WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Ideal amino acid exchange forms for approximating substitution matrices

Piotr Pokarowski,[1]* Andrzej Kloczkowski,[2] Szymon Nowakowski,[1] Maria Pokarowska,[3] Robert L. Jernigan,[2,4] and Andrzej Kolinski[5]

[1] Institute of Informatics, Faculty of Mathematics, Informatics and Mechanics, Warsaw University, 02-097 Warsaw, Poland

[2] Laurence H. Baker Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, Iowa 50011-3020

[3] Faculty of Geodesy and Cartography, Warsaw University of Technology, 00-661 Warsaw, Poland

[4] Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, Iowa 50011-3020

[5] Laboratory of Theory of Biopolymers, Faculty of Chemistry, Warsaw University, 02-093 Warsaw, Poland

ABSTRACT

We have analyzed 29 published substitution matrices (SMs) and five statistical protein contact potentials (CPs) for comparison. We find that popular, 'classical' SMs obtained mainly from sequence alignments of globular proteins are mostly correlated by at least a value of 0.9. The BLOSUM62 is the central element of this group. A second group includes SMs derived from alignments of remote homologs or transmembrane proteins. These matrices correlate better with classical SMs (0.8) than among themselves (0.7). A third group consists of intermediate links between SMs and CPs - matrices and potentials that exhibit mutual correlations of at least 0.8. Next, we show that SMs can be approximated with a correlation of 0.9 by expressions $c_0 + x_ix_j + y_iy_j + z_iz_j$, $1 \leq i, j \leq 20$, where c_0 is a constant and the vectors (x_i), (y_i), (z_i) correlate highly with hydrophobicity, molecular volume and coil preferences of amino acids, respectively. The present paper is the continuation of our work (Pokarowski et al., Proteins 2005;59:49–57), where similar approximation were used to derive ideal amino acid interaction forms from CPs. Both approximations allow us to understand general trends in amino acid similarity and can help improve multiple sequence alignments using the fast Fourier transform (MAFFT), fast threading or another methods based on alignments of physicochemical profiles of protein sequences. The use of this approximation in sequence alignments instead of a classical SM yields results that differ by less than 5%. Intermediate links between SMs and CPs, new formulas for approximating these matrices, and the highly significant dependence of classical SMs on coil preferences are new findings.

Proteins 2007; 69:379–393.

Key words: substitution matrices; sequence alignment; protein contact potentials; protein structure prediction.

INTRODUCTION

In analyses of amino acid (AA) sequences or protein structures researchers often use information that is written as a matrix $\boldsymbol{A} = (a_{ij})$, particularly for substitution matrices (SMs) and contact potentials (CPs). The SMs carry information about the likelihood of replacement of a given AA by another type as a result of evolutionarily driven mutations, and are used in assessments of sequence alignments or similarity searches. Statistical contact potentials contain information about contact propensities among residues in native protein structures taken from the Protein Data Bank (PDB) or within computer generated decoys. The CPs are used in many kinds of protein structure predictions: homology modeling, threading or *ab initio* methods.

In the previous work[1] we systematically studied CPs mostly developed since 1995, and which were used by successful groups in recent Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiments. Each of the CPs is similar to one of two matrices derived by Miyazawa and Jernigan.[2] The CP matrices of the first class can be approximated with a correlation of order 0.9 by formulas $a_{ij} = x_i + x_j$, $1 \leq i, j \leq 20$, where the residue-type-dependent factor (x_i) is highly correlated with the frequency of occurrence of a given AA type within proteins. The major contribution to the potentials of this class is the one-body transfer energy of the AA from water to the protein environment. Potentials belonging to the second class can be approximated with a correlation of 0.9 by the formula $a_{ij} = c_0 - x_ix_j + y_iy_j$, where c_0 is a constant, with the vector (x_i) being strongly correlated with the Kyte-Doolittle hydrophobicity scale,[3] and a new, less dominant, residue-type-dependent factor (y_i) correlated ($\sim$0.9) with amino acid iso-

Grant sponsor: NIH; Grant number: 1R01GM072014-01; Grant sponsor: Polish Research Council KBN; Grant number: PBZ-KBN-088/P04/2003.
*Correspondence to: Piotr Pokarowski, Institute of Informatics, Warsaw University, Banacha 2, 02-097 Warsaw, Poland. E-mail: pokar@mimuw.edu.pl
Received 21 April 2006; Revised 25 January 2007; Accepted 27 February 2007
Published online 10 July 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21509

electric points pI.[4] We interpret potentials of this class as representing energies of contacts between specific AA pairs within an average protein environment. Both of these "one-body" ideal AA interaction formulas, allow us to comprehend different forces included in statistical potentials. This approximation of CPs is more accurate and universal than previous ones.[5–7]

In the present work we have analyzed 29 diverse published substitution matrices derived from sequence or structure alignments, physicochemical properties of AAs, or optimization of alignments or threading force field parameters. We have also included five CP matrices for comparison. We have found three different groups of SMs. The first group of SMs consists of the most popular, "classical" matrices obtained mainly from sequence alignments of globular proteins. These matrices are mutually correlated by at least a value 0.9 and the BLOSUM62 matrix[8] is the central member of this group. Similar clusters were observed by others.[9,10] The second group is formed by matrices derived from transmembrane proteins and matrices obtained from alignments of remote homologs. Interestingly matrices from this group correlate better with SMs from the first group (0.8) than among themselves (0.7). The third group of matrices consists of intermediate links between SMs and CPs—matrices and potentials that are mutually correlated by at least 0.8. These similarities are reported here for the first time.

Both SM and CP matrices are often defined in a similar way as the log-odds ratio between p_{ij}—the observed frequency of aligned pairs of AAs i and j (or an $i - j$ contacts in protein structures), and q_iq_j—the expected probability for a pair (i, j). Then for SMs we have $a_{ij} = \log(p_{ij}) - \log(q_i) - \log(q_j)$ and for CPs $a_{ij} = -\log(p_{ij}) + \log(q_i) + \log(q_j)$. To separate out the effects of observed substitution or contact frequencies (p_{ij}) and the definitions of expected probabilities (q_iq_j), we apply centering of the matrices, a standard method in statistical data analysis, which reduces the dependence of a_{ij} on q_iq_j and then allows us to explain basic differences among the various matrices. For example, the main differences among the three analyzed transmembrane SMs result from different expected probabilities, because the centering increases correlations from 0.66, 0.76, 0.92 to 0.88, 0.89, and 0.99, respectively. Similar operations were used in Refs. 2 and 11.

We show that SMs can be approximated with a correlation of 0.9 by the expression $c_0 + x_ix_j + y_iy_j + z_iz_j$, where c_0 is a constant and the vectors (x_i), (y_i), (z_i) are highly correlated with hydrophobicity, molecular volume, and coil preferences of AAs, respectively. This result suggests that polarity, side chain packing, and secondary structures are the most conserved physicochemical properties during protein evolution. While the strong dependence of SMs on hydrophobicity and volume of residues is well known, a significant (thought not such high or universal) dependence of the exchangeability of AAs on coil preferences is a new finding. The approximation for SMs and CPs allows us to understand the general trends in AA similarities. The new versions of the BLOSUM62 matrix derived by Muller *et al.*,[12] or Crooks and Brenner[13] separate amino acids into six groups: aliphatic hydrophobic (I, L, M, V), aromatic hydrophobic (F, Y, W), polar (D, E, H, K, N, Q, R), small (A, S, T, P), G and C. For comparison, approximating vectors of the newer version of Miyazawa-Jernigan potential[2] contact potential MJ3 lead to the following partition of residues: positively charged (K, R), negatively charged (D, E), neutral polar (G, H, N, P, Q, S, T, Y), neutral hydrophobic (A, F, I, L, M, W, V) and C. Instead of comparing covariance or distance matrices obtained from SMs and CPs by principal component analysis[11,14] or multidimensional scaling,[15,16] our present approach considers the approximating vectors directly. Our approximation of CPs is more accurate than the one given by distance matrices composed of triplets of AA indices.[10]

Our preliminary calculations have shown that the loss in an alignment's quality resulting from replacement of a classical SM by its approximation is below 5%. Moreover, these new approximate formulas, because of their form, can be applied to improve fast threading[17] or multiple sequence alignments based on the fast Fourier transforms (MAFFT).[18,19] Let us take a closer look at the last problem. At the first stage of computations, for each pair of sequences the algorithm MAFFT finds 20 maximal gapless alignments scored by the substitution matrix $\boldsymbol{A} = (h_ih_j + v_iv_j)$ where the vectors (h_i) and (v_i), are hydrophobicity and volume indices, respectively. The above formula for the matrix $\boldsymbol{A}$ enables computation of values of all gapless alignments using the fast Fourier transform. Our approximation [formulas (3)–(4) below] enables a deeper understanding of the MAFFT method for scoring of alignments and allows us to adapt it for each specific case. Indeed, (i) the matrix $\boldsymbol{A}$ correlates at 0.6–0.7 with matrices from the first and the second group, therefore it may be considered as an approximation of a certain substitution matrix; (ii) replacement of $\boldsymbol{A}$ by our approximations increases these correlations to the range 0.8–0.9. For example, to align sequences of transmembrane proteins we could start MAFFT from our approximation for the transmembrane-specific SM derived by Mueller and collaborators.[20]

METHODS

The method of approximation for matrices used here is quite similar to the method used in our previous work.[1] Because these are symmetrical matrices $\boldsymbol{A} = (a_{ij})$, $1 \leq i, j \leq n$, $n = 20$, we use only the upper diagonal part. Our aim is to find the best approximation of $\boldsymbol{A}$ in the form of a simple function of 1–3 vectors $\mathbf{x} = (x_i)$,

DOI 10.1002/prot

$\mathbf{y} = (y_i)$, $\mathbf{z} = (z_i)$, which will be called properties or indices of AAs. Specifically, we will investigate the following four functions $\tilde{\mathbf{A}} = (\tilde{a}_{ij})$, where

$$\tilde{a}_{ij} = x_i + x_j \tag{1}$$

$$\tilde{a}_{ij} = c_0 + c_1 x_i x_j \tag{2}$$

$$\tilde{a}_{ij} = c_0 + c_1 x_i x_j + c_2 y_i y_j \tag{3}$$

$$\tilde{a}_{ij} = c_0 + c_1 x_i x_j + c_2 y_i y_j + c_3 z_i z_j \tag{4}$$

Vectors x, y, z and constants $c_0,\ldots,c_3$ are obtained by the least squares method

$$\sum_{ij:i\le j} (a_{ij} - \tilde{a}_{ij})^2 \to \min_{\mathbf{x},\mathbf{y},\mathbf{z}} \tag{5}$$

Equations (2)–(4) are similar to the spectral decomposition of a symmetric matrix but there are important differences. Eigenvectors need not be the least squares solution of (5) for the functions (2)–(4). On the other hand, there is no assumption about the orthogonality of the approximating vectors. However, the eigenvectors of a matrix A can be used as the initial solution in an iterative procedure for solving the nonlinear least squares problem for functions (2–4) (see below).

