CHAPTER 24

Evolution of Protein Molecules

THOMAS H. JUKES[1]

Space Sciences Laboratory,
University of California,
Berkeley, California

AND

CHARLES R. CANTOR[2]

Columbia University,
Department of Chemistry,
New York, New York

[1] Supported by grant NsG 479 by the National Aeronautics and Space Administration to the University of California.
[2] Supported by grant GM 14825 from the National Institutes of Health.

I. Introduction

Few names have been so appropriately bestowed as the word *protein,* which emphasizes the central importance and primordial origin of this group of compounds. The derivation of the word, and the history of its introduction, were discussed by Munro (Vol. I, Chapter 1, Section III).

The deoxyribonucleic acid (DNA) of an organism contains hereditary information, which is termed the *genotype* of the organism. This is translated into an assembly of protein molecules that are responsible for the visible and functional characteristics of the organism. These characteristics, resulting from the genotype, are termed the *phenotype* and the DNA carrying genotypic information is termed the *genome.* The phenotypic characteristics of living organisms are largely the result of the properties of proteins. Proteins participate in the evolutionary process and have a role in directing it. Natural selection depends on the properties of the phenotype and on its interaction with the environment. Proteins have enzymic and structural functions, and are subject to genetic change as a direct result of changes in DNA. Fortunately, the chemistry of protein molecules is such that evolutionary changes in them can often be clearly perceived; indeed, a major part of the current understanding of molecular evolution stems from studies of homology in the primary structure of proteins.

The study of the evolution of living organisms involves population genetics. Changes in the members of a population, especially in the case of multicellular organisms, take place because chromosomes, carrying DNA molecules, are reapportioned by the diploid-haploid-diploid alternation that occurs during reproduction, so that phenotypic changes between successive generations consequently take place. Environmental forces have a selective effect on the perpetuation of these changes. Mutations are fed into the gene pool and find their way into evolutionary changes. The study of protein evolution provides an important new means for examining the relation between mutations, the occurrence of which is shown by amino acid differences between homologous proteins, and evolutionary differences between species.

As commonly understood (e.g., Neurath *et al.,* 1968), homology in proteins refers to significant similarity between the amino acid sequences of two or more proteins. The similarity, to be useful for the purposes of this review, should be sufficient to indicate a common evolutionary origin. An example of insufficient similarity from this standpoint would be the series of nine enzymes with reactive serines in the active center (Epstein and Motulsky, 1965). These centers have the general formula

for a tetrapeptide Gly-Asp(or Glu)-Ser-Gly(or Ala). Three of the proteins in this list (trypsin, chymotrypsin, and elastase) have a common evolutionary origin as evidenced by extensive homology over large regions of their amino acid sequences, but any tetrapeptide is so short that, in the absence of reasonably complete information concerning the remainder of the molecules of the other six enzymes (thrombin, *Escherichia coli* alkaline phosphatase, a bacterial proteinase, phosphoglucomutase, pseudocholinesterase, and liver aliesterase) it must be concluded (cf. Dixon, 1966) that the similarities based solely on the tetrapeptide may well be coincidental.

The discovery of the amino acid sequences of various insulins, by Sanger and his collaborators in the early 1950's (Sanger, 1952) provided the first evidence for homology in the primary structure of similar proteins originating in different organisms. The insulins of cattle, pigs, sheep, horses, and sperm whales were shown to be identical except for substitutions in three consecutive amino acid residues in the A chain. It was well known that many different organisms contained identical molecules, such as adenosine triphosphate, glutathione, etc., but the finding with insulin showed that the same protein hormone could exhibit slight differences as well as extensive similarities from species to species. In view of the gene-protein relationship, this drew attention to differentiation of molecules at the genetic and evolutionary level.

Tuppy (1958) compared polypeptide sequences in several cytochromes *c* as follows:

		Differences from (a)
(a)	Val-Gln-Lys-Cys-Ala-Gln-Cys-His-Thr-Val-Glu	—
(b)	Val-Gln-Lys-Cys-*Ser*-Gln-Cys-His-Thr-Val-Glu	1
(c)	Val-Gln-*Arg*-Cys-Ala-Gln-Cys-His-Thr-Val-Glu	1
(d)	*Lys-Thr-Arg*-Cys-*Glu-Leu*-Cys-His-Thr-Val-Glu	5

(a) cattle, horse, pig, salmon; (b) chicken; (c) silkworm; (d) yeast.

The divergent nature of the relationship, and the similarity between the excerpts from several cytochromes *c*, powerfully conveyed the suggestion of a common evolutionary origin for organisms as diverse as animals and yeast.

A now well-known book "The Molecular Basis of Evolution," by Anfinsen, appeared in 1959, and brought together the evidence from polypeptide sequences. The information then extant was scanty, but Anfinsen was able to cite the findings mentioned above, together with comparisons of a few polypeptide hormones, as clear examples of homology in the primary structures of polypeptides and proteins.

Further information rapidly appeared during the next two or three years, especially in the field of hemoglobins and cytochromes *c*. In 1961,

Ingram reviewed the evidence for an evolutionary relationship between myoglobin and the four polypeptide chains of three human hemoglobins: A, A_2, and F. He traced these five polypeptide chains back to a common ancestor by postulating four events in which gene duplications led ultimately to five independent genes. Ingram's evolutionary scheme for the hemoglobins has been widely quoted and accepted. His proposal paralleled the investigations and conclusions of Braunitzer and co-workers (1961b), who, in making comparisons of the amino acid sequences of polypeptide chains, noticed that in order to align the sequences for maximum homology, it was necessary to postulate the presence of "gaps" occurring occasionally in one of the two chains in the comparison. This observation was compatible with the well-known genetic phenomenon of unequal crossing-over and recombination in genes. This phenomenon is illustrated by two recently discovered human hemoglobin variants, in each of which deletion of a portion of the β-polypeptide chain evidently occurred during a mutational event (Jones *et al.*, 1966b; Bradley *et al.*, 1967).

Braunitzer *et al.* (1961b) aligned the sequences of the α and β human hemoglobin chains and showed that at least 24 residues in 44 comparisons were identical in both chains. They commented that a duplication of the genetic material must have occurred during evolution, followed by independent differentiation, and that these changes must have occurred a long time ago because of the extensive differences.

Proteins are compared sometimes by means of the total content of their respective amino acids. For obvious reasons, this is only a rough measure of homology: two proteins of identical, or nearly identical, amino acid content could have very different sequences. A second means of indirect comparison is by means of immunological cross-reactions, introduced by Nuttall (1904), and used by him successfully to make comparisons of the serum proteins of various animals so that a rough parallel was obtained between taxonomy and the precipitin reaction. Refinements of this approach have been made by using modifications of the Ouchterlony procedure for producing the precipitin reaction in gels and by the use of microcomplement fixation (Sarich and Wilson, 1967). This review, however, will be devoted principally to comparisons based on homology demonstrable between known polypeptide sequences, because such comparisons have a direct relationship to changes in the base sequences of DNA that take place during evolution.

The majority of evolutionary changes in proteins undoubtedly correspond to mutational changes that have become incorporated in the genome of a species; for example, 10 of the mutations in human hemoglobins consist of single amino acid substitutions that are identical with

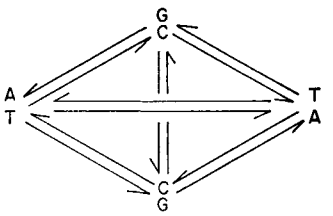
the residue occupying the same site in another globin chain. It has long been evident that mutational changes far outnumber evolutionary changes, and this will be discussed below. The causes that lead a mutation to becoming fixed as an evolutionary change are being explored. An obvious and common explanation is that beneficially adaptive mutations tend to spread through a species as a result of natural selection. It is unlikely that this accounts for all of the many minor differences between homologous proteins (King and Jukes, 1969).

Recent advances in experimental science have provided a molecular basis for the study of genetics. It is now recognized that DNA is the repository of hereditary information in all species, except for a small number of viruses which use ribonucleic acid (RNA) for this purpose. The means by which the aggregate of hereditary information in an organism is translated into phenotypic characteristics are now being revealed by the techniques of biochemistry.

Evolution depends upon the occurrence of occasional changes, large or small, in hereditary characteristics. Molecular genetics gave rise to the new field of molecular evolution, which is currently exploring the changes that take place in proteins and nucleic acids over long periods of time. It is possible to measure these changes by comparing the structures of molecules in various living organisms.

The molecular approach to biological evolution is based principally on the concept that changes in DNA are incorporated into the genome of a species. These changes may be divided into two broad groups:

a. Replacements of Base Pairs by Each Other. Replacements of one DNA base pair by another, which for convenience may be termed "single-base changes," often affect protein molecules by producing structural changes. Such changes are known as *point mutations.* The possible replacements of base pairs by each other are as follows:



It is assumed that base replacements are randomly distributed along the length of all DNA molecules and that they occur quantitatively at an approximately constant rate. It is not certain whether some replacements predominate over others, although there are indications, first emphasized by Sueoka (1961), that some microorganisms have tended

toward DNA of a high GC content and others to DNA of a high AT content. Animals contain DNA of a relatively constant GC content, in the neighborhood of 38%. Fitch (1967) has suggested that G → A changes (equivalent in terms of double-stranded DNA to

$$\begin{matrix} G \\ C \end{matrix} \rightarrow \begin{matrix} A \\ T \end{matrix} \quad \text{and} \quad \begin{matrix} C \\ G \end{matrix} \rightarrow \begin{matrix} T \\ A \end{matrix}$$

changes) predominate in the genes for hemoglobin and cytochrome *c*. Yanofsky *et al.* (1966a) have described a mutator gene which produces a tendency toward

$$\begin{matrix} A \\ T \end{matrix} \rightarrow \begin{matrix} C \\ G \end{matrix} \quad \text{and} \quad \begin{matrix} T \\ A \end{matrix} \rightarrow \begin{matrix} G \\ C \end{matrix}$$

changes, and Speyer (1965) found a mutation in the DNA polymerase of T_4 bacteriophage which led to changes that were reversible by base analogs and hence, were probably transitions, i.e.,

$$\begin{matrix} A \\ T \end{matrix} \rightleftharpoons \begin{matrix} G \\ C \end{matrix} \quad \text{and} \quad \begin{matrix} T \\ A \end{matrix} \rightleftharpoons \begin{matrix} C \\ G \end{matrix}$$

interchanges. Much of the evidence available from comparisons of related proteins has been interpreted (Jukes, 1966) to indicate that purine ⇋ pyrimidine interchanges (transversions) are more frequent than transitions, despite the fact that transitions would be expected to occur with less steric change than is produced in the DNA molecule by transversions during replication.

b. Changes in the Total Amount of DNA in the Genome. The amount of DNA may be increased by *duplication* of portions of the genome. In contrast to the steady and slow evolutionary progress of single-base changes, the duplication may occur suddenly, repeatedly, and at widely separated intervals of time. The evidence for this has been presented by Britten and Kohne (1965–1966).

The amount of DNA may also be increased or decreased by crossing-over and recombination of sections of DNA of various lengths. These events occur typically during meiosis. Such changes may be long enough to duplicate several genes or *cistrons;* the term *cistron* refers to a piece of DNA that contains the information for an entire polypeptide sequence, starting with an N-terminal amino acid and ending with a C-terminal amino acid. Alternatively, the changes may be short, so that they lengthen or shorten a single cistron by one or more base pairs.

In addition to these two types of changes, there are control mechanisms which govern the transcription of portions of the genome, so that large portions of it may be "shut off," or "not read," during all or part of the life of an individual.

The evolution of protein molecules is secondary to changes in DNA, since proteins are formed by transcription of DNA into RNA, followed by the translation of RNA into polypeptide sequences. Following this, proteins are either conserved, or eliminated from the biological scene, depending upon their net usefulness in terms of reproduction, competition, and consequent survival of the species which produces them.

The model described in the preceding outline is consistent with *divergent,* and not with parallel, or convergent, evolution. *Tempora mutantur, nos et mutamur in illis.* Together with the times, proteins change, and we change with both time and proteins. This chapter will explore the types of evolutionary changes in proteins and the mechanisms by which these changes take place.

It is easy by simple arithmetical calculations to show the enormous number of ways in which twenty different amino acids may be arranged to make polypeptide chains. For example, White, Handler, and Smith (1964) cite a calculation by Synge that a hypothetical protein containing only 12 different amino acids and 288 residues has enough possible isomers so that if one molecule of each were put together, the total mass of the aggregate would be more than 10 billion times the mass of the Earth. Proteins, however, are not formed by chance association of amino acids any more than a book is the fortunate (or unfortunate) outcome of a random series of permutations of the alphabet. The proteins represent a continuum; a series whose members have progressively increased in number, size, and complexity. Chance may have initially brought together short sequences of amino acids that joined to form a catalytically active peptide molecule, but as soon as a self-replicating system was established, the peptide was able to lengthen by end-to-end duplications and to multiply by duplicating in entirety, thus giving rise to families of related enzymes whose continuity extends backward for more than a billion years. We shall discuss the evidence for this process and the mechanisms by which it takes place.

A. Methods and Procedures in the Study of Protein Evolution

The evolution of proteins is measured by comparing the primary structures of two or more different proteins, or by comparing portions of the primary structure of a single protein with each other to search for evidence of internal repetition.

The primary structure is defined as the ordered sequence of amino acid residues in a protein or polypeptide. The most common procedure for determining the primary structure is to hydrolyze the protein with enzymes, separate the hydrolyzate into its constituent peptides, and determine the amino acid sequence of each peptide. At least two enzymes

of different specificities must be used in order to produce two different sets of peptides. The sequence of the entire molecule is then deduced by overlaps. The primary structures of two proteins are compared by aligning the two sequences with each other as in the following abbreviated example:

Gly-Phe-Ser-Ala-Gly-Asp-Ser-Lys-Lys-Gly
Glu-Phe-Lys-Ala-Gly- Ser-Ala-Lys-Lys-Gly

The sequences show 60% of similarity and 40% of difference with respect to their amino acid residues. A more precise comparison of the genetic difference between the two proteins is made by expressing it in terms of base differences in the genetic code. There are good reasons for this. First, the genetic difference is actually a difference between two molecules of DNA. Second, some pairs of amino acids are more closely related genetically than are other pairs; for example, a single-base change can convert a codon for glycine, GGA or GGG, into a codon for glutamic acid, GAA or GAG, while the codons for serine and lysine differ by a minimum of two base replacements. Third, single amino acid replacements in proteins, when caused by point mutations, correspond to single-base changes in codons. There is only one exception known in more than a hundred known examples of such mutations. It is therefore concluded that point mutations are the expression of single-base changes. The further conclusion may be drawn that the difference between two homologous polypeptide chains (as shown in the preceding example, which is taken from the cytochromes *c*) has taken place during evolution by a series of single-base changes.

The calculation of base differences must necessarily ignore the question of "silent" differences in the third bases of two codons that are being compared. An example of "silent" difference would be if glycine residues at corresponding positions in two different proteins were coded by GGA and GGC, respectively. The difference would have no genetic effect, however, and hence is not of concern in comparisons of polypeptide sequences. It is therefore usual to express the differences between the amino acid residues in two homologous polypeptide sequences in terms of minimum base differences per codon (MBDC). These differences are expressed under the sequence as follows:

	Gly	Phe	Ser	Ala	Gly	Asp	Ser	Lys	Lys	Gly
	Glu	Phe	Lys	Ala	Gly	Ser	Ala	Lys	Lys	Gly
Minimum base differences per codon	1	0	2	0	0	2	1	0	0	0
Maximum base differences per codon	2	1	3	1	1	3	3	1	1	1

The MBDC for the two sequences is 0.60; a difference of 60% as con-

trasted with only 40% in the comparison of the amino acid residues. Note that the maximum difference could be much higher, 1.70.

The MBDC will often be low if the two sequences have a common evolutionary origin. If not, the MBDC will always be high, provided that about ten or more residues are compared. In two completely randomized sequences of all 20 amino acids, the MBDC will be in the neighborhood of 1.66.

An example of a comparison involving all 20 amino acids is shown below. In the first row the 20 are written in alphabetical order; in the second row the same sequence is shifted two places to the left. The minimum base differences per codon are shown below:

Ala-Arg-Asn-Asp-Cys-Gln-Glu-Gly-His-Ile-Leu-Lys-Met-Phe-Pro-Ser-Thr-Trp-Tyr-Val
Asn-Asp-Cys-Gln-Glu-Gly-His-Ile-Leu-Lys-Met-Phe-Pro-Ser-Thr-Trp-Tyr-Val-Ala-Arg
2 2 2 2 3 2 2 2 1 1 1 3 2 1 1 1 2 2 2 2

In this comparison, the MBDC is 1.80, even though the "polypeptides" are identical in amino acid content!

Gaps, Deletions, and Additions

A protein molecule may lose or gain one or more amino acid residues by genetic crossing-over and recombination during evolution. The crossing-over takes place between two DNA strands, and must occur so that the number of base pairs in the region of overlap is three or a multiple of three. If the protein is compared with a homologous protein obtained from another species, or even from the same species, the loss or gain of the amino acid residue or residues often is absent from the second protein at the corresponding site. The comparison is then "out of register" on the right-hand side of the loss or gain, unless a gap (Braunitzer *et al.*, 1961b) is inserted arbitrarily in one of the sequences, thus:

Gly-Val- -Ser-Ser-Cys-Met-Gly-Asp-Ser-Gly
Gly-Gly-Lys-Asn-Ser-Cys-Gln-Gly-Asp-Ser-Gly

The insertion of such a gap requires specific and statistical justification in each case, and this is discussed in Section V.

The variant termed Gun Hill hemoglobin provides a striking illustration of a mutational event which has produced a gap (Bradley *et al.*, 1967). The sequences of two corresponding regions in the β chains of normal and Gun Hill hemoglobins are as follows:

	90	100
Normal	Leu-Ser-Glu-Leu-His-Cys-Asp-Lys-Leu-His-	-Val-Asp-Pro-Glu-Asn-Phe
Gun Hill	Leu-Ser-Glu-Leu-His-	-Val-Asp-Pro-Glu-Asn-Phe

Five amino acids have been deleted from the β chain, presumably by

crossing-over during meiosis. We can assume that in most cases such a shortening would have lethal consequences, but in this case the hemoglobin molecule was, astonishingly, still functional in oxygen transport because the two abnormal β chains, which did not carry heme groups, became associated with two normal α chains to form a tetrameric molecule represented by the abbreviation $\alpha_2{}^{A}\beta_2{}^{\text{Gun Hill}}$.

Point mutations may also shorten protein molecules during evolution by producing one of the three chain-terminating codons UAA, UAG, or UGA. Such mutations will probably be deleterious or lethal if they occur in the middle region of the molecule. Short sections, however, may occasionally be removed from either end without harmful effects.

B. The Genetic Code

The expression of minimum base differences per codon is based on the genetic code (Table I). Knowledge of the code is based mainly

TABLE I
THE GENETIC CODE

UUU Phenylalanine	CUU Leucine	AUU Isoleucine	GUU Valine
UUC Phenylalanine	CUC Leucine	AUC Isoleucine	GUC Valine
UUA Leucine	CUA Leucine	AUA Isoleucine	GUA Valine
UUG Leucine	CUG Leucine	AUG Methionine	GUG Valine
UCU Serine	CCU Proline	ACU Threonine	GCU Alanine
UCC Serine	CCC Proline	ACC Threonine	GCC Alanine
UCA Serine	CCA Proline	ACA Threonine	GCA Alanine
UCG Serine	CCG Proline	ACG Threonine	GCG Alanine
UAU Tyrosine	CAU Histidine	AAU Asparagine	GAU Aspartic acid
UAC Tyrosine	CAC Histidine	AAC Asparagine	GAC Aspartic acid
UAA chain termn.	CAA Glutamine	AAA Lysine	GAA Glutamic acid
UAG chain termn.	CAG Glutamine	AAG Lysine	CAG Glutamic acid
UGU Cysteine	CGU Arginine	AGU Serine	GGU Glycine
UGC Cysteine	CGC Arginine	AGC Serine	GGC Glycine
UGA chain termn.	CGA Arginine	AGA Arginine	GGA Glycine
UGG Tryptophan	CGG Arginine	AGG Arginine	GGG Glycine

on experiments carried out *in vitro* with cell-free amino acid incorporating systems, usually obtained from *Escherichia coli*. The reactions in such systems differ in many respects from the corresponding processes *in vivo*. Most of the findings *in vitro* depend on the use of synthetic polyribonucleotides, which may differ from natural messenger RNA in their way of attaching to ribosomes and in their initiation of polypeptide synthesis. The single amino acid mutations in proteins, however, provide a source of information about the code that is derived from intact living organisms. More than a hundred examples of such mutations are known, including 48 different amino acid interchanges, and all but one correspond to single-base changes in the code in Table I. The code probably is "universal," i.e., the same in all terrestrial organisms. We shall assume that it has not changed during the period in which divergent evolution took place from a single form of life. This is the period with which we are concerned in discussing the evolution of proteins, although it may be conjectured (Jukes, 1967) that the genetic code evolved from earlier codes that specified fewer amino acids.

Table II contains a list of the minimum base differences per codon for all the amino acids. Assuming that amino acid interchanges are caused by point mutations, the evolutionary interchanges ascribable to single-base changes should be more common than those caused by two-base changes in a codon, according to the laws of chance. This agrees with the information obtained from homologous proteins. For example, in comparisons of the hemoglobins with each other (Table X) there are 343 single-base changes and 222 two-base changes; and in the case of the cytochromes *c* (Table IV) there are 389 single-base changes and 207 two-base changes. Perceptible three-base changes are quite rare in such comparisons, because of the lack of specificity in the third base of most codons, and because only 6% of the amino acid interchanges (Table II) necessitate three-base changes in a codon.

Errors and Variations

The coded information in the DNA is replicated many times and is translated into proteins with astonishing fidelity. It is obvious that if this were not the case, it would be difficult for living organisms to exhibit their usual property of retaining specific characteristics through many generations.

Virtually identical results were found when different preparations of proteins, such as horse hemoglobin and egg white lysozyme, were analyzed for primary structure in different laboratories. The identity of the results is, however, not quite complete. It has been perceived repeatedly that such preparations frequently, perhaps usually, contain minor components whose sequences differ in one or more loci from the predominant form.

TABLE II

MINIMUM BASE DIFFERENCES BETWEEN CODONS

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	0	2	2	1	2	2	1	1	2	2	2	2	2	2	1	1	1	2	2	1
Arg	2	0	2	2	1	1	2	1	1	1	1	1	1	2	1	1	1	1	2	2
Asn	2	2	0	1	2	2	2	2	1	1	2	1	2	2	2	1	1	3	1	2
Asp	1	2	1	0	2	2	1	1	1	2	2	2	3	2	2	2	2	3	1	1
Cys	2	1	2	2	0	3	3	1	2	2	2	3	3	1	2	1	2	1	1	2
Gln	2	1	2	2	3	0	1	2	1	2	1	1	2	3	1	2	2	2	2	2
Glu	1	2	2	1	3	1	0	1	2	2	2	1	2	3	2	2	2	2	2	1
Gly	1	1	2	1	1	2	1	0	2	2	2	2	2	2	2	1	2	1	2	1
His	2	1	1	1	2	1	2	2	0	2	1	2	3	2	1	2	2	3	1	2
Ile	2	1	1	2	2	2	2	2	2	0	1	1	1	1	2	1	1	3	2	1
Leu	2	1	2	2	2	1	2	2	1	1	0	2	1	1	1	1	2	1	2	1
Lys	2	1	1	2	3	1	1	2	2	1	2	0	1	3	2	2	1	2	2	2
Met	2	1	2	3	3	2	2	2	3	1	1	1	0	2	2	2	1	2	3	1
Phe	2	2	2	2	1	3	3	2	2	1	1	3	2	0	2	1	2	2	1	1
Pro	1	1	2	2	2	1	2	2	1	2	1	2	2	2	0	1	1	2	2	2
Ser	1	1	1	2	1	2	2	1	2	1	1	2	2	1	1	0	1	1	1	2
Thr	1	1	1	2	2	2	2	2	2	1	2	1	1	2	1	1	0	2	2	2
Trp	2	1	3	3	1	2	2	1	3	3	1	2	2	2	2	1	2	0	2	2
Tyr	2	2	1	1	1	2	2	2	1	2	2	2	3	1	2	1	2	2	0	2
Val	1	2	2	1	2	2	1	1	2	1	1	2	1	1	2	2	2	2	2	0
UAA	2	2	2	2	2	1	1	2	2	2	1	1	3	2	2	1	2	2	1	2
UAG	2	2	2	2	2	1	1	2	2	3	1	1	2	2	2	1	2	1	1	2
UGA	2	1	3	3	1	2	2	1	3	2	1	2	3	2	2	1	2	1	2	2

The practice of "rounding off" the results of peptide analyses may have obscured the presence of such variants. In some cases, the alternate or minor components are ascribable to multiplicity of genes, or to the presence of alleles in a population of mixed genetic composition. Examples are the light chains of immunoglobulins (Wikler *et al.*, 1967), isocytochromes in yeast (Sherman *et al.*, 1968), a variant ferredoxin in spinach with two amino acid substitutions (Matsubara and Sasaki, 1968), and the β chains of A and B hemoglobins in sheep (Boyer *et al.*, 1966) and rabbits (Galizzi and von Ehrenstein, 1967). A second possible cause, that of translational variation, was discussed by von Ehrenstein (1966), who studied anomalies in the α chain of rabbit hemoglobin and found six positions which contained more than one amino acid. The results are shown below, using the numbering system of Table XI.

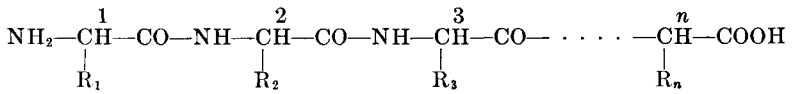
Site No.	Amino acids and ratio
30	0.6 Val:0.4 Leu
50	0.5 Leu:0.5 Phe
51	0.5 Ser:0.5 Thr
77	0.8 Val:0.2 Thr
83	0.8 Leu:0.2 Val
87	0.5 Ser:0.5 Leu

Von Ehrenstein emphasizes the possibility that these variations are caused by a few "ambiguous" codons which are translated by two different tRNA's carrying different amino acids. He does not exclude gene duplication as a cause, however, and he has described a male rabbit with only valine and no leucine in site 30, 0.9 phenylalanine:0.1 leucine in site 50, and 0.9 threonine and 0.1 serine in site 51. The rabbit was mated to a female with 0.5:0.5 ratios of the two amino acids in each of these three sites. Four of the ten offspring inherited the maternal pattern. This suggests the existence of a multiplicity of cistrons for the α chain in rabbits and the differentiation of the cistrons from each other by the occurrence of a small number of point mutations. Genetic variations, rather than ambiguous codons, may therefore be a cause of the existence of "minor components" in proteins.

C. The Biological Synthesis of Proteins

This is only a short outline of the biological synthesis of proteins, a subject complex and important enough for a complete textbook.

Protein molecules consist of one or more polypeptide chains which can be depicted as follows:



The amino acid residues are numbered, as indicated, from left to right. R, R_2, etc., are side chains belonging to any of the 20 amino acids that participate in protein synthesis (Table I). The terminal NH_2— in some proteins is acetylated or formylated. The order in which the amino acids appear is governed by a genetic mechanism: the sequence of bases in specific regions, or cistrons, of one of the two strands of DNA. The cistron is *transcribed* into a complementary single strand of RNA by the action of an enzyme, RNA polymerase. The RNA strand is a molecule of *messenger RNA* (mRNA) so called because it transmits the genetic message. This is translated by means of a code in which the sequence of bases in RNA is "read off" in consecutive groups of three. Each such group is termed a *codon*. There are 64 different codons, corresponding to the number of different ways in which the four RNA bases, A, C, G, and U, can be arranged in groups of three. Sixty-one of the 64 codons each specifies an amino acid, and the other three, UAA, UAG and UGA, are signals for the termination of a polypeptide chain (Table I).

A protein molecule may contain a single polypeptide chain, as in the case of cytochrome *c*, or it may consist of two or more polypeptide chains held together by various linkages, such as by —S—S— bridges between cysteine residues, as in insulin, or by noncovalent bonds, as in the hemoglobins.

The formation of proteins takes place on the surface of intracellular particles termed ribosomes. A ribosome is composed of proteins and RNA. It is formed by two loosely bound components of unequal size; the smaller one is termed the 30 S component and the larger, the 50 S component. The terms 30 S and 50 S refer to the speed of sedimentation of the particles in the ultracentrifuge, measured in Svedberg units (S). Actually, 30 S and 50 S components are obtained from *Escherichia coli*. Yeast and mammalian ribosome subunits are somewhat larger; about 40 S and 60 S. The sequence of events in protein synthesis has not been completely delineated, but as typified by *E. coli*, it is thought to be as follows: the "left-hand end" of messenger RNA molecule starts with a phosphate or triphosphate group in ester linkage to the 5′-OH group of the ribose components of the terminal nucleoside. This end binds to a 30 S ribosomal component. Perhaps several of the nucleotides at

the end of the mRNA molecule participate in the initial binding procedure. When protein synthesis starts, the 30 S particle and a tRNA molecule attach to a 50 S component, forming a 70 S ribosome. There are two sites on the ribosome, the polypeptide site and the amino acid site, each of which binds a molecule of transfer RNA (tRNA). Each molecule of tRNA contains about 75 to 85 bases and becomes covalently attached to a specific amino acid. The amino acid is esterified to the ribose of the adenosine group which terminates each tRNA molecule at the right-hand end (3′-OH end).

Before protein synthesis starts, the tRNA-binding sites on the ribosome are both unoccupied. The first of these sites (the polypeptide site) becomes occupied by a tRNA molecule which may be of a special type, termed a "chain-initiating" tRNA. It is known that there is such a tRNA for methionine in *E. coli*, and probably in some other organisms. This tRNA has the special property of participating in a reaction which formylates the NH_2 group of the methionine molecule carried by the tRNA.

Next, the second site (the amino acid site) on the ribosome binds a tRNA molecule, and a peptide bond is formed by an enzymically catalyzed reaction between the two amino acids on the pair of tRNA molecules that are bound to the ribosome, adjacently to each other. The formation of the peptide bond frees the first tRNA from linkage to its amino acid and it leaves the surface of the ribosome. The second tRNA molecule, carrying a dipeptide in ester linkage, now moves to the peptide site. The amino acid site is therefore vacated. It is promptly reoccupied by another tRNA, which carries the amino acid specified by the next codon on the mRNA strand. A second peptide linkage is formed, connecting the third amino acid to the dipeptide which is attached to the adjoining tRNA. This procedure continues repeatedly until the polypeptide chain is completed, at which point a signal for release is given by a chain-terminating codon in the mRNA.

The selection of tRNA molecules by the mRNA is carried out by codon-anticodon pairing. Each tRNA molecule contains a "loop" of seven unpaired bases, the middle three of which are the anticodon. The second and third bases of the anticodon form complementary (Watson-Crick) pairs with the first two bases of the codon, i.e., A pairs with U, and C with G. The pairing between the first base of the anticodon and the third base of the codon is less specific (Crick, 1966), and it appears to be as follows:

U in the anticodon pairs with	G or C	in the codon;
G in the anticodon pairs with	U or C	in the codon;
C in the anticodon pairs with	G	in the codon;
I in the anticodon pairs with	U,C, or A	in the codon.

It is not yet clear whether A occupies the first position in any anticodon but, if so, it is presumed to pair with U in codons.

The codons UAA, UGA, and UAG do not pair with a tRNA carrying an amino acid. When one of these codons reaches the amino acid-binding site on the ribosome, the ester linkage between the incumbent tRNA and its amino acid is broken, and a carboxyl group is formed, thus terminating the polypeptide chain.

The initiation of polypeptide chains is still under investigation. One of the initiating codons is AUG, which binds with formylmethionine tRNA to start the synthesis of a polypeptide chain in *E. coli* and probably in some other species. A short sequence of bases in mRNA possibly precedes the chain-initiating codon. It is suspected that this sequence may consist of two or more G's. This accords with the finding that clusters of C's in DNA are probably the sites where transcription of mRNA by RNA polymerase is initiated (Szybalski *et al.*, 1966).

Several enzymes participate in protein synthesis and there are at least three factors, apparently proteins, which are involved in polypeptide chain initiation (Iwasaki *et al.*, 1968). The mechanism by which the terminal carboxyl is set free is unknown. Also unknown is the manner in which ribosomes are formed and assembled from their constituent RNA and protein molecules. Takanami (1967) has shown that the 30 S and 50 S components obtained from various bacteria and yeast contain, respectively, one and two molecules of RNA, each with a characteristic 5′-terminal sequence of bases, the first base of which carries a monophosphate group in 5′-OH linkage. A smaller molecule of RNA, 5 S RNA, is also present in ribosomes.

Each ribosomal component contains several different proteins. The tRNA and rRNA molecules contain several bases which have been modified by the addition of substituents such as methyl groups.

D. Mutations and Evolution

It is obvious that mutations and evolution are two separate phenomena. The relationship between the two is complex. Many mutants are present in the existing population of any living species. A very small proportion of these may eventually become incorporated into the genome that typifies the majority of individuals in the species. The processes by which this incorporation takes place, including the effects of environment, are explained in textbooks which deal with population genetics. Stebbins (1966) emphasizes that mutations are rarely, if ever, the direct source of variation upon which evolutionary change is based. Mutations are fed into the gene pool and furnish a source of variability; this is steadily reduced by natural selection, which favors the survival of those

types that are best adapted to the environment. The supply of different types is furnished primarily by recombination of genes or entire chromosomes, occurring by sexual reproduction, polyploidy, aneuploidy, transduction, and other processes of chromosomal shuffling.

