Netw Model Anal Health Inform Bioinforma (2013) 2:175–183
DOI 10.1007/s13721-013-0024-z

ORIGINAL ARTICLE

Dynamically predicting protein functions from semantic associations of proteins

Jingyu Hou · Wei Zhu · Yi-Ping Phoebe Chen

Received: 24 September 2012 / Revised: 2 December 2012 / Accepted: 17 January 2013 / Published online: 6 February 2013

Abstract Predicting functions of un-annotated proteins is a significant challenge in the post-genomics era. Among existing computational approaches, exploiting interactions between proteins to predict functions of un-annotated proteins is widely used. However, it remains difficult to extract semantic associations between proteins (i.e. protein associations in terms of protein functionality) from protein interactions and incorporate extracted semantic associations to more effectively predict protein functions. Furthermore, existing approaches and algorithms regard the function prediction as a one-off procedure, ignoring dynamic and mutual associations between proteins. Therefore, deriving and exploiting semantic associations between proteins to dynamically predict functions are a promising and challenging approach for achieving better prediction results. In this paper, we propose an innovative algorithm to incorporate semantic associations between proteins into a dynamic procedure of protein function prediction. The semantic association between two proteins is measured by the semantic similarity of two proteins which is defined by the similarities of functions two proteins possess. To achieve better prediction results, function similarities are also incorporated into the prediction procedure. The algorithm dynamically predicts functions by iteratively selecting functions for the un-annotated protein and updating the similarities between the un-annotated protein and its neighbour annotated proteins until such suitable functions are selected that the similarities no longer change. The experimental results on real protein interaction datasets demonstrated that our method outperformed the similar and non-dynamic function prediction methods. Incorporating semantic associations between proteins into a dynamic procedure of function prediction reflects intrinsic relationships among proteins as well as dynamic features of protein interactions, and therefore, can significantly improve prediction results.

Keywords Protein function prediction · Interaction network · Semantic association

J. Hou (✉)
School of Information Technology,
Deakin University, Melbourne, Australia
e-mail: jingyu@deakin.edu.au

W. Zhu · Y.-P. P. Chen (✉)
Department of Computer Science and Computer Engineering,
La Trobe University, Melbourne, Australia
e-mail: phoebe.chen@latrobe.edu.au

W. Zhu
e-mail: w6zhu@students.latrobe.edu.au

1 Introduction

Assigning functions to proteins whose functions have not been annotated through biological experiments continues to be one of the challenges in computational biology due to the importance of proteins in various biological processes and the high cost of biological experiments. To tackle this challenge, significant effort has been given for predicting protein functions using computational methods or tools. In general, the development of computational approaches for protein function prediction has undergone two stages. In the early stage, researchers mainly concentrated on searching for the homologous sequences of the un-annotated protein, and the annotated functions of the homologous sequences were selected as the predicted functions of the un-annotated protein. In other words, the approaches in this stage mainly focused on the inner biological structure of individual proteins without considering the external and

mutual interactions among proteins (Abual-Rub et al. 2012). The BLAST (Altschul et al. 1990) and FASTA (Pearson 1990) were two representative tools in this stage.

With the development of biological technology, it has been revealed that proteins involved in biological processes do not work individually, but interact with each other to implement various complex biological processes. It has been demonstrated that a protein may show different functions in different biological processes when it interacts with different proteins (Misteli 2001). These discoveries lead to the second development stage of computational prediction approaches, where predictions are based on the external interactions among proteins, rather than only on the inner structures of individual proteins. Meanwhile, new biological technologies have produced large amounts of high-throughput protein–protein interaction (PPI) data which provide interaction information between proteins. Due to the rich information it carries about the external and mutual interaction information of proteins, PPI data are now widely used for protein function prediction.

Usually, PPI data are modelled as an interaction network with nodes representing proteins and edges representing protein interactions (Kilic and Mehmet 2012; Xiang et al. 2012). With this PPI network model, the neighbour proteins of a protein are defined as those proteins in the network that directly interact with the protein. Based on this model and an assumption that an un-annotated protein would share functions with its neighbour proteins, a number of algorithms have been developed to make use of the known functions of neighbour proteins to predict functions of an un-annotated protein. In other words, this kind of prediction approach exploited the topological, rather than the semantic or biological, information of the PPI network for predictions. The representative algorithms of this approach were the Markov random field (MRF) method (Deng et al. 2003) and the majority rule method (MRM) (Schwikowski et al. 2000). The MRF method made use of the topological structure of a PPI network to determine the probability of a function being assigned to all other proteins in the whole dataset and predicted functions of an un-annotated protein from the probability ranking. Some variations of the MRF method can be found in Letovsky and Kasif (2003) and Vazquez et al. (2003). The MRM, which was also known as the neighbour counting method, predicted functions by selecting functions that have higher occurrence frequencies among the neighbour proteins as the predicted functions. One deficiency of these algorithms was that they did not consider the indirect neighbour proteins within the whole PPI network when predicting functions. To address this issue, Hishigaki et al. (2001) defined the neighbourhood with levels (i.e. the neighbour proteins have direct or indirect interactions with the un-annotated protein), and the top-ranked functions in terms of their occurrence frequency among the levelled neighbour proteins are selected as the predicted functions of an un-annotated protein. Chua et al. (2006) developed another algorithm to calculate the similarity between a pair of proteins when making use of both direct and indirect neighbours of an un-annotated protein to predict functions.

