جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Cross-Species Protein Function Prediction with Asynchronous-Random Walk

Item Type	Article
Authors	Zhao, Yingwen;Wang, Jun;Guo, Maozu;Zhang, Xiangliang;Yu, Guoxian
Citation	Zhao, Y., Wang, J., Guo, M., Zhang, X., & Yu, G. (2019). Cross-Species Protein Function Prediction with Asynchronous-Random Walk. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1–1. doi:10.1109/tcbb.2019.2943342
Eprint version	Post-print
DOI	10.1109/tcbb.2019.2943342
Publisher	Institute of Electrical and Electronics Engineers (IEEE)
Journal	IEEE/ACM Transactions on Computational Biology and Bioinformatics
Rights	(c) 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/ republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.
Download date	2024-04-17 04:19:30
Link to Item	http://hdl.handle.net/10754/656922

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2943342, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Cross-Species Protein Function Prediction with Asynchronous-Random Walk

Yingwen Zhao, Jun Wang, Maozu Guo, Xiangliang Zhang and Guoxian Yu

Abstract—Protein function prediction is a fundamental task in the post-genomic era. Available functional annotations of proteins are incomplete and the annotations of two homologous species are complementary to each other. However, how to effectively leverage *mutually complementary* annotations of different species to further boost the prediction performance is still not well studied. In this paper, we propose a cross-species protein function prediction approach by performing Asynchronous Random Walk on a heterogeneous network (*AsyRW*). AsyRW firstly constructs a heterogeneous network to integrate multiple functional association networks derived from different biological data, established homology-relationships between proteins from different species, known annotations of proteins and Gene Ontology (GO). To account for the intrinsic structures of intra- and inter-species of proteins and that of GO, AsyRW quantifies the individual walk lengths of each network node using the gravity-like theory and performs asynchronous-random walk with the individual length to predict associations between proteins and GO terms. Experiments on annotations archived in different years show that individual walk length and asynchronous-random walk can effectively leverage the complementary annotations of different species, AsyRW has a significantly improved performance to other related and competitive methods. The codes of AsyRW are available at: http://mlda.swu.edu.cn/codes.php?name=AsyRW.

Index Terms—Protein function prediction, Data fusion, Heterogeneous network, Asynchronous random walk, Gene Ontology

1 Introduction

Proteins play crucial functions in many life processes, such as metabolism, signal transduction and hormonal regulation [1]–[3]. Comprehensively and accurately annotating biological functions of proteins greatly advances the development of drug, bio-chemicals and disease analysis. However, rapidly accumulated genomic/proteomic data far exceeds the capability of annotating the functions of proteins by wet-lab techniques, which are very time consuming and expensive. Furthermore, research preferences of biologists and the experimental ethics involving human and animals/plants result in the bias of annotations [4]. Therefore, developing computational models to precisely annotate functions of proteins in large scale is of great importance and necessity.

Gene Ontology (GO) [5], [6] is launched to support computational protein function prediction, and to provide controlled vocabularies (or terms) for describing functions of protein across different species in a species-neutral way. These terms are organized in three sub-ontology, namely biological process ontology (BPO), molecular functions ontology (MFO) and cellular component ontology (CCO). The terms in each ontology form a direct acyclic graph (DAG), which captures the hierarchical relationships between terms and also reflects the current knowledge of biology. GO annotations, another component of GO, record the currently known functional annotations of proteins. Each (negative) annotation associates a protein with a GO term, indicating the protein (not) carrying out the function described by the term.

Because of the wide application of high-throughput biotechnologies, various proteomics/genomics data are used for predicting protein functions, such as protein-protein interaction (PPI) network, gene expression, amino acids and so on [3]. More competitive solutions integrate multiple types of biological data to reach a comprehensive view of proteins and to predict protein function [7]–[10]. To name a few, Lan *et al.* [8] proposed a multi-source k-nearest neighbor method, which employs a kNN classifier on each individual functional association network derived from different data sources and then integrates these classifiers to predict protein function. Yu *et al.* [9] introduced a semantic based data fusion solution (SimNet), which firstly defines a weighted term overlap based semantic similarity kernel, and then aligns multiple functional association networks with the semantic kernel to optimize the weights of these networks. Next, SimNet combines these networks into a composite network with the weights and performs random walk on the composite network to predict protein functions. Cho *et al.* [10] proposed an integrative approach (MashUP), which firstly applies random walk with restarts on each functional network of proteins to explore the topology