We may assume that the constants c_k (k = 1, 2, 3) are equal to –1 or 1, since by multiplying the approximating vector by $\sqrt{|c_k|}$ and updating $c_k = \mathrm{sign}(c_k)$ we get the same result of the approximation. Moreover, note that if all components of the vector are non-positive, then multiplication by –1 does not change the approximation. Therefore we may assume that either all components of the vector are non-negative, or at least two of the components have opposite signs. These observations lead us to a conclusion, that the approximating factors $c_1 x_i x_j$ (and similarly $c_2 y_i y_j$, $c_3 z_i z_j$) may have one of four forms, depending on whether: (i) the constant $c_k = 1$ or $c_k = -1$; (ii) all components of the approximating vectors are non-negative or some components of the vector are negative. Each factor type may be associated with different characteristics of the ideal form of the exchangeability of AAs or the residue–residue interactions as shown in the Conclusions.

Accuracy of the approximation will be measured by a correlation coefficient. Specifically, let us denote the scalar product of vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ as $(\boldsymbol{x}, \boldsymbol{y})$ the mean value of $\boldsymbol{x}$ as $\bar{x}$ and its standard deviation as σ_x. The correlation between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as $cor(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x} - \bar{x}, \boldsymbol{y} - \bar{y})/(n\sigma_x\sigma_y)$. It is worth noting that the solution of (5) maximizes the correlation between the matrix $\boldsymbol{A}$ and its approximation. The proof of this fact is a simple consequence of two observations: (i) formulas (2)–(4) are closed with respect to affine transformations and (ii) $2 - 2cor(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{y}^N - \boldsymbol{x}^N\|^2/n$, where $\|\boldsymbol{z}\|^2 = (\boldsymbol{z}, \boldsymbol{z})$, $\boldsymbol{x}^N = (\boldsymbol{x} - \bar{x})/\sigma_x$ and $\boldsymbol{y}^N = (\boldsymbol{y} - \bar{y})/\sigma_y$.[21]

It is easy to anticipate that by increasing the complexity of approximations (1–4) we can improve their accuracy (see Appendix in Ref. 1). The simplest additive approximation (1) is given by the vector $\boldsymbol{x}$, which is often highly correlated with empirical hydrophobicities. The solution of the least squares problem for this case leads to linear equations that can be solved analytically [see Eq. (5) in Ref. 1]. The other approximations given by Eqs. (2)–(4) lead to nonlinear least squares problems, requiring numerical solutions. We have used the free software R in our computations.[22] Many vectors were used as starting solutions: the vector defined by Eq. (5) in Ref. 1, the diagonal (a_{ii}) and the eigenvectors of $\boldsymbol{A}$ (or matrix obtained from $\boldsymbol{A}$ by subtracting the mean value) corresponding to the dominant eigenvalues. Generally, there is no significant dependence on starting vectors or on the algorithm used in the optimization: (the Levenberg-Marquardt or the Gauss-Newton method). Moreover, the same results were obtained by using Matlab with the optimization toolbox (Mathworks, Inc; www.mathworks.com).

Approximations of CPs were studied earlier by Godzik *et al.*,[5] Li *et al.*,[6] and Tobi *et al.*[7] In Godzik *et al.*[5] formula (1) and the approximating vector given by the main diagonal of the CP matrix were used. Li *et al.*[6] used a more general formula (2) with the approximating vector being the dominant eigenvector of the matrix obtained from the CP by subtracting the mean value. Tobi *et al.*[7] analyzed performance of threading approximating CPs by their two most dominant eigenvectors. For SMs an approximation defined by distance matrices composed of triplets of AA indices was investigated.[10] Recently, Kinjo and Nishikawa[23] analyzed three dominant eigenvectors of the BLOSUM matrices as a function of sequence identity.

We should note, that substitution matrices could be non-symmetric, unlike contact potentials matrices that are symmetric, except for the Maiorov and Crippen[24] and Baker and coworkers[25] potentials. This may be one of the most obvious differences between interactions and substitutions. For example: the substitution of a small residue with a large one may have a different probability than the substitution of a large AA by a small one. However, only six substitution matrices out of the 29 analyzed in this paper are non-symmetric and we replace these by their symmetric forms $\boldsymbol{A}^{sym} = (\boldsymbol{A} + \boldsymbol{A}^T)/2$. Such a symmetrization procedure does not lead to significant differences in the results, but it does simplify computations. Alternatively, analogs to the approximations (1–4) for non-symmetric matrices could be defined by formulas in which there would be twice as many approximating vectors. For example Eq. (2) can be replaced by $\tilde{a}_{ij} = c_0 + c_1 x^{(1)} x^{(2)}$ with the two vectors $\boldsymbol{x}^{(1)} = (x_i^{(1)})$ and $\boldsymbol{x}^{(2)} = (x_i^{(2)})$.

New here, with respect to our previous work,[1] is the centering of the matrices, which is often used in statistical data analysis. For a given matrix A we define the centered matrix $A^c = (a^c_{ij})_{i,j=1}^n$ such that $a^c_{ij} := a_{ij} - \bar{a}_{i\bullet} - \bar{a}_{\bullet j} + \bar{a}_{\bullet\bullet}$, where $\bar{a}_{i\bullet} = \sum_{j=1}^n a_{ij}/n$, $\bar{a}_{\bullet j} = \sum_{j=1}^n a_{ij}/n$, $\bar{a}_{\bullet\bullet} = \sum_{i,j=1}^n a_{ij}/n^2$. Centering is easily formulated in matrix form as: $A^c = HAH$ with $H = I - J/n$, where I is the identity matrix and J is the matrix composed of ones. Similar procedures were used in Refs. 2 and 11.

As we stated in the Introduction, centering of SMs and CPs reduces the dependence of these matrices on the expected probability vectors. Indeed, let us assume, that $A = (a_{ij})$ and $B = (b_{ij})$ are SMs differing only by expected probability vectors $(q_i^{(1)})$, $(q_i^{(2)})$. Then we have $a_{ij} = m_{ij} - \log q_i^{(1)} - \log q_j^{(1)}$ and $b_{ij} = m_{ij} - \log q_i^{(2)} - \log q_j^{(2)}$, where $m_{ij} = \log p_{ij}$ and p_{ij} is the observed frequency of aligned pairs of AAs i and j. A simple computation shows that $a^c_{ij} = m_{ij} - \overline{m}_{i\bullet} - \overline{m}_{\bullet j} + \overline{m}_{\bullet\bullet} = b^c_{ij}$. Hence, after centering the matrices A and B become equal.

RESULTS AND DISCUSSION

Substitution matrices studied

We have analyzed 29 different substitution matrices and five statistical contact potential matrices for comparison. We have studied mostly new SMs but we have included also a few older, historically important SMs such as the PAM and BLOSUM series. The SMs and CPs are listed and abbreviated as follows.

M—one of the oldest SM developed by Mc Lachlan.[26] The relative substitution frequencies were computed from an alignment of 89 sequences.

G—the matrix of distances between AAs derived by Grantham.[27] The distances combine properties that correlate with protein substitution frequencies: hydrophobicity, molecular volume, and composition.

D120, D250, D500—the popular PAM120, PAM250, PAM500 (Percentage of Acceptable point Mutations per 10^8 years) matrices computed from global alignments of closely related proteins by Dayhoff *et al.*[28]

GCB—the matrix obtained by Gonnet, Cohen, and Benner using exhaustive pairwise alignments of the protein databases[29] and normalizing to the PAM distance of 250. Other matrices of this series reported in Ref. [30] are correlated with GCB at the 0.98 level and these have been omitted in our comparative analyses.

JTT, JTTm—other versions of the PAM250 matrix based on the SWISS-PROT protein sequence database developed by Jones, Taylor, and Thornton. The first matrix was computed from alignments of globular proteins,[15] but the second one originates from alignments of transmembrane proteins.[31]

HH—the BLOSUM62 matrix (BLOcks SUbstitution Matrices)—one of the most popular SM series derived from multiply aligned strongly conserved sequence blocks by Henikoff and Henikoff.[8] This matrix is derived from blocks with at most 62% sequence identity. Because the BLOSUM45, the BLOSUM80 and the OPTIMA[32] matrices have a correlation with HH greater than 0.98, we have studied only the HH matrix.

JO—Johnson and Overington substitution matrix[14] constructed from structural alignments of 235 proteins. A multiple sequence alignment for each of 65 families was computed by the program COMPARER using many structural criteria.

MJsm—the SM estimated by Miyazawa and Jernigan using the contact energies and the number of contacts for each type of AA pair[33] together with a genetic model for transitions. A transition probability matrix for codon substitutions was obtained and finally a log-odds matrix.

BCGg—a version of the PAM250 matrix obtained by Benner *et al.*[30] with the assumption that the genetic code is the only constraint for AA pairs with a PAM1 distance.

NN, AN—the matrices developed by the Nussinov group. The first matrix (Naor *et al.*[34]) was derived from the single residue interchanges at spatially, locally conserved structural regions of globally unrelated proteins. The second matrix (Azarya-Srinzak *et al.*[35]) comes from spatially neighboring pair interchanges calculated from the same protein database.

WAC—Wei, Altman and Chang SM based from analyses of AA microenvironments.[36] Twenty one features summarizing atomic, chemical group, residue, and secondary structure properties were considered to define comparative profiles and finally to construct a substitution matrix.