Applying clustering methods on PPI networks to predict protein functions is another representative approach. This approach relies on the assumption that proteins in the same cluster (or complex in biology) should share some common functions. With this approach, proteins in a PPI network including unannotated ones are clustered into clusters based on their similarities derived from their interaction relationships in the PPI network. The known functions (usually the representative or the feature functions) of a cluster are the candidates from which the predicted functions are selected for the un-annotated proteins in the same cluster. A representative clustering-based method of detecting a molecular complex is the MCODE (Bader and Hogue 2003). In the MCODE, each edge of a PPI network is given a weight and the tensest interconnected modules are considered complexes. Similar to the MCODE, other clustering-based algorithms were proposed by Spirin and Mirny (2003), Pereira-Leal et al. (2004) and Dunn et al. (2005) to predict complexes as well as functions. Since clustering-based algorithms rely on protein similarity definitions, Samanta and Liang (2003) defined a P value as the similarity of a pair of proteins and divided a PPI network into clusters according to the similarity values to improve accuracy of prediction. Arnau et al. (2005) and Rives and Galitski (2003) also defined protein similarities to measure the shortest distance between proteins for function prediction purposes. The clustering-based prediction method in Zhu et al. (2010) defined a new protein similarity and excluded the un-annotated proteins in the clustering operations. In addition to the above approaches, other methods were also proposed for function predictions, such as those that use the graph theory and data mining techniques. More details and discussions can be found in Sharan et al. (2007).

It has been observed that existing prediction methods exploit the associations (e.g. in terms of topological similarity) between proteins in various ways to predict functions. Although some algorithms have good prediction results for specific datasets, in most cases the prediction results have been unsatisfactory. Therefore, extracting and exploiting semantic associations between proteins (i.e. protein associations in terms of protein functionality) constitute one of the keys to improving existing prediction results. On the other hand, existing similarity-based prediction approaches regard the prediction as a one-off procedure, i.e. the available known functions in a prediction domain (e.g. a set of neighbour proteins) finally determine the functions of an un-annotated protein. In other words,

the un-annotated protein is passive, and the function prediction is unidirectional and a one-off procedure. However, in real biological processes, proteins have high mobility and dynamically interplay to produce a framework which is ever-changing but overall stable (Misteli 2001). The proteins exchange their biological information and share functions in a dynamic, rather than a static and unidirectional, circumstance. This means that the mutual interactions between pair-wise proteins reveal the reality of biological processes, and this dynamic feature of protein interactions should be taken into account when predicting functions. Unfortunately, existing prediction approaches do not take this dynamic feature into consideration.

The work in this paper addresses the above issues in function prediction. We incorporate semantic association information between proteins into a dynamic prediction procedure to iteratively predict functions for un-annotated proteins. The algorithm is well designed to make significant proteins and functions endorse each other through iterative procedures, and to select the most semantically important functions as the predicted functions for un-annotated proteins. The semantic associations between proteins in the algorithm are expressed in terms of functional similarity between proteins, and the dynamic prediction procedure of the algorithm reflects the dynamic features of protein interactions. Incorporating semantic association between proteins into a dynamic function prediction procedure makes our approach different from existing ones.

This paper is organized as follows. In Sect. 2, we propose and discuss the dynamic prediction algorithm that incorporates semantic association information between proteins. The experimental results of our algorithm and result analyses, as well as comparisons with other algorithms, are given in Sect. 3. Finally, in Sect. 4 we conclude our work and discuss some possible improvements to our approach in the future.

2 Methods

The proposed protein function prediction algorithm is based on dynamic functional voting from the neighbour proteins of the un-annotated protein. The functions that obtain higher votes or voting scores from the neighbour proteins are the predicted functions of the un-annotated protein. The votes a function obtains depend on the importance of not only the neighbour proteins that have the function, but also the function among all available functions possessed by the neighbour proteins. These two kinds of importance are measured in terms of protein semantic similarity (i.e. the protein similarity defined by function similarities) and function similarity, respectively. Suppose we have already defined the semantic similarity between two proteins p and p' as $sim(p, p')$, and the similarity between two functions f and f' as $f\ sim(f, f')$. Let $N(p)$ be the set of neighbour proteins of the protein p, $F(p)$ be the set of functions protein p has, and $NF(p)$ be the set of functions the protein p's neighbour proteins have, i.e. $NF(p) = \bigcup_{p' \in N(p)} F(p')$. The scores a function $f \in NF(p)$ which can be obtained from functional voting of the un-annotated protein p's neighbour proteins are defined as

$$Score(p,f) = \sum_{p' \in N(p)} \left[sim(p,p') \times \sum_{f' \in F(p')} f\ sim(f,f') \right] \quad (1)$$

It can be seen from (1) that the MRM proposed by Schwikowski et al. (2000) is actually a special case of our prediction method, where $sim(p, p')$ in (1) is set to 1 [i.e. $sim(p, p') = 1$], the component $\sum_{f' \in F(p')} f\ sim(f,f')$ in (1) is replaced by a simple indicator function (i.e. if p' has function f then the value is 1, otherwise 0) without considering function associations, and $N(p)$ consists of only the proteins that directly interact with the un-annotated protein p, i.e. the level-1 neighbour proteins of p.

