J. Comput. Biol. 1996, **3**, 2, 213-222
Submitted August 18, 1995

Delaunay Tessellation of Proteins: Four Body Nearest Neighbor Propensities of Amino Acid Residues

Raj K. Singh, Alexander Tropsha and Iosif I. Vaisman*
University of North Carolina at Chapel Hill, Chapel Hill, NC 27599

Abstract

Delaunay tessellation is applied for the first time in the analysis of protein structure. By representing the location of amino acid residues in protein chains by C_α atoms, the protein structure is described as a set of points in three-dimensional space. Delaunay tessellation of a protein structure generates an aggregate of space-filling irregular tetrahedra, or Delaunay simplices. The vertices of each simplex define objectively four nearest neighbor C_α atoms, i.e. four nearest neighbor residues. A simplex classification scheme is introduced in which simplices are divided into five classes based on the relative positions of vertex residues in the protein primary sequence. Statistical analysis of the residue composition of Delaunay simplices reveals nonrandom preferences for certain quadruplets of amino acids to be clustered together. This nonrandom preference may be used to develop a four-body potential that can be used in evaluating sequence-structure compatibility for the purpose of inverted structure prediction.

Introduction

Analysis of known protein topologies is an important component in understanding protein folding. Recent research indicates that all protein structures may be classified using a limited number of protein folds (Chothia 1992), which provides a foundation for inverted structure prediction methods (Bowie et al. 1991). These methods rely on the analysis of sequence-structure relationships in known protein folds and the problem of structure prediction for a protein sequence is formulated in terms of finding existing structural templates (folds) which are most compatible with this sequence. Sequence-structure compatibility is estimated based on empirical potential energy functions (Sippl 1995). An unknown structure is predicted by threading its primary sequence through known protein structural templates and finding a set of templates with the lowest potential energy (Bryant and Altschul 1995).

Potential energy functions for protein folding simulation and structure prediction are derived based on the statistical analysis of nearest neighbors in proteins. The majority of these

* E-mail to ivaisman@email.unc.edu

functions account for pairwise interactions only; however consideration of cooperative interactions of higher order may improve the quality of structure prediction (see recent reviews by Johnson et al. 1994, Sippl 1995).

Most methods of protein structure analysis require a definition of nearest neighbor residues. Existing definitions of nearest neighbors are often based on arbitrary distance criteria (e.g. separation of C_α atoms by no more than 5.5 Å (Yee et al. 1994), the separation between any pair of atoms that belong to different residues by less than 2.8 Å (Behe et al. 1991), the separation in the range of distance from 4.5 to 7.5 Å (Miyazawa and Jernigan 1985) or a set of separations between various pairs of atoms (Crippen and Maiorov 1994)). Because the definition of a contact between two residues is not objective, the results of such analyses strongly depend on the chosen criteria of the contact. Therefore, an objective and robust definition of nearest neighbor residues in proteins would eliminate one of the dominant sources of ambiguity in methods for inverted structure prediction and protein folding simulation.

An objective definition of nearest neighbors in three-dimensional space can be obtained by applying the methods of statistical geometry. The statistical geometry approach for studying structure of disordered systems was introduced by Bernal (Bernal 1959). He suggested characterization of structural disorder using statistical analysis of irregular polyhedra obtained by a specific tessellation in the three-dimensional space. The method, including the design and implementation of practical algorithms, was further developed by Finney for the case of Voronoi tessellation (Finney 1970, 1977). 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 each point of this region is closer to the atom than to any other atom of the system. A group of four atoms whose Voronoi polyhedra meet at a common vertex forms another basic topological object called a Delaunay simplex. The procedure for constructing Voronoi polyhedra and Delaunay simplices in two dimensions is illustrated in Figure 1. The topological difference between these objects is that the Voronoi polyhedron represents the environment of individual atoms whereas the Delaunay simplex represents the ensemble of neighboring atoms. Although the Voronoi polyhedra and the Delaunay simplices are completely determined by each other, there exists a significant difference. Whereas the Voronoi polyhedra may differ topologically (i.e., they may have different numbers of faces and edges), the Delaunay simplices are always topologically equivalent (i.e., in three-dimensional space they are always tetrahedra). Delaunay tessellation has been used for structural analysis of various disordered systems. In most such cases it has served as a valuable tool for structure description (Voloshin et al. 1988, Vaisman et al. 1994). In this paper we report for the first time the use of Delaunay tessellation to define objectively the

