James E. Gentle
Wolfgang Härdle
Yuichi Mori

Editors

Handbook of Computational Statistics

Concepts and Methods

With 236 Figures and 76 Tables

Springer

Statistical and Computational Geometry of Biomolecular Structure

IV.3

Iosif I. Vaisman

Introduction

Recent revolutionary developments in genomics and computational structural biology lead to the rapidly increasing amount of data on biomolecular sequences and structures. The deposition rate for both sequence and structure databases continues to grow exponentially. The efficient utilization of this data depends on the availability of methods and tools for biomolecular data analysis. Significant achievements have been made in DNA and protein sequence analysis, now the focus in bioinformatics research is shifting from sequence to structural and functional data. Accurate prediction of protein three-dimensional structure from its primary sequence represents one of the greatest challenges of modern theoretical biology. Detailed knowledge of protein structure is essential for understanding the mechanisms of biological processes at molecular, cellular, and evolutionary levels. The structures of only a fraction of all known primary sequences have been determined experimentally. Several approaches to protein structure prediction have been developed in recent years. Many of these approaches rely on the knowledge derived from the analysis of significant spatial and compositional patterns in known protein structures and understanding of the role these patterns play in the extremely complex processes, like protein folding or protein function. Such an analysis requires an objective definition of nearest neighbor residues that can be provided by the statistical geometry methods.

In the statistical geometry methods the nearest neighbor atoms or groups of atoms are identified by statistical analysis of irregular polyhedra obtained as a result of a specific tessellation in three-dimensional space. Voronoi tessellation partitions the space into convex polytopes called Voronoi polyhedra (Voronoi, 1908). 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 (Delaunay, 1934). The results of the procedure for constructing Voronoi polyhedra and Delaunay simplices in two dimensions are illustrated in Fig. 3.1. 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. The Voronoi polyhedra and the Delaunay simplices are topological duals and they are completely determined by each other. However the Voronoi polyhedra may have different numbers of faces and edges, while the Delaunay simplices are always tetrahedra in three-dimensional space. These tetrahedra can be used to define objectively the nearest neighbor entities in molecular systems.

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, it can be performed using the Quickhull algorithm developed by Barber et al. (1996). The Quickhull algorithm is a variation of the randomized, incremental algorithm of Clarkson

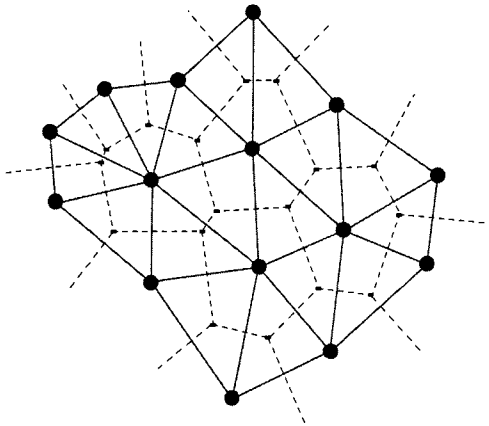

Figure 3.1. Voronoi/Delaunay tessellation in 2D space; Voronoi tessellation – *dashed line*, Delaunay tessellation – *solid line*

and Shor. The algorithm 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.

Statistical Geometry of Molecular Systems 3.2

A statistical geometry approach to study structure of molecular systems was pioneered by John Bernal, who in the late 1950 s suggested that "many of the properties of liquids can be most readily appreciated in terms of the packing of irregular polyhedra" (Bernal, 1959). Bernal pointed out that "it would be most desirable to find the true minimum number of parameters or parametral functions defining the statistical structure of any homogenous irregular assemblage in the way that the lattice vectors define a regular one" (Bernal, 1959). Methods of computational geometry, Voronoi and Delaunay tessellations in particular, may be used to address this problem. This approach, based on the Voronoi partitioning of space (Voronoi, 1908) occupied by the molecule, was further developed by Finney for liquid and glass studies (Finney, 1970). Finney proposed a set of "descriptive parameters" for packing of polyhedra in simple liquids. In the mid-1970 s the statistical geometry approach was first applied to study packing and volume distributions in proteins by Richards (1974), Chothia (1975) and Finney (1975).

Richards applied Voronoi tessellation to calculate atomic volumes in the experimentally solved crystallographic structures of ribonulease C and lysozyme (Richards, 1974) and Chothia extended the calculations to a larger set of proteins

(Chothia, 1975). Standard Voronoi tessellation algorithm treats all atoms as points and allocates volume to each atom regardless of the atom size, which leads to the volume overestimate for the small atoms and underestimate for the large ones. Richards introduced changes in the algorithm (Richards, 1974) that made Voronoi volumes proportional to the atom sizes, creating chemically more relevant partitioning, however it has been done at the expense of the robustness of the algorithm. The Voronoi polyhedra in this case do not fill all available space. In addition to polyhedra around the atoms Richards' method produces so called vertex polyhedra in the unallocated volumes in the neighborhood of each vertex. As a result the accuracy of the tessellation is reduced. An alternative procedure, the "radical plane" partitioning, which is both chemically appropriate and completely rigorous was designed by Gellatly and Finney (1982) and applied to the calculation of protein volumes. The radical plane of two spheres is the locus of points from which the tangent lengths to both spheres are equal. Using the three-dimensional structure of RNAase-C as an example, they have shown that the differences between the results from the radical plane and Richards' methods are generally smaller than the difference between either of those and Voronoi's method. Both radical plane and Richards' methods are relatively insensitive to the treatment of surface, which makes them preferential to other techniques (Gellatly and Finney, 1982). Volume calculation remains one of the most popular applications of Voronoi tessellation to protein structure analysis. It has been used to evaluate the differences in amino acid residue volumes in solution and in the interior of folded proteins (Harpaz et al., 1994), to monitor the atomic volumes in the course of molecular dynamics simulation of a protein (Gerstein et al., 1995), to compare the sizes of atomic groups in proteins and organic compounds (Tsai et al., 1999), to calculate the atomic volumes on protein-protein interfaces (Lo Conte et al., 1999), and to measure sensitivity of residue volumes to the selection of tessellation parameters and protein training set (Tsai and Gerstein, 2002). Deviations from standard atomic volumes in proteins determined through Voronoi tessellation correlate with crystallographic resolution and can be used as a quality measure for crystal structures (Pontius et al., 1996). A modified Voronoi algorithm, where dividing planes between the atoms were replaced by curved surfaces, defined as the set of geometrical loci with equal orthogonal distance to the surfaces of the respective van der Waals spheres, was proposed by Goede et al. (1997). Voronoi cells with hyperbolic surface constructed by this algorithm improve the accuracy of volume and density calculations in proteins (Rother et al., 2003). Another extension of the Voronoi algorithm, the Laguerre polyhedral decomposition was applied to the analysis of the residue packing geometry (Sadoc et al., 2003).

