A Tensor-based Bi-random Walks Model for Protein Function Prediction

Sai Hu

Changsha University

Zhihong Zhang

Changsha University

Huijun Xiong

Changsha University

Meiping Jiang

Hunan Provincial Maternal and Child Health Care Hospital

Yingchun Luo

Hunan Provincial Maternal and Child Health Care Hospital

Wei Yan

Changsha University

Bihai Zhao (✉ bihaizhao@163.com)

Changsha University

Research Article

Keywords: Tensor, Protein-protein interaction, protein function, bi-random walks

Posted Date: January 31st, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1294398/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

A tensor-based bi-random walks model for protein function prediction

Sai Hu^{1†}, Zhihong Zhang^{1,3†}, Huijun Xiong¹, Meiping Jiang^{2,4}, Yingchun Luo^{2,4},
Wei Yan¹, Bihai Zhao^{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

[†]These authors contributed equally to this work

*Corresponding Author, Bihai Zhao (email: bihaizhao@163.com)

Email addresses:

SH: husaiccsu@163.com

ZHZ: zhzhang@ccsu.edu.cn

HJX: xionghuijunccsu@163.com

MPJ: meipingjiang123@163.com

YCL: 807859088@qq.com

WY: 1150273378@qq.com

BHZ: bihaizhao@163.com

Abstract

Background:

The accurate characterization of protein functions is key to understanding life at the molecular level and has a huge impact on biomedicine and pharmaceuticals. Computationally predicting protein functions has been studied in the past decades. Plagued by noise and errors in protein-protein interaction networks, researchers have begun to focus on the fusion of multi-omics data in recent years. A data model that appropriately integrates network topologies with biological data and preserves their intrinsic characteristics is still a bottleneck and an aspirational goal for function prediction.

Results

We propose RWRT to accomplish protein function prediction by applying bi-random walks on the tensor. RWRT firstly constructs a functional similarity tensor by combining protein interaction networks with multi-omics data derived from domain annotation and protein complex information. After this, RWRT extends the bi-random walk algorithm from a two-dimensional matrix to the tensor for scoring functional similarity between proteins. Finally, RWRT filters out possible pretenders based on the concept of cohesiveness coefficient and annotated target proteins with functions of the remaining functional partners. Experimental results indicate that RWRT performs significantly better than the state-of-the-art methods and improves the area under the receiver-operating curve (AUROC) by no less than 18%.

Conclusions:

Functional similarity tensor offers us an alternative, in that it is a collection of networks sharing the same nodes; however, the edges belong to different categories or represent interactions of different natures. We demonstrate that tensor-based random

walk model can not only discover more partners with similar functions, but also effectively free from the constraints of errors in 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: Tensor, Protein-protein interaction, protein function, bi-random walks

Background

As the major components of cells, proteins play important roles in almost all cell functions. Biological functions are not performed by a single protein but by a group of interacting proteins shared with the same or similar functions. Accurate annotation of protein functions is key to understand life at the molecular level, which has far-reaching influences for biomedicine and pharmacy. Therefore, how to accurately determine functions of unknown proteins has become one of the greatest challenges of the post-genomic era. Unfortunately, biological experiments have been unable to meet the need of functional annotation of the growing sequence data hampered by their high costs and inherent difficulties. To solve this dilemma, a number of computational methods have been proposed, which implement functional annotation by discovering interacting proteins with known functions in biological networks. High throughput techniques provided high-quality and large-scale protein-protein interaction data and resulted in the emergence of network-based method [1-3] to predict protein functions. The graph-theoretic algorithm [4, 5] was also applied to annotate functions, due to the nature of PPI networks that can be modelled as graphs. However, incompleteness and errors contained in the PPI network, as well as the low coverage of protein interaction data in most species limited the performance of these approaches mentioned above.

