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

Kentaro Tomii and Minoru Kanehisa

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

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) 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., 1983b; 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...
Table I. The list of 42 amino acid mutation matrices

<table>
<thead>
<tr>
<th>Accession No.</th>
<th>Matrix (reference)</th>
<th>Basis</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALT5910101</td>
<td>The PAM-120 matrix (Altchul, 1991)</td>
<td>Sequence comparison</td>
</tr>
<tr>
<td>BEN5940102</td>
<td>Log-odds scoring matrix collected in 6-8.7 PAM (Bennet et al., 1994)</td>
<td>Sequence comparison</td>
</tr>
<tr>
<td>BEN5940103</td>
<td>Log-odds scoring matrix collected in 22-29 PAM (Bennet et al., 1994)</td>
<td>Sequence comparison</td>
</tr>
<tr>
<td>BEN5940104</td>
<td>Genetic code matrix (Bennet et al., 1994)</td>
<td>Genetic code</td>
</tr>
<tr>
<td>CSE940101</td>
<td>Residue replace ability matrix (Cserzé et al., 1994)</td>
<td>Neighbourhood selectivity</td>
</tr>
<tr>
<td>DAY750101</td>
<td>Log odds matrix for 250 PAMs (Dayhoff et al., 1978)</td>
<td>Sequence comparison</td>
</tr>
<tr>
<td>FIN950101</td>
<td>Structure-Genetic matrix (Feng et al., 1985)</td>
<td>Genetic code and chemical similarity</td>
</tr>
<tr>
<td>FTT660101</td>
<td>Mutation values for the interconversion of amino acid pairs (Fitch, 1966)</td>
<td>Genetic code</td>
</tr>
<tr>
<td>GEO990101</td>
<td>Hydrophobicity scoring matrix (George et al., 1999)</td>
<td>Hydrophobicity index</td>
</tr>
<tr>
<td>GON5920101</td>
<td>A composite log-odds matrix (Gonnet et al., 1992)</td>
<td>Sequence comparison</td>
</tr>
<tr>
<td>GRAR740104</td>
<td>Chemical distance (Grantham, 1974)</td>
<td>Physical property indices</td>
</tr>
<tr>
<td>HEN5920102</td>
<td>BLOSUM45 substitution matrix (Henikoff-Henikoff, 1992)</td>
<td>Sequence comparison by protein blocks</td>
</tr>
<tr>
<td>HEN5920103</td>
<td>BLOSUM62 substitution matrix (Henikoff-Henikoff, 1992)</td>
<td>Sequence comparison by protein blocks</td>
</tr>
<tr>
<td>HEN5920104</td>
<td>BLOSUM80 substitution matrix (Henikoff-Henikoff, 1992)</td>
<td>Sequence comparison by protein blocks</td>
</tr>
<tr>
<td>JOH930101</td>
<td>Structure-based amino acid scoring table (Johnson-Overington, 1993)</td>
<td>Structure-based sequence comparison</td>
</tr>
<tr>
<td>JON940101</td>
<td>The 250 PAM PET91 matrix (Jones et al., 1992)</td>
<td>Sequence comparison</td>
</tr>
<tr>
<td>JON940102</td>
<td>The 250 PAM transmembrane protein exchange matrix (Jones et al., 1994)</td>
<td>Sequence comparison</td>
</tr>
<tr>
<td>KOL5920101</td>
<td>Conformational similarity weight matrix (Kolaskar-Kulkarni-Kale, 1992)</td>
<td>Main-chain folding angles</td>
</tr>
<tr>
<td>LEV580101</td>
<td>The secondary structure similarity matrix (Levin et al., 1986)</td>
<td>Sequence comparison by secondary structure</td>
</tr>
<tr>
<td>LU5910101</td>
<td>Structure-based comparison table for outside other class (Lüthy et al., 1991)</td>
<td>Sequence comparison</td>
</tr>
<tr>
<td>LU5910102</td>
<td>Structure-based comparison table for inside other class (Lüthy et al., 1991)</td>
<td>Sequence comparison</td>
</tr>
<tr>
<td>LU5910103</td>
<td>Structure-based comparison table for inside other class (Lüthy et al., 1991)</td>
<td>Sequence comparison</td>
</tr>
<tr>
<td>LU5910104</td>
<td>Structure-based comparison table for inside other class (Lüthy et al., 1991)</td>
<td>Sequence comparison</td>
</tr>
<tr>
<td>LU5910105</td>
<td>Structure-based comparison table for outside beta class (Lüthy et al., 1991)</td>
<td>Sequence comparison</td>
</tr>
<tr>
<td>LU5910106</td>
<td>Structure-based comparison table for inside beta class (Lüthy et al., 1991)</td>
<td>Sequence comparison</td>
</tr>
<tr>
<td>LU5910107</td>
<td>Structure-based comparison table for other class (Lüthy et al., 1991)</td>
<td>Sequence comparison</td>
</tr>
<tr>
<td>LU5910108</td>
<td>Structure-based comparison table for alpha helix class (Lüthy et al., 1991)</td>
<td>Sequence comparison</td>
</tr>
<tr>
<td>LU5910109</td>
<td>Structure-based comparison table for beta strand class (Lüthy et al., 1991)</td>
<td>Sequence comparison</td>
</tr>
<tr>
<td>MLA710101</td>
<td>The similarity of pairs of amino acids (McLachlan, 1971)</td>
<td>Sequence comparison</td>
</tr>
<tr>
<td>MLA710102</td>
<td>The similarity of pairs of amino acids (McLachlan, 1971)</td>
<td>Sequence comparison</td>
</tr>
<tr>
<td>MLY5920101</td>
<td>Chemical similarity scores (McLachlan, 1972)</td>
<td>Chemical similarity</td>
</tr>
<tr>
<td>MIY5920101</td>
<td>Base-substitution-protein-stability matrix (Miyazawa-Jernigan, 1983)</td>
<td>Genetic code and contact potential</td>
</tr>
<tr>
<td>MIY5790101</td>
<td>Amino acid pair distance (Miyata et al., 1979)</td>
<td>Physical property indices</td>
</tr>
<tr>
<td>MOHR710101</td>
<td>EMPAR matrix (Mohana Rao, 1987)</td>
<td>Structural and physical property indices</td>
</tr>
<tr>
<td>NIE910101</td>
<td>Structure-derived correlation matrix 1 (Niefeld-Schomburg, 1991)</td>
<td>Main-chain folding angles</td>
</tr>
<tr>
<td>NIE910102</td>
<td>Structure-derived correlation matrix 2 (Niefeld-Schomburg, 1991)</td>
<td>Main-chain folding angles</td>
</tr>
<tr>
<td>OVE920101</td>
<td>STR matrix from structure-based alignments (Henikoff-Henikoff, 1992)*</td>
<td>Structure-based sequence comparison</td>
</tr>
<tr>
<td>QU 670102</td>
<td>Cross-correlation coefficients of preference factors (Qu et al., 1993)</td>
<td>Contacts of main chain atoms</td>
</tr>
<tr>
<td>QU 680103</td>
<td>The mutant distance based on spatial preference factors (Qu et al., 1993)</td>
<td>Contacts of side chain atoms</td>
</tr>
<tr>
<td>RISJ80101</td>
<td>Scoring matrix (Risler et al., 1988)</td>
<td>Main+side</td>
</tr>
<tr>
<td>TUD990101</td>
<td>Isomorphism of replacements (Tudós et al., 1990)</td>
<td>Neighbourhood selectivity</td>
</tr>
</tbody>
</table>

