Protein Engineering vol.9 no.1 pp.27–36, 1996

Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins

Kentaro Tomii and Minoru Kanehisa[1]

Institute for Chemical Research, Kyoto University, Uji, Kyoto 611, Japan

[1]To whom correspondence should be addressed

An amino acid index is a set of 20 numerical values representing any of the different physicochemical and biochemical properties of amino acids. As a follow-up to the previous study, we have increased the size of the database, which currently contains 402 published indices, and re-performed the single-linkage cluster analysis. The results basically confirmed the previous findings. Another important feature of amino acids that can be represented numerically is the similarity between them. Thus, a similarity matrix, also called a mutation matrix, is a set of 20×20 numerical values used for protein sequence alignments and similarity searches. We have collected 42 published matrices, performed hierarchical cluster analyses and identified several clusters corresponding to the nature of the data set and the method used for constructing the mutation matrix. Further, we have tried to reproduce each mutation matrix by the combination of amino acid indices in order to understand which properties of amino acids are reflected most. There was a relationship between the PAM units of Dayhoff's mutation matrix and the volume and hydrophobicity of amino acids. The database of 402 amino acid indices and 42 amino acid mutation matrices is made publicly available on the Internet.

Keywords: cluster analysis/database/PAM/sequence alignment/similarity matrix

Introduction

Amino acid sequence analysis often provides important insights into the tertiary structure and biological function of proteins. The basic strategy is first to find the similarity of sequences in the forms of pairwise sequence alignments, multiple sequence alignments and homology searches against the databases, and then to infer 3-D structural similarity and/or functional similarity. The sequence similarity is usually defined by an optimization function based on a measure of similarity between amino acids. Thus, the amino acid similarity matrix, also called the amino acid mutation matrix, which defines this measure, is the basis of various sequence analysis methods.

Dayhoff *et al.* (1978a) were the first to compile such a mutation matrix. They constructed phylogenetic trees from 71 groups of closely related proteins (>85% pairwise sequence identity) and collected the data of accepted point mutations (PAMs) per 100 residues. Their log-odds matrix is still the most widely used scoring scheme. The elements of the mutation matrix compiled from such an observed amino acid exchange frequency represent the degree of physicochemical and biological similarities of amino acids in molecular evolution. In order to identify each accepted point mutation, Dayhoff *et al.* (1978a) used very similar protein amino acid sequences. Hence, there is the indication that 'each alignment will have poor informational content' (Risler *et al.*, 1988) about substitutions between distantly related proteins.

There have been attempts to observe directly exchanges of amino acids from more divergent sequences. Henikoff and Henikoff (1992) derived substitution frequencies from their BLOCK database of protein sequence motifs, where conserved segments were aligned no matter how evolutionarily distant sequences were. Structural comparison methods were incorporated into alignments (Risler *et al.*, 1988; Johnson and Overington, 1993), but they had the limitation that the number of data with known tertiary structures was smaller than the number of available sequence data. Lüthy *et al.* (1991) made separate mutation matrices for different secondary structures by using the profile method (Gribskov *et al.*, 1987). They suggested that, in detecting distantly related sequences with similar folds, using their distinct matrices was better than using Dayhoff's matrix alone.

There was another claim (Risler *et al.*, 1988; George *et al.*, 1990) that it was possible that Dayhoff's matrix was biased because of the size of the data set they had used. The amount of sequence data currently available is much larger than that used by Dayhoff *et al.* (1978a). Thus, the matrix has been updated with larger numbers of amino acid sequences (Gonnet *et al.*, 1992; Jones *et al.*, 1992). It has also been pointed out that substitution tendencies of non-aqueous proteins may be unlike those of soluble proteins (George *et al.*, 1990). Most recently a mutation matrix for transmembrane proteins was constructed (Jones *et al.*, 1994). In order to see the relationships between different matrices, the cluster analysis was made with nine (Risler *et al.*, 1988) or 13 (Johnson and Overington, 1993) mutation matrices.

Mutation matrices can also be constructed from the physicochemical properties of amino acids, such as hydrophobicity, volume and conformational preferences (Grantham, 1974; Miyata *et al.*, 1979; Mohana Rao, 1987). It is known that the volume and hydrophobicity of amino acids contribute significantly to Dayhoff's matrix (French and Robson, 1983; Kidera *et al.*, 1985b; Taylor, 1986). In fact these two properties are the major factors that influence the amino acid substitution during evolution (Grantham, 1974; Miyata *et al.*, 1979).

Kidera *et al.* (1985b) derived 10 orthogonal factors that expressed various amino acid properties, and represented each position of aligned homologous protein sequences by linear combinations of those factors. Kubota *et al.* (1981, 1982) used the correlation coefficient of several amino acid properties as a measure for finding homologous regions between two protein sequences. The 3-D–1-D scores of Bowie *et al.* (1991) can be regarded as kinds of amino acid properties that exhibit the compatibility of 20 amino acids with each of the 18 environments they arranged.

The amino acid property can be represented by the set of 20 numerical values, which we call the amino acid index (Kidera *et al.*, 1985a; Nakai *et al.*, 1988). As reported

```
H   PALJ810101
D   Normalized frequency of alpha-helix from LG (Palau et al., 1981)
R   0805095
A   Palau, J., Argos, P. and Puigdomenech, P.
T   Protein secondary structure
J   Int. J. Peptide Protein Res. 19, 394–401 (1981)
*   LG: a set of protein samples formed by 44 proteins
*   CF: a set of protein samples formed by 33 proteins
C   LEVM780104  0.988   NAGK730101  0.953   GEIM800101  0.951
    PRAM900102  0.943   LEVM780101  0.943   KANM800101  0.928
    TANS770101  0.918   ROBB760101  0.914   CRAJ730101  0.891
    PALJ810102  0.889   MAXF760101  0.889   ISOY800101  0.882
    CHOP780201  0.881   RACS820108  0.872   BURA740101  0.850
    GEIM800104  0.841   KAN800103   0.836   NAGK730103 -0.808
I   A/L   R/K   N/M   D/F   C/P   Q/S   E/T   G/W   H/Y   I/V
    1.30  0.93  0.90  1.02  0.92  1.04  1.43  0.63  1.33  0.87
    1.30  1.23  1.32  1.09  0.63  0.78  0.80  1.03  0.71  0.95
//
```

Fig. 1. An example of the amino acid index entry in the AAindex database. Each record of an entry is identified by the following codes: H, accession number; D, data description; R, LITDB (Seto *et al.*, 1988) literature database identifier; A, author(s); T, title of the article; J, journal reference; C, accession numbers of similar entries with the correlation coefficients of 0.8 (–0.8) or more (less); I, actual data in the specified order; and *, optional comments.

