Protein Engineering vol.8 no.5 pp.417–423, 1995

Independence divergence-generated binary trees of amino acids

Gábor E.Tusnády, Gábor Tusnády[1] and István Simon[2]

Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, Budapest H-1518, PO Box 7 and [1]Mathematical Institute, Hungarian Academy of Sciences, Budapest H-1364, PO Box 127, Hungary

[2]To whom correspondence should be addressed

The discovery of the relationship between amino acids is important in terms of the replacement ability, as used in protein engineering homology studies, and gaining a better understanding of the roles which various properties of the residues play in the creation of a unique, stable, 3-D protein structure. Amino acid sequences of proteins edited by evolution are anything but random. The measure of non-randomness, i.e. the level of editing, can be characterized by an independence divergence value. This parameter is used to generate binary tree relationships between amino acids. The relationships of residues presented in this paper are based on protein building features and not on the physico-chemical characteristics of amino acids. This approach is not biased by the tautology present in all sequence similarity-based relationship studies. The roles which various physico-chemical characteristics play in the determination of the relationships between amino acids are also discussed.

Key words: amino acid distance matrix/homology studies/ protein design/sequence analysis

Introduction

Proteins are constructed from 20 residues: some residues have similar properties while others are rather different. The 20 residues can be grouped in several ways into overlapping or non-overlapping groups on the basis of their physical and chemical characteristics, such as charge, hydrophobicity, etc. (Grantham, 1974; Miyata *et al.*, 1979; Pongor, 1987; Rao, 1987; Stryer, 1988; Creighton, 1993). A limited number of attempts has also been made to discover the relationships between residues using protein building (biological) features, e.g. their replacement ability in phylogenesis or in site-directed mutagenesis (Fitch and Margoliash, 1967; McLachlan, 1971; Dayhoff *et al.*, 1978; Levin *et al.*, 1986; Risler *et al.*, 1988; Tüdős *et al.*, 1990; Altschul, 1991; Gonnet *et al.*, 1992; Henikoff and Henikoff, 1992, 1993; Jones *et al.*, 1992, 1994; Johnson and Overington, 1993; Cserző *et al.*, 1994). Residues with a high degree of similarity in terms of their chemistry are usually found to have similar biological properties too. However, for the majority of the 190 different pairs of amino acids it is a hard task to derive a natural relationship using their physical or chemical characteristics because we do not know whether hydrophobicity, charge, side-chain volume or even certain secondary structure-forming abilities play the dominant role in the structure of a given protein.

The possibility of evaluating the replacement ability from homology studies or protein engineering is only one side of the coin. It is equally plausible that discovering the natural relationship between amino acids might lead to a deeper understanding of the roles played by various physico-chemical properties of residues in the creation of the unique, stable, 3-D structure of a protein.

Properties of the primary structure of proteins have been studied in our laboratory (Vonderviszt *et al.*, 1986; Cserző and Simon, 1989; Simon and Cserző, 1990). The results were used to predict isomorphic amino acid replacement and various protein structure-related parameters; they were also incorporated into conformational energy calculations of proteins (Vonderviszt and Simon, 1986; Tüdős *et al.*, 1990, 1994; Simon *et al.*, 1991; Fiser *et al.*, 1992; Simon, 1993; Cserző *et al.*, 1994). In these studies the frequency of amino acid pairs was analyzed; a correlation coefficient calculation was the main tool in the statistical analysis.

Currently, the sequence database is large enough to perform analyses at the triplet level. Here, protein sequences are analyzed at this level and we introduce the independence divergence calculation, a statistical method used broadly in various areas of science (Shepp and Vardi, 1982; Csiszár and Tusnády, 1984; Smith and Grandy, 1985) but not to date in protein sequence studies. General mathematical methods useful in protein analysis have been collected by Gindikin (1992), and information about the theoretical approach is given by Yockey and Hubert (1992). A comparison of probability distribution based on the 'entropy term', a close relative of the independence divergence value, is discussed by Press *et al.* (1989).

The divergence value is a measure of the extent of the difference between two probability distributions (Kullback, 1959; Gokhale and Kullback, 1978). If the observed relative frequency (measured value of probability) of the *i*th event is p_i and a certain model would predict a probability q_i, then the divergence value D is:

$$D = \sum_i p_i \cdot \ln(p_i/q_i). \qquad (1)$$

For a perfect model $p_i = q_i$, and therefore $D = 0$.

The aim of this paper is 2-fold. We present binary trees describing a hierarchically structured relationship between amino acids as well as a resulting possible grouping of them. We also demonstrate how independence divergence can be used in protein sequence studies.

Materials and methods

Database

Annotated and classified entries of the Protein Identification Resource (PIR) database (release 34.00, September 1992; Protein Identification Resource, NBRF, Georgetown University, Washington, DC) were used after removing homologous sequences by a filtration procedure similar to that proposed by Hobohm *et al.* (1992). Some modifications have been made to their algorithm to overcome the difficulties caused by the size of the data set. Each protein was represented by a 400 dimension vector. The components of the vector were frequencies of the dipeptides AA,AC,AD... ...YV,YW,YY. Proteins were considered related if the correlation coefficient of their dipeptide compositional vectors exceeded a certain limiting value. In the removal procedure, Algorithm 2 of Hobohm *et al.* (1992) was followed; however, instead of using fixed-limit values, this limit decreased in a stepwise fashion from 100 to 0% in steps of 1%. Proteins removed at higher limiting values were neglected in further steps. Because a very large number of almost identical sequences were removed immediately during the first few steps, the procedure was faster than that described by Hobohm *et al.* (1992). From the stepwise procedure, data sets at 100, 80, 60, 40 and 20% were selected. The 100% level represents the unfiltered database; the 20% level represents an overfiltered database where a large number of unrelated proteins have also been removed, since the average correlation coefficient between the dipeptide composition vector of two randomly selected proteins is close to the 20% level. Table I shows the number of sequences and residues remaining after the various filtration levels.

