Analyzing Protein Structure Using Almost-Delaunay Tetrahedra

Deepak Bandyopadhyay[1], Alexander Tropsha[2] and Jack Snoeyink[1]

[1]Department of Computer Science, UNC Chapel Hill
and [2]Laboratory for Molecular Modeling, School of Pharmacy, UNC Chapel Hill

ABSTRACT

Delaunay tessellations and Voronoi diagrams capture proximity relationships among sets of points. When applied to points representing protein atoms or residue positions, they are used to compute molecular surfaces and protein volumes, to define cavities and pockets, to analyze and score packing interactions, and to find structural motifs. Since atom and residue coordinates are known imprecisely, we explore the effect of coordinate perturbation on Delaunay-based scoring and motif finding. We define and compute the *almost-Delaunay tetrahedra*, which are tetrahedra that can become part of a Delaunay tessellation if the point coordinates are perturbed by at most $\epsilon \geq 0$, and the probability that each is Delaunay assuming random Gaussian perturbations of all points. By analyzing these tetrahedra, we show that Delaunay four-body potential functions are robust and derive a new method to detect structural motifs. An implementation in MATLAB is available from http://www.cs.unc.edu/~debug/papers/AlmDel.

INTRODUCTION

The Voronoi diagram and Delaunay tessellation, which are geometric structures defined for sets of points, have found use in many areas of science and engineering (Aurenhammer, 1991; de Berg et al., 2000; Boissonnat and Yvinec, 1998; Okabe et al., 1992). Below, we survey a number of applications in computational molecular biology, including scoring packing interactions and finding structural motifs. It is natural to ask whether these analyses are stable and robust under changes to the input coordinates. In this work we complement the empirical answers given for specific applications with a mathematical approach that considers the possible structures that could be defined by nearby inputs.

For a finite set $P \in R^3$ of point *sites*, the *Voronoi diagram* is the decomposition of space into regions with the same set of closest neighbor sites (Voronoi, 1908). The *Delaunay tessellation* is a decomposition of the same space based on an "empty sphere property:" (Delaunay, 1934) if a subset of sites, $S \subset P$, lie on the boundary of a sphere that is otherwise empty of sites, then the convex hull of S is a region of the Delaunay tessellation.

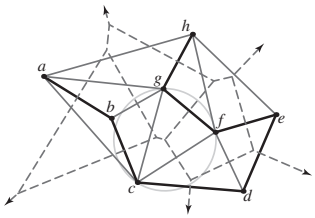

Fig. 1. Two-dimensional Voronoi diagram (dashed) and Delaunay tessellation (solid) for sites *a–h*. Shows circumcircle for $\triangle cfg$.

Figure 1 illustrates the two-dimensional Voronoi and Delaunay for sites *a–h*. Codes for computing both are available in standalone programs (Barber et al., 1996; Watson, 1981) and packages such as MATLAB (www.mathworks.com).

Richards (1974) pioneered the use of Voronoi diagrams to compute protein volumes. This has been an active research area, with more detailed empirical analysis of parameters (Gerstein et al., 1995; Tsai et al., 1999; Tsai and Gerstein, 2002), with refinements on the definition of the surface, often by interaction with randomly placed solvent molecules (Liao et al., 2001; McConkey et al., 2002; Soyer et al., 2000), and with analysis of differential packing in the core and surface regions (Gerstein et al., 1995; Liang and Dill, 2001).

The Delaunay tessellation gives structure that can help define and detect pockets and cavities in proteins (Bakowies and van Gunsteren, 2002; Liang et al., 1998), and even to analyze mechanical properties of proteins (Kobayashi et al., 1997).

Both the Voronoi and Delaunay have been used to score residue interactions in folded proteins and decoys. The contact area between Voronoi regions of residues has been incorporated into 2-body (Zimmer et al., 1998) and geometric potentials (Angelov et al., 2002).

The Delaunay tessellation collects sets of four "neighboring" representative points into tetrahedra, as defined in the next section. Researchers have analyzed the frequency of occurrence of different amino acid types in tetrahedra to develop empirical four-body potentials to score folded

proteins and try to distinguish the native state from decoys (Carter et al., 2001; Krishnamoorthy and Tropsha, 2003; Munson and Singh, 1997; Singh et al., 1996; Weberndorfer et al., 1999). The four-body potentials complement fragment-based methods (Simons et al., 1997) and pairwise potentials (Miyazawa and Jernigan, 1996) to capture favorable or unfavorable packing interactions. Simplicial Neighborhood Analysis of Protein Packing (SNAPP) (http://mmlsun4.pha.unc.edu/psw/3dworkbench.html) was developed to exploit Delaunay tessellation as a computational structural biology tool.

Voronoi and Delaunay structures have also been used in the search for local motifs. Wernisch et al. (1999) use the Voronoi contacts to partition proteins into structural domains with minimal interaction between them. Wako and Yamato (1998) compute the Delaunay tessellation of C_α carbons and find patterns of the backbone sequence among Delaunay neighbors to identify local motifs for helices and sheets.

Stability and robustness

We would like to know if these Delaunay and Voronoi analyses are stable and robust under changes to the input coordinates. The Delaunay and Voronoi themselves are not robust: small changes to the coordinates of nearly cospherical input points can cause large changes to the set of regions in both diagrams.

In this paper, we consider which additional sets of sites could become edges, triangles, or tetrahedra of the Delaunay tessellation if all sites are perturbed by a minimum amount $\epsilon \geq 0$. We can also calculate the probability that a set is in a Delaunay tessellation if the perturbations are random Gaussian distributions.

The Definitions and Methods section gives formal definitions and their properties. The Algorithms section sketches how to compute the almost-Delaunay threshold and Delaunay probability for a tetrahedron. Our general framework of almost-Delaunay simplices in arbitrary dimensions, their geometrical properties and algorithms to compute them are covered in a companion paper in Computational Geometry (Bandyopadhyay and Snoeyink, 2004). In the Discussion, we focus on three questions: how do additional tetrahedra introduced by perturbation affect point sets (from random to proteins), how do they affect statistical potentials, and how can they help recognize structural motifs?

There has been empirical work on estimating the robustness of Voronoi and Delaunay analysis. The best studied is the problem of estimating volumes for surface molecules. Empirical stability analysis has been performed by computing the volumes of many copies of a protein during a molecular dynamics computation (Gerstein et al., 1995), or by checking thirty thousand crystallographic structures for small organic compounds (Tsai et al., 1999). Gerstein *et al.* point out that the analysis of volumes at the surface is sensitive to the radii chosen for water molecules, and to the method of defining bisecting planes for atoms of different sizes.

Empirical analysis is more difficult for fold scoring, because we are unlikely to have a significant number of independent structures to test significance. Carter et al. (2001) correlated empirical folding free energy change $\Delta\Delta G$ with SNAPP scores of mutant vs. wild type proteins. Cammer and Tropsha, in unpublished work, observed that four-body statistical potentials derived with Delaunay tessellation of C_α vs. side chain centroid representations are similar but not identical.

Empirical analysis has another difficulty: even when there is enough evidence to support an observed phenomenon, it can be difficult to assess whether the root cause is biological, or purely geometric. For example, a study of Voronoi faces (regions with two closest sites) observed that the faces on the surface had an average of 5.03 edges (Angelov et al., 2002). In fact, this average is determined by the number of faces f and the genus (number of holes) g according to Euler's relation: $e_{avg} = 6 + (12g - 12)/f$. Thus, only for proteins with about a dozen surface faces and no holes can the average be close to 5.