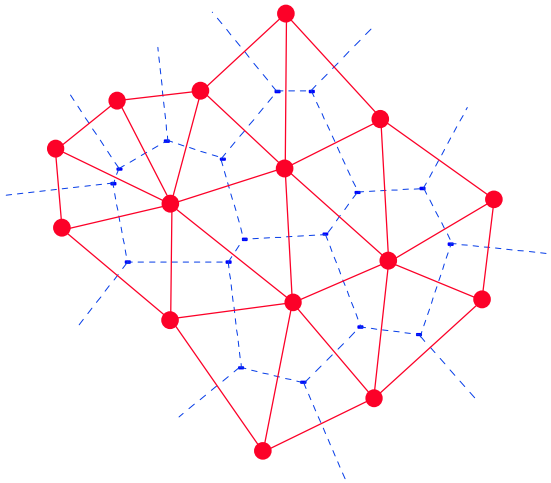

Figure 1. Voronoi/Delaunay tessellation in 2D space (Voronoi tessellation - dashed line, Delaunay tessellation - solid line).

nearest neighbor residues in 3D protein structures. The most significant feature of Delaunay tessellation, as compared with other methods of nearest neighbor identification, is that the number of nearest neighbors in three dimensions is always four, which represents a fundamental topological property of 3D space. Statistical analysis of the amino acid composition of Delaunay simplices provides information about spatial propensities of all quadruplets of amino acid residues to be clustered together in folded protein structures. The empirical four-body contact potentials derived from this analysis may significantly improve the results of protein structure prediction.

Methods

Delaunay tessellation is a canonical tessellation of space based on nearest neighbors (Aurenhammer 1991, Sugihara 1995). A Delaunay tessellation of a set of points is equivalent to a convex hull of the set in one higher dimension (Barber et al. 1993). For example, to determine the Delaunay tessellation of a set of points in 3D, we lift the points to a paraboloid and compute their convex hull in 4D. In general, a (d+1)-dimensional convex hull of a set of points is a simplicial complex which is represented by its vertices, d-dimensional facets (or simplices), and lists of adjacent facets.

Since Delaunay tessellation is performed on a set of points, we represent these points using only the C_α atoms for each residue (it has been shown (Rey 1992) that such reduced representation of proteins is adequate for accurate restoration to the full atomic backbone structure). The first step in this process is extracting the set of 3D coordinates of the C_α atoms from the PDB entry file. Delaunay tessellation of this set of points is then performed using the program **qhull** which implements the Quickhull algorithm developed by Barber et al. (Barber et al. 1993) and is distributed by the University of Minnesota Geometry Center. The Quickhull algorithm is a variation of the randomized, incremental algorithm of Clarkson and Shor. The program **qhull** produces the Delaunay tessellation by computing the convex-hull of this set of points in four dimensions and is shown to be space and time efficient. The results of the tessellation are analyzed with the **qstat** program which computes various geometrical properties and compositional statistics of Delaunay simplices. Computer software was developed using the C programming language. Wall clock time required for analyzing a typical protein structure using an HP-9000/735 workstation is on the order of 10 seconds.

Delaunay tessellation was performed on three representative sets of high-resolution proteins structures with low primary-sequence identity (Jones et al. 1992, Fischer 1994, and Hobohm and Sander 1994). The results of our analysis were nearly identical for all three datasets; the results obtained using the Jones' list are presented in this paper. The dataset contains 103 protein chains with high crystallographic resolution that do not have apparent structural similarity and share less than 30% sequence identity. The entries for the proteins in the Jones' list were extracted from Brookhaven Protein Data Bank (PDB; Bernstein et al. 1977). Several PDB entries form the original list, namely 1abp, 1cd4, 1cy3, 1gcr, 1lrd1, 1pcy, 1sn3, were updated to more recent, 5abp, 3cd4, 2cy3, 4gcr, 1lmb1, 1plc, 2sn3, respectively.

