Computational Biology and Chemistry 43 (2013) 11–16

Contents lists available at SciVerse ScienceDirect

Computational Biology and Chemistry

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

Protein function prediction using neighbor relativity in protein–protein interaction network

Sobhan Moosavi[a,1], Masoud Rahgozar[a,1], Amir Rahimi[b,c,*]

[a] Database Research Group, Control and Intelligent Processing Center of Excellence, School of Electrical and Computer Engineering, University College of Engineering, University of Tehran, Tehran, Iran
[b] Medicinal and Natural Products Chemistry Research Center, Shiraz University of Medical Science, P.O. Box 71345-3388, Shiraz, Iran
[c] Department of Bioinformatics, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran

ARTICLE INFO

Article history:
Received 16 August 2012
Received in revised form 5 December 2012
Accepted 7 December 2012

Keywords:
Protein function prediction
Neighbor Relativity Coefficient
Path connectivity
Protein–protein interaction network

ABSTRACT

There is a large gap between the number of discovered proteins and the number of functionally annotated ones. Due to the high cost of determining protein function by wet-lab research, function prediction has become a major task for computational biology and bioinformatics. Some researches utilize the proteins interaction information to predict function for un-annotated proteins. In this paper, we propose a novel approach called "Neighbor Relativity Coefficient" (NRC) based on interaction network topology which estimates the functional similarity between two proteins. NRC is calculated for each pair of proteins based on their graph-based features including distance, common neighbors and the number of paths between them. In order to ascribe function to an un-annotated protein, NRC estimates a weight for each neighbor to transfer its annotation to the unknown protein. Finally, the unknown protein will be annotated by the top score transferred functions. We also investigate the effect of using different coefficients for various types of functions. The proposed method has been evaluated on *Saccharomyces cerevisiae* and *Homo sapiens* interaction networks. The performance analysis demonstrates that NRC yields better results in comparison with previous protein function prediction approaches that utilize interaction network.

1. Introduction

The increasing number of proteins with unknown function in the post genomic era has opened an important challenge for computational methods to predict function for un-annotated proteins. In the past decade, several approaches have been developed which use protein–protein interaction networks information. These approaches can be categorized into two main groups: direct annotation schemes and module assisted based schemes (Sharan et al., 2007). The first group uses protein connections for function prediction by a general assumption that closer proteins in the network have more chance to have similar functions. The second group tries to identify some modular clusters in the network and assigns function to un-annotated proteins based on the known functions in the related cluster. In direct annotation group, Neighborhood-counting (Schwikowski et al., 2000) is the simplest approach which counts the frequency of functions among direct neighbors of an un-annotated protein and then, selects the top k frequent functions and assigns them to the protein. According to Deng et al. (2003), Ahmed et al. (2011) and Wong (2011), the shortcomings of Neighborhood-counting method are as follows: (a) ignoring the full topology of the network and considering only direct neighbors; (b) considering equal weights for all interactions in the network; (c) lack of significance level for function assignment; and (d) neglecting function frequency in the entire interaction network. In Hishigaki et al. (2001), the authors proposed to use Chi-Square distribution to predict protein functions, based on function density among protein neighbors and provided a significance level for function assignment. They also proposed to use the n-neighborhood to exploit the topology of the network for function prediction. However, they did not consider any difference between the proteins with different network distances. In Chua et al. (2006), the authors used first and second level neighbors of an un-annotated protein and assigned different weights according to the network topology. This method also considers reliability of interactions between protein pairs and function frequency in the entire network and could overcome the Neighborhood-counting and Chi-Square methods in results.