RSh—the SM introduced by Russell *et al.*[37] based on structural alignments of 335 pairs of remote homologous proteins derived from the SCOP database. Two another SMs derived in this reference are not included here because of their low correlations with other matrices.

OOU—the matrix proposed by Ogata *et al.*[38] for homology modeling. An initial matrix derived from structural alignments was optimized by the Markov chain Monte Carlo (MCMC) method. The objective function in the MCMC algorithm was defined as entropy.

NHH—the matrix built by Ng *et al.*[39] from predicted hydrophobic and transmembrane (PHAT) regions of the Blocks database. The authors demonstrated that in searches with transmembrane queries this matrix outperforms the HH, JTT and JTTm matrices.

PDS—the structure-derived substitution matrix (SDM) derived by Prlic *et al.*[40] The formalism of the Henikoffs was applied to 122 proteins grouped into structurally similar subsets using the PROSUP algorithm. The second matrix reported in this article (HSDM) correlates with SDM by more than 0.98, and we do not include it here.

DOI 10.1002/prot

BC—the SM proposed by Blake and Cohen.[41] Amino acid residue exchange coefficients are estimated from structural superposition data by the MINAREA method.

LMT—the non-symmetric matrix family derived by Lin *et al.*[42] An artificial neural network model was used to obtain probabilities of AA substitutions.

DTs, DTt, DTsn, DTtn—non-symmetric matrices computed from threading force fields by Dosztanyi and Torda.[11] DTs was calculated using SAUSAGE—a low resolution scoring function derived for protein fold recognition. DTt was built from a Boltzmann-based potential of mean force taken from THREADER2.5.[43] DTsn and DTtn are centered versions of DTs and DTt.

MRR—a non-symmetric SM derived by Mueller *et al.*[20] for detection of homologous transmembrane proteins. The observed substitution frequencies are the same as in NHH. The asymmetry arises because there are different background frequencies for the query and the subjects.

MSV—the matrix VTML160 proposed by Mueller *et al.*[12] The search for the set of substitution frequencies and the expected probabilities were calculated by iterating a maximum likelihood estimation until convergence, starting from PAM matrices.

QG—the STROMA matrix derived by Qian and Goldstein[32] for accurate alignments between known homologs. Starting from the GCB matrix the iterative optimization of a coordinate root mean square deviation was performed on the DAPS database.

CB—the SM developed by Crooks and Brenner.[13] Entries of the matrix were computed as the log probability ratios, using the dist.20comp Dirichlet Mixture Model of expected AA probabilities. For comparison we have included five CPs chosen from the set we analyzed in our previous paper.[1]

MJ3h, MJ3—the third version of the popular Miyazawa-Jernigan potential published in 1999.[2] These matrices are centers of two groups of CPs representing transfer energy from water and from an average buried environment, respectively.[4]

B2, B3 and B5—the newest version of quasichemical potentials developed in the research group of Baker. Earlier versions of this potential were discussed in Refs. 25 and 44. The potential is distance dependent—distance bins are denoted by increasing integer numbers. The potentials are part of ROSETTA—currently the most successful protocol for ab initio prediction of protein structure from sequence.[45]

Comparative analysis of substitution matrices

Clustering

Table I shows the results of comparative analysis of SMs and a few CPs. In our computations we use CPs with negative signs, so their correlations with SMs will be positive. The entries below the diagonal represent correlations between the original matrices, while the entries above the diagonal show correlations between centered matrices. All entries are multiplied by 100. The coloring scheme (shown below the table) emphasizes the partition of the set of matrices into five groups: $G_1,\ldots,G_5$. The group G_1 consists of "classical" SMs derived mainly from sequence alignments for globular proteins. This is the most consistent group of SMs: almost all pairs correlate at least at the level 0.9. The central element of this group is the matrix ***HH*** (BLOSUM 62) and their newer variants: matrices ***MSV*** and ***CB***. The group G_2 contains mainly matrices derived from transmembrane proteins (***JTTm***, ***NHH,*** and ***MRR***) and matrices obtained from alignments of remote homologs (***RSh***, ***BC***, ***DTt***, ***LMT***). Matrices from this group correlate better with SMs from G_1 (0.8) than among themselves (0.7). The group G_3 contains only four elements ***MJsm***, ***NN***, ***AN***, and ***DTsn*** intermediate between classical SMs and CPs. This group is more homogenous than G_2, and additionally is more distant from G_2, than from G_1. The matrix ***NN*** represents the distinct center of this group. The group G_4 contains CPs: ***B5***, ***B3***, ***B2,*** and ***MJ3***. These matrices have correlations of 0.8 among themselves, and correlations of 0.7 with G_3. The popular transfer energy potential ***MJ3h*** and the matrix ***DTs*** form G_5.

Groups $G_1,\ldots,G_5$ could be uniquely defined in the following way:

- G_1—a maximal set of SMs such that $\boldsymbol{M} \in G_1$, if and only if ***M*** correlates above 0.85 with more than a half of G_1,
- $G_3 \cup G_4$—a maximal set of matrices, which correlates above 0.65 with ***NN*** and ***MJ3***,
- G_5—a maximal set of matrices, which correlates above 0.65 with ***MJ3h***,
- G_2—remaining matrices.

The importance of the matrices ***MJ3***, ***MJ3h***, and ***NN*** in our clustering is motivated by the observation that ***MJ3*** and ***MJ3h*** are centers of two groups of CPs while ***NN*** is the central element of matrices derived by using the structural properties of proteins. Threshold values 0.65 and 0.85 set by us in this analysis represent moderate (minimal essential) and large correlations, respectively. The group G_2 is defined as the complement of the other groups and it consists of various modifications of classical SMs. The elements of G_2 treated as points in 210-dimensional Euclidean space are located in the neighborhood of the cohesive group G_1, but placed in different directions away from G_1 (among 12 elements of G_2 as many as 10 have the nearest neighbor belonging to G_1).

Let us note that the above partition is consistent with a popular agglomerative clustering method in which dis-

Table I

Correlations Between Substitution Matrices (SMs) Given in the Lower Triangular Part and Centered SMs in the Upper Triangular Part (Values Multiplied by 100)

	D500	D120	D250	QG	JTT	GCB	MSV	HH	CB	DTtn	JO	PDS	M	DTt	BC	RSh	JTTm	LMT	NHH	MRR	WAC	G	OOU	BCGg	MJsm	NN	AN	DTsn	B5	B3	B2	MJ3	MJ3h	DTs
D500	100	88	95	84	87	88	84	78	76	76	79	74	74	76	80	70	67	67	65	65	61	71	73	62	64	56	52	48	57	39	33	39	39	59
D120	84	100	97	88	94	92	94	92	90	88	90	85	89	88	84	74	83	82	83	82	79	79	79	81	75	68	66	62	60	48	42	45	45	62
D250	93	97	100	90	93	94	93	89	88	87	89	84	85	87	85	76	79	79	78	78	75	80	79	75	74	68	65	60	61	47	42	46	46	64
QG	84	85	89	100	91	95	92	89	89	89	89	88	86	89	84	82	79	74	79	78	77	81	79	71	78	76	73	70	65	57	51	56	56	67
JTT	86	90	92	90	100	94	93	91	89	89	89	85	88	89	83	78	89	80	84	84	80	79	78	84	82	72	69	65	63	54	48	50	50	63
GCB	87	88	92	95	94	100	97	95	95	94	94	93	91	94	88	83	81	81	84	84	82	85	85	74	81	79	77	73	67	58	52	57	57	71
MSV	83	91	92	92	93	97	100	97	96	95	95	91	93	95	88	80	84	84	87	87	85	85	85	76	78	78	76	71	64	57	50	54	54	67
HH	78	89	88	89	90	95	96	100	98	97	96	93	93	97	89	77	86	91	92	91	88	84	86	77	78	79	77	72	65	58	51	54	54	68
CB	76	84	85	88	88	94	95	97	100	97	96	94	94	97	90	78	83	90	89	88	90	86	86	74	79	83	81	78	70	63	56	61	61	74
DTtn	75	83	84	87	88	93	93	95	96	100	94	94	92	100	88	82	84	85	88	88	89	86	84	72	82	89	87	84	74	68	62	64	64	76
JO	78	89	88	87	86	91	93	94	93	90	100	92	93	94	91	77	80	88	86	86	85	85	86	76	79	78	76	71	70	56	49	55	55	74
PDS	74	81	82	87	85	93	91	92	94	93	89	100	89	94	86	80	80	81	84	84	84	83	82	70	80	84	81	79	69	63	57	60	60	73
M	73	87	85	84	86	89	91	92	91	88	92	86	100	92	83	72	80	85	84	83	87	84	81	78	79	78	76	72	64	57	49	52	52	68
DTt	70	74	76	81	80	86	86	87	88	90	83	85	81	100	88	82	84	85	88	88	89	86	84	72	82	89	87	84	74	68	62	64	64	76
BC	78	83	84	81	79	84	84	86	85	83	91	82	83	75	100	76	74	80	78	78	76	78	82	66	74	75	71	70	75	55	49	57	57	79
RSh	68	65	71	79	76	82	78	75	78	79	71	79	67	72	69	100	71	56	69	69	65	74	67	59	77	77	73	74	67	62	60	63	63	68
JTTm	62	77	75	75	84	77	79	81	78	77	75	76	75	62	67	66	100	75	89	88	80	72	67	81	74	70	71	64	52	54	45	45	45	53
LMT	67	77	76	73	79	80	82	89	87	84	84	79	82	76	76	53	69	100	85	85	84	70	78	72	61	62	63	54	53	41	33	38	38	58
NHH	61	77	73	74	79	78	82	86	82	81	79	78	78	78	72	61	76	79	100	99	87	73	74	76	69	70	72	64	54	56	41	42	42	57
MRR	59	68	67	71	75	76	79	82	81	80	73	76	72	85	65	63	66	77	92	100	86	73	74	75	69	70	71	64	54	55	39	41	41	57
WAC	57	71	68	73	75	77	80	82	85	82	77	79	78	82	67	62	65	76	78	82	100	80	75	70	71	77	77	75	59	61	52	53	53	64
G	65	76	77	73	69	75	75	76	75	74	82	72	81	70	78	59	59	60	62	54	67	100	78	63	74	77	74	73	66	55	54	61	61	73
OOU	61	72	70	69	65	69	72	73	67	69	75	66	71	62	73	45	52	66	63	54	54	68	100	64	66	68	65	60	64	51	43	49	49	69
BCGg	62	78	73	70	83	73	75	76	73	72	74	70	75	64	64	57	78	71	72	66	63	55	52	100	78	56	56	52	48	42	37	37	37	49
MJsm	62	70	71	76	81	79	76	76	77	81	75	79	76	70	70	75	70	60	65	60	62	62	55	76	100	81	75	78	69	66	66	66	66	66
NN	55	65	66	74	71	77	76	77	80	88	74	82	74	72	71	74	70	60	63	56	68	65	56	55	79	100	97	93	77	76	74	72	72	79
AN	47	62	61	67	65	70	70	72	74	79	70	75	70	57	66	66	71	57	59	47	61	62	52	52	71	95	100	89	71	74	69	68	68	75
DTsn	48	59	59	69	64	72	70	70	76	84	68	77	69	75	66	71	59	54	59	57	68	63	51	51	76	91	82	100	75	70	74	73	73	82
B5	54	51	55	62	61	65	61	61	68	72	60	68	56	66	63	66	47	51	49	50	55	45	44	47	66	73	62	73	100	75	74	78	78	92
B3	37	45	45	54	53	57	55	56	61	65	53	61	54	49	51	61	57	39	49	40	50	46	36	42	65	76	74	66	70	100	83	79	79	58
B2	31	39	39	49	47	50	48	49	54	59	46	55	46	42	46	58	48	30	37	29	44	44	33	36	65	75	71	70	68	83	100	90	90	60
MJ3	38	43	45	54	50	56	53	54	60	63	53	60	51	52	55	63	45	37	41	35	50	54	38	37	65	72	66	71	74	78	89	100	97	68
MJ3h	21	8	14	25	23	28	24	22	31	34	18	30	17	56	19	37	-1	20	28	48	38	15	4	16	30	22	2	37	48	21	26	45	100	68
DTs	27	14	20	30	29	34	30	27	35	38	20	35	19	61	19	37	4	29	27	50	37	7	11	20	28	23	4	40	57	9	9	22	81	100

