Amino Acid Substitution Matrices Based on 4-Body Delaunay Contact Profiles

Ahmet Sacan and I. Hakki Toroslu

Computer Engineering Department,
Middle East Technical University
Ankara, TURKEY
Email: [ahmet,toroslu]@ceng.metu.edu.tr

Abstract

Sequence similarity search of proteins is one of the basic and most common steps followed in bioninformatics research and is used in making evolutionary, structural, and functional inferences. The quality of the search and the alignment of the protein sequences depend crucially on the underlying amino-acid substitution matrix. We present a method for deriving amino acid substitution matrices from 4-body contact propensities of amino-acids in 3D protein structures. Unlike current popular methods, our method does not rely on mutational analysis, evolutionary arguments, or alignment of protein sequences or structures. The alignment accuracy of our derived matrices is evaluated using the BAliBASE reference alignment set and is found to be comparable to that of popular matrices from the literature. Notably, the metric subset of our matrices outperform other available metric matrices. Our matrices will be useful especially in the development of empirical potential energy functions and in distance-based sequence indexing.

Supplementary Material: Our substitution matrices and detailed alignment data can be obtained from http://www.ceng.metu.edu.tr/~ahmet/bioinfo/distmat

Keywords: Protein Structure, Delaunay tessellation, Similarity Search, Sequence Alignment, Amino Acid Substitution Matrix, Metric Matrix

I. Introduction

Alignment of protein sequences has been one of the most widely utilized tools of bioinformatics research [8]. Applications of sequence alignment and comparison include finding homologous proteins, predicting protein structure or function, and defining the phylogeny of the species.

An alignment score is defined as the sum of individual scores of the aligned residues as looked up from a residue scoring matrix, and is used in database searches for similar sequences. Optimal alignment score of two sequences can be obtained using the dynamic programming algorithm [19, 27]. Rapid increase in the size of protein sequence databases has rendered calculation of the optimal alignments unfeasible and has prompted the development of near-optimal heuristic approaches like BLAST [1] and FASTA [21].

The quality and significance of database search results and sequence alignments depend strongly on the underlying residue scoring matrix and the gap cost function. For computational convenience, affine gap penalty is used in practice [10] and gap opening and gap extension penalties are determined by statistical optimization on a reference alignment set.

The popular scoring matrices are based on the log-likelihood of residue substitutions obtained from the frequencies of mutations observed in the sequence alignment of similar proteins. The initial alignments were constructed either by hand [4], by automated alignments from large sequence databases [9] or by the alignment of conserved blocks [12].

Structural superpositions have also served as a basis for alignment of sequences and counting of substitutions [14]. Protein structures can be aligned even in the absence of significant sequence similarity. Substitution matrices derived from structural alignments are especially useful in detecting distantly related sequences and similarities that result from convergent evolution.

Other methods of obtaining residue exchangeability include evaluation of engineered mutations either by experimental assay studies [30], or by computational fitness functions such as those based on force fields [5]. Physicochemical properties such as hydrophobicity, volume, and conformational preferences have also been used as a basis

for similarity measures [11, 20].

In this study, we use the multi-body contact propensities of residues in three dimensional protein structures as the basis for amino-acid similarity. Amino-acids have previously been found to have non-random multi-body contact preferences [26] and this property has been exploited in development of statistical pseudo-potentials to discriminate native and non-native protein conformations [15]. We use these non-random preferences to derive an amino-acid scoring matrix to be used in protein sequence alignments. We expect that this scoring matrix will be suitable for detecting remote homologs that share structural similarities. Moreover, the unique features of this matrix make it especially useful in the development of contact-based empirical potential energy functions and in the distance-based indexing of protein sequences.

Substitution matrices that form metric-distance functions are highly desirable in the distance-based indexing of protein sequences. A subset of the matrices developed in this study form metric-distance functions, such that the following properties are satisfied for any three amino acids x, y, and z:

1) *Identity:* $d(x, y) = 0$ iff $x = y$
2) *Positivity:* $d(x, y) \geq 0$
3) *Symmetry:* $d(x, y) = d(y, x)$
4) *Triangle Inequality:* $d(x, y) \leq d(x, z) + d(y, z)$

Sensitive metric matrices are a prerequisite to the development of fast sequence analysis algorithms that are both scalable and sensitive [29]. The metric matrices we derived outperform previous metric metrices in alignment accuracy.

