An Approach for a Substitution Matrix Based on Protein Blocks and Physicochemical Properties of Amino Acids through PCA

Youngki You[1], Inhwan Jang[1], Kyungro Lee[2], Heonjoo Kim[1], Kwanhee Lee[1,*]

[1]School of Life Science, Handong Global University, Pohang, Korea
[2]Department of Biotechnology Yonsei, University, Seoul, Korea

Subject areas; Bioinformatics/Computational biology/Molecular modeling

Author contribution; The research is mainly done by Y. Y., who did most of the work. H. Kim was in charge of statistical work.

***Correspondence** and requests for materials should be addressed to K.L. (khlee@handong.edu).

Editor; Keun Woo Lee, Gyeongsang National University, Korea

Received February 25, 2014;
Accepted August 29, 2014;
Published November 05, 2014

Citation; You, Y., et al. An Approach for a Substitution Matrix Based on Protein Blocks and Physicochemical Properties of Amino Acids through PCA. IBC 2014, 6:3, 1-10. doi: 10.4051/ibc. 2014.6.4.0003

Funding; The research was funded by Handong Global University.

Competing interest; All authors declare no financial or personal conflict that could inappropriately bias their experiments or writing.

SYNOPSIS

Amino acid substitution matrices are essential tools for protein sequence analysis, homology sequence search in protein databases and multiple sequence alignment. The PAM matrix was the first widely used amino acid substitution matrix. The BLOSUM series then succeeded the PAM matrix. Most substitution matrixes were developed by using the statistical frequency of substitution between each amino acid at blocks representing groups of protein families or related proteins. However, substitution of amino acids is based on the similarity of physiochemical properties of each amino acid. In this study, a new approach was used to obtain major physiochemical properties in multiple sequence alignment. Frequency of amino acid substitution in multiple sequence alignment database and selected attributes of amino acids in physiochemical properties database were merged. This merged data showed the major physiochemical properties through principle components analysis. Using factor analysis, these four principle components were interpreted as flexibility of electronic movement, polarity, negative charge and structural flexibility. Applying these four components, BAPS was constructed and validated for accuracy. When comparing receiver operated characteristic (ROC_{50}) values, BAPS scored slightly lower than BLOSUM and PAM. However, when evaluating for accuracy by comparing results from multiple sequence alignment with the structural alignment results of two test data sets with known three-dimensional structure in the homologous structure alignment database, the result of the test for BAPS was comparatively equivalent or better than results for prior matrices including PAM, Gonnet, Identity and Genetic code matrix.

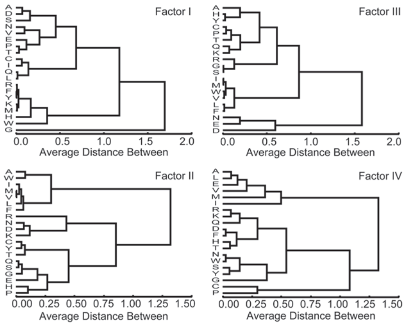

FIG. 2.

Key Words: BAPS; factor analysis; principle component analysis; scoring matrix; sequence alignment

INTRODUCTION

Protein sequence alignment is a method used to find homology among different proteins and is consequently a powerful tool for predicting protein structure and function. Homology was based on specific groups according to specific physiochemical properties of amino acids. Substitution matrices, which describe the relationship between each amino acid quantitatively, were studied to improve accuracy[1-6].

PAM and BLOSUM matrices are commonly used alignment programs. PAM matrix is a mutation probability matrix based on evolutionary distance[1,7]. BLOSUM series was constructed using the substitution frequency between amino acids from PROSITE, which contains information concerning conserved regions between related proteins in a protein family[8]. PAM series, BLOSUM series and most substitution matrices use only statistical information from the sequence alignment database. However, this statistical information is limited to sequence alignment data and can be further improved using structural information. When comparing sequences, a matrix based on structural information is essential for accurate and efficient predictions concerning protein structure and function[9,10].

Structure based matrices were developed to satisfy this requirement[11,12]. However, these matrices were constrained by structural data set size. The structure block matrix was constructed in an attempt to overcome this problem[13]. Although, this matrix considered even more properties involved in protein conformation, the lack of data sets in the structure database in comparison to data sets from the sequence database continued to be a reoccurring problem. New methods using substitution matrices were studied[14-18].

Table 1. Selected attribute of amino acid to the structure and function

Accession number	Data description
BEGF750101	Conformational parameter of inner helix (Beghin-Dirkx, 1975)
BEGF750102	Conformational parameter of beta-structure (Beghin-Dirkx, 1975)
BEGF750103	Conformational parameter of beta-turn (Beghin-Dirkx, 1975)
BHAR880101	Average flexibility indices (Bhaskaran-Ponnuswamy, 1988)
CHAM810101	Steric parameter (Charton, 1981)
CHAM820101	Polarizability parameter (Charton-Charton, 1982)
CHAM830107	A parameter of charge transfer capability (Charton-Charton, 1983)
CHAM830108	A parameter of charge transfer donor capability (Charton-Charton, 1983)
CHOC760102	Residue accessible surface area in folded protein (Chothia, 1976)
DAWD720101	Size (Dawson, 1972)
DAYM780201	Relative mutability (Dayhoff et al., 1978b)
EISD860101	Solvation free energy (Eisenberg-McLachlan, 1986)
FASG760101	Molecular weight (Fasman, 1976)
FASG760103	Optical rotation (Fasman, 1976)
FAUJ880108	Localized electrical effect (Fauchere et al., 1988)
FAUJ880111	Positive charge (Fauchere et al., 1988)
FAUJ880112	Negative charge (Fauchere et al., 1988)
HUTJ700101	Heat capacity (Hutchens, 1970)
HUTJ700102	Absolute entropy (Hutchens, 1970)
PRAM900101	Hydrophobicity (Prabhakaran, 1990)

This study fused the substitution frequency between amino acids from PROSITE and the data of physiochemical properties of amino acids. The fused resulting data reflects key amino acid physicochemical properties to align protein sequences. The number of dimensions decreased through principle component analysis. And the resulting principle components were applied to make a substitution matrix. The new substitution matrix, Block based Amino acid Physiochemical properties Substitution matrix (BAPS), shows the utility of this study. Several major potential physiochemical properties are selected to make the matrix. BAPS was then compared with seven other matrices by receiver operated characteristic (ROC_{50}) values and structural alignment tests.

