https://doi.org/10.1186/s12859-022-05057-3

RESEARCH

Open Access

Protein function annotation based on heterogeneous biological networks

Sai Hu[3†], Yingchun Luo[2,4†], Zhihong Zhang[1,5], Huijun Xiong[3], Wei Yan[1], Meiping Jiang[2,4] and Bihai Zhao[1,5*]

[†]Sai Hu and Yingchun Luo are Equal contributors.

*Correspondence: bihaizhao@163.com

[1] School of Computer Science and Engineering, Changsha University, Changsha 410022, Hunan, China
[2] Department of Ultrasound, Hunan Provincial Maternal and Child Health Care Hospital, Changsha 410008, Hunan, China
[3] School of Mathematics, Changsha University, Changsha 410022, Hunan, China
[4] NHC Key Laboratory of Birth Defect for Research and Prevention, Hunan Provincial Maternal and Child Health Care Hospital, Changsha 410100, Hunan, China
[5] Hunan Provincial Key Laboratory of Industrial Internet Technology and Security, Changsha University, Changsha 410022, Hunan, China

Abstract

Background: Accurate annotation of protein function is the key to understanding life at the molecular level and has great implications for biomedicine and pharmaceuticals. The rapid developments of high-throughput technologies have generated huge amounts of protein–protein interaction (PPI) data, which prompts the emergence of computational methods to determine protein function. Plagued by errors and noises hidden in PPI data, these computational methods have undertaken to focus on the prediction of functions by integrating the topology of protein interaction networks and multi-source biological data. Despite effective improvement of these computational methods, it is still challenging to build a suitable network model for integrating multiplex biological data.

Results: In this paper, we constructed a heterogeneous biological network by initially integrating original protein interaction networks, protein-domain association data and protein complexes. To prove the effectiveness of the heterogeneous biological network, we applied the propagation algorithm on this network, and proposed a novel iterative model, named Propagate on Heterogeneous Biological Networks (PHN) to score and rank functions in descending order from all functional partners, Finally, we picked out top *L* of these predicted functions as candidates to annotate the target protein. Our comprehensive experimental results demonstrated that PHN outperformed seven other competing approaches using cross-validation. Experimental results indicated that PHN performs significantly better than competing methods and improves the Area Under the Receiver-Operating Curve (AUROC) in Biological Process (BP), Molecular Function (MF) and Cellular Components (CC) by no less than 33%, 15% and 28%, respectively.

Conclusions: We demonstrated that integrating multi-source data into a heterogeneous biological network can preserve the complex relationship among multiplex biological data and improve the prediction accuracy of protein function by getting rid of the constraints of errors in PPI networks effectively. PHN, our proposed method, is effective for protein function prediction.

Keywords: Protein function prediction, Heterogeneous biological network, Network propagation

Background

Proteins are the basic organic matter that constitutes all cells and tissues of the living body. Accurately and automatically annotation of protein function is one of the fundamental tasks of bioinformatics, and it has become very hot in recent years. Methods for experimentally determining protein function such as gene expression inhibition [1], targeted mutation [2] and gene knockout [3] require considerable time and cost, and can only deal with one gene or protein at a time. With the increasing number of functional proteins to be labelled, such low-throughput experimental techniques cannot meet practical needs. Consequently, the computational method serves as a more suitable solution for determining protein function.

The rapid developments of high-throughput technologies have generated huge amounts of high-quality, large-scale protein interaction data, which provide fundamental and abundant data for network-based approaches to deduce protein functions. Schwikowski et al. [4] found that proteins interacting with each other generally share the same function, and proposed a method named NC for function prediction based on interacting neighbour voting. Chua et al. [5] proposed a functional similarity measurement method to recalculate the interaction strength of proteins by comprehensively utilizing the global structure of the protein interaction network characterized by direct and indirect neighbours, and improved the NC method on this basis. Since PPI networks can be represented by graph models, graph-theoretic algorithms were naturally applied to protein function prediction as well. Functions were deduced by the global connectivity pattern of the protein physical network, which was determined by minimizing the number of protein interactions between different functional categories [6]. The GLIDER [7] method predicted protein functions from a new graph-based similarity network instead of the PPI network. It can infer missing connections in PPI networks based on local and global graph properties.

Considering the incompleteness of the protein–protein interaction network, researchers combined multiple biological data with the protein interaction network to establish functional similarity networks for function annotation. Through statistical analysis, Liang et al. [8] found that two proteins are likely to perform the same or similar function if they have the same domain composition. Consequently, they constructed the Protein Overlap Network (PON) for protein function annotation. Peng et al. established the protein interaction network, domain co-occurrence network and functional interrelationship network and ran the random walk algorithm on these networks to deduce the function of proteins [9]. Sarker et al. reconstructed the protein interaction network using protein-domain association data and proposed the *GrAPFI* [10, 11] method to predict functions for the target protein by using the label propagation algorithm on this network. Generally speaking, if two genes or genes products have similarities in some context, we can conclude that they have the same or similar annotation terms. Song et al. determined the functions of the unknown protein by exploring its functional partner with the highest domain context similarity derived from their direct neighbours [12]. The DCS (Domain Combination Similarity) [13] calculated domain context similarity by adding domains of the protein itself and improved the performance of prediction of protein function. DeepGOPlus [14] deduced protein functions for the target protein based on the sequence similarity with known functions using deep learning techniques.

Despite effective improvement of these computational methods for function annotation, it is still a challenge to build an appropriate network model for the integration of multiplex biological data and PPI networks. The most prevalent way is to merge multiplex biological data into a single and unique network, in which the role of different types of data is reflected in the form of setting weights or parameters. Therefore, the parameter is an important factor that affects the performance of methods for function prediction, which generally depends on experience or the result of association analysis. Even if the parameter setting model is optimized, different species and even different data sets have different settings. So, how to set the parameter value will become one of the biggest barriers to the application of these prediction models. In addition, the construction of a single network ignores the differences among multiplex biological data and covers up the inherent attributes of different types of biological data. In this paper, we constructed a heterogeneous biological network with the integration of PPI networks and multiple biological data, including protein complexes and protein-domain association data. On this basis, we design a novel protein function prediction method named PHN (Propagate on Heterogeneous Networks) by applying the propagation algorithm [15] on the heterogeneous biological network. To evaluate the performance of PHN, we apply our method on the *Saccharomyces cerevisiae* PPI network. Experimental results show that the PHN method outperforms seven competing methods for prediction of protein function: NC [4], Song [12], DCS [13], DSCP [13], NPF [15], PON [8] and *GrAPFI* [10].