II. Methods

Due to its objective and robust definition and well–defined geometric properties, Delaunay tessellation has been the method of choice for extracting multi-body contacts from protein structures [26]. The protein is modeled by a set of points representing the amino-acids. The region of space around each point that is closer to the enclosed point than any other point defines a Voronoi polyhedron. (See Figure 1). Delaunay tessellation is obtained by connecting points that share a Voronoi boundary. In 2D, each triangular area in the Delaunay tessellation defines a set of 3 points that are in contact. In 3D, each tetrahedra gives a set of 4-body contacts.

There are several ways of representing amino acids of a given protein structure. Here, we use the most commonly used representations: location of the alpha Carbon atom (CA), location of the beta Carbon (CB), or the centroid of the side-chain atoms ($CENT$). Glycine lacks a CB atom, so for Glycine, CA is used instead of CB. For each of these representations, the Delaunay tessellation is computed using the Quickhull algorithm [2].

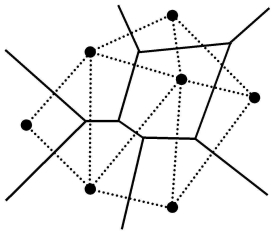

Fig. 1. Delaunay tessellation (dashed lines) and Voronoi diagram (solid lines) of a set of points in 2D. In 3D, Delaunay tessellation gives space-filling tetrahedra.

For a given protein structure, the Delaunay tessellation results in a list of amino-acid quadruplets defining the 4-body contacts. We record the frequency of observing an amino acid type in contact with the remaining three amino acids in the quadruplet. This gives us a frequency matrix of size 20 by 8000, where each row stands for an amino acid type, and each column represents different combinations of the remaining three amino acids. We call each row of this matrix the *4-body contact profile* of the corresponding amino acid.

We postulate that the exchangeability of amino acids in three dimensional structures would be reflected in their Delaunay contact profiles. An amino acid substitution can thus be derived from the contact profiles matrix. We use both the Euclidean distance (EUC) and Pearson's correlation (COR) measures between the rows of the contact profiles matrix in order to quantify the exchangeability of amino-acids. The Euclidean distance is defined as:

$$d_{a,b}^{EUC} = \sqrt{\sum_{i=1}^{8000} (A_i - B_i)^2}$$

where $d_{a,b}^{euc}$ is the calculated distance between amino acids a and b and where A_i and B_i are the i^{th} elements of the corresponding rows of the contact profile matrix. Similarly, the correlation distance is defined as:

$$d_{a,b}^{COR} = 1 - \frac{\sum_{i=1}^{8000} (A_i - \mu_A)(B_i - \mu_B)}{\sqrt{\sum_{i=1}^{8000} (A_i - \mu_A)^2 \sum_{i=1}^{8000} (B_i - \mu_B)^2}}$$

where μ_A denotes the mean value of the row A of the contact matrix. Each of these distance functions define a target 20 by 20 amino acid substitution matrix. We refer to these matrices as the *Euclidean matrices* and the *correlation matrices*, respectively, in this presentation.

III. Experiments

The PDBselect25 [13] representative dataset, which contains a non-redundant set of PDB (Protein Data Bank [3]) structures with less than 25% mutual sequence identity, was used for the derivation of contact profiles and the construction of substitution matrices. The downloaded version of PDBselect25 used in this study was compiled in January 2007 and contained 3080 proteins.

Using three types of amino acid representations (CA, CB, and CENT) and two types of distance measures (EUC and COR), a total of six substitution matrices were obtained. The derived matrices were compared with 16 other matrices from the literature (see Table I for a list of matrices). For completeness, an identity matrix was also included. Comparison and analysis of matrices were performed via principal component analysis and hierarchical clustering. These methods have been noted to be sufficient to highlight the overall relationship between matrices [16].