RESULTS AND DISCUSSION

Factor analysis with principle component analysis

The transform value of the 20 amino acid physiochemical properties (Table 1) for the 387,240 columns was determined by factor analysis. The transform value signifies the influence of the 20 physiochemical properties on each column. Using principle component analysis (PCA), four out of 20 resulting factors were selected according to eigenvalue and scree plot in Figure 1. Factors preceding drastic slope changes on the scree plot and with eigenvalues greater than 1 were chosen. As shown in Table 2, factor loading scores for the four factors were then rotated by varimax. The four factors consist of a set of unique properties based on high absolute loading scores.

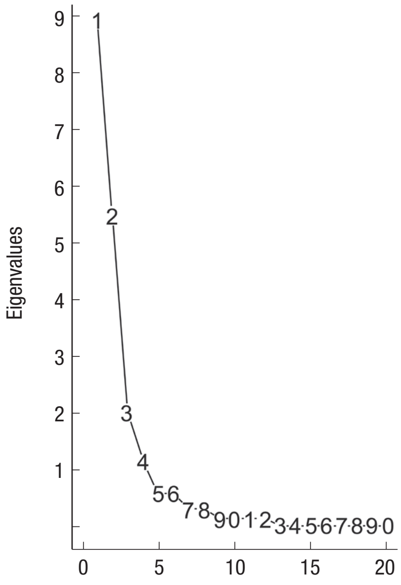

Figure 1. Scree plot of principle component analysis.

Factor I has high factor loading scores in size, absolute entropy, polarizability parameter, and molecular weight. High scores in size and molecular weight indicate association between Factor I and volume. Factor I can also be expected to relate to electron motion within a molecule due to high factor load scores in absolute entropy and polarizability parameter. Accordingly, Factor I represents flexibility of electron movement in relation to volume.

Factor II signifies polarity. Factor II had high absolute scores in residue accessible surface area in folded protein, solvation free energy, conformational parameter of beta-structure, average flexibility indices, positive charge, and negative charge. Residue accessible surface area in folded protein score is higher in value when an amino acid has polar features. Polar amino acids have lower solvation energy values because solvation energy is inversely proportional to polarity. In addition, lower values of hydrophobicity based on the AAindex describe amino acids with hydrophobic characteristics. Polar amino acids also have relatively lower values in conformation parameter of beta-structure due to the inverse relationship between polarity and this parameter.

Factor III relates to negative charge. Negative charge, localized electrical effect, and a parameter of charge transfer capability are attributes with high factor loading scores for Factor III. These properties are strengthened in the abundance of electrons. Although not to the same degree, Factor III also shares attributes with Factor II.

Table 2. The VARIMAX rotated factor loading score

	Factor I	Factor II	Factor III	Factor IV	Com.
A parameter of charge transfer donor capability	0.790	0.345	0.082	-0.259	0.816
Heat capacity	0.814	-0.449	-0.112	0.032	0.877
Size	0.909	0.031	-0.223	0.145	0.899
Absolute entropy	0.948	0.134	-0.139	0.130	0.953
Polarizability parameter	0.954	-0.094	-0.190	-0.006	0.955
Molecular weight	0.970	0.062	0.071	-0.027	0.950
Hydrophobicity	0.161	0.834	0.502	0.096	0.983
Solvation free energy	0.209	-0.837	-0.488	-0.035	0.983
Residue accessible surface area in folded protein	0.233	0.880	0.338	-0.011	0.943
Conformational parameter of beta-structure	0.300	-0.743	-0.469	0.241	0.919
Average flexibility indices	-0.412	0.741	0.407	0.019	0.885
Positive charge	0.418	0.831	-0.061	0.111	0.882
Negative charge	0.024	0.232	0.870	0.176	0.843
Localized electrical effect	-0.038	0.323	0.792	-0.281	0.811
A parameter of charge transfer capability	-0.305	0.377	0.740	-0.047	**0.785**
Relative mutability	-0.466	0.236	0.554	0.320	**0.682**
Steric parameter	0.655	-0.473	-0.055	0.477	0.883
Conformational parameter of beta-turn	-0.083	0.628	0.430	-0.587	0.932
Optical rotation	-0.072	0.270	0.228	0.807	**0.781**
Conformational parameter of inner helix	0.234	-0.560	-0.362	0.635	0.902

Twenty properties with communality (Com.) estimates < 0.8 shown in bold. The interpretation of each factor is given in the text.

Factor IV reflects structural flexibility. Optical rotation, conformational parameter of inner helix, conformational parameter of beta-turn and steric parameter are characteristics directly related to secondary and tertiary structure. Scores for optical rotation and conformational parameter of inner helix are high, and values for conformational parameter of beta-turn (inversely proportion to factor IV) are low, when a structure is flexible.

These four factors are key properties for substitution between amino acids in multiple sequence alignment.

Amino acid factor score

Communality value which is the sum of the squared factor loading score for each physiochemical property represents the variance in a property described by the four factors. As shown in Table 2, most properties had high communality values (> 0.8) indicating that properties had high factor loading scores of at least one factor, and that the four factor model is sufficient. However, among the twenty traits, relative mutability had a low communality value that did not fit our model, and had a large unique component of variability.