From Eq. (1), the scores a function f obtains from neighbour protein voting are determined by two factors: the importance of each neighbour protein to the un-annotated protein p [i.e. $sim(p, p')$ in (1)], and the importance of function f in each neighbour protein [i.e. $\sum_{f' \in F(p')} f\ sim(f,f')$in (1)]. For example, functions that are possessed by the most important neighbour proteins and are the most important among neighbour functions will achieve the highest scores, and therefore be selected as the predicted functions of the un-annotated protein p. It is obvious a function possessed by the less important neighbour proteins but very important among neighbour functions, or possessed by very important neighbour proteins but less important among neighbour functions, is still likely to obtain higher scores.

For a given un-annotated protein p and its neighbour proteins, the importance of a function f in each neighbour protein [i.e. $\sum_{f' \in F(p')} f\ sim(f,f')$ in (1)], and in turn the importance of f among neighbour functions, is fixed. Therefore, the variation of the score depends on the variation of semantic similarities between the un-annotated protein p and its neighbour proteins [i.e. $sim(p, p')$, in (1)]. However, since the functions of the un-annotated protein p are unknown and to be predicted, the semantic similarity $sim(p, p')$ is therefore uncertain. Our prediction method is to iteratively update the semantic similarities $sim(p, p')$ between the un-annotated protein and its neighbour proteins, and in turn to iteratively update function scores until they no longer change. The functions with the higher scores are then selected as the predicted functions. To do this and

to reflect the dynamic features of protein interactions in this iterative semantic similarity updating, it is necessary to define the semantic similarity between proteins from their functional similarities, so the important neighbour proteins and important functions endorse each other through the iterative updated by Eq. (1), with the final selected functions the most important to the un-annotated protein.

In our iterative prediction algorithm, we adopt the function similarity and semantic protein similarity definitions from our previous work, where the protein semantic similarity between proteins is defined from their functional similarities. To ensure the completeness and understanding of these similarity definitions, we present the details of these definitions here. More details can be found in Zhu et al. (2010).

Protein functions can be expressed in the format of an annotation scheme, such as the Gene Ontology (GO) and the Functional Catalogue (FunCat). In our work, we use the FunCat scheme for protein function annotation. The FunCat is a numerical hierarchical annotation scheme developed by the Munich Information Centre for Protein Sequences (MIPS) (Ruepp et al. 2004). The scheme allows a protein function to be expressed numerically by up to six layers. A digital number at each layer defines a specific function category. The deeper a function's layer achieves, the more specific the function is. With the FunCat scheme and our definition, the similarity of two protein functions is determined by the common layers two functions share, i.e. the more layers two functions share, the more similar two functions are. The details of the function similarity definition are as follows.

For two given functions f and f', which are in the FunCat digital layer format, we define $l(f,f') \stackrel{def}{=}$ the number of common sequent layers the f and f' share from the first layer. It is obvious $0 \leq l(f, f') \leq 6$. For example, suppose $f = 10.01.05.03.01.0$ and $f' = 10.01.03.0.0.0$, these two functions share the first two layers 10.01 and therefore $l(f, f') = 2$. Based on this definition, we define the similarity $f\,sim(f, f')$ between two functions f and f' as

$$f\,sim(f,f') = \sum_{i=1}^{l(f,f')} i^2 / \sum_{j=1}^{6} j^2 \tag{2}$$

If $l(f, f') = 0$, then $f\,sim(f, f') = 0$. With the function similarity definition in (2), we define the semantic similarity between two proteins p_1 and p_2 as

$$sim(p_1,p_2) = \sum_{f\in F(p_1)} \sum_{f'\in F(p_2)} f\,sim(f,f') / \sum_{f\in F(p_1)} \sum_{f'\in F(p_2)} w(f,f') \tag{3}$$

where

$$w(f,f') = 2 - l(f,f')/6.$$

To calculate the function scores in (1) with respect to the un-annotated protein p, two questions need to be answered: one is how to choose the neighbour proteins $N(p)$ from which the functions are predicted, and the second question is how to calculate the similarities between the un-annotated protein p and its neighbour proteins. Regarding the first question, we choose level-1 and level-2 neighbour proteins of the un-annotated protein p as the neighbour protein set $N(p)$. Level-1 neighbour proteins are those that directly interact with the un-annotated protein p, while level-2 neighbour proteins are those that directly interact with the level-1 neighbour proteins of p but do not interact with the un-annotated protein p directly. This neighbour protein set $N(p)$ construction is based on the work in (Chua et al. 2006), where it was indicated that in most cases, level-1 and level-2 neighbour proteins contain major functions of the protein p.