Y. Zhao and J. Wang are with the College of Computer and Information Sciences, Southwest University, Chongqing 400715, China.
G. Yu are with the College of Computer and Information Sciences, Southwest University, Chongqing 400715, China; and Computer Science, Electrical and Mathematical Science and Engineering Division, King Abdullah University of Science and Technology, SA.
M. Guo is with the School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China and Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing 100044, China.
X. Zhang is with Computer Science, Electrical and Mathematical Science and Engineering Division, King Abdullah University of Science and Technology, SA.
X. Zhang and G. Yu are the corresponding authors. Email: xiangliangzhang@kaust.edu.sa; gxyu@swu.edu.cn.
Manuscript received January 27, 2018; Revised xx xxx, xxxx

1545-5963 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2943342, IEEE/ACM Transactions on Computational Biology and Bioinformatics

information and add-ups these updated networks into a composite network. Next, MashUP uses singular value decomposition on the composite network to get low-rank feature representation of proteins, and then uses the support vector machine and low-rank features to predict protein function. Increasing studies shows that fusing different biological data can significantly improve the prediction accuracy than using single type data alone [1]. However, most data fusion based solutions focus on predicting functions of proteins from single species [9], [11]–[13], and the curated annotations of proteins in different species are biased, incomplete and imbalanced [4]. It is recognized that homologous species have a large portion of homologous proteins, which share similar (or even the same) GO annotations [4]. However, homologous proteins are annotated with different GO terms, due to resource limitations and preferences of biologists. Therefore, these complementary annotations of proteins across species should be credibly leveraged for large scale protein function prediction.

The sequence (structure) data along with the BLAST, PSI-BLAST [14] are typically used for cross-species protein function prediction, but these homology-based solutions are not reliable when the sequence identity is 25% or less [3]. Mitrofanova *et al.* [15] introduced a Gene Ontology chain-graph-based approach to predict protein functions, which builds on connecting networks of two (or more) different species by links of high interspecies sequence homology and PPIs, but this chain-graph-based approach can only apply for a small amount of GO terms (or labels). Park *et al.* [16] stated that sequence comparison does not directly assess the extent to which two proteins participate in the same biological processes, they additionally used the gene expression data to complement traditional sequence similarity measures and significantly improved the transfer of annotations between organisms [16]. Some other studies try to use more advanced sequence/physical-chemical similarity metrics to fuse different types of feature similarities to boost the performance [17]–[19]. For example, You *et al.* [19] recently introduced an approach called GOLabeler, which first separately extracts sequence features from five different perspectives and individually trains five different classifiers on these five types of features (one classifier for one type of features). GOLabeler then utilizes the principle of 'learning to rank' [20] to ensemble these five classifiers to achieve prediction.

These cross-species solutions canonically borrow annotations of the 'well-annotated' species for reference, while the 'well-annotated' species also lack annotated sequences [21]. Furthermore, they neglect the dynamic *mutually complementary* annotations of two close-homology species. Yu *et al.* [22] investigated semantic similarity based cross-species protein function prediction and observed that annotations of two species with high homology are complementary for each other, and integrating them significantly improves the prediction accuracy. But for two species with low homology, integrating the annotations does not bring in a significant improvement. Wang *et al.* [23] introduced an approach called ProSNet, which first builds an integrated heterogeneous network to include molecular networks of multiple species and link together homologous proteins based on sequence data. Next, ProSNet samples a large number of heterogeneous paths on the heterogeneous network to find the low-dimensional vector for each node. After that, it separately computes an intra-species affinity score and an inter-species affinity score for transferring annotations within species and between species. Finally, it averages these scores to accomplish protein function prediction. However, the low-rank representations of proteins may distort the intrinsic structures of the heterogeneous network. To avoid this issue, Yu *et al.* [24] proposed a bi-random walk based approach called NewGOA, which can account for structural difference between the subnetwork of terms and that of proteins by setting two walk-lengths for nodes of these two subnetworks. NewGOA, alike other random-walk based protein function prediction solutions [25], [26], still ignores the different contributions of different nodes on function transfer.

In this paper, we propose a cross-species protein function prediction approach with Asynchronous Random Walk (AsyRW). AsyRW firstly constructs multiple functional association networks based on different genomic data, and then integrates these functional association networks into a composite network for each species. It utilizes the amino acids data to establish the homology-relationships between proteins of different species, which serve as the bridge for functions transfer between species. Next, AsyRW constructs a heterogeneous network (As shown in Fig. 1), which includes the composite networks, established homology-relationships between proteins, known annotations of proteins and Gene Ontology. To account for the different structures of intra- and inter-species subnetworks, and that of Gene Ontology, it uses a gravity-like theory to quantify the individual walk-length of each node. Finally, AsyRW predicts annotations of proteins based on asynchronous-random walk on the heterogeneous network, where each node has its own individual walk-lengths.