Results and discussion

Delaunay tessellation of folded protein structures. The typical results of Delaunay tessellation of a folded protein are illustrated on Figure 2a for crambin (1crn). The tessellation of this 46-residue protein generates an aggregate of 192 nonoverlapping, space-filling irregular tetrahedra (Delaunay simplices). Each Delaunay simplex uniquely defines four nearest neighbor C_α atoms (vertices of the simplex), i.e., four nearest neighbor amino acid residues.

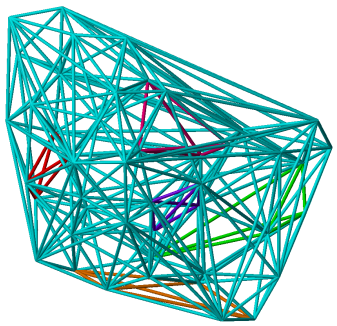

Figure 2a: Delaunay tessellation of Crambin

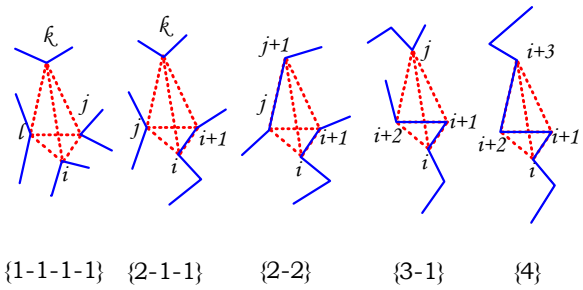

Figure 2b. Five classes of Delaunay simplices

For the analysis of correlations between the structure and sequence of proteins, we introduced a classification of simplices based on the relative positions of vertex residues in the primary sequence. Two residues were defined as distant if they were separated by one or more residues in the protein primary sequence. Simplices were divided into five nonredundant classes: class {4}, where all four residues in the simplex are consecutive in the protein primary sequence; class {3,1}, where three residues are consecutive and the fourth is a distant one; class {2,2}, where two pairs of consecutive residues are separated in the sequence; class {2,1,1}, where two residues are consecutive, and the other two are distant both from the first two and from each other; and class {1,1,1,1} where all four residues are distant from each other (Figure 2b). All five classes usually occur in any particular protein.

We first investigated differences between classes of simplices using geometrical parameters of tetrahedra such as volume and tetrahedrality. Tetrahedrality is a quantitative measure of the degree of distortion of the Delaunay simplices from the ideal tetrahedron:

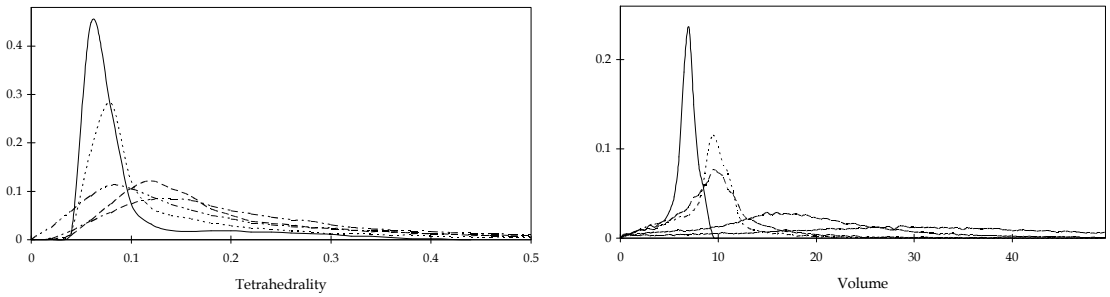

Figure 3. Distribution of tetrahedrality and volume (in $Å^3$) of Delaunay simplices

$$T = \sum_{i>j} (l_i - l_j)^2 \; / \; 15 \, \bar{l}^{\,2} \qquad (1)$$

where l_i is the length of the *i*-th edge, and $\bar{l}$ is the mean length of the edges of the given simplex. Distributions of volume and tetrahedrality for all five classes of simplices is shown in Figure 3. The sharpest peaks correspond to the simplices of classes {4} and {2,2}. They tend to have the lowest volume and lowest distortion of tetrahedrality. These results suggest that tetrahedra of these two classes may occur in regular protein conformations such as α-helices.

