mathrm{d}t\right],$$

where $\phi(p,x;t)$ is given by equation (3.2) of chapter 3. It turns out that this distribution, when the age (t) is measured in units of $4N_e$ generations, is

$$f_x(\lambda) = 2x\sum_{i=1}^{\infty}(2i+1)T_{i-1}^1(z)e^{-i(i+1)\lambda}, \tag{8.37}$$

where $\lambda = t/(4N_e), z = 1-2x$ and $T_{i-1}^1(\cdot)$ is the Gegenbauer polynomial defined by

$$T_{i-1}^1(z) = \frac{i(i+1)}{2}F\left(i+2, 1-i, 2, \frac{1-z}{2}\right), \tag{8.38}$$

in which $F(\cdot,\cdot,\cdot,\cdot)$ stands for the hypergeometric function. Here, the computation is facilitated by noting that for the first two terms, $T_0^1(z) = 0$, $T_1^1(z) = 3z$, and that for the remaining terms, the following recurrence formula is available.

$$iT_i^1(z) = z(2i+1)T_{i-1}^1(z) - (i+1)T_{i-2}^1(z), \tag{8.39}$$

where $i = 2$, 3, etc.

Comparison of equation (8.35) with an equation giving the average time until extinction reveals the interesting fact that the average age before fixation of an allele is equal to the average time until extinction of that allele. In fact, it was shown by Kimura and Ohta (1969a) that the number of generations until loss (excluding the cases of fixation) for a neutral allele having the frequency p is

$$\bar{t}_0(p) = -4N_e\left(\frac{p}{1-p}\right)\log_e p. \tag{8.40}$$

This is not difficult to understand because the random process by which allelic frequency goes from p to x and the reverse process by which it goes from x to p are symmetric if there is no selection or mutation during the processes.

In applying the above formula (8.35) to actual situations, one troublesome problem which we encounter is that we cannot know whether the allele in question (with frequency x) has increased from a very low frequency or has decreased from a previously fixed state (i.e. the other, alternative allele increased from a low frequency to $1-x$). For the latter case, the appropriate age is $\bar{a}(1-x)$ rather than $\bar{a}(x)$. The most reasonable solution to this problem, it seems to me, is to average out these two alternative cases, weighting by the respective probabilities. It turns out that the relative probabilities of these two alternative events (assuming no mutations in the process) are $1-x$ and x respectively. This follows from the consideration that the probability is p/x that a neutral mutant allele with initial frequency p subsequently reaches a higher frequency x before it is either lost from the population or fixed in it. Similarly, the probability is $p/(1-x)$ that the frequency of an allele with initial frequency p increases to $1-x$. Thus, combining these two cases, the probable age which we denote by $\bar{A}(x)$ may be given by $(1-x)\bar{a}(x) + x\bar{a}(1-x)$. This leads to

$$\bar{A}(x) = -4N_e[x\log_e x + (1-x)\log_e(1-x)]. \tag{8.41}$$

For example, if $x = 0.1$, we get $\bar{A}(0.1) \approx 1.3N_e$. In this case, it is only slightly larger than $\bar{a}(0.1) \approx N_e$. It is interesting to note that the probable age given by (8.41) is equal to the average time until either fixation or loss of a neutral allele whose current frequency is x.

So far, we have neglected the possibility that further mutations occur during the process of gene frequency change. If such a possibility is included, the problem becomes much more complicated. In this case we must take into account not only the possibility that further mutations occur before the mutant allele reaches a higher frequency but also the possibility that a mutant allele that has once become fixed in the population eventually has its frequency decreased by further mutation in conjunction with random drift. In such a case it may be appropriate to consider the 'age including fixation' (Maruyama, 1974). Let v be the mutation rate per generation and assume that whenever a mutation occurs it represents a new, not pre-existing allele. Assuming a selectively neutral allele, we can show that if $4N_e v < 1$, the probability of the allele reaching fixation is given by

$$u(p) = 1 - (1-p)^{1-M}, \tag{8.42}$$

where $M = 4N_e v$ and p is the initial frequency of the allele. Note that in this case the fixation is temporary, because the allelic frequency eventually decreases by mutation changing the allele into a new allele. On the other hand, we can show that if $4N_e v \geqq 1$, the probability of the allele reaching the

state of fixation (even temporarily) is zero (assuming that N_e is large so that the diffusion model is valid). In either case, an appropriate measure is the age including fixation. For a selectively neutral allele with current frequency x, this is given by

$$\bar{a}_{\mathrm{inc}}(x) = \frac{4N_e}{1-M}\left\{\int_0^x \frac{1-(1-\xi)^{1-M}}{\xi(1-\xi)^{1-M}}\mathrm{d}\xi + [1-(1-x)^{1-M}]\int_x^1 \frac{\mathrm{d}\xi}{\xi(1-\xi)^{1-M}}\right\}, \tag{8.43}$$

where $M = 4N_e v$. In this formula, it is assumed that the initial frequency p is very low ($p \to 0$). A general formula for $\bar{a}_{\mathrm{inc}}(x)$ is given in Maruyama (1974) who also tabulated its values for various combinations of $4N_e v$ and x. The age including fixation is usually much longer than the age before fixation. For example, when $4N_e v = 0.1$ the mean age including fixation of an allele whose current frequency is 10% is about $5.4N_e$ generations, whereas the age before fixation of the same allele is about N_e generations. The large difference between these two ages comes from the fact that, if we consider mutation of this magnitude, there is a good chance that an allele whose current frequency is 0.1 happens to be the remnant of a once fixed allele, and for such an allele, the length of time required for this allele to first reach fixation plus the time spent during temporary fixation and subsequent decrease must all be taken into account. This augments the average age substantially. Note that the average age of a once-fixed allele is at least $4N_e$ generations.

It will be of interest here to inquire how these ages compare with the average length of time until a mutant allele reaches a certain frequency, say x, for the first time starting from a lower frequency p. A general formula for the average first arrival time was given by Kimura and Ohta (1973b). For a neutral mutant, disregarding further mutation, the average first arrival time (in generations) is given by

$$\bar{t}_x(0) = 4N_e\left[\frac{1-x}{x}\log_e(1-x) + 1\right]. \tag{8.44}$$

In deriving this formula, it was assumed that the initial frequency is very low ($p \to 0$). As an example, let us assume that $x = 0.1$. Then we have $\bar{t}_x(0) \approx 0.2N_e$ so that the average first arrival time is only 1/5 as large as the average age before fixation. When x is much smaller than unity, (8.44) may be approximated by $\bar{t}_x(0) = 2N_e x$.

Let us now consider, following Kimura and Ohta (1973b), the bearing of the above results on the spatial pattern of genetic variation. Here we are

particularly concerned with the question: how much migration is required to keep the frequencies of a 'rare' polymorphic allele essentially uniform among localities when the allele is selectively neutral? First, consider a one-dimensional habitat forming a circle of radius r. Let N_T be the total number of breeding individuals in one generation and assume that they are distributed uniformly with density δ so that $N_T = 2\pi r\delta$. If we denote by σ^2 the mean square distance of individual migration (assumed to be isotropic) in one generation, then the distance of migration during t generations should follow the normal distribution with mean zero and variance $\sigma_t^2 = t\sigma^2$ when t is large. If the abscissa of this distribution is wrapped around a circle of radius $2\pi r$, and if the resulting (superimposed) probability distribution on the circle is approximately uniform, then the frequencies among localities of a mutant allele having age t will become essentially uniform. On the other hand, if the time since the first appearance of the mutant allele is short and if the superimposed probability distribution on the circle still markedly deviates from the uniform distribution, clear local differentiation of allelic frequencies should result.

If we substitute $\bar{a}(x)$ for t in this reasoning and if we note that the superimposed distribution is essentially uniform when $2\pi r \leqq 2\sigma_t$ or

$$\pi r \leqq \sigma_t, \tag{8.45}$$

then we obtain

$$N_T \leqq 2\delta\sigma\sqrt{\bar{a}(x)} \tag{8.46}$$

as a condition for the uniform distribution of frequencies of neutral alleles among localities.

Next, we consider a two-dimensional habitat extending over a sphere of radius r. Let us assume that the individuals are distributed uniformly with density δ. Let σ^2 be the mean square distance of individual migration in one generation, and assume that the migration is isotropic. Wrapping the sphere in a two-dimensional normal distribution for the distance of migration in t generations and considering the resulting probability distribution on the sphere similar to the one considered above, the condition (8.45) leads to

$$\pi N_T \leqq 4\delta\sigma^2 t \tag{8.47}$$

for the two-dimensional case, since $N_T = 4\pi r^2\delta$ in this case.

For a mutant allele whose average frequency in the whole population happens to be 0.1, the condition for uniform distribution reduces roughly to

$$\sigma^2\delta \geqq 1 \tag{8.48}$$

if we put $t = \bar{a}(0.1) \approx N_T$ in (8.47).

If the total population is subdivided into colonies (subpopulations) in each of which the mating is at random, and migration in one generation is restricted to neighboring colonies (two-dimensional stepping stone model of Kimura and Weiss, 1964), condition (8.48) becomes

$$mN \geqq 1, \tag{8.49}$$

where m is the rate at which each colony exchanges individuals with four surrounding colonies each generation and N is the effective size of each colony. However, since the age $\bar{a}(x)$ at $x = 0.1$ has a relatively large standard deviation, it may be safer to use the first arrival time $\bar{t}_x(0)$ at $x = 0.1$ for t in (8.47) to derive the required condition. This leads approximately to

$$mN \geqq 5. \tag{8.50}$$

These results agree essentially with those obtained by Kimura and Maruyama (1971) based on an entirely different method. For a neutral mutant having the frequency $x = 0.05$, the corresponding formula becomes approximately

$$mN \geqq 10. \tag{8.51}$$

This means that uniform distribution of frequencies among colonies can be attained if exchange of at least 2.5 individuals occurs on the average between two neighboring colonies per generation irrespective of the size of each colony. Thus we conclude that when the average frequency of a 'rare' polymorphic allele is a few per cent in the whole population, its frequency among different localities is expected to be essentially uniform if the allele is selectively neutral and if there is migration of at least a few individuals on the average between adjacent colonies each generation. It is possible that in many *Drosophila* species, with their enormous population size and with their high individual mobility in addition to the possibility of long range migration by airborne individuals, virtual panmixia is usually attained even if subpopulations are very far apart.

In addition, the associative overdominance at a neutral locus caused by linked selected loci (Ohta and Kimura, 1970, 1971d; Ohta, 1971) will contribute at least partly to keeping the distribution of neutral alleles uniform among localities. The reason for this is that the associative overdominance creates a sort of inertia so that whenever a local frequency is temporarily disturbed it tends to go back to the original frequency, although there are no definite equilibrium frequencies for the neutral allele to settle back to on the long term basis (Ohta, 1973b).

Prakash, Lewontin and Hubby (1969), in their studies on the pattern of genetic variation among subpopulations of *Drosophila pseudoobscura*, rejected the model of neutral isoalleles on the ground that frequencies of

rarer alleles at several loci (such as malic dehydrogenase locus) are essentially identical among widely separated subpopulations, and that the isoallelic hypothesis cannot explain such identical allelic configurations. The above treatment, I believe, has made it clear that their observations are not incompatible with the neutral theory.

Also I would like to point out that if the observed uniformity of the frequencies of rare alleles among localities is due to weak 'balancing selection' rather than migration, the effective size of the local population (not the whole species) has to be probably at least of the order of a million, not mentioning the fact that the selection coefficients have to be equal among localities. This is because, as first discovered by Robertson (1962) for overdominant alleles, balancing selection actually accelerates fixation rather than retards it if the equilibrium frequencies lie outside the range 0.2–0.8 unless $N_e(s_1 + s_2)$ is very large, where s_1 and s_2 are the selection coefficients against the two homozygotes. In fact, if the equilibrium frequency is 5%, $N_e(s_1 + s_2)$ has to be about 2100 in order to retard fixation by a factor of 100 as compared with the completely neutral case (see also Crow and Kimura, 1970, p. 414).

In the above discussion on the spatial distribution pattern of neutral alleles, based on the concept of the age of an allele, we have tacitly assumed that there is enough migration between neighboring colonies so that the situation is essentially equivalent to a panmictic population as a whole. When migration is highly restricted, the problem becomes more intricate. It has been shown by Maruyama (1970a, 1971) that some quantities, such as the fixation probability of a mutant gene and the sum of heterozygosity due to a mutant in the entire process leading to fixation or loss, are geographically invariant. In other words, these quantities are independent of the mode of population subdivision, provided that no part of the population is completely separated from the others (see also Maruyama, 1977). It is likely that, as compared with these quantities, the age of a mutant is more dependent on the degree of population subdivision, and further studies will be required.

For definitely deleterious alleles, their ages are much younger than the neutral alleles, and therefore, they are more likely to lead to local differentiation of frequencies.

8.4 Sum of a certain quantity along a sample path of gene frequency change

In the previous section, we investigated the problem: what is the age of an allele with a given frequency in a finite population?

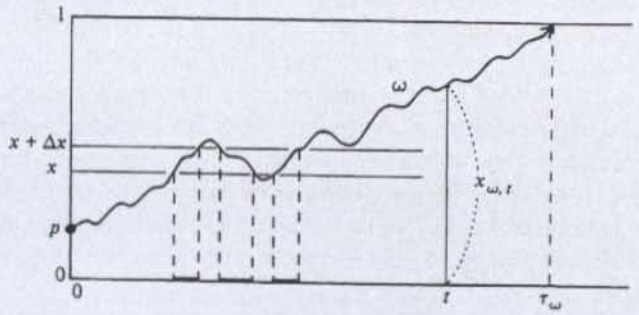
Mathematically speaking, this is a special case of a much more general problem of determining the sum (i.e. integral) of an arbitrary function, say, $f(x)$, of gene frequency (x) along a sample path of gene frequency change. In chapter 3 we presented a few examples of sample paths (see Fig. 3.3) drawn from Monte Carlo experiments simulating the behavior of a neutral allele in a small population.

In the mathematical treatment of the stochastic behavior of mutant alleles in a finite population, it is often convenient to regard the stochastic process involved as a collection of sample paths of gene frequency change. In terms of such a formulation, the problem of determining the average time until fixation of a mutant allele reduces to that of evaluating the average value of the integral along sample paths which reach fixation $(x = 1)$ starting from the initial frequency $(x = p)$, taking $f(x) = 1$ as the integrand. Similarly, the expected value of the sum of the fraction of heterozygotes produced by a mutant allele until fixation or loss in a random mating population is equal to the average of the integral of $f(x) = 2x(1 - x)$ along a sample path over all cases that start from p and reach either fixation or loss.

Although a full treatment of this subject is beyond the scope of this book, I would like to present a short account of the theory with special reference to the 'sojourn time' which a mutant allele takes at various frequency classes in the course of change.

Let us first consider a collection of sample paths representing the frequency of an allele that reaches either fixation or loss starting from an initial frequency p. We shall denote by $x_{\omega,t}$ the frequency of the allele at time t along a particular sample path ω as shown in Fig. 8.5. Let $F(p)$ be the

Fig. 8.5. Diagram illustrating a sample path (denoted by ω) of gene frequency change with time. In this case the path starts from p at time 0 and reaches fixation at time τ_ω. The thick horizontal segments in the figure represent, as a whole, the sojourn time in the frequency interval $(x, x + \Delta x)$. The ordinate; gene frequency. The abscissa; time.



expected value of the sum of $f(x)$ along a sample path. In symbols, we have

$$F(p) = E\left[\int_0^{\tau_\omega} f(x_{\omega,t})\mathrm{d}t\right], \tag{8.52}$$

where E is the operator of taking the expectation (average) over all the sample paths and τ_ω is the time when the path reaches either fixation or loss. In terms of the transition probability density $\phi(p,x;t)$, $F(p)$ can be expressed as

$$F(p) = \int_0^\infty \mathrm{d}t \int_0^1 \phi(p,x;t) f(x)\mathrm{d}x, \tag{8.53}$$

where $\phi(p,x;t)$ is the probability density that the frequency of the allele in question becomes x at time t given that its frequency is p at $t = 0$. Then it can be shown that $F(p)$ satisfies the following ordinary differential equation.

$$\frac{1}{2}V_{\delta p}\frac{\mathrm{d}^2F(p)}{\mathrm{d}p^2} + M_{\delta p}\frac{\mathrm{d}F(p)}{\mathrm{d}p} + f(p) = 0, \tag{8.54}$$

where $M_{\delta p}$ and $V_{\delta p}$ are respectively the mean and the variance of the change of the gene frequency p during one generation.

If the occurrence of mutations in the process can be neglected, the two boundaries, $x = 0$ and $x = 1$, act as absorbing barriers, and the appropriate boundary conditions are

$$F(0) = F(1) = 0. \tag{8.55}$$

The solution of equation (8.54) under these boundary conditions is

$$F(p) = [1 - u(p)]\int_0^p \psi_f(\xi)u(\xi)\mathrm{d}\xi + u(p)\int_p^1 \psi_f(\xi)[1 - u(\xi)]\mathrm{d}\xi, \tag{8.56}$$

where

$$\psi_f(\xi) = 2f(\xi)\int_0^1 G(\lambda)\mathrm{d}\lambda \Big/ [V_{\delta\xi}G(\xi)], \tag{8.56a}$$

$$u(p) = \int_0^p G(\xi)\mathrm{d}\xi \Big/ \int_0^1 G(\xi)\mathrm{d}\xi, \tag{8.56b}$$

and

$$G(\xi) = \exp\left[-2\int_0^\xi (M_{\delta x}/V_{\delta x})\mathrm{d}x\right] \tag{8.56c}$$

(Maruyama and Kimura, 1971).

In these expressions, $u(p)$ is the probability of the ultimate fixation of a

mutant allele whose initial frequency is p (Kimura, 1962), and $\exp[\cdot]$ stands for the exponential function. Note that in the special case of genic selection in which the mutant allele has a selective advantage s, we have $M_{\delta p} = sp(1-p)$, and in conjunction with $V_{\delta p} = p(1-p)/(2N_e)$, we get $G(\xi) = e^{-S\xi}$ from (8.56c), where $S = 4N_e s$. Substituting this in (8.56b), we obtain $u(p) = (1-e^{-Sp})/(1-e^{-S})$ in agreement with equation (3.11) of chapter 3.

The sum of the fraction of heterozygotes over all generations until either fixation or loss for a mutant allele whose initial frequency is p can be obtained from $F(p)$ of equation (8.56) by putting $f(p) = 2p(1-p)$, which is the expected frequency of heterozygotes involving a mutant allele whose frequency is p in a random mating population. Let us suppose that the mutant allele has a selective advantage s and that no further mutations occur in the subsequent process. In this case, $G(\xi) = e^{-S\xi}$, $V_{\delta\xi} = \xi(1-\xi)/(2N_e)$, $f(\xi) = 2\xi(1-\xi)$, and $\psi_f(\xi) = 8N_e e^{S\xi}(1-e^{-S})/S$. Thus, if we denote the expected value of the sum of the fraction of heterozygotes by $H(p)$, we obtain $H(p) = 2[u(p) - p]/s$ in agreement with equation (3.15).

For a selectively neutral mutant, letting $s \to 0$, the equation for $H(p)$ reduces to $H(p) = 4N_e p(1-p)$. Thus the total heterozygosity produced by a single mutant is

$$H(1/2N) \approx 2N_e/N. \tag{8.57}$$

This reduces to 2 if $N_e = N$, that is, if the effective and the actual population sizes are equal.

The average length of time until fixation or loss of a mutant which we denote by $T(p)$ can be obtained by putting $f(p) = 1$ in equation (8.56). For a selectively neutral mutant, we have $M_{\delta x} = 0$ so that $G(\xi) = 1$ and $u(p) = p$, and therefore $\psi_f(\xi) = 4N_e/[\xi(1-\xi)]$. Thus we obtain

$$T(p) = -4N_e[(1-p)\log_e(1-p) + p\log_e p], \tag{8.58}$$

which agrees with the result obtained by Watterson (1962).

We shall now consider the problem of computing the average value of the sum of $f(x)$ along sample paths of gene frequency change conditional on fixation, namely, we take the average only for sample paths reaching fixation and excluding those leading to loss. In symbols, we consider the quantity

$$E\left[\int_0^{\tau_\omega} f(x_{\omega,t})\,dt \,\middle|\, x_{\omega,\tau_\omega} = 1\right].$$

In terms of the transition probability density, this can be expressed as

$F_1(p)/u(p)$, where $u(p)$ is the fixation probability and

$$F_1(p) = \int_0^\infty dt \int_0^1 \phi(p,x;t) f(x) u(x)\,dx. \tag{8.59}$$

In this expression, the subscript 1 denotes that it is concerned with sample paths reaching fixation ($x = 1$).

It can then been shown that $F_1(p)$ satisfies the differential equation

$$\tfrac{1}{2} V_{\delta p} F_1''(p) + M_{\delta p} F_1'(p) + u(p) f(p) = 0, \tag{8.60}$$

where primes indicate differentiation. This equation is similar to (8.54). The only difference is that here the equation contains the term $u(p)f(p)$ instead of $f(p)$ as found in equation (8.54).

Under no mutation, the relevant boundary conditions are

$$F_1(1) = F_1(0) = 0. \tag{8.61}$$

It turns out then that the expected value of the sum of $f(x)$ along a sample path starting from p and reaching fixation is

$$E\left[\int_0^{\tau_\omega} f(x_{\omega,t})dt \,\middle|\, x_{\omega,\tau_\omega} = 1\right] = \frac{1-u(p)}{u(p)} \int_0^p \psi_f(\xi)[u(\xi)]^2 d\xi + \int_p^1 \psi_f(\xi) u(\xi)[1-u(\xi)]\,d\xi, \tag{8.62}$$

where $\psi_f(\xi)$ is given by (8.56a).

The average time until fixation (excluding the cases of eventual loss) of a selectively neutral mutant can be obtained from this equation by putting $f(x) = 1$, and noting that $G(\xi) = 1$ for a neutral allele and therefore $\psi_f(\xi) = 4N_e/[\xi(1-\xi)]$. Substituting these in (8.62) we can immediately see that at the limit $p \to 0$ this reduces to $4N_e$ in agreement with Kimura and Ohta (1969a).

The above treatments are concerned with the mean (the first moment) of the stochastic integral, but a theory has been developed by Maruyama and Kimura (1971) to obtain the higher moments. Those who are interested in this type of problem may consult this paper together with Maruyama and Kimura (1975), Maruyama (1977) and Nagylaki (1974).

Let us now treat the problems related to the sojourn times, namely, the lengths of time which a mutant allele spends at various frequency classes in the course of change. We shall be particularly concerned with the mean sojourn time for sample paths conditional on fixation as depicted in Fig. 8.5. Using the general theory which we have developed in this section, the mean sojourn time at a given gene frequency, say y, may be obtained by letting

$f(x) = \delta(x - y)$ where $\delta(\cdot)$ is the Dirac delta function. This function extracts the probability at $x = y$ but ignores all probabilities for other values of x than y. More generally, for any function $\phi(x)$, we have

$$\int \phi(x)\delta(x - y)\mathrm{d}x = \phi(y) \tag{8.63}$$

provided that the range of integration includes the point y. In the following, the letter x rather than y will be used to represent the allele frequency.

Let $\Phi_1(p,x)$ be the density of the conditional mean sojourn time for sample paths that reach fixation starting from p. This has the meaning that $\Phi_1(p,x)\mathrm{d}x$ represents the average of the total time which an allele spends in the frequency range $x \sim x + \mathrm{d}x$ before it becomes fixed in the population (see the thick horizontal segments on the time-axis in Fig. 8.5). From the above explanation, it will be evident that $\Phi_1(p,x)$ is given by equation (8.62) if we put $f(\xi) = \delta(\xi - x)$ in (8.56a). This leads to the following expressions for $\Phi_1(p,x)$.

$$\left.\begin{aligned} \Phi_1(p,x) &= \frac{2u(x)[1-u(x)]}{V_{\delta x}u'(x)}, \quad (x \geqq p) \\ \Phi_1(p,x) &= \frac{2[1-u(p)]u^2(x)}{u(p)V_{\delta x}u'(x)}, \quad (x < p) \end{aligned}\right\} \tag{8.64}$$

where $u(x)$ is the fixation probability and

$$u'(x) = \mathrm{d}u(x)/\mathrm{d}x = G(x)\Big/\int_0^1 G(x)\mathrm{d}x.$$

For a mutant allele which is represented only once at the moment of appearance, $p = 1/(2N)$, where N is the actual population size. Then only the expression for $x \geqq p$ in equation 8.64 is relevant, and the average number of generations which the mutant allele spends in the interval between x and $x + \mathrm{d}x$ before fixation is $\Phi_1(x)\mathrm{d}x$, where

$$\Phi_1(x) = \frac{2u(x)[1-u(x)]}{V_{\delta x}u'(x)}, \quad \left(1 > x \geqq \frac{1}{2N}\right). \tag{8.64a}$$

The mean sojourn time at each frequency class, $x = i/(2N)(i = 1, 2, \ldots, 1 - 1/2N)$, is given by $\Phi_1(x)/(2N)$.

If the mutant is selectively neutral, $u(x) = x$ and $V_{\delta x} = x(1-x)/(2N_e)$, and the above equation reduces to

$$\Phi_1(x) = 4N_e. \tag{8.65}$$

This means that the neutral mutant spends on the average $\Phi_1(x)/2N = 2N_e/N$ generations at each frequency class until fixation. Summing this for all $(2N - 1)$ frequency classes, we obtain the result that a neutral mutant

takes on the average about $4N_e$ generation until fixation (see equation 3.23). If the actual and effective population sizes are equal, the mutant allele spends on the average two generations at each frequency class on its way to fixation.

Next, consider the case of genic selection. Let s be the selective advantage of the mutant allele over the pre-existing allele. Then $M_{\delta x} = sx(1-x)$, $V_{\delta x} = x(1-x)/(2N_e)$ and $u(x) = (1 - \mathrm{e}^{-Sx})/(1 - \mathrm{e}^{-S})$, so that we obtain.

$$\Phi_1(x) = \frac{(1-\mathrm{e}^{-Sx})[1-\mathrm{e}^{-S(1-x)}]}{sx(1-x)(1-\mathrm{e}^{-S})}, \tag{8.66}$$

where $S = 4N_e s$. In the neighborhood of $x = 0$, the right-hand side of this equation becomes roughly equal to $4N_e$, suggesting that even a mutant allele having a definite selective advantage or disadvantage (assuming $|s|$ to be small, but $4N_e|s|$ may be large) behaves as if selectively neutral when it is still rare. At this stage its fate is largely determined by random sampling drift.

At intermediate frequencies, the mean sojourn times of a definitely advantageous mutant (i.e. $4N_e s \gg 1$) are shorter than the corresponding sojourn times of neutral alleles. For values of x for which both $4N_e sx$ and $4N_e s(1-x)$ are reasonably larger than unity, say, larger than 5, we have approximately

$$\Phi_1(x)\mathrm{d}x = \frac{\mathrm{d}x}{sx(1-x)}. \tag{8.67}$$

It is well-known from the deterministic treatment of gene frequency change (see equation 6.4 of chapter 6) that the rate of change of allelic frequency is given by

$$\frac{\mathrm{d}x}{\mathrm{d}t} = sx(1-x), \tag{8.68}$$

so that the length of time which a mutant allele with selective advantage $s(>0)$ takes to increase from x to $x + \mathrm{d}x$ is $\mathrm{d}t = \mathrm{d}x/[sx(1-x)]$. This means that in this frequency range the behavior of mutant alleles can be described satisfactorily by deterministic theory. In the neighborhood of $x = 1$, the right-hand side of equation (8.66) again becomes roughly equal to $4N_e$. This means that a mutant allele having a definite selective advantage, as it approaches fixation, starts to behave as if it were selectively neutral. Then, the deterministic equation (8.68) again becomes inadequate to describe its behavior, and we must use equation (8.66).

The selection coefficient s may either be positive or negative, but when it is negative, it is convenient to put $s = -s'$, where $s' > 0$. Substituting this in

(8.66) we get

$$\Phi_1(x) = \frac{(1-e^{-S'x})[1-e^{-S'(1-x)}]}{s'x(1-x)(1-e^{-S'})}, \tag{8.69}$$

where $S' = 4N_e s'$. Comparison of the two equations, (8.66) and (8.69), shows that two mutant alleles, one with selection coefficient s and the other with $-s$, have exactly the same mean sojourn time at various frequency classes under the condition that the alleles are destined to fixation in the population. Furthermore, comparison of these equations with (8.65) shows that the mean sojourn times are shorter than those of selectively neutral alleles. Thus, the average time until fixation, i.e. $\int_p^1 \Phi_1(x)dx$ where $p = 1/(2N)$, becomes shorter for deleterious as well as for advantageous alleles. It is obvious that the length of time until fixation should be shorter for an advantageous mutant whose increase is helped by positive natural selection, but it is rather difficult to perceive by common sense, that this is also shorter for a deleterious mutant whose increase tends to be resisted by negative selection. This apparent paradox can be resolved by noting that the probabilities of these two events are very different. In particular, fixation of a deleterious mutant must be a very rare event, and only under concatenation of lucky steps of chance increase, can such a mutant reach fixation. In other words, when a mutant is fixed by chance in the face of adverse selection, the entire process must be finished quickly through a series of lucky jumps.

So far, we have considered the mean sojourn time for the sample paths leading to fixation (essentially the same method can be used to treat the paths leading to extinction). In addition to these, it is desirable to derive expressions for the mean sojourn time for all the paths together, that is, for the paths that go either to fixation or loss starting from p. The density, $\Phi(p,x)$, of such mean sojourn time satisfies equation (8.54) with $f(p) = \delta(x-p)$. If we assume that both $x=0$ and $x=1$ act as absorbing barriers, the boundary conditions are $\Phi(0,x) = \Phi(1,x) = 0$, and the solution can be expressed as follows.

$$\left.\begin{aligned} \Phi(p,x) &= \frac{2u(p)[1-u(x)]}{V_{\delta x}u'(x)} \quad (x \geqq p) \\ \Phi(p,x) &= \frac{2[1-u(p)]u(x)}{V_{\delta x}u'(x)} \quad (x < p) \end{aligned}\right\} \tag{8.70}$$

For a selectively neutral allele, the above reduces to

$$\left.\begin{aligned} \Phi(p,x) &= 4N_e p/x \qquad (x \geqq p) \\ \Phi(p,x) &= 4N_e(1-p)/(1-x) \quad (x < p) \end{aligned}\right\} \tag{8.71}$$

Equations (8.70) are equivalent to 'the transient function' $t(x)$ of Ewens (1969), and also to the steady flux distribution with $v_m = 1$ given by Kimura (1969b). For a mutant allele which is represented only once at the moment of appearance, the density of the mean sojourn time at x before fixation or loss of the allele can be obtained from the first formula of (8.70). In this case, if we note that

$$u\left(\frac{1}{2N}\right) \approx \frac{1}{2N} \cdot \frac{1}{\int_0^1 G(x)dx}, \tag{8.72}$$

we obtain

$$\Phi(x) = \frac{1}{N \cdot V_{\delta x} G(x)} \cdot \frac{\int_x^1 G(\xi)d\xi}{\int_0^1 G(\xi)d\xi}, \quad \left(\frac{1}{2N} \leqq x < 1\right), \tag{8.73}$$

where $G(\xi)$ is given by (8.56c). In the neighborhood of $x=0$, we have approximately $\Phi(x) = 1/(NV_{\delta x})$. Then, if we put $V_{\delta x} \approx x/(2N_e)$, we have

$$\Phi(x) \approx (2N_e/N) \cdot x^{-1}, \quad (0 < x \ll 1). \tag{8.74}$$

Comparison of this expression with the equation for the mean sojourn time of a neutral mutant (see equation (8.71), where we put $p = 1/(2N)$) shows that, in the neighborhood of $x=0$, the behavior of mutants in general, including those having a mild selective advantage or disadvantage, is essentially the same as that of a selectively neutral mutant.

Detailed behavior of overdominant alleles can also be studied using the same general method as explained in this section. Readers who are interested in the subject are invited to read Nei and Roychoudhury (1973).

We have restricted our consideration to the average (first moment) of sojourn times, but we can obtain the higher moments as shown by Maruyama and Kimura (1971) and Nagylaki (1974). The latter went further and worked out not only a general expression for the n-th moment but also the probability distribution of the sojourn time. For details, readers should refer to Nagylaki (1974).

8.5 The stepwise mutation model

The infinite allele model discussed in section 8.2 assumes that whenever a mutation occurs it represents a new, not pre-existing allele. This model would be perfectly realistic if individual variants in the population could be identified at the level of nucleotide or amino acid sites. However, experimental analyses of genetic variability at present still rest mostly on a

variants. A mutation can be detected only when it leads to replacement of an amino acid which causes a change in electrophoretic mobility. The alleles are identified only as a discrete spectrum of broad bands on the electrophoresis gels.

In order to accommodate such a limited resolving power, a model of stepwise production of neutral alleles was proposed by Kimura and Ohta (1973a) and some of its mathematical consequences were first worked out by Ohta and Kimura (1973). In its simplest form the model assumes that the entire sequence of allelic states can be expressed by integers, and that if an allele changes state by mutation (due to replacement of an amino acid that causes a change in electric charge), it moves either one step in the positive direction or one step in the negative direction (see Fig. 8.6). In this model, one positive and one negative change in electric charge cancel each other leading the allele back to the original state.

Previously, Bulmer (1971b) criticized the neutral theory on the ground that there is a definite association between the allelic frequencies and their relative mobilities on the electrophoretic gels. Analysing the data of Prakash *et al.* (1969), he noted the interesting fact that the most frequent allele in the population tends to have intermediate mobility and this is flanked by rare alleles. In other words, if we denote by C each common allele and by R each rare allele, then their gel sequence together with their population frequencies (sometimes called electrophoretic profile) is always

Fig. 8.6. Stepwise mutation model.

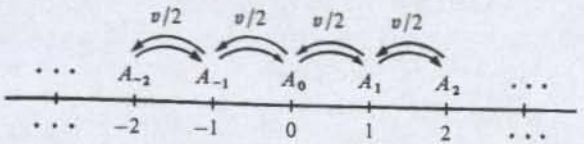
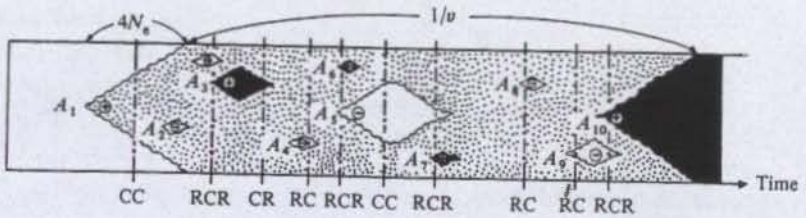


Fig. 8.7. A typical pattern of extinction and multiplication of selectively neutral, electrophoretically detectable mutants in a finite population. It is assumed that mutations occur at the rate of one mutation every ten generations ($4N_e v = 0.2$). Modified from Kimura and Ohta (1973a). For details see text.



RCR and never CRR or RRC. From such observation Bulmer claimed that the alleles involved are not selectively neutral, because, if they were, one would expect all possible orders of sequences to be equally likely.

That Bulmer's criticism is unfounded and that his observation is indeed compatible with the neutral theory was pointed but by Maynard Smith (1972). Also, Kimura and Ohta (1973a), using the stochastic theory of population genetics and based on the nature of electrophoretic mutations, explained why Bulmer's observation is a natural consequence of protein polymorphisms being selectively neutral.

In Fig. 8.7, a typical pattern of extinction and multiplication of selectively neutral mutants is illustrated, where the assumption of stepwise production of mutations is also incorporated. The figure was constructed by taking into account the population dynamics of neutral mutants as explained in the above few sections. Particularly, it has been noted in drawing the figure that if p is the present frequency of a neutral allele in the population, then the probability is p/x that it subsequently reaches a higher frequency x before it becomes either lost from the population or fixed in it. As to the mode of change in mobilities due to mutation, it is assumed that change occurs in such a way that the band moves either one step in the positive direction or one step in the negative direction on the gel. In Fig. 8.7, such changes are marked by plus and minus signs. In more concrete terms, it is assumed that aspartic and glutamic acids are acidic and negatively charged, lysine and argine are basic and positively charged, while the rest are electrically neutral, and that the mobility is solely determined by the difference in net charge of a protein. Furthermore, it is assumed that one positive and one negative charge cancel each other. At various times (abscissae), the sequence on the electrophoresis gels of common (C) and rare (R) alleles are expressed by such representations as CC, CR, RCR, etc. It will be seen that although RCR may occur, CRR, RRC and CRC are unlikely to occur. In particular, CRC presupposes two common alleles separated by at least two mutational steps with a rare intermediate allele between them, and it is clear that such an event is extremely unlikely when the effective number of alleles is 1.2 (i.e. $4N_e v = 0.2$).

Thus, Fig. 8.7 clearly illustrates that Bulmer's prediction regarding the equal occurrence of all sequences is unfounded, and his observation is indeed compatible with the neutral theory. Referring to the assumption that one negative and one positive charge cancel each other, Kimura and Ohta (1973a) stated that this is clearly an oversimplification and that it is possible, if more sensitive tests could be applied, that two variants having charges of the same sign could still be discriminated. In such a circumstance, we

should expect that CRR may be found to occur, giving an exception to Bulmer's rule.

It is expected that under the stepwise production of mutant alleles the level of heterozygosity is lower than when every mutational change can be discriminated. Using the stepwise mutation model as depicted in Fig. 8.6, and assuming a random mating population of the effective size N_e, Ohta and Kimura (1973) showed that the average homozygosity or the expected value of the sum of squares of allelic frequencies at equilibrium is

$$\bar{H}_0 = E\left[\sum_i x_i^2\right] = 1/\sqrt{(1+8N_e v)}, \tag{8.75}$$

where x_i is the frequency of A_i and v is the mutation rate per locus per generation. More generally, they showed that the expected value of the product of the frequencies of alleles which are k steps ($k = 0, 1, 2, \ldots$) apart is

$$C_k = E\left[\sum_i x_i x_{i+k}\right] = \bar{H}_0 \lambda_1^k, \tag{8.76}$$

where

$$\lambda_1 = \frac{1 + 4N_e v - \sqrt{(1+8N_e v)}}{4N_e v} \tag{8.76a}$$

The expected value of the effective number (n_e) of alleles in the population, as defined by $1/\bar{H}_0$, is

$$\bar{n}_e = \sqrt{(1+8N_e v)}. \tag{8.77}$$

It is of interest to compare this with the corresponding formula $\bar{n}_e = 1 + 4N_e v$ (see equation 8.29) for the infinite allele model which assumes that every mutation is new and detectable. Although the two models give very similar $\bar{n}_e$ values when $4N_e v$ is small, the stepwise model gives a much smaller value than the infinite allele model when $4N_e v$ is large. For example, if $4N_e v = 0.2$, we have $\bar{n}_e = \sqrt{(1+8N_e v)} = 1.18$ which is very near to $\bar{n}_e = 1 + 4N_e v = 1.20$. However, if $4N_e v = 400$, we get $\bar{n}_e = 28.3$ for the stepwise mutation model but $\bar{n}_e = 401$ for the infinite allele model.

Since Ohta and Kimura's (1973) report, a large number of papers have been published treating the model in various contexts. Also, it has received several other designations such as the charge-state model (Marshall and Brown, 1975; Ewens, 1977), the charge ladder model (Ramshaw *et al.*, 1979), the Ohta-Kimura model (Li, 1976; Kingman, 1977), or simply the ladder model, etc.

As to the problem of whether the model is realistic in describing the electrophoretically detectable variations, opinions differ considerably. Some authors, such as Johnson (1974), criticize the model as being totally

unrealistic from the biochemical standpoint. Also, Ramshaw *et al.* (1979), based on their experiments using hemoglobin variants, came to the conclusion that the model is clearly incorrect and that the sequential method of electrophoresis under several conditions of pH and gel concentration can detect a substantial fraction of amino acid changes. Fuerst and Ferrell (1980), through a similar study of hemoglobin variants, found that, although the net electrostatic charge is the principal determinant of electrophoretic mobility, significant deviation from strict stepwise mobility occurs for 30 to 40% of the variants studied. They conclude that the infinite-allele model of Kimura and Crow (1964) or the 'mixed model' of Li (1976) may be more appropriate than the stepwise mutation model for the analysis of electrophoretic data from natural populations.

On the other hand, the data of Richardson *et al.* (1975) from *Drosophila mulleri* complex support the assumption of an equally spaced ladder. Also, Marshall and Brown (1975) consider the model valid as a first approximation in predicting the amount and nature of electrophoretic variability in natural populations. Space does not permit us to discuss further pros and cons of the stepwise mutation model as a means of analysing the nature of electrophoretically detectable variants. Readers who are interested in the subject should consult a thoughtful review by Brown *et al.* (1981). In what follows, I shall summarize various results obtained through mathematical studies of the stepwise mutation model.

Moran (1975, 1976) developed a powerful method which enables us to compute the second and the higher moments of allelic frequencies. The formula for the variance of homozygosity (or heterozygosity) turned out to be very complicated, but he obtained a simple approximation formula

$$\mathrm{Var}(H_0) = \theta/(3 + 11.25\theta + 13\theta^2 + 1.7\theta^3), \tag{8.78}$$

where $\theta = 4N_e v$.

Moran (1975) also considered the quantity

$$\sigma^2 = E\left[\sum_i i^2 x_i^2 - \left(\sum_i i x_i\right)^2\right] \tag{8.79}$$

which is the expected value of the variance of the step numbers in the frequency profile. Although the mean position of the frequency profile wanders around in the allele space with time, the variance σ^2 settles to a constant value at equilibrium. He showed that at equilibrium

$$\sigma^2 = 2N_e v. \tag{8.80}$$

For example, if $4N_e v = 0.2$ this becomes $\sigma^2 = 0.1$, so that the standard deviation of step numbers is $\sigma = 0.32$.

The problem of obtaining the distribution of allelic frequency ('frequency spectrum') for the stepwise mutation model is very difficult, and the exact solution has not been obtained. However, Kimura and Ohta (1975, 1978) developed a theory which enables us to obtain approximate formulae which are sufficiently accurate for our purpose when $4N_e v$ does not exceed unity. Without going into the details of the mathematical treatment involved, I shall mention an important property which distinguishes the distribution under stepwise mutation (which we denote by $\Phi_{step}(x)$) from that under the conventional infinite allele model (denoted by $\Phi(x)$; see equation 8.17); the expected numbers of low-frequency alleles is much smaller under stepwise mutation. For example, when $4N_e v = 0.2$, we have $\Phi_{step}(0.001) = 111.3$ but $\Phi(0.001) = 200.1$. Also, when $4N_e v = 1.0$, we get $\Phi_{step}(0.001) = 147.7$, while $\Phi(0.001) = 1000.0$. Such a difference in these two models was noted earlier by Ohta (1975).

A related but a more remarkable feature is that the actual number of alleles can be enormously different in these two models, even when the effective numbers of alleles ($\bar{n}_e = 1/\bar{H}_0$) are similar. This occurs when $4N_e v$ is small but the population size N is very large (assuming that $4N_e v$ is the same for both models). If we denote by $\bar{n}_{a;step}(N)$ the expected value of the actual number of alleles in the population, then

$$\bar{n}_{a;step}(N) = \int_{1/(2N)}^{1} \Phi_{step}(x)\mathrm{d}x. \tag{8.81}$$

This value increases extremely slowly as N increases if we hold $4N_e v$ constant. In fact, Kimura and Ohta (1978), based on their approximation theory, obtained the result that $\bar{n}_a(\infty) = 4.09$ for $4N_e v = 1$. (In this case the effective number of alleles is 1.73.) On the other hand for the conventional infinite allele model, $\bar{n}_a$ increases without limit as N increases even if $4N_e v$ is held constant (see equation 8.28). Rapid increase of $\bar{n}_a$ with N may be suggested from the fact that already in a sample of the size 250 from a population in which $4N_e v = 1$, the number of different alleles contained is 6.79 (see Table 1 of Ewens, 1972). Note that the effective number of alleles, $\bar{n}_e = 4N_e v + 1 = 2.0$, does not differ very much from the corresponding value ($\bar{n}_e = 1.73$) for the stepwise mutation model.