As reported in Table 3, the amino acid factor score was calculated using the physiochemical property score and rotated factor loading score. Numerically, the amino acid factor score shows the similarity of each amino acid according to each factor. Unweighted pair group method with arithmetic mean cluster analysis (UPGMA) elucidate the distance between amino acids at each factor (Figure 2). The distance between amino ac-

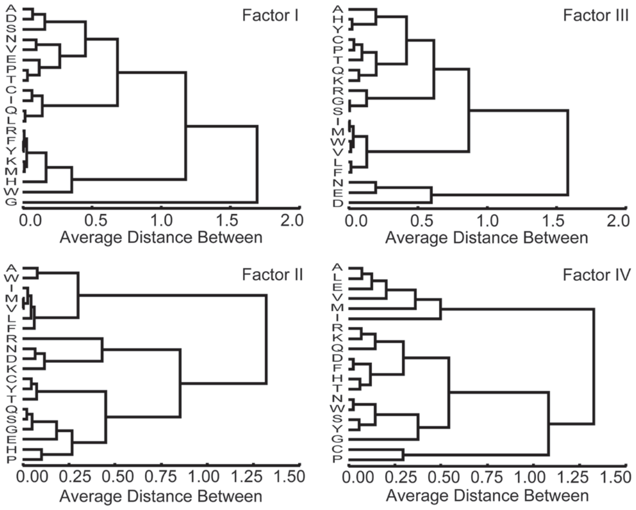

Figure 2. Unweighted pair group method with arithmetic mean cluster analysis of distances computed from the amino acid solution scores from Factors I–IV.

ids at each of the four factors shows various trends. For example, Gly at Factor I has the longest distance to all of the other amino acids, but Pro and Cys are separated by the longest distance at Factor IV. William R. A obtained five factors in his study by factor analysis[19]. Although the interpretation of the five factors is similar to the results of this study, the order of the factors is not the same. Furthermore, the distance between amino acids is different among similarly defined factors. Because all the properties found in the database were not considered in this study, interpretation of aforementioned factors was similar but there are differences in distances. However, Pro was found to be far apart from the other amino acid in both researches.

Construction of substitution matrix

Relative distance between the amino acids was measured by node count between two amino acids resulting from dendrogram of cluster analysis. Short distances in the dendrogram are indistinguishable from each other due to exceptionally long distances. However, node distance clearly represents differences among short distances. The distance from Glu to Pro and the distance from Pro to Thr are the same after normalization because Gly has a very long distance to the other amino acids in Factor I. Using node distances, we can clearly distinguish distances between close amino acids. The distance from Gly to the other amino acids in Factor I is two nodes. However, the node distance from most amino acids to Gly is around seven nodes. In order to use a substitution matrix, these node distances were transformed into effective node distances. Effective node distance matrices of each Factor were calculated with Eq. (3.2). Then the BAPS matrix was constructed using the aforementioned method. BAPS0707 and BAPS0708, among versions ranging from widths of 6 to 25, were optimal in sequence alignment.

Comparison of BAPS with BLOSUM and PAM

BAPS0707 and BAPS0708 were compared with commonly used BLOSUM62 and PAM250 in Figure 3, 4. There are pointed differences between the matrices. Almost all of the amino acid self-pairs have scores more positive in BAPS than in BLOSUM62 and PAM250. Only Trp in BLOSUM62 and Cys, Trp in PAM250 have points more positive than points in BAPS. Furthermore, many amino acid matches have scores more positive in BAPS series than in the other matrices. For example, the scores for elements QC, TP, PC in BLOSUM62, and QC, FD, VW, PC in

Table 3. Four factor solution scores for the 20 selected amino acid attributes

Amino acid	Factor I	Factor II	Factor III	Factor IV	Amino acid	Factor I	Factor II	Factor III	Factor IV
Ala	-4.053	-4.628	-3.396	2.039	Leu	0.717	-6.047	-4.792	1.902
Arg	3.467	4.626	-0.122	0.673	Lys	3.551	2.959	-1.129	0.801
Asn	-1.478	2.272	2.076	-0.562	Met	3.599	-5.896	-4.243	1.331
Asp	-3.716	2.579	4.684	0.202	Phe	3.438	-5.632	-4.744	0.243
Cys	0.220	-1.836	-1.520	-1.730	Pro	-2.800	-0.579	-1.652	-2.287
Gln	0.790	0.131	-0.852	0.475	Ser	-4.660	0.032	0.371	-0.369
Glu	-2.141	1.010	2.812	1.747	Thr	-2.649	-2.281	-1.937	0.065
Gly	-7.666	0.307	0.358	-1.148	Trp	5.385	-4.267	-4.376	-0.608
His	4.339	-1.055	-2.623	-0.044	Tyr	3.442	-2.048	-2.701	-0.268
Ile	-0.100	-5.774	-4.229	2.743	Val	-1.009	-5.891	-4.320	2.261

```
      A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
      6 -2  0  1 -1 -1  1 -3  1  1  3 -1  1  0  0 -2 -1  2  1  1 A
A  8     4 -2  1  1 -1 -3  1  1 -1 -2  1 -1  3  0  0  0  1  3  0 R
R -1  3     4  2  1 -2  1 -2 -3  0  0 -1 -1  1  1 -1 -1  4  1  2 N
N -2 -2  8     3  1 -1 -2 -2  0  0  1  1  0  3 -1 -1  1  2  1  1 D
D -1  0  1  5     0  4  2  1  2  0 -1  2 -2 -1  6  0  3 -1  3 -1 C
C  1  2  2  3 -3     5 -3  2  1  1  1  1 -3  1  2  1  3  0  1 -1 Q
Q -2 -1 -3 -3  6  6     5  0 -2  2  3 -4  1  0  2  0  1  0  0  3 E
E  0 -2  0 -3  3 -3  6     2  2  0  0  0  0  1  1  3  1  0  3 -1 G
G -4  2 -2 -4  1  1 -2  3     2  0  0  2  1  2  3  0  4  1  1  0 H
H  0 -1 -4 -2  2 -2 -3  2  4     5  1  0  3  0  1 -1 -2  3 -2  0 I
I  1 -2 -1 -1  1  0  1 -1 -1  4     6 -1  0  1  1 -1 -2  1 -2  2 L
L  4 -1  0  1  4  1  3  0 -1  1  4     5  0  3  0 -2  2  2  3 -1 K
K -1  0 -2  0  4  1 -3  0  1 -1  0  5     5  2 -1 -2 -2  2  0  3 M
M  1 -2 -1  0  2 -2  1  0  1  2 -2 -1  4     4  1  0  1  0 -2  1 F
F  1  4  1  6  1  3  2  3  3 -1 -1  5  2  1     3  1  4  1  2  1 P
P -2 -2 -1 -1  6  1  2 -1  1  0  1  0 -1  2  4     6 -1  2  3  0 S
S -2 -1 -1 -1 -1  2  0  2  0 -2  0 -2 -1  1 -1  8     5  0  3 -1 T
T -2  0 -1  0  4  3  0 -1  3 -3 -1  1 -2  2  3 -1  7    -2 -2  3 W
W  5 -4  4  5  5  3  4  5  2  5  1  2  5  1  3  1  3 -8     3 -2 Y
Y  2  5  1  2  1  4  2  5  3 -2 -2  5  1 -6  4  4  4  0  0     6 V
V  1 -1  1  0  0 -1  3 -3 -1 -1  1 -1  2  1  0 -1 -1  6 -1  6
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
```