Methods

The outline for the proposed PHN method includes (1) constructing a heterogeneous biological network by integrating the topology of PPI networks, protein-domain association data, and protein complex information, (2) running the propagation algorithm on the heterogeneous biological network to generate a functional similarity partners list for the given target protein, and (3) scoring and ranking functions from the partners list in descending, and picking out top *L* of them to annotate the unknown protein. The flowchart for the PHN method is provided in Fig. 1.

Construction of a heterogeneous biological network

Protein–protein interaction network

It has been observed that more than 70% of proteins perform at least one function with their direct neighbours in networks [16]. We make a statistical analysis of yeast PPI data and observe that 21.3% of proteins share no functions with their direct neighbours, and they display enormous function similarity with some of their level-2 neighbours. In this paper, we evaluate functional similarity between a protein and its neighbours from the two different levels. Given a protein *u*, S_1 and S_2 denote the set of direct neighbours and level-2 neighbours of *u*, respectively. We classify all proteins except *u* into four categories: direct neighbours that are also level-2 neighbours (i.e. $S_1 \cap S_2$), direct neighbours that are not level-2 neighbours (i.e. S_1-S_2), level-2 neighbours that are not direct neighbours (i.e. S_2-S_1) and protein that are not direct neighbours or level-2 neighbours (i.e.$\overline{S_1 \cup S_2}$). In this work, the parameter α ($0 < \alpha < 1$) is adopted to evaluate the importance of direct neighbours in functional analysis based

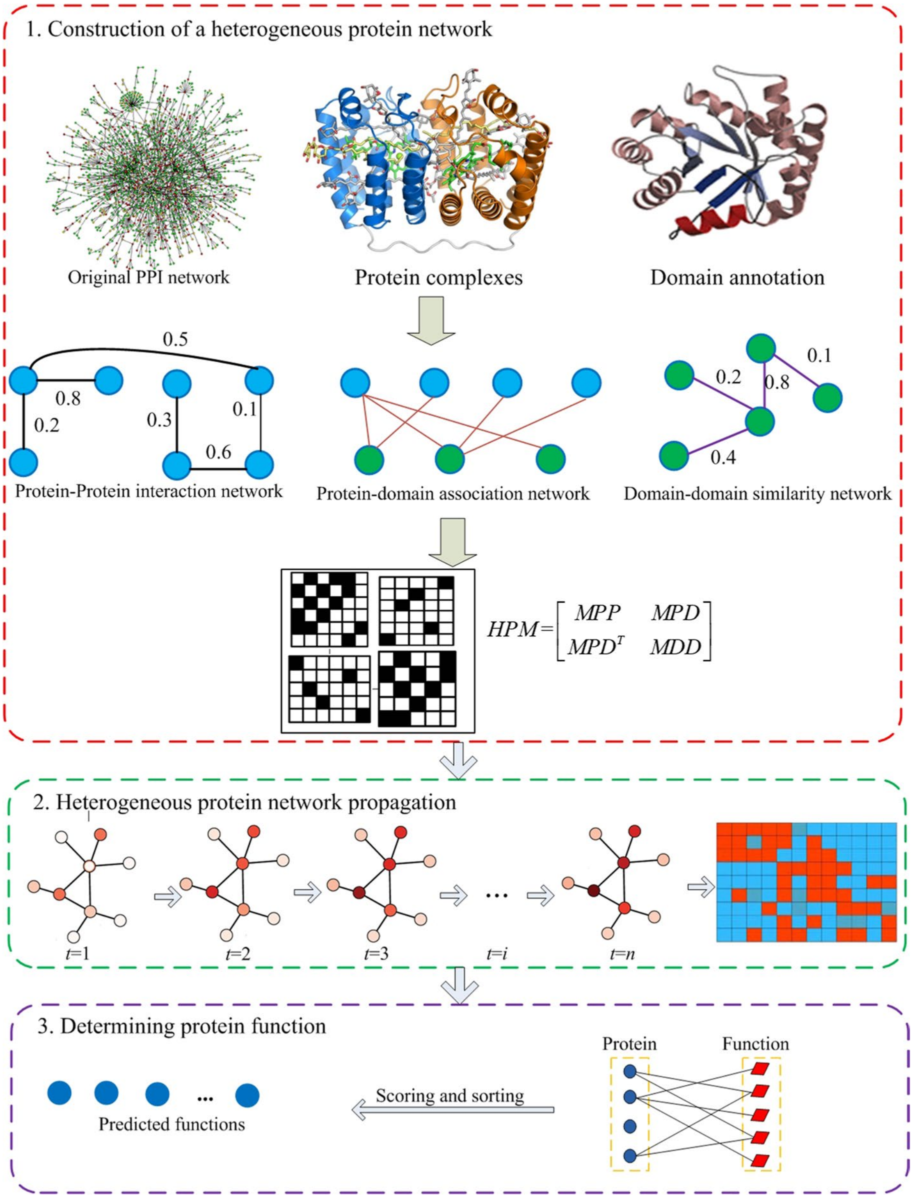

1. Construction of a heterogeneous protein network
Original PPI network
Protein complexes
Domain annotation
0.5
0.8
0.2
0.3
0.1
0.6
Protein-Protein interaction network
Protein-domain association network
0.2
0.8
0.1
0.4
Domain-domain similarity network
HPM = [MPP MPD; MPD^T MDD]
2. Heterogeneous protein network propagation
t=1
t=2
t=3
t=i
t=n
3. Determining protein function
Predicted functions
Scoring and sorting
Protein
Function

