Expert Systems with Applications 36 (2009) 12605–12612

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Exploiting knowledge ontology and software agents for PPI network analysis

Wen-Shyong Tzou [a], Wei-Po Lee [b,*], Bingchiang Jeng [b]

[a] Institute of Bioscience and Biotechnology, National Taiwan Ocean University, Keelung, Taiwan
[b] Department of Information Management, National Sun Yat-sen University, Kaohsiung, Taiwan

ARTICLE INFO

Keywords:
Protein–protein interaction
Motif analysis
Evolutionary conservation
Gene ontology
Mobile agent
Adaptive computing

ABSTRACT

One major goal of functional genomics has been to identify and analyze molecular interactions in a cellular context to better understand the underlying design principles and mechanisms. To investigate into a PPI network from both topological and functional points of view, this work proposes a methodology that exploits ontology-based biological knowledge for network analysis. To speed up the procedure, an agent-based framework is also presented for supporting distributed computing. The preliminary results show that through the knowledge obtained from gene ontology, our work in analyzing building blocks of PPI networks can give a higher resolution than that of previous ones. Also our agent-based framework can successfully speed up the task of network analysis in an adaptive manner.

1. Introduction

Recent advances in network biology indicate that cellular networks are organized and governed by generic principles and mechanisms (Alon, 2006; Barabasi & Oltvai, 2004; Kitano, 2001). Though the traditional biological research has provided a wealth of knowledge about individual cellular components and their functions, it is increasingly clear that a discrete biological function can rarely be attributed to an individual molecule. Instead, most biological characteristics arise from complex interactions between the numerous constituents of a cell. Therefore, it is crucial to understand the structure and the dynamics of the complex intercellular web of interactions that contribute to the structure and function of a living cell.

Protein–protein interaction (PPI) is a specific type of molecular interaction that plays a central role in relaying signals, building molecular machines, engaging in enzyme reactions, and making decisions among multiple biological processes. To uncover the structural design principles of molecular networks, research has been conducted to identify and extract network motifs (Huang, Sun, Cheng, & Hsieh, 2007; Milo et al., 2002; Wernicke, 2006) that are defined as the simple building blocks (subgraphs) of a complex network. It has been shown that certain network motifs occur at a significantly higher frequency than that are expected from a random network (Milo et al., 2002; Shen-Orr, Milo, Mangan, & Alon, 2002). The analysis of network motifs has led to interesting results in many research areas. For example, the motif studies in transcription network of *Escherichia coli* suggest that network motifs play key information processing roles in this type of networks. Also it has been shown that network motifs may be conserved over evolutionary time. The tendency to conserve evolutionarily the protein components of topologically distinct motifs could be indicative of their importance and involvement in specific biological functions. That is, the conservation of these motifs points to their functional regulatory role (Eom, Lee, & Jeong, 2006; Milo et al., 2002; Wuchty, Oltvai, & Barabasi, 2003).

In the study of gene regulation, the transcriptional regulator is considered the master and the regulated gene the slave (Milo et al., 2002; Wuchty et al., 2003). A directed graph is commonly used to represent the regulatory network, in which the nodes are genes and the arcs are regulation relationships. Though the directed graph is appropriate to describe a gene regulatory network, such a dichotomous (master–slave) representation cannot be used for a PPI since proteins have multiple functions. Therefore, an indirected graph is adopted for PPI network. To investigate a PPI network from both topological and functional perspectives, in this work we propose an approach to exploit ontology-based biological knowledge for network analysis. The protein nodes are firstly labeled on the basis of their functional attributes of gene ontology (GO) (Baclawski & Niu, 2006; GO database), and then the recurring patterns of the functional attributes of protein interactions are pursued. With the newly proposed representation of a PPI network, we can categorize the repertoire of network motifs of the same topology into "motif modes", in which each motif mode features a special topological combination of molecular functions (see Section 2.1). This leads us to the finding that when the functional attributes of the motifs are taken into account, the evolutionary constraints on the motifs of the same topology are certainly not the same. On the contrary, the constraints vary a great deal. Through the knowledge obtained from GO, our work in analyzing building blocks of PPI networks can give a higher resolution and

* Corresponding author.
E-mail address: wplee@mail.nsysu.edu.tw (W.-P. Lee).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.05.040

new insight than others reported before in which only topological information is considered.

As the proposed approach now involves using functional information of PPI recorded in the gene ontology, and counting the appearing frequency of each possible network building block, its computational cost is therefore much higher than other work on network motif analysis, such as (Huang et al., 2007; Milo et al., 2002; Wernicke, 2006). To reduce the time needed, in this work we also present an agent-based framework in which multiple agents interoperate on the same platform to support distributed computing. In this platform, mobile agents play the most important roles that are able to dynamically migrate from machine to machine to discover currently available resources to share the computation load. To verify the proposed approach we implement a prototype system on a middleware platform JADE (Java Agent Development Framework) (Bellifemine, Caire, & Greenwood, 2007) in which agents can control their own thread of execution so that they can easily be programmed to initiate the execution of actions autonomously. Experiments are conducted to evaluate their corresponding performance, and the preliminary results show the promise and efficiency of our agent-based approach.