More recently, Kesten (1980) developed a rigorous mathematical theory to estimate the number of different alleles in a large sample based on the Ohta-Kimura model. According to him, the asymptotic value of $\Lambda(n, N)$, the number of different alleles in a sample of size n from a population of N breeding individuals, does not depend on the mutation rate. For large values of n and N ($n \to \infty$, $N \to \infty$ with $2N \geq n$), he showed that

$$\Lambda(n, N) \to h^{-1}(p_2^+ + p_1^-)\lambda(n), \tag{8.82}$$

where $h^{-1}(p_2^+ + p_1^-)$ is a constant, which for the simplest case illustrated in Fig. 8.6 appears to reduce to 2. What is remarkable is the behavior of $\lambda(n)$. A few values of $\lambda(n)$ are as follows: $\lambda(n) = 1$ for $3 \leqq n < 16$, $\lambda(n) = 2$ for $16 \leqq n < 3814280$, and $\lambda(n) = 3$ for $3814280 \leqq n < 10^{1656520}$. This shows that $\lambda(n)$ grows slower than any repeatedly iterated logarithm of n. Yet $\lambda(n)$ should become indefinitely large as $n \to \infty$ so that $n_a(\infty) = \infty$, contrary to Kimura and Ohta's (1978) assertion that $n_a(\infty)$ remains finite at $n \to \infty$. Although not valid in the strictest mathematical sense, the approximation of Kimura and Ohta (1978) may be justified since the sample size in the real world can never be larger than 10^{100}. In fact, as is well known, the total number of proton-electron pairs in our universe is estimated to be of the order of 10^{80}, and we seldom imagine taking a sample of size larger than $n = 10^4$. The stepwise mutation model has also been investigated by Kingman (1976, 1977) because of its intrinsic mathematical interest.

In a more concrete vein, Weir *et al.* (1976) developed a statistical theory to test if polymorphisms detected by electrophoresis are in accordance with the stepwise, neutral mutation model. They applied the test to analyse Ayala and Tracey's (1974) data on 27 allozyme loci in six populations of *Drosophila willistoni*, and obtained the following results. Of the 27 loci, about 15% (approximately 1/4 of all polymorphic loci) significantly and consistently departed from the prediction of the model. A further 15% departed significantly in some populations but not in others. The remaining loci showed an acceptable fit to the model. In the first two cases, selection appeared to be acting against the alleles associated with extreme electrophoretic mobilities.

Maruyama and Kimura (1978) investigated the amount of genetic variability maintained in a finite population assuming that both selectively neutral and slightly deleterious mutations occur in stepwise fashion. They showed, using two examples in which very slightly deleterious mutations occur 100 times as frequently as strictly neutral mutations and have a selective disadvantage only 10 times as large as the total mutation rate, that the effective number of alleles (n_e) increases much more slowly as the population size (N_e) increases than is the case when all the mutations are strictly neutral.

Takahata (1980) considered a composite stepwise mutation model in which alleles in each charge state (or an 'electromorph' in the sense of King and Ohta, 1975) are members of K allelic states (i.e. each electromorph consists of K allelic states). In this model it is assumed that mutations occur at the rate v_e for electrophoretically detectable alleles and at the rate v_h for electrophoretically hidden (silent) alleles. In the former case, a change occurs in such a way that the allele moves one step either in the positive

direction or in the negative direction with equal frequency. In the latter case of a silent change, the allele changes to any of the remaining $K-1$ alleles in the same charge state with equal frequency. Thus, an allele changes to one of the other electrophoretically distinct alleles at the rate $v_e/(2K)$, or it changes to one of the other electrophoretically identical alleles at the rate $v_h/(K-1)$. Assuming selective neutrality of alleles, and taking into account both electrophoretically detectable and silent alleles, he showed that at equilibrium the average net homozygosity is

$$\bar{H}_0 = \frac{1}{K}\left[\frac{1}{\sqrt{(1+2\theta_e)}} + \frac{(K-1)^2}{K(1+\theta_T)-(1+\theta_e)}\right], \tag{8.83}$$

where $\theta_T = 4N_e v_T$, $\theta_e = 4N_e v_e$ and $v_T = v_e + v_h$. The average homozygosity when only electrophoretic mobility is taken into account is

$$C_0 = 1/\sqrt{(1+2\theta_e)}. \tag{8.84}$$

This equation is equivalent to equation (8.75), and it is essentially a special case of equation (8.83) for $K = 1$. On the other hand, for $\theta_e = 0$, equation (8.83) reduces to $\bar{H}_0 = (K-1+\theta_T)/(K-1+K\theta_T)$, which is equivalent to (8.33a).

Takahata's model was motivated by Singh *et al.*'s (1976) report that, by using four different electrophoretic conditions and heat stability test, they could detect 37 alleles at the xanthine dehydrogenase (Xdh) locus in *Drosophila pseudoobscura* where previously only 6 had been found. This means that the additional conditions enabled them to detect six times as much polymorphism as the simple (standard) electrophoresis. According to them, the average heterozygosity at the Xdh locus was 0.44 when the standard method of electrophoresis was used alone, but this rose to 0.73 when all five criteria were applied. If we use Takahata's model to analyse their data for which $C_0 = 0.56$, $\bar{H}_0 = 0.27$ and $K = 5$, we get $\theta_e = 1.09$, and $\theta_T = 3.47$. This gives the ratio $\theta_e/\theta_T = 0.32$, which is the fraction of mutations to electrophoretically detectable alleles. Takahata also pointed out that, when θ_e is small and $K \geqq 2$, the net genetic variability in the population is similar to that predicted by the infinite allele model ('Kimura-Crow' model) rather than to that predicted by the stepwise mutation model.

8.6 Infinite site model

This model was originally formulated (Kimura, 1969b) based on the consideration that the total number of nucleotide sites making up the human genome is very large, amounting to $(3 \sim 4) \times 10^9$ base pairs, while the mutation rate per site must be very low, possibly of the order of 10^{-8} or 10^{-9} per generation. The model assumes that the total number of sites

available for mutation is so large while the mutation rate per site is so low that whenever a mutant appears, it represents a mutation at a new, previously homallelic site, that is at a site where mutant forms are not segregating.

Since a mutant which appears in a finite population is either fixed or lost from the population within a finite length of time, the number of segregating sites at any given moment remains finite even if the total number of available sites is infinite and if a certain (finite) number of new mutations are constantly being fed into the population in each generation.

Using this model, a mathematical theory was developed (Kimura, 1969b) which enables us to estimate the number of heterozygous nucleotide sites maintained in a finite population under steady flux of mutations. Then it was shown, for example, that in a mammalian species having the variance effective number of 10 000, if the mutations (nucleotide change) occur at the rate of two per gamete per generation and if they are selectively neutral, the average number of heterozygous nucleotide sites per individual due to such mutations turns out to be about 8×10^4.

This model was later called 'the model of infinite sites' (Kimura, 1971), and has since been proven to be very useful in treating problems of evolution and variation at the molecular level. Without going into the details of the mathematical treatment involved, I shall summarize, in what follows, the main results obtained on the number of heterozygous sites, the number of segregating sites, and the distribution of mutant frequencies in the population (for details, see Kimura, 1969b).

Consider a random mating, diploid population of the effective size N_e, and assume that new mutations occur in v_m sites in the population in each generation. Let s be the selective advantage of the mutant over the pre-existing form so that the advantage in homozygotes and heterozygotes are respectively $2s$ and s (the case of 'semidominance' in fitness). We denote by p the initial frequency of the mutant. Let us assume that the parameters s and p are the same among mutations at different sites. Then, it can be shown that at equilibrium in which mutational input and random extinction balance each other, the number of heterozygous sites per individual is

$$H(p) = (2v_m/s)[u(p) - p], \tag{8.85}$$

where $u(p) = (1 - e^{-Sp})/(1 - e^{-S})$, in which $S = 4N_e s$. Here, $u(p)$ represents the probability of ultimate fixation of a mutant. If the mutants are selectively neutral ($s \to 0$), this reduces to

$$H(p) = 4N_e v_m p(1-p). \tag{8.86}$$

In a population consisting of N individuals, if we assume the a mutant

form is represented only once at the moment of its appearance (a realistic assumption for molecular mutants) we have $p = 1/(2N)$. Then the number of heterozygous sites per individual is given by $H(1/2N)$.

In the special case of selectively neutral mutants, we have, from (8.86),

$$H(1/2N) \approx 2v_m N_e/N. \tag{8.87}$$

If we denote by v_{gam} the mutation rate per gamete per generation so that

$$v_m = 2Nv_{gam}, \tag{8.88}$$

equation (8.87) may be expressed as

$$H(1/2N) \approx 4N_e v_{gam}. \tag{8.89}$$

Later, when we restrict our consideration to a certain region of the genome (such as a coding region consisting of 1000 nucleotide sites) rather than the whole genome, we shall use the letter v rather than v_{gam}.

In the previous section (see section 8.4), the same symbol, $H(p)$, was used to represent the expected value of the sum of the frequencies of heterozygotes made by a mutant over all generations until fixation or loss starting from the initial frequency p. There, it was shown (see equation 8.57) that for a selectively neutral mutant which is represented only once at the moment of appearance, this quantity reduces to $2N_e/N$. Since $2Nv_{gam}$ new, different mutations occur per generation in the population, the expected number of heterozygous sites at equilibrium is $(2N_e/N) \times 2Nv_{gam} = 4N_e v_{gam}$ in agreement with equation (8.89).

As an example, let us assume that in the human population selectively neutral mutations occur on the average at the rate 10^{-8} per site per generation. With 3.5×10^9 nucleotide pairs making up the genome (the haploid chromosome set), $v_{gam} = 3.5 \times 10^9 \times 10^{-8} = 35$. If the effective population number of human species during its main course of evolution is $N_e = 10^5$, then, from (8.89), we have $H(1/2N) = 4 \times 10^5 \times 35 = 1.4 \times 10^7$, namely, the number of heterozygous nucleotide sites due to neutral mutations amounts to 14 million. Although this is large, it amounts to only 0.4% of the total number of nucleotide sites making up the genome.

Let us consider a more general case with natural selection. From equation (8.85) we obtain

$$H(1/2N) = 4N\left(\frac{v_{gam}}{s}\right)\left[\frac{1-\exp(-Sp)}{1-\exp(-S)} - \frac{1}{2N}\right] \tag{8.90}$$

where $S = 4N_e s$. This reduces to

$$H(1/2N) \approx 4N_e\left(\frac{v_{gam}}{s}\right)\left[\frac{2s}{1-\exp(-4N_e s)} - \frac{1}{2N_e}\right] \tag{8.91}$$

if $|s| \ll 1$.

If each mutant has a definite selective advantage such that $4N_e s \gg 1$, this formula is further simplified to give

$$H(1/2N) \approx 8N_e v_{gam}, \tag{8.92}$$

which shows that the definitely advantageous mutants contribute twice as much to heterozygosity as the selectively neutral mutants (see equation 8.89), provided that they occur with equal frequencies. On the other hand, if the mutant is definitely deleterious such that $4N_e s' \gg 1$, where $s' = -s$, equation (8.91) reduces to

$$H(1/2N) \approx 2v_{gam}/s'. \tag{8.93}$$

Ohta (1973, 1974) proposed a hypothesis stating that the majority of protein polymorphisms are caused by mutational production of very slightly deleterious alleles and suggested the possibility that such polymorphisms represent mutation-selection balance where the selection intensity involved is not much larger than the mutation rate. A similar possibility has also been suggested by Wright (1978).

On the whole, I think that mutations having a definite advantage or disadvantage cannot contribute greatly to the heterozygosity of an individual because of the rare occurrence of advantageous mutations and the rapid elimination of deleterious ones.

Next, let us consider the average number of segregating sites in the population. This will be denoted by $I_1(p)$, where we assume $p = 1/(2N)$. If each mutant is definitely advantageous such that $4N_e s \gg 1$, we have

$$I_1(1/2N) \approx 4N_e v_{gam}[\log_e(8NN_e s) + \gamma + 1], \tag{8.94}$$

where $\gamma = 0.5772\ldots$ is the Euler's constant. If, on the other hand, each mutant is definitely deleterious so that $4N_e s' \gg 1$, where $s' = -s$, we have

$$I_1(1/2N) \approx 4N_e v_{gam}[-\log_e(2N_e s'/N) - \gamma + 1]. \tag{8.95}$$

For the case of selectively neutral mutations, the average number of segregating sites is

$$I_1(1/2N) \approx 4N_e v_{gam}[\log_e(2N) + 1]. \tag{8.96}$$

For example, in a population with $N_e = 10^5$ and $N = 10^6$, if the neutral mutations occur at the rate of 10^{-8} per site per generation and if the total number of nucleotide sites per genome is 3.5×10^9, we get $I_1(1/2N) = 2.2 \times 10^8$. This amounts to roughly 6% of the total sites, although the frequencies of mutants at segregating sites are mostly in the low-frequency range.

Finally, let us consider the distribution of allelic frequencies at equilibrium. We denote by $\Phi(p, x)$ the expected number of sites where the frequencies of mutants lie within the range between x and $x + dx$. Under the

condition that a mutant at each individual site is represented only once at the moment of occurrence, if we denote $\Phi(1/2N, x)$ by $\Phi(x)$, we obtain

$$\Phi(x) = \frac{2v_{gam}}{V_{\delta x}G(x)} \frac{\int_x^1 G(x)dx}{\int_0^1 G(x)dx}, \tag{8.97}$$

where $G(\cdot)$ is given by equation (8.56c) and $M_{\delta x}$ and $V_{\delta x}$ are the mean and the variance in the change of the mutant frequency (x) during one generation. This formula was obtained by Kimura (1964) as an extension of Wright's distribution for irreversible mutation (Wright, 1938, 1942, 1945). Note that this coincides with equation (8.73) except for the factor $2Nv_{gam}$, which represents the number of sites at which mutations occur in each generation distributed throughout the population.

A more exact and detailed study on the number of segregating sites was made by Watterson (1975) using the infinite site model (see also Ewens, 1979).

Although the infinite site model was originally formulated with the whole genome comprising all nucleotide sites in mind, the model may be applied to treat a gene locus or cistron (a segment of DNA coding for a protein), because each cistron consists of a large number, say 1000 of nucleotide sites. Li (1977b) investigated the distribution of nucleotide differences between two randomly chosen cistrons in a finite population applying the infinite site model to cistrons.

Ohta and Kimura (1971c) investigated linkage disequilibrium (i.e. non-random association of mutant forms) between two segregating nucleotide sites in a finite population using the infinite site model. Based on the results obtained, they pointed out that strong linkage disequilibrium must be very common between segregating nucleotide sites within a cistron.

8.7 Model of effectively neutral mutations

In our discussion on the relationship between the rate of evolution and selective constraint (see chapter 7), we often found it convenient to classify mutations into two distinct types, strictly neutral and definitely deleterious, and argued that the proportion of neutral mutations decreases with increasing functional constraint. In reality, however, there may be a continuum between these two types and the possibility cannot be excluded that the intermediate types between these two extremes are important. In other words, at the molecular level, a substantial proportion of new mutations could be very slightly deleterious as emphasized by Ohta (1973, 1974). Such mutations behave as if selectively neutral in a small population, but are unequivocally selected against in a large population.

Ohta (1977b) was the first to investigate the population genetical consequences of very slightly deleterious mutations by assuming that selection coefficients against the new mutations follow an exponential distribution. She has shown that under this assumption, the rate of evolution per generation in terms of mutant substitutions is inversely proportional to the effective population size of the species.

From the standpoint of the neutral theory, however, Ohta's model has a drawback in that it cannot accommodate enough mutations that behave effectively as neutral when the population size gets large. To overcome this difficulty, I proposed (Kimura, 1979) a model which assumes that the selection coefficients among mutants follow a Gamma (Γ) distribution and this was termed the model of effectively neutral mutations. This model is based on the idea that selective neutrality is the limit when the selective disadvantage becomes indefinitely small (Kimura and Ohta, 1974).

More specifically, the model assumes that the frequency distribution of the selective disadvantage (denoted by s') of mutants among different sites within a gene (cistron) follows the Γ distribution

$$f(s') = \alpha^\beta e^{-\alpha s'} s'^{\beta-1}/\Gamma(\beta), \tag{8.98}$$

where $\alpha = \beta/\bar{s}'$, in which $\bar{s}'$ is the mean selective disadvantage, and β is a parameter such that $0 < \beta \leqq 1$. If we measure the selective disadvantage in terms of the Malthusian parameter (Fisher, 1930b), s' has the range $(0, \infty)$. On the other hand, if we measure it, in terms of the conventional selection coefficient, the true range of s' is restricted to the interval (0, 1). However, because we assume that $\bar{s}'$ is small, with a typical value of 10^{-3}, $f(s')$ is negligible beyond $s' = 0.1$ so that we can take the entire positive axis as the range of integration without serious error.

In this formulation, we disregard beneficial mutants, and restrict our consideration only to deleterious and neutral mutations. Admittedly, this is an oversimplification, but as I shall show later, a model assuming that beneficial mutations also arise at a constant rate independent of environmental changes leads to unrealistic results.

Let us consider a diploid population of effective size N_e. We apply the infinite site model (see section 8.6) to a gene locus or cistron, and denote by v_e the effectively neutral mutation rate that is defined by the relationship

$$v_e = v \int_0^{1/(2N_e)} f(s')ds', \tag{8.99}$$

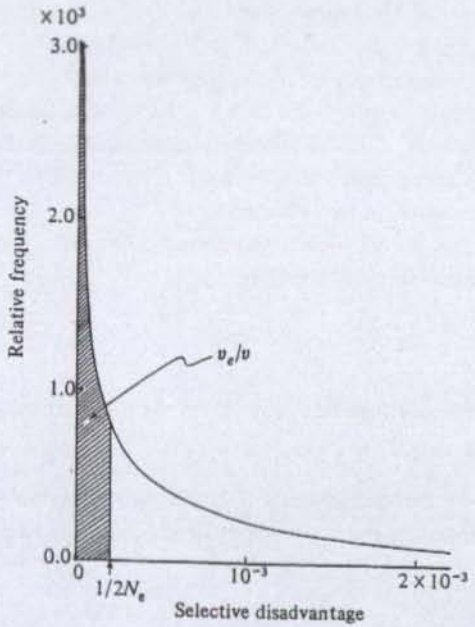
where v is the mutation rate per cistron. In other words, we call the

mutations effectively neutral if their selective disadvantage is less than $1/(2N_e)$. For $2N_e\bar{s}' \gg 1$, (8.99) is approximated by

$$v_e = \frac{v}{\Gamma(1+\beta)}\left(\frac{\beta}{2N_e\bar{s}'}\right)^{\beta}. \tag{8.100}$$

Fig. 8.8 illustrates the distribution $f(s')$ for the case $\beta = 0.5$ and $\bar{s}' = 10^{-3}$. In this figure, the shaded area represents the fraction of effectively neutral mutations (v_e/v) when the effective population size (N_e) is 2500. This fraction becomes smaller as the population size increases. Note that even if the frequency of strictly neutral mutations (for which $s' = 0$) is zero in the present model, a large fraction of mutations can be effectively neutral if β is small. Note also that $f(0) = \infty$ for $0 < \beta < 1$. We may regard β as representing the degree of physiological homeostasis, while $\bar{s}'$ represents the degree of functional constraint of the molecule. In the limiting situation $\beta \to 0$, all mutations become neutral. On the other hand, if $\beta = 1$, the model reduces to Ohta's model for which $v_e/v \approx 1/(2N_e\bar{s}')$ when $2N_e\bar{s}' \gg 1$.

Fig. 8.8. Frequency distribution of selection coefficients among mutants at different sites within a cistron (gene). The shaded area represents the fraction of effectively neutral mutations. Parameter values assumed for this case in the model of effectively neutral mutations are $\beta = 0.5$ and $\bar{s}' = 0.001$. From Kimura (1979).



The rate of evolution in terms of the number of mutant substitutions per generation, which we denote by k_g, is

$$k_g = \int_0^{\infty} 2Nvuf(s')\,ds', \tag{8.101}$$

where subscript g indicates that it refers to the rate per generation rather than per year In this formula, u is the probability that a mutant with selective disadvantage s' become eventually fixed in the population, and it is given by

$$u = \left[1 - e^{2s'(N_e/N)}\right] \Big/ \left(1 - e^{4N_e s'}\right) \tag{8.102}$$

(this formula follows from equation 3.11, if we put $p = 1/2N$ and $s = -s'$).

It can then be shown that, as a rough approximation,

$$k_g \approx v_e, \tag{8.103}$$

namely, the rate of evolution per generation is equal to the rate of effectively neutral mutations. This may be regarded as a generalization of the well-known relationship $k = v$ which is valid for strictly neutral mutations (see equation 3.20). In the case of $\beta = 0.5$ which is illustrated in Fig. 8.8, the rate of evolution per generation is inversely proportional to $\sqrt{N_e}$, provided that $N_e\bar{s}'$ is large as can be seen from (8.100).

Let $\bar{H}_n$ be the average number of heterozygous sites per cistron, so that

$$\bar{H}_n = \int_0^{\infty} H(1/2N)f(s')\,ds', \tag{8.104}$$

where $H(1/2N)$ is given by equation (8.90) in which we put $v_{gam} = v$ and $s = -s'$. For numerical calculation of $\bar{H}_n$, the following approximation formula is useful.

$$\bar{H}_n = 8N_e vR^{\beta}\left[\frac{(1+R)^{1-\beta} + (1.5+R)^{1-\beta} - R^{1-\beta}}{1-\beta} - (1+R)^{-\beta} - \frac{(1.5+R)^{2-\beta} - (0.5+R)^{2-\beta}}{(1-\beta)(2-\beta)}\right], \tag{8.105}$$

where $R = \beta/(4N_e\bar{s}')$ and $0 < \beta < 1$. In the limiting situation either for $\beta \to 0$ or $\bar{s}' \to 0$, this equation reduces to $\bar{H}_n = 4N_e v$, which agrees with equation (8.89) obtained in the previous section for strictly neutral mutations.

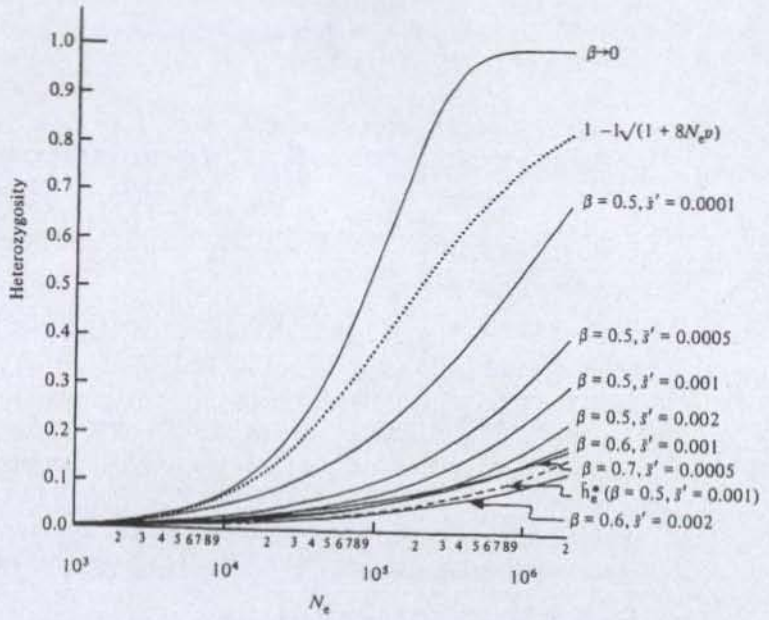
Let $\bar{h}_e$ be the expected heterozygosity of the gene under consideration. Then, assuming that different sites behave independently, we have

$$\bar{h}_e = 1 - e^{-\bar{H}_n}, \tag{8.106}$$

because this represents the probability that the gene is heterozygous at least in one of the sites. (This assumes independence of segregating sites within a cistron and therefore, strictly speaking, is not valid. For a small value of $\bar{H}_n$, however, it should give a satisfactory approximation.) In Fig. 8.9, the expected heterozygosity is shown as a function of N_e for various parameter values (β and $\bar{s}'$) assuming $v = 2 \times 10^{-6}$. Note that as compared with the situation in which all the mutations are neutral ($\beta \to 0$), a case such as $\beta = 0.5$ and $\bar{s}' = 0.001$ is interesting because of the slow rate at which heterozygosity increases as N_e gets large. In the same figure, the dotted line represents $\bar{h}_e = 1 - 1/(1 + 8N_e v)^{1/2}$, the heterozygosity expected under the stepwise mutation model (see section 8.5). Similarly, the broken line represents $\bar{h}_e^* = 1 - 1/(1 + 8N_e v_e)^{1/2}$ in which v_e is the effectively neutral mutation rate with $\beta = 0.5$ and $\bar{s}' = 0.001$. In all these cases the mutation rate $v = 2 \times 10^{-6}$ is assumed.

In the above analysis, we have assumed that all the mutations are either deleterious or at best, selectively neutral. To make our analysis complete, however, it is important to investigate how the evolutionary rate is

Fig. 8.9. The mean heterozygosity per locus as a function of the effective population size for various combinations of parameters, β and $\bar{s}'$, assuming the mutation rate per locus $v = 2 \times 10^{-6}$. From Kimura (1979).



influenced if a certain fraction of mutations are advantageous. So, let us assume that advantageous mutations occur at the rate v_a per generation and that the selection coefficient s for such mutations follows the Γ distribution with mean $\bar{s}$ and parameter γ;

$$f_a(s) = \alpha^\gamma e^{-\alpha s} s^{\gamma-1}/\Gamma(\gamma), \tag{8.107}$$

in which $\alpha = \gamma/\bar{s}$ and $\gamma > 0$.

Noting that the probability of ultimate fixation of a single mutant with selective advantage $s(>0)$ is $u = 2s(N_e/N)/(1 - e^{-4N_e s})$ (see equation 3.12), the rate of evolution due to advantageous mutations is

$$k_g = \int_0^\infty 2Nv_a u f_a(s) ds$$

$$= 4N_e v_a \bar{s} R^{\gamma+1} \sum_{j=0}^{\infty} (j+R)^{-\gamma-1}, \tag{8.108}$$

in which $R = \gamma/(4N_e\bar{s})$. This can be approximated by

$$k_g = v_a\{4N_e\bar{s} + [\gamma/(6N_e\bar{s} + \gamma)]^\gamma\}. \tag{8.109}$$

For $N_e\bar{s} \gg 1$, we have $k \approx 4N_e\bar{s}v_a$. This means that the rate of evolution can become enormously high in a very large population, k_g being directly proportional to N_e, contrary to actual observations.

Let us go back to the original assumption that the occurrence of favorable mutations can be neglected and that the frequency of disadvantageous mutations follows the Gamma distribution (8.98). With this model, we have shown that the rate of evolution per generation (k_g) decreases as N_e increases. Of particular interest is the case $\beta = 0.5$, which is an intermediate situation between $\beta = 0$ (i.e. all the mutations are strictly neutral) and $\beta = 1$ (Ohta's model of exponential distribution). This appears to be more realistic than either of these two extremes.

In this case of $\beta = 0.5$, as mentioned already, the rate of evolution is inversely proportional to $\sqrt{N_e}$, that is,

$$k_g \propto 1/\sqrt{N_e}, \tag{8.110}$$

if $4N_e\bar{s}' \gg 1$.

Thus the evolutionary rate per year which we denote by k_1 is

$$k_1 = k_g/g \propto 1/(g\sqrt{N_e}), \tag{8.111}$$

where g is the generation span in years. If, in addition, the generation span is inversely proportional to the square root of the effective population size among various organisms, i.e.

$$g \propto 1/\sqrt{N_e}, \tag{8.112}$$

we have, from (8.111), the relationship,

$$k_1 = \text{constant}, \tag{8.113}$$

namely, the evolutionary rate per year is constant, provided that the mutation rate per generation (v) is constant among different organisms.

As discussed extensively in chapter 4 (see section 4.2), there is strong evidence that for each protein the actual rate of evolution in terms of amino acid substitutions is approximately constant per year, but not per generation, among diverse lineages. This is particularly evident in the α hemoglobin among mammals.

One of the criticisms against the neutral theory has been that the observed constancy of the evolutionary rate per year rather than per generation contradicts the prediction of the neutral theory, because the mutation rate per generation but not per year appears to be roughly equal among animals whose generation spans are very different. For example, observed rates for lethals and visible mutations per locus per generation are roughly equal among *Drosophila*, the mouse and the human. Unless the mechanism of mutation differs fundamentally for neutral alleles from other type of mutations, we should expect that the frequency of neutral mutations is roughly constant per generation, and if we apply the formula $k = v$ (equation 3.20), the rate of evolution should be constant per generation, contrary to observations.

I must admit that this is a difficult problem for the neutral theory to cope with (and even more so for the 'selectionist' theory). The original explanation (see, for example Kimura and Ohta, 1971a) that the mutation rate for neutral alleles is constant per year could still be valid. If this is indeed so, we need to go more deeply to inquire why the occurrence of neutral and nearly neutral mutations should be proportional to years. This question is particularly pertinent in the face of the observations on pseudogenes (see section 7.4), for which previously deleterious mutations have changed into neutral ones.

According to Kondo (1977) there is little doubt that spontaneous mutations are mostly produced as copy errors during DNA replication. He tried to estimate per locus per replication rates of visible mutations in *Drosophila melanogaster*, the mouse and man, and obtained respectively $(2 \sim 50) \times 10^{-8}$, 4×10^{-8} and $(2 \sim 20) \times 10^{-8}$ for them. These values are obtained by dividing the mutation rates per locus per gamete with the effective number of germ cell generations, assumed to be respectively 11, 16 and 20 for *Drosophila*, mice and humans. Considering the very large difference in generation spans among these organisms, we must conclude

that the spontaneous mutation rates are comparable among them per generation but not per year.

The assumption that a species with a longer generation span tends to have a smaller effective population size appears to be realistic, because organisms with a large body size (such as the elephant) tend to have a longer generation time and at the same time have a smaller number of individuals per species as compared with those having smaller body size (such as the mouse). However, whether such a simple relationship as $g \propto 1/\sqrt{N_e}$ really holds (even approximately) remains to be investigated. As an example, let us consider the elephant vs. the mouse. The former has a generation span some 40 times as long as the latter, because the average generation time of the elephant is probably 20 years or so while that of the mouse is half a year. If the relationship $g \propto 1/\sqrt{N_e}$ applies to these two animals, the effective population size of the mouse is expected to be some 1600 times as large as that of the elephant, and this is a not wholly unrealistic proposition. Note that the effective size here refers to that of the whole species or subspecies rather than that of a local population or deme.

One exception to the rate constancy per year rule of protein evolution is the apparent slow molecular evolution in higher primates. This anomaly may be explained by assuming that in the line of higher primates, increase of g had not been accompanied by a corresponding decrease of N_e so that the product $g\sqrt{N_e}$ in equation (8.111) became significantly larger in them than in other mammals. This led to the slow down of the rate per year. Note, however, that, as mentioned in chapter 4, there is still disagreement as to whether the rate of molecular evolution has really slowed down, so the present discussion is only tentative.

Generally speaking, it is likely that the value of the parameter β is smaller in mammals than in insects, because of higher physiological homeostasis in the mammals. The possibility of more mutations being neutral in higher forms such as mammals with advanced homeostasis has been suggested by Kondo (1977). Low physiological homeostasis and frequent local extinction of colonies must be the main reason why the virtual heterozygosity (or 1 minus the sum of squares of allelic frequencies) does not get very high in organisms having immense apparent population sizes such as neotropical *Drosophila* (Ayala *et al.*, 1972) and *Escherichia coli* (Milkman, 1973; Selander and Levin, 1980). More will be discussed on this problem in the next chapter (see section 9.2).

The model of effectively neutral mutations (Kimura, 1979) discussed in this section represents my attempt to make the neutral theory more precise and realistic. The model assumes that molecular evolution and polymor-

phism are mainly caused by random drift of very slightly deleterious but effectively neutral mutations. In this respect, the theory resembles Ohta's theory of slightly deleterious mutations (Ohta, 1973, 1974, 1976b). But there are some important differences. Ohta (1976b) claims that, in very large populations, a stable mutation-selection balance will be realized with heterozygosity reaching the upper limit, while molecular evolution should have stopped or at least have slowed down. Then, fixation of mutants is mainly restricted to population bottlenecks at the time of speciation. On the other hand, I assume that, even in very large populations, alleles at intermediate frequencies, as often found in *Drosophila* species (see Selander, 1976, table II), represent effectively neutral mutations carried by random drift and that evolution by drift is unlikely to be stopped in these species.

Finally, there is one biological problem that we have to consider. Under the present model, effectively neutral, but, in fact, very slightly deleterious mutants accumulate continuously in every species. The selective disadvantage of such mutants (in terms of an individual's survival and reproduction – i.e. in Darwinian fitness) is likely to be of the order of 10^{-5} or less, but with 10^4 loci per genome coding for various proteins and each accumulating the mutants at the rate of 10^{-6} per generation, the rate of loss of fitness per generation may amount of 10^{-7} per generation. Whether such a small rate of deterioration in fitness constitutes a threat to the survival and welfare of the species (not to the individual) is a moot point, but this can easily be taken care of by adaptive gene substitutions that must occur from time to time, say once every few hundred generations (see also section 6.7, where random fixation of slightly deleterious mutants under stabilizing selection is discussed).

8.8 Genetic distance

For estimating the amount of evolutionary divergence between populations (species, races, etc.) based on gene frequency data, a suitable measure of the distance between populations is required. In their pioneering attempt to construct a phylogenetic tree of human races through data on blood-group frequencies, Cavalli-Sforza and Edwards (1967) used a measure of distance derived from 'square-root' transformation of gene frequencies. This measure was proposed earlier by Cavalli-Sforza and Conterio (1960) who analysed the distribution of blood group genes in the Parma river valley, and it is based on the following consideration. Let A_1, $A_2, \ldots, A_m$ be m alleles at a locus, and assume m-dimensional Cartesian co-ordinates with each axis representing the square root of one of the allelic

frequencies. Then the genetic composition of a population is located on a hypersphere with radius 1 (actually we need only the $1/2^m$th part of the space in which all the co-ordinates are non-negative). For two populations each with allele frequencies $(p_1, p_2, \ldots, p_m)$ and $p'_1, p'_2, \ldots, p'_m)$, the genetic distance between them is given by θ, where

$$\cos\theta = \sum_{i=1}^{m} \sqrt{p_i}\sqrt{p'_i}. \tag{8.114}$$

Geometrically, θ is the angle (in radians) made by the two vectors, $(\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_m})$ and $(\sqrt{p'_1}, \sqrt{p'_2}, \ldots, \sqrt{p'_m})$, which represent the two populations. One advantage of this distance is that if gene frequencies change solely by random drift, the expected value of the variance of change in θ per generation is a constant which depends on the effective population size but is independent of gene frequencies. Since $\theta = \pi/2$ corresponds to a complete gene substitution, it may be convenient to work in terms of $2\theta/\pi$, the unit distance being one gene substitution.

However, in applying the method of maximum likelihood, this measure turned out to be intractable (owing to the curved space and difficulties with the co-ordinate system), so Cavalli-Sforza and Edwards used a more convenient measure $(2\sqrt{2}/\pi)\sqrt{(1-\cos\theta)}$ which is the length of the chord joining the two points (populations). Such a measure is appropriate if divergence of gene frequencies between populations reflects the direct action of random drift over a relatively short period, and if the effect of mutation is small and negligible.

On the other hand, for the purpose of treating the long-term evolution in which the cumulative effect of mutations is significant, a different measure of genetic distance may be required. Probably the best one so far proposed, and certainly the most widely used for such a purpose is the distance measure proposed by Nei (1972). No doubt, its popularity stems from the simplicity of formulation, facility of application, and clarity of biological meaning as I shall explain below.

Consider two populations 1 and 2, and suppose that the frequency of allele A_i at a locus in populations 1 and 2 is respectively x_i and y_i $(i = 1, 2, \ldots)$. Let $j_1 = \Sigma x_i^2$, $j_2 = \Sigma y_i^2$ and $j_{12} = \Sigma x_i y_i$, where j_1 is the probability of identity of two randomly chosen homologous genes from population 1, j_2 is that of population 2, and j_{12} is the probability of identity of two randomly chosen homologous genes one each from populations 1 and 2. These are defined for each locus, so we take averages over all the relevant loci for which gene frequency data are available. Let $J_1 = \overline{j_1}$,

$J_2 = \overline{j_2}$ and $J_{12} = \overline{j_{12}}$ be such averages. Then, Nei's 'standard genetic distance' is given by

$$D = -\log_e \frac{J_{12}}{\sqrt{J_1}\sqrt{J_2}}. \tag{8.115}$$

This may also be expressed as $D = -\log_e I$, where

$$I = J_{12}/\sqrt{(J_1 J_2)}. \tag{8.116}$$

This quantity I is called the normalized identity of genes between two populations.

The standard genetic distance, D, is intended to measure the mean number of net codon differences, and under complete isolation it gives the accumulated number of gene substitutions per locus after the populations diverged from a common ancestral population. The fact that the rate of molecular evolution in terms of mutant substitutions is roughly constant per year (see chapter 4) makes this distance measure particularly useful for molecular evolutionary studies.

Let v be the mutation rate per locus for neutral alleles, and suppose that this rate is equal among loci. Then the expected value of identity in the t-th generation is

$$E(I_t) = e^{-2vt} I_0, \tag{8.117}$$

where I_0 is the identity probability (or homozygosity) in the ancestral population ($t = 0$). This formula follows immediately from the consideration that if we use the infinite allele model which assumes that every mutation leads to a new not pre-existing allele (see section 8.2), and if two populations are isolated from each other, the identity probability between two genes from the two populations decreases at the rate $2v$ per generation so that

$$E(I_t) = (1 - 2v)E(I_{t-1}) = (1 - 2v)^t I_0 \approx e^{-2vt} I_0. \tag{8.118}$$

This means that the expected value of the standard distance is

$$E(D) \approx -2vt, \tag{8.119}$$

if the population is in equilibrium so that $J_1 = J_2 = I_0$. On the other hand, if isolation is not complete as in many local populations of a species, progress of divergence is prevented by migration, and genetic distance cannot become very large (the identity coefficient between subpopulations under mutation and migration was studied by Nei and Feldman, 1972). Note that as a measure of the net codon difference, D must in general be estimated by using a large number of loci taken as a random sample from the whole genome, including monomorphic as well as polymorphic loci.

There is a possibility that the mutation rate (v) for neutral alleles differs from locus to locus, as we discussed in relation to the distribution of allelic frequency and the mean heterozygosity over many loci (see section 8.2). According to Nei *et al.* (1976a), the distribution of mutation rate per locus is well approximated by a gamma distribution, and the empirical data suggests that the coefficient of variation $\alpha = \bar{v}^2/V_v$ is about 1. Then, an appropriate formula for the distance is

$$D_v = (1 - I)/I. \tag{8.120}$$

For the rationale of this formula, readers may refer to Nei (1978b).

The single locus identity ($j_{12}/\sqrt{j_1 j_2}$) between two species varies a great deal among loci. It usually shows a U-shaped distribution in vertebrate species, and often takes a value of either 1 or 0 (Nei, 1978a). This is easily understandable by noting that alleles at a given locus are either fixed or lost most of the time in the course of evolution with random drift playing a predominant role in determining their frequencies (as claimed by the neutral theory). This means that a large number of loci must be studied in order to estimate the true genetic distance (D) between species. The situation is similar in estimating the heterozygosity per locus (H) of a species. Nei and Roychoudhury (1974) and Nei (1978a), investigated the variances of these estimates by decomposing each of these variances into two components; one due to variation from locus to locus (variance among loci) and the other due to sampling of genes within loci. It turns out that the inter-locus variance is much larger than the intra-locus variance, unless n (the number of individuals sampled) is extremely small. Thus, for estimating the average heterozygosity and the true genetic distance, a large number of loci rather than a large number of individuals per locus should be used (see also Li and Nei, 1975).

Nei (1975) computed the standard genetic distances between races, species, genera etc. in various organisms using published data. The results show that the distance is generally $0.00 \sim 0.05$ between races, $0.02 \sim 0.20$ between subspecies, $0.1 \sim 2.0$ between species, and more than 1 between genera.

Recently, Nei (1978b) reviewed his work on genetic distance as applied to the problems of human evolution. I shall mention some of the interesting findings reported by him. At the protein level the racial differences in man correspond to those between local races in other organisms. The mean number of codon differences between man and chimpanzee is about 1/19 of that between man and horse. Among the three major races, Caucasoid, Negroid and Mongoloid, the first divergence is estimated to have occurred

about 120 000 years ago between the Negroid group on the one hand and the joint Caucasoid and Mongoloid group on the other. Then the latter group split into Caucasoid and Mongoloid around 60 000 years ago. The genetic identity between man and chimpanzee corresponds to a divergence time of 4 ~ 6 million years if the assumption of constant rate of amino acid substitution is valid (see section 4.1 for the paleontological estimate of this divergence time, which may be older than this). As to a similar study with special reference to human populations in the Asian-Pacific area, the readers are invited to consult Omoto (1981, 1982).

9

Maintenance of genetic variability at the molecular level

9.1 The problem, the facts, and the neutralist solution

What is the mechanism by which genetic variability at the molecular level, such as manifested in the form of protein polymorphism, is maintained? This is the outstanding problem confronting present-day population genetics.

In the previous chapter (chapter 8) I have presented some models of molecular population genetics which were developed through the use of the diffusion equation method. It represents the neutralist program to treat this difficult but important problem. Furthermore, I have shown in quantitative terms, taking a hypothetical mammal as an example, how a realistic level and pattern of genetic variability can be predicted by assuming a set of parameter values derived from independent sources, such as the base substitution rate as estimated from the evolutionary study of globin pseudogenes, the fraction of neutral mutants among all mutations in the average protein as estimated by the frequency of rare electrophoretic variants, the average size of genes that code enzymes, and a reasonable guess about the effective population size in the long term evolution of a typical mammal. The central theme in this endeavor is that protein polymorphism and protein evolution are not two independent phenomena. Rather, protein polymorphism represents a transient phase of molecular evolution (Kimura and Ohta, 1971a). Admittedly, each estimate or informed guess for these parameter values cannot be very accurate at the moment. Nevertheless, the fact that a consistent and all-round picture can be drawn in this way represents a real strength of the neutral theory.

In this chapter, I shall discuss how the neutral theory can stand in the face of hard facts coming from numerous observations in nature and diverse experiments in the laboratories. I shall also discuss how the neutral theory

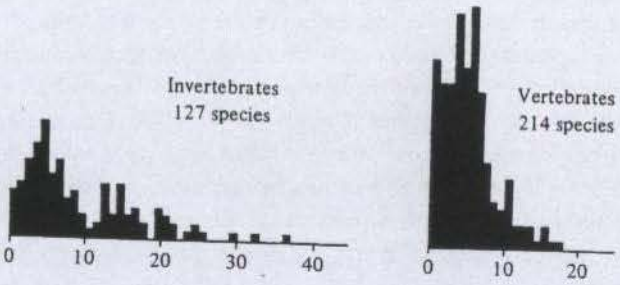
fares well as compared with rival 'selectionist' theories in explaining various observations on the genetic variability at the molecular level.