We utilized GO annotations archived in 2016 to train AsyRW and then evaluated its predictions based on GO annotations archived in 2018. Experimental results on three model species (Homo sapiens, Mus musculus and Rattus norvegicus) show that AsyRW generally achieves better performance than other related methods across various evaluation metrics. The individual walk-lengths used by asynchronous-random walk can utilize the mutually complementary annotations of proteins and contribute to a better performance than existing relevant solutions [9], [10], [19], [22], [24]. In addition, AsyRW is robust to input parameters and thus readily applicable for other species.

1545-5963 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2943342, IEEE/ACM Transactions on Computational Biology and Bioinformatics

2 The Proposed Method

In this section, we firstly construct a heterogeneous network to fuse different types of genomics/proteomics data, Gene Ontology and known GO annotations of proteins, and then predict GO annotations of proteins based on asynchronous-random walk on the heterogeneous network. We elaborate on these two steps in the following subsections.

2.1 Constructing a heterogeneous network

The heterogeneous network is illustrated in Fig. 1, it is composed with two types of nodes (proteins and GO terms), and four types of edges (functional associations between GO terms and proteins, hierarchical relationships between GO terms, intra-species and inter-species protein similarities).

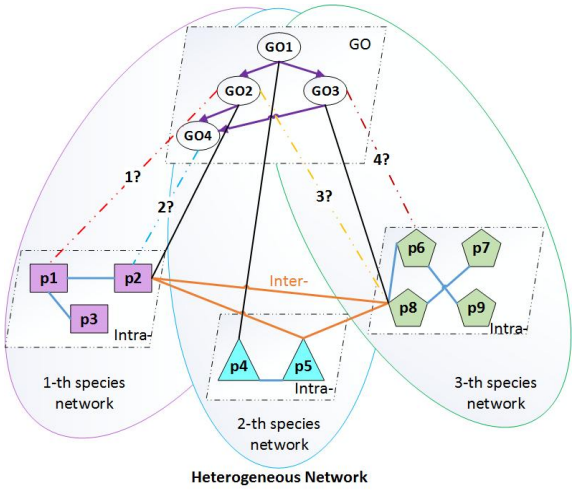


Fig. 1. A heterogeneous network includes proteins of three species, GO terms and relationships between them. Multiple protein functional similarity networks derived from different genomics data of each species are fused into a intra-species network of proteins. The inter-species networks (orange edges) encode the PSI-BLAST established homology relationships between proteins. Direct solid edges between proteins and GO terms encode the known GO annotations of proteins, while dashed edges marked with '*?' indicate the potential new annotations of proteins.

2.1.1 Constructing intra-species network for functional transfer within the species

With the continuous development and wide application of high-throughput biotechnologies, the accumulated biological data related to protein functions are increasingly rich, such as PPI networks, amino acids, gene micro-array, RNA-Seq data. Increasing evidences show that effectively integrating these heterogeneous data contributes to a higher accuracy and coverage than only using single type of data alone [9], [10]. The reason is that multi-type data describe proteins from different perspectives and complement with each other, and combining them can achieve a more comprehensive view of proteins in various complex biological activities, and thus improve the prediction.

Suppose we are given K species and M types of biological data. For each biological data type of the k-th species ($k = 1, 2, \cdots, K$), following the conventional similarity metrics for this data type, we can construct a functional association network of proteins using their respective feature vectors, and thus obtain M functional association networks, which are encoded by $\mathbf{A}_m^k \in \mathbb{R}^{n_k \times n_k} (m = 1, 2, \cdots, M)$, where n_k indicates the number of proteins for the k-th species. For example, we can separately apply PSI-BLAST to construct a functional association network of proteins from amino acids data, Pearson correlation coefficients to construct another functional network from gene micro-array data, and the Jaccard similarity to construct a network from protein domain data. Note, AsyRW is also applicable when some types of biological data are not available for a species.