Regarding the second question above, the issue is that functions of protein p are unknown. However, our semantic protein similarity $sim(p, p')$ is calculated from protein function similarities. Therefore, to kick off the prediction, we need to assign initial functions to the un-annotated protein p. To do this, we select the level-1 neighbour proteins of p and assume $sim(p, p') = 1$ in (1) to calculate the scores of level-1 functions. The level-1 functions are then ranked according to their scores. We select the average number of functions each level-1 neighbour protein has as the cut-off rate for selecting ranked level-1 functions to initialize the functions of the un-annotated protein p. For example, if on average each level-1 neighbour protein has three functions, we then select the functions with the first three highest scores as the initial functions of p.

With the initial functions assigned to the un-annotated protein p, we can now calculate the similarities between the un-annotated protein p and its neighbour proteins in (1). At this stage, the neighbour protein set $N(p)$ consists of level-1 and level-2 neighbour proteins rather than just level-1 proteins. For a level-2 neighbour protein, its similarity with the un-annotated protein p is calculated a little differently. In fact, suppose a level-2 protein is p_2, then the similarity between p and p_2 is calculated as

$$sim(p,p_2) = \mathrm{Max}\left[sim(p,p_2), \max_{p_1\in N1(p_2)} (sim(p,p_1) \times sim(p_1,p_2))\right] \tag{4}$$

where $N1(p_2) = \{\, p_1 | p_1$ is the level 1 neighbour of p and directly interacts with $p_2\}$

With the issues of calculating the function score in Eq. (1) being resolved as described above, function prediction can be conducted dynamically based on (1). In fact, with the initial functions being assigned to the un-annotated protein p, we set the $N(p)$ in Eq. (1) as the set of

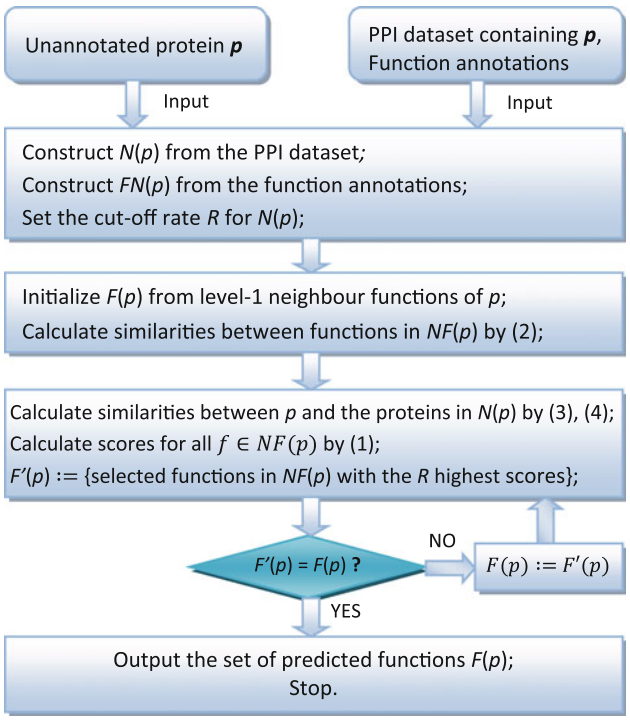

Fig. 1 The diagram of the dynamic function prediction algorithm

level-1 and level-2 neighbour proteins of p, and then calculate the scores of all functions in the neighbour and rank the functions according to their scores. The functions with higher scores are selected as the first-round predicted functions. The number of selected functions depends on the cut-off rate in the prediction. At this prediction stage, we define the cut-off rate as the *average number* of functions each protein has in the neighbour set $N(p)$. If the current-round prediction results are the same as the previous ones, i.e. the prediction results do not change anymore, the prediction operation is then stopped and the current-round prediction results are the final predicted functions of the un-annotated protein (for the first-round of prediction, the previous-round prediction results are those initial functions of the un-annotated protein). Otherwise the current-round predicted functions are assigned to the un-annotated protein and the prediction based on (1) is conducted again until the results do not change anymore. This dynamic prediction algorithm is described with a diagram in Fig. 1.

Regarding the convergence of the algorithm, as mentioned in the above algorithm discussion, the dynamic prediction procedure makes the protein similarities and function scores endorse each other by Eq. (1) iteratively. In other words, if a function with the highest functional score is selected as a predicted function after the first round of iteration, it will be kept as a predicted function in the following iteration rounds as well, and each iteration round will select new functions with the highest functional score while the previously selected functions are still kept. This means that after finite rounds of iteration, the prediction results will no longer change or the algorithm is convergent. Our experiments on real protein nitration datasets also demonstrated the convergence of the algorithm.

