NPF：Network propagation for protein function prediction

bihai zhao (bihaizhao@163.com)
Changsha University https://orcid.org/0000-0003-0870-7468
Zhihong Zhang
Changsha uinversity
Meiping Jiang
Changsha University
Sai Hu
Changsha University
Yingchun Luo
Changsha University
Lei Wang
changsha university

Research article

Keywords: Network propagation, Protein-protein interaction, prediction of protein function Background

Posted Date: March 10th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-16452/v1

Version of Record: A version of this preprint was published on August 12th, 2020. See the published version at https://doi.org/10.1186/s12859-020-03663-7.

NPF：Network propagation for protein function prediction

Bihai Zhao[1,3,5*], Zhihong Zhang[1,3*], Meiping Jiang[2,4], Sai Hu[1], Yingchun Luo[2,4], Lei Wang[1,3§]

[1] College of Computer Engineering and Applied Mathematics, Changsha University, Changsha, Hunan 410022, China

[2] Department of Ultrasound, Hunan Provincial Maternal and Child Health Care Hospital, Changsha, Hunan 410008, China

[3] Hunan Provincial Key Laboratory of Industrial Internet Technology and Security, Changsha University, Changsha, Hunan 410022, China

[4] NHC Key Laboratory of Birth Defect for Research and Prevention (Hunan Provincial Maternal and Child Health Care Hospital), Changsha, Hunan 410100, China

[5] Hunan Provincial Key Laboratory of Nutrition and Quality Control of Aquatic Animals, Changsha University, Changsha, Hunan 410022, China

[*] These authors contributed equally to this work

[§]Corresponding Author, Lei Wang (email: wanglei@xtu.edu.cn)

Abstract

Background:

The accurate annotation of protein functions is of great significance in elucidating the phenomena of life, disease treatment and new drug development. Various methods have been developed to facilitate the prediction of functions by combining protein interaction networks (PINs) with multi-omics data. However, how to make full use of multiple biological data to improve the performance of functions annotation is still a dilemma.

Results

We presented NPF (Network Propagation for Functions prediction), an integrative protein function predicting framework assisted by network propagation and functional module detection, for discovering interacting partners with similar functions to target proteins. NPF leverages knowledge of the protein interaction network architecture and multi-omics data, such as domain annotation and protein complex information, to augment protein-protein functional similarity in a propagation manner. We have verified the great potential of NPF for accurately inferring protein functions. Comprehensive evaluation of NPF indicates that NPF archived higher performance than competing methods in terms of leave-one-out cross-validation and ten-fold cross validation.

Conclusions:

We demonstrated that network propagation combined with multi-omics data can not only discover more partners with similar function, but also effectively free from the constraints of the "small-world" feature of protein interaction networks. We conclude that the performance of function prediction depends greatly on whether we can extract and exploit proper functional similarity information from protein correlations.

Keywords: Network propagation, Protein-protein interaction, prediction of protein function

Background

Proteins are the main component of cells and play an essential role in nearly all cell functions such as composing cellular structure etc. Biological functions are performed by groups of interacting and functionally associated proteins, instead of individual proteins. The accurate characterization of protein functions is a key to understanding life at the molecular level and has a huge impact on biomedicine and pharmaceuticals. With the advent of high-throughput sequencing technology, the number of proteins with available sequence data and unknown functions has grown dramatically. Thus, accurately inferring functions for unknown proteins has become one of the great challenges in the post-gene era. However, due to the inherent difficulty and high cost, experimental techniques to determine protein functions has been unable to meet the growing genomic sequence data. The availability of an increasing number of protein-protein interaction data translates into an urgent need for computational methods to predict protein functions. A protein interaction network (PIN) can be modelled as an undirected graph, in which a vertex represents a protein and an edge denotes an interaction between a pair of proteins. Intuitively, many network-based [1, 2, 3] or graph-based [4, 5] approaches are applied to predict protein functions from PINs. These methods are based on the observation that proteins often possess similar or identical biochemical functions with their interaction partners in the PINs [1]. Unfortunately, these methods are often plagued by noise and errors**,** which can result in biases and low confidence in PINs. Analysis based on the concordance of protein

interaction data suggests that only 30–50% of the high-throughput interactions are biologically relevant [6].

To provide an accurate prediction results, the integration of different types of biological data has become an important and popular strategy. A number of approaches have been developed to facilitate the prediction of protein functions by combining PPIs with multi-source biological data. Cozzetto *et al*. [7] proposed an effective method to infer protein functions by integrating PINs and a wide variety of biological information, such as sequence, gene expression, etc. Zhang *et al*. [8] developed the domain context similarity for the prediction of protein functions using protein domain composition and PINs. As an improvement on Zhang's method, two algorithms, named DCS (domain combination similarity) [9] and DSCP (context of protein complexes) were proposed to annotate unknown proteins by combining PINs with proteins' domain information and protein complexes information. For the annotation of protein functions, the protein overlap network (PON) [10] was constructed using the protein domain information and PIN topology. Sarker *et al*. [11] initially reconstructed a protein-protein network based on PINs and protein domains, and then presented the *GrAPFI* method for the annotation of protein functions. INGA [12] and INGA 2.0 [13] web servers were developed to infer protein functions by combining protein interaction networks, domain assignments and sequence similarity. PANNZER2 [14] was another functional annotation web server based on sequence similarity practical. In spite of the advances in these methods, it was a central challenge to the integration of multiple biological data categories within a single analysis framework.