50 60 70 80 90 100

For comparison we also include five protein contact potentials (CPs). The coloring key is given below the table.

tances between clusters are computed using McQuitty's formula (for details see the Help to the function hclust in the statistical package **R**).[22] Indeed, Figure 1 shows that groups G_1, G_3,...,G_5 are branches of the dendrogram generated by McQuitty's clustering method.

Automatic clustering of 42 SMs by the single linkage and the complete linkage agglomerative clustering algorithms was studied 10 years ago by Tomii and Kanehisa.[10] The largest cluster from their partition is, in a certain sense, a predecessor of our group of classical

DOI 10.1002/prot

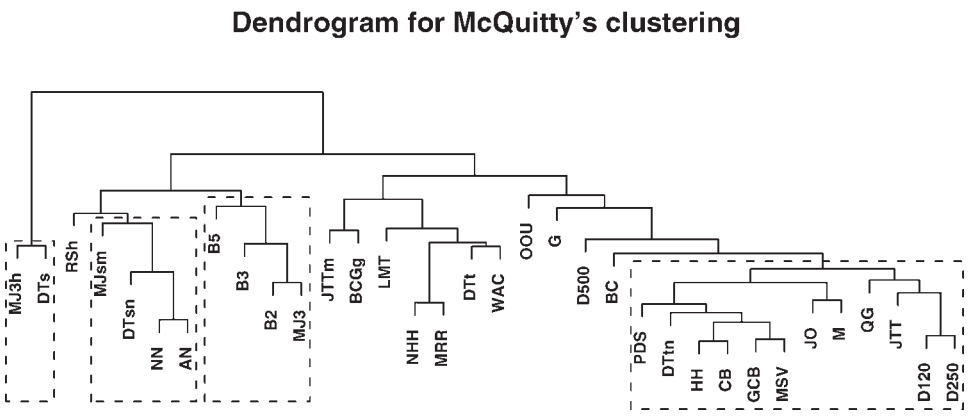

Figure 1

The dendrogram for SMs and CPs matrices generated by agglomerative clustering algorithm using McQuitty's formula for intercluster distances. Branches contained in dashed-line rectangles correspond to the same clusters of matrices as given in Table I.

SMs. However the other groups are different. In another paper by May[9] a comparison of 18 SMs derived mostly before 1993 (except the ***RSh*** matrix) was made. This original, two-step hierarchical clustering leads to a group composed of many classical SMs, but unexpectedly the popular ***GCB*** matrix is left outside this group. For comparison, our clustering is more laborious, but provides a more insightful partition of the SMs. Additionally, 16 of the SMs analyzed here were published after 1996 and were not included in those earlier studies.[9,10]

Intermediate links

The intermediate links between SMs and CPs are provided by the matrices ***NN***, ***DTsn***, ***DTs*** and the potentials ***B2***, ***B3,*** and ***B5***. Figure 2 illustrates these relationships. We should note that the "transition" between SMs and CPs is more "continuous" (correlations ~0.8) than the transition between these two groups of CPs centered around ***MJ3h*** and ***MJ3***, respectively (correlations ~0.5).

Centering

We use centering of matrices to understand the effects of databases of substitutions or contacts, and on the other hand, differences in expected probabilities. It appears that the expected probability is the main factor differentiating contact potentials. Indeed after centering, each potential from the first group of CPs (transfer energy dependent) moves to the second group of the average protein environment dependent potentials (data not shown). It is worth remembering that the matrix ***MJ3*** was originally derived as the result of "physical" centering of ***MJ3h***, which depended not only on this matrix but also on the number of contacts. Since after the centering here, ***MJ3h*** is correlated with ***MJ3*** with a value of 0.97, we can clearly interpret the centered ***MJ3h*** (***HMJ3hH***) as an approximation of ***MJ3***, which is based only on ***MJ3h***.

Table I shows also that the intra-group correlations calculated before and after centering for the matrices of the first and the fourth group are almost the same. This

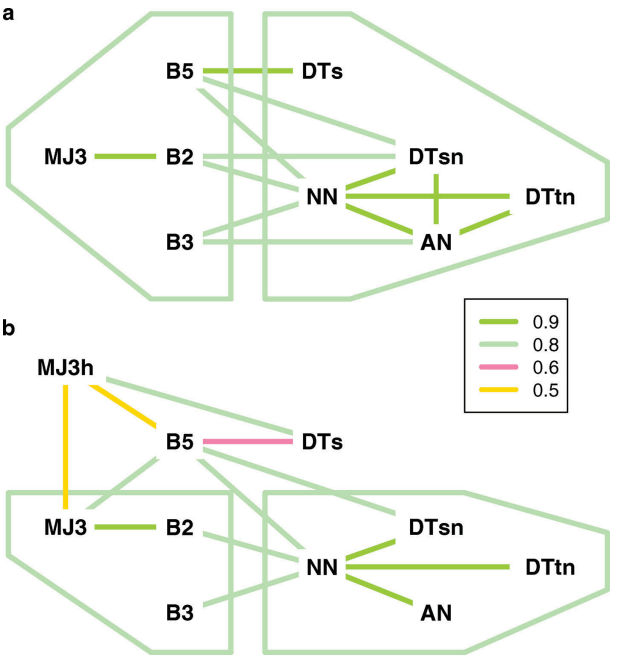

Figure 2

*Illustration of intermediate links between CPs and SMs. Bonds in the graph denote correlations according to the same coloring scheme as in Table I. Part (**a**) corresponds to centered and part (**b**) to original SMs, respectively. To simplify the figure, only the strongest correlations for each node have been shown. Groups of matrices correlated with each other by more than 0.75 are enclosed by boxes.*

means that the original matrices are already centered. Otherwise, correlations among centered SMs from the second group would be higher (∼0.8) than among original ones (∼0.7). Centering enhances especially the correlations among the transmembrane matrices ***JTTm***, ***NHH,*** and ***MRR,*** increasing them from 0.66, 0.76, 0.92 to 0.88, 0.89 and 0.99, respectively. Additionally, correlations between the first and the second group have also improved after centering (from about 0.8–0.85). This is easily seen for ***DTt***, ***G,*** and ***OOU***. Probably the oddest behavior is seen for the matrix ***DTs***, where before centering ***DTs*** is most correlated with the transfer energy matrix ***MJ3h***, but after centering ***DTs*** shows the highest correlation with ***B5***. The substitution matrix ***DTs*** derived from the SAUSAGE force-field shows more similarity to CPs than to SMs.

Ideal substitution forms

Accuracy of one-body approximations

Table II shows in columns 2–5 correlations between the matrices and their one-body approximations given by Eqs. (1)–(4). It is evident that correlations between SMs belonging to $G_1 \cup G_2$ and their approximations (2) are roughly 0.7. The second and the third approximating vectors increase the accuracy of the approximation of elements of $G_1 \cup G_2$ to 0.8 and 0.9, respectively. For comparison, intermediate links and CPs are easier to approximate, because the obtained correlations between matrices belonging to $G_3 \cup G_4 \cup G_5$ and their approximations (2) are at the level 0.9. It is interesting that by comparing correlations in columns 2–5 of Table II, we may group matrices (rows of Table II) similarly to the groupings shown in Table I. Additionally we should notice that vectors ***x*** approximating given matrix in various formulas (2)–(4) may differ [similarly as do vectors ***y*** in formulas (3)–(4)].

Interpretation of approximating vectors

To interpret our approximations we used the AAindex database[46] of amino acid properties developed by the Kanehisa laboratory (http://www.genome.jp/aaindex/) currently containing 494 AA indices, supplemented by an additional 26 hydrophobicity scales. All numerical data and detailed references are available as Supplementary Materials to this article and can be found at http://www.mimuw.edu.pl/∼pokar. For all approximated matrices and all approximating vectors ***x***, ***y***, ***z*** we identified the most closely correlated AA properties in the AAindex, and tried to divide this set of observable scales, denoted as A, into homogenous and easily interpretable groups. At the very beginning we have observed that certain scales from A frequently occur in our computations.