Table I. The list of 42 amino acid mutation matrices

Accession No.	Matrix (reference)	Basis
ALTS910101	The PAM-120 matrix (Altschul, 1991)	Sequence comparison
BENS940101	Log-odds scoring matrix collected in 6.4-8.7 PAM (Benner *et al.*, 1994)	Sequence comparison
BENS940102	Log-odds scoring matrix collected in 22-29 PAM (Benner *et al.*, 1994)	Sequence comparison
BENS940103	Log-odds scoring matrix collected in 74-100 PAM (Benner *et al.*, 1994)	Sequence comparison
BENS940104	Genetic code matrix (Benner *et al.*, 1994)	Genetic code
CSEM940101	Residue replace ability matrix (Cserzö *et al.*, 1994)	Neighbourhood selectivity
DAYM780301	Log odds matrix for 250 PAMs (Dayhoff *et al.*, 1978)	Sequence comparison
FEND850101	Structure-Genetic matrix (Feng *et al.*, 1985)	Genetic code and chemical similarity
FITW660101	Mutation values for the interconversion of amino acid pairs (Fitch, 1966)	Genetic code
GEOD900101	Hydrophobicity scoring matrix (George *et al.*, 1990)	Hydrophobicity index
GONG920101	A composite log-odds matrix (Gonnet *et al.*, 1992)	Sequence comparison
GRAR740104	Chemical distance (Grantham, 1974)	Physical property indices
HENS920101	BLOSUM45 substitution matrix (Henikoff-Henikoff, 1992)	Sequence comparison by protein blocks
HENS920102	BLOSUM62 substitution matrix (Henikoff-Henikoff, 1992)	Sequence comparison by protein blocks
HENS920103	BLOSUM80 substitution matrix (Henikoff-Henikoff, 1992)	Sequence comparison by protein blocks
JOHM930101	Structure-based amino acid scoring table (Johnson-Overington, 1993)	Structure-based sequence comparison
JOND920103	The 250 PAM PET91 matrix (Jones *et al.*, 1992)	Sequence comparison
JOND940101	The 250 PAM transmembrane protein exchange matrix (Jones *et al.*, 1994)	Sequence comparison
KOLA920101	Conformational similarity weight matrix (Kolaskar-Kulkarni-Kale, 1992)	Main-chain folding angles
LEVJ860101	The secondary structure similarity matrix (Levin *et al.*, 1986)	Sequence comparison by secondary structure
LUTR910101	Structure-based comparison table for outside other class (Lüthy *et al.*, 1991)	Sequence comparison
LUTR910102	Structure-based comparison table for inside other class (Lüthy *et al.*, 1991)	Sequence comparison
LUTR910103	Structure-based comparison table for outside alpha class (Lüthy *et al.*, 1991)	Sequence comparison
LUTR910104	Structure-based comparison table for inside alpha class (Lüthy *et al.*, 1991)	Sequence comparison
LUTR910105	Structure-based comparison table for outside beta class (Lüthy *et al.*, 1991)	Sequence comparison
LUTR910106	Structure-based comparison table for inside beta class (Lüthy *et al.*, 1991)	Sequence comparison
LUTR910107	Structure-based comparison table for other class (Lüthy *et al.*, 1991)	Sequence comparison
LUTR910108	Structure-based comparison table for alpha helix class (Lüthy *et al.*, 1991)	Sequence comparison
LUTR910109	Structure-based comparison table for beta strand class (Lüthy *et al.*, 1991)	Sequence comparison
MCLA710101	The similarity of pairs of amino acids (McLachlan, 1971)	Sequence comparison
MCLA720101	Chemical similarity scores (McLachlan, 1972)	Chemical similarity
MIYS930101	Base-substitution-protein-stability matrix (Miyazawa-Jernigan, 1993)	Genetic code and contact potential
MIYT790101	Amino acid pair distance (Miyata *et al.*, 1979)	Physical property indices
MOHR870101	EMPAR matrix (Mohana Rao, 1987)	Structural and physical property indices
NIEK910101	Structure-derived correlation matrix 1 (Niefind-Schomburg, 1991)	Main-chain folding angles
NIEK910102	Structure-derived correlation matrix 2 (Niefind-Schomburg, 1991)	Main-chain folding angles
OVEJ920101	STR matrix from structure-based alignments (Henikoff-Henikoff, 1993)[a]	Structure-based sequence comparison
QU_C930101	Cross-correlation coefficients of preference factors (Qu *et al.*, 1993)	Contacts of main chain atoms
QU_C930102	Cross-correlation coefficients of preference factors (Qu *et al.*, 1993)	Contacts of side chain atoms
QU_C930103	The mutant distance based on spatial preference factor (Qu *et al.*, 1993)	Main+side
RISJ880101	Scoring matrix (Risler *et al.*, 1988)	Structure-based sequence comparison
TUDE900101	Isomorphicity of replacements (Tüdös *et al.*, 1990)	Neighbourhood selectivity

[a]The substitution data were obtained by Overington *et al.* (1992).

previously (Nakai *et al.*, 1988) we constructed and maintain the database of amino acid indices. Here we present the revised format of this database, now called AAindex, and the results of the single-linkage hierarchical cluster analysis of 402 amino acid indices. We then report a new addition to the AAindex database, which is a collection of 42 reported mutation matrices.

```
H   MOHR870101
D   EMPAR matrix (Mohana Rao, 1987)
R   1304091
A   Mohana Rao, J.K.
T   New scoring matrix for amino acid residue exchanges based on residue characteristic physical parameters
J   Int. J. Peptide Protein Res. 29, 276-281 (1987)
C   LEVJ860101        0.813       HENS920101        0.801
I   Data ordered by I+J*(J-1)/2 where I,J = ARNDCQEGHILKMFPSTWYV
        16.    8.   16.    9.   10.   16.    9.   10.   11.   16.
        11.    8.    9.    8.   16.   11.   10.   11.   11.   10.
        16.   10.    9.   10.   11.    9.   11.   16.    8.    7.
        10.    9.    8.    8.    6.   16.   11.   10.   10.    9.
        10.   11.   11.    7.   16.    9.    4.    5.    3.    8.
         6.    4.    6.    8.   16.   11.    6.    7.    6.   11.
         9.    7.    6.   10.   10.   16.   10.   11.   11.   11.
         9.   12.   11.    7.   11.    4.    7.   16.   11.    6.
         6.    5.   10.    9.    8.    4.   10.    9.   11.    8.
        16.   10.    5.    6.    4.   10.    7.    6.    7.    9.
        12.   11.    6.   10.   16.    6.    6.    9.    8.    7.
         7.    5.   11.    5.    3.    4.    6.    2.    4.   16.
        10.    9.   11.   10.   10.   10.    9.   11.   10.    8.
         8.   10.    7.    8.   10.   16.   10.    9.   10.    9.
        10.   10.    8.   10.   10.   10.    9.    9.    8.   10.
         8.   11.   16.   11.    7.    8.    6.   11.    9.    7.
         8.   10.   11.   11.    7.   10.   11.    6.   10.   11
        16.    9.    7.    8.    7.   10.    8.    6.   10.    9.
        10.    9.    7.    8.   10.    8.   11.   11.   11.   16.
         9.    5.    5.    3.    8.    6.    4.    6.    9.   12.
        10.    5.    9.   11.    3.    8.   10.   11.   10.   16.
//
```

Fig. 2. An example of the amino acid mutation matrix entry in the AAindex database. The data format is the same as described in Figure 1.

The relationships among these matrices are analysed by hierarchical cluster analyses and each of the matrices is reconstructed from the combination of amino acid indices in order to find which properties of amino acids are reflected most.

Materials and methods

Amino acid index database

The amino acid index database, AAindex, now contains 402 published indices as compared with the previous version of 222 indices (Nakai *et al.*, 1988). It is organized in a flat-file format with one entry corresponding to one index, i.e. a set of 20 numerical values and associated reference information. A sample entry of the database is shown in Figure 1 and the complete list of the 402 indices has been made publicly available by the Japanese GenomeNet database service at the following addresses:

FTP	ftp.genome.ad.jp
Gopher	gopher.genome.ad.jp
WWW	http://www.genome.ad.jp/

In Gopher and WWW a database entry may be obtained by using the DBGET Integrated Database Retrieval System. The entire database may be downloaded by the anonymous FTP from the directory /db/genomenet/aaindex with the file name aaindex.

