CABIOS

Vol.8. no.3. 1992
Pages 275 – 282

The rapid generation of mutation data matrices from protein sequences

David T.Jones[1,2], William R.Taylor[2] and Janet M.Thornton[1]

Abstract

An efficient means for generating mutation data matrices from large numbers of protein sequences is presented here. By means of an approximate peptide-based sequence comparison algorithm, the set sequences are clustered at the 85% identity level. The closest relating pairs of sequences are aligned, and observed amino acid exchanges tallied in a matrix. The raw mutation frequency matrix is processed in a similar way to that described by Dayhoff et al. (1978), and so the resulting matrices may be easily used in current sequence analysis applications, in place of the standard mutation data matrices, which have not been updated for 13 years. The method is fast enough to process the entire SWISS-PROT databank in 20 h on a Sun SPARCstation 1, and is fast enough to generate a matrix from a specific family or class of proteins in minutes. Differences observed between our 250 PAM mutation data matrix and the matrix calculated by Dayhoff et al. are briefly discussed.

Introduction

Despite the great diversiy of methods devised for the alignment and comparison of protein sequences, all of these depend at some point on the simple comparison of two amino acid residues. The most popular method for measuring the similarity between amino acids is to use a scoring matrix of some form. At its simplest, a typical scoring matrix comprises 20 × 20 elements, each element representing some metric that relates two residues.

The least sophisticated matrix is the 'Unitary Protein Matrix' (UPM), also known as the 'identity matrix'. The UPM scores a 1 for exactly matching residues and a 0 for every other combination. Obviously this matrix lacks sensitivity, as it is unable to detect the possibility of phenotypically silent mutational events between two sequences. One advantage of the UPM is that it is wholly unbiased, providing a very easily understood alignment metric. The 'percentage identity' between two sequences is often offered as a universal means of describing the mutual degree of 'homology' between them. Although a low identity score can in no way prove or disprove the existence of homology, it has proved easier to provide rules of thumb for identity scoring than for any other scheme. In general, for two sequences of reasonable length (say 50 residues or more), a percentage identity of >25% points to a significant structural homology between them. Feng and Doolittle have described a fuzzy region around 20% identity which they call the 'Twilight Zone'. Within this zone and below, it is not possible to tell the difference between real sequence similarity implying a common structural framework, and accidental similarity providing no useful structural information.

Probably the next simplest amino acid scoring matrix is the 'Genetic Code Matrix' (GCM). This matrix scores amino acid similarity by the maximum number of common nucleotide bases between their closest matching representative codons. Identical residues of course share a maximum of 3 bases, whereas non-identical residues may have only 0, 1 or 2 bases in common. This matrix has a pleasantly 'genetic flavour' to it, but it must be realized that the bulk of the selection pressure is on the protein sequence and not on the underlying DNA sequence. Although there does seem to be a reasonable correlation between the nucleotide codons associated with amino acids and the degree of chemical similarity between them (Woese, 1969, for example), the rather limited range of match-scores puts the GCM somewhat in the shade. To detect weak homologies between sequences a more accurate amino acid comparison table is required.

McLachlan (1972) published a scoring matrix that attempted to quantify explicitly the degree of chemical similarity between amino acids. This matrix, known as the 'Structure-Genetic Matrix' (SGM), incorporated two sources of information in evaluating the similarities of amino acids. The first source was a statistical analysis of observed amino acid exchanges in available families of proteins, the second was from the assignment of transition values for each pair of amino acids depending on the number of overlapping physico-chemical properties between them. These data were used to 'bias' the UPM in such a way that only 20 of the 190 possible interchanges were significantly preferred (Feng *et al.*, 1985). The problem with the SGM and other matrices that attempt to incorporate 'real' amino acid similarities is that the groupings used are artificial, there is no guarantee that an arbitrary common amino acid property is at all important for strucural and functional conservation between proteins. A better approach is to concentrate on the observed exchanges between amino acids

[1]*Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT and* [2]*Laboratory of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London, NW7 1AA, UK*

in very similar aligned sequences. Evidently amino acids that share the appropriate properties will exchange more frequently than ones that do not. McLachlan's earlier attempt to compare amino acids (McLachlan, 1971) was based entirely on such a statistical approach.

Recently, matrices based on the principles of structural comparison have been described (Risler *et al.*, 1988; Overington *et al.*, 1990). These matrices essentially contain statistics on the pairwise substitutions observed at structurally equivalent positions in aligned families of protein structures. In the case of Overington *et al.*, a range of matrices are calculated, one from each class of structural environment, an example of one such class being 'buried coil' for example. These matrices show great promise in increasing the accuracy of sequence-to-sequence, and sequence-to-structure alignments, though the sparsity of structural data presently available is a significant disadvantage of this approach.