Changes in DNA are essential for the introduction of new characteristics to this general pool of genetic material. Probably the most drastic changes are the saltatory multiplications of segments of DNA (Britten and Kohne, 1965–1966), in which as many as a million copies of a short section of DNA may suddenly appear. Duplications of genes or segments of genes also take place. These changes, followed by differentiation caused by a steady barrage of point mutations, produce genes with new functions. Stebbins stresses the conclusion that there is no evidence for a relationship between rate of mutation and rate of evolution, because the supply of mutations is always more than sufficient to furnish the requirements for variants.

Point mutations are replacements of one base pair in DNA by another. It may be concluded that this process takes place each time a DNA molecule is replicated, and that the rate is that of the "error rate" in replication. This has been estimated to be in the neighborhood of 10^{-8} for

$$\begin{matrix} A \\ T \end{matrix} \rightarrow \begin{matrix} A \\ C \end{matrix}$$

and lower values for errors involving

$$\begin{matrix} G \\ C \end{matrix} \rightarrow \text{other pairs}$$

(Watson, 1965). The rate would correspond to about one to ten replacements per replication in the DNA of a eukaryotic organism containing about 10^9–10^{10} base pairs.

Watson's estimate of errors depends upon mispairing that results from the presence of the rare enolic forms of the purine and pyrimidine bases, as originally postulated by Watson and Crick (1953b). Such mispairing would produce transitional substitutions in DNA, but not transversions. Studies by Yanofsky and his co-workers (Yanofsky *et al.*, 1966a; Cox and Yanofsky, 1967) with a mutator strain of *E. coli* (Treffers *et al.*, 1954) showed that transversions of

$$\begin{matrix} A \\ T \end{matrix} \rightarrow \begin{matrix} C \\ G \end{matrix}$$

are produced by this strain so that the mutation rate is increased 1000-fold, from 10^{-8} to 10^{-5}. A possible explanation for the effect of the mutator gene is that it is itself a mutation rather than a gene, and that the

mutation is in DNA polymerase, so that the altered enzyme occasionally selects a pyrimidine to pair with a purine on the template. The existence of the Treffers mutator gene shows that a mechanism can exist which would account for the predominance of transversions in the evolutionary divergence of proteins (Jukes, 1966). The occurrence of mutations in proteins, such as DNA polymerase, that control the DNA replication mechanisms could provide a means for accelerating the mutation rate.

Base replacements in genetically functional regions of DNA are point mutations. The effects of these may be deleterious, or, quite rarely, beneficial, as is well recognized in genetics. However, the effect of a base replacement may also be *neutral,* as in the case of the replacement of one amino acid by another that produces no functional change in a protein. Finally, some mutations may have *zero* effect, as in the case of a change of a codon to a second codon which is synonymous with the first. A mutation of the codon UCG (serine) to UAG (chain termination) could in some locations be deleterious or even lethal, but the mutation UCG to UCA (serine) will have zero effect. A third possibility, UCG to ACG (threonine), could perhaps be a neutral mutation.

The hemoglobins furnish good illustrations of both mutations and evolution. Ninety-two mutations in human hemoglobin A have been chemically identified. Most of these are in the α and β chains of hemoglobin A. The variants were discovered mainly as a result of searching for electrophoretic changes in samples of hemoglobin obtained from patients. Variants not producing such changes may remain undetected; only 32 of the 75 single-base changes in the amino acid codons produce amino acid substitutions that are accompanied by a change in charge. Based on this proportionality, the 92 known variants could possibly indicate the existence of a total of over 200 actual mutant hemoglobins in the patients who were sampled. The number of different mutant hemoglobins in the total human population is obviously far greater than 92 or even than 200. No figures are available for the number of samples of blood that have been examined for the presence of variant hemoglobins, but the number must correspond to only a small fraction of the population, perhaps to only one person in 100,000.

There is, of course, an upper limit on the number of mutations that can occur in the molecule of hemoglobin A. The maximum theoretical number of amino acid replacements resulting from single-base changes in the genetic code is about 7 per site, corresponding to about 2000 for the combined α and β chains. It is difficult to estimate how many of these 2000 are possibilities which could exist in the form of mutants. Presumably at least half of these possible replacements could be accommodated in the secondary and tertiary structure of the hemoglobin mole-

cule. One thousand seems a large number, but it may be within the bounds of possibility, because 180 replacements have been found in α chains of various species and 103 replacements in β chains. Only 17 species of animals have been examined, and in some cases, only the α chain was analyzed. In the light of these findings at the evolutionary level, it is evident that many changes in hemoglobin are possible. It therefore seems likely that there could be 1000 different mutational variants in human hemoglobin A. A large pool of variants is evidently available for evolutionary selection, and it is possible to make some rough estimates of the rate at which this occurs.

The hemoglobin A of the gorilla differs from human hemoglobin A by amino acid substitutions corresponding to two single-base replacements. The α chains of orangutan and human differ by one single-base replacement, and the human and chimpanzee α chains are identical Sarich and Wilson (1967) have recently estimated five million years as the time of separation of man from the African apes.

Evolutionary changes separated the α and β chains of human hemoglobins from those of the horse. The differences between these two pairs of polypeptides have all been identified. They apparently resulted from point mutations which produced amino acid changes. These changes correspond to a minimum of 56 base substitutions. This is equivalent to 1 base change per chain in 3 million years if the rate of change is the same in each species, and if the common ancestor lived 75 million years ago. The human:carp α chain difference is 92 base differences in 140 amino acid residues compared. The common ancestor may have lived 400 million years ago, so the rate of change is about one base in 8 million years for one of the two chains, roughly corresponding to twice this rate, one base per 4 million years in the two polypeptide chains that are present in hemoglobin A. Similarly, a comparison of the human and rhesus monkey β chains shows 7 base replacements in 146 codons compared (Matsuda *et al.*, 1968). This difference accumulated during 30 million years (Sarich and Wilson, 1967).

The comparisons seem to be in the same order of magnitude for a rate of evolutionary change in the hemoglobin gene corresponding to about one base pair replacement in 3 or 4 million years, selected from a large pool of mutants. It should not be assumed that the rate is constant; indeed, it seems that certain eras are marked by bursts of increased evolutionary rates. An abrupt change in environment could produce a rapid selection of new forms from the genetic pool. Such an event is suggested by Degens *et al.* (1967) as having occurred at the Precambrian-Cambrian boundary as a result of an increase in the pH of the ocean. They postulate that the increase produced more favorable

conditions for the deposition of $CaCO_3$ and changed the composition of proteins in the integument of the ancestors of mollusks. As a consequence, the integument became calcified, and, simultaneously, there occurred an enormous diversification in shell forms. Perhaps the evolutionary flexibility of hemoglobin has, by an analogous procedure, contributed to the rapid evolution of vertebrates.

II. Evolutionary Changes in Protein Structure

A. The Cytochromes c

The cytochromes *c* are a family of homologous proteins found in all aerobic nucleated (eukaryotic) organisms. They consist of a single polypeptide chain containing between 103 and 112 amino acids. The chain is longest in the cytochrome *c* of wheat and shortest in that of tuna. Most of the variations in length arise from the fact that additional amino acid residues are attached to the amino-terminal end in a number of invertebrate organisms. The sequences are given in Table III.

All the cytochromes *c* have a molecular weight near 12,300, an isoelectric point near pH 10, a single heme prosthetic group per molecule, and an oxidation-reduction potential near +0.250 V (Margoliash and Fitch, 1967). All react with mammalian cytochrome oxidase and are interchangeable with each other in the terminal oxidation chain of mitochondria. Their activities and physical and chemical properties differ from those of various *c*-type cytochromes which have been obtained from bacteria, although there are indications that both types have a common evolutionary origin, as deduced from a comparison of *Neurospora* cytochrome *c* and *Pseudomonas* cytochrome C-551 (Cantor and Jukes, 1966a,b), and of horse cytochrome *c* with *Rhodospirillum rubrum* cytochrome c_2 (Dus and Sletten, 1968).

The cytochromes *c* from various mammals do not exhibit the property, commonly associated with "isozymes," of showing different primary structures that are organ-specific. It was found by Stewart and Margoliash (1965) that samples of hog cytochrome *c* from kidney, liver, heart muscle, and brain all had identical amino acid sequences. There are, however, two forms of cytochrome *c* in yeast, termed "iso-1-cytochrome *c*" and "iso-2-cytochrome *c*" (Sherman *et al.*, 1968). These are under the control of two separate genes.

The similarity between the amino acid sequences of the cytochrome *c* from widely differing organisms is such that this protein, more than any other, has been used to support the thesis that all organisms which synthesize it have a common evolutionary origin. The differences between

the various cytochromes *c* bear a roughly quantitative relationship to the phylogenetic separation of the various species from which they are derived.

The sequences of 18 vertebrate cytochromes *c* are shown in Table III. They have been listed in terms of differences from a common vertebrate type, which is a sequence written to include the amino acid residues that predominate at each of the 104 sites. The differences from the vertebrate type are given in Table IV. Since the "vertebrate type" sequence is based principally on mammals, it does not represent a modal composition for the Vertebrata; obviously, if more fishes had been included, we should expect the type molecule to be more like tuna cytochrome *c*.

No two cytochromes *c* selected from all those that have been analyzed are found to differ from each other in more than 44% of their homologous sites, but the total number of amino acid residues which remains constant at such sites in all cytochromes *c* so far examined is only 35 residues, equivalent to 33%. A calculation based on the Poisson distribution indicates that this residuum might contain about 6 residues that are potentially variable and might be shown to be so if more cytochromes *c* were examined. Examples of two widely differing cytochromes are those of *Neurospora crassa* and wheat (*Triticum vulgare*). These are identical in 58 of 107 comparable homologous sites. Of these 58 sites in the sequences of other cytochromes *c*, 24 differ from either *Neurospora* or wheat and are therefore potentially variable sites. This may be taken as evidence that the divergence between *Neurospora* and wheat has not yet reached a maximum, so that the evolutionary differentiation of these two species is presumably incomplete.

It may seem surprising that the evolutionary difference between the cytochromes *c* of two species that can reproduce in a few days or even less, such as *Neurospora* and bakers' yeast, appears to be no greater than the difference between either one of these and the cytochrome *c* of human beings which usually take 15 to 30 years per generation. The comparisons are as follows:

	No. of sites compared	Identical sites
Neurospora: bakers' yeast	107	66
Neurospora: human	104	59
Bakers' yeast: human	104	68

A conceivable explanation might be that the divergence between the

cytochromes *c* of *Neurospora* and yeast reached a maximum value long ago, and that subsequent changes are reflected merely in changes of amino acids that have already been differentiated. This explanation is not valid, because 34 residues that are identical in corresponding sites in the cytochromes *c* of *Neurospora* and yeast vary in other species; therefore, equilibrium has not been reached in the *Neurospora*:yeast divergence.

A second possible explanation is that point mutations reflect base replacements in DNA and that these base replacements follow molecular events that occur at a constant rate with respect to time. If this explanation is valid, the mean generation time of a species would not affect the rate at which its proteins exhibited evolutionary divergence from their homologs in another species. The changes in DNA per unit of time would go on regardless of the number of generations and this would be translated into a rate of change in proteins that would be independent of the length of the reproductive cycle, but would perhaps be related to the number of replication cycles of DNA. The findings with the cytochromes *c* tend to support such a model.

As pointed out by Margoliash and Smith (1965), and Margoliash and Fitch (1967), the presence of cytochromes *c* in a wide variety of animals and plants, including yeasts and molds, together with the similarities in the primary structures of all these cytochromes, leads to the conclusion that there has been only one effective emergence of eukaryotic life on earth. As mentioned above, the cytochromes *c* of eukaryotic organisms and bacteria are sufficiently similar to betoken a common origin, so that this conclusion may be extended to organisms without visible nuclei.

Comparisons of the cytochromes *c* on the basis of differences at homologous sites, expressed either as amino acid differences or as minimum base differences per codon, have matched with the phylogenetic relationships of the species involved. Fitch and Margoliash (1967a) have emphasized this procedure as a method for phylogenetic comparison, but it is our conclusion that the correlation between the cytochrome *c* data and systematics is only a rough one. The obvious reason for this low order of accuracy is that a single small protein such as cytochrome *c* is derived from only about one ten-millionth of the possible genetic information, calculated as DNA, that is present in a higher organism. The frequent tendency of biochemists to overlook this point is sometimes a source of understandable irritation to classical taxonomists (Simpson, 1964). Comparisons of several vertebrate cytochromes *c* are shown in Table V. From the standpoint of taxonomy, a number of interesting anomalies are obvious. There is more than twice as much difference

TABLE III

AMINO ACID RESIDUES IN CYTOCHROMES *c*[a]

Site No.:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Neurospora				NH_2Gly		Phe	Ser	Ala	Gly	Asp	Ser	Lys	Lys	Gly	Ala	Asn
Bakers' yeast				NH_2Thr	Glu	Phe	Lys	Ala	Gly	Ser	Ala	Lys	Lys	Gly	Ala	Thr
Candida krusei		NH_2Pro		Ala	Pro	Phe	Glu	Gln	Gly	Ser	Ala	Lys	Lys	Gly	Ala	Thr
Debaromyces kloeckeri[b]						Tyr		Lys			Glu					Asn
Wheat	AcNHAla	Ser	Phe	Ser	Glu	Ala	Pro	Pro	Gly	Asn	Pro	Asp	Ala	Gly	Ala	Lys
Moth (Samia cynthia)				NH_2Gly		Val	Pro	Ala	Gly	Asn	Ala	Glu	Asn	Gly	Lys	Lys
Vertebrate type (v.t.)								AcNHGly		Asp	Val	Glu	Lys	Gly	Lys	Lys
Vertebrates (differences from v.t.)											c,p-Ile	t-Ala				

[a] Sources of information: Bahl and Smith (1965); Chan and Margoliash (1966 a,b); Chan *et al.* (1966); Fitch (1966); Fitch and Margoliash (1967 a, b,); Heller and Smith (1966); Kreil (1963, 1965); Margoliash (1963); Margoliash and Smith (1965); Matsubara and Smith (1963); McDowall and Smith (1965); Narita and Sugeno (1968); Narita and Titani (1965); Needleman and Margoliash (1966); Nolan and Margoliash (1966, 1968); Rothfus and Smith (1965); Smith and Margoliash (1964); Stevens *et al.* (1967); Stewart and Margoliash (1965); Yaoi *et al.* (1966).

[b] Differences from *C. krusei* (Narita and Sugeno, in preparation).

[c] Abbreviations are as follows: b = beef, sheep, and pig; h = horse; r = rabbit; w = whale; hu = human; c = chicken and turkey; k = kangaroo; rs = rattlesnake; * = yeast C_2; st = snapping turtle; d = dog; m = rhesus monkey; t = tuna; bf = bullfrog; x = unspecified (from Nolan and Margoliash, 1968); du = duck; df = dogfish; p = penguin.

(*continued*)

TABLE III—*Continued*

Site No.:	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Neurospora	Leu	Phe	Lys	Thr	Arg	Cys	Ala	Glu	Cys	His	Gly	Glu	Gly	Gly	Asn	Leu
Bakers' yeast	Leu	Phe	Lys	Thr	Arg	Cys	Glu	Leu	Cys	His	Thr	Val	Glu	Lys, Glu*	Gly	Gly
Candida krusei	Leu	Phe	Lys	Thr	Arg	Cys	Ala	Glu	Cys	His	Thr	Ile	Glu	Ala	Gly	Gly
Debaromyces kloeckeri[b]							Glu	Leu				Val		Glx		
Wheat	Ile	Phe	Lys	Thr	Lys	Cys	Ala	Gln	Cys	His	Thr	Val	Asp	Ala	Gly	Ala
Moth (Samia cynthia)	Ile	Phe	Val	Gln	Arg	Cys	Ala	Gln	Cys	His	Thr	Val	Glu	Ala	Gly	Gly
Vertebrate type (v.t.)	Ile	Phe	Val	Gln	Lys	Cys	Ala	Gln	Cys	His	Thr	Val	Glu	Lys	Gly	Gly
Vertebrates (differences from v.t.)	t-Thr df-Val		hu,m-Ile; x-Thr	hu,m-Met; rs-Thr			hu,c,rs,m,bf,du p-Ser					x-Cys		df,t-Asn; bf-Ala	x-Ala	

Site No.:	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
Neurospora	Thr	Gln	Lys	Ile	Gly	Pro	Ala	Leu	His	Gly	Leu	Phe	Gly	Arg	Lys	Thr
Bakers' yeast	Pro	His, Asn*	Lys	Val	Gly	Pro	Asn	Leu	His	Gly	Ile	Phe	Gly	Arg	His	Ser
Candida krusei	Pro	His	Lys	Val	Gly	Pro	Asn	Leu	His	Gly	Ile	Phe	Ser	Arg	His	Ser
Debaromyces kloeckeri[b]											Val	-	-		Thr	
Wheat	Gly	His	Lys	Gln	Gly	Pro	Asn	Leu	His	Gly	Leu	Phe	Gly	Arg	Gln	Ser
Moth (Samia cynthia)	Lys	His	Lys	Val	Gly	Pro	Asn	Leu	His	Gly	Phe	Tyr	Gly	Arg	Lys	Thr
Vertebrate type (v.t.)	Lys	His	Lys	Thr	Gly	Pro	Asn	Leu	His	Gly	Leu	Phe	Gly	Arg	Lys	Thr
Vertebrates (differences from v.t.)				t-Val					df,t-Trp, st,k-Asn; bf-Tyr; x-Ser, Gln		k,p-Ile	bf, st-Ile				

(continued)

TABLE III—*Continued*

Site No.:	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64
Neurospora	Gly	Ser	Val	Asp	Gly	Tyr	Ala	Tyr	Thr	Asp	Ala	Asn	Lys	Gln	Lys	Gly
Bakers' yeast	Gly	Gln	Ala	Glu	Gly	Tyr	Ser	Tyr	Thr	Asp	Ala	Asn	Ile	Lys	Lys	Asn
Candida krusei	Gly	Gln	Ala	Glu	Gly	Tyr	Ser	Tyr	Thr	Asp	Ala	Asn	Lys	Arg	Ala	Gly
Debaromyces kloeckeri[b]						Phe								Lys	Lys	
Wheat	Gly	Thr	Thr	Ala	Gly	Tyr	Ser	Tyr	Ser	Ala	Ala	Asn	Lys	Asn	Lys	Ala
Moth (Samia cynthia)	Gly	Gln	Ala	Pro	Gly	Phe	Ser	Tyr	Ser	Asn	Ala	Asn	Lys	Ala	Lys	Gly
Vertebrate type (v.t.)	Gly	Gln	Ala	Pro	Gly	Phe	Ser	Tyr	Thr	Asp	Ala	Asn	Lys	Asn	Lys	Gly
Vertebrates (differences from v.t.)				p,st,c, t,du-Glu; r,rs, w-Val; bf-Ala; df-Gln		hu, rs,m,-Tyr	k,h-Thr			df,hu, rs,m, -Ala; st,x-Glu				df,t-Ser		

Site No.:	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
Neurospora	Ile	Thr	Trp	Asp	Glu	Asn	Thr	Leu	Phe	Glu	Tyr	Leu	Glu	Asn	Pro	Lys
Bakers' yeast	Val	Leu	Trp	Asp	Glu	Asn	Asn	Met	Ser	Glu	Tyr	Leu	Thr	Asn	Pro	Lys
Candida krusei	Val	Glu	Trp	Ala Thr	Glu	Pro Glx	Thr Asp	Met Leu	Ser	Asp	Tyr	Leu	Glu	Asn	Pro	Lys
Debaromyces kloeckeri[b]																
Wheat	Val	Glu	Trp	Glu	Glu	Asn	Thr	Leu	Tyr	Asp	Tyr	Leu	Leu	Asn	Pro	Lys
Moth (Samia cynthia)	Ile	Thr	Trp	Gly	Asp	Asp	Thr	Leu	Phe	Glu	Tyr	Leu	Glu	Asn	Pro	Lys
Vertebrate type (v.t.)	Ile	Thr	Trp	Gly	Glu	Asp	Thr	Leu	Met	Glu	Tyr	Leu	Glu	Asn	Pro	Lys
Vertebrates (differences from v.t.)		hu,k, st, rs, -Ile x- Val		h- Lys; t- Asn df- Gln	t- Asn; rs- Asp df- Gln	w,st, h,b, d,df- Glu			df- Arg	x- Ile, Val						

(continued)

TABLE III—*Continued*

Site No.:	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96
Neurospora	Lys	Tyr	Ile	Pro	Gly	Thr	Lys	Met	Ala	Phe	Gly	Gly	Leu	Lys	Lys	Asp
Bakers' yeast	Lys	Tyr	Ile	Pro	Gly	Thr	Lys	Met	Ala	Phe	Gly	Gly	Leu	Lys	Lys	Glu
Candida krusei	Lys	Tyr	Ile	Pro	Gly	Thr	Lys	Met	Ala	Phe	Gly	Gly	Leu	Lys	Lys	Ala
Debaromyces kloeckeri[b] Wheat	Lys	Tyr	Ile	Pro	Gly	Thr	Lys	Met	Val	Phe	Pro	Gly	Leu	Lys	Lys	Pro
Moth (Samia cynthia)	Lys	Tyr	Ile	Pro	Gly	Thr	Lys	Met	Val	Phe	Ala	Gly	Leu	Lys	Lys	Ala
Vertebrate type (v.t.)	Lys	Tyr	Ile	Pro	Gly	Thr	Lys	Met	Ile	Phe	Ala	Gly	Ile	Lys	Lys	Lys
Vertebrates (differences from v.t.)									rs-Val		hu, m-Val; rs-Thr		df, rs-Leu	rs-Ser		d-Thr

Site No.:	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112
Neurospora	Lys	Asp	Arg	Asn	Asp	Ile	Ile	Thr	Phe	Met	Lys	Glu	Ala	Thr	Ala	-
Bakers' yeast	Lys	Asp	Arg	Asn	Asp	Leu	Ile	Thr	Tyr	Leu	Lys	Lys	Ala	Cys	Glu	-
Candida krusei	Lys	Asp	Arg	Asn	Asp	Leu	Val	Thr	Tyr	Met	Leu	Glu	Ala	Ser	Lys	-
Debaromyces kloeckeri[b]							Ile			Leu	Val	Lys		Thr		
Wheat	Gln	Asp	Arg	Ala	Asp	Leu	Ile	Ala	Tyr	Leu	Lys	Lys	Ala	Thr	Ser	Ser
Moth (Samia cynthia)	Asn	Glu	Arg	Ala	Asp	Leu	Ile	Ala	Tyr	Leu	Lys	Glu	Ser	Thr	Lys	-
Vertebrate type (v.t.)	Gly	Glu	Arg	Ala	Asp	Leu	Ile	Ala	Tyr	Leu	Lys	Lys	Ala	Thr	Asn	Glu
Vertebrates (differences from v.t.)	bf, hu, m- Glu; p,c,df,du- Ser r- Asp; rs- Lys; st- Ala h- Thr			h,b- Glu; c- Val df,t- Gln; rs- Thr; bf- Gly; x- Asp, Lys	rs- Asn		t- Val					bf, st, p,c,du- Asp; rs- Glu t- Ser x- Gln	df- Thr; rs- Lys	bf- Glu df- Ala	d- Lys; st, c,p t- Ser; df,du, rs- Ala; bf- Cys	bf, st, c,du-p Lys; df,t- Gap; rs- Ala

TABLE IV

MINIMUM BASE DIFFERENCES BETWEEN THE CODONS OF INDIVIDUAL VERTEBRATE CYTOCHROMES *c* AND THE GENERALIZED "VERTEBRATE TYPE" CYTOCHROME *c*

Human	9	Kangaroo	4
Rhesus monkey	8	Chicken and turkey	10
Horse	7	Duck	10
Cattle, sheep, and pig	2	Rattlesnake	24
Rabbit	3	Snapping turtle	12
Whale	7	Bullfrog	16
Dog	4		

between bullfrog and rattlesnake as between bullfrog and chicken cytochrome *c*. The difference between bullfrog and rabbit cytochrome *c* is almost identical with that between two mammals, horse and human. The rabbit is more similar than the horse to all the other vertebrates; and the rattlesnake and snapping turtle, both of which are reptiles, contrast astonishingly in the extent of their respective disparities from the other eight vertebrates. The totals at the foot of each column serve to emphasize the position of each of the nine species relative to the others, using cytochrome *c* as a measure of differentiation. It seems to indicate that one should not rely on cytochrome *c* comparisons as a phylogenetic measurement in vertebrates.

Less inconsistency is evident when the comparison is made over a range of phyla. Table VI shows the minimum base differences per codon for the cytochromes *c* of two mammals, a bird, a reptile, a fish, a moth,

TABLE V

BASE DIFFERENCES BETWEEN CODONS OF HOMOLOGOUS AMINO ACID RESIDUES IN SOME VERTEBRATE CYTOCHROMES *c*

	1	2	3	4	5	6	7	8	9	10
1 Human	—	14	8	8	10	16	19	13	22	32
2 Horse	14	—	9	9	10	14	30	15	20	25
3 Rabbit	8	9	—	6	7	9	22	11	15	21
4 Whale	8	9	6	—	11	12	22	13	18	27
5 Kangaroo	10	10	7	11	—	14	26	13	18	26
6 Chicken	16	14	9	12	14	—	26	8	13	25
7 Rattlesnake	19	30	22	22	26	26	—	24	29	35
8 Snapping turtle	13	15	11	13	13	8	24	—	13	24
9 Bullfrog	22	20	15	18	18	13	29	13	—	27
10 Tuna	32	25	21	27	26	28	35	24	27	—
Totals:	142	146	108	126	135	140	233	134	175	242
MBDC:	0.152	0.156	0.116	0.135	0.144	0.150	0.25	0.143	0.187	0.26

TABLE VI
MINIMUM BASE DIFFERENCES BETWEEN CODONS (MBDC) OF HOMOLOGOUS AMINO ACID RESIDUES IN NINE CYTOCHROMES *c*

	1	2	3	4	5	6	7	8	9
1 Human	—	14	16	19	32	35	50	58	59
2 Horse	14	—	14	30	25	32	55	62	61
3 Chicken	16	14	—	26	28	31	59	61	60
4 Rattlesnake	19	30	26	—	35	38	54	62	59
5 Tuna	32	25	28	35	—	40	63	64	70
6 Moth (*Samia cynthia*)	35	32	31	38	40	—	61	58	60
7 Bakers' yeast	50	55	59	54	63	61	—	37	55
8 Yeast (*Candida krusei*)	58	62	61	62	64	58	37	—	56
9 Mold (*Neurospora*)	59	61	60	59	70	60	55	56	—
Totals:	283	293	295	323	357	355	434	458	480
MBDC:	0.34	0.35	0.36	0.39	0.44	0.43	0.52	0.55	0.59

two yeasts, and a mold. A steady trend is evident as the phylogenetic difference increases. This is seen in the MBDC averages at the foot of the table, which express the difference between each organism and the other eight.

1. Rhodospirillum rubrum Cytochrome c_2

An analysis of the sequence of *Rhodospirillum rubrum* cytochrome c_2 was published by Dus and Sletten (1968). Its homology to horse cytochrome *c* is shown in Table VII. The average MBDC when the two are compared is 0.76. This is a difference higher than the value of 0.60 for a comparison of horse cytochrome *c* and *Candida krusei* cytochrome *c*. The latter shows the greatest difference from horse cytochrome *c* of any of the sequences in Table VI.

2. Internal Repetition in the Molecules of Cytochromes c

An event of crossing-over and recombination within the cistron for a polypeptide chain can produce a repetition of a part of its sequence, thus lengthening the molecule of a protein. Such repetitions will be subject to evolutionary differentiation as a result of subsequent point mutations, so that they may be difficult to detect by visual inspection of the polypeptide chain. However, it is possible to search for repetitions by means of a computer program which examines the amino acid residues in terms of minimum differences per codon between each amino acid and all 19 others (Fitch, 1966a). The computerized procedure then scans the polypeptide chains for evidence of repetitions of varying lengths. By this means, Cantor and Jukes (1966a) detected the presence of a possible

TABLE VII

SEQUENCES OF HORSE CYTOCHROME *c* (A) (NUMBERED AS IN TABLE III) AND *Rhodospirillum rubrum* CYTOCHROME c_2 (B) ALIGNED TO SHOW HOMOLOGY

```
                1                                  10                                  20
A        NH2Glu-Gly-Asp-Ala-Ala-Ala-Gly-Glu-Lys-Val-Ser-    -Lys-Lys-Cys-Leu-Ala-Cys-His-Thr-Phe-Asp-Gln-Gly-Gly-
                                                   10                                  20
B          AcNHGly-Asp-Val-Glu-Lys-Gly-Lys-Lys- Ile-Phe-Val-Gln-Lys-Cys-Ala-Gln-Cys-His-Thr- Val-Glu-Lys-Gly-Gly-
   MBDC                    1  1  2       1      1  1       1           2  2                 1  1  1
                      30                                  40                                  50
A  Ala-Asn-Lys- Val-Gly-Pro-Asn-Leu-Phe-Gly- Val-Phe-Glu-Asn-Thr- Ala- Ala- His-Lys-Asp-Asn-Tyr- Ala-Tyr- Ser-Glu-Ser-Tyr-
                      30                                  40                                  50
B  Lys- His-Lys-Thr-Gly-Pro-Asn-Leu- His-Gly-Leu-Phe-Gly-Arg-Lys-Thr-Gly-Gln-Ala-Pro-Gly-Phe-Thr-Tyr-Thr-Asp-Ala-Asn-
    2  1       2                    2       1      1  2  1  1  1  1  2  2  2  1  1        1  1  1  1
                            60                                  70                                  80
A  Thr-Glu-Met-Lys- Ala-Lys-Gly-Leu-Thr-Trp-Thr-Glu- Ala-Asn-Leu- Ala- Ala-Tyr- Val-Lys-Asn-Pro-Lys-Ala-Phe-Val-Leu-Glu-
                                  60                                  70
B     -  -    -Lys-Asn-Lys-Gly- Ile-Thr-Trp-Lys-Glu-Glu-Thr-Leu-Met-Glu-Tyr-Leu-Glu-Asn-Pro-Lys-Lys-Tyr- Ile-Pro-Gly-
                2          1       1       1  1       2  1       1  1                 2  1  1  1  1
                                  90                                  100
A  Lys-Ser-Gly-Asp-Pro-Lys-Ala-Lys- Ser-Lys-Met-Thr-Phe-    -Lys-Leu-Thr-Lys-Asp-Asp-Glu- Ile-Glu-Asn- Val-Ile-Ala-Tyr-
                                                                      90
B     -  -  -  -  -  -  -  -  -Thr-Lys-Met- Ile-Phe-Ala-Gly- Ile-Lys-Lys-Lys-Thr-Glu-Arg-Glu-Asp-Leu-Ile-Ala-Tyr-
                                  1          1          2  1  1       2  2       1       1
          110
A  Leu-Lys-Thr-Leu-LysCOOH
          100          104
B  Leu-Lys-Lys- Ala-Thr-Asn-GluCOOH
             1  2  1
```

internal repetition in the cytochrome *c* molecule of *Neurospora crassa* which was not perceptible in the cytochromes *c* of other species. This repetition is illustrated below:

Residue No. (from Table III)

	6													20
	Lys	Gly	Ala	Asn	Leu	Phe	Lys	Thr	Arg	Cys	Ala	Glu	Cys	His-Gly
	21													35
	Glu	Gly	Gly	Asn	Leu	Thr	Gln	Lys	Ile	Gly	Pro	Ala	Leu	His-Gly
MBDC	1	0	1	0	0	2	1	1	1	1	1	1	2	0 0

Evidence of the internal repetition has disappeared from the other cytochromes *c* as a result of evolutionary differentiation, as follows:

Human cytochrome *c*:

Residue No.

	6													20
	Lys	Gly	Lys	Lys	Ile	Phe	Ile	Met	Lys	Cys	Ser	Gln	Cys	His-Thr
	21													35
	Val	Glu	Lys	Gly	Gly	Lys	His	Lys	Thr	Gly	Pro	Asn	Leu	His-Gly
MBDC	2	1	0	2	2	3	2	1	1	1	1	2	2	0 2

Tuna cytochrome *c*:

Residue No.

	6													20
	Lys	Gly	Lys	Lys	Thr	Phe	Val	Gln	Lys	Cys	Ala	Gln	Cys	His-Thr
	21													35
	Val	Glu	Asn	Gly	Gly	Lys	His	Lys	Val	Gly	Pro	Asn	Leu	Trp-Gly
MBDC	2	1	1	2	2	3	2	1	2	1	1	2	2	3 2

The repetition existed at one time in all the cytochromes *c*, as shown by the fact that homology exists between *Neurospora crassa* cytochrome *c* and the other cytochromes *c* over the region of 30 consecutive amino acids, residues 6 to 35, corresponding to the two repetitive sequences. It is possible that the entire molecule of cytochrome *c* has evolved by repeated end-to-end duplication of a short piece of DNA, somewhat in the manner described by Britten and Kohne (1965–1966) for the short repetitious segment of mouse satellite DNA.

