22 A Model of Evolutionary Change in Proteins

M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt

In the eight years since we last examined the amino acid exchanges seen in closely related proteins,[1] the information has doubled in quantity and comes from a much wider variety of protein types. The matrices derived from these data that describe the amino acid replacement probabilities between two sequences at various evolutionary distances are more accurate and the scoring matrix that is derived is more sensitive in detecting distant relationships than the one that we previously derived.[2,3] The method used in this chapter is essentially the same as that described in the *Atlas*, Volume 3[4] and Volume 5.[1]

Accepted Point Mutations

An accepted point mutation in a protein is a replacement of one amino acid by another, accepted by natural selection. It is the result of two distinct processes: the first is the occurrence of a mutation in the portion of the gene template producing one amino acid of a protein; the second is the acceptance of the mutation by the species as the new predominant form. To be accepted, the new amino acid usually must function in a way similar to the old one: chemical and physical similarities are found between the amino acids that are observed to interchange frequently.

Any complete discussion of the observed behavior of amino acids in the evolutionary process must consider the frequency of change of each amino acid to each other one and the propensity of each to remain unchanged. There are 20 X 20 = 400 possible comparisons. To collect a useful amount of information on these, a great many observations are necessary. The body of data used in this study includes 1,572 changes in 71 groups of closely related proteins appearing in the *Atlas* volumes through Supplement 2.

The mutation data were accumulated from the phylogenetic trees and from a few pairs of related sequences. The sequences of all the nodal common ancestors in each tree are routinely generated. Consider, for example, the much simplified artificial phylogenic tree of Figure 78. The matrix of accepted point mutations calculated from this tree is shown in Figure 79. We have assumed that the likelihood of amino acid X replacing Y is the same as that of Y replacing X, and hence 1 is entered in box YX as well as in box XY. This assumption is reasonable, because this likelihood should depend on the product of the frequencies of occurrence of the two amino acids and on their chemical and physical similarity. As a consequence of this assumption, no change in amino acid frequencies over evolutionary distance will be detected.

By comparing observed sequences with inferred ancestral sequences, rather than with each other, a sharper

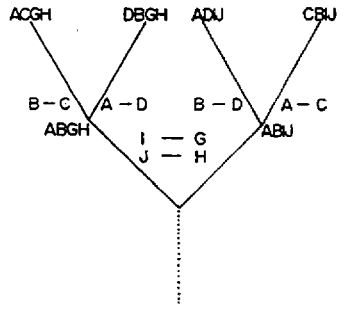

Figure 78. Simplified phylogenetic tree. Four "observed" proteins are shown at the top. Inferred ancestors are shown at the nodes. Amino acid exchanges are indicated along the branches.

	A	B	C	D	G	H	I	J
A			1	1				
B			1	1				
C	1	1						
D	1	1						
G							1	
H								1
I					1			
J						1		

Figure 79. Matrix of accepted point mutations derived from the tree of Figure 78.

picture of the acceptable point mutations is obtained. In the first amino acid position of the illustration in Figure 78, A changes to D and A changes to C, but C and D do not directly interchange. If we had compared the observed sequences directly, we would have inferred the change of C to D also. (In practice, some of the positions in the nodal sequences are blank [ambiguous]. For these, we have treated the changes statistically, distributing them among all observed alternatives.)

The total numbers of accepted point mutations observed between closely related sequences from 34 superfamilies, grouped into 71 evolutionary trees, are shown in Figure 80. In order to minimize the occurrence of changes caused by successive accepted mutations at one site, the sequences within a tree were less than 15% different from one another and ancestral sequences were even closer. Of the 190 possible exchanges shown in Figure 80, 35 were never observed. These usually involved the amino acids that occur infrequently and are not highly mutable and exchanges where more than one nucleotide of the codon must change. Of the 1,572 exchanges the largest number, 83, was observed between Asp and Glu, two chemically very similar amino acids with codons differing by one nucleotide. About 20% of the interchanges, far more than one would expect for such similar sequences, involved amino acids whose codons differed by more than one nucleotide. Presumably, in any one tree, changes at some of the amino acid positions are rejected by selection and multiple changes at the mutable sites are favored. Many of the changes expected from the mutations of one nucleotide in a codon are seldom observed. Presumably these mutations have occurred often but have been rejected by natural selection acting on the proteins. For example, there were no exchanges between Gly and Trp.