Considering the diversity, systematicness and dynamics of protein functions, as well as the poor quality of PPI networks, a variety of approaches have been proposed to promote the prediction of protein functions by incorporating multi-source biological information. The typical processing mode of these methods is to integrate and represent functionally associated attributes of proteins in the form of biological network, and then carry out functions annotation using diffusion algorithm [6], clustering algorithm [7]. Liang et al. constructed the Protein Overlap Network (PON) [8] for functions annotation based on the observation that two proteins are likely to perform the same or similar functions if they share same domain compositions. Sarker et al. proposed a method named *GrAPFI* [9], which reconstructed a protein-protein network based on the network topology and protein domain information, and used the label propagation algorithm to annotate functions for unknown proteins. Peng et al. constructed three biological networks: protein interaction network, domain co-occurrence network and functional interrelationship network, performed function prediction by using unbalanced random walk algorithm in these networks [10]. In our previous studies [11], we designed a dynamic network model for functions annotation by integrating PPI networks, gene expression profile and proteins domain information. Another common used processing way of these methods is to seek the most functional similar partners for unknown proteins based in the context of protein interaction networks. Zhang et al. [12] annotated unknown proteins with all functions of the neighbor which has the highest domain context similarity in the PPI network. On this basis, Peng et al. optimized calculation of domain context similarity by supplementing the domain compositions of proteins themselves and proposed DCS (Domain Combination Similarity) [13] method. Moreover, they designed another protein function similarity measure DSCP (Domain combination Similarity in Context of

Protein Complex) according to the domain compositions of both proteins and complexes including them. Rehman et al. developed the FP (Functional Potential) [14] method to calculate the similarity between interacting proteins based on motif similarity, homology similarity and sequence similarity. Piovesan et al. proposed function prediction methods named INGA [15] and INGA 2.0 [16], which integrated sequence similarity, domain architecture search and PPI networks. After comparing ligand similarity, sequence similarity and functional genomic similarity of proteins, O'Meara et al [17]. found that ligand similarity and functional genomic similarity are complementary for protein function prediction. Stavros et al. proposed a new co-expression measure MLC (Metric Learning for Co-expression) [18] instead of the Pearson correlation to assigns a GO term-specific weight to each expression sample for gene function prediction. Gligorijević et al. designed a novel graph convolution network-based protein function prediction method DeepFRI [19], which extracted sequence features from protein language models and protein structures.

These methods attempted to improve the quality of PPI networks by assigning different weights to edges at different levels corresponding to multi-omics data. They classically aggregated multiple biological data into a composite and reliable network, which tends to eliminate the topologies and attributions of the individual protein interaction networks. Researches and experimental results indicate that each type of biological data has its own property or correlation and can play a different role in prediction of protein functions. The way of representing different types of data source in a system with a single type of link is no longer a magic cure-all for network science. In this context, the very pressing need for protein function prediction is to find a suitable data model. Intuitively, a proper data model describing function relevance in multi-omics data should satisfy two properties: it should not only be able to describe

the hierarchy and heterogeneity of biological network, but also reflect the complex relationship between multi-omics data, and it should be supported by diversified solutions and rigorous theoretical system, which is conducive to generalization to other research fields. We formalize these two properties with a multidimensional tensor model integrating the topology of PPI networks with multiple biological data and develop the RWRT (Random Walks with Restart on Tensor) method to infer protein functions. The RWRT method not only improves its performance but also preserves the functional relevance between multi-omics data and their own attributes. We apply the RWRT to the yeast protein interaction network and combine it with multiple biological data, including protein complexes and domain information. Experimental results show that our proposed RWRT method outperformed six types of methods, including NC [1], Zhang [12], DCS [13], DSCP [13], PON [8] and *GrAPFI* [9].

Methods

Our RWRT method consists of three major stages. First, integrating the topology of PPI networks, protein domains, and protein complexes information, a functional similarity tensor T is constructed. The similarity tensor preserves and reflects multiple relationships between proteins derived from multi-omics data. In the second stage, an iterative procedure calculates functional correlation scores for protein pairs in the network. The iterative procedure is an extension of bi-random walks algorithm on the tensor model, which simulates a high-order Markov chain by means of two state transition tensors. In the third stage, scoring and sorting all functions of their “similar” partners (neighbors), target proteins are annotated by top K of these predicted functions. The flowchart for the RWRT approach is provided in Figure 1.

Construction of the functional similarity tensor T

The success of functional genomics is associated with the rapid accumulation of various biological data about genes, proteins or other macromolecules [20]. We now have access to multiple types of physical or functional interactions between proteins. These different interactions with their own features are better represented as a multi-graph framework. Figure 2 (a) shows an example of a multi-graph with five nodes and three types of edges. The multi-graph can also be represented as a tensor, shown in Figure 2 (b). The tensor, as an extension of a matrix in high-order space, has many advantages for the representation and processing of complex relationships between proteins or genes [21]. For our purpose, we construct a functional similarity tensor $T \in \mathbb{R}^{n \times n \times m}$, where n and m represents the number of proteins and types of connections between proteins, respectively. If there is a k -th type of interaction between two proteins i and j , then $t(i, j, k) \in T > 0$, otherwise $t(i, j, k) \in T = 0$. In this paper, we consider three types of physical or functional associations between proteins, including physical associations based on the topology of PPIs, co-structure associations derived from domain contexts and co-module associations coming from protein complex information.

