STATISTICAL GEOMETRY ANALYSIS OF PROTEINS: IMPLICATIONS FOR INVERTED STRUCTURE PREDICTION

A. TROPSHA[1], R. K. SINGH[2], I. I. VAISMAN[1], and W. ZHENG[1]

[1]*Laboratory for Molecular Modeling, School of Pharmacy, and*
[2]*Microelectronic Systems Laboratory, Department of Computer Science,*
University of North Carolina at Chapel Hill, Chapel Hill, NC 27599

The topology of folded proteins from the representative dataset of well-defined three-dimensional protein structures is studied using a statistical geometry approach. Amino acid residues in protein chains are represented by C_α atoms, thus reducing the protein three-dimensional structure to a set of points in three dimensional space. The Delaunay tessellation of a protein structure generates an aggregate of space-filling irregular tetrahedra, or Delaunay simplices. Each simplex objectively defines four nearest neighbor C_α atoms, i.e. four nearest neighbor residues. The statistical analysis of residue composition of Delaunay simplices reveals nonrandom preferences for certain quadruplets of amino acids. These nonrandom preferences are used to develop a fitness function that evaluates sequence-structure compatibility. Using this fitness function, several tested native proteins score the highest among 100,000 random sequences with average protein amino acid composition. The statistical geometry approach, based solely on first principles, provides a unique means for protein structure analysis and has direct implications for inverted protein structure prediction.

1 Introduction

Accurate prediction of protein three-dimensional (3D) structure from its primary sequence represents one of the greatest challenges of modern theoretical biology. It is still experimentally much easier to determine a protein's primary sequence than its 3D structure. As a result, the size of protein primary sequence databases (e.g. PIR and Swiss-Prot[1]) exceeds that of the experimentally determined 3D protein structure database[2] by at least an order of magnitude. This disparity will likely increase over time. Therefore, the most attractive means for obtaining information about 3D protein structure is to predict it from the protein primary sequence.

The goal of predicting 3D protein structures from primary sequences has given rise to a number of techniques which can be divided into three major categories: potential energy based analysis, lattice simulations of protein folding, and knowledge based approaches[3]. Potential energy based methods include molecular mechanics optimizations and molecular dynamics (MD)[4]. MD has proven extremely useful in refinement of experimentally determined structures[5]. However, even the MD equilibration of systems already near equilibrium (e.g. starting from crystallographically or spectroscopically obtained coordinates) requires substantial computational resources. The prediction of folded protein structures by dynamic simulations is currently computationally prohibitive. Furthermore, it has been shown that deliberately misfolded structures often have much lower molecular mechanics potential energies than the native structures[6].

Monte Carlo lattice simulations of protein folding with simplified potentials using single point (C_α atoms) representation of amino acid residues have led to reasonable predictions of approximate protein folds and, in a few cases, to accurate predictions of several simple protein structures[7,8]. Due to high computational cost and inadequacy of the simplified potentials, the predictions are currently limited to fairly small proteins or stable structural motifs such as coiled coils[9]. However, it has been shown that the C_α-based representation of protein 3D structure is sufficient for reliable restoration of the complete backbone structure and, with a reasonable accuracy, a full atom structure, including the side chains[10].

Knowledge-based methods of protein 3D structure prediction rely on the analysis of sequence-structure relationships in known protein folds. According to a recent evaluation, the number of different protein folds may be limited to about 1000[11]. A significant amount of effort by several research groups has been focused on the area of inverted protein structure prediction[12-14]. These methods are based on the statistical analysis of amino acid preferences for particular secondary structures, combined with two-body and, in some cases, three-body[12] propensities of amino acids to be clustered together in folded proteins. From this analysis, sequence-structure compatibility scores are derived for each amino acid, and the prediction is achieved by "threading" new protein sequence through known protein structural templates in order to locate the most compatible template. These methods lead in principle to full atom predictions of protein architecture and have been shown in several cases to outperform other methods in the accuracy of structure prediction.