Mutability of Amino Acids

A complete picture of the mutational process must include a consideration of the amino acids that did not

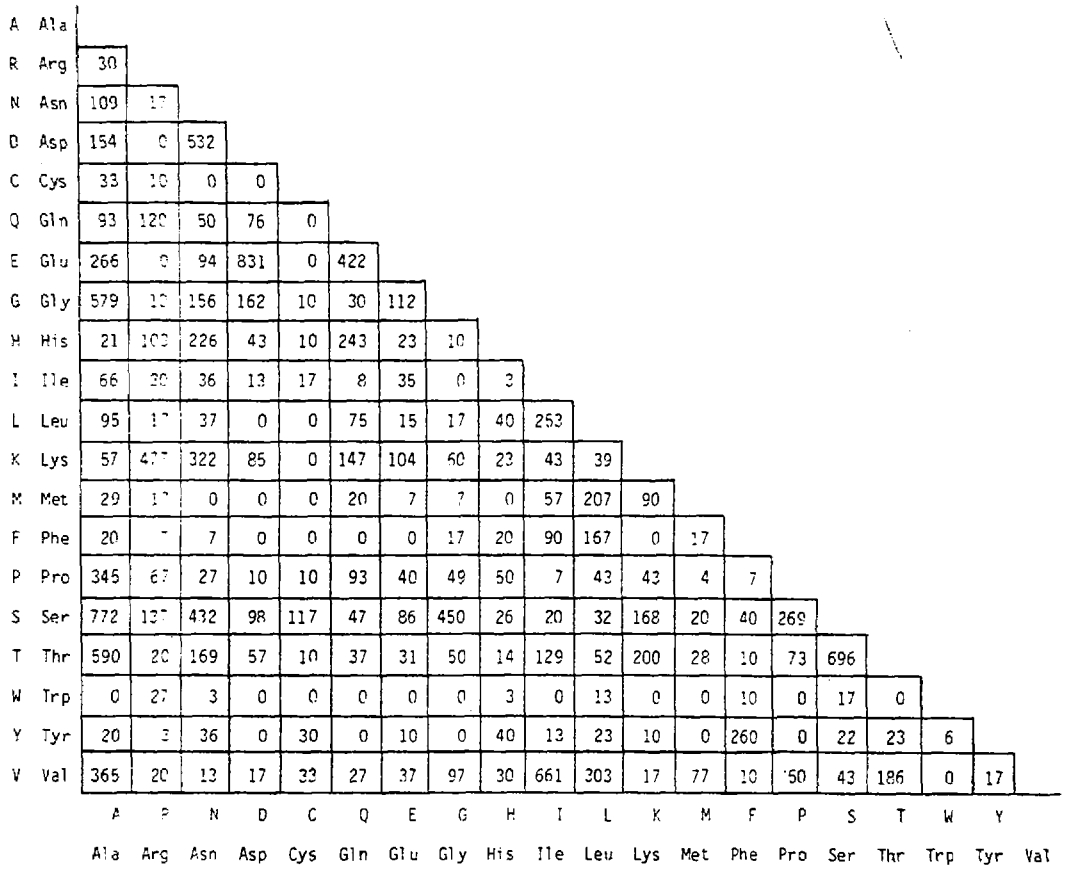

Figure 80. Numbers of accepted point mutations (X10) accumulated from closely related sequences. Fifteen hundred and seventy-two exchanges are shown. Fractional exchanges result when ancestral sequences are ambiguous.

change, as well as those that did. For this we need to know the probability that each amino acid will change in a given small evolutionary interval. We call this number the "relative mutability" of the amino acid.

In order to compute the relative mutabilities of the amino acids, we simply count the number of times that each amino acid has changed in an interval and the number of times that it has occurred in the sequences and thus has been subject to mutation. The relative mutability of each amino acid is proportional to the ratio of changes to occurrences. Figure 81 illustrates the computation for a simple case in which B changes relatively often, A less often, and D never.