DEFINITIONS AND METHODS

Consider, by way of example, a finite set of points, or *sites*, representing the C_α positions of the residues of a protein. We assume, to make description easier, that the sites are in general position—no four lie on a common plane and no five lie on a common sphere. We begin with some definitions that are standard in geometry (de Berg et al., 2000; Boissonnat and Yvinec, 1998).

A *k-simplex* is the convex hull of $k + 1$ affinely independent points, called the *vertices* of the simplex. In 3D, we have 3-simplices (tetrahedra), 2-simplices (triangles), 1-simplices (edges), and 0-simplices (vertices). The *Delaunay tessellation* consists of simplices with the empty sphere property: a simplex is Delaunay if and only if some sphere circumscribing the vertices is empty of other sites. The vertices of a Delaunay simplex are "neighbors" in the sense that some point (the circumcenter) is closer to them than the rest of the sites—thus, the vertices of a Delaunay simplex define a region in the Voronoi diagram.

Any simplex that is not Delaunay could become Delaunay if the sites move to satisfy the empty sphere property. Suppose that by moving sites $P = \{p_1, p_2, \ldots, p_n\}$ to $P' \subset \{p'_1, \ldots, p'_n\}$, we can make the simplex $\{p'_1, \ldots, p'_4\}$ be empty. We measure this motion by the maximum distance any p'_i is from p_i, and say that simplex $\{p'_1, \ldots, p'_4\}$ is in $AD(\epsilon)$. The minimum distance ϵ is denoted the *AD threshold* of the simplex. For example,

Delaunay simplices have threshold 0. Figure 2(a) shows a small movement in our two-dimensional example that makes bf a Delaunay edge. Note that we allow all sites to move—not just the vertices of the simplex.

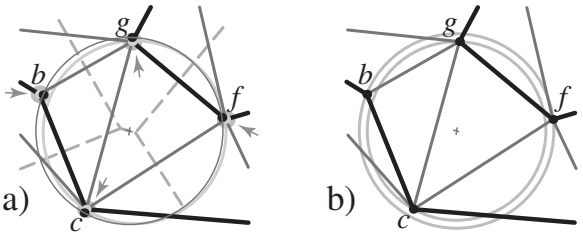

Fig. 2. a) Movement can make bf into a Delaunay edge, and triangle bcf and bfg into Delaunay triangles. b) The smallest movement can be found from a minimum-width annulus.

ALGORITHM

The computation of almost-Delaunay thresholds can be expressed as a computational metrology problem of measuring roundness of a manufactured object (Garcia-Lopez et al., 1998).Two parameters help us speed up our algorithm for proteins:

Edge Length Prune: In proteins, only residues within about 10 Å are close enough to be considered neighbors for evaluating contacts, packing, and volume occupancy. Thus, we prune all edges longer than an *edge-length prune* parameter, and all simplices that contain a long edge. We follow the work on SNAPP (Carter et al., 2001) and use a 10 Å prune in this paper; we have experimented with values from 8 Å, where the first few AD tetrahedra typically appear, to 15 Å.

Threshold Cutoff: To study the effects of perturbation on the Delaunay tessellation of protein coordinates from the PDB (www.rcsb.org), certain small ranges of threshold ϵ are of interest. Rounding in the coordinates to two decimal places is captured by perturbations of at most 0.01 Å. Possible errors in the last digit are captured by perturbations of 0.1 Å. Uncertainties of atom position due to thermal motion, X-ray refinement resolution, choice of representative point for a residue, or configurational change could make ranges up to 0.5 Å, 1 Å, or 2 Å worthy of study. We use 2 Å as a *threshold cutoff* in our experiments unless otherwise stated, covering all the above ranges of perturbation.

We run our experiments with a MATLAB implementation of this algorithm, available on http://www.cs.unc.edu/~debug/papers/AlmDel. It takes a few seconds to a minute on a 2.0GHz computer to compute almost-Delaunay thresholds for a typical protein chain with 100–600 residues for typical values of cutoff and prune parameters. The companion paper (Bandyopadhyay and Snoeyink, 2004) has more details of the algorithm and a detailed analysis of running time.

Delaunay probability: The almost-Delaunay simplices $AD(\epsilon)$, as defined, capture a worst-case scenario for perturbations of size ϵ. For the typical case, assuming known, independent probability distributions for the positional uncertainty of each site, we may compute the probability of each tetrahedron occurring in the Delaunay tessellation. We are able to express this *Delaunay probability* for each tetrahedron as an integration problem that can be approximated by Monte Carlo methods (Bandyopadhyay and Snoeyink, 2004). We can show that the almost-Delaunay tetrahedra with small ϵ are the only ones that need to be considered; an efficient procedure for generating all $AD(\epsilon)$ tetrahedra allows us to scale our analysis to proteins.

SNAPP

Simplicial Neighborhood Analysis of Protein Packing (SNAPP) scores protein structures using the likelihood of neighboring four-tuples of residues from the Delaunay tessellation of their sidechain centroids. Carter et al. (2001) observed the frequencies of the 8855 unique combinations from choosing 4 of 20 amino acids with replacement in a training set of 1100 proteins selected to span different folds and families. Krishnamoorthy and Tropsha (2003) additionally divide the tetrahedra into five classes, based on adjacency of their residues along the backbone. A *four-body potential table* records the log-likelihood of each four-tuple in the Delaunay tessellations of proteins in the training set. Each new protein structure is scored by summing the potentials of its own Delaunay tetrahedra. Krishnamoorthy and Tropsha (2003) weight each tetrahedron's score by tetrahedron type (specifically, by the ratio of the type's frequency to its frequency in the training set), which improves discrimination between native proteins and decoys.

We evaluated the sensitivity of the SNAPP scores to a change in the Delaunay tessellation in two ways, which correspond roughly to average and worst-case perturbations. For average case, we estimated the Delaunay probability (Bandyopadhyay and Snoeyink, 2004) of each tetrahedron, and evaluated new potentials and new scores by weighting each tetrahedron by its probability. We assumed that the average radius of perturbation was 0.1 Å. For the worst case, we used the subset of the AD tetrahedra with threshold at most 0.3 Å, in addition to the Delaunay tetrahedra (*D+AD-SNAPP*) or instead of the Delaunay tetrahedra (*AD-SNAPP*). This threshold cutoff was chosen so that almost every residue is touched by some AD tetrahedron.

When comparing scores, one must be aware of their sensitivity to the number of tetrahedra. The SNAPP scores as defined in Carter et al. (2001) tend to increase with the number of Delaunay tetrahedra, since more Delaunay tetrahedra indicate better packing. The sum of

the Delaunay probabilities of AD(0.3) tetrahedra for all proteins in our training set are within 99–101% of the number of Delaunay tetrahedra at any value of edge length prune. Thus, *Delaunay-probability SNAPP* and original SNAPP scores can be compared directly.

When we augment the Delaunay with additional almost-Delaunay tetrahedra, however, the SNAPP score should not automatically increase. In fact, we have seen that more AD tetrahedra may indicate lower stability and worse packing. To perform comparisons between Delaunay, AD or D+AD-SNAPP, therefore, we may divide each residue's score by the number of tetrahedra it appears in (per-residue *local averaging*), or divide the total score for a protein by the total number of Delaunay, AD or D+AD tetrahedra used to compute it (*global averaging*). Global averaging has the advantage of providing a single number as a SNAPP score, which allows for easier comparison than the profile of the per-residue scores from local averaging.

