A SIMILARITY RING FOR AMINO ACIDS BASED ON THEIR EVOLUTIONARY SUBSTITUTION RATES

EDWARD ARGYLE

Dominion Radio Astrophysical Observatory, National Research Council Canada, Herzberg Institute of Astrophysics, Penticton, B.C., Canada V2A 6K3

(Received 6 August, 1979; in revised form 26 March, 1980)

Abstract. A similarity ring for amino acids is obtained by reordering the rows and columns of their substitution probability matrix so as to minimize the variance of the matrix elements about the main diagonal.

During evolution the amino acid sequences in typical proteins have changed slowly. By examining several proteins in a family of modern proteins it is possible to estimate ancestral sequences by use of the assumption that evolution proceeded through a minimum number of amino acid substitutions. From a phylogenetic tree based on the results of this procedure it is possible to construct a transition probability matrix of amino acid substitutions. Hasegawa and Yano (1975) have published such a substitution matrix for 12 families of proteins, and the portion of it relevant to this paper is reproduced in Figure 1. (Their matrix included other changes, such as deletions.)

This amino acid substitution matrix, F_{ij}, is characterized by very large probabilities down the main diagonal – testimony to the well known conservatism of evolution. Looking down the first column it is seen that alanine is most frequently substituted by serine, then threonine, and so on. However serine and threonine are not located near alanine in the column – obviously because the amino acids are in alphabetical order. This observation raises the question: Could the amino acids be listed in some kind of similarity order so that their substitution matrix would be more heavily diagonalized?

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	9224	12	64	55	0	13	59	92	0	27	10	24	0	11	89	110	89	0	25	82
ARG	5	9629	21	0	0	51	8	23	48	13	13	59	49	0	10	19	15	33	0	0
ASN	21	0	9258	95	0	0	8	23	79	0	0	41	0	10	10	84	37	0	0	0
ASP	16	0	106	9422	0	0	269	41	16	0	0	12	0	0	30	44	6	0	25	0
CYS	0	0	0	0	9984	0	0	0	0	27	0	0	0	0	0	26	15	0	0	13
GLN	21	48	11	8	0	9517	118	9	79	27	5	47	0	0	20	13	6	0	0	21
GLU	53	12	32	206	0	89	9286	32	0	0	0	47	0	0	30	26	0	0	0	21
GLY	100	24	64	32	0	64	50	9592	0	0	10	24	0	0	20	58	22	0	25	27
HIS	5	24	42	24	0	64	0	5	9666	0	10	18	0	72	10	26	0	33	37	0
ILE	5	36	21	8	0	0	8	5	0	9295	87	6	0	51	20	26	37	33	0	178
LEU	21	12	11	0	0	25	25	5	0	186	9596	6	146	62	69	6	6	0	37	69
LYS	16	108	95	24	0	25	17	18	16	40	5	9556	0	0	10	0	22	0	0	6
MET	0	12	0	0	0	25	0	0	0	40	63	18	9563	10	0	0	6	33	0	55
PHE	11	0	0	0	0	0	0	0	0	13	58	6	49	9547	10	19	6	100	258	13
PRO	58	24	11	0	0	13	0	0	16	27	10	6	0	0	9524	39	6	0	0	0
SER	211	24	170	32	16	25	0	73	16	0	13	36	49	21	69	9245	186	0	25	27
THR	137	0	32	16	0	38	42	23	48	80	5	41	97	21	30	200	9441	0	0	34
TRP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9601	12	0
TYR	11	12	21	8	0	0	8	0	16	13	5	6	0	154	0	19	6	133	9533	13
VAL	53	0	32	16	0	13	50	13	0	212	91	24	49	31	30	19	52	33	12	9477

Fig. 1 Substitution probability matrix with amino acids in alphabetical order. The elements have been multiplied by 10 000.

Origins of Life **10** (1980) 357–360. 0302–1688/80/0104–0357$00.60.

In mathematical terms can the sum of the elements in the kth lesser diagonal

$$D_k = \sum_{j=1}^{20} F_{n+1,j} \quad \text{where} \quad n \equiv (j + k - 1) \bmod 20$$

be made large for small values of k, the distance of the lesser diagonal from the main diagonal?

(Because all lesser diagonals are short it is necessary to continue them out of the bottom of the matrix into the top of the matrix in the next column. This is equivalent to supposing that the amino acids are arranged in a ring so that the last one in the list is adjacent to the first.)

It is natural to require that the standard deviation, σ, of the D's about the main diagonal be minimized, i.e., that

$$\sigma^2 = \sum_{k=-9}^{10} k^2 D_k \Big/ \sum_{k=-9}^{10} D_k, \qquad k \neq 0,$$

be a minimum.