In the second group, as stated in Sharan et al. (2007), some of module assisted based methods (Brun et al., 2003; Milenkovic and Przulj, 2008) just used the information extracted from network

* Corresponding author at: Medicinal and Natural Products Chemistry Research Center, Shiraz University of Medical Science, P.O. Box 71345-3388, Shiraz, Iran. Tel.: +98 711 2303872; fax: +98 711 2332225.
E-mail addresses: Sobhan.moosavi@ut.ac.ir (S. Moosavi), Rahgozar@ut.ac.ir (M. Rahgozar), RahimiAmir@sums.ac.ir, AmirRahimi@ut.ac.ir (A. Rahimi).
[1] Tel: +98 21 82089718.

1476-9271/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiolchem.2012.12.003

Table 1
The statistics of *Saccharomyces cerevisiae* and *Homo sapiens* interaction networks.

	Number of proteins	Number of interactions	GO terms	Cellular component	Molecular function	Biological process
Yeast	2112	4392	2541	500	713	1328
Human	1081	1291	3961	464	847	2650

topology; and some others (Hanisch et al., 2002; Ideker et al., 2002; Luscombe et al., 2004; Balazsi et al., 2005; Wachi et al., 2005) used further sources of information such as gene expression measurements. According to some primary evaluations, the prediction performance of the direct methods is more accurate than the module-assisted methods (Sharan et al., 2007). However, it needs a comprehensive and systematic study to confirm the results.

In this paper, we propose a novel approach named "Neighbor Relativity Coefficient" (NRC) which can be considered as a direct annotation scheme, according to the aforementioned categorization. NRC is an indicator which determines the influence of each neighbor for ascribing function to an unknown protein. Our results show that the performance of this approach is higher than other state of the art methods in protein function prediction that utilize protein–protein interaction network.

2. Materials and methods

2.1. The datasets

The Core datasets of the molecular interaction networks of *Saccharomyces cerevisiae* and *Homo sapiens* (released on 10/27/2011) were downloaded from the Database of Interacting Proteins (DIP)[1] (Xenarios et al., 2000, 2001). We also used the Gene Ontology (GO)[2] description of proteins functions which was obtained from the UniProt website.[3] GO system is a hierarchical set of functions which contains three categories: Cellular-component, Molecular-function and Biological-process (Ashburner et al., 2000). In this system, each protein may be annotated by several GO terms in each category. The functionally un-annotated proteins and the interactions with other organisms' proteins were filtered out from the original datasets to make them suitable for algorithm assessments. In the prepared datasets, there were 4392 interactions between 2112 proteins in Yeast network and 1291 interaction between 1081 proteins in Human network. Table 1 shows the statistics of the purified datasets.

2.2. Assessment of proteins function similarity

In order to measure functional similarity between two interacting proteins p_1 and p_2, we used Eq. (1), where the P_i_Function_set is the set of functions for the ^{ith} protein:

$$\text{Function\_Similarity} = \frac{P_1\text{\_Function\_set} \cap P_2\text{\_Function\_set}}{P_1\text{\_Function\_set} \cup P_2\text{\_Function\_set}} \quad (1)$$

2.3. Algorithm

We propose a novel approach for assigning function to an unknown protein based on its neighbor's functional annotation by estimating the proteins relativity. The main idea of our approach is based on the assumption that strongly linked proteins are more likely to have common functional properties than those which are less connected. NRC determines the weight of each neighbor in assigning its function to an un-annotated protein as defined by Eq. (2). In this equation, NRC is calculated for each protein pairs in the network using some graph-based features including their distance, their common neighbors and the number of paths between them. For assigning function to an un-annotated protein, each of its neighbor scores up its functions as candidate functions with regard to its NRC value. Finally the candidate functions with higher scores will be assigned to the un-annotated protein.

Table 2
The values for tuning parameters of NRC equation (α and β) which are determined by grid search.