Independence divergence of random sequences

The short-range non-randomness of amino acid sequences can be measured by calculating the independence divergence value (IDV) at the triplet level in the following way:

$$D_3 = \sum_{i=1}^{20} \sum_{j=1}^{20} \sum_{k=1}^{20} P_{i,j,k} \cdot \ln(P_{i,j,k}/Q_{i,j,k}), \qquad (2)$$

where $P_{i,j,k}$ is the observed frequency of the amino acid triplet i,j,k using a fixed frame, and $Q_{i,j,k}$ is the expected value of $P_{i,j,k}$ assuming perfect randomness in the amino acid sequences. The frame means two specified distances x and y between first and second, and first and third elements of the triplets, respectively. For example, when $x = 2$ and $y = 5$, blocks of six subsequent residues are considered, and i represents the first, j represents the third $(3 = 1 + 2)$, and k represents the sixth $(6 = 1 + 5)$ element of the block.

In one approach, $Q_{i,j,k}$ is the product of the observed relative frequencies of the residues in the corresponding positions of the triplet (P_i, P_j and P_k) and the total number of residues (NR):

$$Q^{(1)}_{i,j,k} = NR \cdot P_i \cdot P_j \cdot P_k. \qquad (3)$$

[Note that this is slightly different from the product of the relative frequencies of the three residues in the whole database. For instance, many sequences start with the residue methionine, the amino acid coded by the start codon. Therefore in triplets the relative frequency of Met is larger in the first position than in the second and third. Methionine is not the only residue that exhibits such an uneven distribution along the amino acid sequence, but its frequent appearance at the N-terminus has by far the largest effect on the statistical calculations (Cserző and Simon, 1989). This is also the reason why Monte Carlo mixing of the sequences was not used, as employed in most related studies.]

Table I. The effect of filtration on the size of the database

Filtration level (FL)	Number of sequences (NS)	Number of residues (NR)
100%	10 550	3 591 370
80%	7553	2 439 205
60%	5597	1 617 370
40%	2809	548 912
20%	357	28 709

Our earlier study showed that the short-range non-randomness did not, in general, extend for 10 residues in the sequences (Cserző and Simon, 1989). Consequently, another reference value can be calculated from the average frequency of residues i, j and k when they are separated by at least 20 residues in the sequence:

$$Q^{(2)}_{i,j,k} = (1/100) \sum_{x=21}^{30} \sum_{y=51}^{60} P_{i,j,k;x,y}, \qquad (4)$$

where $P_{i,j,k;x,y}$ is the frequency of the triplet in which j and k residues are separated from the i residue by x and y elements in the sequence. The advantage of this approach is that it helps to separate short-range regularities from a possible long-range regularity on which the short-range one is superimposed. [Strictly speaking, only the divergence between $P_{i,j,k}$ and $Q^{(1)}_{i,j,k}$ measures the independence of residues. The second case, $P_{i,j,k}$ versus $Q^{(2)}_{i,j,k}$, compares deviation from independence in close and distant elements.]

D_3 values were calculated not only on the subsequent three residues (triplet segment), but also on all possible triplets within a sequential interval up to 10 residues using both $Q_{i,j,k}$ values. Let $D_{3,3}$ denote the D_3 value calculated according to Equation 2 for subsequent residues. For other triplets, $D_{3,S}$ is the mean of the D_3 values calculated for all triplets where there is less than S residues between the first and the third residues of the triplets. For example, $D_{3,10}$ is the mean of 36 individually calculated D_3 values for triplets where the relative positions of the residues are: 1,2,3; 1,2,4; ...; 1,2,10; 1,3,4; ...; ...; 1,9,10.

Relationships derived from independence divergence calculations

The calculated IDVs showed a significant deviation from independence. This means that the 20 amino acids as a whole are not independent. It may be that some of them are independent. One way to find independent pairs is the unification or amalgamation of pairs.

Any two amino acids can be unified into one pair and an IDV can be recalculated for 19 new units, i.e. 18 of the original residues and one artificial element formed by the unified pair. If the distinction between any two of the residues is exactly random, then their unification will not result in a significant effect on the IDV. If there is no independent pair, then the unification of the most similar two residues into one element will cause the smallest decrease in the IDV. There are 190 ways to unify two different amino acids into one element. The largest calculated IDV represents the smallest decrease relative to that calculated from all 20 elements. Consequently, it pinpoints the most similar pair out of the 20 elements. We unified this pair first. After this first step the procedure was repeated with the 19 remaining residues (18 individual amino acids and one unified pair); the process was repeated until only one unified element remained. At each step the pair of elements (individual amino acid or group of amino acids) leading to the maximum IDV was unified. This procedure resulted in a binary tree that represents a sequence of groups of amino acids; it also shows a hierarchical relationship between individual residues. When the three residues were not adjacent but came from a segment of four to 10 residues, IDVs were calculated for all possible triplets of the given segment length. The maximum IDV pinpointed the pair with the most similar elements.

Advantage of binary trees

One may ask for the best grouping of amino acids into h groups, where h is any given number between 1 and 20. Combinatorial calculations show that n different elements can be grouped into h (non-empty) groups in k different ways, where:

$$k = \left[\sum_{i=0}^{h-1} (-1)^i \left[\begin{matrix} h \\ i \end{matrix} \right] (h-i)^n \right] \Big/ h!\,. \tag{5}$$

For example, when $n = 20$ (the number of amino acids), k is $>0.5\times10^6$ for $h = 2$ and k is $\sim4.5\times10^{10}$ for $h = 4$. Therefore it is practically impossible to check all possible groupings. Our partial optimization procedure is based on the assumption that from the best h group the best $(h - 1)$ group can be obtained by integrating two of the h groups into one and leaving the other $(h - 2)$ groups unchanged. Making several thousand random unifications of the elements, a Monte Carlo simulation demonstrated that partial optimization is a reasonable approach because IDVs resulting from partial optimization are on the top of the distribution of IDVs generated by random groupings. Partial optimization is computationally manageable because at each level of the tree it involves only $[h(h - 1)]/2$ different unifications, where h is the number of elements. (Nevertheless, considering the five levels of filtration, the two kinds of reference $Q_{i,j,k}$ values and the eight different segment lengths representing one to 36 different triplets combinations, altogether $\sim10^5$ IDVs were calculated.)