One of the problems in constructing Voronoi diagram for the molecular systems is related to the difficulty of defining a closed polyhedron around the surface atoms, which leads to ambiguities in determining their volumes and densities (a recent example in Quillin and Matthews, 2000). This problem can be addressed by the "solvation" of the tessellated molecule in the at least one layer of solvent or by using computed solvent-accessible surface for the external closures of Voronoi polyhedra. The analysis of atomic volumes on the protein surface can be used to adjust parameters of the force field for implicit solvent models, where the solvent is represented by the generalized Born model of electrostatic salvation which require knowledge of the volume of individual solute atoms (Schaefer et al., 2001). Interactions between residues in proteins can be measured using the contact area between atoms defined as the area of intersection of the van der Waals sphere and the Voronoi polyhedron of an atom (Wernisch et al., 1999). Examining the packing of residues in proteins by Voronoi tessellation revealed a strong fivefold local symmetry similar to random packing of hard spheres, suggesting a condensed matter character of folded proteins (Soyer et al., 2000). Correlations between the geometrical parameters of Voronoi cells around residues and residue conformations were discovered by Angelov et al. (2002). Another recent study described application of Voronoi procedure to study atom-atom contacts in proteins (McConkey et al., 2002).

A topological dual to Voronoi partitioning, the Delaunay tessellation (Delaunay, 1934) has an additional utility as a method for non-arbitrary identification of neighboring points in the molecular systems represented by the extended sets of points in space. Originally the Delaunay tessellation has been applied to study model (Voloshin et al., 1989) and simple (Medvedev and Naberukhin, 1987) liquids, as well as water (Vaisman et al., 1993) and aqueous solutions (Vaisman and Berkowitz, 1992; Vaisman et al., 1994). The Delaunay tessellation proved to be a particularly convenient and efficient descriptor of water structure, where a natural tetrahedral arrangement of molecules is present in the first hydration shell (Vaisman et al., 1993).

The first application of the Delaunay tessellation for identification of nearest neighbor residues in proteins and derivation of a four-body statistical potential was developed in the mid-1990 s (Singh et al., 1996). This potential has been successfully tested for inverse protein folding (Tropsha et al., 1996), fold recognition (Zheng et al., 1997), decoy structure discrimination (Munson et al., 1997; Krishnamoorthy and Tropsha, 2003), protein design (Weberndorfer et al., 1999), protein folding on a lattice (Gan et al., 2001), mutant stability studies (Carter et al., 2001), computational mutagenesis (Masso and Vaisman, 2003), protein structure similarity comparison (Bostick and Vaisman, 2003), and protein structure classification (Bostick et al., 2004). Statistical compositional analysis of Delaunay simplices revealed highly nonrandom clustering of amino acid residues in folded proteins when all residues were treated separately as well as when they were grouped according to their chemical, structural, or genetic properties (Vaisman et al., 1998). A Delaunay tessellation based alpha-shape procedure for protein structure analysis was developed by Liang et al. (Liang et al., 1998a). Alpha shapes are polytopes that represent generalizations of the convex hull. A real parameter alpha defines the "resolution" of the shape of a point set (Edelsbrunner et al., 1983). Alpha shapes proved to be useful for the detection of cavities and pockets in protein structures (Liang et al., 1998b; 1998c). Several alternative Delaunay and Voronoi based techniques for cavity identification were described by Richards (1985), Bakowies and van Gunsteren (2002) and Chakravarty et al. (2002). Delaunay tessellation has been also applied to compare similarity of protein local substructures (Kobayashi and Go, 1997) and to study the mechanical response of a protein under applied pressure (Kobayashi et al., 1997).

Tetrahedrality of Delaunay Simplices as a Structural Descriptor in Water

Quantitative measurement of local structural order in the computational models of liquid water (and other associated liquids) is an intrinsically difficult problem. Conventional structure descriptors, such as radial distribution functions cannot be used to adequately evaluate structure in the specific regions of complex model systems like multicomponent solutions or solutions of large biological molecules (Vaisman and Berkowitz, 1992). Another set of structural descriptors, the geometric and thermodynamic parameters of hydrogen bond network depend on arbitrary values in the hydrogen bond definition (Vaisman et al., 1994). Statistical geometry enables a robust and accurate approach to addressing the problem of structure characterization in water. The snapshot conformations from the molecular simulation of water or aqueous solution by molecular dynamics, Monte Carlo, or other method can be easily tessellated and geometrical parameters of the resulting Delaunay simplices can be measured. Tetrahedrality is a quantitative measure of the degree of distortion of the Delaunay simplices from the regular tetrahedra, that was introduced by Medvedev and Naberukhin (1987) for simple liquids, but can be easily extended to water and other systems. Tetrahedrality is calculated as:

