Artificial Intelligence in Medicine 71 (2016) 62–69

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Predicting overlapping protein complexes from weighted protein interaction graphs by gradually expanding dense neighborhoods

Christos Dimitrakopoulos[a,*], Konstantinos Theofilatos[a], Andreas Pegkas[b], Spiros Likothanassis[a,b], Seferina Mavroudi[a,b,c]

[a] InSyBio Ltd, 109 Uxbridge Road, W5 5TL London, UK
[b] Department of Computer Engineering and Informatics, University of Patras, Building B University Campus, Rio, 26500, Greece
[c] Department of Social Work, School of Sciences of Health and Care, Technological Educational Institute of Western Greece, Megalou Aleksandrou 1, Koukouli, 26334 Patra, Greece

ARTICLE INFO

Article history:
Received 17 October 2015
Received in revised form 30 May 2016
Accepted 30 May 2016

Keywords:
Computational prediction of protein complexes
Clustering of protein–protein interaction networks
Functionally homogeneous protein clusters
Parkinson differentially expressed network modules

ABSTRACT

Objective: Proteins are vital biological molecules driving many fundamental cellular processes. They rarely act alone, but form interacting groups called protein complexes. The study of protein complexes is a key goal in systems biology. Recently, large protein–protein interaction (PPI) datasets have been published and a plethora of computational methods that provide new ideas for the prediction of protein complexes have been implemented. However, most of the methods suffer from two major limitations: First, they do not account for proteins participating in multiple functions and second, they are unable to handle weighted PPI graphs. Moreover, the problem remains open as existing algorithms and tools are insufficient in terms of predictive metrics.

Method: In the present paper, we propose gradually expanding neighborhoods with adjustment (GENA), a new algorithm that gradually expands neighborhoods in a graph starting from highly informative "seed" nodes. GENA considers proteins as multifunctional molecules allowing them to participate in more than one protein complex. In addition, GENA accepts weighted PPI graphs by using a weighted evaluation function for each cluster.

Results: In experiments with datasets from *Saccharomyces cerevisiae* and human, GENA outperformed Markov clustering, restricted neighborhood search and clustering with overlapping neighborhood expansion, three state-of-the-art methods for computationally predicting protein complexes. Seven PPI networks and seven evaluation datasets were used in total. GENA outperformed existing methods in 16 out of 18 experiments achieving an average improvement of 5.5% when the maximum matching ratio metric was used. Our method was able to discover functionally homogeneous protein clusters and uncover important network modules in a Parkinson expression dataset. When used on the human networks, around 47% of the detected clusters were enriched in gene ontology (GO) terms with depth higher than five in the GO hierarchy.

Conclusions: In the present manuscript, we introduce a new method for the computational prediction of protein complexes by making the realistic assumption that proteins participate in multiple protein complexes and cellular functions. Our method can detect accurate and functionally homogeneous clusters.

1. Introduction

Proteins are considered the most important players in molecular interactions. They play a significant role in all cellular functions (e.g. transmission of regulatory signals in the cell) and they catalyze a huge number of chemical reactions. They rarely act isolated, but they are combined in functional modules with one example of them being the protein complexes. The prediction of protein complexes is crucial for understanding the cellular mechanisms and for predicting the functions of uncharacterized proteins. The experimental prediction of protein complexes is mainly limited to tandem affinity purification (TAP) [1], which provides erroneous data and is not cost-effective and time-efficient. TAP results have raised the human

* Corresponding author.
E-mail addresses: c.dimitrakopoulos@insybio.com (C. Dimitrakopoulos), k.theofilatos@insybio.com (K. Theofilatos), pegkas@ceid.upatras.gr (A. Pegkas), likothan@ceid.upatras.gr (S. Likothanassis), mavroudi@ceid.upatras.gr (S. Mavroudi).

http://dx.doi.org/10.1016/j.artmed.2016.05.006
0933-3657/

interactome's coverage, but also include many false positives and false negatives [2].

In virtue of the experimental approaches' limitations, researchers have recently emphasized in the computational prediction of protein complexes from protein–protein interaction (PPI) data by using unsupervised clustering algorithms [3,4]. The assumption behind most of the methods is the detection of strongly connected components of proteins that are sparsely connected to the rest of the graph [5,6]. These algorithms are based on very different approaches. Most of them require the specification of a considerable number of parameters, some of which drastically affect the results.

Clustering as a modelling approach can address the problem of detecting protein complexes in PPI graphs, however, standard clustering is not ideal for PPI networks. Proteins may have multiple functions, and therefore their corresponding graph nodes may belong to more than one cluster. For instance, 17 pairs of complexes overlap in the Aloy dataset [7], 40 pairs in the BT_409 dataset [8] and 215 pairs in the Pu dataset [9]. Such nodes present a challenge to traditional PPI clustering algorithms and recently, algorithms that detect overlapping clusters have been proposed [10,11]. Moreover, the state of the art methods for clustering PPI graphs are usually applied to weighted PPI graphs only after 'binarizing' them by removing weighted edges below a given threshold. The idea of using the original weighted PPI graphs was introduced recently [12] and demonstrated a significant improvement in the detection of protein complexes.