Mutation matrix database

There have been reports of different kinds of amino acid mutation matrices for use of sequence alignments and similarity searches. We have collected 42 published amino acid mutation matrices, listed in Table I. Each entry contains the actual data and the reference information as shown in Figure 2. The order of the 210 elements (20 diagonal and 20×19/2 off-diagonal elements) as they appear in the entry is shown in Figure 3. The collection of amino acid mutation matrices is stored in the same FTP directory with the file name aaindex2. This file is not a part of the DBGET system; use the FTP option from Gopher and WWW as well.

Cluster analysis

We first analysed the relationships among the 402 amino acid indices by the single-linkage hierarchical cluster analysis. We then analysed the relationships among the 42 amino acid mutation matrices by both the single-linkage and the complete-linkage hierarchical cluster analyses. To perform a cluster analysis, we defined the distance d between each pair of indices or matrices in the same manner as Nakai *et al.* (1988):

$$d = 1 - |c|$$

where c is the correlation coefficient:

$$c = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) / [\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2]^{1/2}$$

Here x_i and y_i represent an element of amino acid indices or mutation matrices to be compared. The mean value is denoted by $\bar{x}$ and $\bar{y}$, and the number of elements by n, which is 20 in the case of an amino acid index and 210 in the case of an amino acid mutation matrix. The result of a hierarchical cluster analysis is often represented by a dendrogram, but here we show the result by a minimum spanning tree (Nakai *et al.*, 1988) because it is easier to conceive the overall groupings when the number of data points is large.

Deriving mutation matrices from amino acid indices

In order to construct a mutation matrix from amino acid indices, we proceeded as follows. When a mutation matrix is derived from a single amino acid index, each element of the matrix is the normalized value of the difference between two index values of the corresponding amino acids. When a matrix is reproduced by combining multiple amino acid indices, we

```
I    Data ordered by I+J*(J-1)/2 where I,J = ARNDCQEGHILKMFPSTWYV
     AA   AR   RR   AN   RN   NN   AD   RD   ND   DD
     AC   RC   NC   DC   CC   AQ   RQ   NQ   DQ   CQ
     QQ   AE   RE   NE   DE   CE   QE   EE   AG   RG
     NG   DG   CG   QG   EG   GG   AH   RH   NH   DH
     CH   QH   EH   GH   HH   AI   RI   NI   DI   CI
     QI   EI   GI   HI   II   AL   RL   NL   DL   CL
     QL   EL   GL   HL   IL   LL   AK   RK   NK   DK
     CK   QK   EK   GK   HK   IK   LK   KK   AM   RM
     NM   DM   CM   QM   EM   GM   HM   IM   LM   KM
     MM   AF   RF   NF   DF   CF   QF   EF   GF   HF
     IF   LF   KF   MF   FF   AP   RP   NP   DP   CP
     QP   EP   GP   HP   IP   LP   KP   MP   FP   PP
     AS   RS   NS   DS   CS   QS   ES   GS   HS   IS
     LS   KS   MS   FS   PS   SS   AT   RT   NT   DT
     CT   QT   ET   GT   HT   IT   LT   KT   MT   FT
     PT   ST   TT   AW   RW   NW   DW   CW   QW   EW
     GW   HW   IW   LW   KW   MW   FW   PW   SW   TW
     WW   AY   RY   NY   DY   CY   QY   EY   GY   HY
     IY   LY   KY   MY   FY   PY   SY   TY   WY   YY
     AV   RV   NV   DV   CV   QV   EV   GV   HV   IV
     LV   KV   MV   FV   PV   SV   TV   WV   YV   VV
//
```

Fig. 3. The order of the matrix elements as stored in the AAindex database. The amino acid types are given in the standard one-letter codes.

adopted the method of Grantham (1974). For example, when combining three indices p, q and r, an element of the derived mutation matrix D_{ij} for the pair of amino acids i and j is given by the following equation:

$$D_{ij} = [\alpha(p_i - p_j)^2 + \beta(q_i - q_j)^2 + \gamma(r_i - r_j)^2]^{1/2}$$

where

$$\alpha = (1/\overline{Dp})^2,\ \beta = (1/\overline{Dq})^2,\ \gamma = (1/\overline{Dr})^2$$

are the scaling factors which are calculated from the mean value $\overline{D}$ of 190 off-diagonal elements. With the use of the 402 indices in the database, we search an index or indices in combination that give the best correlation coefficient with each of the 42 mutation matrices.

Results

Minimum spanning tree of amino acid indices

The minimum spanning tree of the 402 amino acid indices is shown in Figure 4, where an index corresponds to a node represented by a circle. Each index can be identified in the enlarged drawing of Figure 5 by the number that corresponds to the listing in the AAindex on the Internet. The linkage between two indices was made by the single-linkage cluster analysis. The shaded area denotes that the distance between two indices is 0.1 or less, i.e. the absolute value of the correlation coefficient is 0.9 or larger. For the sake of convenience, we divided the minimum spanning tree into six regions: α and turn propensities, β propensity, amino acid composition, hydrophobicity, physicochemical properties, and other properties such as the frequency of left-handed helix (Maxfield and Scheraga, 1976; Tanaka and Scheraga, 1977). The six regions were identified, respectively, by the letters A, B, C, H, P and O as shown in Figure 5. The boundaries of the regions were determined by the largest distance among relevant node connections; for example, B168 was nearer to B257 than to H170. Of course, the assignment to each of the six regions is not very meaningful for the outlying indices. In the previous study Nakai *et al.* (1988) classified the minimum spanning tree of the 222 indices into four regions: α and turn propensities, β propensity, hydrophobicity and physicochemical properties.

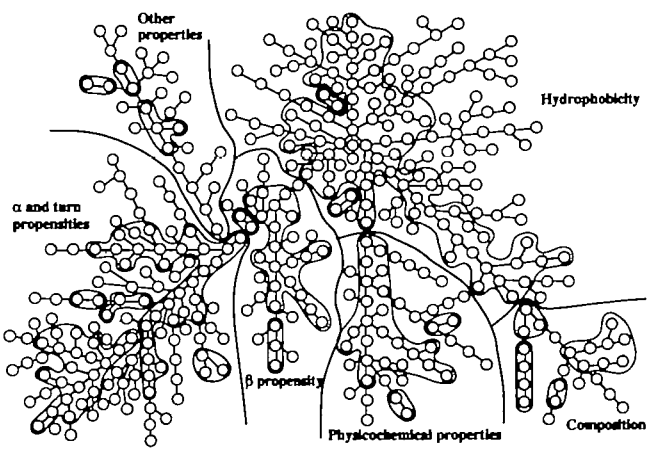

Fig. 4. The minimum spanning tree of 402 amino acid indices. The shaded areas correspond to clusters identified by single-linkage with a threshold distance of 0.1. The tree is conveniently divided into six regions.

Here the last physicochemical properties region was further subdivided into three regions.

The result of clustering was generally consistent with the previous study. The subgroups of the hydrophobicity cluster that had been observed with the threshold distance of 0.05 were still present (data not shown). However, some minor differences were also observed. When individual indices were examined, there were instances of repositioning within a large cluster. The helix-coil equilibrium constant (Ptitsyn and Finkelstein, 1983, A256) used to be located in between the α subgroup and the turn subgroup within the α and turn propensities region. It is now a member of the mostly turn propensity subgroup which also contains neural network weights (Qian and Sejnowski, 1988) for coil at the window positions −1 to 3 (A289–A293) as well as for helix at the window positions 1 to 4 (A265–A268).