In the context of functions prediction, most network analysis methods depended on the principle of 'guilt by association', which is based on observations that a protein

shares many functional features with its direct interacting partners in PINs. A simple and generic method might be to characterize unknown proteins with functions of all direct neighbours in PINs. Nevertheless, such a straightforward way would potentially yield false positives that are linked to proteins by irrelevant interactions; it would also introduce false negatives that do not directly connect to proteins with known functions. It is verified by our statistics on the yeast PINs. We investigated the relationships between pairs of proteins with common functions and the distance between them, as shown in Figure 1. Figure 1 reveals an interesting phenomenon that proteins seem to co-annotate with their level-3 or level-4 neighbours instead of direct interacting partners, due to the incompleteness and fault of the PINs. To address this hurdle, as a proxy to a 'functional distance' between proteins, the short-path distance instead of Euclidean distance was adopted in some approaches to infer protein functions. However, most of proteins can arrive at other proteins within a few steps because of the small-world feature of the PIN. Although these approaches can effectively suppress false negatives, it will also return many spurious functions by including irrelevant interactions. Network propagation provides us with a more refined approach by using the flow of information through network connections as a means to establish relationships between nodes [15]. There are various guises of network propagation, such as random walks on graphs, the Google PageRank search algorithm, heat diffusion processes, graph kernels, etc. In biological network, plenty of methods based on network propagation have been widely applied to herb targets identification [16], drug synergy prediction [17], tumors classification [18, 19], disease associated genes identification [20, 21] and drug-disease associations inference [22], which demonstrated that network propagation is a powerful data transformation method of broad utility in genetic research [23]. Inspired by these findings, we developed an

unsupervised network propagation based method, named NPF, for prediction of protein functions. Our model initially simulates the random walk with restart algorithm and constructs a propagation network by integrating knowledge of the protein interaction network architecture, protein domains and protein complexes. Using it as a base, we detect functional modules with high coupling to infer functions for unknown proteins. To evaluate the performance of NPF, we apply our method and six other state-of-the-art methods for prediction of protein functions on yeast PINs. Experimental results demonstrated that NPF outperformed these competing methods, including Neighbourhood-counting (NC) [1], Zhang [8], DCS [9], DSCP [9], PON [10] and *GrAPFI* [11].

Methods

The NPF method is divided into three stages: (1) Constructing three protein-protein correlation networks by integrating knowledge of the protein interaction network architecture, protein-domain associations and protein-complex associations. (2) Building a propagation network by applying an improved random walk with restart algorithm to multiple protein correlation networks. (3) Detecting functional modules with high coupling in the propagation network and annotating functions for target proteins. The flowchart for the NPF method is shown in Figure 2.

Construction of multiple protein correlation networks

Biological functions are performed by a group of genes or proteins which are related to one or more cellular interactions, e.g. protein-protein interaction, co-regulation, co-expression or membership of a protein complex. Physical PINs directly indicate the cooperation of proteins to drive a biological process [24]. Moreover, computational approaches had successfully detected stable functional modules from co-expression networks [25]. We suspect that tightly interacting and functionally dependent proteins

may co-express, co-regulate or share a common protein complex, etc. Therefore, we constructed multiple protein-protein correlation networks with integration of knowledge of protein interaction network architecture, protein domain annotation and protein complexes information.

Co-Neighbor network

Molecular functions are performed by groups of proteins interacting to each other. So, a straightforward strategy is to annotate proteins for target proteins using knowledge of the protein interaction network architecture. In this study, we used the overlapping interacting partners between a pair of proteins as an estimate of their functional correlation. In the Co-Neighbor network, two proteins are connected if they have a physical interaction and link to one or more common proteins simultaneously. Given a pair of proteins p_i and p_j in the Co-Neighbor network, their correlation value was calculated as follow:

$$P\_N(p_i, p_j) = \frac{2|N_{p_i} \bigcap N_{p_j}|}{|N_{p_i}| + |N_{p_i} \bigcap N_{p_j}|} \times \frac{2|N_{p_i} \bigcap N_{p_j}|}{|N_{p_j}| + |N_{p_i} \bigcap N_{p_j}|} \quad (1)$$

where, N_{pi} and N_{pj} represents the set of direct neighbors of p_i and p_j respectively. $N_{pi} \cap N_{pj}$ is an intersection of N_{pi} and N_{pj}.