The most widely used comparison matrix today is the 'Log-Odds Matrix' and the very closely related 'Mutation Data Matrix' (MDM) published by Dayhoff *et al.* (1978). The MDM was calculated from a study of the exchange probabilities (or odds) derived from an analysis of the evolutionary changes seen in groups of very similar proteins. A strictly Markovian model (i.e. the current probabilities are independent of previous events) of amino acid exchange is assumed in the Dayhoff model. This model has been criticized (see George *et al.*, 1990, for a review), but comparisons of different scoring schemes have tended hesitantly to recommend the MDM over other matrices (Feng *et al.*, 1985).

In this paper we show a straightforward and automatic procedure for generating mutation data matrices, in order that very large sets of sequences can be processed without using inordinate amounts of computing resources. In particular we are able to improve the generality of the MDM, in that we now have access to a much greater variety of protein sequences than were available to Dayhoff and her workers in 1978, and it is our hope that the matrices presented here will more clearly express the general nature of the underlying amino acid similarities.

The original mutation data matrix (MDM68) was presented in the original *Atlas of Protein Sequence and Structure* (1968), and the method (outlined below) remained virtually unaltered through each of the subsequent updates. There are five main steps required for the creation of a mutation data matrix:

1. Construction of the raw PAM matrix

The basic unit of molecular evolution expressed in a MDM is the 'accepted point mutation', or with a little license to ease pronunciation: PAM. One PAM is simply the mutation of a single amino acid in a sequence such that the new amino acid may be accommodated in the structure and function of the protein. In general, therefore, amino acid residues that are frequently seen to exchange in a PAM matrix typically have similar physico-chemical properties.

The raw PAM substitution matrix is created by considering the possible mutational events that could have occurred between two closely related sequences. Ideally we would like to compare every present-day sequence with its own immediate predecessor and thus accurately map the evolutionary history of each sequence position. Of course this is impossible, and so two main courses of action may be taken to approximate this information. The method used by Dayhoff was the 'common ancestor' method. Here closely homologous pairs of present-day sequences are taken and a common ancestral sequence inferred. Given only a pair of present-day sequences, an unambiguous inferred common ancestor cannot be generated. A complete phylogenetic tree is required in this case to allow the most probable common ancestors to be inferred for each tree node. The important thing to realize is that the inference of common ancestors must consider the overall topology of the tree. Every suggested common ancestor must be traced back to higher level nodes and evaluated in order to determine whether or not that ancestral sequence is the most probable for the tree as a whole.

An alternative to the common ancestor method is to relate present-day sequences by their pairwise alignment distances, estimating a possible phylogenetic tree from this distance matrix. This method was first described by Fitch and Margoliash (1967). Although construction of the distance matrix is a trivial exercise, the generation of an optimal phylogenetic tree from this data again requires an exhaustive iterative analysis such that the total number of mutations required to produce the present day set of sequences is minimized. Although both of the above methods have advantages and disadvantages, matrix methods are now most widely used.

No matter which method is finally used to infer the phylogenetic tree, construction of the PAM matrix is the same. The raw matrix is generated by taking pairs of sequences, either a present-day sequence and its inferred ancestor, or two present-day sequences, and tallying the amino acid exchanges that have apparently occurred. Given the following alignment:

```
ACDEFL
AGDEAL
```

we count four PAMs (C → G, G → C, F → A and A → F). The raw PAM matrix is obviously symmetric given the fact that we cannot know whether for example C mutated to G or G mutated to C; there is no harm in this as we are interested in discerning the extent of similarity between amino acids here, and 'similarity' is generally thought of as being symmetric. Treatment of gaps/insertions in an alignment is arbitrary: one possibility is to count gap characters as another type of amino acid; another possibility that is probably the safer of the two is simply to ignore gaps. We are after all only interested in

the exchange of amino acids, the deletion of a particular amino acid tells us nothing of its relative similarity to other amino acids, though it does provide information as to the amino acid's characteristic 'mutability'.

2. Calculation of relative mutabilities

Evidently if we are to estimate the probability of a given mutation event, we must consider two pieces of information. Firstly how likely is it that a given amino acid A changes at all, secondly how likely is it that the given amino acid changes to amino acid B given that A does change? We are therefore interested in the conditional probability that amino acid A changes to amino acid B given that A is seen to change. The probability of amino acid A changing at all in a given unit of time is usually expressed as the 'relative mutability' of A. Relative mutability is simply calculated as the number of observed changes of an amino acid divided by its frequency of occurrence in the aligned sequences. From the alignment shown earlier, A is seen to change once, but occurs three times in the alignment. The relative mutability of A from this alignment alone is therefore calculated as ⅓. An overall measure of relative mutability must allow for the different evolutionary distances and different sequence lengths found in a non-specific collection of sequences. Mutability is normalized by defining the basic unit of evolutionary distance as being a single accepted point mutation in a sequence of length 100. The average relative mutability of an amino acid given this definition is therefore the total number of changes observed for this amino acid in all the families of proteins considered, divided by the total sum of all local frequencies of occurrence of the amino acid multiplied by the numbers of mutations per 100 residues in each of the branches of all the family trees.

