EFFICIENT PROTEIN TERTIARY STRUCTURE RETRIEVALS AND CLASSIFICATIONS USING CONTENT BASED COMPARISON ALGORITHMS

Pin-Hao Chi

Dr. Chi-Ren Shyu, Dissertation Supervisor

ABSTRACT

Functionally important sites of proteins are potentially conserved to specific three-dimensional structural folds. To understand the structure-to-function relationship, life sciences researchers and biologists have a great need to retrieve similar structures from protein databases and classify these structures into the same protein fold. Traditional protein structure retrieval and classification methods are known to be either computationally expensive or labor intensive. In the past decade, more than 35000 protein structures have been identified. To meet the needs of fast retrieval and classifying high-throughput protein data, our research covers three main subjects: (1) Real-time global protein structure retrieval: We introduce an image-based approach that extracts signatures of three-dimensional protein structures. Our high-level protein signatures are then indexed by multi-dimensional indexing trees for fast retrieval. (2) Real-time global protein structure classification: An advanced knowledge discovery and data mining (KDD) model is proposed to convert high-level protein signature into itemsets for mining association rules. The advantage of this KDD approach is to effectively reveal the hidden knowledge from similar protein tertiary structures and quickly suggest possible SCOP domains for a newly-discovered protein. In addition, we develop a non-parametric classifier, E-Predict, that can rapidly assign known SCOP folds and recognize novel folds for newly-discovered proteins. (3) Efficient local protein structure retrieval and classification: We propose a novel algorithm, namely, the Index-based Protein Substructure Alignment (IPSA), that constructs a two-layer indexing tree to capture the obscured similarity of protein substructures in a timely fashion. Our research works exhibit significantly high efficiency with reasonably high accuracy and will benefit the study of high-throughput protein structure-function evolutionary relationships.