

A New Method for Protein Secondary Structure Assignment Based on a Simple Topological Descriptor

Todd Taylor and Iosif I. Vaisman¹

Keywords: protein topology, secondary structure assignment, computational geometry, Delaunay tessellation

1 Introduction.

We have analyzed protein structures with a geometrical construction known as the Delaunay tessellation [1]. Each amino acid is abstracted to a point, in this case the coordinates of the alpha carbon. These points are joined by edges in a unique way to form a set of non-overlapping, irregular, space-filling tetrahedra (Figure 1). The tetrahedra have the property that a sphere on the surface of which all four vertices reside does not contain a vertex from any other tetrahedron (the “empty sphere property”). The tetrahedra are called Delaunay simplices and the process by which they are generated is called tessellation.

Simplices resulting from the tessellation of protein structures can be divided into 5 classes based on the way the main chain threads through them [1]. In type 0 simplices, none of the four residues at the vertices of the simplex are consecutive in primary sequence. In type 1 simplices, exactly one pair are consecutive in primary sequence. In type 2, two pairs are consecutive, but the pairs are separated from each other in primary sequence. In type 3, exactly three residues are consecutive. In type 4, all four residues are consecutive. One can tally the number of simplices of each type that a given residue belongs to, and we call these sums *t-numbers*. For example, if residue i is a vertex in 10 type 0 simplices, 7 type 1 simplices, 8 type 2, 0 type 3, and 4 type 4 simplices, its t -numbers are $t_0(i)=10$, $t_1(i)=7$, $t_2(i)=8$, $t_3(i)=0$ and $t_4(i)=4$. The total number of simplices of all types in which a residue can participate varies greatly, ranging from 1 to 72 in the data set examined here.

There is a strong relationship between t -numbers and secondary structure (Figure 2). We have mapped t -numbers to DSSP secondary structure assignments via several methods [3] for a set of approximately 224,000 residues from a non-homologous set of 996 PDB protein x-ray structures obtained from PISCES [2]. The results of these mappings were compared to assignments from the existing methods DEFINE, DSSP, P-SEA, secstr, STRIDE, VoTAP, and XTLsstr. The degree of agreement between the t -number based methods and DSSP was in line with other secondary structure assignment methods as was the degree of self-consistency in assignment (Table 1). It is surprising that our assignment agrees well with previous methods since, unlike them, it is based solely on main chain topology and does not explicitly depend on any angles, lengths, areas, putative hydrogen bonds, or internal residue geometry.

2 Software and files.

All secondary structure assignments are available at http://binf.gmu.edu/struct_binf_group/sec_str/

¹ George Mason University, School of Computational Sciences, 10900 University Blvd., Manassas, VA, USA. E-mail: ttaylor@gmu.edu, ivaisman@gmu.edu

3 Figures and tables.

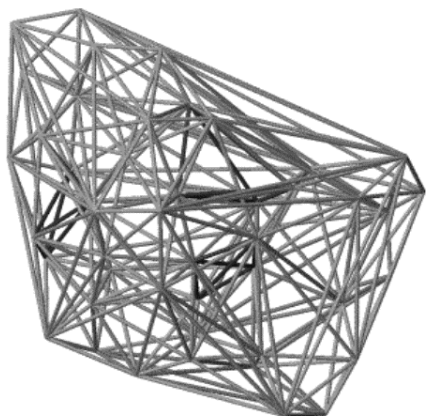


Figure 1: Delaunay tessellation of crambin

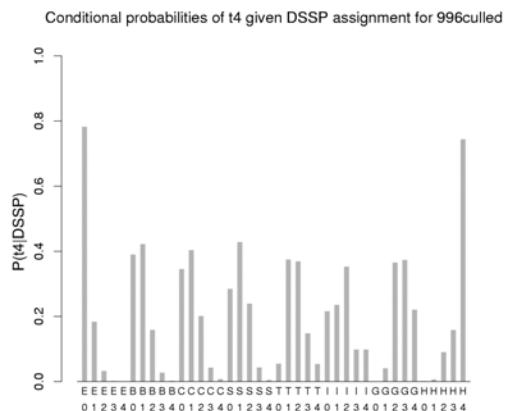


Figure 2: Conditional probability t4|DSSP

Assignment method	C (DSSP)	E (DSSP)	H (DSSP)
t4_C4.5 sensitivity	0.699	0.849	0.917
t4_C4.5 specificity	0.894	0.885	0.948
DEFINE sensitivity	0.354	0.928	0.915
DEFINE specificity	0.933	0.734	0.902
P-SEAsensitivity	0.778	0.773	0.853
P_SEA specificity	0.823	0.909	0.969
secstr sensitivity	0.934	0.821	0.999
secstr specificity	0.931	0.999	0.958
STRIDE sensitivity	0.920	0.988	0.978
STRIDE specificity	0.982	0.987	0.965
XTLsstr sensitivity	0.773	0.651	0.943
XTLsstr specificity	0.836	0.941	0.922

Table 1: Accuracy of assignment with respect to DSSP.

4 Selected References.

- [3] Quinlan R 1993. *C4.5: Programs for Machine Learning*, San Mateo: Morgan Kaufmann.
- [1] Singh RK, Tropsha A, Vaisman II 1996. Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. *J Comput Biol.* 3(2):213-221.
- [2] Wang G, Dunbrack RL, Jr. 2003. PISCES: a protein sequence culling server. *Bioinformatics* 19(12):1589-1591.