The number of indices for the amino acid composition was increased. As shown in Figure 5C the composition indices of mitochondrial proteins (Nakashima *et al.*, 1990, C199, C201, C203 and C208) and membrane proteins (Nakashima and

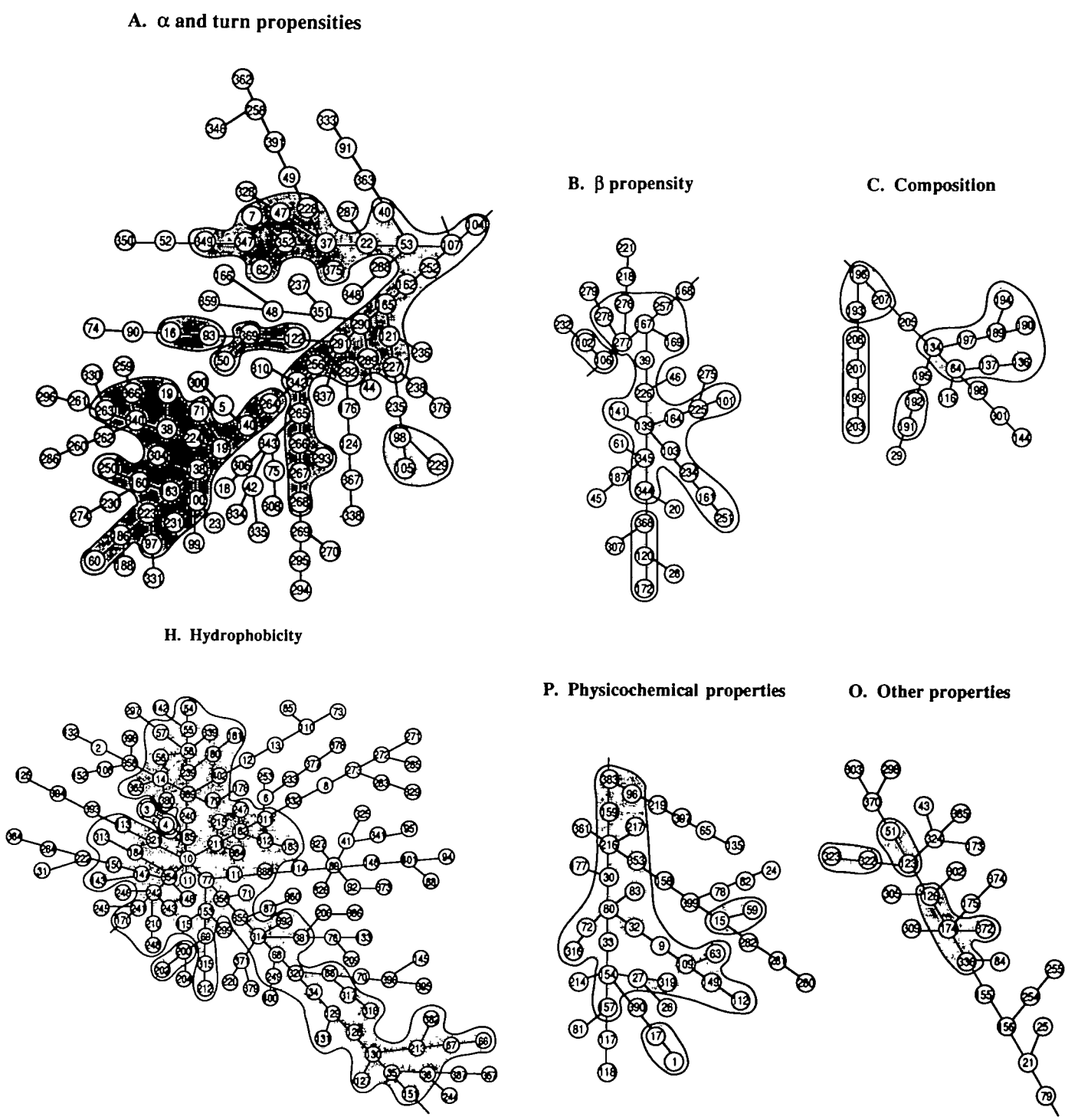

Fig. 5. Enlarged drawing of the minimum spanning tree of amino acid indices. Each amino acid index is identified in the text by the single-letter classification code, A, B, C, H, P or O, followed by the number listed in the AAindex, available on the Internet.

Nishikawa, 1992, C193 and C196) are separate from the composition index of Dayhoff *et al.* (1978b, C64), when the distance of 0.1 was used as the threshold.

It was interesting to observe in the lower right region of Figure 5A that the helix index for alpha proteins (Geisow and Roberts, 1980, A98) and the normalized frequency of helix in all alpha class (Palau *et al.*, 1981, A229) are highly correlated (the correlation coefficient is 0.92) with each other. However, except for the aperiodic index for alpha-proteins (Geisow and Roberts, 1980, A105) there was no index that exhibited a correlation coefficient of 0.8 or more with either of them.

Minimum spanning tree of amino acid mutation matrices

Figure 6 shows the minimum spanning tree of 42 amino acid mutation matrices. The shaded areas denote a distance of 0.04 or less between two matrices, while the outer contours denote a distance of 0.08 or less. When the distance is >0.3, the linkage is represented by a dashed line. These are the results

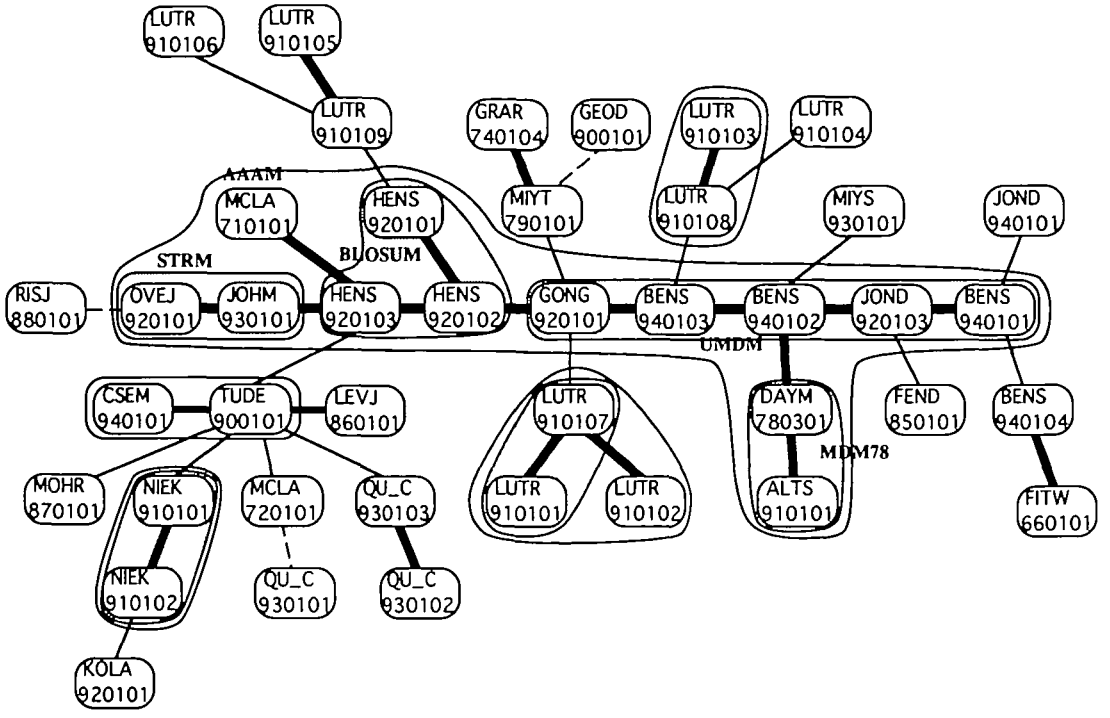

Fig. 6. The minimum spanning tree of 42 amino acid mutation matrices. The shaded areas and the outer contours correspond to clusters identified by single-linkage with a threshold distances of 0.04 and 0.08, respectively. The thick lines denote clusters identified by complete-linkage with a threshold distance of 0.18. See Table II for identification of each matrix.

obtained by the single-linkage hierarchical cluster analysis. In addition, the clusters identified by the complete-linkage hierarchical cluster analysis with a threshold distance of 0.18 are shown by the thick lines in Figure 6.