One of the advantages of this binary tree representation of amino acid relationships is that it is not based on the comparison of homologous sequences. Risler *et al.* (1988) pointed out that there is a certain level of tautology in those calculations: the homologous sequences are defined on the basis of sequence similarities (not only on the basis of the same residues but also on the similar ones) and then similarities among the residues are derived from this data set. Our procedure, similar to those we have published previously (Tüdős *et al.*, 1990; Cserző *et al.*, 1994), is not biased by such tautology.

Finally, some disadvantages of the procedure must also be mentioned. At every level of the tree, two groups (or single residues) are unified according the highest IDV, regardless of whether the second highest value, etc., is much lower or almost the same. To overcome such errors we considered binary trees generated by slightly different IDVs. One may argue that there is a 'tie-up sale' in the lower level of the tree because the best candidates for pairing to a certain residue or residue groups might already be grouped with other residues and they are therefore not available as a single residue any longer. However, our Monte Carlo simulation (mentioned above) indicates that this 'tie-up sale' effect is present only in trees which are far from the optimal ones.

One may also criticize the method leading to binary trees by saying that triplets cannot accumulate all the stochastic relationships existing in proteins. A possible extension of our method to larger blocks is based on the random sets formed by residues representative in the investigated block. Small-sized sets occur significantly much more often in proteins than one would expect assuming independence.

The relationship of individual amino acids given by the trees can be converted into an amino acid distance matrix in several ways. In one of the simplest cases the (i,j) element of the distance matrix is defined as:

$$M(i,j) = \mathrm{A} + \mathrm{B} - 2\mathrm{C} \tag{6}$$

if residues i and j appear in the tree at levels A and B respectively (branch tops) and they become part of a unified element at level C (the closest junction). These distance matrices can be compared with other similarity or distance matrices from the literature by calculating correlation coefficients. For this comparison all matrices were prepared according to Johnson and Overington (1993).

One obvious disadvantage of this very simple transformation is that it does not distinguish between unifications taking place near the top or near the bottom of the tree. Nevertheless, despite the scaling uncertainty due to the limited optimization during tree building and the simplicity of the transformation, these matrices give reasonable correlations with other amino acid distance and similarity matrices and thus can be used in homology and protein engineering studies.

Results and discussion

Binary trees obtained after various levels of data filtration and for various segment lengths have been created.

Figure 1 shows the binary trees calculated for decapeptide segments at the 100, 60 and 20% filtration levels, and those calculated for three-, six- and nine-membered oligopeptide segments at the 60% filtration level using $Q^{(1)}_{i,j,k}$ as a reference. Figure 2 shows the binary trees calculated for heptapeptide segments using both $Q^{(1)}_{i,j,k}$ and $Q^{(2)}_{i,j,k}$ references at the 60% filtration level. These were selected as representative trees because the matrices calculated from them gave the highest

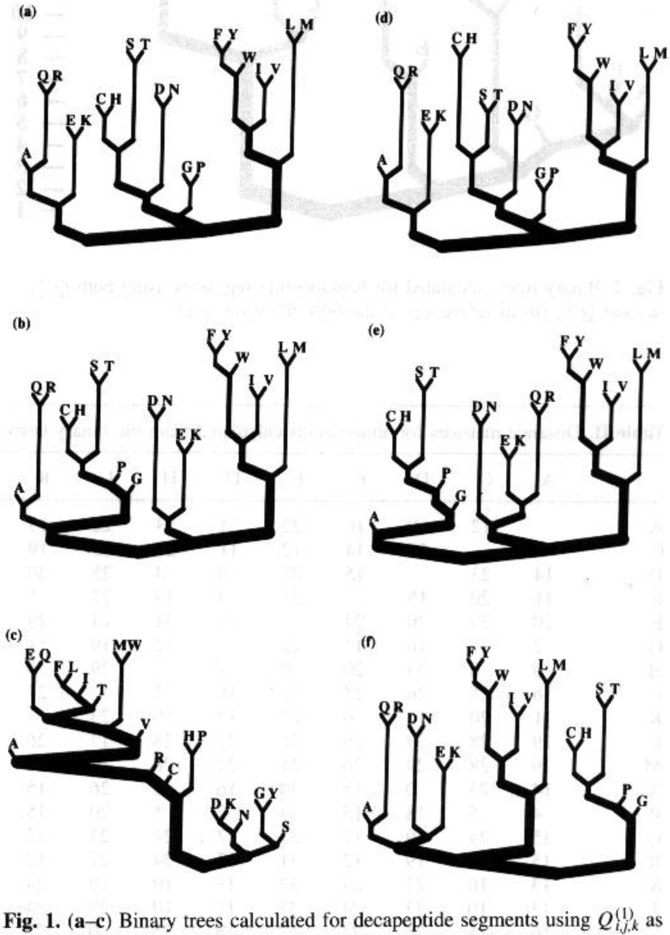

Fig. 1. (**a–c**) Binary trees calculated for decapeptide segments using $Q^{(1)}_{i,j,k}$ as the reference at the 100, 60 and 20% filtration levels. (**d–f**) Binary trees calculated for three-, six- and nine-membered oligopeptide segments at the 60% filtration level using $Q^{(1)}_{i,j,k}$ as the reference.

