Analyzing Protein Structure Using Almost-Delaunay Tetrahedra

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

Keywords: almost-Delaunay, neighbors, protein structure, Delaunay probability, SNAPP, four-body statistical potential, motif detection, secondary structure

1 Introduction.

The Delaunay tessellation (DT) of a protein structure [6] collects representative points of four "neighboring" residues into tetrahedra. The DT has many applications in the analysis of protein structure, of which we consider two in detail – scoring folded proteins to distinguish the native state from decoys [2, 4] and detecting motifs of local structure [7].

The Delaunay tessellation is defined using an "empty sphere" criterion (Delaunay, 1934) – the circumspheres of Delaunay tetrahedra contain no other points. However, protein atom coordinates are subject to uncertainties from rounding, measurement, conformational change and motion, and small changes in the coordinates may cause large changes in the DT.

The almost-Delaunay tetrahedra [1] expand the set of Delaunay tetrahedra to account for perturbation or motion of point coordinates. We define a quadruple of points to be in the set of *almost-Delaunay tetrahedra* with parameter ϵ, denoted AD(ϵ), if there is a perturbation of all points by at most ϵ that makes its circumsphere empty. We denote the minimum such perturbation for a quadruple its *AD threshold*. The Delaunay tetrahedra have threshold 0.

2 Experiments and Results

Our implementation in MATLAB and C++ can calculate the AD tetrahedra for typical proteins of 100-1000 residues in a few seconds to a few minutes, for typical values of two parameters: the maximum edge length (*prune*) and maximum perturbation allowed (*cutoff*). By studying a large number of point sets with different structure, we observed that there are fewer AD tetrahedra at low thresholds in proteins than in random point sets; hence the DT is more stable in proteins.

Simplicial Neighbor Analysis of Protein Packing (SNAPP) [2, 4] scores protein structures by summing the frequencies with which the four-tuples of amino acids observed as Delaunay neighbors, occur in native protein structures. We calculated SNAPP scores using AD tetrahedra for 6 proteins and their decoys from the *4state_reduced* [5] set. Overall, the modified scores were as successful as the original at distinguishing proteins from decoys, and made a slightly stronger distinction between proteins and highest-scoring decoys. Thus, decoy discrimination using the DT is robust.

We observed that the histogram distribution of AD tetrahedra vs. threshold for an ideal α-helix has sharp peaks at $\epsilon = 0.3$, 0.7 and 1.2. These values of ϵ correspond to specific patterns in the residue sequence numbers, as shown in Figure 1. Histograms for proteins containing α-helices reveal the same peaks and patterns; we can mark the corresponding tetrahedra as α-helical, and determine the residues in α-helical conformation using a heuristic. We may identify β-sheets and β-turns similarly by decoding patterns present in their AD tetrahedra. For details of the methods, see http://www.cs.unc.edu/~debug/

[1]Department of Computer Science, University of North Carolina at Chapel Hill. E-mail: {debug, snoeyink}@cs.unc.edu

[2]Laboratory of Molecular Modeling, School of Pharmacy, University of North Carolina at Chapel Hill. E-mail: tropsha@email.unc.edu

papers/AlmDel. Secondary structures assigned using the AD method match the widely used DSSP [3] assignments in most cases (a few are shown in Table 1). They are consistent and robust in cases where DSSP is not, and are closer to visual assignments done by a human expert.

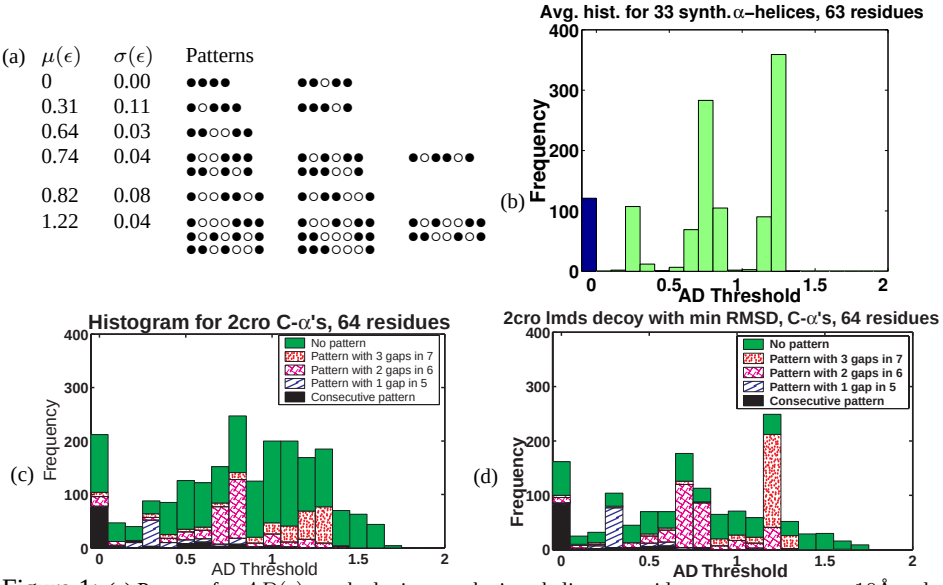

Figure 1: (a) Patterns for $AD(\epsilon)$ tetrahedra in a synthetic α-helix. $\bullet$=residue, $\circ$=gap, prune $= 10$Å and cutoff $\epsilon < 2$Å. (b) Histogram showing α-helical peaks, also seen in (c) 2cro and (d) a decoy with same secondary structure.

PDB ID	**#**	α-helix		β-sheet		β-turn	
/chain	**resid**	**DSSP**	**AD**	**DSSP**	**AD**	**PRO**	**AD**
1brx	209	158	158	10	8	12	10
1lrv	233	90	100	0	0	29	36
1timA	247	106	101	42	51	15	18
1bg5	254	70	102	0	12	68	32
1ejdA	418	128	138	105	134	43	40
1oen	524	133	112	126	138	86	94

Table 1: α-helical, β-sheet and β-turn residues assigned by DSSP [3] and by our AD patterns for 6 protein chains with varying lengths and CATH architectures.

References

[1] D. Bandyopadhyay and J. Snoeyink. Almost-Delaunay simplices : Nearest neighbor relations for imprecise points. In *ACM-SIAM Symposium On Discrete Algorithms*, pages 403–412, 2004.

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

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

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

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

[6] R. Singh, A. Tropsha, and I. Vaisman. Delaunay tessellation of proteins. *J. Comput. Biol.*, 3:213–222, 1996.

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