The mutation matrices can now be grouped into several clusters corresponding with the method and the data set used for construction. When the distance of 0.08 was applied to the threshold, a large cluster emerged containing most of the matrices that are widely used in sequence alignments, such as the Dayhoff PAM250 matrix (Dayhoff *et al.*, 1978a, DAYM780301) and the BLOSUM series matrices (Henikoff and Henikoff, 1992, HENS920101-03). These matrices are constructed from the observation of amino acid exchanges in related proteins. The same clusters were also obtained by the complete-linkage hierarchical cluster analysis with a threshold distance of 0.18, which is illustrated by the thick lines in Figure 6. Thus, the distance between any pair of the 13 matrices constituting this cluster is ≤ 0.18 (complete linkage) and any matrix has the closest one with a distance of 0.1 or smaller (single linkage).

Among the matrices based on observed substitution data, the mutation matrices for the different protein secondary structure classes, α-helix, β-strand and others, as well as inside and outside (Lüthy *et al.*, 1991, LUTR910101-09) and the matrix for transmembrane proteins (Jones *et al.*, 1994, JOND940101) are distinct and not included in the cluster of 13 matrices. The matrices for residues in the other secondary structure class were classified into the same cluster (lower middle of Figure 6) irrespective of whether inside or outside of the globule (LUTR910101,02,07). Especially the matrix for outside (LUTR910101) and inside and outside combined (LUTR910107) are very close with a distance of only 0.01. The three matrices for residues in β-strands (LUTR-910105,06,09) could be combined into a single cluster (upper left of Figure 6) when a threshold distance of 0.09 was used, although the other class matrices would then be merged into the above cluster of 13 matrices. For the three matrices for α-helices, the matrix for all alpha (LUTR910108) and the matrix for outside alpha (LUTR910103) are similar, but the matrix for inside alpha (LUTR910104) is somewhat different (upper middle of Figure 6).

The rest of the single member clusters in Figure 6 are the matrices mainly based on physicochemical properties of amino acids (McLachlan, 1972, MCLA720101; Grantham, 1974, GRAR740104; Miyata *et al.*, 1979, MIYT790101; George *et al.*, 1990, GEOD900101), the matrices predominantly based on conformational preferences (Mohana Rao, 1987, MOHR870101; Kolaskar and Kulkarni-Kale, 1992, KOLA920101), the matrices based on indices that individual authors had developed (Miyazawa and Jernigan, 1993, MIYS930101; Qu *et al.*, 1993, QU_C930101-03), and the matrices dependent on the genetic code (Fitch, 1966, FITW660101; Benner *et al.*, 1994, BENS940104; Feng *et al.*, 1985, FEND850101). The matrix by Risler *et al.* (1988, RISJ880101) is based on observed substitution data obtained by using structural comparison of homologous proteins, but the matrix is different because it is converted to the χ^2 distance matrix.

There are two small clusters in the lower left region of Figure 6. One of them (Niefind and Schomburg, 1991, NIEK910101,02) is based on main chain conformational preferences. The difference between the two members is due to the treatment of the data as discrete or Gaussian distribution. In the other cluster, one member (Cserzö *et al.*, 1994, CSEM940101) is a refined version of the other (Tüdös *et al.*, 1990, TUDE900101) based on their developed method of neighbourhood selectivity (Cserzö and Simon, 1989). We noted that the correlation coefficients reported in their subsequent

Table II. The result of reproducing amio acid mutation matrices from the combination of amino acid indices[a]

Matrix	Correlation coeff.	One index	Correlation coeff.	Two indices	Correlation coeff.	Three indices
GRAR740104	0.80	H110	0.91	H110 P353	1.00	H110 H111 P112
GEOD900101	–0.99	H153	–0.99	H115 H153	–0.97	H 77 H115 H153
MIYT790101	0.75	H 55	1.00	H111 P112	0.97	H 55 H111 P112
MOHR870101	–0.69	B277	–0.84	H 68 A252	–0.92	H355 B164 A162
*GONG920101	–0.74	H 55	–0.84	H111 P112	–0.87	H111 H211 P149
MIYS930101	–0.85	H185	–0.86	H 71 H185	–0.87	H 71 H185 H384
*BENS940103	–0.73	H 55	–0.84	H111 P112	–0.86	H111 H211 P149
LUTR910108	–0.80	H212	–0.83	H211 P383	–0.85	H151 P383 P353
QU(C930102	–0.70	H388	–0.80	H355 B 28	–0.85	H 94 H111 B 28
*HENS920102	–0.65	H211	–0.79	H111 P112	–0.84	H210 H389 P154
*HENS920103	–0.64	H111	–0.80	H111 P112	–0.84	H111 P217 B278
*JOHM930101	–0.64	H210	–0.80	H127 P353	–0.84	H127 H212 A165
*BENS940102	–0.69	H 55	–0.81	H111 P177	–0.83	H111 H365 P177
CSEM940101	–0.77	H111	–0.79	H 55 H213	–0.83	H 67 H111 P112
LUTR910104	–0.76	P216	–0.81	H212 P217	–0.83	H111 P217 P383
*HENS920101	–0.58	H211	–0.75	H151 P112	–0.82	H210 H402 P 33
LUTR910109	–0.67	H 13	–0.74	H127 H212	–0.82	H127 H402 P216
*DAYM780301	–0.65	H212	–0.77	H365 P177	–0.81	H111 H212 P177
LEVJ860101	–0.75	H365	–0.79	H212 H365	–0.81	H212 H365 A289
*OVEJ920101	–0.60	H210	–0.76	H127 P353	–0.81	H127 H212 A165
QU(C930103	–0.62	H388	–0.72	H 94 H111	–0.81	H 87 H 94 A121
LUTR910106	–0.56	H389	–0.68	H 13 H247	–0.80	H 13 P319 B168
RISJ880101	–0.55	O303	–0.70	P216 O303	–0.80	P219 A308 O303
TUDE900101	–0.73	H111	–0.77	H111 H211	–0.80	H 67 H111 P158
*JOND920103	–0.63	H 55	–0.76	H111 P177	–0.79	H111 H365 P177
*MCLA710101	–0.63	H147	–0.74	H 14 H150	–0.79	H150 H365 P112
*ALTS910101	–0.60	H212	–0.75	H243 P216	–0.78	H111 H212 P177
KOLA920101	–0.51	A290	–0.66	P 1 A119	–0.77	P 1 B120 A264
LUTR910102	–0.57	P353	–0.69	H181 A252	–0.77	H 12 P177 A291
*BENS940101	–0.60	H365	–0.72	H111 P177	–0.76	H151 H402 P177
LUTR910107	–0.60	H384	–0.69	H147 P353	–0.76	H384 H402 P319
JOND940101	–0.54	H 71	–0.67	H151 P319	–0.73	H151 H365 P319
FEND850101	–0.64	H111	–0.69	H111 P353	–0.72	H111 H185 P177
LUTR910101	–0.54	H147	–0.66	H 10 P319	–0.72	H354 P158 A 75
LUTR910105	–0.52	H 13	–0.67	H110 P383	–0.71	H 77 H110 P216
NIEK910101	–0.56	A224	–0.66	B161 A224	–0.70	B161 A 19 A140
NIEK910102	–0.54	A224	–0.65	B161 A224	–0.69	B161 A 19 A140
LUTR910103	–0.56	H212	–0.62	H365 P112	–0.68	H 14 H148 P361
MCLA720101	–0.43	H111	–0.61	H111 P157	–0.65	H241 H402 P157
BENS940104	–0.38	H 71	–0.47	H 71 P177	–0.48	H 71 H111 P177
QU(C930101	–0.36	A 50	–0.41	H271 A 90	–0.46	H 94 A 90 O255
FITW660101	0.32	H243	0.39	H111 A 52	0.44	H184 H388 A 52