We adapted the C++ code of Krishnamoorthy and Tropsha (2003) to evaluate log-likelihood potentials and compute SNAPP scores using Delaunay probability and almost-Delaunay tetrahedra. The training set was picked from the CulledPDB and WHATIF databases as described by Krishnamoorthy and Tropsha (2003). We generated tables of 4-body potentials using Delaunay probabilities, and using AD or D+AD tetrahedra, which show the same trend in log-likelihood scores for individual four-tuples as SNAPP, with some noise. In the Discussion section we compare the SNAPP scores for native proteins and their decoys from the *4state_reduced* decoy set in the Decoys 'R' Us database (Samudrala and Levitt, 2000), on which the original SNAPP scores best distinguish decoys from the native structure.

RESULTS AND DISCUSSION

We sought to answer three questions:

1. We relax the definition of Delaunay tetrahedra to include almost-Delaunay tetrahedra up to some threshold ϵ. What effects does this have on data ranging from random point sets to synthetic chains to native protein structures? Are the effects different when C_α or sidechain centroids are used as representative points?
2. Does the SNAPP analysis of protein packing give similar results when applied to this enlarged set of tetrahedra (whether weighted by probability or unweighted)?
3. Can patterns of almost-Delaunay tetrahedra be used to recognize structural motifs?

Distribution of almost-Delaunay tetrahedra

To compare almost-Delaunay tetrahedra on proteins and non-protein data structures, we selected a small protein size range (62–68 residues) for which we knew we had a good set of decoys from the Decoys 'R' Us database (Samudrala and Levitt, 2000). We chose an AD threshold cutoff of 2 Å and pruned edges that were longer than 10 Å. Because the packing density is important in Voronoi and Delaunay analysis (Gerstein et al., 1995; Liang and Dill, 2001), we selected point sets that fit in a bounding box with sides of 20–25 Å, with the exception of chains from random walks.

Uniform Random: 33 instances of 64 randomly generated points, uniformly distributed in a 20 Å cube.

Non-Colliding Random Walks: 33 instances of 64-step random walks generated by removing packing potentials from a Monte Carlo chain-growing algorithm (Gan et al., 2000, 2001), reimplemented by David O'Brien. The C_α carbons of the chain backbone are grown on a 3-1-1 lattice with angle constraints on each step and inter-C_α distances greater than unity. These do not respect the bounding box.

Folded Chains with MJ Potential: 33 instances from the above chain-growing algorithm, using a pairwise potential (Miyazawa and Jernigan, 1996) for the 65 residue protein 2cro and enforcing a 20–25 Å bounding box to generate more protein-like structures. This method does not do a good job of growing secondary structure.

Decoys 'R' Us: 33 instances from the *4state_reduced* decoy data set of Samudrala and Levitt (2000) that pack into a box with 20–25 Å sides. These are built to have good secondary structure, but may have suboptimal packing.

Protein Represented by C_αs: 33 small proteins with 60 to 69 residues that lie in a similar bounding volume, chosen from the CulledPDB database (Wang and Dunbrack, 2002) for less than 25% sequence identity, better than 2.4 Å resolution and R-factor 0.3(listed in Table 3).

Protein Represented by Sidechain Centroids: the same 33 proteins.

Synthetic α-helix: 33 instances of C_α atoms of residues along the helical path of radius of 2.3 Å, with a pitch of 5.4 Å and 3.6 residues per turn, with ± 0.125 Å uniform random noise applied tangentially and radially.

For each of these point sets, we produced histograms of the distribution of the almost-Delaunay tetrahedra for thresholds from 0 to 2 Å, with buckets at every 0.1 Å. In Figure 3 we plot the mean values in each bucket, with error bars for the standard deviations. The 0 bucket contains the Delaunay tetrahedra only, so it is drawn darker.

We can make some observations from these graphs.

1.) Random walk had the smallest number of Delaunay and almost-Delaunay tetrahedra. This set did not respect the 20–25 Å bounding box, so many of its tetrahedra had edges longer than 10 Å and were pruned.

2.) The α-helix had a similar low number of tetrahedra, with a striking distribution of positive almost-Delaunay thresholds: three sharp peaks at $\epsilon = 0.3$, 0.7 and 1.2,

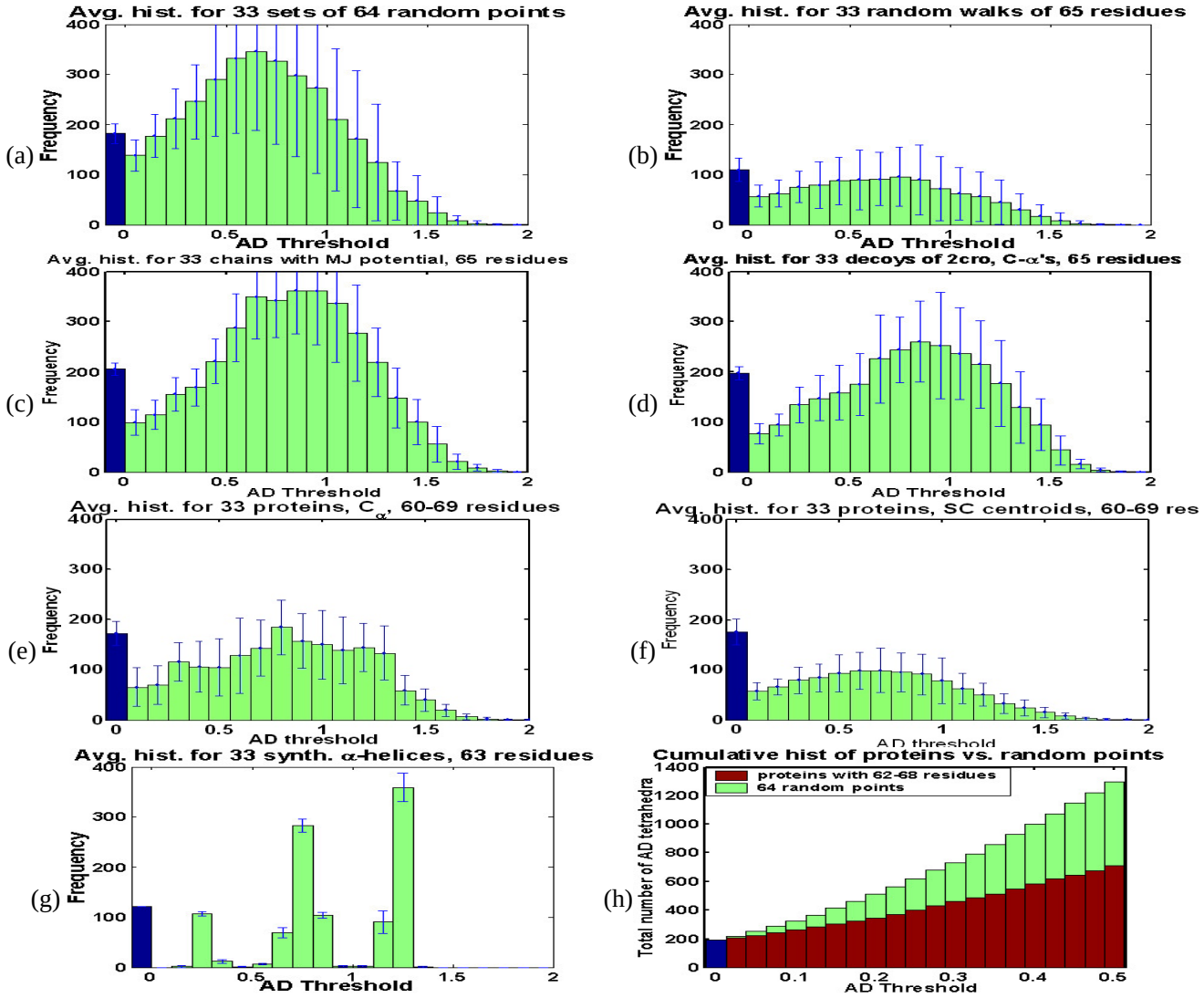

Fig. 3. Mean histograms of AD threshold for different structures. (a) random points (b) random walks lattice (c) chains folded with MJ potential (d) Decoys 'R' Us chains (e) proteins represented by C_α (f) proteins represented by sidechain centroids (g) synthetic α-helix. (h) shows that the cumulative number of tetrahedra in the average protein grows slower than in the average random point set.

which arise from the regular geometric pattern. Although not seen in the summary graph, individual C_α histograms also reveal α-helix peaks; we will show that they characterize residues in α-helices.

