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

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

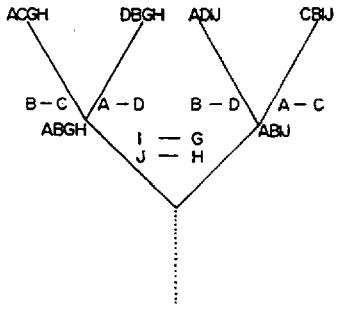


Figure 78. Simplified phylogenetic tree. Four "observed" proteins are shown at the top. Inferred ancestors are shown at the nodes. Amino acid exchanges are indicated along the branches.

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

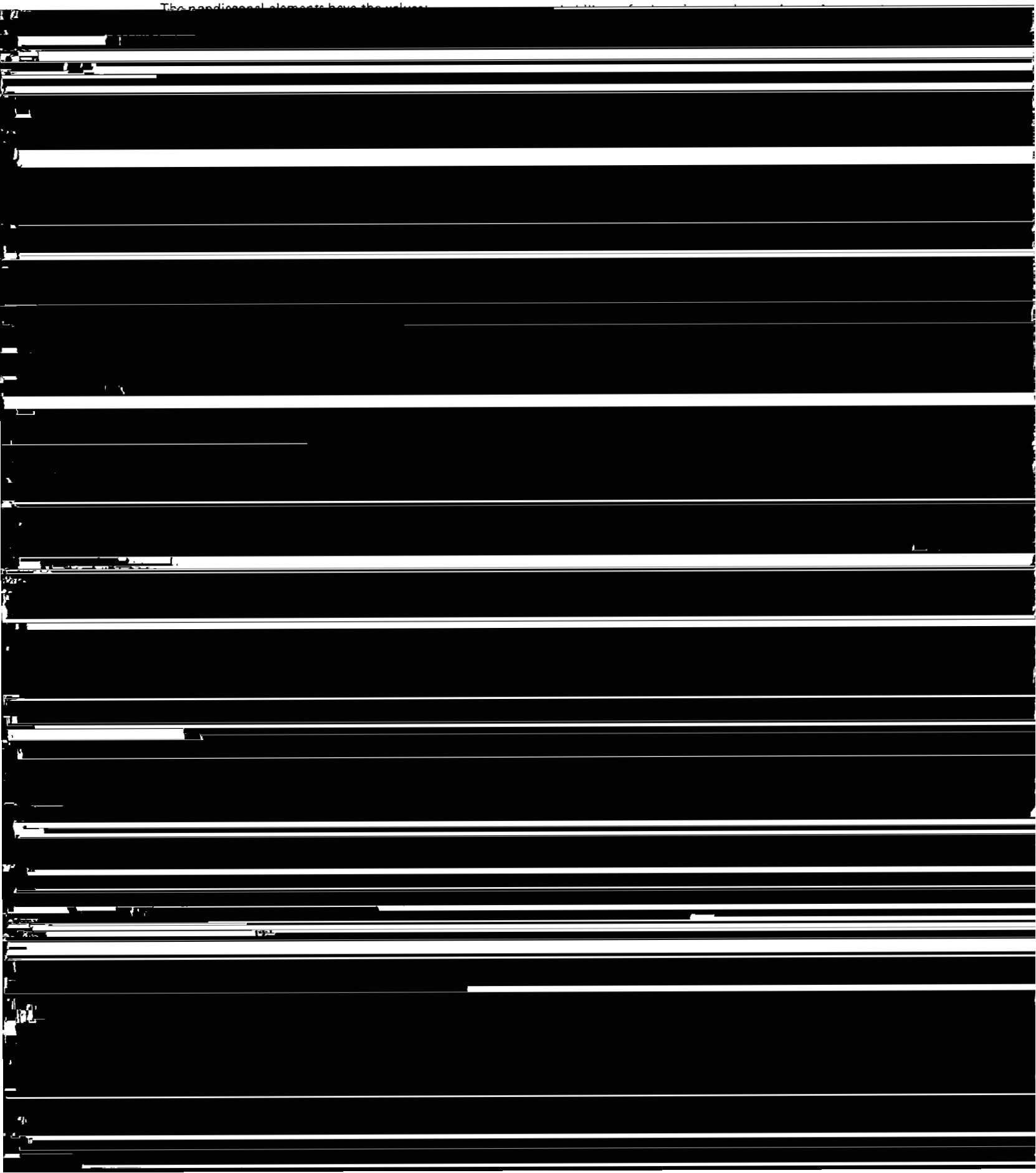
change, as well as those that did. For this we need to know the probability that each amino acid will change in a given small evolutionary interval. We call this number the "relative mutability" of the amino acid.

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

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

The nondiagonal elements have the values:



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

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

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

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 evolu-

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

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 muta-

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

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