[a]The amino acid index is represented by the classification code shown in the AAindex, available on the Internet.

paper (Tusnády *et al.*, 1995) are somewhat different from ours because the assignment of diagonal elements that are missing in the original matrices is different.

When a distance of 0.04 is used as the threshold, the large cluster of 13 matrices can be further divided into five subclusters, which are named respectively Dayhoff's mutation data matrix (MDM78) group, the updated mutation data matrix (UMDM) group, Henikoff's BLOSUM group, the structure-based matrix (STRM) group and McLachlan's alternative amino acids-based matrix (AAAM) group. As mentioned, these matrices are all obtained from the observed frequency of amino acid substitution data, but apparently there are differences due to the size and the nature of the data, which are reflected in the five subclusters.

The MDM78 cluster contains the log odds matrices for 250 PAM units (Dayhoff *et al.*, 1978a, DAYM780301) and for 120 PAM units (Altschul, 1991, ALTS910101). Both matrices are based on the same sequence data (Dayhoff *et al.*, 1978a); namely, the mutation probability matrix for any PAM units can be obtained by making the power of the matrix for 1 PAM unit. Compared with Dayhoff's data set, the five matrices constituting the UMDM group are made by using a larger amount of sequence data. While the matrices of MDM78 and UMDM groups are obtained by using closely related protein sequences, the matrices in the other groups are constructed by directly observing substitutions without respect to the evolutionary distance among sequences being compared. The BLOSUM group contains three matrices, BLOSUM 45 (HENS920101), BLOSUM 62 (HENS920102) and BLOSUM 80 (HENS920103), representing different levels of clustering percentages to adjust contributions from closely related sequences when they measured the amino acid replacement frequency from the aligned segments of the BLOCK database. The matrix JOHM930101 (Johnson and Overington, 1993) is based on the amino acid exchange frequency observed in structurally aligned sets of homologous proteins. The matrix OVEJ920101 was computed in the same manner as the BLOSUM series (Henikoff and Henikoff, 1993) from different substitution data (Overington *et al.*, 1992). It seems that McLachlan's scoring scheme (McLachlan, 1971) is somewhat peculiar; each score is assigned an integer between 0 and 9 from the relative substitution frequencies.

Reproducing matrices from amino acid indices

According to the procedure described in Materials and methods, we searched the best combination of up to three amino acid indices to represent a mutation matrix. The result is summarized in Table II where the best correlation coefficient for each of the 42 published matrices is shown when the derived matrix is obtained from a single amino acid index (column 2), two indices in combination (column 4) and three indices in combination (column 6). The best combination of two indices was calculated from 80 601 ($= {}_{402}C_2$) possibilities and the best combination of three indices was searched from 10 746 800 ($= {}_{402}C_3$) possibilities.

Here the correlation coefficients were calculated from 190 off-diagonal elements. When the calculation was made from all 210 elements, the correlation coefficient for the amino acid index-based matrices NIEK910101,02, KOLA920101 and QU_C930101 or the genetic code-based matrices FITW660101 and BENS940104 showed a marked improvement, ~0.1 or more (data not shown). This is mainly due to the fact that all diagonal elements of such matrices are equal, namely, the difference is 0. The matrices RISJ880101, FEND850101, MCLA710101 whose diagonal elements are 8 or 9 and the matrix MCLA720101 whose diagonal elements are 5 or 6 also have a similar bias. Except for these matrices the result with 190 elements conformed well to that with 210 elements.

Table II is sorted according to the value for the three index combination. The top ones are the matrices calculated from the indices stored in our database, so it was natural to observe a perfect correlation. Concerning the 13 matrices that are grouped into the same cluster in Figure 6, which are identified by the asterisks in Table II, they exhibit a similar tendency. When a single index was used to represent a matrix, all the selected indices belonged to the large hydrophobicity cluster (shaded area) shown in Figure 5H. When the combination of two indices was used, all the selected pairs except for MCLA710101 consisted of the hydrophobicity and the size of the amino acid side chain. When 210 elements were used to calculate the correlation coefficient, the matrix MCLA710101 also had a similar combination of the hydrophobicity and the size. The refractivity index (McMeekin *et al.*, 1964, P177) which often appeared here is also highly correlated with the amino acid size indices in the physicochemical properties region shown in Figure 5P. When the combination of three indices was examined, only a slight improvement of the correlation coefficient was observed over the two index combination. Thus, the elements of the 13 published mutation matrices reflect mostly the similarity of the volume and hydrophobicity of amino acids. This suggests that for each amino acid replacement during protein evolution the volume needs be conserved to retain the packing of the globule and the hydrophobicity needs be conserved to keep the properties of inside and outside residues.

The matrices that take into account the main chain torsion angles (Niefind and Schomburg, 1991, NIEK910101,02; Kolaskar and Kulkarni-Kale, 1992, KOLA920101) are correlated with the conformational preference indices. The matrix MOHR870101 (Mohana Rao, 1987) is mostly explained by only three indices in combination, despite the fact that the matrix was established by using five parameters, i.e. three conformational preference parameters, polarity and hydrophobicity. Although Levin *et al.* (1986) had empirically determined their matrix to optimize the secondary structure matching, the matrix LEVJ860101 was highly correlated with a single hydrophobicity index and no significant improvement was observed in the combination of two or three indices. This is consistent with the result of Risler *et al.* (1988), who found an eigenvalue that could mostly represent the matrix of Levin *et al.* (1986).

The genetic code-based matrices FITW660101 (Fitch, 1966) and BENS940104 (Benner *et al.*, 1994) did not have a good correlation with any amino acid indices, which is consistent with the observation by Nakai *et al.* (1988). The correlation coefficients with the derived matrices were <0.5. When we performed a search of best combinations using 400 pseudo-indices that had sets of random values, the mean of the correlation coefficients was 0.49. This implies that the genetic code-based matrices cannot be represented by any amino acid properties. Compared with these two matrices, the matrix FEND850101 (Feng *et al.*, 1985) which is considered both genetic code-based and physicochemical similarity-based did exhibit correlations with some indices.