Evidence for more than one repetition of the sequence of 15 amino acid residues was found by Cantor and Jukes (1966b) when *Neurospora* cytochrome *c* was compared with cytochrome C-551 of *Pseudomonas*. This comparison is shown in Table VIII. This comparison indicates that *Neurospora* cytochrome *c* and *Pseudomonas* cytochrome C-551 have a common evolutionary origin. The comparison of the two shows also that the internal repetition within *Pseudomonas* cytochrome *c* may have occurred six times. This repeated occurrence of internal duplication

within the *Pseudomonas* cytochrome *c* molecule is detectable only by comparing it with the sequence in *Neurospora* cytochrome *c* as shown in Table VIII. The observations are a good illustration of the fact that evolutionary relationships in proteins eventually become obscured by the slow process of allogenic differentiation defined below. The most interesting phenomena in molecular evolution, those which carry evidence of the very earliest relationships in evolutionary history, are, therefore, found at the borderline that separates statistical significance from insignificance. Beyond this borderline, of course, the traces of primordial evolution have vanished.

TABLE VIII

INTERNAL REPETITIONS IN AMINO ACID SEQUENCES OF *Neurospora* (N) AND *Pseudomonas* (Ps) CYTOCHROMES *c*

Neurospora

```
          1
(a)  Glu-Phe-Ser-Ala-Gly-Asp-Ser-Lys-Lys-Gly-Ala-Asn-Leu-Phe-Lys
          16                              23
(b)  Thr-Arg-Cys-Ala-Glu-Cys-His-Gly-Glu-Gly-Gly-Asn-Leu-Thr-Gln
          31                              38
     Lys-Ile-Gly-Pro-Ala-Leu-His-Gly
```

Pseudomonas

```
                                  1
                                 Glu-Asp-Pro-Glu-Val-Leu-Phe-Lys-Asn
                                  16              17
     Lys-Gly-Cys-Val-Ala-Cys-His-  -   -   -Ala-Ile-Asp-Thr-Lys
      22
     Met-Val-Gly-Pro-Ala-Tyr-Lys-Asp-Val-Ala-  -   -   -   -
      32
(c)  Ala-Lys-Phe-Ala-Gly-Gln-Ala-Gly-Ala-Glu-Ala-Glu-Leu-Ala-Gln
      47
(d)  Arg-Ile-Lys-Asn-Gly-Ser-Gln-Gly-Val-Trp-Gly-Pro-Ile-Pro-Met
      62                                                        76
(e)  Pro-Pro-Asn-Ala-Val-Ser-Asp-Asp-Glu-Ala-Gln-Thr-Leu-Ala-Lys
```

Sequences compared	MBDC
(a) : (b)	1.13
N16-23 : N31-38	0.78
(b) : (c)	1.00
(b) : (d)	1.20
(b) : (e)	0.93
(c) : (d)	1.40
(c) : (e)	1.20
(d) : (c)	1.27
N26-38 : Ps17-29	0.85
N 7-22 : Ps 1-16	1.12
Ps5-10 : Ps17-28	1.17

3. Invariant Residues in the Cytochromes c

Margoliash and Smith (1965) and Margoliash (1963) have repeatedly called attention to a consecutive sequence of 11 amino acid residues present in all the cytochromes *c* of the type found in vertebrates, yeast, etc. This sequence extends from residues 78 to 88 in Table III, identical with residues 70 to 80 in the vertebrate cytochromes *c*. It is presumed that this polypeptide sequence interacts with cytochrome *c* oxidase and that this interaction is essential for the functioning of the cytochrome system. Presumably this sequence of 11 amino acid residues is essential in the sense that any mutations occurring within it are lethal and disappear from the scene, thus accounting for its preservation in all species that contain this type of cytochrome. There are 23 other amino acid residues that have not been found to vary in any of the cytochromes *c* so far analyzed. These are indicated by asterisks in Table III. The question presents itself as to how many of these other 23 amino acid residues are invariant and how many are subject to possible evolutionary variations that have not yet been detected or that have not yet occurred.

An attempt has been made to answer this question by means of the following assumptions:

1. Amino acid replacements in the cytochrome *c* polypeptide chain occur on a random basis but replacements of invariant residues have effects that are highly deleterious and therefore replacements of these are not found in living organisms.
2. The number of changes at variable sites follows the Poisson distribution, which describes mathematically the chances of a number of events occurring at the same locus, such as base changes that occur as a result of a random series of point mutations.
3. One of the amino acids at each site is the amino acid that originally occupied the site before evolutionary differentiation started.
4. The number of changes at a site must be calculated in terms of the genetic code. For example: Gly, Gly, Asn, Gly, Gly represents two changes; Gly, Gly, Gly, Val, Gly, one change; Gly, Gly, Asn, Val, Gly, three changes.

These assumptions enable Table IX to be constructed. The number of unchanged sites will therefore represent the number of invariant sites plus the number of sites that should be unchanged in terms of examining the changed sites by the Poisson distribution.

A comparison of the Poisson distribution with the calculated results is shown in Fig. 1. The calculation leads to the estimate that there are 29 residues within the cytochrome *c* sequence that are invariant.

TABLE IX
DISTRIBUTION OF AMINO ACID CHANGES IN 110 SITES COMPARED IN 23 CYTOCHROMES

Interchanges per site	0	1	2	3	4	5	6	7	8
Distribution found	35	17	18	19	10	6	3	0	2
Minus 29 invariable sites	6	17	18	19	10	6	3	0	2
Poisson distribution for $m = 2.6$	6.0	15.7	20.4	17.7	11.5	6.0	2.6	1.0	0.3

The six remaining unchanged residues are therefore susceptible to evolutionary changes which have not yet been discovered because not enough cytochromes *c* have yet been analyzed.

Using a somewhat different approach, Fitch and Margoliash (1967b) arrived at a similar numerical conclusion. They compared the known cytochrome *c* sequences with that of "various ancestral forms" of cytochrome *c* whose amino acid sequences they did not specify. The calculation by Fitch and Margoliash is based on the elimination of five sites as being highly mutable and therefore aberrant. No such exclusion was made in the calculation used in the present article, because we found that the frequency of mutations at these sites was not in conflict with the variation inherent in the Poisson distribution. Fitch and Margoliash concluded that there were 27 to 29 invariant sites.

The calculations shown in Fig. 1 support the hypothesis that amino acid replacements during the evolution of homologous proteins are distributed on a random basis.

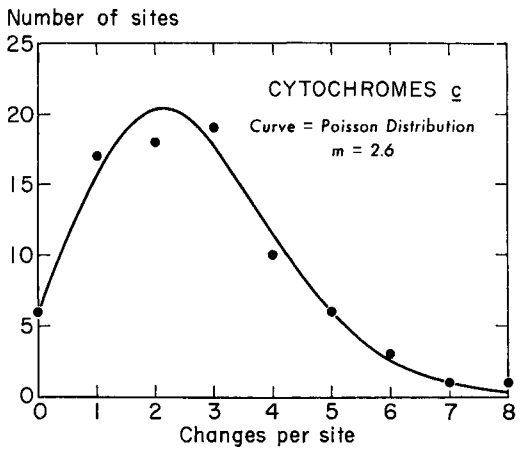


FIG. 1. Comparison of Poisson distribution with calculated results for cytochromes *c*.

B. The Globins

The globins are a group of proteins which include the hemoglobins and myoglobins. They are composed of polypeptide chains containing about 140 to 160 amino acid residues. Each chain is bound to a molecule of heme, to which is attached an atom of iron. The binding of heme to globin is loose, and the two may be separated by mild acidification.

The globins evolved from a single archetypal form by duplication and translocation, which separated the genes for myoglobin and the various forms of hemoglobin from each other, and by differentiation, which produced the distinctions that characterize myoglobin and each of the different hemoglobin polypeptide chains. Differentiation occurred both with and without speciation, as shown by comparing, for example, the α and β hemoglobin chains of human beings and horses. The archetypal globin molecule evolved from a shorter molecule (Fitch, 1966b; Cantor and Jukes, 1966a).

Duplication of the genes for the globins occurred at widely spaced intervals, some of which were hundreds of millions of years apart. The process is diagrammatically illustrated in Fig. 2. A comparison of the globin polypeptide chains of various species indicated that their differentiation proceeded steadily, accompanied by base replacements scattered more or less randomly along the DNA cistrons that code for each chain. These base replacements originated in point mutations, but the reservation must be made that the great majority of the mutations did not result in evolutionary changes. The number of mutations may be as-

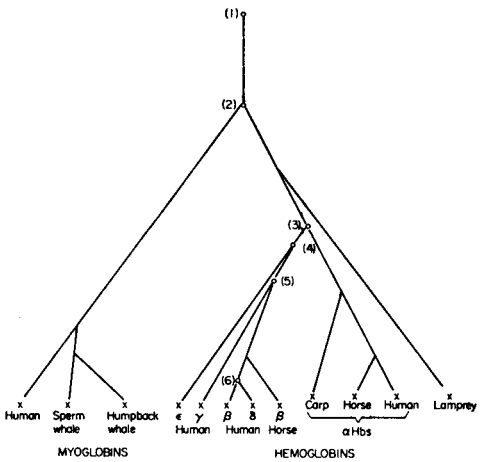


FIG. 2. Evolution of some of the globins. ○, event of gene duplication. (1), internal repetition of 21-peptide. (2–6), separation by gene duplication of lines leading to various hemoglobins.

sumed to occur roughly at a linear rate with respect to time, with certain exceptions.

The amino acid sequences of the globins are shown in Table X. The numbering is arranged to include the gaps, so that homologous sites have identical numbers. The numbers are underscored to distinguish them from those used in numbering systems which are consecutive for each chain and do not assign numbers to gaps.

The globin molecules have the following features in common: they are single polypeptide chains usually containing 141 to 156 amino acids, and they occur in nature in combination with heme, a ferroprotoporphyrin. The heme group lies within a crevice of the folded and convoluted globin molecule, much of which consists of right-handed α-helical conformations of the general type described by Pauling and Corey (1951). The α helices are formed so that one helical turn averages 3.6 amino acid residues in length. The nonhelical regions lie on the surface of the globin molecule, and polar residues are almost entirely excluded from the interior sites, at which the α helices make van der Waals contacts. Perutz and co-workers (1965) stated that replacements of a nonpolar interior site by an amino acid with a polar side chain are probably lethal and such replacements have not been noted in any of the hemoglobin variants. However, β-chain mutants, Val to Glu, at sites 69 and 126 are known.

The globin molecules have become extensively differentiated from each other during evolution, and only seven sites have remained constant. These are marked by asterisks in Table X. It is concluded that these sites are occupied by amino acid residues whose identity is essential to the function of the globin molecule. Their functions were listed by Perutz (1965) as follows: proline at 38 is at the corner between the B and C helices; phenylalanine at 44, histidines at 65 and 94, and leucine at 90 are linked to heme; and tyrosine at 147 is hydrogen-bonded with isoleucine or valine at 100. The role of lysine at 134 is unknown. Mutant hemoglobins containing changes at 44, 65, 94, 134 and 147 are known, but these may be deleterious changes that interfere with the role of the molecule, and in some cases this is apparent.

The extensive changes in the primary structure that have taken place in the globin molecules during evolution have not altered its essential properties of oxygen transport, solubility, etc. The secondary and tertiary structures have remained unchanged. It almost seems that every amino acid residue in the sequence has been subjected to changes to find out which molecules could be retained and which discarded! An examination of the distribution of the changes, however, indicates that they could be the result of the random occurrence of point mutations, a certain proportion of which have passed into the genome. The evolutionary

TABLE X

AMINO ACID RESIDUES IN GLOBIN POLYPEPTIDE CHAINS[a,b,c]

Helix	Site Number	Myoglobin	Hemoglobins			Others
			α	β	γ	
	1	NH_2Gly w-NH_2Val	NH_2Val ca-AcNHSer	NH_2Val c-AcNHVal;	NH_2Gly	
	2	—	—	His; c-Thr h-Gln; le-Leu; s, b-NH_2Met	His	βmu-Tyr; δmu-Arg; x-Asp
	3	Leu	Leu	Leu	Phe	
A	4	Ser	Ser;	Thr; h, r, b, le, Ser	Thr	
	5	Glu; hw Asp	Pro; h, s, la-Ala ca-Asp; ms-Gly	Pro; h-Gly; p, b, s-Ala; r-Ser ms-Asp	Glu	
	6	Gly; hw-Ala	Ala; ca-Lys ms-Glu	Glu; sC-NH_2 Pro	Glu	βmu-Val, Lys; αmu-Asp γmu-Lys
	7	Glu	Asp; la-Glu ca-Lys	Glu; sC-Asn ms-Ala	Asp	βmu-Lys, Gly
	8	Trp	Lys	Lys; le-Asp	Lys	
	9	Gln	Thr; ms, mo, s-Ser ca-Ala; b-Gly	Ser; h, ms, s, b-Ala mo-Asn	Ala	δ-Thr βmu-Cys
	10	Gap; w-Leu	Asn; la-Lys; ca-Ala	Ala; le-His; sC-Leu	Thr	
	11	Val	Val; la, ms, r, -Ile	Val; sC-Ile	Ile	
	12	Leu	Lys; la-Arg	Thr; h-Leu; Il-Asx; ms-Ser	Thr	x-Gly, δ-Asn
	13	Asn; w-His	Ala; la-Ser; r-Thr; ca-Ile	Ala; s-Gly; b-Ser; ms-Cys; mo-Thr	Ser	αmu-Asp
	14	Val; hw-Ile	Ala	Leu; s-Phe	Leu	βmu-Arg
	15	Trp	Trp	Trp, b-Phe	Trp	
	16	Gly; w-Ala	Gly; h-Ser; r-Glu; la, ca-Ala	Gly; h-Asp; sC-Ser, b-Ala	Gly	αmu; βmu-Asp; β, δmu-Arg; x-Lys
	17	Lys	Lys; le-Asp; la-Pro	Lys	Lys	αmu-Glu
	18	Val	Val; ms, r, ca-Ile	Val	Val	x-Leu
A	19	Glu	Gly; la-Tyr; ca-Ser	—	—	
AB	20	Pro; w-Ala	Ala; h, mo, ms-Gly;	—	—	

(*continued*)

TABLE X—*Continued*

Helix	Site Number	Myoglobin	Hemoglobins α	Hemoglobins β	Hemoglobins γ	Others
	20 (*cont'd*)		la, r-Ser; ca-Pro			
B	21	Asp	His; la-Asn; s-Asp; ca-Lys	Asn, s, b-Lys; sf-Gln; bB-His	Asn	
	22	Ile; w-Val	Ala; la-Tyr; r-Gly	Val; h-Glu	Val	
	23	Ala	Gly; mo, ca-Asp; la-Glu; b-Ala	Asp; h-Glu	Glu	αmu-Asp
	24	Gly	Glu; g, ca-Asp; la-Thr; s-Gly	Glu	Asp	αmu-Gln; δ-Ala; δmu-Glu; βmu-Lys, Ala
	25	His	Tyr; le-Thr; la-Ser; hV-Phe; ca-Ile; p-Gly	Val	Ala	βmu-gap
	26	Gly	Gly; NBms-Val	Gly	Gly	
	27	Gln	Ala; la-Val	Gly; s-Ala	Gly	βmu-Arg
	28	Glx; w-Asp	Glu; la-Asp	Glu	Glu	βmu-Lys
	29	Val; w-Ile	Ala; le-Ser; la-Ile	Ala	Thr	
	30	Leu	Leu; rV-Val	Leu	Leu	βmu-Pro
	31	Ile	Glu; la-Val; ca-Gly	Gly	Gly	αmu-Gln
	32	Arg	Arg; la-Lys	Arg	Arg	βmu-Ser
	33	Leu	Met; la-Phe	Leu	Leu	
	34	Phe	Phe; ca-Leu	Leu	Leu	
	35	Lys	Leu; la, ca-Thr; ms-Ala	Val; mo-Leu	Val	
B	36	Gly; w-Ser	Ser; h, r-Gly; ca-Val	Val	Val	
C	37	His	Phe; la-Thr; ca-Tyr	Tyr	Tyr	βmu-Phe
	38*	Pro	Pro	Pro	Pro	
	39	Glu	Thr; la-Ala; ca-Gln	Trp	Trp	
	40	Thr	Thr; la-Ala	Thr	Thr	
	41	Leu	Lys; la-Gln	Gln; Il-Arg	Gln	
	42	Glu	Thr; la-Glu	Arg	Arg	
C	43	Lys	Tyr; la-Phe	Phe	Phe	
CD	44*	Phe	Phe	Phe	Phe	αmu-Val; βmu-Ser

TABLE X—*Continued*

Helix	Site Number	Myoglobin	Hemoglobins			Others
			α	β	γ	
	45	Asp	Pro; ca-Ala	Glu, h-Asp	Asp	βmu-Ala
	46	Lys; w-Arg	His; la-Lys	Ser, s-His	Ser	
	47	Phe	Phe; ca-Trp	Phe; mo-Leu	Phe	
	48	Lys	-; ca-Ala; la-Lys	Gly	Gly	βmu-Glu
	49	His	Asp; la:Gly	Asp	Asn	αmu-Gly, His; βmu-Asn
	50	Leu	Leu; la-Met; ms-Val; rV-Phe	Leu	Leu	
CD	51	Lys	Ser; la, r-Thr	Ser; mo-Glu	Ser	
D	52	Ser; w-Thr	His; la-Ser; ca-Pro	Thr; h-Gly; mo, r, b, sA-Ser; sB-Asn	Ser	δ-Ser; αmu-Asp
	53	Glu la-Ala	—	Pro; r, s, b-Ala	Ala	
	54	Asp; w-Ala	—	Asp; mo-Glu; r-Asn, rV-His	Ser	
	55	Glu	—	Ala	Ala	
	56	Met; la-Leu	—	Val; b-Ile	Ile	
	57	Lys	—	Met; sC, b-Leu	Met	
D	58	Ala	Gly; la-Lys	Gly; r, sA, B, Asn; rV-Ser	Gly	βmu-Asp; αmu-Arg
E	59	Ser	Ser	Asn	Asn	
	60	Glu;	Ala; ca, le-Gly; r-Glu	Pro; sC-Ala	Pro	βmu-Arg
	61	Asp	Gln; ca-Pro; la-Asp	Lys	Lys	αmu-Arg, Glu
	62	Leu	Val; r-Ile	Val	Val	x-Phe
	63	Lys	Lys; la-Arg	Lys	Lys	βmu-Glu; Asn
	64	Lys	Gly; la-Trp; ou-Asp; h, r-Ala; ca-Gap	Ala	Ala	αmu-Asp
	65*	His	His	His	His	αmu-Tyr; βmu-Tyr, Arg
	66	Gly	Gly; la-Ala	Gly; r-Ala	Gly	
	67	Ala; w-Val	Lys; la, s, b-Glu; hV-Gln	Lys; Il-Thr	Lys	
	68	Thr	Lys; la-Arg	Lys	Lys	x-Asn
	69	Val	Val; la-NBms-Ile	Val	Val	βmu-Glu, Ala
	70	Leu	Ala; la, ca-Ile; r-Ser	Leu	Leu	
	71	Thr	Asp; r-Glu; la-	Gly; mo-Glu;	Thr	βmu-Asp

(*continued*)

TABLE X—*Continued*

Helix	Site Number	Myoglobin	Hemoglobins α	Hemoglobins β	Hemoglobins γ	Others
	71 (*cont'd*)		Asn; s-Ala; ca-Gly	h-His; le-Ser; r-Ala; b-Asp		
	72	Ala	Ala; h-Gly	Ala; h, s, b-Ser	Ser	x-Glu, Asp
	73	Leu	Leu; la, ca-Val	Phe	Leu	
	74	Gly	Thr; la-Asn; ms-Ala; ca-Gly	Ser; h-Gly; b-Cys	Gly	
	75	Gly, w-Ala	Asn; h, mo-Leu; la, ca-Asp; NBms-Ser; r, s-Lys; mV-Thr	Asp; h, b, r-Glu	Asp	αmu-Lys, Asp; βmu-Asn
	76	Ile	Ala	Gly; mo-Glu; Asp	Ala	
	77	Leu	Val; ms-Gly; rV-Thr	Leu; mo, h, sA-Val; Thr, sB-Met	Ile	
E	78	Lys	Ala; r, h-Gly; s-Asp; ca-Ser; b-Glu	Ala; mo, sA-Glu; Pro; h-His; b, sB, p-Lys; r-Ser; mo, rV-Asn	Lys	βmu-Glu
EF	79	Lys	His; la-Ser; ca-Lys	His, b-Gln	His	βmu-Asp
	80	Lys	Val; h, r, s, ms-Leu; la-Met, ca-Ile	Leu	Leu	
	81	Gly	Asp	Asp	Asp	βmu-Asn
	82	His	Asp	Asn; b-Asp	Asp	
	83	His	Met; h, s, ms, r, ca-Leu; la-Thr; rV-Val	Leu	Leu	
	84	Glu	Pro; la-Glu; ca-Val	Lys	Lys	
	85	Ala	Asn; h, r, s, ms, ca-Gly; la-Lys	Gly	Gly	
	85.5	—	la-Met	—	—	
EF	86	Glu	Ala; la-Ser; ca, s-Gly	Thr; b-Ala	Thr	
F	87	Ile; w-Leu	Leu; la-Met; rV-Ser	Phe; ll-Tyr	Phe	αmu-Arg
	88	Lys	Ser; la-Lys; ca-Ala	Ala	Ala	δ-Ser

TABLE X—*Continued*

Helix	Site Number	Myoglobin	Hemoglobins			Others
			α	β	γ	
	89	Pro	Ala; h-Asn; la-Asp; r, s-Thr; ca-Ser	Thr; p, r-Lys; h-Ala; s-Gln; b-Ser	Gln	βmu-Lys, δ-Gln
	90*	Leu	Leu	Leu	Leu	βmu-Pro
	91	Ala	Ser	Ser	Ser	αmu-Arg
	92	Gln	Asp; la-Gly; ca-Glu	Glu; ll-Gln	Gln	αmu-Asn; βmu-Lys
	93	Ser	Leu; la-Lys	Leu	Leu	βmu-Pro
	94*	His	His	His	His	αmu, βmu-Tyr
F	95	Ala	Ala	Cys	Cys	βmu-gap
FG	96	Thr	His; la-Lys; ca-Ser	Asp; le-Val	Asp	βmu-Asn, gap
	97	Lys	Lys; la-Ser	Lys; p-Glu	Lys	αmu-Asn; βmu-Glu, gap
	98	His	Leu; la-Phe	Leu	Leu	βmu-gap
	99	Lys	Arg; la-Gln	His	His	αmu-Leu, Gln; βmu-gap
FG	100	Val; w-Ile	Val	Val	Val	βmu-Met
G	101	Pro	Asp	Asp	Asp	βmu-His, Asn
	102	Ile	Pro	Pro	Pro	
	103	Lys	Val; la-Gln; ca-Ala	Glu	Glu	
	104	Tyr	Asn; la-Tyr	Asn	Asn	x-Asp, βmu-Thr
	105	Leu	Phe	Phe	Phe	
	106	Glu	Lys	Arg; g-Leu; sB, mo-Lys	Lys	
	107	Phe	Leu, ca-Ile	Leu	Leu	
	108	Ile	Leu	Leu	Leu	x-Val
	109	Ser	Ser; ca-Ala	Gly	Gly	αmu-Arg
	110	Glu	His; ca-Asn	Asn	Asn	
	111	Ala	Cys; ca-His	Val	Val	
	112	Ile	Leu; ca-Ile	Leu	Leu	
	113	Val; w-Ile	Leu; ca-Val	Val; h-Ala; le-Ser	Val	
	114	Asp; w-His	Val; h-Ser	Cys; h-Leu; le-Asp; r-Ile; s, b-Val	Thr	
	115	Val	Thr; ca-Gly	Val	Val	βmu, x-Glu
	116	Leu	Leu; ca-Ile	Leu; h-Val	Leu	x-Ser
	117	Glu; w-His	Ala; ca-Met	Ala; r-Ser	Ala	
	118	Ser	Ala; h-Val; r, ms-Ser; ca-Phe; rV-Asn	His; h, s, b-Arg	Ile	δ-Arg
G	119	Lys; w-Arg	His; ca-Tyr	His, b-Arg	His	δ-Asn; αmu-Gln

(*continued*)

TABLE X—*Continued*

Helix	Site Number	Myoglobin	Hemoglobins			Others
			α	β	γ	
GH	120	His	Leu; ms-His; r-Val; His	Phe, sA, B-His	Phe	
	121	Pro	Pro	Gly	Gly	αmu-Arg
	122	Gly	Ala; r, h-Asn; ca-Gly; r-Ser	Lys; b, sA-Ser; sB-Asn	Lys	x-His; αmu-Asp; βmu-Glu
	123	Asn	Glu; h, ca, ms-Asp	Glu; h-Asp	Glu	αmu, βmu-Lys, Gln; γmu-Lys
GH	124	Phe	Phe	Phe	Phe	x-Lys
H	125	Gly	Thr; ca-Pro	Thr, b-Ser	Thr	x-Asp
	126	Ala	Pro	Pro	Pro	
	127	Asp	Ala; ca-Glu	Pro; h, b-Glu; mo, sC, r-Gln; sA, B-Val	Glu	δ-Glu
	128	Ala	Val	Val; h, b-Leu	Val	δ-Met, x-Phe βmu-Glu
	129	Gln	His	Gln	Gln	
	130	Gly	Ala; ca-Met	Ala	Ala	
	131	Ala	Ser	Ala; h, b-Ser; sA, C-Glu; sB-Asp	Ser	
	132	Met	Leu; ca-Val	Tyr; s, b-Phe	Trp	βmu-Asp
	133	Asn	Asp	Gln	Gln	
	134*	Lys	Lys	Lys	Lys	βmu-Gln
	135	Ala	Phe	Val	Met	
	136	Leu	Leu; ca-Phe	Val	Val	
	137	Glu	Ala; h-Ser; ca-Gln	Ala; b-Thr	Thr	
	138	Leu	Ser; r-Asp; ca-Asn	Gly	Gly	βmu, δmu-Asp
	139	Phe	Val; ca-Leu	Val	Val	
	140	Arg	Ser; ca-Ala	Ala	Ala	
	141	Lys	Thr; ca-Leu	Asn; le-Asp; sC-Ser	Ser	
	142	Asp	Val; ca-Ala	Ala	Ala	
	143	Met; w-Ile	Leu	Leu	Leu	αmu-Pro
	144	Ala	Thr; la-Arg; ca-Ser	Ala	Ser	
	145	Ser; w-Ala	Ser; ca-Glu	His	Ser	βmu-Asp
	146	Asp; w-Lys	Lys; la-Ala	Lys; b, sA, sC-Arg	Arg	
	147*	Tyr	Tyr; la-TyrCOOH	Tyr	Tyr	βmu-His
	148	Lys	Arg-COOH	His-COOH	His-COOH	αmu-gap, Pro

TABLE X—*Continued*

Helix	Site Number	Myoglobin	Hemoglobins α	Hemoglobins β	Hemoglobins γ	Others
H	149	Glu	—	—	—	
HC	150	Leu	—	—	—	
	151	Gly	—	—	—	
	152	Phe; w-Tyr	—	—	—	
	153	Gln	—	—	—	
	154	Gly-COOH	—	—	—	

Lamprey hemoglobin has the following N-terminal sequence preceding the N-terminal amino acid that is present in other globins:

	−9		−5		−1
NH	Pro-Ile-Val-Asp-	Ser-	Gly,	Ser,	Ala, Pro

[a] Abbreviations: b = bovine; bB = B variant of bovine β chain; c = chicken; ca = carp; g = gorilla; h = horse; la = lamprey; le = lemur; ll = llama; mo = monkey (various); ms = mouse; mu = human mutants; NBms = NB mouse; ou = orangutan; p = pig; s = sheep; sA, sB, sC = A, B, and C variants of sheep β chain; sf = sheep fetal; w = sperm whale; hw = humpbacked whale; x = undesignated (Perutz, 1965); δ = human delta chain; r = rabbit; rV = rabbit variants; hV = horse α variant; mV = mouse variant.

[b] Sources of information: Babin *et al.* (1966); Bonaventura and Riggs (1967); Boyer *et al.* (1966); Bradley *et al.* (1967); Braunitzer *et al.* (1961a); Braunitzer and Matsuda (1963); Braunitzer and Hilschmann (1964); Braunitzer and Koehler (1966); Braunitzer (1967); Buettner-Janusch and Hill (1965); Eck and Dayhoff (1966); Edmundson (1965); Galizzi and von Ehrenstein (1967); Goldstein *et al.* (1963); Hill and Konigsberg (1962); Hill *et al.* (1963, 1969); Jones *et al.* (1966); Kilmartin and Clegg (1967); Labie *et al.* (1966); Lehmann *et al.* (1966); Lennox and Cohn (1967); Matsuda *et al.* (1968); Miyaji *et al.* (1968); Perutz (1965); Popp (1965); Rifkin *et al.* (1966); Rudloff *et al.* (1966); Schroeder *et al.* (1963); Schroeder and Jones (1965); von Ehrenstein (1966); Yamaguchi *et al.* (1965); Zuckerkandl and Schroeder (1961).

[c] *Mutations in Human Hemoglobins*—Sources of information: Allan *et al.* (1965); Baglioni and Ingram (1961a,b); Baglioni (1962, 1965); Baglioni and Lehmann (1962); Beale and Lehmann (1965); Blackwell and Liu (1966); Bonaventura and Riggs (1967); Bookchin *et al.* (1966); Botha *et al.* (1966); Bowman *et al.* (1964); Buettner; Janusch and Hill (1965); Carrell *et al.* (1966, 1967); Chernoff and Perrillie (1964)-Clegg *et al.* (1966); Crookston *et al.* (1965); Dacie *et al.* (1967); Gerald and Efron (1961); Gottlieb *et al.* (1963); Hanadu and Rucknagel (1963); Huehns and Shooter (1965); Hill and Schwartz (1959); Hunt and Ingram (1960, 1961); Ingram (1957, 1962); Iwasaki *et al.* (1968); Jones *et al.* (1963, 1964, 1966a,b); Kleihauer *et al.* (1968); Krause (1965); Krause *et al.* (1966); Lehmann *et al.* (1964); Lehmann and Carrell (1969); Liddell *et al.* (1964); Lisker *et al.* (1963); Minnich *et al.* (1965); Miyaji *et al.* (1963, 1966, 1968a,b); Muller and Kingma (1961); Munkres and Richards (1965); Nakajima *et al.* (1963); Pierre *et al.* (1963); Reynolds and Huisman (1966); Salomon *et al.* (1965); Sansome *et al.* (1967); Schneider *et al.* (1964); Schneider and Jones (1965); Shibata *et al.* (1963, 1964); Shim and Bearn (1964); Stamatoyannopoulos *et al.* (1968); Swenson *et al.* (1962); Tuchinda *et al.* (1965); Watson-Williams *et al.* (1965); Yoshida (1967).

TABLE XI
DISTRIBUTION OF AMINO ACID CHANGES IN 148 SITES COMPARED IN GLOBIN CHAINS

Changes per site	0	1	2	3	4	5	6	7	8	9
No. of sites having listed number of changes	7	21	23	33	29	20	7	5	2	1
Minus 3 invariable sites	4	21	23	33	29	20	7	5	2	1
Poisson distribution for $m = 3.5$	4.4	15.3	26.8	31.2	27.4	19.2	11.1	5.6	2.5	1.0

changes have taken place at a rate which is correlated with the passage of time, although whether or not the relationship is linear is not known.

The amino acid interchanges that differentiate the globin molecules from each other are restricted in certain sites to interchanges between amino acids with hydrophobic side chains, such as alanine, leucine, isoleucine, valine, methionine, phenylanine, and tryptophan. Such sites are in the regions of α helices that face the interior of the globin molecule, as pointed out by Perutz (1965). Despite this restriction, the distribution of amino acid interchanges among the various globin chains shows distinct evidence of randomness (Table XI).

Two cistrons may differentiate if they are separated by gene duplication, which will place them in different chromosomal regions. Each of the two will then be subjected to a different pattern of point mutations as a result of the randomness of the mutational process. Differentiation will also occur if two cistrons are separated by speciation rather than by duplication, again because each of the two will undergo a different random sequence of point mutations, and the two are no longer part of a common genetic pool which would merge them into a single type. These considerations suggest that two homologous globins, e.g., cattle and sheep β hemoglobin chains, differ from each other to an extent that is proportional to the time of evolutionary separation of the two species and is unrelated to any difference in physiological requirements between cattle and sheep. This phenomenon, the differentiation of two pieces of DNA which were identical prior to duplication or to separation by speciation, will be termed *allogenic differentiation*. It is a molecular phenomenon and probably results primarily from errors made by DNA polymerase during replication.

This line of reasoning does not exclude the possibility that an event which produces an advantage may appear during the differentiation of two protein molecules. This advantage may be great enough to ensure a step forward in evolution. Evidently, this happened after the cistron

for the α chain duplicated so that differentiation of the duplicate α cistron was initiated. The differentiation by random point mutations eventually produced a new globin, the γ chain, which had the property of forming dimers that associated with α_2 dimers to form tetramers, $\alpha_2\gamma_2$. The tetrameric molecule had properties of oxygen transport which were markedly superior to those of monomers. The function of hemoglobin is to carry oxygen from the lungs, where the concentration of oxygen is high, to tissues, such as the muscles, where oxygen is consumed rapidly by metabolic processes. An improvement in oxygen transport and transfer can improve the efficiency of the muscles and hence enable animals to move more rapidly. The new tetrameric hemoglobin must have greatly accelerated the evolution of vertebrates. The change illustrates the far-reaching possibilities of the effects produced by a molecular event that affects a single protein. The formation of the tetramer from two dimers depends upon the matching of α_2 and γ_2 dimers to produce noncovalent bonds. The tetramer differs from the monomeric myoglobins in having a sigmoid oxygen dissociation curve, so that oxygen is liberated readily at low oxygen tensions such as those encountered in internal tissues where foodstuffs are undergoing rapid metabolism.