Before we go into detailed discussion on the nature of polymorphism at the molecular level, I would like to mention two types of observations that are fundamental for such a discussion. The first is concerned with the distribution of the average heterozygosities (per locus per individual) among diverse organisms, and the second is concerned with the distribution of allelic frequencies at a given enzyme locus among related but different species.

As to the distribution of the average heterozygosities among organisms, very extensive data are now available. Fig. 9.1 illustrates such distributions drawn separately for vertebrates (214 species) and invertebrates (128 species). Note that in this figure, the abscissa represents heterozygosity, or more precisely, one minus the sum of squares of allelic frequencies per locus (in %). In the terminology of this book, it represents virtual heterozygosity (or 'gene diversity' in the terminology of Nei, 1975) which coincides with heterozygosity in the ordinary sense for a random mating diploid species. A very similar set of histograms are also given in Fig. 2 of Nevo (1978) who compiled data for 93 invertebrate and 135 vertebrate species in which 14 and more loci have been studied. It is clear from these figures that the average heterozygosity seldom exceeds 30%, that is, $\bar{H}_e$ in natural populations is restricted to the range $0.0 \sim 0.3$.

This may be compared with the statement of Lewontin (1974) who claimed, based on the data then available (consisting of a dozen species, i.e. human, mouse, peromyscus, horseshoe crab and eight species of *Drosophila*), that the observed range of heterozygosities over all the species

Fig. 9.1. Distributions of the average heterozygosities per locus among invertebrate and vertebrate species (Nei and Graur, personal communication). Abscissa: heterozygosity, or more precisely 'virtual heterozygosity' (i.e. one minus the sum of squares of allelic frequencies) per locus (in %). Ordinate: number of species.



lies in the 'sensitive region', between 0.056 and 0.184, and that this range corresponds to values of $N_e v$, the product of the effective population size and mutation rate (see equation 8.15 in the previous chapter), between 0.015 and 0.057, so that $N_e v$ values are only within a factor of four of each other. Considering the possibility that mutation rate v is comparable among diverse organisms, he points out that this must imply an extraordinary invariance of population size of all multicellular organisms. He goes on to state that 'the patent absurdity of such a proposition is strong evidence against a neutralist explanation of observed heterozygosity'.

It is clear, however, thanks to much more extensive data now available, that the average heterozygosity can be extremely low or even zero for some organisms. Particularly, among mammals with a large body size, very low heterozygosity appears to be quite common (see Cameron and Vyse, 1978). I shall mention a few remarkable examples. Allendorf *et al.* (1979) investigated 13 enzyme loci for 52 individuals of the polar bear (*Thalarctos maritimus*) from Greenland and found no variation in the mobility of electrophoretic bands, that is, all the loci investigated were monomorphic ($\bar{H}_e = 0$). Also, no polymorphisms were found among 21 proteins encoded by 24 loci for the elephant seal (*Mirounga angustirostris*) investigated by Bonnell and Selander (1974), who ascribe the complete monomorphism to a consequence of fixation of alleles brought about by the decimation of this species by sealers in the last century. However, as I explained at the end of section 8.1, rapid decrease of the population size for a small number of generations is unlikely to cause significant reduction in heterozygosity. Therefore, the extremely low heterozygosity must be genuine in this species. From the standpoint of the neutral theory, we can explain such a low heterozygosity by assuming that large mammals in general have very small effective population sizes, so that $N_e v$ are very small for them. If we assume that the mutation rate for neutral electromorphs is $v_{0(E)} \approx 10^{-6}$ per generation (assumed to be several times as large as that of the hypothetical mammal with generation time of one year), the expected average heterozygosity becomes less than 0.004 if N_e is less than 1000.

Therefore, the only serious problem that remains to be explained in relation to Lewontin's criticism is why the average heterozygosity seldom exceeds 0.3, even for species which have immense apparent population sizes. According to Ayala (1972), the neotropical fruit fly, *D. willistoni* has an immense population size with geographical distribution extending over several million square kilometers, and the total number of individuals at a given time is doubtless much larger than 10^9. He criticized the neutral theory by saying that if we take $N_e = 10^9$ as a lower estimate and if we

assume (following Kimura and Ohta, 1971a) $v = 10^{-7}$ for the mutation rate, we get $4N_e v + 1 = 401$ as the effective number of alleles. In other words, the predicted heterozygosity, $\bar{H}_e = 4N_e v/(4N_e v + 1)$, should be practically 100%. On the other hand, the observed average heterozygosity of *D. willistoni* is only about 18%, contrary to the prediction from the neutral theory (see also Ayala *et al.*, 1972).

Ohta and Kimura (1973) countered this criticism by pointing out a real possibility that $N_e = 10^9$ is an overestimate, not as the number of breeding flies per generation, but as the number to be used in obtaining the expected heterozygosity. Not only is this number controlled by the minimum population size when population number fluctuates from generation to generation (equation 3.5 in chapter 3), but also it takes a length of time in the order of the reciprocal of the mutation rate (if $1/v \ll N_e$) or the population size (if $1/v \gg N_e$) for the equilibrium state with respect to heterozygosity to be established after the population expands rapidly from a much smaller population size with low heterozygosity (see equation 8.13). As suggested by Ohta and Kimura (1973), it is likely that the effect of small population number during the last continental glaciation (between some 30 000 and 10 000 years ago) still remains in the genetic composition of tropical fruit flies. According to Calder (1974) the large areas of tropical forests were cool and dry during the ice ages. Therefore, it is quite likely that the population sizes of neotropical fruit flies then must have been very much smaller. This means that the equilibrium heterozygosity as expected from the present flourishing condition has not yet been reached. It is possible that relative abundance of various species of *Drosophila* changes rather rapidly (in the time scale pertinent to a discussion of evolutionary change), so that correlation between heterozygosity and the contemporary population size can be very low.

Furthermore, as pointed out by Maruyama and Kimura (1980), local extinction and recolonization of subpopulations must be rather common in *Drosophila* species and this will reduce their effective population sizes drastically (see (v) in section 3.3). More specifically, if we assume in equation (3.10) that the rate of local extinction (λ) of colonies is much higher than the rate (m) of the equal exchange of individuals between colonies, the effective population size (N_e) of the species becomes much smaller than the product of the number of colonies (n) and the effective size ($\tilde{N}$) of the individual colony. In fact if m/λ is small, $N_e \approx n\tilde{N}m/\lambda$ from equation (3.10). In *D. willistoni*, it is possible that the appropriate effective size N_e is two orders of magnitude less than the lower estimate of N_e by Ayala. An additional point to be made regarding Ayala's criticism is that $v = 10^{-7}$ of Kimura and Ohta (1971a) refers to the neutral mutation rate per year but not per generation. For the fruit flies that breed all the year round in tropical forest, there may be some 10 generations per year, so that an appropriate mutation rate for neutral electromorphs is $v_{0(E)} = 10^{-8}$ if we accept the neutralist paradigm (but, see section 8.7 for deeper discussion on the relationship between mutation rate for the 'effectively neutral' alleles and N_e). Then, with $N_e = 10^7$ and $v_{0(E)} = 10^{-8}$, we get $\bar{H}_e = 0.29$ which is still larger than the observed value (0.177), but close enough to make Ayala's criticism lose much of its force.

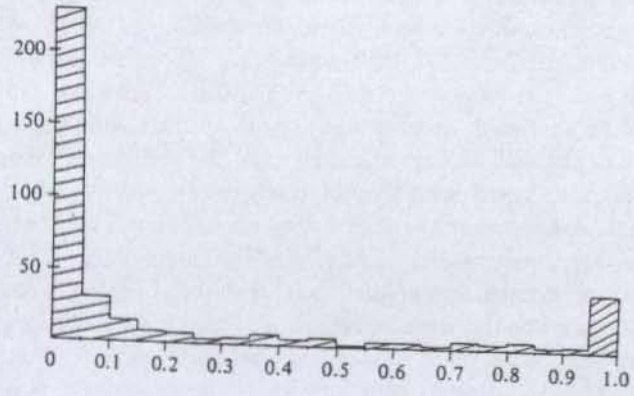
That the effective size of many *Drosophila* species must have been rather small in the course of evolution can also be inferred from the fact that they are usually separated by a number of chromosomal changes (see, for example, Dobzhansky, 1970). Since chromosome rearrangements such as translocations and inversions are generally deleterious when heterozygous but have normal fitness when homozygous, the situation is formally equivalent to negative overdominance, and a small effective population size is required for their fixation in evolution. Lande (1979) presented an excellent mathematical treatment of the process involved. He noted that species are usually composed of numerous demes which fluctuate in size and undergo local extinction and recolonization. He computed the probability of fixation (denoted by U) of a newly arisen chromosome mutant within a deme and noted that if R is the fixation rate, this is given by $R = 2N_a \mu U$, where μ is the rate of appearance of chromosome mutants per gamete and N_a is the actual number of breeding individuals per deme. Then, he pointed out that if n is the number of semi-isolated demes that make up the species, the probability that a given deme (or its descendants) will spread throughout the species is $1/n$, so that the number of demes cancels and fixation rate for the species as a whole is the same as the fixation rate for a single deme, R.

It is likely that, in many organisms with very high reproductive rate, random local extinction and recolonization predominate over interdeme migration, and as mentioned above, this will drastically reduce the effective population size of the species in the long-term evolution. Slatkin (1977) made a pioneering study of the effect of local extinction and recolonization on genetic variability of the species. This was followed by a more detailed study by Maruyama and Kimura (1980) which I mentioned already. I believe that these studies must be taken into account, in addition to Nei *et al.*'s (1975) work on population bottlenecks, when we discuss the genetic variability of organisms with immense apparent population sizes. There are other causes (such as inequality of the number of breeding males and

females) which reduce the effective population size. Thus in general, the pertinent value of N_e to be used in our formula $\bar{H}_e = 4N_e v/(4N_e v + 1)$ for computing the average heterozygosity must be much smaller than the number of breeding individuals at a given moment. As noted by Nei (1980), whenever a rough estimate of effective population size is obtainable, the observed heterozygosity is almost always lower than the expected, but, such a discrepancy does not pose any serious problem to the neutral theory. In other words, there is no difficulty for the neutral theory to explain 'an enormous amount' of genic variability, as often referred to. In fact, the amount as predicted from the neutral mutations is almost too much, and no additional mechanisms seem to be required. In my opinion this poses some difficulty to the selectionist theories assuming various causes (such as 'marginal overdominance' due to genotype–environment interaction in heterogeneous environment and 'frequency-dependent selection' that favors less frequent alleles) which positively help to maintain the allelic variants in the population; these will certainly inflate the amount of heterozygosity still more unless mutational input is neglected.

Next, let us examine briefly the second type of observation, that is, the distribution of allelic frequencies at a specific enzyme locus among different species. For this purpose observations from *Drosophila* and other organisms are available. Fig. 9.2 which is due to Gojobori (1979) illustrates distribution of allelic frequencies at xanthine dehydrogenase (*Xdh*) locus (or more precisely, *Xdh* locus group, because exact genetic identity of this locus among remotely related organisms is not assured). The distribution is U-

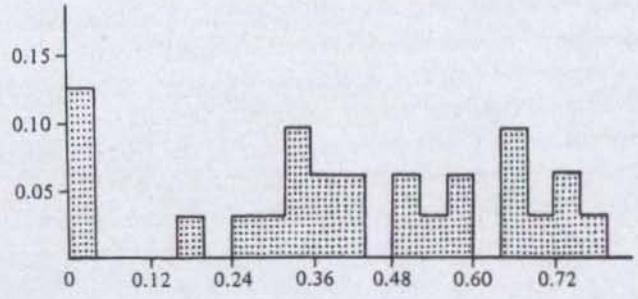
Fig. 9.2. Distribution of allelic frequencies at the *Xdh* locus (or more precisely, locus group) compiled by Gojobori (1979) using data from various organisms. Abscissa: allele frequency. Ordinate: number of alleles.



shaped with frequencies extending from very low to very high frequencies. Such U-shaped distributions are found in almost all other enzyme loci. This is consistent with the fact that distribution of allelic frequencies among different loci within a single species is also U-shaped, similar to Fig. 8.3 which we showed in the previous chapter for the marine flatfish. Usually, a good fit to such a distribution can be achieved (save for the excess of rare alleles) by using equation (8.17) in which $4N_e v$ is estimated by the observed average heterozygosity. This constitutes supporting evidence for the neutral theory. From the neutralist viewpoint, a particular allele in a given species has simply been brought to that frequency by random drift but not by adaptation to the living condition of that species. In other words, it is there by 'chance' but not by 'necessity'. Therefore, as time goes on the frequency will drift away with a calculable speed. That a reasonably good fit can also be obtained with respect to the distribution of genetic distance between different species as we shall discuss later (see section 9.3 below) also supports this viewpoint.

Similarly to the distribution of allelic frequencies, we can examine the distribution of heterozygosity at a given locus among related but different species. Fig. 9.3 shows such a distribution at *Xdh* locus among *Drosophila* species. Note that heterozygosity varies from almost zero to nearly 80%. A similar distribution is presented by Selander (1976) in his table II for ten proteins in 28 *Drosophila* species. As to *Xdh*, his table indicates that about 32% of species are monomorphic while heterozygosity is higher than 0.60 in 4% of species, with the average value of 27.5% (the slight difference between the Gojobori and Selander data is probably due to sampling of species). The situation is somewhat similar for other enzyme loci. The fact that at the same (homologous) loci heterozygosity can differ widely (ranging from monomorphism to a very high polymorphism) even between different

Fig. 9.3. Distribution of heterozygosity at *Xdh* locus among *Drosophila* species (redrawn from Gojobori, 1979). The mean is $\bar{H}_e = 0.345$. Abscissa: heterozygosity. Ordinate: relative frequency.



species within genus *Drosophila* suggests that a mere mutation–selection balance, as considered by King and Ohta (1975), can not be the main cause of enzyme polymorphism and that a much more dynamic process must be involved.

9.2 Experimental check on selective neutrality

As a scientific hypothesis, the neutral theory of protein polymorphism is subject to experimental tests. It should be possible to test selective neutrality of isozyme alleles involved, particularly using *Drosophila*, through such procedures as setting up a cage population containing two or more isozyme alleles in various frequencies, or using controlled environmental conditions to measure fitness components of various genotypes. Thus, one should be able to settle the issue of whether protein polymorphisms are selectively neutral or actively maintained by balancing selection. The neutralist side appears to be particularly vulnerable to such tests.

In fact, numerous studies have been published on the subject, and papers reporting the existence of balancing and other types of selection (thereby intending to refute selective neutrality) have appeared, and are still appearing frequently. Yet, it has turned out that the resolution of the issue through such a route is much more difficult than originally thought. There are at least two main reasons why this is so. Firstly, it is extremely difficult to make sure that the observed selective differences really come from polymorphic alleles themselves rather than from the effect of surrounding loci that are held in linkage disequilibrium with the locus in question. Secondly, there is no assurance that the experimental conditions under which the test is made really reflect the prevailing natural conditions of the species.

Because of these difficulties (particularly the first), most of the results have been regarded as inconclusive. Critical appraisal of the experimental results in the laboratory pertaining to the issue have been made by Lewontin (1974) and Wright (1978). Lewontin (1974, pp. 250–60) explains eloquently how easily a false pattern suggesting overdominance and other selection can arise in laboratory experiments and how difficult it is to detect weak selection even with an intensity of a few per cent. Similarly, Wright (1978, see p. 298), after carefully evaluating various reports on selection experiments, came to the following conclusions. If the experiments are based on a few separate stocks, strong selective differences between allozymes are usually found, and manifested in the form of rarity advantage, where the term allozymes refer to different enzyme forms produced by

different alleles. (Strictly speaking, this means genetically determined 'isozymes', or multiple molecular forms of enzymes. However, the two terms 'isozymes' and 'allozymes' are often used interchangeably in population genetics, so I shall follow this practice.) However, further study has usually indicated that such differences can be separated from the allozymes in question, suggesting that they depend on linked genes. In carefully designed experiments in which genetic backgrounds of the allozymes are well homogenized using a large number of independent sources, little or no selective difference has been detected. He states that 'these results do not, however, demonstrate complete neutrality of the tested allozymes, but merely that differential selection is usually weak (of the order of 10^{-3} or less)'.

Alcohol dehydrogenase

Among protein polymorphisms so far investigated, evidence for selection appears to be strongest for alcohol dehydrogenase (*Adh*) alleles. In fact, polymorphism at this locus in *Drosophila melanogaster* has been studied intensively by Clarke and his associates to prove that selection is acting directly on this locus rather than indirectly through other loci held in linkage disequilibrium with it (see, for review, Clarke, 1975 and 1979). The strategy adopted by him for demonstrating the direct action of natural selection is as follows. First, through biochemical and physiological study of enzymes, we postulate (and, if possible, identify) the selective agent involved and relate the agent and gene products. Then, using such knowledge, we set up experimental populations and manipulate gene frequencies to see if predictable results can be obtained. We also examine distributions of gene frequencies and ecological factors to see if a consistent explanation can be given for the observed pattern of gene frequencies in natural populations. In other words, correlation between gene or genotype frequencies and particular environmental factors is actively sought. According to Clarke (1975) this strategy is firmly rooted on the principles of ecological genetics, and he stated, on an optimistic note, that if this strategy could be applied to demonstrate natural selection at only seven or eight polymorphic loci in one organism, then 'the arguments between "neutralists" and "selectionists" would finally be resolved'.

In *D. melanogaster*, two polymorphic alleles denoted by *F* (meaning 'fast' in electrophoretic mobility) and *S* ('slow') commonly occur at the *Adh* locus (these are often denoted as Adh^F and Adh^S). According to Clarke (1975), the *F* enzyme is about twice as active as the *S* enzyme when ethanol is used as a substrate. On the other hand, *F* enzyme is less stable to heat. The main

selective agent was readily discovered to be alcohols, and Clarke (1975) assumed that *Adh* plays the role of a detoxifing enzyme for *D. melanogaster* which lives naturally in rotting fruit and which 'exists in a virtual bath of ethanol and other alcohols'. This led to the prediction that a high concentration of ethanol in the medium should favor the *F* allele, and this prediction was quickly confirmed in his laboratory (see Morgan, 1976 who also found that selection involved was frequency-dependent). Supporting evidence also came from Briscoe *et al.* (1975) who found that *D. melanogaster* flies caught in a Spanish wine cellar contain Adh^F in higher frequency than those caught at neighboring rubbish heaps. In addition, these authors investigated experimentally if mortality differs among genotypes on medium containing ethanol (12.5%) and found that Adh^S/Adh^S homozygote is significantly less resistant to alcohol than the heterozygote or Adh^F/Adh^F homozygote. Thompson and Kaiser (1977) used a *S* allele called $Adh^S(ve)$ which has only half the enzyme activity of the normal *S* allele, and they investigated relative egg to adult viability together with Adh^S and Adh^F under conditions with or without ethanol and *N*-butanol. They observed a positive correlation between survival in alcohol-supplemented media and enzyme activity.

At the time when these papers were published evidence for the existence of direct selection at the *Adh* locus appeared to be strong. However, later, more careful studies have given results that are either contradictory to or fundamentally different from these papers. Extensive studies by Australian geneticists have revealed that ethanol tolerance of *Drosophila melanogaster* is polygenically determined rather than by alleles at the *Adh* locus, and that they could not detect differences in the frequencies of *F* between the populations inside and outside of Australian winery cellars (McKenzie and McKechnie, 1978; Oakeshott and Gibson, 1981). Gibson *et al.* (1979) carried out selection experiments for increased tolerance to ethanol in two populations of *D. melanogaster* starting from the condition that the frequency of Adh^F was 0.5. They obtained the results showing that successful selection for increased ethanol tolerance does not necessarily depend on either an increase in the frequency of the *F* allele or on any increase in the level of alcohol dehydrogenase activity. The authors note that these results cast doubts on the validity of the claim of Clarke (1975) and others that differences of *Adh* activity between *F* and *S* alleles in *D. melanogaster* have adaptive significance in natural populations to alcohol tolerance, but support the conclusions of McKenzie and McKechnie (1978) that the adaptation to the alcohol concentrations in the environment of a winery cellar is independent of the *Adh* system.

More recently, Oakeshott and Gibson (1981) reported the results from five experiments that they performed to settle the issue. It turned out that none of the results support the claim that environmental ethanol specifically selects for the high activity *F/F* and *F/S* flies and against the lower activity *S/S* flies. On the contrary, in their experiments, *S/S* flies showed higher survival percentage under exposure to ethanol and there was indication that the advantage was dependent upon their low (rather than high) *Adh* activities. This may be explained by noting that acetaldehyde produced by *Adh* acting on ethanol is highly toxic to the organisms and therefore high activity of *Adh* is disadvantageous rather than advantageous. Although this appears strained, such an explanation was indeed adopted to explain the higher sensitivity of Japanese to alcohol as compared with Caucasoids (see Stamatoyannopoulos *et al.*, 1975). Yet, Daly and Clarke (1981) recently reported that the proportion of surviving *FF* flies, relative to *SS*, was higher on 2% ethanol than on distilled water and claimed that their experimental results support the view that selection acts directly on *Adh* locus, in agreement with Clarke (1975). In the light of the above experimental results by Australian workers, however, Clarke's (1975) claim is much in doubt.

Frequency-dependent selection and other forms of natural selection

Kojima and Yarbrough (1967) appear to be the first who proposed frequency-dependent selection with approximate selective neutrality at equilibrium as a possible general mechanism responsible for a large amount of enzyme polymorphisms observed in natural populations. They reported experimental results showing that very strong rarity advantage exists between *F* and *S* alleles at the esterase 6 (*Est*-6) locus in *Drosophila melanogaster*. The method used by them to estimate the relative fitnesses of the three genotypes was to take the ratio of the observed proportion of each genotype at the adult stage to the proportion expected from gene frequencies in the previous generation. This method obviously cannot give correct estimate of fitnesses if fertility differences also occur among genotypes. In this connection, it is known (see Prout, 1965; Christiansen *et al.*, 1977) that if post-observational selection exists, the estimated fitness will give a false impression of frequency-dependent selection even though the true fitnesses are constant. Furthermore, the false selection most often takes the form of rare genotype advantage. Yarbrough and Kojima (1967) also performed, to strengthen their case for frequency-dependent selection, cage experiments involving *F* and *S* alleles at *Est*-6 locus over 30

generations (replicated with two different media), starting from two different initial frequencies. They obtained the result that frequencies converged rapidly to the neighborhood of the frequencies of the base population (30% *F* and 70% *S*) in three of four cages, although in the remaining cage population that started from 10% *F* and 90% *S* no consistent change occurred from the initial set of frequencies.

Since then, a large number of experimental studies have been done to examine if frequency-dependent selection of the type postulated by Kojima and Yarbrough exist at other polymorphic enzyme loci. Kojima and Tobari (1969) investigated *F* and *S* alleles at *Adh* locus of *D. melanogaster* and found evidence for frequency-dependent selection with rarity advantage as in *Est*-6 locus. To clarify the mechanism which underlies frequency-dependent selection, Huang *et al.* (1971) investigated the *Est*-6 locus using a 'conditioned' media method. This is a method of preconditioning the medium by allowing larvae of a given genotype to develop before introducing larvae of the same or other genotype whose survival was to be measured. They found that such a conditioning reduces the relative survival of larvae of the same genotype, suggesting that this was caused by either depletion of nutrients needed by the same genotype or production of substance harmful to its own genotype. However, no such effect due to conditioning the medium was found in the later, more carefully designed experiments by Dolan and Robertson (1975) in relation to their examination of frequency-dependent selection at *Est*-6 and *Adh* loci in *Drosophila*. On the other hand, Morgan (1976) reported experimental results clearly showing frequency-dependent selection operating at both of these loci.

To settle the issue, Yoshimaru and Mukai (1979) carried out carefully designed, large-scale experiments (with respect to Adh^F and Adh^S in *D. melanogaster*). Their experiments consist of two parts. In the first, 563 second chromosomes were extracted from flies sampled in the Ogasawara natural population, and later, 124 chromosome lines having the *F* allele and 120 lines having the *S* allele were established, both consisting of inversion-free chromosomes, randomly selected for study. Using these lines, competition experiments were performed between *F*/*F* and *S*/*S* genotypes at five gene frequency levels (frequencies of $F = 0.82, 0.67, 0.5, 0.33$ and 0.17). First-instar larvae of these two homozygous types in various ratios were placed in a vial, and the frequency of adult flies of each genotype was determined. The relative viabilities of these two types did not change with the change in the frequencies of genes. The average viability of *S*/*S* relative to *F*/*F* was 1.0825 ± 0.0144 in the 9-*ml* culture media and 1.0879 ± 0.0127 in the 1.5-*ml* culture media. These results completely contradict those of Morgan (1976). Also in the second experiment, which made use of the results of viability test using the *Cy* method (with variation of gene frequencies among the vials in the viability test of individuals made by randomly combining two second chromosomes), no evidence of frequency dependent selection was obtained. This is in accord with the earlier work of Yamazaki (1971) who conducted a comprehensive series of experiments with respect to sex-linked *Est*-5 polymorphism (involving *S* and *F* alleles) in *D. pseudoobscura*. He could not detect differences in components of fitness among genotypes; no evidence of frequency-dependent selection was obtained for that matter.

Of course, negative results do not prove that the alleles involved are completely neutral with respect to natural selection. It is possible that very weak selection may still be acting. Usually, the resolving power of experiments in detecting natural selection is limited, and ordinarily it is not possible to measure selection intensity less than 0.01. However, Mukai *et al.* (1980), utilizing the inversion-carrying chromosome *In*(2*L*)*t* in an isolated *Drosophila* population could show that the intensity of selection per enzyme locus must be of the order of 10^{-4} or less if it exists. This finding and the argument that led to this conclusion were as follows. In the Ishigakijima population, the frequency of the second chromosome inversion *In*(2*L*)*t* is about 60%. Three isozyme loci are nearly monomorphic in the inversion-carrying chromosomes but are highly polymorphic in other (standard) chromosomes. The inversion carrying chromosomes were introduced into the population about 1000 generations ago, as determined by the known rate of crossing over and the slight polymorphism. This is long enough for lethal and mildly deleterious mutants to reach equilibrium frequencies, but not long enough for very weakly selected mutants. Actually, the distribution of homozygous viabilities in the inverted chromosome is indistinguishable from that of the standard chromosomes. It is inferred that the *In*(2*L*)*t* chromosome was monomorphic at the time it entered the population about 1000 generations ago. This is not long enough for polymorphic equilibrium to be established. If such loci are maintained by heterotic selection there should be clear difference in viability between homozygotes and heterozygotes for the standard chromosome which is highly polymorphic with respect to enzyme loci, but not within nearly monomorphic inversion-carrying chromosome. In reality, no such difference was found within the limit of about 1%. Assuming some 350 polymorphic loci with 20% heterozygosity each, the estimated selection coefficient against homozygotes per locus turned out to be -0.00013.

Incidentally, the population cage method has frequently been used to

demonstrate selection at enzyme loci. Claims for the existence of balancing selection (such as minority advantage and overdominance) have been made by many authors based on this method (e.g. Ayala and Anderson, 1973; Fontdevila *et al.*, 1975; van Delden, Boerema and Kamping, 1978). A typical pattern presented in these experiments to demonstrate selection is that by starting from a certain set of allelic frequencies that differ widely from those in natural populations, the cage population rapidly converges to an equilibrium state in which allelic frequencies are roughly equal to those in natural populations. Mukai and Yamazaki (1980) attempted to estimate the selection coefficients involved by examining the published experimental data (by authors mentioned above). It turned out that the estimated selection coefficients are generally rather large, of the order of 0.1 and sometimes even $0.2 \sim 0.4$. They pointed out that selection coefficients of 10% or more are much too large for the effect of normal, polymorphic alleles at a single isozyme locus, and concluded that these experimental results do not reflect the effects of the single loci in question.

As to the supposed prevalence of frequency-dependent selection, a possibility is often suggested that usually each genotype is adapted to a slightly different environment or utilizes a food resource within a niche somewhat in a different way as compared with other genotypes. Then, stable polymorphism with rarity advantage would result if a suitable set of additional conditions are satisfied. In fact, a number of models of maintenance of polymorphism of this sort have been proposed. However, Maynard Smith and Hoekstra (1980) showed through mathematical analysis that, in order for such models to work, the relative niche sizes must lie within a narrow range.

Clarke and Allendorf (1979) proposed a model which is based on the idea that relative velocities of the reactions catalysed by two allozymes vary with substrate concentration. Under a set of assumptions (such as fitness is a monotonically increasing function of reaction velocity, etc.), they showed that frequency-dependent selection will result and suggested that this may be responsible for maintaining a significant proportion of enzyme polymorphisms (see also Clarke, 1979). On the other hand, Maynard Smith and Hoekstra (1980) showed that Clarke and Allendorf's model has the drawback that polymorphism can be maintained only for a narrow range of environmental conditions, and they pointed out that because of the lack of robustness, it is unlikely to be of wide relevance. They also pointed out that generally speaking, the models assuming heterogeneous environment for the maintenance of polymorphism require that, unless selective advantages per locus are large, the fitness must nicely be adjusted to the niche sizes,

and that this fact casts serious doubts on the plausibility of this type of mechanism as an important cause of widespread polymorphism.

I would like to add one more point. It is often stated that frequency-dependent selection with minority advantage is not accompanied by genetic load. However, this is not necessarily true. As shown in section 6.4, only when a certain additional condition is satisfied (see equation 6.27 and discussion on it) do the three genotypes at equilibrium become selectively equivalent leading to no genetic load. Otherwise, genetic load is unavoidable, and indeed always occurs in a finite population.

Haploid organisms

Overdominance or heterozygote advantage has been one of the most popular mechanisms assumed by 'selectionists' to explain widespread protein polymorphism. Therefore, studies of genic variability in haploid organisms are of special relevance to the neutralist-selectionist controversy: in haploid organisms, heterozygotes do not exist and therefore overdominance can not work to maintain genetic variability. Can we find protein polymorphism in haploid organisms?

A pioneering investigation along this line was carried out by Milkman (1973) using *Escherichia coli*. He studied electrophoretic variation at five enzyme loci (ADH, AP, G-6-PDH, 6-PGDH and MDH) in 829 *E. coli* clones from 156 samples from diverse natural sources, and found that the amount of genetic variation is not very different from that in many diploid species. Since *E. coli* is a haploid organism, heterozygotes do not exist, and the term heterozygosity is not relevant in the ordinary sense. However, we can compute 'virtual heterozygosity' at each locus as one minus the sum of squares of allelic frequencies. This is the proportion of heterozygotes among imaginary zygotes that would be produced by randomly combining two haploid individuals. The average value of virtual heterozygosity for five loci as computed from Milkman's (1973) Table 1 turns out to be $\bar{H}_e = 0.242$ or about 24%. This value is slightly higher than the average heterozygosity of a typical diploid organism but not very much. This observation was a fatal blow to the notion that overdominance is the universal cause of genic polymorphism. In fact, from this time on, enthusiasm toward overdominance has gradually subsided among selectionists, and this has been replaced by the popularity of frequency-dependent selection.

While denying heterosis as the universal cause of genic polymorphism, Milkman (1973) also claimed that his observation was inconsistent with the neutral hypothesis. His arguments were based on the relationship between the predicted and the observed values of $\bar{n}_e - 1$, where $\bar{n}_e$ is the mean

effective number of alleles per locus (see equation 8.29). According to him (see also Milkman, 1976), *E. coli* has been at a population size well over 10^{10} for at least 4×10^{10} generations or 4×10^7 years, as inferred from the fact that it is found in most mammals all over the world. Since the mutation rate v for the neutral alleles could hardly be less than 10^{-8}, the effective number of alleles minus one ($\bar{n}_e - 1$), which is equal to $4N_e v$ if the neutral theory is valid, must be about 400. On the other hand, the observed values of $n_e - 1$ at five loci are between 0 and 1. Actually, the observed values of n_e, or the reciprocal of sum of squares of allelic frequencies, are listed in Table 1 of Milkman (1973), and the average over five loci turns out to be 1.38. This means that the observed value of $\bar{n}_e - 1$ is only 0.38.

Recently, genetic diversity and structure of *E. coli* populations were re-examined by Selander and Levin (1980), and they brought out a new picture which is much less contradictory to the neutral theory. They investigated 20 enzymes from 109 clones of *E. coli* from various natural sources and obtained the average virtual heterozygosity of 0.4718 which is approximately twice as high as the value estimated by Milkman (1973). A much more important finding (and indeed, a remarkable one) is that the number of distinctive genotypes appear to be rather limited. They found only 98 distinctive electromorph types among the 109 clones examined. In fact, identical pairs of clones were isolated from hosts that were separated by large geographic distances and were recovered at different times. A remarkable example of this is that one clone from an infant in a Massachusetts hospital nursery was electrophoretically indistinguishable from the common *E. coli* K-12 stock (possibility of contamination here was ruled out). Furthermore, several cases were found where a pair of clones differ at only one or two loci among 20 loci examined. They pointed out that these observations can most easily be explained by assuming that recombination is extremely rare in natural populations of *E. coli*, contrary to the earlier claim of Milkman (1973) who thought that recombination takes place regularly within hosts. Actually, if alleles at different loci recombine at random, the number of distinctive combinations becomes astronomical, and according to the calculation of Selander and Levin (1980), even the combination with the highest expectation would occur at the frequency of only 5.35×10^{-5}: there would be very little possibility of recovering clones with identical or similar electromorphs, contrary to actual observations. Because of very rare occurrence of recombination, any combination of genes at different loci would be broken up at exceedingly low rates, and clonal and 'periodic selection' processes would play the dominant role in determining the amount of genetic variability. Here, 'periodic selection' refers to the phenomenon in which a mutant clone with a high fitness appears from time to time and this is followed by rapid expansion of this clone to replace the pre-existing clone.

In this connection, Maruyama and Kimura's (1980) study on the effective population size assuming local extinction and recolonization is pertinent (see section 3.3). They showed that if a species (population) consists of n haploid lines (subpopulations) which reproduce asexually and each of which is subject to random extinction and subsequent replacement, the average virtual heterozygosity at equilibrium is $\bar{H}_e = 2N_e v/(1 + 2N_e v)$, where N_e is given by equation (3.9). Using this model we can show that if local extinction and recolonization occur frequently, not only is the effective population size of the species much reduced but also divergence of subpopulations is largely prevented. As an example, let us suppose that a haploid species consists of 100 000 lines ($n = 10^5$), and that each line starts from a single individual and although it may grow into a line comprising an immense number of individuals it becomes extinct on the average in 1000 generations ($\lambda = 10^{-3}$) so that its harmonic mean size is only $\tilde{N} = 100$. Then, assuming the mutation rate per generation of $v = 10^{-8}$, we obtain $N_e \approx 5 \times 10^7$ and $\bar{H}_e \approx 0.5$. This means that the average virtual heterozygosity is only 50% similar to that found in *E. coli*. Yet at any moment, this species may comprise an immense number, say 10^{20}, of individuals. It can also be shown that the genetic variability is almost entirely due to line differences.

It is now clear, from Selander and Levin's (1980) observation together with Maruyama and Kimura's (1980) mathematical analysis, that the genetic structure of *E. coli* population can most easily be understood by regarding it as a collection of asexual lines. Random sampling of such lines occurs frequently through periodic selection. We can think of each line being derived from a single individual (bacterium) in which a mutation that endows higher competitive ability happens to occur followed by rapid expansion of its progeny by asexual means to form countless individuals. Eventually such a line may become extinct to be replaced by a new 'periodic' line. We may regard intestines of mammals as a sort of chemostat in which periodic selection goes on continuously. It is likely that the effective size of *E. coli* is not really very large. Therefore Milkman's criticism against the neutral theory is not warranted. In this respect, Nei (1976) had a remarkable insight when he suggested that the effective size of *E. coli* in the long evolutionary history must be much smaller than 10^{10}, contrary to

Milkman's claim that *E. coli* has been at a population size well over 10^{10} for at least 4×10^{10} generations. 'This is because an *E. coli* colony rapidly grows under certain circumstances, while in other circumstances it easily becomes extinct' (Nei, 1976).

Recently, using the chemostat, Dykhuizen and Hartl (1980) performed elegant experiments to test selective neutrality of 6PGD (6-phosphogluconate dehydrogenase) allozymes which are due to alleles at the *gnd* locus of *E. coli*. They investigated four naturally occurring alleles (i.e. W^+, *F2*, *S4*, *S8*) and used transduction to transfer these alleles into the genetic background of strain *E. coli* K-12. They took great care to produce isogenic strains carrying these alleles. The limit of detectability of selection with their procedure of investigating competition in the chemostat is a selection coefficient of 0.5% per hour or about 1% per generation (the limit is set by sampling error). In their experiments they controlled gluconate in the medium based on the idea that gluconate is phosphorylated by the cell and becomes the substrate of 6PGD so that limiting gluconate may amplify the selective difference between allozymes.

It turned out that these alleles are selectively neutral within the limit of resolution (i.e. 0.5% per hour) in the genetic background of the normal *E. coli* K12. Also, no evidence of frequency-dependent selection was obtained. These tests were performed in the genetic background of edd^+, where *edd* codes for phosphogluconate dehydratase and it is known that this provides an alternative metabolic route for 6-phosphogluconate through bypassing 6PGD.

It was thought that small differences in the functional efficiency of the allozymes might be compensated metabolically by small changes in the rate of flow of substrate through the alternative pathway. Thus, selection experiments in the genetic background of edd^- (which cuts off the alternative metabolic route) were performed, and they detected selection between allozymes. Furthermore the selection was in the direction expected from the measured apparent K_m values of the allozymes (note that the smaller the Michaelis constant K_m, the stronger the affinity of the enzyme to the substrate). These results suggest, as noted by Dykhuizen and Hartl (1980), that the allozymes of 6PGD are selectively neutral or nearly neutral in natural populations, but that they have a latent potential for selection. Such a potential may on occasion be realized under the appropriate conditions of environment or genetic background, and under these conditions allozymes can become the raw material for adaptive evolution. These authors also refer to a statistical test made at their laboratory using data supplied by Selander and Levin on 109 naturally occurring *E. coli* strains. There were 15 distinguishable alleles of *gnd*, and a test was made if the observed distribution of allele frequencies conform to Ewens' formula for selective neutrality (see chapter 8 where Ewens' sampling theory is discussed). It turned out that the observed distribution is not incompatible with the hypothesis of selective neutrality. Hartl and Dykhuizen (1981) studied selection for two more alleles at the *gnd* locus using the chemostat, but their conclusion was unaltered.

From the standpoint of the neutral theory, their finding that otherwise neutral or nearly neutral alleles may have a latent potential for selection which can be realized under the appropriate conditions has a far-reaching implication. I would like to call their finding the 'Dykhuizen–Hartl effect' and I shall discuss this again in the next chapter. If this effect is general, as I assume it must be, even extreme 'selectionists' must admit that allozyme alleles can be selectively neutral (or nearly so) without being meaningless in evolution.

According to Wills (1981, p. 210), selection at the octanol dehydrogenase (*Odh*) locus in *D. pseudoobscura*, which is undetectable in a normally polymorphic outbreeding population, can be detected by applying 'the stress media' (octanol and ethanol) and at the same time by using inbred flies. Such a claim may no longer be regarded as contradictory to the neutral theory if we accept the Dykhuizen–Hartl effect.

As an additional example to show that the level of genetic variability in haploid organism is not very different from that of diploid organism, I would like to mention the recent work of Yamazaki (1981) on haploid liverwort *Conocephalum conicum*. He examined, using starch gel electrophoresis, 11 loci of this plant sampled from 7 different populations near Fukuoka, in the southern part of Japan. Of the 11 loci studied, 7 turned out to be polymorphic. The average virtual heterozygosity was 0.167. A similar result has earlier been obtained by Spieth (1975) using another haploid organism, *Neurospora intermedia*. However, Yamazaki (1981) points out that this fungus often endures and survives in nature in a diploid-like heterokaryon state, so *Conocephalum* can be regarded as a more ideal haploid organism. Therefore, it more definitely rules out overdominance as the general mechanism for the maintenance of genetic variability at the molecular level.

9.3 Statistical tests of selective neutrality

The neutral theory of protein polymorphism is based on the mathematical theory of population genetics, and as explained in chapter 8, models of neutral mutations have been extensively developed. We have a

better understanding of the behavior of neutral alleles in a finite population than any other class of selected alleles. Therefore, it appears that one can readily test the neutral theory by examining the distribution pattern of allelic frequencies in natural populations. Using the terminology of statistics, we treat the neutral theory as the null hypothesis and test it against observations. In fact, a number of tests have been proposed and applied, but, it has turned out that each test has its shortcomings. So far, no definite evidence against the neutral theory has been obtained through such tests. Rather, favorable cases for the theory have been added through Nei's test as I shall explain in this section. Before we proceed, however, I must emphasize that the existence of deleterious alleles by no means contradicts the neutral theory. As I explained in chapter 8, it is possible that deleterious alleles occur some ten times as frequently as the selectively neutral alleles at the time of mutation, although these deleterious alleles are restricted to a lower frequency range (e.g. less than 10%) in the populations. This means that simply testing strict neutrality against the existence of selection is not appropriate to the resolution of the neutralist-selectionist controversy.

In this context, Li's (1979) designations are helpful to clarify the issue. He distinguishes two hypotheses, the pan-neutrality hypothesis (denoted by H_p) which postulates that all alleles in a population are selectively neutral, and the neutral mutation hypothesis (H_n) which postulates that the genetic variation or the heterozygosity of a population is mainly due to neutral or almost neutral mutations. Needless to say, the neutral theory which I advocate in this book is H_n but not H_p.

Single-locus tests of Ewens and Watterson

I have already explained the sampling theory of Ewens (1972) and mentioned the proposal by him of testing selective neutrality with an approximate statistical procedure (see equation 8.31 and subsequent discussion in chapter 8). Pursuing the problem along a similar line, Watterson (1977, 1978a, b) developed a statistical theory to test selective neutrality more exactly in terms of sample homozygosity,

$$\hat{F} = \sum_{i=1}^{k} n_i^2/n^2, \tag{9.1}$$

where n is the sample size (i.e. total number of genes), k is the number of distinct alleles in the sample, and $n_1, n_2, \ldots, n_k$ are the numbers of representatives of these alleles in the sample ($n = n_1 + n_2 + \ldots + n_k$). As in Ewens (1972), he uses the number of alleles (k) to estimate the unknown parameter $\theta = 4N_e v$, and test $\hat{F}$ based on the estimate of θ. One of the examples given by Watterson (1977) is a sample of $n = 582$ for the esterase-2 locus from *Drosophila willistoni*. It contains seven alleles with numbers $n_1 = 559$, $n_2 = 11$, $n_3 = 7$, $n_4 = 2$, $n_5 = n_6 = n_7 = 1$. Applying his theory, he found that $\hat{F}$ is considerably larger than what is expected under selective neutrality, and the exact probability of getting such a large $\hat{F}$ occurs only with probability $P = 0.0069$. Then he concludes that 'neutrality was not the case and the selection appears to be in favor of homozygotes'. This means, however, that overdominance is not the case and the existence of slightly deleterious alleles is suggested. A similar test for *D. tropicalis* turned out to be not significantly different from the hypothesis of neutrality.

Although Watterson's test (and also Ewens') is well worked-out mathematically, it contains some problems with respect to its applicability to real data. It belongs to a type of test which treats each individual locus (and also each individual population) separately. Li (1979) pointed out that such a single-locus approach generally has the following three drawbacks: (i) large mean square error of the estimator for $\theta\,(=4N_e v)$, (ii) low statistical power, (iii) inability to use data from monomorphic loci.