The accuracy of knowledge based 3D structure prediction can be improved by a systematic application of statistical and pattern matching techniques to the comparison, alignment, and classification of known protein structures. In this paper, we employ the Delaunay tessellation of folded proteins for unambiguous identification of all clusters of four nearest neighbor residues in any protein structure. The statistical analysis of the amino acid composition of the nearest neighbor quadruplets provides a novel set of tetrabody residue potentials and a new sequence-structure compatibility scoring function. Thus, the results of this study have direct implication for inverted protein structure prediction.

2 Methods

The statistical geometry approach for studying structure of disordered systems was introduced by Bernal[15]. He suggested characterization of structural disorder by statistical analysis of irregular polyhedra obtained as a result of a specific tessellation in three-dimensional space. The method, including the design and implementation of practical algorithms, was further developed by Finney for the case of Voronoi tessellation[16]. A Voronoi tessellation partitions the space into convex polytopes called Voronoi polyhedra. For a molecular system the Voronoi polyhedron is the region of space around an atom, such that all points of this region are closer to this atom than to any other atom of the system. A group of four atoms, whose Voronoi polyhedra meet at one vertex, forms another basic topological object, the Delaunay simplex. The topological difference between these tessellations is that the Voronoi polyhedron describes the coordination of the nearest atomic environment while the Delaunay simplex describes the ensemble of neighboring atoms. Although the Voronoi polyhedra and the Delaunay simplices are completely determined by each other, Voronoi polyhedra may differ topologically (having different number of faces and edges), while the Delaunay simplices are always topologically equivalent (they are always tetrahedra in three-dimensional space) and can be compared quantitatively. The Delaunay tessellation was used for structural analysis of various disordered systems and in most cases served as a valuable tool for structure description[17,18].

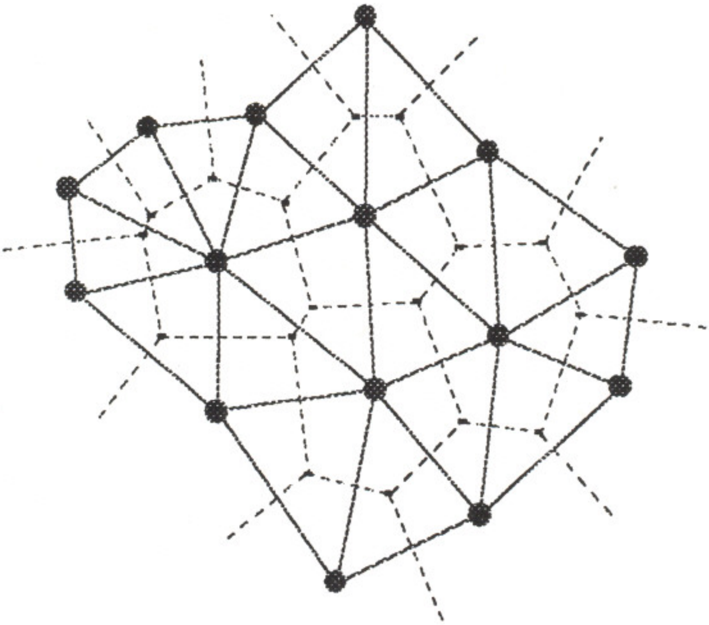

Figure 1: Voronoi (dashed line) and Delaunay (solid line) tessellations in two dimensions

The Delaunay tessellation was performed on the dataset of unique protein structures identified by Jones et al[19]. This dataset contains 103 protein chains with high crystallographic resolution that do not have apparent structural similarity and carry low sequence identity. This dataset was successfully used by the authors to identify unique protein folds and provides an adequate database for statistical analysis of protein structure.

The proteins in the Jones' list were analyzed in a pipeline fashion as follows: (i) preprocessing of raw PDB files; (ii) Delaunay tessellation in 3D space; (iii) statistical analysis of residue composition of Delaunay simplices. First step is the extraction of the necessary 3D coordinates of C_α atoms from the PDB entry file. Then, the Delaunay tessellation is performed using the **qhull** program developed by Barber et al[20] and distributed by the University of Minnesota Geometry Center. The program produces the Delaunay tessellation from the convex hull of a set of points in general N-dimensions by computing a convex hull using a randomized incremental algorithm. After the tessellation is done, the **pdb** program takes the PDB entry file and the tessellation results from the **qhull** program as input and computes various characteristics of tetrahedra and their constituent residues. For this work we were interested mainly in the amino acid composition and the geometry of the simplices which was analyzed using the **qfc** program. Both **pdb** and **qfc** programs were written in the C programming language. All calculations were performed on a HP-9000/735 workstation running HP-UX operating system. The cumulative wall clock time for analyzing a protein structure (through all the phases) for a typical protein was on the order of 10 seconds.