3.) Proteins represented by sidechain centroids produce the same number of Delaunay tetrahedra, but fewer AD tetrahedra than those represented by C_αs. With sidechain centroids, the residue positions are more widely spaced and the number of short edges is likely to be smaller. The edge-length prune can be increased to compensate. Individual side-chain centroid histograms do not show differences in structure as much as C_αs.

4.) The progression from random points, to chains with MJ potential, to the decoys with good secondary structure to the proteins shows that as the well-packed point sets become more structured, the number of AD tetrahedra decreases. In proteins there is a noticeable drop in the number of almost-Delaunay tetrahedra at low threshold values relative to the number of Delaunay tetrahedra, and the number of tetrahedra do not grow as quickly as they do in random points or chains. (See Figure 3(h).)

The last observation suggests that fewer tetrahedra can change under geometric perturbation in proteins than in random point sets. This is reassuring, since we depend on PDB coordinates for geometric analysis of proteins. Rather than place undue significance on this, however, we go on to explore how statistical potentials can change when the almost-Delaunay tetrahedra are added.

Robustness of SNAPP analysis

Figure 4 plots SNAPP scores for the protein 2cro and a tenth of its *4state_reduced* decoy set, in order of increasing RMSD (similar data and figures are available for 1sn3, 3icb, 4rxn and 4pti on the web and in the appendix). We used globally averaged scores computed on the AD(0.3) tetrahedra of the sidechain centroids for most of these experiments. A "weighted" score means we weight each tetrahedron type by its frequency (Krishnamoorthy and Tropsha, 2003), in addition to weighting by Delaunay probabilities. We summarize our findings below:

1.) The Delaunay-probability SNAPP score is within $\pm 5\%$ of the SNAPP score for 99% of the decoys, while the weighted Delaunay-probability score also closely follows the SNAPP score. Both unweighted and weighted methods are able to distinguish the decoys that SNAPP can distinguish.

2.) For most of the proteins in this decoy set, the weighted Delaunay probability SNAPP tends to increase the distinction between the score of the native structure and some decoys with SNAPP scores close to it. Figure 7 shows one of the exceptions, where many decoy scores higher than the native structure are worsened.

3.) Globally averaged AD-SNAPP and D+AD-SNAPP scores loosely follow SNAPP, though SNAPP itself shows only a weak decreasing trend with increasing RMSD. Weighted SNAPP scores distinguish decoys from the native structure better than unweighted scores (Krishnamoorthy and Tropsha, 2003), but both scores are less successful when averaged for comparison, since the number of tetrahedra does play a part in the distinction. D+AD-SNAPP and AD-SNAPP scores are almost equally successful or unsuccessful as the averaged SNAPP scores.

4.) The per-residue averaged Delaunay and Delaunay-probability profile scores are within $\pm 6\%$ for 99% of the residues in our decoy set. Per-residue averaged scores of AD and D+AD SNAPP loosely follow those for SNAPP, though there are outliers. 88% of residue AD-SNAPP scores and 98.5% of residue D+AD-SNAPP scores in the set were within $\pm 20\%$ of SNAPP.

5.) Among scores computed from C_αs and sidechain centroids, the centroids are better able to distinguish decoys from the native state for all our variants, as observed by Krishnamoorthy and Tropsha (2003) for SNAPP.

These comparisons allow us to make the following claim: the Delaunay tessellation is a robust measure of the quality of protein packing as evidenced by the invariance of relative SNAPP scores between proteins and decoys and the numeric similarity of total and profile Delaunay-probability SNAPP and Delaunay SNAPP scores. Further analysis of the discrepancies between the scores may indicate structures and residues where using the Delaunay to calculate scores may lead to errors.

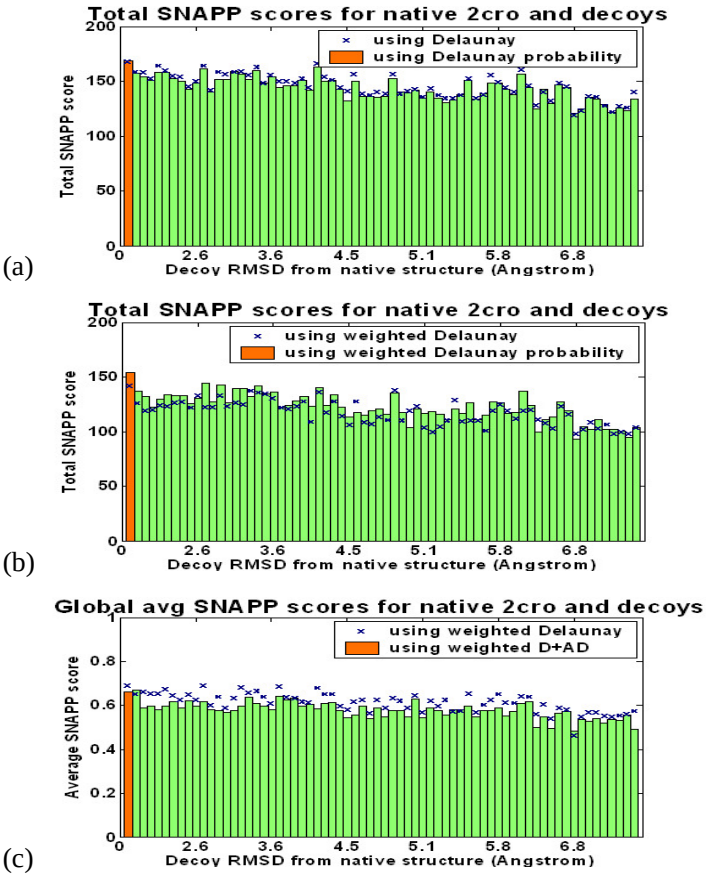

Fig. 4. Comparing scores of some SNAPP variants for 2cro native state (darker bar at extreme left) and some decoys in order of increasing RMSD, against Delaunay based SNAPP scores (x's). (a) Delaunay-probability vs. Delaunay (b) weighted Delaunay-probability vs. weighted Delaunay (c) Weighted D+AD vs. globally averaged weighted Delaunay

Determining secondary structure motifs

Wako and Yamato (1998) suggested that the Delaunay tessellation of backbone C_αs gives a framework to recognize structural motifs in proteins. The almost-Delaunay tetrahedra extend this framework and make it more discerning.

Each tetrahedron will use a set of residues that can be denoted by their sequence numbers, such as $i + (1\,2\,4\,5)$, or by a vertex use/gap pattern, ●●○●●. This pattern and the sequential pattern, $i + (1\,2\,3\,4)$ or ●●●●, occur in helical regions of a protein (Singh et al., 1996).

Wako and Yamato (1998) define a code for each Delaunay tetrahedron τ based on relative ordering of the vertices of τ and its up to four neighbors. They show example superpositions of common structures that have the same codes. Not all common structures will have the same code, however; changes to the Delaunay due to perturbation of coordinates can change the codes.