3. Calculation of the mutation probability matrix

The basic matrix in the generation of MDM type matrices is the 'mutation probability matrix'. Elements of this matrix give the probability that a residue in column j will mutate to the residue in row i in a specified unit of evolutionary time. Evidently a diagonal element of this matrix represents the probability of residue $i = j$ remaining unchanged, and hence being easily calculated according to the following formula:

$$M_{ji} = 1 - \lambda m_j \quad (1)$$

where m_j is the average relative mutability of residue j, and λ is a proportionality constant.

Non-diagonal elements are given by:

$$M_{ij} = \frac{\lambda m_j A_{ij}}{\sum_i A_{ij}} \quad (2)$$

where A_{ij} is a (non-diagonal) element of the raw PAM matrix.

The value of λ relates to the evolutionary distance represented by the probability matrix, accordingly:

$$\sum_i f_i M_{ii} = 1 - \frac{P}{100} \quad (3)$$

where f_i is the normalized frequency of occurrence of residue i, and P approximates the evolutionary distance (in PAMs) represented by the matrix. This relationship breaks down for $P >> 5$.

P is usually given the value 1 so that the basic mutation probability matrix represents a distance of 1 PAM. Matrices representing larger evolutionary distances may be derived from the 1 PAM matrix by matrix multiplication. Squaring the 1 PAM matrix gives a 2 PAM matrix, cubing it a 3 PAM matrix and so forth.

4. Calculating the log-odds matrix

Of more use than the mutation probability matrix in the alignment of protein sequences is the 'relatedness odds matrix'. This symmetric matrix represents the probability of residue j being replaced by residue i per occurrence of i, and is derived from the mutation probability matrix simply by dividing each element M_{ij} by the normalized frequency of occurrence of i, f_i. For the purposes of sequence comparison the relatedness odds for each alignment position should be multiplied together in order to arrive at a total 'alignment odds' value. To avoid slow floating-point multiplications, the relatedness odds matrix is usually converted to the log odds-matrix (also known as the mutation data matrix) thus:

$$MDM_{ij} = 10 \log_{10} R_{ij} \quad (4)$$

where R_{ij} are elements of the relatedness odds matrix (MDM_{ij} values are rounded to the nearest integer).

Automating the procedure

Although computational tools were used in constructing the original MDMs, in particular for the inference of common ancestral sequences and the generation of phylogenetic trees, the whole process was only partially automated. This was hardly of consequence considering the small number of available sequences in the 1970s, but as at the time of writing some 23 000 protein sequences are available for analysis, it is evident that a more streamlined approach is now required.

Our method for generating MDMs is in fact very similar in essence to that described by Dayhoff *et al.* (1978). The method involves three steps: (i) clustering the sequences into homologous families, (ii) tallying the observed mutations between highly similar sequences and (iii) relating the observed mutation frequencies to those expected by pure chance. The main difference here is in our use of an approximate method (a

pairwise present-day ancestor scheme) for inferring the phylogenetic relationships among the sequences in the data set. A program was written to compute all the relevant data automatically from a file of protein sequences.

In view of the relative inefficiency of standard methods for inferring maximum parsimony phylogenetic trees it was found to be necessary to implement an approximate method to find the reasonable family trees by means of cluster analysis of the sequence data. Although the limitations of using such simple means alone for the inference of phylogenetic trees are well known (Czelusniak *et al.*, 1990), and the large-scale structure of such crude phylogenetic trees tends to be somewhat incorrect, the relationships between closely related sequences are inferred correctly. To verify our methodology, we attempted to re-create the set of sequences used to construct MDM78. Using these sequences we found our mutation data closely approximated those in the original work with 164 of the 400 mutation frequencies (number of mutations occurring per 10000 observations) being identical, and 350 differing by five or less. It should be pointed out that though our results very closely match those of Dayhoff *et al.*, our matrices are not derived from the same explicit evolutionary model outlined in the original work. The practical significance of this fact depends on the intended application of the matrices. In terms of sequence analysis applications, a derivation independent of the choice of evolutionary model might well be preferred due to the reduced possibility of bias (in particular, maximum parsimony nucleotide substitution methods will tend to produce results biased towards the exchanges expected from the genetic code rather than generally observed amino acid similarities). A further justification for determining relationships via a pairwise scheme is that of the 2621 families of proteins in the current release of SWISS-PROT, 79% contain fewer than five sequences. With such small families the results of simple clustering and those of rigorous maximum parsimony analysis are indistinguishable with respect to the present application.