The first type of association is guided by the ‘Guilt by Association’ principle. Given two proteins p_i, p_j that share common interacting partners, their functional similarity is estimated as follow:

$$t(i, j, 1) = \frac{4 |N_{p_i} \cap N_{p_j}|^2}{(|N_{p_i}| + |N_{p_i} \cap N_{p_j}|) \times (|N_{p_j}| + |N_{p_i} \cap N_{p_j}|)} \quad (1)$$

where N_{p_i} and N_{p_j} represent a set that include p_i and p_j themselves and their direct neighbors, respectively.

Research shows that certain sets of domains are frequently found together and cooperate to perform cellular functions [22]. Based on the observation, we established co-structure associations between proteins using domain content similarity. For any two proteins p_i and p_j , let D_i represents the set of distinct domain types in the neighbors of p_i with itself included, and D_j is that of p_j as well of its neighbors. DC denotes the set of distinct domain types contained by both neighbors of p_i and neighbors of p_j , while DT is the set of domain types in the whole PPI networks. Then the domain content similarity between p_i and p_j is calculated using the following equations:

$$DS(p_i, p_j) = -\log \frac{\binom{|DT|}{|DC|} \binom{|DT| - |DC|}{|D_i| - |DC|} \binom{|DT| - |D_i|}{|D_j| - |DC|}}{\binom{|DT|}{|D_i|} \binom{|DT|}{|D_j|}} \quad (2)$$

Figure 3 shows an example of calculating the domain content similarity of two proteins. Protein A and its four neighbors contain five different domains, while protein B involves three types of domains along with its two neighbors. The functional similarity of co-structure associations between p_i and p_j is measured by the normalization processing of their domain content similarity, which is formally described as follows:

$$t(i, j, 2) = \frac{DS(p_i, p_j) - \min_{1 \leq i \leq n, 1 \leq j \leq n} (DS(p_i, p_j))}{\max_{1 \leq i \leq n, 1 \leq j \leq n} (DS(p_i, p_j)) - \min_{1 \leq i \leq n, 1 \leq j \leq n} (DS(p_i, p_j))} \quad (3)$$

Most cellular functions are carried out through interactions between multiple functional modules at various levels [23]. It seems clear that the functional module or protein complex information is important for protein function prediction. If two proteins participate in the same module, they are likely to perform the same or similar

functions. Firstly, we calculate the density score of experimental detection functional modules, which can be expressed as:

$$Score(M_k) = \frac{2 \times |E_k|}{|V_k| \times (|V_k| - 1)} \quad (4)$$

where E_k and V_k denotes the set of physical interactions and proteins of the module M_k , respectively. After getting the scores of benchmark functional modules obtained from Equation (4), we can evaluate the reliability of co-module associations between proteins. For any two proteins p_i and p_j in networks, their functional similarity of their co-module associations is calculated as follow:

$$t(i, j, 3) = \frac{(\sum_{k=1, p_i \in M_k, p_j \in M_k}^{|M|} Score(M_k))^2}{\sum_{k=1, p_i \in M_k}^{|M|} Score(M_k) \times \sum_{k=1, p_j \in M_k}^{|M|} Score(M_k)} \quad (5)$$

Random walk with restart on the tensor T

In this paper, multiple functional associations are introduced in the constructed tensor T . Therefore, the significance of proteins is taken into account in the iterative process as well as different type of interactions. Furthermore, our model is based two hypotheses: proteins with high significance values connect to each other through significant interactions, and interactions with high significance values frequently joined between many significant proteins. In this stage, our model performs iteration on the tensor T for a target protein to obtain its functional association scores with other annotated proteins. Given a target protein p_i , $X_i = [x_{i1}, x_{i2}, \dots, x_{in}]^T \in \mathbb{R}^n$ and $Y_i = [y_{i1}, y_{i2}, \dots, y_{in}]^T \in \mathbb{R}^n$ is the vector representing its functional similar scores with known proteins and significance scores of different categories of interactions between proteins, respectively. We can thus extend the classical random walk with restart

