Computers in Biology and Medicine 40 (2010) 300–305

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Prediction of protein functions based on function–function correlation relations

Ka-Lok Ng [a,*], Jin-Shuei Ciou [a,1], Chien-Hung Huang [b]

[a] Department of Bioinformatics, Asia University, 500 Lioufeng Road, Wufeng Shiang, Taichung 41354, Taiwan
[b] Department of Computer Science and Information Engineering, National Formosa University, 64, Wen-Hwa Road, Hu-wei, Yun-Lin 632, Taiwan

ARTICLE INFO

Article history:
Received 11 September 2008
Accepted 1 January 2010

Keywords:
Protein–protein interaction
Protein function
Protein function–function correlation
Yeast protein function prediction

ABSTRACT

A protein function pair approach, based on protein–protein interaction (PPI) data, is proposed to predict protein functions. Randomization tests are performed on the PPI dataset, which resulted in a protein function correlation scoring value which is used to rank the relative importance of a function pair. It has been found that certain classes of protein functions tend to be correlated together. Scoring values of these correlation pairs allow us to predict the functionality of a protein given that it interacts with proteins having well-defined function annotations.

The jackknife test is used to validate the function pair method. The protein function pair approach achieves a prediction sensitivity comparable to an approach using more sophisticated method. The main advantages of this approach are as follows: (i) a set of function–function correlation relations are derived and intuitive biological interpretation can be achieved, and (ii) its simplicity, only two parameters are needed.

1. Introduction

Knowing the biological functions of proteins is fundamental to many studies of biological processes. Biological functions include transcription, gene regulation, metabolism and cell cycle processing and so on. Assigning functions to novel proteins is one of the most important problems in proteomic study, and several methods have been developed to assign functions to an unknown protein. The conventional way is based on the homology search, such as PSI- BLAST [1]. Non-homology-based methods have recently been introduced to assign putative functions to unknown proteins, such as the Rosetta stone method [2,3], the phylogenetic method [4] and the combined method [5–7], the protein–protein interaction method [8–10], and the integrative approach [11–13].

Most of the methods developed to infer protein functions have many parameters, so it hindered these models' applications in practice. The computational method we considered, the so-called protein function pair approach, is carried forward from the protein domain pair approaches [14,15], with protein functions substituted for protein domains in the present study. In our formulation, Kim et al. approach is modified by incorporating a randomization procedure in order to assign function–function correlation score for a protein function pair, which could facilitate protein function prediction.

In Section 2 we give a description of the input data, the method used, and the randomize procedure employed in this paper. In Section 3, a set of relations for the putative protein function–function correlation (FFC) is reported, and a scoring value for each function pair is derived as well. This set of relations can be used to predict the biological functions of a protein given that it interacts with proteins having well-defined functions. In Section 4 we present the conclusion.

2. Materials and methods

2.1. Input data

The yeast proteins information is obtained from the Saccharomyces Genome Database (SGD) database [16]. Protein–protein interaction (PPI) data are downloaded from the database DIP [17,18]. Functional annotation of each PPI record in DIP is obtained from the protein database Munich Information Center for Protein Sequences (MIPS) [19]. The MIPS functional catalogue database classifies protein functions into 27 different catalogues, for instance, numbers 01 and 11 denote metabolism and transcription, respectively. See Appendix A for a complete list of the MIPS functional catalogues.

2.2. Protein function pair approach

Assuming proteins P_i and P_j contain M and N functions, respectively, then given an interacting protein pair (P_i, P_j), one

* Corresponding author. Tel.: +886 423323456x1856; fax: +886 4 23321019.
E-mail address: ppiddi@gmail.com (K.-L. Ng).
[1] moved to Taiwan Agricultural Research Institute, Taiwan, after May 22, 2009.

0010-4825/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2010.01.001

considers that there are MN possible function pairs. The set of function pairs of two proteins P_i and P_j, S_{ij}, is defined by

$$S_{ij} = s(P_i) \times s(P_j) \quad (1)$$

where $s(P_i)$ denotes a set of M protein functions in protein P_i, i.e. $\{\alpha_1,\ldots\alpha_M\}$, and $\times$ denotes the Cartesian product of two sets $s(P_i)$ and $s(P_j)$. Since a protein can either has a single function or multiple functions, combinations of possible function pairs can be derived from each of the interacting protein pair obtained from the DIP database.

To measure the likelihood of a function–function combination, the function pair interaction matrix I is introduced. The element $I_{\alpha\beta}$ denotes the weighted combination probability of a function pair (α, β) for a given protein pair (P_i, P_j), and its value is given by

$$I_{\alpha\beta} = \sum_{(\alpha,\beta)\in S_{ij}} \frac{1}{|s(P_i)| \times |s(P_j)|} \quad (2)$$

where $|s|$ denotes the cardinality of set s, the summation is over all possible pairs of (P_i, P_j) such that α and β is an element of $s(P_i)$ and $s(P_j)$, respectively. For self-interacting protein, i equals to j in Eq. (2). Then, the element of the normalized function pair score $F_{\alpha\beta}$ is defined by

$$F_{\alpha\beta} = \frac{I_{\alpha\beta}}{\sum_{\rho,\sigma} I_{\rho,\sigma}} \quad (3)$$

where (ρ, σ) denotes the function pair element for the set of interaction protein pairs. As an illustration, suppose that three proteins P_A, P_B and P_C, with functions $s(P_A)=\{\alpha1, \alpha2\}$, $s(P_B)=\{\beta1, \beta2\}$, $s(P_C)=\{\beta1\}$ and assume that the set of interaction protein pairs $\{P_A-P_B, P_A-P_C\}$. Then, $S_{AB}=\{(\alpha1, \beta1), (\alpha1, \beta2), (\alpha2, \beta1), (\alpha2, \beta2)\}$, $|s(P_A)| \times |s(P_B)|$ and $|s(P_A)| \times |s(P_C)|$ equal to four and two, respectively. Therefore, the element $I_{\alpha1,\beta1}$ equals to 3/4 (i.e. 1/4+1/2), and $\sum_{\rho,\sigma} I_{\rho,\sigma}$ equals to 2 (i.e. total number of PPIs). Thus, $F_{\alpha1,\beta1}$ equals to 3/8. The scoring schema of our calculation is determined by a randomization process, and it is described in the following sub-section.