Aligned sequences	A D A A D B		
Amino acids	A	B	D
Changes	1	1	0
Frequency of occurrence (total composition)	3	1	2
Relative mutability	.33	1	0

Figure 81. Sample computation of relative mutability. The two aligned sequences may be two experimentally observed sequences or an observed sequence and its inferred ancestor.

In calculating relative mutabilities from many trees, the information from sequences of different lengths and evolutionary distances is combined. Each relative mutability is still a ratio. The numerator is the total number of changes of this amino acid on all branches of all protein trees considered. The denominator is the total exposure of the amino acid to mutation, that is, the sum for all branches of its local frequency of occurrence multiplied by the total number of mutations per 100 links for that branch.

The relative mutabilities of the amino acids are shown in Table 21. On the average, Asn, Ser, Asp, and Glu are most mutable and Trp and Cys are least mutable.

The immutability of cysteine is understandable. Cysteine is known to have several unique, indispensible functions. It is the attachment site of heme groups in cytochrome and of FeS clusters in ferredoxin. It forms cross-links in other proteins such as chymotrypsin or ribonuclease. It seldom occurs without having an important function.

The substitution of one of the larger amino acids of distinctive shape and chemistry for any other is rather uncommon. At the other extreme, the low mutability of glycine must be due to its unique smallness that is advantageous in many places. Even though serine sometimes functions in the active center, it much more often performs a function of lesser importance, easily mimicked by several other amino acids of similar physical and chemical properties. On the average it is highly mutable.

Table 21
Relative Mutabilities of the Amino Acids[a]

Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Tyr	41
Ile	96	Phe	41
Met	94	Leu	40
Gln	93	Cys	20
Val	74	Trp	18

[a]The value for Ala has been arbitrarily set at 100.

Amino Acid Frequencies in the Mutation Data

The relative frequencies of exposure to mutation of the amino acids are shown in Table 22. These frequencies, f_i, are approximately proportional to the average composition of each group multiplied by the number of mutations in the tree. The sum of the frequencies is 1.

Mutation Probability Matrix for the Evolutionary Distance of One PAM

We can combine information about the individual kinds of mutations and about the relative mutability of the amino acids into one distance-dependent "mutation probability matrix" (see Figure 82). An element of this matrix, M_{ij}, gives the probability that the amino acid in column j will be replaced by the amino acid in row i after a given evolutionary interval, in this case 1 PAM.

Table 22
Normalized Frequencies of the Amino Acids in the Accepted Point Mutation Data

Gly	0.089	Arg	0.041
Ala	0.087	Asn	0.040
Leu	0.085	Phe	0.040
Lys	0.081	Gln	0.038
Ser	0.070	Ile	0.037
Val	0.065	His	0.034
Thr	0.058	Cys	0.033
Pro	0.051	Tyr	0.030
Glu	0.050	Met	0.015
Asp	0.047	Trp	0.010

The nondiagonal elements have the values:

$$M_{ij} = \frac{\lambda m_j A_{ij}}{\sum_i A_{ij}}$$

where

A_{ij} is an element of the accepted point mutation matrix of Figure 80,
λ is a proportionality constant, and
m_j is the mutability of the jth amino acid, Table 21.

The diagonal elements have the values:

$$M_{jj} = 1 - \lambda m_j$$

Consider a typical column, that for alanine. The total probability, the sum of all the elements, must be 1. The probability of observing a change in a site containing alanine (the sum of all the elements except M_{AA}) is proportional to the mutability of alanine. The same proportionality constant, λ, holds for all columns. The individual nondiagonal terms within each column bear the same ratio to each other as do the observed mutations in the matrix of Figure 80.

The quantity $100 \times \Sigma f_i M_{ii}$ gives the number of amino acids that will remain unchanged when a protein 100 links long, of average composition, is exposed to the evolutionary change represented by this matrix. This apparent evolutionary change depends upon the choice of λ, in this case chosen so that this change is 1 mutation. Since there are almost no superimposed changes, this also represents 1 PAM of change. If λ had been four times as large, the initial matrix would have represented 4 PAMs; the discussion which follows would not be changed noticeably.