Only 21 of the 147 comparable amino acid residues in myoglobin are identical with the corresponding residues in human α, β, and γ hemoglobin chains. The similarity of the myoglobin to the three hemoglobin chains is more evident when the comparison is made on the basis of minimum base differences per codon (Table XII). In terms of MBDC, the comparisons of myoglobin: α chain; myoglobin: β chain; and myoglobin: γ chain, have values of 1.10, 1.14, and 1.11, respectively, as compared with a random value of about 1.41 for nonhomologous sequences in globin chains (Cantor and Jukes, 1966a). The similarity on the basis of MBDC is thus clearly significant.

This comparison of myoglobin with the α, β, and γ hemoglobin chains affords a fine example of divergent evolution, for, although the three comparisons, myoglobin: α, myoglobin: β, and myoglobin: γ, differ over a range of only 0.04 MBDC, the α and β chains differ by 0.70 and the α and γ chains by 0.76 MBDC, while the β and γ chains differ by 0.34 MBDC. Another way of expressing this is as follows: there are only 21 sites at which myoglobin and all three hemoglobin chains are identical, but myoglobin shares 35 or 36 identical sites with any of the three hemoglobin chains taken singly. The α chain is identical with the β and γ chains at 40 sites, but it shares 64 identical sites with the β chain and 59 with the γ chain. Evidently allogenic differentiation has proceeded in all four cistrons at roughly the same rates.

Ingram (1963) pointed out that the α chain forms a dimer, α_2, which

TABLE XII

MINIMUM DIFFERENCES PER CODON IN THE RELATIONSHIP BETWEEN HUMAN α, β, γ, AND δ HEMOGLOBIN CHAINS AND MYOGLOBIN, BETWEEN HUMAN AND HORSE α AND β HEMOGLOBIN CHAINS, HUMAN AND CARP α HEMOGLOBIN CHAINS, AND HUMAN AND RHESUS MONKEY β CHAINS

Comparison	Minimum differences per codon					
	Sites compared	None	1	2	3	Average
Myoglobin:α human	139	38	56	44	1	1.06
Myoglobin:β human	143	37	56	48	2	1.09
Myoglobin:γ human	143	37	59	46	1	1.08
α human:β human	139	64	53	22	0	0.70
α human:γ human	139	59	55	25	0	0.76
α human:α carp	140	73	42	25	0	0.66
β human:γ human	146	106	29	10	0	0.34
α human:α horse	142	126	12	5	0	0.16
β human:β horse	146	120	18	8	0	0.23
β human:δ human	146	136	9	1	0	0.08
β human:β rhesus monkey	146	139	6	1	0	0.05

must fit with either of two different partners, β_2 and γ_2, so that less variation is permitted in the α chain than in the β or γ chains. However, Ingram also added the proviso that the "conservation" of the α chain may be more apparent than real and may be confined to the N-terminal sequence. The contacts between the α and β chains were listed by Perutz (1965) as shown in Table XIII, and 14 α chain residues are involved, only 8 of which are invariable in different α chains. Therefore, if the residues in contact are the ones which are conserved, the requirements for a tetrameric structure restrict only a small amount of the α chain from undergoing changes.

Variations in the α chain of rabbit hemoglobin were mentioned on page 33. Similar variations were reported for the β chain by Galizzi and von Ehrenstein (1967) who found asparagine and histidine at site 52; and asparagine and serine at 56 and 76. Kilmartin and Clegg (1967) found heterogeneity in horse α hemoglobin; phenylalanine or tyrosine at site 24; and glutamine or lysine at site 60; and Rifkin *et al.* (1966) reported serine/threonine heterogeneity at site 68 in the α chain of strain SEC mice.

1. Internal Homology in the Globins

When internal repetition occurs in polypeptide chains, it may be possible to use the repetition as a measure of the extent to which the chain

TABLE XIII
RESIDUES AT CONTACTS BETWEEN UNLIKE HEMOGLOBIN CHAINS[a]

Possible nonpolar contact between residues in				Residues in the region of contact between			
α_1 and β_1				α_1 and β_2			
						Pro	38
		Pro or Ala	53	Thr	39	Trp	39
Pro	126	Met	57	Thr	42	Arg	42
		Val	35	Tyr	43	Phe	43
Val or Ala	118	Gly	121	Arg	99	His	99
				Val or Ala	103	Glu	103
Ala	117	Arg or His	118				
Phe	124	Ala	117				
Ser or Val	114						
Arg	32	Phe	124				
Ser or Val	114						
Leu or Ala	35	Pro	126				
His	129	Val	36				
		Arg	42				
Possible polar contacts							
Asp	133	Tyr	37				
Phe (CO)	124	Arg or His	118				

[a] From Perutz (1965).

has been conserved in the repeated regions. Such repetitions are the vestiges of recombinational events by which protein molecules were formed from duplications of shorter polypeptide chains (Cantor and Jukes, 1966a).

A repeating sequence, separated by 66 residues, was detected in the α and β chains of hemoglobin by Fitch (1966b) who constructed a table of minimum base differences per codon. He used a computer to scan the amino acid sequences of the hemoglobins after converting them into base differences. This procedure showed the presence of a repeating sequence as two zones in which the MBDC value was markedly lower than those obtained by comparing any two regions which were separated by more or less than 66 residues. Cantor and Jukes (1966a) proposed that the repeating sequences contained 21 residues and that the first of the two sequences was preceded by a partial repetition. The findings are shown in Table XIV. A gap between residues 52 and 53 is necessary to show homol-

TABLE XIV

COMPARISONS OF SEQUENCES IN POLYPEPTIDE CHAINS OF CERTAIN GLOBINS[a]

												50					53				58
												Leu-	Lys-	Ser-			-Glu-	Asp-	Glu-	Met-	Lys-Ala
M	59																				79
	Ser-	Glu-	Asp-	Leu-	Lys-	Lys-	His-	Gly-	Val-	Thr-	Val-	Leu-	Thr-	Ala-	Leu-	Gly-	Ala-	Ile-	Leu-	Lys-	Lys
	125																				145
	Gly-	Ala-	Asp-	Ala-	Gln-	Gly-	Ala-	Met-	Asn-	Lys-	Ala-	Ser-	Glu-	Leu-	Phe-	Arg-	Lys-	Asp-	Met-	Ala-	Ser
												50									58
												Leu-	Ser-	His-	-	-	-	-	-	-	-Gly
α	59																				79
	Ser-	Ala-	Gln-	Val-	Lys-	Gly-	His-	Gly-	Lys-	Lys-	Val-	Ala-	Asp-	Ala-	Leu-	Thr-	Asn-	Ala-	Val-	Ala-	His
	125																				145
	Thr-	Pro-	Ala-	Val-	His-	Ala-	Ser-	Leu-	Asp-	Lys-	Phe-	Leu-	Ala-	Ser-	Val-	Ser-	Thr-	Val-	Leu-	Thr-	Ser
												50									58
												Leu-	Ser-	Thr-			-Pro-	Asp-	Ala-	Val-	Met-Gly
β	59																				79
	Asn-	Pro-	Lys-	Val-	Lys-	Ala-	His-	Gly-	Lys-	Lys-	Val-	Leu-	Gly-	Ala-	Phe-	Ser-	Asp-	Gly-	Leu-	Ala-	His
	125																				145
	Thr-	Pro-	Pro-	Val-	Gln-	Ala-	Ala-	Tyr-	Gln-	Lys-	Val-	Val-	Ala-	Gly-	Val-	Ala-	Asn-	Ala-	Leu-	Ala-	His
												50									58
												Leu-	Ser-	Ser-			-Ala-	Ser-	Ala-	Ile-	Met-Gly
γ	59																				79
	Asn-	Pro-	Lys-	Val-	Lys-	Ala-	His-	Gly-	Lys-	Lys-	Val-	Leu-	Thr-	Ser-	Leu-	Gly-	Asp-	Ala-	Ile-	Lys-	His
	125																				145
	Thr-	Pro-	Glu-	Val-	Gln-	Ala-	Ser-	Trp-	Gln-	Lys-	Met-	Val-	Thr-	Gly-	Val-	Ala-	Ser-	Ala-	Leu-	Ser-	Ser
												50									58
												Met-	Thr-	Ser-			-Ala-	Asp-	Glu-	Leu-	Lys-Lys
Lamprey																					
	59																				79
	Ser-	Ala-	Asp-	Val-	Arg-	Trp-	His-	Ala-	Glu-	Arg-	Ile-	Ile-	Asn-	Ala-	Val-	Asn-	Asp-	Ala-	Val-	Ala-	Ser
Prototype																					
	Ser-	Pro-	Glu-	Val-	Lys-	Ala-	His-	Gly-	Lys-	Lys-	Val-	Leu-	Thr-	Ala-	Val-	Ala-	Asp-	Ala-	Leu-	Ala-	His

[a] M = human myoglobin; α = α chain of hemoglobin A; β = β chain; γ = γ chain; Lamprey = lamprey hemoglobin. The numbering system is identical with that in Table X.

ogy between residues 50–58 and 70–79 or 50–58 and 136–145. The sequence designated as "prototype" in Table XIV is a hypothetical sequence made by selecting the predominant amino acid in each vertical column. The base sequences in codons were used as a means of selecting amino acids from columns in which no amino acid predominated.

Substantial homology in different parts of the molecules is shown when the regions in different parts of the polypeptide chains are compared with each other. The short sequence of nine amino acids consisting of residues 50 to 58 should, on a random basis, contain not more than 1 or 2 identities with sequences 70–79 and 136–145, but there are three comparisons which contain four identical sites. Similarly sequences 59–79 would be expected to contain only 2 or 3 sites identical to corresponding residues in sequences 125–145. Instead, there are five comparisons in which 6 to 8 sites are identical as shown below. The comparisons of all

Comparison			Identities	Comparison			Identities
Myo	50–58:	γ70–79	4	α 59–79:	β125–145		7
Lamprey	50–58:	myo70–79	4	α 125–145:	β59–79		7
γ	50–58:	γ70–79	4	α 125–145:	β59–79		6
				β 59–79:	β125–145		8
				γ 59–79:	β125–145		7
				γ 59–79:	γ125–145		6

regions involved in the repetitive homology are shown in terms of MBDC in Table XV. The repetitive homology is more easily discernible in comparisons involving the β and γ chains than in comparisons which do not include these chains. It would therefore seem probable that, as first noted by Fitch (1966b), the myoglobin and α chains have differentiated more from the archetypal globin sequence than have the β and γ chains. Table XV indicates also that lamprey hemoglobin, as discussed below, is not differentiating more slowly than are the β and γ chains.

It is possible to compare the MBDC values in Table XV with the corresponding values obtained when the sequences are compared with the same rather than different regions of the various globin chains. Thus the average MBDC for myoglobin 59–79 compared with α 59–79 is 1.14 as contrasted with 1.19 when myoglobin 59–79 is compared with α 125–145. The first value expresses the differentiation after duplications of the archetypal globin gene took place (Fig. 2). The second value expresses this differentiation *plus* the differentiation that occurred following the internal repetition and preceding the duplication. Table XVI summarizes the values, grouped as homologous and translocated comparisons. The translocated comparisons show significant homology, 1.10 as contrasted with a random value of 1.41. The homologous regions, of course, show even greater homology, expressing the differentiation after the events of duplication designated as (2) (3) and (5) in Fig. 2.

TABLE XV

COMPARISONS OF SEQUENCES IN DIFFERENT REGIONS OF GLOBIN POLYPEPTIDE CHAINS IN TERMS OF MINIMUM BASE DIFFERENCES BETWEEN AMINO ACID CODONS IN EACH PAIR OF SEQUENCES THAT ARE COMPARED[a]

		Myo			α			β			γ			Lamprey
		1	2	3	1	2	3	1	2	3	1	2	3	1
Myo	2	9												
	3	11	31											
α	1													
	2	14		22										
	3	10	25			26								
β	1		12	13		11	9							
	2	12		25			21	10						
	3	12	29			16		11	16					
γ	1		10	13		12	7		11	9				
	2	6		28			22	8		18	7			
	3	11	26			21		11	20		7	19		
Lamprey	1		6	15		13	9		10	9		6	9	
	2	10		25			23	8		24	11		22	11

[a] Headings 1, 2, and 3 refer, respectively, to sequences of residues 50–58, 59–79, and 125–145, shown in Table XIV. For sequences picked at random, comparisons involving sequences 1 (9 amino acids) should average $9 \times 1.41 = \sim 14$ base differences and comparisons between sequences 2 and 3 (21 amino acids) should average $21 \times 1.41 = \sim 30$ base differences. Comparisons that show marked homology are underlined for emphasis.

TABLE XVI

MBDC VALUES IN COMPARISONS OF SEQUENCES 50–58, 59–79, AND 125–145 IN GLOBIN CHAINS SHOWN IN TABLE XIV

Sequences compared	No. of sites compared	MBDC
Homologous	390	0.87
Translocated	744	1.10

2. Lamprey Hemoglobin

The first 114 residues in the amino acid sequence of a lamprey hemoglobin, which is monomeric, have been placed in order by Rudloff *et al.* (1966). This portion of the chain contains no gaps, and an extra residue is present between sites 85 and 86. The molecule has an N-terminal sequence of nine residues which is not found in the other globins (Table X). The 105 residues that follow this sequence are comparable with the

TABLE XVII
COMPARISON OF RESIDUES 1–105 IN LAMPREY HEMOGLOBIN AND THE FOUR HUMAN GLOBIN CHAINS

Comparison	Sites compared	Minimum base differences per codon				
		None	1	2	3	Average
Lamprey:myoglobin	105	28	46	28	1	1.00
Lamprey:α chain	100	38	40	22	0	0.84
Lamprey:β chain	105	30	45	30	0	1.00
Lamprey:γ chain	105	35	38	32	0	0.97

corresponding regions in other globin chains, as shown in Table XVII. Within the limits of the incomplete comparison, lamprey hemoglobin is more similar to the α chain than to the other three. Comparisons of the corresponding regions *inter alia* give MBDC values of myo:α = 1.11; α:β = 0.66, and α:γ = 0.74.

A comparison of these values with those in Table XII indicates that the separation of lamprey hemoglobin from the α chain occurred prior to the duplication that resulted in the separation of the archetypal γ chain from the α chain. These values provide no indication that lamprey Hb is "more primitive" than the other chains in the sense of having diverged more slowly than these from the archetype. If this were the case, the MBDC values between lamprey Hb and the β and γ chains should be equal to or less than the α:β or α:γ differences, which is not the case.

There is another way of examining the rate of evolution of a globin. This is by comparing the repeating sequences in the chain (Table XV). In a rapidly evolving chain, there will be a greater difference between the two repetitive sections than in the case of a slowly evolving chain. The comparison shows that the β and γ chains are differentiating more slowly than are lamprey Hb, or the α chain, or myoglobin. The primitive morphology of the lamprey therefore is not matched by a slow rate of evolutionary change in its hemoglobin, although it has evidently followed a different evolutionary pathway which did not lead to the formation of a tetrameric hemoglobin. The absence of a tetrameric hemoglobin may have slowed the evolution of the line of inheritance that led to the Cyclostomata, which include the lampreys.

Briehl (1963) has reported that lamprey hemoglobin, which is a monomer when oxygenated, forms aggregates when it is deoxygenated. He suggested that the oxygenation proceeded in two steps, deaggregation

followed by oxygenation. Reaggregation was brought about by deoxygenation, and was favored by decreased pH. This behavior is distinct from that shown by the hemoglobins of the mammalian type, in which the oxygenated and deoxygenated forms are both stable tetramers.

C. The Haptoglobins

Haptoglobin is a globin present in human blood plasma which binds free hemoglobin. This reaction detoxifies hemoglobin when it is set free from erythrocytes during hemolysis. An enzyme present in the liver, heme α-methenyl oxygenase, readily attacks the heme present in haptoglobin-hemoglobin combination and converts heme to a precursor of biliverdin. Haptoglobin contains two polypeptides, α and β; the α polypeptide chains are genetically variable. Several types of α chain are present in human beings.

Type 1α chains contains 83 residues. The chains can be either fast (F) or slow (S), referring to their speed of electrophoretic migration; F and S differ by an interchange between lysine and glutamic acid at position 54 (Black *et al.*, 1967).

Type 2α chains usually consist of type 1αF and type 1αS chains linked by crossing-over and recombination with the elimination of 24 amino acid residues; 13 from the C-terminal end of the fast chain and 11 from the N-terminal end of the slow chain, as shown in Fig. 3. This presents a scheme for recombination in which the following steps might take place: Each cistron breaks. The broken strands are attacked by exonucleases, and short complementary sequences are exposed. These become bonded by Watson-Crick pairing, and the two strands are closed by DNA ligase. The crossing-over region may be depicted as containing →AGCCG and TCGGC← in the two DNA strands. There are two other kinds of type 2α chains, formed by crossing-over between two fast and two slow chains, respectively. All three 2α chains are formed by crossing-over at the same point.

An even larger α chain, the Johnson phenotype, was discovered by Smithies (1965). This is produced by crossing-over between two type 2α chains to form a polypeptide almost three times as long as that of type 1α.

Studies of the world-wide distribution of the haptoglobins (Shim and Bearn, 1964) showed that the 2α polypeptide was present in India, where hemolytic diseases are common, in frequencies as high as 82% in groups of individuals. It has been suggested that the predominance of 2α in India and Southeast Asia is evidence of natural selection, because haptoglobin 2-2, containing two long (2α) chains is more effective than either 1-1 (two short chains) or 2-1 (one long and one short

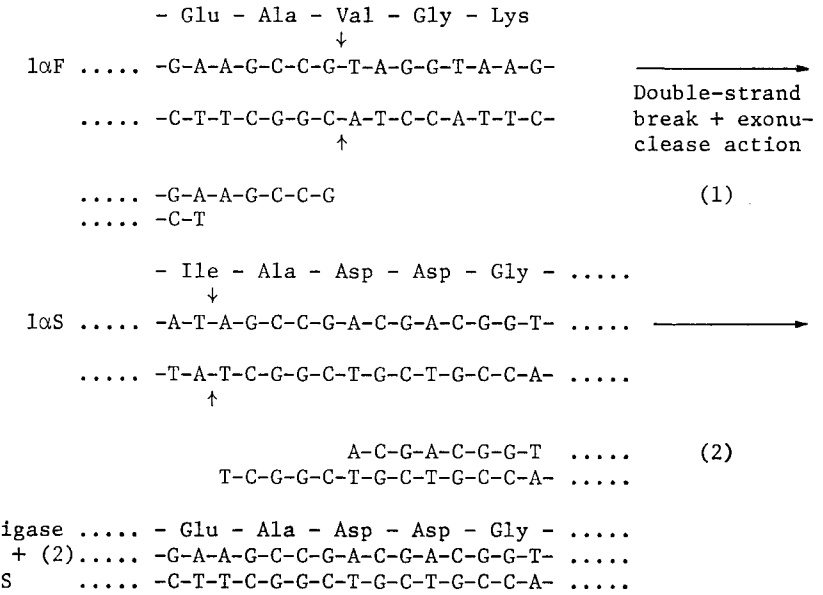


FIG. 3. Hypothetical depiction of cross-over region in strands of cistrons for haptoglobin 1α-peptide chains. The union is postulated as taking place by Watson-Crick pairing between complementary single-stranded regions, and joining the free ends by DNA ligase. CT, chain-termination codon. The amino acid sequences in the region of the junction are described by Black *et al.* (1967).

chain) in facilitating the action of liver heme α-methenyl oxygenase (Nakajima *et al.*, 1963). Therefore, the elongated haptoglobin variants may represent an incident of evolution that is still in active progress. It is interesting to reflect on the possible effects of public health measures. Will such measures change the natural course of human evolution by decreasing the incidence of diseases that would otherwise confer a selective advantage on inherently resistant individuals? Conversely, will the use of DDT in Africa, by reducing the incidence of malaria, remove the advantage (Allison, 1964) conferred by the sickle-cell trait against malaria?

D. Fibrinopeptides

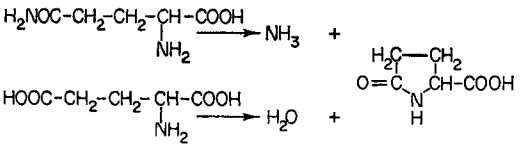
During the clotting of blood, two peptides are released from fibrinogen by the action of thrombin, a proteolytic enzyme that splits fibrinogen at a linkage between arginine and glycine to yield fibrinopeptides:

$$\text{fibrinogen} \xrightarrow{\text{thrombin}} \text{two A-peptides} + \text{two B-peptides} + \text{fibrin monomer}$$

The fibrin monomer forms a clot and the fibrinopeptides, which served to keep fibrinogen in a soluble and nonclottable form, are discarded. The process is reminiscent of the activation of trypsinogen by the liberation of the hexapeptide Val-$(Asp)_4$-Lys from its N-terminal end. If the only function of fibrinopeptides is to stabilize the fibrinogen molecule against coagulation, this may be a function that could be carried out by any of a large number of different polypeptides. In such a case, it would be perhaps not surprising that their composition varies greatly in different animal species. This variation might be explained by allogenic differentiation, unrestrained by conservation of structure.

Fibrinopeptides A and B have lengths of 19 or 21 amino acid residues or less; the length varying in different species (Table XVIII). A great deal of variation exists between the fibrinopeptides as obtained from different animal species. Extensive studies of these peptides have been made by Blombäck and Doolittle (1963; Blombäck *et al.*, 1965, 1966; Doolittle, *et al.*, 1967; Doolittle and Blombäck, 1964), following the first sequential analysis of peptide A from bovine fibrinogen by Folk and co-workers (1959).

The B fibrinopeptides in most cases start with pyrrolidone carboxylic acid (PCA) shown as "Pyr" in Table XVIII. This substance is formed by cyclization of glutamine or glutamic acid as follows:



The presence of PCA at the N-terminus caused difficulties in analyzing the fibrinopeptides B. These were resolved by the use of a new enzyme, pyrollidonyl peptidase, prepared from a strain of *Pseudomonas fluorescens* (Mross and Doolittle, 1967). The tyrosine residue at position 16 in the fibrinopeptides B, counting from the carboxyl end, is sulfated.

The A and B peptides are N-terminal fragments of longer polypeptide chains, present in fibrinogen, which are broken by thrombin as indicated above, so that the C-terminal residues of fibrinopeptides are always arginine. It is likely that partial deletions have been frequent in the fibrinopeptide series; it is not always easy to guess where the deletions should be placed. No attempt has been made to estimate the location of the deletions (Table XVIII), except for the obvious gap in Cape buffalo A.

Mross and Doolittle compared the amino acid sequences of fibrinopeptides A and B from 17 ruminants, many of which were obtained from

the San Diego Zoo. In some cases, the analyses of the peptides were incomplete but enough information was obtained to indicate the probable amino acid sequences in these peptides by comparing them with closely related fibrinopeptides of known sequence. For convenience in comparing the various ruminant fibrinopeptides, "ruminant types" have been added to the list in Table XVIII. These are hypothetical sequences which are written by listing the predominant amino acid at each site, using the genetic code to resolve the choice between amino acids at certain highly variable sites. The ruminant type is predominantly derived from sequences of fibrinopeptides of various deer, so that the fibrinopeptides of individual species of deer deviate less from it than do those of cattle, camels, etc.

Mross and Doolittle (1967) point out the agreement between classical taxonomy and the divergences shown in their study of the fibrinopeptides of ruminants. Comparisons of nonruminant fibrinopeptides with the ruminant type are somewhat difficult because of the problem of locating the gaps (Table XVIII).

Two interpretations are possible of the differences between fibrinopeptides. The first is that expressed by Doolittle (Doolittle and Blombäck, 1964; Doolittle *et al.*, 1967; Mross and Doolittle, 1967). It states that these molecules may be used as taxonomic tools and as models for studying the rates of molecular evolution. Mross and Doolittle carry the concept of the taxonomic significance of the fibrinopeptides to considerable lengths; for example, they use fibrinopeptide comparisons as evidence that the subfamily Bovinae are "very much more different from the subfamily Caprinae (sheep and goats) than classical taxonomy would have led us to believe." It is also true, however, that cattle and sheep have identical cytochromes *c*! Mross and Doolittle (1967) cite the Lys/Glu difference between llama and camels at residue 12 (numbering from the C-terminal residue) in fibrinopeptide A in support of the conclusion that llamas are more primitive than camels, because the pig also has Lys at this locus. Since only five different amino acids are present at this locus in 14 fibrinopeptides A, it is obvious that the coincidence of Lys at this site in the fibrinopeptides A of pig and llama may well be due to chance.

Mross and Doolittle express such a viewpoint as follows: ". . . are we to assume that threonine is a better amino acid at position A7 for ox and muntjak, but alanine is better for the other artiodactyls? Or are we to ascribe a certain amount of this change to accident and 'genetic drift'?"

The second interpretation of the differences between fibrinopeptides, which is preferred by us, is that they result principally from allogenic

TABLE XVIII

AMINO ACID SEQUENCES OF FIBRINOPEPTIDES A AND B IN VARIOUS MAMMALS[a]

Fibrinopeptides A:	
Human	NH_2-Ala-Asp-Ser-Gly-Gly-Gly-Asp-Phe-Leu-Ala-Glu-Gly-Gly-Gly-Val-Arg-COOH
Green and Rhesus monkeys	Ala-Asp-Thr-Gly-Glu-Gly-Asp-Phe-Leu-Ala-Glu-Gly-Gly-Gly-Val-Arg
Horse	Thr-Glu-Glu-Gly-Glu-Phe-Leu-His-Glu-Gly-Gly-Gly-Val-Arg
Donkey	Thr-Lys-Thr-Glu-Glu-Gly-Glu-Phe-Ile-Ser-Glu-Gly-Gly-Gly-Val-Arg
Dog, fox	Thr-Asn-Ser-Lys-Glu-Gly-Glu-Phe-Ile-Ala-Glu-Gly-Gly-Gly-Val-Arg
Cat	Gly-Asp-Val-Glu-Glu-Gly-Glu-Phe-Ile-Ala-Glu-Gly-Gly-Gly-Val-Arg
Mink	Thr-Asn-Val-Lys-Glu-Ser-Glu-Phe-Ile-Ala-Glu-Gly-Ala-Gly-Ala-Arg
Badger	Thr-Asn-Val-Lys-Glu-Ser-Glu-Phe-Ile-Ala-Glu-Gly-Ala-Val-Gly-Arg
Rabbit	Val-Asp-Pro-Gly-Gly-Thr-Ser-Phe-Leu(Thr,Glu,Gly,Gly)Asp-Ala-Arg
Guinea pig	Thr-Asp-Thr-Glu-Phe-Glu-Ala-Ala-Gly-Gly-Gly-Val-Arg
Rat	Ala-Asp-Thr-Gly-Thr-Thr-Ser-Glu-Phe-Ile-Asp-Glu-Gly-Ala-Gly-Ile-Arg
Pig	Ala-Glu-Val-Gln-Asp-Lys-Gly-Glu-Phe-Leu-Ala-Glu-Gly-Gly-Gly-Val-Arg
Bovine	Glu-Asp-Gly-Ser-Asp-Pro-Pro-Ser-Gly-Asp-Phe-Leu-Thr-Glu-Gly-Gly-Gly-Val-Arg

[a] Sources of information: Doolittle and Blombäck (1964); Doolittle et al. (1967); Mross and Doolittle (1967); Blombäck and Doolittle (1963); Blombäck et al. (1965, 1966).

Fibrinopeptides A (continued):

Bison	Glu-Asp-Gly-Ser-Asp-Pro-Ala-Ser-Gly-Asp-Phe-Leu-Ala-Glu-Gly-Gly-Gly-Val-Arg
Water buffalo	Glu-Asp-Gly-Ser-Asp-Ala-Val-Gly-Gly-Glu-Phe(Leu,Ala,Glu,Gly,Gly,Gly,Val)Arg
Cape buffalo	Glu-Asp-Gly-Ser- - - - -Gly-Glu-Phe-Leu(Ala,Glu,Gly,Gly,Gly,Val)Arg
Sheep and goat	Ala-Asp-Asp-Ser-Asp-Pro-Val-Gly-Gly-Glu-Phe-Leu-Ala-Glu-Gly-Gly-Gly-Val-Arg
Red deer, elk	Ala-Asp-Gly-Ser-Asp-Pro-Ala-Ser-Ser-Asp-Phe-Leu-Ala-Glu-Gly-Gly-Gly-Val-Arg
Sika deer	Ala-Asp-Gly-Ser-Asp-Pro-Ala-Ser-Ser-Glu-Phe-Leu-Ala-Glu-Gly-Gly-Gly-Val-Arg
Muntjak	(Ala,Asp,Gly,Ser,Asp,Pro,Ala,Ser,Gly,Glu)Phe(Leu,Thr,Glu,Gly,Gly,Gly,Val)Arg
Reindeer	Ala-Asp-Gly-Ser-Asp-Pro-Ala-Gly-Gly-Glu-Phe(Leu,Ala,Glu,Gly,Gly,Gly,Val)Arg
Pronghorn	Ala-Asp-Gly-Ser-Asp-Pro-Ala-Gly-Gly-Glu-Ser(Leu,Pro,Asp,Gly,Thr,Gly,Ala)Arg
Mule deer	Ser-Asp-Pro-Ala-Gly-Gly-Glu-Phe(Leu,Ala,Glu,Gly,Gly,Gly,Val)Arg
Llama, vicuna	Thr-Asp-Pro-Asp-Ala-Asp-Lys-Gly-Glu-Phe-Leu-Ala-Glu-Gly-Gly-Gly-Val-Arg
Camel	Thr-Asp-Pro-Asp-Ala-Asp-Glu-Gly-Glu-Phe(Leu,Ala,Glu,Gly,Gly,Gly,Val)Arg
Ruminant type	Ala-Asp-Gly-Ser-Asp-Pro-Ala-Ser-Gly-Glu-Phe-Leu-Ala-Glu-Gly-Gly-Gly-Val-Arg

(continued)

TABLE XVIII—(*Continued*)

Fibrinopeptides B:	
Human	Pyr-Gly-Val-Asn-Asp-Asn-Glu-Glu- - - - - - -Gly-Phe-Phe-Ser-Ala-Arg-COOH
Green monkey	Pyr-Gly-Val-Gly-Asp-Asn-Glu-Glu- - - - - - -Gly-Leu-Phe-Gly-Gly-Arg
Rhesus monkey	Asn-Glu-Glu- - - - - - -Ser-Pro-Phe-Ser-Gly-Arg
Dog	His-Tyr-Tyr-Asp-Asp-Thr-Asp-Glu-Glu-Glu-Arg-Ile-Val-Ser-Thr-Val-Asp-Ala-Arg
Fox	Glu-Tyr-Tyr-Asp-Asp-Thr-Asp-Glu-Glu-Glu-Arg-Ile-Val-Ser-Thr-Val-Asp-Ala-Arg
Cat	Ile-Ile-Asp-Tyr-Tyr-Asp-Glu-Gly-Glu-Glu-Asp-Arg-Asp-Val-Gly-Val-Val-Asp-Ala-Arg
Pig	Ala-Ile-Asp-Tyr-Asp-Glu-Asp-Glu-Asp-Gly-Arg-Pro-Lys-Val-His-Val-Asp-Ala-Arg
Bovine, bison	Pyr-Phe-Pro-Thr-Asp-Tyr-Asp-Glu-Gly-Gln-Asp-Asp-Arg-Pro-Lys-Val-Gly-Leu-Gly-Ala-Arg
Water buffalo	Pyr(Phe,Pro,Thr)(Asp,Tyr,Asp,Glu,Gly,Gln,Asp,Asp,Arg,Pro,Lys)(Leu,Gly,Leu,Gly,Ala)Arg
Cape buffalo	Pyr(Phe,Pro,Thr,Asp,Tyr,Asp,Glu,Gly,Gln,Asp,Asp,Arg,Pro,Lys)(Ser,Gly,Leu,Gly,Ala)Arg
Sheep, goat	Gly-Tyr-Leu-Asp-Tyr-Asp-Glu-Val-Asp-Asp-Asn-Arg-Ala-Lys-Leu-Pro-Leu-Asp-Ala-Arg
Red deer, elk, sika deer	Pyr-His-Ser-Thr-Asp-Tyr-Asp-Glu-Glu-Glu-Asp-Asp-Arg-Ala-Lys(Leu,His,Leu,Asp,Ala)Arg
Muntjak	Pyr(His,Ser,Thr)(Asp,Tyr,Asp,Glu,Val,Glu,Asp,Asp)Arg-Ala-Lys(Leu,His,Leu,Asp,Ala)Arg
Reindeer	Pyr-His-Leu-Ala-Asp-Tyr-Asp-Glu-Val(Glu,Asp,Asp)Arg-Ala-Lys-Leu-His-Leu-Asp-Ala-Arg
Mule deer	Pyr-His-Leu(Ala,Asp,Tyr,Asp,Glu,Val)Asp-Asp-Asp-Arg-Ala-Lys(Leu,His,Leu)Asp-Ala-Arg
Pronghorn	Pyr(Pro,Ser)(Tyr,Asp,Tyr,Asp,Glu,Glu,Glu,Asp,Asp)Arg-Ala-Lys-Leu-Arg(Leu,Asp,Ala)Arg
Llama, Camel, Vicuna	Ala-Thr-Asp-Tyr-Asp-Glu-Glu-Glu-Asp-Asp-Arg-Val-Lys-Val-Arg-Leu-Asp-Ala-Arg
Ruminant type	Pyr-His-Ser-Thr-Asp-Tyr-Asp-Glu-Glu-Glu-Asp-Asp-Arg-Ala-Lys-Leu-Arg-Leu-Asp-Ala-Arg

differentiation following the separation of species, and that many of the amino acid substitutions are "neutral" changes. Such differences would be roughly proportional to the lengths of time that have elapsed since the separation, as in the case of the hemoglobins and cytochromes *c*, but the proportionality cannot be used as a substitute for classical taxonomy. The sample of the genome contained in the two fibrinopeptide sequences is too small for the purposes of systematics. It is evident that the fibrinopeptides can vary greatly in length and composition. Their function, as suggested above, may well be confined to providing a recognition site for thrombin prior to the hydrolysis of the Arg-Gly linkage to form fibrin.