In generating the initial distance matrix, we do not assume that the input sequences are in any way pre-clustered into family groups, and are therefore forced to calculate the entire distance matrix to sort the sequences into families, and thereafter produce trees for each family. Evidently the vast majority of pairwise comparisons are unnecessary, so some simple (and quick) means is needed to filter out sequence pairs that have no chance of producing alignment identity scores >85%. We propose here a simple approximate algorithm for 'estimating' the percentage identity between two protein sequences without prior alignment. Our algorithm considers the distribution of residue triplets (or 3-tuples) between the two sequences. If there are sufficient identical triplets between both sequences we assume that the sequences show a potential homology. The longest sequence is taken and a hash table constructed containing the frequencies of occurrence of the constituent triplets. The triplet frequencies

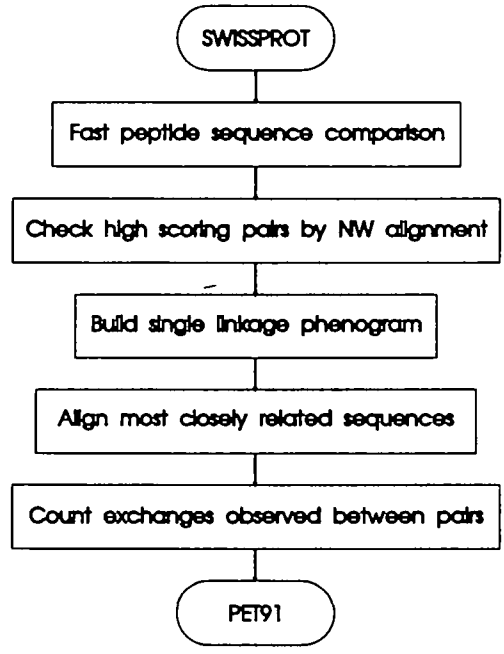

Fig. 1. An outline of the described method for generating mutation data matrices.

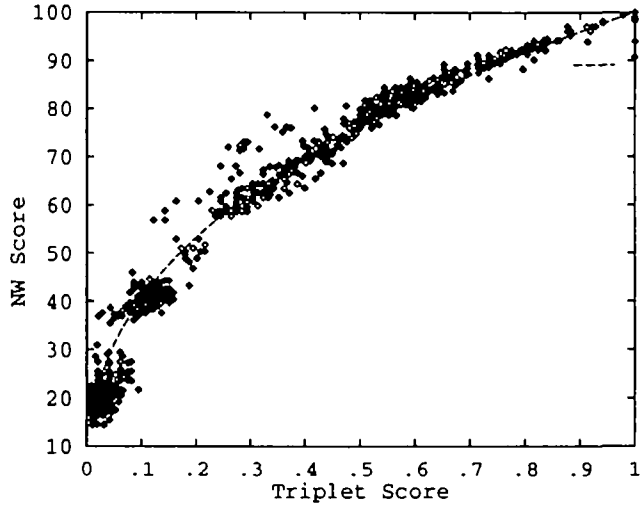

Fig. 2. Relationship between triplet scores and per cent identity after Needleman–Wunsch alignment with constant gap penalty.

of the shorter sequence are then compared with those of the longer. A comparison score is calculated thus:

$$S = \frac{\sum_{pqr=AAA}^{VVV} \min(f_a^{pqr}, f_b^{pqr})}{\min(n_a, n_b) - 2} \quad (5)$$

where f_a^{pqr} and f_b^{pqr} are the frequencies of occurrence of triplet pqr in sequences a and b, and n_a and n_b are the respective sequence lengths.