By considering the almost-Delaunay tetrahedra we can search for patterns in the backbone sequence and the threshold values at which they arise, and detect motifs more accurately and robustly. We wrote a MATLAB program to tabulate the frequent patterns for tetrahedra and the associated distributions of AD threshold, and applied it initially to synthetic models of secondary structure motifs, such as the α-helices described earlier. For each motif and its associated patterns, we modified the histogram plot of AD thresholds to draw a stacked bar chart of the AD tetrahedra classified according to the pattern they fall into. These *pattern histograms* can reveal the structural motifs in regular histograms. We use them to discriminate three basic secondary structure elements: α-helices, β-sheets and β-turns.

Discriminating the α-helix The Delaunay tessellation of the α-helix is built on two repeating patterns mentioned above. The almost-Delaunay tetrahedra add several patterns with characteristic threshold values that are detailed in Table 1, and visible in Figure 3(g).

$\mu(\epsilon)$	$\sigma(\epsilon)$	Patterns
0	0.00	●●●● ●●○●●
0.31	0.11	●○●●● ●●●○●
0.64	0.03	●●○○●●
0.74	0.04	●○○●●● ●○●○●● ●○●●○● ●●○●○● ●●●○○●
0.82	0.08	●○○●●○● ●○●●○○●
1.22	0.04	●○○○●●● ●○○●○●● ●○●○○●● ●○●○●○● ●●○○○●● ●●○○●○● ●●○●○○● ●●●○○○●

Table 1. Patterns for $AD(\epsilon)$ tetrahedra in a synthetic α-helix. Prune $= 10$ Å and cutoff $\epsilon < 2$ Å.

In a qualitative analysis, we studied pattern histograms for 30 proteins with varying degrees of α-helical content, and for decoys from the Kesar and Levitt (1999) local minima decoy set. Residues were represented by C_αs or sidechain centroids, the threshold cutoff was set to 2.0 Å, and the edge length prune was varied between 9.0 and 12.0 Å. Figure 5 shows typical pattern histograms for the protein 2cro. The α-helical peaks are present but somewhat diffuse for C_α histograms of proteins with α-helical content, and are lacking for sidechain centroids and for proteins with no significant α-helical content, e.g. immunoglobulin and γ-chymotrypsin. For decoys built by fixing the helix structure, the α-helical peaks are sharp, but there are noticeably fewer *non-pattern tetrahedra* (tetrahedra whose corresponding threshold values do not fall into the associated patterns). This indicates poorer packing of secondary structure. It will be interesting to investigate non-pattern tetrahedra as a tool for distinguishing decoys from the native state.

For more quantitative analysis, we can isolate individual α-helices using the patterns and AD thresholds. We partition the AD tetrahedra by pattern and keep only the tetrahedra whose thresholds are in a range that is characteristic of each pattern for an α-helix. That is, we

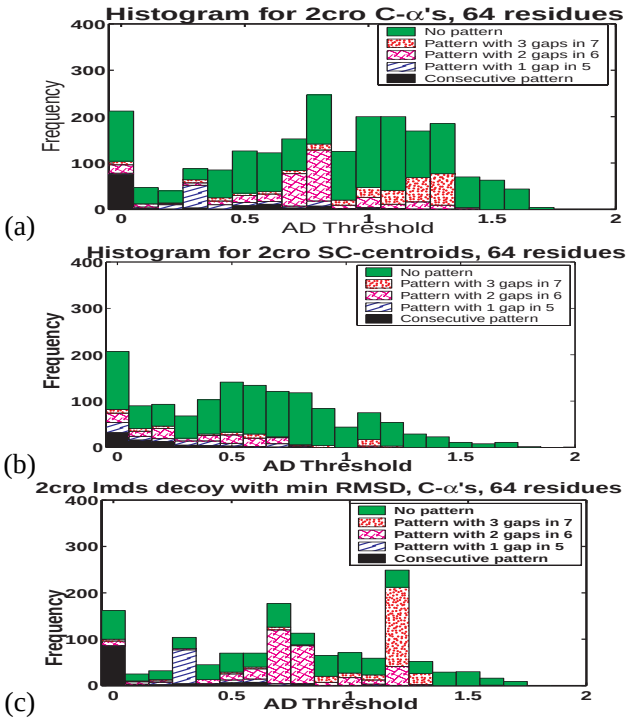

Fig. 5. Comparing the pattern histograms of 2cro and its decoys. (a) 2cro C_αs (b) 2cro sidechain centroids (c) LMDS decoy with minimum RMSD

keep thresholds between 0.2–0.4 for the patterns with one gap in 5 residues, between 0.6–0.9 for patterns with 2 gaps in 6 residues, and between 0.9–1.3 for patterns with 3 gaps in 7 residues. Next we count, for each pattern, how many Delaunay and AD tetrahedra use each residue. Based on these counts we decide at which residues an α-helix can start or end. Empirically, 4, 8 and 8 tetrahedra in two out of the three patterns with 1, 2 and 3 gaps in sequence is enough to start a helix, a total of 10 tetrahedra in all patterns is required to maintain it, and any of the 3 counts becoming 1 or zero is low enough to end it.

We observe from Table 4 that the Delaunay patterns alone cannot distinguish α-helices from 3_{10}-helices. However, filtering using AD thresholds removes most of the tetrahedra in 3_{10} and π helices, and the empirical rules for α-helix start and stop eliminate the remaining. Thus our AD patterns are capable of distinguishing α- from 3_{10}-helices purely using geometric criteria.

Evaluation of α-helix assignments We compared the assignments made by our algorithm with DSSP (Kabsch and Sander, 1983) for a subset of 45 proteins chosen to span the different architectures in CATH (Orengo et al., 1997). The numbers of residues in α-helices were generally within 5–10% of DSSP, as seen in Table 2. Most individual helices were correctly detected, up to an error of two residues in the start or end positions. These numbers indicate a good match between our method and a standard

method for α-helix detection.

The largest deviations from DSSP secondary structure assignments can be explained by non-robustness of DSSP's measures of hydrogen bonding. A sequence of residues (118–133) in the unstructured protein 1bg5 that looks convincingly like an α-helix is classified as β-turn by DSSP since it is distorted and missing a few H-bonds. The PDB header records indicate a helix structure, and the AD method finds this helix. Thus our assignment has 50% more α-helix than DSSP on 1bg5.

PDB ID	#	α-helix		β-strand		β-turn	
/chain	resid	DSSP	AD	DSSP	AD	PRO	AD
1aa8A	340	91	83	96	100	49	46
1ahl	49	0	0	6	5	14	10
1aorA	605	246	247	82	97	82	77
1b2p	238	0	0	110	97	40	35
1bg5	254	70	102	0	12	68	32
1bp1	456	100	87	204	196	48	43
1brx	209	158	158	10	8	12	10
1cem	363	168	162	8	28	52	47
1div	149	48	54	46	38	13	13
1dlc	584	177	191	174	167	73	56
1dze	225	164	179	10	4	14	9
1ejdA	418	128	138	105	134	43	40
1era	62	0	0	23	30	10	8
1f8d	388	5	4	171	177	71	59
1gab	53	35	29	0	0	0	2
1gmc	240	17	19	78	72	55	46
1havA	216	11	11	98	78	29	37
1hcd	118	0	0	55	35	25	16
1ilg	270	142	143	21	14	30	21
1kapP	470	66	73	98	153	91	74
1lrv	233	90	100	0	0	29	36
1lxa	262	40	44	70	96	60	51
1mbn	153	118	115	0	0	6	8
1npoA	81	0	0	26	15	16	17
1oen	524	133	112	126	138	86	94
1ospO	251	8	8	131	123	42	40
1pdc	45	0	0	6	8	13	14
1plq	258	37	37	111	117	43	30
1pprM	312	220	220	0	4	16	10
1rie	127	8	10	43	29	28	24
1rthA	543	157	157	127	125	70	46
1timA	247	106	101	42	51	15	18
1tl2	235	20	4	96	63	41	51
1tsg	98	10	17	4	12	39	20
1vdf	230	185	185	0	0	8	13
1ytf	100	34	37	40	22	24	10
2acy	98	24	24	41	18	10	6
2bnh	456	188	171	52	85	58	62
2hgf	97	9	10	28	25	26	19
2imm	114	0	0	58	57	33	17
2vsgA	358	166	181	17	23	58	46
3daaA	277	79	82	81	57	40	32
4bcl	350	55	62	171	169	44	30
4jdwA	360	82	80	71	77	55	51
8gch	240	21	24	78	76	55	41