algorithm to the tensor model. Our method performs a two-step iteration operation to update significance scores of proteins and different types of interactions by

$$\begin{aligned} X_i^{t+1} &= \alpha \mathcal{T} X_i^t Y_i^t + (1-\alpha) X_i^0 \\ Y_i^{t+1} &= \mathcal{T}^o X_i^t X_i^{t+1} \end{aligned} \quad (6)$$

where $\mathcal{T}$ and $\mathcal{T}^o$ are got from the tensor T constructed in the first stage by normalizing so that entries in each row sum to 1, and they are calculated as follow:

$$\mathcal{T}_{i,j,k} = \begin{cases} \frac{t_{i,j,k}}{\sum_{i=1}^n t_{i,j,k}} & \text{if } \sum_{i=1}^n t_{i,j,k} > 0 \\ 1/n & \text{otherwise} \end{cases} \quad (7)$$

$$\mathcal{T}_{i,j,k}^o = \begin{cases} \frac{t_{i,j,k}}{\sum_{k=1}^m t_{i,j,k}} & \text{if } \sum_{k=1}^m t_{i,j,k} > 0 \\ 1/m & \text{otherwise} \end{cases} \quad (8)$$

, and the parameter $\alpha \in [0,1]$ balances between the iteration information and initial significant scores, which is set to 0.5 [24, 25]. In the equation (6), $X_i^0 = [x_{i1}^0, x_{i2}^0, \dots, x_{in}^0]^T \in \mathbb{R}^n$ is the vector of initial functional similar scores, and its element $x_{i,j}^0$ can be calculated as:

$$x_{i,j}^0 = \frac{|D_i \cap D_j|}{\sqrt{|D_i| \times |D_j|}} + \frac{|C_i \cap C_j|}{\sqrt{|C_i| \times |C_j|}} \quad (9)$$

where D_i and D_j denotes the set of domains contained by protein p_i and p_j , respectively. C_i and C_j represents the set of protein complexes in which p_i and p_j is involved, respectively. Following iterations for all proteins, we obtain a functional similarity matrix M_{fs} , which is formally described as follows:

$$M_{fs} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nn} \end{bmatrix} \quad (10)$$

The proof of the convergence for the random walk on tensor algorithm is related to our previous work [26].

Computational annotation of proteins with unknown functions

Benefiting from the iteration on the constructed tensor, we are able to discovery more partners with similar functions to target proteins, which are ignored by PPI networks. These partners as well as the target protein interact with each other to perform biological functions within multiple functional modules. Intuitively, members within the same functional module are often more densely connected than those across functional modules [27]. Unfortunately, some of these partners are pretenders who have closer connections to the outside of the module than to the inside. Therefore, those pretenders should be removed from the functional module. We introduce the concept of cohesiveness coefficient (CC) to determine whether a partner is false. Let $fs^{in}(p_i)$ denote the total functional similarity scores between all other proteins inside the functional module and the protein p_i , and let $fs^{out}(p_i)$ denote the total functional similarity scores between all proteins outside the functional module and the protein p_i . The cohesiveness coefficient of p_i is then given by

$$CC(p_i) = \frac{fs^{in}(p_i)}{fs^{in}(p_i) + fs^{out}(p_i)} \quad (11)$$

where $fs^{in}(p_i) = \sum_{p_j \in M} x_{ij}$, $fs^{out}(p_i) = \sum_{p_k \notin M} x_{ik}$, and M is a set of partners that have same

or similar functions as the target protein. Cohesiveness coefficient provides an efficient way to assess whether a functional module satisfies the two hypotheses mentioned on the second stage. A well-separated module consisted of many proteins with similar functional has a high fs^{in} and a low fs^{out} . For a partner with similar functions to the target protein, its $CC \leq 1/3$ [28] implies that it has more external functional similarity than internal functional similarity and should be removed from

the partners group. After the cohesiveness-filter processing with Equation (11), we are able to get a functional similarity proteins set $FSP = \{fsp_1, fsp_2, \dots, fsp_l\}$. Let $F = \{f_1, f_2, \dots, f_m\}$ be a list of distinct functions of proteins in FSP . We score and rank functions within F in descending order to annotate the target protein with top K of them. Given a function f_i for the target protein p_t , its ranking score is calculated by the following formula:

$$RS(f_i) = \sum_{k=1}^l x_{p_t, fsp_k}, f_i \in GO(fsp_k) \quad (12)$$

Where x_{p_t, fsp_k} is the functional similarity score between partner fsp_k and the target protein p_t , $GO(fsp_k)$ is a set of functions belonging to fsp_k . In this paper, the parameter K was assigned as the number of functions of the protein within FSP , which had the highest functional similarity score to the target protein.