Incorporating semantic protein association information and function similarities into the function prediction via the function score calculation (1) differentiates our algorithm from existing ones because this incorporation makes it possible to iteratively predict functions of an un-annotated protein and reflect the dynamic features of protein interactions in the prediction procedure. The existing similarity-based prediction methods, however, try to define the protein similarity by making use of other available information, such as the PPI network topological structure information or gene microarray information, rather than the information about the protein and function associations. Therefore, existing prediction methods cannot predict functions iteratively as the protein similarities are separated from function similarities in these methods, thus ignoring dynamic features of protein interaction.

3 Experimental results

The proposed dynamic prediction algorithm was evaluated on two real PPI datasets of budding yeast *Saccharomyces cerevisiae*. The first dataset was obtained from the BIOGRID PPI database (http://thebiogrid.org). This dataset contained a total of 232,238 protein–protein interactions. For evaluation purposes, 172,001 interactions that contained annotated proteins (i.e. the functions of these proteins were known) were selected for the experiments, and the other 60,237 interactions that contained un-annotated proteins were removed from the dataset. For the selected 172,001 interactions, there were 5,702 annotated proteins involved in the interactions. Another dataset was obtained from the MIPS PPI database (ftp://ftpmips.gsf.de/yeast/PPI). There were a total of 15,456 PPIs, from which we selected 8,050 PPIs for the experiments, while removing 7,406 PPIs that referred to un-annotated proteins. The selected 8,050 PPIs contained 1,172 annotated proteins. The summary of these two datasets is shown in Table 1.

In the experiments, we used the MIPS FunCat scheme (Ruepp et al. 2004) for function annotation. To evaluate the effectiveness of the algorithm, we chose some annotated proteins from the dataset and recorded their functions. We regarded these chosen proteins as un-annotated proteins and predicted their functions with our dynamic prediction algorithm. For each un-annotated protein, if a predicted function was the same as a recorded one, the prediction was correct for that function. We did not consider layered predictions in the experiments.

Table 1 Datasets for evaluation

Descriptions	BIOGRID	MIPS
Total number of PPIs	232,238	15,456
Number of PPIs with annotated proteins	172,001	8,050
Number of PPIs with un-annotated proteins	60,237	7,406
Total number of proteins for evaluation	5,702	1,172

The effectiveness of the algorithm was evaluated by the prediction *precision*, *recall* and *F value*. Their definitions are as follows. Let N_A be the number of real functions a protein actually processes, N_C be the number of correctly predicted functions, and N_P be the number of all predicted functions. The prediction precision and recall for a protein are calculated as:

$$\text{Precision} = \frac{N_C}{N_P},\ \text{Recall} = \frac{N_C}{N_A}. \quad (5)$$

To avoid the trade-off between the precision and recall in the evaluation, we adopted the F value (Kiritchenko et al. 2005) to evaluate the overall performance of a prediction, which is defined as follows:

$$F\ \text{value} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

Due to the lack of existing similar iterative prediction algorithms, we compared our iterative algorithm with the MRM (Schwikowski et al. 2000). We denote this method as MRM, and our proposed method as Dynamic Prediction Algorithm (DPA). The MRM is a non-dynamic method and its prediction also relies on neighbour function voting from which our dynamic prediction algorithm stems. The comparison was to show that the dynamic prediction approach with the incorporation of semantic protein associations can significantly improve function predictions.

Since the neighbour of an un-annotated protein in our DPA consists of level-1 and level-2 neighbour proteins, to make the prediction results comparable, we also extended the MRM method to level-1 and level-2 neighbour proteins of an un-annotated protein. We named the original MRM method as *MRM_1*, and the extended MRM method as *MRM_2*.

We compared the prediction precisions, recalls and F values of three methods on both datasets. The purpose was to demonstrate that the DPA outperformed non-dynamic algorithms on different datasets that had different data sizes and came from different data resources. We first evaluated the algorithms on the BIOGRID PPI dataset. We randomly selected ten groups of proteins from the dataset. The number of functions to be predicted for each group was between 15 and 26, and the total number of functions to be predicted for ten groups was 211. We predicted the functions for each protein in each group using three methods, and calculated the corresponding precisions, recalls and F values. Then, we calculated the average precision, recall and F values for each group. The overall (i.e. the average of all groups) precision, recall and F value of three methods across ten groups are listed in Table 2. It can be seen from Table 2 that the DPA achieved 60.6, 40.2 and 48.3 % of average precision, recall and F value, respectively, which were much higher than the methods MRM_1 and MRM_2.