Fig. 1 The flowchart of PHN method. **1** Constructing the heterogeneous biological network based on original PPI network, domain annotation and protein complex information. **2** Given a testing protein *u*, running the propagation algorithm on the heterogeneous biological network to obtain the functional similarity score vector *pr* between *u* and the other proteins. **3** Scoring and ranking functions in descending order coming from all functional partners and selecting the first *L* of them as candidates to annotate the target protein

on network topology. Accordingly, the functional similarity between nodes of the sets S_1-S_2 and u is defined asα. In particular, if a protein appears in both S_1 and S_2 ($S_1 \cap S_2$), it is considered to have a necessary functional association with u and the functional similarity between them was set to 1. So, the functional similarity between a node in the sets S_2-S_1 and u is assigned $1-\alpha$. Given a protein v in the network, the functional similarity between u and v based on the local topology of PPI networks can be defined as follow:

$$fs(u,v) = \begin{cases} 1, & if \quad v \in S_1 \cap S_2 \\ \alpha, & if \quad v \in S_1 - S_2 \\ 1-\alpha, & if \quad v \in S_2 - S_1 \\ 0, & otherwise \end{cases} \tag{1}$$

Figure 2 illustrates these four sets of protein pairs. We are able to calculate the functional similarity between these neighbours and the target protein u according to Eq. 1: $fs(u, P_1) = \alpha$, $fs(u, P_2) = 1$, $fs(u, P_3) = 1$, $fs(u, P_4) = \alpha$, $fs(u, P_5) = 1 - \alpha$, $fs(u, P_6) = 0$.

Protein complexes are functional units of macro-molecular organization consisting of interacting proteins that perform cellular biological functions [17]. Considering the fact that associated experimental techniques may cause a false positive result in protein–protein interaction assays, we apply protein complex data to evaluate the module similarity between proteins for the construction of a more reliable protein interaction network. Let $ms(u, v)$ denote the module similarity of the interaction between u and v. Then, the module similarity of interaction (u, v) is calculated using the following equations:

$$ms(u,v) = \frac{|C_u \cap C_v|^2}{|C_u| \times |C_v|} \tag{2}$$

where C_u and C_v denote the set of protein complexes containing u and v, respectively. $C_u \cap C_v$ denotes the set of protein complexes containing both u and v. Finally, we construct a weighted protein–protein interaction network with high confidence by integrating the topology of PPI networks and protein complexes. The weight between protein u and v in the newly constructed protein–protein interaction network is the linear combination of their functional similarity and module similarity, and is defined as:

$$mpp(u,v) = \beta * fs(u,v) + (1-\beta) * ms(u,v). \tag{3}$$

Protein-domain association network

Domain refers to the dense spherical region of protein subunit structure, which is composed of 100–200 amino acid residues, each with unique spatial conformation and different biological functions. If protein u contains domain d, u connects domain dn with

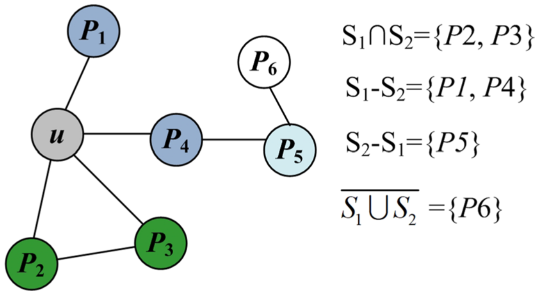


Fig. 2 Example to illustrate four sets of protein pairs. Exhibition four sets of neighbour pairs of the target protein u, including $S_1 \cap S_2$, $S_1 – S_2$, $S_2 – S_1$ and $\overline{S_1 \cup S_2}$

an edge in the protein-domain association network and $mpd(u, dn) = 1$, otherwise, there is no edge between them and $mpd(u, dn) = 0$.

Domain-domain similarity network

In this work, we evaluate the similarity between domains through their associated protein interaction information. The domain-domain similarity network is constructed based on the above constructed weighted protein–protein interaction network and the protein-domain association network. Let $PL(dn)$ represents the list of proteins that contain the domain dn. We quantitatively analyse the association between protein u and $PL(dn)$ of the domain dn according to weighted protein interaction data, which is expressed as follows:

$$S\_PD(u, PL(dn)) = \max_{pn_i \in PL(dn)} (mpp(u, pn_i)) \tag{4}$$

Then, for a given pair of domain dn_i and domain dn_j, the similarity between them can be calculated as follows:

$$mdd(dn_i, dn_j) = \frac{\sum_{pn_x \in PL(dn_i)} S\_PD(pn_x, PL(dn_i)) + \sum_{pn_y \in PL(dn_j)} S\_PD(pn_y, PL(dn_j))}{|PL(dn_i)| + |PL(dn_j)|} \tag{5}$$

where $PL(dn_i)$ and $PL(dn_j)$ are the protein set containing domain dn_i and domain dn_j, respectively and $S\_PD(pn_x, PL(dn_i))$ denotes the association between protein pn_x and the set of protein $PL(dn_i)$.$|PL(dn_i)|$ and $|PL(dn_j)|$ is the size of $PL(dn_i)$ and $PL(dn_j)$, respectively.

Based on the newly established protein–protein interaction network, protein-domain association network and domain-domain similarity network, a heterogeneous biological network is constructed and formally expressed by the adjacency matrix as follows:

$$HBM = \begin{bmatrix} MPP & MPD \\ MPD^T & MDD \end{bmatrix} \tag{6}$$

where MPP, MPD and MDD is the adjacency matrix corresponding to the protein–protein interaction network, protein-domain association network and domain-domain similarity network, respectively. MPD^T is a transport matrix of the matrix MPD.

Heterogeneous biological network propagation

Restricted to the small-world and scale-free features of the protein–protein interaction network, existing distance measures such as shortest distance and Euclidean distance are not suitable for evaluating the functional similarity between proteins [18, 19]. In order to prioritize functional partners in the network for a target unknown protein, the propagation algorithm is run on the heterogeneous biological network. Firstly, we established a transition probability matrix HBM_T based on the matrix MPD by normalized operation, which is formalized as follows:

$$HBM\_T = \begin{bmatrix} MPP\_T & MPD\_T \\ MPD\_T^T & MDD\_T \end{bmatrix} \tag{7}$$

The transition probability from protein pn_i to protein pn_j is expressed as:

$$mpp\_t(i,j) = p(pn_j|pn_i) = \begin{cases} mpp(i,j)/\sum_j mpp(i,j), & if \quad \sum_j mpd(i,j) = 0 \\ (1-\lambda)mpp(i,j)/\sum_j mpp(i,j), & otherwise \end{cases} \tag{8}$$