E. The Immunoglobulins

One of the distinctive characteristics of vertebrate organisms, especially the mammals, is the vast array of different antibodies which appear in response to the parenteral presence of foreign proteins or antigens. Any one of many thousands of antigens can stimulate the formation of a specific antibody protein. The study of the formation and localization of antibodies is a major task of the science of immunology. Only the briefest outline of the general topic of antibodies will be presented here. Antibodies are proteins in the γ-globulin fraction of blood plasma. They combine with specific antigens in the antigen-antibody reaction, which is a major defense mechanism against the injurious effects of proteins formed by invasive pathogenic microorganisms.

Studies on the primary structure of antibody proteins have included an intensive series of investigations into the structure of the immunoglobulins (Cohen and Milstein, 1967; Lennox and Cohn, 1967; Porter, 1967; Putnam and Easiley, 1965; Hood *et al.*, 1968).

It is customary to depict the immunoglobulin molecules from all species as being formed of two light and two heavy chains joined together by disulfide bonds (Fig. 4). Both the light and heavy chains may exist in several forms, thus making possible a large assortment of different immunoglobulin molecules. There are two general types of light chains, termed κ and λ, which are present in each individual. Both of these types possess antibody specificity for many different antigens. The immunoglobulins of human beings have been studied by means of the Bence-Jones proteins, which are present in the urine of about 50% of myeloma patients. The Bence-Jones proteins are light chains of either κ or λ type. There are three general types of heavy chains, termed G, A, and M.

Fragmentation of the immunoglobulin molecules by papain or pepsin takes place at specific sites. The fragmentation has given rise to a nomen-

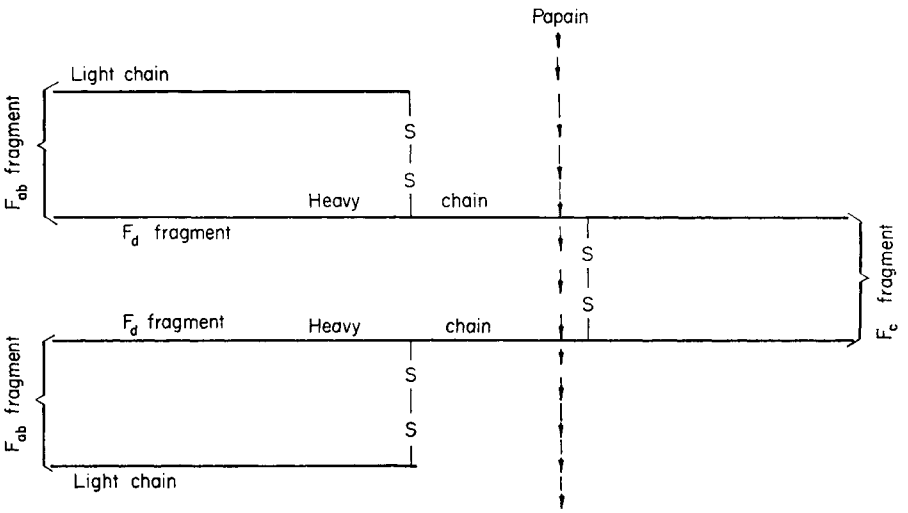


FIG. 4. Diagrammatic representation of immunoglobulin molecule showing nomenclature of various regions.

clature for the fragments as shown in Fig. 4. One papain fragment is termed "F_c" and has a molecular weight of about 48,000. It contains two C-terminal portions of the two heavy chains linked by disulfide bridges. The other two fragments obtained from the action of papain are termed "F_{ab}." Each consists of a light chain in disulfide linkage with a portion (F_d) of heavy chain. F_{ab} has a molecular weight of about 42,000.

One of the most remarkable phenomena in protein chemistry is the fact that the light chain of the immunoglobulins is composed of equal parts of variable and constant regions. The first half of either a κ or λ light chain, containing about 107 amino acid residues, is subject to extensive variations. In contrast, the C-terminal half of the chain, residues 108–214, is quite constant in composition in each species, although it varies from species to species in a manner analogous to, for example, the α chains of hemoglobins in various mammals. To explain the variable and constant portions, Gray (1966) has described the light chain as representing "two genes, one polypeptide chain." Dreyer and Bennett (1965) have presented a theory for explaining the manner in which the joining of the two halves of the light chain takes place. They proposed that the variable portion results from a segment of DNA which pairs with another piece of DNA, coding for the constant portion, during the differentiation

of immunologically competent cells. They suggested that this could take place by a process similar to that encountered when a lysogenic bacteriophage is incorporated into a bacterial chromosome.

The first half of light chains, termed the S (specificity) region, varies greatly in amino acid sequence among the array of light chains found within each species. The sequence variations in a single species are analogous to that seen when examples of gene duplication, such as the α, β, γ, and δ hemoglobins, are compared (Hood *et al.*, 1968). The S regions of κ chains can also be subdivided into two main classes, $S\kappa_I$ and $S\kappa_{II}$, on the basis of the presence or absence of a gap at residues 31 to 34 in both human beings and mice. It is therefore concluded that the $S\kappa$ cistron duplicated prior to the evolutionary separation of mice and men. The $S\kappa_I$ and $S\kappa_{II}$ classes have identical C regions of κ chains; the C regions are species-specific. The mouse and the human being therefore each have two types of κ chain; mouse $S\kappa_I$-$C\kappa$, mouse $S\kappa_{II}$-$C\kappa$, human $S\kappa_I$-$C\kappa$, and human $S\kappa_{II}$-$C\kappa$. This difference clearly implies separate gene duplication for the $S\kappa$ region, but not for the $C\kappa$ region, prior to the evolutionary separation of mice and human beings.

The manner of crossing-over between S and C was the subject of further speculation by Dreyer *et al.* (1967), who suggested a mechanism that included a "splice region" at the beginning of the C cistron, which was postulated to pair with a hypothetical "replisomal anticodon" that participated in the splicing of C and S cistrons.

Dreyer *et al.* (1967) have outlined a proposed evolutionary course of events for the immunoglobins as follows: The first event was the appearance of a mechanism that permitted the joining of "specific" and "common" cistrons to enable the production of hybrid proteins containing variable and constant regions, as discussed above for the light chains. Next, the heavy and light chains diverged from the respective pathways for the variable and constant regions. These two divergences gave rise to further gene duplication and differentiation so that two families of cistrons were formed. The first of these families formed the specific regions of the heavy and light chains, including the heavy chain genes for SG, SA, and SM and the light chain genes $S\kappa_I$, $S\kappa_{II}$, and $S\lambda$. The second family formed the genes for the common regions of κ and λ (light chains) and G_1, G_2, G_3, G_4, A, and M (heavy chains). Much subsequent branching took place in the first of these families so that the large array of different specific-region cistrons was formed.

The homology within the variable portion of the light chain polypeptides obtained from various antibodies suggests that the cistron for this portion of the immunoglobulin molecule exists in thousands of copies

which arose by saltatory multiplication of a short segment of DNA in the manner proposed by Britten and Kohne (1965–1966) for mouse satellite DNA. A limited degree of differentiation of these multiple segments could conceivably give rise to cistrons for a very large number of polypeptide sequences, each of which has a combining power for a different antigen resulting from the polypeptide configuration of the individual antibody.

It is necessary to explain how the correct antibody becomes produced in response to a specific antigen, but for discussion of this point, the reader is referred to articles and textbooks of immunology. We shall concern ourselves only with the possibilities of duplication and differentiation during the evolution of the immunoglobulins. This is discussed in Section V.

F. Insulin

Insulin consists of two short and complete polypeptide chains which are linked by two —S—S— bridges. The insulins of various species of animals show a characteristic phylogenetic divergence in terms of amino acid sequences. One of the more variable regions is residues 8, 9, and 10 of the A chain. Interchanges of these amino acid residues in different mammalian species were discovered by Sanger (1952). This was an early finding in protein evolution. Subsequent studies showed that many residues in the two chains are variable. Smith (1966) has analyzed the insulins of various fishes and he has compiled the information from various sources on the sequence of amino acids in these and other insulins.

Multiple insulins are found in the rat and in fishes. The rat has two allelic insulins which differ by a single interchange, lysine to methionine, at residue B29 (Fig. 5). Only 23 of the 51 amino acid residues in insulins have not been found to be subject to replacement.

According to a proposal by Eck (1964), a common evolutionary origin for the two chains was followed by the appearance of several gaps. With the gaps inserted, the average of the minimum base differences per codon in the comparison of the two chains of bovine insulin is 0.77.

One of the proposed gaps preceding the first residue in the B chain, has, since Eck's proposal, been shown to be occupied by methionine or valine in insulins of the toadfish and angler fish. The information on the fish insulins is not complete enough to make quantitative comparisons of their sequences with those of mammalian insulins.

The variability in residues of 8, 9, and 10 of the A chain of mammalian

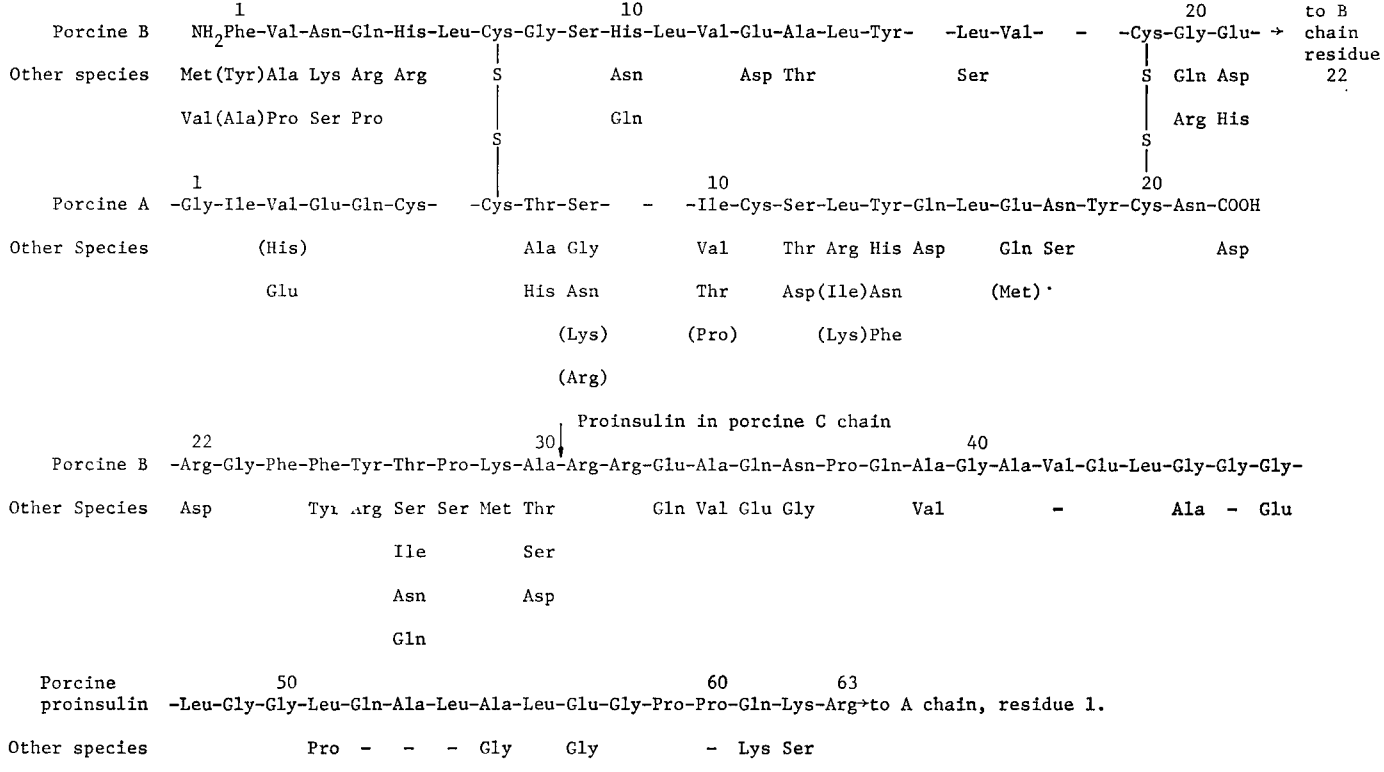


FIG. 5. Porcine insulin B and A chains, with gaps as proposed by Eck (1964); compared with insulins of other species (Smith, 1966); and connected with the C chain as in proinsulin (Chance and Ellis, 1968; Margoliash and Steiner, 1968). The amino acid residues in parentheses are tentative assignments (Smith, 1966).

insulins suggests that these are "neutral" changes. Each of the following groups of mammals has insulins that are identical in this region:

(A)	(B)	(C)
Pig	Sei whale	Elephant
Dog	Human	Rabbit
Sperm whale		
Fin whale		

These identities show no phylogenetic relationship, and thus support the concept of selectively neutral amino acid replacements in evolution (King and Jukes, 1969).

Evidence was presented by Steiner and Oyer (1967) and Steiner *et al.* (1967) that insulin is derived biologically from a larger precursor protein termed *proinsulin*. They found that radioactive amino acids were incorporated earlier into proinsulin than into insulin in islet cells, and that the radioactivity could be "chased" into insulin. They suggested that proinsulin consisted of a single polypeptide chain beginning at its N-terminal end with the B chain sequence of insulin, followed by an additional polypeptide, and terminating with the A chain sequence.

These predictions were realized by Chance and Ellis (1968) and Margoliash and Steiner (1968). The latter investigators found that proinsulin was present to the extent of about 1% in crude crystalline insulin as obtained from the commercial manufacturing process. Porcine proinsulin (Chance and Ellis, 1968) may be diagrammatically represented as follows:

B chain—33-residue peptide—A chain

The 33-residue peptide, which for convenience may be termed the "C" chain, is apparently removed intracellularly in the pancreatic islet tissue by specific proteases, thus converting proinsulin to insulin. Margoliash and Steiner found that the bridge in bovine proinsulin contained 27 residues.

Proinsulin has little or no activity in the biological assay for insulin. It may be inferred that the two disulfide bridges (Fig. 5) are formed in the proinsulin molecule prior to its conversion to insulin. The conversion may be produced *in vitro* by chymotrypsin or cocoonase.

III. Functional Differentiation of Proteins as a Result of Evolutionary Divergence

The duplication of genes, followed by differentiation, leads to increases in the functional complexity of living organisms. This concept was stated

TABLE XIX

HOMOLOGOUS REGIONS IN FOUR PITUITARY HORMONES[a]

	Positions	Sequence
A[b]	1, 10, 16	Ser-Tyr-Ser-Met-Glu-His-Phe-Arg-Trp-Gly-Lys-Pro-Val-Gly-Lys-Lys-
B	1, 13	Ser-Tyr-Ser-Met-Glu-His-Phe-Arg-Trp-Gly-Lys-Pro-Val
C	1, 10, 20, 22	Ala-Glu-Lys-Lys-Asp-Glu-Gly-Pro-Tyr-Arg-Met-Glu-His-Phe-Arg-Trp-Gly-Ser-Pro-Pro-Lys-Asp
D	35, 50, 59	-Gln-Ala-Ala-Glu-Lys-Lys-Asp-Ser-Gly-Pro-Tyr-Lys-Met-Glu-His-Phe-Arg-Trp-Gly-Ser-Pro-Pro-Lys-Asp-Lys-

[a] Sources of information: Harris and Lerner (1957); Harris (1959); Li *et al.* (1965); Shepherd *et al.* (1956).

[b] A, corticotropin; B and C, α and β (human) melanocyte-stimulating hormones; D, sheep lipotropic hormone.

TABLE XX

COMPLETE AMINO ACID SEQUENCES OF BOVINE CHYMOTRYPSINOGEN A, B, AND TRYPSINOGEN (LINES 2, 3, AND 4) TOGETHER WITH PARTIAL SEQUENCES OF PORCINE ELASTASE (LINE 5)[a,b]

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Chymotrypsinogen A	Cys	Gly	Val	Pro	Ala	Ile	Gln	Pro	Val	Leu	Ser	Gly	Leu	Ser	Arg	Ile	Val	Asn	Gly	Glu	Glu	Ala	Val
Trypsinogen															Val	Asp	Asp	Asp	Asp	Lys	Ile	Val	Gly
Chymotrypsinogen B	Cys	Gly	Val	Pro	Ala	Ile	Gln	Pro	Val	Leu	Ser	Gly	Leu	Ala	Arg	Ile	Val	Asn	Gly	Glu	Asp	Ala	Val
Elastase																Val	Val	Gly	Gly	Thr	Glu		

	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46
Chymotrypsinogen A	Pro	Gly	Ser	Trp	Pro	Trp	Gln	Val	Ser	Leu	Gln	Asp	Lys	Thr	Gly	Phe	His	Phe	Cys	Gly	Gly	Ser	Leu
Trypsinogen	Ala	Asn	Thr	Val	Pro	Tyr	Gln	Val	Ser	Leu	0	0	Asn	Ser	Gly	Tyr	His	Phe	Cys	Gly	Gly	Ser	Leu
Chymotrypsinogen B	Pro	Gly	Ser	Trp	Pro	Trp	Gln	Val	Ser	Leu	Gln	Asp	Ser	Thr	Gly	Phe	His	Phe	Cys	Gly	Gly	Ser	Leu
Elastase															Trp	Ala	His	Thr	Cys	Gly	Gly	Thr	Leu

	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69
Chymotrypsinogen A	Ile	Asn	Glu	Asn	Trp	Val	Val	Thr	Ala	Ala	His	Cys	Gly	Val	Thr	Thr	Ser	Asp	Val	Val	Val	Ala	Gly
Trypsinogen	Ile	Asn	Ser	Gln	Trp	Val	Val	Ser	Ala	Ala	His	Cys	Tyr	Lys	Ser	Gly	Ile	Gln	Val	Arg	Leu	0	Gly
Chymotrypsinogen B	Ile	Ser	Glu	Asp	Trp	Val	Val	Thr	Ala	Ala	His	Cys	Gly	Val	Thr	Thr	Ser	Asp	Val	Val	Val	Ala	Gly
Elastase								Thr	Ala	Ala	His	Cys	Val	Asp	Arg	Glu							

[a] Sources of information: Brown et al. (1967); Hartley et al. (1965); Walsh and Neurath (1964); Smillie et al. (1968).

[b] Asx = either Asp or Asn; Glx = either Glu or Gln. The numbering corresponds to the total number of sites, including gaps (shown by 0), as postulated for chymotrypsinogen A and trypsinogen. Gaps have been placed in terms of the minimum number of base changes at homologous loci.

```
                        70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92
Chymotrypsinogen A  Glu-Phe-Asp-Gln-Gly-Ser-Ser-Ser-Glu-Lys-Ile-Gln-Lys- 0 -Leu-Lys-Ile-Ala-Lys-Val-Phe-Lys-Asn-

Trypsinogen         Gln -0- Asp-Asn-Ile-Asn-Val-Val-Glu-Gly-Asn-Gln-Gln-Phe-Ile-Ser-Ala-Ser-Lys-Ser-Ile-Val-His-

Chymotrypsinogen B  Glu-Phe-Asp-Gln-Gly-Leu-Glu-Thr-Glu-Asp-Thr-Gln-Val- 0 -Leu-Lys-Ile-Gly-Lys-Val-Phe-Lys-Asn-

Elastase

                        93  94  95  96  97  98  99  100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115
Chymotrypsinogen A  Ser-Lys-Tyr-Asn-Ser-Leu-Thr-Ile-Asn-Asn-Asn-Ile-Thr-Leu-Leu-Lys-Leu-Ser-Thr-Ala-Ala-Ser-Phe-

Trypsinogen         Pro-Ser-Tyr-Asn-Ser-Asn-Thr-Leu-Asn-Asn-Asp-Ile-Met-Leu-Ile-Lys-Leu-Lys-Ser-Ala-Ala-Ser-Leu-

Chymotrypsinogen B  Pro-Lys-Phe-Ser-Ile-Leu-Thr-Val-Arg-Asn-Asp-Ile-Thr-Leu-Leu-Lys-Leu-Ala-Thr-Pro-Ala-Gln-Phe-

Elastase

                        116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138
Chymotrypsinogen A  Ser-Gln-Thr-Val-Ser-Ala-Val-Cys-Leu-Pro-Ser-Ala-Ser-Asp-Asp-Phe-Ala-Ala-Gly-Thr-Thr-Cys-Val-

Trypsinogen         Asn-Ser-Arg-Val-Ala-Ser-Ile-Ser-Leu-Pro-Thr-Ser-Cys-Ala- 0 - 0 -Ser-Ala-Gly-Thr-Gln-Cys-Leu-

Chymotrypsinogen B  Ser-Glu-Thr-Val-Ser-Ala-Val-Cys-Leu-Pro-Ser-Ala-Asp-Glu-Asp-Phe-Pro-Ala-Gly-Met-Leu-Cys-Ala-

Elastase                                                                                    -Ala-Asn-Asn-Ser-Pro-Cys-Tyr

                        139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161
Chymotrypsinogen A  Thr-Thr-Gly-Trp-Gly-Leu-Thr-Arg-Tyr-Thr-Asn-Ala-Asn-Thr-Pro-Asp-Arg-Leu-Gln-Gln-Ala-Ser-Leu-

Trypsinogen         Ile-Ser-Gly-Trp-Gly-Asn-Thr-Lys-Ser-Ser-Gly-Thr-Ser-Tyr-Pro-Asp-Val-Leu-Lys-Cys-Leu-Lys-Ala-

Chymotrypsinogen B  Thr-Thr-Gly-Trp-Gly-Lys-Thr-Lys-Tyr-Asn-Ala-Leu-Lys-Thr-Pro-Asp-Lys-Leu-Gln-Gln-Ala-Thr-Leu-

Elastase
```

(*continued*)

TABLE XX—(*Continued*)

```
                    162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184
Chymotrypsinogen A  Pro-Leu-Leu-Ser-Asn-Thr-Asn-Cys-Lys-Lys-Tyr-Trp-Gly-Thr-Lys-Ile-Lys-Asp-Ala-Met-Ile-Cys-Ala-

Trypsinogen         Pro-Ile-Leu-Ser-Asp-Ser-Ser-Cys-Lys-Ser-Ala-Tyr-Pro-Gly-Gln-Ile-Thr-Ser-Asn-Met-Phe-Cys-Ala-

Chymotrypsinogen B  Pro-Leu-Val-Ser-Asn-Thr-Asp-Cys-Arg-Lys-Tyr-Trp-Gly-Ser-Arg-Val-Thr-Asp-Val-Met-Ile-Cys-Ala-

Elastase                                -Ala-Ile-Cys-Ser-Ser-Ser-Ser-Ser-Tyr-                          -Met-Val-Cys-Ala-

                    185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207
Chymotrypsinogen A  Gly-Ala-Ser-Gly-Val- 0 --0--Ser-Ser-Cys-Met-Gly-Asp-Ser-Gly-Gly-Pro-Leu-Val-Cys-Lys-Lys-Asn-

Trypsinogen         Gly-Tyr-Leu-Glu-Gly-Gly-Lys-Asn-Ser-Cys-Gln-Gly-Asp-Ser-Gly-Gly-Pro-Val-Val-Cys- 0 - 0 - 0-

Chymotrypsinogen B  Gly-Ala-Ser-Gly-Val- 0 - 0 -Ser-Ser-Cys-Met-Gly-Asp-Ser-Gly-Gly-Pro-Leu-Val-Cys-Gln-Lys-Asn-

Elastase            Gly-                    -Arg-Ser-Gly-Cys-Gln-Gly-Asp-Ser-Gly-Gly-Pro-Leu-His-Cys-Leu-Val-Asn-

                    208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230
Chymotrypsinogen A  Gly-Ala-Trp-Thr-Leu-Val-Gly-Ile-Val-Ser-Trp-Gly-Ser-Ser-Thr-Cys-Ser- 0 -Thr-Ser-Thr-Pro-Gly-

Trypsinogen         -0 -Ser-Gly-Lys-Leu-Gln-Gly-Ile-Val-Ser-Trp-Gly-Ser-Gly- 0 -Cys-Ala-Gln-Lys-Asn-Lys-Pro-Gly-

Chymotrypsinogen B  Gly-Ala-Trp-Thr-Leu-Ala-Gly-Ile-Val-Ser-Trp-Gly-Ser-Ser-Thr-Cys-Ser- 0 -Thr-Ser-Thr-Pro-Ala-

Elastase            Gln-Tyr-                                -Val-Ser-Arg-Leu-Gly-Cys-Asn-Val-Thr-Arg-Lys-Pro-Thr

                    231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249
Chymotrypsinogen A  Val-Tyr-Ala-Arg-Val-Thr-Ala-Leu-Val-Asn-Trp-Val-Gln-Gln-Thr-Leu-Ala-Ala-Asn-COOH

Trypsinogen         Val-Tyr-Thr-Lys-Val-Cys-Asn-Tyr-Val-Ser-Trp-Ile-Lys-Gln-Thr-Ile-Ala-Ser-Asn-COOH

Chymotrypsinogen B  Val-Tyr-Ala-Arg-Val-Thr-Ala-Leu-Met-Pro-Trp-Val-Gln-Glu-Thr-Leu-Ala-Ala-Asn-COOH

Elastase
```

by Horowitz (1945) and Lewis (1951), who postulated that a biosynthetic pathway may depend on a series of enzymes which have been formed by duplication and differentiation.

Myoglobin and hemoglobin are good examples of an evolutionary divergence which has resulted in the existence of two proteins with slightly different functions, although both proteins still combine with the same oxygen-transporting prosthetic group. A more marked divergence in function is evident when corticotropin, the lipotropic hormone and the melanocyte-stimulating hormones are compared. This homology is shown in Table XIX. These hormones consist of polypeptide chains with varying lengths, but all contain a region which appears to have a single evolutionary origin. Homology between other regions of the molecules of corticotropin and the lipotropic hormone is not perceptible.

Functional differentiation is evident in the differences in pH optima and substrate specificities shown by the proteases, trypsinogen, chymotrypsinogen and elastase. These have evidently diverged following gene duplication, since trypsinogen and chymotrypsinogen show substantial homology of their primary structures (Table XX and XXI), and suffi-

TABLE XXI
COMPARISONS OF VARIOUS PROTEINS FOR HOMOLOGY

Comparison	Sites Compared	Minimum base differences per codon				
		0	1	2	3	Average
Chymotrypsinogens A and B	245	195	38	12	0	0.24
Trypsinogen: chymotrypsinogen A	225	100	69	54	2	0.82
Trypsinogen: chymotrypsinogen B	225	90	79	54	2	0.86
Glucagon: secretin	27	14	7	6	0	0.70
Chicken lysozyme: bovine α-lactalbumin (incomplete)	99	42	31	23	3	0.87
Chicken lysozyme: bovine pancreatic RNase	124	16	51	53	4	1.38
Bovine pancreatic RNase: bovine α-lactalbumin (incomplete)	98	10	39	45	4	1.44

cient information on the sequences of chymotrypsinogen B and elastase is available to indicate that all four enzymes are homologous. A fifth enzyme, cocoonase, has preliminary indications of homology with these four (Kafatos *et al.*, 1967) since it has a similar amino acid composition; however, its primary structure is not known.

Glucagon (Behrens and Bromer, 1958) and secretin (Mutt and Jorpes, 1966) (Table XII) are two small polypeptide hormones with entirely different functions. They apparently have been formed by evolutionary differentiation from a common archetype (Eck and Dayhoff, 1966). Glucagon increases the level of blood glucose by accelerating the activity of liver phosphorylase kinase, which increases the formation of glucose 1-phosphate. Secretin stimulates the flow of pancreatic juice. The homology between the primary structures of the two hormones was discussed by Weinstein (1968). The MBDC is 0.70 (Table XXII).

Homology between chicken lysozyme from hen's egg white and bovine α-lactalbumin was discovered by Brew *et al.* (1967). A comparison of the complete sequence of chicken lysozyme and the partial sequence of bovine α-lactalbumin is given in Table XXIII. The comparison is summarized in Table XXI.

A second homology to chicken lysozyme, that of bovine pancreatic ribonuclease, was proposed by Manwell (1967), and the comparison is in Table XXIII. Manwell inserted a 4-residue gap between residues 3 and 4 and a single residue gap between residues 90 and 91 of the ribonuclease. The first of these gaps does not improve the homology sufficiently to justify its inclusion and it has been omitted from Table XXIII. The homology between chicken lysozyme and bovine pancreatic

TABLE XXII
HOMOLOGY BETWEEN GLUCAGON AND SECRETIN

	1									10					
Glucagon	His-	Ser-	Gln-	Gly-	Thr-	Phe-	Thr-	Ser-	Asp-	Tyr-	Ser-	Lys-	Tyr-	Leu-	Asp-
Secretin	His-	Ser-	Asp-	Gly-	Thr-	Phe-	Thr-	Ser-	Glu-	Leu-	Ser-	Arg-	Leu-	Arg-	Asp-
MBDC	0	0	2	0	0	0	0	0	1	2	0	1	2	1	0

					20									29	
Glucagon	Ser-	Arg-	Arg-	Ala-	Gln-	Asp-	Phe-	Val-	Gln-	Trp-	Leu-	Met-	Asn-	Thr-	COOH
Secretin	Ser-	Ala-	Arg-	Leu-	Gln-	Arg-	Leu-	Leu-	Gln-	Gly-	Leu-	Val-	$CONH_2$		
MBDC	0	2	0	2	0	2	1	1	0	1	0	1			

TABLE XXIII

COMPARISONS OF PRIMARY STRUCTURES OF HENS EGG WHITE LYSOZYME (A), BOVINE α-LACTALBUMIN (B), AND BOVINE PANCREATIC RIBONUCLEASE (C)

```
          1                   10                  20
A   Lys-Val-Phe-Gly-Arg-Cys-Glu-Leu-Ala-Ala-Ala-Met-Lys-Arg-His-Gly-Leu-Asp-Asn-Tyr-Arg-Gly-Tyr-Ser-Leu-

B   Glu-Glu-Leu-Thr-Lys-Cys-Glu-Val-Phe-Arg-Glu-Leu-Lys-       Asp-Leu-Lys-Gly-Tyr-Gly-Gly-Val-Ser-Leu-

C                   Lys-Glu-Thr-Ala-Ala-Ala-Lys-Phe-Glu-Arg-Gln-His-Met-Asn-Ser-Ser-Thr-Ser-Ala-Ala-Ser-

                30                  40                  50
A   Gly-Asn-Trp-Val-Cys-Ala-Ala-Lys-Phe-Glu-Ser-Asn-Phe-Asn-Thr-Gln-Ala-Thr-Asn-Arg-Asn-Thr-Asp-Glu-Ser-Thr-

B   Pro-Glu-Trp-Val-Cys-Thr-Thr-Phe-His-Thr-Ser-Gly-Tyr-Asx-Thr-Glx(Ala,Ile,Val,Glx)Asx-   -Asx(Glx,Ser,Thr,

C   Ser-Ser-Asn-Tyr-Cys-Asn-Gln-Met-Met-Lys-Ser-Arg-Asn-Leu-Thr-Lys-Asp-Arg-Cys-Lys-Pro-Val-Asn-Thr-Phe-Val-

                60                  70
A   Asp-Tyr-Gly-Ile-Leu-Gln-Ile-Asn-Ser-Arg-Trp-Trp-Cys-Asp-Asn-Gly-Arg-Thr-Pro-Gly-Ser-Arg-Asn-Leu-Cys-Asn-

B   Asx)Tyr-Gly-Leu-Phe(Glx,Ile,Asx,Asx)Lys-Ile-Trp-Cys-Lys-Asx-Asx-Glx-Asx-Pro-His-Ser-Ser-Asx-Ile-Cys-Asn-

C   His-Glu-Ser-Leu-Ala-Asp-Val-Gln-Ala-Val-Cys-Ser-Gln-Lys-Asn-Val-Ala-Cys-Lys-Asn-Gly-Gln-Thr-Asn-Cys-Tyr-

            80                  90                  100
A   Ile-Pro-Cys-Ser-Ala-Leu-Leu-Ser-Ser-Asp-Ile-Thr-Ala-Ser-Val-Asn-Cys-Ala-Lys-Lys-Ile-Val-Ser-Asp-Gly-Asp-

B   Ile-Ser-Cys-Asp-Lys-Phe-Leu-Asx-Asx-Asx-Leu-Thr(Asx,Asx,Ile)Met-Cys-Val-Lys-Lys-Ile-Leu-   -Asp-Lys-Val-

C   Gln-Ser-Tyr-Ser-Thr-Met-Ser-Ile-Thr-Asp-Cys-Arg-Glu-Thr-Gly-Ser-Ser-   -Lys-Tyr-Pro-Asn-Cys-Ala-Tyr-Lys-

                100                 120
A   Gly-Met-Asn-Ala-Trp-Val-Ala-Trp-Arg-Asn-Arg-Cys-Lys-Gly-Thr-Asp-Val-Gln-Ala-Trp-Ile-Arg-Gly-Cys-Arg-Leu

B   Gly-Ile-Asn-Tyr-Trp-Leu-Ala-His-Lys-Ala-Leu-Cys-Ser-Glu-Lys-Leu-Asp-Gln-   -Trp-Leu-Cys-Glu-   -Lys-Leu

C   Thr-Thr-Gln-Ala-Asn-Lys-His-Ile-Ile-Val-Ala-Cys-Glu-Gly-Asn-Pro-Tyr-Val-Pro-Val-His-Phe-Asp-Ala-Ser-Val
```

ribonuclease (RNase) is not substantial when measured in terms of MBDC, but is supported by the fact that 16 of the 122 pairs of residues in the comparison are identical at homologous sites.

