Biochemical and Biophysical Research Communications 440 (2013) 388–392

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

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

Prediction of regulation relationship between protein interactions in signaling networks

Wei Liu *, Hongwei Xie

The College of Mechanical & Electronic Engineering and Automatization, National University of Defense Technology, 410073 Changsha, China

ARTICLE INFO

Article history:
Received 3 September 2013
Available online 1 October 2013

Keywords:
Regulation relationship
Protein interaction
Signaling network
Logistic regression

ABSTRACT

The discovery of regulation relationship of protein interactions is crucial for the mechanism research in signaling network. Bioinformatics methods can be used to accelerate the discovery of regulation relationship between protein interactions, to distinguish the activation relations from inhibition relations. In this paper, we describe a novel method to predict the regulation relations of protein interactions in the signaling network. We detected 4,417 domain pairs that were significantly enriched in the activation or inhibition dataset. Three machine learning methods, logistic regression, support vector machines(SVMs), and naïve bayes, were explored in the classifier models. The prediction power of three different models was evaluated by 5-fold cross-validation and the independent test dataset. The area under the receiver operating characteristic curve for logistic regression, SVM, and naïve bayes models was 0.946, 0.905 and 0.809, respectively. Finally, the logistic regression classifier was applied to the human proteome-wide interaction dataset, and 2,591 interactions were predicted with their regulation relations, with 2,048 in activation and 543 in inhibition. This model based on domains can be used to identify the regulation relations between protein interactions and furthermore reconstruct signaling pathways.

1. Introduction

With the development of high-throughput technologies, large-scale protein–protein interaction (PPI) data for multiple species has been produced, which provided the basis for the investigation of protein function and dynamics [1–6]. An important investigation area is discovering the potential signaling pathways from protein interactions to understand their roles in signal transduction, gene regulation and disease. The typical experimental method to infer the regulation relations between pathway components is perturbing the cells with molecular interventions [7,8]. It needs many experiments to determine their molecular mechanism and regulation relationships, which is expensive, time-consuming and error-prone.

Several groups have made efforts to develop bioinformatics methods to infer signaling pathways [9–14]. For example, Steffen, et al. developed a computational approach to generate static models of signal transduction networks from large-scale two-hybrid screens and expression profiles [9]. Silverbush et al. [10] and Gitter et al. [11] presented several algorithms to discover high-confidence pathways. Shlomi et al. presented a comprehensive framework, Qpath, using homologous pathway queries to identify biologically significant pathways and their functions [12]. We have also proposed two methods to predict the directionality in pairwise proteins, based on the domains and functional annotations [13,14]. These methods can achieve good performance in a part of protein interaction datasets. However, it was still difficult to determine the regulation relationship of protein interactions in the signaling pathways. Giving a pair of interacting proteins, we can predict the direction of signal flow through it using the methods proposed in [13,14], but we cannot distinguish whether its regulation relation is activation or inhibition. Therefore, it is necessary to develop new bioinformatics methods to predict the regulation relations between protein interactions.

In this paper, we introduced a novel method to predict the regulation relationship between protein interactions in the signaling network according to their constituent domains. Firstly, we proposed a measure, Enrichment_ratio, to identify the domain pairs significantly enriched in the activation/inhibition dataset. Then, we trained the classifiers based on three machine learning methods (logistic regression, SVM and naïve bayes) with the activation dataset and the inhibition dataset. Furthermore, we evaluated these methods based on 5-fold cross-validation and the independent test dataset. Finally, we applied the logistic regression method to predict the regulation relations in the human proteome-wide interactions.

* Corresponding author.
E-mail addresses: angel_nudt@126.com, liuwei314@nudt.edu.cn (W. Liu).

0006-291X/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.bbrc.2013.09.093

2. Materials and methods

2.1. Extraction of signaling networks in multiple species

As a classical and well-known pathway database, KEGG (Kyoto Encyclopedia of Genes and Genomes) contains manually annotated pathways based on biochemical evidence from the literature, including a large amount of signaling and metabolic pathways [15]. All the signaling networks of human, mouse, rat, fly and yeast were downloaded from KEGG. From these signaling networks, 1,893 protein interactions are extracted with their regulation relationship, including 1,554 in the category of activation and 339 in the inhibition, which are used as the golden standard positive set. In human, rat, mouse, fly and yeast, 76.40% proteins have one or more Pfam domains. Interaction between two proteins typically involves binding between specific domains.

2.2. Transforming human signaling pathways to protein interactions

By transforming protein interactions in signaling pathways into binary model, we transformed the pathways in 7 databases, including PID, BioCarta, Reactome, NetPath, INOH, SPIKE and KEGG and established the human protein interaction dataset with known regulation relations. This dataset include 6,791 protein interactions (Additional file 1), with 5,261 in activation and 1,530 in inhibition. Abandoning the interactions recorded in the golden standard positive set, the rest can be used as the independent test dataset to evaluate the performance of our classifier.