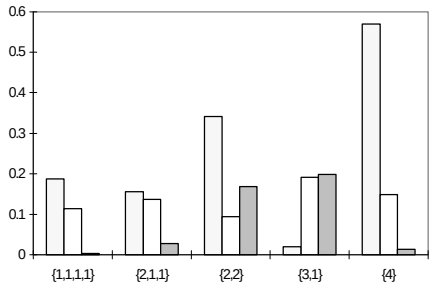

Figure 4. Classes of Delaunay simplices and secondary structures
(helices - dots, b-sheets - stripes, coils - empty)

In order to verify this hypothesis, we analyzed the possible correlations between the classes of simplices and the conventional secondary structure assignment of the constituent residues. The secondary structure assignments of individual residues were extracted from the headers of PDB files. We have considered three conformational states for all residues: helical (H), β-strand (S), and random coil (C). There are 15 possible combinations of these three conformational states in sets of four residues. Figure 4 presents the frequency of occurrence of each Delaunay simplex class where all four residues are found in the same conformational state. As can be seen from this Figure, there are certain correlations between secondary structure and simplex class. For example, simplices of classes {4} and {2,2} are formed mainly by residues in helical conformations. On the other hand, the residues in α-helical conformation almost never form simplices of class {3,1}. Residues in β-sheet conformation almost never form classes {4} and {1,1,1,1} simplices but frequently form simplices of classes {2,2} and {3,1}.

The observed correlations between regular conformations of protein backbone and classes of tetrahedra (Figure 4) suggest that the ratio of tetrahedra of different classes in a protein may be characteristic of a protein fold family. We have calculated the relative frequency of occurrence of tetrahedra of each class in each protein in the Jones' dataset and plotted the results in Figure 5. The proteins were sorted in the ascending order of fraction of tetrahedra of class {4}. Interestingly, the content of simplices of class {3,1} (which are indicative of β-sheet conformation; cf. Figure 4) decreases with the increase of the content of class {4} simplices. According to common classifications of protein fold families (Richardson and Richardson 1989, Orengo 1995), at the top level of hierarchy most proteins can be characterized as all-alpha, all-beta, or alpha/beta. The fold families for the proteins in the Jones' dataset are also shown in Figure 5. The results of Figure 4 suggest that proteins having a high content of tetrahedra of classes {4} and {2,2} (i.e., proteins in the right-most part of the plot of Figure 5) belong to the family of all-alpha proteins. A comparison between conventional protein fold family assignment and relative frequency of tetrahedra of different classes confirms this hypothesis (cf. Figure 5). Similarly, proteins having a low content of tetrahedra of classes {4} and {2,2} but a high content of tetrahedra of classes {2,2} and {3,1} (i.e., proteins in the left-most part of the plot in Figure 5) belong to the all-beta protein fold family. Finally, proteins in the middle of the plot belong to the alpha/beta fold family. Thus, the results of this analysis show that the ratio of tetrahedra of

different classes is indicative of the protein fold family. Further systematic analysis using clustering techniques may lead to more rigorous classification of the fold families based solely on the classes of Delaunay simplices.