Figure 2 shows a gray-scale depiction of matrix correlations based on sample correlations of their 400 elements. An unweighted average distance (UPGMA) clustering of matrices from these correlations is also obtained (Figure 3). The type of amino acid representation (CA, CB, and CENT) does not have a significant effect on the resulting matrix as can be seen from the high correlations among the corresponding matrices. This is due to only a small fraction of the tetrahedrons differing among the tessellations obtained from different amino acid representations.

On the other hand, the choice of distance measure gives qualitatively different matrices. The Euclidean measure is sensitive to the background frequencies of amino acids in the initial protein structure dataset, and the *Euclidean matrices* reflect this bias. Whereas, the correlation coefficient results in exchange values normalized for the background frequencies of amino acids.

The *correlation matrices* (CA-COR, CB-COR, and CENT-COR) are found to be most closely correlated to the NAOR [18] substitution matrix with an average correlation coefficient of 0.76. NAOR has been derived from amino acid interchanges observed at spatially, locally conserved regions in globally dissimilar and unrelated proteins. Although Delaunay tetrahedra form a more granular motif, we conjecture that the tetrahedra contacts derived in this study share common overall characteristics with the conserved substructural motifs studied by Naor et al. [18]. Note that Delaunay tessellations have, in fact, been found useful in discovering locally conserved structural sites [25].

Unlike the *correlation matrices*, the *Euclidean matrices* (CA-EUC, CB-EUC, and CENT-EUC) do not show significant correlation with any other substitution matrix. We attribute this to the inherent bias of the Euclidean measure

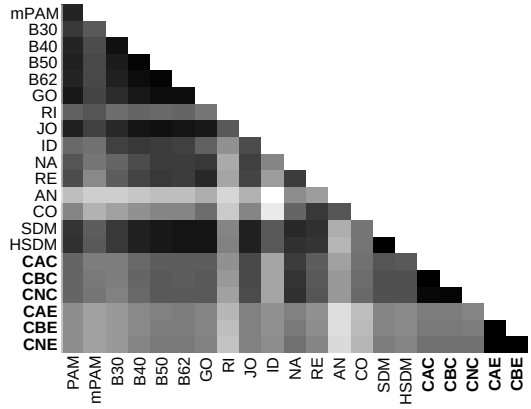

Fig. 2. Correlation of matrices based on the pairwise sample correlation of matrix elements. The higher the correlation between a pair of matrices, the darker the corresponding cell.

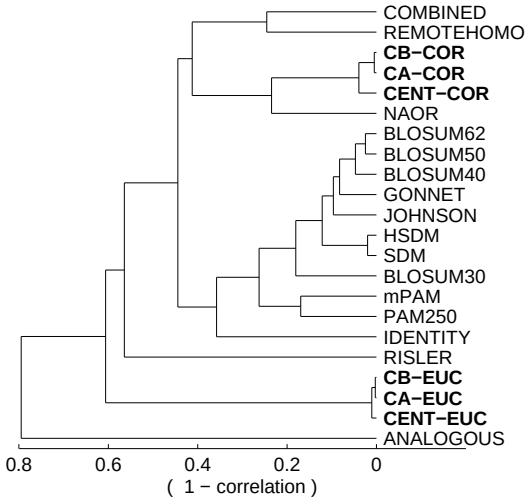

Fig. 3. UPGMA clustering of matrices based on correlation of matrix elements

to the background amino acid frequencies; this bias is not present in the other matrices.

Analysis of the matrices at the amino acid level can help characterize the physico-chemical properties underlying the amino acid exchanges. We observed that the exchange values defined in the *correlation matrices* are strongly related to the hydrophobicity of the amino acids. In Figure 4, the substitution matrix CA-COR is represented as a projection on the first two principal components. The first principal component is essentially a hydrophobicity scale with hydrophobic residues on the left, and hydrophilic and charged residues clustered on the right.