In the present paper, we propose gradually expanding neighborhoods with adjustment (GENA), a fully unsupervised clustering algorithm, which consists of two steps. In the first step, a greedy approach is used to initialize the clusters. We used initial "seed" vertices with a high potential for cluster formation based on the clustering coefficient metric. The seed nodes are absorbing neighboring nodes (and gradually forming a growing cluster) based on an evaluation function, which is defined as a generalized version of the connectivity in the weak sense (Section 2.2). Each cluster grows independently from the other clusters and as a result they can overlap. The clusters stop growing when no neighboring node can improve their evaluation function. In the second step of "adjustment", random moves are performed between the clusters to optimize the clustering solution of the initialization step.

In experiments using public datasets of protein complexes from *Saccharomyces cerevisiae* and human, GENA outperformed restricted neighborhood search (RNSC), Markov clustering (MCL) and clustering with overlapping neighborhood expansion (ClusterONE), which are the state of the art algorithms for predicting protein complexes. We performed an extensive analysis by using several input networks and evaluation datasets. In specific, we used three weighted PPI graphs and four evaluation datasets for *Saccharomyces cerevisiae* as well as two weighted PPI graphs and three evaluation datasets for human (Section 2.1).

2. Materials and methods

2.1. Datasets

2.1.1. Protein-protein interactions datasets

In the present paper, six PPI datasets have been used as inputs for the prediction algorithms. They originate from different organisms (*Saccharomyces cerevisiae* and human). Based on the PPI datasets, we created weighted PPI graphs. For *Saccharomyces cerevisiae*, we used Krogan et al. core and extended datasets [13], Collins et al. [14] dataset as described in [12]. For human, we built a network by using the interactions reported in the human protein reference database (HPRD).

Krogan et al. [13] have combined results from matrix-assisted laser desorption/ionization – time of flight (MALDI-TOF) mass spectrometry and liquid chromatography coupled with tandem mass spectrometry (LC–MS/MS) experiments to identify protein–protein interactions. The reason for using data from two independent experimental settings was based on the observations that a single mass spectrometry method often fails to identify all proteins. Hence, using data from two independent methods was expected to increase the coverage and confidence of the obtained interactome. The results of the two methods were combined by using a supervised machine learning approach using hand-curated protein complexes from the Munich information center for protein sequences (MIPS) reference database [15] as a gold standard dataset. A two round learning phase framework was encountered coupling the output of Bayesian networks and decision trees with the stacked generalization algorithm [16]. In the first round, Bayesian inference networks and 28 different kinds of decision trees were tested finally settling on three methods: Bayesian networks and C4.5-based and boosted decision stumps. The output of these three methods was used as the input for a second round of learning with the stacked generalization algorithm. The output of the stacked generalization algorithm (i.e. a probability value between 0 and 1) was then thresholded at two different levels to obtain the core and extended datasets. The Krogan core dataset included all interactions with posterior probability higher than 0.273, while the extended dataset included all interactions with posterior probability higher than 0.101 [12].

Collins et al. [14] have combined the experimentally derived PPI networks of Krogan et al. [13] and Gavin et al. [17] by re-analyzing the raw primary affinity purification data of these experiments using a novel scoring technique called purification enrichment (PE). The PE scores are motivated by the probabilistic socio-affinity scoring framework of Gavin et al. [17], but also take into account negative evidence (i.e. pairs of proteins where one of them fails to appear as a prey when the other one is used as a bait).

The first PPI dataset for the human organism consists of the protein interactions from the HPRD database [18]. These protein interactions were filtered using the evolutionary Kalman mathematical modelling (EVOKALMAMODEL) method proposed in the human interactome knowledge base (HINT-KB) [19]. EVOKALMAMODEL predicts protein–protein interactions (PPIs) by fusing sequential, functional and structural PPI data. The extracted PPI graph consists of 7450 proteins and 21.475 interactions. The main idea of EVOKALMAMODEL is to construct an optimal mathematical predictor equation by exploring a pool of given mathematical terms. It combines Kalman filtering, an adaptive filtering technique with a genetic algorithm, a heuristic method based on the process of natural selection. The genetic algorithm detects the optimal subset of terms for the predictor's mathematical equation and then applies extended Kalman filters to compute its optimal parameters. The final equation is used to score and filter the protein interactions.

The second human PPI network is the entire HINT-KB network, which is constructed based on protein interactions included in the IrefIndex database and predicted as positive by the EVOKALMAMODEL method [19]. It contains 20845 unique proteins and 211367 unique interactions and was selected because it provides the highest coverage of the human interactome, while at the same time it is comprised of only confidently predicted interactions. Consequently, it enables the prediction of a high number of high quality, not previously reported protein complexes.

2.1.2. Evaluation datasets

For *Saccharomyces cerevisiae*, four well-studied datasets of protein complexes were used. The first is the one proposed and described in [8], which is named BT_409 and consists of 409 pro-