The transition probability from domain dn_i to domain dn_j is expressed as:

$$mdd\_t(i,j) = p(dn_j|dn_i) = \begin{cases} mdd(i,j)/\sum_j mdd(i,j), & if \quad \sum_j mpd(j,i) = 0 \\ (1-\lambda)mdd(i,j)/\sum_j mdd(i,j), & otherwise \end{cases} \tag{9}$$

The transition probability from protein pn_i to domain dn_j is expressed as:

$$mpd\_t(i,j) = p(dn_j|pn_i) = \begin{cases} \lambda mpd(i,j)/\sum_j mpd(i,j), & if \sum_j mpd(i,j) \neq 0 \\ 0, & otherwise \end{cases} \tag{10}$$

The transition probability from domain dn_i to protein pn_j is expressed as:

$$mpd\_t(j,i) = p(pn_j|dn_i) = \begin{cases} \lambda mpd(j,i)/\sum_j mpd(j,i), & if \quad \sum_j mpd(j,i) \neq 0 \\ 0, & otherwise \end{cases} \tag{11}$$

The parameter λ is the moving probability of the movement from the weighted protein–protein interaction network to the domain-domain similarity network and is assigned as 0.2 [20]. And then, we perform an iteration operation to calculate aggregated functional similarity scores between the given target protein u with other proteins by the following equation:

$$pr^{t+1} = (1-\gamma) * HBM\_T * pr^t + \gamma * pr^0 \tag{12}$$

The parameter $\gamma \in [0,1][0,1]$ is balanced between the propagation information and initial scores, which is set to 0.5 [21, 22]. pr^0=$[h(P);\ h(D)]$ denotes the initial functional similarity score vector, which is derived from the protein–protein interaction network corresponding to the matrix MPP and protein-domain association network. For a given protein p_i, its initial functional similarity score between the target protein u is expressed by the weight of interaction between p_i and u in the protein–protein interaction network, that is:

$$h(p_i) = mpp(u, p_i) \tag{13}$$

As for domains, their initial functional similar scores are derived from scores of their relevant proteins. Given a domain d_j, its initial score is computed by the following formula:

$$h(d_j) = \max_{p_x \in PL(d_j)} (h(p_x)) \tag{14}$$

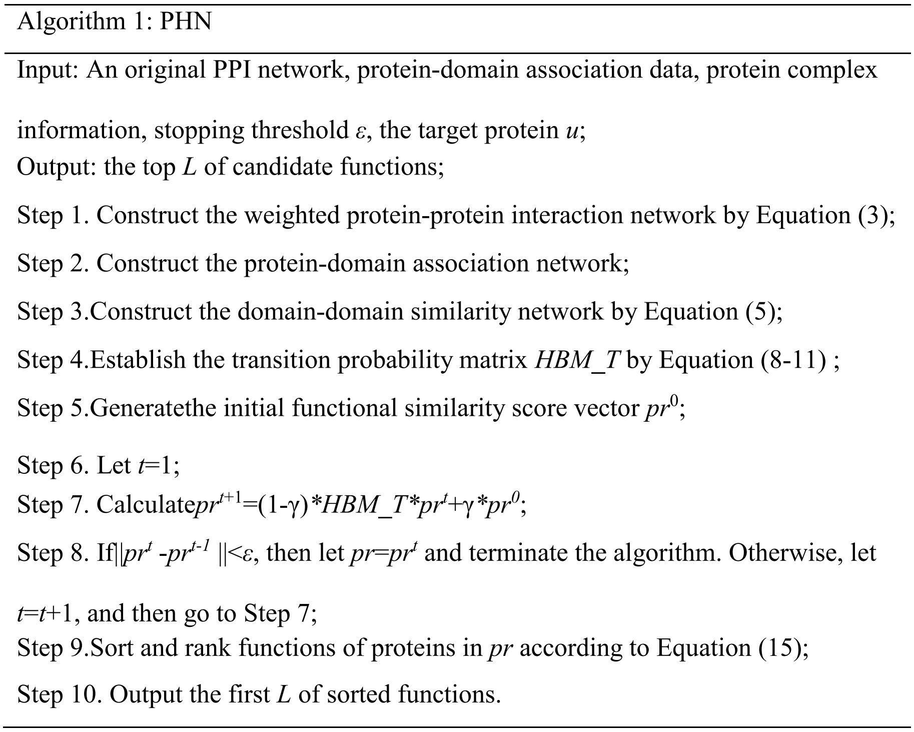
where $PL(d_i)$ is the protein set of domain d_i. In Eq. (12), if $\left\|pr^{t+1} - pr^t\right\|_1 \geq \varepsilon$, then t=$t+1$ and return to the previous step to continue the iteration, otherwise, the iteration end. When the propagation converges, we can obtain an aggregated scores vector pr, in which proteins are arranged in descending according to their functional similarity to the target protein u.

Determining protein function

MethodS typified by Song et al. [11] assigned all functions of the protein with the highest similarity value to the target protein with unknown functions. However, our statistical results on recent PPI data indicate that the function overlaps of more than half of protein pairs fell into [0.4, 0.6] and that of only 11.99% of protein pairs is above 0.6. So, functions are scored and ranked in descending order coming from all functional partners and the top L of them are picked out as candidates to annotate the target protein in this work. Let $FN=\{fn_1, fn_2, \ldots, fn_m\}$ be a list of distinct functions of proteins in pr that have a functional similarity score greater than 0 to the target protein u. For a given function fn_i in FN, its ranking score is obtained using the following formula:

$$S(fn_i) = \sum_{j=1}^{n} pr(j) * t_{ij} \tag{15}$$

In Eq. (15), if pn_i contains function fn_i, then $t_{ij}=1$, otherwise $t_{ij}=0$. The parameter L is assigned the number of functions of the protein within pr, which has the highest functional similarity score to the target protein u. Algorithm 1 gives the overall framework of the proposed PHN method.



```
Algorithm 1: PHN
Input: An original PPI network, protein-domain association data, protein complex
information, stopping threshold $\varepsilon$, the target protein $u$;
Output: the top $L$ of candidate functions;
Step 1. Construct the weighted protein-protein interaction network by Equation (3);
Step 2. Construct the protein-domain association network;
Step 3.Construct the domain-domain similarity network by Equation (5);
Step 4.Establish the transition probability matrix $HBM\_T$ by Equation (8-11) ;
Step 5.Generatethe initial functional similarity score vector $pr^0$;
Step 6. Let $t$=1;
Step 7. Calculate$pr^{t+1}$=(1-γ)*$HBM\_T$*$pr^t$+γ*$pr^0$;
Step 8. If||$pr^t$ -$pr^{t-1}$ ||<$\varepsilon$, then let $pr$=$pr^t$ and terminate the algorithm. Otherwise, let
$t$=$t$+1, and then go to Step 7;
Step 9.Sort and rank functions of proteins in $pr$ according to Equation (15);
Step 10. Output the first $L$ of sorted functions.
```