Figure 3. The difference matrix (Upper) obtained by subtracting the BLOUSM62 from BPAS0707 entry by entry. The difference matrix (Lower) obtained by subtracting the PAM250 from BPAS0707 entry by entry.

```
      A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
      7 -2  0  1 -1  0  2 -3  1  1  3 -1  1  0  0 -1 -1  2  1  2 A
A  9     5 -2  1  1 -1 -3  2  1 -1 -2  2  0  4  0  1  0  2  3  0 R
R -1  4     5  2  1 -2  2 -2 -3  0  0 -1 -1  1  1 -1  0  4  1  3 N
N -2 -2  9     4  1 -1 -2 -2  0  0  1  1  0  3  0  0  1  2  1  1 D
D -1  0  1  6     1  4  2  1  3  0  0  2 -2 -1  6  0  3 -1  3 -1 C
C  1  2  2  3 -2     6 -3  2  1  1  1  1 -3  2  2  2  3  0  1 -1 Q
Q -1 -1 -3 -3  6  7     6  0 -1  2  3 -4  1  0  2  0  1  0  0  4 E
E  1 -2  1 -3  3 -3  7     3  2  0  0  1  0  1  1  3  1  1  3 -1 G
G -4  3 -2 -4  1  1 -2  4     3  0  0  3  1  2  3  1  4  1  1  0 H
H  0 -1 -4 -2  3 -2 -2  2  5     6  2  0  3  0  1 -1 -2  3 -2  1 I
I  1 -2 -1 -1  1  0  1 -1 -1  5     7 -1  1  1  1 -1 -2  1 -2  3 L
L  4 -1  0  1  5  1  3  0 -1  2  5     6  1  3  1 -2  2  2  3 -1 K
K -1  1 -2  0  4  1 -3  1  2 -1  0  6     6  2 -1 -2 -2  2  0  3 M
M  1 -1 -1  0  2 -2  1  0  1  2 -1  0  5     5  2  0  1  1 -2  1 F
F  1  5  1  6  1  4  2  3  3 -1 -1  5  2  2     4  2  5  2  3  1 P
P -2 -2 -1  0  6  1  2 -1  1  0  1  1 -1  3  5     7 -1  2  3  0 S
S -1  0 -1  0 -1  3  0  2  1 -2  0 -2 -1  1  0  9     6  0  3 -1 T
T -2  0  0  0  4  3  0 -1  3 -3 -1  1 -2  2  4 -1  8    -1 -1  4 W
W  5 -3  4  5  5  3  4  6  2  5  1  2  5  2  4  1  3 -7     4 -2 Y
Y  2  5  1  2  1  4  2  5  3 -2 -2  5  1 -6  5  4  4  1  1     7 V
V  2 -1  2  0  0 -1  4 -3 -1  0  2 -1  2  1  0 -1 -1  7 -1  7
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
```

Figure 4. The difference matrix (Upper) obtained by subtracting the BLOUSM62 from BPAS0708 entry by entry. The difference matrix (Lower) obtained by subtracting the PAM250 from BPAS0708 entry by entry.

PAM250 are four bits and six bits more negative than in BAPS matrices. On the other hand, KE in BLOSUM62, and YF in PAM250 have scores six bits more negative than in BAPS matrices. Substitution between Pro and Cys was more tolerable in BPAS series than in BLOSUM62 and PAM250. Although the majority of scores are more positive in BAPS than in other matrices, as shown in Figure 5, these scores are located in the central portion of the distribution of elements. Thus, higher scores in BAPS indicate relevant and important highly conserved regions. On the other hand, low scores signify substitutions that disrupt conformation and structure. The correlation between BAPS and BLOSUM62 was 0.84 and the correlation between BAPS and PAM was 0.70. These values indicate high correlation.

Validation of new substitution matrices

Detection of homologous pairs

Homologous pairs were searched with 103 yeast queries in 6,433 yeast sequences. As shown in Table 4, the results of this search showed that BAPS matrices were sufficiently better than the Genetic code and Identity matrix. However, the BAPS matrices had a slightly lower ROC_{50} value than BLOSUM and PAM matrices. Although in this study the PAM70 ROC_{50} value was a little better than the PAM250 value, this result was contrary to the results found in the Weija study[15], which used their own database search program. This means that the ROC_{50} value depends on the specific database search program. Therefore, the gap between BAPS and PAM is not significant, and this suggests similar efficiency.

Table 4. Comparison of average ROC_{50} value

Alignment Method	Average ROC_{50}				
Matrix name	BLOSUM62	BLOSUM45	GONNET	PAM70	PAM250
Striped	0.8382	0.8329	0.8350	0.8064	0.7986
BLAST	0.8524	0.8603	-	0.7842	0.7830
Matrix name	Genetic code	Identity	BAPS0707	BAPS0708	
Striped	0.5223	0.5914	0.7673	0.7587	
BLAST	-	-	-	-	

Striped (Smith-Waterman implementations) program was used to compare average ROC_{50} values. BLAST program results similar to Striped results. BLSOUSM62 was better than BLSOUM45 using Striped, but the opposite result can be when using BLAST. BLOSUM62, BLOSUM45, PAM70, and PAM250 were tested with BLAST using the BLAST program option and were not modified with median zero.