The functional association networks of each species are derived from different data sources, and each data source may have incomplete (or noisy) features of proteins, or even omit some proteins. Therefore, it is advisable to integrate them to form a composite network. Although these networks are of the same size, they may have different scales. As such, we firstly normalize these networks into the same scale [0,1] as follows:

$$\tilde{\mathbf{A}}_m^k(i,j) = \frac{\mathbf{A}_m^k(i,j) - min(\mathbf{A}_m^k)}{max(\mathbf{A}_m^k) - min(\mathbf{A}_m^k)} \quad (1)$$

where $min(\mathbf{A}_m^k)$ (or $max(\mathbf{A}_m^k)$) is the min (or max) value of $\mathbf{A}_m^k$. Since the respective relevances of these networks for function prediction are unknown and varying across functions, we integrate these networks by simply averaging them $\mathbf{A}^k = \frac{1}{M}\sum_{m=1}^{M} \tilde{\mathbf{A}}_m^k$. We want to remark that more advanced integration of these networks can be made but with more computational costs. In practice, the results in Table S6 and Section 3 of the supplementary file show that combining these networks by different weights does not significantly improve the prediction performance, since our collected datasets are indeed complementary for each other. To facilitate the follow-up random walk process on the network and to ensure that the transitional probability from one protein to other proteins of the same species is summed up to 1, $\mathbf{A}^k$ is further normalized as follows:

$$\mathbf{W}_{pp}^k = (\mathbf{D}^k)^{-1}\mathbf{A}^k \quad (2)$$

where $\mathbf{D}^k$ is a diagonal matrix with $\mathbf{D}^k(i,i) = \sum_{j=1}^{n_k} \mathbf{A}^k(i,j)$.

2.1.2 Constructing inter-species networks for function transfer between species

Compared with other biological data, most species have only sequence data [17], [19]. Some methods utilizes sequence data to establish the cross-species function transfer bridges [19], [23]. PSI-BLAST [14] is widely-used to

1545-5963 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2943342, IEEE/ACM Transactions on Computational Biology and Bioinformatics

measure the sequence similarity between proteins from amino acids, which can be used to construct a weighted interspecies network of proteins and enables functional knowledge transfer between species. Here, we use PSI-BLAST with E-Value= $1e^{-5}$ as threshold to establish the cross-species relationships, and $\mathbf{H}^o \in \mathbb{R}^{N\times N}(N = n_1+n_2+\cdots+n_K)$ to store those inter-species information. Particularly, if the E-Value between protein i and protein j is lower than $1e^{-5}$, $\mathbf{H}^o(i,j)$ is set as the PSI-BLAST score between them, otherwise $\mathbf{H}^o(i,j) = 0$. Similar as the normalization on the intra-species network, $\mathbf{H}^o$ is normalized to ensure the transitional probability from a protein to other proteins is 1 as follows:

$$\mathbf{W}^o = (\mathbf{D}^o)^{-1} * \mathbf{H}^o \tag{3}$$

where $\mathbf{D}^o$ is a diagonal matrix with $\mathbf{D}^o(i,i) = \sum_{j=1}^{N} \mathbf{H}^o(i,j)$.

2.1.3 Constructing inter-association network between proteins and GO terms

Unlike the above two types of subnetworks, which are often undirect and embody the same object type, the inter-association network is direct and embody two types of nodes (proteins and GO terms). The associations between proteins and GO terms can be specified based on the GO annotations of proteins. Let $\mathbf{Y} = [\mathbf{Y}^1; \mathbf{Y}^2; \cdots; \mathbf{Y}^K] \in \mathbb{R}^{N\times|\mathcal{T}|}$ store all the available annotations of proteins, $\mathbf{Y}^k \in \mathbb{R}^{n_k\times|\mathcal{T}|}$ encodes the known annotations of n_k proteins of the k-th species with respect to $|\mathcal{T}|$ GO terms. If protein i is annotated with term t or its descendants, $\mathbf{Y}^k(i,t) = 1$, otherwise $\mathbf{Y}^k(i,t) = 0$. In this way, we can initialize the inter-association network of proteins and GO terms as $\mathbf{W}_{pg} = \mathbf{Y}$.

2.1.4 Measuring the transitional probability between GO terms

It is observed that given a missing term (or annotation) of a protein, the prediction of this missing term from its parental terms is more accurate than that from other ancestor terms [27]–[29]. As such, the patterns of GO annotations and hierarchy can be used to predict the missing annotations of proteins [25], [30]. But due to the incomplete GO annotations of many proteins, the underlying conditional probability (or taxonomic similarity) between two GO terms cannot be well estimated from available annotations alone. Based on the species-neutral characteristics of GO, we use a composite transitional probability between parent term t and its child term s as follows:

$$p^k(s|t) = \frac{n_s^k}{n_t^k} + \frac{IC(s)}{\sum_{v\in ch(t)} IC(v)} \tag{4}$$