3 Results and discussion

3.1 Delaunay tessellation of folded protein structures

The typical result of the Delaunay tessellation of a folded protein is shown in Figure 2 for crambin (the Brookhaven code 1crn). The tessellation of this 46-residue protein generates an aggregate of 192 nonoverlapping, space-filling irregular tetrahedra or Delaunay simplices. Each Delaunay simplex uniquely defines four nearest neighbor C_α atoms, i.e., four nearest neighbor amino acid residues, as vertices of this simplex. A vertex may be shared by several tetrahedra. Thus, individual amino acid residues may have different number of neighbors. For instance, in crambin, as many as 15 edges may originate from a common vertex.

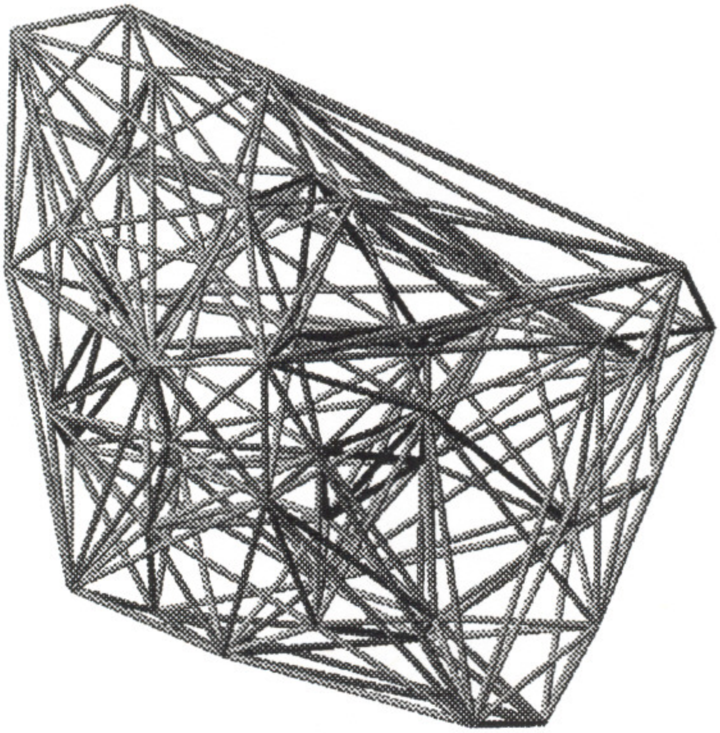

Figure2: Delaunay tessellation of Crambin

Thus, the Delaunay tessellation may in principle define all neighbors of a given residue. However, the Delaunay tessellation emphasizes the fundamental property of a set of three-dimensional point objects where four and only four nearest neighbors could be identified unambiguously.

3.2 Statistical analysis of the composition of Delaunay simplices

The Delaunay tessellation of 103 protein chains in the dataset generates a total of 114,617 simplices. The composition of these simplices was analyzed in terms of statistical likelihood of occurrence of four nearest neighbor amino acid residues for all observed quadruplet combinations of 20 natural amino acids. The agglomeration factor q was calculated for each quadruplet from the Eq.1:

$$q_{ijkl} = log\frac{f_{ijkl}}{p_{ijkl}} \tag{1}$$

where i,j,k,l are any of the 20 natural amino acid residues, f_{ijkl} is the observed normalized frequency of occurrence of a given quadruplet, and p_{ijkl} is the expected frequency of occurrence of a given quadruplet. The q_{ijkl} shows the likelihood of finding four particular residues in one simplex. The f_{ijkl} is calculated by dividing the total number of occurrence of each quadruplet type by the total number of observed quadruplets of all types. The p_{ijkl} was calculated from the Eq. 2:

$$p_{ijkl} = Ca_i a_j a_k a_l \tag{2}$$

where a_i, a_j, a_k,and a_l denote the individually observed frequency of occurrence of each amino acid residue (i.e. total number of occurrences of each residue type divided by the total number of amino acid residues in the dataset), and C is the combination factor, defined as

$$C = \frac{4!}{\prod_{i}^{n}(t_i!)} \tag{3}$$

where n is the number of distinct residue types in a quadruplet and t_i is the number of amino acids of type i, where i ranges from 1 to n. The factor C accounts for the underestimation due to permutability of replicated residue types.

We have first analyzed the composition of the Delaunay simplices in terms of well known chemical classes of the amino acid side chains. The amino acid residues were classified as hydrophobic (F), hydrophilic (L), and polar (P) types[21]; hydrophobic amino acids include Ala, Val, Phe, Ile, Leu, Pro, Met, hydrophilic amino acids include Asp, Glu, Lys, Arg, and polar amino acids include Ser, Thr, Tyr, Cys, Asn, Gln, His, Trp; this consideration reduces the 20-letter amino acid alphabet to a three-letter code. Since we have been interested in the analysis of amino acid contacts that may produce physico-chemical interaction, we

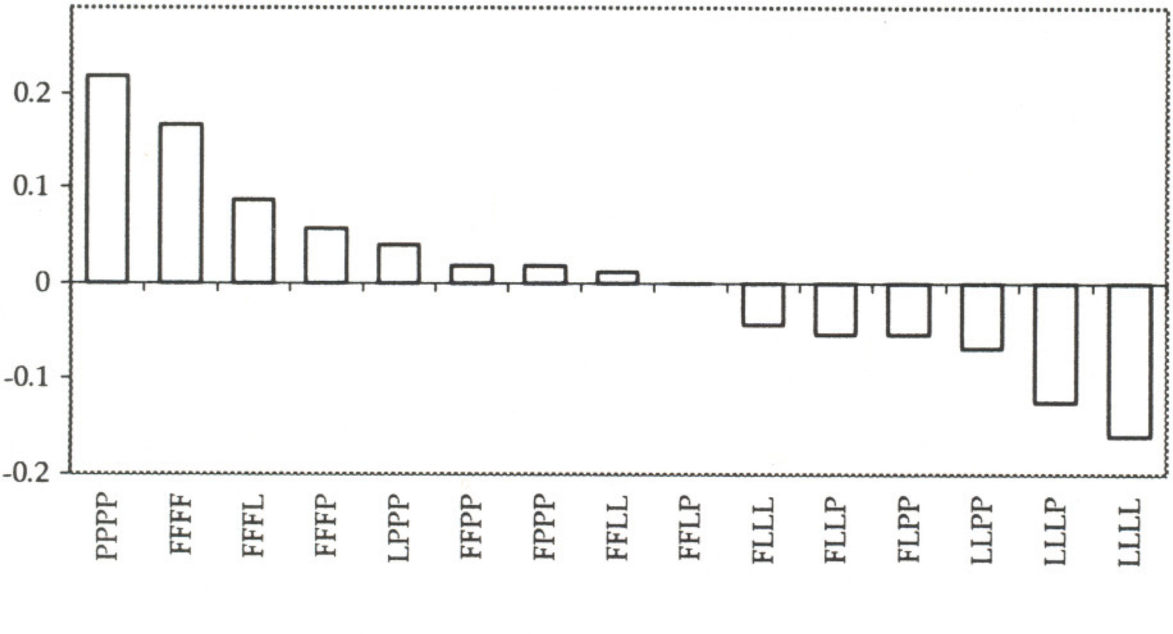

Figure 3 : Delaunay simplices with distinct composition (3 letter alphabet)

have introduced a 7 Å distance cutoff to qualify amino acid residues in one tetrahedron as nearest neighbors; thus, tetrahedra with at least one edge exceeding 7 Å were excluded from the analysis. (For comparison, 5.5 to 7.5 Å cutoff distances between Cα atoms are customarily used in the distance geometry based classification of amino acids as nearest neighbors). Figure 3 shows the log-likelihood ratio for fifteen possible quadruplet combinations of the three types of amino acids among all simplices in tessellated proteins of the dataset. This ratio is calculated as the observed frequency of occurrence of each quadruplet divided by the expected (theoretical) frequency of occurrence. Quadruplets containing four or three residues of types P and F are much more likely to occur than the ones with four or three type L residues.