Results and discussion

Experimental data

The *Saccharomyces cerevisiae* (yeast) protein interaction networks are widely used in the research of network-based function prediction algorithms as a gold standard data because of their completeness and convincement. Here, we apply PHN and seven

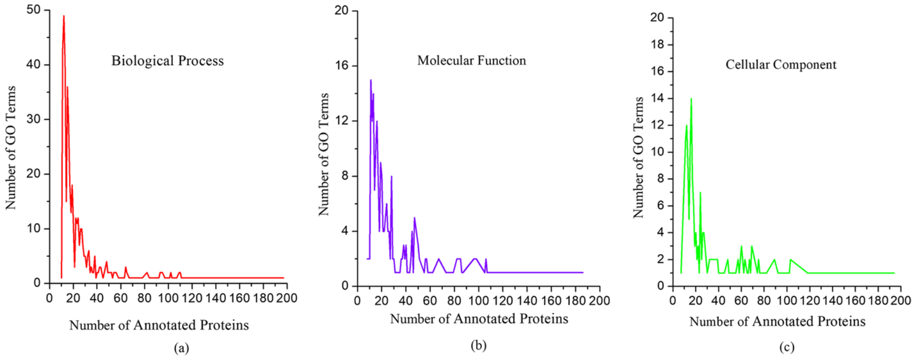


Fig. 3 Distribution of GO term in the BioGRID dataset. This Figure shows the distribution of GO term in the BioGRID dataset. X-axis represents the number of annotated protein. Y-axis represents the number of GO terms. **a** Biological process, **b** molecular function, **c** cellular component

competing methods on yeast interaction networks to test the performance of our new method. The original protein interaction data used in this work is downloaded from the BioGRID (Biological General Repository for Interaction Datasets) database [23], compiled on Aug. 25, 2022. The BioGRID PPI network contains 3145 proteins and 15,070 experimentally detected interactions with self-interactions and repeated interactions removed. The experimentally detected protein complex set for construction of the weighted PPI network is obtained from the CYC2008 database [24], which consists of 408 complexes involving 1408 proteins in the BioGRID database. The function annotation of proteins used for validation is downloaded from the Gene Ontology Consortium (GOC) [25]. The GO terms maintain three structured controlled vocabularies, which describe gene products in terms of their associated biological processes (BP), cellular components (CC) and molecular functions (MF). In the BioGRID network, 2957, 2250 and 2130 out of 3145 proteins are annotated by BP, MF and CC, respectively. The gold standard GOC consists of 518, 219 and 174 GO terms for BP, MF and CC respectively. Figure 3 depicts the distribution of GO terms in BP, MF and CC, respectively. We obtain 4936 protein-domain association data with invalid and duplicate relationships removed from the PFAM [26] database. It involves 906 distinct domain types related to 2044 proteins of the PPI networks. Figure 4 shows the distribution of Domain types in the BioGRID network. Figure 4 reveals that more than 63% of the domain types are associated with less than 5 proteins.

Evaluation criteria

In this work, the performance of PHN is evaluated by two cross validations, including leave-one-out cross validation and ten-fold cross validation. To measure the quality of predicted functions by our method, we adopt the ROC (Receiver-Operating Curve) [27, 28] as an assessment. The ROC curve is plotted with FPR (False Positive Rates) and TPR (True Positive Rates) [29], which is widely used in performance evaluation for protein function prediction. In addition, we also calculate the Precision, Recall and F-measure of the proposed PHN method. Precision is the fraction of predicted functions that are matched with benchmark functions while Recall is the fraction of benchmark functions that are matched with predicted functions. As the harmonic mean of precision and

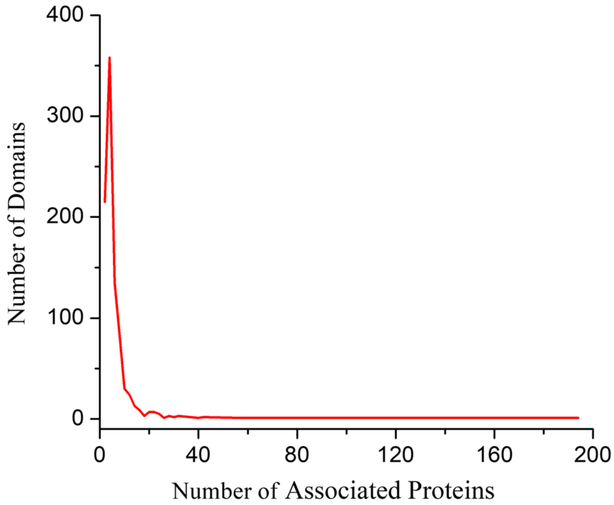


Fig. 4 Distribution of domain in the BioGRID dataset. This Figure shows the distribution of domain in the BioGRID dataset. X-axis represents the number of associated protein. Y-axis represents the number of domains

recall, F-measure is another assessment to evaluate the performance of the protein function prediction method synthetically.

Effect of parameter β

On the construction of the weighted protein–protein interaction network, we employ a user-defined parameter β to balance functional similarity and module similarity in Eq. 3. With different values of parameter β, the performances of prediction might differ greatly. As a result, we investigate the effect of parameter β on the PHN method by running 11 times with equal intervals of β from 0 to 1. The corresponding values of Recall, Precision and F-measure at different values of β are calculated. Table 1 shows how these performance evaluation criteria of our method fluctuate under various values of β based on GO terms in BP, MF and CC. From Table 1, we can easily see that the comprehensive evaluation criteria F-measure of PHN in BP reached the maximum value when β is assigned to 0.8. Similarly, the PHN method obtains optimal performance in MF and CC when β is set to 0.7 and 0.9, respectively. Therefore, we set the default value of parameter β in BP, MF and CC to 0.8, 0.7 and 0.9 respectively in all the following experiments.