The popular Kyte-Doolittle[3] hydrophobicity scale (abbreviated as Hp, No. 151 in AAindex) was usually the closest one to the approximating vector ***x***. On the other hand we know that hydrophobicity is a fuzzily defined concept: e.g. Hp scale correlates with the Zimmerman *et al.*[4] hydrophobicity scale (No. 400 in AAindex) at 0.67 and with the Levitt scale[47] (No. 153) at 0.68. Because of that we assumed that the hydrophobicity scale must be correlated with Hp at a level of at least at 0.65.

Frequently, the scales that were the closest to the vector ***y*** appeared to be the scales referring to the size of residues or AA volumes. Similarly to hydrophobicity, AA size is also a fuzzy concept. AAindex contains several dozen scales that can be interpreted as AA size, but correlate with each other by no more than 0.7. For example, the scale that counts the number of atoms in side chains (No. 28) correlates with the molecular volume of the side chain (abbreviation Vol, No. 150) at 0.74.[48] This last scale appears to be the best candidate for the center of the group of scales referring to AA size, since all the scales that definitely measure the AA size are highly correlated with it. Similarly to the choice of the hydrophobicity scale we assumed that the scale referring of the AA size must be correlated with Vol by at least 0.65.

The most difficult task was to group scales belonging to A that have the largest correlation with the third approximating vector ***z***. The most frequent (occurring six times in Table II) was the scale of the frequency of AAs occurring as the fourth residue in a turn (No. 52),[49] however this scale is not highly correlated with other elements of A, and therefore cannot be chosen as the center of the cluster. Nevertheless we found, that among other scales a homogenous group of scales is formed by those correlated at 0.65 or more with the scale measuring normalized, non-weighted frequency of reverse turns (abbreviation Cl, No. 165).[50] Thus, the third approximating vector ***z*** correlates with various definitions of coil. The observation of coil preferences in SMs is an important new result in the present paper, and therefore Table III contains a short definition of all scales measuring coil preferences in the A set. Figure 3 illustrates the highest correlations among these scales.

One should notice that six scales out of the total 14 in Table III have the word "turn" in their descriptions, and other two scales define the frequency of terminal residues in helices or β-sheets. Additional four scales are defined based on the frequency of the occurrence of amino acids in β-sheets (these scales are marked in red in Fig. 3). Because these scales were derived for β-proteins the change of the sign of the scale reverses the preference to β-sheets to the preference to coil (since helices are absent in β-proteins). The sign of the AA scale is often a problem of a convention; the scales that are negatively correlated with Cl in Figure 3 are enclosed in boxes.

In summary, in all matrices the third vector reflects the absence of a regular secondary structure for the corresponding residues.

Because some scales contained in A are correlated weakly with all centers, while certain other correlate

 DOI 10.1002/prot

Table II
One-Body Approximations of SMs

	add	x	xy	xyz	x		y		z		dist	
Matrix				cor	cor	No.	cor	No.	cor	No.	cor	No.
D500	9	74	90	97	72	489	95	159	77	110	80	111, 212, 177
D120	31	63	78	86	86	151	77	474	86	434	78	111, 212, 177
D250	22	68	82	91	96	365	78	22	78	52	82	111, 212, 177
QG	15	74	83	90	84	151	85	474	89	511	82	111, 150, 427
JTT	14	65	82	90	80	365	73	52	90	483	79	505, 111, 177
GCB	15	74	85	91	95	151	92	353	78	52	87	211, 111, 150
MSV	18	67	82	88	80	151	95	63	80	212	84	506, 390, 80
HH	17	62	76	83	92	486	87	29	84	61	84	210, 390, 154
CB	18	67	79	85	96	489	87	399	85	361	87	151, 158, 431
DTtn	5	76	84	89	96	111	84	28	88	416	87	487, 436, 112
JO	26	63	78	85	92	519	86	28	91	181	85	489, 212, 165
PDS	13	74	83	88	97	211	73	29	75	52	83	477, 111, 353
M	24	66	76	84	84	151	89	353	86	489	79	505, 147, 112
DTt	43	78	85	90	92	519	84	28	91	497	82	448, 388, 319
BC	29	68	83	89	91	489	95	212	77	23	82	489, 147, 353
RSh	26	73	83	87	96	111	85	316	86	228	79	151, 58, 363
JTTm	37	72	86	91	92	111	85	316	71	347	73	151, 505, 319
LMT	7	40	54	63	85	151	89	212	86	404	67	505, 489, 316
NHH	36	57	75	84	82	127	92	397	81	360	79	252, 131, 319
MRR	42	66	78	86	89	388	88	316	79	127	87	127, 111, 41
WAC	39	66	77	83	94	388	88	154	90	52	78	388, 154, 39
G	48	72	92	97	96	239	91	157	89	110	100	111, 112, 110
OOU	54	60	81	89	85	130	92	212	91	136	72	523, 212, 303
BCGg	13	60	76	87	85	111	67	63	72	52	49	436, 71, 177
MJsm	19	87	94	97	97	185	71	346	72	236	87	509, 185, 71
NN	20	91	96	98	97	211	83	48	73	313	92	151, 10, 167
AN	39	87	96	98	94	111	93	79	85	424	88	151, 241, 14
DTsn	5	96	98	99	97	210	88	79	86	240	90	519, 446, 426
B5	26	87	98	99	96	489	86	361	84	464	89	67, 211, 151
B3	31	86	93	96	98	151	91	401	84	464	82	436, 466, 490
B2	34	86	94	98	89	151	90	401	94	131	84	71, 209, 241
MJ3	21	80	90	94	94	151	86	401	93	146	80	36, 321, 390
MJ3h	86	93	96	98	97	495	94	401	97	495	73	299, 400, 324
DTs	88	97	99	100	95	210	90	110	95	210	52	356, 66, 55

50 60 70 80 90 100

Columns 2–5 contain values of correlations between particular SMs and their approximations in the forms (1)–(4) denoted here as ***add***, ***x***, ***xy***, and ***xyz***, respectively. Columns 6, 8, and 10 contain values of correlations for the individual optimal approximating vectors with respect to Eq. (4) and the closest amino acids indices from the subset A of the AA index database. Numbers denoting scales are included in columns 7, 9, and 11. Column 12 contains the correlations between SMs and an approximation given by Grantham's formula (6). The last column indicates the numbers of scales used in the last approximation. We denote identifying numbers of hydrophobic scales in blue, volume scales in red, coil preferences scales in green and unclassified scales in black. The scaling factor and the coloring scheme is the same as in Table I.

highly with a few of them (e.g. with Hp and Cl), we need an additional condition to properly identify scales and therefore the approximating vectors. The scale S belonging to A will be called hydrophobic if $|\mathrm{cor}(S,\mathrm{Hp})| > \max(|\mathrm{cor}(S,\mathrm{Vol})|, |\mathrm{cor}(S,\mathrm{Cl})|, 0.65)$. Similarly defined are scales referring to the AA size and the coil preferen-

Table III
The List of Scales in A Which are Classified as Coil Preferences

No.	Accession no.	Short description of the index and the reference
39	CHOP780202	Normalized frequency of beta-sheet[49]
48	CHOP780210	Normalized frequency of N-terminal non beta region[49]
52	CHOP780215	Frequency of the 4th residue in turn[49]
79	FAUJ880102	Smoothed upsilon steric parameter[51]
165	LEVM780106	Normalized frequency of reverse turn, unweighted[50]
167	LIFS790101	Conformational preference for all beta-strands[52]
228	PALJ810106	Normalized frequency of turn from CF[58]
236	PALJ810114	Normalized frequency of turn in all-beta class[53]
346	ROBB760108	Information measure for turn[54]
347	ROBB760109	Information measure for N-terminal turn[59]
363	SUEM840101	Zimm-Bragg parameter s at 20 C[55]
416	AURR980114	Normalized positional residue frequency at helix termini C2[56]
426	VINM940102	Normalized flexibility parameters (B-values) for each residue surrounded by none rigid neighbours[57]
431	MUNV940103	Free energy in beta-strand conformation[58]

Accession numbers and descriptions are taken from the AAindex.

ces. If the given scale S doesn't belong to any of these three groups we call it unclassifiable. Additionally, by definition the scale No. 52 refers to coil preferences. Table IV illustrates clustering of all scales belonging to A. It contains absolute values of correlations among these scales and the centers Hp, Vol, and Cl.

Columns 6–11 of Table II show correlations between approximating vectors and AA indices collected in the AAindex. Our major conclusions from the analysis of this data are:

1. SMs can be accurately (∼0.9) approximated by Eq. (4). Since the constants c_1–c_3 are positive for almost all SMs this expression can be simplified to $c_0 + x_i x_j + y_i y_j + z_i z_j$. The vectors (x_i), (y_i), (z_i) are highly correlated with hydrophobicity, molecular volume, and coil preferences of AAs, respectively. This result suggests that polarity, side chain packing and secondary structure are the most conserved physicochemical properties during protein evolution.
2. CPs are easier to approximate than SMs. However, the second vector **y** for CPs, instead of being associated with molecular volume (Vol), correlates better with isoelectric points (pI, scale No. 401 in AA index).
3. Intermediate links from the third group of SMs can be approximated, similarly to CPs by Eqs. (4) and (6). Among the intermediate links in the fourth group, the potential ***B5*** has its vector ***y*** correlated with Vol – similarly to the SMs belonging to groups 1–2. On the contrary, the matrix ***DTs*** is associated with the transfer energy ***MJ3h***.

Approximation of SMs by distance matrices

The substitution matrices have been approximated in the past by using distance matrices between AAs scales.[10,27] This has a sound foundation, because most of the classical substitution matrices M are positive definite, and therefore might be obtained by centering of matrices of the squares of Euclidean distances $\boldsymbol{D}_2$ ($-2\boldsymbol{M} = \boldsymbol{HD_2H}$, see Th. 14.2.1 in Ref. 59). Columns 12–13 of Table II show correlations between SMs and their corresponding approximations given by the distances $\boldsymbol{D} = (d_{ij})$ defined by the formula introduced by Grantham[27]

$$d_{ij} = \sqrt{\alpha_x(x_i - x_j)^2 + \alpha_y(y_i - y_j)^2 + \alpha_z(z_i - z_j)^2} \qquad (6)$$

where $\alpha_w = (\sum_{i>j} |\mathbf{w}_i - \mathbf{w}_j|/190)^{-2}$ and $w \in \{x, y, z\}$.