Comparison of structural alignment and sequence alignment

When the substitution matrix is sensitive to structural conformation components, we can expect sequence alignment results to be closer to structural alignment results. BAPS matrices were compared with the structural alignment data from homologous structure alignment database (HOMSTRAD) for further validation. Average ratio of miss counts in a set was measured with 88 data sets (Figure 6). ClustalW2 and Muscle, commonly used to align sequences, were used to align test set sequences. ClustalW2[20] is a popular MSA tool that is easy to use and is suitable for medium alignments[21]. This program was recently upgraded to version 2.1[22]. Muscle[23] on the other hand is known for fast and accurate results[21]. Both programs allow the utilization of user matrices. In ClustalW2, BAPS matrices have a better or similar accuracy in comparison to other matrices excluding BLOSUM62. BAPS matrices, having fewer miss counts, were especially more accurate than PAM250 (Figure 6A). PAM250 is the substitution matrix mainly used within the PAM series[24]. However, when using Muscle, the accuracy of the PAM series was slightly greater than the BAPS matrices (Figure 6B).

Miss percentage is dependent on the length of the consensus region in each test set. Therefore, as shown in Table 5, BAPS matrices were correctly evaluated by aligning sequences against other test sets with consensus columns above 100. BLOSUM series obtained positive results from both programs.

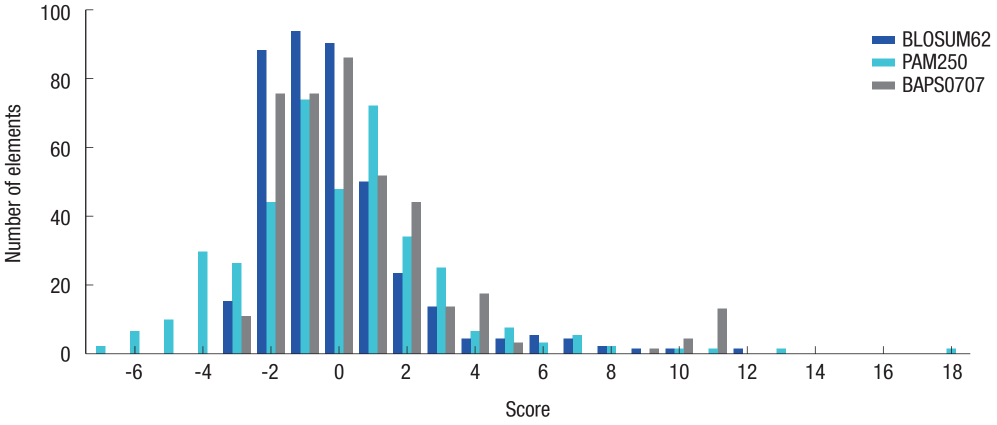

Figure 5. Distribution of elements in BLOSUM62, PAM250, and BAPS0707 matrices. The median for the three matrices was zero. Scores ranged from -3 to 12 in BLOSUM62, -7 to 18 in PAM250, and -3 to 11 in BAPS0707.

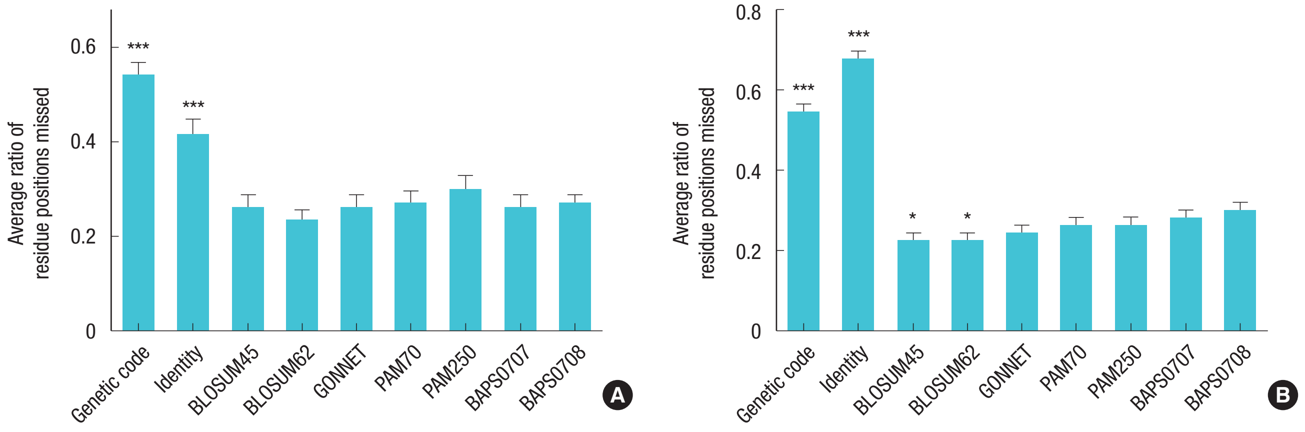

Figure 6. Average miss ratio when comparing with structural alignment. (A) Comparing matrices using ClustalW2.1 (B) Comparing matrices using Muscle. Data represents the mean ± SEM of 88 independent experiments. Asterisks denote statistically significant differences ($P < 0.05$) between BAPS0707 and each of the other matrices.

Table 5. Average miss ratio when comparing structural alignment

Matrix name	Genetic code	Identity	BLOSUM45	BLOSUM62	GONNET
ClustalW2.1	0.58	0.45	0.27	0.24	0.30
MUSCLE	0.58	0.74	0.24	0.24	0.25
Matrix name	PAM70	PAM250	BAPS0707	BAPS0708	
ClustalW2.1	0.31	0.34	0.29	0.29	
MUSCLE	0.28	0.28	0.32	0.33	

The test set contained families selected from the HOSTRAD database with number of structures greater than 2, average length above 50, average identity ranging from 25 to 50, and consensus regions above 100. The test set is composed of 89 families and the total number of consensus regions is 14,410.