This normalized score (S) is effectively the fractional area

Table I. The 250 PAM PET91 matrix ($\log_{10}$ relatedness odds), based on 59 190 accepted point mutations found in 16 130 protein sequences

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2	247	216	386	106	208	600	1183	46	173	257	200	100	51	901	2413	2440	11	41	1766
R	-1	5	116	48	125	750	119	614	446	76	205	2348	61	16	217	413	230	109	46	69
N	0	0	3	1433	32	159	180	291	466	130	63	758	39	15	31	1738	693	2	114	55
D	0	-1	2	5	13	130	2914	577	144	37	34	102	27	8	39	244	151	5	89	127
C	-1	-1	-1	-3	11	9	8	98	40	19	36	7	23	66	15	353	66	38	164	99
Q	-1	2	0	1	-3	5	1027	84	635	20	314	858	52	9	395	182	149	12	40	58
E	-1	0	1	4	-4	2	5	610	41	43	65	754	30	13	71	156	142	12	15	226
G	1	0	0	1	-1	-1	0	5	41	25	56	142	27	18	93	1131	164	69	15	276
H	-2	2	1	0	0	2	0	-2	6	26	134	85	21	50	157	138	76	5	514	22
I	0	-3	-2	-3	-2	-3	-3	-3	-3	4	1324	75	704	196	31	172	930	12	61	3938
L	-1	-3	-3	-4	-3	-2	-4	-4	-2	2	5	94	974	1093	578	436	172	82	84	1261
K	-1	4	1	0	-3	2	1	-1	1	-3	-3	5	103	7	77	228	398	9	20	58
M	-1	-2	-2	-3	-2	-2	-3	-3	-2	3	3	-2	6	49	23	54	343	8	17	559
F	-3	-4	-3	-5	0	-4	-5	-5	0	0	2	-5	0	8	36	309	39	37	850	189
P	1	-1	-1	-2	-2	0	-2	-1	0	-2	0	-2	-2	-3	6	1138	412	6	22	84
S	1	-1	1	0	1	-1	-1	1	-1	-1	-2	-1	-1	-2	1	2	2258	36	164	219
T	2	-1	1	-1	-1	-1	-1	-1	-1	1	-1	-1	0	-2	1	1	2	8	45	526
W	-4	0	-5	-5	1	-3	-5	-2	-3	-4	-2	-3	-3	-1	-4	-3	-4	15	41	27
Y	-3	-2	-1	-2	2	-2	-4	-4	4	-2	-1	-3	-2	5	-3	-1	-3	0	9	42
V	1	-3	-2	-2	-2	-3	-2	-2	-3	4	2	-3	2	0	-1	-1	0	-3	-3	4

Values have been multiplied by 10 and rounded to the nearest integer. The upper half of the matrix shows the actual numbers of exchanges observed.

Table II. Mutation probability matrix for an evolutionary distance of 1 PAM. Values are scaled by a factor of 10^5

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	98759	27	24	42	12	23	66	129	5	19	28	22	11	6	99	264	267	1	4	193
R	41	98962	19	8	21	125	20	102	74	13	34	390	10	3	36	69	38	18	8	11
N	43	23	98707	284	6	31	36	58	92	26	12	150	8	3	6	344	137	0	23	11
D	63	8	235	98932	2	21	478	95	24	6	6	17	4	1	6	40	25	1	15	21
C	44	52	13	5	99450	4	3	41	17	8	15	3	10	28	6	147	28	16	68	41
Q	43	154	33	27	2	98955	211	17	130	4	64	176	11	2	81	37	31	2	8	12
E	82	16	25	398	1	140	99042	83	6	6	9	103	4	2	10	21	19	2	2	31
G	135	70	33	66	11	10	70	99369	5	3	6	16	3	2	11	129	19	8	2	32
H	17	164	171	53	15	233	15	15	98867	10	49	31	8	18	58	51	28	2	189	8
I	28	12	21	6	3	3	7	4	4	98722	212	12	113	31	5	28	149	2	10	630
L	24	19	6	3	3	29	6	5	12	122	99328	9	90	101	53	40	16	8	8	117
K	28	334	108	14	1	122	107	20	12	11	13	99101	15	1	11	32	57	1	3	8
M	36	22	14	10	8	19	11	10	8	253	350	37	98845	18	8	19	123	3	6	201
F	11	3	3	2	14	2	3	4	11	41	230	1	10	99357	8	65	8	8	179	40
P	150	36	5	7	3	66	12	16	26	5	97	13	4	6	99278	190	69	1	4	14
S	297	51	214	30	44	22	19	139	17	21	54	28	7	38	140	98548	278	4	20	27
T	351	33	100	22	9	21	20	24	11	134	25	57	49	6	59	325	98670	1	6	76
W	7	65	1	3	23	7	7	41	3	7	49	5	5	22	4	21	5	99684	24	16
Y	11	12	30	23	43	10	4	4	134	16	22	5	4	222	6	43	12	11	99377	11
V	226	9	7	16	13	7	29	35	3	504	161	7	71	24	11	28	67	3	5	98772

of overlap between the two triplet histograms. Scatter plots based on all possible pairwise alignment scores in a set of 200 protein sequences (containing a mixture of related and unrelated sequences) plotted against our scoring metric were produced (a subset of this data is shown in Figure 2). The raw triplet scores were thus compared with Needleman–Wunsch scores

(>40% ID), and the following relationship (correlation coefficient 0.986) was observed:

$$I \approx 100S^{0.3912}$$

where S is the normalized triplet frequency score, and the result I is in units of percentage identity.

