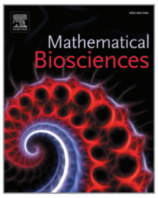
Mathematical Biosciences 274 (2016) 25–32

Contents lists available at ScienceDirect

Mathematical Biosciences

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



Protein function prediction based on data fusion and functional interrelationship

Jun Meng[a], Jael-Sanyanda Wekesa[a], Guan-Li Shi[a], Yu-Shi Luan[b,*]

[a] *School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning 116023, China*
[b] *School of Life Sciences and Biotechnology, Dalian University of Technology, Dalian 116023, China*

ARTICLE INFO

Article history:
Received 22 September 2015
Revised 8 January 2016
Accepted 1 February 2016
Available online 9 February 2016

Keywords:
Bi-relational graph
Transductive learning
Multi-label classification
PPI
Classifier integration

ABSTRACT

One of the challenging tasks of bioinformatics is to predict more accurate and confident protein functions from genomics and proteomics datasets. Computational approaches use a variety of high throughput experimental data, such as protein-protein interaction (PPI), protein sequences and phylogenetic profiles, to predict protein functions. This paper presents a method that uses transductive multi-label learning algorithm by integrating multiple data sources for classification. Multiple proteomics datasets are integrated to make inferences about functions of unknown proteins and use a directed bi-relational graph to assign labels to unannotated proteins. Our method, bi-relational graph based transductive multi-label function annotation (Bi-TMF) uses functional correlation and topological PPI network properties on both the training and testing datasets to predict protein functions through data fusion of the individual kernel result. The main purpose of our proposed method is to enhance the performance of classifier integration for protein function prediction algorithms. Experimental results demonstrate the effectiveness and efficiency of Bi-TMF on multi-sources datasets in *yeast, human* and *mouse* benchmarks. Bi-TMF outperforms other recently proposed methods.

1. Introduction

Protein function prediction is an essential task for disease analysis and drug development. High throughput experimental data has been used for protein function prediction. Many studies have been conducted on function annotation of proteins, since proteins play a major role in biological processes. There are two main prediction strategies widely used: sequence-similarity based and structure-similarity based function prediction [1]. Homology based methods such as BLAST and PSI-BLAST [2] are based on sequence similarity, thus proteins derived evolutionary from common ancestral sequence have similar functions. Further, homology based methods are classified as orthology based (genes derived through speciation from single ancestral sequence) or analogy based.

Function inference from protein–protein interaction (PPI) network constructed by pairs of interacting proteins whereby classification depends on the proteins structural similarity to predict protein functions. There are several high throughput experimental data sources for protein functions such as gene expression data, phylogenetic profile, PPI and protein domain structures. Traditional approaches use a single data source such as amino acid sequence of protein and predict one function at a time. These methods use single biological dataset hence the predicted functions are of low confidence and are less accurate [3].

A number of studies have integrated data from multiple sources [4–7], by combining interactions from independent data sources into one. Also, integration can be done via a joint probabilistic model or kernels combined using their optimal relative weights [8]. Integration of multiple networks from multiple data sources enhances predictive performance since each network is composed of different information of different quality [9]. Marcotte et al. [7] proposed a method of protein prediction in *yeast Saccharomyces cerevisiae* from different data sources based on heuristic combination. Lee et al. [6] introduced a method that integrates different classes of data and perform functional coupling.

Approaches that integrate heterogeneous data use kernel or network based classification algorithms to generate functions of unannotated proteins [10]. Methods that have used protein networks [5] include: majority vote [11], graph based [12], Bayesian [4], discriminative learning [13] and probability integration by log-likelihood scores [6]. Integration of the networks can be classified as: vector space where proteins are classified into vectors, classifier integration where each data source is trained on a classifier such

[*] Corresponding author. Tel.: +86 411 84706356; fax: +86 411 84706365.
E-mail addresses: mengjun@dlut.edu.cn (J. Meng), js_wekesa@yahoo.com (J.-S. Wekesa), 1289531794@qq.com (G.-L. Shi), luanyush@dlut.edu.cn (Y.-S. Luan).

http://dx.doi.org/10.1016/j.mbs.2016.02.001
0025-5564/© 2016 Elsevier Inc. All rights reserved.

as KNN and combine their predictions into one and kernel integration where each source is combined into a new network [14].

Since most methods that integrate data focus on the creation of function specific composite network, assignment of network weights is a complex task due to the existence of a vast number of function categories. Also, it is time consuming to combine the networks with a big number of nodes. For this reason, we ensure scalability and speed by performing prediction of protein functions at network level and combining the results rather than creating a composite network. Many algorithms have been proposed for identification of protein complexes and prediction of protein functions using PPI data. Chen et al. [15] presented the use of cliques in a PPI network to generate clusters with higher accuracy. Protein complexes are identified from the cliques which improve the performance of the entropy-based algorithm.

