ANALYSIS OF AMINO ACID SUBSTITUTION DURING DIVERGENT EVOLUTION: THE 400 BY 400 DIPEPTIDE SUBSTITUTION MATRIX

Gaston H. Gonnet, Mark A. Cohen, and Steven A. Benner

Institutes of Scientific Computation and Organic Chemistry
Swiss Federal Institute of Technology, CH-8092 Zurich, Switzerland

Received January 3, 1994

Most formal methods for analyzing the divergent evolution of protein sequences assume a Markov model where position *i* in a polypeptide chain undergoes amino acid substitution independently from position *i*+1. The large number of aligned homologous sequence pairs available from the exhaustive matching of the protein sequence database makes it possible to examine this assumption empirically. We have constructed a 400 by 400 matrix that reports empirical probabilities for the interconversion of all pairs of dipeptides in proteins undergoing divergent evolution. Comparison of these probabilities with those expected if substitution at adjacent positions in a protein sequence were independent reveals interesting patterns that arise through the breakdown of this assumption. Several of these are useful in extracting conformational information from patterns of conservation and variation in homologous protein sequences.

In analyzing alignments of homologous protein sequences undergoing divergent evolution subject to functional constraints, matches and mismatches are generally scored using a 20 by 20 "log-odds" or "Dayhoff" matrix (1,2). Each element of this matrix reports the logarithm of the probability that the index amino acids are matched by reason of ancestry divided by the probability that they are matched by random chance. These probabilities come from empirical data, the frequencies with which each of the 210 possible matches of the 20 natural proteinogenic amino acids are found in a sample of aligned homologous protein sequences.

Such a matrix can be used to score an alignment if a Markov model for amino acid substitution is assumed. Two premises of the Markov model are: (a) amino acid substitutions subsequent in time are independent of preceding substitutions, and (b) substitutions at specific positions in the protein sequence are independent of substitutions elsewhere in the sequence, in particular, substitution at position *i* is independent of substitution at positions *i*+1 and *i*-1.

Even in Dayhoff's most advanced matrix (\2), the number of sequences available was small, and there was an insufficient number of matched amino acid pairs to sustain an analysis of amino acid substitution any more sophisticated than that implied by the Markov model. Today, sequence information is no longer scarce. Some 2 million aligned protein sequence pairs are now available from the exhaustive matching of the protein sequence database (3). Even after redundant alignments are removed, the number of aligned positions in pairwise alignments, which provide the raw data

0006-291X/94 $5.00

needed for the construction of empirical mutation matrices, is more than 100 fold greater than that available to Dayhoff. Thus, it is timely to examine empirically the Markov model in greater detail.

We recently reported that the first assumption of the Markov model, the independence of substitution patterns as a function of time, breaks down in an interesting way with real protein sequences (4). Therefore, we have examined more closely the second assumption, the assumed independence of amino acid substitution at positions *i* and *i*+1. We report here the first analysis of substitution matrices describing the probabilities for interconversion of all 400 possible dipeptides during divergent evolution. Substitution at adjacent positions proves not to be independent, and in an interesting way.

METHODS

Sequence data were manipulated using the DARWIN system (5), available in a version that operates on Sun, DEC or MIPS workstations under Unix. Many of DARWIN's routines are available on a server via computer mail to cbrg@inf.ethz.ch. The analysis here is based on 1.7 million pairs of aligned protein sequences found by the exhaustive matching of Version 64 of the MIPS protein sequence data base (\3), aligned using gap scoring penalties reported elsewhere (6). Subsequent analyses have used the SWISS-PROT database (7). These aligned sequence pairs are also available to the public in computer readable form.

The procedures used to obtain alignments of indisputable quality are described in detail elsewhere (\4,\6). In summary, all alignments had scores of 150 or greater, where the score is ten times the log of the probability that the aligned sequences are related by common ancestry divided by the probability that their similarities arose by chance. Sequence pairs had PAM distances (the number of Point Accepted Mutations per 100 amino acids separating the two sequences)(\1) greater than 4 and fewer than 100. All alignments had more than 80 matched amino acids.