Co-Domain network

Domains are sequential and structural motifs found independently in different proteins and play as the stable functional block of proteins. We now generalize the idea to construct a protein correlation network based on the protein domain annotation information. For a pair of proteins p_i and p_j, let M denotes the total number of domain categories in PINs, and let x and y represent the number of domain categories of p_i and p_j, respectively. Let z expresses the number of overlapping domain categories

between p_i and p_j. Then, we measured the functional correlation between two proteins p_i and p_j in the Co-Domain network with the follow formula:

$$P\_D(p_i, p_j) = -\log \frac{M^z (M-s)^{x-z} (M-a)^{y-z}}{M^x M^y} \quad (2)$$

Finally, the correlation score between p_i and p_j was obtained by the normalization processing, which was described as follows:

$$P\_D(p_i, p_j) = \frac{P\_D(p_i, p_j) - \min_{1\le k\le n|,1\le l\le n|}(P\_D(p_k, p_l))}{\max_{1\le k\le n|,1\le l\le n|}(P\_D(p_k, p_l)) - \min_{1\le k\le n|,1\le l\le n|}(P\_D(p_k, p_l))} \quad (3)$$

Co-Complex network

Protein complexes consisting of molecular aggregations of proteins assembled by multiple protein interactions are fundamental units of macro-molecular organization and play crucial roles in integrating individual gene products to perform useful cellular functions. Studies [9] have revealed that if two proteins are consisted of the same protein complexes, they tend to perform the same or similar biological functions. As much, incorporating quality-controlled protein complexes and analysing functional associations are both essential for accurate function annotation. We therefore proposed to construct the protein correlation network Co-Complex, where the functional correlation between two proteins is measured using the Equation (4).

$$P\_C(p_i, p_j) = \frac{|C_{p_i} \cap C_{p_j}|}{|C_{p_i}| * |C_{p_j}|} \quad (4)$$

In Equation (4), C_{pi} and C_{pj} represents the set of protein complexes in which p_i and p_j is involved respectively. $C_{pi} \cap C_{pj}$ denotes the set of protein complexes containing both p_i and p_j.

Network propagation algorithm

The network propagation algorithm involved a random walk with restart process on multiple protein correlation networks to generate an aggregated protein functional

similarity network with high confidence. This process considered the global connectivity patterns of the PIN for annotating target proteins. Moreover, this algorithm took the structural feature and modular feature of proteins into account for measuring functional similarity by performing a two-step propagation operation. The output of the network propagation algorithm is a propagated protein functional matrix, which could be used as input for protein function prediction.

At the first step of the network propagation algorithm, we established the transition matrix, *H*, based on the Co-Neighbor network. The transition probability from protein *i* to protein *j* was computed using the following equation:

$$h(i,j)=\begin{cases}\dfrac{P\_N(p_i,p_j)}{\sum_{k=1}^{n}P\_N(p_i,p_k)} & , \quad if \ \sum_{k=1}^{n}P\_N(p_i,p_k)>0 \\ 0 & , \quad otherwise\end{cases} \tag{5}$$

Intuitively, we wish to calculate functional similarity between proteins by propagation that takes both structural feature and modular feature of proteins into account. These two features are derived from domain annotation and protein complex information, respectively. Therefore, we performed a two-step propagation operation to calculate functional similarity between the protein p_i with other proteins by

$$VD_i^{t+1}=\alpha HVC_i^t+(1-\alpha)RV\_D_i \tag{6}$$

$$VC_i^{t+1}=\alpha H^T VD_i^t+(1-\alpha)RV\_C_i \tag{7}$$

where the parameter $\alpha\in[0,1]$ balances between the propagation information and restart scores, VD_i^t and VC_i^t are two vectors at the t step to measure structural correlation and modular correlation between protein p_i with the remaining proteins, respectively. Elements of the two vectors are initialized to $1/n$ (i.e., $VD_i^0=[1/n,1/n,\cdots,1/n]^T$, $VC_i^0=[1/n,1/n,\cdots,1/n]^T$). It was to note that it is possible to

tune the functional similarity scores by defining two restart vectors RV_D_i and RV_C_i by

$$RV\_D_i = [P\_D(i,1), P\_D(i,2), \cdots P\_D(i,n)]^T \tag{8}$$

$$RV\_C_i = [P\_C(i,1), P\_C(i,2), \cdots P\_C(i,n)]^T \tag{9}$$

In this study, we set α to 0.5 [26, 27]. When the propagation converges, we can obtain an adjacency matrix responding to the propagation network, which is formally described as follows:

$$PN = \begin{bmatrix} VC_{11}+VD_{11} & VC_{12}+VD_{12} & \cdots & VC_{1n}+VD_{1n} \\ VC_{21}+VD_{21} & \ddots & & \vdots \\ \cdots & \cdots & \ddots & \vdots \\ VC_{n1}+VD_{n1} & \cdots & \cdots & VC_{nn}+VD_{nn} \end{bmatrix} \tag{10}$$

The overall framework of network propagation algorithm can be illustrated as the Algorithm 1.