Here coefficients α_w normalize contribution of the scales **w** at the distance d_{ij}. It is unclear why Grantham didn't use classical normalization, dividing **w** by its standard deviation, which would be equivalent to the use of the formula $\alpha_w = (\sum_{i>j}(w_i - w_j)^2/400)^{-1}$.

The approximation (6) is slightly less accurate (∼0.8), but could be improved by choosing fitted solutions of

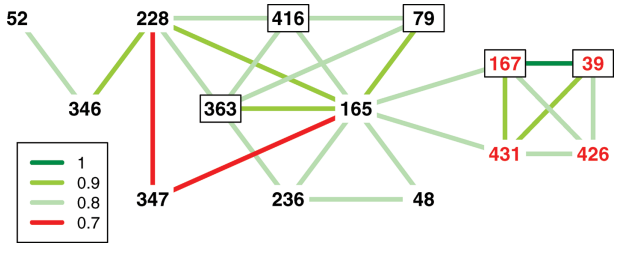

Figure 3
Coil preferences scales in the subset A described as in Table III. Edges of the graph depict the strongest correlations for each node. Scales that are marked in red were developed based on the frequency of the occurrence of amino acids in β-sheets. Scales that are enclosed in boxes are negatively correlated with the central scale No. 165.

DOI 10.1002/prot

Table IV
Classification of Scales in A

No.	151	150	165	No.	151	150	165	No.	151	150	165	No.	151	150	165	No.	151	150	165
10	83	38	61	110	30	48	37	209	87	3	47	356	88	14	37	448	89	21	59
14	67	62	51	111	86	37	55	210	81	32	70	360	60	34	4	464	59	31	44
22	9	96	45	112	4	99	62	211	80	51	74	361	2	74	25	466	68	21	31
23	17	29	32	127	85	20	27	212	52	73	55	363	44	73	88	474	3	99	52
28	38	74	28	130	87	3	49	228	47	54	85	365	75	57	62	477	84	41	77
29	34	88	35	131	87	11	37	236	55	49	79	388	79	31	50	483	75	54	55
36	89	32	33	136	13	55	22	239	73	64	65	390	78	53	54	486	81	39	51
39	67	54	71	146	5	32	19	240	77	55	63	397	26	66	44	487	85	31	73
41	37	25	46	147	71	22	67	241	85	31	73	399	44	76	64	489	87	2	50
48	49	32	79	150	4	100	59	252	85	1	34	400	67	9	13	490	84	39	69
52	22	3	33	151	100	4	56	299	76	15	50	401	20	32	15	495	84	30	61
55	60	73	65	154	37	88	35	303	3	24	19	404	41	4	52	497	77	20	29
58	72	62	67	157	33	90	37	313	1	35	56	416	16	62	75	505	75	57	62
61	70	42	70	158	38	86	64	316	26	87	46	424	11	29	59	506	85	26	65
63	17	89	45	159	11	86	31	319	9	88	63	426	61	62	69	509	86	44	71
66	84	5	62	165	56	59	100	321	84	35	63	427	65	45	65	511	74	33	60
67	90	7	56	167	69	58	81	324	42	51	61	431	66	54	82	519	80	44	77
71	90	15	48	177	5	81	48	346	54	20	69	434	50	55	38	523	84	20	36
79	29	68	87	181	62	57	43	347	48	12	65	436	84	18	59				
80	14	96	43	185	82	50	69	353	27	92	67	446	92	16	56				

50 60 70 80 90 100

Columns 1, 5, 9, 13, and 17 contain the numbers of scales. We denote identifying numbers of hydrophobic scales in blue, volume scales in red, coil preferences scales in green and unclassified scales in black. Remaining columns contain absolute values of correlations between a given scale and centroids of hydrophobic (151), volume (150) and coil preferences (165) scales. The scaling factor and the coloring scheme is the same as in Table I.

(5) instead of scales selected from AA index. The former one was applied by Tomii and Kanehisa[10] in the statistical analysis of SMs. Here we have performed the same analysis but for a different set of matrices, and by employing a significantly expanded database of AA properties.

Trends in amino acid similarities shown in CPs and SMs

Figure 4 compares AA scales (Hp, pI, Vol) with the vectors ***x***, ***y***, ***z*** in formula (4) that are the solution to the least squares problem (5) for the matrices ***MJ3***, ***B5***, ***NN***, ***CB*** and ***MSV***, respectively. It is an illustration of the results from Table II that the first approximating vector ***x*** for ***MJ3*** is closely correlated with Hp, while the second one correlates well with pI. For comparison, the first approximating vectors for ***CB*** and ***MSV***, newer versions of the BLOSUM62 matrix, correlate also with hydrophobicity. This is, however, a different aspect of hydrophobicity. Indeed, in ***MJ3*** aliphatic AAs are more hydrophobic than aromatic ones, contrary to the case for ***CB*** and ***MSV***. The ***B5*** and ***NN*** matrices represent an intermediate case, with aliphatic and aromatic residues being closer to one another than for typical SMs, but more distant than for typical CPs.

The approximation of SMs and CPs allows us to understand general trends in AA similarity. Classical SMs separate AAs into six groups: aliphatic hydrophobic (I, L, M, V), aromatic hydrophobic (F, Y, W), polar (D, E, H, K, N, Q, R), small (A, S, T, P), G and C. For comparison, the approximating vectors of the potential ***MJ3*** lead to the following partition of residues: positively charged (K, R), negatively charged (D, E), neutral, polar (G, H, N, P, Q, S, T, Y), neutral, hydrophobic (A, F, I, L, M, W, V) and C.

Similar results were obtained earlier by principal component analysis[11,14] or multidimensional scaling of substitution matrix data. In the case of the first, this fact is easily explicable, since for centered symmetric matrices

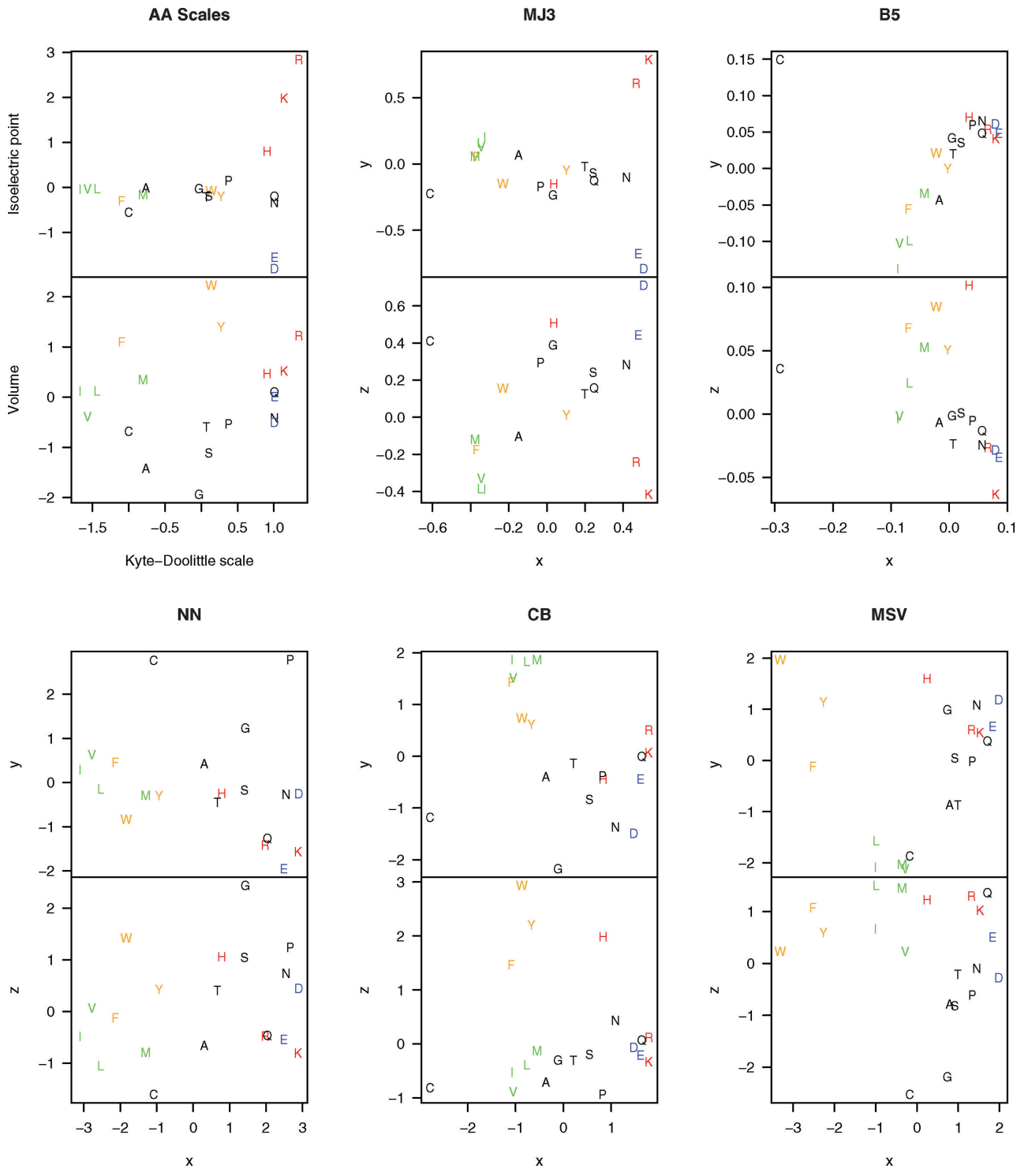

Figure 4

Amino acid similarity shown by the three indices (Hp, pI, Vol) in comparison with the optimal approximating vectors x, y, z in formula (4) for representative CPs and SMs. AAs are colored as follows: aliphatic in green, aromatic in orange, positively charged in red, negatively charged in blue and the remainder in black.