$$T = \sum_{i>j} (l_i - l_j)^2 / 15\bar{l}^2 , \qquad (3.1)$$

where l_i is the length of the i-th edge, and l is the mean length of the edges of the given simplex. For a regular tetrahedron with four equilateral triangular faces, $T = 0$. For any irregular tetrahedron, $T > 0$. In case of a simulated molecular system the tessellation produces a large number of Delaunay simplices for each snapshot conformation, and a number of such conformations can be very large as well. If the simulated system is at equilibrium, the ergodic theorem applies, and time averages along a system trajectory can be combined with ensemble averages over the phase space. Such a combination increases the number of simplices for the analysis by several orders of magnitude (10^3–10^4 simplices in a conformation multiplied by 10^3–10^4 conformations), which affords high statistical reliability of the results. The nature of this descriptor allows to calculate it separately in confined or limited regions of the simulation system, e.g., in concentric spherical layers around a solute.

The distribution of water tetrahedrality in different layers around solutes depend on the nature of the solute. In the case of charged ions, like ammonium, the difference between tetrahedrality of bulk water and the ammonium hydration water is particularly strong due to the strong hydrogen bonding between water and solute. The peak of the distribution of the tetrahedrality of the ammonium hydration water is shifted to the right which indicates that the hydration water is less tetrahedral than bulk water (Fig. 3.2). Conversely, water in the first hydration shell of methane is just slightly more tetrahedral than the bulk water. Thus,

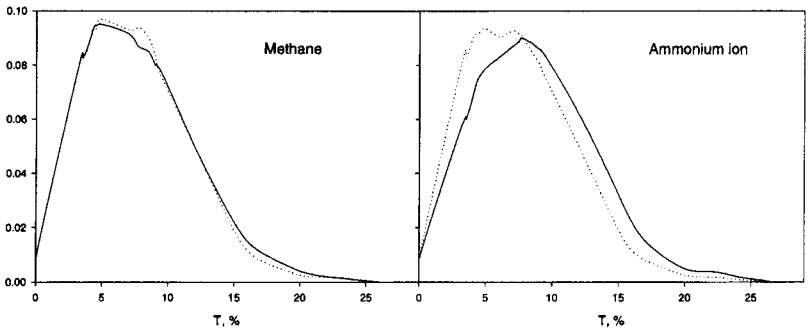

Figure 3.2. Distribution of tetrahedrality of water around solutes; *solid line* – first hydration shell, *dotted line* – bulk water

the hydration water around ammonium is significantly more distorted than that around methane as one could expect in the case of hydrophilic and hydrophobic hydration, respectively (Vaisman et al., 1994).

It is worth to note that the presence of hydrophobic solute changes the distribution of water tetrahedrality in the same way as the decrease of temperature. This observation is consistent with the concept of the decrease of 'structural temperature' of water, surrounding hydrophobic particles, that has been discussed in the literature for a long time. The decrease in the structural temperature corresponds to the increased structural order of water, because any structural characteristic of liquid must be a monotonically decreasing function of temperature. Distribution of tetrahedrality entirely complies with this requirement.

The influence of both solutes on the water tetrahedrality is almost unobservable beyond the first hydration shell. The distribution of tetrahedrality for both solutions is similar at both cutoff radii (Vaisman et al., 1994). This result indicates that the distribution of tetrahedrality is not sensitive to the treatment of long-range interactions. Distribution of tetrahedrality beyond the first hydration shell is similar to that in pure water.

Spatial and Compositional Three-dimensional Patterns in Proteins 3.4

Delaunay simplices obtained as a result of the tessellation can be used to define objectively the 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 sim-

plices provides information about spatial propensities of all quadruplets of amino acid residues clustered together in folded protein structures. The compositional statistics can be also used to construct four-body empirical contact potentials, which may provide improvement over traditional pairwise statistical potentials (e.g., Miyazawa and Jernigan, 2000) for protein structure analysis and prediction.

To perform the tessellation protein residues should be represented by single points located, for example, in the positions of the C_α atoms or the centers of the side chains. Tessellation training set includes high-quality representative protein structures with low primary-sequence identity (Wang and Dunbrack, 2003). The tessellated proteins are analyzed by computing various geometrical properties and compositional statistics of Delaunay simplices.

An example of Delaunay tessellation of a folded protein is illustrated on Fig. 3.3 for crambin (1 crn). 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 and thus four nearest neighbor amino acid residues.

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 (Singh et al., 1996). 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 (Fig. 3.4). All five classes usually occur in any given protein.

The differences between classes of simplices can be evaluated using geometrical parameters of tetrahedra such as volume and tetrahedrality (3.1). Distributions of