In addition, the existence of slightly deleterious alleles and the incomplete detectability of allelic differences in electrophoretic data cause difficulty to the tests of Ewens and Watterson, because these tests are based on the assumptions that all alleles are selectively neutral (i.e. pan-neutralist hypothesis H_p) and that every mutation leads to a new allele that can be discriminated from pre-existing alleles (i.e. the infinite allele model). Li (1979) showed by numerical examples that when the mean heterozygosity of a population is mainly due to selectively neutral alleles but the population contains many deleterious alleles each having a low frequency, Ewens' method of estimating θ using the number of different alleles turns out not to be appropriate to test 'generalized neutrality' H_n, although it might be appropriate to test 'strict neutrality' H_p.

As compared with this, pooling data from all available loci, and estimating θ using the average heterozygosity ($\bar{H}_e$) as done by Nei (1975) and Ohta (1976), can overcome such difficulties and show many advantages over Ewens–Watterson's approach in testing the neutral hypothesis H_n in which existence of deleterious alleles is admitted, particularly if variation of the mutation rate for neutral alleles among loci is taken into account (Nei *et al.*, 1976a).

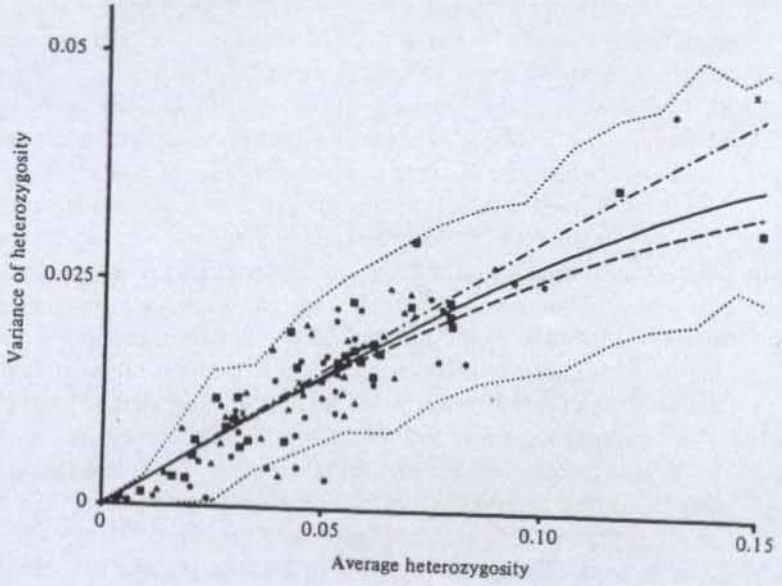
Nei's test

This test is based on the idea (Nei, 1975) that if we estimate the parameter $\theta = 4N_e v$ from the average heterozygosity ($\bar{H}_e$) by the formula $\theta = \bar{H}_e/(1 - \bar{H}_e)$ (see equation 8.20) and use this to predict the variance of

heterozygosity (σ_H^2) by using Stewart's formula (equation 8.16), then this predicted variance can be compared with the corresponding observed variance to test the neutral theory. The variance of heterozygosities among loci within species may be used to make this test. This belongs to a class of methods of pooling data from a large number of loci as just referred to above. It has been used extensively by Nei and his associates (Nei *et al.*, 1976b; Fuerst *et al.*, 1977) and has brought forward quite favorable evidence for the neutral theory.

Note that the observed variance computed by using data on heterozygosities at a number of loci contains sampling variance at the time of the gene frequency survey, so caution is needed to subtract this sampling variance (see Nei and Roychoudhury, 1974). On the other hand, Stewart's formula for the theoretical variance (σ_H^2) does not include the difference of mutation rates among loci, so it is desirable to take the variation of mutation rate into account (Nei *et al.*, 1976a) in computing the predicted variance.

Fig. 9.4. Relationship between average heterozygosity and the interlocus variance for vertebrate species. ——, theoretical relationship for the infinite allele model; — —, theoretical relationship for the stepwise mutation model; —·—, theoretical relationship for the infinite allele model with varying mutation rate (coefficient of variation of mutation rate = 1.0); . . . . ., 95% significance intervals of the variance. ■, mammals; ●, reptiles; ▲, fishes; ×, amphibians. From Fuerst *et al.* (1977).



Fuerst *et al.* (1977) applied Nei's test to data from 129 animal species and 6 differentiated subspecies for which at least 20 loci were examined with a sample of at least 30 genomes.

Fig. 9.4 shows the result of their analysis using data from a group of vertebrates consisting of 95 species and one differentiated subspecies. In this figure, the observed relationship between the average heterozygosity and the interlocus variance is plotted by various dots (the square dots for mammals, round dots for reptiles, triangular dots for fishes and cross dots for amphibians). The corresponding relationship obtained by using Stewart's formula for the variance and Kimura-Crow formula (equation 8.15) for the average heterozygosity is shown by the solid line. It will be seen that the agreement between observations and theoretical predictions is satisfactory. The broken line (— —) in the same figure represents the theoretical relationship expected under the stepwise mutation model (see section 8.5), and the chain-like line (—·—) the theoretical relationship under the infinite allele model in which the assumption of varying mutation rate (with coefficient of variation $\alpha = 1$) is incorporated (see equation 8.22 and subsequent explanations). Note that the observed points are mostly included within the 95% significance intervals delineated by dotted lines. These authors also analysed data from 34 species and 5 differentiated subspecies of invertebrates using the same method. Again the agreement between the observed values and theoretical predictions are good, but there is some tendency for the observed values to deviate upward from theoretical values predicted under the assumption of constant mutation rate. This is because the average heterozygosity is higher in this case than in the case of vertebrates and therefore the inflationary effect on the observed variance of differences in mutation rate from locus to locus is manifested more prominently. In fact, theoretical values derived from the model of varying mutation rates did much better, showing an excellent fit to the actual data.

The variance of heterozygosity among different species for a homologous locus can also be tested in a similar way. In this case, the mutation rate v should be essentially the same for related species, but the effective population size N_e may vary with species. In addition, there is the possibility that heterozygosities in related species are correlated if the divergence time is short, and this reduces the variance among species. Keeping such difficulties in mind, Fuerst *et al.* (1977) analysed data for a total of 125 loci from various vertebrates, and for these animals very good agreement with the theoretical prediction was obtained. For invertebrates, information on 47 loci, all from *Drosophila*, was available for study. It turned out that the agreement with the theoretical value was generally

good, but for some groups of *Drosophila* species, there was significant discrepancy, showing lower than expected variance, the real cause of which is unknown.

In order to test the neutral theory, Nei and his associates also investigated the pattern of genetic differentiation between populations (Chakraborty *et al.*, 1978). They investigated the relationship between genetic identity (I) and the correlation (r) of heterozygosities of different populations or species to see if the observed pattern agreed with expectations from the neutral theory. As explained in the previous chapter (see equation 8.116) the index I serves to measure the genetic distance between populations and under complete isolation it decreases at the rate $2v$ per generation. Since the mutation rate may differ from locus to locus, Chakraborty *et al.* (1978) used a formula for I which takes the variation of v among loci into account. The correlation coefficient r represents the correlation of single-locus heterozygosity over loci between two populations, and it decreases at the rate $(4v + 1/N_e)$ per generation under complete isolation (Li and Nei, 1975), i.e.

$$r = \exp[-(4v + 1/N_e)t], \tag{9.2}$$

where t is the time in generations since divergence.

The observed relationship between I and r was examined separately for four group of animals, mammals, teleost fishes, reptiles and *Drosophila*. In computing the theoretical relationship based on the neutral theory, $\bar{M} = 0.06$ and $\bar{H} = 0.05$ were assumed for the first three groups (vertebrate groups), and $\bar{M} = 0.28$ and $\bar{H}_e = 0.17$ for the *Drosophila* group, where $\bar{M}$ stands for the mean value of $4N_e v$ and $\bar{H}_e$ the average heterozygosity per locus. It turned out that agreement between observations and predictions was good for data from mammals and teleosts. For data from reptiles, the relationship between I and r was not clear, although a reasonable agreement between observations and prediction was found for many cases. On the other hand, for data from *Drosophila*, observed correlations were definitely higher than predicted by the neutral theory. They could give no definite explanation for such discrepancy. Chakraborty *et al.* also examined the relationship between the mean and the variance of genetic distance, and in this case, they found surprisingly good agreement between observed and expected values in all of the four groups of organisms.

Lewontin–Krakauer test

This test (Lewontin and Krakauer, 1973) is based on the idea that although selection may act differently on different loci, the effect of

inbreeding should be uniform over all loci. More specifically they suggested the following method for testing the neutrality of polymorphism. Suppose a pair of alleles A_1 and A_2 are segregating at a locus. We compute the inbreeding coefficient F using the mean $(\bar{p})$ and the variance (σ_p^2) of the frequency of A_1 among populations within a species by the formula

$$F = \sigma_p^2/[\bar{p}(1-\bar{p})]. \tag{9.3}$$

This corresponds to Wright's F_{ST} (Wright, 1943), and we apply this formula to obtain F for each of loci under investigation. Then, we compute the mean $\bar{F}$ and the variance s_F^2 of such F values among different loci. Suppose that F is estimated from n populations. Lewontin and Krakauer claim that under selective neutrality a conservative value of s_F^2 is $2\bar{F}^2/(n-1)$, and this can be used to test selective neutrality: if $k = (n-1)s_F^2/\bar{F}^2$ significantly exceeds 2, the neutral theory can be rejected. In other words, if s_F^2/σ_F^2 is much larger than unity, where $\sigma_F^2 = 2\bar{F}^2/(n-1)$ is the theoretical variance, F values are considered to be too heterogeneous to be explained without selection.

The idea of using uniformity of F values among different loci to assess the role of random drift appears to have originated from Cavalli-Sforza (1966), who calculated F values for various blood group polymorphisms over a wide range of human groups. Lewontin and Krakauer applied their method to the data collected by Cavalli-Sforza obtaining $s_F^2/\sigma_F^2 = 10$, and claimed that they had shown highly significant heterogeneity in F values for human polymorphic genes over the world, thus demonstrating that a significant fraction of human polymorphisms owe their current gene frequencies to the action of natural selection.

The validity of the Lewontin-Krakauer test has been questioned by Robertson (1975) who showed that $k = 2$ may be a serious underestimate if there are correlations between populations (such as due to migration or recent branch off) as is likely in human populations. Robertson pointed out that for a sample of 60 populations, as used by Lewontin and Krakauer, a correct value of σ_F^2 would be about six times the value given by them, and as the observed variance between loci was ten times their expectation, this would throw some doubt on their claim that heterogeneity of F values is highly significant (for more detailed account on the expected value of k under correlation between populations, readers may refer to Ewens, 1977). The Lewontin-Krakauer test was also criticized by Nei and Maruyama (1975) who performed a Monte Carlo simulation of structured populations and found that observed values for k are usually much larger than two. In response to these criticisms, Lewontin and Krakauer (1975) listed 'do's' and 'don'ts' for testing the heterogeneity of F values. Among the correct uses,

they include the application of their method to temporal (rather than geographical) variation of gene frequencies within populations, as they did to data of Krimbas and Tsakas (1971) on two polymorphic loci in the olive fruit fly *Dacus oleae*. In this case, the difference of the temporal inbreeding coefficients for the two gene loci was not statistically significant. More recently, Gaines and Whittam (1980) applied the Lewontin-Krakauer test to temporal variation in gene frequency at five electrophoretic loci in each of four fluctuating populations of the prairie vole (*Microtus ochrogaster*) over a three-year period. The observed values of the variance ratio, s_F^2/σ_F^2 (assuming $k = 2$) were all near to unity and therefore statistically non-significant. These authors concluded that 'nonselective forces', especially genetic drift, are the major determinant of gene frequency changes in the vole population.

Other tests

Kimura and Ohta (1971a, b) examined the relationship between the average heterozygosity ($\bar{H}_e$) and the proportion of polymorphic loci (P_{poly}) using the formula $P_{poly} = 1 - q^{H/(1-H)}$, where $H = \bar{H}_e$. This formula was derived by using the infinite allele model, and we usually take $q = 0.01$ or $q = 0.05$ (see equation 8.9 and discussion on this formula in chapter 8). They applied this formula to data from several organisms and found that observed values agree reasonably well with theoretical expectations. Later, the same test was performed by Fuerst *et al.* (1977) using much more extensive data, and the results were obtained wholly consistent with the neutral theory.

Johnson (1972) proposed to test the neutral theory by examining the relationship between the actual number of different alleles n_a and the ratio of the effective to the actual numbers of alleles, i.e. n_e/n_a, where n_e is the reciprocal of the sum of squares of the allelic frequencies. He reasoned that if the neutral theory is correct, n_e/n_a decreases as n_a increases since the equilibrium frequency distribution of isoalleles must be very distorted (with larger number of less frequent alleles for larger n_a). He investigated this relationship for three sets of data from *Drosophila*, and presented 'quite clear-cut' results showing that in all three cases n_e/n_a increases as n_a increases, contrary to the prediction of the neutral theory. Yamazaki and Maruyama (1973) re-examined the same problem using much more extensive data together with the exact theoretical values based on the infinite neutral allele model. Unlike Johnson, they found excellent agreement between observation and the neutralist prediction, Yamazaki and Maruyama also pointed out that Johnson excluded those alleles with a frequency of less than 0.01 and this may have led him to a false conclusion.

Yamazaki and Maruyama (1972, 1974) devised their own approach to test selective neutrality of protein polymorphism. It is based on Maruyama's (1972a) finding that the expected value of the sum of heterozygotes made by a mutant from its appearance until fixation or loss is 'invariant', that is, independent of population structure. It also makes use of the theory that under irreversible mutation (with the infinite site model) the steady state distribution of neutral mutants is given by $\Phi(x) = 4N_e v/x$ which is a special case of equation 8.97 for $G(x) = 1$ and $v_{gam} = v$, where v is the mutation rate per locus and x is the frequency of a mutant allele. Then, multiplying $2x(1-x)$ to $\Phi(x)$, the relative frequencies of heterozygotes for various mutant frequency classes are given by $8N_e v(1-x)$. Yamazaki and Maruyama also considered cases with selection. They showed that if the mutant is unconditionally advantageous, the corresponding frequency of heterozygosity is roughly constant, while if it is definitely deleterious the distribution has minimum at $x = 0.5$ and increases to a maximum at $x = 0$ or 1. If the mutant alleles are overdominant, it is not easy to predict an unequivocal distribution, but they made a conservative assumption that all equilibrium frequencies are equally likely, that is, distributed uniformly (in reality, alleles with equilibrium frequencies near 0.5 should survive more often). Then the distribution of heterozygosity becomes bell shaped with maximum at $x = 0.5$. One difficulty of this approach is that when we observe a heterozygote, say AA', we do not know which of A and A' is the mutant allele. To avoid this difficulty, Yamazaki and Maruyama suggested reflecting the gene frequency scale around the value of 0.5 so that the ordinate values for x and $1-x$ are added. With this device, the distribution of heterozygosity for neutral mutations becomes constant over the gene frequency range $x = 0.0 \sim 0.5$, namely, the distribution becomes uniform as in the case of advantageous mutations.

They then plotted the sum of the observed heterozygosities (ordinate) against the gene frequency (abscissa). In other words, they considered relative contribution of alleles at each frequency class to the amount of heterozygosity. In doing so, they eliminated one allele at random from each locus to correct the bias from the nonindependence of allele frequencies at a single locus. Using all available data (442 alleles of 145 proteins from 16 different species in their 1972 paper), they obtained the result which fits well to the expectation of uniform distribution of heterozygosity, that is, to the case of either neutral or definitely advantageous mutants. Of these two, they concluded that the hypothesis of selective neutrality is more appealing because it is unlikely that most new mutations are more fit than the pre-existing alleles.

Yamazaki and Maruyama's method has been criticized by Ewens and

Feldman (1974) who pointed out that if $4N_e v$ is large, the formula used in the above analysis is not exact. However, this point is minor since $4N_e v$ for neutral alleles can be assumed to be much smaller than unity. The same analysis was repeated later with more data (1530 alleles of 600 polymorphic loci) by Yamazaki and Maruyama (1975), but the observed distribution pattern of heterozygosity turned out to be essentially the same, again supporting selective neutrality.

Ayala and Gilpin (1973) criticized Yamazaki and Maruyama (1972) saying that the theoretical distribution of the relative amount of heterozygosity calculated by Yamazaki and Maruyama under the assumption of overdominance is not valid, and that if all configurations of allelic frequencies are equally likely, the patterns of heterozygosity observed agree well with the overdominant model. Clearly, Ayala and Gilpin's criticism comes from their insufficient knowledge of the stochastic theory of population genetics: the probability distribution of allelic frequencies must be derived from the basic assumptions of overdominant mutations in a finite population, and it is not allowed, as they did, to make such an arbitrary assumption that all allelic configurations for a given number of alleles, n, are equally likely, that is, to assume that all points in the hyperplane defined by $Y_1 + Y_2 + \ldots + Y_n = 1$ (where Y_i is the frequency of the i-th allele) have equal probability.

Ayala and Gilpin (1974) also criticized the neutral theory as follows: if the theory is correct the distribution of genetic distance D between taxa must be approximately normal or bell shaped (with the probability of $D = 0$ vanishingly small), whereas the observed distribution of the genetic distance between taxa in the *D. willistoni* group is U-shaped in all cases and this refutes the neutral theory (see also Ayala, 1974). In deriving the theoretical prediction, they applied a numerical solution of the Fokker-Plank diffusion equation involved and computed the probability distribution of distance D at $t = 0.1N$, $0.2N$ etc. between two isolated populations each of size N, assuming that the two populations diverged from a single population at $t = 0$; here $D = [\Sigma(X_{1i} - X_{2i})^2/2]^{1/2}$ in which X_{1i} and X_{2i} are the frequencies of the i-th allele in the first and second populations and summation is over all the alleles at a given locus.

However, as pointed out by Nei and Tateno (1975), Ayala and Gilpin's (1974) computation is based on two unrealistic assumptions. First, they assumed that the initial gene frequency at the time of population splitting are the same for all loci. In reality, however, gene frequencies vary greatly from locus to locus: some loci are monomorphic while others are polymorphic in different extent. Secondly, they assumed no mutation, and

this too is unrealistic. In Ayala and Gilpin's study, they started from such a state that the population contains, for example, three alleles each with equal frequencies (1/3, 1/3, 1/3) and considered the process in which one of the alleles tends to fix in the population as time goes on. It seems to me that Ayala and Gilpin overlooked the possibility that monomorphic loci remain the same after splitting (i.e. $D = 0$) until new, successful mutants occur and spread to reach appreciable frequencies in either of the two populations. For this to occur may take a long time. Furthermore, even at polymorphic loci, there is usually a predominant allele at each of them and it will take a long time for different alleles to replace the previously predominant allele.

Nei and Tateno (1975) incorporated, in their simulation studies, a realistic distribution of gene frequencies among loci at the time of population splitting and investigated how the D value (as used by Ayala and Gilpin) changes after the splitting under mutation and random drift, assuming $M = 4N_e v = 0.1$. It turned out that distributions of D values thus obtained at $0.5N_e$ and N_e generations are L-shaped with the maximum in the neighborhood of $D = 0$ and are strikingly different from the corresponding theoretical distributions obtained by Ayala and Gilpin, whose distributions are bell-shaped. Nei and Tateno also investigated the process of change of the genetic identity coefficient I (see equation 8.116 in chapter 8) under the neutral mutation hypothesis and obtained distributions which agree quite well with the observed distributions of interracial and interspecific gene differences for enzyme loci in *Drosophila*. These results indicate that the gene substitutions at enzyme loci can be explained by the neutral theory, contrary to the claim of Ayala and Gilpin (1974).

9.4 On some rival hypotheses

As shown in the above two sections, the neutral theory of protein polymorphism has been able to withstand various experimental and statistical tests so far performed. Together with its success in explaining an increasing number of observations regarding evolutionary changes at the molecular level, it is likely that the neutral theory will survive into the future, although some modifications and refinements will no doubt be made. In this section, I would like to discuss some rival hypotheses on the maintenance of genetic variability at the molecular level.

Overdominance hypothesis

I have already enumerated enough evidence showing that overdominance can not be the major cause of protein polymorphism. Especially, the finding that genic variability or virtual heterozygosity is just

as high in haploid organisms as in diploid organisms must be a fatal blow against the claim that overdominance is the universal cause for the maintenance of protein polymorphism. Here, I shall present an additional piece of evidence against the overdominance hypothesis. Through numerical and analytical approaches, Lewontin *et al.* (1978) showed that heterozygote superiority in fitness alone can not maintain many alleles at a locus (they considered only deterministic changes). According to them, even when all heterozygotes are more fit than all homozygotes, the proportion of fitness arrays that will lead to stable, feasible equilibrium of more than six or seven alleles is vanishingly small: as the number of alleles increases from three, the conditions for stable equilibrium become rapidly very severe. Even when such conditions hold, the distribution of allelic frequencies at equilibrium becomes nearly uniform. On the other hand, the observed distributions of alleles at a polymorphic locus where a large number of alleles are segregating (such as *Xdh* locus in *Drosophila*; see Coyne, 1976 and Singh, Lewontin and Felton, 1976) are very asymmetrical.

Evidence against the overdominance hypothesis of protein polymorphism has accumulated so much now that only blind faith can maintain it.

Franklin–Lewontin theory involving large-scale linkage disequilibrium

In their paper entitled 'Is the gene the unit of selection?', Franklin and Lewontin (1970) investigated a model in which a large number of overdominant loci are tightly packed within a chromosome with multiplicative fitness effect among loci. They disclosed a remarkable property of the model that, under a suitable condition, an extremely strong linkage disequilibrium develops in such a way that a population consists mainly of two complementary chromosome types. In one series of simulation experiments, for example, they assumed 36 overdominant loci on a chromosome with recombination fraction $r = 0.0025$ between adjacent loci and with 10% fitness decrease in homozygous condition at individual loci. The effective population size was assumed to be 400. Starting from the condition that gene frequencies are equal (i.e. 0.5) at all the loci and combinations of genes between loci are completely at random, they observed, for instance, in one of replicate trials, that, in generation 300, the population consisted mainly of two chromosome types, denoted by

011 010 110...101

and

100 101 001...010,

in roughly equal frequencies. These two chromosomes amounted to some

83% while the remaining types were only 17% in all. Once such an equilibrium is reached, the average fitness of the population ($\bar{w}$) becomes 0.4 which is much higher than $\bar{w} = 0.16$ expected under random combination of genes between loci. No such remarkable polarization effect occurred if the recombination fraction between adjacent loci was $r = 0.005$ or more. Thus, linkage must be tight for the effect to appear. However, no appreciable increase of fitness occurred for $r = 0.0$ (complete linkage).

What makes this model attractive is that $r = 0.0025$ can be a reasonable map distance between two adjacent segregating loci in *Drosophila*. They also found a very interesting property of the model: by increasing the number of segregating loci within a chromosome while holding the total map length and the inbreeding effect (i.e. the decrease of fitness in the homozygous state) constant, they found that the average correlation between genes on the chromosome remained essentially unchanged (for analytical treatments, see Slatkin, 1972). In other words, the property of the system does not depend on the number of genes; what matters is the chromosome as a whole but not individual loci. If a very large number of segregating loci are involved, the marginal fitness effect of making one locus homozygous is very small, and this gives a realistic feature to the model. On the other hand, in this model, allelic combinations are polarized into only two groups in an extreme situation so that it becomes somewhat analogous to the balanced lethal system. If such a polarization really occurs in natural populations, the argument of 'heavy segregational load' against the assumption of a large number of overdominant loci breaks down; the load under random mating does not necessarily become much larger than 0.5.

Franklin and Lewontin (1970) also investigated various cases such as asymmetrical fitness between two homozygotes at each loci, unequally spaced loci on a chromosome and truncation selection. They confirmed that their new finding is robust to such a change. This led them to the conclusion that the theory of population genetics which treats the chromosome as a whole is more suitable and realistic than the traditional theory which is framed in terms of gene frequencies at individual loci. Such a claim was again repeated by Lewontin (1974) in his book *The Genetic Basis of Evolutionary Change*, who closed it by pointing out an analogy between the theory treating the chromosome as a whole and the well-known law in thermodynamics $PV = nRT$, which describes in simple terms the relationship among pressure, temperature and volume, in spite of the fact that an immense number of molecules are involved each undergoing chaotic motion. It looked as if he considered that a new era in population genetics would be opened by this type of approach.

Although interesting, whether such marked linkage disequilibrium really exists in nature is another matter. In fact, as more and more actual observations accumulate, it has become increasingly clear that there is virtually no linkage disequilibrium among enzyme loci in *Drosophila*, unless gene loci are associated with inversions (a structural change with a chromosome segment inverted).

That linkage disequilibrium between isozyme loci seldom exists in natural populations of *Drosophila* has been shown by extensive studies of Mukai and his associates (Mukai *et al.*, 1971, 1974; Mukai and Voelker, 1977), and also by Langley *et al.* (1978) and Loukas *et al.* (1979, 1981). Even between such closely linked loci as *Odh* (octanol dehydrogenase) and *Est-C* (esterase-C) with recombination fraction 0.0058 between them, no linkage disequilibrium was observed by Mukai and Voelker (1977). Thus, even though the possibility can not be excluded that the Franklin-Lewontin theory may be valid in some cases to explain linkage disequilibrium between genes within a very short chromosome segment, as a general theory for the maintenance of protein polymorphism, it has been a failure. This is a good example where a beautiful dream in science has been destroyed by brutal facts.

Incidentally, significant linkage disequilibria have been found in some cases between isozyme genes and polymorphic inversions. However, these are most appropriately explained by founder effect, that is, historical accident and random drift. Nei and Li (1980) investigated mathematically the problem of non-random association between isozyme alleles and inversion chromosomes in finite populations. Using the diffusion equation method, they estimated the probable magnitude of non-random association, that is, the difference in the frequencies of a given allele between inversion and non-inversion chromosomes. According to their analysis, significant non-random association will develop if $4N_1r$ is small (say, less than 1), where N_1 is the effective number of the inversion chromosome and r is the rate by which the allele in the inversion chromosome is exchanged with the allele in the non-inversion chromosome. They revealed that the standard error of non-random association often becomes larger than the mean.

Previously, Prakash and Lewontin (1968) noted that, at the $Pt-10$ and *Amylase* loci, there is a strong association of different but characteristic alleles with different inversion types (such as 'Standard' and 'Santa Cruz') in the third chromosome of *Drosophila pseudoobscura*, irrespective of the geographical origin of the strains. Furthermore, in the sibling species, *D. persimilis* whose third chromosome gene arrangements belong to (or are directly derived from) Standard type, they found that the allelic constitution is very similar to Standard type of *D. pseudoobscura*. Considering the possibility that the separation of different inversion types dates back at least several million generations, they conclude that the gene contents of the inversions are 'coadapted', in agreement with Dobzhansky (1951), so that any introduction of foreign alleles is rejected. However, Nei and Li (1980), through a more meticulous re-examination of the problem, point out that Prakash and Lewontin's (1968) observation can easily be accommodated by the neutral theory.

Ecological theories assuming adaptive strategies

If protein and other molecular polymorphisms are adaptive, as claimed by the selectionists, one should expect to find correlation between genetic variability and some environmental factors. In fact, a number of selectionists adhere to the view that genic diversity is an adaptive strategy for the species to exploit temporally and spatially heterogeneous environments.

One popular hypothesis, which was proposed by Levins (1968) predicts that genetic variability should be higher in unstable environments than in stable ones. It is a plausible hypothesis and similar predictions have also been made by other authors (see Soulé, 1976 and Valentine, 1976, for review). Since the deep sea is the most stable and homogeneous habitat on earth, it was predicted that genetic variability of organisms living there must be very low. It turned out, however, that many organisms living at the bottom of the deep sea have very high genetic variability (Gooch and Schopf, 1972; Ayala *et al.*, 1975). Thus, this specific prediction was clearly falsified, with observations having come in an opposite direction. Also, Somero and Soulé (1974) tested this hypothesis (which they call 'niche-variation hypothesis') by examining if there is a tendency for fishes from thermally variable habitats to display higher levels of protein polymorphism than those from habitats with stable thermal regimes, but no such tendency was found.

Failure of Levins' hypothesis led some to adopt the 'trophic resource stability' hypothesis (see Valentine, 1976). According to this hypothesis, if the food supply is abundant and stable, the best strategy for the species is to have allelic variants which are specialized to different foods, so that high genetic variability is predicted. On the other hand, if the food supply is unpredictable, genetic uniformity is predicted because the best strategy is to have a generalist genotype. On another occasion, a hypothesis was entertained (Ayala and Valentine, 1979) that, in species that inhabit

temporally variable environments or where individual organisms range widely through a variety of environments, functionally flexible alleles are favored, and that in the more extreme cases, the species tend to be monomorphic for the most flexible alleles. These ecological theories assuming various adaptive strategies are so flexible, it seems to me, that almost any observation on genetic variability can be explained by using some of them (or by inventing still more).

Returning to Levins' hypothesis, both Powell (1971) and McDonald and Ayala (1974) reported results from *Drosophila* population cage experiments which showed that the average heterozygosity was higher in populations maintained in heterogeneous environments than in populations in more constant environments. Since then, their experiments have been cited as evidence for the view that protein polymorphism is maintained by balancing selection. However, Nei (1980), who made a close examination of their data, points out that, in their experiments, the average heterozygosity declined faster in both environments than the rate expected under pure genetic drift. In McDonald and Ayala's work, the original sample was obtained from a natural population of *Drosophila pseudoobscura* which had the average heterozygosity of 0.212 (for 20 polymorphic loci), and the size of their experimental population was never smaller than 500. Therefore, the maximum reduction in average heterozygosity due to random sampling drift for the period of $12 \sim 15$ generations of their experiment must be such that 0.212 changes to $0.212 \times (1 - 1/1000)^{15}$ or 0.209. As compared with this theoretical prediction, the average heterozygosity observed in heterogeneous environments at the end of the experiment was 0.195, while the average heterozygosity in homogeneous environments was 0.146. Clearly, the heterogeneous environments did not retard but actually accelerated the reduction of heterozygosity as pointed out by Nei. In a similar experiment carried out by Powell (1971) using *Drosophila willistoni*, the reduction of average heterozygosity in heterogeneous environments was more pronounced. Note, however, that in both experiments, reduction of heterozygosity was larger in the homogeneous environment. The reason for this is unknown. Minawa and Birley (1978) investigated the same problem using cage populations of *Drosophila melanogaster*. In contrast to the results of Powell (1971) and McDonald and Ayala (1974), no significant difference in the average heterozygosities was detected between flies kept in the constant and variable environments. Thus, environmental heterogeneity had no effect on maintaining heterozygosity. On the other hand, significant differences of electromorph frequencies between different environments (as compared with that between replicate populations) were found. Although

these results were interpreted by the authors as inconsistent with the neutral theory, in my opinion, these are also inconsistent with the adaptive strategy hypothesis. Significant shifts of electromorph frequencies were probably caused by selection acting on unrecognized loci in linkage disequilibrium. Clearly, more careful and detailed experiments on the problem are required to settle the issue. For additional problems on the relationship between genetic polymorphism and environmental heterogeneity, readers may refer to such review articles as Hedrick *et al.* (1976), Nevo (1978), Soulé (1976) and Valentine (1976).

9.5 Effects of molecular structure and functional constraint on genetic variability

From the standpoint of the neutral theory, protein and other molecular polymorphisms represent a transient phase of molecular evolution. If so, we should expect that these polymorphisms are strongly influenced by the structure and the functional constraint of the molecules involved, as we found in molecular evolution (see chapter 7). In this context, an observation made by Zouros (1976) is of special importance. He compared the amounts of electrophoretic variability of loci coding for multimeric enzymes (consisting of more than one polypeptide subunit) and those of loci coding for monomeric enzymes in four animal and two plant species. In all cases, he found that heterozygosities are lower for multimers than for monomers. He also found that within a given species or population, multimeric enzymes are less polymorphic than monomeric enzymes of similar function. That monomeric proteins show, on the average, significantly higher levels of heterozygosity (and also polymorphism) than multimeric proteins has been confirmed by other workers (Ward, 1977; Harris *et al.*, 1977; see also Koehn and Eanes, 1978). As pointed out by Zouros (1976), this observation necessitates rejection of the hypothesis (Fincham, 1966) that heteromultimers formed by association of subunits coded by two different alleles provide a molecular basis of overdominance. On the other hand, this observation is consistent with the neutral theory, because, structural constraint for proper function must be greater for multimers (due to intersubunit interactions) than for monomers, and therefore, other things being equal, the fraction of mutations that are neutral (i.e. not harmful) is smaller for multimeric enzyme loci than monomeric ones.

Another topic of interest is the relationship between the subunit molecular weight of protein and the level of intraspecific variability. Koehn and Eanes (1977, 1978) made an extensive analysis of data on enzyme

polymorphism in *Drosophila* and found that, for 11 dimeric enzymes, both heterozygosity and the number of alleles show significant correlation with subunit molecular weight. The estimated correlation coefficient (r) was 0.413 if the subunit molecular weight was correlated with observed heterozygosity per sample for each locus, but was 0.753 if this was correlated with the average heterozygosity per locus (Koehn and Eanes, 1977). Both correlations were statistically significant. According to the neutral theory, or more precisely, if we assume the infinite neutral allele model, the average heterozygosity is $\bar{H}_e = 4N_e v_0/(1 + 4N_e v_0)$. In this formula, v_0 is the mutation rate for neutral alleles and this is expected to be proportional to the gene size. Therefore v_0 should be proportional to the molecular weight of the protein coded by the gene. This means that the average heterozygosity should be highly correlated with the molecular weight. Thus, the observed higher correlation, i.e. $r = 0.753$ is consistent with the neutral theory. Also, according to the neutral theory, the observed heterozygosity at a given protein locus for an individual species may deviate a great deal from the expected value $\bar{H}_e$ due to random genetic drift, as is evident from the wide ranging distribution of per-locus heterozygosity among related species (see, for example, Fig. 9.3 in this chapter). This stochastic change will certainly reduce the correlation coefficient when it is computed for each sample separately as compared to the case when it is computed by using the average heterozygosity over a number of species. This also explains the lower correlation coefficient, i.e. $r = 0.463$, obtained by Koehn and Eanes.

Nei *et al.* (1978) made a more detailed and careful analysis of the relationship between subunit molecular weight and genetic variability of proteins using data from six groups of organisms: primates, rodents, reptiles, salamanders, fishes and *Drosophila*. Proteins used for their analysis were mostly dimers and monomers, but included some trimers and tetramers. These authors found that the correlation coefficient between molecular weight and the heterozygosity was positive in all groups, and that the magnitude of correlation was roughly in agreement with what was expected under the neutral theory when the incomplete correlation between molecular weight and the mutation rate was taken into account. Furthermore, they found that the correlation was higher when the average heterozygosity was high than when this was low. This is also consistent with the expectation of the neutral theory. Harris *et al.* (1977) made a similar study using data on 87 enzyme loci in human populations. They compared the average subunit size of the enzymes between polymorphic and nonpolymorphic loci, but found no significant difference between them.

(Also no significant differences were found when loci coding for monomeric and multimeric enzymes were considered separately.) The real reason for such lack of correlation between subunit molecular weight and the level of polymorphism is not known. Since the average heterozygosity is low in man (about 8%), and since heterozygosity at a particular locus varies a great deal from species to species by random drift, it is possible that this lack of correlation in man is accidental (but see Nei *et al.*, 1978). On the other hand, Harris *et al.* (1977) found that multimeric enzymes show a significantly lower incidence of polymorphism than do monomeric enzymes in agreement with Zouros' earlier finding mentioned above. Furthermore, they disclosed a remarkable fact that incidence of polymorphism is particularly low among multimeric enzymes in which interlocus molecular hybrids occur. According to Harris *et al.* (see their Table 4), the incidence of polymorphism is 56% for 27 loci coding for monomeric enzymes and 42% for 38 loci coding for multimeric enzymes not forming interlocus hybrid molecules, but it is only 9% for 22 loci coding for multimeric enzymes which form interlocus hybrid molecules. This finding is particularly favorable to the neutral theory, because formation of interlocus hybrid molecules imposes structural constraint on the protein molecules so that the probability of mutational change being selectively neutral is reduced.

The neutralists' proposition that an increase of functional constraint on a protein decreases its intraspecific variability appears to have wide applicability. In fact, it can be used to reinterpret various observations or hypotheses which were originally put forward to support the selectionist's cause.

Gillespie and Kojima (1968) compared variabilities of two groups of enzymes, one known to be active in energy metabolism (Group I) and the other which have broad substrate specificities (Group II). They found that Group I enzymes exhibit much less variation than Group II enzymes, namely, the mean heterozygosity per locus of Group I enzymes is only $9 \sim 24\%$ of that of Group II enzymes in *Drosophila ananassae*. Gillespie and Langley (1974) made a similar study using more extensive data from human, mouse and *Drosophila* populations. They introduced a slight modification in classifying enzymes into two groups, namely, (i) enzymes characterized by a singular physiological substrate (largely coinciding with the Group I enzymes of Gillespie–Kojima) and (ii) those with multiple physiological substrates (almost the same as Group II). With this redefinition of enzyme classes, a clear and consistent difference between Group I and Group II enzymes with respect to the average heterozygosity was shown to exist not only for *Drosophila* but also for the human and the

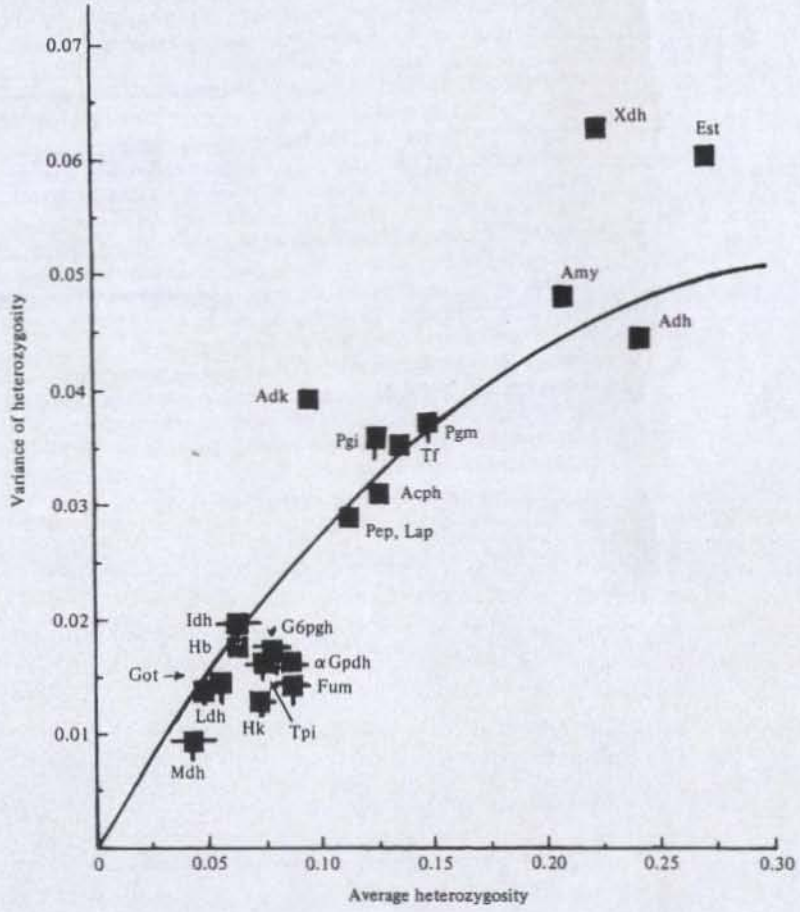
mouse. The ratio of the average heterozygosity of Group I to that of Group II is 0.38 for man, 0.65 for the mouse and 0.17 for *Drosophila*. It was also noted that the average heterozygosity of the Group I enzymes is roughly constant and is approximately 0.05 in these animals. Gillespie and Langley (1974) presented these observations to support their claim that enzyme polymorphism is primarily due to selection acting on environmental variation in gene function. Their observations, however, can also be explained easily by the neutral theory, namely, functional constraint is stronger for substrate-specific enzymes (Group I) than substrate-nonspecific enzymes (Group II), and therefore the probability of a mutation change (amino acid replacement) being not harmful (i.e. selectively neutral) is smaller for the Group I than for Group II enzymes.

Yamazaki (1977) went further and showed, using data from various *Drosophila* species, that not only do the substrate-specific enzymes have lower heterozygosity (with an average heterozygosity $\bar{H}_e = 0.075$) than the nonspecific enzymes ($\bar{H}_e = 0.200$), but also that these two enzyme groups differ in the mean-variance relationship of heterozygosity. He computed the expected variance of heterozygosity for each enzyme locus using Stewart's formula (see equation 8.16 in the previous chapter), where parameter θ or $4N_e v$ was estimated from the average heterozygosity over species. He also computed the observed variance for each enzyme over different species, and compared the observed variances with the expected variances at 22 enzyme loci, of which 13 code for substrate-specific enzymes and 9 substrate-nonspecific enzymes. In the 13 substrate-specific enzymes, the observed variances were always smaller than expected, and the difference was statistically significant in 9 enzymes. On the other hand no such tendency was noticeable in the 9 substrate-nonspecific enzymes; the observed variance was smaller in 5 cases and larger in 4 cases. Furthermore, only in 2 cases were the differences statistically significant.

A similar analysis was carried out by Gojobori (1979, 1982) using data on 20 different proteins (mostly enzymes) from 14 *Drosophila* species, 14 *Anolis* species and 31 other species. Fig. 9.5 illustrates the results of his analysis with observed values shown by solid squares. Various bars attached to these squares represent functional constraints: a downward bar indicates that the enzyme is substrate specific, a left-hand bar denotes that the enzyme is involved in a single pathway and a right-hand bar indicates that the enzyme is involved in main pathways (for details, see Gojobori, 1982). In the same figure, the curve represents the theoretical relationship between the average heterozygosity and the variance of heterozygosity based on the infinite neutral allele model (see equations 8.15 and 8.16 in the previous

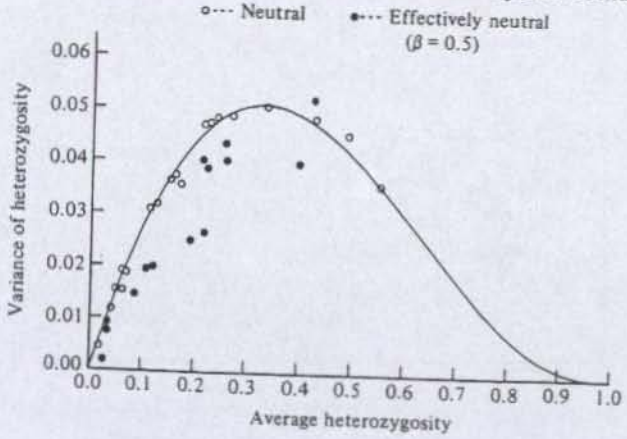
chapter). It is evident from the figure that enzymes with various functional constraints (i.e. squares with one or more bars) tend to have low heterozygosity. Furthermore, these enzymes have variances of heterozygosity markedly lower than expected on the neutral infinite allele model.

Fig. 9.5. Relationship between the average heterozygosity and the variance of heterozygosity for 20 different proteins (mostly enzymes) plotted as solid squares (observed values). The curve represents the theoretical relationship based on the neutral, infinite allele model. Bars attached to individual squares designate various functional constraints: a downward bar indicates that it is a substrate-specific enzyme, a left-hand bar denotes that it is an enzyme involved in a single metabolic pathway and a right-hand bar denotes that it is an enzyme involved in main pathways. From Gojobori (1979) with slight modification.



