Future Generation Computer Systems 26 (2010) 434–440

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

IMPRECO: Distributed prediction of protein complexes

Mario Cannataro, Pietro H. Guzzi *, Pierangelo Veltri

Bioinformatics Laboratory, Department of Experimental Medicine and Clinic, University Magna Graecia, Viale Europa, 88100 Catanzaro, Italy

ARTICLE INFO

Article history:
Received 22 December 2008
Received in revised form
18 June 2009
Accepted 3 August 2009
Available online 7 August 2009

Keywords:
Protein-to-protein interactions
Protein complexes prediction
Graph-clustering
Meta-predictor

ABSTRACT

Proteins interact among themselves, and different interactions form a very huge number of possible combinations representable as protein-to-protein interaction (PPI) networks that are mapped into graph structures. Protein complexes are a subset of mutually interacting proteins. Starting from a PPI network, protein complexes may be extracted by using computational methods. The paper proposes a new complexes meta-predictor which is capable of predicting protein complexes by integrating the results of different predictors. It is based on a distributed architecture that wraps predictor as web/grid services that is built on top of the grid infrastructure. The proposed meta-predictor first invokes different available predictors wrapped as services in a parallel way, then integrates their results using graph analysis, and finally evaluates the predicted results by comparing them against external databases storing experimentally determined protein complexes.

1. Introduction

Proteins are elementary building blocks of biological processes occurring within cells. They play their role via mutual interaction, composing a very broad network of interactions known as *interactome* [1]. Biological research has therefore focused on the determination of the complete set of Protein-to-Protein Interactions (PPI) that occur in various organisms. From this work, different experimental assays have accumulated a large quantity of data about protein interactions in cells [2,3]. There exist many different typologies of interaction among proteins considering the biochemical nature of the interactions. The common interaction involves the direct contact of molecules, but proteins may also interact through a medium or even through the exchange of ions. The set of all binary interactions is spread across different repositories, such as BIND [4], DIP [5], and MIPS [6]. These databases usually contain interaction information determined in wet labs via one or more experimental technologies.

The high number of protein interactions taking place in a cell makes the manual analysis unfeasible, e.g. the individuation of global or local properties even for a simple organism such as yeast. The need for the introduction of computer-based tools for PPI data modeling, management, and analysis therefore arises.

The basic protein interaction (**binary interaction**) involves only two proteins and can be modeled by the couple of the involved proteins and by an information describing the kind and if necessary the direction of the interaction. It should be noted that usually binary interactions are not named, i.e. there is not a naming standard for interactions yet. A **PPI network**, being the set of all the protein interactions in an organism, is commonly represented as a (if applicable directed) graph [7,8], where nodes represent proteins and edges represent the interactions among them.

Nonetheless such model does not capture the differences among interactions. Edges, in fact, are usually not labelled, so the kind of interaction is usually an unknown parameter in the analysis phase.

The modeling of PPI networks as graphs has enabled the investigation of biological properties of an organism through the use of graph-based algorithms [9] that aim to discover biologically meaningful facts by exploring structural properties of the underlying graph.

Initial attempts tried to discover the global properties of such networks and the individuation of theoretical models to explain these. In addition to the analysis of global properties, the study of recurring local topological features, such as the overrepresented subgraphs, has found an increasing interest. Finally, a recent trend in protein interaction analysis aims to the comparative investigation of PPI networks, discovering conserved subgraphs among them.

For instance, small dense regions in a PPI network, i.e. regions with a number of interactions higher than the average of the networks, may represent a *protein complex*, that is a group of two or more associated proteins that interact to achieve a common biological goal. Proteins bound in a complex act as a single functional unit via non-covalent interactions. Each complex has a different lifetime, i.e., the time over which it remains stable. Moreover, the formation of protein complexes acts as an activator or an inhibitor of one or more of the members of the complex.

* Corresponding author.
E-mail address: hguzzi@unicz.it (P.H. Guzzi).

0167-739X/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2009.08.001

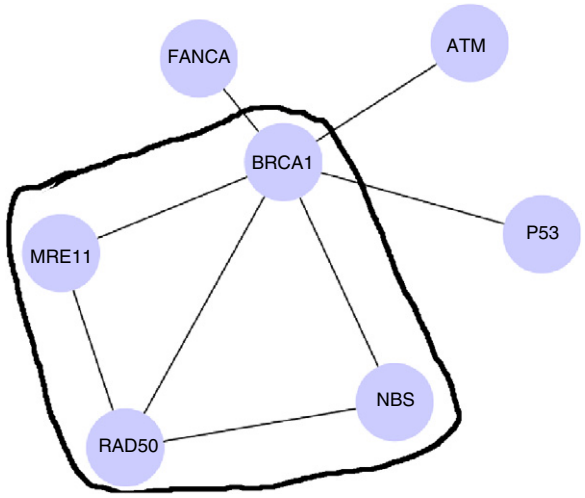

Fig. 1. An example of a protein complex comprising BRCA1.