These criteria ensured that the alignments paired homologous positions in the sequence pairs (that is, that aligned amino acids are descendents of a specific codon in an ancestral gene). Two independent methods were used to demonstrate this (\4,\6). First, artificial sequence pairs were generated by a process that simulated evolutionary divergence from a single authentic sequence. The generated sequences were then aligned, and correspondence between the alignment and the true evolutionary relation between the two sequences, known from the process by which the sequence pair was generated, was used to measure the quality of the alignment procedure. Second, all pairs of aligned protein sequences where crystallographic secondary structures were available for both proteins were extracted from the database. The quality of the alignment was then assessed by its ability to orient the secondary structural units. Details of this comparison are reported elsewhere (\4).

Dipeptide substitution data were collected for pairwise alignments in four PAM bands (6.25-12.5 PAM, 12.5-25 PAM, 25-50 PAM, and 50-100 PAM). The parameters of the data sets are given in Table 1. To avoid having the substitution data biased by protein families that happen to be heavily represented in the protein sequence database, only a single aligned pair was examined between connected components joined by a pairwise alignment within each PAM window, where a connected component was defined as a set of protein sequences joined with each other by pairwise alignments with PAM distances less than the lower bound of each window (see references \4 and \6 for details. Data from pairwise alignments within each band yielded mutation matrices where each element is the sum of the individual counts of each transition within that PAM band. Each data set was normalized (so that the terms sum to 1) to generate a mutation matrix. The resulting transition matrices (M_a, M_b, M_c and M_d) were then treated to give an average matrix ($\bar{M}$) which could be further manipulated as described elsewhere (\4).

$$\bar{M} = \frac{M_a^8 + M_b^4 + M_c^2 + M_d}{4} .$$

The resulting matrix was compared with a 400 by 400 dipeptide matrix calculated from the 20 x 20 matrix that was built assuming that substitutions at adjacent positions occur independently. This was examined as a matrix of elements R, such that:

$$R = 10^{*}\log_{10}\left(\frac{\text{observed mutation probability}}{\text{probability of two independent mutations}}\right).$$

RESULTS AND DISCUSSION

The elements (R) of the dipeptide matrix are of the general form αX->βZ, where α, β, X and Z each represent any of the 20 naturally encoded amino acids. The matrix is presented as 400 individual 20 by 20 submatrices of the form αX->αZ. A sample of these is shown in Figure 1.

Log-Relative mutation matrix for (Ala-X) --> (Ala-Z)

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

The diagonal average value is -1.0 (or 80.1% of random)
The off-diagonal average value is -1.1 (or 77.4% of random)

Log-Relative mutation matrix for (α-Ala) --> (β-Ala)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	-1	-1	1	0	-2	0	1	1	-2	1	-1	-1	-1	-1	0	0	2	-2	-3	1
R	-2	-1	-1	-3	2	0	-2	0	1	-2	0	0	0	-2	-1	-1	-3	2	-3	-3
N	-1	-2	-1	0	-1	-2	-1	-2	0	0	-2	-1	-1	-1	-3	0	-1	-4	-1	-2
D	-1	-3	0	-1	-2	-2	0	-1	-2	1	-2	-3	-3	0	-5	-3	-3		-1	2
C	-4	2	-1	-2	0	-1	-3	-1	0	-4	-4	1	-6	1	-1	-1	-3	-3	2	-3
Q	-1	1	-1	-1	1	0	-1	-2	2	-1	-1	0	-1	-4	-1	-2	-3	1	-2	-1
E	0	-1	0	1	-2	0	-1	1	-2	0	-1	-1	-2	-4	-3	-2	-2	-3	-4	0
G	-1	0	-2	0	0	-3	1	-1	-3	1	-2	-1	-1	-3	-3	-1	0	-4	-4	2
H	-4	1	0	-3	0	0	-3	-3	-1	-2	0	-3	-4	-1	0	-4	-6	-3	1	-2
I	0	-2	0	1	-3	-2	-1	1	-1	-1	-1	-2	0	-1	-3	0	1	-2	-4	0
L	-2	0	-1	-1	-3	-1	-1	-1	1	-1	-1	-2	0	-1	1	1	-1	-1	-3	0
K	-2	0	-1	-2	1	-1	-2	-1	-2	-1	-2	-1	-1		-3	-2	-1	1	-3	-2
M	-1	1	0	-2	-4	-1	-2	0	-2	1	1	0	0	-2	-1	-2	3	-5	-2	2
F	-2	-2	-1	0	1	-5	-5	-3	-1	-1	-1		-3	-1	-2	2	-3	-3	0	0
P	0	0	-2	-4	0	-1	-4	-2	1	-3	1	-3	-1	-1	-2	0	-1			-2
S	0	0	0	-2	0	-2	-3	0	-3	0	1	-2	-3	3	0	-2	-1	0	-2	0
T	1	-3	0	-3	-2	-3	-3	0	-5	1	-1	-1	2	-2	-2	-1	-1	-4	-4	-1
W	-3	2	-4		-3	0	-4	-4	-2	-3	-1	1	-6	-3		-1	-4	0	-2	-1
Y	-4	-4	-1	-2	1	-3	-6	-5	1	-4	-4	-4	-4	-1		-3	-5	-2	0	-3
V	0	-2	-1	2	-2	-1	-1	2	-1	1	0	-2	1	0	-2	-1	-1	0	-2	-1