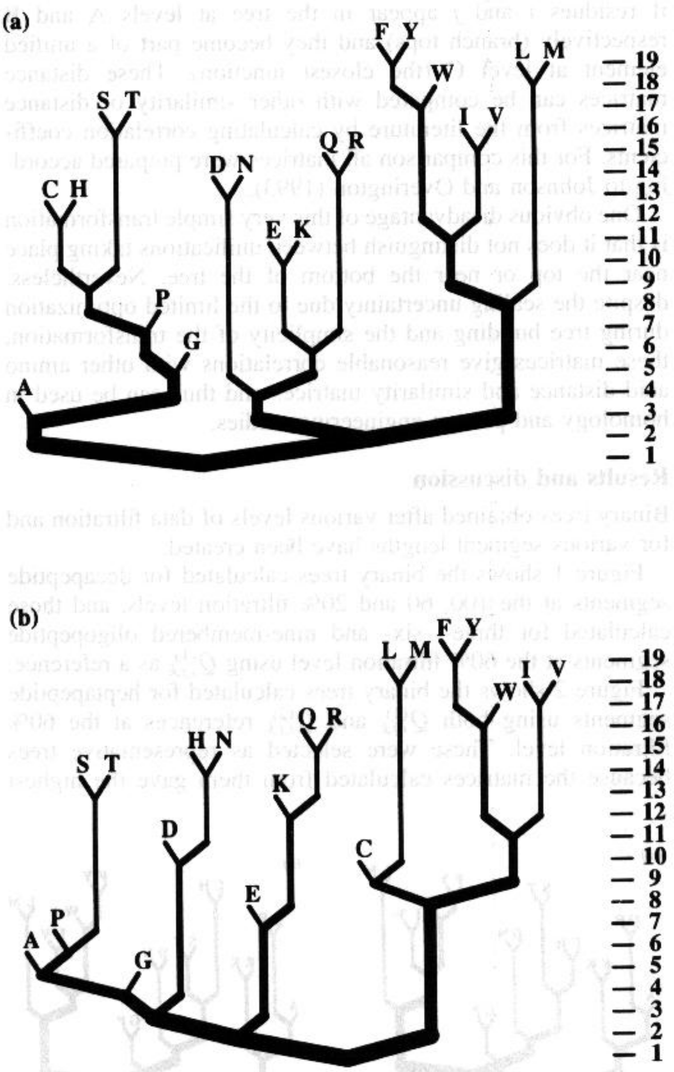

Fig. 2. Binary trees calculated for heptapeptide segments using both $Q^{(1)}_{i,j,k}$ (a) and $Q^{(2)}_{i,j,k}$ (b) as references at the 60% filtration level.

average correlation coefficients with the other matrices obtained from trees representing various segment sizes and various levels of filtration. Table II shows the two distance matrices calculated from the trees of Figure 2 by Equation 6. We would like to emphasize that the trees represent the primary information and Equation 6 is only one of the many ways of transforming trees into matrices.

It can be seen that the lack of filtration (100% level) results in rather similar trees to those obtained after 60% filtration, while the trees generated after 20% filtration (keeping only a fraction of the original database) differ significantly from them (Figure 1). Thus, homologous proteins presented in such a large database do not influence the results significantly. This discovery has great practical importance because, as we discussed earlier (Vondervizst *et al.*, 1986; Cserzö and Simon, 1989), there is no reliable way to obtain a perfectly filtered database. Even if only one protein remains from the whole database, it might contain repetitive sequence elements due to gene duplication. However, to avoid any bias, the results from the unfiltered database will not be discussed further here. Therefore we shall discuss the results obtained after the 80, 60 and 40% levels of filtration.

The binary trees in Figure 1 also show that the results obtained from triplets in which the residues are separated by a few residues are rather similar, regardless of the lengths of the segment (the sequential distance between the first and the third residues); however, they differ slightly from the results obtained using adjacent residues (tripeptides). This justifies our decision to use statistics not only on tripeptide segments.

To understand the roles that various physico-chemical characteristics play in the determination of the relationships between the amino acids, all 48 trees were analyzed. A decomposition of these trees is shown in Table III. One obvious result that can be seen in all 48 trees obtained from the eight oligopeptides with different lengths after the 80, 60 and 40% levels of filtration and using $Q^{(1)}_{i,j,k}$ and $Q^{(2)}_{i,j,k}$ as references, was that seven hydrophobic residues (F, I, L, M, V, W and Y) were in one group in 47 of the cases. When $Q^{(1)}_{i,j,k}$ was used as a reference basis, only one out of the 24 cases showed different patterns; in 23 of the cases the residues mentioned above

Table II. Distance matrices for amino acids calculated from the binary trees of Figure 2a (lower left of table) and b (upper right of table)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A		12	9	10	22	1	13	20	15	21	21	13	1	18	18	8	8	20	19	22
C	9		17	14	12	11	21	10	19	9	9	21	13	22	22	20	20	10	9	12
D	14	23		15	27	8	4	25	20	26	26	4	10	23	23	17	17	25	24	27
E	11	20	15		24	9	19	22	5	23	23	19	11	8	8	18	18	22	21	24
F	20	29	30	27		21	31	14	29	21	21	31	23	32	32	30	30	14	3	0
G	2	7	16	13	22		12	19	14	20	20	12	2	17	17	9	9	19	18	21
H	9	0	23	20	29	7		29	24	30	30	0	14	27	27	21	21	29	28	31
I	16	25	26	23	12	18	25		27	19	19	29	21	30	30	28	28	0	11	14
K	11	20	15	0	27	13	20	23		28	28	24	16	3	3	23	23	27	26	29
L	19	28	29	26	21	21	28	17	26		0	30	22	31	31	29	29	19	18	21
M	19	28	29	26	21	21	28	17	26	0		30	22	31	31	29	29	19	18	21
N	14	23	0	15	30	16	23	26	15	29	29		14	27	27	21	21	29	28	31
P	4	5	18	15	24	2	5	20	15	23	23	18		19	19	7	7	21	20	23
Q	15	24	19	12	31	17	24	27	12	30	30	19	19		0	26	26	30	29	32
R	15	24	19	12	31	17	24	27	12	30	30	19	19	0		26	26	30	29	32
S	13	10	27	24	33	11	10	29	24	32	32	27	9	28	28		0	28	27	30
T	13	10	27	24	33	11	10	29	24	32	32	27	9	28	28	0		28	27	30
V	16	25	26	23	12	18	25	0	23	17	17	26	20	27	27	29	29		11	14
W	18	27	28	25	2	20	27	10	25	19	19	28	22	29	29	31	31	10		3
Y	20	29	30	27	0	22	29	12	27	21	21	30	24	31	31	33	33	12	2	