Transductive learning, a semi-supervised learning approach proposed by Vapnik [16] has been studied widely by researchers. Some studies [17–19] have applied this method in the form of graph-based techniques such as label propagation and adaptive graphs. In this paper we focus on transductive multi-label learning because the structure of the test dataset is exploited and proteins are annotated with multiple functions simultaneously.

Recently, the development of approaches that analyze data in the form of graphs for prediction of protein functions on a network in computational biology has been on the rise [8]. The graph consists of vertices representing instances and edges denote as similarities between the instances. Graph-based semi-supervised learning algorithms utilize a limited amount of labeled data to explore information on a large volume of unlabeled data [17,20]. Label propagation a state-of-the-art approach is a typical example of graph based learning algorithm that assumes connected nodes have similar functions [9].

Other machine learning algorithms most commonly used for classification and function prediction are support vector machines (SVMs) [21,22], artificial neural networks [23], and Naïve Bayesian classifiers [24,25].

In our method we studied how to explore network propagation on multiple networks from different data sources to predict functions of proteins. We also consider functional interrelationships within the function class network to improve performance of function prediction. We use transductive multi-label classification (TRAM) [26] algorithm to predict protein functions on individual networks and implement data fusion method to get the final annotations for the unlabeled proteins. By using data fusion to combine data from different sources we exploit interdependencies in order to increase accuracy of the prediction of functions. The fusion process is represented by a cost equation which is a quadratic functional representation of smoothness and consistency condition.

The rest of the paper is organized as follows. The methods including framework of Bi-TMF is briefly introduced in Section 2. In Section 3, the results of the experiments are analyzed and compared with other existing methods. Section 4 concludes this paper. The related datasets and codes of our method Bi-TMF are available in supporting website https://github.com/ML-DM-DUT/Bi-TMF.

2. Methods

In this section, we introduce transductive multi-label learning on protein-function bi-relational graph approach using data fusion to predict protein functions. We describe the methodology used in the following sub-sections.

2.1. Framework of prediction model

Firstly, functional similarities are formulated explicitly, and then correlation between the protein interaction network and the functions is modeled to form the bi-relational graph. The protein functions from each bi-relational graph are integrated to get final function annotations for proteins. Fig. 1 shows the prediction process of our proposed method Bi-TMF.

The detailed steps in the framework are described as follows.

(1) Creation of similarity matrix
Functional similarities are formulated in form of interfunctional similarities and protein interaction similarities. For a set of proteins $P = p_i$ $(i = 1, 2, \ldots, n)$ and K functional categories $C = c_i$ $(i = 1, 2, \ldots, K)$. Suppose each of the first l proteins has a set of labels $y_i \subseteq F$ represented by a binary vector $y_i \in \{0, 1\}^K$ where $y_{ik} = 1$ if p_i belongs to kth functional class and 0 otherwise. We generate a matrix W_{pp} for proteins, W_{FF} for the similarities between functions and W_{PF} for similarity matrix between proteins and functions. Since we use a directed bi-relational graph with directed edges as shown in Fig. 2, information propagation is from function to protein nodes in the intra-subgraphs (W_{pp} and W_{FF}) and inter-subgraph (W_{PF}).

(2) Protein function prediction
The main goal of our method is to predict $\{Y_i\}_{i=l+1}^{N}$ protein functions. To achieve this, we use TRAM algorithm on the protein-function bi-relational graph generated from integration of GO functional classification scheme. Protein function prediction is done on each of the N independent datasets. We determine the prediction likelihood score using a vector $\{F(j)\}_{j=l+1}^{N}$ for the N datasets.

(3) Data fusion
The vector scores of the different independent data sources are integrated to obtain a comprehensive score used to obtain final annotations. Unlike most methods that integrate heterogeneous data by combining the networks to create a composite network, in our method we integrate the results of each network to get a final result.

2.2. Functional correlation from heterogeneous data

Proteins from the same biological process have direct and indirect links with their neighbors within the network. Interacting proteins belong to common functional class since the interacting pairs share functions. The user defines the training and testing proteins from which confidence scores are used to determine reliability of predicted functions based on the relationship between proteins and their functions. Correlation coefficients on each dataset is converted into scores $F(j)$ which are combined across the network's datasets. The pair-wise relationship between proteins and their functions is modeled. Global optimization [12] is used to determine optimal protein function annotations by implementing data fusion of annotations from each dataset. This result is the precise and accurate annotations of proteins in the network. The error of false positives and negatives is inevitable in the network topology. We minimize the magnitude of this error by implementing global optimization of the results to get more refined and optimal annotations.

We use functional correlation matrices to denote the relationship between proteins and functions. Incorporation of semantic similarity into the functional interrelationship for neighboring proteins is also a feature that contributes toward accuracy in our method. The functional interrelationships contribute viable biological knowledge useful in function annotation of unannotated proteins.

2.3. Construction of bi-relational graph

Our method is graph-based prediction method and considers both the interrelationship between functional classes and the PPI