The diagonal average value is -0.9 (or 81.9% of random)
The off-diagonal average value is -1.4 (or 73.2% of random)

Figure 1. Examples of submatrices of the log-relative substitution matrices for substitutions of the type αX->αZ or Xα->Zα abstracted from the 400 x 400 dipeptide substitution matrix. Only entries represented by ≥2 transitions are shown. The matrix elements are R values, where R = 10*log$_{10}$(observed mutation probability/predicted mutation probability), where the predicted mutation probability is based on a log-odds matrix calculated for the entire protein sequence database (Gonnet *et al.*, 1992) assuming that adjacent positions undergo substitution independently.

Table 1. Population of Data within the 400 x 400 Dipeptide Substitution Matrix

Observed Events	Number of matrix elements	percentage
0	36674	45.7
1	13588	16.9
2	6870	8.6
3	4110	5.1
4	2809	3.5
5	1932	2.4
6	1506	1.9
7	1104	1.4
8	925	1.2
9	792	1.0
10	686	0.9
>10	9204	11.5
Total	80200	100.

A matrix element represents the probability that any one of 80,200 possible dipeptide pairs will be matched in a pairwise alignment of two homologous sequences. Counts represent the number of times such a pairing was observed in the database of 1,743,134 dipeptides considered here.

Elements of the matrix are ten times the logarithm of the probability that the dipeptide match appears in the database divided by the expected probability of the match were adjacent positions in a polypeptide chain to undergo substitution independently. This latter probability was calculated from the broadly based log-odds matrix reported in Cohen *et al.* (\4). The 400 x 400 matrix is available to interested individuals in computer readable form.

To derive the matrix, 1,743,134 matched dipeptides were considered. Of these, 1,071,219 were unsubstituted, 506,251 had one substitution, and 165,664 had two substitutions. Since every substitution participated in two dipeptide matchings, 418,789 (24%) of the aligned positions had suffered point mutation, while 1,324,344 (76%) of these positions were not mutated. As there are 80,200 distinct matches between all possible dipeptides, the average number of substitutions per matrix element was 5.2, very similar to that in the 20 by 20 matrix provided by Dayhoff (\2).

As is clear from Table 1, the 80,200 possible dipeptide pairs are not equally represented in the sequence database. Nearly 46 percent of the possible dipeptide matches are unrepresented. Conversely, some 11 percent of the possible dipeptide matches are found more than ten times in the database. This behavior is far from that expected if amino acid substitution were to occur independently at adjacent positions. Therefore, we asked whether information might be obtained in the way that the Markov assumption of independent adjacent substitution breaks down.

The most important observation is that variability at a position in a sequence correlates with variability at adjacent positions, with the degree of correlation depending on amino acid type in a rationalizable way. For example, matches where only one amino acid in a dipeptide is conserved are uniformly less probable (the relative substitution probability is negative) than would be expected if adjacent positions underwent substitutions independently. Showing this, Table 2 reports the average diagonal terms in the submatrices of the form αX->αZ (representing probabilities where both amino acids of the dipeptide are conserved), the average off-diagonal terms in these submatrices (where the first amino acid in a dipeptide is conserved and the second is not), and the average difference (diagonal minus off-diagonal) for the submatrices. In most cases, this difference is positive, implying that residue *i*+1 is more likely to be conserved if residue *i* is conserved.