Table 2 Overall precisions, recalls and F values of three methods on the BIOGRID dataset

	MRM_1	MRM_2	DPA
Overall precision	0.344	0.588	0.606
Overall recall	0.312	0.301	0.402
Overall F value	0.327	0.398	0.483

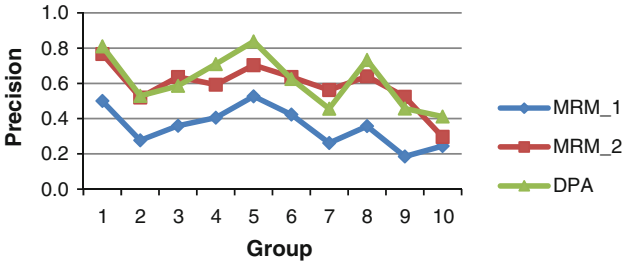

Fig. 2 Comparison of precisions on the BIOGRID dataset

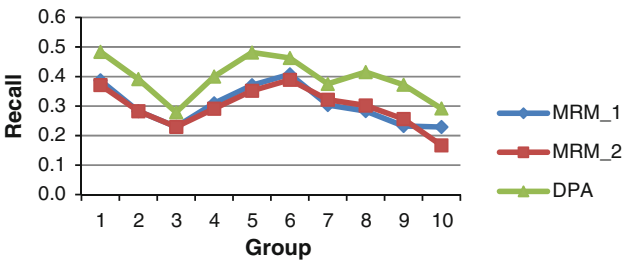

Fig. 3 Comparison of recalls on the BIOGRID dataset

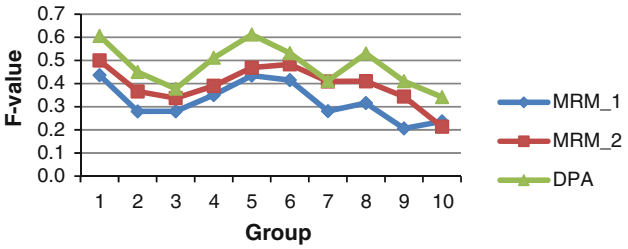

Fig. 4 Comparison of F values on the BIOGRID dataset

Table 3 Overall precisions, recalls and F values of the three methods on the MIPS dataset

	MRM_1	MRM_2	DPA
Overall precision	0.377	0.471	0.547
Overall recall	0.313	0.308	0.384
Overall F value	0.342	0.372	0.451

As the data quality varies for different data groups, we present Figs. 2, 3 and 4 to examine the detailed prediction performance of three methods on each individual group. In terms of precision (see Fig. 2), the dynamic method DPA outperformed the MRM_1 across all groups, and outperformed the MRM_2 for most groups. In terms of recall and F value (see Fig. 3, 4), the DPA outperformed the MRM_1 and MRM_2 across all groups. Therefore, overall, the dynamic method DPA outperformed the non-dynamic methods on the BIOGRID dataset.

In addition to the evaluations on the BIOGRID PPI dataset, we also evaluated the effectiveness of the methods on the MIPS PPI dataset, due to the usual differences in data quality between different datasets. Our evaluation was to demonstrate that for different datasets with different data quality, our dynamic prediction algorithm still outperformed the non-dynamic algorithms. To this end, and similar to the evaluation on the BIOGRID PPI dataset, we also randomly selected ten protein groups from the MIPS PPI dataset for evaluation. The number of functions to be predicted in each selected group ranged from 50 to 62. The total number of functions to be predicted for all groups was 532. The overall prediction effectiveness of the three methods in terms of precision, recall and F value across all groups is shown in Table 3. It is clear from Table 3 that the DPA still outperformed the MRM_1 and MRM_2.

The performance of the three methods on individual groups in terms of precision, recall and F value is presented in Figs. 5, 6 and 7, respectively. The evaluations on the MIPS dataset, as well as the BIOGRID dataset, showed that the predictions on the prediction domain that consisted of level-1 and level-2 neighbour proteins of an un-annotated protein produced better prediction results, i.e. the MRM_2 and DPA methods outperformed the MRM_1 method. For the MRM_2 and DPA, which both relied on the level-1 and level-2 neighbour proteins of an un-annotated protein for prediction, the dynamic method DAP outperformed the MRM_2 on the MIPS dataset in terms of precision, recall and F value as shown in Figs. 5–7.

Figure 8 presents the precision-recall curves of the three methods, which show the effectiveness of the methods from another perspective. It clearly shows that the dynamic method DPA is much more effective than the non-dynamic methods MRM_1 and MRM_2 across the whole range of recall.