	Saccharomyces cerevisiae		*Homo sapiens*	
	α	β	α	β
Cellular-component	2	1.5	1.2	0.4
Molecular-function	1.1	1.2	1.1	0.3
Biological-process	1.1	1.4	1.0	1.1

$$\text{NRC}_x(u,v) = F_x(d) \times \left(\sum_{p \in U_x} \frac{1}{L(p)}\right) \times \frac{a \times |N_u \cap N_v| + 1}{|N_u| + |N_v| - a \times |N_u \cap N_v| + 1} \quad (2)$$

In the above equation, u and v are two arbitrary proteins which have the maximum x-step distance in the interaction network. $F_x(d)$ is a function of the distance between two interacting proteins and is defined by Eq. (3):

$$F_x(d) = \begin{cases} 1 & \text{if } d = 1 \\ \frac{1}{\beta} & \text{if } d > 1 \end{cases} \quad (3)$$

In which β is a parameter that varies for three GO categories (Table 2) due to different reduction rates of function similarity upon distance in each category (Fig. 2). We have proposed "protein path connectivity" as a new measure for network connectivity which is used in the NRC equation. *Protein_Path_Connectivity* makes use of paths between two proteins in interaction network and is calculated as follows:

$$\text{Proteins\_Path\_Connectivity}(u,v) = \sum_{p \in U_x} \frac{1}{L(p)} \quad (4)$$

Here, U_x is the set of all paths between u and v with maximum length x, and p is a path with length $L(p)$ as a member of U_x. According to this definition, proteins with more paths and shorter

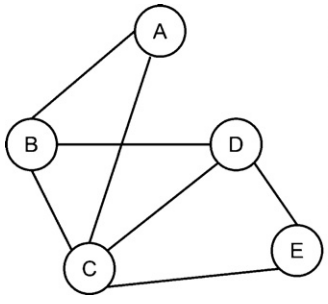

Fig. 1. Calculation of Proteins_Path_Connectivity for two pairs of nodes in a sample graph.

[1] http://dip.doe-mbi.ucla.edu/dip/Download.cgi?SM=7.
[2] http://www.geneontology.org/.
[3] http://www.uniprot.org/.

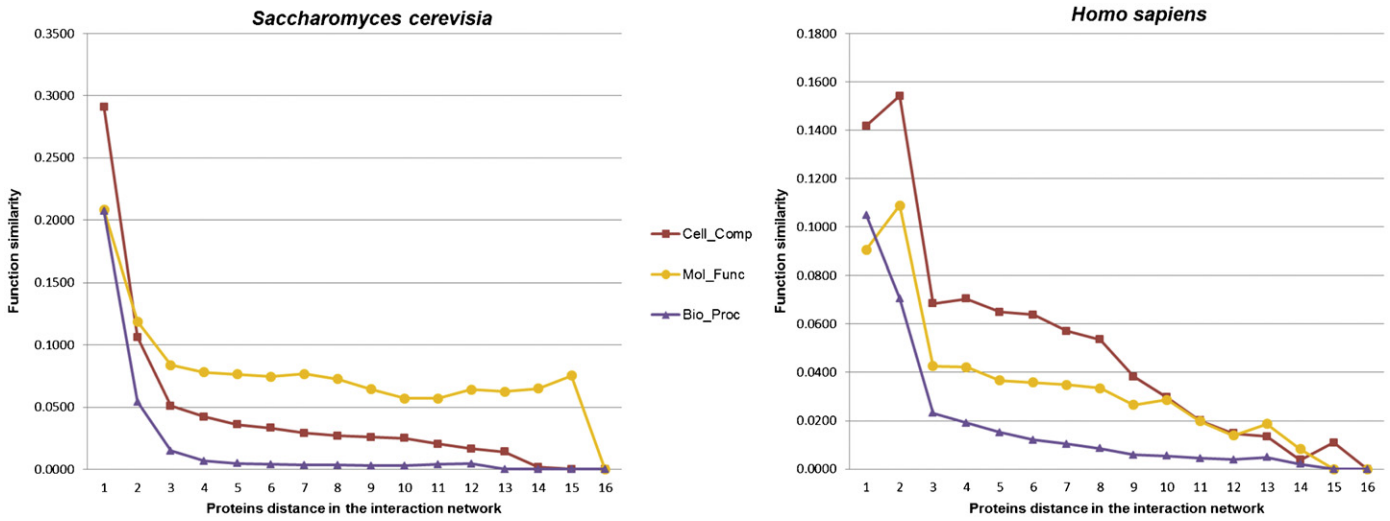

Fig. 2. The relationship between function similarity and proteins distance in the interaction networks of *Saccharomyces cerevisiae* and *Homo sapiens*. Cell_Comp, Mo_Func and Bio_Proc are stand for Cellular-component, Molecular-function and Biological-process respectively.