Theoretically, the maximum number of all possible quadruplets of natural amino acid residues is 8,855 whereas only 8,351 occur in the dataset. The agglomeration factor q is plotted in Figure 4 for all observed quadruplets of amino acids. Each quadruplet is thus characterized by a certain value of the q factor which describes the nonrandom bias for the four amino acid residues to be found in the same Delaunay simplex. This value can therefore be interpreted as a four-body potential for the quadruplets of amino acid to be nearest neighbors in 3D protein space. Hence, based on the data of Figure 4, for each native tessellated protein the total score can be calculated as the sum of individual scores for all composing Delaunay simplices. The resulting value is considered as an estimate of the sequence-structure compatibility score for the native protein.

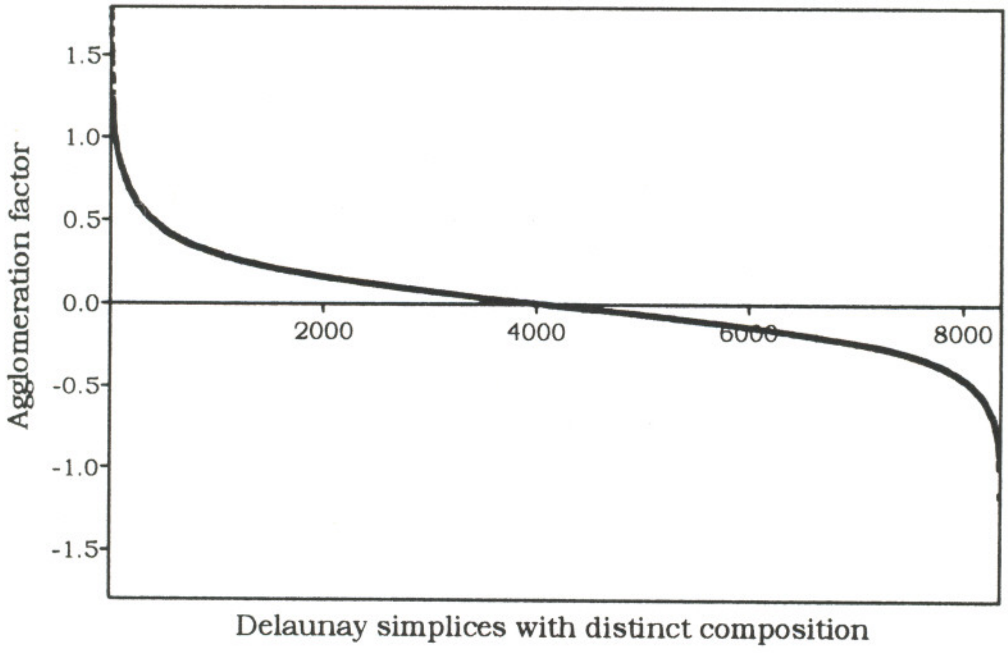

Figure 4: Agglomeration factor for all simplices in the dataset

3.3 Implications for inverted structure prediction

We have evaluated the proposition that the agglomeration factor accurately reflects the compatibility of native sequence with native structure; this is an essential step towards inverted structure prediction. We have chosen three proteins of different length from the dataset: hemoglobin (1eca), flavodoxin (4fxn), and papain (9pap).