By aligning only those sequence pairs with corrected triplet scores indicating sequence identity ≥45% and subsequently excluding sequence pairs with alignment scores of ≤85% identity we were able rapidly to generate a sparse distance matrix complete enough for our purposes. By combining this very rapid heuristic measure of identity with an efficiently coded dynamic programming algorithm as a 'second level filter' we were able to construct the distance matrix at an average rate of over 1000 similarity score calculations per second on a Sun SPARCstation 1 (standard Sun C compiler). Out of the 130 million pairwise alignments that would normally be required, only 559 692 passed the initial similarity filter, speeding up the process nearly 200-fold.

Using this matrix of identity scores, the sequences were subjected to an efficient single-linkage clustering algorithm, with mutation statistics being generated for each sequence by aligning it with the sequence that offers the highest pairwise alignment score. For each sequence pair, amino acid substitutions are tallied with alignment positions containing at least one non-standard residue code (B, Z, X or 'Gap') being ignored.

Implementation

The matrix generation program MAKEPET is coded in standard Sun C, and should be portable to most platforms supporting a C compiler. The required matrix PAM distance and other control parameters are specified as command line arguments. MAKEPET takes as input a single file of sequences in 'compact PIR' format, where each sequence is preceded by two description lines and terminated by a '*' character. A simple keyword searching program SEQGREP allows specific sets of sequences to be compiled from the complete sequence databank, permitting the easy generation of matrices biased towards particular structural or functional classes (membrane-bound proteins for example).

Results

The upper half of Table I shows how many of each of the possible 190 exchanges were observed, with the lower half of Table I showing our equivalent of the widely used MDM78 matrix ($\log_{10}$ relatedness-odds matrix for 250 PAMs), which we call PET91 (Pairwise Exchange Table 1991). The 1 PAM mutation probability matrix required to generate mutation data matrices for evolutionary distances other than 250 PAMs is shown in Table II. PET91 was generated from Release 15.0 of the SWISS-PROT protein sequence database (Bairoch, 1990),

Table III. Relative mutabilities and normalized frequencies of occurrence for the 20 amino acid residues, calculated from the PET91 data set, compared with the values from Dayhoff *et al.* (1978)

	Relative Mutability* (1991)	Relative Mutability* (1978)	Relative Frequency of Occurrence (1991)	Relative Frequency of Occurrence (1978)
Ala (A)	100	100	0.077	0.087
Arg (R)	83	65	0.051	0.041
Asn (N)	104	134	0.043	0.040
Asp (D)	86	106	0.052	0.047
Cys (C)	44	20	0.020	0.033
Gln (Q)	84	93	0.041	0.038
Glu (E)	77	102	0.062	0.050
Gly (G)	50	49	0.074	0.089
His (H)	91	66	0.023	0.034
Ile (I)	103	96	0.053	0.037
Leu (L)	54	40	0.091	0.085
Lys (K)	72	56	0.059	0.081
Met (M)	93	94	0.024	0.015
Phe (F)	51	41	0.040	0 040
Pro (P)	58	56	0.051	0.051
Ser (S)	117	120	0.069	0.070
Thr (T)	107	97	0.059	0.058
Trp (W)	25	18	0.014	0.010
Tyr (Y)	50	41	0.032	0.030
Val (V)	98	74	0.066	0.065

* Relative to Ala which is arbitrarily assigned a mutability of 100.

containing 16 941 sequences, though sequences <20 residues were excluded to avoid insignificant alignments. It should be noted that the 250 PAM matrix is shown here for reasons of comparison with the most common variant of the original matrix, and that matrices calculated for evolutionary distances other than 250 PAMs are often found to perform better for some sequence comparisons. The recently described sequence databank search program, BLAST (Altschul *et al.*, 1990), for example, uses a 120 PAM Dayhoff matrix by default.

Of particular interest here are the differences between these results and those of the original work, a rough impression of which may be gained from a comparison of the relative mutabilities shown in Table III with those observed by Dayhoff (1978). A value of 0.76 is obtained for the Spearman rank correlation coefficient between the old and new relative mutabilities, indicating little overall change. Ser (serine) and Thr (threonine) are found to be the most mutable residues in this work, as opposed to asparagine and serine in the 1978 table. Trp (tryptophan) and Cys (cysteine) are found to be least mutable here, which agrees with the earlier findings, though the mutability of Cys found here is double the original value. The frequencies of occurrence of the amino acid residues (Table I) show no significant differences from the earlier values.