Table III. Decomposition of binary trees obtained after various levels of filtration (FL), various lengths of oligopeptide (LO) using two kinds of reference, $Q^{(1)}$ and $Q^{(2)}$

	LO	FL	Tree decomposition
$Q^{(1)}$	3	40	(((A,(Q,R)),(E,K)),(((((C,H),(S,T)),(D,N)),(G,P)),((((F,Y),W),(I,V)),(L,M))))
$Q^{(1)}$	4	40	(((A,(Q,R)),(E,K)),(((((C,H),(S,T)),(D,N)),(G,P)),((((F,Y),W),(I,V)),(L,M))))
$Q^{(1)}$	5	40	((((A,(Q,R)),((D,N),(E,K))),((((F,Y),W),(I,V)),(L,M))),(((C,H),(S,T)),(G,P)))
$Q^{(1)}$	6	40	((((A,(Q,R)),((D,N),(E,K))),((((F,Y),W),(I,V)),(L,M))),(((C,H),(S,T)),(G,P)))
$Q^{(1)}$	7	40	((((A,(Q,R)),((D,N),(E,K))),((((F,Y),W),(I,V)),(L,M))),(((C,H),(S,T)),(G,P)))
$Q^{(1)}$	8	40	(((A,(Q,R)),(E,K)),((((C,G),P),((D,N),(H,(S,T)))),((((F,Y),W),(I,V)),(L,M))))
$Q^{(1)}$	9	40	(((A,(Q,R)),(E,K)),((((C,G),P),((D,N),(H,(S,T)))),((((F,Y),W),(I,V)),(L,M))))
$Q^{(1)}$	10	40	(((A,(Q,R)),(E,K)),((((C,G),P),((D,N),(H,(S,T)))),((((F,Y),W),(I,V)),(L,M))))
$Q^{(1)}$	3	60	(((A,(Q,R)),(E,K)),(((((C,H),(S,T)),(D,N)),(G,P)),((((F,Y),W),(I,V)),(L,M))))
$Q^{(1)}$	4	60	((((A,(Q,R)),((D,N),(E,K))),((((F,Y),W),(I,V)),(L,M))),(((C,H),(S,T)),(G,P)))
$Q^{(1)}$	5	60	((((A,(Q,R)),((D,N),(E,K))),((((F,Y),W),(I,V)),(L,M))),(((C,H),(S,T)),(G,P)))
$Q^{(1)}$	6	60	((A,((((C,H),(S,T)),P),G)),(((D,N),((E,K),(Q,R))),((((F,Y),W),(I,V)),(L,M))))
$Q^{(1)}$	**7**	**60**	**((A,((((C,H),(S,T)),P),G)),(((D,N),((E,K),(Q,R))),((((F,Y),W),(I,V)),(L,M))))**
$Q^{(1)}$	8	60	((A,((((C,H),(S,T)),P),G)),(((D,N),((E,K),(Q,R))),((((F,Y),W),(I,V)),(L,M))))
$Q^{(1)}$	9	60	((((A,(Q,R)),((D,N),(E,K))),((((F,Y),W),(I,V)),(L,M))),((((C,H),(S,T)),P),G))
$Q^{(1)}$	10	60	(((A,(Q,R)),((((C,H),(S,T)),P),G)),(((D,N),(E,K)),((((F,Y),W),(I,V)),(L,M))))
$Q^{(1)}$	3	80	(((A,((E,K),(Q,R))),((I,V),(L,M))),((((C,H),((F,Y),W)),(D,N)),(G,(P,(S,T)))))
$Q^{(1)}$	4	80	((((A,(Q,R)),((D,N),(E,K))),((((F,Y),W),(I,V)),(L,M))),(((C,H),(S,T)),(G,P)))
$Q^{(1)}$	5	80	((A,(((C,H),(S,T)),(G,P))),(((D,N),((E,K),(Q,R))),((((F,Y),W),(I,V)),(L,M))))
$Q^{(1)}$	6	80	((A,((((C,H),(S,T)),P),G)),(((D,N),((E,K),(Q,R))),((((F,Y),W),(I,V)),(L,M))))
$Q^{(1)}$	7	80	((A,((((C,H),(S,T)),P),G)),(((D,N),((E,K),(Q,R))),(((((F,Y),W),I),V),(L,M))))
$Q^{(1)}$	8	80	(((A,G),(((C,H),(S,T)),P)),(((D,N),((E,K),(Q,R))),(((((F,Y),W),I),V),(L,M))))
$Q^{(1)}$	9	80	(((A,G),(((C,H),(S,T)),P)),(((D,N),((E,K),(Q,R))),(((((F,Y),W),I),V),(L,M))))
$Q^{(1)}$	10	80	(((A,(Q,R)),((((C,H),(S,T)),P),G)),(((D,N),(E,K)),(((((F,Y),W),I),V),(L,M))))
$Q^{(2)}$	3	40	((((A,(K,(Q,R))),E),((((C,S),T),(D,(H,N))),(G,P))),((((F,Y),V),(I,W)),(L,M)))
$Q^{(2)}$	4	40	((((A,(K,(Q,R))),E),(((C,(S,T)),(D,(H,N))),(G,P))),((((F,Y),W),(I,(M,V))),L))
$Q^{(2)}$	5	40	(((A,(E,(K,(Q,R)))),((((C,H),(S,T)),(D,N)),(G,P))),((((F,Y),W),(I,V)),(L,M)))
$Q^{(2)}$	6	40	((((A,((C,H),(S,T))),(G,P)),((D,N),(E,(K,(Q,R))))),((((F,Y),W),(I,V)),(L,M)))
$Q^{(2)}$	7	40	((((A,((C,H),(S,T))),(G,P)),((D,N),(E,(K,(Q,R))))),((((F,Y),W),(I,V)),(L,M)))
$Q^{(2)}$	8	40	((((A,((C,H),(S,T))),(G,P)),((D,N),(E,(K,(Q,R))))),((((F,Y),W),(I,V)),(L,M)))
$Q^{(2)}$	9	40	((((A,(P,(S,T))),(C,G)),((D,(H,N)),(E,(K,(Q,R))))),((((F,Y),W),(I,V)),(L,M)))
$Q^{(2)}$	10	40	((((A,((C,H),(S,T))),(G,P)),((D,N),(E,(K,(Q,R))))),((((F,Y),W),(I,V)),(L,M)))
$Q^{(2)}$	3	60	((((A,(K,(Q,R))),E),(((((C,H),N),(S,T)),D),(G,P))),((((F,Y),W),(I,V)),(L,M)))
$Q^{(2)}$	4	60	(((A,(E,(K,(Q,R)))),((((C,(H,N)),D),(S,T)),(G,P))),((((F,Y),W),(I,V)),(L,M)))
$Q^{(2)}$	5	60	(((((A,(C,G)),(P,(S,T))),(D,(H,N))),(E,(K,(Q,R)))),((((F,Y),W),(I,V)),(L,M)))
$Q^{(2)}$	6	60	(((A,(E,(K,(Q,R)))),(((D,(H,N)),G),(P,(S,T)))),((C,(L,M)),(((F,Y),W),(I,V))))
$Q^{(2)}$	**7**	**60**	**(((((A,(P,(S,T))),G),(D,(H,N))),(E,(K,(Q,R)))),((C,(L,M)),(((F,Y),W),(I,V))))**
$Q^{(2)}$	8	60	(((((A,(P,(S,T))),G),(D,(H,N))),(E,(K,(Q,R)))),((C,(L,M)),(((F,Y),W),(I,V))))
$Q^{(2)}$	9	60	(((((A,(P,(S,T))),G),(D,(H,N))),(E,(K,(Q,R)))),((C,(L,M)),(((F,Y),W),(I,V))))
$Q^{(2)}$	10	60	(((((A,(P,(S,T))),G),(D,(H,N))),(E,(K,(Q,R)))),((C,(L,M)),(((F,Y),W),(I,V))))
$Q^{(2)}$	3	80	(((A,(E,(K,(Q,R)))),(((D,(H,N)),(S,T)),(G,P))),((C,((F,Y),W)),((I,V),(L,M))))
$Q^{(2)}$	4	80	(((A,(E,(K,(Q,R)))),((((C,(H,N)),D),(S,T)),(G,P))),((((F,Y),W),(I,V)),(L,M)))
$Q^{(2)}$	5	80	((((A,(S,T)),((C,G),P)),((D,(H,N)),(E,((K,R),Q)))),((((F,Y),W),(I,V)),(L,M)))
$Q^{(2)}$	6	80	(((A,(E,((K,R),Q))),((D,(H,N)),(G,(P,(S,T))))),((C,(((F,Y),W),(I,V))),(L,M)))
$Q^{(2)}$	7	80	((((A,(P,(S,T))),G),((D,(H,N)),(E,((K,R),Q)))),((C,(((F,Y),W),(I,V))),(L,M)))
$Q^{(2)}$	8	80	((((A,(P,(S,T))),G),((D,(H,N)),(E,((K,R),Q)))),((C,(((F,Y),W),(I,V))),(L,M)))
$Q^{(2)}$	9	80	((((A,(P,(S,T))),G),((D,(H,N)),(E,((K,R),Q)))),((C,(((F,Y),W),(I,V))),(L,M)))
$Q^{(2)}$	10	80	((((A,(P,(S,T))),G),((D,(H,N)),(E,((K,R),Q)))),((C,(((F,Y),W),(I,V))),(L,M)))