2. PPI network analysis

As mentioned above, the task of network analysis with functional annotation needs considerably large amount of time, a distributed computing environment is thus desired to speed up the computation. Though using network-based machines to realize parallelism is much cheaper than using a parallel computer, some machines have to be pre-specified to contribute their computation power. This is not practical for a computing laboratory environment in which the computational resources are shared by many end-users. Alternatively, we develop an agent-based framework to manage the execution and communication for tasks with high computational load in an adaptive manner. Our system mainly includes three types of agents: information agents for retrieving domain knowledge from databases, mobile agents for carrying and executing code, and planning agents for dispatching mobile agents to achieve subtasks. In the following, we firstly present our methodology for PPI network analysis and then concentrate on describing how we use a mobile agent-based approach to speed up the execution of the target task.

2.1. Using biological knowledge for motif extraction and analysis

Instead of calculating network motif by only considering network topology as other previous research, our work integrates domain knowledge into the PPI network so that motifs can be further analyzed from the perspective of molecular function. The biological knowledge can be obtained from different scientific literatures or public databases. Here we use the most popular and prominent biological knowledge source, gene ontology, for network analysis. GO aims at delivering a shared, consistent, structured, and controlled vocabulary that can be applied to any organism. It is specifically intended for annotating gene products with these vocabulary terms, based on their functions in the cell. The GO source contains three independent ontologies: molecular function, biological process, and cellular component. Each of which serves as an organizing principle for describing gene products, and is a hierarchical classification scheme structured as a direct acyclic graph (DAG). In the ontology, each GO node designing a biological class in the GO graph, and each edge represents the relationship. Use of GO in analysis of experimental data enables integration of biological background data in a controlled manner.

The building blocks of the GO are the terms, each of which has a unique numerical identifier (e.g., GO:00004867) and a term name (e.g., signal transduction). Since our focus is on the recurring patterns of the functional attributes of protein interactions, we choose to use the molecular function ontology (hereafter GO terms) to annotate each node of a PPI network. Fig. 1 shows an example of annotation of GO terms. GO terms are organized in DAGs as mentioned above, and each term can be traced to different depths in the hierarchies. In this work, the molecular function of the GO terms for each protein in our dataset is traced from its original position in the tree hierarchy upwards to the nodes at depths five and six. If there are several paths towards the chosen GO level, the shortest path is picked. Fig. 2 illustrates the above procedure. Also if the GO term for a protein is located at a depth lower than or equal to five or six (i.e., depth $\leqslant 5$ or 6) in the GO tree, the GO term remains unchanged. The reason for choosing depths five and six is that, based on the statistics, the average depth of the GO terms in

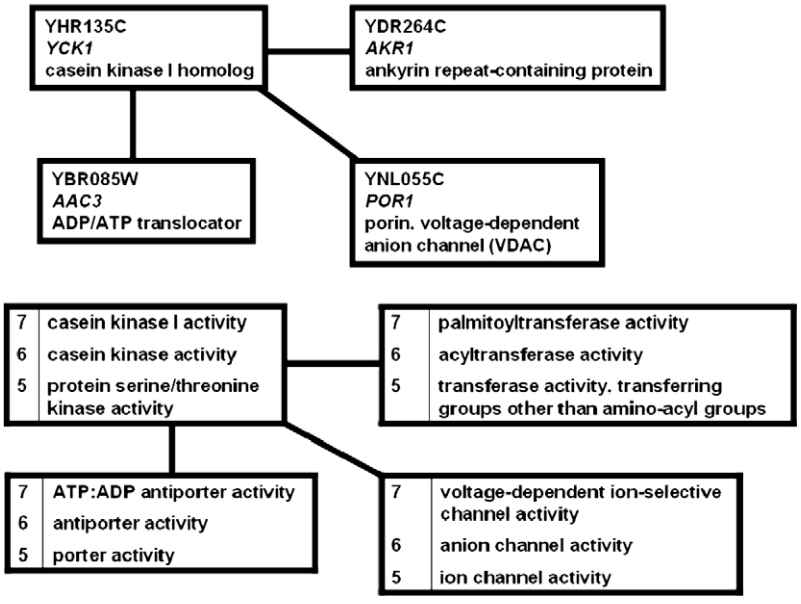

Fig. 1. An example of annotation of GO terms for a network motif (GO terms for each node are provided at three GO levels).