Table IV. The difference matrix ($PET91_{ij}$ − $MDM78_{ij}$) between the 250 PAM PET91 matrix and the MDM78 matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	0	+1	0	0	+1	-1	-1	0	-1	+1	+1	0	0	+1	0	0	+1	+2	0	+1
R	+1	-1	0	0	+3	+1	+1	+3	0	-1	0	+1	-2	0	-1	-1	0	-2	+2	-1
N	0	0	+1	0	+2	-1	0	0	-1	0	0	0	0	+1	0	0	+1	-1	+1	0
D	0	0	0	+1	+2	-1	+1	0	-1	-1	0	0	0	+1	-1	0	-1	+2	+2	0
C	+1	+3	+3	+2	-1	+2	+1	+2	+3	0	+3	+2	+3	+4	+1	+1	+1	+3	+2	0
Q	-1	+1	-1	-1	+2	+1	0	0	-1	-1	0	+1	-1	+1	0	0	0	+2	+2	-1
E	-1	+1	0	+1	+1	0	+1	0	-1	-1	-1	+1	-1	0	-1	-1	-1	+2	0	0
G	0	+2	0	0	+2	0	0	0	0	0	0	+1	0	0	0	0	-1	+3	+1	-1
H	-1	0	-1	-1	+3	-1	-1	0	0	-1	0	+1	0	+2	0	0	0	0	+4	-1
I	+1	-1	0	-1	0	-1	-1	0	-1	-1	0	-1	+1	-1	0	0	+1	+1	-1	0
L	+1	0	0	0	+3	0	-1	0	0	0	-1	0	-1	0	+3	+1	+1	0	0	0
K	0	+1	0	0	+2	+1	+1	+1	+1	-1	0	0	-2	0	-1	-1	-1	0	+1	-1
M	0	-2	0	0	+3	-1	-1	0	0	+1	-1	-2	0	0	0	+1	+1	+1	0	0
F	+1	0	+1	+1	+4	+1	0	0	+2	-1	0	0	0	-1	+2	+1	+1	-1	+3	+1
P	0	-1	0	-1	+1	0	-1	0	0	0	+3	-1	0	+2	0	0	+1	+2	+2	0
S	0	-1	0	0	+1	0	-1	0	0	0	+1	-1	+1	+1	0	0	0	-1	+2	0
T	+1	0	+1	-1	+1	0	-1	-1	0	+1	+1	-1	+1	+1	+1	0	-1	+1	0	0
W	+2	-2	-1	+2	+3	+2	+2	+5	0	-1	0	0	-1	-1	+2	-1	-1	-2	0	+3
Y	0	+2	-1	+2	+2	+2	0	+1	+4	-1	0	+1	0	-3	+2	+2	0	0	-1	-1
V	+1	-1	0	0	0	-1	0	-1	-1	0	0	-1	0	+1	0	0	0	+3	-1	0

A positive matrix element indicates that the PET91 value is higher than the related value in MDM78. Absolute differences ≥ 2 are shown shaded.

Table IV shows the pattern of changes between the MDM78 and the PET91 matrices. Both Cys and Trp show very different patterns of mutability, both now showing a much greater tendency to exchange with other amino acid residues than in the previous study. This can be attributed mainly to the paucity of mutational events involving Cys and Trp in the original data set. Overall, in Dayhoff's data 35 amino acid exchanges were never observed at all (e.g. Cys and Trp); here, however, all possible exchanges have been observed (Cys and Trp exchanging 38 times in the current data set). PET91 incorporates 442 Trp exchanges and 1292 Cys exchanges, where only 7 Trp exchanges and 28 Cys exchanges were recorded for the MDM78 matrix. Interestingly, however, the average absolute change of the Cys matrix elements is higher than that of Trp, even though the Cys sample was larger than that of Trp in the 1978 data set. This anomaly is attributable to the fact that Cys residues occur in three very different chemical roles in proteins: as free sulphydryl groups (−S−H), in disulphide bridges (−S−S−), and as ligands for metals (−S..X). The number of observed cys exchanges in the original work would have been insufficient to sample these three situations effectively. In addition, the Cys residue exchanges observed in the original work were mostly from the metallothionein sequences included in the data set.