Matrix name	Short name	Reference	Based on
PAM250	PAM	[4]	sequence alignment of similar proteins
mPAM	mPAM	[29]	expected time between mutations in aligned sequences
BLOSUM30,40,50,62	B30,40,50,62	[12]	sequence alignment of conserved blocks in related proteins
GONNET	GO	[9]	exhaustive automated sequence alignments
RISLER	RI	[23]	structural alignment of related proteins
JOHNSON	JO	[14]	structure based sequence comparison
MIYAZAWA	MJ	[17]	base substitution – protein stability
NAOR	NA	[18]	structural alignment of spatially conserved substructural motifs
REMOTEHOMO	RE	[24]	structural alignment of remote homologs
ANALOGOUS	AN	[24]	structural alignment of analogous proteins
COMBINED	CO	[24]	structural alignment of analogous and remote homologs
SDM	SDM	[22]	structurally equivalent residues of analogous proteins
HSDM	HSDM	[22]	structurally equivalent residues of homologous proteins
CA-COR	CAC	present study	correlation of Delaunay contact profiles from CA atoms
CB-COR	CBC	present study	correlation of Delaunay contact profiles from CB atoms
CENT-COR	CNC	present study	correlation of Delaunay contact profiles from side chain centers
CA-EUC	CAE	present study	Euclidean distance of Delaunay contact profiles from CA atoms
CB-EUC	CBE	present study	Euclidean distance of Delaunay contact profiles from CB atoms
CENT-EUC	CNE	present study	Euclidean distance of Delaunay contact profiles from side chain centers
IDENTITY	ID	present study	identity matrix

TABLE I. Substitution matrices used for comparison

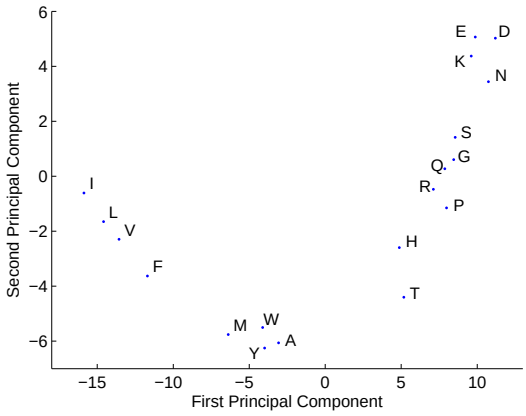

Fig. 4. Principal component analysis of the matrix CA-COR. The first and second principal components account for the 72.7% and 24.7% of the variation in the matrix values. Cysteine residue with coordinates -18.2,20.2 is omitted from the figure for illustration purposes. The analysis for the other matrices can be found in the Supplementary Material.

The first eigenvector of the CA-COR matrix and the hydrophobicity scale of Fasman [6] are indeed highly correlated, with a coefficient of 0.93. The strong correlation with hydrophobicity scales is no surprise because protein folding and, as a result, the Delaunay contacts are guided by hydrophobic interactions among amino acids residues.

In order to evaluate the sequence alignment performance of our derived matrices, we used the BAliBASE [28, version 3] suite of reference alignments. Pairwise alignments for each multiple alignment were extracted to result in a total of 155,550 pairs of sequences. For each substitution matrix, pairwise alignment of sequences was performed using Gotoh's algorithm [10] with affine gap penalties. The optimal gap penalties were used as found by Prlic et al. [22]. For matrices where optimal gap penalties were not available, we used parameters interpolated from those of the PAM250 matrix.

Substitution Matrix	Alignment accuracy
SDM	63.0
HSDM	62.6
GONNET	61.8
BLOSUM30	60.6
PAM250	60.3
MIYAZAWA	58.9
RISLER	58.7
NAOR	58.6
CA-COR	**58.5**
CB-COR	**58.3**
BLOSUM40	58.2
CENT-COR	**57.9**
CENT-EUC	**56.0**
CA-EUC	**55.7**
CB-EUC	**55.7**
BLOSUM50	54.9
REMOTEHOMO	54.9
BLOSUM62	54.8
JOHNSON	53.7
mPAM	52.5
IDENTITY	22.5
COMBINED	19.1
ANALOGOUS	14.4

TABLE II. Alignment accuracy of matrices

The performance of a matrix is defined as the percentage of the correctly aligned residues compared to the reference alignment. The summary results of the sequence aligments are tabulated in Table II. The break-down of the results to individual BAliBASE subsets can be obtained

from the Supplementary Material. The ranking of the matrices obtained here for the BAliBASE dataset are comparable to those found by Prlic et al. [22] on their smaller data set of 122 protein pairs. Only PAM250 had a significantly higher performance ranking on the BAliBASE database, compared to the rankings in [22].