Results

Experimental data

To estimate the performance of RWRT for function prediction, we perform computational analysis on our method as well as other six competing algorithms, such as NC, Zhang, DCS, DSCP, PON and *GrAPFI* on two benchmark datasets from yeast, including DIP [29] and BioGRID [30]. The DIP dataset and BioGRID dataset updated to February 5, 2017 and Oct.28, 2017, respectively. The DIP dataset consists of 4,912 proteins and 22,129 interactions among the proteins, and the BioGRID dataset consists of 4,114 proteins and 26,112 physical interactions. Self-interactions and repeated interactions are removed from the two benchmark datasets. The annotation data of proteins used for validation is downloaded from GO official website [31]. We primarily investigate and analyze the annotations in the Biological Process (BP) category in this manuscript. To avoid too special and too general, only those GO

terms that annotate at least 10 and at most 200 proteins will be kept in our experiments [12]. The protein domain data is retrieved from the PFAM database [32], which involves 1064 and 1026 distinct domain types related to 2945 and 2566 proteins of the DIP dataset and BioGRID dataset, respectively. The experimental detection functional module set comes from CYC2008 [33], which consists of 408 modules referring to 1465 and 1600 proteins in DIP and BioGRID, respectively.

Assessment criteria

To evaluate the effectiveness of RWRT and other six competing methods in protein function prediction, we applied cross validation using two different strategies to split proteins of PPI networks into the training set and the testing set. The two strategies are leave-one-out cross validation and ten-fold cross validation. To measure the quality of predicted functions by each method, we use two assessment criteria: AUROC (area under the receiver-operating curve) and average F-score. AUROC is widely used in performance evaluation for protein function prediction. As the harmonic mean of Precision and Recall, F-score is another measure to evaluate the performance of a method synthetically.

Leave-one-out cross-validation

In our first set of evaluations, we apply leave-one-out cross validation to assess the quality of predicted functions predicted by RWRT, as well as other six competing methods: NC, Zhang, DCS, DSCP, PON and *GrAPFI*. The performance is averaged with only one protein into the testing set and rest of proteins used as the training set. We first evaluate the performance of RWRT and six competing methods on these target proteins by the average Precision, TPR (True Positive Rate), FPR (False Positive Rate) and F-Score. Table 1 lists the prediction results of RWRT and other competing methods. RWRT achieves the highest average Precision, TPR and F-Score

values, and the lowest FPR values among the seven methods. Take the DIP dataset as an example, the average F-Score of RWRT is 108.96%, 137.29%, 42.86%, 19.32%, 220.61% and 94.44% higher than NC, Zhang, DCS, DSCP, PON and *GrAPFI*, respectively. For comprehensive performance comparison between RWRT and competing methods, we employed piecewise statistics of the predicted results according to functional size of target proteins. The results, shown in Figure 4 and Figure 5, suggest that these methods get different performance for different size of proteins, each with its own unique advantages. The performance of our RWRT had obvious advantages with size falls into [2, 8], while the prediction accuracy of DSCP on BioGRID data sets dropped sharply when size is in [6, 9]. Note that only four proteins have ten or more functional annotations. Therefore, the results of these methods in [10, 14] are not statistically significant and are not included in this analysis.

To further investigate the performance of RWRT and six competing methods, we apply the ROC curve, whose vertical and horizontal coordination are the values of TPR and FPR, respectively. Figure 6 (a) and Figure 6 (b) depicts the ROC curve of seven methods on the BioGRID dataset and DIP dataset, respectively. For intuitive evaluation of the performance of various methods, we calculate the area under all curves and list the results in Table 2. The AUROC of RWRT on BiosGRID is 18.43%, 102.36%, 51.18%, 47.70%, 576.32% and 283.58% higher than that of NC, Zhang, DCS, DSCP, PON and *GrAPFI*, respectively. As for the DIP dataset, the AUROC of RWRT increases by no less than 60% compared with other competing methods. Comparison results also reveal a phenomenon that the performance of these methods in the DIP dataset is significantly inferior to that of them on the BioGRID dataset. For NC, Zhang, DCS and DSCP, which mainly depend on neighbors, the gap is even

more obvious. In our opinion, it is due to the fact that the DIP network is sparser than the BioGRID network. However, this does not seem to have much effect on the RWRT method. Experimental comparison results strongly prove the effectiveness and robustness of our method.