In their comparison, Brew *et al.* (1967) inserted two gaps in lysozyme and five in a α-lactalbumin. However, the two gaps postulated in lysozyme produce insufficient gain in homology and these have been omitted from Table XXIII. They are not supported by the lysozyme-RNase comparison proposed by Manwell (1967). This comparison includes six identities in lysozyme and RNase that are present also in bovine α-lactalbumin. The mean probability is for less than one coincidence in the three sequences. The comparison of lysozyme and α-lactalbumin is noteworthy for the coincidence of seven pairs of cysteine residues. It was the identity in numbers of disulfide bonds that led Brew and Campbell (1967) to suggest a common homology for α-lactalbumins and lysozymes from higher animals.

Brew *et al.* (1967) noted that α-lactalbumin is known to participate in lactose synthesis in the mammary gland. This involves formation of a β1→4 glucopyranosyl linkage. Lysozyme is concerned with the breakage of similar linkages during dissolution of the cell wall of bacteria. Brew *et al.* (1967) therefore suggested that a primitive gene for lysozyme became duplicated. One of the two genes continued to code for lysozyme and the second differentiated to provide the information for an enzyme involved in lactose synthesis. The MBDC value for the comparison of the two proteins is low enough to show clear evolutionary homology and sufficiently high to indicate a separation that could have occurred well before the divergence of mammals and birds.

IV. Taxonomic Serology in the Study of Evolution

One of the first biochemical methods used for comparing different organisms with respect to their phylogenetic relationships was the precipitin test (Nuttall, 1904). The quantitative nature of the test encouraged its application to comparisons of many plants and animals, from which protein extracts were obtained and used for immunizing rabbits. As newer techniques became available, they were applied to these comparative studies. Extensive reviews of research in this field are in "Taxonomic Biochemistry and Serology," edited by Leone (1964).

Immune responses obtained from tissue extracts will reflect the response to a mixture of antigens and, while quantitative studies can be made by this method, they cannot be accurately interpreted in terms of evolutionary divergence at the DNA or protein level. In terms of present-day concepts of the evolutionary divergence of proteins, further information is needed on the precise relationship between immunological responses

and amino acid differences in protein sequences. The first step in this direction is to make a quantitative immunological comparison between two different species on the basis of antibodies for homologous proteins obtained in a pure or crystalline form from both species. Studies relating immunological cross-reactivity to the structure of proteins were made by Reichlin, Wilson and their collaborators. Reichlin *et al.* (1966) prepared antibodies against the α and β chains of human hemoglobin by intramuscularly injecting the antigen mixed with Freund's adjuvant, followed by intravenous injections of the antigen. Complement fixation tests were by the micro method of Wasserman and Levine (1961). The procedure enabled the detection of differences in complement fixation between the normal chains and the following mutants containing single amino acid replacements as indicated: $\alpha^{16\text{Lys}\rightarrow\text{Glu}}$; $\alpha^{6\text{Glu}\rightarrow\text{Val}}$; $\beta^{6\text{Glu}\rightarrow\text{Lys}}$; and $\beta^{7\text{Glu}\rightarrow\text{Lys}}$. The depression in complement fixation by the heterologous antigen varied from 25% to 48%. Wilson *et al.* (1964) also briefly described differential responses to some of these hemoglobin variants.

These investigations laid the groundwork for establishing a correlation between immunological cross-reactions and the primary structure of protein by means of the micro complement fixation (MCF) test. Sarich and Wilson (1967) used the MCF test to compare serum albumin from human beings and various apes and monkeys. The procedure led to their devising an index of dissimilarity. Some typical findings are in Table XXIV.

TABLE XXIV
REACTIVITY OF VARIOUS PRIMATE SERUM ALBUMINS WITH ANTISERA PREPARED AGAINST HOMINOID ALBUMINS[a]

Species	Index of dissimilarity		
	Antiserum to man	Antiserum to chimpanzee	Antiserum to gibbon
Hominoidea (apes and man)			
Man	1.0	1.09	1.29
Chimpanzee	1.14	1.00	1.40
Pygmy chimpanzee	1.14	1.00	1.40
Gorilla	1.09	1.17	1.31
Orangutan	1.22	1.24	1.29
Siamang	1.30	1.25	1.07
Gibbon	1.28	1.25	1.00
Cercopithecoidea (Old World monkeys)			
Six species (mean ± S.D.)	2.46 ± 0.16	2.22 ± 0.27	2.29 ± 0.10

[a] The index of dissimilarity is based upon the value of 1.00 representing complete reactivity and no dissimilarity. Taken from Sarich and Wilson (1967).

From these results, compared with the primate fossil record, they concluded that the lines of descent leading to the hominoids and Old World monkeys divided about 30 million years ago. Using this scale, they calculated that the time of divergence of man from the African apes (gorilla and chimpanzee) was 5 million years ago. Sarich and Wilson (1967) point out that extensive immunological studies of proteins of known amino acid sequences are necessary to establish a relationship between immunological cross-reactions and degree of primary structure resemblance.

A sensitive means for making immunological comparisons of the cytochromes *c* was described by Margoliash *et al.* (1968). Antisera specific to cytochrome *c*, which had previously been reported to be nonantigenic, were produced in rabbits by any of the three following methods: (a) repeated injections of cytochrome *c* for long periods, (b) injection of cytochrome *c* coupled covalently to acetylated bovine γ globulin, (c) injection of polymers of cytochrome *c* produced by treating it with ethanol or glutaraldehyde. They found that some sera did not distinguish between certain cytochromes *c* which had different amino acid sequences.

Using these antisera, they found important relationships between structure and immunological specificity in the cytochromes *c*. Human cytochrome *c* differs from that of the rhesus monkey solely in having isoleucine rather than threonine at position 58. This substitution has a marked effect on the immunological response. Horse cytochrome *c* resembles that of the monkey in that both have threonine at 58, while kangaroo cytochrome *c*, like human cytochrome *c*, has isoleucine at site 58. Elsewhere in the molecule, the horse and kangaroo cytochromes *c* differ markedly from the human and rhesus monkey cytochromes *c*.

When antihuman cytochrome *c* sera were treated with an excess of rhesus cytochrome *c*, the treated sera, containing residual antibodies, were still able to react with human or kangaroo cytochrome *c* but not significantly with any of the other cytochromes *c* that were tested. The presence of isoleucine at position 58 therefore conferred on human and kangaroo cytochromes *c* a specificity that was different from those of a number of other cytochromes *c* even though these included species that were more closely related than kangaroos to human beings on a taxonomic basis.

In contrast, a similar comparison of horse and donkey cytochromes *c* which differ from each other at position 47, where the residue in the horse is threonine and in the donkey serine, were indistinguishable by the immunological test. The threonine to serine interchange at 47 therefore conferred no immunological differentiation, thus differing from the threonine to isoleucine interchange at 58. The conclusion is that each

amino acid in a protein molecule may have its own specific effect upon antigenic behavior.

The most prolific example of specificity in serology is, of course, the immunoglobulins themselves. The amino-terminal halves of the light chains (S-regions) recognize different antigens, in many cases on the basis of single amino acid differences in the light chain sequences. The valine and leucine interchange in the C-region of human light chains is distinguishable by antisera against human immunoglobulin (Hood *et al.*, 1968).

V. Statistical Procedures and Computer Techniques

Once the amino acid sequences of a few proteins had been determined, it became possible to search for similarities in sequence which might indicate a common function or evolutionary origin for two or more proteins. Early work in this area was performed by hand and often concentrated on sequences thought to be near the active sites of enzymes. In recent years, as the rate of discovery of new sequences accelerated, the search has been intensified. The study of amino acid sequences of homologous proteins from many different organisms has led to the conclusion that evolution on a molecular level fairly closely parallels the classical evolutionary pathways known for organisms (Fitch and Margoliash, 1967a). Suggestions for homology within a set of similar proteins such as the cytochromes or globins are easily justified because very few amino acid residues are found to differ when the proteins of two closely related species are compared. A more serious problem arises when attempts are made to compare the amino acid sequences of two very different proteins or to compare different parts of the sequence of the same protein to determine if internal duplication has occurred. Here the difficulty arises that very few amino acids are identical when the two sequences are examined side by side. There exists no simple objective criterion which is capable of deciding when two sequences are similar enough to be declared related by an evolutionary pathway. A second problem is the large number of comparisons which can be made. First is the difficulty of comparing all of the possible alignments of two fairly dissimilar sequences in an attempt to look for any traces of homology. Even more serious is the fact that the number of possible comparisons between two distinct protein sequences increases as the square of the number of proteins of known sequence. Thus the advantages of using a modern digital computer to compare amino acid sequences are overwhelming. Computer comparisons are rapid and thorough. They also permit semiquantitative criteria to be developed which will allow the significance of a suspected homology to be estimated.

Consider, first, the problem of comparing two arbitrary continuous protein sequences chosen at random from two different proteins or from different parts of the same protein. A typical selection is as follows:

Sequence

	1																			20
1	Thr-	Tyr-	Pro-	Gly-	Asp-	Gln-	Gln-	Met-	Glu-	Arg-	Lys-	Val-	Trp-	Ser-	Thr-	Gly-	Glu-	His-	Leu-	Pro
2	Phe-	Glu-	Pro-	His-	Gly-	Asp-	His-	Ile-	Cys-	His-	Ile-	Gly-	Ser-	Thr-	Lys-	Glu-	Leu-	Leu-	Val-	Thr

(1)

If we knew the a priori probability of choosing each of the twenty normal amino acids for inclusion in the above sets, we could calculate the frequency of finding the same amino acid at the same site on both chains. In the above example the amino acids were selected with the aid of a random number table. Each was given a statistical weight of $\frac{1}{20}$. In this case, the probability of finding m identical amino acids in the same site on the two chains of length L is given by a simple binominal expression

$$P(L,m) = \frac{L!}{m!(L-m)!}\left(\frac{1}{20}\right)^m\left(\frac{19}{20}\right)^{L-m} \tag{2}$$

A major difficulty exists, however, when one tries to extend the above argument to a comparison involving sequences chosen from two real protein chains. In this case, the probability of occurrence of each of the amino acids is not the same. Certain amino acids, for example histidine, methionine, and tryptophan, are found rather infrequently in most of the proteins of known sequence. In contrast, others such as alanine and glycine are almost always among the most prevalent. Thus the binomial expression used above must be corrected to reflect our knowledge about the a priori probability of amino acid occurrence. Unfortunately this knowledge is insufficient. We cannot yet say with any substantial grounds why a given amino acid should occur in a protein more often than any other one.

What is needed is an approximate way of estimating the a priori probability of each amino acid being found in a given protein. The simplest and most convenient way to do this is to consider each protein as a set of amino acids whose sequence is random. Then the probability of finding each amino acid at a given site in the protein is just the fractional composition of that amino acid in the protein (Fitch, 1966a). If X_{1i} and X_{2i} are the mole fractions of the ith amino acid in proteins 1 and 2, respectively, the probability that the same amino acid will appear at a given site in both proteins is $P(1)$.

$$P(1) = \sum_{i=1}^{20} X_{1i}X_{2i} \tag{3}$$

The probability of m identical residues then becomes

$$P(L,m) = \frac{L!}{m!(L-m)!} P(1)^m [1 - P(1)]^{L-m} \quad (4)$$

Until several years ago this expression was about the best one could use to calculate the probability of random occurrence of two similar protein chains. The assumption of fixing amino acid probabilities from the amino acid composition of the proteins means, of course, that no homology can be claimed for two proteins that merely have similar proportions of the 20 amino acids. Only if specific sequences are similar should a large set of comparisons show substantially more coincidences than calculated by Eq. (4). The major disadvantage associated with the direct comparison of amino acids is that the probability distribution which occurs is rather coarse-grained. This means that no consideration is given to the occurrence of "similar" amino acids in the same site. To circumvent this difficulty Pauling and Zuckerkandl (1963) considered, for example, interchanges of two hydrophobic amino acids to be less drastic than, say, exchange of a hydrophobic for an ionic amino acid. While such an approach is certainly reasonable from the point of view of the function of these amino acids in the protein, it is less direct than a knowledge of the relative probabilities of interchanging two amino acids by mutations. Our understanding of the molecular mechanisms of mutation is not nearly complete enough yet to permit these probabilities to be accurately determined. However, the genetic code permits an approximation to be made (Jukes, 1966). From the triplet nature of the code one can see that certain amino acid interchanges are much more likely than others in the limit of small number of base changes. Depending on the amino acids involved, it can take either 1, 2, or 3 base changes in DNA (or RNA) to convert one amino acid to another. In general, amino acids with similar chemical structures have codons that are related by relatively small numbers of base changes. Thus Gly-Ala can be effected by changing only a single base in the codon while Asp-Trp requires three base changes. This provides support for the approach used by Pauling and Zuckerkandl (1963) and it also means that one can compare amino acid sequences on the nucleotide level instead of directly. The approximation one must make is to say that all single base changes are equally probable. It will be possible to relax this assumption once more extensive experimental data on the frequency of amino acid interchanges are available.

If we knew the nucleotide sequences of the messenger RNAs which directed the synthesis of a given pair of proteins, we could make a very good estimate of the number of base changes needed to convert

one amino acid into the other. This would be a minimum estimate since many alternative routes would exist through which a less direct interconversion could be possible. It can be shown that the mean number of base differences at a single position on the mRNA, μ, is related to the observed fraction of residues with single base differences, p, by the expression

$$\mu = \frac{3}{4} \ln \frac{3}{3 - 4p}$$

The difficulty presented by the genetic code is its synonymities. Given a codon, we can tell the amino acid it codes for, but the reverse transformation cannot be performed uniquely. In some cases, 6 possible codons correspond to the same amino acid. This means that we cannot say with certainty how many base changes are needed to convert, say serine into an arginine. There are two possible ways to resolve this problem. The first, and simplest, is to compare two amino acids on the basis of the minimum number of base changes which would be needed to interconvert them. Thus, Asp → Asn (codons: Asp, GAU, GAC; Asn, AAU, AAC) would be counted as one base difference even though it could have occurred by two observable base changes or two or more actual mutations. This is the approach we have used in most of the calculations to be discussed in this chapter. The alternative is to consider the average number of base changes needed to convert two amino acids. There are four ways to change Asp to Asn; two of these involve one base change, the others involve two. The average is 1.5 base differences per codon. This method seems reasonable if two randomly selected sequences are being compared, but it is open to question when two very similar sequences are under study. The additional complications introduced by considering average base differences per codon do not seem to be warranted at present, although they are necessary when one tries to estimate the actual number of base changes that may have occurred in the evolution of one protein sequence from another. If minimum base differences per codon are used, all the necessary information for protein sequences can be summarized in a 23 × 23 matrix. This symmetric array, shown in Table II, permits at a glance, a number from 0 to 3 to be assigned to the relative differences between two amino acids. Codons which lead to chain termination are included in the matrix.

If two fairly similar proteins are to be compared, a visual inspection of the sequences will often permit an arrangement which seems to maximize the homology between them. There is, of course, no guarantee that this orientation aligns the maximum number of similar amino acids. When the genetic code is used as a measure of amino acid similarity,

visual analysis requires a memorization of the code. The use of digital computers easily permits arrangements of sequences to be located which maximize homology. It is also relatively simple to use minimum codon differences to compare sequences. The major advantage of computational methods is that every possible arrangement of sequences chosen from two proteins can be tested. The result is a distribution of comparisons involving varying numbers of total minimum base differences. The comparisons involving the least number of minimum base differences may indicate possible sites of homology. Their frequency of occurrence can be tested against statistical models to ascertain whether these comparisons could simply be due to chance. To permit efficient use of statistics it is desirable to have a very large number of total comparisons. An easy way to do this was first suggested by Fitch (1966a). Choose all possible sequences of length L from the total sequences of a much larger protein of length N_1. These can then be compared with all sequences of length L from the second protein of length N_2. This results in $(N_1 - L + 1)$ $(N_2 - L + 1)$ total comparisons. If a protein is compared with itself, neglecting comparisons of identical sequences, there are only $(N_1 - L + 1)$ $(N_1 - L)/2$ comparisons. L is called the comparison length. For most sequence comparisons we have found it convenient to choose L from 20 to 30 amino acids. Working with short sequences offers an additional advantage over comparisons of entire protein chains. Deletions or additions of amino acid residues are often thought to occur in proteins. If two whole proteins are compared, the effect of one of these would be to throw any homologous residues after the deletion out of register. Comparisons of short sequences will pick up homology on both sides of the deletion or addition. The relative orientations of the sets of homologous peptide sequences can be used to find the location and size of the gap in the sequence. A detailed example of this will be presented later in the chapter.

It is appropriate at this point to give a more detailed description of the computer routines we have found useful in comparing protein sequences. Two protein sequences are numerically coded. For example Ala-Gly-Ser-Thr-Cys would be represented as the vector (1,8,16,17,5). All possible sequences of length L chosen from the two proteins are aligned. The act of comparing the two sequences consists of summing the appropriate elements of the matrix, M, shown in Table II. Thus the result of comparing sequences (1,8,16,17,5) and (2,8,4,12,6) would be $M_{1,2} + M_{8,8} + M_{16,4} + M_{17,12} + M_{5,6} = 8$. The number of minimum base differences found for each comparison is tabulated. If this number is less than some previously defined value, the location of the two sequences, the actual amino acids compared, and the number of base differences can be printed

out. For the above comparison the following output would result:

First residue = 54	Ala-Gly- Ser-Thr-Cys	Minimum differences
First residue = 39	Arg-Gly-Asp-Lys-Gln	8
	2 0 2 1 3	

When all possible comparisons have been completed, the total number of comparisons yielding each possible number of minimum base differences is printed along with normalized values for this distribution and any other statistical parameters of interest. The frequency distribution and base changes for sequences of length L must then be compared with values calculated for random sequences.

Earlier we demonstrated how to compute the probability of finding m identical amino acids in the same positions of two protein sequences chosen at random. These calculations must now be extended to take into account the fact that we are evaluating comparisons by minimum base differences. This was first done by Fitch (1966a). Assume for the moment that the a priori probabilities of picking pairs of amino acids which differ in their codons by 0, 1, 2, or 3 bases are known. These are defined as $P(0)$, $P(1)$, $P(2)$, and $P(3)$. Suppose that for a given comparison we find that N_0 amino acid pairs which differ by 0 base differences N_1 by 1 base difference, etc. $N_0 + N_1 + N_2 + N_3 = L$. There are $L!/N_0!N_1!N_2!N_3!$ ways in which this set of comparisons can be permuted among a set of L amino acid pairs. The frequency of occurrence of such sets is given by

$$P(L,N_0,N_1,N_2,N_3) = L! \prod_{I=0}^{3} P(I)^{N_I}/N_I!$$

We are interested in the frequency of finding any comparison with n total base differences out of L residues. The number of base differences is $n = N_1 + 2N_2 + 3N_3$. Thus the quantity we want is

$$P(L,n) = \Sigma P(L,N_0,N_1,N_2,N_3)$$

where the sum is taken over all possible choices of N_0, N_1, N_2, and N_3 which lead to the desired value of n. For values of L between 20 and 30 a very large number of terms contribute to the above sum. These must be evaluated by computer. The distribution $P(L,n)$ is approximately Gaussian. In the past, we have sometimes used this approximation to simplify calculations of $P(L,n)$ (Cantor and Jukes, 1966a), but for more precise calculations the exact values should be used.

The problem which remains is to obtain estimates for the probability, $P(I)$, of finding amino acid pairs characterized by 0, 1, 2, or 3 base changes. The simplest approximation is to assume each amino acid is

a priori equally probable. Then $P(I)$ is simply the fraction of elements in the matrix of Table II which contains the value I. But as discussed earlier, this is not a very accurate representation of protein sequences. The second approach, first used by Fitch (1966a), is to weight the probability of finding amino acids according to their frequency in the amino acid composition of the two proteins. Then the probability of a comparison between the ith type of amino acid of protein 1 and the jth type of protein 2 is $X_{1i}X_{2j}$. Using this information we can estimate $P(I)$.

$$P(I) = \Sigma X_{1i}X_{2j}$$

The above sum is carried out over all of the coefficients which correspond to elements of the comparison matrix whose value is I. Calculations of $P(L,n)$ using $P(I)$ estimated in this way have usually been found to give excellent agreement with observed frequencies when two completely unrelated proteins are compared. However, the above procedure overcounts certain amino acids and for the most exact work a slight modification may sometimes be useful.

In the process of comparing two sequences, we are going to test each sequence of length L from protein 1 with all sequences of length L from protein 2. Only *one* oligomer of length L contains the N-terminal residue of protein 1. This will be involved in a total of $N_2 - L + 1$ comparisons with sequences from protein 2. But there are L oligomers which contain a residue located far from the ends of sequence one. Therefore one of these residues will be involved in $L(N_2 - L + 1)$ comparisons. The same argument, of course, holds for residues on sequence 2. If the amino acid composition of the ends of the proteins is quite similar to the overall composition, this end effect will not be important. But for some proteins this is not a good assumption. In these cases, it is necessary to modify the probability of finding a given amino acid pair according to the distribution of these residues near the ends of the protein chains. This correction essentially consists in comparing the two proteins with a comparison length of one. This then serves as the control for comparisons of length L. Thus the same program which compares sequences can also be used to calculate the necessary $P(I)$ needed to estimate the probability of occurrence of a given single amino acid comparison.

Suppose that two proteins have been compared by the computer methods discussed above. In addition, the probabilities of observing comparisons with various minimum base differences have been calculated. Two distributions of number of comparisons as a function of minimum base differences result. We shall call the first of these the observed distribution and the second the calculated distribution. How should they be compared to see if there is any homology between the two sequences? A standard

method of comparing two probability distributions involves the use of the chi-square test. This test weights elements of a distribution according to the number of events which fall into a particular category, such as number of base changes. However, the comparisons which tend to indicate protein homology are often just a few isolated points near one tail of the distribution. These may not contribute strongly to a calculation of chi-square. A second problem is that use of the chi-square test presupposes a knowledge of the number of independent determinations which led to the observed frequency distribution. The methods of protein comparison we have used do not lead to statistically independent events. If 20 residues of protein 1 are compared with a set of protein 2, and the neighboring set of 20 is then compared with the neighboring set on the second sequence, 19 out of the 20 comparisons are identical. These comparisons are obviously not independent. Since we are starting with a limited set of amino acids, and are forming a large number of possible subsets from these and comparing them, it is not quite clear just how many independent measurements have been made. So instead of using tests which compare the shapes of the distributions, it has proven convenient simply to compare the parts of the calculated and observed distributions which are of interest.

What is needed is a way of comparing the observed and calculated frequencies of minimum base differences which focuses attention on the parts of the distribution that involve relatively small numbers of base differences. A simple linear plot of the two distributions *vs.* base differences would tend to obscure any small deviations which occur near the "tails." Two convenient methods exist which can magnify these deviations. The first, which was employed by Fitch (1966a), involves the use of probability paper. Here the cumulative frequency of comparisons is plotted *vs.* minimum base changes on a special scale which is constructed such that a Gaussian distribution will produce a straight line. The slope of the line is related to the standard deviation. If the tail of the Gaussian curve is distorted this will appear as a deviation from the straight line which is greatly accentuated. Two representative plots of this kind are shown in Fig. 6. These have been chosen to illustrate relatively clear-cut cases. The first example is a comparison of sequences of 20 residues of *Candida krusei* cytochrome *c* (Narita and Titani, 1965) with other sequences chosen from the same protein. The comparison length is 20 residues. It can be seen that the observed frequency distribution can be fitted very well by a straight line. This indicates that no significant self-homology can be claimed for this protein. The second case has been chosen to demonstrate what happens when two proteins which are closely homologous are compared. Here human Bence-Jones

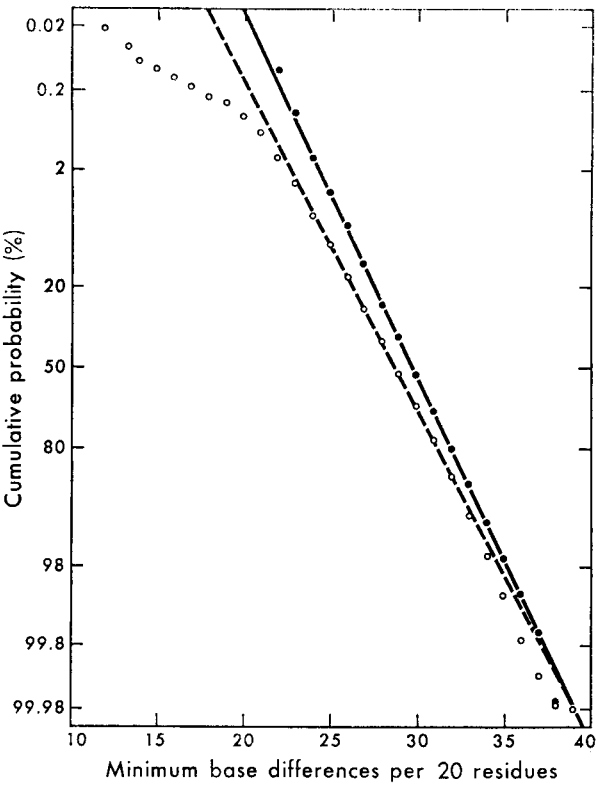


FIG. 6. Comparisons of protein sequences by computer. *C. krusei* cytochrome *c* shows no self-homology while κ and λ immunoglobulins demonstrate considerable similarity. ——●—— cytochrome *c*, *C. Krusei vs.* self. ---○--- IgG, κ *vs.* λ.

protein λ (Wikler *et al.*, 1967) has been compared with κ (Titani *et al.*, 1967) using a length of 20 residues. In the region of large numbers of differences, the data falls close to a straight line. As smaller values for differences are approached, the observed curve sharply deviates from the line. This is evidence for homology between the two sequences.

The method we have just described is very simple to use but two difficulties often arise. In many cases the data do not seem to fit a straight line in any region of the curve. This makes it hard to determine whether any homology is present. A second problem is that the use of probability paper does not provide any quantitative estimate of the significance of the homology between two proteins. In an attempt to circumvent these problems, we have devised an alternative method of plotting the results of protein comparisons. This is based on the premise that what one really is interested in comparing is the difference between

the observed distribution and the results calculated from a knowledge of the amino acid composition. We have chosen to plot two ratios of the observed and calculated frequencies, P_{obs}/P_{calc}, *vs.* minimum base differences on a logarithmic scale. This scale has not been chosen for any statistical reasons. It simply permits a convenient representation of the data. If two sufficiently large sets of random sequences are compared, a plot of log (P_{obs}/P_{calc}) *vs.* minimum base differences would be a straight line of zero slope. P_{obs}/P_{calc} would have a value of near 1.0 for every number of minimum base differences. If the set from which the sequences are chosen is smaller, the data should still fall about the same strength line although deviations due to random fluctuations would be seen. Any homology between two proteins would be evidenced by a pronounced deviation from the line in the region of small numbers of minimum base differences. In this case, the magnitude of the deviation is clearly related to the probability that observed frequency exceeds the frequency that would have arisen by chance. Two representative plots obtained by this method are shown in Fig. 7. These have again been chosen to depict extreme examples. The first, a comparison with no apparent homology, is sheep pituitary β-lipotropic hormone (Li *et al.*, 1965) *vs. Clostridium pasteurianum* ferredoxin (Tanaka *et al.*, 1964). The second, a case in which homology is fairly well established (Fitch, 1966b), is human β-hemoglobin *vs.* itself (Braunitzer *et al.*, 1961a,b; Goldstein *et al.*, 1963), with identical comparisons excluded. For both examples, a comparison length of 20 was used. From the results shown for β-hemoglobin in Fig. 7, it can be seen that comparisons of 20 residues involving 15 minimum base differences occurred almost 100 times as often as predicted by our statistical model. This is clear evidence for self-homology.

At this point a specific example may help to clarify the methods we have used. All possible sets of 19 amino acids from *C. butyricum* ferredoxin (Benson *et al.*, 1966) were compared with all sequences of the same length from alfalfa ferredoxin (Keresztes-Nagy *et al.*, 1968). Since the lengths of these proteins are 55 and 97 residues, respectively, this resulted in a total of 2923 comparisons. Of all of these, one particular comparison showed an unusually low number of minimum base changes. There were a total of 13 base changes for the 19 residues compared. The amino acids involved are shown below:

Alfalfa	-Gly-	Ser-	Cys-	Ser-	Ser-	Cys-	Ala-	Gly-	Lys-	Val-	Ala-	Ala-	Gly-	Glu-	Val-	Asn-	Gln-	Ser-	Asp-
C. butyricum	-Val-	Ser-	Cys-	Gly-	Ala-	Cys-	Ala-	Gly-	Glu-	Cys-	Pro-	Val-	Ser-	Ala-	Ile-	Thr-	Gln-	Gly-	Asp-
MBDC	1	0	0	1	1	0	0	0	1	2	1	1	1	1	1	1	0	1	0

From the amino acid compositions of the two proteins, the probabilities of finding 0, 1, 2, or 3 minimum mutations when two amino acids picked

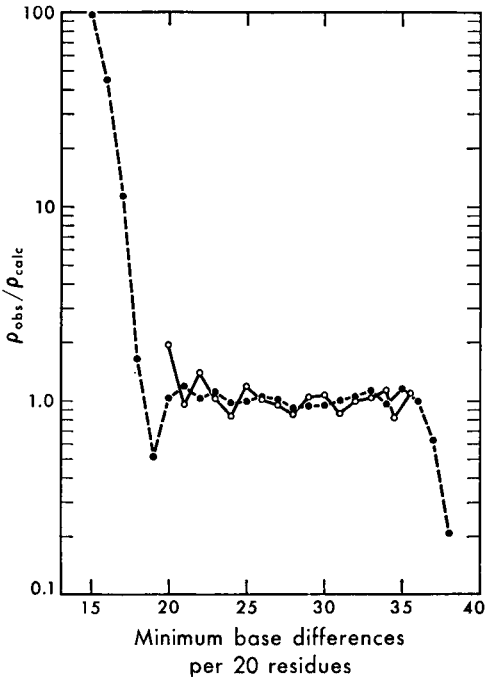


FIG. 7. Comparisons of protein sequences by computer. The vertical scale indicates deviations from randomness. No homology is found when β lipotropic hormone is compared with *C. pasteuranium* ferredoxin. Substantial self-homology is indicated for the human hemoglobin β chain. ---●--- βHb *vs.* self. ——○—— βLH *vs.* ferredoxin.

at random are compared could be calculated. The results are: $P(0) = 0.068$; $P(1) = 0.419$; $P(2) = 0.502$; $P(3) = 0.011$. This now permits us to estimate that the probability of 13 or less minimum mutations out of 19 residues is 5.85×10^{-7}. Since 2932 comparisons were made, the average number of times 13 out of 19 would be observed is 5.85×10^{-7} multiplied by $2923 = 1.7 \times 10^{-3}$. Thus we have observed an homology which would occur only once in 588 times by chance. This is fairly substantial evidence that the two ferredoxins evolved from a common ancestor.

Thus far the methods of searching for protein homology we have discussed all involve the comparison of a large number of short sequences chosen from the two proteins under consideration. An alternative approach has recently been taken by Needleman and Wunsch (1968). They have developed a scheme by which the entire chains of proteins are compared at once. A matrix is constructed with rows and columns repre-

senting the sequences of the two proteins to be compared. The elements of the matrix are determined by the criteria used to compare the sequences. If these are minimum base differences, then each matrix element would be either a 0, 1, 2 or 3. For two short peptides, Lys-Ile-Val-Ser-Asp- and Lys-Leu-Glu-Asp-Lys, the comparison matrix is shown below:

	Lys	Ile	Val	Ser	Asp
Lys	[0]	1	2	2	2
Leu	2	[1]	1	1	2
Glu	1	2	[1]	2	1
Asp	2	2	1	2	[0]
Lys	0	1	2	2	2

The computer attempts to find a path through the matrix which minimizes the sum of the elements through which it passes; for the matrix above the residues on this path are enclosed in boxes. This leads to the comparison

	Lys-	Ile-	Val-	Ser-	Asp
	Lys-	Leu-	Glu-	-Asp-	Lys
MBDC	0	1	1		0

In many cases there may not be a uniquely short path. After comparing the two protein sequences, Needleman and Wunsch randomize them and run matrix comparisons again on 10 sets of scrambled sequences. The resulting values for minimum base differences are compared with that found for the unaltered sequence to test for any evidence of homology. This method is indeed capable of finding ways of aligning the sequences to maximize homology. It does this, though, at the possible expense of putting a large number of gaps into the sequences such as the one shown in the case above. The statistical effects of these gaps and ways of testing for their significance will be discussed later.

