DISTANCE-BASED IDENTIFICATION OF STRUCTURE MOTIFS IN PROTEINS USING CONSTRAINED FREQUENT SUBGRAPH MINING

Jun Huan[1], Deepak Bandyopadhyay[1], Jan Prins[1],
Jack Snoeyink[1], Alexander Tropsha[2], Wei Wang[1]
[1]Computer Science Department
[2]The Laboratory for Molecular Modeling, School of Pharmacy
University of North Carolina at Chapel Hill

Structure motifs are amino acid packing patterns that occur frequently within a set of protein structures. We define a labeled graph representation of protein structure in which vertices correspond to amino acid residues and edges connect pairs of residues and are labeled by (1) the Euclidian distance between the $C\alpha$ atoms of the two residues and (2) a Boolean indicating whether the two residues are in physical/chemical contact. Using this representation, a structure motif corresponds to a labeled clique that occurs frequently among the graphs representing the protein structures. The pairwise distance constraints on each edge in a clique serve to limit the variation in geometry among different occurrences of a structure motif. We present an efficient constrained subgraph mining algorithm to discover structure motifs in this setting. Compared with contact graph representations, the number of spurious structure motifs is greatly reduced. Using this algorithm, structure motifs were located for several SCOP families including the Eukaryotic Serine Proteases, Nuclear Binding Domains, Papain-like Cysteine Proteases, and FAD/NAD-linked Reductases. For each family, we typically obtain a handful of motifs within seconds of processing time. The occurrences of these motifs throughout the PDB were strongly associated with the original SCOP family, as measured using a hyper-geometric distribution. The motifs were found to cover functionally important sites like the catalytic triad for Serine Proteases and co-factor binding sites for Nuclear Binding Domains. The fact that many motifs are highly family-specific can be used to classify new proteins or to provide functional annotation in Structural Genomics Projects.