To this end a computer program was written to calculate σ for the substitution matrix and minimize its value by rearranging the amino acids. The ordering of the amino acid list was altered by selecting two amino acids at random and interchanging them. The corresponding pair of rows and pair of columns of the transition matrix were then interchanged, and σ was recalculated. If the new standard deviation was smaller than the old the interchange was allowed to stand. Otherwise the previous ordering of the list and matrix was restored. This procedure was iterated until none of the 190 possible interchanges produced any further reduction of σ. The entire algorithm was run repeatedly, using a different initialization of the random number generator each time. Although there are 19!/2 significant permutations of the amino acid list only one half of all 'solutions' were distinct, and of these, two thirds were repeats of the best solution. Furthermore, ninety percent of the poor solutions were sufficiently similar to the best to serve the purposes of this paper. In other words, good solutions seemed to be highly accessible to the algorithm.

	ASP	ASN	GLY	SER	ALA	THR	PRO	CYS	VAL	ILE	LEU	MET	PHE	TRP	TYR	ARG	HIS	GLN	LYS	GLU
ASP	9422	106	41	44	16	6	30	0	0	0	0	0	0	0	25	0	16	0	12	269
ASN	95	9258	23	84	21	37	10	0	0	0	0	0	10	0	0	0	79	0	41	8
GLY	32	64	9592	58	100	22	20	0	27	0	10	0	0	0	25	24	0	64	24	50
SER	32	170	73	9245	211	186	69	16	27	0	13	49	21	0	25	24	16	25	36	0
ALA	55	64	92	110	9224	89	89	0	82	27	10	0	11	0	25	12	0	13	24	59
THR	16	32	23	200	137	9441	30	0	34	80	5	97	21	0	0	0	48	38	41	42
PRO	0	11	0	39	58	6	9524	0	0	27	10	0	0	0	0	24	16	13	6	0
CYS	0	0	0	26	0	15	0	9984	13	27	0	0	0	0	0	0	0	0	0	0
VAL	16	32	13	19	53	52	30	0	9425	212	91	49	31	33	12	0	0	13	24	50
ILE	8	21	5	26	5	37	20	0	178	9295	87	0	51	33	0	36	0	0	6	8
LEU	0	11	5	6	21	6	69	0	69	186	9596	146	62	0	37	12	0	25	6	25
MET	0	0	0	0	0	6	0	0	55	40	63	9563	10	33	0	12	0	25	18	0
PHE	0	0	0	19	11	6	10	0	13	13	58	49	9547	100	258	0	0	0	6	0
TRP	0	0	0	0	0	0	0	0	0	0	0	0	0	9601	12	0	0	0	0	0
TYR	8	21	0	19	11	6	0	0	13	13	5	0	154	133	9533	12	16	0	6	8
ARG	0	21	23	19	5	15	10	0	0	13	13	49	0	33	0	9629	48	51	59	8
HIS	24	42	5	26	5	0	10	0	0	0	10	0	72	33	37	24	9666	64	18	0
GLN	8	11	9	13	21	6	20	0	21	27	5	0	0	0	0	48	79	9517	47	118
LYS	24	95	18	0	16	22	10	0	6	40	5	0	0	0	0	108	16	25	9556	17
GLU	206	32	32	26	53	0	30	0	21	0	0	0	0	0	0	12	0	89	47	9286

Fig. 2. Substitution probability matrix with amino acids reordered to provide minimum variance about the main diagonal.

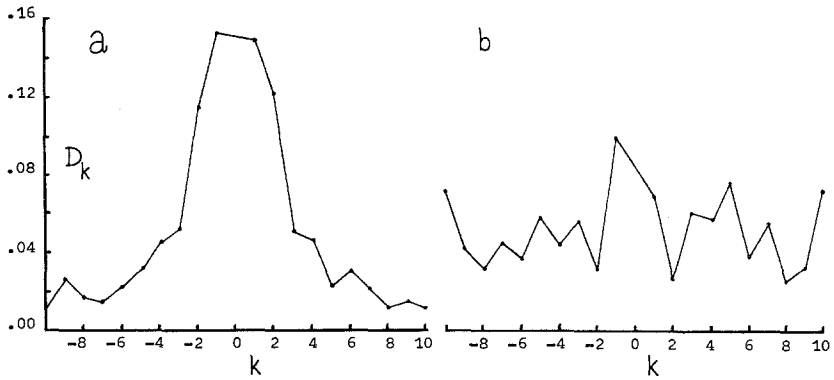

Fig. 3. (a) In the best solution the low order lesser diagonals are very strong. The main diagonal, D_0, is not plotted. (b) In the alphabetic ordering the values of the lesser diagonals are randomly distributed.