Leave-one-out cross-validation

In this part, we use leave-one-out cross validation to evaluate the quality of the functions predicted by PHN and seven other competing methods. In each round, there is only one protein in the testing set and the rest in the training set. Firstly, we evaluate the comprehensive performance of PHN and seven other competing algorithms, such as NC, Song, DCS, DSCP, NPF, PON and *GrAPFI* by the average Precision, Recall and F-measure. Figure 5 shows the overall performance of the above eight methods in the matter of Precision, Recall and F-measure. PHN is the only method with F-measure above 50% in BP, MF and CC. Compared with NC, a classic network-based function prediction method, the F-measure of PHN for MF, CC and BP category is improved by 36.98%, 74.14% and

Table 1 Effect of parameter β on the performance of HPN

	0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
BP											
Recall	0.301	0.511	0.514	0.518	0.521	0.525	0.529	0.530	**0.536**	0.537	0.135
Precision	0.309	0.492	0.494	0.497	0.501	0.508	0.510	0.512	**0.516**	0.513	0.104
F-measure	0.305	0.501	0.504	0.507	0.511	0.516	0.519	0.521	**0.526**	0.525	0.118
MF											
Recall	0.441	0.547	0.549	0.550	0.554	0.552	0.555	**0.556**	0.549	0.545	0.238
Precision	0.446	0.551	0.553	0.554	0.556	0.557	0.558	**0.561**	0.561	0.561	0.239
F-measure	0.443	0.549	0.551	0.552	0.555	0.554	0.557	**0.559**	0.555	0.553	0.238
CC											
Recall	0.556	0.588	0.589	0.592	0.599	0.604	0.606	0.607	0.607	**0.612**	0.175
Precision	0.551	0.576	0.579	0.584	0.590	0.593	0.592	0.595	0.596	**0.598**	0.181
F-measure	0.554	0.582	0.584	0.588	0.594	0.598	0.599	0.601	0.602	**0.605**	0.178

Bold values represent the optimal value of parameter β set in BP, MF and CC

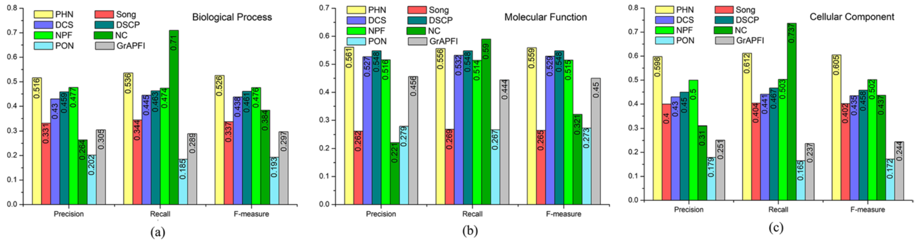


Fig. 5 The predicted results of various methods using leave-one-out cross validation. Numbers of each bar are the values for each score, including average Precision, Recall and F-measure. **a** Biological process, **b** molecular function, **c** cellular component

38.44% respectively. While compared with NPF which is the latest proposed function prediction algorithm with protein domain and complex information integrated, PHN also shows remarkable performance. Particularly, for GO terms in CC, the F-measure of PHN is 20% higher than that of NPF. Figure 5 indicates that PHN obtains the highest prediction precision of all the methods and the second-highest recall after NC. The recall of PHN is inferior to that of NC, due to the functions annotation strategy that PHN only selects the top part of the predicted functions to annotate the unknown protein, while the NC method assigns all the functions of neighbours to the target protein. This treatment of the NC method causes a lot of noise to emerge in its predicted functions, resulting in a sharp drop in precision. In this experiment, the recall of NC for the BP, MF and CC category is 32.31%, 6.11% and 20.38% higher than that of PHN, respectively. While its precision in BP, MF and CC is 95.83%, 154.1% and 92.49% lower than that of PHN.

The choice of the number of functions predicted by various methods is an important factor affecting their performance. We try our best to select a unified candidate functions selection strategy for each method to comprehensively and objectively compare and analyse the performance of different methods. The predicted functions are arranged in descending order based on the functional similarity score values derived by PHN, NPF, NC, PON and *GrAPFI*, respectively. And then top *L* of candidate functions

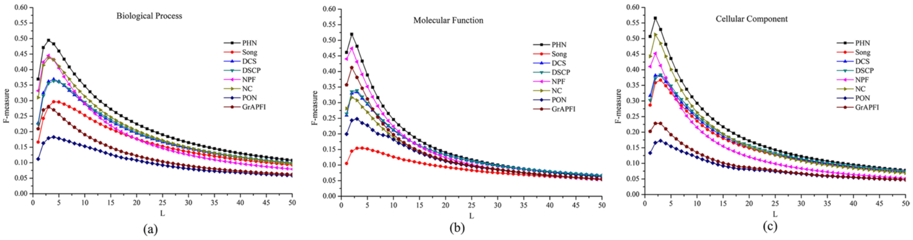


Fig. 6 F-measure curves as the number of selected functions *L* varies. This Figure depicts the F-measure of our method and other competing methods fluctuate under various value of the parameter *L*. The X-axis and Y-axis represents of the curve are the values of the parameter *L* and average F-measure, respectively. **a** Biological process, **b** molecular function, **c** cellular component