Discussion

Since the original efforts of Dayhoff and Eck (1968) and McLachlan (1971) who studied amino acid substitutions in homologous protein sequences, and of Fitch (1966) who employed a matrix derived from the genetic code, there have been reports of various mutation matrices to search for sequence similarity. Among them Dayhoff's PAM 250 matrix (Dayhoff *et al.*, 1978a, DAYM780301) has long been used as a standard similarity measure in protein sequence comparison. On the other hand, Dayhoff's matrix has also been criticized because of, for instance, the possible bias due to the limited size of the data set, the influence of observing amino acid mutations only in closely related proteins and their assumptions on the evolutionary model of proteins. According to our analysis, at least the first one is not really critical. That is, the updated versions with larger sets of sequence data, JOND920103 (Jones *et al.*, 1992), GONG920101 (Gonnet *et al.*, 1992) and BENS940101-03 (Benner *et al.*, 1994) are all very similar to the original Dayhoff matrix. For the second one, we have shown that the matrices derived from sequence data of varying evolutionary distances (MCLA720101, HENS920101-03, OVEJ920101 and JOHM930101) are also correlated with the original Dayhoff matrix. In practice, however, there may be some differences in detecting sequence similarity.

Concerning the model of protein evolution, Benner *et al.* (1994) suggested that the amino acid substitution patterns are not uniform at any evolutionary distance between sequences, by separately constructing matrices (BENS940101-03) with specific divergence ranges of sequences. They concluded that at low divergence the genetic code strongly affected amino acid mutations, but chemical characters of amino acids were influential at high divergence (Gonnet *et al.*, 1992; Benner *et al.*, 1994). Our results in Table II also suggest that when more divergent sequence data are used in constructing matrices, these matrices have higher correlations with the size and hydrophobicity of amino acids. If this is the case, why can Dayhoff's matrix detect distantly related sequences despite the fact that they only observed substitutions in closely related (low divergent) sequences? Schwartz and Dayhoff (1978) empirically found that the PAM 250 unit matrix was effective to do so. Here we uncover another clue. Figure 7 shows the correlation coefficients between the Dayhoff matrix calculated for every 10 PAM units from 10 to 490 and the matrix derived from the size (Grantham, 1974, P112) and hydrophobicity

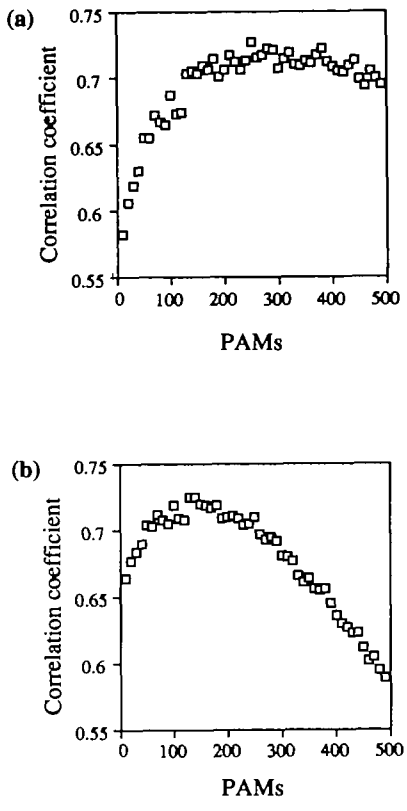

Fig. 7. The absolute value of the correlation coefficient between the matrix constructed from the volume (P112) and hydrophobicity (H365) indices and the Dayhoff matrix calculated at every 10 PAM units. The correlation coefficient was obtained either from 190 off-diagonal elements (a) or from all 210 elements (b).

(Sweet and Eisenberg, 1983, H365) indices. The correlation coefficients were obtained from 190 off-diagonal elements (Figure 7a) or from all 210 elements (Figure 7b). The PAM unit range of 70 to 250 that includes widely used PAM 120 and PAM 250 matrices exhibits higher correlations with the derived matrix of the size and hydrophobicity than the other ranges. This indicates that the PAM units in this range indeed reflect the size and hydrophobicity of amino acids. Dayhoff *et al.* (1978a) were thus able to construct the matrix for substitution patterns in distantly related proteins by extending the PAM units of their mutation probability matrix.

There is, however, a large difference between the asymptotic behaviors of Benner's and Dayhoff's matrices for longer evolutionary distances. On the one hand, in Benner's matrix the element of a pair of amino acids that are physicochemically dissimilar but similar in the genetic code, e.g. Cys and Trp, decreases in value as the evolutionary distance increases. On the other hand, because of their assumption, i.e. a Markovian model, the off-diagonal elements increase monotonically with increasing distances in Dayhoff's matrix. Our analysis also indicated that the matrices based on the genetic code (FITW660101 and BENS940104) did not sufficiently reflect any properties of amino acids. This may be the reason why such matrices are not suited for searching distantly related proteins (Schwartz and Dayhoff, 1978; Feng *et al.*, 1985).

The diversity of amino acid properties is the key to the structure, function and evolution of protein molecules. Figure 8 is an illustration of how amino acid indices are related to other parameters of amino acids. As shown in this paper the amino acid mutation matrix is a manifestation of amino acid indices, notably the hydrophobicity and the side-chain size. While the mutation matrix is the scoring scheme for sequence comparison, the so-called structural parameters are

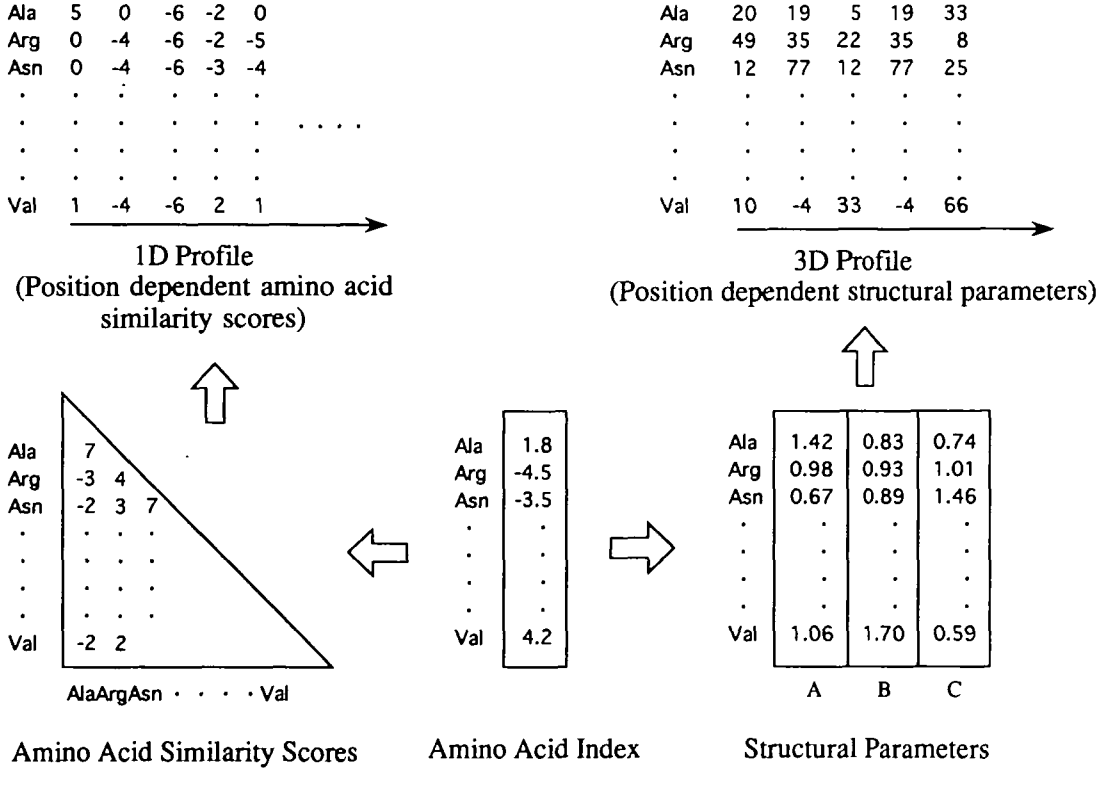

Fig. 8. An illustration showing the relationships among amino acid indices, mutation matrices (similarity scores) and profiles.

the scoring scheme for structure prediction or sequence/structure comparison. For example, the conformational parameters of Chou and Fasman (1978) represent empirical relationships between 20 amino acid residues and three secondary structure classes. The 1-D profile and the 3-D profile are, respectively, the position dependent scoring schemes for sequence/sequence comparison and sequence/structure comparison. The 1-D profile of Gribskov *et al.* (1987) is derived from Dayhoff's PAM 250 matrix, while the 3-D–1-D scores of Bowie *et al.* (1991) can be regarded as a refined form of conformational parameters. The amino acid index database AAindex, which currently contains various amino acid indices, structural parameters, and mutation matrices, can thus be a useful resource for sequence and structure analyses of proteins.