ORIGINAL AMINO ACID

REPLACEMENT AMINO ACID	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R Arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N Asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D Asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C Cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q Gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E Glu	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G Gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H His	1	2	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I Ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L Leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K Lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F Phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S Ser	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W Trp	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y Tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

Figure 82. Mutation probability matrix for the evolutionary distance of 1 PAM. An element of this matrix, M_{ij}, gives the probability that the amino acid in column j will be replaced by the amino acid in row i after a given evolutionary interval, in this case 1 accepted point mutation per 100 amino acids. Thus, there is a 0.56% probability that Asp will be replaced by Glu. To simplify the appearance, the elements are shown multiplied by 10,000.

Simulation of the Mutational Process

For evaluating statistical methods of detecting relationships, for developing methods of measuring evolutionary distances between proteins, and for determining the accuracy of programs to construct evolutionary trees, we need to have examples of proteins at known evolutionary distances. The mutation probability matrix provides the information with which to simulate any amount of evolutionary change in an unlimited number of proteins. Further, we can start with one protein and simulate its separate evolution in duplicated genes or in divergent organisms. By considering many groups of sequences related by the same evolutionary history, a measure is readily obtained of the expected deviations due to random fluctuations in the evolutionary process.

If we only require that, on the average, one mutation takes place in the evolutionary interval of 1 PAM, we can use a simulation requiring one random number for each amino acid in the sequence, as follows: To determine the fate of the first amino acid, say Ala, a uniformly distributed random number between 0 and 1 is obtained. The first column of the mutation probability matrix (Figure 82) gives the relative probability of each possible event that may befall Ala (neglecting deletion for simplicity). If the random number falls between 0 and .9867, Ala is left unchanged. If the number is between .9867 and .9868, it is replaced with Arg, if it is between .9868 and .9872, it is replaced with Asp, and so forth. Similarly, a random number is produced for each amino acid in the sequence, and action is taken as dictated by the corresponding column of the matrix. The result is a simulated mutant sequence. Any number of these can be generated; their average distance from the original is 1 PAM although some may have no mutations and some may have two or more. The effects on the sequence of a longer period of evolution may be simulated by successive applications of the matrix to the sequence resulting from the last application.

For simulations in which a predetermined number of changes are required, a two-step process involving two random numbers for each mutation can be used. Starting with a given sequence, the first amino acid that will mutate is selected: the probability that any one will be selected is proportional to its mutability (Table 21). Then the amino acid that replaces it is chosen. The probability for each replacement is proportional to the elements in the appropriate column of Figure 82. Starting with the resultant sequence, a second mutation can be simulated, and so on, until a predetermined number of changes have been made. In this process, superimposed and back mutations may occur.

The 1 PAM matrix can be multiplied by itself N times to yield a matrix that predicts the amino acid replacements to be found after N PAMs of evolutionary change in a sequence of average composition. On the average, the results of the simulations above match the predictions of the corresponding matrices.

Mutation Probability Matrices for Other Distances

The mutation probability matrix M_1, corresponding to 1 PAM, has a number of interesting properties (see Figure 82). If, in a simulation, it is applied to a protein with the average amino acid composition given in Table 22, on the average, the composition of the resulting mutated proteins will be unchanged. Repeated applications of the matrix to proteins of any other composition will give mutants that change toward average composition; any such matrix has implicit in it some particular asymptotic composition.

There is a different mutation probability matrix for each evolutionary interval. These can be derived from the one for 1 PAM by matrix multiplication. If the 1-PAM matrix is multiplied by itself an infinite number of times, each column of the resulting matrix approaches the asymptotic amino acid composition:

$$M_\infty = \begin{matrix} f_A & f_A & f_A & f_A & \cdots \\ f_R & f_R & f_R & f_R & \cdots \\ f_N & f_N & & & \\ \vdots & & & & \end{matrix}$$

At a great distance, there is very little relationship information left in the matrix. For example, at a distance of 2,034 PAMs, all of the matrix values are within 5% of their limiting values except for the Trp-Trp element, which is 75% higher than the limit, and the Cys-Cys element, which is 11% higher.