Ten-fold cross validation

For comprehensive comparison between the novel method RWRT and the six other competing methods, we adopted the ten-fold cross validation to examine the stability of these methods on the training dataset. Proteins are randomly divided into ten subsets, a single subset is retained for the testing set, and the remaining nine subsets are used as the training set. The cross validation process is then repeated ten rounds, each of which uses different folds as the training and validation data. The ten results from the folds are then averaged to produce the final performance. We run the functional annotation methods of RWRT as well as other six competing methods on the BioGRID and DIP networks. Figure 7 presents the average Precision, TPR, FPR and F-score of seven methods on two datasets. Additionally, we drew the ROC curves of all methods on the two PPI networks, which are shown in Figure 8. The AUROC of RWRT on BiosGRID is 19.38%, 604.49%, 271.73%, 96.42%, 51.75% and 48.28% higher than that of NC, PON, *GrAPFI*, Zhang, DCS and DSCP, respectively. On the DIP dataset, AUROC of RWRT increases by 110.79%, 559.46%, 241.54%, 322.87%, 103.87% and 61.80% , respectively, compared to the above six methods. All of these experimental results indicate that RWRT still outperforms other six competing methods using other competing methods

Discussions

Accurate annotation of protein functions is still a big challenge for understanding life in the post-genomic era. In spite of the advances in computational methods for protein

function prediction, it still does not achieve satisfactory prediction accuracy plagued by the incompleteness and errors in the original PPI data. To overcome this problem, the integration of different types of biological data has become an important and popular strategy, which has led to the emergence of various interactions between proteins. Each type of biological data has its own property or correlation and can play a different role in prediction of protein functions. Motivated by it, we set up a multidimensional data model and describe it formally with the tensor. To get rid of constraints of the small-world and scale-free properties of PPI networks, we extended the bi-random walks algorithm to the tensor model. In this way, we can discovery more potential proteins with similar functions to target proteins and improve the true positive rate of prediction. However, the enlargement of the traversal range of similar functional neighbors will inevitably lead to the increase of noise. The NC method is a typical example to illustrate the fact. For this purpose, we run the functional module detection algorithm based on the concept of cohesiveness coefficient to filter out pretenders and reduce the false positive rate of prediction. We are of the opinion that there are both overlaps and differences in functional annotation between proteins. It is not appropriate to annotate a target protein with all the functions of the protein most similar to it.

Conclusions

In this paper, we design a novel protein function prediction method named RWRT by applying a tensor-based bi-random walks model. The RWRT method constructs a functional similarity tensor according to the original PPI network as well as multi-omics data firstly. And then, it extends the random walk with restart algorithm to the tensor by simulating a high-order Markov chain. After this phase, RWRT can discover more functional similarity partners ignored by original protein interactions

data. However, it also inevitably introduces some spurious nodes. Therefore, a pretender-filtering procedure is applied to remove possible pretenders loosely connected to the target protein and finally generate predicted functions from the remaining functional partners. Experimental comparison results on two PPI networks indicate that RWRT performs significantly better than the state-of-the-art methods and the proposed model can provide more insights for future study in PPI networks.

Declarations

Funding

This work was supported in part by the National Natural Science Foundation of China (No. 61772089), Natural Science Foundation of Hunan Province (No. 2020JJ4648), National Scientific Research Foundation of Hunan Province (No. 19A048, 21A0548), 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).

Availability of data and materials

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

Authors' contributions

SH and ZHZ obtained the protein-protein interaction data, domain data, and the protein complexes information. SH, ZHZ and BHZ designed the new method, RWRT, and analysed the results. SH, ZHZ and BHZ drafted the manuscript together. HJX, MPJ, WY 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. Chua H N, Sung W K, Wong L. **Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions.** *Bioinformatics*, 2006, **22**(13): 1623-1630.
3. Moosavi S, Rahgozar M, Rahimi A. **Protein function prediction using neighbor relativity in protein-protein interaction network.** *Computational biology and chemistry*, 2013, **43C**(1):11-16.
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. Cao R, Cheng J. **Integrated protein function prediction by mining function associations, sequences, and protein–protein and gene–gene interaction networks.** *Methods*, 2016, **93**:84-91.