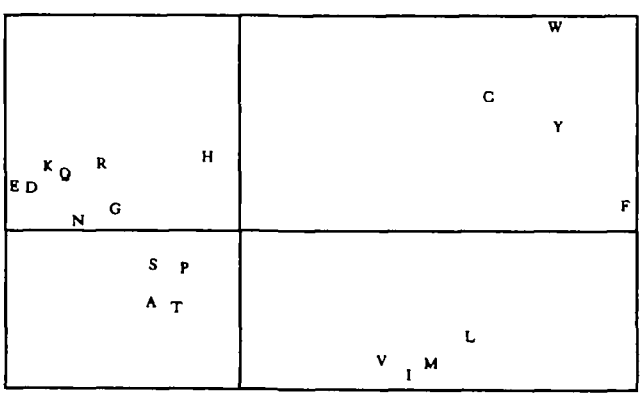

Fig. 3. The general trends in amino acid residue similarity shown in the PET91 relatedness odds matrix, visualized by means of multidimensional scaling.

It is also interesting to note that even with the very large amount of data collected here, some amino acid exchanges are still very seldom observed: Trp and Asn (asparagine), for example, were only seen to exchange twice. Indeed it is hard to be certain whether these highly infrequent exchanges are real observations or artefacts caused by errors in the sequence database.

A common method for interpreting the complex trends in a similarity matrix is to project the 20 × 20 = 400-dimensional

pattern onto a plane via multidimensional scaling (French and Robson, 1983). The plot in Figure 3 shows such a projection, which clearly delineates the relationships between the 20 amino acids found in PET91. The general trends shown in the PET matrix are essentially those found in the original Dayhoff matrix: hydrophobicity and size being the most significant factors.

Discussion

In general, the most significant differences (PET91 matrix elements differing from MDM78 elements by ±2 or more) correspond almost exactly to exchanges that were observed no more than once in Dayhoff's sequence alignments. Despite these few anomalous differences, however, it is interesting to see how little the bulk of PET91 differs from MDM78. The fundamental amino acid similarities remain unchanged, and given that we have now collected enough data to iron out the residual sampling errors in the mutation data matrix, we feel confident that PET91 represents a relatively unbiased measure of amino acid similarity in sequence data and should be used in preference to the MDM78 in sequence analysis applications. Investigation is currently under way as to the performance of our matrices compared to others with regard to sequence alignment and databank searching. We are also developing matrices biased to particular protein classes and residue environments, and a dipeptide mutability matrix (400 × 400 elements) which has enabled us to investigate short-range sequence neighbourhood effects on residue mutability.

The matrix generation programs and the complete data, including all intermediate matrices and tables required for constructing matrices for evolutionary distances other than 250 PAMs, may be obtained from the authors in printed or machine-readable form.

Acknowledgements

D.T.J. acknowledges receipt of a SERC CASE studentship with the MRC.

References

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **214**, 403–410.

Bairoch,A. (1990) PC/Gene: a protein and nucleic acid sequence analysis microcomputer package, PROSITE: a dictionary of sites and patterns in proteins and SWISS-PROT: a protein sequence data bank. Ph.D. thesis, University of Geneva.

Czelusniak,J., Goodman,M., Moncrief,N.D. and Kehoe,S.M. (1990) Maximum parsimony approach to construction of evolutionary trees from aligned homologous sequences. *Methods Enzymol.*, **183**, 601–615.

Dayhoff,M.O., Schwartz,R.M and Orcutt,B.C. (1978) In *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC, Vol. 5 Suppl. 3, pp. 345–352.

Feng,D.-F., Johnson,M.S. and Doolittle,R.F. (1985) Aligning amino acids sequences: comparison of commonly used methods. *J. Mol. Evol.*, **21**, 112–125.

Fitch,W.M. and Margoliash,E. (1967) Construction of phylogenetic trees. *Science*, **115**, 279–284.

French,S. and Robson,B. (1985) What is a conservative substitution? *J. Mol. Evol.*, **19**, 171–175.

George,D.G., Barker,W.C. and Hunt,L.T. (1990) Mutation data matrix and its uses. *Methods Enzymol.*, **188**, 333–351.

McLachlan,A.D. Test for comparing related amino acid sequences. Cytochrome c and cytochrome c551. (1971) *J. Mol. Biol.*, **61**, 409–424.

McLachlan,A.D. Repeating sequences and gene duplication in proteins. (1972) *J. Mol. Biol.*, **64**, 417–437.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Overington,J., Johnson,M.S., Šali,A. and Blundell,T.L. (1990) Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc. R. Soc. Lond. B*, **241**, 132–145.

Risler,J.L., Delorme,M.O., Delacroix,H. and Henaut,A. (1988) Amino acid substitutions in structurally related proteins. A pattern recognition approach. *J. Mol. Biol.*, **210**, 181–193.

Woese,C.R. (1969) Models for the evolution of codon assignments. *J. Mol. Biol.*, **43**, 235–240.

Received on October 21, 1991; accepted on December 6, 1991