The performance of substitution matrices depend on the degree of similarity of the aligned sequences, with lower scores for sequences that have lower sequence identity. However, the ranking of matrix performance is found to be similar across different BAliBASE subsets. The performance of our derived matrices are comparable to that of other matrices. Among our derived matrices, the *correlation matrices* perform slightly better than the *Euclidean matrices*.

The Euclidean set of similarity matrices (CA-EUC, CB-EUC, and CENT-EUC) have a notable feature of being metric; their corresponding distance matrices satisfy *positivity*, *symmetry*, and *triangle inequality*. This is a natural outcome of the underlying metric Euclidean measure used for obtaining these matrices. The *Euclidean matrices* outperform mPAM [29] in alignment accuracy, which was previously shown to be more sensitive than other available metric matrices.

IV. Discussion

We have generated 4-body Delaunay contact profiles from a non-redundant set of protein structures. The contact profiles of amino acid residues were then used to derive an amino acid substitution matrix. We have investigated the effects of using different amino acid representations and different contact profile distance measures on the resulting matrix.

The *correlation matrices* were closely related to the NAOR [18] substitution matrix which is derived from amino acid substitutions observed at locally conserved but globally unrelated protein structures. Furthermore, principal component analysis shows a strong correlation of the *correlation matrices* with the hydrophobicity scale of amino acids. This is of no surprise, because hydrophobicity inherently guides the contacts formed in the protein structures.

Alignment accuracies of our matrices have also been illustrated using the BAliBASE multiple alignment dataset as reference. To the best of our knowledge, this is the first study to compare the alignment accuracies of the popular matrices on the comprehensive BAliBASE dataset. The performance of our matrices was comparable to that of other matrices. It is interesting to see that the Delaunay contact profile matrices we have derived, which do not rely on any evolutionary arguments or on observed substitution rates, can perform so well.

We believe that the matrices we have derived based on the unique principles of Delaunay contact profiles make an important contribution to the set of available amino acid substitution matrices. Multiple scoring matrices can be used to increase the reliability and significance of sequence alignments [7]. Another advantage of the availability of various scoring matrices based on different principles is the ability to select appropriate matrices for specialized problems.

In applications where distance, rather than similarity, between sequences is relavant, the similarity matrix is converted to a dissimilarity matrix by subtracting each matrix element from the maximum value of the matrix. However, commonly used matrices fail to meet the conditions of metric distance function. Unequal values along the diagonal of the commonly used matrices violate the *identity* condition, resulting in positive distance values of a sequence to itself, which is undesirable in distance-based similarity measures. The elements of the matrices also violate the triangle inequality, which is a prerequisite to sensitive sequence indexing methods.

A unique feature of the *Euclidean matrices* is the satisfaction of the metric-distance conditions. Moreover, the *Euclidean matrices* perform better than other metric matrices and approach the commonly used non-metric matrices in alignment accuracy. This makes the *Euclidean matrices* especially suitable in providing sensitive distance measures between sequences and in scalable distance-based indexing of protein databases for fast retrieval of similar sequences.

V. Acknowledgements

We would like to thank Heidi Nemeth for her valuable comments and suggestions on the draft of this manuscript.

References

[1] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. Gapped BLAST and PSIBLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[2] C.B. Barber, D.P. Dobkin, and H.T. Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22(4):469–483, 1996.

[3] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. E. Meyer Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The protein data bank: A computer-based archival file for macromolecular structures. *J. of. Mol. Biol.*, 112:535, 1977.

[4] M. O. Dayhoff, R. M. Schwartz, and B.C. Orcutt. A model of evolutionary change in proteins. *In Atlas of Protein Sequence and Structure 5(3) M.O. Dayhoff (ed.), National Biomedical Research Foundation, Washington.*, pages 345–352, 1978.

[5] Z. Dosztanyi and A. E. Torda. Amino acid similarity matrices based on force fields. *Bioinformatics*, 17(8): 686, 2001.