path-lengths are more tightly connected in the interaction network. Fig. 1 shows an example for calculation of this new measure.

Similar to the propositions in Brun et al. (2003) and Chua et al. (2006), we propose another measure for proteins relativity, called *Common_Neighbor_Ratio* which is defined as follows:

$$\text{Common\_Neighbor\_Ratio}(u, v) = \frac{a \times |N_u \cap N_v| + 1}{|N_u| + |N_v| - a \times |N_u \cap N_v| + 1} \quad (5)$$

Here, N_u is a set of proteins containing u and all direct neighbors of u, and N_v is a set which includes v and all direct neighbors of v. The parameter α tunes the effect of the number of common neighbors and was determined experimentally by grid search (Table 2).

2.4. Assessment of algorithms

The Leave-One-Out procedure was used to compare function prediction performance of NRC with three well-known methods including: Neighborhood-counting (Schwikowski et al., 2000), Chi-Square (Hishigaki et al., 2001) and FS-weight (Chua et al., 2006, 2007). The first two methods are basic approaches that show general predictability on any dataset. The third method, FS-weight, is a state-of-the art approach that has been tested on interaction networks of different organisms by Chua et al. (2007) and resulted higher function prediction performance in comparison with other methods (Chua and Wong, 2009; Chua et al., 2011). Since NRC uses both direct and indirect neighbors' information for function prediction, analysis on other methods have been executed in two levels: in the first level (tagged with #1), only direct neighbors have been considered and in the second level (tagged with #1&2), both direct and indirect neighbors were used to predict function by the aforementioned methods. Three comparison measures: Precision, Recall and *F*-score values were calculated for different methods using the following equations:

$$\text{Precision} = \frac{\sum_{p \in V} K_p}{\sum_{p \in V} m_p} \quad (6)$$

$$\text{Recall} = \frac{\sum_{p \in V} K_p}{\sum_{p \in V} n_p} \quad (7)$$

$$F = \frac{2(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \quad (8)$$

where k_p is the number of correctly predicted functions for protein p, m_p is the total number of functions predicted for protein p and n_p is the number of all known functions of protein p.

3. Results

3.1. The relationship between function similarity and network distance

The overall relationship between proteins network distances and their functional similarities were assessed for different GO function categories and the result is shown in Fig. 2. In interaction network of *S. cerevisiae*, functional similarity decreases strongly by increasing distance between proteins. However, decreasing ratios vary among different function types. For example, decreasing rate of function similarity for Molecular-function is less than other types of functions in yeast dataset.

Fig. 2 also shows that the Cellular-component similarity of the direct neighbor proteins is higher than Molecular-function and Biological-process in both datasets. In Human network, proteins are more functionally similar to indirect neighbors than direct ones in Cellular-component and Molecular-function. It might be the consequence of this point that the human interaction network is not as saturated as yeast interaction network (Hart et al., 2006) and more direct interactions might be discovered by future researches.

3.2. Function prediction assessment

The ability of NRC method in assigning function to unknown proteins was assessed on *S. cerevisiae* and *H. sapiens* networks and the same evaluation circumstances were implemented for the other three methods (Schwikowski et al., 2000; Hishigaki et al., 2001; Chua et al., 2006).