Algorithm 1: network propagation algorithm

Input: multiple protein correlation networks ; Stopping threshold ∂;
Output: An adjacency matrix *PN* responding to the propagation network

1. Construct a transition matrix H with Equation (5)
2. FOR each protein p_i
3. Initialize $VD_i^0 = [1/n, 1/n, \cdots, 1/n]^T$; $VC_i^0 = [1/n, 1/n, \cdots, 1/n]^T$
4. Calculate restart vectors RV_D_i and RV_C_i with Equation (8) and (9)
5. Let t=1
6. Calculate $VD_i^{t+1} = \alpha H VC_i^t + (1-\alpha) RV\_D_i$
7. Calculate $VC_i^{t+1} = \alpha H^T VD_i^t + (1-\alpha) RV\_C_i$
8. If $\| VD_i^t - VD_i^{t-1} \| + \| VC_i^t - VC_i^{t-1} \| < \partial$ then let $VD_i = VD_i^t$, $VC_i = VC_i^t$ and terminate the algorithm. Otherwise, let t=t+1, and then go to Step6.
9. EDN FOR
10 $PN = [VD_1 + VC_1, VD_2 + VC_2, \cdots VD_n + VC_n]$
11. Output *PN*

Prediction of protein functions

Intuitively, interacting partners are helpful to characterize target proteins. However, members of the same functional module are often more densely connected than those across functional modules [28]. Therefore, at the final stage of our work, we threw out loosely connected neighbours and annotated target proteins with the remaining partners in the newly constructed propagation network. Given a target protein *v*, *M_V* is a module of the propagation network *PN*, which is composed of all neighbour nodes of *v*. The module fitness [29] was introduced to quantitative describe the cohesion of *M_V*.

$$f_{M\_V} = \frac{WD^{in}_{M\_V}}{(WD^{in}_{M\_V} + WD^{out}_{M\_V})^{\beta}} \tag{11}$$

where $WD^{in}_{M\_V}$ and $WD^{out}_{M\_V}$ are the total internal and external weighted degree of *v* in the module *M_V*, β is a positive real-valued parameter, controlling the size of the module. To simplify operation, we set β to 1. The aim of this stage was to determine a module starting from protein *v* such that the inclusion of a new neighbour or the elimination of one neighbor from the module would lower f_{M_V}. So to do this, we introduced the concept of neighbour fitness. Given a *v*'s neighbour *u*, the neighbour fitness of *u* in reference to the module *M_V* was calculated as follows:

$$f^{u}_{M\_V} = f_{M\_V+\{u\}} - f_{M\_V-\{u\}} \tag{12}$$

In equation (12), *M_V*+ {*u*} and *M_V*-{*u*} represents the module obtained from *M_V* with neighbour *u* inside and outside, respectively.

First, neighbours of *v* were ranked in descending order according to the functional similarity to *v*. And then, all neighbours of *v* were visited and nodes with neighbour fitness greater than 0 were selected to form a candidate proteins set $P = \{p_1, p_2, \ldots, p_l\}$.

Let F= {f_1, f_2 ,…, f_m} be a list of functions of all proteins in P. The score of a candidate function f_j in F can be calculated as follows:

$$Score\_F(f_j) = \sum_{u=1}^{l} PN(v,u) \times t_{uj} \quad (13)$$

where $PN(v, u)$ represents the functional similarity between u and v in the newly constructed propagation network. If u contains function f_j, then t_{uj} =1, otherwise t_{uj}=0. Finally, candidate functions were ranked in descending order according to their scores and TOP K of them were selected to characterize the target protein v. In this study, the parameter K was set to the number of functions of the protein with the greatest functional similarity to v in the propagation network PN. The Algorithm 2 gave the overall framework of the proposed NPF method.

Algorithm 2: NPF

Input: A PIN network, domain annotation information, protein complex information, Stopping threshold Ɛ, target protein v;
Output: Top K functions

Step 1. Construct three protein correlation networks according to Equations (1)-(4)

Step 2. PN= Algorithm1(three protein correlation networks, Ɛ)

Step 3. Generate the candidate proteins set P= {p_1, p_2 ,…, p_l} according to Equation (11) and (12)

Step 4. Sort and rank functions of proteins in P according to Equation (13)

Step 5. Output top K of sorted functions

Results

Experimental data

To test the performance of NPF, we applied our method and six competing methods to infer protein functions in the protein interaction network of *Saccharomyces cerevisiae* (Baker's yeast), because of their completeness, convincement, and widespread used in function prediction algorithms as gold standard data. The PIN data is derived from DIP database [30], updated to Feb.28, 2012, which consists of 5023 proteins and 22570 interactions among the proteins with self-interactions and repeated

interactions removed. The annotation data of proteins used for validation was downloaded from GO official website [31]. The protein domain data was downloaded from Pfam database [32], which contains 1079 different types of domains associated with 3035 proteins in the DIP network. The benchmark protein complexes set was adopted from CYC2008 [33], which consists of 408 complexes involving 1492 proteins in the DIP dataset. The above four dataset were uniformly transformed to use the Ensemble Genomes Protein labelling system.

Assessment criteria

Two assessment criteria were adopted to compare function prediction performance of the NPF with six competing methods, including NC [1], ZhangDC [8], DCS [9], DSCP [9], PON [10] and *GrAPFI* [11]. The NC method is a classic protein function annotation method, which is only based on the PIN. Zhang and DCS infered protein functions using protein domain composition and PINs, and DSCP extends the protein functional similarity definition in DCS by combining the domain compositions of both proteins and complexes including them. PON and *GrAPFI* constructed a protein correlation network and characterized unknown proteins by integrating PINs and protein domain information.