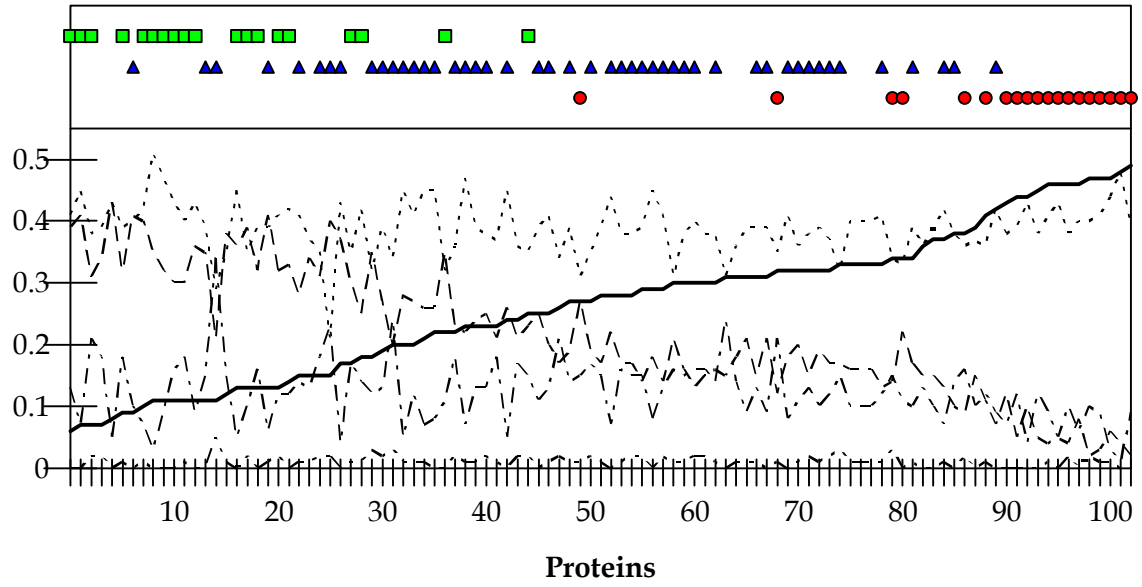

Figure 5. Classes of Delaunay simplices and protein fold families. Contents of simplices of class {4} (solid line), class {3,1} (dash), class {2,1} (dot), class {2,1} (dash-dot), class {1,1,1,1} (dash-dot-dot). Upper part of the figure displays fold family assignment: all-alpha (circles), all-beta (squares), and alpha-beta (triangles). Proteins are ordered as the following: 1-1hoe, 2-1tnfa, 3-2gcr, 4-2sodo, 5-2hlab, 6-2cna, 7-8atcb, 8-2fb4h, 9-1i1b, 10-2fb4l, 11-2rhe, 12-2sga, 13-3cd4, 14-4sgbi, 15-9wgaa, 16-2ltna, 17-2paba, 18-2pcy, 19-2stv, 20-3fxc, 21-4ptp, 22-2er7e, 23-1tgsi, 24-2ca2, 25-2sn3, 26-4cpai, 27-1ubq, 28-2azaa, 29-1paz, 30-5pti, 31-2ssi, 32-1csee, 33-4dfra, 34-7rsa, 35-3dfr, 36-2rnt, 37-4rxn, 38- 8adh, 39-1dhfa, 40-1gd1o, 41-2ovo, 42-9pap, 43-1csei, 44-5acn, 45-1hip, 46-2sns, 47-3grs, 48-2tmne, 49-1rhd, 50-2cdv, 51-3pgm, 52-2lbp, 53-3cla, 54-3icd, 55-5cpa, 56-1ctf, 57-1phh, 58-5abp, 59-1crn, 60-1fx1, 61-1pfka, 62-1wsyb, 63-4mdha, 64-1fd2, 65-3blm, 66-3gapa, 67-6ldh, 68-7cata, 69-1ccr, 70-1ypia, 71-2aat, 72-2gbp, 73-3pgk, 74-8atca, 75-1lz1, 76-1wsya1, 77-1wsya2, 78-1wsya3, 79-4fxn, 80-1cc5, 81-2cy3, 82-4xiaa, 83-2cyp, 84-2cpp, 85-3adk, 86-1bp2, 87-351c, 88-1lrp, 89-2cro, 90-1l01, 91-4cpv, 92-2ccya, 93-4tnc, 94-1mba, 95-1mbd, 96-1utg, 97-3hhba, 98-3icb, 99-1lh1, 100-256ba, 101-2mhr, 102-2wrpr, 103-1eca.