Many additional methods of comparing protein sequences will undoubtedly be proposed in the future. It is too early to say which types of methods are likely to be the most successful. We hope, of course, that when different methods are applied to the same systems they will yield similar conclusions. Manwell (1967) has recently demonstrated several different methods which are consistent in indicating a certain amount of homology between bovine pancreatic ribonuclease and chicken egg white lysozyme. Any useful method should fulfill several requirements: it should be capable of finding any homology that may exist; it should involve a large enough number of comparisons to produce statistically significant results; and it should be based on a set of well-

defined criteria, rather than just the intuitive principles that have frequently been used in the past.

Often when two protein chains are compared, there are extensive areas of homology in widely spaced regions of the protein, but if the two protein sequences are placed side by side the number of residues between the homologous regions is different on the two chains. This implies that some amino acid residues have been genetically inserted into one chain or deleted from the other. This phenomenon was first discussed by Braunitzer *et al.* (1961b). Several well-known examples found in the globulin chains will be discussed in later sections of this chapter. Various genetic mechanisms which account for this phenomenon were discussed earlier in the text. Our concern here is how to collect evidence that these gaps actually exist. In cases like the hemoglobins, where the homology between the chains on both sides of the gap is very extensive, a great deal of confidence can be placed in the existence of the gap. When two fairly dissimilar proteins are compared, however, the issue can become much more controversial. Many workers have attempted to improve the apparent homology between two protein chains by the insertion of gaps into one or both of the sequences. Often very large numbers of gaps have been proposed. The difficulty is that by judicious choice of gaps it is always possible to improve the homology between two protein chains. Consider what happens when two gaps are placed into the randomly chosen sequences used earlier in this chapter (Eq. 1) and the most homologous arrangements are found. Without the gaps, a minimum of four identical amino acids can be aligned. Through the use of two gaps, three more can be made to coincide. Using the Poisson distribution one can estimate that the probability of an alignment with four identical residues is $e^{-1}/4!$, while seven residues will be identical only once in $e^{-1}/7!$ times. Have we found a significant homology? Obviously we have not, since the sequences were chosen at random.

What must be calculated is the number of times a given method of comparing two protein chains would be expected to produce an alignment containing a specified number of identical amino acids or minimum base differences. When gaps are allowed as part of the comparison protocol, a much larger total number of comparisons is possible than without the gaps. Recall that the calculated frequency of observing a given comparison is equal to the product of the a priori probability of picking such a comparison and the total number of comparisons that have been made. When we compare sequences of length L chosen from two chains of length N_1 and N_2, this second factor, which we derived previously, is $(N_1 - L + 1)\ (N_2 - L + 1)$. If gaps are to be included, a much

larger number of comparisons is possible. For example, if a gap of one residue is placed on a chain of 20 amino acids, there are 19 different places where the gap can be located. Thus the sequence with a gap can be compared with other sequences in 19 times as many ways as an uninterrupted sequence. With a gap in each sequence, as in the above example, 19^2 times as many comparisons are possible. Instead of comparing ungapped and gapped events with frequencies of $e^{-1}/4!$ and $e^{-1}/7!$, respectively, we must compare events with relative occurrences of $e^{-1}/4!$ and $19^2e^{-1}/7!$ Thus our comparison with gaps will occur on a random basis about twice as often as the ungapped comparison. This suggests that there is little or no significance to the gaps in the above example.

The approach used above can be extended to many cases of general interest. Roughly, the number of possible comparisons increases as L^{Ng}, the length of the comparison to the power of the number of gaps. To be significant, a series of gaps must drastically decrease the a priori probability of finding the comparison by a random process.

Even with the strict constraints outlined above, some gaps that have been proposed can readily be shown to be significant. An example chosen from the immunoglobulins is shown in Table XXV. The 20 C-terminal residues of Bence-Jones λ (Wikler *et al.*, 1967) were compared with the 22 C-terminal amino acids of Bence-Jones κ (Titani *et al.*, 1967). With no gaps, the most favorable alignment shown below results in 20 minimum base differences in the 20 codons compared. If a single gap of 2 residues is inserted into the λ chain between residues 201 and 202, only 13 minimum base differences are needed for the 20 residues compared. The distributions of minimum base differences and a priori prob-

TABLE XXV

MINIMUM BASE DIFFERENCES PER CODON (MBDC) IN COMPARISONS OF SEGMENTS OF IMMUNOGLOBULIN CHAINS, WITH AND WITHOUT A GAP

λ-no gaps	Ser-	Cys-	Gln-	Val-	Thr-	His-	Glu-	Gly-	Ser-	Thr-	Val	
MBDC	1	0	1	0	0	0	1	0	1	1	2	
κ	Ala-	Cys-	Glu-	Val-	Thr-	His-	Gln-	Gly-	Leu-	Ser-	Ser	
MBDC	1	0	1	0	0	0	1	0			0	
λ-gap	Ser-	Cys-	Gln-	Val-	Thr-	His-	Glu-	Gly-			-	Ser
λ-no gaps	Glu-	Lys-	Thr-	Val-	Ala-	Pro-	Thr-	Glu-	Cys			
MBDC	2	2	0	2	1	2	1	2	1			
κ	Pro-	Val-	Thr-	Lys-	Ser-	Phe-	Asn-	Arg-	Gly-	Glu-	Cys	
MBDC	1	0	2	0	1	1	2	1	2	0	0	
λ-gap	Thr-	Val-	Glu-	Lys-	Thr-	Val-	Ala-	Pro-	Thr-	Glu-	Cys	

Totals: κ *vs.* λ-no gaps: 20 MBDC/20 codons

κ *vs.* λ-1 gap: 13 MBDC/20 codons

ability calculated from the amino acid compositions of the two proteins are shown below:

Comparison	Minimum base differences per codon				$P(L,n)$
	0	1	2	3	
λ *vs.* κ no gaps	6	8	6	0	1.47×10^{-4}
λ *vs.* κ 1 gap	10	7	3	0	1.04×10^{-8}

By inserting one gap, we arrive at a comparison which is more than 10^4 times as rare. The gap in a chain of 22 amino acids permits 22 times as many comparisons. The net result is a comparison which should occur 640 times less frequently than the ungapped comparison. Even if one inserts the gap on both strands, the homology with the gap is still hundreds of times as rare. This is fairly clear evidence that the postulated gap has some statistical significance.

Since computer searches have been very successful in studying the homology of uninterrupted protein chains, one might expect that similar methods should prove useful in searching for the most favorable locations of possible gaps. We have developed the programs for handling such an approach, but they have two serious limitations. Since the number of comparisons increases as L^{Ng}, the computation time does too. Present computer speeds usually restrict one to inserting only 1 gap in proteins of up to 100 residues. Even so, it is sometimes possible to evaluate the location and size of gaps by using a computer. Several calculations for illustrative purposes were performed on the first 40 residues of human hemoglobin α and β chains. A comparison length of 16 residues was used. A gap of either 1, 2, or 3 residues was allowed to occur in all possible positions of either the α or the β chain. Plots of the cumulative frequency of comparisons as a function of minimum mutations are shown in Fig. 8; for two of these cases, a gap of 2 residues is placed in α or β. Both distributions shown in Fig. 8 are nonlinear, which reflects the well-known homology between the α and β chains. The possibility of a 2-residue gap in β results in a large number of comparisons with very small numbers of minimum mutations, as evidenced by the shift in the distribution to the left. When gaps of 1 or 3 residues are used, no such shift of β relative to α is observed. The magnitude of the shift is evidence of the definite existence of a gap. Since the shift occurred only with a gap of 2 in β, this identifies the length of the gap and the chain on which it occurred. An examination of the detailed output from the programs permits the actual location of the gap to be found.

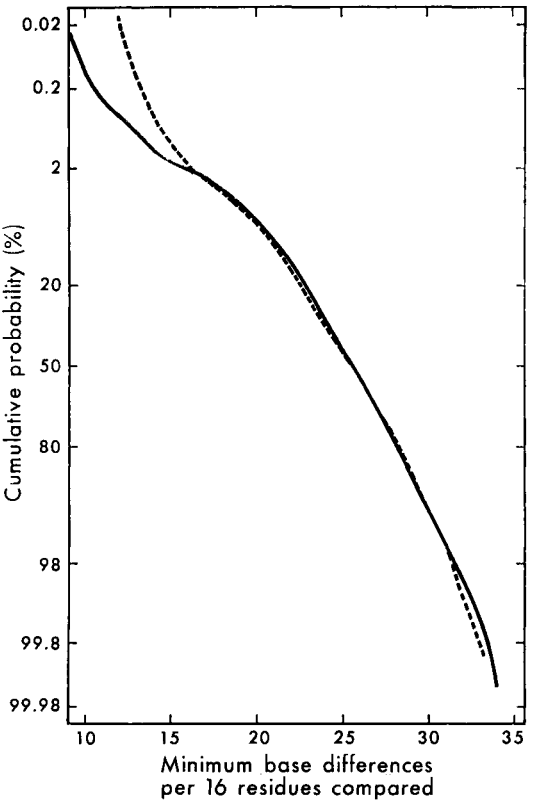


FIG. 8. Computer comparison of the first 40 amino acids of α and β human hemoglobin chains including a gap of 2 residues on either one of the chains. The shift in the β distribution is evidence for a gap of 2 on the β chain.——gap of 2 on β----gap of 2 on α.

This is consistent with the results found by hand, as shown later in the text. In conclusion, we have shown that there is often considerable evidence for the presence of gaps or insertions in protein chains. Each gap, however, must be judged on its own merits. The indiscriminate placement of a large number of gaps into protein sequences will almost always increase the apparent homology but it may lead to comparisons that are not statistically significant.

Our discussion of protein sequence comparison thus far has been very general. Now we would like to consider a specific set of protein sequences. From the results of computer comparisons, we should be able to show the alignment of several protein chains which leads to maximum homology. Some of the data will clearly point to the occurrence of gaps in

certain of the protein chains. The examples we shall consider consist of three relatively long protein chains chosen from the immunoglobulins. These are two complete light chain sequences, human Bence-Jones λ with 214 residues (Titani *et al.*, 1967), human Bence-Jones κ with 213 (Wikler *et al.*, 1967), and the rabbit F_c fragment of length 161 from the heavy immunoglobulin chain (Hill *et al.*, 1966). With three possible protein chains, there are six sets of comparisons which have to be made. Three compare one part of the same sequence with another. The others search for intersequence homology. All six comparisons were carried out on an IBM 7094 digital computer, using a comparison length of 20 residues. Only about 20 minutes of computer time were needed for this project. This included calculations of probability distributions as well as the actual sequence comparisons. It may be helpful in the discussion that follows to refer to the final proposed arrangement of sequences (Fig. 9).

Two general rules were followed to permit a self-consistent approach in assembling the immunoglobulin chains to maximize homology. The regions showing maximum homology were selected by a preliminary computer search. These were matched in order of decreasing homology as determined by minimum base differences per codon for the regions involved. At first, gaps were introduced only when clearly indicated by the computer comparisons. Later, gaps were added only after it could be shown that these were statistically reasonable. The computer results indicated that there was strong homology between several regions of κ and λ Bence-Jones. Additional clear evidence for homology was found between sections of both κ and λ chains and the immunoglobulin F_c fragment. There was some evidence for self-homology in the F_c fragment, but since this was weak we shall ignore it at first. Little or no evidence could be found for self-homology in the κ and λ Bence-Jones protein sequences. These results are summarized in Fig. 10.

The sequence comparisons showing the greatest homology are summarized in Fig. 11. The results shown include virtually all of the significant comparisons between κ and λ and some of the homology found between F_c and either κ or λ. The format of Fig. 11 is essentially a spectrum of homology. The horizontal axis gives the absolute location of the first-named sequence. The numbers below the sets of comparisons indicate the stagger between the two sequences. For example, a stagger of 2 might indicate that the sequence beginning with residue 10 on the κ chain was compared with the sequence beginning with residue 8 on the λ chain. One can see from the vertical axis that this comparison resulted in 17 minimum base changes per 20 codons. As a rough rule of thumb, one can estimate that single comparisons of 20 residues which involve 18


1 10 20
λ Ser-Glu-Leu-Thr-Gln-Asp-Pro-Ala-Val-Ser-Val-Ala-Leu-Gly-Gln-Thr-Val-Arg-Ile-Thr-Cys-Gln-Gly-
Fc
Fc
1 10 20
κ(HAg) Asp-Ile-Gln-Met-Thr-Gln-Ser-Pro-Ser-Ser-Leu-Ser-Ala-Ser-Val-Gly-Asp-Arg-Val-Thr-Ile-Thr-Cys-Gln-Ala-
30 40
λ -Asp-Ser-Leu-Arg-Gly-Tyr-Asp-Ala-Ala-Trp-Tyr-Gln-Gln-Lys-Pro-Gly-Gln-Ala-Pro-Leu-Leu-Val-Ile-Tyr-Gly-
Fc
Fc
30 40 50
κ(HAg) -Ser-Gln-Asx-Ile-Asx-Ser-Phe-Leu-Asn-Trp-Tyr-Gln-Gln-Gly-Pro-Lys-Lys-Ala-Pro-Lys-Ile-Leu-Ile-Tyr-Asp-
50 60 70
λ -Arg-Asn-Asn-Arg-Pro-Ser-Gly-Ile-Pro-Asp-Arg-Phe-Ser-Gly-Ser-Ser-Ser-Gly-His-Thr-Ala-Ser-Leu-Thr-Ile-
1 10 20
Fc Thr-Ala-Arg-Pro-Pro-Leu-Arg-Glu-Gln-Gln-Phe-Asp-Ser-Thr-Ile-Arg-Val-Val-Ser-Thr-Leu-Pro-
Fc
60 70
κ(HAg) -Ala-Ser-Asn-Leu-Glu-Thr-Gly-Val-Pro-Ser-Arg-Phe-Ser-Gly-Ser-Gly-Phe-Gly-Thr-Asp-Phe-Thr-Phe-Thr-Ile-

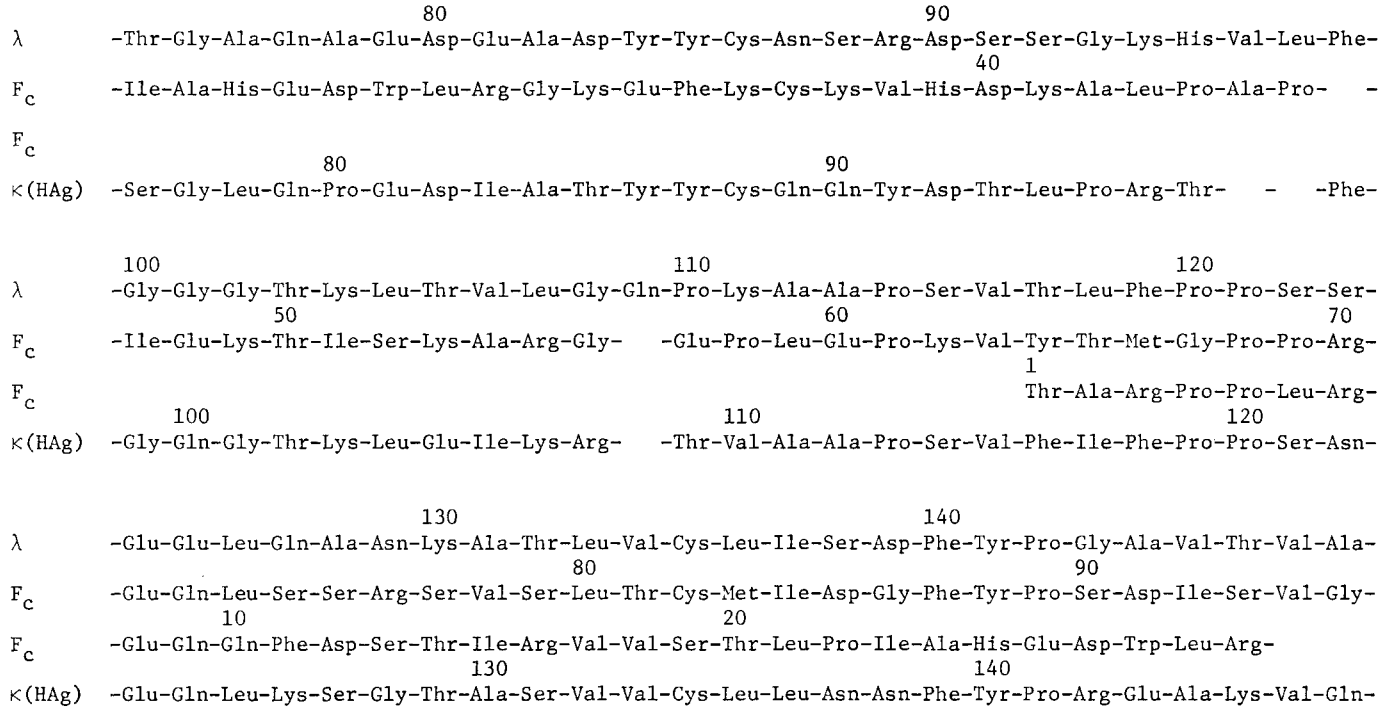


FIG. 9. See p. 116 for legend. *(continued)*

FIG. 9. Comparison of sequences of λ Bence-Jones, F_c, and κ(HAg) segments of immunoglobulins.

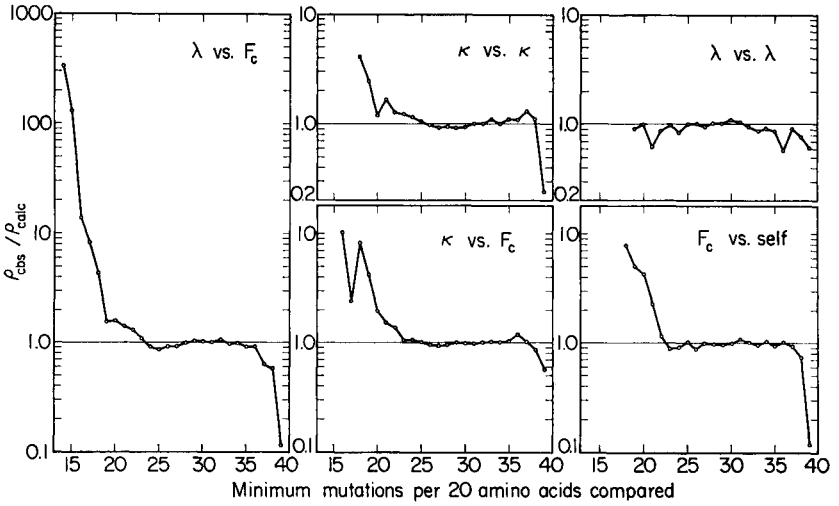


FIG. 10. Comparisons of κ and λ Bence-Jones proteins and the F_c fragment of heavy human IgG.

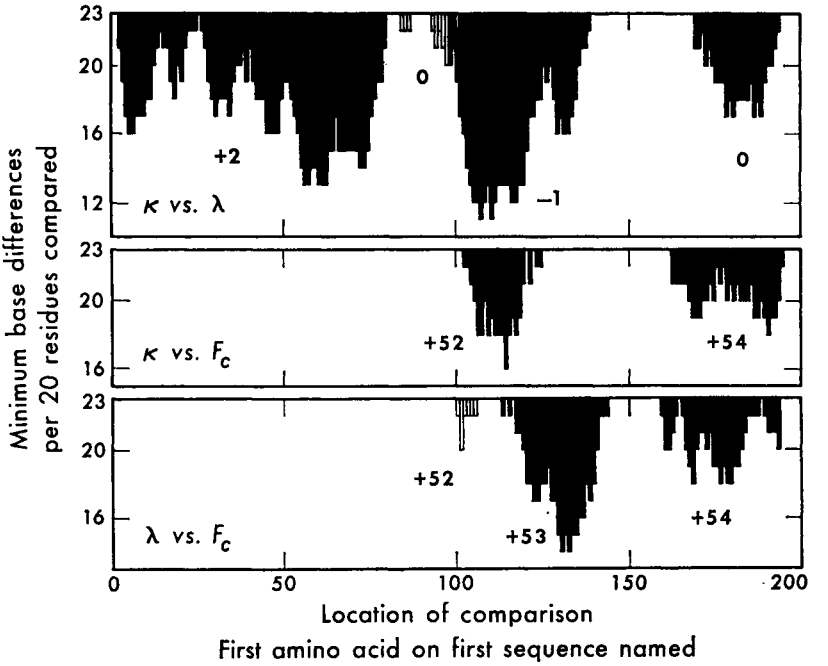


FIG. 11. Details of some of the comparisons given in Fig. 10. Each black vertical bar is the result of comparing 20 residues of one sequence with 20 of another. Details given in the text.

or fewer minimum mutations are usually evidence of homology. A long string of successive comparisons which averages 20 or less minimum mutations per comparison is usually also meaningful. Using these criteria, all of the comparisons plotted in Fig. 11 are very good evidence of homology. It is apparent, from Fig. 11, that the greatest immunoglobulin homology exists between κ and λ Bence-Jones chains. The data in Fig. 11 can be used to align the two sequences. Since the stagger between the sequences which leads to maximum homology is not constant throughout the sequences, gaps must be introduced in the κ and λ chains. For example, between residues 105 and 140 the homology between λ and κ is maximized if the nth amino acid on κ is compared with the $(n+1)$th residue of λ. Between residues 1 and 90 on κ, the nth residue of the κ chain must be aligned with the $(n-2)$th residue of λ in order to preserve this homology. Thus it is clear that somewhere between residues 90 and approximately 115 there must be 3 extra residues on the λ chain. Gaps corresponding to these residues must be introduced into the κ chain in order to maintain an alignment between the two sequences that maximizes homology. From Fig. 11 it can be seen that several sets of sequences in the region compare favorably when there is no stagger between them. This suggests that the gap of 3 in κ occurs in 2 pieces. The first, a gap of two, will permit some residues between about 95 and 121 to be aligned with considerable homology. The second, a gap of one, is necessary to complete the total of three which must come between 90 and 115. An examination of the actual sequences leads after some trial and error to the conclusion that there is a gap of 2 residues between amino acids numbers 97 and 98 of the κ chain, and an additional gap of 1 residue probably between 108 and 109 of the κ chain. This gap is further justified by homology with F_c discussed later. The remaining comparisons between κ and λ, shown in Fig. 11 demonstrate the presence of homology between sequences near the C-termini of these proteins with the two sequences aligned perfectly in phase. Since we have just shown that the centers of the proteins are out of phase by 1 residue, there must be a gap in the λ chain somewhere between residues 150 and 170. It is not possible to locate this gap unequivocally from the data given in Fig. 11, but a visual comparison of the sequences along with some of the results from the F_c fragment suggests that it is near the end of this region and we have tentatively decided to place it between residues 169 and 170 of the λ chain. This completes the extent to which we can align the λ and κ chains from the direct results of the computer calculations.

The only place where no significant homology has been found thus far is the C-terminal ends of the proteins. That there is probably homology here too has been shown earlier in this chapter. If a gap of 2 residues

is placed between amino acids 202 and 203 of λ, 4 out of the 10 terminal amino acids of λ and κ are identical. The statistical significance of this gap has already been shown. The tentative arrangement of the λ and κ chains which we have now reached is shown in Fig. 12a. A total of 4 gaps was needed. With these, there are 188 minimum base differences for 209 codons compared. This corresponds to a ratio of 0.90 base differences per codon, a very low number, indicative of substantial homology. This homology is extensive enough to be easily seen by inspection; 78 of the 209 amino acids are identical.

Next, we turn our attention to the comparison between the F_c fragment and the λ and κ chains shown in Fig. 11. These will be used to align the F_c fragment relative to the arrangement of κ and λ chains we have already postulated. First examine the comparison between κ and F_c, and recall that there is thus far no evidence for a gap in the κ chain between residues 106 and the C-terminus. Two regions of homology are found between κ and F_c. The first, with a stagger of 52 between the chains, occurs between residues 102 and 145 of κ. The second, with a stagger of 54 occurs between residue 162 and the C-terminus. If both of these are to occur simultaneously, there must be a gap of two residues on the F_c chain, somewhere between amino acids number 73 and 108 or so. From a comparison of κ and F_c alone it would be difficult indeed to locate this gap any more closely. Consider, however, the comparisons shown in Fig. 11 between the F_c fragment and λ Bence-Jones protein. There are three regions of homology: a very weak section near λ-100 with a stagger of 52, and stronger regions with a stagger of 53 from λ-113 to 154 and a stagger of 54 from about λ 162 to the C-terminus. We had previously placed a gap in the λ chain between residues 169 and 170. But the homology between λ and F_c shown in Fig. 11 is uninterrupted in this region. To maintain the homology between λ and F_c this gap must also have occurred in F_c. This means that there is a gap of 1 between residues 115 and 116 of the F_c chains. This is one of the two gaps that we needed in the F_c chain to align it with the κ Bence-Jones protein amino acid sequence. Where is the second one? The other region of homology between λ and F_c can permit us to locate this gap. This occurs from λ-113 and 154 which corresponds to F_c-63 to 101. The second gap therefore lies between λ 101 and 115. An examination of the actual sequences suggests that the most likely position for the gap is between residues 107 and 108 of the F_c chain. One more piece of evidence remains: the weak homology between residues 100 to 120 of λ and 48 to 68 of the F_c fragment. Since there is a stagger of 52 for this comparison and a stagger of 53 for the neighboring comparison of λ and F_c there must be an additional gap in the F_c chain somewhere in this region.

However, the comparison between this region of F_c and the κ chain shows an uninterrupted homology. But we have already shown that there is a gap in κ between residues 108 and 109. It is quite reasonable to suppose that the gap on F_c we are looking for corresponds to this gap. This occurs between residues 56 and 57 on the F_c chain. It is the last gap which can easily be supported by the direct computer comparisons shown in Fig. 11. We have now established the relative orientation of the λ, κ, and F_c chains which maximizes homology for all regions save for the C-terminal sequences. The only problem which remains is whether there is a gap on the F_c chain which corresponds to the gap of 2 which we have already placed on the λ chain between residues 202 and 203. The complete evidence will be presented later, during our discussion of the self-homology of the F_c fragment, so let us assume for the moment that there is no gap on the F_c chain in this region. This completes our alignment of the κ, λ and F_c amino acid sequences. The results thus far are summarized in Fig. 12b.

We can now use our arrangement of the immunoglobulin sequences to search for possible internal homology in some of the polypeptide chains. As mentioned earlier, there is little evidence for such homology with the λ and κ chains but some homology seems to exist between different sequences chosen from the F_c fragment. Two regions in which self-homology seems possible appear when F_c residues 1 to 21 are compared with 64 to 88, and residues 12–47 are compared with 116–150. The homology in these regions is not very strong, involving 21 base differences in 21 codons and 34 base differences in 36 codons, respectively. If no other evidence existed, one might easily be tempted to ignore the possibility of self-homology. Additional hints for this homology can be found, however, when the sequences of the F_c fragment are compared with the λ and κ chains. We have already discussed the most striking results of these comparisons which were used to align the three immunoglobulin chains. Other relatively strong homology exists, which is summarized below, and in Fig. 12c. Since we have already aligned F_c

Additional Immunoglobulin Homology

F_c 12 – 46	F_c 4 – 24
κ 170–204	κ 119–139
31 base differences	21 base differences
35 codons	21 codons

relative to λ and κ we can use this additional homology to try to orient one part of the F_c chain relative to another part. This assumes that the homology presented above is entirely due to internal homology in the

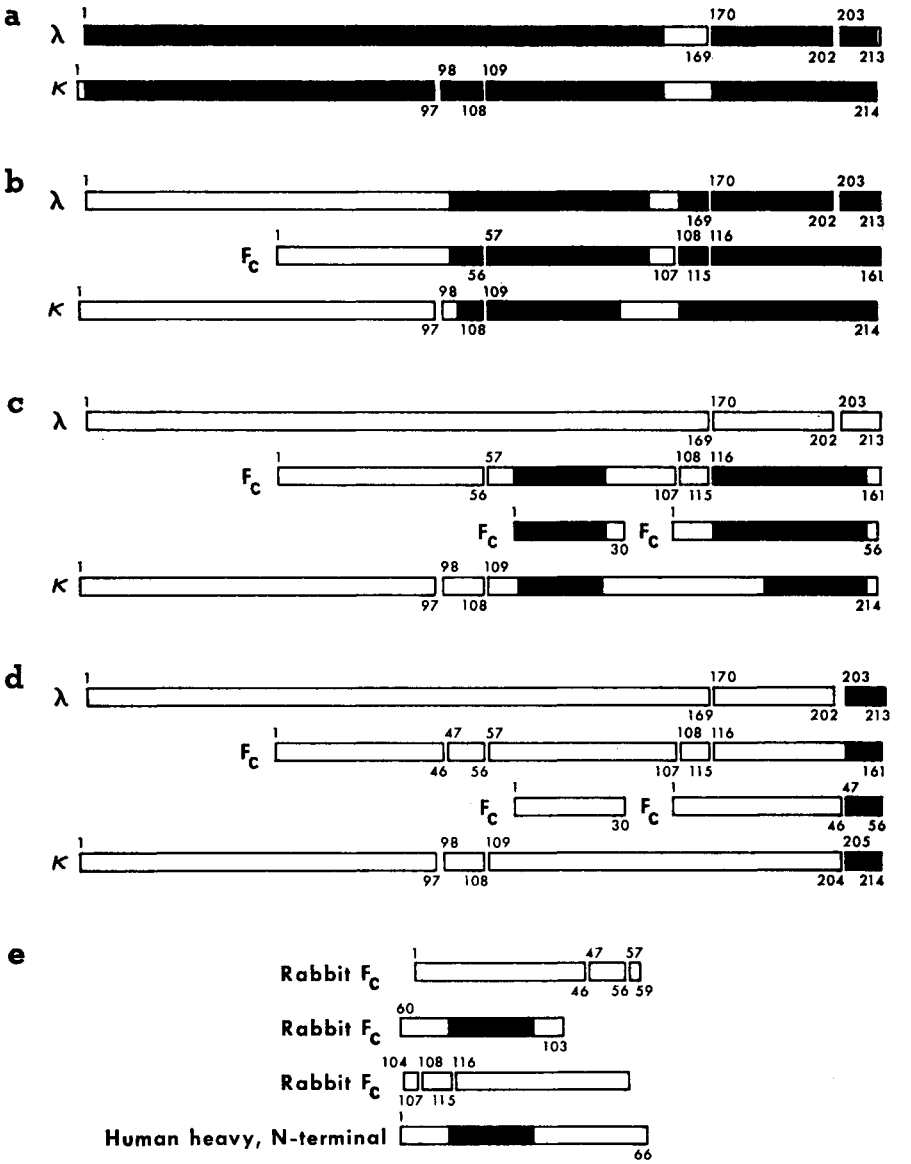


FIG. 12. Stages in elaborating the homologies among the immunoglobulin chains. (a) Comparisons of λ and κ. (b) Comparisons between F_c and λ or κ. (c) Comparisons between FC residues 1 to 56 and other regions of the F_c fragment. (d) C-terminal comparisons. (e) Comparison of F_c with the N-terminal sequence of human heavy immunoglobulin chain. The F_c sequences are from rabbit immunoglobulin. The shaded areas show regions of homology discovered or used at each stage.

F_c chain. We found previously that residues 64 to 160 of F_c were homologous to 116 to 214 of the κ Bence-Jones protein amino acid sequence. Now we also see that residues 4 to 24 are homologous to 119 to 139 of the κ chain. These two results are perfectly consistent and suggest that residues 1 to 25 of the F_c chain may be homologous to residues 64 to 88 of the same chain. Are they? This is the very set of residues for which self-homology was suspected as the result of the direct self-comparison of the F_c chain discussed above. Thus the homology between the F_c and κ sequences reinforces the possibility of self-homology in the F_c chain. Similar evidence has been used in the past to find self-homology in the cytochromes. The other set of comparisons between F_c and κ tabulated above is likewise completely consistent with the possibility of the self-homology in the F_c chain occurring between residues 12–53 and 116–157. These results are summarized in Fig. 12c.