Table 2. Numbers of α-helical and β-sheet residues assigned by DSSP (Kabsch and Sander, 1983) and β-turn residues assigned by PROMOTIF (Hutchison and Thornton, 1996), and by our AD patterns and thresholds for 45 protein chains with different CATH architectures.

We conclude that peaks in the distribution of the AD tetrahedral thresholds are a reliable, *sequence-independent* means of determining that a protein has α-helix and quantifying the number and location of the helices.

Discriminating β-sheets We investigated the AD threshold distribution of several proteins that are classified as mainly β under the CATH classification (Orengo et al., 1997). No single threshold value seems to be characteristic of β-sheets, and all patterns based on sequence interval have a standard deviation of around 0.3 Å, indicating a large spread and no peaks. The relative flatness of β-sheets makes their AD tetrahedra dependent on the positions of neighbors rather than on local geometry, so that picking a tetrahedron from residues on two or three strands with varying horizontal offset (*skew*) and inter-strand *separation* does not yield a pattern in the AD threshold.

However, the *existence* of AD tetrahedra that span two strands does yield a pattern, as long as they have a relatively low threshold (we used 1.0 Å) and are evaluated at a low value of the edge length prune, at most 10 Å. This pattern is seen in the *maximum gap in sequence* of the tetrahedron, which as we observed was 1, 2 or 3 in the case of α-helical patterns. For two parallel β-strands, tetrahedra consecutive in sequence along the strands have the same maximum sequence gap, so that allowing for skew in the tetrahedra, the histogram distribution of the gap shows a sharp peak with width equal to the skew, corresponding to the sequence separation between the strands.For anti-parallel β-strands, the sequence gap is distributed in a consecutive interval as the tetrahedra step from one end of the strand to another, leading to a plateau in the histogram with a gentle rise towards the center of the interval where skewed tetrahedra from both sides tend to converge. Thus we can conservatively isolate the β-sheet tetrahedra by detecting ranges in this histogram lying in plateaus or peaks, as shown in Figure 6.

We can isolate the beta strands from the tetrahedra by searching for the residue on each strand that is "most connected" with a residue on another strand, much like we did for α-helices. We implement this search as a mutual maximum frequency of occurrence search for each residue. If residue a occurs most frequently in tetrahedra with residue b, and vice versa, then a and b are most connected, and we choose them as neighbors in adjacent β-strands. In this way we build up a list of candidate strand neighbors, and then cluster those that are in parallel or anti-parallel sequences to complete the β-sheet.

Evaluation of β-sheet assignment: We tested the β-sheet residue count as well as individual strand positions against the DSSP values for the same 45 proteins. The results are summarized in Table 2.

In general, β-sheet determination was less accurate than α-helix determination, since we could not find a signature

based on the AD thresholds to restrict the search, and often segments that were parallel to each other but did not have the geometry of a β-sheet were misclassified, for example, two β turns in front of adjacent parallel strands in the β-α-β protein 2bnh. To avoid this kind of error, we modified our method to take the α-helix and β-turn determinations to be more accurate, and reject detected "sheets" that overlap with an α-helix or a β-turn. We also reject sheets with less than two residues in each strand. Note that this method does not detect isolated β-bridges or strands.

Detecting β-turns The primary criteria for defining beta turns (Lewis et al., 1973; Richardson, 1981; Hutchison and Thornton, 1994) have been that the first and fourth residues in the sequence have their C_αs less than 7.0 Å apart, and the residues involved are not helical. According to the Richardson classification that is most widely used at present, there are 6 distinct types of turns (I,I',II,II',VIa and VIb) based on geometry and chirality, along with a miscellaneous category IV that captures anything not fitting into the previously defined categories.

The approach of Wako and Yamato (1998) does not work for all beta-turns, since beta-turns do not differ significantly from other motifs in their sequence ordering and nearest-neighbor tetrahedra codes. Tropsha and others (Singh et al., 1996) have used tetrahedrality as an additional geometric discriminator. Tetrahedrality is a measure of deviation of edge lengths of a tetrahedron from those of an ideal tetrahedron. It is defined as below, where l_i is the length of edge i, $i \in \{1..6\}$:

$$T = \sum_{i<j} (l_i - l_j)^2 / 15\bar{l}^2 \quad (1)$$

Tetrahedra that include the C_αs of four residues in a beta-turn have a high degree of similarity in their edge lengths, which leads to a low value of tetrahedrality (less than 0.2).

We present a hybrid approach using AD tetrahedra, where we examine the AD thresholds of tetrahedra with four residues in sequence and with C_α^i–C_α^{i+3} distance being less than 7.0 Å. The pattern that corresponds to a β-turn is, broadly speaking, a low or zero threshold and low tetrahedrality at the tetrahedron that is in the turn, surrounded by higher thresholds on either one or both sides of the turn. More precisely, we have determined 6 different categories of turns, corresponding to threshold ranges as given in Table 5. Turns that overlap detected helices are rejected.

Evaluation of β-turn detection Our categories of residues in β-turns do not coincide with the Richardson classification (Richardson, 1981), though usually the β-turns that we miss lie in the miscellaneous (IV) category whose geometry is hardest to classify. The AD method performs somewhat better than combining the C_α distance (<7.0 Å) and tetrahedrality (< 0.2) criteria, and significantly better than Delaunay-based methods (Wako and Yamato, 1998).

We do not compare the accuracy of β-turn assignment to DSSP directly, since DSSP classification of β-turns as pairs of residues in S and T conformation (turns without and with hydrogen bonding) is observed to be inaccurate. Instead, we compare with the PROMOTIF program (Hutchison and Thornton, 1996), which detects turns based on phi and psi angles and classifies each using the Richardson classification (Richardson, 1981).

Modeling pseudo-visual secondary structure assignment We wanted to compare the results of our method, along with DSSP, against a method that models assignment of secondary structures by visual inspection of the C_α trace by a human expert, as suggested by David and Jane Richardson (personal communication). Our "pseudo-visual" helix assignment is based on the idea of fitting residues on the surface of a cylinder, as described by (Drennan et al.). We implemented Kahn's method to find the helix axis (Kahn, 1989), taking the cross product of the bisectors of the angles formed by three consecutive C_αs as the local axis direction, and using least-squares fitting. By checking that other parameters are within normal ranges for an α-helix (local curvature 94–104°, pseudo-dihedral angle 35–58°, rise per residue 1.25–1.85 Å, and local radius 2.4 ± 0.3 Å), we can approximate pretty well the visual intuition of residues lying on a cylinder and discriminate 3_{10} helices from α-helices. Tolerance in the axis direction and a pass that smoothes isolated gaps in helices ensure that helices with imperfections and even bends are assigned correctly, e.g. helix G of 1MBN.