In Muscle, PAM and BLOSUM series had a lower average miss ratio than BAPS series. However, BAPS series had better accuracy in comparison to the Gonnet matrix, PAM series, genetic code matrix and identity matrix in ClustalW2. Gonnet matrix, the default matrix in ClustalW2, had a higher miss ratio than in BAPS series.

Comparing the structural alignment validations produced different results according to the alignment program used (Figure 6 and Table 5). BAPS had a miss ratio higher in ClustalW2 than in Muscle. ClustalW2 and Muscle differ in default gab-open and gab-extension penalty values. The default gab-open value is -10, and gab-extension penalty is -0.20 in ClustalW2. Muscle, on the other hand, uses the profile scoring function (log-expectation score is defaulted) for gab-open and gab-extension penalty. After the sequence alignment tools were made and developed, the programs were evaluated by using general scoring matrices like BLOSUM or PAM series. Therefore, many kinds of option values are suitable for these matrices. As accuracy improves according to the alignment program, BLOSUM and PAM have more of advantages. To evaluate these matrices, impartial alignment programs may be needed in the future.

In the ROC$_{50}$ test, BAPS scored lower than the PAM and BLOSUM series. However, comparing the result of the structural alignment test BAPS scored similarly or better than the PAM series. This means that the BLOSUM series and PAM series, which uses statistical information, is better suited for sequence homology search in the database, but is lacking when comparing protein structure. For this reason, the BAPS had a lower score in the ROC$_{50}$ test, but since BAPS is based on the physiochemical properties of amino acid, it is more appropriate for protein structure comparison. Thus there may be a need to modify the method of James O.[25] to improve the accuracy of sequence homology search.

CONCLUSION AND PROSPECTS

A novel statistical approach was developed that rearranged amino acid relationships using PCA frequency information from the BLOCKS database[25]. New principle components were interpreted and the similarities in amino acids were measured. It was a new approach that applied PCA but only part of the results interpreted. Another method that proposed a solution for the metric problem concerning sequence alignment was introduced later[19]. New attribute physicochemical properties factors were defined with factor analysis using the physicochemical properties of amino acids from the AAindex database. Then, similarities in factor scores for each amino acid were analyzed. Though this research used factor analysis to find new attributes, it was not directly applied to sequence analysis.

Our research defines similarities between amino acids in multiple sequence alignment by adopting the statistical methods used in the previous study. These similarities were then applied to the developing substitution matrix. The physiochemical properties, Factor I, II, III, and IV, greatly influence the result of sequence alignment. The matrices, made by using these relationships between amino acids, not only have statistical information on substitution but also represent each amino acid's physiochemical properties. The BAPS series contain information that is different than the most widely used BLOSUM series and PAM series. In future studies, the chosen key attributes need to be redefined in more detail using other selected properties and a new filtering method of effective counts from blocks.

Our study introduces an approach to extract key amino acid properties for sequence alignment by merging multiple sequence alignment databases with physiochemical properties database. However, due to novel techniques used in this study, further improvements can be made in our filtering methods. In the future, we expect to fine tune BAPS and use it in junction with sequence alignment programs.

MATERIALS AND METHODS

Effective Amino Acid Type Frequencies from the BLOCKS Database

There are 774,565 multiple sequence alignment columns in the BLOCKS database (Version 14.3, April 2007)[26]. Effective counts of amino acid type in the columns of each family were estimated using the method applied from independent counts[25]. Briefly, the Effective count, $N_{eff}(c, a)$, in each family is calculated according to Eq. (1.1), with column position and amino acid type represent by the letter *c* and *a*, respectively. $N_{type}(c, a)$ is the average amino acid number in a position within a block when the amino acid *a* is in position *c* in the block within a family.

$$N_{eff}(c, a) = \frac{ln\left(1.0 - \dfrac{N_{type}(c, a)}{20}\right)}{ln\,(0.95)} \qquad (1.1)$$

Effective frequencies $f_{c,\,a}$ of each amino acid type a at the column position c is calculated by Eq. (1.2) [25].

$$f_c^a = \frac{N_{eff}(c, a)}{\sum_{a=1}^{20} N_{eff}(c, a)} \qquad (1.2)$$

Selection of meaningful aligned columns

Significant columns were selected from the 774,565 columns using methods stated in the previous article[25]. The 774,565 columns were sorted by total N_{eff}, and then divided into 70 groups each with 11,064 columns (70th group has 11,149 columns). Two graphs were created in order to access the characteristics of each column group: the plot$_{hydro}$ graph was constructed by plotting mean hydrophobicity values versus mean total N_{eff} values and the plo$_{tsize}$ graph was constructed by plotting mean size values versus mean total N_{eff} values. These plots were divided objectively into three regions. Divisions were based on three least-squared linear fits on all points. The columns in the thickest region of the plot have high total N_{eff} values. This signifies that these columns have a varied array of amino acids and are subsequently less conserved. On the other hand, columns with low N_{eff} values, found in the thinnest region, lack amino acid diversity and are well conserved. Data selected by significant information concerning amino acid substitution, was the basis for continued analysis. Low N_{eff} columns were excluded because these columns are well conserved yielding minimal amino acid substitution information. Additionally, the high N_{eff} columns were excluded because substitution information is vague and can contribute to noise that can hinder future analysis due to varying amino acid content. The column groups residing in the common middle region of both the plot$_{hydro}$ and plot$_{size}$ graphs were selected and analyzed. There are 387,240 columns in the 35 groups, and total N_{eff} values ranged from 20.49 to 439.36.