Table 2. Average diagonal and off-diagonal terms for selected submatrices of the 400 x 400 dipeptide substitution matrix

Index Amino Acid	Submatrix		Diagonal	Off-Diagonal	Difference
Pro	Pro-X	Pro-Z	-10.7	1.8	-12.5
Gly	Gly-X	Gly-Z	-7.4	-3.5	-3.9
Glu	Glu-X	Glu-Z	-6.3	-4.2	-2.1
Lys	Lys-X	Lys-Z	-6.1	-6.1	0.0
Asp	Asp-X	Asp-Z	-9.3	-9.9	0.6
Ser	Ser-X	Ser-Z	-9.7	-10.9	1.2
Leu	Leu-X	Leu-Z	-4.4	-5.9	1.5
Ala	Ala-X	Ala-Z	-9.6	-11.1	1.5
Asn	Asn-X	Asn-Z	-11.3	-15.1	3.8
Arg	Arg-X	Arg-Z	-7.3	-12.1	4.8
Gln	Gln-X	Gln-Z	-2.9	-7.9	5.0
Thr	Thr-X	Thr-Z	-9.7	-15.1	5.4
Phe	Phe-X	Phe-Z	-2.7	-8.4	5.7
Ile	Ile-X	Ile-Z	-7.8	-14.9	7.1
Tyr	Tyr-X	Tyr-Z	-1.3	-9.3	8.0
Val	Val-X	Val-Z	-7.2	-15.5	8.3
Cys	Cys-X	Cys-Z	-2.7	-11.2	8.5
Trp	Trp-X	Trp-Z	1.5	-9.0	10.5
His	His-X	His-Z	-7.4	-23.7	16.3
Met	Met-X	Met-Z	-3.7	-20.5	16.8

Diagonal and off-diagonal terms are 10 x log$_{10}$(probability / expected probability). The difference (diagonal - off-diagonal) is most negative when the amino acid at position *i*+1 is most likely to be variable (when compared with the probability that it will be conserved) when the index amino acid at position *i* is conserved.

This implication is, in itself, not surprising. A residue at position *i* may be conserved because it lies inside the folded structure of the protein (8,9). The probability is therefore increased that position *i*+1 also lies inside, and therefore will also be conserved. Conversely, if residue *i* is conserved because it is at the active of an enzyme, then residue *i*+1 is also likely to be near the active site, and therefore is also more likely to be conserved.

What is surprising, however, is that the rule depends strongly on *which* amino acid is conserved at position *i*. In particular, conservation at position *i*+1 is greatest when the conserved residue at position *i* is hydrophobic (e.g. Val), or is likely to be conserved at an active site (e.g. His). In contrast, when the conserved residue at position *i* is hydrophilic, the generalization applies only slightly (Ser, Asp) or not at all (Lys, Glu). This is consistent with the explanation above; if residue *i* is indeed inside the folded structure, it is more likely to be hydrophobic.

Proline and glycine provide the most striking exceptions to this the rule. If either Pro or Gly is conserved at position *i*, position *i*+1 is considerably more likely to be *variable*. This observation suggests that a conserved Gly or, more strongly, a conserved Pro indicates a "parse", a segment of the polypeptide that separates standard secondary structural elements (\8,9). Because secondary structure disruption generally occurs on the surface of the folded protein, position *i*+1 is more likely to be on the surface, and therefore more likely to tolerate variation, if position *i* contains a conserved Pro or Gly.

Conversely, matchings where both amino acids in a dipeptide undergo substitution are uniformly more probable than expected if independent substitution at positions *i* and *i*+1 is assumed

Log-Relative mutation matrix for (Asp-X) --> (Glu-Z)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	0	1	3	3		1	2	3	2	4	0	1	2	1	5	3	5		2	3
R	1	0	2	1		3	1	3	4	1	2	2	3	3	4	2	2		0	-1
N	3	2	0	2		1	2	4	3	3		2	3	3	3	4	3		2	3
D	3	1	3	1		2	5	5	2		4	1	4		2	2	1		5	5
C	-1		4	6	1			4		0			1	3	7	2	1		6	-1
Q	1	3	1	2		1	1	1	4	1	3	2	0	2	4	2	1			1
E	3	1	3	5		2	1	7	2	3	3	2	4	3	3	3	2		3	3
G	2	3	3	5	4	2	5	1	3	4	2	2	4	3	2	3	4			4
H	1	4	3	1		3	0	0	0			1	0	2	4	2	0		3	1
I	4	3	4	4	3	2	1	6	3	0	2	1	3	2	4	4	4			3
L	1	3	2	3	1	3	1	2	5	2	1	0	3	3	4	4	3	1	2	2
K	1	3	4	1		2	1	4	3	1	0	1	3		2	2	3	4		1
M	1	4	4		2	1			3	2	2	2	0	-2		2	5		1	3
F	1	2	2	4	1			3	3	1	2	2	-1	1		6	1		4	2
P	5	4	3	0		4	1	3	5	3	4	1			0	4	3		5	3
S	3	3	4	3	3	1	2	4	4	4	3	2	2	6	5	0	3		4	3
T	4	1	2	2	2	1	1	5	2	4	2	2	4	1	4	1	-1		0	2
W		0					4	5			2			0				2	-3	5
Y	2	1	5	3	5	4	0		5	1	2	2	0	3		2	1	-1	2	1
V	2	0	2	6	2	2	2	6	3	2	3	1	4	3	4	4	2		1	0