Statistical analysis of the composition of the Delaunay simplices and implications for inverted structure prediction. Delaunay tessellation of 103 protein chains in the Jones' dataset (Jones et al 1992) generates a total of 114,617 simplices. The composition of these simplices was first analyzed in terms of unbiased preferences for four amino acid residues to be clustered together. We analyzed the results of the Delaunay tessellation of these proteins 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 log-likelihood factor, q, for each quadruplet was calculated from the following equation:

$$q_{ijkl} = \log \frac{f_{ijkl}}{p_{ijkl}} \qquad (2)$$

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 randomly 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 following equation:

$$p_{ijkl} = C a_i a_j a_k a_l \quad (3)$$

where $\mathbf{a_i}$, $\mathbf{a_j}$, $\mathbf{a_k}$,and $\mathbf{a_l}$ are the observed frequencies of occurrence of individual 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 permutation factor, defined as

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

where n is the number of distinct residue types in a quadruplet and t_i is the number of amino acids of type i. The factor C accounts for the permutability of replicated residue types.

Theoretically, the maximum number of all possible quadruplets of natural amino acid residues is 8,855 whereas only 8,351 actually occur in the dataset. The log-likelihood factor q is plotted in Figure 6 for all observed quadruplets of natural 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 be interpreted as a four-body potential energy function. This function can be applied both for inverted structure prediction and in simulations of protein folding. This work is currently in progress, and the results will be described elsewhere.

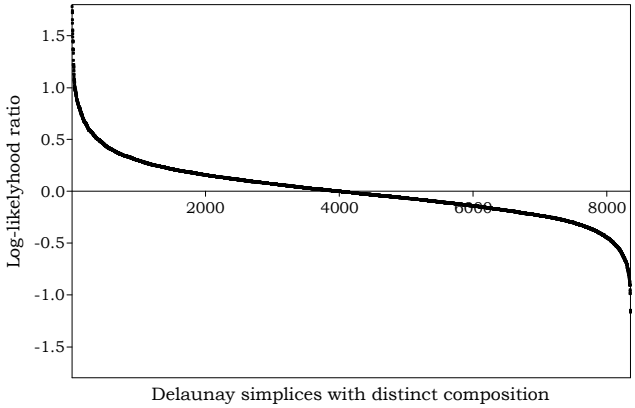

Figure 6. Log-likelihood ratio for the Delaunay simplices.

Conclusions

Existing methods of protein three-dimensional structure prediction would greatly benefit from a rigorous definition for the representation of nearest neighbor amino acid residues. In this paper, we introduced Delaunay tessellation as such a definition, and use it in the analysis of protein architecture. Delaunay tessellation is an objective and robust algorithm for identifying sets of four nearest neighbors in 3D space. We have identified all sets of nearest neighbor residues in a number of structurally diverse proteins and classified the corresponding Delaunay simplices on the basis of sequence proximity of the constituent residues. We have shown that simplices of each class are characterized by specific distributions of their geometrical characteristics (volume and tetrahedrality) and that this classification correlates with the conventional secondary structure assignment of composing residues. We have also shown that the ratio of tetrahedra of different classes in individual proteins may be used for protein classification.

This novel approach, based on first principles, provides a unique means for protein structure analysis and has direct implications for protein structure prediction. Statistical analysis of the amino acid composition of Delaunay simplices affords calculation of nonrandom preferences for all observed quadruplets of amino acids clustered together in folded proteins. These results provide a basis for calculation of four body potentials that can be used in simulations of protein folding and inverted structure prediction. These ideas are currently being implemented in our laboratories.

Acknowledgments

This research was supported in part by the NSF grant MIP-9024585 and the UNC School of Pharmacy seed grants. The authors greatly appreciate fruitful and stimulating discussions with Weifan Zheng and Doug Hoffman.

References

Aurenhammer F. 1991. Voronoi diagrams -- a survey of a fundamental geometric data structure. *ACM Computing Surveys*, 23, 345-405.

Barber C. B., Dobkin D. P., Huhdanpaa H. 1993. The quickhull algorithm for convex hull, Tech. Rep. GCG53, Geometry Center, University of Minnesota, Minneapolis, MN 55454.

Behe, M.J., Lattman, E.E., and Rose, G.D. 1991. The protein folding problem: the native fold determines packing, but does packing determine the native fold? *Proc. Natl. Acad. Sci. USA*. 88, 4195-4199.

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. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol* 112,535-542.

Bernal, J.D. 1959. *Nature* 183, 141-147.