2.3. The computation of Enrichment_ratio and P-value of domain pairs

To investigate the enrichment extent of a domain pair appearing in the activation dataset or inhibition dataset, compared to the whole protein interaction dataset, we proposed a novel measure Enrichment_ratio. It is defined as:

$$\text{Enrichment\_ratio} = \frac{\frac{m}{M}}{\frac{n}{N}} \tag{1}$$

where N is the number of protein interactions in the whole standard dataset, M is the number of protein interactions in the activation/inhibition dataset. For a specific pair of domains, n is defined as the number of protein interactions containing this pair in the whole standard dataset, and m is the number of protein interactions containing this pair in the activation/inhibition dataset. For a given pair of domains, we can calculate two Enrichment_ratio values, one of which represents the enrichment extent in the activation dataset and the other represents the enrichment extent in the inhibition dataset. If the Enrichment_ratio in the activation dataset is larger than a certain cutoff, such as 1, then this pair is relatively enriched in this dataset. If the Enrichment_ratio is smaller than 1, it is relatively lacked. The enriched domain pairs in the inhibition dataset can be extracted by the similar method.

Furthermore, to investigate whether two domains always appear in pairs to introduce protein interaction or they only appear accidentally, we made a hypothesis testing. We used the hypergeometric cumulative distribution to analyze the enrichment significance of domain pairs appearing in the activation/inhibition dataset. For a specific pair of domains, its *P*-value is defined as:

$$P - value = \sum_{m'=m}^{n} \frac{\binom{M}{m'}\binom{N-M}{n-m'}}{\binom{N}{n}} \quad (\text{Enrichment\_ratio} \geq 1) \tag{2}$$

$$P - value = \sum_{m'=0}^{m} \frac{\binom{M}{m'}\binom{N-M}{n-m'}}{\binom{N}{n}} \quad (\text{Enrichment\_ratio} < 1) \tag{3}$$

Through setting the cutoff the *P*-value, for instance 0.05, we can discover the domain pairs significantly enriched in the activation/inhibition dataset. If the Enrichment_ratio > 1 and *P*-value < 0.05, this pair of domains is regarded as significantly enriched in the activation/inhibition dataset.

2.4. Machine learning

Three machine-learning algorithms were investigated: logistic regression, and support vector machine (SVM) based on PolyKernel, and naïve bayes, all of which have been widely used for pattern classification and regression problems. For a specific domain pair selected by the enrichment analysis, if it appears in the protein interactions of the training dataset, the corresponding feature is set as its Enrichment_ratio, otherwise this feature is set as zero. The WEKA package [16] was used to build classifiers that could distinguish the activation relations from inhibition, using selected features.

We can evaluate the performance of three classifiers using 5-fold cross-validation. During the test process, 20% of the interactions in the positive and negative datasets were singled out in turn to become the test sample, and the remaining interactions were used as the training set to predict the class of the interactions in the test sample. The performance was measured by the analysis of receiver operating characteristic (ROC) curves. A ROC curve can show the efficacy of one test by presenting both sensitivity and specificity for different cutoff points [17]. Sensitivity and specificity can measure the ability of a test to identify true positive and false positives in a data set. These two indexes can be calculated as Sensitivity = TP/T and Specificity = 1 − (FP/F) where TP and FP are the number of identified true and false positives, respectively, whereas T and F are the total number of positives and negatives in a test. The area under the ROC curve (AUC) provided the metric for the overall performance of the classifier. The closer the AUC of a test was to 1.0, the higher the overall efficacy of the test.

3. Results

3.1. Extraction of domain pairs enriched in activation/inhibition dataset

Domains are elements of proteins in a sense of structure and function. Most proteins interact with each other through their domains. Therefore, it is crucial and useful to understand PPIs based on the domains [18]. In Fig. 1, we gave an example to demonstrate the domain pairs contained in the protein interactions. Protein A contains three domains D_1, D_2 and D_3, and Protein B contains two domains E_1 and E_2. In principle, the domains contained in Protein A and the domains contained in Protein B can compose 6 domain pairs. In fact, only few domain pairs will be significantly enriched in protein interactions of the activation or inhibition dataset. These domain pairs significantly enriched in the activation or inhibition dataset may suggest the regulation relations between protein interactions, and can be used as the valid features to build classifiers in order to distinguish the activation relations from the inhibition relations.

According to the constituent domains of interacting proteins, we computed the Enrichment_ratios and P-values of domain pairs in the activation/inhibition dataset (see Section 2). We extracted 7,805 pairs of domains significantly enriched with their Enrichment_ratio > 1 and P-value < 0.05, in which 5,796 pairs are enriched in the activation dataset and 2,009 pairs enriched in the