Proteins in PINs were divided into two categories: the training set and the testing set. The testing set consists of target proteins with unknown proteins, which were annotated with proteins with known proteins in the training set. The validation process is repeated multiple times until each protein has a chance to become a member of the testing set. The final performances were evaluated by the average of all rounds. The first assessment criterion was leave-one-out cross-validation [9] which put one target protein into the testing set and the rest of proteins into the training set per round. However, the leave-one-out cross-validation was often plagued by many

unannotated proteins in the network. Another assessment criterion used in this study was ten-fold cross validation [34], in which the proteins set was randomly divided into ten subsets, a single subset was retained for the testing set, and the remaining nine subsets were used as the training set. The cross-validation process was then repeated ten times, with each of the ten subsets used exactly once as the testing set. The ten results from the folds were then averaged to produce the final performance.

To assess the quality of predicted functions, we matched inferred functions with actual functions of target proteins. Precision and Recall were the commonly used measures to test the performance of function prediction methods. Precision is the fraction of predicted functions that are matched with known proteins while Recall is the fraction of known functions that are matched with predicted functions. In this study, true positive (TP), false positive (FP) and false negative (FN) represents the number of matched predicted functions, incorrectly matched predicted functions and missing matched known functions, respectively. Therefore, these two measures can be defined as follows:

$$\mathrm{Pr}\,ecison = \frac{TP}{TP + FP} \tag{14}$$

$$\mathrm{Re}\,call = \frac{TP}{TP + FN} \tag{15}$$

F-measure, as the harmonic mean of Precision and Recall, was another measure to evaluate the performance of a method synthetically, which was calculated as follows:

$$F - measure = \frac{2 \times \mathrm{Pr}\,ecision \times \mathrm{Re}\,call}{\mathrm{Pr}\,ecision + \mathrm{Re}\,call} \tag{16}$$

Leave-one-out cross-validation

First, the leave-one-out cross validation was applied to verify the quality of predicted functions inferred by our NPF methods, as well as a representative set of competing

methods: NC, Zhang, DCS, DSCP, PON and *GrAPFI*. To avoid too special or general, we filtered out those GO terms whose number of annotated proteins are less than 10 or more than 200 proteins. After processing by this step, the number of GO terms is 267. Out of all the 5023 proteins in the PINs, 2860 proteins are annotated. We first assess the performance of NPF and six other competing methods on these target proteins by the average Precision, Recall and F-measure. The basic information about predicted functions by NPF and six other competing methods was presented in Table 1. In Table 1, *MP* was the number of proteins successfully matching at least one known function, while *PMP* represented the number of proteins perfectly matching the known functions, and *ZP* is the number of proteins with zero-error prediction. MMP denotes the number of proteins completely mismatching the known functions. In other words, none of the predicted functions match the known functions. From Table 1, we can see that NPF contained the second-biggest number of perfect matching proteins (959) after NC (1105), while *ZP* of our method (1555) is far more than NC's (218). Figure 3 showed the overall comparison in terms of Precision, Recall and F-measure. It illustrated that NPF archives the largest value of Precision and F-measure, the second-largest value of Recall after NC. This is due to the maximum number of perfect matching proteins with NC. F-measure of NPF was 124.38%, 98.11%, 40.71%, 22.94%, 201.37% and 93.07% higher than NC, Zhang, DCS, DSCP, PON and *GrAPFI*, respectively.

To further investigate the performance of NPF and six other competing methods, we applied the Precision-Recall (PR) curve, whose vertical and horizontal coordination are the values of Precision and Recall, respectively. The PR curve is a standard for evaluation of the comprehensive performance of all methods in terms of different strategies of function selection. Predicted functions were ranked in

descending order according to the values of functional similarity calculated by NPF, NC, PON and *GrAPFI*, respectively. Then, the top *K* functions were selected and annotated target proteins. The Parameter *K* changed from 1 to 50. As for the methods of Zhang, DSC and DSCP, top *N* (*N*<=*K*) proteins which had the highest similarity value with target proteins were selected and *K* functions in these fell out proteins were selected in turn to characterize target proteins. For a given target protein and the parameter *K*, the precision and recall values can be calculated according to the definition in Equations (14) and (15). The final PR curves of NPF and six other competing methods were drew according to the average precision and recall values over all target proteins. The PR curves of seven methods were illustrated in Figure 4. Numbers in brackets represent the maximum F-measures for these seven methods. As shown in Figure 4, NPF archived the first maximum F-measures in all methods. The PR curves of our method was above that of six other competing methods, which means that the NPF has a higher number of true positives and at the same time a smaller number of false positives when selecting different parameters. With the constant increase of *K*, the PR curve of NPF did not show drastic fluctuations. Even in the worst case, the precision value of NPF can still archive 0.274. However, the precision values of DSCP and DCS drop sharply with the emergence of a large number of similar proteins. Additionally, the corresponding areas under the curves (AUC) were calculated for quantitative evaluation of all these methods. The AUC of NPF exhibits improvements of 60.38%, 159.34%, 30.77%, 9.42%, 163.80% and 164.79% compared to the values achieved by NC, Zhang, DCS, DSCP, PON and *GrAPFI*, respectively.