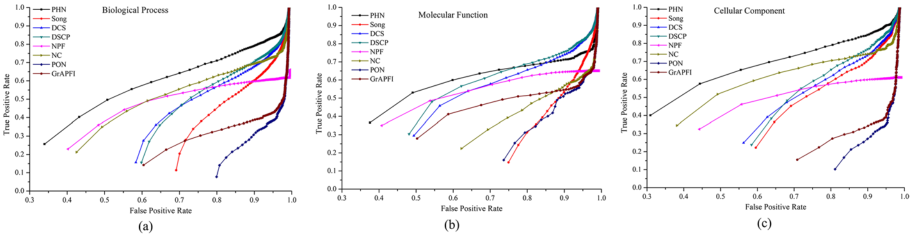


Fig. 7 ROC curves of eight methods using leave-one-out cross validation. The figure denotes the ROC (receiver-operating) curves of PHN and other seven competing methods (NC, Song, DCS, DSCP, NPF, PON and *GrAPFI*) based on the average prediction performance over all testing proteins. The X-axis and Y-axis of the ROC curves are the values of false positive rate and true positive rate, respectively. **a** Biological process, **b** molecular function, **c** cellular component

are picked out to annotate the target protein. For the three methods of Song, DCS and DSCP, the top M ($M \leq L$) of proteins with the highest function similarity to the target protein are selected, and the top L of functions from these M proteins are selected as predicted functions. A more valuable comparison between algorithms is presented by plotting F-measure curves as the value of L ($L \leq 50$) varies. Figure 6 shows the F-measure of our method and other competing methods fluctuates under various values of L in BP, MF and CC. From Fig. 6, we can see that the setting of L in the interval [2, 5] is the optimal solution for all methods. The experimental results also show that PHN achieves the best performance of all methods, regardless of the value of L.

Moreover, we also employ Receiver-Operating Curve (ROC) curves and the corresponding areas under the ROC curve (AUROC) values to evaluate the overall performance of each method. Firstly, functions are ranked in descending order according to the functional similarity scores to the target proteins computed by each method. After that, the top K functions are picked out and put into positive data set as candidate functions, and then the remaining functions are stored in negative data set. The upper limit values of K in BP, MF and CC are 518, 219, and 174, respectively. With different values of K selected, the values of TPR (False Positive Rates) and FPR (True Positive Rates) are computed for each method, respectively. Then, the values of TPR and FPR are plotted in ROC curves with different cut-off values. The experimental results are illustrated in Fig. 7. From Fig. 7, we can see intuitively that the ROC of PHN in BP, MF and CC is clearly above those of all other methods.

Table 2 The results of PHN and seven competing methods using ten-fold cross validation

Categories	Methods	Recall	Precision	F-measure
BP	PHN	0.523	0.504	0.513
	Song	0.335	0.323	0.329
	DCS	0.431	0.415	0.423
	DSCP	0.450	0.446	0.448
	NPF	0.470	0.475	0.472
	NC	0.678	0.259	0.374
	PON	0.175	0.192	0.183
	GrAPFI	0.289	0.306	0.297
MF	PHN	0.547	0.553	0.549
	Song	0.266	0.259	0.263
	DCS	0.522	0.519	0.520
	DSCP	0.536	0.536	0.536
	NPF	0.507	0.509	0.508
	NC	0.563	0.217	0.313
	PON	0.261	0.273	0.267
	GrAPFI	0.444	0.455	0.449
CC	PHN	0.604	0.584	0.594
	Song	0.398	0.394	0.396
	DCS	0.428	0.418	0.423
	DSCP	0.458	0.442	0.449
	NPF	0.503	0.555	0.520
	NC	0.706	0.304	0.425
	PON	0.160	0.175	0.167
	GrAPFI	0.237	0.251	0.244

For further comparison, we calculate the AUROC of all these methods. For the BP category, the AUROC of PHN is 154.03%, 75.33%, 75.91%, 40.29%, 28.06%, 548.92% and 209.91% larger than that of Song, DCS, DSCP, NPF, NC, PON and *GrAPFI*, respectively. Compared with NC which has the best performance in ROC curves among seven competing methods, the AUROC of PHN for the MF and CC category is improved by 123.4% and 23.56%, respectively.

Ten-fold cross validation

TO avoid the possible deviation caused by leave-one-out cross validation, we further evaluate the prediction performance of the PHN method using the ten-fold cross 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. The results of ten folds are averaged to generate the final performance. Table 2 lists the prediction results of eight methods, including the average Precision, Recall and F-measure. Table 2 shows that PHN still performs the best, in terms of precision and F-measure. Taking the BP category as an example, the F-measure of PHN is 55.93%, 21.28%, 14.51%, 8.69%, 37.17%, 180.33% and 72.73% higher than that of Song, DCS, DSCP, NPF, NC, PON and *GrAPFI*, respectively. In addition, we plot ROC curves of all methods for the MF, CC and BP category as shown in Fig. 8. The AUROC of PHN in BP is 152.21%, 77.71%, 76.27%, 35.47%, 33.10%, 570.28% and

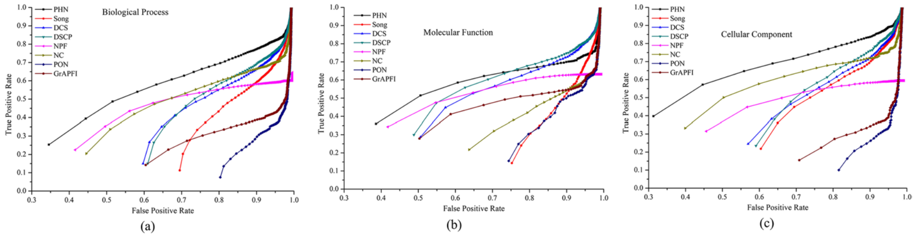


Fig. 8 ROC curves of eight methods using ten-fold cross validation. This Figure shows the ROC curves of PHN and other seven methods using ten-fold validation. Proteins are randomly divided into ten equal sets, one set for testing, and the rest for training. And then, the cross validation is repeated process for ten rounds, using various folds as the training and validation data in each round. The results of ten folds are averaged to generate the final performance. **a** Biological process, **b** molecular function, **c** cellular component

199.27% larger than that of Song, DCS, DSCP, NPF, NC, PON and *GrAPFI*, respectively. As for the MF and CC category, PHN increases the AUROC by no less than 15% and 28%, respectively, compared with other competitive comparison methods.

