Recognition of Tertiary Packing Motifs in Protein Structures using Delaunay Tessellation

Stephen A. Cammer[1] and Alexander Tropsha[2]

The Laboratory for Molecular Modeling, Division of Medicinal Chemistry and Natural Products, School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599

email: (1) cammer@email.unc.edu; (2) tropsha@email.unc.edu

A novel approach to recognizing recurrent sequence-structure patterns, or packing motifs, in proteins has been developed on the basis of Delaunay tessellation of protein structure. A structure is modeled using a united residue approach where each residue is represented by a point located at the center of its side chain. The tessellation partitions a structure into irregular tetrahedra, or simplices whose vertices correspond to clusters of four nearest-neighbor residues. Tetrahedral clusters with vertex residues not adjacent along the polypeptide chain were considered. These clusters were classified on the basis of their amino acid composition and three distances separating four vertex residues along the sequence; these distances are defined as the number of residues from the first to second, second to third, and third to fourth residues. A tertiary packing motif is defined as a quadruplet of specific nearest neighbor residues with three values of sequence separation between them. The analysis of three large databases of diverse protein structures (less than 30% sequence identity between any pair of sequences, 1922 structures total) identified 226 examples of recurrent motifs found in 391 proteins. To further evaluate the methodology, three groups of proteins from unique SCOP-defined families have been analyzed and tertiary packing motifs common to these protein groups have been identified. The proposed methodology termed Simplicial Neighborhood Analysis of Protein Packing (SNAPP) can be used to locate recurrent tertiary contacts as well as active site sequence patterns similar to *Prosite* signatures. Furthermore, the SNAPP approach will be useful in automating the analysis of structures determined in structural genomics projects.