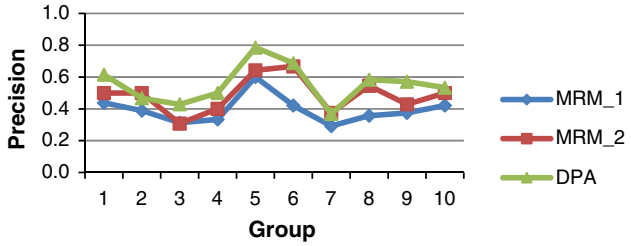

Fig. 5 Comparison of precisions on the MIPS dataset

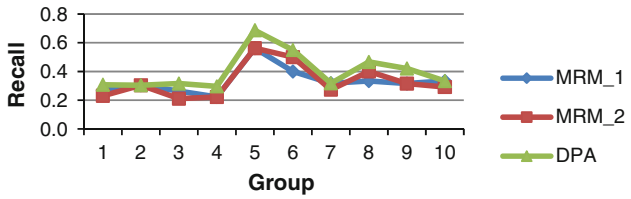

Fig. 6 Comparison of recalls on the MIPS dataset

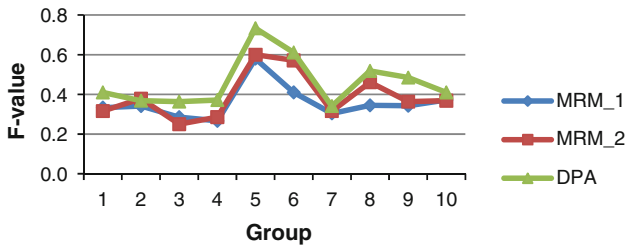

Fig. 7 Comparison of F values on the MIPS dataset

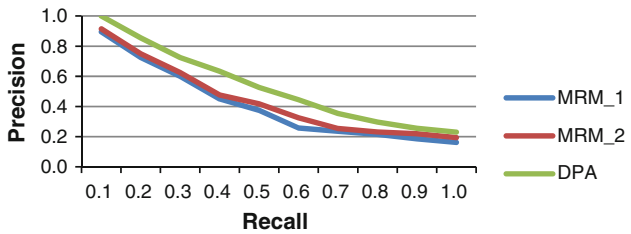

Fig. 8 Precision–recall curves of the MRM_1, MRM_2 and DPA methods

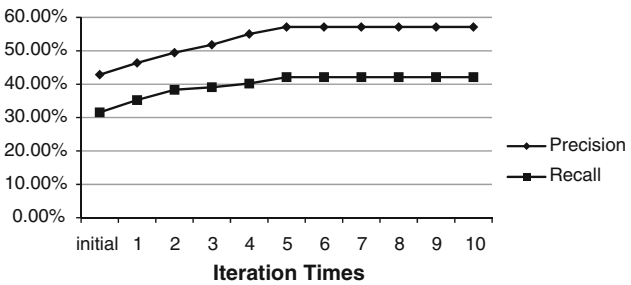

Fig. 9 Convergence evaluation of the DPA algorithm

In our experiments, we also evaluated the convergence of our iterative prediction algorithm DPA. For each evaluation data group, we calculated the average precision and recall for each round of iteration, and observed that usually after two to five rounds of iteration the prediction results became stable, i.e. the results no longer changed. In other words, the DPA algorithm was convergent. It was also observed that after the initialization, the first round of iteration and the following iterations significantly improved the initial prediction results. Figure 9 shows the evaluation results on a data group. As shown in Fig. 9, the precision and recall increased gradually until the results became stable after five iterations.

Finally, we randomly selected some sample prediction results produced from our dynamic prediction algorithm and present them in Table 4. As shown in Table 4, most cut-off rates used in the DPA were the same as or close to the real number of functions the "un-annotated" proteins have. It demonstrated that our cut-off rate selection for the predictions was reasonable. It can be seen from the table that most real functions of the "un-annotated" proteins were predicted correctly. It was also noticed that some

Table 4 Randomly selected sample prediction results from the DPA method

ID	Protein	No. of functions	Cut-off rate	FunCat	FunCat description	Predicted
1	Q0045	3	3	02.11.0.0.0.0	Electron transport and membrane-associated energy conservation	√
				02.13.03.0.0.0	Aerobic respiration	√
				20.01.15.0.0.0	Electron transport	√
2	YAL062w	5	3	01.01.03.02.01.0	Biosynthesis of glutamate	×
				01.01.03.02.02.0	Degradation of glutamate	√
				01.02.0.0.0.0	Nitrogen, sulphur and selenium metabolism	√
				02.10.0.0.0.0	Tricarboxylic-acid pathway (citrate cycle, Krebs cycle, TCA cycle)	√
				42.01.0.0.0.0	Cell wall	×
3	YBL002w	3	3	10.01.09.05.0.0	DNA conformation modification (e.g. chromatin)	√
				11.02.03.04.0.0	11.02.03.04 Transcriptional control	√
				16.03.01.0.0.0	DNA binding	√
4	YBL004w	2	2	11.04.01.0.0.0	rRNA processing	√
				16.03.03.0.0.0	RNA binding	√
5	YBL017c	4	3	14.04.0.0.0.0	Protein targeting, sorting and translocation	√
				16.01.0.0.0.0	Protein binding	×
				20.09.07.0.0.0	Vesicular transport (Golgi network, etc.)	√
				20.09.13.0.0.0	Vacuolar/lysosomal transport	√
6	YBL052c	4	3	10.01.09.05.0.0	DNA conformation modification (e.g. chromatin)	√
				11.02.03.04.0.0	Transcriptional control	√
				14.07.04.0.0.0	Modification by acetylation, deacetylation	√
				34.11.03.0.0.0	Chemoperception and response	×
7	YBL082c	2	2	01.05.0.0.0.0	C-compound and carbohydrate metabolism	√
				14.07.02.02.0.0	*N*-directed glycosylation, deglycosylation	√
8	YBL090w	3	2	12.01.01.0.0.0	Ribosomal proteins	√
				12.04.01.0.0.0	Translation initiation	×
				42.16.0.0.0.0	Mitochondrion	√
9	YBL037w	3	2	14.10.0.0.0.0	Assembly of protein complexes	×
				20.09.13.0.0.0	Vacuolar/lysosomal transport	√
				20.09.18.0.0.0	Cellular import	√
10	YAL055w	2	2	14.04.0.0.0.0	Protein targeting, sorting and translocation	√
				20.09.03.0.0.0	Peroxisomal transport	√