Acknowledgements

We thank our colleague Hiroyuki Ogata who had maintained the Amino Acid Index Database and Dr Kenta Nakai, Osaka University, for helpful discussions. This work was supported in part by a grant-in-aid for scientific research on the priority area 'Genome Informatics' from the Ministry of Education, Science and Culture. The computation time was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

References

Altschul,S.F. (1991) *J. Mol. Biol.*, **219**, 555–565.
Benner,S.A., Cohen,M.A. and Gonnet,G.H. (1994) *Protein Engng*, **7**, 1323–1332.
Bowie,J.U., Lüthy,R. and Eisenberg,D. (1991) *Science*, **253**, 164–170.
Chou,P.Y. and Fasman,G.D. (1978) *Adv. Enzymol.*, **47**, 45–148.
Cserzö,M. and Simon,I. (1989) *Int. J. Peptide Protein Res.*, **34**, 184–195.
Cserzö,M., Bernassau,J.-M., Simon,I. and Maigret,B. (1994) *J. Mol. Biol.*, **243**, 388–396.
Dayhoff,M.O. and Eck,R.V. (1968) In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Spring, MD, Vol. 3, pp. 33–41.
Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978a) In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC, Vol. 5, Suppl.3, pp. 345–352.
Dayhoff,M.O., Hunt,L.T. and Hurst-Calderone,S. (1978b) In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, D.C., Vol.5, Suppl.3, p. 363.
Feng,D.F., Johnson,M.S. and Doolittle,R.F. (1985) *J. Mol. Evol.*, **21**, 112–125.
Fitch,W.M. (1966) *J. Mol. Biol.*, **16**, 9–16.
French,S. and Robson,B. (1985) *J. Mol. Evol.*, **19**, 171–175.
Geisow,M.J. and Roberts,R.D.B. (1980) *Int. J. Biol. Macromol.*, **2**, 387–389.
George,D.G., Barker,W.C. and Hunt,L.T. (1990) *Methods Enzymol.*, **188**, 333–351.
Gonnet,G.H., Cohen,M.A. and Benner,S.A. (1992) *Science*, **256**, 1443–1445.
Grantham,R. (1974) *Science*, **185**, 862–864.
Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
Henikoff,S. and Henikoff,J.G. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
Henikoff,S. and Henikoff,J.G. (1993) *Proteins*, **17**, 49–61.
Johnson,M.S. and Overington,J.P. (1993) *J. Mol. Biol.*, **233**, 716–738.
Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) *Comput. Appl. Biosci.*, **8**, 275–282.
Jones,D.T., Taylor,W.R. and Thornton,J.M. (1994) *FEBS Lett.*, **339**, 269–275.
Kidera,A., Konishi,Y., Oka,M., Ooi,T. and Scheraga,H.A. (1985a) *J. Protein Chem.*, **4**, 23–55.
Kidera,A., Konishi,Y., Ooi,T and Scheraga,H.A. (1985b) *J. Protein Chem.*, **4**, 265–297.
Kolaskar,A.S. and Kulkarni-Kale,U. (1992) *J. Mol. Biol.*, **223**, 1053–1061.
Kubota,Y., Takahashi,S., Nishikawa,K. and Ooi,T. (1981) *J. Theor. Biol.*, **91**, 347–361.
Kubota,Y., Nishikawa,K. and Ooi,T. (1982) *Biochim. Biophys. Acta*, **701**, 242–252.
Levin,J.M., Robson,B. and Garnier,J. (1986) *FEBS Lett.*, **205**, 303–308.
Lüthy,R., McLachlan,A.D. and Eisenberg,D. (1991) *Proteins*, **10**, 229–239.
McLachlan,A.D. (1971) *J. Mol. Biol.*, **61**, 409–424.
McLachlan,A.D. (1972) *J. Mol. Biol.*, **64**, 417–437.
McMeekin,T.L., Groves,M.L. and Hipp N.J. (1964) In Stekol,J.A. (ed.), *Amino Acids and Serum Proteins*. American Chemical Society, Washington, DC, p. 54.
Maxfield,F.R. and Scheraga,H.A. (1976) *Biochemistry*, **15**, 5138–5153.
Miyata,T., Miyazawa,S. and Yasunaga,T. (1979) *J. Mol. Evol.*, **12**, 219–236.
Miyazawa,S. and Jernigan,R.L. (1993) *Protein Engng*, **6**, 267–278.
Mohana Rao,J.K. (1987) *Int. J. Pept. Protein Res.*, **29**, 276–281.
Nakai,K., Kidera,A. and Kanehisa,M. (1988) *Protein Engng*, **2**, 93–100.
Nakashima,H. and Nishikawa,K. (1992) *FEBS Lett.*, **303**, 141–146.
Nakashima,H., Nishikawa,K. and Ooi,T. (1990) *Proteins*, **8**, 173–178.
Niefind,K. and Schomburg,D. (1991) *J.Mol.Biol.*, **219**, 481–497.
Overington,J.P., Donnelly,D., Sali,A., Johnson,M.S. and Blundell,T.L. (1992) *Protein Sci.*, **1**, 216–226.
Palau,J., Argos,P. and Puigdomenech,P. (1981) *Int. J. Peptide Protein Res.*, **19**, 394–401.
Ptitsyn,O.B. and Finkelstein,A.V. (1983) *Biopolymers*, **22**, 15–25.
Qian,N. and Sejnowski,T.J. (1988) *J. Mol. Biol.*, **202**, 865–884.
Qu,C., Lai,L., Xu,X. and Tang,Y. (1993) *J. Mol. Evol.*, **36**, 67–78.
Risler,J.L., Delorme,M.O., Delacroix,H. and Henaut,A. (1988) *J. Mol. Biol.*, **204**, 1019–1029.
Schwartz,R.M. and Dayhoff,M O. (1978) In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC, Vol. 5, Suppl.3, pp. 353–358.
Seto,Y., Ihara,S., Kohtsuki,S., Ooi,T. and Sakakibara,S. (1988) In Lesk,A.M. (ed.), *Computational Molecular Biology*. Oxford University Press, New York, pp. 27–37.
Sweet,R.M. and Eisenberg,D. (1983) *J. Mol. Biol.*, **171**, 479–488.
Tanaka,S. and Scheraga,H.A. (1977) *Macromolecules*, **10**, 9–20.
Taylor,W.R (1986) *J. Theor. Biol.*, **119**, 205–218.
Tüdös,E., Cserzo,M. and Simon,I. (1990) *Int. J. Peptide Protein Res.*, **36**, 236–239.
Tusnády,G.E., Tusnády,G. and Simon,I. (1995) *Protein Engng*, **8**, 417–423.

Received September 7, 1995; revised October 9, 1995; accepted October 11, 1995