The decomposition of binary trees from Figure 2 is printed in bold.

formed a separate group. When $Q^{(2)}_{i,j,k}$ was used as the reference, the seven hydrophobic residues were together in all 24 cases, but in 11 of these cases they were also accompanied by Cys. Indeed cysteine is a very special residue. In proteins it appears partly as a rather polar cysteine and partly as an apolar half cystine. Further, because –SH is the most reactive group, a high percentage of cysteine appears in the active center of the protein and it can bind, like His, a lot of ligands and metal ions, e.g. the zinc finger protein family. Our sequence-level analyses cannot distinguish between the large variety of forms, and this may cause uncertainty in the position of the Cys residue in the binary trees. It is also worth mentioning here that Y is always in the hydrophobic branch of the trees, and in all 48 cases F and W are its closest relative. This suggests that from the point of view of editing the primary structures, the aromatic and not the weak polar character of Y plays a significant role. Another interesting feature is that Ala, Gly and Pro never join with the other hydrophobic residues. These amino acids usually appear in trees at very low levels, indicating that they are at a large distance from most of the other residues. Note that their appearance at such a low level is due to the differences of these residues from the other large groups of residues rather than their similarities with each other.

S and T, as well as L and M, appear 47 times as closest relatives. They are reported as the best replacing pairs in the literature. The next most common pair is Q and R, which can be seen together a total of 42 times. R is charged, but because it appears very often on the surface of the protein and its charge is neutralized by counter ions, this pairing is not strange. However, the pairs K and R, E and Q, and N and Q would be less surprising.