Conclusions

ProteinS are one of the most important and diverse macromolecules in cell life activities. Precise labelling of protein function information is important for promoting the research and development of protein mechanism analysis, disease mechanism analysis and control, new drug research and development, crop production promotion, bio-energy development and so on. The rapid developments of high-throughput technologies have generated large quantities of protein–protein interaction (PPI) data, which prompts the emergence of computational methods to determine protein function. Despite the effective improvement of these computational methods, building a suitable network model to integrate multiplex biological data remains a challenge due to the incomplete and error-prone raw PPI data. How to construct an effective network model that integrates multiplex biological data and network topology remains a challenge. Current methods aggregated multiple biological data into a single network, in which the role of different types of data is reflected in the form of setting weights or parameters. The choice of weighting parameters and the inherent properties of different biological data restrict the further development of these methods. In this work, we construct a heterogeneous biological network with two categories of nodes: protein and domain. To shake off the bound of the small-world and scale-free of PPI networks, innovatively, we use the propagation algorithm on the heterogeneous biological network and obtain a functional partners list with aggregated similarity to the target protein. Finally, we score and rank functions from the partners list in descending order. The number of candidate functions we selected is equal to the number of functions of the neighbour most similar to the target protein. To assess the overall performance of PHN, we use the leave-one-out cross validation and ten-fold cross validation. The F-measure and AUROC of our method improved by more than 17% and 15%, respectively, compared with other approaches. The experimental results also indicate that PHN is a specific and effective method that can predict protein function.

Acknowledgements
Not applicable.

Author contributions

SH and BHZ obtain the protein–protein interaction data, domain data, GO Terms, and the protein complexes information. SH, ZHZ, BHZ and YCL designed the new method, PHN, and analysed the results. SH, ZHZ, BHZ, YCL HJX, WY and MPJ draft the manuscript together. SH and BHZ participate in revising the manuscript. All authors have read and approved the manuscript.

Funding

This work is 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 (Nos. 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 are analysed in this study. This data and the PHN program can be found here: https://github.com/husaiccsu/PHN.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent to publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 3 August 2022 Accepted: 15 November 2022
Published online: 18 November 2022

References

1. Lewis DL, Hagstrom JE, Loomis AG, et al. Efficient delivery of siRNA for inhibition of gene expression in postnatal mice. Nat Genet. 2002;32(1):107–8.
2. Li E, Bestor TH, Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell. 1992;69(6):915–26.
3. Tsien JZ, Chen DF, Gerber D, et al. Subregion-and cell type–restricted gene knockout in mouse brain. Cell. 1996;87(7):1317–26.
4. Schwikowski B, Uetz P, Fields S. A network of protein–protein interactions in yeast. Nat Biotechnol. 2000;18(12):1257–61.
5. Chua HN, Sung WK, Wong L. Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. Bioinformatics. 2006;22(13):1623–30.
6. Vazquez A, Flammini A, Maritan A, et al. Global protein function prediction from protein-protein interaction networks. Nat Biotechnol. 2003;21(6):697–700.
7. Devkota K, Schmidt H, Werenski M, et al. GLIDER: function prediction from GLIDE-based neigborhoods. Bioinformatics. 2022;38:3395.
8. Liang S, Zheng D, Standley DM, et al. A novel function prediction approach using protein overlap networks. BMC Syst Biol. 2013;7(1):61.
9. Peng W, Li M, Chen L, et al. Predicting protein functions by using unbalanced random walk algorithm on three biological networks. IEEE/ACM Trans Comput Biol Bioinf. 2015;14(2):360–9.
10. Sarker B, Rtichie D W, Aridhi S. Exploiting complex protein domain networks for protein function annotation. In: International conference on complex networks and their applications, Springer, Cham, 2018. p. 598–610.
11. Sarker B, Ritchie DW, Aridhi S. GrAPFI: predicting enzymatic function of proteins from domain similarity graphs. BMC Bioinf. 2020;21(1):1–15.
12. Zhang S, Chen H, Liu K, et al. Inferring protein function by domain context similarities in protein–protein interaction networks. BMC Bioinf. 2009;10:1.
13. Peng W, Wang J, Cai J, et al. Improving protein function prediction using domain and protein complexes in PPI networks. BMC Syst Biol. 2014;8(1):35.
14. Kulmanov M, Zhapa-Camacho F, Hoehndorf R. DeepGOWeb: fast and accurate protein function prediction on the (semantic) web. Nucleic Acids Res. 2021;49:140.
15. Zhao B, Zhang Z, Jiang M, et al. NPF: network propagation for protein function prediction. BMC Bioinf. 2020;21(1):1–21.
16. Titz B, Schlesner M, Uetz P. What do we learn from high-throughput protein interaction data? Expert Rev Proteomics. 2004;1(1):111–21.
17. Tang X, Li X, Hu S, et al. A framework for identifying functional modules in dynamic networks. Int J Data Min Bioinform. 2018;21(1):1–17.
18. Zhao B, Han X, Liu X, et al. A novel method to predict essential proteins based on diffusion distance networks. IEEE Access. 2020;8:29385–94.

19. Cao M, Zhang H, Park J, et al. Going the distance for protein function prediction: a new distance metric for protein interaction networks. PLoS ONE. 2013;8(10):e76339.
20. Zhao B, Zhao Y, Zhang X, et al. An iteration method for identifying yeast essential proteins from heterogeneous network. BMC Bioinf. 2019;20(1):1–13.
21. Hwang TH, Sicotte H, Tian Z, et al. Robust and efficient identification of biomarkers by classifying features on graphs. Bioinformatics. 2008;24(18):2023–9.
22. Vanunu O, Magger O, Ruppin E, et al. Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol. 2010;6(1):e1000641.
23. Rose O, Chris S, Bobby-Joe B, et al. The BioGRID interaction database: 2019 update. Nuclc Acids Res. 2018;D1:D1.
24. Pu S, Wong J, Turner B, et al. Up-to-date catalogues of yeast protein complexes. Nucleic acids Res. 2009;37(3):825–31.
25. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25–9.
26. Sara EG, Jaina M, Alex B, et al. The Pfam protein families database in 2019. Nuclc Acids Res. 2018;D1:D1.
27. Peng W, Tang Q, Dai W, et al. Improving cancer driver gene identification using multi-task learning on graph convolutional network. Briefings Bioinf. 2022;23(1):bbab432.
28. Peng W, Yi S, Dai W, et al. Identifying and ranking potential cancer drivers using representation learning on attributed network. Methods. 2021;192:13–24.
29. Hu S, Zhang Z, Xiong H, et al. A tensor-based bi-random walks model for protein function prediction. BMC Bioinf. 2022;23:199.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.