where $ch(t)$ includes the set of direct child terms of t, and n_t^k is the number of proteins annotated with t or t's descendants in the k-th species (if $n_t^k = 0$, then $p^k(s|t) = 0$). $IC(t)$ is the information content of term t and calculated as:

$$IC(t) = 1 - \frac{log(|desc(t)|)}{log(|\mathcal{T}|)} \tag{5}$$

where $|desc(t)|$ is the cardinality of $desc(t)$, which includes all the descendants of t and itself, and $|\mathcal{T}|$ is the total number of terms. The compositional transitional probability is a composition of the known annotations of proteins and the hierarchical structure of GO, thus it is less affected by incomplete and imbalanced annotations, and it is also applicable to a new species, whose functional annotations are completely unknown. In fact, Eq. (5) has been successfully used in [24], [31]. In addition, to account for the intrinsic pattern of GO annotations of each species, the transitional probability is separately computed for each species.

To make the transitional probability from a term to its direct child terms equal to 1, we use a matrix $\mathbf{W}_{gg}^k \in \mathbb{R}^{|\mathcal{T}|\times|\mathcal{T}|}$ to encode the transitional probability between GO terms for the k-th species as follows:

$$\mathbf{W}_{gg}^k(t,s) = \frac{p^k(s|t)}{\sum_{v\in ch(t)} p^k(v|t)} \tag{6}$$

To this end, we construct a composite heterogeneous network composed of proteins from multiple species, GO terms and the associations between them. We then can propagate the annotations of proteins via random walk on the network to accomplish cross-species protein function prediction.

2.2 Asynchronous random walk for protein function prediction

Research priority of biologists and particular wet-lab experiments conducted on different model species cause that GO annotations of proteins across species are very imbalanced and complementary to each other [4], [6]. To use the mutually complementary annotations of close-homology species, we perform random walk on the constructed heterogeneous network to explore the network topology and to predict the missing associations between proteins and GO terms. Various random walk based solutions [32] can be adopted here, but these solutions generally apply random walk with restart under a *fixed* walk-length for all nodes. In practice, two homologous proteins may be annotated with some different GO terms, and a protein may should borrow the annotations from its homologous proteins, but not vice versa. In other words, these solutions ignore the different contributions of different nodes in the network. Some variant solutions prefix two or three walk-lengths in different subnetworks [24], [33], but these variants still do not assign individual walk-lengths to different nodes. Given that, we advocate to perform *asynchronous random walk* with each node having its own walk-length to infer the associations between proteins and GO terms. In the following subsections, we firstly quantify the individual walk-lengths of nodes and then formalize the asynchronous random walk under the quantified walk-lengths.

1545-5963 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2943342, IEEE/ACM Transactions on Computational Biology and Bioinformatics

2.2.1 Quantifying individual walk-lengths on three subnetworks

A network node has its own characteristics, such as the unique topology and information entroy. The number of nodes affected by a node also have certain restrictions. As such, a node should (can) have its own walk-length. The walk-length of a node generally depends on its influence (or information) in the network [34]. Newton's law of universal gravitation measures the gravitation between two objects by their masses and distance as follows:

$$G_{ij} = g\frac{M_i M_j}{r^2} \tag{7}$$

where M_i and M_j represent the masses of two objects, r represents the distance between them, and g is the gravitation constant. This equation means that the gravitation of two objects is proportional to the product of their masses and inversely proportional to their distance. Several gravity-like algorithms have been proposed based on the Newton's law of universal gravitation and successfully applied in different fields [35]–[37].

Inspired by these applications [37], we assume that the total gravitational value of a node relative to its surrounding nodes can be viewed as its influence, which determines its walk-length. To realize this assumption, we take the inverse of similarity between nodes of the network as the topological distance, $\mathcal{T}_i$ as the annotated GO terms to protein i and the mass of this protein, and set the gravitation constant g as 1. The product between $\mathcal{T}_i$ and $\mathcal{T}_j$ is defined as the number of elements in their intersection set $|\mathcal{T}_i \cap \mathcal{T}_j|$. In this way, we can quantify the individual walk-length of a node as follows:

$$\mathbf{L}_p^k(i) = \sum_{j \in \mathcal{N}_i^{intra}} \frac{|\mathcal{T}_i \cap \mathcal{T}_j|}{(\frac{1}{\mathbf{W}_{pp}^k(i,j)})} \tag{8}$$

$$\mathbf{L}^o(i) = \sum_{k \in \mathcal{N}_i^{inter}} \frac{|\mathcal{T}_i \cap \mathcal{T}_k|}{(\frac{1}{\mathbf{W}^o(i,k)})} \tag{9}$$