7. Liao B, Li Y, Jiang Y, et al. **Using multi-instance hierarchical clustering learning system to predict yeast gene function.** *PloS one*, 2014, **9**(3): e90962.
8. 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.
9. 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.
10. Wei P , Min L , Lu C , et al. **Predicting Protein Functions by Using Unbalanced Random Walk Algorithm on Three Biological Networks.** *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, 2017, **14**(2):360.
11. Zhao B, Wang J, Li M, et al. **A new method for predicting protein functions from dynamic weighted interactome networks.** *IEEE transactions on nanobioscience*, 2016, **15**(2): 131-139.
12. 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.
13. 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.
14. Rehman H U, Benso A, Di Carlo S, et al. **Combining homolog and motif similarity data with Gene Ontology relationships for protein function prediction.** *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on. IEEE*, 2012: 1-4.

15. 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.
16. 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.
17. O'Meara M J, Ballouz S, Shoichet B K, et al. Ligand Similarity Complements Sequence, Physical Interaction, and Co-Expression for Gene Function Prediction. *PLoS One*, **2016**, 11(7): e0160098.
18. Makrodimitris S, Reinders M J T, Van Ham R C H J. **Metric learning on expression data for gene function prediction.** *Bioinformatics*, 2020, **36**(4): 1182-1190.
19. Gligorijević V, Renfrew P D, Kosciolk T, et al. **Structure-based protein function prediction using graph convolutional networks.** *Nature communications*, 2021, **12**(1): 1-14.
20. Novoa-del-Toro E M, Mezura-Montes E, Vignes M, et al. **A multi-objective genetic algorithm to find active modules in multiplex biological networks.** *PLoS computational biology*, 2021, **17**(8): e1009263.
21. Wang X, Yang L T, Kuang L, et al. **A tensor-based big-data-driven routing recommendation approach for heterogeneous networks.** *IEEE Network*, 2019, **33**(1): 64-69.
22. Forslund K, Sonnhammer E L L. **Predicting protein function from domain content.** *Bioinformatics*, 2008, **24**(15): 1681-1687.

23. Li J, Zhao P X. **Mining functional modules in heterogeneous biological networks using multiplex PageRank approach.** *Frontiers in plant science*, 2016, **7**: 903.
24. Taehyun H , Hugues S , Tian Z , et al. **Robust and efficient identification of biomarkers by classifying features on graphs.** *Bioinformatics*, 2014, **24**(18):2023-2029.
25. Vanunu O, Magger O, Ruppin E, et al. **Associating genes and protein complexes with disease via network propagation.** *PLoS computational biology*, 2010, **6**(1): e1000641.
26. Zhao B , Zhang Z , Jiang M , et al. **NPF:network propagation for protein function prediction.** *BMC Bioinformatics*, 2020, **21**(1): 355.
27. Hartwell L , Hopfield J , Leibler S . **Murray AW: From molecular to modular cell biology.** *Nature*, 1999, **402**(6761).
28. Nepusz T , Yu H , Paccanaro A . **Detecting overlapping protein complexes in protein-protein interaction networks.** *Nature Methods*, 2012, **9**(5):471-472.
29. 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
30. Oughtred R, Stark C , Breitkreutz B J, et al. **The BioGRID interaction database: 2019 update.** *Nucleic acids research*, 2019, **47**(D1), D529-D541.
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.

Figure Legends

Figure 1 - Flowchart of RWRT method

(A) Constructing the functional similarity tensor T according to functional association analysis based on original PPI network, domain context and protein complex information. (B) Given a testing protein p_i , running a bi-random walks algorithm on the tensor T to obtain the functional similarity vector X_i between p_i and the other proteins. (C) Removing redundant partners that have more external functional similarity than internal functional similarity within the module consisted of the target protein and its partners, and picking out top K of all functions of the remaining partners.

Figure 2 - Example of a multi-graph

(a) Showing a multi-graph consisted of five nodes and three types of connections between nodes. (b) Representing the multi-graph with a tensor model formally.

Figure 3 - Illustration of domain content similarity

The figure gives an example of the domain content similarity of protein A and B, in which rectangles in different colors are drawn to represent different types of domains.

$$D_A = \{1, 2, 3, 4, 5\}, D_B = \{5, 6, 7\}, DC = D_A \cap D_B = \{5\}.$$

Figure 4 - The average FPR of seven methods according to different functional size of target proteins.