principal components and eigenvectors coincide. However, classical SMs are almost centered and our approximating vectors do not differ significantly from eigenvectors of SMs, which are a starting solution for Eq. (5). Similarities with the results of the second method follow from the equivalence of principal component analysis

DOI 10.1002/prot

Table V
The effect of gap penalties and different approximations on sequence alignments using the Mueller, Spang and Vingrom (MSV) matrix[12]

	MSV_0				Two-vector approximation					Three-vector approximation				
No	α	β	r_0	a_0	α_2	β_2	r_2	a_2	e_2	α_3	β_3	r_3	a_3	e_3
1	36	2.05	4.1	73	14	1.45	4.3	69	6.6	20	2.50	4.4	71	3.7
2	24	3.40	4.2	75	12	1.75	4.4	70	6.3	18	1.75	4.5	73	3.6
3	20	2.80	4.4	77	12	1.60	4.4	70	6.3	16	1.75	4.5	75	3.4
4	18	2.05	4.5	80	11	1.45	4.7	72	5.4	15	1.75	4.6	76	3.2
5	12	2.80	4.8	82	11	1.00	5.0	74	4.7	12	1.90	4.9	78	1.5
6	12	2.20	5.0	83	9	1.30	5.3	76	4.8	11	1.60	5.2	81	1.2
7	9	2.35	5.3	84	9	1.15	5.5	77	4.5	9	1.90	5.4	83	1.1

The loss of precision of alignments for seven different values of the gap penalty parameters (α, β) for MSV_0 is shown in columns 2–3. Columns 4 and 5 show $r_0(\alpha, \beta)$ and $a_0(\alpha, \beta)$. The first two columns of the two-vector approximation section of Table 5 show gap penalty parameters (α_2,β_2) for MSV_2 , which are optimal according to Eq. (7). The next three columns of this section present $r_2(\alpha_2, \beta_2)$, $a_2(\alpha_2, \beta_2)$ and the minimal error computed from Eq. 7 $e_2(\alpha_2, \beta_2)$. The last section of Table V displays the corresponding data for the three-vector approximation MSV_3.

and classical multidimensional scaling (Th. 14.3.1 and Th. 14.4.1 in Ref. 59). Experimental confirmation of this fact is evidenced by the high correlation between a given substitution matrix and its covariance or distance matrix.

One-body approximation does not significantly affect alignments

To test the performance of the one-body approximations [Eqs. (3) and (4)] we aligned 1033 pairs of protein sequences (one random pair for each homologous family) from the HOMSTRAD database (release April 1, 2005).[60] We used the Needleman-Wunsch global alignment algorithm with an affine gap penalty[61,62] $p(l) = -\alpha - \beta l$, where l denotes the length of a gap and $\alpha, \beta > 0$ are the gap opening and the gap extension penalties, respectively. End gaps were not penalized. Alignment quality was measured by the root-mean-square distance (RMSD) of aligned fragments, and the *alignment percentage* (AP), defined as $100^{*}N/(N_1 + N_2 - N)$, where N is the number of aligned residues and N_1 and N_2 are the lengths of two sequences. We have carried out, as an example, about 500 experiments for the ***MSV*** matrix derived by Mueller *et al.*,[12] and its approximations [Eqs. (3) and (4)], changing the gap penalties (α = 9,10, . . .,45, β = 0.85, 1.00,. . .,4.15) around the previously[63] used values (α, β) = (14, 2). For each experiment the results were averaged over the HOMSTRAD database.[60]

To precisely quantify the loss of alignment quality resulting from the replacement of the ***MSV*** matrix by its approximations we introduce the following definitions. Let $\boldsymbol{MSV}_0$, $\boldsymbol{MSV}_2$, and $\boldsymbol{MSV}_3$ denote the original ***MSV*** matrix and its two- and three-vector approximations, respectively. Let $r_i(\alpha, \beta)$ and $a_i(\alpha, \beta)$ denote the averaged RMSD and the averaged AP of alignments of the $\boldsymbol{MSV}_i$, i =0,2,3, matrices for the gap penalty parameters (α, β). We will call parameters (α, β) *admissible* for the matrix $\boldsymbol{MSV}_i$, if there are no (α', β') such that $r_i(\alpha', \beta') < r_i(\alpha, \beta)$ and $a_i(\alpha', \beta') > a_i(\alpha, \beta)$. We define the proportional loss of the alignment quality produced by the two-vector approximation $\boldsymbol{MSV}_2$ and the gap penalty parameters (α', β') relative to the original $\boldsymbol{MSV}_0$ matrix and (α, β) penalties as

$$err_2(\alpha, \beta, \alpha', \beta') = 100\sqrt{\frac{[r_0(\alpha, \beta) - r_2(\alpha, \beta)]^2}{2r_0^2(\alpha, \beta)} + \frac{[a_0(\alpha, \beta) - a_2(\alpha', \beta')]^2}{2a_0^2(\alpha, \beta)}} \quad (7)$$

Let $e_2(\alpha, \beta) = \min_{\alpha',\beta'} err_2(\alpha, \beta, \alpha', \beta')$ and (α_2, β_2) be the optimal penalty parameters for $\boldsymbol{MSV}_2$ given $\boldsymbol{MSV}_0$ and (α, β), i.e. $e_2(\alpha, \beta) = err_2(\alpha, \beta, \alpha_2, \beta_2)$. Similarly we define $err_3(\alpha, \beta, \alpha', \beta')$, $e_3(\alpha, \beta)$, α_3 and β_3 for the three-vector approximation $\boldsymbol{MSV}_3$.

Table V shows the loss of precision of alignments for seven different values of the gap penalty parameters (α, β) for $\boldsymbol{MSV}_0$ displayed in columns 2–3. Columns 4 and 5 show the values $r_0(\alpha, \beta)$ and $a_0(\alpha, \beta)$ of the average RMSD and the average AP. The first two columns of the two-vector approximation section of Table V show gap penalty parameters (α_2, β_2) for $\boldsymbol{MSV}_2$, which are optimal according to Eq. (7). The next three columns of this section of the Table present $r_2(\alpha_2, \beta_2)$, $a_2(\alpha_2, \beta_2)$, and $e_2(\alpha_2, \beta_2)$. The last section of Table V displays the corresponding data for the three-vector approximation $\boldsymbol{MSV}_3$. The fifth row of Table V shows probably the most interesting statistical data on sequence alignments obtained for $\boldsymbol{MSV}_0$ with (α, β) = (12, 2.80), These gap penalty parameters are better than (α, β) = (14, 2) used earlier by Green and Brenner.[63] From the columns e_2 and e_3 in this row it is evident that the most successful two- and three-vector approximations give alignments worse only by 4.7% and 1.5% from alignments obtained with the original matrix ***MSV***. For comparison among all admissible parameters (α, β) the maximal loss of the alignment quality resulting from the replacement of ***MSV*** by its two-vector approximation is 7.9% and for the three-vector approximation is 4.6%.

Figure 5 shows alignment percentage AP (ordinate) vs. RMSD (abscissa) plot of points $(r_i(\alpha, \beta), a_i(\alpha, \beta))$ for all admissible parameters (α, β) for $\boldsymbol{MSV}_i$, $i = 0, 2, 3$. The similarity between Figure 5 and the ROC (Receiver Operating Characteristics) curves is not accidental. As a matter of fact, the methodology used in both cases is quite similar. Because there is no single universally accepted measure of the quality of alignments, we have chosen two popular, negatively correlated with one another measures: AP and RSDM, corresponding to sensitivity and 1—specificity, respectively, which taken together allow us to make a quite comprehensive comparison of alignments.

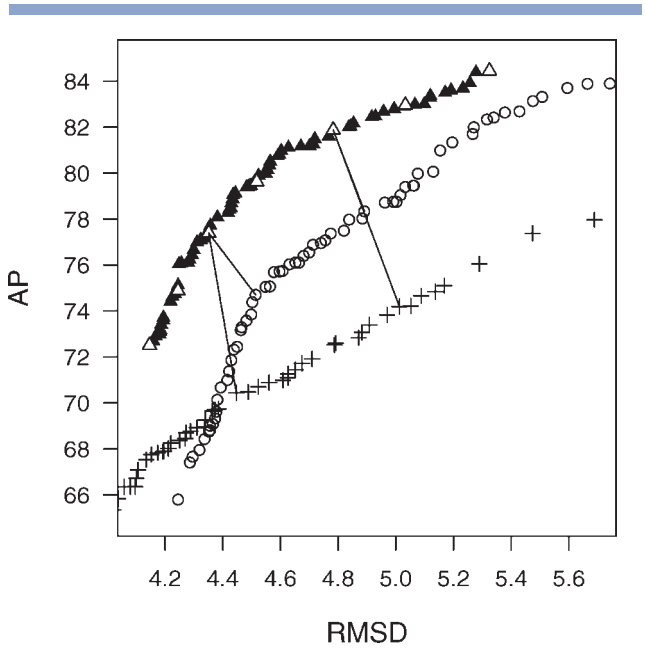

Figure 5

Effects of gap penalties (values not shown) and different approximations to the Mueller, Spang and Vingron matrix.[12] *Ordinate values are alignment percentage AP and abscissa values are RMSD values. Points* $(r_i(\alpha, \beta), a_i(\alpha, \beta))$ *for all admissible parameters* (α, β) *for the* MSV *matrix and its two- and three-vector approximations are shown. Triangles denote scores for the* MSV *matrix, pluses for the two-vector approximation and circles for the three-vector approximation, respectively. Open triangles denote seven points listed in columns 4–5 of Table V. The segments between data points indicate loss of alignment quality due to the approximations for the 3rd and the 5th rows in Table V.*

CONCLUSIONS

1. The approximations (2)–(4) allow us to comprehend better the exchangeability of AAs as well as the residue-residue interactions in proteins. In the Methods section we have shown that there are four different types of approximation factors. Now we would like to give them a physical interpretation. The four types of approximating factors could be treated as ideal forms of these phenomena and could be conveniently named as follows: *positive additivity* if $\boldsymbol{x} \geq 0$ (all components of the vector are non-negative) and $c = 1$; *negative additivity* if $\boldsymbol{x} \geq 0$ and $c = -1$; s*eparability* if not $\boldsymbol{x} \geq 0$ and $c = 1$; c*oupling ability* if not $\boldsymbol{x} \geq 0$ and $c = -1$.