$$\mathbf{L}_g^k(t) = \sum_{s \in ch(t)} \frac{|\mathbf{Y}^k(\cdot,t) \cap \mathbf{Y}^k(\cdot,s)|}{(\frac{1}{\mathbf{W}_{gg}^k(t,s)})} \tag{10}$$

where $\mathcal{N}_i^{intra}$ includes directly connected proteins of the i-th protein in the intra-species subnetwork, $\mathcal{N}_i^{inter}$ includes the neighbour proteins of the i-th protein in the inter-species subnetwork; $\mathbf{L}_p^k \in \mathbb{R}^{n_k}$, $\mathbf{L}^o \in \mathbb{R}^N$ and $\mathbf{L}_g^k \in \mathbb{R}^{|\mathcal{T}|}$ separately store the individual walk-lengths of n_k proteins in the intra-species subnetwork, N proteins in the inter-species subnetwork, and $|\mathcal{T}|$ terms in the GO subnetwork of the heterogenous network. $\frac{1}{\mathbf{W}_{pp}^k(i,j)}$, $\frac{1}{\mathbf{W}^o(u,q)}$ and $\frac{1}{\mathbf{W}gg^k(t,s)}$ are inverse similarity (distance) between respective nodes. The more similar the two nodes are, the smaller the distance between them is. If a node embody more annotation information and is densely connected with other nodes, then it will have a large individual walk-length. In this way, we can quantify the individual walk-lengths of a node by summing the gravity-like values (according to Eqs. (8-10)) with respect to its neighbors.

The quantified individual walk-lengths are in different scales, we further re-scale these individual walk-lengths into the same scale $[0, l]$ ($l \in \mathbb{N}_+$) as follows:

$$\tilde{\mathbf{L}}_p^k(i) = \frac{\mathbf{L}_p^k(i) - min(\mathbf{L}_p^k)}{max(\mathbf{L}_p^k) - min(\mathbf{L}_p^k)} \times l \tag{11}$$

$\tilde{\mathbf{L}}^o(i)$ and $\tilde{\mathbf{L}}_g^k(i)$ are similarly re-scaled using Eq. (11).

2.2.2 Asynchronous random walk

Based on the individual walk-lengths of different nodes, we introduce an asynchronous random-walk algorithm on the heterogeneous network to accomplish cross-species protein function prediction. The asynchronous random walks can be divided into three cases.

Case 1: random walk on the intra-species subnetwork. From Fig. 1, we can find that $GO2$ is a missing annotation of $p1$ and but a known annotation of $p2$. A random walker on the intra-species network can start from $p1$ to $p2$ and then to $GO2$. As a result, we can recover the association between $p1$ and $GO2$. This process can be formulated as below:

$$\mathbf{F}1_\tau^k(i,t) = \begin{cases} \alpha \sum_{j=1}^{n_k} \mathbf{W}_{pp}^k(i,j)\mathbf{F}_{\tau-1}^k(j,t) \\ \quad +(1-\alpha)\mathbf{Y}^k(i,t) & \text{if } \tau \leq \lceil \tilde{\mathbf{L}}_p^k(i) \rceil, \\ \mathbf{F}_{\tau-1}^k(i,t) & \text{otherwise.} \end{cases} \tag{12}$$

where $\mathbf{F}1_\tau^k \in \mathbb{R}^{n_k \times |\mathcal{T}|}$ stores the predicted association probabilities between proteins (k-th species) and GO terms in the τ-th iteration, $\mathbf{F}_{\tau-1}^k \in \mathbb{R}^{n_k \times |\mathcal{T}|}$ (see Eq. (15)) stores the predicted association probabilities between proteins (k-th species) and GO terms in the $(\tau-1)$-th iteration ($\mathbf{F}_0^k = \mathbf{Y}^k$), and $\alpha \geq 0$ controls the probability for a walker staying at the starting node. $\lceil \tilde{\mathbf{L}}_p^k(i) \rceil$ is the integer round up operator. If $\tau > \lceil \tilde{\mathbf{L}}_p^k(i) \rceil$, the random walker starting from protein i will not walk on the network any more and $\mathbf{F}1_\tau^k(i,t) = \mathbf{F}_{\tau-1}^k(i,t)$. In this way, asynchronous random walk is realized.