From the standpoint of the neutral theory these observations may most naturally be explained by assuming that for such enzymes, effectively neutral but very slightly deleterious mutations are largely responsible for polymorphic variations. To investigate this, I have performed Monte Carlo simulation experiments using the model of effectively neutral mutations. As explained in the previous chapter (see section 8.7), this model assumes that the selective disadvantage, denoted by s', of mutant alleles follows a gamma distribution. The simulation was designed to study the case with $\beta = 0.5$ (see equation 8.98), and I assumed a population consisting of 500 breeding individuals ($N = N_e = 500$). Both the average selection coefficient ($\bar{s}'$) and the mutation rate (v) were varied from experiment to experiment: values of

Fig. 9.6. Relationship between the average heterozygosity and the variance of heterozygosity, as studied by Monte Carlo simulation experiments. Solid circles represent experimental results obtained by assuming the model of effectively neutral mutations with $\beta = 0.5$ (see equation 8.98). Each point (solid circle) is an outcome of 1000 observations made at intervals of 20 generations. In the simulation experiments, a population consisting of 500 breeding individuals ($N_e = N = 500$) was assumed together with the following parameters: the average selection coefficient against the mutants ($\bar{s}'$) was chosen from the range between 0.1 and 0.5, but mostly I assumed $\bar{s}' = 0.2$. As to the mutation rate (v), a value was chosen from the range between 0.00025 and 0.01. For each experiment, $\bar{s}'$ and v were fixed. Open circles in the figure represent experimental results obtained by assuming completely neutral mutations ($\bar{s}' = 0$), and they agreed quite well with the theoretical prediction (represented by the curve in the figure) based on the infinite neutral allele model. For the completely neutral case, simulation experiments were carried out by assuming a population consisting of 100 breeding individuals, in which mutation rate was varied from experiment to experiment to obtain different values of the average heterozygosity. These serve as controls of the simulation experiments to test the model of effectively neutral mutations.



$\bar{s}'$ were chosen from the range $0.1 \sim 0.5$ but mostly I assumed $\bar{s}' = 0.2$. The assumed mutation rates range from 0.00025 to 0.01. Note that the mean and the variance of heterozygosity are not determined independently by values of N_e, $\bar{s}'$ and v, but by the products, $N_e\bar{s}'$ and $N_e v$, so that the present simulation allows us to make inferences regarding large populations (as we usually find in nature) even if we assume a small population size ($N_e = 500$) in the experiments.

The results of the simulation experiments are plotted as solid circles (dots) in Fig. 9.6. For example, the solid dot at the bottom left corner represents the outcome of an experiment with $N_e = 500$, $\bar{s}' = 0.2$ and $v = 0.0004$ so that $4N_e\bar{s}' = 400$ and $4N_e v = 0.8$. Note that this is equivalent to the case in which $N_e = 10^6$, $\bar{s}' = 10^{-4}$ and $v = 2 \times 10^{-7}$. From the figure, it will be seen that if polymorphism is caused by very slightly deleterious mutations that follow the Γ distribution, the observed variance of heterozygosity must be significantly lower than expected from strict neutrality (as shown by open circles and the solid curve), provided that $4N_e\bar{s}'$ is reasonably large, say 400. If $4N_e\bar{s}'$ becomes larger, the downward departure of the observed points from the theoretical curve will become still more pronounced. Although much remains to be clarified, it looks as if the model of effectively neutral mutations with $N_e\bar{s}' = 10 \sim 100$ can explain the observed pattern of heterozygosity of some of the enzymes such as Hk (hexokinase) and Fum (fumarase) which show low heterozygosity and which are subject to strong functional constraints.

It has been pointed out by several authors that soluble enzymes can tolerate mutational changes much more readily than structural proteins and membrane-bound enzymes. Langridge (1974) investigated the effect of amino acid replacements on enzyme function in β-galactosidase of *Escherichia coli*. By isolating mutants possessing 50% or less of normal enzyme activity, he estimated that only 11 out of 733 amino acid substitutions reduced activity of this enzyme below 50%. Furthermore, β-galactosidase appears to be protected by a potential capacity for increased synthesis to counteract partial inactivation. Langridge concluded that most mutations in genes coding for soluble enzymes are nearly neutral to natural selection under customary conditions, although this does not mean that a particular amino acid substitution would be indifferent to natural selection under all conceivable conditions; many of the substitutions would be expressed under different conditions of substrate type or concentration, temperature, pH, etc.

As compared with such soluble enzymes, structural proteins and membrane-bound enzymes are generally subject to stronger functional constraint; structural proteins are closely interlocked with each other, and

membrane-bound enzymes are heavily dependent on accurate steric relations with other molecules in the membrane. Therefore these two classes of proteins are highly invariant and tolerate only a small amount of polymorphism (Jones, 1980). Klose and Feller (1981) investigated the genetic variability of membrane proteins (structure-bound proteins) and cytosol proteins (water-soluble proteins) by comparing two inbred strains of the mouse using a two-dimensional electrophoresis method. They found a large amount of genetic variation in proteins (mostly in the amount of proteins) in cytosol, but no genetic variation was detected among membrane proteins.

Previously, Brown and Langley (1979) investigated, using a two-dimensional electrophoretic technique, the level of genetic variability of the most abundant proteins in *Drosophila melanogaster* adults from a natural population. They obtained the result that the average heterozygosity per locus over approximately 54 loci screened was 4%. This was much lower than the corresponding value, i.e. the average heterozygosity of 14%, obtained for enzyme loci in *D. melanogaster* using the standard method of starch gel electrophoresis. A similar study of the heterozygosity of abundant proteins, using two-dimensional electrophoresis, was made by Racine and Langley (1980) for a wild population of the house mouse. The average heterozygosity estimated over approximately 72 structural loci turned out to be 2% which was much lower than the corresponding value (17.2%) estimated by using starch gel electrophoresis.

The finding that abundant proteins in the tissue show much lower heterozygosity than soluble enzymes can also be explained in part by noting that proteins which exist in major quantities in the cell will have a much larger effect on fitness when their structure is altered by mutation than those existing in minor quantities (see equation 7.11). Note also that these abundant proteins are structural proteins including actin, tublin etc. (McConkey *et al.*, 1979) which are known to be highly conservative in evolution.

In chapter 7, we have discussed extensively how negative selection due to structural and functional constraints of the molecule restricts the occurrence of selectively neutral mutations, thereby slowing down the rate of molecular evolution. If the neutral theory is valid, we should expect that the same set of constraints would reduce intraspecific genetic variability. Various observations discussed in this section appear to be wholly consistent with this expectation. At the moment, our knowledge on the detailed nature of molecular constraints is quite limited. In this respect, Novotny's (1973) account, in detailed physico-chemical terms, of how such

constraints arise through interactions among side chains of amino acid residues and how variability of the protein is influenced by them is instructive, although 'membranin', a molecule used in his explanation, is a fictitious proteins of his invention. It is hoped that we shall soon be able to give such explanations with respect to actual proteins such as hemoglobins, cytochrome *c* etc. (For some recent progress along this line, readers are invited to consult Lesk and Chothia, 1980, and Richardson, 1981.)

I conclude that, if the neutral theory is valid, it will be found that molecules or parts of one molecule which are more important in function, and which therefore evolve more slowly, will show a lower level of heterozygosity. On the other hand, a great deal of polymorphism (not only with respect to base substitutions but also additions and losses of bases) will be found in noncoding regions (particularly in spacer regions), and also in introns even for genes coding for highly conserved proteins. Thus, the factor which governs the level of heterozygosity is not balancing selection, but mutation, population size and negative selection.

In this connection, it is appropriate to mention the work of Latter (1981a, b). He examined published data on 25 enzyme loci for nine species of *Drosophila* (five temperate and four tropical species), and analysed the data separately for group I and group II enzymes (following the classification of Kojima *et al.*, 1970). The contribution to heterozygosity C_t by alleles in various frequency classes (see Latter, 1981b, for details) was treated as a function of allelic frequency $\bar{q}_i$, and this functional relationship was investigated using observations and compared with the corresponding theoretical predictions. Based on computer simulation, he found that the model assuming a combination of loci producing either neutral or very slightly deleterious mutations can adequately account for all aspects of the data. According to him, the consistent difference between group I and group II enzymes can be explained by difference in the average intensity of selection. Also, a highly significant difference between the temperate and tropical species in the distribution of heterozygosity is likely to be due to the smaller effective population sizes of temperate species. He emphasized that the selective disadvantage of mutants in the model of best fit to the *Drosophila* data is of an extremely low order of magnitude, the highest value being $kNs = 3$, where kN is the total species breeding population size (namely, k subpopulations each with the effective size N) and s is the selection coefficient against the mutants. He concluded that the neutral model proposed by Kimura (1968a) and modified by Ohta (1974, 1976) to include the accumulation of slightly disadvantageous mutations, is capable of explaining all features of the data, and that, from an evolutionary point of

view, much of the electrophoretic variation observed in natural populations may justifiably be considered to be effectively neutral.

9.6 Frequency distribution of rare variant alleles

Rare variant alleles, whose frequencies in the population are too low to be regarded as a member of polymorphic systems, can nevertheless supply valuable information on the mechanism by which polymorphism at the molecular level is maintained. This is because, as explained in chapter 8 (see particularly section 8.4), even a mutant allele having a definite selective advantage or disadvantage (assuming $|s|$ to be small but $4N_e|s|$ may be large) behaves as if selectively neutral when its frequency is very low, and therefore we can estimate the mutation rate by using data on the distribution of rare alleles (see equation 8.24 and the explanation that follows). In fact, I have shown, using observations from the plaice and Japanese macaque, that only about 1/10 of amino acid altering mutations on the average are selectively neutral and that the rest are deleterious contributing neither to polymorphism nor evolution. In this section, I shall examine the same problem using observations mainly from human populations for which much more extensive data are available.

Harris *et al.* (1974) reported the incidence of rare alleles determining electrophoretic variants at 43 enzyme loci in Europeans. From their Table 1, I have chosen 26 loci for which the sample size is larger than 1000. The average sample size for them is $\bar{n} = 4058.04$. The average number of rare alleles per locus for them has turned out to be 49/26 or 1.89. Since Harris *et al.* defined rare alleles as those alleles whose individual frequency in the sample was less than 0.005 this corresponds to $q = 0.005$ of equation (8.26), i.e. $M_q = \bar{n}_a(x<q)/\log_e(2\bar{n}q)$. Then, substituting $\bar{n}_a(x<q) = 1.89$, $\bar{n} = 4058.04$ and $q = 0.005$ in this equation, we get $M_q = 0.509$. This gives an estimate for $4N_e v_T$ where v_T is the total mutation rate per enzyme locus for electrophoretically detectable alleles. On the other hand, from equation (8.20), i.e. $M = \bar{H}_e/(1-\bar{H}_e)$, we obtain $M = 0.0753$ by assuming $\bar{H}_e = 0.07$ which is an approximate value for the average heterozygosity per locus due to common polymorphic alleles (Harris and Hopkinson, 1972). This gives an estimate for $4N_e v_0$ where v_0 is the mutation rate for selectively neutral alleles. Therefore, as shown in equation (8.27), we can estimate the fraction of mutations that are selectively neutral among all electrophoretically detectable mutations as follows.

$$P_{neut} = M/M_q \approx 0.15.$$

This value is very close to the corresponding estimate obtained for the plaice in chapter 8.

There is another method for estimating $4N_e v_T$ or M_q which also makes use of the theory that behavior of mutant alleles in general is similar to neutral alleles when their frequency is very low. Let $F_T(x<q)$ be the sum of frequencies per locus of alleles whose frequencies are less than q. Then, assuming that sample size n is large enough so that $q \gg 1/(2n)$, we have

$$F_T(x<q) = \int_0^q x\Phi(x)dx \approx 4N_e v_T q = M_q \cdot q. \tag{9.3}$$

Note that as compared with $\Phi(x)$ which becomes indefinitely large as x approaches 0, the integrand, i.e. $x\Phi(x)$, in the above formula is finite at $x = 0$, so that we can choose 0 as the lower limit of integration without serious error. Therefore, we can estimate $4N_e v_T$ by the formula

$$M'_q = F_T(x<q)/q, \tag{9.4}$$

where M'_q stands for $4N_e v_T$ as estimated by this method. According to Harris *et al.* (1974), the average heterozygosity per locus due to rare alleles is 1.76 per 1000 so that $F_T(x<q) = 1.76/2000$. Noting that $q = 0.005$ in this case, we obtain $M'_q = 0.176$. As compared with $M_q = 0.509$ which was obtained by using equation (8.26), this value is only 1/3 as large. Since equation (9.4) must be very sensitive to the choice of q in the denominator, it is likely that M'_q is more liable to error than M_q and therefore less reliable. So, I shall regard M_q as a better estimator of $4N_e v_T$ than M'_q.

One important observation made by Harris *et al.* (1974) is that 'polymorphic' and 'monomorphic' loci do not differ in the average heterozygosity for rare alleles if the placental alkaline phosphatase, an unusually variable locus, is excluded. This is easy to understand from the neutral theory: the intrinsic mutation rates (v_T) at these two classes of loci are essentially the same, and therefore $F_T(x<q)$ must be nearly equal as is evident from equation (9.3).

Extensive studies of rare variants in human populations have also been done by Neel and his associates for Amerindians, and valuable data have been obtained. I use the data presented in Table 1 of Neel (1978) which lists the occurrence of rare variants at 28 loci in 21 Amerindian tribes. His definition of rare variant alleles corresponds to $q = 0.01$ in our terminology, and from his table we obtain $\bar{n}_a(x<q) = 1.29$ and $\bar{n} = 6442.07$, giving $M_q = 0.266$. An interesting feature of his data is that some of the variants represent what he called 'private polymorphisms', that is, they are concentrated in a single or several related tribes where their frequencies are well above the minimum for a polymorphism. For example, an allele called YAN-2 at the albumin (*Alb*) locus is present in more than 6% of the members of the Yanomama tribe but absent in other tribes. We can

calculate the value of M using the average heterozygosity at 23 loci over 12 tribes as listed in Table 5 of Neel (1978), where we find $\bar{H}_e = 0.054$. Thus, we obtain $M = 0.0571$. Therefore, the fraction of neutral mutations, as estimated by M/M_q, turns out to be $P_{neut} = 0.21$ which is not very different from the corresponding value obtained for European populations (i.e. $P_{neut} = 0.15$).

Incidently, these values of P_{neut} are consistent with the results obtained from the studies on the evolutionary rates of pseudogenes which we discussed in section 7.4. If we use the estimates given in Table 3 of Li *et al.* (1981), the average rate of nucleotide substitutions which cause amino acid changes in the normal globin genes turns out to be roughly 0.14 of the corresponding rate of the globin pseudogenes. From the standpoint of the neutral theory, this value (i.e. 0.14) should represent the fraction of neutral mutations among all the mutations in globins, and considering the fact that the evolutionary rates of globins are near the median of the evolutionary rates of proteins (see chapter 4), such consistency is still another piece of evidence in favor of the neutral theory.

A similar analysis may be made on the data from five species of *Drosophila willistoni* group studied by Ayala and his associates (1974), although the size of sample per locus per species is not as large as in the human data. Of the five species studied, only the *D. willistoni* data is extensive enough for the average sample size per locus to be larger than 500 (in terms of the gene number, i.e. $2n$), so I shall concentrate on this species. From their table 1, which lists allelic frequencies at 31 loci, I have chosen alleles whose frequencies are less than 1% ($q = 0.01$). There are 85 such alleles, so that $\bar{n}_a(x < q) = 85/31 \approx 2.74$. The average sample size per locus is $2\bar{n} = 568.06$. From these values, we obtain $M_q = 1.60$. The average heterozygosity per locus ($\bar{H}_e$) as listed in Table 6 of Ayala *et al.* (1974) is 0.177 from which we obtain $M = 0.215$. Therefore, our estimate for the fraction of neutral mutations among all electrophoretic mutations at the time of occurrence is $P_{neut} = M/M_q \approx 0.13$. It is interesting to note that this estimate is not very different from the corresponding estimates obtained for human populations.

Ohta (1975) was probably the first who noticed that although the observed and theoretical distributions agree quite well under the neutral theory with respect to polymorphic alleles, there is a marked excess of rare alleles in the observed distribution. The present analysis confirms her results. Ohta (1976) went further and showed that the excess of rare alleles is more pronounced in the substrate-specific enzymes than the substrate-nonspecific enzymes. Actually, if we apply the present method for each of these classes of enzymes separately, we obtain $P_{neut} = 0.070$ for the substrate specific group and $P_{neut} = 0.204$ for the substrate nonspecific group, in agreement with her analysis. From these two P_{neut} values, it is evident that the probability of a mutational change being selectively neutral is much smaller for the substrate-specific enzymes than for the nonspecific enzymes. This substantiates the neutralist explanation (as given in the previous section) on the observation that the heterozygosity per locus is much lower for substrate-specific than for the nonspecific enzymes.

Clearly, a detailed study of rare variant alleles is just as important for understanding the mechanism of the maintenance of genetic variability as that of polymorphic alleles. It is hoped that more data on rare variants will be obtained for wild species whose ecologies, particularly the population sizes, are well known.

9.7 Neutral alleles in a geographically structured population

Very often, a mode of selection is inferred from an observed pattern of polymorphism in natural populations. Thus, if the same pair of alleles are found in uniform frequencies over a wide distribution range of the species, it is claimed that natural selection is actively maintaining these frequencies. If, on the other hand, different alleles are fixed in different local populations, or if there is a cline, these are often considered to be the result of local adaptation of these alleles. Furthermore, if the frequencies of alleles are uniform within each locality but different among localities, this is also assumed to indicate the existence of some form of 'balancing selection'.

Actually, selection can be invoked to explain any pattern of polymorphism in natural populations. Often, such presumed selection is used to refute the neutral polymorphism theory. As pointed out by Kimura and Maruyama (1971), these patterns of polymorphism can equally be explained in terms of migration and random frequency drift of selectively neutral mutations.

In the previous chapter, I have explained (see section 8.3) that in the two-dimensional stepping stone model, if migration of at least a few individuals occurs on the average between adjacent colonies or demes in each generation, the frequencies of a polymorphic allele among different colonies becomes essentially uniform. More precisely, it was shown by Maruyama (1970b), based on the mathematical analysis of the two-dimensional stepping stone model, that if N is the effective size of each colony and m is the rate at which each colony exchanges individuals with four surrounding colonies, then the whole population tends to become practically panmictic if

$$mN \geqq 4. \tag{9.5}$$

Actually, an additional condition is needed to ensure the uniformity of

allelic frequencies among colonies. This is

$$N_T v \ll 1, \tag{9.6}$$

where N_T is the effective size of the whole population and v is the mutation rate (see Kimura and Maruyama, 1971). This condition is likely to be met by most organisms with respect to protein polymorphism, because the average heterozygosity seldom exceeds 30%. On the other hand, if

$$mN < 1 \tag{9.7}$$

marked local differentiation of gene frequencies can occur.

It is a well-known observation that, in *Drosophila*, local populations of the same species generally have similar frequencies of polymorphic alleles (see, for example, Ayala *et al.*, 1974). Kimura and Ohta (1971a) explained this from the standpoint of the neutral theory by noting that *Drosophila* species generally have high migrating ability so that condition (9.5) is easily met. Against this explanation, Ayala *et al.* (1974) pointed out that if the whole species becomes practically panmictic, the population sizes of some of the neotropical *Drosophila* species (such as *D. willistoni*) must be immense, amounting to at least 10^9, and probably 10^{10} or even greater. They therefore claimed that even if the mutation rate is as low as $v = 10^{-7}$, the value of $4N_T v + 1$ becomes at least 401, which means that the expected heterozygosity reaches nearly 100%, contrary to observations. I have already discussed this point in section 9.1 of this chapter, so I shall not repeat it here.

In a similar vein, but in more detailed mathematical terms, Bulmer (1973) criticized the neutral theory of protein polymorphism by saying that the infinite neutral allele hypothesis of Kimura and Crow (1964), when combined with Malécot's (1967) two-dimensional dispersal model, leads to results which do not fit actual data. Bulmer claims that there are two contradictions. First, if we use Malécot's formulae for a continuous two-dimensional dispersal model, we get a rather strong geographical localization of allele frequencies, whereas considerable uniformity is observed in the data. He uses Malécot's two quantities: the inbreeding coefficient (f_0) defined as the probability that two homologous loci in an individual are identical by descent, and the coefficient of kinship (f_x) defined in a similar way, between two individuals at a distance x apart. Using Malécot's formulae for f_0 and f_x, Bulmer calculates the ratio, $\rho_x = f_x/f_0$. The numerical values of ρ_x presented in his examples certainly decrease quite rapidly as x increases. On the other hand, the estimates of this ratio using data from *Drosophila pseudoobscura* (Prakash *et al.*, 1969) are very close to unity even for very distant localities, such as Texas, Colorado and

California. He claims then that such a discrepancy cannot be resolved except under the absurd assumption that the root mean square dispersal distance, σ, is of the order of 2000 km.

Secondly, Bulmer pointed out that the expected degree of homozygosity (f_0) would be much less than that of the data, unless the gene dispersion is very restricted. The average homozygosity found by Prakash *et al.* (1969) is 0.877. He claims, using Malécot's formula for f_0, that to make f_0 be about this magnitude, the mean square dispersal distance between the breeding places of parent and offspring (σ^2) must be an absurdly small value. For example, if the mutation rate v is 10^{-5}, then the number of flies within a circle of radius σ, that is, $\pi\sigma^2\delta$ must be about 0.38, where δ is the density of the breeding population. This seems to be not only untrue, but if the dispersion is restricted to this degree, ρ_x should decrease very rapidly.

Maruyama and Kimura (1974), in their defence of the neutral theory, emphasized the important point that Malécot's formulae used by Bulmer are valid only for cases where the total population number (N_T) is much greater than the reciprocal of the mutation rate (v) and the habitat is large. The size of habitat alone, however, is not sufficient, as long as the population is finite. Of course, if the population is truly infinite, his formulae are valid. On the other hand, if N_T and $1/v$ are about the same order of magnitude, it is necessary to use formulae valid for a finite population. For such a population, Malécot's formulae cannot be used even as an approximation, because they give entirely different values. Exact analysis for a finite torus-like space of size $L_1 \times L_2$ was made by Maruyama (1972b). For the more realistic case of a rectangular habitat of size $L_1 \times L_2$, although exact solutions were not obtained, it turned out that solutions of the torus-like space of size $2L_1 \times 2L_2$ provided approximate answers. Thus, Maruyama and Kimura (1974) assumed the torus-like space and used Maruyama's formulae to re-analyse the problem. In their analysis, they assumed $L_1 = L_2$ and carried out extensive numerical calculations by a computer.

Contrary to Bulmer's (1973) claim, they found that the theoretical expectations and data are consistent. If $N_T v$ is 1 or less and $\sigma^2\delta$ is of the order of 100 or 10 or even less, ratio $f(x,y)/f_0$ stays very close to 1, and it is roughly independent of distance and habitat size, where $f(x,y)$ is the probability that two homologous genes separated by distances x and y along the first and second axes of the Cartesian co-ordinate system are identical by descent. Furthermore, if $4N_T v = 0.1 \sim 0.2$, the value of f_0 turns out to be about the right order of magnitude consistent with the neutral theory. Maruyama and Kimura (1974) emphasized that the finiteness of

population makes the situation completely different from that of a truly infinite case. Thus, Bulmer is unwarranted in using Malécot's formulae which assumes an infinite population size. This led him to erroneous conclusions which seemed to be inconsistent with data: Bulmer's dilemma can be simply resolved by using more appropriate formulae.

Baker (1981) investigated a song dialect population of white-crowned sparrow (*Zonotrichia leucophrys nuttalli*) occupying a natural environment of scrub vegetation along the central California coast. Roughly speaking, this population consists of 2000 birds, and he estimated that the neighborhood size in the sense of Wright (1946) is $N = 36$ consisting of about 100 birds. According to him, each dialect group of this species is significantly isolated from other dialects, and calculation of Wright's statistic F_{ST} from observed isozyme frequencies within dialects gives an average $F_{ST} = 0.0046$. He also computed the theoretical F_{ST} using Wright's (1951b) neighborhood model and obtained a value which is an order of magnitude larger than the observed F_{ST} (for example, if there are 20 neighborhoods each with the effective size 36, one gets $F_{ST} = 0.0475$). He reasoned that natural selection may therefore be maintaining greater homogeneity within a dialect than expected by assuming that alleles are neutral. However, he also noted, that the torus model of Maruyama (1972b), by assuming the mutation rate $v = 2 \times 10^{-6}$, gives the expected F_{ST} values which are close to the observed value suggesting no major departure from panmixia within dialect populations.

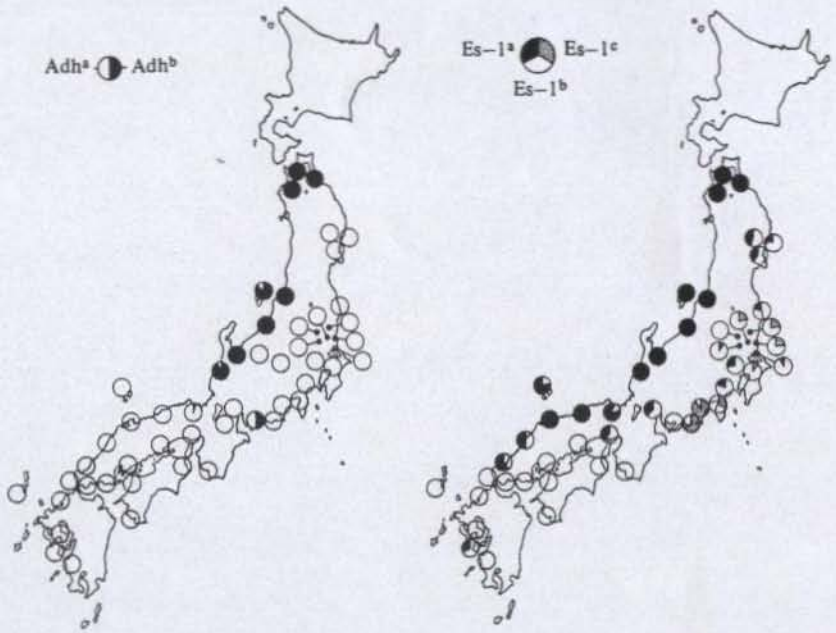
I would like to close this section by giving a few additional examples which suggest that isolation and random drift are the major determining factors for the geographic distribution of allelic frequencies. Sakaizumi *et al.* (1980) studied the geographic distribution of alleles at 25 enzyme loci in *Oryzias latipes*. This is a tiny freshwater fish native to Japan and nearby regions, living in brooks, ponds and paddy fields. Of the 25 loci studied, five showed clearcut differentiation of allelic composition between populations along the coast of the Sea of Japan on the one hand and those along the Pacific coast on the other. Fig. 9.7 illustrates such a pattern of allelic distribution for alcohol dehydrogenase (*Adh*) and esterase 1 (*Est*-1). These patterns can be understood by noting that the two sets of populations are isolated from each other by the mountain range forming the backbone of mainland Japan. It is clear that this mountain range has acted as an unsurpassable barrier for this freshwater fish for a very long time, and that random fixation or near fixation of different alleles has occurred at these five loci.

In sharp contrast, no such differentiation of allelic composition was found for the fruit fly *Drosophila virilis* investigated by Ohba (1977) between

the two coastal regions. He scored allelic frequencies extensively over these regions at seven enzyme loci including *Est-α*, *Est-β*, *Adh* etc. The result is easily understandable if we note that this insect is well able to migrate so that no local population is sufficiently isolated from others.

Finally, I intend to discuss an example where the condition (9.6) does not seem to apply. Some years back, Robertson (1968) commented that if a large proportion of mutations at a locus is selectively neutral, we should find either very many alleles segregating in large populations, or a small number of different sets of alleles in different isolated small populations. He considered that because neither of these alternatives is found, the great majority of polymorphisms have, at some time, been actively maintained by selection. Kimura and Ohta (1971a) pointed out that both the situations suggested by Robertson are typical of the heterochromatic pattern polymorphism (as revealed by cold-treatment) of chromosomes in wild populations of the perennial plant *Trillium kamtschaticum*. Extensive cytological studies of this plant by Haga and his associates (Haga and

Fig. 9.7. Geographical distribution of alleles at alcohol dehydrogenase locus (left-hand map) and esterase 1 locus (right-hand map) of the 'medaka' fish (*Oryzias latipes*) in Japan (quoted from Sakaizumi, 1982 with slight modification).



Kurabayashi, 1954; Haga, 1969) have shown that several chromosome types are segregating within a large population, while different types are fixed in small isolated populations. Indeed, Robertson's suggestion is pertinent if isolation between subpopulations is nearly complete and if the mutation rate for neutral isoallelic variations is sufficiently high so that more than one new mutant appears within a large population each generation. The chromosome polymorphism in *Trillium* can be explained by assuming a relatively high mutation rate per chromosome for heterochromatic variation and very low migration rate per generation for this plant.

10

Summary and conclusion

Traditionally, studies of evolution were concerned only with the visible structures of organisms, such as birds' wings, giraffes' necks, horses' teeth and the like, and their function. The guiding principle by which evolutionary changes in such phenotypes were interpreted was Darwin's theory of natural selection, or the survival of the fittest. In Darwin's time, the mechanism of inheritance and the nature of heritable variations were unknown and this caused Darwin difficulty, but with his great insight he was able to grasp the real significance of natural selection in evolution.

With the rise of Mendelian genetics, the difficulty which troubled Darwin was gradually resolved, and with the development of population genetics, the synthesis of Darwinism and Mendelism was achieved. This led to the 'synthetic theory' of evolution which flourished and grew into an impressive edifice toward the end of 1950s. Although various factors such as mutation, recombination, and migration were taken into account, the dominating feature of the 'synthetic theory' was, and still is, its great emphasis on natural selection. It claimed that the speed and direction of evolution are predominantly determined by positive selection with mutation playing only a subsidiary role. Because of this character, it has also been called the neo-Darwinian theory.

By the early 1960s the consensus seems to have been reached that every biological character can be interpreted in the light of adaptive evolution by natural selection. Also, it was generally agreed that almost no mutant genes are selectively neutral, and that random genetic drift does not play any significant role in evolution, except as a factor in Wright's 'shifting balance' theory. From such a theoretical viewpoint, and based on various studies such as genetical analyses of natural populations, paleontological investigations, ecological genetics and speciation studies, a great deal has

been written on how gene pools of species are organized and how they change in evolution. It looked as if the mechanism of evolution could be understood exclusively in neo-Darwinian terms.

In reality, however, these statements were inferences based on observations at the phenotypic level, and there was no way of actually knowing what is going on in evolution at the molecular level, that is, at the level of the internal structure of the gene. At that time, much importance was attached to epistasis and overdominance in explaining the genetic structure of natural populations. However, little was actually known as to how much variability exists within the species at the molecular level.

With the advent of molecular genetics, and through the introduction of its methods and concepts to studies of evolution, the situation has changed. It became possible, through comparative studies of amino acid sequences of homologous proteins among related organisms (such as comparison of hemoglobins among vertebrates) and the use of paleontological data, to estimate the rate and pattern of amino acid substitutions in evolution. This enabled us to estimate the evolutionary rates of nucleotide substitutions inside the genes. These rates could then be extrapolated to the whole DNA content of the organisms. The resulting figure suggested that the rate of accumulation of mutations in the species in the course of evolution is much higher than previously thought. The development of new electrophoretic techniques permitted the rapid detection of enzyme variability among individuals, and these studies have disclosed a wealth of polymorphic variations at the enzyme level in many organisms. They suggested that genetic variability is much higher than previously presumed.

All these developments brought us many fascinating observations and at the same time some puzzling questions. The proposal of the neutral theory (or more precisely the neutral mutation-random drift hypothesis) represented my attempt to answer these questions using the stochastic theory of population genetics. Unlike the traditional synthetic theory (or the neo-Darwinian view), the neutral theory claims that the great majority of evolutionary mutant substitutions are not caused by positive Darwinian selection but by random fixation of selectively neutral or nearly neutral mutants. The theory also asserts that much of the intraspecific genetic variability at the molecular level, such as is manifested in the form of protein polymorphism, is selectively neutral or nearly so, and maintained in the species by the balance between mutational input and random extinction or fixation of alleles. From the standpoint of the neutral theory, evolutionary mutant substitutions and molecular polymorphisms are not two independent phenomena, but simply two aspects of a single phenomenon. In

other words, protein polymorphism merely represents a transient phase of molecular evolution. The neutral theory is accompanied by a well-developed mathematical theory, and it attempts to treat quantitatively various problems of molecular evolution and polymorphism from the standpoint of population genetics.

Against the neutral theory, strong opposition and criticism have been leveled by the 'selectionists' who adhere to the traditional synthetic theory. They consider that evolutionary mutant substitutions must be adaptive and are caused by positive Darwinian selection. They also regard protein polymorphism as adaptive and claim that it is maintained in the species by some form of balancing selection. In other words, they invoke two different types of natural selection to explain these two phenomena. I also know that not a few biologists question the validity of the neutral theory, and that the great majority are not yet convinced of its correctness.

Some of the criticisms of the neutral theory are based on misunderstandings. For example, the theory does not assume that all the mutations at the time of their occurrence are selectively neutral (equivalent). Rather it assumes that a certain fraction of mutations are deleterious and the rest are selectively neutral, and that this fraction depends on the functional constraint on a molecule: the stronger the constraints, the larger the fraction of deleterious mutants, and therefore the smaller the chance of a mutation being selectively neutral. Favorable mutations no doubt occur in the course of evolution, but the theory assumes that these are too rare to be of significance in explaining the major features of molecular evolution as revealed by comparative studies of protein and DNA sequences, and intraspecific genetic variability as detected by electrophoretic and other refined techniques of molecular biology.

In this book, I have endeavored to show that, not only has the neutral theory survived various tests designed to disprove it, but has suddenly gained support through data coming from the recent molecular revolution of cloning DNA segments and rapidly determining their nucleotide sequences. Let me mention a few pieces of such evidence. It has been established beyond doubt that synonymous substitutions (that cause no amino acid changes) within protein-coding regions of the genome occur at much higher rates in evolution than amino acid-altering substitutions, sometimes in a dramatic fashion for highly conserved proteins such as histone H4 and tubulin. Since natural selection acts through phenotypes of the organism for which the structure and function of proteins play a decisive role, one should expect that the mutations that do not cause amino acid changes in proteins, other things being equal, are much less subject to

natural selection than those that cause amino acid changes. Yet, the undeniable facts which have emerged during the last few years are that synonymous changes and also 'silent' nucleotide changes in the non-coding regions of DNA are among the most prevalent evolutionary changes at the molecular level. These observations are quite similar in nature to those made on amino acid substitutions in proteins. For example, when active insulin is formed from proinsulin, the middle segment (C) of proinsulin is removed and discarded. It is known that for this peptide C the rate of evolution in terms of amino acid substitutions is several times as fast as that of insulin.

More recently, it has been discovered that most eukaryotic genes contain intervening sequences or 'introns' which are removed when the mature messenger RNA is formed and which therefore do not participate in protein formation. We have seen that evolutionary nucleotide substitutions are very rapid in introns, Still more remarkable is the finding that in pseudogenes ('dead' genes), which are homologous to normal counterparts but have lost their function as genes, evolutionary nucleotide substitutions are very rapid, and their evolutionary rates are roughly equal to or even slightly higher than those of synonymous substitutions. A general rule that has emerged from these observations is that *molecular changes that are less likely to be subject to natural selection occur more rapidly in evolution.* This empirical rule can readily be understood from the standpoint of the neutral theory, because such molecular changes have a higher chance of being selectively neutral (i.e. selectively equivalent), and therefore neutral evolution by random drift occurs at higher rates. As mentioned already, the neutral theory assumes that a certain fraction of mutations are deleterious, depending on the extent of functional constraint. Such constraint must be very weak or non-existent for synonymous changes and also nucleotide changes in the pseudogenes, so that most or practically all mutations are neutral. According to the neutral theory, the weaker the functional constraint, the larger the fraction of mutations that are selectively neutral and therefore, the higher the evolutionary rate. On the other hand, this empirical rule clearly contradicts the neo-Darwinian interpretation of mutant substitutions, because for a more rapid accumulation of mutants to occur in the species, stronger selection for the mutants is required. The only solution in neo-Darwinian terms is to insist that these rapidly evolving parts have some unknown functions, and that they are undergoing rapid adaptive evolution by accumulating advantageous mutants.

Traditional studies of evolution have amply demonstrated that evolution at the phenotypic level is characterized by adaptation and opportunism,

irregularity in pace, and inequality of rates among lineages. In contrast, studies of molecular evolution have revealed quite different features characterized by changes that are conservative in nature, random in pattern (independent of phenotypic characters), and quite regular in pace with equal rates among diverge lineages for a given protein.

The last mentioned feature, that is, approximate constancy of the evolutionary rate in terms of amino acid substitutions, has become one of the most controversial subjects in molecular evolution. That the constancy is in terms of years rather than in generations makes this an especially remarkable feature of molecular evolution. Reactions among students of evolution to this phenomenon have been diverse. Some flatly deny the very existence of constancy in the rate of amino acid substitutions. In sharp contrast, there are those who strongly believe in its existence in the strict sense and consider it as axiomatic, but pay little attention to the underlying mechanism.

In this book I have tried to show that the phenomenon of rate-constancy for a given protein does exist, albeit the constancy is an approximate one. For a few proteins including globins, the observed variance of the evolutionary rate among lines (mostly among mammals) for a given protein is slightly larger than what is expected from statistical fluctuations. For example, the ratio (R) of observed to expected variance of the evolutionary rate for myoglobin turned out to be 1.7 among mammals, including human, harbor seal, badger, horse, bovine, and sperm whale. In this case, deviation from the assumed rate-constancy is not statistically significant. In other cases, we have seen that values of R range from 1.3 to 3.3. It is remarkable that evolutionary amino acid substitutions have proceeded at a constant rate among diverse lines of mammals whose living conditions and habits are very different. This constancy may be explained from the neutral theory by assuming that, for a given protein, the rate of occurrence of neutral mutations is constant per year among diverse lineages.

This explanation is based on the following theorem applicable to neutral alleles: the rate (k) by which mutant alleles are substituted at a given locus (or site) in the course of evolution within the species is equal to the mutation rate (v). In other words, if we restrict our consideration to selectively neutral mutants, we have $k = v$, provided that both quantities are measured in the same time unit. On the other hand, if mutant substitions are carried out by natural selecting acting on definitely advantageous mutants, it can be shown that $k = 4N_e\bar{s}_A v_A$, where N_e is the effective population size of the species, v_A is the mutation rate for advantageous alleles and $\bar{s}_A$ is the average selective advantage for such alleles. In this case, in order to have a constant

k value, the product $N_e\bar{s}_A v_A$ must remain constant in the course of evolution among diverse lineages. Considering the possibility that N_e, $\bar{s}_A$ and v_A may each vary widely among different organisms such as human, harbor seal, badger, horse, bovine and sperm whale, it is highly unlikely that this product remains constant for all of them. This poses a difficulty for the selectionist explanation of the rate-constancy phenomenon. Sometimes it is stated that the rate-constancy can be explained by assuming an equal amount of 'selection pressure' among different lineages, but this is clearly wrong – constant $\bar{s}_A$ does not guarantee constant k. Furthermore, in each generation the main part of the selection pressure for each species comes from stabilizing selection, and this has nothing to do with the rate of adaptive mutant substitutions.

The problem of rate-constancy in molecular evolution becomes much more complicated if this is extended to include organisms having widely different generation spans such as *Drosophila*, mouse and man. Whether the rate-constancy rule is applicable to such a case is not absolutely certain, but it is likely that the rule holds approximately. This has been used by some to criticize the neutral theory as follows: mutational studies suggest that the spontaneous mutation rate per generation, but not per year, appears to be roughly equal among animals whose generation spans are very different, and in view of the theorem $k = v$ for neutral mutations, we should expect, if the neutral theory is valid, that the evolutionary rate is constant per generation rather than per year, contrary to observations.

I must admit that this is a difficult problem for the neutralists, although an explanation in terms of positive selection is even more difficult. There are at least two ways to solve this problem within the framework of the neutral theory. One is to assume that the occurrence of neutral mutations is proportional to years, even if visible and lethal mutations occur at the rates which are comparable per generation between organisms with very different generation spans. For me, this appears to be unsatisfactory in the light of observations on pseudogenes, because the extraordinarily high evolutionary rates observed in them must be explained by assuming that mutations which had previously been definitely deleterious changed into selectively neutral mutations after the original genes lost their function to become pseudogenes. The rates of evolution of pseudogenes, however, do not appear to be proportional to the generation span, although data are still insufficient to make sure of this.

Another explanation for the rate-constancy per year phenomenon is to resort to Ohta's hypothesis of very slightly deleterious mutations. According to this hypothesis, what we call neutral mutants are not strictly

neutral, but are in fact very slightly deleterious, with values of the selection coefficient against them (denoted by s') comparable to the mutation rate or only slighter larger. Such very slightly deleterious mutants behave as if selectively neutral and are subject to extensive random drift in relatively small populations, whereas they are effectively selected against in very large populations. From this standpoint, the truly neutral mutants are the limit of selectively disadvantageous mutants when their disadvantages become indefinitely small, rather than the limit of advantageous mutants. Suppose that the frequency distribution of selective disadvantage of newly arising mutants is highly skewed such that those with smaller s' values occur more frequently. Then, a large fraction of mutations are effectively neutral in a species whose effective population size is small. As the population size becomes larger, more and more mutants become selected against so that a smaller fraction of mutations are effectively neutral. There is a general tendency for an animal species with a larger body size to have a longer generation span and at the same time a smaller effective population size (N_e). This means that an animal species with a longer generation span tends to have a smaller population size and therefore show a higher rate of production of mutant alleles which behave as if selectively neutral. Thus, the mutation rate for effectively neutral alleles per year becomes more or less constant among different animals. My model of 'effectively neutral mutations' assuming that s' follows a Gamma distribution is an extention and quantification of Ohta's idea. In this new formulation, the rate of evolution per generation is equal to the rate of 'effectively neutral mutations'. I have particularly considered the case with the parameter $\beta = 0.5$ for which the rate of evolution per generation (k_g) turns out to be inversely proportional to the square root of the effective population size, i.e. $k_g \propto 1/\sqrt{N_e}$. If the generation span (g) is inversely proportional to the square root of the effective population size, then the evolution rate per year is $k_1 = k_g/g \propto (1/\sqrt{N_e})/(1/\sqrt{N_e}) = 1$, which means that the evolutionary rate per year is constant, provided that the mutation rate per site per generation is constant. Needless to say, whether such an explanation is valid or not has to be examined in the future. Note that N_e here refers to the effective population size relevant to long-term evolution, because the rate of mutant substitutions is measured in terms of tens of millions of years. Therefore, testing of this explanation against *Drosophila* data using current population sizes of various species is not adequate. This is because not only are reliable estimates of long-term evolutionary rates for proteins not available for *Drosophila*, but the relative abundance of various species is changeable, and estimation of the relevant effective population sizes of

various species using their current population sizes can be highly misleading as estimates of past values.

Perhaps, comparison of mammalian species which have very different generation spans will be more suitable for this purpose. For example, we may compare the elephant and the mouse. It is likely that the evolutionary rates (per year) at the molecular level in these two lineages are the same. The elephant probably has the generation span of about 20 years, while that of the mouse is about half a year, so that the average generation span of the elephant is roughly 40 times as long as that of the mouse. If the relationship $g \propto 1/\sqrt{N_e}$ applies to these two animals, the effective population size of the mouse is expected to be some 1600 times as large as that of the elephant, and this appears to be a realistic figure. One additional question which we should not forget in relation to this comparison is at what stage of the evolutionary history of these two animals such a wide difference of generation span evolved. In other words, if ancestors of these animals had retained similar generation spans for a long time after divergence, this has to be taken into account. Whatever the generality of the rate-constancy hypothesis, if the neutral theory is correct, it should be able to supply an adequate explanation of why the rate-constancy hypothesis holds when it is valid, and at the same time it should be able to give an explanation when it fails (if such cases exist).