The results show that NRC significantly improves the prediction results compared with all the other methods in the prediction of Cellular-component in both networks (Figs. 3A and 4-A). In predicting Molecular-function, NRC overcomes other methods in the Human network (Fig. 4B). But, there is no significant difference between the performance of NRC and FS-weight #1&2 method for the yeast dataset (Fig. 3B).

In predicting of Biological-process, NRC has no major difference compared with FS-weight #1&2 in *S. cerevisiae* dataset (Fig. 3C). However, both methods have significantly higher performance

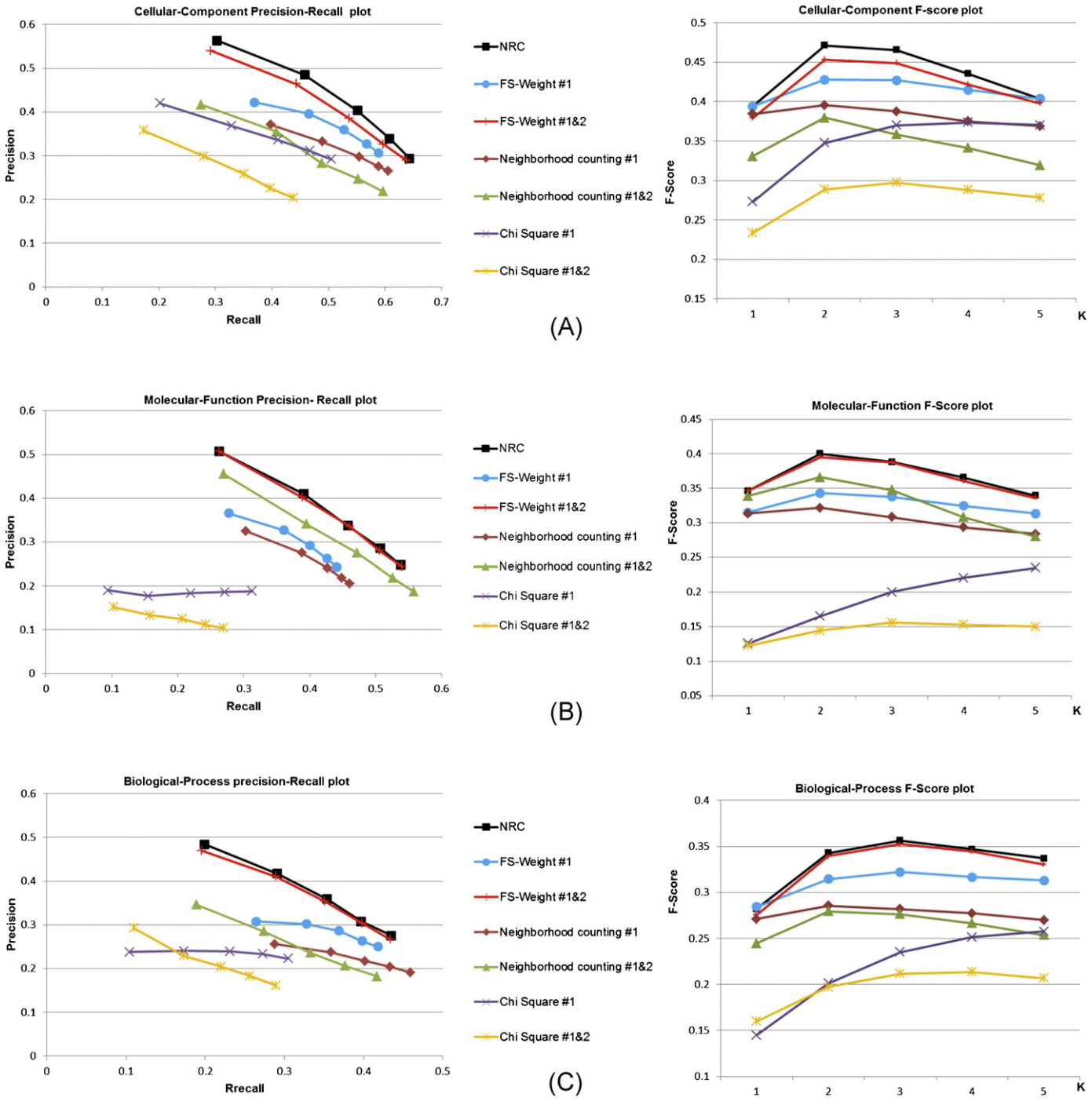

Fig. 3. Analysis of function prediction performance on yeast interaction network by Precision–Recall and *F*-score plots for three types of functions': (A) Cellular-component, (B) Molecular-function, and (C) Biological-process. The analysis comprise different methods including Neighbor Relativity Coefficient (NRC), FS-weight using only direct neighbors (FS-weight #1), FS-weight using both direct and indirect (level 2) neighbors (FS-weight #1&2), Neighborhood-counting using direct neighbors (Neighborhood-counting #1), Neighborhood-counting using both level 1 and 2 neighbors (Neighborhood-counting #1&2), Chi-Square with $n = 1$ (Chi Square #1) and Chi-Square with $n = 2$ (Chi Square #1&2).

than the other methods. For the human interaction network, NRC has competing results in comparison with FS-weight #1&2 and Neighborhood-counting #1&2 (Fig. 4C).

4. Discussion

In this study, we proposed an approach for protein function prediction based on protein–protein interaction network. This method ascribes the functions of neighbor proteins to un-annotated proteins regarding their graph-based relativity. The results demonstrate that using NRC as an influencing coefficient for each neighbor in assigning function to un-annotated proteins increases the prediction performance. In order to estimate functional relativity of two proteins, NRC combines various topological aspects of protein interaction network including distance of proteins in the network, common neighbor ratio and the number of paths between them. Each of these factors represents a different aspect of relation between two proteins in the network.