*The substitution data were obtained by Overington et al. (1992).
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 Genome Net database service at the following addresses:

<table>
<thead>
<tr>
<th>FTP</th>
<th>ftp.genome.ad.jp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gopher</td>
<td>gopher.genome.ad.jp</td>
</tr>
<tr>
<td>WWW</td>
<td><a href="http://www.genome.ad.jp/">http://www.genome.ad.jp/</a></td>
</tr>
</tbody>
</table>

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/aindex with the file name ainindex2. 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):

\[
d = 1 - c I
\]

where c is the correlation coefficient:

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

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...
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:

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

where

\[
\alpha = (1/Dp)^2, \quad \beta = (1/Dq)^2, \quad \gamma = (1/Dr)^2
\]

are the scaling factors which are calculated from the mean value \( D \) 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: \( \alpha \) and turn propensities, \( \beta \) 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: \( \alpha \) and turn propensities, \( \beta \) propensity, hydrophobicity and physicochemical properties. 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 \( \alpha \) subgroup and the turn subgroup within the \( \alpha \) and turn propensities region. It is now a member of the mostly turn propensity subgroup which also contains neural network propensity subgroup which also contains neural network propensities region. It is now a member of the mostly turn propensity subgroup which also contains neural network propensities.

The number of indices for the amino acid composition was ordered by \( 1 + J(1-J)/2 \) where \( IJ = \text{ARNDCEQGHILKMFPSTWYV} \).
A. α and turn propensities

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...
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 (LUTR910105,06,09) 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 Jermigan, 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 $\chi^2$ 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 (Cserzô et al., 1994, CSEM940101) is a refined version of the other (Tüdös et al., 1990, TUDE900101) based on their developed method of neighbourhood selectivity (Cserzô and Simon, 1989). We noted that the correlation coefficients reported in their subsequent
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 regard to the evolutionary distance among sequences being compared. The BLOSUM group contains three matrices, BLOSUM 45 (HEN920101), BLOSUM 62 (HEN920102) and BLOSUM 80 (HEN920103), 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 (\(= 402^2\)) possibilities and the best combination of three indices was searched from 10,746,800 (\(= 402^3\)) 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 to be conserved to retain the packing of the globule and the hydrophobicity needs to 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 (Mohanra 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
Amino acid indices and mutation matrices

Fig. 7. The absolute value of the correlation coefficient between the matrix constructed from the volume (P12) 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).

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 Hironuki 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 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


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