The matrix for 0 PAMs is simply a unit diagonal; no amino acid would have changed:

$$M_0 = \begin{matrix} 1 & 0 & 0 & \cdots \\ 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & \\ \vdots & & & \end{matrix}$$

The mutation probability matrix for 250 PAMs is shown in Figure 83. At this evolutionary distance, only one amino acid in five remains unchanged. However, the

amino acids vary greatly in their mutability; 55% of the tryptophans, 52% of the cysteines and 27% of the glycines would still be unchanged, but only 6% of the highly mutable asparagines would remain. Several other amino acids, particularly alanine, aspartic acid, glutamic acid, glycine, lysine, and serine are more likely to occur in place of an original asparagine than asparagine itself at this evolutionary distance! This is understandable from the data giving the preferred mutations and the relative mutabilities. Asparagine is highly mutable, therefore it changes to other amino acids. These are less mutable and may not change again. This effect is much more conspicuous in the case of methionine. Surprisingly, a methionine originally present would have changed to leucine in 20% of the cases, but would remain methionine in only 6%. Over one-third of the mutations in methionine are specifically to leucine (Figure 80). Leucine is less than one-half as mutable as methionine (Table 21).

From the series of distance-dependent mutation probability matrices, we can compute detailed answers to the question "How does the evolutionary process affect the similarity of related protein sequences?"

Estimation of Evolutionary Distance

There is a different mutation probability matrix for each evolutionary interval measured in PAMs. For each such matrix, we can calculate the percentage of amino acids that will be observed to change on the average in the interval by the formula:

$$100\left(1 - \sum_i f_i M_{ii}\right)$$

Table 23 shows the correspondence between the observed percent difference between two sequences and the evolutionary distance in PAMs. We use this scale to estimate

ORIGINAL AMINO ACID

REPLACEMENT AMINO ACID	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	13	6	9	9	5	8	9	12	5	8	6	7	7	4	11	11	11	2	4	9
R Arg	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N Asn	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D Asp	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C Cys	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q Gln	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E Glu	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G Gly	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H His	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I Ile	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L Leu	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K Lys	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M Met	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F Phe	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P Pro	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S Ser	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T Thr	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W Trp	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y Tyr	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V Val	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

Figure 83. Mutation probability matrix for the evolutionary distance of 250 PAMs. To simplify the appearance, the elements are shown multiplied by 100. In comparing two sequences of average amino acid frequency at this evolutionary distance, there is a 13% probability that a position containing Ala in the first sequence will contain Ala in the second. There is a 3% chance that it will contain Arg, and so forth. The relationship of two sequences at a distance of 250 PAMs can be demonstrated by statistical methods.

Table 23
Correspondence between Observed Differences and the Evolutionary Distance

Observed Percent Difference	Evolutionary Distance in PAMs
1	1
5	5
10	11
15	17
20	23
25	30
30	38
35	47
40	56
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246
85	328

evolutionary distances from matrices of percent difference between sequences. These estimated distances were used in the computations of evolutionary trees in this book. The differences predicted for a given PAM distance differ by up to 2.5% from those that we reported in Volume 5. A more complete scale is given in Table 36 of the Appendix.

Relatedness Odds Matrix

The elements, M_{ij}, of the mutation probability matrix for each distance give the probability that amino acid j will change to i in a related sequence in that interval. The normalized frequency f_i gives the probability that i will occur in the second sequence by chance.

The terms of the relatedness odds matrix are then:

$$R_{ij} = \frac{M_{ij}}{f_i}$$

The odds matrix is symmetrical. Each term gives the probability of replacement per occurrence of i per occurrence of j.

Amino acid pairs with scores above 1 replace each other more often as alternatives in related sequences than in random sequences of the same composition whereas those with scores below 1 replace each other less often.

The information in the 250-PAM odds matrix has proven very useful in detecting distant relationships between sequences. When one protein is compared with another, position by position, one should multiply the odds for each position to calculate an odds for the whole protein. However, it is more convenient to add the logarithms of the matrix elements. The log of the 250-PAM odds matrix is shown in Figure 84.

