Chapter 45
Incorporating Correlations among Gene Ontology Terms into Predicting Protein Functions

Pingzhao Hu
York University, Canada & University of Toronto, Canada

Hui Jiang
York University, Canada

Andrew Emili
University of Toronto, Canada

ABSTRACT

One of the key issues in the post-genomic era is to assign functions to uncharacterized proteins. Since proteins seldom act alone, but rather interact with other biomolecular units to execute their functions, the functions of unknown proteins may be discovered through studying their associations with proteins having known functions.

In this chapter, the authors discuss possible approaches to exploit protein interaction networks for automated prediction of protein functions. The major focus is on discussing the utilities and limitations of current algorithms and computational techniques for accurate computational function prediction. The chapter highlights the challenges faced in this task and explores how similarity information among different gene ontology (GO) annotation terms can be taken into account to enhance function prediction.

The authors describe a new strategy that has better prediction performance than previous methods, which gives additional insights about the importance of the dependence between functional terms when inferring protein function.

DOI: 10.4018/978-1-4666-3604-0.ch045

INTRODUCTION

Currently the sequencing of many genomes has brought to light the discovery of thousands of putative open reading frames which are all potentially transcribed and translated into protein products. For many of these proteins, little is known beyond their primary sequences, and for the typical proteome, between one-third and one-half of all proteins remains functionally uncharacterized. For example, despite being the most highly studied model bacterium, a comprehensive community annotation effort indicated that only half (~54%) of the protein-coding gene products of *E. coli* currently have experimental evidence indicative of a biological role (Riley, 2006). The remaining genes have either only generic (homology-derived) functional attributes (e.g. 'predicted DNA-binding') or no discernable physiological role. Some of these functional 'orphans' may have eluded characterization because they lack obvious mutant phenotypes, are expressed at low or undetectable levels, or have no obvious homology to annotated proteins. Moreover, since proteins often perform different roles in alternate biological contexts, due to the complexity of biological systems, many functions of these alternate functions may not have yet been discovered. As a result, a major challenge in modern biology is to develop efficient methods for determining protein function at the genomic scale (Eisenberg, 2000; Brun, 2003; Barabasi, 2004; Chen 2006; Hu, 2009a).

Given the slow, laborious and expensive nature of experimenttation, computational procedures to systematically predict the functions of uncharacterized proteins from their molecular relationships are increasingly seen to be useful (Vazquez, 2003; Zhou, 2005; Zhao, 2007 and 2008; Hu, 2009a). The most handy and well-known computational method for function prediction is based on the detection of significant sequence similarity to gene products of known function, using such basic bioinformatic software tools as BLAST (Basic Local Alignment Search Tool) (Altschul, 1997). The assumption is that proteins that are similar in sequence likely have similar biological properties. A major caveat with this simplistic approach is that only those functions are obviously and directly tied to sequence, such as enzymatic activity, can be predicted accurately.

However, proteins seldom act alone, but rather interact with other biomolecular units to execute their biological functions. For example, physical interactions operate at almost every level of cellular functions (Chien, 1991; Jansen, 2003; Wodak, 2004). Thus, implications about function can often be made via the study of such molecular interactions. These inferences are based on the premise that the function(s) of unknown proteins may be gleaned from their interaction partners having a known function. In fact, it has been postulated that protein function and the higher-level organization of proteins into biological pathways can be reliably deduced by studying protein interaction networks generated via proteomic, genomic and bioinformatic approaches, providing insights into the molecular mechanisms underlying biological processes (Huynen, 2000; Gavin, 2002 and 2006; Jansen, 2003; Asthana, 2004; Altaf-Ul-Amin, 2006; Chua, 2007; Hu, 2009a). Systematic functional predictions based on computational integration of high-throughput interaction datasets have gained popularity among computational biologists for investigating gene action in model organisms such as yeast (Chen, 2004) and prokaryote such as *E. coli* (Hu, 2009a). For example, a recent integrative analysis of large-scale phenotypic, phylogenetic and physical interaction data in bacteria revealed an evolutionarily conserved set of novel motility-related proteins (Rajagopala, 2007).

In this chapter, we introduce some state-of-the-art computational procedures that allow for the automated prediction of protein functions based on the analysis of the patterns of functional associations of both known and unannotated proteins in the context of interaction networks. We discuss the potential and caveats of existing algorithms for accurate function prediction and describe new