Thus far, we have used the available homologies among the κ, λ, and F_c chains to align the three chains, find evidence for gaps, and propose the possibility of internal duplication in the F_c sequence. The problem which remains is to work out the details of this self-homology. To do this, we must first reexamine the troublesome C-terminal regions of these proteins. It is always difficult to obtain clear-cut evidence from computer comparisons for homologies at the ends of protein chains because the number of comparisons that the ends can be involved in is necessarily small. Thus more detailed comparison of the sequences and statistical estimates of their significance must be relied upon. It has already been shown that there is a gap in the λ chain between residues 202 and 203. Let us return to reexamine this point. In Table XXVIA the C-terminal sequences of κ, γ, and two self-homologous regions of the F_c chain are compared as aligned in Fig. 12c but with the gap in λ absent. There is substantial homology at the beginning of these sequences, but toward the end little evidence remains. There are 133 minimum base differences in 123 amino acid residues that can be compared among the four chains. Consider now the effect of the placement of two gaps as shown in Table XXVIB. The first of these occurs only on the λ chain and is the same gap between residues 202 and 203 that we postulated before. The second occurs in all the chains except for the C-terminal fragment of the F_c chain. Thus this gap may be considered to be an insertion of tyrosine 151 between a histidine and a threonine on the F_c chain. With these two gaps, the total sequence comparison becomes 102 base changes out of 120 codons compared. This is a striking improvement for only two gaps. The probability of random occurrence of the comparison is diminished by roughly 5^{14}. In return, we have a

TABLE XXVIA
IMMUNOGLOBULINS: C-TERMINAL HOMOLOGY—NO GAPS

Sequences				Comparisons					
λ 194	F_c 36	F_c 140	κ 194	λ κ	F_c36 F_c140	λ F_c36	κ F_c36	λ F_c140	κ F_c140
Cys	Cys	Cys	Cys	0	0	0	0	0	0
Gln	Lys	Ser	Glu	1	2	1	1	2	2
Val	Val	Val	Val	0	0	0	0	0	0
Thr	His	Met	Thr	0	3	2	2	1	1
His	Asp	His	His	0	1	1	1	0	0
Glu	Lys	Glu	Gln	1	1	1	1	0	1
Gly	Ala	Ala	Gly	0	0	1	1	1	1
Ser	Leu	Leu	Leu	1	0	1	0	1	0
Thr	Pro	His	Ser	1	1	1	1	2	2
Val	Ala	Asn	Ser	2	2	1	1	2	1
Glu	Pro	His	Pro	2	1	2	0	2	1
Lys	Ile	Tyr	Val	2	2	1	1	2	2
Thr	Glu	Thr	Thr	0	2	2	2	0	0
Val	Lys	Glu	Lys	2	1	2	0	1	1
Ala	Thr	Lys	Ser	1	1	1	1	2	2
Pro	Ile	Ser	Phe	2	1	2	1	1	1
Thr	Ser	Ile	Asn	1	1	1	1	1	1
Glu	Lys	Ser	Arg	2	2	1	1	2	1
Cys	Ala	Arg	Gly	1	2	2	1	1	1
Ser	Arg	Ser	Glu	2	1	1	2	0	2
	Gly	Pro	Cys		2		1		2
				21	26	24	19	21	22

Total MBDC: 133
Comparisons: 123 Average = 1.08

situation where there are about 20^4 more possible comparisons. The net result, however, is to produce a homology which is far less likely to have occurred by a random process. The result is shown in Fig. 12d, when the gaps discussed above are incorporated.

With the placement of this last gap, we now have an extended region of homology between residues 17 to 56 of the F_c and 121 to 161 of F_c. When this is added to the evidence of homology between residues 1–25 and 64–88, the result shown in Fig. 12e is produced. This alignment scheme, consistent with all of our evidence, suggests that the F_c fragments evolved as a trimer from a smaller polypeptide chain. The homology between residues 1–25 and 64–88 is rather weak, however, so it may be that this comparison is simply fortuitous. In this case the internal

TABLE XXVIB
IMMUNOGLOBULINS: C-TERMINAL HOMOLOGY—GAPS INSERTED

Sequences				Comparisons					
λ 194	F_c 36	F_c 140	κ 194	λ κ	F_c36 F_c140	λ F_c36	κ F_c36	λ F_c140	κ F_c140
Cys	Cys	Cys	Cys	0	0	0	0	0	0
Gln	Lys	Ser	Glu	1	2	1	1	2	2
Val	Val	Val	Val	0	0	0	0	0	0
Thr	His	Met	Thr	0	3	2	2	1	1
His	Asp	His	His	0	1	1	1	0	0
Glu	Lys	Glu	Gln	1	1	1	1	0	1
Gly	Ala	Ala	Gly	0	0	1	1	1	1
Ser	Leu	Leu	Leu	1	0	1	0	1	0
Thr	Pro	His	Ser	1	1	1	1	2	2
—	Ala	Asn	Ser	—	2	—	1	—	1
—	Pro	His	Pro	—	1	—	0	—	1
—	—	Tyr	—						
Val	Ile	Thr	Val	0	1	1	1	2	2
Glu	Glu	Glu	Thr	2	0	0	2	0	2
Lys	Lys	Lys	Lys	0	0	0	0	0	0
Thr	Thr	Ser	Ser	1	1	0	1	1	0
Val	Ile	Ile	Phe	1	0	1	1	1	1
Ala	Ser	Ser	Asn	2	0	1	1	1	1
Pro	Lys	Arg	Arg	1	1	2	1	1	0
Thr	Ala	Ser	Gly	2	1	1	1	1	1
Glu	Arg	Pro	Glu	0	1	2	2	2	2
Cys	Gly	Gly	Cys	0	0	1	1	1	1
				13	16	17	19	17	19

Total MBDC: 101
Comparisons: 120 Average = 0.84

repetition in the F_c chain would appear to be a chain doubling with subsequent mutations obscuring most of the internal homology. A choice between these two possible models cannot be clearly made at present. The resolution of this problem awaits more sequence data on other immunoglobulins, and especially on other parts of the heavy immunoglobulin chain. All of our results to date can be summarized by the detailed alignment of sequences presented in Fig. 9.

Recently the amino acid sequence of the N-terminal fragment of human Daw heavy immunoglobulin has been determined (Press, 1967). Residues 14 to 36 of this fragment are homologous to 74 to 96 of the F_c fragment. For 23 residues compared there are only 19 minimum base differences. This permits us to align the N-terminal fragment as shown

in Fig. 12e. It provides significant support to the idea that the F_c chain was formed by several duplications of a much smaller polypeptide chain.

VI. Conclusion

The molecular basis of heredity rests on the sequence of four bases in DNA molecules. This provides a direct key to the molecular mechanism of evolution. Since all hereditary information is carried as linear sequences of four variables, evolution can take place only through changes in such sequences, consisting of repetition, shortening, and replacements in DNA. These alterations produce changes in the organism that are termed phenotypic changes. Most of these changes occur in proteins. The net result is change in the fitness of the organism for its environment. This leads to natural selection: the emergence of some species, and the extinction of others.

The new field of molecular evolution measures evolutionary changes in DNA molecules by comparing their base sequences. One method for comparing them is by the annealing procedure, in which the ability of single strands of DNA from two different species to form hybrids is quantitatively measured (Britten and Kohne, 1965–1966). Another is the indirect method of determining and comparing the amino acid sequences in proteins. In a very few cases, it has been possible to analyze and directly compare the base sequences in molecules of RNA.

It is fortunately possible, although the procedures are laborious, to determine the amino acid sequences of proteins. The genetic code enables the sequences to be translated back into the base sequences in DNA that code for the proteins. There are some limitations to this procedure because of certain ambiguities in the code. However, the method is sufficiently accurate to provide a lot of information on the evolution of proteins. Such information is deduced from comparisons of proteins with identical or analogous functions, and with similar but not identical sequences, obtained in many cases from different species of organisms.

The evolutionary changes in proteins have their counterparts in genetic changes described as point mutations, recombination, and duplication. The changes take place at the DNA level, and appear in proteins as a result of changes in a region of DNA that forms the structural gene, or cistron, for a polypeptide chain. A small proportion of such changes find their way into the genome of an entire species, and affect its evolution. In any one protein, amino acid replacements corresponding to point mutations appear to take place steadily during evolution, although proteins differ from each other with respect to the rates at which such changes appear. The accumulation of evolutionary changes produces differentiation in proteins often leading to modifications of their function,

or even to new functions, as in the case of lysozyme and lactose synthetase, which are two proteins with apparently a common molecular antecedent.

Lengthening and, more commonly, shortening of protein molecules take place during evolution as a result of recombinational events that occur in cistronic DNA. These events may be detected if they occur in only one member of a pair of homologous proteins. They may be partially or completely obscured by subsequent differentiation. As a consequence, their detection is often difficult. Computerized statistical methods may be helpful in deciding whether proposals for such events are valid. Examples of the use of such methods are given for several proteins, including the hemoglobins and immunoglobulins.

REFERENCES

Allan, N., Beale, D., Irvine, D., and Lehmann, H. (1965). *Nature* **208,** 658.

Allison, A. C. (1964). *Cold Spring Harbor Symp. Quant. Biol.* **29,** 137.

Anfinsen, C. B. (1959). "The Molecular Basis of Evolution." Wiley, New York.

Babin, D. R., Schroeder, W. A., Shelton, J. R., Shelton, J. B., and Robberson, B. (1966). *Biochemistry* **5,** 1297.

Baglioni, C. (1962). *J. Biol. Chem.* **237,** 69.

Baglioni, C. (1965). *Nature* **207,** 259.

Baglioni, C., and Ingram, V. M. (1961a). *Nature* **189,** 465.

Baglioni, C., and Ingram, V. M. (1961b). *Biochim. Biophys. Acta.* **48,** 253.

Baglioni, C., and Lehmann, H. (1962). *Nature* **196,** 229.

Bahl, O. P., and Smith, E. L. (1965). *J. Biol. Chem.* **240,** 3585.

Beale, D., and Lehmann, H. (1965). *Nature* **207,** 159.

Behrens, O. K., and Bromer, W. W. (1958). *Vitamins Hormones* **16,** 263.

Benson, A. M., Mower, H. F., and Yasunobu, K. T. (1966). *Proc. Natl. Acad. Sci. U.S.* **55,** 1532.

Black, J. A., Kauffman, D. L., and Dixon, G. H. (1967). Quoted by Dayhoff and Eck (1967–1968).

Blackwell, R. Q., and Liu, C. S. (1966). *Biochem. Biophys. Res. Commun.* **24,** 732.

Blombäck, B., and Doolittle, R. F. (1963). *Acta Chem. Scand.* **17,** 1819.

Blombäck, B., Blombäck, M., and Grondahl, N. J. (1965). *Acta Chem. Scand.* **19,** 1789.

Blombäck, B., Blombäck, M., Grondahl, N. J., and Holmberg, E. (1966). *Arkiv Kemi* **25,** 411.

Bonaventura, J., and Riggs, A. (1967). *Science* **158,** 800.

Bookchin, R. M., Nagel, R. L., Ranney, H. M., and Jacobs, A. S. (1966). *Biochem. Biophys. Res. Commun.* **23,** 122.

Botha, M., Beale, D., Alsaacs, W., and Lehmann, H. (1966). *Nature* **212,** 792.

Bowman, B. H., Oliver, C. P., Barnett, D. R., Cunningham, J. E., and Schneider, R. G. (1964). *Blood* **23,** 193.

Boyer, S. H., Hathaway, P., Pascasio, F., Ortin, C., Bordley, J., and Naughton, M. A. (1966). *Science* **153,** 1539.

Bradley, T. B., Wohl, R. C., and Reider, R. F. (1967). *Science* **157,** 1581.

Braunitzer, G. (1967). *Naturwissenschaften* **54,** 407.
Braunitzer, G., and Hilschmann, N. (1964). *Advan. Protein Chem.* **19,** 34.
Braunitzer, G., and Koehler, H. (1966). *Z. Physiol. Chem.* **343,** 290.
Braunitzer, G., and Matsuda, G. (1963). *J. Biochem. (Toyko)* **53,** 262.
Braunitzer, G., Gehring-Mueller, R., Hilschmann, N., Hilse, K., Hobom, G., Rudloff, V., and Wittmann-Liebold, B. (1961a). *Z. Physiol. Chem.* **325,** 283.
Braunitzer, G., Hilschmann, N., Rudloff, V., Hilse, K., Liebold, B., and Mueller, R. (1961b). *Nature* **190,** 480.
Braunitzer, G., Best, J. S., Flamm, U., and Schrank, B. (1966). *Z. Physiol. Chem.* **347,** 207.
Brew, K., and Campbell, P. N. (1967). *Biochem. J.* **102,** 258.
Brew, K., Vanaman, T. C., and Hill, R. L. (1967). *J. Biol. Chem.* **242,** 3747.
Briehl, R. H. (1963). *J. Biol. Chem.* **238,** 236.
Britten, R. J., and Kohne, D. E. (1965–1966). Annual Report of the Director, Department of Terrestrial Magnetism, Carnegie Institution, Washington, D.C., p. 78.
Brown, J. R., Kauffman, D. L., and Hartley, B. S. (1967). *Biochem. J.* **103,** 497.
Buettner-Janusch, J., and Hill, R. J. (1965). *Science* **147,** 866.
Cantor, C., and Jukes, T. H. (1966a). *Proc. Natl. Acad. Sci. U.S.* **56,** 177.
Cantor, C., and Jukes, T. H. (1966b). *Biochem. Biophys. Res. Commun.* **23,** 319.
Carrell, R. W., Lehmann, H., and Hutchinson, H. E. (1966). *Nature* **210,** 915.
Carrell, R. W., Lehmann, H., Larkin, P. A., Ralh, E., and Hunter, E. (1967). *Nature* **215,** 626.
Chan, S. K., and Margoliash, E. (1966a). *J. Biol. Chem.* **241,** 335.
Chan, S. K., and Margoliash, E. (1966b). *J. Biol. Chem.* **241,** 507.
Chan, S. K., Tulloss, I., and Margoliash, E. (1966). *Biochemistry* **5,** 2586.
Chance, R. E., and Ellis, R. M. (1968). *Federation Proc.* **27,** 392; *Science* **160,** (in press).
Chernoff, A. K., and Perrillie, P. E. (1964). *Biochem. Biophys. Res. Commun.* **16,** 368.
Clegg, J. B., Naughton, M. A., and Weatherall, D. J. (1966). *J. Mol. Biol.* **19,** 91.
Cohen, S., and Milstein, C. (1967). *Advan. Immunol.* **7,** 1.
Cox, E. C., and Yanofsky, C. (1967). *Proc. Natl. Acad. Sci. U.S.* **58,** 1895.
Crookston, J. H., Beale, D., Irvine, D., and Lehmann, H. (1965). *Nature* **208,** 1059.
Crick, F. H. C. (1966). *J. Mol. Biol.* **9,** 548.
Dacie, J. V., Shinton, N. K., Gaffniey, P. J., Carrell, R. W., and Lehmann, H. (1967). *Nature* **216,** 663.
Dayhoff, M. O., and Eck, R. V. (1967–1968). "Atlas of Protein Sequence and Structure."
Degens, E. T., Johanneson, B. W., and Meyer, R. W. (1967). *Naturwissenschaften* **54,** 638.
Dixon, G. H. (1966). *In* "Essays in Biochemistry" (P. N. Campbell and G. D. Greville, eds.), Vol. 2, p. 147. Academic Press, New York.
Doolittle, R. F., and Blombäck, B. (1964). *Nature* **202,** 147.
Doolittle, R. F., Schubert, D., and Schwartz, S. A. (1967). *Arch. Biochem. Biophys.* **118,** 456.
Dreyer, W. J., and Bennett, J. C. (1965). *Proc. Natl. Acad. Sci. U.S.* **54,** 864.
Dreyer, W. J., Gray, W. R., and Hood, L. (1967). *Cold Spring Harbor Symp. Quant. Biol.* **32,** 353.

Dus, K., and Sletten, K. (1968). *In* "Symposium on Cytochromes—Structural and Chemical Aspects of Cytochromes," (K. Okunuki, M. D. Kamen, and I. Sekuzu, eds.), p. 293. Univ. of Tokyo Press, Japan.
Eck, R. (1964). *Proc. Conf. Biol. Med.* **17,** 115.
Eck, R., and Dayhoff, M. (1966). "Atlas of Protein Structure." Natl. Biomed. Res. Federation, Silver Spring, Maryland.
Edmundson, A. B. (1965). *Nature* **205,** 883.
Epstein, C. J., and Motulsky, A. G. (1965). *Progr. Med. Genet.* **4,** 94.
Fitch, W. M. (1966a). *J. Mol. Biol.* **16,** 1.
Fitch, W. M. (1966b). *J. Mol. Biol.* **16,** 17.
Fitch, W. M. (1967). *J. Mol. Biol.* **26,** 499.
Fitch, W. M., and Margoliash, E. (1967a). *Science* **155,** 279.
Fitch, W. M., and Margoliash, E. (1967b). *Biochem. Genet.* **1,** 65.
Folk, J. E., Gladner, J. A., and Levin, Y. (1959). *J. Biol. Chem.* **234,** 2317.
Galizzi, A., and von Ehrenstein, G. (1967). Personal communication.
Gerald, P. A., and Efron, M. L. (1961). *Proc. Natl. Acad. Sci. U.S.* **47,** 1758.
Goldstein, J., Konigsberg, W., and Hill, R. J. (1963). *J. Biol. Chem.* **238,** 2016.
Gottlieb, A. J., Restrepe, A., and Itano, H. A. (1963). *Federation Proc.* **23,** 172.
Gray, W. R. (1966). *Proc. Roy. Soc.* **B166,** 146.
Hanadu, M., and Rucknagel, D. (1963). *Biochem. Biophys. Res. Commun.* **11,** 229.
Heller, J., and Smith, E. L. (1966). *J. Biol. Chem.* **241,** 3165.
Huehns, E. R., and Shooter, E. M. (1965). *J. Med. Genet.* **2,** 48.
Hill, R. L., and Konigsberg, W. (1962). *J. Biol. Chem.* **237,** 3151.
Hill, R. L., and Schwartz, H. C. (1959). *Nature* **184,** 641.
Hill, R. L., Buettner-Janusch, J., and Buettner-Janusch, V. (1963). *Proc. Natl. Acad. Sci. U.S.* **50,** 885.
Hill, R. L., Delaney, R., Fellows, R. E., Jr., and Lebovitz, H. E. (1966). *Proc. Natl. Acad. Sci. U.S.* **56,** 1762.
Hill, R. L., Harris, C. M., Naylor, J. F., and Sams, W. M. (1969). *J. Biol. Chem.* **244,** 2182.
Hood, L., Gray, W. R., Sanders, B. G., and Dreyer, W. J. (1967). *Cold Spring Harbor Symp. Quant. Biol.* **32,** 133.
Horowitz, N. H. (1945). *Proc. Natl. Acad. Sci. U.S.* **31,** 153.
Hunt, J. A., and Ingram, V. M. (1960). *Biochim. Biophys. Acta* **42,** 409.
Hunt, J. A., and Ingram, V. M. (1961). *Biochim. Biophys. Acta* **49,** 520.
Ingram, V. M. (1957). *Nature* **180,** 326.
Ingram, V. M. (1961). *Nature* **189,** 704.
Ingram, V. M. (1962). *In* "The Molecular Control of Cellular Activity" (J. M. Allen, ed.), p. 179. McGraw-Hill, New York.
Ingram, V. M. (1963). "The Hemoglobins in Genetics and Evolution." Columbia Univ. Press, New York.
Iwasaki, K., Sabol, S., Wahba, A. J., and Ochoa, S. (1968). *Arch. Biochem. Biophys.* **125,** 542.
Jones, R. T., Koler, R. D., and Lisker, R. G. (1963). *Clin. Res.* **11,** 105.
Jones, R. T., Coleman, R. B., and Heller, P. (1964). *Federation Proc.* **23,** 173.
Jones, R. T., Brimhall, B., Huehns, E. R., and Barnicot, N. A. (1966a). *Science* **151,** 1406.
Jones, R. T., Brimhall, B., Huisman, T. H. J., Kleihauer, E., and Betke, K. (1966b). *Science* **154,** 1024.
Jones, R. T., Brimhall, B., Huehns, E. R., and Motulsky, A. G. (1968). *Biochim.*

Biophys. Acta **154,** 278.
Jukes, T. H. (1966). "Molecules and Evolution." Columbia Univ. Press, New York.
Jukes, T. H. (1967). *Biochem. Biophys. Res. Commun.* **27,** 573.
Kafatos, F. C., Law, J. H., and Tartakoff, A. M. (1967). *J. Biol. Chem.* **242,** 1458.
Keresztes-Nagy, S., Perini, F., and Margoliash, E. (1968). Personal communication.
Kilmartin, J. V., and Clegg, J. B. (1967). *Nature* **213,** 269.
King, J. L., and Jukes, T. H. (1969). *Science* **163** (in press).
Kleihauer, E. F., Reynolds, C. A., Dozy, A. M., Wilson, J. B., Moores, R. R., Berenson, M. P., Wright, C., and Huisman, T. H. J. (1968). *Biochim. Biophys. Acta* **154,** 220.
Krause, L. M. (1965). *Nature* **207,** 159.
Krause, L. M., Miyaji, T., Iuchi, I., and Kraus, A. P. (1966). *Biochemistry* **5,** 3701.
Kreil, G. (1963). *Z. Physiol. Chem.* **334,** 154.
Kreil, G. (1965). *Z. Physiol. Chem.* **340,** 86.
Labie, D., Schroeder, W. A., and Huisman, T. H. J. (1966). *Biochim. Biophys. Acta* **127,** 428.
Lehmann, H., and Carrell, R. W. (1969). *Brit. Med. Bull.* **25,** 14.
Lehmann, H., Beale, D., and Bio-Daku, F. S. (1964). *Nature* **203,** 363.
Lehmann, H., Huntsman, R. G., and Ager, J. A. M. (1966). "The Metabolic Basis of Inherited Disease," 2nd ed., Chapter 49, p. 1100. McGraw-Hill, New York.
Lennox, E. S., and Cohn, M. (1967). *Ann. Rev. Biochem.* **36,** 365.
Leone, C. A., ed. (1964). "Taxonomic Biochemistry and Serology." Ronald, New York.
Lewis, E. B. (1951). *Cold Spring Harbor Symp. Quant. Biol.* **16,** 159.
Li, C. H., Barnafi, L., Chretien, M., and Chung, D. (1965). *Nature* **208,** 1093.
Liddell, J., Brown, D., Beale, D., Lehmann, H., and Huntsman, R. G. (1964). *Nature* **204,** 269.
Lisker, R. G., Ruiz-Reyes, J. V., and Loria, A. (1963). *Blood* **22,** 342.
McDowall, M. A., and Smith, E. L. (1965). *J. Biol. Chem.* **240,** 4635.
Manwell, C. (1967). *Comp. Biochem. Physiol.* **23,** 383.
Margoliash, E. (1963). *Proc. Natl. Acad. Sci. U.S.* **50,** 672.
Margoliash, E., and Schejter, A. (1966). *Advan. Protein Chem.* **21,** 113.
Margoliash, E., and Fitch, W. M. (1968). *Ann. N.Y. Acad. Sci.* **151,** 359.
Margoliash, E., and Smith, E. L. (1965). "Evolving Genes and Proteins" (V. Bryson and H. Vogel, eds.), p. 221. Academic Press, New York.
Margoliash, E., and Steiner, D. F. (1968). *Proteins Nucleic Acids Symp. Univ. Houston, Houston, Texas, April 9.*
Margoliash, E., Reichlin, M., and Nisonoff, A. (1968). *In* "Symposium on Cytochromes—Structural and Chemical Aspects of Cytochromes" (K. Okunuki, M. D. Kamen, and I. Sekuzu, eds.), p. 269. Univ. of Tokyo Press, Japan.
Matsubara, H., and Sasaki, R. (1968). *J. Biol. Chem.* **243,** 732.
Matsubara, H., and Smith, E. L. (1963). *J. Biol. Chem.* **238,** 2732.
Matsuda, G., Maita, T., Yamaguchi, M., Ota, H., Migita, M., and Miyauchi, T. (1968). *J. Biochem. (Tokyo)* **63,** 136.
Minnich, V., Hill, R. L., Khuri, P. D., and Anderson, M. E. (1965). *Blood* **25,** 830.
Miyaji, T., Iuchi, I., Shibata, S., Takeda, I., and Tamura, A. (1963). *Acta Haematol. Japon.* **26,** 538.
Miyaji, T., Suzuki, H., Ohba, Y., and Shibata, S. (1966). *Clin. Chim. Acta* **14,** 624.
Miyaji, T., Ohba, K., Yamamoto, K., Shibata, S., Iuchi, I., and Hamilton, H. B. (1968a). *Science* **159,** 204.

Miyaji, T., Ohba, K., Yamamoto, K., Shibata, S., Iuchi, I., and Takenaka, M. (1968b). *Nature* **217,** 89.
Mross, G. A., and Doolittle, R. F. (1967). *Arch. Biochem. Biophys.* **122,** 674.
Muller, C. J., and Kingma, S. (1961). *Biochim. Biophys. Acta* **50,** 595.
Munkres, K. D., and Richards, F. M. (1965). *Arch Biochem. Biophys.* **109,** 457.
Mutt, V., and Jorpes, J. E. (1966). *4th Intern. Symp. Chem. Nat. Products, Stockholm.*
Nakajima, H., Takemura, T., Nakajima, O., and Yamaoka, K. (1963). *J. Biol. Chem.* **238,** 3784.
Narita, K., and Titani, K. (1965). *Proc. Japan Acad.* **41,** 831.
Narita, K., and Sugeno, K. (1968). In preparation.
Needleman, S. B., and Margoliash, E. (1966). *J. Biol. Chem.* **241,** 853.
Needleman, S. B., and Wunsch, C. D. (1968). *J. Mol. Biol.* (in press).
Neurath, H., Walsh, K. A., and Winter, W. P. (1968). *Science* **158,** 1638.
Nolan, C., and Margoliash, E. (1966). *J. Biol. Chem.* **241,** 1049.
Nolan, C., and Margoliash, E. (1968). *Ann. Rev. Biochem.* **37,** 727.
Nuttall, G. H. F. (1904). "Blood Immunity and Blood Relationship." Macmillan, New York.
Pauling, L., and Corey, R. B. (1951). *Proc. Natl. Acad. Sci. U. S.* **37,** 235.
Pauling, L., and Zuckerkandl, E. (1963). *Acta Chem. Scand.* **17,** 9.
Perutz, M. F. (1965). *J. Mol. Biol.* **13,** 646.
Perutz, M. F., Kendrew, J. C., and Watson, H. C. (1965). *J. Mol. Biol.* **13,** 669.
Pierre, L. E., Rath, C. E., and McCoy, K. (1963). *New Engl. J. Med.* **268,** 862.
Popp, R. A. (1965). *Federation Proc.* **24,** 1252.
Porter, R. R. (1967). *Biochem. J.* **105,** 417.
Press, C. M. (1967). *Biochem. J.* **104,** 30C.
Putnam, F. W., and Easiley, C. W. (1965). *J. Biol. Chem.* **240,** 1626.
Reichlin, M., Bucci, E., Fronticelli, C., Wyman, J., Antonini E., Ioppolo, C., and Rossi-Fanelli, A. (1966). *J. Mol. Biol.* **17,** 18.
Reynolds, C. A., and Huisman, T. H. J. (1966). *Biochim. Biophys. Acta* **130,** 541.
Rifkin, D. B., Rifkin, M. R., and Konigsberg, W. (1966). *Proc. Natl. Acad. Sci. U.S.* **55,** 586.
Rothfus, J. A., and Smith, E. L. (1965). *J. Biol. Chem.* **240,** 4277.
Rudloff, V., Zelenik, M., and Braunitzer, G. (1966). *Z. Physiol. Chem.* **344,** 284.
Salomon, H., *et al.* Cited by Schroeder and Jones (1965).
Sanger, F. (1952). *Advan. Protein Chem.* **7,** 1.
Sansome, G., Carrell, R. W., and Lehmann, H. (1967). *Nature* **214,** 877.
Sarich, V. M., and Wilson, A. C. (1967). *Science* **158,** 1200.
Schneider, R. G., and Jones, R. T. (1965). *Science* **148,** 240.
Schneider, R. G., Haggard, M. E., McNutt, C. W., Johnson, J. E., Bowman, B. H., and Barnett, D. R. (1964). *Science* **143,** 697.
Schroeder, W. A., and Jones, R. T. (1965). *Progr. Chem. Org. Nat. Products* **23,** 113.
Schroeder, W. A., Shelton. J. R., Shelton, J. B., Cormick, J., and Jones, R. T. (1963). *Biochemistry* **2,** 992.
Sherman, F., Stewart, J. W., Parker, J., Futterman, G. J., and Margoliash, E. (1968). *In* "Symposium on Cytochromes" (K. Okunuki, M. D. Kamen, and I. Sekuzu, eds.), 257. Univ. of Tokyo Press, Japan.
Shibata, S., Iuchi, I., Miyaji, T., and Takeda, I. (1963). *Bull. Yamaguchi Med. School* **10,** 1.

Shibata, S., Miyaji, T., Iuchi, I., Ueda, S., and Takeda, I. (1964). *Clin. Chim. Acta* **10,** 101.
Shim, B., and Bearn, A. G. (1964). *Am. J. Hum. Genet.* **16,** 477.
Simpson, G. G. (1964). *Science* **146,** 1535.
Smillie, L. B., Furka, A., Nagabhusha, N., Stevenson, K. J., and Parkes, C. O. (1968). *Nature* **218,** 343.
Smith, E. L., and Margoliash, E. (1964). *Federation Proc.* **23,** 1243.
Smith, L. F. (1966). *Am. J. Med.* **40,** 662.
Smithies, O. (1965). *Proc. Brook Lodge Conf. Proteins Polypeptides.*
Speyer, J. F. (1965). *Biochem. Biophys. Res. Commun.* **21,** 6.
Stamatoyannopoulos, G., Yoshida, A., Adamson, J., and Heinenburg, S. (1968). *Science* **159,** 741.
Stebbins, G. L. (1966). "Processes of Organic Evolution," p. 29. Prentice-Hall, Englewood Cliffs, New Jersey.
Steiner, D., and Oyer, P. E. (1967). *Proc. Natl. Acad. Sci. U.S.* **57,** 473.
Steiner, D. F., Cunningham, D., Spigelman, L., and Aten, B. (1967). *Science* **157,** 697.
Stevens, F. C., Glazer, A. N., and Smith, E. L. (1967). *J. Biol. Chem.* **242,** 2764.
Stewart, J. W., and Margoliash, E. (1965). *Can. J. Biochem.* **43,** 1187.
Sueoka, N. (1961). *Proc. Natl. Acad. Sci. U.S.* **47,** 1141.
Swenson, R. T., Hill, R. L., Lehmann, H., and Jim, R. T. S. (1962). *J. Biol. Chem.* **237,** 1517.
Szybalski, W., Kubinski, H., and Sheldrick, P. (1966). *Cold Spring Harbor Symp. Quant. Biol.* **31,** 123.
Takanami, M. (1967). *J. Mol. Biol.* **23,** 135.
Tanaka, M., Nakashima, T., Benson, A. M., Mower, H. F., and Yasunobu, K. T., (1964). *Biochem. Biophys. Res. Commun.* **16,** 422.
Titani, K., Wikler, M., and Putnam, F. W. (1967). *Science* **155,** 828.
Treffers, H. P., Spinell, V., and Belser, N. O. (1954). *Proc. Natl. Acad. Sci. U.S.* **55,** 274.
Tuchinda, S., Beale, D., and Lehmann, H. (1965). *Brit. Med. J.* **I,** 1583.
Tuppy, H. (1958). *In* "Symposium on Protein Structure" (A. Neuberger, ed.), Wiley, New York.
von Ehrenstein, G. (1966). *Cold Spring Harbor Symp. Quant. Biol.* **31,** 705.
Walsh, K. A., and Neurath, H. (1964). *Proc. Natl. Acad. Sci. U.S.* **52,** 595.
Wasserman, E., and Levine, L. (1961). *J. Immunol.* **87,** 290.
Watson, J. D. (1965). "Molecular Biology of the Gene," p. 268. Benjamin, New York.
Watson, J. D., and Crick, F. H. C. (1953a). *Nature* **171,** 737.
Watson, J. D., and Crick, F. H. C. (1953b). *Nature* **171,** 964.
Watson-Williams, E. S., Beale, D., Irvine, D., and Lehmann, H. (1965). *Nature* **205,** 1273.
Weinstein, B. (1968). *Abst. Am. Chem. Soc. Meeting, San Francisco.*
White, A., Handler, P., and Smith, E. L. (1964). "Principles of Biochemistry," p. 145. McGraw-Hill, New York.
Wikler, M., Titani, K., Shinoda, T., and Putnam, F. W. (1967). *J. Biol. Chem.* **242,** 1668.
Wilson, A. C., Kaplan, N. O., Levine, L., Pesce, A., Reichlin, M., and Allison, W. S. (1964). *Federation Proc.* **23,** 1257.

Yamaguchi, Y., Horie, H., Matsuo, A., Sasakawa, S., and Satake, K. (1965). *J. Biochem.* (*Tokyo*) **58,** 186.

Yanofsky, C. (1965). *Biochem. Biophys. Res. Commun.* **18,** 898.

Yanofsky, C., and Helinski, D. R. (1962). *Proc. Natl. Acad. Sci. U.S.* **48,** 173.

Yanofsky, C., Carlton, B. C., Guest, J. R. Helinski, D. R., and Henning, U. (1964). *Proc. Natl. Acad. Sci. U.S.* **51,** 266.

Yanofsky, C., Cox, E. C., and Horn, V. (1966a). *Proc. Natl. Acad. Sci. U.S.* **55,** 274.

Yanofsky, C., Ito, J., and Horn, V. (1966b). *Cold Spring Harbor Symp. Quant. Biol.* **31,** 151.

Yaoi, Y., Titani, K., and Narita, K. (1966). *J. Biochem.* (*Tokyo*) **59,** 247.

Yoshida, A. (1967). *Proc. Natl. Acad. Sci. U.S.* **57,** 835.

Zuckerkandl, E., and Pauling, L. (1962). *In* "Horizons in Biochemistry" (M. Kasha and B. Pullman, eds.), p. 189. Academic Press, New York.

Zuckerkandl, E., and Pauling, L. (1965). *In* "Evolving Genes and Proteins" (V. Bryson and H. J. Vogel, eds.), pp. 97–166. Academic Press, New York.

Zuckerkandl, E., and Schroeder, W. A. (1961). *Nature* **192,** 984.