Complexes are a fundamental building block of many biological processes, so the analysis of their conservation during evolution or the eventual correlation among complexes and various diseases are important research areas [10]. For example, Breast Cancer Protein 1 (BRCA1) is known to participate in multiple cellular processes involving multiple protein complexes that play an important role in the mechanisms for DNA repair [11]. Fig. 1 depicts a fragment of human PPI network evidencing a complex in which BRCA1 participates. The complex comprises the proteins BRCA1, RAD50, Mre11 and NBS.

Protein complexes can be determined in wet labs using various techniques such as Mass-Spectrometry (MS) [3] or yeast-two-hybrid (Y2H) [12], but these experiments are usually time-consuming.

The main idea underlying the application of MS is the use of a protein as a *bait* to capture all the possible interacting partners. This protein is initially inserted into a sample, then it is purified from other proteins through a series of subsequent cleavages that aim to separate the investigated protein from other proteins that are not interacting. Finally all the proteins that are bound to the *bait*, are analyzed through the mass spectrometer. Data generated from the spectrometer are mined and the interacting proteins are identified. The *yeast-two-hybrid* technique aims to verify the existence of an interaction among two selected proteins. This assay uses a protein as a bait to identify the interaction with another protein called *prey*. In summary MS is able to directly identify protein complexes, while Y2H is able only to check the existence of binary interactions, so it requires many experiments using the same bait to find the complex. A protein complex prediction algorithm (*complex predictor*) tries to find highly connected regions in a PPI network that may reveal a protein complex. Protein complexes can be extracted from PPI networks by searching for small dense regions, i.e., regions containing many interactions compared with the average degree of PPI network, i.e. a higher ratio of edges with respect to the number of nodes. After the early work of David A. Bader [13], a number of algorithms for the prediction of protein complexes [14,15,13,16,17] have been introduced.

A protein complex predictor can be evaluated taking into account the percentage of discovered subnetworks that correspond to real complexes, against the meaningless ones. Currently, there is not a common accepted benchmark and there is no *gold standard*. To estimate the quality of prediction, a set of databases of experimentally verified complexes can be used as a benchmark. Currently, only a few of such databases exist, including the MIPS catalog of protein complexes in yeast [18], and the CORUM Complexes Database [19]. These databases store experimentally verified complexes, i.e., complexes that have been determined or verified by using experimental assays. The performance of a prediction algorithm is therefore influenced by: (i) the kind and the initial configuration of the used algorithm, and (ii) the validity of the initial protein protein interactions (i.e., edges in the graph) of the input interaction network (i.e., the graph) [20].

We have developed a tool (IMPRECO, for IMproving PREdiction of COmplexes) that combines different predictor results using an integration algorithm which is able to gather (partial) results from different predictors and eventually produce novel predictions.

In this paper, we present a distributed architecture that implements the IMPRECO prediction algorithm and demonstrates its ability to predict protein complexes. The proposed meta-predictor first invokes different available predictors wrapped as services in a parallel way, then integrates their results using graph analysis, and finally evaluates the predicted results by comparing them against external databases storing experimentally determined protein complexes.

The remainder of the paper is organized as follows. Section 2 introduces protein complex prediction. Section 3 discusses related work. Section 4 presents the IMPRECO algorithm. Section 5 discusses the distributed architecture of IMPRECO. Section 6 presents a case study and evaluates performance of the resulting predictions with respect to those of basic predictors. Finally, Section 7 concludes the paper and outlines future work.

2. Protein complex prediction

The prediction of protein complexes from experimental data is still a challenge. This paper focuses on algorithmic approaches that rely on two main ideas: (i) modeling the whole set of interactions as a graph, and (ii) the use of clustering for finding complexes. The workflow for complex prediction comprises three main steps: (i) building a PPI network from binary interaction data, (ii) algorithmic analysis of the network, and (iii) result evaluation, as depicted in Fig. 2. After an algorithm has shown its ability to correctly predict known complexes, it can be used as a predictor for other complexes.

First step. The prediction of protein complexes starts with the collection of protein-to-protein interaction data stored in various databases. The interaction databases can be categorized based on: (i) the type of stored interactions, for example, databases of verified interactions such as BIND [4], or (ii) databases that store predicted interactions such as OPHID [21]. Databases in the first category store interactions that have been determined via in vitro experiments, whereas those in the second category store interactions predicted via computational methods.

Second step. After that data have been collected and modeled in a graph, researchers can apply different available algorithms that predict complexes from graphs, such as MCODE [13], MCL [22], or RNSC [15]. These algorithms usually try to identify highly connected regions in a graph, where each region is defined as a set of nodes of the graph whose local density is greater than the average density of the graph. Density is defined as the ratio: $\frac{2E}{V*V-1}$, where E is the number of the existing edges, i.e. *the degree*, and V is the number of vertices. During algorithm execution, no biological knowledge is used to guide cluster identification or cluster selection.

Third step. Predicted complexes can be compared with those stored in a reference database, such as a set of catalogs of verified complexes, which enables evaluation of the predictor. A recent work [20] has compared many algorithms for protein complex prediction. The authors evaluated the algorithms in terms of clustering parameters such as cluster separation, for existing datasets, and thereby found optimal parameters for each algorithm.