The conservative nature of mutant substitutions is also a feature of molecular evolution. This is the property that substitutions which are less disruptive to the existing structure and function of a molecule occur more frequently in evolution than more disruptive ones. This has most clearly been demonstrated by studies of protein evolution, showing that chemically similar amino acids are substituted more frequently in evolution than more dissimilar ones. From the standpoint of the neutral theory this can readily be explained by noting that the probability of a mutation not being harmful and therefore selectively neutral is higher if the mutation leads to a similar amino acid. The synonymous substitutions which I have mentioned already represent that limiting situation in which the amino acid difference vanishes.

On the other hand, the selectionist's explanation saying that a mutational change with smaller phenotypic effects has a higher chance of being advantageous as shown by Fisher, and is therefore incorporated more frequently in evolution, is clearly false. Such an explanation overlooks the crucial point that a mutation, if its effects are very small, is likely to have a very small selective advantage and therefore has a correspondingly small probability of fixation in the species. This is one good example to show that some knowledge of the stochastic theory of population genetics is needed

for meaningful discussion of the mechanism of molecular evolution. In particular, the theory treating the behavior of mutant alleles in a finite population is important: in any population of appreciable size, a large number of mutants arise in each generation, but a great majority of them, including those which have a small selective advantage, are lost from the population within a few generations (say, 10). Only a small fraction can spread to the whole population to reach fixation, and this takes a very long time (say 100 000 generations). In many discussions of evolution, it is tacitly assumed that every advantageous mutation that appears in the population is inevitably incorporated in the species, but this is not true. We must be careful not to be misled by such an assumption when we discuss the mechanism of molecular evolution.

Let us now consider the relationship between functional constraint and the rate of evolution. It has become increasingly clear that the weaker the functional constraint to which a molecule or a part of a molecule is subject, the higher the evolutionary rate of mutant substitutions. Roughly speaking, this means that functionally less important molecules or parts of a molecule evolve faster than more important ones. We have already touched on this problem when we discussed the rule that molecular changes which are less likely to be subject to natural selection occur more rapidly in evolution. We have seen such examples as the middle segment (C) of proinsulin which is removed when active insulin is formed, and intervening sequences (introns) which are not included in the mature messenger RNA. In these parts, the rate of evolution in terms of mutant (amino acid or nucleotide) substitutions is very high. We have also seen that in hemoglobins the surface part evolves about 10 times as fast as the heme pocket, in accordance with the known fact that the surface part generally has many fewer constraints in maintaining the structure and function of the molecules than the heme pocket which is vitally important.

Let us consider the problem in more quantitative terms based on the neutral theory. Suppose that a certain fraction f_0 of the mutants are selectively neutral and that the rest of the mutants $(1 - f_0)$ are definitely deleterious. Although advantageous mutations may occur in rare occasions, we assume that their frequency is negligible in determining the overall rate of molecular evolution. If we denote by v_T the total mutation rate per site, then the mutation rate for neutral alleles is $v = v_T f_0$, and from the theorem, $k = v$, applicable to neutral mutants, we have

$$k = v_T f_0. \tag{10.1}$$

We now assume that the probability of a mutational change being neutral (i.e. not harmful) depends strongly on functional constraints. The weaker

the functional constraint, the higher the probability f_0 of a random change being neutral, with the result that k in equation (10.1) increases. Clearly, the maximum evolutionary rate is attained when $f_0 = 1$, namely, when all the mutations are neutral ($k = v_T$). Here, we assume that the total mutation rate per site v_T is the same among different molecules. In my opinion, the very high evolutionary rate observed for pseudo globin genes in mammalian evolution (roughly $k = 5 \times 10^{-9}$/nucleotide site/year) must be very near to this limit. This is roughly 10 times as high as the amino acid altering substitution rate (expressed per nucleotide site) in the evolution of normal globin genes. Also, this is nearly twice as high as the synonymous substitution rate, and this suggests that even synonymous substitutions are subject to negative selection, although the intensity of selection involved must be exceedingly weak. This is consistent with the observation that synonymous codons are used in 'non-random' or unequal fashion.

In fact, non-random codon usage is a quite common phenomenon, and furthermore there is a consistent pattern of usage for various genes within the genome of a species. There is good evidence now which suggests that choice of a synonymous codon is largely constrained by tRNA availability. This leads to a hypothesis that preferential codon usage represents the optimum state in which the population of synonymous codons matches that of cognate tRNA available in the cell. Then, the concept of stabilizing selection in quantitative genetics can be applied to treat the problem using the diffusion equation method. The results indicate that the observed pattern of non-random usage of synonymous codons and its effect on retarding the fixation of synonymous mutants can be explained reasonably well in quantitative terms by this model. It is reassuring that the universal phenomenon of non-random codon usage can be explained satisfactorily within the framework of the neutral theory in terms of selective constraint. This is particularly so, because this phenomenon has often been mentioned as evidence against the neutral theory.

More generally, it has been shown that, if a very large number of segregating loci (or sites) are involved in a quantitative character which is subject to stabilizing selection, the average selection coefficient per mutant may be exceedingly small. These mutants are very slightly deleterious but nearly neutral, so that mutant substitutions are mainly controlled by random drift, although the rate of evolution may be lower than when all the mutations are strictly neutral. This demonstration, I believe, increases the credibility of the view that neutral evolution can occur extensively at the molecular level even when natural selection predominates at the phenotypic level. In our treatment of the problem of non-random usage of synonymous codons, we assumed that the pattern of codon usage changes

in the course of evolution in such a way that they match those of cognate tRNAs available in the cell. There is also the reverse possibility of the relative frequencies of cognate tRNAs (as determined by the numbers of their genes) being adjusted in the course of evolution so that they match the relative frequencies of synonymous codons in the genome of the species. We do not know at present which one of these two may be regarded as the primary cause. It is quite likely that the influence is reciprocal, each regulating the other's change, and they shift very slowly in unison. This serves as a good example at the molecular level of 'mutual evolution', the concept which Komai (1970) emphasized as important for our understanding of the mechanism of evolution. Although he mainly considered the aspect of mutual evolution in accelerating the speed of phenotypic evolution, the concept must apply equally to cases showing unusual stability in evolution. This probably explains why a given pattern of synonymous codon usage characterizes a large taxonomic group, such as found in the similarity of usage pattern among mammals.

These considerations suggest that the amount of selective constraint against synonymous changes is essentially the same for different genes within a genome. This explains why the rates of synonymous substitutions are not only high, but are similar to each other even between genes whose amino acid substitution rates are widely different. For example, α-tubulin in vertebrates is an extremely conservative protein, with the estimated rate of evolutionary amino acid substitution $k_{aa} = 1.15 \times 10^{-11}$ per amino acid site per year, which is only 1/100 as fast as that of hemoglobins. Yet, the synonymous component of the rate of nucleotide substitutions at the third positions of the codons is essentially the same between α-tubulin and β-globin. A similar observation has been made with respect to histone H4, which is known as the most conservative protein so far studied. It is remarkable that synonymous substitutions have occurred in histone H4 at a very high rate comparable to those of other proteins whose amino acid substitution rates are much higher.

Existence of a maximum evolutionary rate set by the total mutation rate, in accordance with the formula $k \leqq v_T$, is one of the very successful predictions of the neutral theory; this prediction is now vindicated by comparative studies of DNA sequences which have suddenly become possible on a large scale during the last few years, thanks to the new revolution in molecular biology. Incidentally, the claim was once made by selectionists, based on early results obtained by the DNA hybridization technique, that synonymous substitutions are not as frequent as predicted by the neutral theory.

During the last few years, observations showing the prevalence of

duplicated or repeated structures within the genome of higher organisms have accumulated rapidly. It looks as if almost every gene in eukaryotes has its duplicated counterpart somewhere in the genome. These observations support the earlier notion that gene duplication plays an important role in progressive evolution: in order to produce a more complicated organization, more genetic information is needed, and gene duplication is the only way to achieve it. Furthermore, what is crucial is that duplications enable genes to make evolutionary experiments which have previously been forbidden, liberating them from incessant natural selection whose overwhelming activity is eliminating variants. In other words, gene duplications create conditions which enable random genetic drift to operate much more prominently on mutants, namely, many mutations which would have been definitely deleterious before duplication become harmless (i.e. selectively neutral) or only very slightly deleterious after duplication, thus enabling them to spread in the population by random drift. Some of them will then turn out to be useful under new environmental conditions. Thus, the neutral theory has a bearing on the problem of progressive evolution. This process that facilitates the production of new genes will, at the same time, cause degeneration of one of the duplicated copies. In fact, the probability of gene duplication leading to degeneration must be very much higher than that leading to production of a new gene having some useful function.

In relation to gene duplication, there is an important problem of how the duplication itself became fixed in the population in the first place. A popular explanation invoking permanent heterozygote advantage as its cause is unsatisfactory. A much more plausible explanation is to resort to neutral evolution, namely, a tandem repeat of the form AA (where A is a gene locus) is created first by intra-chromosomal duplication, and when this is selectively neutral or only very slightly deleterious, it becomes fixed in the species by random drift. Like ordinary point mutations, gene duplications are constantly occurring in the population, and the continued mutational pressure for intra-chromosomal duplication in conjunction with random drift, during a long course of evolution, must be the main cause for the widespread occurrence of repeated structures in the genome of higher organisms. The prevalence of pseudogenes and other nonfunctional 'junk DNA' can be understood in similar terms; so also can the evolutionary behavior of 'selfish DNA'. (For additional new information on these topics, see Dover and Flavell, 1982.)

The multigene family, consisting of a large number of tandemly repeated genes, is an extreme form of gene duplication. Notable examples of multigene families are: immunoglobulin genes, ribosomal RNA genes and

histone genes. They exhibit a newly discovered evolutionary phenomenon known as coincidental evolution, which is also called horizontal or concerted evolution. This is a phenomenon such that members of one multigene family evolve in unison, while retaining intraspecific homogeneity among themselves. The most plausible explanation of this phenomenon is to assume that, through repeated intra-chromosomal unequal crossing-over, a mutation which occurred in one of the members spreads through the entire family by chance (horizontal spreading of a mutation). This is analogous to random fixation of a neutral mutant within a finite population by sampling drift. Thus, the evolutionary process of a multigene family can be treated as a double diffusion process: diffusion of mutants on one chromosome by unequal crossing-over (or by gene conversion) and diffusion of such chromosomes in the population by ordinary random drift caused by sampling of gametes.

Let us now turn to the problem of intraspecific genetic variability. Through extensive electrophoretic studies of protein polymorphism during the last decade and a half, it has been amply demonstrated that natural populations of various organisms contain a large amount of genetic variability at the molecular level. Then, the problem arises: what is the mechanism by which such abundant genetic variability is maintained? This has been regarded by many population geneticists to be the outstanding problem in the field. Opinions differ widely among those concerned with the problem, and in fact, they are more or less polarized into two opposing stands: the selectionist side insists that the majority of polymorphic alleles are actively maintained by some form of balancing selection, while the neutralist side maintains that these alleles are selectively neutral and drifting randomly in the population. Numerous experiments have been designed and performed to examine whether polymorphic alleles are selectively neutral. Also a number of statistical tests have been devised and applied to examine whether the observed distribution of polymorphic alleles in natural populations conforms to the predictions of the neutral theory. Not only has the neutral theory of protein polymorphism thus been subjected to severe tests of many kinds, but various rival hypotheses have been proposed which incorporate natural selection. As I have detailed in the previous chapter, with increasing data, apparent difficulties of the neutral theory have gradually been resolved, and supporting cases have accumulated.

Many of the laboratory results suggesting clearcut selective differences of polymorphic alleles appear not to be valid, and often, when more careful studies were repeated later, they failed to confirm the reported selection. In

other cases, it is not clear whether the selective effects come from the polymorphic loci themselves or from surrounding linked loci. A claim that the adaptive nature of protein polymorphism can be unambiguously revealed, case after case, by applying the methods of ecological genetics has not been fulfilled. So far, the case for natural selection appears to be strongest at the alcohol dehydrogenase locus (*Adh*) in the fruit fly *Drosophila*, but experimental results are conflicting, and recent careful studies indicate that the true difference in fitness between alleles *F* and *S* at this locus, if it exists, must be exceedingly small. Furthermore, the existence of frequency-dependent selection is much in doubt.

In the past, overdominance or heterozygote advantage was proposed enthusiastically by many selectionists as the major cause of widespread protein polymorphism. However, evidence against the overdominance hypothesis has accumulated so much in recent years, including the finding of widespread polymorphism in haploid organisms, that this can no longer be regarded as a valid scientific hypothesis.

As to testing selective neutrality of protein polymorphism by examining the distribution pattern of allelic frequencies, a number of statistical procedures have been devised and applied, but it has turned out that each test has its shortcomings. So far, no definite evidence against the neutral theory has been obtained. Rather, favorable cases for the theory have been added through Nei's test of examining the relationship between the mean and the variance of heterozygosity; observations from diverse organisms conform quite well to the prediction of the neutral theory.

At one time some authors claimed that linkage disequilibrium (i.e. non-random association of alleles between loci) is prevalent among isozyme loci in natural populations, and these authors emphasized its significance for the maintenance of widespread protein polymorphism. However, as data from actual observations accumulate, it has become increasingly evident that there is virtually no linkage disequilibrium among enzyme loci in *Drosophila*, unless the gene loci are associated with inversions. Furthermore, the last mentioned association can most adequately be explained by assuming historical accidents and random drift: it can be shown mathematically that, if the rate by which alleles are exchanged between the inversion and non-inversion chromosomes is much lower than the reciprocal of the effective number of inversion chromosomes, significant non-random association between the alleles and inversions will develop.

It is likely that, if cases of strong linkage disequilibrium are found in natural populations, loci involved are confined within a very small chromosomal region, within which occurrence of recombination is exceedingly rare. Also, it is likely that such a disequilibrium is caused by random drift rather than by epistatic interaction in fitness. It has been suggested (Ohta and Kimura, 1971) that linkage disequilibrium should be common among segregating nucleotide sites within a cistron. More recently, Ohta (1982a) showed that marked linkage disequilibrium between tightly linked loci will develop by random drift if a species is subdivided into demes and if migration between them is limited. She has suggested that linkage disequilibrium observed in the major histocompatibility complex (MHC) of man and the mouse should be understood in these terms rather than by epistatic interaction in fitness.

From the standpoint of the neutral theory, intraspecific genetic variability at the molecular level, such as manifested by protein polymorphism, is simply a transient phase of molecular evolution. Then, such polymorphism should be controlled by the same causes which govern molecular evolution. More specifically, the level of polymorphism should be strongly influenced by the functional constraint of the molecule, but almost independent of environmental conditions in which the species is placed. This is in sharp contrast to one of the selectionist hypotheses claiming that polymorphism represents an 'adaptive strategy' for the species to cope with environmental variability. Although different versions exist, they agree in having predicted that a parallelism should exist between environmental variability (or heterogeneity) and genetic variability. This prediction has not been fulfilled. In particular, the prediction that, because of the highly stable environment at the deep-sea bottom, organisms living there would have very low genetic variability, has been falsified, with observations coming in completely the other way round; genetic variability is generally extremely high among organisms living at the bottom of oceans. Then, after this failure, an alternative hypothesis was proposed, saying that 'trophic resource stability' of temporally stable environment induces organisms to adapt to luxury of food types by having high genetic variability. On other occasions, a hypothesis was entertained that an allele having high genetic plasticity (flexibility) is favored under environmental heterogeneity, thus promoting monomorphism. Such ecological hypotheses assuming adaptive strategies are quite elusive and difficult to quantify, so I have not discussed them in detail.

The neutral theory predicts that loci which code for proteins that are subject to stronger functional constraint tend to show lower genetic variability. It has been found that, both in man and *Drosophila*, multimeric enzymes which form interlocus hybrid molecules have a markedly reduced level of polymorphism as compared with monomeric enzymes. This is

consistent with the neutralist prediction, because subunit interactions in such multimeric enzymes increase the functional constraint, and therefore reduce the possibility that a mutation becomes harmless, that is, selectively neutral. On the other hand this finding rejects the hypothesis that hybrid molecules can be the basis of overdominance, adding still one more piece to the already accumulated evidence against the claim that overdominance is the universal cause of protein polymorphism. In *Drosophila* and other organisms, substrate-specific enzymes are on the average less polymorphic than substrate-nonspecific ones. This too can be understood by noting that functional constraint is stronger for the substrate-specific enzymes. Recently, it has been reported (based on two-dimensional electrophoresis) that the abundant proteins in tissues show much lower heterozygosity than soluble enzymes. This can likewise be explained by noting that proteins which exist in major quantities in the cell will have a much larger effect on fitness when their structure is altered by mutation than those existing in minor quantities. In fact, these abundant proteins are structural proteins such as actin, tubulin, etc., some of which are known to be highly conservative in evolution.

Eventually, it will be found, if the neutral theory is valid, that molecules or parts of one molecule which are more important in function, and which therefore evolve more slowly, will show a lower level of heterozygosity. This means that a great deal of polymorphism will be found in noncoding regions (particularly in spacer regions), and also in introns even for genes coding for highly conserved proteins.

The neutral theory enables us to predict the amounts of heterozygosity and polymorphism by using simple formulae. Let us consider an ideal diploid population which is mating at random, has an equal number of males and females, and which keeps the same population size from generation to generation. In such a population, the effective population size (denoted by N_e) is equal to the number of breeding individuals in one generation. Consider a particular locus, and assume that whenever mutation occurs at this locus, it leads to a new, not pre-existing allele. This assumption is known as the infinite allele model. Under these simplifying assumptions, it can be shown that, at statistical equilibrium in which mutational production of new alleles is balanced by random extinction of pre-existing alleles, the expected heterozygosity is

$$\bar{H}_e = \frac{4N_e v_0}{4N_e v_0 + 1}, \tag{10.2}$$

where v_0 is the mutation rate for neutral alleles per locus per generation.

Under the same set of assumptions, the probability that this species is polymorphic with respect to this locus is

$$P_{\text{poly}} = 1 - q^V \tag{10.3}$$

where $V = 4N_e v_0$. In this formula, q stands for the level of gene frequency that is used to define polymorphism; i.e. we call a population polymorphic with respect to a given locus if the frequency of the most common allele is $1 - q$ or less. It is customary to choose $q = 0.01$. In other words, a population is polymorphic if the total frequency of less common alleles is higher than 1%.

To see the use of these formulae, let us assume a hypothetical mammal whose generation length is exactly one year and which comprises, as a species, 100 000 breeding individuals each generation ($N_e = 10^5$). Consider a gene locus that codes for an imaginary protein consisting of 333 amino acids (similar to *Adh*) which correspond to 999 nucleotide sites. Suppose that the total mutation rate per nucleotide site per generation (or year) is 5×10^{-9} which is the value we estimated for the evolutionary rate of pseudo globin genes in mammals. It is known that, roughly speaking, on the average, 1/3 of amino acid replacements in a protein can be detected under standard conditions of electrophoresis. Then, the rate for electrophoretically detectable mutations is 1.3×10^{-6} for this imaginary protein. A large fraction of such mutational changes are likely to be deleterious, so let us assume that on the average only 1/10 of amino acid altering mutations are essentially neutral, while the remaining 9/10 are definitely deleterious neither contributing to molecular evolution nor to polymorphism. This assumption is consistent with the observation that the rate of evolutionary amino acid substitutions in normal globins is about 1/10 of the corresponding rate of pseudo globins. (Note that the globins show the standard evolutionary rate among proteins.) This is also consistent with the corresponding value estimated by comparing the frequencies of rare electrophoretic variants with those of polymorphic variants in the populations of a marine flatfish and the Japanese macaque. Then the neutral mutation rate per locus per generation is $v_0 = 1.3 \times 10^{-7}$ if we restrict our consideration to electrophoretically detectable mutations. With $N_e = 10^5$ as we have assumed for our fictitious mammal, this gives $4N_e v_0 = 5.2 \times 10^{-2}$. Using equations (10.2) and (10.3) above, we obtain $\bar{H}_e \approx 0.049$ or 4.9% for the average heterozygosity, and $P = 0.213$ or 21.3% for the fraction of polymorphic loci. These agree fairly well with the corresponding observed values from a large number of species of mammals ($\bar{H}_e \approx 0.036 \pm 0.025$ and $P \approx 0.15 \pm 0.1$ from Nevo's estimates). We can also derive

theoretical distributions of allelic frequencies, and these too agree well with the observed pattern of allelic distributions.

Because of its biological interest the average heterozygosity has been studied extensively, and a large body of data are now available for animals (including vertebrates and invertebrates) and also for plants: the observed values of $\bar{H}_e$ in natural populations are restricted to the range $0.0 \sim 0.3$ and they seldom exceed 30%. At one time, when the observed heterozygosities were restricted to a very small number of species (consisting of only human, mouse, peromyscus, horseshoe crab and several *Drosophila* species), there was a popular criticism against the neutral theory pointing out that $\bar{H}_e$ is confined to the range between 0.056 and 0.184, and that this casts doubt on the plausibility of the theory for the following reason: this range of $\bar{H}_e$ corresponds to the range of values of $N_e v_0$ between 0.015 and 0.057 if we use the neutralist model, and this means that $N_e v_0$ values are within a factor of 4 of each other. Considering the possibility that the mutation rate is comparable among diverse organisms, this implies an extraordinary invariance of the population size among these organisms, which is a highly absurd proposition.

With much more data now available, it is clear that this claim has lost its main force. For one thing, the average heterozygosity can be extremely low or even zero in some organisms. Particularly, among mammals with a large body size, low heterozygosity appears to be quite common. One of the most remarkable examples is the polar bear from Greenland, for which all the 13 enzyme loci investigated were monomorphic, i.e. $\bar{H}_e = 0$. This is easy to understand from our standpoint, because the effective population size of such mammals must be quite small. In fact, the expected average heterozygosity becomes less than 0.004 if N_e is less than 1000.

The remaining problem to be explained in relation to the above criticism is why the average heterozygosity seldom exceeds 0.3 even for species with immense population sizes. This can be explained by noting that the relevant effective size (N_e) for discussing genetic variation at the molecular level must usually be much smaller, sometimes by several orders of magnitude, than the apparent population size. This is because of a number of factors which are responsible for reducing the long term effective size, such as inequality of the number of breeding males and females, fluctuation of population size from generation to generation, inheritance of fertility, local extinction of colonies followed by recolonization, etc. Particularly, if the population size goes through a sequence of bottlenecks in the course of evolution and if recovery from reduced population size after each bottleneck is slow, the average heterozygosity will be much reduced. In

nature, even if some species are distributed widely covering an enormous area and comprise an immense number of individuals, it is unlikely that they have been so for millions of years and will remain so in the coming millions of years. Sooner or later, such a state will be disrupted by the process of speciation. Thus, in general, the pertinent value of N_e to be used in equation (10.2) for computing the average heterozygosity must be much smaller than N_e estimated from the number of breeding individuals at a given moment. In fact, extensive studies of this problem by Nei and his associates show that whenever a rough estimate of effective population size is obtainable, the observed heterozygosity is almost always lower than expected from this estimate. In other words, there is no difficulty for the neutral theory to explain the 'enormous amount' of genetic variability often referred to.

The last mentioned observation poses a difficulty for selectionist theories that assume various causes (including frequency-dependent selection favoring less frequent alleles) which positively help to maintain the allelic variants in the population; these will certainly inflate the amount of heterozygosity still more unless mutational input is neglected. On the other hand, from the standpoint of the neutral theory, we should expect a significant correlation between $\bar{H}_e$ and N_e, if this is investigated broadly over different type of animals.

At the moment, molecular biology is undergoing a second round of revolutionary development with new knowledge being added at remarkable pace. This will certainly deepen as well as broaden our understanding of the mechanism of evolution. Particularly noteworthy is the discovery of the intronic/exonic structure of the eukaryotic genes.

Referring to this structure, Gilbert (1978) suggested that it might facilitate rapid evolutionary changes not otherwise possible. For example, a single mutational change in the splicing signal region could lead to quite novel proteins. Furthermore, he suggested that if exons correspond to polypeptides with different functions, recombination within introns will assort these functions independently. Consistent with this suggestion is the finding by Tonegawa and his group (Sakano *et al.*, 1979) that in the constant region of immunoglobulin heavy chain there are four separate exons which encode four different units, three of which have a clear domain structure.

In hemoglobins, the situation is not as clearcut as in immunoglobins, but a recent prediction by Gō (1981) and its later confirmation support this idea. Gō investigated the distances between the α-carbon atoms in the three-dimensional structure of mouse β-globin, and identified four segments, designated F1, F2, F3 and F4. It is well known that in the mouse

gene coding for β-globin, there are three exons separated by two introns, with the central segment (F2 + F3) enclosing the haem group. Based on the conformational characteristics Gō considered that it is more natural for the F2 and F3 regions to be coded by separate exons, and suggested that there was an intron in the ancestor of the present α- and β-globins. Presumably it was removed because of being unnecessary. A remarkable confirmation is seen in the work of Jensen *et al.* (1981), who found three introns in the leghaemoglobin gene from the soybean.

It is also interesting to note that the phenomenon known as 'exon shuffling' or 'domain transfer' may be occurring regularly, albeit at a low frequency, between members of a multigene family, as suggested by comparative studies of the constant region of immunoglobulin heavy chain (Miyata *et al.*, 1980a). This phenomenon may occur either through double unequal crossing-over or gene conversion during pre-meiotic divisions of germ cells. Recently, Ohta (1982b) proposed a model incorporating this phenomenon to explain the unusually high polymorphism at the major histocompatibility complex of man and the mouse. These represent just a few examples to show how new molecular findings lead us to previously unsuspected vistas of evolution. They will help reveal the history of life in its early stage of evolution which is unattainable by conventional methods of comparative morphology and paleontology.

The origin of life on the earth has been, and still is, one of the greatest enigmas in science. So far, no definite clue has been obtained as to how the first self-replicating molecule came to be (but see van Roode and Orgel, 1980, who recently have done important experimental work on enzyme-free replication of a template by metal ions). With the development of molecular genetics, there is now a better hope that light will be thrown on this problem by pursuing the evolutionary implications of newer molecular findings. Already, a good beginning has been made by Eigen (1971) by the proposal and mathematical investigations of the 'hypercycle' model. (For a lucid account of his theory on the origin of genetic information readers are invited to read Eigen *et al.*, 1981.) Although the concept of Darwinian natural selection is central in his treatment, it is possible that stochastic elements also play important roles in this problem.

Returning to the main topic of this book, the mathematical methods that I have used to establish the neutral theory contains various parameters such as mutation rates, selection coefficients, population size and migration rates. Although they often appear as products such as $N_e v$, the individual parameters must be estimated in order to check theoretical predictions with observations and experiments. Among students of evolutionary biology,

there has been a strong tendency to claim that these population genetical parameters will never be known accurately and therefore theories which contain such parameters are of little use. I take the opposite view; these parameters have to be investigated and measured if we really want to understand the mechanism of evolution at the molecular level. Can astronomers and cosmologists claim that theories which contain various astronomical parameters should be avoided because such parameters are difficult to estimate accurately? This reminds me of an aphorism, which I understand is due to Galileo, and which in effect says: what we can measure we should measure; what we cannot measure at present, we should endeavor to make measurable.

Many people, directly and indirectly, have told me that the neutral theory is not biologically important, since neutral genes are by definition not concerned with adaptation. The term 'evolutionary noise' has often been used to describe the role of neutral mutations in evolution, with such a contention in mind. I believe that this is a too narrow view. First, what is important in science is to find out the truth, so the neutral theory should be of value if it is valid as a scientific hypothesis. Second, even if the 'neutral alleles' are functionally equivalent or nearly so under a prevailing set of environmental conditions of a species, it is possible that some of them, when a new environmental condition is imposed, will become selected. Experiments strongly suggesting this possibility have been reported, and I have used the designation 'Dykhuizen-Hartl effect' to refer to this, namely, otherwise neutral or nearly neutral alleles may have a latent potential for selection which can be realized under the appropriate conditions. Thus, neutral mutants can be the raw material for adaptive evolution. In his popular book on evolution *The Life Game*, Calder (1973) wrote 'A gene that is neutral today may, in new circumstances a thousand generations hence, come to assume great significance.' He also remarked that 'if trivial changes can accumulate in a molecule over many millions of years, without interference from natural selection, the day may come when a peculiar molecule thus produced goes through one more mutation that suddenly gives it new importance–just as changing one card can alter a worthless poker hand into a royal straight flush'. In general, material for adaptive evolution comes from slightly deleterious mutant genes which have been floating in the species and which have turned out to be advantageous because of a change of environment. As the limit of very slightly deleterious mutants, neutral mutants may also be considered as such. In this connection, Ohta's thesis is instructive. It says that a mutant can sometimes be advantageous under a restricted condition; it is almost impossible for a

mutant to find itself advantageous under all conditions. Thus, a mutant has a better chance of being advantageous when the environment is uniform.

Although this book is concerned with molecular evolution, I need to make some comments on evolution at the phenotypic level. I believe that Darwinian (positive) selection mainly acts through phenotypes which are the end results of the action of many genes. Here, ecological conditions must play a decisive role in determining what phenotypes will be selected for; positive selection cares little how a given phenotype is determined by genotypes. Certainly, consideration of 'ecologic opportunity' is essential for interpreting fossil records; it is highly unlikely that adaptive, phenotypic evolution has occurred at a uniform rate. I emphasize also that laws governing molecular evolution are clearly different from those governing phenotypic evolution.

It may be asked, then, why natural selection is so prevalent at the phenotypic level and yet random fixation of selectively neutral or nearly neutral alleles prevails at the molecular level. The answer to this question, I think, comes from the fact that the most common type of natural selection at the phenotypic level is 'stabilizing' selection. Unlike the type of natural selection that Darwin had in mind when he tried to explain evolutionary change, stabilizing selection eliminates phenotypically extreme individuals and preserves those that are near the population mean. Since the early work of Bumpus on the house sparrow and Weldon on the land snail, many examples of stabilizing selection have been reported. Probably the best example in man is the relationship between the birth-weights of babies and their neo-natal mortality, as studied by Karn and Penrose. As I have mentioned already, if a large number of segregating loci or sites are involved in a quantitative character, the average selection coefficient per mutant under stabilizing selection must be exceedingly small. It has also been shown that such mutants are very slightly deleterious or almost neutral, so that mutant substitutions are mainly controlled by random drift, although the rate of 'neutral evolution' is lower than when all the mutations are strictly neutral. When applied to the problem of 'non-random' or unequal usage of synonymous codons, this model gives a satisfactory solution of the problem within the framework of the neutral theory.

The picture of evolution that emerges from this analysis is as follows. From time to time, the position of the optimum of a phenotypic character shifts due to change of environment, and the species tracks such a change rapidly by altering its mean. During this short period of change, an extensive shift of gene frequencies is expected to occur at many loci, but *this process itself will seldom cause gene substitutions.* But, most of the time,

stabilizing selection predominates, under which neutral evolution or *random fixation of mutant alleles occurs extensively*, transforming all genes including those of living fossils. Gould and Eldredge (1977) have argued that the evolutionary record is one of long periods of essentially no change interspersed with rare periods of sudden change ('punctuated equilibrium'). This is entirely consistent with the picture I have just given of long periods of stabilizing selection, interspersed with periods of rapid directional selection to track environmental changes. However, it does not seem necessary to invoke a different mechanism for the periods of very rapid change, as Gould and Eldredge appear to do. Thus, we can bridge the gap between the mode of evolution at the phenotypic level and that at the molecular level by assuming a preponderance of stabilizing selection at the phenotypic level and a very large number of nucleotide sites involved at the molecular level.

Toward the end of his book, *The Origin of Species*, Darwin contemplates an entangled bank with complex ecological interactions among various creatures, and he notes that they have all been produced by 'laws acting around us', including inheritance, variability, struggle for life and natural selection. Production of wonderful higher forms from 'the war of nature, from famine and death' seemed to be most impressive to Darwin. Then, he closes the book with a well-known passage:

> There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved.

We now know that underneath this remarkable procession of life and indeed, deep down at the level of the genetic material, an enormous amount of evolutionary change has occurred, and is still occurring. What is remarkable, I think, is that the overwhelming majority of such changes are not caused by natural selection but by random fixation of selectively neutral or nearly neutral mutants. Although such random processes are slow and insignificant for our ephemeral existence, in the span of geological times, they become colossal. In this way, the footprints of time are evident in all the genomes on the earth. This adds still more to the grandeur of our view of biological evolution.

References

Abramowitz, M. & Stegun, I.A. (eds.) (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Washington, DC: US Department of Commerce.

Allendorf, F.W. (1978). Protein polymorphism and the rate of loss of duplicate gene expression. *Nature, Lond.* **272**, 76–8.

Allendorf, F.W., Christiansen, F.B., Dobson, T., Eanes, W.F. & Frydenberg, O. (1979). Electrophoretic variation in large mammals I. The polar bear, *Thalarctos maritimus*. *Hereditas*, **91**, 19–22.

Allison, A.C. (1955). Aspects of polymorphism in man. *Cold Spring Harbor Symp. Quant. Biol.* **20**, 239–55.

Alvarez, L.W., Alvarez, W., Asaro, F. & Michel, H.V. (1980). Extraterrestrial cause for the Cretaceous-Tertiary extinction. *Science*, **208**, 1095–108.

Ayala, F.J. (1972). Darwinian versus non-Darwinian evolution in natural populations of *Drosophila*. In *Proceedings of the Sixth Berkeley Symp. Math. Stat. Prob.*, vol. 5, ed. L.M. Le Cam, J. Neyman & E.L. Scott, pp. 211–36. University of California Press.

Ayala, F.J. (1974). Biological evolution: natural selection or random walk? *American Scientist*, **62**, 692–701.

Ayala, F.J. & Anderson, W.W. (1973). Evidence of natural selection in molecular evolution. *Nature New Biol.* **241**, 274–6.

Ayala, F.J. & Gilpin, M.E. (1973). Lack of evidence for the neutral hypothesis of protein polymorphism. *Jour. Heredity*, **64**, 297–8.

Ayala, F.J. & Gilpin, M.E. (1974). Gene frequency comparisons between taxa: support for the natural selection of protein polymorphisms. *Proc. Natl. Acad. Sci. USA*, **71**, 4847–9.

Ayala, F.J., Powell, J.R., Tracey, M.L., Mourão, C.A. & Pérez-Salas, S. (1972). Enzyme variability in the *Drosophila willistoni* group. IV. Genetic variation in natural populations of *Drosophila willistoni*. *Genetics*, **70**, 113–39.

Ayala, F.J. & Tracey, M.L. (1974). Genetic differentiation within and between species of the *Drosophila willistoni* group. *Proc. Natl. Acad. Sci. USA*, **71**, 999–1003.

Ayala, F.J., Tracey, M.L., Barr, L.G., McDonald, J.F. & Pérez-Sala, S. (1974). Genetic variation in natural populations of five Drosophila species and the hypothesis of the selective neutrality of protein polymorphisms. *Genetics*, **77**, 343–84.

Ayala, F.J. & Valentine, J.W. (1979). Genetic variability in the pelagic environment: a paradox? *Ecology*, **60**, 24–9.

Ayala, F.J., Valentine, J.W., Hedgecock, D. & Barr, L.G. (1975). Deep-sea asteroids: high genetic variability in a stable environment. *Evolution*, **29**, 203–12.

Baker, M.C. (1981). Effective population size in a songbird: some possible implications. *Heredity*, **46**, 209–18.

Baltimore, D. (1981). Gene conversion: some implications for immunoglobulin genes. *Cell*, **24**, 592–4.

Barnard, E.A., Cohen, M.S., Gold, M.H. & Kim, J-K. (1972). Evolution of ribonuclease in relation to polypeptide folding mechanism. *Nature, Lond.* **240**, 395–8.

Bell, G.I., Pictet, R.L., Rutter, W.J., Cordell, B., Tischer, E. & Goodman, H.M. (1980). Sequence of the human insulin gene. *Nature, Lond.* **284**, 26–32.

Bird, A.P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research*, **8**, 1499–504.

Birky, C.J. & Skavaril, R.V. (1976). Maintenance of genetic homogeneity in systems with multiple genomes. *Genet. Res., Camb.* **27**, 249–65.

Black, J.A. & Gibson, D. (1974). Neutral evolution and immunoglobulin diversity. *Nature, Lond.* **250**, 327–8.

Blundell, T.L., Cutfield, J.F., Cutfield, S.M., Dodson, E.J., Dodson, G.G., Hodgkin, D.C., Mercola, D.A. & Vijayan, M. (1971). Atomic positions in rhombohedral 2-zinc insulin crystals. *Nature, Lond.* **231**, 506–11.

Blundell, T.L. & Wood, S.P. (1975). Is the evolution of insulin Darwinian or due to selectively neutral mutation? *Nature, Lond.* **257**, 197–203.

Bogardt, Jr., R.A., Jones, B.N., Dwulet, F.E., Garner, W.H., Lehman, L.D., & Gurd, F.R.N. (1980). Evolution of the amino acid substitution in the mammalian myoglobin gene. *J. Mol. Evol.* **15**, 197–218.

Bonnell, M.L. & Selander, R.K. (1974). Elephant seals: genetic variation and near extinction. *Science*, **184**, 908–9.

Boyer, S.H., Crosby, E.F., Thrmon, T.F., Noyes, A.N., Fuller, G.F., Leslie, S.E., Shepard, M.K. & Herndon, C.N. (1969). Hemoglobins A and A_2 in new world primates: comparative variation and its evolutionary implications. *Science*, **166**, 1428–1431.

Boyer, S.H., Scott, A.F., Kunkel, L.M. & Smith, K.D. (1978). The proportion of all point mutations which are unacceptable: an estimate based on hemoglobin amino acid and nucleotide sequences. *Can. J. Genet. Cytol.* **20**, 111–37.

Briscoe, D.A., Robertson, A. & Malpica, J.-M. (1975). Dominance at *Adh* locus in response of adult *Drosophila melanogaster* to environmental alcohol. *Nature, Lond.* **255**, 148–9.

Brown, A.J.L. & Langley, C.H. (1979). Reevaluation of level of genic heterozygosity in natural population of *Drosophila melanogaster* by two-dimensional electrophoresis. *Proc. Natl. Acad. Sci. USA*, **76**, 2381–4.

Brown, A.H.D., Marshall, D.R. & Weir, B.S. (1981). Current status of the charge state model for protein polymorphism. In *Genetic Studies of Drosophila Populations*, ed. J.B. Gibson & J.G. Oakeshott, pp. 15–43. Australian National University Press.

Bulmer, M.G. (1971a). The stability of equilibria under selection. *Heredity*, **27**, 157–62.

Bulmer, M.G. (1971b). Protein polymorphism. *Nature, Lond.* **234**, 410–11.

Bulmer, M.G. (1972). The genetic variability of polygenic characters under optimizing selection, mutation and drift. *Genet. Res., Camb.* **19**, 17–25.

Bulmer, M.G. (1973). Geographical uniformity of protein polymorphisms. *Nature, Lond.* **241**, 199–200.

Calder, N. (1973). *The Life Game*. London: British Broadcasting Corporation.

Calder, N. (1974). *The Weather Machine*. London: British Broadcasting Corporation.

Cameron, D.G. & Vyse, E.R. (1978). Heterozygosity in *Yellowstone Park Elk, Cervus canadensis*. *Biochemical Genetics*, **16**, 651–7.

Carlson, E.A. (1981). *Genes, Radiation, and Society. The Life and Work of H.J. Muller*. Cornell University Press.

Cavalli-Sforza, L.L. (1966). Population structure and human evolution. *Proc. Roy. Soc. London Ser. B*, **164**, 362–79.
Cavalli-Sforza, L.L. & Bodmer, W.F. (1971). *The Genetics of Human Populations*. San Francisco: W.H. Freeman & Co.
Cavalli-Sforza, L.L. & Conterio, F. (1960). Analisi della fluttuazione di frequenze geniche nella populazione della Val Parma. *Atti Assoc. Genet. Ital.* **5**, 333–44.
Cavalli-Sforza, L.L. & Edwards, A.W.F. (1967). Phylogenetic analysis: models and estimation procedures. *Amer. J. Hum. Genet.* **19**, 233–57.
Chakraborty, R., Fuerst, P.A. & Nei, M. (1978). Statistical studies on protein polymorphism in natural populations II. Gene differentiation between populations. *Genetics*, **88**, 367–90.
Charlesworth, B. (1979). Evidence against Fisher's theory of dominance. *Nature, Lond.* **278**, 848–9.
Charlesworth, B. & Charlesworth, D. (1979). Dominance modification in a fluctuating environment. *Genet. Res., Camb.* **34**, 281–5.
Childs, G., Maxson, R., Cohn, R.H. & Kedes, L. (1981). Orphons: dispersed genetic elements derived from tandem repetitive genes of eucaryotes. *Cell*, **23**, 651–63.
Christiansen, F.B., Bundgaard, J. & Barker, J.S.F. (1977). On the structure of fitness estimates under post-observational selection. *Evolution*, **31**, 843–53.
Clarke, B. (1970). Selective constraints on amino-acid substitutions during the evolution of proteins. *Nature, Lond.* **228**, 159–60.
Clarke, B. (1971). Natural selection and the evolution of proteins. *Nature, Lond.* **232**, 487.
Clarke, B. (1975). The contribution of ecological genetics to evolutionary theory: detecting the direct effects of natural selection on particular polymorphic loci. *Genetics*, **79**, 101–13.
Clarke, B. (1979). The evolution of genetic driversity. *Proc. Roy. Soc. Lond. B.* **205**, 453–74.
Clarke, B. & Allendorf, F.W. (1979). Frequency-dependent selection due to kinetic differences between allozymes. *Nature, Lond.* **279**, 732–4.
Clarke, B. & O'Donald, P. (1964). Frequency-dependent selection. *Heredity*, **19**, 201–6.
Cordell, B., Bell, G., Tischer, E., DeNoto, F.M., Ullrich, A., Pictet, R., Rutter, W.J. & Goodman, H.M. (1979). Isolation and characterization of a cloned rat insulin gene. *Cell*, **18**, 533–43.
Coyne, J. (1976). Lack of genic similarity between two sibling species of *Drosophila* as revealed by varied techniques. *Genetics*, **84**, 593–607.
Crick, F. (1979). Split genes and RNA slicing. *Science* **204**, 264–71.
Crow, J.F. (1957). Possible consequences of an increased mutation rate. *Eugenics Quarterly*, **4**, 67–80.
Crow, J.F. (1958). Some possibilities for measuring selection intensities in man. *Human Biology*, **30**, 1–13.
Crow, J.F. (1964). More on the heterozygous effects of lethals in *Drosophila* populations. *Amer. Natur.* **98**, 447–9.
Crow, J.F. (1968). The cost of evolution and genetic loads. In *Haldane and Modern Biology*, ed. K.R. Dronamraju, pp. 165–78. Baltimore: Johns Hopkins Press.
Crow, J.F. (1969). Molecular genetics and population genetics. *Proc. XII International Congress of Genetics* (Tokyo), vol. 3, pp. 105–13.
Crow, J.F. (1970). Genetic loads and the cost of natural selection. In *Mathematical Topics in Population Genetics*, ed. K. Kojima, pp. 128–77. Berlin: Springer-Verlag.
Crow, J.F. (1972). The dilemma of nearly neutral mutations: How important are they for evolution and human welfare? *J. Hered.* **63**, 306–16.
Crow, J.F. (1979). Minor viability mutants in Drosophila. *Genetics*, **92**, (supplement), s165–s172.