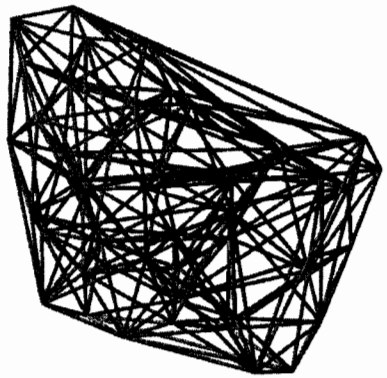

Figure 3.3. Delaunay tessellation of Crambin

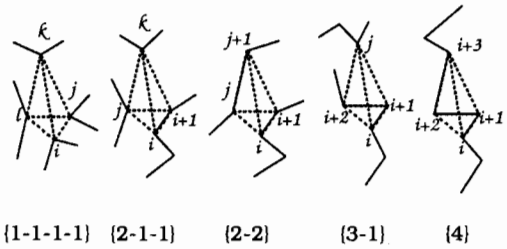

Figure 3.4. Five classes of Delaunay simplices

volume and tetrahedrality for all five classes of simplices is shown in Fig. 3.5. The sharp narrow peaks correspond to the simplices of classes {4} and {2, 2}. They tend to have well defined distributions of volume and distortion of tetrahedrality. These results suggest that tetrahedra of these two classes may occur in regular protein conformations such as α-helices and may be indicative of a protein fold family. We have calculated the relative frequency of occurrence of tetrahedra of each class in each protein in a small dataset of hundred proteins from different families and plotted the results in Fig. 3.6. The proteins were sorted in the ascending order of fraction of tetrahedra of class {4}. Noticeably, the content of simplices of class {3, 1} decreases with the increase of the content of class {4} simplices. According to common classifications of protein fold families (Orengo et al., 1997), 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 dataset are also shown in Fig. 3.6. These results suggest that proteins having a high content of tetrahedra of classes {4} and {2, 2} (i.e., proteins in the right part of the plot in Fig. 3.6) belong to the family of all-alpha proteins. 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 part of the plot in Fig. 3.6) 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.

Identification of significant patterns in biomolecular objects depends on the possibility to distinguish what is likely from what is unlikely to occur by chance (Karlin et al., 1991). Statistical analysis of amino acid composition of the Delaunay simplices provides information about spatial propensities of all quadruplets of amino acid residues to be clustered together in folded protein structures. 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 using the following equation:

$$q_{ijkl} = \log \frac{f_{ijkl}}{p_{ijkl}} \tag{3.2}$$

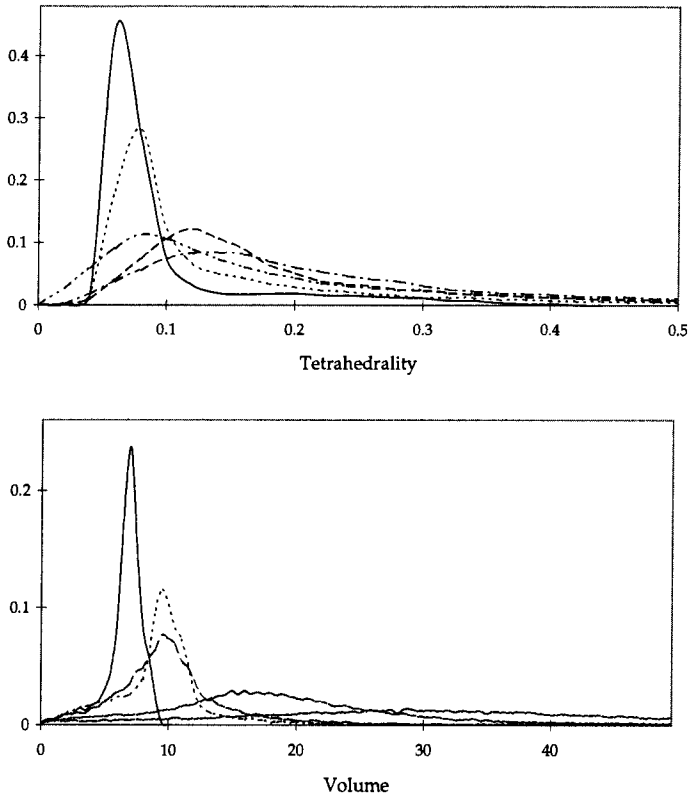

Figure 3.5. Distribution of tetrahedrality and volume (in Å^3) of Delaunay simplices in proteins

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:

$$p_{ijkl} = Ca_i a_j a_k a_l \tag{3.3}$$

where a_i, a_j, a_k, and 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

$$C = \frac{4!}{\prod_i^n (t_i!)} \tag{3.4}$$

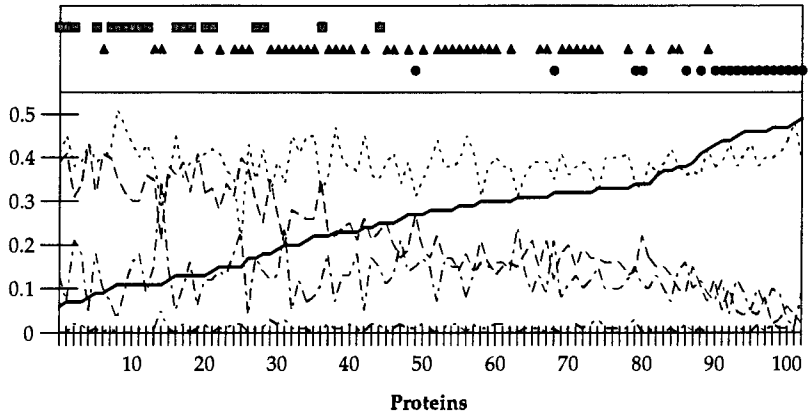

Figure 3.6. Classes of Delaunay simplices and protein fold families. Contents of simplices of class {4} (*solid line*), class {3, 1} (*dashed line*), class {2, 1} (*dotted line*), class {2, 1} (*dash-dotted line*), class {1, 1, 1, 1} (*dash-dot-dotted line*). *Upper part* of the figure displays fold family assignment: all-alpha (*circles*), all-beta (*squares*), and alpha-beta (*triangles*)

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 20 natural amino acid residues is 8855 ($C_{20}^4 + 3C_{20}^3 + 2C_{20}^2 + C_{20}^2 + C_{20}^1$). The first term accounts for simplices with four distinct residue types, the second – three types in 1 – 1 – 2 distribution, the third – two types in 1 – 3 distribution, the fourth – two types in 2 – 2 distribution, and the fifth – four identical residues. The log-likelihood factor q is plotted in Fig. 3.7 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 statistical potential energy function. The statistical potential can be used in a variety of structure prediction, protein modeling, and computational mutagenesis applications.