The symbol "√"means the corresponding function was predicted correctly by our algorithm, and the symbol "×" means the corresponding function was not predicted correctly by our algorithm

functions could not be predicted by the DPA method. There might be other factors that affected the prediction accuracy of the algorithm, such as the noises in the dataset and the selection of cut-off rate. These factors will be addressed and investigated in our future work.

4 Conclusions

In this paper, we propose an innovative algorithm to incorporate the semantic association information between proteins for dynamically or iteratively predicting functions of un-annotated proteins. The dynamic prediction procedure of our algorithm reflects dynamic features of protein interactions. Protein semantic similarities derived from protein function similarities guarantee the dynamics of the prediction procedure and the convergence of the algorithm. These two major contributions make our method different from other existing methods. The evaluations on real protein–protein interaction datasets demonstrated the effectiveness of our new method. It is concluded that incorporating semantic protein association information into dynamic prediction can significantly improve the prediction quality.

It was observed from our evaluations on real PPI datasets that the selection of a cut-off rate in the prediction had some impact on the prediction quality. If the cut-off rate was low (i.e. the number of predicted function was small), it may result in some real functions being excluded from the final predicted functions and a low recall value. But if the cut-off rate was too high (i.e. the number of predicted functions was too big), it may result in many unrelated functions being included in the final predicted functions and a low precision. Selecting cut-off rates dynamically to achieve a better balance between the precision and recall of the prediction is a possible way to improve dynamic prediction results. Our evaluations on different PPI datasets from different resources also demonstrated that the data quality varied across different datasets. Choosing reliable protein interactions from a dataset or many datasets to increase the accuracy of predictions is a further challenge. These issues will be addressed in our future work to improve the quality and effectiveness of function prediction.

References

Abual-Rub MS, Al-Betar MA, Abdullah R, Khader AT (2012) A hybrid harmony search algorithm for ab initio protein tertiary structure prediction. Netw Model Anal Health Inform Bioinform 1(3):69–85

Altschul SF, GishW Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Arnau V, Mars S, Marn I (2005) Iterative cluster analysis of protein interaction data. Bioinformatics 21:364–378

Bader G, Hogue C (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinform 4:2

Chua HN, Sung WK, Wong L (2006) Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. Bioinformatics 22:1623–1630

Deng M, Zhang K, Mehta S et al (2003) Prediction of protein function using protein–protein interaction data. J Comp Biol 10:947–960

Dunn R, Dudbridge F, Sanderson C (2005) The use of edge-betweenness clustering to investigate biological function in protein interaction networks. BMC Bioinform 6:39

Hishigaki H, Nakai K, Ono T et al (2001) Assessment of prediction accuracy of protein function from protein–protein interaction data. Yeast 18:523–531

Kilic C, Mehmet Tan (2012) Positive unlabelled learning for deriving protein interaction networks. Netw Model Anal Health Inform Bioinform 1(3):87–102

Kiritchenko S, Matwin S, Famili F (2005) Functional annotation of genes using hierarchical text categorization. In: Proceedings of the BioLINK SIG: linking literature, information and knowledge for biology

Letovsky S, Kasif S (2003) Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics 19(Suppl. 1):i197–i204

Misteli T (2001) Protein dynamics: implications for nuclear architecture and gene expression. Science 291:843

Pearson WR (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 183:63–98

Pereira-Leal JB, Enright AJ, Ouzounis CA (2004) Detection of functional modules from protein interaction networks. Proteins 54:49–57

Rives AW, Galitski T (2003) Modular organization of cellular networks. PNAS 100:1128–1133

Ruepp A, Zollner A, Maier D et al (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res 32:5539–5545

Samanta MP, Liang S (2003) Predicting protein functions from redundancies in large-scale protein interaction networks. PNAS 100:12579–12583

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

Sharan R, Ulitsky I, Shamir R (2007) Network-based prediction of protein function. Mol Syst Biol 3:88

Spirin V, Mirny LA (2003) Protein complexes and functional modules in molecular networks. PNAS 100:12123–12128

Vazquez A, Flammini A, Maritan A et al (2003) Global protein function prediction from protein–protein interaction networks. Nature Biotech 21:697–700

Xiang Y, Fuhry D, Kaya K, Jin R, Çatalyürek UV, Huang K (2012) Merging network patterns: a general framework to summarize biomedical network data. Netw Model Anal Health Inform Bioinform 1(3):103–116

Zhu W, Hou J, Chen YP (2010) Semantic and layered protein function prediction from PPI networks. J Theor Biol 267:129–136