For overall comparison, we counted the number of true positive and false positive functions predicted by NPF and competing methods. A more valuable comparison

between these methods was presented by plotting FP/TP curves as parameter *K* varies. Figure 5 showed the FP/TP of our method and six other competing methods fluctuated under various value of the parameter *K* (ranging from 1 to 50). The smaller slope of the FP/TP curve of a method was, the lower the noise ratio was, which resulted in a greater predicted accuracy of the method. From this figure we can see that, FP/TP curve of NPF has consistently been covered with that of all other methods. That is, NPF generated the fewest false positives among all the methods when matching the same number of known functions.

To further analyze the difference between NPF and six other competing methods, we selected YNL262W, YBR278W and YPR175W as examples and infered proteins using the seven methods. Table 2 lists the basic information of these target proteins, including degree, number of domains and number of involving complexes. Figure 6 showed the predicted functions by various methods and the benchmark set. In Figure 6, red elliptic nodes are target proteins, and red edges represent interactions between target proteins. Green round rectangle and gray rectangle nodes represent matched functions and false matched functions, respectively. Solid edges and dash edges between proteins and functions denote correct and false associations. Table 3 showed the description of seven known functions of the three selected proteins. Take the protein YBR278W as an example, which does not contain any domains. For the three domain-based methods Zhang, PON and *GrAPFI*, no one function was inferred, let alone matched a known function. DCS and DSCP generated two predicted functions with one function matched by including neighbors or complex members for calculation of domain context similarities. The NC method annotated the protein YBR278W with functions of its all neighbors. Although the method successfully matched five functions, it introduced a large number of false-positive functions. Out

of seven functions predicted by NPF, five functions are matched with known functions. This is due to the fact that we discovered more partners with similar functions through network propagation and got rid of some functionally unrelated proteins by detecting functional modules with high coupling. The example exhibits the highest predicting accuracy of NPF, compared to the results archived by other competing methods.

Ten-fold cross validation

In the previous section, we applied the leave-one-out cross-validation to exhibit the NPF's improvement on function prediction compared to the state-of-the-art methods. However, in real-world applications, there are usually much more unknown proteins than just one. To do this we adopted the ten-fold validation to verify the validity of our method on PINs with less function information. The entire set of proteins was divided into ten equal sets randomly, nine of which were used for training and the remaining part was used for testing. We ran the functional annotation methods of NPF, Zhang, DCS, DSCP, PON and *GrAPFI* on PINs to get average values of precision, recall and F-measure, as shown in Table 4. Additionally, predicted functions were ranked in descending order according to the values obtained by various method and the top K functions were selected to annotate target proteins. A more valuable comparison between these methods was presented by plotting PR curves and F-measure curves as the parameter K varies using the ten-fold validation. Figure 7 and 8 illustrated the PR curves and F-measure curves of various methods, respectively. Table 4, Figure 7 and 8 exhibited the performance improvement of NPF compared to six other competing methods. Therefore, NPF seemed to be an effective method for characterizing unknown proteins.

Discussions

The accurate annotation of protein functions is the key to understanding life at the molecular level and plays an important role in disease treatment, new drug development. Limited by the quality of protein interaction data generated by high-throughput technologies, network-based methods are progressing slowly and have not produced satisfactory results. A popular optimization scheme for the problem is to infer protein functions by combining PINs with multiple biological data. Despite the advances in these methods, designing efficient algorithms to fuse these multi-source biological data remains challenging. Additionally, the topology of the PINs, such as the “small world”, is also one of the factors that affect the prediction performances. Here, we presented the NPF, a network propagation-based method to annotate functions for target proteins. To overcome the problem of incomplete and false interaction data, we constructed a propagation network by integrating knowledge of the protein interaction network architecture, protein-domain associations and protein-complex associations. By propagating functional similarities across the networks, we can obtain more functionally relevant interacting partners to characterize the target proteins, which effectively free from the constraints of the "small-world" characteristic. Additionally, we take out those redundant function-independent partners by forming functional modules with high cohesion. Comprehensive comparisons among the state-of-the-art methods and our method have been made in terms of the leave-one-out cross-validation and the ten-fold cross validation. Experimental results demonstrated that our method outperforms other competing methods. Based on these results, we can conclude that the network propagation is useful for the study of protein interaction networks.

Conclusions

In this study, we proposed a novel protein functions annotation method based on network propagation, named NPF, which incorporates the topology of PINS and multiple biological data, such as domain annotation information, protein complexes information. Furthermore, we guarantee the NPF against false functions by detecting functional modules based on the neighbour fitness. Experimental comparison results between NPF and six state-of-the-art methods on yeast PINs showed that NPF significantly outperforms other competing methods. In our future study, we will take the hierarchical structure of GO Terms into account for further improvement of the performance of function prediction.

Declarations

Funding

This work was supported in part by the National Natural Science Foundation of China (61772089, 61873221, 61672447), Natural Science Foundation of Hunan Province (No. 2019JJ40325, No. 2018JJ3566, No. 2018JJ3565, No. 2018JJ4058), National Scientific Research Foundation of Hunan Province (19A048), Major Scientific and Technological Projects for collaborative prevention and control of birth defects in Hunan Province (2019SK1010), Hunan Provincial Key Laboratory of Industrial Internet Technology and Security (2019TP1011), and Hunan Provincial Key Laboratory of Nutrition and Quality Control of Aquatic Animals (2018TP1027).