Computational mutagenesis is based on the analysis of a protein potential profile, which is constructed by summing the log-likelihood scores from (3.2) for all simplices in which a particular residue participates. A plot of the potential profile for a small protein, HIV-1 protease, is shown in Fig. 3.8. The shape of the potential profile frequently reflects important features of the protein, for example, the residues in local maxima values of the profile are usually located in the hydrophobic core of the protein and these residues play an important role in maintaining protein stability.

A potential profile can be easily calculated for both wild type and mutant proteins, assuming that the structural differences between them are small and that their tessellation results are similar. In this case the difference between the

profiles is defined only by the change in composition of the simplices involving the substitution site. The resulting difference profile provides important insights into the changes in protein energetics due to the mutation.

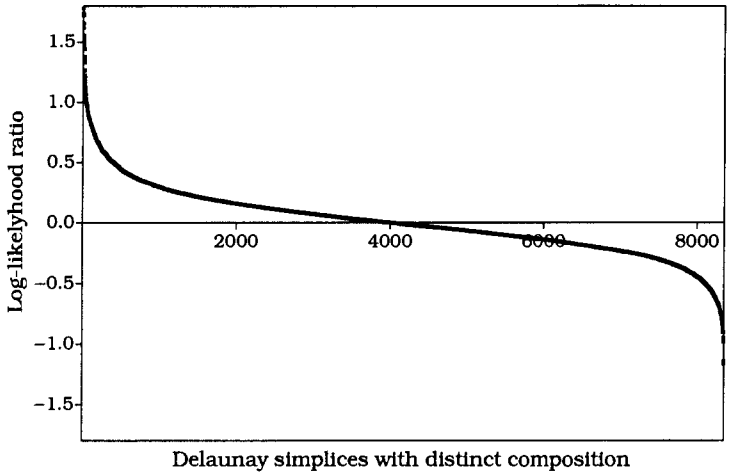

Figure 3.7. Log-likelihood ratio for the Delaunay simplices

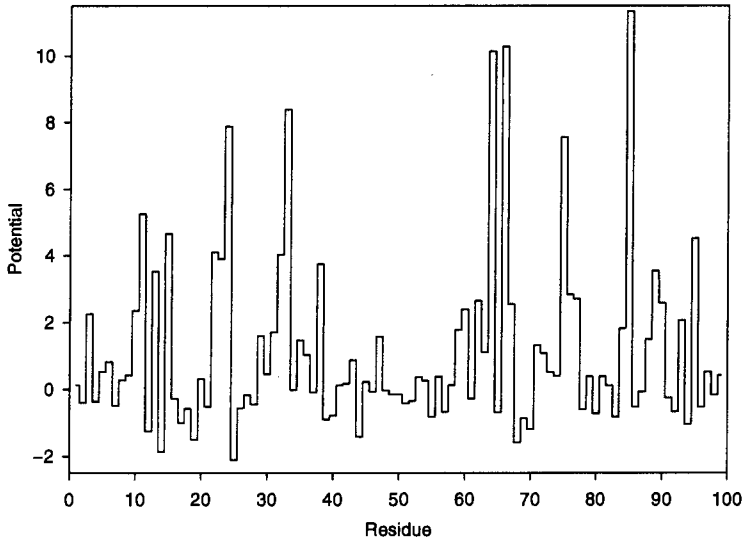

Figure 3.8. Potential profile of HIV-1 protease

Protein Structure Comparison and Classification

3.5

Using the information from the Delaunay tessellation of a protein's backbone, it is possible to build a statistical representation of that protein, which takes into account the way its sequence must "twist and turn" in order to bring each four-body residue cluster into contact. Each residue – i, j, k, and l of a four-body cluster comprising a simplex are nearest neighbors in Euclidean space as defined by the tessellation, but are separated by the three distances – d_{ij}, d_{jk}, and d_{kl} in sequence space. Based on this idea, we build a 1000-tuple representation of a single protein by making use of two metrics: (1) the Euclidean metric used to define the Delaunay tessellation of the protein's C_α atomic coordinates and (2) the distance between residues in sequence space.

If we consider a tessellated protein with N residues integrally enumerated according to their position along the primary sequence, the length of a simplex edge in sequence space can be defined as $d_{ij} = j - i - 1$, where d_{ij} is the length of the simplex edge, $\overline{ij}$, corresponding to the ith and jth α-carbons along the sequence. If one considers the graph formed by the union of the simplex edge between the two points i and j and the set of edges between all d_{ij} points along the sequence between i and j, it is seen that the Euclidean simplex edge, $\overline{ij}$, can generally be classified as a far edge (Pandit and Amritkar, 1999). Every simplex in the protein's tessellation will have three such edges associated with its vertices: i, j, k, and l where i, j, k, and l are integers corresponding to C_α atoms enumerated according to their position along the primary sequence. Thus, we proceed to quantify the degree of "farness" in an intuitive way, by applying a transformation, T, which maps the length, d, of each edge to an integer value according to

$$T : d \mapsto \begin{cases} 1 & \text{if } d = 0 \\ 2 & \text{if } d = 1 \\ 3 & \text{if } d = 2 \\ 4 & \text{if } d = 3 \\ 5 & \text{if } 4 \leq d \leq 6 \\ 6 & \text{if } 7 \leq d \leq 11 \\ 7 & \text{if } 12 \leq d \leq 20 \\ 8 & \text{if } 21 \leq d \leq 49 \\ 9 & \text{if } 50 \leq d \leq 100 \\ 10 & \text{if } d \geq 101 \end{cases} \tag{3.5}$$