Crow, J.F. (1981). The neutralist-selectionist controversy: an overview. In *Population and Biological Aspects of Human Mutation*, ed. E.B. Hook, pp. 3–14. New York: Academic Press.
Crow, J.F. & Kimura, M. (1970). *An Introduction to Population Genetics Theory*. New York: Harper & Row.
Crow, J.F. & Kimura, M. (1972). The effective number of a population with overlapping generations: A correction and further discussion. *Amer. J. Hum. Genet.* **24**, 1–10.
Crow, J.F. & Kimura, M. (1979). Efficiency of truncation selection. *Proc. Natl. Acad. Sci. USA*, **76**, 396–9.
Crow, J.F. & Temin, R.G. (1964). Evidence for the partial dominance of recessive lethal genes in natural populations of Drosophila. *Amer. Natur.* **98**, 21–33.
Daly, K. & Clarke, B. (1981). Selection associated with the alcohol dehydrogenase locus in *Drosophila melanogaster*: differential survival of adults maintained on low concentrations of ethanol. *Heredity*, **46**, 219–26.
Darwin, C. (1859). *The Origin of Species by Means of Natural Selection*. London: John Murray.
Dawkins, R. (1976). *The Selfish Gene*. Oxford University Press.
Dayhoff, M.O. (ed.) (1969). *Atlas of Protein Sequence and Structure*, vol. 4. Silver Spring, Maryland: National Biomedical Research Foundation.
Dayhoff, M.O. (ed.) (1972). *Atlas of Protein Sequence and Structure*, vol. 5. Washington, DC: National Biomedical Research Foundation.
Dayhoff, M.O. (ed.) (1976). *Atlas of Protein Sequence and Structure*, vol. 5, supplement 2. Washington, DC: National Biomedical Research Foundation.
Dayhoff, M.O. (ed.) (1978). *Atlas of Protein Sequence and Structure*, vol. 5, supplement 3. Washington, DC: National Biomedical Research Foundation.
Dayhoff, M.O. & Eck, R.V. (1968). *Atlas of Protein Sequence and Structure*, 1967–68. Silver Spring: National Biomedical Research Foundation.
Deevey, E.S., Jr. (1960). The human population. *Scientific American*, **203**(3), 195–204.
Dickerson, R.E. (1971). The structure of cytochrome c and the rates of molecular evolution. *J. Mol. Evol.* **1**, 26–45.
Dickerson, R.E. & Geis, I. (1969). *The Structure and Action of Proteins*. Menlo Park: W.A. Benjamin.
Dobzhansky, Th. (1937). *Genetics and the Origin of Species*, 1st edn. Columbia University Press.
Dobzhansky, Th. (1951). *Genetics and the Origin of Species*, 3rd edn. Columbia University Press.
Dobzhansky, Th. (1955). A review of some fundamental concepts and problems of population genetics. *Cold Spring Harbor Symp. Quant. Biol.* **20**, 1–15.
Dobzhansky, Th. (1965). Genetic diversity and fitness. In *Genetics Today* (*Proc. XI International Congr. Genetics*) ed. S.J. Geerts, pp. 541–52. Oxford: Pergamon Press.
Dobzhansky, Th. (1970). *Genetics of the Evolutionary Process*. Columbia University Press.
Dolan, R. & Robertson, A. (1975). The effect of conditioning the medium in *Drosophila*, in relation to frequency-dependent selection. *Heredity*, **35**, 311–16.
Doolittle, W.F. & Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature, Lond.* **284**, 601–3.
Dover, G.A. & Flavell, R.B. (ed.) (1982). *Genome Evolution*. London: Academic Press.
Dover, G. (1982). Molecular drive: a cohesive mode of species evolution. *Nature* **299**, 111–117.
Dykhuizen, D. & Hartl, D.L. (1980). Selective neutrality of 6PGD allozymes in *E. coli* and the effects of genetic background. *Genetics*, **96**, 801–17.
Efstratiadis, A., Kafatos, F.C. & Maniatis, T. (1977). The primary structure of rabbit β-globin mRNA as determined from cloned DNA. *Cell*, **10**, 571–85.

Eigen, M. (1971). Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, **58**, 465–523.
Eigen, M., Gardiner, W., Schuster, P. & Winkler-Oswatitsh, R. (1981). The origin of genetic information. *Scientific American*, **244(2)**, 78–94.
Epstein, C.J. (1967). Non-randomness of amino-acid changes in the evolution of homologous proteins. *Nature, Lond.* **215**, 355–9.
Ewens, W.J. (1969). *Population Genetics*. London: Methuen.
Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, **3**, 87–112.
Ewens, W.J. (1977). Population genetics theory in relation to the neutralist-selectionist controversy. *Adv. Hum. Genet.* **8**, 67–134.
Ewens, W.J. (1979). *Mathematical Population Genetics*. Berlin, Heidelberg: Springer-Verlag.
Ewens, W.J. & Feldman, M.W. (1974). Analysis of neutrality in protein polymorphism. *Science*, **183**, 446–8.
Felsenstein, J. (1971). On the biological significance of the cost of gene substitution. *Amer. Natur.* **105**, 1–11.
Ferris, S.D. & Whitt, G.S. (1977). Loss of duplicate gene expression after polyploidization. *Nature, Lond.* **265**, 258–60.
Fincham, J.R.S. (1966). *Gene Complementation*. Amsterdam: Benjamin.
Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. Roy. Soc. Edinburgh*, **52**, 399–433.
Fisher, R.A. (1922). On the dominance ratio. *Proc. Roy. Soc. Edinburgh*, **42**, 321–41.
Fisher, R.A. (1930a). The distribution of gene ratios for rare mutations. *Proc. Roy. Soc. Edinburgh*, **50**, 205–20.
Fisher, R.A. (1930b). *The Genetical Theory of Natural Selection*. Oxford University Press. 2nd revised edition (1958), Dover Publications.
Fisher, R.A. (1936). The measurement of selective intensity. *Proc. Roy. Soc. London, Ser. B.* **121**, 58–62.
Fisher, R.A. (1953). Population genetics. *Proc. Roy. Soc. London, Ser. B*, **141**, 510–23.
Fisher, R.A. & Ford, E.B. (1950). The "Sewall Wright" effect. *Heredity*, **4**, 117–19.
Fisher, W.K., Nash, A.R. & Thompson, E.O.P. (1977). Haemoglobins of the shark, *Heterodontus portusjacksoni* III. Amino acid sequence of the β-chain. *Aust. J. Biol. Sci.* **30**, 487–506.
Fitch, W.M. (1971a). Evolutionary variability in hemoglobins. In *Synthese, Struktur und Funktion des Hämoglobins*, ed. H. Martin & L. Nowicki, pp. 199–215. München: J.F. Lehmanns Verlag.
Fitch, W.M. (1971b). Rate of change of concomitantly variable codons. *J. Mol. Evol.* **1**, 84–96.
Fitch, W.M. (1971c). Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.* **20**, 406–16.
Fitch, W.M. (1980). Estimating the total number of nucleotide substitutions since the common ancestor of a pair of homologous genes: comparison of several methods and three beta hemoglobin messenger RNA's. *J. Mol. Evol.* **16**, 153–209.
Fitch, W.M. & Langley, C.H. (1976). Protein evolution and the molecular clock. *Fed. Proc.* **35**, 2092–7.
Fitch, W.M. & Markowitz, E. (1970). An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem. Genet.* **4**, 579–93.
Fontdevila, A., Mendez, J., Ayala, F.J. & McDonald, J. (1975). Maintenance of allozyme polymorphisms in experimental populations of *Drosophila*. *Nature, Lond.* **255**, 149–51.

Ford, E.B. (1964). *Ecological Genetics*, 1st edn., London: Methuen. 2nd & 3rd edns. (1965, 1971). London: Chapman & Hall.
Ford, E.B. (1965). *Genetic Polymorphism*. London: Faber & Faber.
Franklin, I. & Lewontin, R.C. (1970). Is the gene the unit of selection? *Genetics*, **65**, 707–34.
Fuerst, P.A., Chakraborty, R. & Nei, M. (1977). Statistical studies on protein polymorphism in natural populations I. Distribution of single locus heterozygosity. *Genetics*, **86**, 455–83.
Fuerst, P.A. & Ferrell, R.E. (1980). The stepwise mutation model: an experimental evaluation utilizing hemoglobin variants. *Genetics*, **94**, 185–201.
Gaines, M.S. & Whittam, T. (1980). Genetic changes in fluctuating vole populations: selective vs. nonselective forces. *Genetics*, **96**, 767–78.
Gatlin, L.L. (1976). Counter-examples to a neutralist hypothesis. *J. Mol. Evol.* **7**, 185–95.
Gauss, D.H. & Sprinzl, M. (1981). Compilation of tRNA sequences. *Nucl. Acids Res.* **9**, rl–r23.
Gibson, J.B., Lewis, N., Adena, M.A. & Wilson, S.R. (1979). Selection for ethanol tolerance in two populations of *Drosophila melanogaster* segregating alcohol dehydrogenase allozymes. *Aust. J. Biol. Sci.* **32**, 387–98.
Gilbert, W. (1978). Why genes in pieces? *Nature, Lond.* **271**, 501.
Gillespie, J.H. (1977). Sampling theory for alleles in a random environment. *Nature, Lond.* **266**, 443–5.
Gillespie, J.H. & Kojima, K. (1968). The degree of polymorphisms in enzymes involved in energy production compared to that in nonspecific enzymes in two *Drosophila ananassae* populations. *Proc. Natl. Acad. Sci. USA*, **61**, 582–5.
Gillespie, J.H. & Langley, C.H. (1974). A general model to account for enzyme variation in natural populations. *Genetics*, **76**, 837–48.
Ginzburg, I., Behar, L., Givol, D. & Littauer, U.Z. (1981). The nucleotide sequence of rat α-tubulin: 3'-end characteristics, and evolutionary conservation. *Nucl. Acids Res.* **9**, 2691–7.
Gō, M. (1981). Correlation of DNA exonic regions with protein structure units in haemoglobin. *Nature, Lond.* **291**, 90–2.
Gojobori, T. (1979). Studies of protein polymorphism by data analyses with respect to loci. Doctorial Dissertation, Faculty of Science, Kyushu University.
Gojobori, T. (1982). Means and variances of heterozygosity and protein function. In *Molecular Evolution, Protein Polymorphism and the Neutral Theory*, ed. M. Kimura, pp. 137–48. Tokyo: Japan Scientific Societies Press Berlin: Springer-Verlag.
Gooch, J.L. & Schopf, T.J.M. (1972). Genetic variability in the deep sea: relation to environmental variability. *Evolution*, **26**, 545–52.
Goodman, M. (1978). Substitutional trends and non-random changes in rates during protein evolution. In *Evolution of Protein Molecules*, ed. H. Matsubara & T. Yamanaka, pp. 17–32. Tokyo: Japan Scientific Societies Press.
Goodman, M., Moore, G.W., Barnabas, J. & Matsuda, G. (1974). The phylogeny of human globin genes investigated by the maximum parsimony method. *J. Mol. Evol.* **3**, 1–48.
Goodman, M., Moore, G.W. & Matsuda, G. (1975). Darwinian evolution in the genealogy of hemoglobin. *Nature, Lond.* **253**, 603–8.
Goossens, M., Dozy, A.M., Embury, S.H., Zachariades, Z., Hadjiminas, M., Stamatoyannopoulos, G. & Kan, Y.W. (1980). Triplicated α-globin loci in humans. *Proc. Natl. Acad. Sci. USA*, **77**, 518–21.
Gould, S.J. & Eldredge, N. (1977). Punctuated equilibria: the tempo and mode of evolution reconsidered. *Paleobiology*, **3**, 115–51.

Grant, V. (1977). *Organismic Evolution*. San Francisco: W.H. Freeman.
Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science*, **185**, 862–4.
Grantham, R. (1980). Workings of the genetic code. *Trends in Biochemical Sciences*, **5**, 327–31.
Grantham, R., Gautier, C. & Gouy, M. (1980a). Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucl. Acids Res.* **8**, 1893–912.
Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. & Mercier, R. (1981). Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucl. Acids Res.* **9**, r43–r74.
Grantham, R., Gautier, C., Gouy, M., Mercier, R. & Pavé, A. (1980b). Codon catalog usage and the genome hypothesis. *Nucl. Acids Res.* **8**, r49–r62.
Grunstein, M. & Grunstein, J.E. (1978). The histone H4 gene of *Strongylocentrotus purpuratus*: DNA and mRNA sequences at the 5′ end. *Cold Spring Harbor Symp. Quant. Biol.* **42** (part 2), 1083–92.
Grunstein, M., Schedl, P. & Kedes, L. (1976). Sequence analysis and evolution of sea urchin (*Lytechinus pictus* and *Strongylocentrotus purpuratus*) histone H4 messenger RNAs. *J. Mol. Biol.* **104**, 351–69.
Guess, H.A. & Ewens, W.J. (1972). Theoretical and simulation results relating to the neutral allele theory. *Theor. Popul. Biol.* **3**, 434–47.
Haga, T. (1969). Structure and dynamics of natural populations of a diploid *Trillium*. In *Chromosomes Today*, ed. C.D. Darlington & K.R. Lewis, pp. 207–17. Edinburgh: Oliver & Boyd.
Haga, T. & Kurabayashi, M. (1954). Genome and polyploidy in the genus *Trillium*. V. Chromosomal variation in natural populations of *Trillium kamtschaticum Pall.* *Mem. Fac. Sci. Kyushu Univ.*, Ser. E, Biol. **1**, 159–85.
Haldane, J.B.S. (1924). A mathematical theory of natural and artificial selection. Part I. *Trans. Camb. Phil. Soc.* **23**, 19–41.
Haldane, J.B.S. (1927). A mathematical theory of natural and artificial selection. Part V. Selection and mutation. *Proc. Camb. Phil. Soc.* **23**, 838–44.
Haldane, J.B.S. (1930). A note on Fisher's theory of the origin of dominance. *Amer. Natur.* **64**, 87–90.
Haldane, J.B.S. (1931). A mathematical theory of natural and artificial selection. Part VIII. Metastable populations. *Proc. Camb. Phil. Soc.* **27**, 137–42.
Haldane, J.B.S. (1932). *The Causes of Evolution*. New York: Harper and Row.
Haldane, J.B.S. (1937). The effect of variation on fitness. *Amer. Natur.* **71**, 337–49.
Haldane, J.B.S. (1938). Forty years of genetics. In *Background to Modern Science*, ed. J. Needham & W. Pagel, pp. 225–43. Cambridge University Press.
Haldane, J.B.S. (1949). Suggestions as to quantitative measurement of rates of evolution. *Evolution*, **3**, 51–6.
Haldane, J.B.S. (1954). The statics of evolution. In *Evolution as a Process*, ed. J. Huxley *et al.*, pp. 109–21. London: George Allen & Unwin.
Haldane, J.B.S. (1957a). Karl Pearson, 1857–1957. *Biometrika*, **44**, 303–13.
Haldane, J.B.S. (1957b). The cost of natural selection. *J. Genet.* **55**, 511–24.
Haldane, J.B.S. (1958). *The Unity and Diversity of Life*. (Sardar Vallabhbhai Patel Lectures). Delhi: Ministry of Information and Broadcasting, Government of India.
Haldane, J.B.S. (1959). Natural selection. In *Darwin's Biological Work*, ed. P.R. Bell, pp. 101–49. Cambridge University Press.
Haldane, J.B.S. (1964). A defense of beanbag genetics. *Perspect. Biol. Med.* **7**, 343–59.
Haldane, J.B.S. & Jayakar, S.D. (1963). Polymorphism due to selection of varying direction. *J. Genet.* **58**, 237–42.

Hardy, G.H. (1908). Mendelian proportions in a mixed population. *Science*, **28**, 49–50.
Harris, H. (1966). Enzyme polymorphism in man. *Proc. Roy. Soc. London, Ser. B* **164**, 298–310.
Harris, H. (1976). Molecular evolution: the neutralist-selectionist controversy. *Fed. Proc.* **35**, 2079–82.
Harris, H. & Hopkinson, D.A. (1972). Average heterozygosity per locus in man: an estimate based on the incidence of enzyme polymorphisms. *Ann. Hum. Genet.* **36**, 9–20.
Harris, H., Hopkinson, D.A. & Edwards, Y.H. (1977). Polymorphism and subunit structure of enzymes: a contribution to the neutralist-selectionist controversy. *Proc. Natl. Acad. Sci. USA*, **74**, 698–701.
Harris, H., Hopkinson, D.A. & Robson, E.B. (1974). The incidence of rare alleles determining electrophoretic variants: data on 43 enzyme loci in man. *Ann. Hum. Genet.* **37**, 237–53.
Hartl, D.L. & Dykhuizen, D.E. (1981). Potential for selection among nearly neutral allozymes of 6-phosphogluconate dehydrogenase in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA*, **78**, 6344–8.
Hayashi, A. & Stamatoyannopoulos, G. (1972). Role of penultimate tyrosine in haemoglobin β subunit. *Nature, New Biol.* **235**, 70–2.
Hedrick, P.W., Ginevan, M.E. & Ewing, E.P. (1976). Genetic polymorphism in heterogeneous environments. *Ann. Rev. Ecol. Syst.* **7**, 1–32.
Heindell, H.C., Liu, A., Paddock, G.V., Studnicka, G.M. & Salser, W.A. (1978). The primary sequence of rabbit α-globin mRNA. *Cell*, **15**, 43–54.
Hendrickson, W.A. & Love, W.E. (1971). Structure of lamprey haemoglobin. *Nature, New Biol.* **232**, 197–203.
Hildreth. P.E. (1956). The problem of synthetic lethals in *Drosophila melanogaster*. *Genetics*, **41**, 729–42.
Hill, W.G. (1972). Effective size of populations with overlapping generations. *Theor. Prop. Biol.* **3**, 278–89.
Hill, W.G. (1979). A note on effective population size with overlapping generations, *Genetics*, **92**, 317–22.
Ho, M.W. & Saunders, P.T. (1979). Beyond neo-Darwinism – an epigenetic approach to evolution. *J. Theor. Biol.* **78**, 573–91.
Holmquist, R. (1979). The method of parsimony: an experimental test and theoretical analysis of the adequacy of molecular restoration studies. *J. Mol. Biol.* **135**, 939–58.
Holmquist, R., Jukes, T.H. & Pangburns, S. (1973). Evolution of transfer RNA. *J. Mol. Biol.* **78**, 91–116.
Hood, L., Campbell, J.H. & Elgin, S.C.R. (1975). The organization, expression, and evolution of antibody genes and other multigene families. *Annu. Rev. Genet.* **9**, 305–53.
Hori, H. & Osawa, S. (1979). Evolutionary change in 5S RNA secondary structure and a phylogenetic tree of 54 5S RNA species. *Proc. Natl. Acad. Sci. USA*, **76**, 381–5.
Horuk, R., Goodwin, P., O'Connor, K., Neville, R.W.J., Lazarus, N.R. & Stone, D. (1979). Evolutionary change in the insulin receptors of hystricomorph rodents. *Nature, Lond.* **279**, 439–40.
Huang, S.L., Singh, M. & Kojima, K. (1971). A study of frequency-dependent selection observed in the esterase-6 locus of *Drosophila melanogaster* using a conditioned media method. *Genetics*, **68**, 97–104.
Hunt, L.T., Hurst-Calderone, S. & Dayhoff, M.O. (1978). Globins. In *Atlas of Protein Sequence and Structure*, vol. 5, supplement 3, ed. M.O. Dayhoff. Washington, DC: National Biomedical Research Foundation.
Ikemura, T. (1980). The frequency of codon usage in *E. coli* genes: correlation with abundance of cognate tRNA. In *Genetics and Evolution of RNA Polymerase*, tRNA and

Ribosomes, ed. S. Osawa, H. Ozeki, H. Uchida & T. Yura, pp. 519–23. University of Tokyo Press.
Ikemura, T. (1981a). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.* **146**, 1–21.
Ikemura, T. (1981b). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J. Mol. Biol.* **151**, 389–409.
Ingram, V.M. (1963). *The Hemoglobins in Genetics and Evolution.* Columbia University Press.
Isenberg, I. (1979). Histones. *Annu. Rev. Biochem.* **48**, 159–91.
Jahn, C.L., Hutchison, III. C.A., Phillips, S.J., Weaver, S., Haigwood, N.L., Volvia, C.F. & Edgell, M.H. (1980). DNA sequence organization of the β-globin complex in the BALB/c mouse. *Cell*, **21**, 159–68.
Jensen, E. Ø., Paudan, K., Hyldig-Nielsen, J.J., Jørgensen, P. & Marcker, K.A. (1981). The structure of a chromosomal leghaemoglobin gene from soybean. *Nature, Lond.* **291**, 677–9.
Johnson, G.B. (1972). Evidence that enzyme polymorphisms are not selectively neutral. *Nature, New Biol.* **237**, 170–1.
Johnson, G.B. (1974). On the estimation of effective number of alleles from electrophoretic data. Genetics, **78**, 771–6.
Jones, J.S. (1980). How much genetic variation? *Nature, Lond.* **288**, 10–11.
Jukes, T.H. (1969). Recent advances in studies of evolutionary relationships between proteins and nucleic acids. *Space Life Sci.* **1**, 469–90.
Jukes, T.H. (1971). Comparison of the polypeptide chains of globins. *J. Mol. Evol.* **1**, 46–62.
Jukes, T.H. (1978a). Codons and nearest-neighbor nucleotide pairs in mammalian messenger RNA. *J. Mol. Evol.* **11**, 121–7.
Jukes, T.H. (1978b). Neutral changes during divergent evolution of hemoglobins. *J. Mol. Evol.* **11**, 267–9.
Jukes, T.H. (1979). Dr. Best, insulin, and molecular evolution. *Can. J. Biochem.* **57**, 455–8.
Jukes, T.H. (1980a). Silent nucleotide substitutions and the molecular evolutionary clock. *Science*, **210**, 973–8.
Jukes, T.H. (1980b). Silent nucleotide substitutions during evolution. *Naturwissenschaften*, **67**, 534–9.
Jukes, T.H. & Cantor, C.R. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism* III. ed. H.N. Munro, pp. 21–132. New York: Academic Press.
Jukes, T.H. & Holmquist, R. (1972). Evolutionary clock: nonconstancy of rate in different species. *Science*, **177**, 530–2.
Jukes, T.H. & King, J.L. (1971). Deleterious mutations and neutral substitutions. *Nature, Lond.* **231**, 114–15.
Jukes, T.H. & King, J.L. (1975). Evolutionary loss of ascorbic acid synthesizing ability. *J. Hum. Evol.* **4**, 85–8.
Jukes, T.H. & King, J.L. (1979). Evolutionary nucleotide replacement in DNA. *Nature, Lond.* **281**, 605–6.
Karlin, S. & Lieberman, U. (1974). Random temporal variation in selection intensities: case of large population size. *Theor. Pop. Biol.* **6**, 355–82.
Karn, M.N. & Penrose, L.S. (1951). Birth weight and gestation time in relation to maternal age, parity and infant survival. *Ann. Eugen.* **16**, 147–64.
Kedes, L.H. (1979). Histone genes and histone messengers. *Annu. Rev. Biochem.* **48**, 837–70.

Kerr, R.A. (1980). Asteroid theory of extinctions strengthened. *Science*, **210**, 514–17.
Kerr, W.E. (1967). Multiple alleles and genetic load in bees. *J. Apicult. Res.* **6**, 61–4.
Kesten, H. (1980). The number of alleles in electrophoretic experiments. *Theor. Pop. Biol.* **18**, 290–4.
Kettlewell, H.B.D. (1955). Selection experiments on industrial melanism in the *Lepidoptera. Heredity*, **9**, 323–42.
Kettlewell, H.B.D. (1956). Further selection experiments on industrial melanism in the *Lepidoptera. Heredity*, **10**, 287–301.
Kettlewell, H.B.D. (1958). A survey of the frequencies of *Biston betularia* (L.) (Lep.) and its melanic forms in Great Britain. *Heredity*, **12**, 51–72.
Kim, S.H., Quigley, G.J., Suddath, F.L., McPherson, A., Sneden, D., Kim, J.J., Weinzierl, J. & Rich, A. (1973). Three-dimensional structure of yeast phenylalanine transfer RNA: folding of the polynucleotide chain. *Science*, **179**, 285–8.
Kim, S.H., Suddath, F.L., Quigley, G.J., McPherson, A., Sussman, J.L., Wang, A.H.J., Seeman, N.C. & Rich, A. (1974). Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science*, **185**, 435–40.
Kimura, M. (1955). Solution of a process of random genetic drift with a continuous model. *Proc. Natl. Acad. Sci. USA*, **41**, 144–50.
Kimura, M. (1957). Some problems of stochastic processes in genetics. *Ann. Math. Stat.* **28**, 882–901.
Kimura M. (1960a). Optimum mutation rate and degree of dominance as determined by the principle of minimum genetic load. *J. Genet.* **57**, 21–34.
Kimura, M. (1960b). Genetic load of a population and its significance in evolution (Japanese with English summary). *Jap. J. Genet.* **35**, 7–33.
Kimura, M. (1961). Some calculations on the mutational load. *Jap. J. Genet.* **36** (supplement), 179–90.
Kimura, M. (1962). On the probability of fixation of mutant genes in a population. *Genetics*, **47**, 713–19.
Kimura, M. (1964). Diffusion models in population genetics. *J. Appl. Probab.* **1**, 177–232.
Kimura, M. (1968a). Evolutionary rate at the molecular level. *Nature, Lond.* **217**, 624–6.
Kimura, M. (1968b). Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genet. Res., Camb.* **11**, 247–69.
Kimura, M. (1969a). The rate of molecular evolution considered from the standpoint of population genetics. *Proc. Natl. Acad. Sci. USA*, **63**, 1181–8.
Kimura, M. (1969b). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations, *Genetics*, **61**, 893–903.
Kimura, M. (1970). The length of time required for a selectively neutral mutant to reach fixation through random frequency drift in a finite population. *Genet. Res., Camb.* **15**, 131–3.
Kimura, M. (1971). Theoretical foundation of population genetics at the molecular level. *Theor. Pop. Biol.* **2**, 174–208.
Kimura, M. (1974). Gene pool of higher organisms as a product of evolution. *Cold Spring Harbor Symp. Quant. Biol.* **38**, 515–24.
Kimura, M. (1977). Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature, Lond.* **267**, 275–6.
Kimura, M. (1978). Change of gene frequencies by natural selection under population number regulation. *Proc. Natl. Acad. Sci. USA*, **75**, 1934–7.
Kimura, M. (1979a). Model of effectively neutral mutations in which selective constraint is incorporated. *Proc. Natl. Acad. Sci. USA*, **76**, 3440–4.
Kimura, M. (1979b). The neutral theory of molecular evolution. *Scientific American*, **241** (5), 94–104.

Kimura, M. (1980a). Average time until fixation of a mutant allele in a finite population under continued mutation pressure: studies by analytical, numerical, and pseudo-sampling methods. *Proc. Natl. Acad. Sci. USA*, **77**, 522–6.
Kimura, M. (1980b). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–20.
Kimura, M. (1981a). Was globin evolution very rapid in its early stages?: a dubious case against the rate-constancy hypothesis. *J. Mol. Evol.* **17**, 110–13.
Kimura, M. (1981b). Doubt about studies of globin evolution based on maximum parsimony codons and the augmentation procedure. *J. Mol. Evol.* **17**, 121–2.
Kimura, M. (1981c). Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA*, **78**, 454–8.
Kimura, M. (1981d). Possibility of extensive neutral evolution under stabilizing selection with special reference to non-random usage of synonymous codons. *Proc. Natl. Acad. Sci. USA*, **78**, 5773–7.
Kimura, M. & Crow, J.F. (1964). The number of alleles that can be maintained in a finite population. *Genetics*, **49**, 725–38.
Kimura, M. & Crow, J.F. (1978). Effect of overall phenotypic selection on genetic change at individual loci. *Proc. Natl. Acad. Sci. USA*, **75**, 6168–71.
Kimura, M. & King, J.L. (1979). Fixation of a deleterious allele at one of two "duplicate" loci by mutation pressure and random drift. *Proc. Natl. Acad. Sci. USA*, **76**, 2858–61.
Kimura, M. & Maruyama, T. (1966). The mutational load with epistiatic gene interactions in fitness. *Genetics*, **54**, 1337–51.
Kimura, M. & Maruyama, T. (1969). The substitutional load in a finite population. *Heredity*, **24**, 101–14.
Kimura, M. & Maruyama, T. (1971). Pattern of neutral polymorphism in a geographically structured population. *Genet. Res., Camb.* **18**, 125–31.
Kimura, M., Maruyama, T. & Crow, J.F. (1963). The mutational load in small populations. *Genetics*, **48**, 1303–12.
Kimura, M. & Ohta, T. (1969a). The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, **61**, 763–71.
Kimura, M. & Ohta, T. (1969b). The average number of generations until extinction of an individual mutant gene in a finite population. *Genetics*, **63**, 701–9.
Kimura, M. & Ohta, T. (1970). Genetic load at a polymorphic locus which is maintained by frequency-dependent selection. *Genet. Res., Camb.* **16**, 145–50.
Kimura, M. & Ohta, T. (1971a). Protein polymorphism as a phase of molecular evolution. *Nature, Lond.* **229**, 467–9.
Kimura, M. & Ohta, T. (1971b). *Theoretical Aspects of Population Genetics*. Princeton: Princeton University Press.
Kimura, M. & Ohta, T. (1972). On the stochastic model for estimation of mutational distance between homologous proteins. *J. Mol. Evol.* **2**, 87–90.
Kimura, M. & Ohta, T. (1973a). Mutation and evolution at the molecular level. *Genetics*, 73 (supplement), 19–35.
Kimura, M. & Ohta, T. (1973b). The age of a neutral mutant persisting in a finite population. *Genetics*, **75**, 199–212.
Kimura, M. & Ohta, T. (1974). On some principles governing molecular evolution. *Proc. Natl. Acad. Sci. USA*, **71**, 2848–52.
Kimura, M. & Ohta, T. (1975). Distribution of allelic frequencies in a finite population under stepwise production of neutral alleles. *Proc. Natl. Acad. Sci. USA*, **72**, 2761–4.
Kimura, M. & Ohta, T. (1977). Further comments on "Counter-examples to a neutralist hypothesis". *J. Mol. Evol.* **9**, 367–8.

Kimura, M. & Ohta, T. (1978). Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proc. Natl. Acad. Sci USA*, **75**, 2868–72.
Kimura, M. & Ohta, T. (1979). Population genetics of multigene family with special reference to decrease of genetic correlation with distance between gene members on a chromosome. *Proc. Natl. Acad. Sci. USA*, **76**, 4001–5.
Kimura, M. & Weiss, G.H. (1964). The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, **49**, 561–76.
King, J.L. (1967). Continuously distributed factors affecting fitness. *Genetics*, **55**, 483–92.
King, J.L. & Jukes, T.H. (1969). Non-Darwinian evolution. *Science*, **164**, 788–98.
King, J.L. & Ohta, T. (1975). Polyallelic mutational equilibria. *Genetics*, **79**, 681–91.
King, M.C. & Wilson, A.C. (1975). Evolution at two levels: Molecular similarities and biological differences between human and chimpanzees. *Science*, **188**, 107–16.
Kingman, J.F.C. (1976). Coherent random walks arising in some genetical models. *Proc. Roy. Soc. London, Ser. A*, **351**, 19–31.
Kingman, J.F.C. (1977). A note on multidimensional models of neutral mutation. *Theor. Pop. Biol.* **11**, 285–90.
Klose, J. & Feller, M. (1981). Genetic variability of proteins from plasma membranes and cytosols of mouse organs. *Biochem. Genet.*, **19**, 859–70.
Klug, A., Ladner, J. & Robertus, J.D. (1974). The structural geometry of co-ordinated base changes in transfer RNA. *J. Mol. Biol.* **89**, 511–16.
Koch, R.E. (1971). The influence of neighboring base pairs upon base-pair substitution mutation rates. *Proc. Natl. Acad. Sci. USA*, **68**, 773–6.
Koehn, R.K. & Eanes, W.F. (1977). Subunit size and genetic variation of enzymes in natural populations of *Drosophila*. *Theor. Pop. Biol.* **11**, 330–41.
Koehn, R.K. & Eanes, W.F. (1978). Molecular structure and protein variation within and among populations. In *Evolutionary Biology*, vol. 11, ed. M.K. Hecht, W.C. Steere & B. Wallace, pp. 39–100.
Kojima, K., Gillespie, J.H. & Tobari, Y.N. (1970). A profile of *Drosophila* species' enzymes assayed by electrophoresis, I. Number of alleles, heterozygosities, and linkage disequilibrium in glucose metabolizing systems and some other enzymes. *Biochem. Genet.* **4**, 627–37.
Kojima, K. & Tobari, Y.N. (1969). The pattern of viability changes associated with genotype frequency at the alcohol dehydrogenase locus in a population of *Drosophila melanogaster*. *Genetics*, **61**, 201–9.
Kojima, K. & Yarbrough, K.M. (1967). Frequency-dependent selection at the esterase 6 locus in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA*, **57**, 645–9.
Komai, T. (1970). Theory of "mutual selection". *Proc. Japan Acad.* **46**, 383–6.
Kondo, S. (1977). Evolutionary considerations on DNA repair and mutagenesis. In *Molecular Evolution and Polymorphism*, ed. M. Kimura, pp. 313–31. Mishima: National Institute of Genetics.
Konkel, D.A., Tilghman, S.M. & Leder, P. (1978). The sequence of the chromosomal mouse β-globin major gene: homologies in capping, splicing and poly (A) sites. *Cell*, **15**, 1125–32.
Kortlandt, A. (1972). *New Perspectives on Ape and Human Evolution*. University of Amsterdam.
Krimbas, C.B. & Tsakas, S. (1971). The genetics of *Dacus oleae*. V. Changes of esterase polymorphism in a natural population following insecticide control – selection or drift? *Evolution*, **25**, 454–60.
Kurtén B. (1971). *The Age of Mammals*. London: Weidenfeld & Nicolson.
Lacy, E. & Maniatis, T. (1980). The nucleotide sequence of a rabbit β-globin pseudogene, *Cell*, **21**, 545–53.

Laki, K. (1965). Enzymatic effects of thrombin. *Fed. Proc.* **24**, 794–9.
Lande, R. (1979). Effective deme sizes during long-term evolution estimated from rates of chromosomal rearrangement. *Evolution*, **33**, 234–51.
Lande, R. (1980). Genetic variation and phenotypic evolution during allopatric speciation. *Amer. Natur.* **116**, 463–79.
Langley, C.H. & Fitch, W.M. (1974). An examination of the constancy of the rate of molecular evolution. *J. Mol. Evol.* **3**, 161–77.
Langley, C.H., Smith, D.B. & Johnson, F.M. (1978). Analysis of linkage disequilibria between allozyme loci in natural populations of *Drosophila melanogaster*. *Genet. Res., Camb.* **32**, 215–29.
Langridge, J. (1974). Mutation spectra and the neutrality of mutations. *Aust. J. Biol. Sci.* **27**, 309–19.
Latter, B.D.H. (1975). Influence of selection pressures on enzyme polymorphisms in *Drosophila*. *Nature, Lond.* **257**, 590–1.
Latter, B.D.H. (1981a). Gene frequency distributions in species of *Drosophila*. In *Genetic Studies of Drosophila Populations*, ed. J.B. Gibson & J.G. Oakeshott, pp. 3–13. Australian National University Press.
Latter, B.D.H. (1981b). The distribution of heterozygosity in temperate and tropical species of Drosophila. *Genet. Res., Camb.* **38**, 137–56.
Lawn, R.M., Efstratiadis, A., O'Connell, C. & Maniatis, T. (1980). The nucleotide sequence of the human β-globin gene. *Cell*, **21**, 647–51.
Leder, P., Hansen, J.N., Konkel, D., Leder, A., Nishioka, Y. & Talkington, C. (1980). Mouse globin system: a functional and evolutionary analysis. *Science*, **209**, 1336–42.
Lerner, I.M. (1954). *Genetic Homeostasis*. Edinburgh: Oliver & Boyd.
Lesk, A.M. & Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136**, 225–70.
Levins, R. (1968). *Evolution in Changing Environments*. Princeton University Press.
Lewontin, R.C. (1974). *The Genetic Basis of Evolutionary Change*. Columbia University Press.
Lewontin, R.C., Ginzburg, L.R. & Tuljapurkar, S.D. (1978). Heterosis as an explanation for large amounts of genic polymorphism. *Genetics*, **88**, 149–70.
Lewontin, R.C. & Hubby, J.L. (1966). A molecular approach to the study of genic heterozygosity in natural populations, II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics*, **54**, 595–609.
Lewontin, R.C. & Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, **74**, 175–95.
Lewontin, R.C. & Krakauer, J. (1975). Testing the heterogeneity of F values. *Genetics*, **80**, 397–8.
Li, W.-H. (1976). A mixed model of mutation for electrophoretic identity of proteins within and between populations. *Genetics*, **83**, 423–32.
Li, W.-H. (1977a) *Stochastic Models in Population Genetics*, Benchmark Papers in Genetics Series Vol. 7. Stroudsburg, Pennsylvania: Dowden, Hutchinson and Ross.
Li, W.-H. (1977b). Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. *Genetics*, **85**, 331–7.
Li, W.-H. (1978). Maintenance of genetic variability under the joint effect of mutation, selection and random drift. *Genetics*, **90**, 349–82.
Li, W.-H. (1979). Maintenance of genetic variability under the pressure of neutral and deleterious mutations in a finite population. *Genetics*, **92**, 647–67.
Li, W.-H. Gojobori, T. & Nei, M. (1981). Pseudogenes as a paradigm of neutral evolution. *Nature, Lond.* **292**, 237–9.

Li, W.-H. & Nei, M. (1975). Drift variances of heterozygosity and genetic distance in transient states. *Genet. Res., Camb.* **25**, 229–48.
Lomedico, P., Rosenthal, N., Efstratiadis, A., Gilbert, W., Kolodner, R. & Tizard, R. (1979). The structure and evolution of the two nonallelic rate preproinsulin genes. *Cell*, **18**, 545–58.
Loukas, M., Krimbas, C.B. & Vergini, Y. (1979). The genetics of *Drosophila subobscura* populations. IX. Studies on linkage disequilibrium in four natural populations. *Genetics*, **93**, 497–523.
Loukas, M., Vergini, Y. & Krimbas, C.B. (1981). The genetics of *Drosophila subobscura* populations. XVII. Further genic heterogeneity within electromorphs by urea denaturation and the effect of the increased genic variability on linkage disequilibrium studies. *Genetics*, **97**, 429–41.
Lueders, K., Leder, A., Leder, P. & Kuff, E. (1982). Association between a transposed α-globin pseudogene and retrovirus-like elements in the BALB/c mouse genome. *Nature, Lond.* **295**, 426–8.
MacArthur, R.H. & Wilson, E.O. (1967). *The Theory of Island Biogeography*. Princeton University Press.
McConkey, E.H., Taylor, B.J. & Phan, D. (1979). Human heterozygosity: a new estimate. *Proc. Natl. Acad. Sci. USA*, **76**, 6500–4.
McDonald, J.F. & Ayala, F.J. (1974). Genetic response to environmental heterogeneity. *Nature, Lond.* **250**, 572–4.
McFarland, W.N., Pough, F.H., Cade, T.J. & Heiser, J.B. (1979). *Vertebrate Life*. New York: Macmillan, London: Collier Macmillan.
McKenzie, J.A. & McKechnie, S.W. (1978). Ethanol tolerance and the *Adh* polymorphism in a natural population of *Drosophila melanogaster*. *Nature, Lond.* **272**, 75–6.
McLachlan, A.D. (1972). Repeating sequences and gene duplication in proteins. *J. Mol. Biol.* **64**, 417–37.
McLaughlin, P.J. & Dayhoff, M.O. (1972). Evolution of species and proteins: a time scale. In *Atlas of Protein Sequence and Structure* 1972, ed. M.O. Dayhoff, pp. 47–52. Washington, DC: National Biomedical Research Foundation.
Maita, T., Tanaka, E., Goodman, M. & Matsuda, G. (1977). Amino acid sequences of the α and β chains of adult hemoglobin of the tupai, *Tupai aglis*. *J. Biochem.* **82**, 603–5.
Malécot, G. (1967). Identical loci and relationship. In *Proc. Fifth Berkeley Symp. Math. Sta. Prob.* vol. 4, pp. 317–32. Berkeley: University of California Press.
Margoliash, E. & Smith, E.L. (1965). Structure and functional aspects of cytochrome c in relation to evolution. In *Evolving Genes and Proteins*, ed. V. Bryson & H.J. Vogel, pp. 221–42. New York: Academic Press.
Marotta, C.A., Wilson, J.T., Forget, B.G. & Weissman, S.M. (1977). Human β-globin messenger RNA. III. Nucleotide sequences derived from complementary DNA. *J. Biol. Chem.* **252**, 5040–53.
Marshall, D.R. & Brown, A.H.D. (1975). The charge-state model of protein polymorphisms in natural populations. *J. Mol. Evol.* **6**, 149–63.
Martial, J.A., Hallewell, R.A., Baxter, J.D. & Goodman, H.M. (1979). Human growth hormone: Complementary DNA cloning and expression in bacteria. *Science*, **205**, 602–7.
Maruyama, T. (1970a). On the fixation probability of mutant genes in a subdivided population. *Genet. Res., Camb.* **15**, 221–6.
Maruyama, T. (1970b). Effective number of alleles in a subdivided population. *Theor. Pop. Biol.* **1**, 273–306.
Maruyama, T. (1971). An invariant property of a structured population. *Genet. Res., Camb.* **18**, 81–4.
Maruyama, T. (1972a). Some invariant properties of a geographically structured finite

population: Distribution of heterozygotes under irreversible mutation. *Genet. Res., Camb.* **20**, 141–9.
Maruyama, T. (1972b). Rate of decrease of genetic variability in a two-dimensional continuous population of finite size. *Genetics*, **70**, 639–51.
Maruyama, T. (1974). The age of an allele in a finite population. *Genet. Res., Camb.* **23**, 137–43.
Maruyama, T. (1977). *Stochastic Problems in Population Genetics* (Lecture Notes in Biomathematics 17). Berlin: Springer-Verlag.
Maruyama, T. & Kimura, M. (1971). Some methods for treating continuous stochastic processes in population genetics. *Jap. J. Genet.* **46**, 407–10.
Maruyama, T. & Kimura, M. (1974). Geographical uniformity of selectively neutral polymorphisms. *Nature, Lond.* **249**, 30–2.
Maruyama, T. & Kimura, M. (1975). Moments for sum of an arbitrary function of gene frequency along a stochastic path of gene frequency change. *Proc. Natl. Acad. Sci. USA*, **72**, 1602–4.
Maruyama, T. & Kimura, M. (1978). Theoretical study of genetic variability, assuming stepwise production of neutral and very slightly deleterious mutations. *Proc. Natl. Acad. Sci. USA*, **75**, 919–22.
Maruyama, T. & Kimura, M. (1980). Genetic variability and effective population size when local extinction and recolonization of subpopulations are frequent. *Proc. Natl. Acad. Sci. USA*, **77**, 6710–14.
Mather, K. (1953). The genetical structure of populations. *Symp. Soc. Exp. Biol.* **7**, 66–95.
Mather, K. (1973). *Genetic Structure of Populations*. London: Chapman & Hall.
Matsubara, H., Hase, T., Wakabayashi, S. & Wada, K. (1978). Gene duplications during evolution of chloroplast-type ferredoxins. In *Evolution of Protein Molecules*, ed. H. Matsubara & T. Yamanaka, pp. 209–20. Tokyo: Japan Scientific Soc. Press, Center for Academic Publications.
Matsuda, G. (1978). Molecular evolution of hemoglobins. In *Evolution of Protein Molecules*, ed. H. Matsubara & T. Yamanaka, pp. 175–85. Tokyo: Japan Scientific Soc. Press.
Matsuda, H., Gojobori, T. & Takahata, N. (1978). Theoretical study on protein polymorphism and its bearing on the evolution of protein molecules. In *Evolution of Protein Molecules*, ed. H. Matsubara & T. Yamanaka, pp. 89–100. Tokyo: Japan Scientific Soc. Press.
Maynard Smith, J. (1968). "Haldane's dilemma" and the rate of evolution. *Nature, Lond.* **219**, 1114–16.
Maynard Smith, J. (1972). Protein polymorphism. *Nature, New Biol.* **237**, 31.
Maynard Smith, J. & Hoekstra, R. (1980). Polymorphism in a varied environment: how robust are the models? *Genet. Res., Camb.* **35**, 45–57.
Mayr, E. (1955). Integration of genotypes: synthesis. *Cold Spring Harbor Symp. Quant. Biol.* **20**, 327–33.
Mayr, E. (1959). Where are we? *Cold Spring Harbor Symp. Quant. Biol.* **24**, 1–14.
Mayr, E. (1963). *Animal Species and Evolution*. Harvard University Press.
Mayo, O. (1970). The role of duplications in evolution. *Heredity*, **25**, 543–53.
Maxam, A.M. & Gilbert, W. (1977). A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA*, **74**, 560–4.
Milkman, R. (1967). Heterosis as a major cause of heterozygosity in nature. *Genetics*, **55**, 493–5.
Milkman, R. (1973). Electrophoretic variation in *Escherichia coli* from natural sources. *Science*, **182**, 1024–6.
Milkman, R. (1976). Selection is the major determinant. *Trend Biochem. Sci.* **1**, N152–N154.