2.3. Protein function–function correlation score

Randomized tests are performed in order to justify the protein function pair calculation. Tests are performed for the PPI data set, in which assignment of protein function is randomized while keeping the number of function assignments for each protein, and the percentage of each protein functional class in the randomized PPI set is the same as the original set. This randomization process is employed in a previous work on studying domain–domain interactions [20]. The correlation score of two protein functions is compared to its randomized counterpart, and it is defined by,

$$R_{\alpha\beta} = \frac{F_{\alpha\beta}}{\langle F_{\rho\sigma}^{rand} \rangle} \quad (4)$$

where $\langle F_{\rho\sigma}^{rand} \rangle$ and $R_{\alpha\beta}$ denote the ensemble average of the randomize counterpart of $F_{\rho\sigma}$, and the function–function correlation score of a function pair (α, β), respectively. This result provides a criterion to rank the function pairs. If the ratio $R_{\alpha\beta}$ is larger than one, it implies that the correlation is stronger than the randomized counterpart. Protein function pairs with higher score are preferred FFC relations.

2.4. Protein function prediction

The jackknife test is used to validate the protein function pair method. An annotated protein pair is selected from the yeast DIP data set. One of the proteins in the protein pair is assumed as 'unannotated', and the rest PPI data are used as the training set. The jackknife test is repeated for all the protein pairs. The performance of the prediction results is evaluated by four statistical measures, which are defined in the last paragraph of this sub-section.

Suppose the 'unannotated' protein A has m unknown functions, and it interacts with k proteins $B_1,\ldots B_k$, where these k proteins have n functions collectively, i.e. $\Psi = \{\beta_1, \ldots \beta_j^{(2)}, \beta_j, \ldots \beta_n\}$, where $\beta_j^{(2)}$ means β_j occurs twice. The correlation score $R_{\alpha\beta}$ is used to predict unknown protein functions. Let $\Phi=\{\alpha_1,\ldots\alpha_m\}$ be the set of unknown functions of A, α_j is determined by β_j with the highest $R_{\alpha_j\beta_j}$ score, where $1 \le j \le n$. In general, m is less than n, and this is because A can have more than one PPI partner. Another criterion in the model is to set the threshold of $R_{\alpha_j\beta_j}$ equal to or greater than one. A FFC pair (α_j, β_j) is removed if $R_{\alpha_j\beta_j}$ is less than one.

Since some of the β in Ψ can be repeated, it may be possible that the number of times a β_j repeated, so-called multiplicity, can possibly affect protein function prediction, i.e. multiplicity may imply more weight is needed to assign to β_j in the calculation. A parameter ρ is defined to denote the average frequency of occurrence for any functional class, and ρ is given by,

$$\rho = \frac{|\Psi|}{n} E \quad (5)$$

where $|\Psi|$ and n denote the cardinality of set Ψ and the total number of functional classes are found in $\{\beta_1,\ldots\beta_n\}$, respectively.

Ratio $|\Psi|/n$ is the expected occurrence of a functional class among the interaction proteins functional classes. For example, suppose protein A has some unknown functions, and it interacts with three proteins B_1, B_2 and B_3, where these three proteins have five functional classes, i.e. $\Psi = \{\beta_1, \beta_2^{(2)}, \beta_3^{(2)}, \beta_4, \beta_5\}$, where $|\Psi| = 7$, and $n=5$. Therefore, the ratio $|\Psi|/n$ is 1.4. An E value equals to one, i.e. $\rho=1.4$, means only the two functional classes β_2 and β_3 are selected in the model. The E value serves as a threshold in selecting the number of distinct functional class. A zero E value implies that all functional classes in Ψ are taken into consideration in predicting Φ. An E value of one implies only functional class with frequency of occurrence equal to or above the average value will be selected.

Four statistical measures are defined to characterize the prediction performance, that is, the sensitivity, S_N, specificity, S_P, accuracy, Q, and $F1$-measrue, $F1$, which are defined as $S_N=TP/(TP+FN)$, $S_P=TN/(TN+FP)$, $Q=(TP+TN)/(TP+TN+FP+FN)$ and $F1=2/(1/S_N + 1/S_P)$, respectively. $F1$-measure is the harmonic mean of S_N (recall) and S_P (precision) [21]. TP, TN, FP and FN stand for true positive, true negative, false positive, and false negative events, respectively.

3. Results

3.1. Function–function correlation pairs and scores

The MIPS database classifies some of the protein functions into a class called the unclassified catalogue. This is omitted in our calculation as we consider only the other 17 functional classes (among the 26 classes only 17 are found in the SGD database). After eliminating the unclassified class and combining with the DIP data, there are a total of 9383 functional annotations among 4710 proteins. Some of the MIPS functional classes occur more often than others among the yeast proteome and the percentage of the ten highest functional classes' composition are given in Fig. 1. The functional class statistics indicate that the top ten functional classes account for 85% of the yeast proteome.

The function pair calculation determined a set of FFC pairs. In Table 1, the rank of the top ten $F_{\alpha\beta}$ values is presented. The $F_{\alpha\beta}$