The reasoning behind the design of the transformation is described by Bostick and Vaisman (2003). This transformation is used to construct an array that is representative of the distribution of combinations of segment lengths along the protein backbone forming four-residue clusters within the protein's structure as defined by the tessellation of its C_α atomic coordinates. Each simplex in the protein's

tessellation contributes to a 3D array, M, where M_{npr} is the number of simplices whose edges satisfy the following conditions:

1. The Euclidean length of any one simplex edge is not greater than 10 Å.
2. $d_{ij} = n$
3. $d_{jk} = p$
4. $d_{kl} = r$

Condition 1 is provided because simplices with a Euclidean edge length above 10 Å are generally a result of the positions of α-carbons on the exterior of the protein. We filter out contributions from these simplices to M, because they do not represent physical interactions between the participating residues. The simplices with the long edges are formed due to the absence of solvent and other molecules around the protein in the tessellation, they would not have existed if the protein was solvated. The data structure, M, contains 1000 elements. The number of elements is invariant with respect to the number of residues of the protein. In order to more easily conceptualize the mapping of the protein topology to the data structure, M, we rewrite it as a 1000-tuple vector $\vec{M} = \{M_{000}, M_{001}, \ldots, M_{010}, M_{011}, \ldots, M_{999}\}$.

Given that each element of this vector represents a statistical contribution to the global topology, a comparison of two proteins making use of this mapping must involve the evaluation of the differences in single corresponding elements of the proteins' 1000-tuples. We define, therefore, a raw topological score, Q, representative of the topological distance between any two proteins represented by data structures, $\vec{M}$ and $\vec{M}'$, as the supremum norm,

$$Q \equiv \left\| \vec{M} - \vec{M}' \right\|_{\sup} \equiv \sum_{i=0}^{999} \left| M_i - M_i' \right| . \tag{3.6}$$

This norm is topologically equivalent to the Euclidean norm and has the added advantage that it is less computationally expensive to calculate.

This topological score has an obvious dependence on the sequence length difference between the two proteins being compared due to the following implicit relation for a single protein representation,

$$N_s = \sum_{i=0}^{999} M_i , \tag{3.7}$$

where i is the number of simplices with no edge having a Euclidian length greater than 10 Å, and the M_i are the elements of the protein representation. In other words, since N_s is proportional to the number of residues in the protein, the difference in the length between two compared proteins might provide a systematic extraneous contribution to their score, Q, in (3.6). This is not to say that the sequence length of a protein does not play a role in its topology. In fact, the length should be quite crucial (Bostick et al, 2004). However, the length dependence of our score implied by (3.7) is endemic to our protein representation (derived from its tessellation),

and not due to protein topology itself. This length dependence may be removed by first normalizing the vector representation as follows:

$$\underset{\rightarrow}{M} = \frac{\vec{M}}{\left\| \vec{M} \right\|} \tag{3.8}$$

resulting in the unit-vector representation, $\underset{\rightarrow}{M}$. The corresponding normalized topological score,

$$\underline{Q} \equiv \left\| \underset{\rightarrow}{M} - \underset{\rightarrow}{M}' \right\|_{\sup} \tag{3.9}$$

can be expected to be less sensitive to the chain length difference between the two proteins being compared. Despite normalization, however, this score should still have an inherent dependence on the length difference between the compared proteins. A protein's structure must be dependent on the length of its sequence, because the number of configurational degrees of freedom in a polymer's structure is proportional to the number of residues it possesses. Such a dependence on the size of compared proteins is present in geometric methods of comparison such as structural alignment as well, and in some cases, has been accounted for (Carugo and Pongor, 2001).

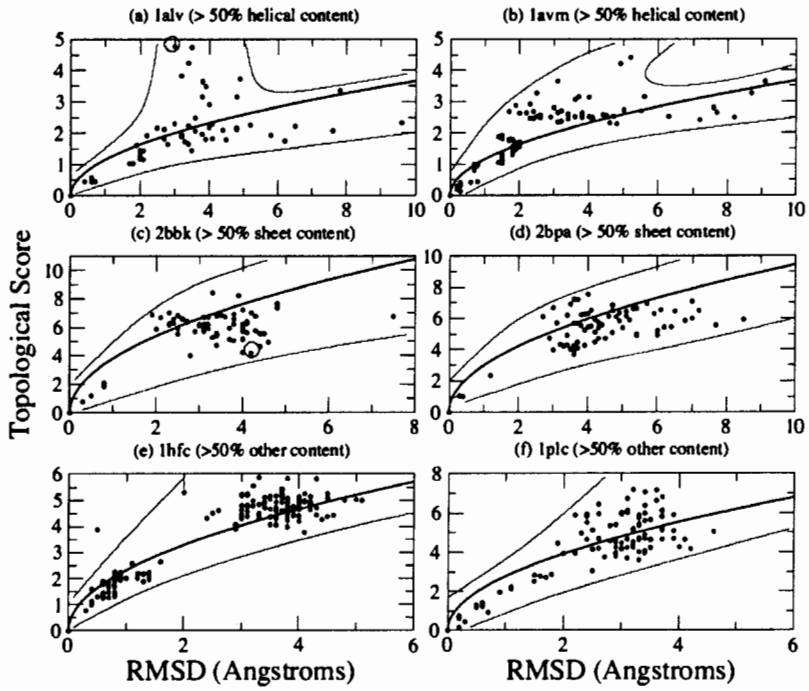

Figure 3.9. Topological and geometric comparison within 6 protein families

The results of topological protein structure comparison can be illustrated using an example of proteins that belong to the same family. Six protein families were selected from the FSSP (Families of Structurally Similar Proteins) database for topological evaluation. We selected families that span various levels of secondary structural content. The representatives of these families are as follows: 1alv and 1avm (having greater than 50% α-helical content), 2bbk and 2bpa (having greater than 50% β-sheet content), and 1hfc and 1plc (having at least 50% content that is classified as neither α-helical nor β-sheet). The FSSP database contains the results of the alignments of the extended family of each of these representative chains. Each family in the database consists of all structural neighbors excluding very close homologs (proteins having a sequence identity greater than 70%). The topological score was calculated for each representative in a one-against-all comparison with its neighbors. All of the scores are plotted against RMSD for each of the families in Fig. 3.9. A strong correlation between the topological score and structure similarity and the power-law trend can be seen for all families.