X is the vertical index; Z is the horizontal index.
The diagonal average value is 0.6 (or 115.2% of random)
The off-diagonal average value is 2.2 (or 164.1% of random)

Log-Relative mutation matrix for (α-Asp) --> (β-Glu)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	0	1	5	4	0	1	2	3	2	3	1	2	1	0	5	3	4		2	2
R	2	0	3	1	5	3	2	4	5	2	2	3	2	2	2	3	2	2	1	2
N	3	2	-1	2		1	2	2	2		1	1			2	4	2			1
D	4	2	3	0		2	5	4	3		4	2			2	3	2		3	6
C	0	3	3		1		4			1	-2	5		3	5	4	5		3	0
Q	1	3	2	2		0	1	2	5	1	3	2	2	1	4	1	1	4	2	3
E	4	2	4	6	5	2	1	7	3	3	3	3	3	3	4	3	3			4
G	2	4	2	4	6	2	6	1	1		1	3			3	4	4			5
H	1	4	1	0		3	-1	0	-1		3	1	1	2	4	2	0		3	1
I	4	1	3	4	2		2		1	1	1	1	2	3	4	4	4		1	2
L	2	2	3		3	4	2	4	3	2	1	0	2	1	6	4	3	-1	1	3
K	3	4	3	3		2	2	4	2	2	3	1	2		2	2	3		0	3
M	2	3	6		2	0	1		1	3	2	0	0	0	5	3	4		0	4
F		2	2		2	2	1	3	2	3	1	1	0	1	4	6	0	-2	3	1
P	5	3		1		4	1	4	6	5	5	3	6		1	5	4			3
S	3	2	4	2	4	1	2	4	4	4	4	1	2	5	4	1	3		1	3
T	4	2	4	1	2	1	1	4	0	5	2	1	3	0	2	2	0		-1	1
W		3									2	4	3	0			5	1	-1	
Y	2	-1	4	4	4	3	2		3	1	0	-1		3	2	2		-2	1	
V	3	2	4	6	-1	2	2	5		2	2	2	4	1	3	3	2		-1	0

α is the vertical index; β is the horizontal index.
The diagonal average value is 0.5 (or 112.6% of random)
The off-diagonal average value is 2.1 (or 160.9% of random)

Figure 2. Examples of log-relative substitution matrices for substitutions of the type αX->βZ or Xα->Zβ abstracted from the 400 x 400 dipeptide substitution matrix. Only entries represented by ≥2 transitions are shown. The matrix elements are R values, where R = 10*$\log_{10}$(observed mutation probability/predicted mutation probability), where the predicted mutation probability is based on a log-odds matrix calculated for the entire protein sequence database (Gonnet *et al.*, 1992) assuming that adjacent positions undergo substitution independently.

(Figure 2). This fact can be extracted from submatrices of type αX-βZ, where the first amino acid of the dipeptide is not conserved. The average diagonal term of the submatrix (where X and Z are the same amino acid) is more negative than the off-diagonal term, with an average difference (diagonal - off diagonal) of -13.2. This implies that if residue *i* is variable, then residue *i*+1 is more likely to be variable than would be expected if substitutions at position *i* and position *i*+1 were independent. Because the regions in a protein most likely to undergo change lie on the surface of the folded structure, and because an amino acid on the surface is more likely to have a neighbor that is also on the surface, these results are not counterintuitive. Again, the exceptions to this rule are most interesting. The strongest exceptions are observed when α and β are Phe and Tyr or Val and Ile. This is, of course, hydrophobic variation, a good indicator of an interior location (\6,8,10,11) at position *i* and *i*+1.