The Chemical Meaning of Amino Acid Mutations

Patterns have been visible in the accepted point mutations since the beginning of protein sequence work. Isoleucine-valine and serine-threonine were frequently observed alternatives. It was obvious that this interchangeability had something to do with their chemical similarities. In the large amount of information that now exists, far more detailed correlations are visible, and many more functional inferences can be made.

In the log odds matrix of Figure 84, the order of the amino acids has been rearranged to show clearly the groups of chemically similar amino acids that tend to replace one another: the hydrophobic group; the aromatic group; the basic group; the acid, acid-amide group; cysteine; and the other hydrophilic residues. Some groups overlap: the basic and acid, acid-amide groups tend to replace one another to some extent, and phenylalanine interchanges with the hydrophobic group more often than chance expectation would predict. These patterns are imposed principally by natural selection and only secondarily by the constraints of the genetic code: they reflect the similarity of the functions of the amino acid residues in their weak interactions with one another in the three-dimensional conformation of proteins. Some of the properties of an amino acid residue that determine these interactions are: size, shape, and local concentrations of electric charge; the conformation of its van der Waals surface; and its ability to form salt bonds, hydrophobic bonds, and hydrogen bonds.

Computing Relationships between Sequences

We use log odds matrices as scoring matrices for detecting very distant relationships between proteins. Such scoring matrices, based ultimately on accepted point mutations, can discriminate significant relationships from

C	Cys	12																		
S	Ser	0	2																	
T	Thr	-2	1	3																
P	Pro	-3	1	0	6															
A	Ala	-2	1	1	1	2														
G	Gly	-3	1	0	-1	1	5													
N	Asn	-4	1	0	-1	0	0	2												
D	Asp	-5	0	0	-1	0	1	2	4											
E	Glu	-5	0	0	-1	0	0	1	3	4										
Q	Gln	-5	-1	-1	0	0	-1	1	2	2	4									
H	His	-3	-1	-1	0	-1	-2	2	1	1	3	6								
R	Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6							
K	Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5						
M	Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6					
I	Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5				
L	Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6			
V	Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4		
F	Phe	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9	
Y	Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10
W	Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
		Cys	Ser	Thr	Pro	Ala	Gly	Asn	Asp	Glu	Gln	His	Arg	Lys	Met	Ile	Leu	Val	Phe	Tyr	Trp

Figure 84. Log odds matrix for 250 PAMs. Elements are shown multiplied by 10. The neutral score is zero. A score of -10 means that the pair would be expected to occur only one-tenth as frequently in related sequences as random chance would predict, and a score of +2 means that the pair would be expected to occur 1.6 times as frequently. The order of the amino acids has been arranged to illustrate the patterns in the mutation data.

random coincidences better than simpler scoring systems. Mere counts of identities and matrices based only on the changes predicted by the genetic code are not sufficiently complex. It is obvious that there is a good deal of information in the detailed nature of both the nonidentities and the identities. Certain combinations of different amino acids are positive evidence of relatedness, and others are contraindications. The log odds matrix for 250 PAMs, which we have found to be a very effective scoring matrix for detecting distant relationships, is compared with other matrices in chapter 23.

References

1. Dayhoff, M.O., Eck, R.V., and Park, C.M., *in* Atlas of Protein Sequence and Structure 1972, Vol.5, ed. Dayhoff, M.O., pp.89-99, Nat. Biomed. Res. Found., Washington, D.C., 1972
2. Schwartz, R.M., and Dayhoff, M.O., *in* Evolution of Protein Molecules, ed. Matsubara, H., and Yamanaka, T., pp.1-16, Japan Sci. Soc. Press, Tokyo, 1978
3. Schwartz, R.M., and Dayhoff, M.O., *in* Origin of Life, ed. Noda, H., pp.457-469, Center for Academic Pub. Japan/Japan Sci. Soc. Press, Tokyo, 1978
4. Dayhoff, M.O., and Eck, R.V., Atlas of Protein Sequence and Structure 1967-68, pp.33-45, Nat. Biomed. Res. Found., Silver Spring, Md., 1968