Case 2: random walk on the inter-species subnetwork. From Fig. 1, we can observe that $GO2$ is a missing annotation of $p8$ but already annotated to $p2$, which is an ortholog of $p8$. A random walker can jump from $p2$ to $p8$ based on the established cross-species bridges and annotate $GO2$ to $p8$. This process can be simulated as follows:

$$\mathbf{F}2_\tau(i,t) = \begin{cases} \alpha \sum_{j=1}^{N} \mathbf{W}^o(i,j)\mathbf{F}_{\tau-1}(j,t) \\ \quad +(1-\alpha)\mathbf{Y}(i,t) & \text{if } \tau \leq \lceil \tilde{\mathbf{L}}^o(i) \rceil, \\ \mathbf{F}_{\tau-1}(i,t) & \text{otherwise.} \end{cases} \tag{13}$$

where $\mathbf{F}2_\tau^k(i,t)$ represents the predicted association probability between protein i and term t. If $\tau > max(\lceil \tilde{\mathbf{L}}^o \rceil)$, the walker will not walk on the inter-species subnetwork any more.

1545-5963 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2943342, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Case 3: random walk on the GO hierarchy
From Fig. 1, we can see that $GO4$ is a missing annotation of $p2$ and its ancestor $GO2$ is annotated to this protein, then a random walker can jump from $p2$ to $GO2$ and then to $GO4$. This process can be formulated as below:

$$\mathbf{F}3^k_\tau(i,t) = \begin{cases} \alpha \sum_{v \in par(t)} \mathbf{F}^k_{\tau-1}(i,v)\mathbf{W}^k_{gg}(v,t) \\ +(1-\alpha)\mathbf{Y}^k(i,t) & \text{if } \tau \leq \lceil \tilde{\mathbf{L}}^k_g(t) \rceil, \\ \mathbf{F}^k_{\tau-1}(i,t) & \text{otherwise.} \end{cases} \tag{14}$$

where $par(t)$ includes the direct parental terms of t. $\mathbf{F}3^k_\tau(i,t)$ represents the predicted association probability between protein i and term t based on the random walk on the GO hierarchy in the τ-th iteration. If $\tau > \lceil \tilde{\mathbf{L}}^k_g(t) \rceil$, the random walker on GO hierarchy will not jump any more, and $\mathbf{F}3^k_\tau(i,t) = \mathbf{F}^k_{\tau-1}(i,t)$.

After each iteration, we further fuse $\mathbf{F}1^k_\tau(i,t)$, $\mathbf{F}2^k_\tau(i,t)$ and $\mathbf{F}3^k_\tau(i,t)$ into $\mathbf{F}^k_\tau$ as follows:

$$\mathbf{F}^k_\tau = max(\mathbf{F}1^k_\tau(i,t), \mathbf{F}2^k_\tau(i,t), \mathbf{F}3^k_\tau(i,t)) \tag{15}$$

In practice, we tried to integrate the above three types of predictions by average (or point multiplication). The obtained results in Table S7 and Section 4 of the supplementary file are generally lower than those obtained by Eq. (15). A more advanced fusion is a future pursue. By iteratively applying Eqs. (12-15), we can gradually predict the associations between N proteins and $|\mathcal{T}|$ GO terms, and the asynchronous random walk is realized with the individual walk-lengths on different subnetworks. We want to remark that, the larger the value of $\mathbf{F}^k_\tau(i,t)$ is, the larger the probability protein i annotated with term t is. In Section 5 and Table S8 of the supplementary file, we also recorded the prediction results based on random walk on the whole heterogeneous network, we find that AsyRW gives better results than random walk on the whole network.

3 Results and Analysis

3.1 Experimental Protocol

3.1.1 Datasets

To investigate the effectiveness of AsyRW, we conduct experiments on three model species (Homo sapiens, Mus musculus and Rattus norvegicus). The wide application of high-throughput bio-technologies produce various proteomics/genomics data, which have been used for predicting protein functions. We choose seven representative protein datasets described in Table 1 for experiments, and report the statistics of these data sources in Table S1 of the supplementary file. Other types of proteomic/genomics data can also be adopted. Our main focus in this work is how to achieve cross-species protein function prediction by leveraging the intra-species, inter-species relationships between proteins, associations between proteins and GO terms, and the hierarchical relationships between GO terms. These biological data are then used to construct functional association networks $\mathbf{A}^k_m (m = 1, \cdots, M; k = 1, \cdots, K)$. For the PPI data, available interactions between proteins stored in the respective databases are directly used to specify $\mathbf{A}^k_m$. For the gene expression data, the network is constructed using Pearson correlation coefficients. For the miRTarBaes, miRWalk and Pfam data sources, the network is constructed using Jaccard similarity. For the amino acids data collected from UniPort, we utilize PSI-BLAST (E-value = $1e^{-5}$) score [14] to construct the network. The weighted adjacency matrices of these networks are in different scales. Here, we use Eq. (1) to normalize $\mathbf{A}^k_m$ into the same scale $[0, 1]$. The inter-species relationships between proteins are also specified using the PSI-BLAST score [14].