Instances where the Markov assumption of independent substitution at adjacent positions breaks down have become central to methods for predicting the conformation of polypeptide chains from alignments of homologous protein sequences (\6-9). For example, assigning both parses and interior positions from sequence alignments have been used in several *bona fide* predictions of protein structures (those made and published before an experimental structure is available). Recent examples of predictions made using this methods, where subsequently determined crystal structures permit the evaluation of the quality of these predictions, may be found elsewhere (9,12-14). These examples illustrate the immediate value of an empirical and detailed analysis of amino acid substitution outside of the Markov model.

Two comments are appropriate concerning possible future applications of a 400 x 400 dipeptide scoring matrix of the type discussed here. First, the dipeptide substitution matrices might be used to improve the quality of the alignments themselves, as indicated by Jones *et al.* (15). Recently, van Heel suggested that histograms based on the dipeptide content of protein sequences may improve sequence searches (16). In its present form, our matrix is too poorly populated to offer a clear advantage over existing matrices (\3,4,17-22) for simply doing sequence alignments within the standard Markov model. This will, however, soon change as the database expands.

Second, longer range correlations in protein sequences are almost certainly also important in controlling the pattern of divergent evolution. For example, interactions between position *i* and positions *i*+2, *i*+3, and *i*+4 may arise from interresidue interactions within beta strands and alpha helices. These have proven to be less easy to detect by an analysis of the type presented here, presumably because the database used to derive the present dipeptide substitution matrix combines both helical and extended segments, and therefore confuses signals that might arise seperately from these structures. However, should the database size be expanded, we believe that such signals will become apparent.

REFERENCES

1. Dayhoff, M. O., Eck, F. V., and Park, C. M. (1972) in Atlas of Protein Sequence and Structure. (Dayhoff. M. O., ed.), vol. 5, pp. 89-99, National Biomedical Research Foundation Washington, DC.
2. Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978) in Atlas of Protein Sequence and Structure. (Dayhoff. M. O., ed.), vol. 5, suppl. 3, pp. 345-352, National Biomedical Research Foundation Washington, DC.
3. Gonnet, G. H., Cohen, M. A., and Benner, S. A. (1992) Science, 256, 1443-1445.
4. Cohen, M. A., Benner, S. A., and Gonnet, G. H. (1994) J. Mol. Biol. accepted.
5. Gonnet, G. H., and Benner. S. A. (1991) Computational Biochemistry at the ETH, Technical Report 154, Departement Informatik, E.T.H., Zürich.
6. Benner, S. A., Cohen, M. A., and Gonnet, G. H. (1993) J. Mol. Biol. 229, 1065-1082.
7. Bairoch, A., and Boeckmann, B. (1992) Nucl. Acids Res. 20, 2019-2022.
8. Benner, S. A. (1989) Advan. Enzym. Regulat. 28, 219-236.
9. Benner, S. A., and Gerloff, D. (1991) Advan. Enzym. Regulat. 31, 121-181.
10. Benner, S. A. (1992) Curr. Opin. Struct. Biol. 2, 402-412.
11. Benner, S. A., Badcoe, I., Cohen, M. A., and Gerloff, D. L. (1994) *J. Mol. Biol.* in press.

12. Gerloff, D. L., Jenny, T. F., Knecht, L. J., Gonnet, G. H., and Benner, S. A. (1993) FEBS Lett. 318, 118-124.
13. Gerloff, D. L., Jenny, T. F., Knecht, L. J., and Benner, S. A. (1993) Biochem. Biophys. Res. Comm. 194, 560-565.
14. Benner, S. A., Cohen, M. A., and Gerloff, D. L. (1993) J. Mol. Biol. 229, 295-305.
15. Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992) CABIOS. 8, 275-282.
16. van Heel, M. (1991) J. Mol. Biol. 220, 877-887.
17. Taylor, W. R. (1986) J. Theoret. Biol. 119, 205-218.
18. Rao, J. K. M. (1987) Int. J. Peptide Protein Res. 29, 276-281.
19. Kelly, L., and Holladay, L. A. (1987) Protein Engineering 1, 137-140.
20. Risler, J. L., Delorme, M. O., Delacroix, H., and Henaut, A. (1988) J. Mol. Biol. 204, 1019-1029.
21. Altschul, S. R. (1991) J. Mol. Biol. 219, 555-565.
22. Henikoff, S., and Henikoff, J. G. (1992) Proc. Nat. Acad. Sci.USA 89, 10915-10919.