Availability of data and materials

Publicly available datasets were analyzed in this study. This data and the NGF program can be found here: https://github.com/husaiccsu/NPF.

Authors' contributions

BHZ, ZHZ and LW obtained the protein-protein interaction data, domain data, and the protein complexes information. BHZ, ZHZ and LW designed the new method, NPF, and analysed the results. BHZ and ZHZ drafted the manuscript together. MPJ, SH and YCL participated in revising the draft. All authors have read and approved the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent to publish

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

Not applicable

References

1. Schwikowski B, Uetz P, Fields S: **A network of protein–protein interactions in yeast.** *Nature biotechnology*. 2000, **18**(12): 1257-1261.
2. Bogdanov P, Singh A K: **Molecular function prediction using neighborhood features**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2009, **7**(2): 208-217.
3. Cho Y R, Zhang A.: **Predicting protein function by frequent functional association pattern mining in protein interaction networks**. *IEEE Transactions on information technology in biomedicine*, 2009, **14**(1): 30-36.
4. Vazquez A, Flammini A, Maritan A, et al.: **Global protein function prediction from protein-protein interaction networks**. *Nature biotechnology*, 2003, **21**(6): 697-700.

5. Nabieva E, Jim K, Agarwal A, et al.: **Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps**. *Bioinformatics*, 2005, **21**(suppl_1): i302-i310.

6. Bader J S, Chaudhuri A, Rothberg J M, et al. **Gaining confidence in high-throughput protein interaction networks**. *Nature biotechnology*, 2004, **22**(1): 78-85.

7. Cozzetto D, Buchan D W A, Bryson K, et al: **Protein function prediction by massive integration of evolutionary analyses and multiple data sources**. *BMC bioinformatics*, 2013, **14**(Suppl 3): S1.

8. Zhang S, Chen H, Liu K, et al. **Inferring protein function by domain context similarities in protein-protein interaction networks**. *BMC bioinformatics*, 2009, **10**(1): 395.

9. Peng W, Wang J, Cai J, et al: **Improving protein function prediction using domain and protein complexes in PPI networks.** *BMC systems biology*, 2014, **8**(1): 35.

10. Liang S, Zheng D, Standley D M, et al: **A novel function prediction approach using protein overlap networks**. *BMC systems biology*, 2013, **7**(1): 61.

11. Sarker B, Rtichie D W, Aridhi S. **Exploiting complex protein domain networks for protein function annotation**. *International Conference on Complex Networks and their Applications*. Springer, Cham, 2018: 598-610.

12. Piovesan D, Giollo M, Leonardi E, et al. **INGA: protein function prediction combining interaction networks, domain assignments and sequence similarity**. *Nucleic acids research*, 2015, **43**(W1): W134-W140.

13. Piovesan D, Tosatto S C E. **INGA 2.0: improving protein function prediction for the dark proteome**. *Nucleic acids research*, 2019, **47**(W1): W373-W378.

14. Törönen P, Medlar A, Holm L. **PANNZER2: a rapid functional annotation web server**. *Nucleic acids research*, 2018, **46**(W1): W84-W88.

15. Martiniano H F M C, Asif M, Vicente A M, et al. **Network Propagation-Based Semi-supervised Identification of Genes Associated with Autism Spectrum Disorder**. *International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics*. Springer, Cham, 2018: 239-248.

16. Yang K, Liu G, Wang N, et al. **Heterogeneous network propagation for herb target identification**. *BMC medical informatics and decision making*, 2018, **18**(1): 17.

17. Li H, Li T, Quang D, et al. **Network propagation predicts drug synergy in cancers**. *Cancer research*, 2018, **78**(18): 5446-5457.

18. Zhang W, Ma J, Ideker T. **Classifying tumors by supervised network propagation**. *Bioinformatics*, 2018, **34**(13): i484-i493.

19. Hofree M, Shen J P, Carter H, et al. **Network-based stratification of tumor mutations**. *Nature methods*, 2013, **10**(11): 1108-1115.

20. Gottlieb A, Magger O, Berman I, et al. **PRINCIPLE: a tool for associating genes with diseases via network propagation**. *Bioinformatics*, 2011, **27**(23): 3325-3326.

21. Qian Y, Besenbacher S, Mailund T, et al. **Identifying disease associated genes by network propagation**. *BMC Systems Biology*. BioMed Central, 2014, **8**(S1): S6.

22. Huang Y F, Yeh H Y, Soo V W. **Inferring drug-disease associations from integration of chemical, genomic and phenotype data using network propagation**. *BMC medical genomics*, 2013, **6**(S3): S4.

23. Cowen L, Ideker T, Raphael B J, et al. **Network propagation: a universal amplifier of genetic associations**. *Nature Reviews Genetics*, 2017, **18**(9): 551.

24. Liang L, Chen V, Zhu K, et al. **Integrating data and knowledge to identify functional modules of genes: a multilayer approach**. *BMC bioinformatics*, 2019, **20**(1): 225.