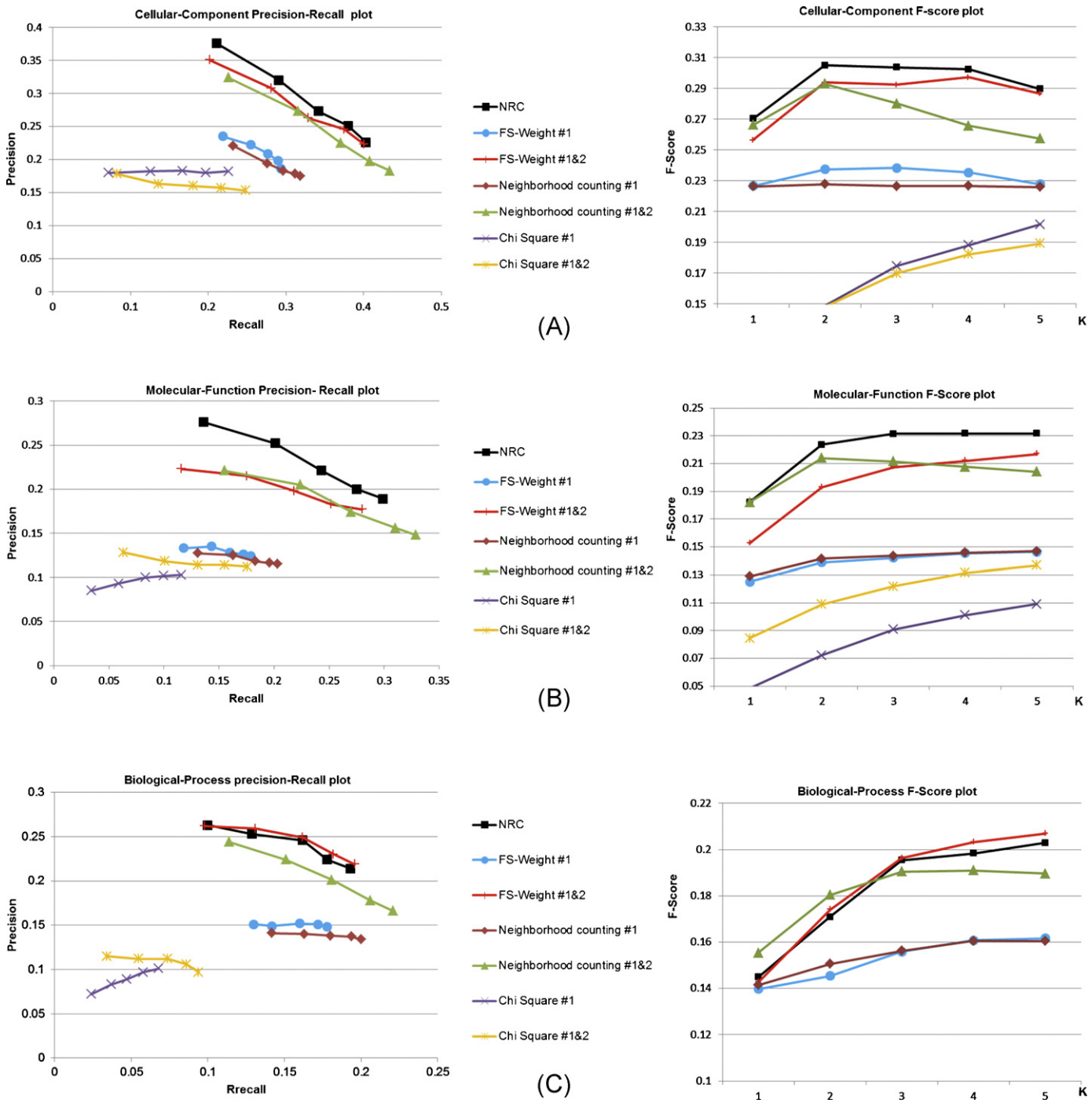

Fig. 4. Analysis of function prediction performance on human interaction network by Precision–Recall and *F*-score plots for three types of functions': (A) Cellular-component, (B) Molecular-function, and (C) Biological-process. The analysis comprise different methods including Neighbor Relativity Coefficient (NRC), FS-weight using only direct neighbors (FS-weight #1), FS-weight using both direct and indirect (level 2) neighbors (FS-weight #1&2), Neighborhood-counting using direct neighbors (Neighborhood-counting #1), Neighborhood-counting using both level 1 and 2 neighbors (Neighborhood-counting #1&2), Chi-Square with $n = 1$ (Chi Square #1) and Chi-Square with $n = 2$ (Chi Square #1&2).

Combination of these factors in the NRC equation provides an appropriate estimate of functional relativity of two proteins.

The distribution of functional similarity in the network differs for various GO function types. So, using different tuning parameters (α and β) for each GO function type is an advantage of NRC over the other methods. The appropriate performance of NRC on both human and yeast interaction networks reveals the generality and the stability of this method.

For Chi-Square methods (Chi-Square #1 and Chi-Square #1&2), the weak prediction outcomes may be the result of network sparseness in this investigation. As claimed in Hishigaki et al. (2001), the Chi-Square method works better on dense parts of interaction network.

Except Chi-Square, using level 1 and 2 neighbors simultaneously increases the prediction performance for all the other methods. The Neighborhood-counting method, despite of its simplicity, has

notable performance when uses both level 1 and 2 neighbors. However, since it does not consider any difference between direct and indirect neighbors, it produces lower performance than NRC and FS-weight #1&2 in most cases.

NRC exploits topological information of interaction networks. This information in conjunction with other complementary sources of information (such as frequency of each function in the network, co-occurrence of functions in neighbors, compatibility of functions that are assigned to a protein and the effect of interaction type on annotation transferring from a known protein to an unknown interacting protein) may improve the results. The best combination of this information for function prediction can be achieved using the state-of-the-art data fusion techniques.