[6] G.D. Fasman. *Prediction of Protein Structure and the Principles of Protein Conformation*. Plenum, New York, 1989. Table XVII.

[7] F. Frommlet, A. Futschik, and M. Bogdan. On the significance of sequence alignments when using multiple scoring matrices. *Bioinformatics*, 20(6):881–887, 2004.

[8] R. S. C. Goble, P. Baker, and A. Brass. A classification of tasks in bioinformatics. *Bioinformatics*, 17: 180188, 2001.

[9] G. H. Gonnet, M. A. Cohen, and S. A. Benner. Exhaustive matching of the entire protein sequence database. *Science*, 256:1443–1445, 1992.

[10] O. Gotoh. Optimal sequence alignment allowing for long gaps. *Bull. Math. Biol.*, 52:359–373, 1990.

[11] R. Grantham. Amino acid difference formula to help explain protein evolution. *Science*, 185:862–864, 1974.

[12] S. Henikoff and J.G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, 89:10915–10919, 1992.

[13] U. Hobohm and C. Sander. Enlarged representative set of protein structures. *Protein Science*, 3:522–524, 1994.

[14] M.S. Johnson and J.P. Overington. A structural basis for sequence comparisons. an evaluation of scoring methodologies. *J. Mol. Biol.*, 233:716–738, 1992.

[15] B. Krishnamoorthy and A. Tropsha. Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. *Bioinformatics*, 18(12):1540–1548, 2003.

[16] A.C. May. Towards more meaningful hierarchical classification of amino acid scoring matrices. *Protein Eng.*, 12:707–712, 1999.

[17] S. Miyazawa and R.L. Jernigan. A new substitution matrix for protein sequence searches based on contact frequencies in protein structures. *Protein Eng.*, 6: 267–278, 1993.

[18] D. Naor, D. Fischer, R.L. Jernigan, H.J. Wolfson, and R. Nussinov. Amino acid pair interchanges at spatially conserved locations. *J. of Mol. Biol.*, 256 (5):924–938, 1996.

[19] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol Biol.*, 48:443, 1970.

[20] K. Niefind and D. Schomburg. Amino acid simiarity coefficients for protein modeling and sequence alignment derived from main-chain folding angles. *J. Mol. Biol.*, 219:481–497, 1991.

[21] W. A. Pearson. Rapid and sensitive sequence comparison with fastp and fasta. *in Methods in Enzymology ed. R. Doolittle Academic Press, San Diego 183*, pages 63–98, 1990.

[22] A. Prlic, F.S. Domingues, and M.J. Sippl. Structure-derived substitution matrices for alignment of distantly related sequences. *Protein Eng.*, 13:545–550, 2000.

[23] J.L. Risler, M.O. Delorme, H. Delacroix, and A. Henaut. Amino acid substitutions in structurally related proteins. a pattern recognition approach. determination of a new and efficient scoring matrix. *J. Mol. Biol.*, 204:1019–1029, 1988.

[24] R.B. Russell, M.A.S. Saqi, R.A. Sayle, P.A. Bates, , and M.J.E. Sternberg. Recognition of analogous and homologous protein folds: Analysis of sequence and structure conservation. *Journal of Molecular Biology*, 269:423–439, 1997.

[25] A. Sacan, O. Ozturk, H. Ferhatosmanoglu, and Y. Wang. Lfm-pro: A tool for detecting significant local structural sites in proteins. *Bioinformatics*, 2007.

[26] R. K. Singh and A. Tropsha. Delaunay tessellation of proteins: Four body nearest neighbor propensities of amino acid residues. *J. Comput. Biol.*, 3:213–221, 1996.

[27] T. Smith and M. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[28] J.D. Thompson, F. Plewniak, and O. Poch. Balibase: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, 15(1): 87–88, 1999.

[29] W. Xu and D. P. Miranker. A metric model of amino acid substitution. *Bioinformatics*, 20(8):1214–21, 2004.

[30] L. Y. Yampolsky and A. Stoltzfus. The exchangeability of amino acids in proteins. *Genetics*, 170: 1459–1472, 2005.