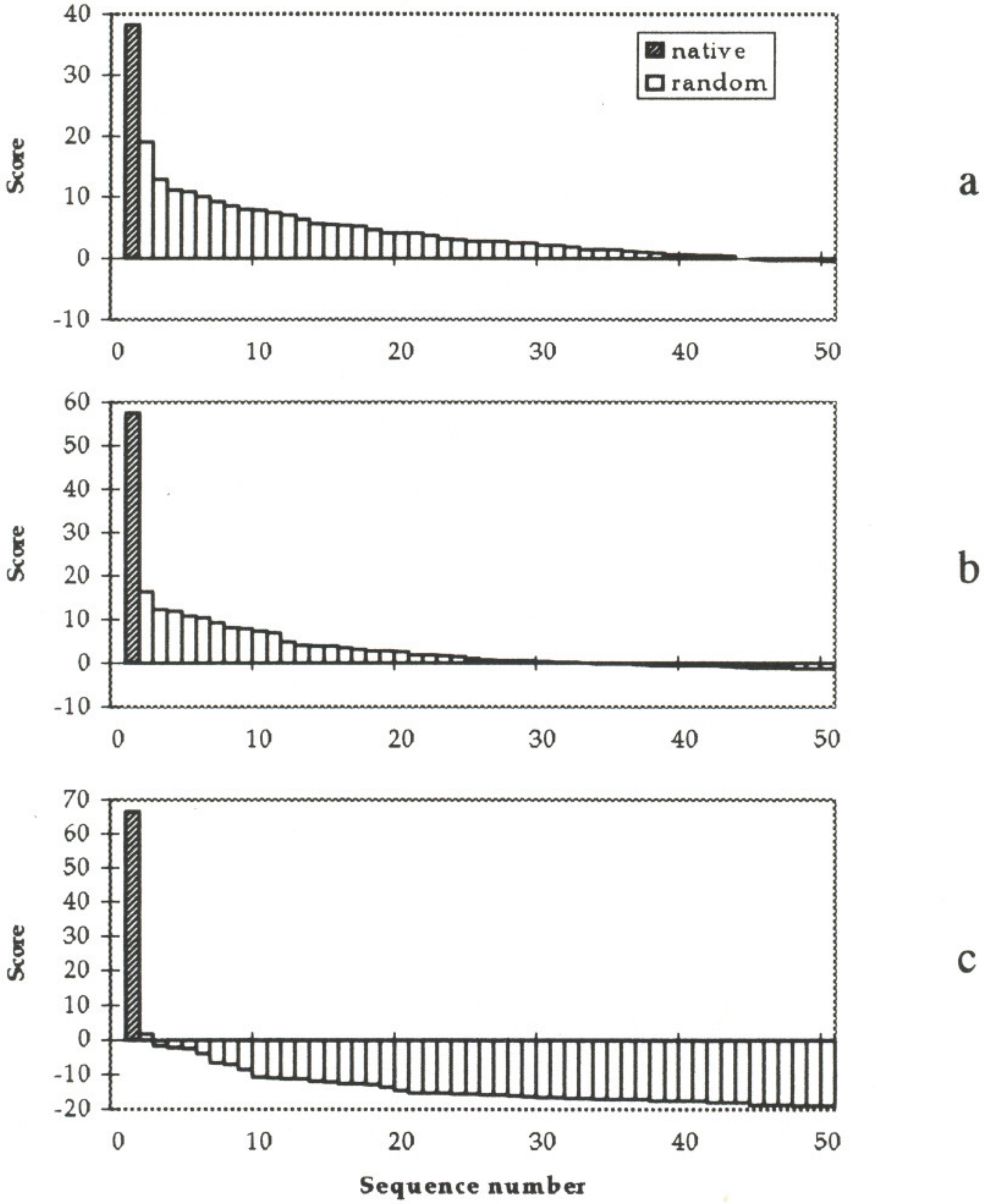

Figure 5: Scores for native and random sequences fitted into Delaunay-based template of hemoglobin (a), flavodoxin (b), and papain (c).

For each of these proteins, 100,000 random sequences of the same length as the native protein were generated as follows. An amino acid was assigned randomly to each position in a sequence according to its observed frequency of occurrence in the training dataset. This procedure generates amino acid sequences of different composition, but ensures that the ensemble of generated sequences has the same observed frequency of occurrence for each amino acid on average. Once a random sequence has been generated and assigned (threaded) to a template, the new amino acid composition and corresponding agglomeration factor for each Delaunay simplex of the template was determined. For the Delaunay simplices that were not observed in the training dataset, the value of agglomeration factor was set to zero. The total sequence/structure compatibility score was calculated as the sum of the agglomeration factors for all compositions of the Delaunay simplices of the random sequence. The results of experiments are presented in Figure 5. As can be seen in this Figure, in all cases the native protein has the highest score. One may hypothesize that the protein sequences that scored close to the native structure may in fact have similar fold. This hypothesis may be further tested experimentally.