Conclusions

Methods of statistical and computational geometry, Voronoi and Delaunay tessellation in particular, play an increasingly important role in exploration of complex nature of molecular and biomolecular structure. The range of applications of statistical geometry for biomolecular structural studies has grown significantly in the past decade. As the new experimental structural information on biomolecules becomes available, the need for sophisticated and robust tools for its interpretation will further increase. At the same time more known structure will enable the creation of larger and better training sets for pattern identification. Existing and new statistical geometry algorithms may prove instrumental in future developments of methods for protein structure analysis, ab initio structure prediction, and protein engineering.

References

Angelov, B., Sadoc, J.F., Jullien, R., Soyer, A., Mornon, J.P. and Chomilier, J. (2002). Nonatomic solvent-driven Voronoi tessellation of proteins: an open tool to analyze protein folds. *Proteins*, 49(4)446–456.

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

Bakowies, D. and van Gunsteren, W.F. (2002). Water in protein cavities: A procedure to identify internal water and exchange pathways and application to fatty acid-binding protein. *Proteins*, 47(4)534–545.

Barber, C.B., Dobkin, D.P. and Huhdanpaa, H.T. (1996). The Quickhull algorithm for convex hulls. *ACM Trans on Mathematical Software*, 22(4)469–483.

Bernal, J.D. (1959). A geometrical approach to the structure of liquids. *Nature*, 183(4655)141–147.

Bostick D. and Vaisman, I.I. (2003). A new topological method to measure protein structure similarity. *Biochem Biophys Res Commun*, 304(2)320–325.

Bostick, D., Shen, M. and Vaisman, I.I. (2004). A simple topological representation of protein structure: Implications for new, fast, and robust structural classification. *Proteins*, 55.

Carter, C.W. Jr, LeFebvre, B.C., Cammer, S.A., Tropsha, A. and Edgell, M.H. (2001). Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. *J Mol Biol*, 311(4)625–638.

Carugo, O. and Pongor, S. (2001). A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Sci*, 10(7)1470–1473.

Chakravarty, S., Bhinge, A. and Varadarajan, R. (2002). A procedure for detection and quantitation of cavity volumes proteins. Application to measure the strength of the hydrophobic driving force in protein folding. *J Biol Chem*, 277(35)31345–31353.

Chothia, C. (1975). Structural invariants in protein folding. *Nature 1975*, 254(5498)304–308.

Delaunay, B. N. (1934). Sur la sphere vide, Izv Akad Nauk SSSR, Otd Mat Est Nauk, 7:793–800.

Edelsbrunner, H., Kirkpatrick, D.G. and Seidel, R. (1983). On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, IT-29(4)551–559.

Finney, J.L. (1970). Random packing and the structure of simple liquids. I. The geometry of random close packing. *Proc Roy Soc Lond*, A319479–493 and 495–507.

Finney, J.L. (1975). Volume occupation, environment and accessibility in proteins. The problem of the protein surface. *J Mol Biol*, 96(4)721–732.

Gan, H.H., Tropsha, A and Schlick, T. (2001) Lattice protein folding with two and four-body statistical potentials. *Proteins*, 43(2)161–174.

Gellatly, B.J. and Finney, J.L. (1982). Calculation of protein volumes: an alternative to the Voronoi procedure. *J Mol Biol*, 161(2)305–322.

Gerstein, M., Tsai, J. and Levitt, M. (1995). The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra. *J Mol Biol*, 249(5)955–966.

Goede, A., Preissner, R. and Frömmel, C. (1997). Voronoi cell: New method for allocation of space among atoms: Elimination of avoidable errors in calculation of atomic volume and density. *J Comput Chem*, 18(9)1113–1123.

Harpaz, Y., Gerstein, M. and Chothia, C. (1994). Volume changes on protein folding. *Structure*, 2(7)641–649.

Karlin, S., Bucher, P. and Brendel, V. (1991). Altschul S.F., Statistical methods and insights for protein and DNA sequences. *Annu Rev Biophys Biophys Chem*, 20:175–203.

Kobayashi, N. and Go N. (1997). A method to search for similar protein local structures at ligand binding sites and its application to adenine recognition. *Eur Biophys J*, 26(2)135–144.

Kobayashi, N., Yamato, T. and Go N. (1997). Mechanical property of a TIM-barrel protein. *Proteins*, 28(1)109–116.

Krishnamoorthy, B. and Tropsha, A. (2003). Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. *Bioinformatics*, 19(12)1540–1548.

Liang, J., Edelsbrunner, H., Fu, P., Sudhakar, P.V. and Subramaniam, S. (1998a). Analytical shape computation of macromolecules: I. Molecular area and volume through alpha shape. *Proteins*, 33(1)1–17.

Liang, J., Edelsbrunner, H., Fu, P., Sudhakar, P.V. and Subramaniam, S. (1998b). Analytical shape computation of macromolecules: II. Inaccessible cavities in proteins. *Proteins*, 33(1)18–29.

Liang, J., Edelsbrunner, H., Woodward, C. (1998c). Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci*, 7(9)1884–1897.

Lo Conte, L., Chothia, C. and Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J Mol Biol*, 285(5)2177–2198.

Masso, M. and Vaisman, I.I. (2003). Comprehensive mutagenesis of HIV-1 protease: a computational geometry approach. *Biochem Biophys Res Commun*, 305(2)322–326.