The pair E, K appeared together in all cases when the

Table IV. Correlation coefficients between the two distance matrices (see Table II) in bold and any pairs of other similarity and distance matrices

Reference[a]	$Q^{(1)}$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	**63**																		
2	**75**	73																	
3	**63**	97	72																
4	**56**	79	68	73															
5	**66**	88	81	92	70														
6	**61**	76	70	77	57	75													
7	**66**	88	83	87	78	94	75												
8	**64**	88	77	87	75	90	81	92											
9	**67**	89	78	88	76	91	82	93	98										
10	**63**	88	78	90	76	92	69	88	83	84									
11	**63**	77	72	75	72	77	59	81	74	75	82								
12	**77**	79	83	77	71	83	69	86	82	83	78	75							
13	**68**	87	81	85	74	89	81	89	90	91	84	74	83						
14	**66**	78	83	78	58	83	83	79	76	77	77	68	78	79					
15	**65**	64	87	63	56	73	66	74	69	71	70	68	75	71	75				
16	**61**	65	81	63	63	75	59	80	76	76	70	69	81	76	69	78			
17	**31**	66	40	61	54	55	61	61	67	66	50	46	50	59	48	35	41		
18	**46**	39	73	41	23	52	49	44	42	42	49	45	49	49	63	67	49	15	
$Q^{(2)}$	**86**	**60**	**78**	**59**	**57**	**66**	**50**	**67**	**63**	**64**	**63**	**57**	**75**	**65**	**65**	**69**	**62**	**25**	**51**

[a] References are as follows: 1, Altschul (1991); 2, Cserző *et al.* (1994); 3, Dayhoff *et al.* (1978); 4, Fitch and Margoliash (1967); 5, Gonnet *et al.* (1992); 6, Grantham (1974); 7, Henikoff and Henikoff (1992); 8, Henikoff and Henikoff (1993); 9, Johnson and Overington (1993); 10, Jones *et al.* (1992); 11, Jones *et al.* (1994); 12, Levin *et al.* (1986); 13, McLachlan (1971); 14, Miyata *et al.* (1979); 15, Pongor (1987); 16, Rao (1987); 17, Risler *et al.* (1988); 18, Tüdős *et al.* (1990).

product of the frequencies of the single residues [$Q^{(1)}_{i,j,k}$] was used as a reference but in none of the 24 cases when $Q^{(2)}_{i,j,k}$ was used as reference; however, they remained close to each other in the latter case. As we have mentioned, charges are usually neutralized by counter ions when they appear on the surface of the protein. There is a similar but less significant asymmetry in the case of the C, H pair. These residues appeared next to each other in 21 out of the 24 cases when $Q^{(1)}_{i,j,k}$ was used as the reference, and only six times in the other 24 cases when $Q^{(2)}_{i,j,k}$ was used. Note that these are the cases where C is joined to the hydrophobic residues 11 times. The C, H pair is a very complicated case, not just because there are two covalent forms of Cys but also because the p*K* of His is so close to neutral that both charged and neutral His are common in the proteins.

In general, while the hydrophobic residues (except A, G and P) are separated from the polar and charged residues, which are rather mixed up in the various trees, indicating that charge is less important in the editing of the amino acid sequences than hydrophobicity.

Table IV compares the two distance matrices calculated from the binary trees in Figure 2 by Equation 6 and some other matrices from the literature. This matrix shows that there are significant differences between any pairs of distance (or similarity) matrices. Some high correlation coefficient values are the consequence of using very similar methods and/or databases. Just as various secondary structure predictions produce different quality results in different proteins, one should expect that certain distance matrices used in alignment or protein engineering studies may work better in one case, while a second matrix may give a better result in another case. In general it is hard to rank the matrices. From a simple statistical point of view, those matrices based on the analysis of larger databases are usually better.

Table II shows two distance matrices. They were calculated from the same database using similar methods. Both matrices were generated by calculating IDVs, with the only difference being two different 'random' residue triplet distributions as a reference. Therefore it is not surprising that the two matrices $Q^{(1)}$ and $Q^{(2)}$ give the highest correlation coefficient value with each other. $Q^{(1)}$ and $Q^{(2)}$ exhibit the highest similarity with the matrix of Cserző *et al.* (1994), 75 and 78% respectively, which was calculated from the non-random pairing of amino acids in a data set similar to that used in this study.

It is worth mentioning that a matrix calculated by the same method as used by Cserző *et al.* (1994) but on a smaller data set (Tüdős *et al.*, 1990) gives one of the lowest correlation coefficient values with $Q^{(1)}$ and $Q^{(2)}$ (46 and 51% respectively), indicating the importance of an appropriate database (see Table IV).

This is one of the reasons why we analyzed the large PIR database instead of the Protein Data Bank [which contains much more detailed information but on a relatively small and specialized subset of proteins, based mainly on small crystallizable (water-soluble) proteins]. It has been discussed previously that the results of statistical analyses may not be applied to members of a larger data set (Simon and Cserző, 1990). It was demonstrated recently that only a similarity matrix based exclusively on the analyses of a protein sequence database can be used to predict transmembrane helices in membrane-bound proteins (Cserző *et al.*, 1994).

In this work, two rather similar distance matrices have been suggested to add to the list of several similar matrices in the literature. The question as to which matrix should be used in a homology or protein engineering study cannot be answered *a priori*. In practice, several approaches should be applied and the result provided by most of the unrelated approaches has the highest probability of being valid. There are a large number of distance matrices in the literature, but the number of methods by which these matrices were calculated are rather limited. From this viewpoint it is important to note that the two matrices in Table II were generated in a completely new way.