Selection and standardization of amino acid properties for protein structure

The amino acid index is a set of 20 numerical values that represent distinct physiochemical and biochemical properties of an amino acid[27,28]. The properties for this study were selected from the 544 amino acid indices indicated in the online database (AAindex, www.genome.jp/aaindex). The online database contains amino acid indices describing general properties including hydrophobicity, size and polarity, as well as specific measured properties such as average side chain orientation angle. For this study, as shown in Table 1, twenty properties pertaining to protein conformation and structure were selected from the aforementioned database. In order to understand the influence of each amino acid on secondary structure formation and protein structure, the conformational parameters for inner helix, beta-structure and beta-turn were selected. Average flexibility and optical rotation were chosen because of their effects on protein folding. Indices related to energy and structure stability such as heat capacity, absolute entropy, solvation free energy and localized electrical effect were selected. Other properties that effect protein structure conformation including steric parameter, polarizability, charge transfer capability, charge transfer donor capability, residue accessible surface area in folded protein[29], were also selected[30-33]. Additionally, basic amino acid characteristics like size, molecular weight, positive charge and negative charge were selected[34]. Finally, relative mutability, values concerning substitution possibility, was chosen. Selected properties were standardized due to the different units corresponding to each property with min-max normalization taking into consideration transform values from Eq. (2.1). The minimum and maximum values were 0.05 and 1 respectively, in order to create an effective frequency range.

Transfer of amino acid frequency to physiochemical property form

The physiochemical properties that influence substitution of amino acids were accessed through calculations using Eq. (2.1). Transform value, *T(c, i)*, was calculated by the effective frequencies $f_{c,\,a}$ of each amino acid in column and the selected physiochemical properties $p_{i,\,a}$.

$$T_i^a = \sum_{a=1}^{20} f_c^a \, p_i^a \qquad (2.1)$$

When substitution of an amino acid in a column occurred, transform values were either high or low depending on the selected twenty amino acid properties. The transform score reflects key characteristics of a particular amino acid substitution in a multiple sequence alignment.

Factor analysis with principle components analysis

Factor analysis was used to make a subset that describes numerically the multitude of variables filtered by the columns[35]. Meaningful relationships, amidst numerous data, can be determined by comparing factors with fewer dimensions. Factors were calculated using PCA. The principle components were derived from the eigenvector of a 20 by 20 correlation matrix calculated according to the transform value of each column and its selected properties. SAS program (version 9.1) was utilized to operate factor analysis with PCA. Factor loading scores were obtained using factor analysis results rotated by varimax. Each factor was interoperated based on factor loading scores.

Construction of substitution matrix with amino acid factor scores

The amino acid factor score, representing the similarity of each amino acid according to each factor, *S(a,i)*, was calculated with Eq. (3.1). Physiochemical property score, *P(j,a)*, and the rotated factor loading score of each property, were used to calculate *S(a,i)*. Physiochemical property scores were normalized by min-max factor loading scores ranging from -1 to +1.

$$S_i^a = \sum_{j=1}^{n} P_a^j F_i^j \qquad (3.1)$$

A dendrogram was constructed with amino acid factor scores using UPGMA. The number of nodes, node distances D_n, was counted from one amino acid to either another amino acid or a group containing another amino acid. Then, the effective node distance, D_{eff}, for each amino acid was calculated with Eq. (3.2). $\tilde{D}_n$ denotes the median of *i* or *j* row.

$$D_{\mathrm{eff}}(i, j) = -\frac{1}{2}\left(\frac{D_n(i, j)}{\tilde{D}_n(i)} + \frac{D_n(j, i)}{\tilde{D}_n(j)}\right) \qquad (3.2)$$

Four effective node distance matrices of each factor were averaged and normalized with min-max normalization and rounded to an integer scale to finally produce our BAPS matrix.

Matrix evaluation

Homologues sequence search

BAPS matrices were compared with other matrices by local alignment for validation. The medians of matrices were subtracted from all elements in each matrix to compare the matrices[15]. The test was progressed by using Cygwin (version 2.320.2.2) with yeast database containing 6,433 protein sequences. Query sequence set has 103 sequences, and their true positives were identified and used by experts from the previous studies[15,36]. Query set that includes true positive and yeast database were downloaded from ftp.ncbi.nlm.nih.gov/pub/impala/blast test which was updated on April, 2004.

The homologous sequences of the 103 queries were searched from the yeast database by using Striped Smith–Waterman local alignment[37] with the following matrices: genetic code matrix, identity matrix, PAM70, PAM250, BLOSUM45, BLOSUM62, BAPS0707, and BAPS0708. After the local alignment scores of each query were calculated, ROC score was computed to compare the accuracy[38]. ROC_{50} value was calculated with Eq. (4.1) according to the true positive hits lists from alignment score sorted in descending order[15]. t_i is the number of the true positive that comes out until the *i*th false positive.

$$ROC_n = \frac{1}{n\mathrm{T}} \sum_{i=0}^{n} t_i \qquad (4.1)$$

Structural alignment comparison

Sequence alignment results were compared with the alignment result of three-dimensional structure to validate the sequence alignment accuracy. Sequence alignment test compares multiple alignments of known structures with sequence alignment results. This test was used by Lipman *et al.* to validate the multiple alignment programs, MSA, with three serine proteases which have a known three-dimensional structure[39]. BLOSUM matrix was also tested with serine proteases[8].

Two structural alignment test sets were brought from the HOMSTRAD which contains protein structure alignments of homologous families (http://tardis.nibio.go.jp/ homstrad/)[40]. One test set was selected using the following conditions: average length greater than fifty, number of sequences from three to eight, and average identity ranging from 31% to 40%. The other test set had different condition; number of sequences greater than two, average identity ranging from 25% to 50%, and length of consensus region greater than 100. The consensus secondary structure was labeled on the structural alignment for each family of HOMSTRAD. Unmatched columns in the consensus region were counted by comparing sequence and structure alignment[40]. The matrices evaluated by the ROC test were compared. Two multiple sequence alignment, ClustalW2 (version 2.520.2.2)[20,22] and Muscle[23], programs that can use user matrices were operated for validation.