We have not yet implemented a pseudo-visual method for assignment of β-sheets. The GAS-P program (, Drennan et al.) identifies individual residues in β-strand geometry using a strategy much like Kahn's method, though it does not search for residues on adjacent strands or model the distance and twist between strands.

The "pseudo-visual" β-turn assignment currently uses the distance criterion that is the definition of a β-turn, and the condition of low tetrahedrality (< 0.2) as described above. Though tetrahedrality is not a visually intuitive parameter, the results from this method look surprisingly accurate, so we include it as an alternative method to compare with the AD and DSSP (PROMOTIF) assignments.

Comparison between AD, DSSP and pseudo-visual assignment To compare AD and DSSP assignments alone, as noted above, we generated side-by-side structural assignments for the residues in 45 protein chains representing almost all the CATH (Orengo et al., 1997) architectures, and summed up the assignments in α-helix, β-sheet and β-turn categories in Table 2. We also developed tools to convert the AD tetrahedra and secondary structures assigned us-

ing AD, DSSP and pseudo-visual methods into the popular Kinemage format (Richardson and Richardson, 1994) for 3D visualization and comparison. The structural assignments and kinemages are available on our web site. We observed that AD, DSSP and pseudo-visual helix assignments agree in most cases, while there are often discrepancies in the other assignments but for the most part they are visually plausible. In some cases as discussed for 1bg5, AD assignments match the pseudo-visual assignments while DSSP does not due to robustness issues.

In AD β-sheet assignments, often the distance between the two strands is non-uniform or differs too much from the ideal hydrogen bonding distance; thus DSSP does not find a β-sheet while AD does. Thus augmenting AD with inter-strand distance and other geometric criteria may improve its accuracy. In some cases (e.g. THR153–ALA155 and ASP108–GLY110 in 4aah chain A) AD finds a sequence to be β-sheet, and DSSP does not, though visually the inter-C_α distances and geometry seem right for hydrogen bonding. There may be differences from the canonical β-sheet (e.g. inter and intra-strand twist or differences in side-chain positions) that are not detected by AD. But in some cases DSSP assignment may not be robust, and AD may lead to a structure assignment that conforms to the visual by tuning geometric parameters.

CONCLUSIONS AND FUTURE WORK

We have introduced the tools of almost-Delaunay tetrahedra and Delaunay probability to give worst-case and average-case analysis of the Delaunay tessellation under perturbation of the input sites, and we have given two initial applications of these tools. Our experiments indicate that SNAPP scores are robust, quantify this to some extent, and our variants of SNAPP may be used in applications where robustness is critical. One goal of our future work would be to pinpoint the location and magnitude of the error such an application makes by using the Delaunay.

Our method for α-helix detection performs well quantitatively and qualitatively, distinguishing different types of helices purely from geometric criteria. We can detect β-sheets and β-turns too, with somewhat less accuracy than helices. We have compared our secondary structure assignments numerically and visually with established methods to show that they are plausible. We plan to extend our approach to detect other structural motifs, and to consider the sensitivity of other analyses to the perturbation of coordinates in Voronoi diagrams and Delaunay tessellations.

ACKNOWLEDGEMENTS

We thank the biogeometry research group and members of the Tropsha lab for discussions. In particular, we acknowledge David O'Brien for his implementation of chain folding, Bala Krishnamoorthy for his SNAPP scoring, and Robert Paul Berretty for work on the annulus problem. We also thank David and Jane Richardson for helpful discussions. DB and JS gratefully acknowledge support from NSF grants 9988742 and 0076984, and AT appreciates the support from the NSF grant ITR/MCB 011289 and a grant from North Carolina – Israel Research Partnership NCI 1999032.

REFERENCES

Angelov, B., J. Sadoc, R. Jullien, A. Soyer, J. Mornon, and J. Chomilier (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 Comput. Surv. 23*(3), 345–405.

Bakowies, D. and W. F. van Gunsteren (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.

Bandyopadhyay, D. and J. Snoeyink (2004). Almost-Delaunay simplices : Robust neighbor relations for imprecise points. In *ACM-SIAM Symposium On Discrete Algorithms*. to appear.

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

Boissonnat, J.-D. and M. Yvinec (1998). *Algorithmic Geometry*. UK: Cambridge University Press.

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

de Berg, M., M. van Kreveld, M. Overmars, and O. Schwarzkopf (2000). *Computational Geometry: Algorithms and Applications* (2nd ed.). Berlin, Germany: Springer-Verlag.

Delaunay, B. (1934). Sur la sphère vide. A la memoire de Georges Voronoi. *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskih i Estestvennyh Nauk 7*, 793–800.

Drennan, D., F. M. Richards, and P. C. Kahn. A geometrical analysis of the structure of proteins, with implications for protein folding. manuscript in preparation.

Gan, H. H., A. Tropsha, and T. Schlick (2000). Generating folded protein structures with a lattice chain growth algorithm. *J. Chem. Phys 113*, 5511–5524.

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

Garcia-Lopez, J., P. Ramos, and J. Snoeyink (1998). Fitting a set of points by a circle. *Discrete and Computational Geometry 20*, 389–402.

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

Hutchison, G. and J. Thornton (1994). A revised set of potentials for β-turn prediction in proteins. *Protein Science* (3), 2207–2216.

Hutchison, G. and J. Thornton (1996). Promotif-a program to identify and analyze structural motifs in proteins. *Protein Science* (5), 212–220.

Kabsch, W. and C. Sander (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*(12), 2577–2637.

Kahn, P. C. (1989). Defining the axis of a helix. *Computers and Chemistry 13*, 185–189.

Kesar, C. and M. Levitt (1999). lmds decoys from dd.stanford.edu.

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

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

Lewis, P., F. Momany, and H. A. Scheraga (1973). Chain reversals in proteins. *Biochim. Biophys. Acta 303*, 211–229.

Liang, J. and K. A. Dill (2001). Are proteins well-packed? *Biophys J. 81*(2), 751–766.

Liang, J., H. Edelsbrunner, P. Fu, P. Sudhakar, and S. Subramaniam (1998). Analytical shape computing of macromolecules II: identification and computation of inaccessible cavities inside proteins. *Proteins 33*, 18–29.

Liao, Y. C., D. J. Lee, and B.-H. Chen (2001). Description of multi-particle systems using Voronoi polyhedra. *Powder Technology 119*(2–3), 81–88.

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

Miyazawa, S. and R. L. Jernigan (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of Molecular Biology 256*, 623–644.

Munson, P. J. and R. K. Singh (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.

Okabe, A., B. Boots, and K. Sugihara (1992). *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. Chichester, UK: John Wiley & Sons.

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

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

Richardson, D. and J. Richardson (1994). Kinemages : simple macromolecular graphics for interactive teaching and publication. *Trends Biochem. Sci. 19*, 135–138.

Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Adv. Protein Chem. 34*, 167–339.

Samudrala, R. and M. Levitt (2000). Decoys 'R' Us: A database of incorrect conformations to improve protein strucure prediction. *Protein Science 9*, 1399–1401. http://dd.stanford.edu.

Simons, K. T., C. Kooperberg, E. Huang, and D. Baker (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology 268*, 209–225.

Singh, R., A. Tropsha, and I. Vaisman (1996). Delaunay tessellation of proteins. *J. Comput. Biol. 3*, 213–222.