The figure depicts the average false positive rate of our method and other competing methods fluctuate under different functional size of target proteins. Here functional size means the number of GO terms in each target protein. Size ranges from 1 to 14,

except for 13. (a) Predicting results of seven methods on the BioGRID dataset. (b) Predicting results of seven methods on the DIP dataset.

Figure 5 - The average F-Score of seven methods according to different functional size of target proteins.

The figure shows the average of RWRT and other competing methods fluctuate under different functional size of target proteins. (a) Predicting results of seven methods on the BioGRID dataset. (b) Predicting results of seven methods on the DIP dataset.

Figure 6 - ROC curves of seven methods using leave-one-out cross validation

The figure denotes the ROC (receiver-operating) curves of RWRT and other six 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 ROC curves are the values of true positive rate and false positive rate, respectively. (a) ROC curves of seven methods on the BioGRID dataset. (b) ROC curves of various methods on the DIP dataset.

Figure 7 - The predicted results of various methods using ten-fold validation

Numbers of each bar are the values for each score, including average precision, TPR (True positive rate), FPR (False positive rate) and F-score. (a) Results of seven methods on the BioGRID dataset. (b) Results of various methods on the DIP dataset.

Figure 8 - ROC curves of seven methods using ten-fold cross validation

This Figure shows the ROC curves of RWRT and other six 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 1000 times, each time using another testing set. (a) ROC curves of seven methods on the BioGRID dataset. (b) ROC curves of various methods on the DIP dataset.

Table Legends

Table 1 - The results of RWRT and six competing methods on the DIP and BioGRID dataset

Table 2 - AUROC of RWRT and other competing methods on the DIP and BioGRID dataset

Figures

Figure 1

Flowchart of RWRT method

(A) Constructing the functional similarity tensor T according to functional association analysis based on original PPI network, domain context and protein complex information. (B) Given a testing protein p_i , running a bi-random walks algorithm on the tensor T to obtain the functional similarity vector X_i between p_i and the other proteins. (C) Removing redundant partners that have more external functional similarity than internal functional similarity within the module consisted of the target protein and its partners, and picking out top K of all functions of the remaining partners.

Figure 2

Example of a multi-graph

(a) Showing a multi-graph consisted of five nodes and three types of connections between nodes. (b) Representing the multi-graph with a tensor model formally.

Figure 3

Illustration of domain content similarity

The figure gives an example of the domain content similarity of protein A and B, in which rectangles in different colors are drawn to represent different types of domains. $D_A = \{1, 2, 3, 4, 5\}$, $D_B = \{5, 6, 7\}$, $DC = D_A \cap D_B = \{5\}$.

(a) BioGRID

(b) DIP

Figure 4

The average FPR of seven methods according to different functional size of target proteins.

The figure depicts the average false positive rate of our method and other competing methods fluctuate under different functional size of target proteins. Here functional size means the number of GO terms in each target protein. Size ranges from 1 to 14, except for 13. (a) Predicting results of seven methods on the BioGRID dataset. (b) Predicting results of seven methods on the DIP dataset.

(a) BioGRID

(b) DIP

Figure 5

The average F-Score of seven methods according to different functional size of target proteins.

The figure shows the average of RWRT and other competing methods fluctuate under different functional size of target proteins. (a) Predicting results of seven methods on the BioGRID dataset. (b) Predicting results of seven methods on the DIP dataset.

Figure 6

ROC curves of seven methods using leave-one-out cross validation

The figure denotes the ROC (receiver-operating) curves of RWRT and other six 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 ROC curves are the values of true positive rate and false positive rate, respectively. (a) ROC curves of seven methods on the BioGRID dataset. (b) ROC curves of various methods on the DIP dataset.

Figure 7

The predicted results of various methods using ten-fold validation

Numbers of each bar are the values for each score, including average precision, TPR (True positive rate), FPR (False positive rate) and F-score. (a) Results of seven methods on the BioGRID dataset. (b) Results of

various methods on the DIP dataset.

Figure 8

ROC curves of seven methods using ten-fold cross validation

This Figure shows the ROC curves of RWRT and other six 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 1000 times, each time using another testing set. (a) ROC curves of seven methods on the BioGRID dataset. (b) ROC curves of various methods on the DIP dataset.