The best solution corresponds to the reordered substitution matrix show in Figure 2. Looking again at the alanine column it is seen that serine and threonine are now adjacent to the diagonal element, as desired. However, some columns are less well ordered. For example, the larger elements in the proline column are somewhat scattered; but even here there is considerable clustering towards the diagonal element. An overall view of the improved diagonalization is given in Figure 3, where D_k is plotted against k – for the best solution in (a), and the alphabetic ordering in (b).

The standard deviation for the best solution is $\sigma = 3.98$, which may be compared with 5.68 for the alphabetic ordering and a theoretical 5.94 for random ordering. The minimum possible value for any substitution matrix would be σ=1.00 but this case would require that each amino acid be substituted only by its two nearest neighbors. Clearly, this case is excluded by the data. A data-limited minimum value for σ can be calculated by supposing the elements of each column to be shifted vertically so as to minimize the variance about the diagonal element in that column, independently of all other columns. The result for the actual data in the amino acid substitution matrix would be σ=2.70. Considering all these possibilities it can be said, on the one hand, that the algorithm has significantly reduced σ, and that amino acids in their substitution probabilities must obey at least two independent similarity principles that permit them to be ordered in a similarity ring. On the other hand the failure of σ to fall near its minimum data-limited value of 2.70 implies that still other factors, not yet discovered, also influence the way in which amino acid substitutions occur during evolution.

The similarity principles affecting amino acid substitution can be inferred from Figure 4, which displays the amino acids in cyclic order of minimum variance. Two axes are visible: one ranging from hydrophobic to hydrophilic; the other molecular weight. Nearly 56% of all substitutions for an amino acid involve its nearest or second nearest neighbors. In nearly two thirds of the nearest neighbor substitutions only one nucleotide in the codon need be changed. The four amino acids taken by Crick *et al.* (1976) and

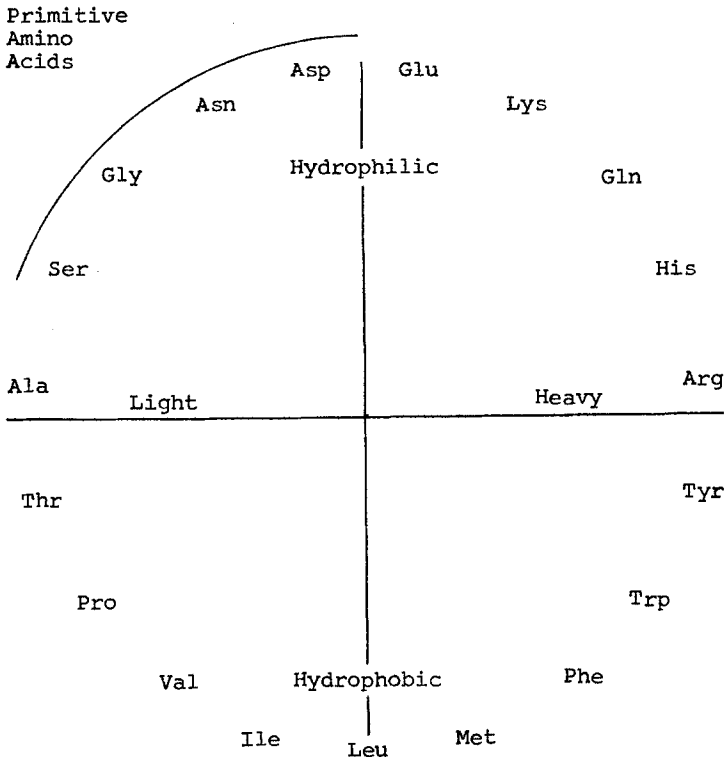

Fig. 4 A similarity ring for amino acids based on a substitution matrix of minimum variance. Cys has been left out of the ring because it fits equally well anywhere between Pro and Arg in the lower portion of the ring.

by Argyle (1977) to be of probable importance to the origin of life are seen together in the second quadrant of the similarity ring. No doubt other regularities are present.

The existence of a similarity ring for amino acids, based on their historical substitution probabilities, testifies to the inherent conservatism of the evolutionary process. The arrangement of amino acids in the ring may have significance for the origin of life and for the origin and evolution of the genetic code.

References

Argyle, E.: 1977, *Origins of Life* 8, 287.
Crick, F. H. C., Brenner, S., Klug, A., and Pieczenik, G.: 1976, *Origins of Life* 7, 389.
Hasegawa, M. and Yano, T.: 1975, *Origins of Life* 6, 219.