Polypeptides with random amino acid sequences do not generally fold into a unique, stable conformation; proteins have edited amino acid sequences. If the level of editing, i.e. the measure of non-randomness, is an important feature of the amino acid sequence of a protein, one can expect that those amino acid replacements found to be acceptable to the protein cause little alteration in the level of editing. Our study argues in favor of this expectation. Our data presented in Figure 2 and Table II may be used in homology studies, protein engineering and all other cases where the grouping of amino acids is necessary. We also wish to demonstrate the usefulness of the independence divergence calculation in sequence analysis and call the reader's attention to this method.

Acknowledgements

We thank Dr G.J.Barton and A.Fiser for their critical comments on this manuscript. This work was supported by research grants OTKA 1361 and T012890. This research was sponsored in part by the US–Hungarian Science and Technology Fund in collaboration with the Cornell University, Ithaca, NY, and the Institute of Enzymology, Budapest, Hungary under project J.F. no. 332, by the Hungarian–British Science and Technology Program No. 2/1993, as well as by the Hungarian–Israel Science and Technology Program No. 1/1994. Additional support from the International Centre for Genetic Engineering and Biotechnology (ICGEB, Trieste, Italy), UNIDO contract no. 93/232 (CRP/HUN 93-03) is also acknowledged. Part of the work was performed under I.S.'s Associate Membership at the International Centre for Theoretical Physics (ICTP, Trieste, Italy). G.E.T. is also grateful for support from the Foundation for Hungarian Science.

References

Altschul,S.F. (1991) *J. Mol. Biol.*, **219**, 555–565.
Creighton,T.E. (1993) *Proteins: Structure and Molecular Principles.* 2nd edition, Freeman, New York.
Cserző,M. and Simon,I. (1989) *Int. J. Peptide Protein Res.*, **34**, 184–195.
Cserző,M., Bernassau,J.-M., Simon,I. and Maigret,B. (1994) *J. Mol. Biol.*, **243**, 388–396.
Csiszár,I. and Tusnády,G. (1984) *Statist. Decis.*, **1**, 205–237.
Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC, Vol. 5. Suppl. 3., pp. 345–358.
Fiser,A., Cserző,M., Tüdős,É. and Simon,I. (1992) *FEBS Lett.*, **302**, 117–120.
Fitch,W.M. and Margoliash,E. (1967) *Science*, **155**, 279–284.
Gindikin,S. (1992) *DIMACS Series in Discrete Mathematics and Theoretical Computer Science.* Volume 8. American Mathematical Society, Providence, RI.
Gokhale,D.V. and Kullback,S. (1978) *The Information in Contingency Tables.* Marcel Dekker Inc., New York.
Gonnet,G.H., Cohen,M.A. and Benner,S.A. (1992) *Science*, **256**, 1443–1445.
Grantham,R. (1974) *Science*, **185**, 862–864.
Henikoff,S. and Henikoff,J.G. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
Henikoff,S. and Henikoff,J.G. (1993) *Proteins: Struct. Funct. Genet.*, **17**, 49–61.
Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) *Protein Sci.*, **1**, 409–417.
Johnson,M.S. and Overington,J.P. (1993) *J. Mol. Biol.*, **233**, 716–738.
Jones,D.T., Taylor,W.R. and Thorton,J.M. (1992) *CABIOS*, **8**, 275–282.
Jones,D.T., Taylor,W.R. and Thorton,J.M. (1994) *FEBS Lett.*, **339**, 269–275.
Kullback,S. (1959) *Information Theory and Statistics.* Wiley, New York.
Levin,J.M., Robson,B. and Garnier,J. (1986) *FEBS Lett.*, **205**, 303–308.
McLachlan,A.D. (1971) *J. Mol. Biol.*, **61**, 409–424.
Miyata,T., Miyazawa,S. and Yasunaga,T. (1979) *J. Mol. Evol.*, **12**, 219–236.
Pongor,S. (1987) *Methods Enzymol.*, **154**, 450–473.
Press,W.H., Flannery,B.P., Teukolsky,S.A. and Vetterling,W.T. (1989) In *Numerical Recipes*. Cambridge University Press, New York, pp. 480–484.
Rao,J.K.M. (1987) *Int. J. Peptide Protein Res.*, **29**, 276–281.
Risler,J.L., Delorme,M.O., Delacroix,H. and Henaut,A. (1988) *J. Mol. Biol.*, **204**, 1019–1029.
Shepp,L.A. and Vardi,Y. (1982) *Maximum-Likelihood Reconstruction for Emission Tomography.* Bell Laboratories, Murray Hill, New York.
Simon,I. (1993) *Peptide Res.*, **6**, 260–262.
Simon,I. and Cserző,M. (1990) *Trends Biochem. Sci.*, **15**, 135–136.
Simon,I., Glasser,L. and Scheraga,H.A. (1991) *Proc. Natl Acad. Sci. USA*, **88**, 3661–3665.
Smith,C.R. and Grandy,W.T. (1985) *Maximum Entropy and Bayesian Methods in Inverse Problems.* D.Reidel Publishing Co., Dordrecht, The Netherlands.
Stryer,L. (ed.) (1988) *Biochemistry.* W.H.Freeman and Co., New York.
Tüdős,É., Cserző,M. and Simon,I. (1990) *Int. J. Peptide Protein Res.*, **36**, 236–239.
Tüdős,É., Fiser,A. and Simon,I. (1994) *Int. J. Peptide Protein Res.*, **43**, 205–208.
Vonderviszt,F. and Simon,I. (1986) *Biochem. Biophys. Res. Commun.*, **139**, 11–17.
Vonderviszt,F., Mátrai,G. and Simon,I. (1986) *Int. J. Peptide Protein Res.*, **27**, 483–492.
Yockey,H.P. (1992) *Information Theory and Molecular Biology.* Cambridge University Press, Cambridge, UK.

Received September 28, 1994; revised February 9, 1995; accepted March 2, 1995