Soyer, A., J. Chomilier, J. Mornon, R. Jullien, and J. Sadoc (2000). Voronoï tessellation reveals the condensed matter character of folded proteins. *Physical Review Letters 85*(16), 3532–3535.

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

Tsai, J., R. Taylor, C. Chothia, and M. Gerstein (1999). The packing density in proteins: Standard radii and volumes. *Journal of Molecular Biology 290*(1), 253–266.

Voronoi, G. M. (1908). Nouvelles applications des paramètres continus à la théorie des formes quadratiques. deuxième Mémoire: Recherches sur les parallélloèdres primitifs. *J. Reine Angew. Math. 134*, 198–287.

Wako, H. and T. Yamato (1998). Novel method to detect a motif of local structures in different protein conformations. *Protein Engineering 11*, 981–990.

Wang, G. and R. L. Dunbrack (2002). Pisces: a protein sequence culling server. *Bioinformatics*. Sumitted, \small{http://www.fccc.edu/research/labs/dunbrack/pisces/culledpdb.html%}.

Watson, D. F. (1981). Computing the n-dimensional Delaunay tesselation with applications to Voronoi polytopes. *Comput. J. 24*(2), 167–172.

Weberndorfer, G., I. L. Hofacker, and P. F. Stadler (1999). An efficient potential for protein sequence design. In *Computer Science in Biology. Univ. Bielefeld, D. GCB'99 Proceedings*, Hannover, Germany, pp. 107–112. http://www.tbi.univie.ac.at/papers/Abstracts/GCB026.ps.gz.

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

Zimmer, R., M. Wöhler, and R. Theiele (1998). New scoring schemes for protein fold recognition based on Voronoi contacts. *Bioinformatics 14*(3), 295–308.

APPENDIX

1bxyA	1d0dA	1h2sB	1i2tA	2igd	1dtdB	1nxb	
1isu	1vie	1f94	1kveA	1svfA	1c8c	1gl2C	
1icfI	1c9oA	2nllA	1b3aA	1b67A	1gutA	1fe0A	.
1e0bA	1tafA	1c4qA	1kliL	1dul	1r69	1gcqC	
1a8o	1napA	1f9rA	1ku5A	1f9sA			

Table 3. The 33 small protein chains whose C_αs and sidechain centroids were used to compare their AD threshold distributions against different datasets. These were chosen from the CulledPDB database (Wang and Dunbrack, 2002) to have 60–69 residues, less than 25% sequence identity, better than 2.4 Å resolution and R-factor 0.3.

$\mu(\epsilon)$	$\sigma(\epsilon)$	Patterns			
0	0.00	●●●●	●●○●●		
0.38	0.03	●○●●●	●●●○●		
0.63	0.03	●●○○●●			
1.06	0.03	●○○●●● ●●○●○●	●○●○●● ●●●○○●	●○●●○●	
0.82	0.06	●○○●●○●	●○●●○○●		
1.34	0.04	●●○●○○●	●●○○●○●	●○●○○●●	●○○●○●●
1.55	0.06	●○○○●●●	●●●○○○●		
1.65	0.06	●●○○○●●	●○●○●○●		

Table 4. Patterns for $AD(\epsilon)$ tetrahedra in a synthetic 3_{10}-helix. Prune $= 13$ Å and cutoff $\epsilon < 2$ Å.

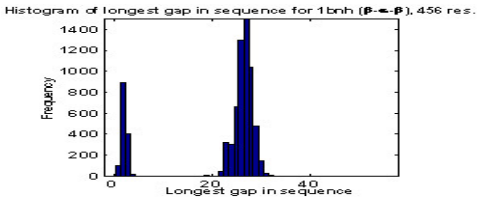

Fig. 6. The histogram of maximum gap in sequence for AD(1.0) tetrahedra in 1bnh shows two sharp peaks: the first at 3–4 corresponds to the α-helices, while the one at $\sim$ 25 indicates parallel β-sheets.

Category	**ϵ in turn**	**lower ϵ bordering turn**	**higher ϵ bordering turn**
1	$(0, 0.25)$	> 0.29	> 0.5
2	0	0	> 0.25
3	0	> 0.09	> 0.3
4	$[0.25, 0.4)$	> 0.5	> 0.7
5	$(0, 0.2)$	$(0, 0.2)$	> 0.3
6	$[0, 0.2)$	$[0, 0.2)$	$(0, 0.2)$

Table 5. Ranges of AD thresholds for tetrahedra in β-turns.

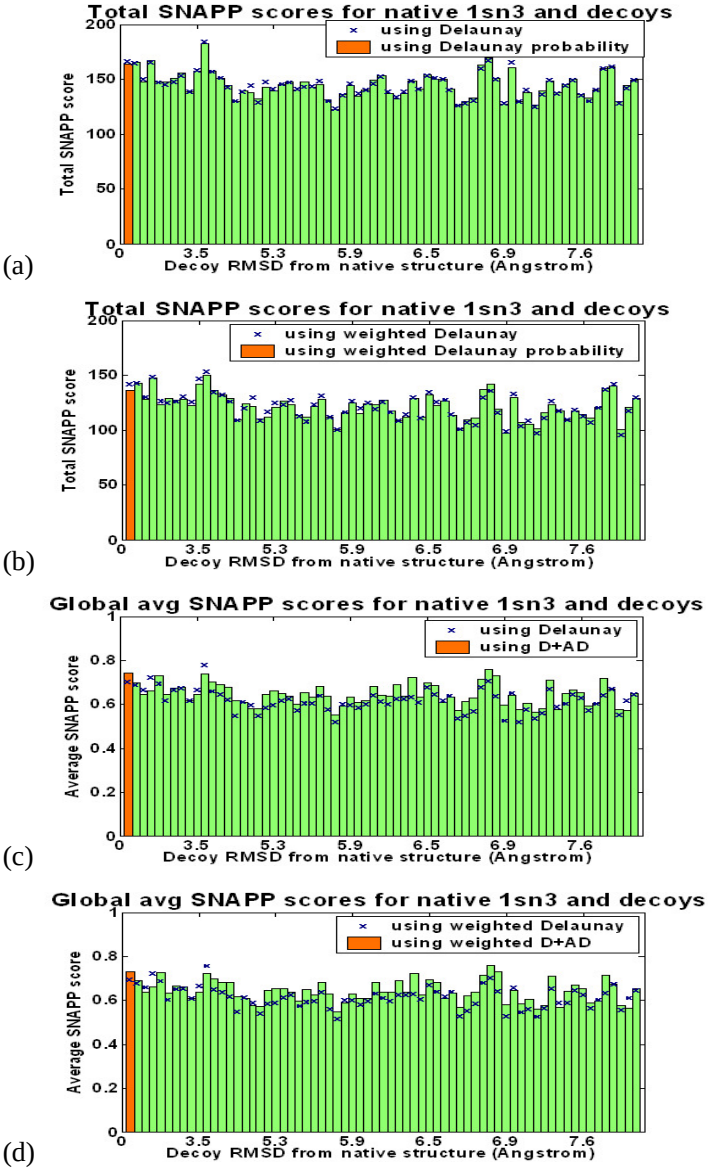

Fig. 7. Comparing scores of SNAPP variants for 1sn3 native state (darker bar at extreme left) and some decoys in order of increasing RMSD, against Delaunay based SNAPP scores (x's). (a) Delaunay-probability SNAPP vs. Delaunay SNAPP (b) Weighted Delaunay-probability SNAPP vs. weighted Delaunay SNAPP (c) D+AD-SNAPP vs. globally averaged Delaunay SNAPP (d) Weighted D+AD-SNAPP vs. globally averaged weighted Delaunay SNAPP