The use of the term *additivity* can be explained by the fact that the approximating factor that satisfies these conditions is always highly correlated with $x_i + x_j$. The case of positive additivity holds for the energy of transfer of a residue from water to a protein environment (we should note again that, in order to compare CPs and SMs we have changed the signs of the former matrices). The negative additivity case has not been observed here. *Separability* was the most frequently occurring form, since it is typical both for hydrophobic interactions in protein environments and for all approximating factors for SMs. The last ideal form, *coupling ability* rewards pairs *i–j* corresponding to components of the approximating vectors which have opposite signs. This form has been found for electrostatic interactions only.

To summarize, the approximating factors for substitution matrices are almost independent (mutual correlations are almost zero) and they divide all AAs into the following groups: hydrophobic/polar, large/small and occurring/absent in loops. On the other hand, for statistical potentials additionally to the separation of residues with respect to their hydrophobicity, we observe the coupling ability of electrostatic interactions.

2. From a practical point of view, our method of approximation can be useful for protein structure and function prediction. Specifically, it can significantly enhance the performance of algorithms which are based on alignments of the physicochemical profiles of protein sequences, such as fast threading[17] or multiple sequence alignment based on a fast Fourier transform (MAFFT).[18,19] We plan to study this problem in near the nearest future.

REFERENCES

1. Pokarowski P, Kloczkowski A, Jernigan RL, Kothari NS, Pokarowska M, Kolinski A. Inferring ideal amino acid interaction forms from statistical protein contact potentials. Proteins: Struct Funct Bioinformatics 2005;59:49–57.
2. Miyazawa S, Jernigan RL. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. Proteins: Struct Funct Genet 1999;34:49–68.
3. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–132.
4. Zimmerman JM, Eliezer N, Simha R. Characterization of amino acid sequences in proteins by statistical methods. J Theor Biol 1968;21:170–201.
5. Godzik A, Kolinski A, Skolnick J. Are proteins ideal mixtures of amino acids—analysis of energy parameter sets. Protein Sci 1995;4:2107–2117.
6. Li H, Tang C, Wingreen NS. Nature of driving force for protein folding: a result from analyzing the statistical potential. Phys Rev Lett 1997;79:765–768.
7. Tobi D, Shafran G, Linial N, Elber R. On the design and analysis of protein folding potentials. Proteins: Struct Funct Genet 2000;40:71–85.
8. Henikoff S, Henikoff JG. Amino-acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
9. May ACW. Towards more meaningful hierarchical classification of amino acid scoring matrices. Protein Eng 1999;12:707–712.
10. Tomii K, Kanehisa M. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng 1996;9:27–36.
11. Dosztanyi Z, Torda AE. Amino acid similarity matrices based on force fields. Bioinformatics 2001;17:686–699.
12. Muller T, Spang R, Vingron M. Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. Mol Biol Evol 2002;19:8–13.
13. Crooks GE, Brenner SE. An alternative model of amino acid replacement. Bioinformatics 2005;21:975–980.
14. Johnson MS, Overington JP. A structural basis for sequence comparisons—an evaluation of scoring methodologies. J Mol Biol 1993;233:716–738.

 DOI 10.1002/prot

15. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 1992;8:275–282.
16. Williams G, Doherty P. Inter-residue distances derived from fold contact propensities correlate with evolutionary substitution costs. BMC Bioinformatics 2004;5:153.
17. Cao HB, Ihm Y, Wang CZ, Morris JR, Su M, Dobbs D, Ho KM. Three-dimensional threading approach to protein structure recognition. Polymer 2004;45:687–697.
18. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 2002;30:3059–3066.
19. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 2005;33:511–518.
20. Muller T, Rahmann S, Rehmsmeier M. Non-symmetric score matrices and the detection of homologous transmembrane proteins. Bioinformatics 2001;17 (Suppl 1):S182–S189.
21. Rao CR. Linear statistical inference and its applications. New York: Wiley; 1973.
22. R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2005. http://www.R-project.org.
23. Kinjo AR, Nishikawa K. Eigenvalue analysis of amino acid substitution matrices reveals a sharp transition of the mode of sequence conservation in proteins. Bioinformatics 2004;20:2504–2508.
24. Maiorov VN, Crippen GM. Contact potential that recognizes the correct folding of globular proteins. J Mol Biol 1992;227:876–888.
25. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins: Struct Funct Genet 1999;34:82–95.
26. Mclachlan AD. Tests for comparing related amino-acid sequences cytochrome-C and cytochrome-C551. J Mol Biol 1971;61:409–424.
27. Grantham R. Amino-acid difference formula to help explain protein evolution. Science 1974;185:862–864.
28. Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure, Vol. 5, Suppl. 3. Washington, DC: National Biomedical Research Foundation; 1978. p 345–352.
29. Gonnet GH, Cohen MA, Benner SA. Exhaustive matching of the entire protein-sequence database. Science 1992;256:1443–1445.
30. Benner SA, Cohen MA, Gonnet GH. Amino-acid substitution during functionally constrained divergent evolution of protein sequences. Protein Eng 1994;7:1323–1332.
31. Jones DT, Taylor WR, Thornton JM. A mutation data matrix for transmembrane proteins. FEBS Lett 1994;339:269–275.
32. Qian B, Goldstein RA. Optimization of a new score function for the generation of accurate alignments. Proteins: Struct Funct Genet 2002;48:605–610.
33. Miyazawa S, Jernigan RL. A new substitution matrix for protein sequence searches based on contact frequencies in protein structures. Protein Eng 1993;6:267–278.
34. Naor D, Fischer D, Jernigan RL, Wolfson HJ, Nussinov R. amino acid pair interchanges at spatially conserved locations. J Mol Biol 1996;256:924–938.
35. Azarya-Sprinzak E, Naor D, Wolfson HJ, Nussinov R. Interchanges of spatially neighbouring residues in structurally conserved environments. Protein Eng 1997;10:1109–1122.
36. Wei L, Altman RB, Chang JT. Using the radial distributions of physical features to compare amino acid environments and align amino acid sequences. Pac Symp Biocomput 1997;5:465–476.
37. Russell RB, Saqi MAS, Sayle RA, Bates PA, Sternberg MJE. Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. J Mol Biol 1997;269:423–439.
38. Ogata K, Ohya M, Umeyama H. Amino acid similarity matrix for homology modeling derived from structural alignment and optimized by the Monte Carlo method. J Mol Graph Modell 1998; 16:178–254.
39. Ng PC, Henikoff JG, Henikoff S. PHAT: a transmembrane-specific substitution matrix. Bioinformatics 2000;16:760–766.
40. Prlic A, Domingues FS, Sippl MJ. Structure-derived substitution matrices for alignment of distantly related sequences. Protein Eng 2000;13:545–550.
41. Blake JD, Cohen FE. Pairwise sequence alignment below the twilight zone. J Mol Biol 2001;307:721–735.
42. Lin K, May ACW, Taylor WR. Amino acid substitution matrices from an artificial neural network model. J Comput Biol 2001;8:471–481.
43. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
44. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
45. Bradley P, Chivian D, Meiler J, Misura KMS, Rohl CA, Schief WR, Wedemeyer WJ, Schueler-Furman O, Murphy P, Schonbrun J, Strauss CEM, Baker D. Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. Proteins: Struct Funct Genet 2003;53:457–468.
46. Kawashima S, Kanehisa M. AA index: amino acid index database. Nucleic Acids Res 2000;28:374.
47. Levitt M. Simplified representation of protein conformations for rapid simulation of protein folding. J Mol Biol 1976;104:59–107.
48. Krigbaum WR, Komoriya A. Local interactions as a structure determinant for protein molecules. II. Biochim Biophys Acta 1979; 576:204–228.
49. Chou PY, Fasman GD. Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol Relat Areas Mol Biol 1978;47:45–148.
50. Levitt M. Conformational preferences of amino-acids in globular proteins. Biochemistry 1978;17:4277–4284.
51. Fauchere JL, Charton M, Kier LB, Verloop A, Pliska V. Amino-acid side-chain parameters for correlation studies in biology and pharmacology. Int J Pept Protein Res 1988;32:269–278.
52. Lifson S, Sander C. Antiparallel and parallel beta-strands differ in amino-acid residue preferences. Nature 1979;282:109–111.
53. Palau J, Argos P, Puigdomenech P. Protein secondary structure—studies on the limits of prediction accuracy. Int J Pept Protein Res 1982;19:394–401.
54. Robson B, Suzuki E. Conformational properties of amino-acid residues in globular proteins. J Mol Biol 1976;107:327–356.
55. Sueki M, Lee S, Powers SP, Denton JB, Konishi Y, Scheraga HA. Helix coil stability-constants for the naturally-occurring amino-acids in water. 22. Histidine parameters from random poly[(hydroxybutyl)-glutamine-*co*-L-histidine]. Macromolecules 1984;17:148–155.
56. Aurora R, Rose GD. Helix capping. Protein Sci 1998;7:21–38.
57. Vihinen M, Torkkila E, Riikonen P. Accuracy of protein flexibility predictions. Proteins: Struct Funct Genet 1994;19:141–149.
58. Munoz V, Serrano L. Intrinsic secondary structure propensities of the amino-acids, using statistical phi–psi matrices—comparison with experimental scales. Proteins: Struct Funct Genet 1994;20:301–311.
59. Mardia KV, Kent JT, Bibby JM. Multivariate analysis. New York: Academic Press; 1979.
60. Mizuguchi K, Deane CM, Blundell TL, Overington JP. HOMSTRAD: a database of protein structure alignments for homologous families. Protein Sci 1998;7:2469–2471.
61. Gotoh O. An improved algorithm for matching biological sequences. J Mol Biol 1982;162:705–708.
62. Needleman SB, Wunsch CD. A general method applicable to search for similarities in amino acid sequence of 2 proteins. J Mol Biol 1970;48:443–453.
63. Green RE, Brenner SE. Bootstraping and normalization for enhanced evaluations of pairwise sequence comparison. Proc IEEE 2002;90:1834–1847.