Chapter 21
Protein Homology Analysis for Function Prediction with Parallel Sub-Graph Isomorphism

Alper Küçükural
University of Kansas, USA & Sabanci University, Turkey

Andras Szilagyi
University of Kansas, USA

O. Uğur Sezerman
Sabanci University, Turkey

Yang Zhang
University of Kansas, USA

ABSTRACT

To annotate the biological function of a protein molecule, it is essential to have information on its 3D structure. Many successful methods for function prediction are based on determining structurally conserved regions because the functional residues are proved to be more conservative than others in protein evolution. Since the 3D conformation of a protein can be represented by a contact map graph, graph matching, algorithms are often employed to identify the conserved residues in weakly homologous protein pairs. However, the general graph matching algorithm is computationally expensive because graph similarity searching is essentially a NP-hard problem. Parallel implementations of the graph matching are often exploited to speed up the process. In this chapter,the authors review theoretical and computational approaches of graph theory and the recently developed graph matching algorithms for protein function prediction.

INTRODUCTION

Computational assignment of protein function from the *3D protein structure* is one of the important open problems in *structural proteomics*. Currently, many proteins deposited in the Protein Data Bank (PDB) have limited or no biological function annotation. Protein functions are usually derived from evolutionarily related proteins. Evolutionary association can be determined from sequence and structural similarities. The methods using sequence information are based on the detection of functional motifs (Huang and Brutlag, 2001; Hulo, et al., 2006; Stark and Russell, 2003), global sequence similarity search (Conesa, et al., 2005; Hawkins, et al., 2006; Martin, et al.,

DOI: 10.4018/978-1-4666-3604-0.ch021

2004), determination of similar loci (Hawkins, et al., 2006), and similarities in phylogeny (Engelhardt, et al., 2005; Storm and Sonnhammer, 2002). However, only around 30% of the protein pairs with less than 50% sequence identity have a similar function. Therefore, sequence similarity itself is not sufficient to develop a robust function prediction (Rost, 2002). In addition, several studies indicate that the inclusion of structural information increases the accuracy of predictions (Devos and Valencia, 2000; Thornton, et al., 2000; Wilson, et al., 2000), because structural features are usually more conserved than sequence.

Similarities between protein structures can be identified by structural alignment methods such as DALI (Holm, et al., 2008), CE (Shindyalov and Bourne, 1998), and TM-align (Zhang and Skolnick, 2005). Several function prediction methods employ structural alignment programs to identify the structurally closest proteins and transfer the functional annotation to the target protein. However, the correlation between function and overall protein fold is weak (Martin, et al., 1998). This can in part be explained by the fact that global structural alignment methods do not always capture locally conserved regions, and the biochemical function of a protein is usually determined by the local structure of a few active residues. Therefore, algorithms that aim to extract local structural information should achieve more robust function prediction (Laskowski, et al., 2005; Weinhold, et al., 2008).

The structures and sequences of remotely homologous protein pairs may have diverged during evolution while local structures involved in protein function may have been preserved. The aim of searching for local structural similarities is to detect these preserved, functionally important structural patterns. To discover local structural motifs, the following methods have been described in the literature. In a method based on 3D templates (Laskowski, et al., 2005), the specific 3D conformations of sets of 2-5 residues were extracted from the structures of functionally significant units. This template set was manually compiled to include four types of templates: the enzyme active site, ligand-binding residues, DNA-binding residues, and reverse templates. Given a target protein, the template set is searched for structures locally matching some part of the target protein, within spheres of a 10 Å radius. The matches are ranked using the SiteSeer scoring function. The degree of overlap between target and template residues is calculated, and the algorithm maximizes the sum of the overlap scores of the matched residues in all possible configurations. The method was tested on various distantly related protein pairs with widely divergent sequences. Significant functional matches were found, e.g. two TIM-barrel proteins with very low sequence identity were found to have a high SiteSeer score, and their functional sites were correctly matched. Moreover, some of the predictions for newly released structures with unknown function have later been experimentally verified. In another study, the combination of sequence and structural features were employed to identify functional similarities, based on the assumption that the preserved amino acids at key sites in similar local structures hint at a functional similarity as well (Friedberg, 2006).

Conserved local regions may contain residues that are not adjacent in sequence. Structurally adjacent residues, however, are preserved in most cases. These structurally conserved patterns have been explored by various tools such as JESS (Barker and Thornton, 2003), PINTS (Stark and Russell, 2003), PDBSiteScan (Ivanisenko, et al., 2004), and PAR-3D (Goyal, et al., 2007). Local structural similarities can be detected with search algorithms based on contact map networks. The algorithms can be described in terms of three major characteristics: representation, scoring, and searching. The contact maps are searched for similar regions with graph matching algorithms. However, the possible mutations, insertions and deletions in the protein structures yield very different contact maps. To discover similarities in the presence of such conformational differences,

Related Content