25. Stuart J M, Segal E, Koller D, et al. **A gene-coexpression network for global discovery of conserved genetic modules**. *Science*, 2003, **302**(5643): 249-255.

26. Hwang T H, Sicotte H, Tian Z, et al. **Robust and efficient identification of biomarkers by classifying features on graphs**. *Bioinformatics*, 2008, **24**(18): 2023-2029.

27. Vanunu O, Magger O, Ruppin E, et al. **Associating genes and protein complexes with disease via network propagation**. *PLoS computational biology*, 2010, **6**(1).

28. Hartwell L H, Hopfield J J, Leibler S, et al. **From molecular to modular cell biology**. **Nature**, 1999, **402**(6761): C47-C52.

29. Lancichinetti A, Fortunato S, Kertész J. **Detecting the overlapping and hierarchical community structure in complex networks**. *New journal of physics*, 2009, **11**(3): 033015.

30. Xenarios I, Salwinski L, Duan X J, et al. **DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions**. *Nucleic acids research*, 2002, **30**(1): 303-305.

31. Ashburner M, Ball C A, Blake J A, et al. **Gene Ontology: tool for the unification of biology.** *Nature genetics*, 2000, **25**(1): 25-29.

32. Bateman A, Coin L, Durbin R, et al. **The Pfam protein families database**. *Nucleic acids research*, 2004, 32(suppl 1): D138-D141.

33. Pu S, Wong J, Turner B et al.: **Up-to-date catalogues of yeast protein complexes.** *Nucleic Acids Res* 2009, **37**: 825-831.

34. Moreno-Torres J G, Sáez J A, Herrera F. **Study on the impact of partition-induced dataset shift on *k*-fold Cross-validation.** IEEE Transactions on Neural Networks and Learning Systems, 2012, 23(8): 1304-1312.

Figure Legends

Figure 1 - Distribution of shortest-path distances in the pairs of proteins sharing functions

This figure illustrates the relationship between pairs of proteins with common functions and the distance between them. The curve is plotted by short-path distances between proteins on the horizontal axis and the number of pairs of co-annotation proteins on the vertical.

Figure 2 - Flowchart of NPF method

(A) Three protein correlation networks Co_Neighbor, Co-Domain and Co-Complex are derived from original PIN, protein domain data, as well as protein complex information, respectively. (B) The propagation network *PN* is generated by running an improved random walk with restart algorithm on multiple functional similarity networks. The propagation process is illustrated at different steps until convergence. Changes in the color of nodes in the graph indicate the progress of the iterative process. (C) Annotation for target proteins. Taking the target node as the seed node, a highly cohesive functional module can be obtained. Functions of neighbors in the detected functional module are used to characterize the target node.

Figure 3 - Overall comparisons of various methods

Numbers of each bar are the values for each score, including precision, recall and F-measure.

Figure 4 - The precision-recall curves of NPF compared to six other competing methods

The figure denotes the precision-recall (PR) curves of NPF and six other competing methods (Zhang, DCS, DSCP, PON and *GrAPFI*) based on the average prediction performance over all testing protein. The vertical and horizontal coordination of the PR curves are the values of Precision and Recall, respectively. Numbers in brackets represent the maximum F-measures for these seven methods.

Figure 5 - The FP/TP curves of various methods

This Figure depicts the FP/TP of our method and other competing methods fluctuate under various value of the parameter K. The vertical and horizontal coordination of the curve are the values of FP/TP and K, respectively.

Figure 6 - Functions of three selected proteins predicted by various methods

Red elliptic nodes denote target proteins, and red edges represent interactions between them. Green round rectangle and gray rectangle nodes represent matched functions and false matched functions respectively. Solid edges and dash edges between proteins and functions denote correct and false associations. (a) Benchmark results (b)-(h) is the result generated by NPF, Zhang, DCS, DSCP, PON and *GrAPFI*, respectively.

Figure 7 - The precision-recall curves of various methods using ten-fold validation

This Figure shows the PR curves of NPF and six other methods using ten-fold validation. The entire set of proteins is divided into ten equal sets randomly, nine of which are used for training and the remaining part is used for testing. The process is repeated ten times, each time using another testing set.

Figure 8 - The F-measure curves of various methods using ten-fold validation
This Figure depicts the F-measure of seven methods fluctuate under various value of the parameter K. The vertical and horizontal coordination of the curve are the values of F-measure and K, respectively.

Table Legends

Table 1 - Basic information of prediction by various algorithms
This table shows the basic information of the results predicted by NPF, NC, Zhang, DCS, DSCP, PON and *GrAPFI*. MP is the number of proteins successfully matching at least one known function. PMP represents the number of proteins perfectly matching the known functions. MMP denotes the number of proteins completely mismatching the known functions. ZP is the number of proteins with zero-error prediction. That is, all the predicted functions in these proteins match the known functions.

Table 2 - Basic information of selected target proteins
This table shows the basic information of three target proteins. The second column represents the number of its direct neighbors in the original PINs, while the third column is the number of domains it contains. The last column denotes the number of complexes involved.

Table 3 - Description of selected GO Terms
The underscored text represents the name of GO Term.

Table 4 – The prediction results using ten-fold validation

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Tables.pdf