5. Conclusion

Protein interaction networks contain helpful information for understanding the role of proteins in cells and predicting function for un-annotated proteins. In this research, we proposed a new approach that uses topology of protein–protein interaction network for assigning function to un-annotated proteins. This method combines three topological feature of interaction network to predict function relativity of each protein pairs. The proposed approach provides a general concept of relativity in the networks which can be used in defining relativity of two nodes in various graph-based problems.

References

Ahmed, K.S., Saloma, N.H., Kadah, Y.M., 2011. Improving the prediction of yeast protein function using weighted protein–protein interactions. Theoretical Biology and Medical Modelling 8, 11–17.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nature Genetics 25, 25–29.

Balazsi, G., Barabasi, A.L., Oltvai, Z.N., 2005. Topological units of environmental signal processing in the transcriptional regulatory network of *Escherichia coli*. Proceeding of the National Academy of Sciences of the United States of the America 102, 7841–7846.

Brun, C., Chevenet, F., Martin, D., Wojcik, J., Guenoche, A., Jacq, B., 2003. Functional classification of proteins for the prediction of cellular function from a protein–protein interaction network. Genome Biology 5, 6–17.

Chua, H., Liu, G., Wong, L., 2011. Protein function prediction using protein–protein interaction networks. In: Protein Function Prediction for Omics Era. Springer, Netherlands, pp. 243–270.

Chua, H.N., Sung, W.K., Wong, L., 2006. Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. Bioinformatics 22, 1623–1630.

Chua, H.N., Sung, W.K., Wong, L., 2007. Using indirect protein interactions for the prediction of gene ontology functions. BMC Bioinformatics 8, 8–20.

Chua, H.N., Wong, L., 2009. Predicting protein functions from protein interaction networks. In: Biological Data Mining in Protein Interaction Networks. IGI Global, Pennsylvania, pp. 203–222.

Deng, M., Zhang, K., Mehta, S., Chen, T., Sun, F., 2003. Prediction of protein function using protein–protein interaction data. Journal of Computational Biology 10, 947–960.

Hanisch, D., Zien, A., Zimmer, R., Lengauer, T., 2002. Co-clustering of biological networks and gene expression data. Bioinformatics 18, 145–154.

Hart, G.T., Ramani, A.K., Marcotte, E.M., 2006. How complete are current yeast and human protein-interaction networks? Genome Biology 7, 120–128.

Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., Takagi, T., 2001. Assessment of prediction accuracy of protein function from protein–protein interaction data. Yeast 18, 523–531.

Ideker, T., Ozier, O., Schwikowski, B., Siegel, A.F., 2002. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 18, 233–240.

Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A., Gerstein, M., 2004. Genomic analysis of regulatory network dynamics reveals large topological changes. Nature 431, 308–312.

Milenkovic, T., Przulj, N., 2008. Uncovering biological network function via graphlet degree signatures. Cancer Informatics 6, 257–273.

Schwikowski, B., Uetz, P., Fields, S., 2000. A network of protein–protein interactions in yeast. Nature Biotechnology 18, 1257–1261.

Sharan, R., Ulitsky, I., Shamir, R., 2007. Network-based prediction of protein function. Molecular Systems Biology 3, 88–100.

Wachi, S., Yoneda, K., Wu, R., 2005. Interactome–transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. Bioinformatics 21, 4205–4208.

Wong, L., 2011. Using biological networks in protein function prediction and gene expression analysis. Internet Mathematics 7, 274–298.

Xenarios, I., Fernandez, E., Salwinski, L., Duan, X.J., Thompson, M.J., Marcotte, E.M., Eisenberg, D., 2001. Dip: the database of interacting proteins: 2001 update. Nucleic Acids Research 29, 239–241.

Xenarios, I., Rice, D.W., Salwinski, L., Baron, M.K., Marcotte, E.M., Eisenberg, D., 2000. Dip: the database of interacting proteins. Nucleic Acids Research 28, 289–291.