Milkman, R. (1978). Selection differentials and selection coefficients. *Genetics*, **88**, 391–403.
Minawa, A. & Birley, A.J. (1978). The genetical response to natural selection by varied environments. 1. Short-term observations. *Heredity*, **40**, 39–50.
Miyata, T. & Hayashida, H. (1981). Extraordinarily high evolutionary rate of pseudogenes: evidence for the presence of selective pressure against changes between synonymous codons. *Proc. Natl. Acad. Sci. USA*, **78**, 5739–43.
Miyata, T., Miyazawa, S. & Yasunaga, T. (1979). Two types of amino acid substitution in protein evolution. *J. Mol. Evol.* **12**, 219–36.
Miyata, T. & Yasunaga, T. (1980). Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J. Mol. Evol.* **16**, 23–36.
Miyata, T. & Yasunaga, T. (1981). Rapidly evolving mouse α-globin-related pseudogene and its evolutionary history. *Proc. Natl. Acad. Sci. USA*, **78**, 450–3.
Miyata, T., Yasunaga, T. & Nishida, T. (1980b). Nucleotide sequence divergence and functional constraint in mRNA evolution. *Proc. Natl. Acad. Sci. USA*, **77**, 7328–32.
Miyata, T., Yasunaga, T., Yamawaki-Kataoka, Y., Obata, M. & Honjo, T. (1980a). Nucleotide sequence divergence of mouse immunoglobulin γ_1 and γ_{2b} chain genes and the hypothesis of intervening sequence-mediated domain transfer. *Proc. Natl. Acad. Sci. USA*, **77**, 2143–7.
Modiano, G., Battistuzzi, G. & Motulsky, A.G. (1981). Nonrandom patterns of codon usage and of nucleotide substitutions in human α- and β-globin genes: an evolutionary strategy reducing the rate of mutations with drastic effects? *Proc. Natl. Acad. Sci. USA*, **78**, 1110–14.
Moran, P.A.P. (1975). Wandering distributions and the electrophoretic profile. *Theor. Pop. Biol.* **8**, 318–30.
Moran, P.A.P. (1976). Wandering distributions and the electrophoretic profile. II. *Theor. Pop. Biol.* **10**, 145–9.
Morgan, P. (1976). Frequency-dependent selection at two enzyme loci in *Drosophila melanogaster*. *Nature, Lond.* **263**, 765–6.
Mross, G.A. & Doolittle, R.F. (1967). Amino acid sequence studies on artiodactyl fibrinopeptides II. Vicuna, elk, muntjak, pronghorn antelope, and water buffalo. *Arch. Biochem. Biophys.* **122**, 674–84.
Mukai, T. (1964). The genetic structure of natural populations of *Drosophila melanogaster*. I. Spontaneous mutation rate of polygenes controlling viability. *Genetics*, **50**, 1–19.
Mukai, T. (1977). Lack of experimental evidence supporting selection for the maintenance of isozyme polymorphisms in *Drosophila melanogaster*. In *Molecular Evolution and Polymorphism*, ed. M. Kimura, pp. 103–26. Mishima: National Institute of Genetics.
Mukai, T., Chigusa, S.I., Mettler, L.E. & Crow, J.F. (1972). Mutation rate and dominance of genes affecting viability in *Drosophila melanogaster*. *Genetics*, **72**, 335–55.
Mukai, T., Mettler, L.E. & Chigusa, S.I. (1971). Linkage disequilibrium in a local population of *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA*, **68**, 1065–9.
Mukai, T., Schaffer, H.E. & Cockerham, C.C. (1972). Genetic consequences of truncation selection at the phenotypic level in *Drosophila melanogaster*. *Genetics*, **72**, 763–9.
Mukai, T., Tachida, H. & Ichinose, M. (1980). Selection for viability at loci controlling protein polymorphisms in *Drosophila melanogaster* is very weak at most. *Proc. Natl. Acad. Sci. USA*, **77**, 4857–60.
Mukai, T. & Voelker, R.A. (1977). The genetic structure of natural populations of *Drosophila melanogaster*. XIII. Further studies on linkage disequilibrium. *Genetics*, **86**, 175–85.
Mukai, T., Watanabe, T.K. & Yamaguchi, O. (1974). The genetic structure of natural populations of *Drosophila melanogaster*. XII. Linkage disequilibrium in a large local

population. *Genetics*, **77**, 771–93.
Mukai, T. & Yamaguchi, O. (1974). The genetic structure of natural populations of *Drosophila melanogaster*. XI. Genetic variability in a local population. *Genetics*, **76**, 339–66.
Mukai, T. & Yamazaki, T. (1964). Position effect of spontaneous mutant polygenes controlling viability in *Drosophila melanogaster*. *Proc. Japan Academy*, **40**, 840–5.
Mukai, T. & Yamazaki, T. (1980). Test for selection on polymorphic isozyme genes using the population cage method. *Genetics*, **96**, 537–42.
Muller, H.J. (1922). Variation due to change in the individual gene. *Amer. Natur.* **56**, 32–50.
Muller, H.J. (1927). Artificial transmutation of the gene. *Science*, **66**, 84–7.
Muller, H.J. (1929). The gene as the basis of life. *Proc. Int. Congr. Plant Sci.* **1**, 897–921.
Muller, H.J. (1933). Further studies on the nature and causes of gene mutations. *Proc. Sixth Inter. Congr. Genet.*, **1**, 213–55.
Muller, H.J. (1949). Redintegration of the symposium on genetics, paleontology, and evolution. In *Genetics, Paleontology, and Evolution*, ed. G.L. Jepsen, G.G. Simpson & E. Mayr, pp. 421–45. Princeton University Press.
Muller, H.J. (1950). Our load of mutations. *Amer. J. Hum. Genet.* **2**, 111–76.
Muller, H.J. (1958). The mutation theory reexamined. *Proc. X Inter. Congr. Genet.* **1**, 306–17.
Muller, H.J. (1959). The guidance of human evolution. *Perspectives in Biology and Medicine*, **3** (1), 1–43.
Muller, H.J. (1960). Evolution and genetics. *Evoluzione e Genetica*, pp. 15–37. Rome: Accademia Nationale dei Lincei.
Muller, H.J. (1962). *Studies in Genetics*. Indiana University Press.
Muller, H.J. (1966). The gene material as the initiator and the organizing basis of life. *Amer. Natur.* **100**, 493–518.
Muller, H.J. (1967). The gene material as the initiator and the organizing basis of life. In *Heritage from Mendel*, ed. R.A. Brink, pp. 417–47. University of Wisconsin Press.
Nagylaki, T. (1974). The moments of stochastic integrals and the distribution of sojourn times. *Proc. Natl. Acad. Sci. USA*, **71**, 746–9.
Nagylaki, T. & Petes, T.D. (1982). Intrachromosomal gene conversion and the maintenance of sequence homogeneity among repeated genes. *Genetics* **110**, 315–337.
Narain, P. (1970). A note on the diffusion approximation for the variance of the number of generations until fixation of a neutral mutant gene. *Genet. Res., Camb.* **15**, 251–5.
Neel, J.V. (1978). Rare variants, private polymorphisms, and locus heterozygosity in Ameridian populations. *Amer. J. Hum. Genet.* **30**, 465–90.
Nei, M. (1969). Gene duplication and nucleotide substitution in evolution. *Nature, Lond.* **221**, 40–2.
Nei, M. (1972). Genetic distance between populations. *Amer. Natur.* **106**, 283–92.
Nei, M. (1975). *Molecular Population Genetics and Evolution*. Amsterdam: North-Holland.
Nei, M. (1976). The cost of natural selection and the extent of enzyme polymorphism. *Trend Biochem. Sci.* **1**, N247–N248.
Nei, M. (1977). Estimation of mutation rate from rare protein variants. *Amer. J. Hum. Genet.* **29**, 225–32.
Nei, M. (1978a). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, **89**, 583–90.
Nei, M. (1978b). The theory of genetic distance and evolution of human rares. *Jap. J. Hum. Genet.* **23**, 341–69.
Nei, M. (1980). Stochastic theory of population genetics and evolution. In *Vito Volterra Symposium on Mathematical Models in Biology* (*Lecture Notes in Biomathematics* 39), ed. C. Barigozzi, pp. 17–47. Berlin: Springer-Verlag.

Nei, M. & Chakraborty, R. (1976). Empirical relationship between the number of nucleotide substitutions and interspecific identity of amino acid sequences in some proteins. *J. Mol. Evol.* **7**, 313–23.
Nei, M., Chakraborty, R. & Fuerst, P.A. (1976a). Infinite allele model with varying mutation rate. *Proc. Natl. Acad. Sci. USA*, **73**, 4164–8.
Nei, M. & Feldman, M.W. (1972). Identity of genes by descent within and between populations under mutation and migration pressures. *Theoret. Pop. Biol.* **3**, 460–5.
Nei, M., Fuerst, P.A. & Chakraborty, R. (1976b). Testing the neutral mutation hypothesis by distribution of single locus heterozygosity. *Nature, Lond.* **262**, 491–3.
Nei, M., Fuerst, P.A. & Chakraborty, R. (1978). Subunit molecular weight and genetic variability of proteins in natural populations. *Proc. Natl. Acad. Sci. USA*, **75**, 3359–62.
Nei, M. & Imaizumi, Y. (1966). Genetic structure of human populations II. Differentiation of blood group gene frequencies among isolated human populations. *Heredity*, **21**, 183–90.
Nei, M. & Li, W.-H. (1980). Non-random association between electromorphs and inversion chromosomes in finite populations. *Genet. Res., Camb.* **35**, 65–83.
Nei, M. & Maruyama, T. (1975). Lewontin-Krakauer test for neutral genes. *Genetics*, **80**, 395.
Nei, M., Maruyama, T. & Chakraborty, R. (1975). The bottleneck effect and genetic variability in populations. *Evolution*, **29**, 1—10.
Nei, M. & Murata, M. (1966). Effective population size when fertility is inherited. *Genet. Res., Camb.* **8**, 257–60.
Nei, M. & Roychoudhury, A.K. (1973). Probability of fixation and mean fixation time of an overdominant mutation. *Genetics*, **74**, 371–80.
Nei, M. & Roychoudhury, A.K. (1974). Sampling variances of heterozygosity and genetic distances. *Genetics*, **76**, 379–90.
Nei, M. & Tateno, Y. (1975). Interlocus variation of genetic distance and the neutral mutation theory. *Proc. Natl. Acad. Sci. USA*, **72**, 2758–60.
Nevo, E. (1978). Genetic variation in natural populations: patterns and theory. *Theor. Pop. Biol.* **13**, 121–77.
Nishioka, Y., Leder, A. & Leder, P. (1980). Unusual α-globin-like gene that has clearly lost both globin intervening sequences. *Proc. Natl. Acad. Sci. USA*, **77**, 2806–9.
Nichols, B.P. & Yanofsky, Ch. (1979). Nucleotide sequences of *trpA* of *Salmonella typhimurium* and *Escherichia coli*: An evolutionary comparison. *Proc. Natl. Acad. Sci. USA*, **76**, 5244–8.
Novotný, J. (1973). Genealogy of immunoglobulin polypeptide chains: a consequence of amino acid interactions, conserved in their tertiary structure. *J. Theor. Biol.* **41**, 171–80.
Nozawa, K., Shotake, T., Kawamoto, Y. & Tanabe, Y. (1982). Population genetics of Japanese monkeys: II. blood protein polymorphisms and population structure. *Primates*, **23**, 252–71.
Nozawa, K., Shotake, T. & Okura, Y. (1975). Blood protein polymorphisms and population structure of the Japanese macaque, *Macaca fuscata fuscata*. In *Isozymes IV. Genetics and Evolution*, ed. C.L. Markert, pp. 225–41. New York: Academic Press.
Oakeshott, J.G. & Gibson, J.B. (1981). Is there selection by environmental ethanol on the alcohol dehydrogenase locus in *Drosophila melanogaster*? In *Genetics Studies of Drosophila Populations*, ed. J.B. Gibson & J.G. Oakeshott, pp. 103–20. Australian National University Press.
O'Donald, P. (1969). "Haldane's dilemma" and the rate of natural selection. *Nature, Lond.* **221**, 815–16.
Ohba, S. (1977). *Genetics of Populations* (in Japanese). University of Tokyo Press.
Ohno, S. (1970). *Evolution by Gene Duplication*. Berlin: Springer-Verlag.

Ohno, S. (1972). So much "junk" DNA in our genome. In *Brookhaven Symposia in Biology No.* 23, ed. H.H. Smith, pp. 366–70. New York: Gordon & Breach.
Ohno, S. (1973). Ancient linkage groups and frozen accidents. *Nature, Lond.* **244**, 259–62.
Ohno, S., Wofl, U. & Atkin, N.B. (1968). Evolution from fish to mammals by gene duplication. *Hereditas*, **59**, 169–87.
Ohta, T. (1971). Associative overdominance caused by linked detrimental mutations. *Genet. Res., Camb.* **18**, 277–86.
Ohta, T. (1972). Population size and rate of evolution. *J. Mol. Evol.* **1**, 305–14.
Ohta, T. (1973a). Slightly deleterious mutant substitutions in evolution. *Nature, Lond.* **246**, 96–8.
Ohta, T. (1973b). Effect of linkage on behavior of mutant genes in a finite population. *Theor. Pop. Biol.* **4**, 145–62.
Ohta, T. (1974). Mutational pressure as the main cause of molecular evolution and polymorphism. *Nature, Lond.* **252**, 351–4.
Ohta, T. (1975). Statistical analyses of *Drosophila* and human protein polymorphisms. *Proc. Natl. Acad. Sci. USA*, **72**, 3194–6.
Ohta, T. (1976a). A simple model for treating the evolution of multigene families. *Nature, Lond.* **263**, 74–6.
Ohta, T. (1976b). Role of very slightly deleterious mutations in molecular evolution and polymorphism. *Theor. Pop. Biol.* **10**, 254–75.
Ohta, T. (1977a). On the gene conversion model as a mechanism for maintenance of homogeneity in systems with multiple genomes. *Genet. Res., Camb.* **30**, 89–91.
Ohta, T. (1977b). Extension to the neutral mutation random drift hypothesis. In *Molecular Evolution and Polymorphism*, ed. M. Kimura, pp. 148–67. Mishima: National Institute of Genetics.
Ohta, T. (1978a). Theoretical study on genetic variation in multigene families. *Genet. Res., Camb.* **31**, 13–28.
Ohta, T. (1978b). Theoretical population genetics of repeated genes forming a multigene family. *Genetics*, **88**, 845–61.
Ohta, T. (1979). An extension of a model for the evolution of multigene families by unequal crossing-over. *Genetics*, **91**, 591–607.
Ohta, T. (1980a). Linkage disequilibrium between amino acid sites in immunoglobulin genes and other multigene families. *Genet. Res., Camb.* **36**, 181–97.
Ohta, T. (1980b). *Evolution and Variation of Multigene Families* (*Lecture Notes in Biomathematics* 37). Berlin: Springer-Verlag.
Ohta, T. (1981a). Genetic variation in small multigene families *Genet. Res., Camb.* **37**, 133–49.
Ohta, T. (1981b). Population genetics of selfish DNA. *Nature, Lond.* **292**, 648–9.
Ohta, T. (1982a). Linkage disequilibrium due to random drift in finite subdivided populations. *Proc. Natl. Acad. Sci. USA*, **79**, 1940–4.
Ohta, T. (1982b). Allelic and non-allelic homology of a supergene family. *Proc. Natl. Acad. Sci. USA*, **79** (in press).
Ohta, T. & Kimura, M. (1970). Development of associative overdominance through linkage disequilibrium in finite populations. *Genet. Res., Camb.* **16**, 165–77.
Ohta, T. & Kimura, M. (1971a). Functional organization of genetic material as a product of molecular evolution. *Nature, Lond.* **233**, 118–19.
Ohta, T. & Kimura, M. (1971b). On the constancy of the evolutionary rate of cistrons. *J. Mol. Evol.* **1**, 18–25.
Ohta, T. & Kimura, M. (1971c). Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. *Genetics*, **68**, 571–80.

Ohta, T. & Kimura, M. (1971d). Behavior of neutral mutants influenced by associated overdominant loci in finite populations. *Genetics*, **69**, 247–60.
Ohta, T. & Kimura, M. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet. Res., Camb.* **22**, 201–4.
Ohta, T. & Kimura, M. (1981). Some calculations on the amount of selfish DNA. *Proc. Natl. Acad. Sci. USA*, **78**, 1129–32.
Omoto, K. (1981). The genetic origins of Philippine Negritos. *Current Anthropology*, **22**, 421–2.
Omoto, K. (1982). Allelic diversity of human populations in the Asian-Pacific area. In *Molecular Evolution, Protein Polymorphism and the Neutral Theory*, ed. M. Kimura, pp. 193–214. Tokyo: Japan Scientific Societies Press/Berlin: Springer-Verlag.
Orgel, L.E. & Crick, F.H.C. (1980). Selfish DNA: the ultimate parasite. *Nature, Lond.* **284**, 604–7.
Orgel, L.E., Crick, F.H.C. & Spienza, C. (1980). Selfish DNA. *Nature, Lond.* **288**, 645–6.
Ozeki, H., Inokuchi, H., Yamao, F., Kodaira, M., Sakano, H., Ikemura, T. & Shimura, Y. (1980). Genetics of nonsense suppressor tRNAs in *Escherichia coli*. In *Transfer RNA: Biological Aspects*, ed. D. Söll, J. Abelson & P. Schimmel, pp. 341–62. Cold Spring Harbor Laboratory.
Pearson, K. (1906). Walter Frank Raphael Weldon, 1860–1906. *Biometrika*, **5**, 1–52.
Pearson, K. (1934). *Tables of the Incomplete Beta-Function*. Cambridge University Press.
Perier, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. & Dodgson, J. (1980). The evolution of genes: the chicken preproinsulin gene. *Cell*, **20**, 555–66.
Perutz, M.F. & Lehmann, H. (1968). Molecular pathology of human haemoglobin. *Nature, Lond.* **219**, 902–9.
Pilbeam, D. (1982). New hominoid skull material from the Miocene of Pakistan. *Nature*, **295**, 232–4.
Post, L.E., Strycharz, G.D., Nomura, M., Lewis, H. & Dennis, P.P. (1979). Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit β in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA*, **76**, 1697–701.
Powell, J.R. (1971). Genetic polymorphisms in varied environments. *Science*, **174**, 1035–6.
Powell, J.R. (1975). Isozymes and non-Darwinian evolution: a re-evaluation. In *Isozymes IV Genetics and Evolution*, ed. C.L. Markert, pp. 9–26. New York: Academic Press.
Prakash, S. & Lewontin, R.C. (1968). A molecular approach to the study of genic heterozygosity in natural populations, III. Direct evidence of coadaptation in gene arrangements of *Drosophila*. *Proc. Natl. Acad. Sci. USA*, **59**, 398–405.
Prakash, S., Lewontin, R.C. & Hubby, J.L. (1969). A molecular approach to the study of genic heterozygosity in natural populations. IV. Patterns of genic variation in central, marginal and isolated populations of *Drosophila pseudoobscura*. *Genetics*, **61**, 841–58.
Proudfoot, N. (1980). Pseudogenes. *Nature, Lond.* **286**, 840–1.
Proudfoot, N.J. & Maniatis, T. (1980). The structure of a human α-globin pseudogene and its relationship to α-globin gene duplication. *Cell*, **21**, 537–44.
Prout, T. (1965). The estimation of fitnesses from genotypic frequencies. *Evolution*, **19**, 546–51.
Provine, W.B. (1971). *The Origins of Theoretical Population Genetics*. The University of Chicago Press.
Racine, R.R. & Langley, C.H. (1980). Genetic heterozygosity in a natural population of *Mus musculus* assessed using two-dimensional electrophoresis. *Nature, Lond.* **283**, 855–7.
Ramshaw, J.A.M., Coyne, J.A. & Lewontin, R.C. (1979). The sensitivity of gel electrophoresis as a detector of genetic variation. *Genetics*, **93**, 1019–37.

Reeder, R.H., Brown, D.D., Wellauer, P.K. & Dawid, I.B. (1976). Patterns of ribosomal DNA spacer lengths are inherited. *J. Mol. Biol.* **105**, 507–16.
Rensch, B. (1960). *Evolution above the Species Level.* Columbia University Press.
Richards, R.I., Shine, J., Ullrich, A., Wells, J.R.I. & Goodman, H.M. (1979). Molecular cloning and sequence analysis of adult chicken β globin cDNA. *Nucl. Acids Res.* **7**, 1137–46.
Richardson, J.S. (1981). The anatomy and taxonomy of protein structure. *Advances in Protein Chemistry*, **34**, 168–339.
Richardson, R.H., Richardson, M.E. & Smouse, P.E. (1975). Evolution of electrophoretic mobility in the *Drosophila mulleri* complex. In *Isozymes*, vol. 4, *Genetics and Evolution*, ed. C.L. Markdert, pp. 533–45. New York: Academic Press.
Robertson, A. (1956). The effect of selection against extreme deviants based on deviation or on homozygosis. *J. Genet.* **54**, 236–48.
Robertson, A. (1961). Inbreeding in artificial selection programmes. *Genet. Res., Camb.* **2**, 189–94.
Robertson, A. (1962). Selection for heterozygotes in small populations. *Genetics*, **47**, 1291–300.
Robertson, A. (1967). The nature of quantitative genetic variation. In *Heritage from Mendel* ed. R.A. Brink, pp. 265–80. University of Wisconsin Press.
Robertson, A. (1968). The spectrum of genetic variation. In *Population Biology and Evolution*, ed. R.C. Lewontin, pp. 5–16. Syracuse University Press.
Robertson, A. (1970). The reduction in fitness from genetic drift at heterotic loci in small populations. *Genet. Res., Camb.* **15**, 257–9.
Robertson, A. (1975). Remarks on the Lewontin–Krakauer test. *Genetics*, **80**, 396.
Romer, A.S. (1967). Major steps in vertebrate evolution. *Science*, **158**, 1629–37.
Romer, A.S. (1968). *The Procession of Life.* London: Weidenfeld & Nicholson.
Romero-Herrera, A.E., Lieska, N. & Nasser, S. (1979). Characterization of the myoglobin the lamprey *Petromyzon marinus*. *J. Mol. Evol.* **14**, 259–66.
Roughgarden, J. (1979). *Theory of Population Genetics and Evolutionary Ecology: An Introduction.* New York: Macmillan.
Ruffié, J. (1976). *De la Biologie a la Culture.* Paris: Flammarion.
Ruffié, J. (1982). *Traité du Vivant.* Paris: Fayard.
Sakaizumi, M. (1982). Biochemical genetics of *Oryzias* populations. Doctoral Dissertation, Faculty of Science, University of Tokyo.
Sakaizumi, M., Egami, N. & Moriwaki, K. (1980). Allozymic variation in wild populations of the fish, *Oryzias latipes*. *Proc. Japan Acad. Ser. B*, **56**, 448–51.
Sakano, H., Rogers, J.H., Hüppi, K., Brack, C., Traunecker, A., Maki, R., Wall, R. & Tonegawa, S. (1979). Domains and the hinge regions of an immunoglobulin heavy chain are encoded in separate DNA segments. *Nature, Lond.* **277**, 627–33.
Salser, W. (1978). Globin mRNA sequences: analysis of base pairing and evolutionary implications. *Cold Spring Harbor Symp. Quant. Biology*, **42**, ("Chromatin", part 2), 985–1002.
Sarich, V.M. & Wilson, A.C. (1967). Immunological time scale for hominid evolution. *Science*, **158**, 1200–2.
Schaffner, W., Kunz, G., Daetwyler, H., Telford, J., Smith, H.O. & Birnstiel, M.L. (1978). Genes and spacers of cloned sea urchin histone DNA analyzed by sequencing. *Cell*, **14**, 655–71.
Schopf, T.J.M. (1977). Patterns of evolution: a summary and discussion. In *Patterns of Evolution*, ed. A. Hallam, pp. 547–561. Amsterdam: Elsevier.
Seeburg, P.H., Shine, J., Martial, J.A., Baxter, J.D. & Goodman, H.M. (1977). Nucleotide sequence and amplification in bacteria of structural gene for rat growth hormone. *Nature, Lond.* **270**, 486–94.

Selander, R.K. (1976). Genic variation in natural populations. In *Molecular Evolution*, ed. F.J. Ayala, pp. 21–45. Sunderland: Sinauer.
Selander, R.K. & Levin, B.R. (1980). Genetic diversity and structure in *Escherichia coli* populations. *Science*, **210**, 545–7.
Shaw, C.S. (1970). How many genes evolve? *Biochemical Genetics*, **4**, 275–83.
Shibata, S., Iuchi, I. & Miyaji, T. (1966). Abnormal hemoglobins discovered in Japan. *Acta Haematologica Japonica*, **29**, 115–27.
Simmons, M.J. & Crow, J.F. (1977). Mutations affecting fitness in Drosophila populations. *Annu. Rev. Genet.* **11**, 49–78.
Simons, E.L. (1981). Man's immediate forerunners. *Philosophical Transactions of the Royal Society* (London) Vol. B 292, p. 21–41.
Simpson, G.G. (1944). *Tempo and Mode in Evolution.* Columbia University Press.
Simpson, G.G. (1949). *The Meaning of Evolution.* Yale University Press.
Simpson, G.G. (1953). *The Major Features of Evolution.* Columbia University Press.
Simpson, G.G. (1964). *This View of Life.* New York: Harcourt, Brace & World.
Singh, R.S., Lewontin, R.C. & Felton, A.A. (1976). Genic heterogeneity within electrophoretic "alleles" of xanthine dehydrogenase in *Drosophila pseudoobscura*. *Genetics*, **84**, 609–29.
Slatkin, M. (1972). On treating the chromosome as the unit of selection. *Genetics*, **72**, 157–68.
Slatkin, M. (1977). Gene flow and genetic drift in a species subject to frequent local extinctions. *Theor. Pop. Biol.* **12**, 253–62.
Smith, G.P. (1974). Unequal crossover and the evolution of multigene families. *Cold Spring Harbor Symp. Quant. Biol.* **38**, 507–13.
Sneath, P.H.A. (1966). Relations between chemical structure and biological activity in peptides. *J. Theor. Biol.* **12**, 157–95.
Sneath, P.H.A. (1980). The estimation of differences in protein evolution rates. *Proc. Geol. Assoc.* **91** (1 & 2), 71–9.
Somero, G.N. & Soulé, M. (1974). Genetic variation in marine fishes as a test of the niche-variation hypothesis. *Nature, Lond.* **249**, 670–2.
Soulé M. (1976). Allozyme variation: its determinants in space and time. In *Molecular Evolution*, ed. F.J. Ayala, pp. 60–77. Sunderland: Sinauer.
Spiess, E.B. (1977). *Genes in Populations.* New York: John Wiley & Sons.
Spiess, E.B. & Allen, A.C. (1961). Release of genetic variability through recombination. VII. Second and third chromosomes of *Drosophila melanogaster*. *Genetics*, **46**, 1531–53.
Spieth, P.T. (1975). Population genetics of allozyme variation in *Neurospora intermedia*. *Genetics*, **80**, 785–805.
Spofford, J.B. (1969). Heterosis and the evolution of duplications. *Amer. Natur.* **103**, 407–32.
Stamatoyannopoulos, G., Chen, S.-H. & Fukui, M. (1975). Liver alcohol dehydrogenase in Japanese: high population frequency of atypical form and its possible role in alcohol sensitivity. *Amer. J. Hum. Genet.* **27**, 789–96.
Stebbins, G.L. & Lewontin, R.C. (1972). Comparative evolution at the levels of molecules, organisms and populations. In *Proc. of the Sixth Berkeley Symp. on Math. Stat. and Prob. Vol. V: Darwinian, Neo-Darwinian, and Non-Darwinian Evolution*, ed. L.M. LeCam, J. Neyman & E.L. Scott, pp. 23–42. University of California Press.
Steele, E.J. (1979). *Somatic Selection and Adaptive Evolution.* Toronto: Williams-Wallace.
Stenzel, P. (1974). Opossum Hb chain sequence and neutral mutation theory. *Nature, Lond.* **252**, 62–3.
Stewart, F.M. (1976). Variability in the amount of heterozygosity maintained by neutral mutations. *Theor. Pop. Biol.* **9**, 188–201.
Sturtevant, A.H. (1965). *A History of Genetics.* New York: Harper & Row.

Subak-Sharpe, J.H., Bürk, R.R., Crawford, L.V., Morrison, J.M., Hay, J. & Keir, H.M. (1966). An approach to evolutionary relationships of mammalian DNA viruses through analysis of the pattern of nearest neighbor base sequences. *Cold Spring Harbor Symp. Quant. Biol.* **31**, 737–48.

Sures, I., Goeddel, D.V., Gray, A. & Ullrich, A. (1980). Nucleotide sequence of human preproinsulin complementary DNA. *Science*, **208**, 57–9.

Sures, I., Lowry, J. & Kedes, L.H. (1978). The DNA sequence of sea urchin (*S. purpuratus*) H2A, H2B and H3 histone coding and spacer regions. *Cell*, **15**, 1033–44.

Sved, J.A. (1968). Possible rates of gene substitution in evolution. *Amer. Natur.* **102**, 283–93.

Sved, J.A. (1976). The relationship between genotype and fitness for heterotic models. In *Population Genetics and Ecology*, ed. S. Karlin & E. Nevo, pp. 441–63. New York: Academic Press.

Sved, J.A., Reed, T.E. & Bodmer, W.F. (1967). The number of balanced polymorphisms that can be maintained in a natural population. *Genetics*, **55**, 469–81.

Takahata, N. (1980). Composite stepwise mutation model under the neutral mutation hypothesis. *J. Mol. Biol.* **15**, 13–20.

Takahata, N. & Kimura, M. (1979). Genetic variability maintained in a finite population under mutation and autocorrelated random fluctuation of selection intensity. *Proc. Natl. Acad. Sci. USA*, **76**, 5813–17.

Takahata, N. & Kimura, M. (1981). A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. *Genetics*, **98**, 641–57.

Takahata, N. & Maruyama, T. (1979). Polymorphism and loss of duplicate gene expression: a theoretical study with application to tetraploid fish. *Proc. Natl. Acad. Sci. USA*, **76**, 4521–5.

Tartof, K. (1974). Unequal mitotic sister chromatid exchange and disproportionate replication as mechanisms regulating ribosomal RNA gene redundancy. *Proc. Cold Spring Harbor Symp. Quant. Biol.* **37**, 491–500.

Tateno, Y. & Nei, M. (1978). Goodman et al.'s method for augmenting the number of nucleotide substitutions. *J. Mol. Evol.* **11**, 67–73.

Temin, R.G., Meyer, H.U., Dawson, P.S. & Crow, J.F. (1969). The influence of epistasis on homozygous viability depression in *Drosophila melanogaster*. *Genetics*, **61**, 497–519.

Thoday, J.M. (1975). Non-Darwinian "evolution" and biological progress. *Nature, Lond.* **255**, 675–7.

Thoday, J.M. (1976). Wrong assumption. *Trends in Biochemical Sciences*, **1**, N247.

Thompson, J.N. & Kaiser, T.N. (1977). Selection acting upon slow-migrating ADH alleles differing in enzyme activity. *Heredity*, **38**, 191–5.

Tomita, S. & Riggs, A. (1971). Studies of the interaction of 2, 3-diphosphoglycerate and carbon dioxide with hemoglobins from mouse, man and elephant. *J. Biol. Chem.* **246**, 547–54.

Tsujimoto, Y., Hirose, S., Tsuda, M. & Suzuki, Y. (1981). Promoter sequence of fibroin gene assigned by *in vitro* transcription system. *Proc. Natl. Acad. Sci. USA*, **78**, 4838–42.

Valentine, J.W. (1976). Genetic strategies of adaptation. In *Molecular Evolution*, ed. F.J. Ayala, pp. 78–94. Sunderland: Sinauer.

Valenzuela, P., Quiroga, M., Zaldivar, J., Rutter, W.J., Kirschner, M.W. & Cleveland, D.W. (1981). Nucleotide and corresponding amino acid sequences encoded by α and β tubulin mRNAs. *Nature, Lond.* **289**, 650–5.

van Delden, W., Boerema, A.C. & Kamping, A. (1978). The alcohol dehydrogenase polymorphism in populations of *Drosophila melanogaster*. I. Selection in different environments. *Genetics*, **90**, 161–91.

van den Berg, J., van Ooyen, A., Mantei, N., Schamböck, A., Grosveld, G., Flavell, R.A. & Weissmann, Ch. (1978). Comparison of closed rabbit and mouse β-globin genes showing strong evolutionary divergence of two homologous pairs of introns. *Nature, Lond.* **276**, 37–44.

Vanin, E.F., Goldberg, G.I., Tucker, P.W. & Smithies, O. (1980). A mouse α-globin-related pseudogene lacking intervening sequences. *Nature, Lond.* **286**, 222–6.

van Ooyen, A., van den Berg, J., Mantei, N. & Weissmann, C. (1979). Comparison of total sequence of a cloned rabbit β-globin gene and its flanking regions with a homologous mouse sequence. *Science*, **206**, 337–44.

van Roode, J.H.G. & Orgel, L.E. (1980). Template-directed synthesis of oligoguanylates in the presence of metal ions. *J. Mol. Biol.* **144**, 579–585.

Vogel, F. (1969). Point mutations and human hemoglobin variants. *Humangenetik*, **8**, 1–26.

Vogel, F. & Motulsky, A.G. (1979). *Human Genetics*. Berlin: Springer-Verlag.

Waddington, C.H. (1957). *The Strategy of the Genes*. London: George Allen & Unwin.

Wallace, B. (1958). The average effect of radiation-induced mutations on viability in *Drosophila melanogaster*. *Evolution*, **12**, 532–52.

Wallace, B. (1968). *Topics in Population Genetics*. New York: Norton.

Ward, R.D. (1977). Relationship between enzyme heterozygosity and quaternary structure. *Biochem. Genet.* **15**, 123–35.

Ward, R.D. & Beadmore, J.A. (1977). Protein variation in the plaice, *Pleuronectes platessa* L. *Genet. Res., Camb.* **30**, 45–62.

Washburn, S. (1978). The evolution of man. *Scientific American*, **239** (3), 146–54.

Wasylyk, B., Derbyshire, R., Guy, A., Molko, D., Roget, A., Téoule, R. & Chambon, P. (1980). Specific *in vitro* transcription of conalbumin gene is drastically decreased by single-point mutation in T-A-T-A box homology sequence. *Proc. Natl. Acad. Sci. USA*, **77**, 7024–8.

Watson, J.D. (1970). *Molecular Biology of the Gene*. New York: Benjamin.

Watterson, G.A. (1962). Some theoretical aspects of diffusion theory in population genetics. *Annu. Math. Stat.* **33**, 939–57.

Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination. *Theor. Pop. Biol.* **7**, 256–76.

Watterson, G.A. (1977a). Heterosis or neutrality? *Genetics*, **85**, 789–814.

Watterson, G.A. (1977b). Reversibility and the age of an allele II. Two-allele models, with selection and mutation. *Theor. Pop. Biol.* **12**, 179–96.

Watterson, G.A. (1978a). An analysis of multi-allelic data. *Genetics*, **88**, 171–9.

Watterson, G.A. (1978b). The homozygosity test of neutrality. *Genetics*, **88**, 405–17.

Weinberg, W. (1908). Über den Nachweis der Vererbung beim Menschen. *Jahresh. Verein. f. Vaterl. Naturk. Württem.* **64**, 368–82.

Weir, B.S., Brown, A.H.D. & Marshall, D.R. (1976). Testing for selective neutrality of electrophoretically detectable protein polymorphisms. *Genetics*, **84**, 639–59.

Weldon, W.F.R. (1901). A first study of natural selection in *Clausilia laminata* (Montagu). *Biometrika*, **1**, 109–24.

White, T.J., Mross, G.A., Osserman, E.F. & Wilson, A.C. (1977). Primary structure of rat lysozyme. *Biochemistry*, **16**, 1430–6.

Wills, C. (1966). The mutational load in two natural populations of *Drosophila pseudoobscura*. *Genetics*, **53**, 281–94.

Wills, C. (1978). Rank-order selection is capable of maintaining all genetic polymorphisms. *Genetics*, **89**, 403–17.

Wills, C. (1981). *Genetic Variability*. Oxford University Press.

Wilson, E.O. (1975). *Sociobiology*. The Belknap Press of Harvard University Press.

Wilson, A.C., Carlson, S.S. & White, T.J. (1977). Biochemical evolution. *Annu. Rev. Biochem.* **46**, 573–639.

Wright, S. (1921). Systems of mating. *Genetics*, **6**, 111–78.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, **16**, 97–159.
Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc. VI. Intern. Congr. Genet.* **1**, 356–66.
Wright, S. (1934). Physiological and evolutionary theories of dominance. *Amer. Natur.* **68**, 25–53.
Wright, S. (1935). The analysis of variance and the correlations between relatives with respect to deviations from an optimum. *J. Genet.* **30**, 243–56.
Wright, S. (1937). The distribution of gene frequencies in populations. *Proc. Natl. Acad. Sci. USA*, **23**, 307–20.
Wright, S. (1938a). The distribution of gene frequencies under irreversible mutation. *Proc. Natl. Acad. Sci. USA*, **24**, 253–9.
Wright, S. (1938b). Size of population and breeding structure in relation to evolution. *Science*, **87**, 430–1.
Wright, S. (1942). Statistical genetics and evolution. *Bull. Amer. Math. Soc.* **48**, 223–46.
Wright, S. (1943). Isolation by distance. *Genetics*, **28**, 114–38.
Wright, S. (1945). The differential equation of the distribution of gene frequencies. *Proc. Natl. Acad. Sci. USA*, **31**, 382–9.
Wright, S. (1946). Isolation by distance under diverse systems of mating. *Genetics*, **31**, 39–59.
Wright, S. (1949). Adaptation and selection. In *Genetics, Paleontology, and Evolution*, ed. G.L. Jepsen *et al.*, pp. 365–89. Princeton University Press.
Wright, S. (1951a). Fisher and Ford on "the Sewall Wright effect." *American Scientist*, **39**, 452–8.
Wright, S. (1951b). The genetical structure of populations. *Ann. Eugen.*, **15**, 323–54.
Wright, S. (1966). Polyallelic random drift in relation to evolution. *Proc. Natl. Acad. Sci. USA*, **55**, 1074–81.
Wright, S. (1969). *Evolution and the Genetics of Populations*, vol. 2, *The Theory of Gene Frequencies*. University of Chicago.
Wright, S. (1970). Random drift and the shifting balance theory of evolution. In *Mathematical Topics in Population Genetics*, ed. K. Kojima, pp. 1–31, Berlin: Springer-Verlag.
Wright, S. (1977). *Evolution and the Genetics of Populations*, vol. 3, *Experimental Results and Evolutionary Deductions*. University of Chicago Press.
Wright, S. (1978). *Evolution and the Genetics of Populations*, vol. 4, *Variability Within and Among Natural Populations*. University of Chicago Press.
Wright, S. & Dobzhansky, Th. (1946). Genetics of natural populations. XII. Experimental reproduction of some of the changes caused by natural selection in certain populations of *Drosophila pseudoobscura*. *Genetics*, **31**, 125–50.
Yamazaki, T. (1971). Measurement of fitness at the esterase-5 locus in *Drosophila pseudoobscura*. *Genetics*, **67**, 579–603.
Yamazaki, T. (1977). Enzyme polymorphism and functional difference: mean, variance, and distribution of heterozygosity. In *Molecular Evolution and Polymorphism*, ed. M. Kimura, pp. 127–47. Mishima: National Institute of Genetics.
Yamazaki, T. (1981). Genetics variabilities in natural population of haploid plant, *Conocephalum conicum*. I. The amount of heterozygosity. *Jap. J. Genet.* **56**, 373–83.
Yamazaki, T. & Maruyama, T. (1972). Evidence for the neutral hypothesis of protein polymorphism. *Science*, **178**, 56–8.
Yamazaki, T. & Maruyama, T. (1973). Evidence that enzyme polymorphisms are selectively neutral. *Nature New Biol.* **245**, 140–1.
Yamazaki, T. Maruyama, T. (1974). Evidence that enzyme polymorphisms are selectively neutral, but blood group polymorphisms are not. *Science*, **183**, 1091–2.
Yamazaki, T. & Maruyama, T. (1975). Isozyme polymorphism maintenance mechanisms viewed from the standpoint of population genetics. In *Isozymes IV. Genetics and Evolution*, ed. C.L. Markert, pp. 103–14. New York: Academic Press.
Yanase, T. (1971). Basis of medical genetics – Molecular disease Part 6. *Clinic All-Round* (*Sōgō Rinshō*), **20** (12), 2964–76 (in Japanese).
Yanofsky, C. (1967). Gene structure and protein structure. *Scientific American*, **216** (5), 80–2.
Yanofsky, C., Horn, V. & Thorpe, D. (1964). Protein structure relationships revealed by mutational analysis. *Science*, **146**, 1593–4.
Yarbrough, K. & Kojima, K. (1967). The mode of selection at the polymorphic esterase 6 locus in cage populations of *Drosophila melanogaster*. *Genetics*, **57**, 677–86.
Yoshimaru, H. & Mukai, T. (1979). Lack of experimental evidence for frequency-dependent selection at the alcohol dehydrogenase locus in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA*, **76**, 876–8.
Zimmer, E.A., Martin, S.L. Beverley, S.M., Kan, Y.W. & Wilson, A.C. (1980). Rapid duplication and loss of genes coding for the α chains of hemoglobin. *Proc. Natl. Acad. Sci. USA*, **77**, 2158–62.
Zouros, E. (1976). Hybrid molecules and the superiority of the heterozygote. *Nature, Lond.* **262**, 227–9.
Zuckerkandl, E. (1976). Evolutionary process and evolutionary noise at the molecular level II. A selectionist model for random fixations in proteins. *J. Mol. Evol.* **7**, 269–311.
Zuckerkandl, E. (1978). Multilocus enzymes, gene regulation, and genetic sufficiency. *J. Mol. Evol.* **12**, 57–89.
Zuckerkandl, E. & Pauling, L. (1965). Evolutionary divergence and convergence in proteins. In *Evolving Genes and Proteins*, ed. V. Bryson & H.J. Vogel, pp. 97–166. New York: Academic Press.