Bowie, J.U., Luthy, R., and Eisenberg, D. 1991. A method to identify protein sequences that fold into a known three dimensional structure. *Science* 253, 164-70.

Bryant, S.H. and Altschul, S.F. 1995. Statistics of sequence-structure threading. *Cur. Opin. Struct. Biol.* 5, 236-244.

Bryant, S.H. and Lawrence, C.E. 1993. An empirical energy function for threading protein sequence through the folding motif. *Proteins*. 16, 92-112.

Chothia, C. 1992. One thousand families for the molecular biologist. *Nature*. 357, 543-544.

Crippen, G.M. and Maiorov, V.N. 1994. Contact potential for Global Identification of Correct Protein Folding. In *The Protein Folding Problem and Tertiary Structure Prediction*, Birkhäuser, Boston,

Eisenhaber F., Persson, B., and Argos, P. 1995. Protein Structure Prediction: Recognition of Primary, Secondary, and Tertiary Structural Features from Amino Acid Sequence. *Crit. Rev. in Biochem. and Mol. Biol.* 30, 1-94.

Finney, J.L. 1970. *Proc.R.Soc.* A319, 479-493.

Finney, J.L. 1977. *Nature* 266, 309-314.

Fisher, D. 1994. List of representative protein structures. *URL: http://www-lmmb.ncifcrf.gov/~nicka/lrs.html*

Godzik, A., Kolinski, A., and Skolnick, J. 1992. Topology fingerprint approach to the inverse protein folding problem. *J. Mol. Biol*. 227, 227-238.

Gutin, A.M., Abkevich, V.I., and Shakhnovich, E.I. 1995. Is burst hydrophobic collapse necessary for protein folding? Biochemistry. 34, 3066-76.
Hobohm, U. and Sander, C. 1994. Enlarged representative set of protein structures. *Protein Science*. 3, 522-524
Jones, D.T., Taylor W.R., and Thornton, J.M. 1992. A new approach to protein fold recognition. *Nature*, 358, 86-89.
Johnson, M.S., Srinivasan, N, Sowdhamini, R., and Blundell, T.L. 1994. Knowledge-based protein modeling. *Crit. Rev. in Biochem. and Mol. Biol.* 29, 1-68.
Miyazawa, S. and Jernigan, R.L. 1985. Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*. 18, 534-552.
Orengo, C.A., Michie, A.D., Thornton, J.M., and Taylor, W.R. CATH: Protein structure classification. *URL: http://www.biochem.ucl.ac.uk/bsm/cath/CATHintro.html*
Rackovsky, S. 1995. On the existence and implications of an inverse folding code in proteins. *Proc. Natl. Acad. Sci. USA*. 92, 6861-6863.
Rey, A. and Skolnick, J. 1992. Efficient algorithm for the reconstruction of a protein backbone from the alpha-carbon coordinates. *J. Comput. Chem*. 13, 443.
Richards, F.M. 1974. The interpretation of protein structures: total volume, group volume distribution and packing density. *J.Mol.Biol.* 82, 1-14.
Richardson, J.S. and Richardson, D.C. 1989. Principles and Patterns of Protein Conformation in: *Prediction of Protein Structure and the Principles of Protein Conformation*. Plenum Press, New York.
Sippl, M. 1995. Knowledge-based potentials for proteins. *Cur. Opin. Struct. Biol.* 5, 229-235.
Sugihara K. and Inagaki H. 1995. Why is the 3D Delaunay Triangulation Difficult to Construct? *Information Processing Letters*, 54, 275-280.
Vaisman, I.I., Brown, F.K., and Tropsha A. 1994. Distance dependence of water structure around model solutes. *J.Phys.Chem.* 98, 5559-5564.
Voloshin, V.P., Naberukhin, Y.I. and Medvedev, N.N. 1989. Can various classes of atomic configurations (Delaunay simplices) be distinguished in random dense packings of spherical particles? *Molec. Simulation*. 4, 209-227
Yee, D.P., Chan, H.S., Havel, T.F., and Dill, K.A. 1994. Does compactness induce secondary structure in proteins? *J. Mol. Biol*. 241, 557-573