4 Conclusions

The analysis of residue contacts in folded proteins provides important information about the topology and stability of protein structures. In order to identify all sets of nearest neighbor residues in proteins we have employed the Delaunay tessellation of protein structure, where each amino-acid residue was represented by its C_α atom. Delaunay tessellation ensures unambiguous definition of the sets of four nearest neighbors. Statistical analysis of residue composition of Delaunay simplices reveals nonrandom preferences for certain combinations of residues. We calculated the log likelihood for all observed quadruplets of amino acid residues. Based on the values of log likelihood we have derived a novel sequence-structure compatibility scoring function. This function is used to discriminate between the native and any random sequences for a given native 3D structure template. The results of this work should aid in further development of methods for the analysis and prediction of protein structure from sequence.

Acknowledgments

This research was supported in part by the NSF grant MIP-9024585, HCC/UNC-CH Partnership Discovery Grant, and by a grant from the Pharmacy Foundation of N.C., Inc.

References

1. Bairoch A. and Boeckmann B., *Nucl. Acid Res.* **21**, 3093 (1993)
2. Bernstein F.C., Coetzle T.F., Williams G.J.B., Meyer E.F. Jr, Brice M.D., Rogers J.R., Kennard O., Shimanouchi T., and Tasumi M., *J Mol Biol* **112**, 535 (1977)
3. Eisenhaber F., Persson, B., and Argos, P., *Crit. Rev. in Biochem. and Mol. Biol.* **30**, 1 (1995)
4. Karplus, M. and Petsko, G.A., *Nature* **347**, 631 (1990)
5. Brunger, A.T. and Nilges, M., *Q. Rev. Biophys.* **26**, 49 (1993)
6. Le Grande, S. M. and Mertz, K. in The Protein Folding Problem and Tertiary Structure Prediction (Birkhäuser, Boston, 1994)
7. Yue, K., Fiebig, K.M., Thomas, P.D., Chan, H.S., Shakhnovich, E.I., and Dill, K.A., *Proc. Natl. Acad. Sci. U.S.A.* **92**, 325 (1995)
8. Gutin, A.M., Abkevich, V.I., and Shakhnovich, E.I., *Biochemistry* **34**, 3066 (1995)
9. Vieth, M., Kolinski, A., Brooks, C.L., and Skolnick, J., *J. Mol. Biol.* **237**, 361 (1994)
10. Rey, A. and Skolnick, J. *J. Comput. Chem.* **13**, 443 (1992)
11. Chothia, C. *Nature*. **357**, 543 (1992)
12. Godzik, A., Kolinski, A., and Skolnick, J., *J. Mol. Biol.* **227**, 227 (1992)
13. Bowie, J.U., Luthy, R., and Eisenberg, D., *Science* **253**, 164 (1991)
14. Bryant, S.H. and Lawrence, C.E. *Proteins*. **16**, 92 (1993)
15. Bernal, J.D. *Nature* **183**, 141 (1959)
16. Finney, J.L., *Proc.R.Soc.* **A319**, 479 (1970); Finney, J.L., *Nature* **266**, 309 (1977)
17. Medvedev, N.N., Voloshin, V.P., and Naberukhin, Y.I., *J.Phys.A:Math.Gen.* **21**, L247 (1988)
18. Vaisman, I.I., Brown, F.K., and Tropsha A., *J.Phys.Chem.* **98**, 5559 (1994)
19. Jones, D.T., Taylor W.R., and Thornton, J.M. *Nature* **358**, 86 (1992)
20. Barber, C. B., Dobkin, D. P., and Huhdanpaa, H. *Tech. Rep.* **GCG53**, (Geometry Center, University of Minnesota, Minneapolis, 1994)
21. Branden, C and Tooze, J. Introduction to Protein Structure (Garland Publishing, New York and London, 1991)