REFERENCES

1. Dayhoff, M. O. E. R. V. N. B. R. F. (1968). Atlas of protein sequence and structure. Silver Spring, Md.: National Biomedical Research Foundation.
2. McLachlan, A. D. (1971). Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c 551. *Journal of molecular biology* 61, 409-424.
3. Feng, D. F., Johnson, M. S., and Doolittle, R. F. (1985). Aligning amino acid sequences: comparison of commonly used methods. *Journal of molecular evolution* 21, 112-125.
4. Mohana Rao, J. K. (1987). New scoring matrix for amino acid residue exchanges based on residue characteristic physical parameters. *International journal of peptide and protein research* 29, 276-281.
5. Risler, J. L., Delorme, M. O., Delacroix, H., and Henaut, A. (1988). Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. *Journal of molecular biology* 204, 1019-1029.
6. Smith, R. F., and Smith, T. F. (1990). Automatic generation of primary sequence patterns from sets of related protein sequences. *Proceedings of the National Academy of Sciences of the United States of America* 87, 118-122.
7. Dayhoff, M. O. N. B. R. F. (1978). Atlas of protein sequence and structure. Washington, D.C.: National Biomedical Research Foundation.
8. Henikoff, S., and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America* 89, 10915-10919.
9. Bowie, J., Luthy, R., and Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253, 164-170.
10. Liu, X., Zhang, L. M., Guan, S., and Zheng, W. M. (2003). Distances and classification of amino acids for different protein secondary structures. *Physical review E, Statistical, nonlinear, and soft matter physics* 67, 051927.
11. Johnson, M. S., and Overington, J. P. (1993). A structural basis for sequence comparisons. An evaluation of scoring methodologies. *Journal of molecular biology* 233, 716-738.
12. Prlic, A., Domingues, F. S., and Sippl, M. J. (2000). Structure-derived substitution matrices for alignment of distantly related sequences. *Protein engineering* 13, 545-550.
13. Liu, X., and Zheng, W. M. (2006). An amino acid substitution matrix for protein conformation identification. *Journal of bioinformatics and computational biology* 4, 769-782.
14. Teodorescu, O., Galor, T., Pillardy, J., and Elber, R. (2004). Enriching the sequence substitution matrix by structural information. *Proteins* 54, 41-48.
15. Xu, W., and Miranker, D. P. (2004). A metric model of amino acid substitution. *Bioinformatics* 20, 1214-1221.
16. Eyal, E., Frenkel-Morgenstern, M., Sobolev, V., and Pietrokovski, S. (2007). A pair-to-pair amino acids substitution matrix and its applications for protein structure prediction. *Proteins* 67, 142-153.
17. Liu, X., and Zhao, Y. P. (2010). Substitution matrices of residue triplets derived from protein blocks. *Journal of computational biology: a journal of computational molecular cell biology* 17, 1679-1687.
18. Xu, H., Ren, W., Liu, X., and Li, X. (2010). Aligning protein sequence and analysing substitution pattern using a class-specific matrix. *J Biosci* 35, 295-314.
19. Atchley, W. R., Zhao, J., Fernandes, A. D., and Druke, T. (2005). Solving the protein sequence metric problem. *Proceedings of the National Academy of Sciences of the United States of America* 102, 6395-6400.
20. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research* 22, 4673-4680.
21. Edgar, R. C., and Batzoglou, S. (2006). Multiple sequence alignment. *Current opinion in structural biology* 16, 368-373.
22. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., *et al.* (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947-2948.
23. Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics* 5, 113.
24. Wheeler, D. (2002). Selecting the right protein-scoring matrix. *Current protocols in bioinformatics/editoral board, Andreas D Baxevanis [et al]* Chapter 3, Unit 3 5.
25. Wrabl, J. O., and Grishin, N. V. (2005). Grouping of amino acid types and extraction of amino acid properties from multiple sequence alignments using variance maximization. *Proteins: Structure, Function, and Bioinformatics* 61, 523-534.
26. Henikoff, J. G., Greene, E. A., Pietrokovski, S., and Henikoff, S. (2000). Increased coverage of protein families with the blocks database servers. *Nucleic acids research* 28, 228-230.
27. Kawashima, S., Ogata, H., and Kanehisa, M. (1999). AAindex: Amino Acid Index Database. *Nucleic acids research* 27, 368-369.
28. Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. *Nucleic acids research* 36, D202-205.
29. Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. *Journal of molecular biology* 105, 1-12.
30. Charton, M., and Charton, B. I. (1982). The structural dependence of amino acid hydrophobicity parameters. *Journal of theoretical biology* 99, 629-644.
31. Charton, M., and Charton, B. I. (1983). The dependence of the Chou-Fasman parameters on amino acid side chain structure. *Journal of theoretical biology* 102, 121-134.
32. Prabhakaran, M. (1990). The distribution of physical, chemical and conformational properties in signal and nascent peptides. *The Biochemical journal* 269, 691-696.
33. Komatsu, D. (2001). Protein folding recognition based on amino acid physicochemical property profiles. *Genome Informatics* 12, 358-359.

34. Biro, J. C. (2006). Amino acid size, charge, hydropathy indices and matrices for protein structure analysis. *Theoretical biology & medical modelling* 3, 15.

35. Johnson, R. A. W. D. W. (2002). Applied multivariate statistical analysis. Upper Saddle River, N.J.: Prentice Hall.

36. Schaffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I., Koonin, E. V., and Altschul, S. F. (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic acids researc*h 29, 2994-3005.

37. Farrar, M. (2007). Striped Smith-Waterman speeds database searches six times over other SIMD implementations. *Bioinformatics* 23, 156-161.

38. Colliver, J. A., Barnhart, A. J., Marcy, M. L., and Verhulst, S. J. (1994). Using a receiver operating characteristic (ROC) analysis to set passing standards for a standardized-patient examination of clinical competence. *Academic Medicine* 69, S37-39.

39. Lipman, D. J., Altschul, S. F., and Kececioglu, J. D. (1989). A tool for multiple sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America* 86, 4412-4415.

40. Stebbings, L. A., and Mizuguchi, K. (2004). HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database. *Nucleic acids research* 32, D203-207.