Medvedev, N.N. and Naberukhin, Yu.I. (1987). Analysis of structure of simple liquids and amophous solids by statistical geometry method. *Zh Strukt Khimii*, 28(3)117–132.

Miyazawa, S. and Jernigan, R.L. (2000). Identifying sequence-structure pairs undetected by sequence alignments. *Protein Eng*,13(7)459–475.

McConkey, B.J., Sobolev, V. and Edelman, M. (2002). Quantification of protein surfaces, volumes and atom-atom contacts using a constrained Voronoi procedure. *Bioinformatics*, 18(10)1365–1373.

Munson, P.J. and Singh, R.K. (1997). Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment. *Protein Sci*, 6(7)1467–1481.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M. (1997). CATH – a hierarchic classification of protein domain structures. *Structure*, 5(8)1093–1108.

Pandit, S. A. and Amritkar, R. E. (1999). Characterization and control of small-world networks. *Phys Rev E*, 60:1119–1122.

Pontius, J., Richelle, J. and Wodak, S.J. (1996). Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J. Mol. Biol.*, 264(1)121–136.

Quillin, M.L. and Matthews, B.W. (2000). Accurate calculation of the density of proteins. *Acta Crystallogr D Biol Crystallogr*, 56(Pt 7)791–794.

Richards, F.M. (1974). The interpretation of protein structures: total volume, group volume distributions and packing density. *J Mol Biol*, 82(1)1–14.

Richards, F.M. (1985). Calculation of molecular volumes and areas for structures of known geometry. *Methods Enzymol*, 115:440–464.

Rother, K, Preissner, R, Goede, A and Frömmel, C. (2003). Inhomogeneous molecular density: reference packing densities and distribution of cavities within proteins. *Bioinformatics*, 19(16)2112–2121.

Sadoc, J.F., Jullien, R. and Rivier, N. (2003). The Laguerre polyhedral decomposition: application to protein folds. *Eur Phys J B*, 33(3)355–363.

Schaefer, M., Bartels, C., Leclerc, F. and Karplus M. (2001). Effective atom volumes for implicit solvent models: comparison between Voronoi volumes and minimum fluctuation volumes. *J Comput Chem*, 22(15)1857–1879.

Singh, R.K., Tropsha, A. and Vaisman, I.I. (1996). Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. *J Comput Biol*, 3(2)213–222.

Soyer, A., Chomilier, J., Mornon, J.P., Jullien, R. and Sadoc, J.F. (2000). Voronoi tessellation reveals the condensed matter character of folded proteins. *Phys Rev Lett*, 85(16)3532–3535.

Sugihara, K. and Inagaki, H. (1995). Why is the 3D Delaunay triangulation difficult to construct? *Information Processing Letters*, 54:275–280.

Tropsha, A., Singh, R.K., Vaisman, I.I. and Zheng, W. (1996). Statistical geometry analysis of proteins: implications for inverted structure prediction. *Pac Symp Biocomput*, 614–623.

Tropsha, A., Carter, C.W. Jr, Cammer, S. and Vaisman, I.I. (2003). Simplicial neighborhood analysis of protein packing (SNAPP): a computational geometry approach to studying proteins. *Methods Enzymol*, 374:509–544.

Tsai, J., Taylor, R., Chothia, C. and Gerstein, M. (1999). The packing density in proteins: standard radii and volumes. *J Mol Biol*, 290(1)253–266.

Tsai, J. and Gerstein, M. (2002). Calculations of protein volumes: sensitivity analysis and parameter database. *Bioinformatics*, 18(7)985–995.

Naberukhin, Y.I., Voloshin, V.P. and Medvedev, N.N. (1991). Geometrical analysis of the structure of simple liquids: percolation approach. *Mol Physics*, 73:917–936.

Vaisman, I.I., Perera, L. and Berkowitz, M.L. (1993). Mobility of stretched water. *J Chem Phys*, 98(12)9859–9862.

Vaisman, I.I. and Berkowitz, M.L. (1992). Local structural order and molecular associations in water-DMSO mixtures. Molecular dynamics study. *J Am Chem Soc*, 114(20)7889–7896.

Vaisman, I.I., Brown, F.K. and Tropsha, A. (1994). Distance Dependence of Water Structure Around Model Solutes. *J Phys Chem*, 98(21)5559–5564.

Vaisman, I.I., Tropsha, A. and Zheng W. (1998). Compositional preferences in quadruplets of nearest neighbor residues in protein structures: Statistical geometry analysis. *Proc. of the IEEE Symposia on Intelligence and Systems*, 163–168.

Voloshin, V.P., Naberukhin, Y.I. and Medvedev, N.N. (1989). Can various classes of atomic configurations (Delaunay simplices) be distinguished in random close packing of spherical particles?, *Molec Simulation*, 4:209–227.

Voronoi, G.F. (1908). Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Deuxième Mémorie: Recherches sur les parallélo-èdres primitifs. *J Reine Angew Math*, 134:198–287.

Wang, G. and Dunbrack, R.L. Jr. (2003). PISCES: a protein sequence culling server. *Bioinformatics*, 19(12)1589–1591.

Weberndorfer, G., Hofacker, I.L. and Stadler, P.F. (1999). An efficient potential for protein sequence design. *Proc German Conf Bioinformatics*, 107–112.

Wernisch, L., Hunting, M. and Wodak, S.J. (1999). Identification of structural domains in proteins by a graph heuristic. *Proteins*, 35(3)338–352.

Zheng, W., Cho, S.J., Vaisman, I.I. and Tropsha, A. (1997). A new approach to protein fold recognition based on Delaunay tessellation of protein structure. *Pac Symp Biocomput*, 486–497.