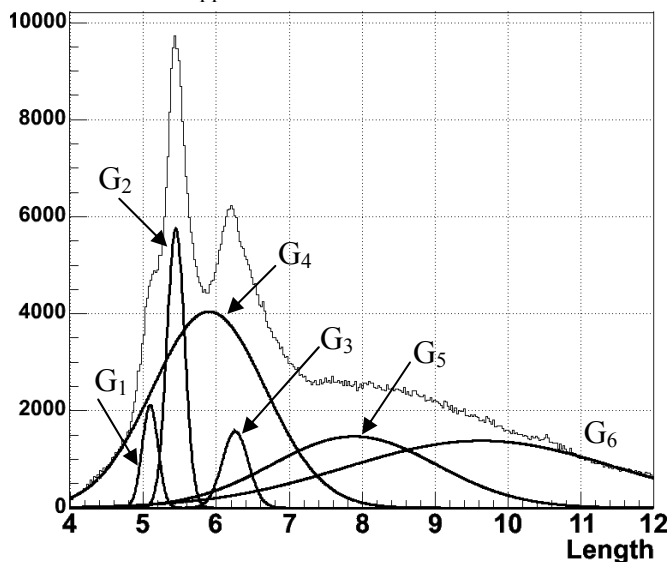


# Statistical characteristics of the Delaunay Tessellation of protein structure

Alexej Abyzov<sup>1</sup> and Valentin A. Ilyin<sup>1</sup>

**Keywords:** Delaunay Tessellation, protein structure, Voronoi Tessellation

Delaunay Tessellation (DT) appears to be useful in many protein structure studies. Here we present an analysis of statistical characteristics of DT in proteins on a large data set using non-redundant PDB. The presented work shows that features of protein structure, such as secondary structure, surface, residues mainchain contacts, etc., correlates with the features of the tessellation of proteins represented by  $C\alpha$  atoms. The hierarchical classification of the elements of DT is introduced. It is demonstrated that elements of various classes/types exhibit preferences to a particular geometrical shapes. It is shown that Delaunay Tessellation quite robust to experimental errors in protein structure and can be used as a definition of contacting residues as an alternative to the definition by distance. Impact of the tessellation on the protein surface is analyzed. The results suggest that the Delaunay Tessellation can be used as a robust model for description of protein structure at the backbone level of approximation.



**Figure 1. Distribution of the DT edge lengths along with 6 Gaussian functions sum of which describes the distribution.**

Bulk DT calculations on the reference set produces 1,172,149 edges, 1,616,548 faces and ~972,000 tetrahedrons, or in other words approximately 6.98 edges, 9.63 faces and 5.79 tetrahedrons per amino acid residue. Thus, it is possible to evaluate the average coordination

---

<sup>1</sup> Department of Biology, Northeastern University, 360 Huntington ave., Boston, MA, USA. E-mail: ilyin@mozart.bio.neu.edu

number of neighboring residues in DT. Since every edge connects two residues but has been counted only once then the average number of edges per residues has to be doubled, i.e. 14. Or in other words residue in protein has 14 neighboring residues defined by DT.

It has been found that the distribution (see Figure 1) of the contact lengths can be very well approximated as a sum of just a few Gaussian functions, i.e. consists of six different distributions: three of them are sharp and three of them are wide. The DT contacts between residues described by sharp  $G_1, G_2, G_3$  distributions are located mainly in regular elements of protein structure. For example, in  $\alpha$ -helices edges between residues  $(i, i+2)$  contribute to the 5.44 Å pick, edges between residues  $(i, i+3)$  contribute to the 5.09 Å pick, and edges between  $(i, i+4)$  residues contribute to the 6.26 Å pick. Interesting is that the area under the Gaussian curves, describing the population, is different, i.e. there are more edges with the length around 5.44 Å and 6.26 Å than edge with length around 5.09 Å. This difference is result of the contribution of edges from  $\beta$ -strands. Most of the edge in  $\beta$ -strands also contribute to the  $G_4$  distribution and just some fraction of them contributes to the distribution described by  $G_2$  and  $G_3$ . Even though the existence of the first three sharp picks is not surprising the fact that DT captures regular elements of protein structure is remarkable.  $G_4$  and  $G_5$  distribution are populated by the edges in  $\beta$ -strands, loops, and proteins interior. The  $G_6$  distribution rises from edges on convex hull produced by DT which represent protein surface. Since the proteins do not necessarily obey convex and have rather complicated surface, those edges do not represent neighboring relationship between residues but the curvature of the whole protein shape. Interestingly, the overlap between pairs of distributions  $G_4/G_5$  and  $G_5/G_6$  is large and this observation does not allow their clear separation. Thus the overall distribution of residue-residue contacts shows non-trivial shape and can be described as a sum of elementary distributions associated with particular features of protein structure.

Major element of DT are tetrahedrons, previously 5 different types of tetra have been introduced (Singh *et al.*, 1996). Here we have extended the previous classification. Taking into account the directionality of the polypeptide bond, three of previously defined types have been mirrored. Tetrahedrons in those pairs have different chirality, are not superimposable and therefore have different structural sense and also have different statistical characteristics. The tetrahedrons of different types exhibit preferences for particular geometrical shapes and not all of them are equally populated. Tetrahedrons of type X and C are found in  $\alpha$ -helices while tetrahedrons of type I and O are abundant of protein surface and in cavities.

Analysis of DT robustness demonstrates that it is indeed stable and only circa 6% of tessellation can change as a result of atom movements either due to protein flexibility or as a result of experimental errors.

A comparison of two definitions of the contacting residues has been performed, namely definition by edges of DT via traditional definition by cutoff distance between residues. It was shown that both definitions agree at most up to 70%, which suggest using definition by DT contacts as a useful alternative to the definition by distance.

The analysis presented here shows that many different features of protein primary, secondary, and tertiary structure result in specific aspects of DT. And DT is the parameter free way of making tessellation for a set of points. This suggests but does not prove that DT may be a common model for representation and analysis of proteins. The challenging will be a study in reverse direction, when analysis of DT allows one to make conclusions about protein structures. The results presented here can be useful in: protein secondary structure studies; protein surface definition; studies of protein flexibility; protein structural comparison; studies of protein folding kinetics.