Following the experimental protocol in Critical Assessment of protein Function Annotation algorithms (CAFA) [1], [2], we train AsyRW using the historical GO annotations and validate the predictions using the recent annotations. The recent GO file and GO annotation files were archived on 2018-11-24, and the historical GO file and annotation files were archived on 2016-05-07. GO file contains the hierarchical relationship between GO terms, which are organized in three sub-ontology, namely BPO, MFO and CCO. The GO annotations file stores the direct associations between proteins and GO terms. We excluded annotations with evidence code 'IEA' (Inferred from Electronic Annotation) to avoid bias toward circular prediction. Next, we augment annotations based on direct annotations and GO hierarchy. Particularly, if a term is annotated to a protein, then ancestor terms of this term are inherently annotated to the protein. Based on these augmented annotations, we initialize the protein-term association matrix $\mathbf{Y}^k$. We applied the above preprocess on two versions of GO file and annotations file. GO terms available in both archived GO files are used in the following experiments to avoid the influence of GO change. Myers *et al.* [38] suggested that terms annotated to very few proteins are hard to be validated by wet-lab experiments and of no interests to biologists. Given that, we selected terms annotated to at least 3 proteins in each species for experiments. The statistics of processed annotations of proteins in Homo sapiens, Mus musculus and Rattus norvegicus are listed in Table 2. We see that during the two years interval, a number of new annotations are appended to the proteins. The three species not only share many GO terms, but also have its distinct GO terms. For example, on CCO branch, the number of intersect (shared) terms of three species is 654, and the number of all involved terms is 935.

3.1.2 Comparing Methods

We compare the performance of AsyRW against SW [39], MashUP [10], SimNet [9], NewGOA [24], PSI-BLAST, InterGFP [22], HPhash [40], GOLabeler [19] and PrimAlign [41]. The first four comparing methods mainly work for single species protein function predictions and the left comparing methods are for cross-species. The inclusion of the first four comparing methods is to investigate the effectiveness of cross-species protein function

1545-5963 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2943342, IEEE/ACM Transactions on Computational Biology and Bioinformatics

TABLE 1
Information of seven biological data sources for experiments.

Describe	Data source	Version
protein-protein interaction	BioGRID (https://thebiogrid.org/)	3.4.159
protein expression	GEO (https://www.ncbi.nlm.nih.gov/geo/)	GDS5282,GDS6248,GDS5269
protein-miRNA interaction	miRTarBase (http://mirtarbase.mbc.nctu.edu.tw/php/download.php)	7.0
protein-miRNA interaction	miRWalk (http://zmf.umm.uni-heidelberg.de/apps/zmf/mirwalk/)	3.0
protein domain	Pfam (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/)	31.0
protein sequence	UniProt (https://www.uniprot.org/)	2018_04
protein-protein interaction	STRING (https://string-db.org/)	10.5

TABLE 2
Statistics of GO annotations. ♯proteins (♯terms) is the number of involved proteins (GO terms). 'history' counts the numbers of annotations in the historical annotation file (archived date: 2016-05-07) and 'recent' counts the numbers in the recent annotation file (archived date: 2018-11-24).

	♯proteins	CCO			MFO			BPO		
		history	recent	♯terms	history	recent	♯terms	history	recent	♯terms
Homo sapiens	19953	254397	298667	844	108673	135590	1671	492663	585671	5868
Mus musculus	16504	227801	252991	889	97782	122601	1690	531694	646476	6527
Rattus norvegicus	7847	81530	137088	723	44810	67512	1311	209687	397123	5660
Intersect terms				654			1235			4996
Union terms				935			1812			6896

prediction. MashUP, SimNet, NewGOA and GOLabeler were introduced in the Introduction Section. NewGOA [24] originally only uses PPI network, here it uses the composite functional network (intra-species network) to fuse multiple data. SW [39] fuses multiple functional association networks to predict protein function. It firstly optimizes the network weight coefficients with respect to a group of related GO terms. Next, it combines multiple networks with the optimized weights into a composite network and predicts protein functions by propagating annotations on the composite network. PSI-BLAST retrieves homologous proteins from multiple species and then adopt the majority vote for function prediction. InterGFP [22] is an interspecies solution that makes use of Best Match Average metric [42] to measure the semantic similarity between proteins and then adopt majority vote for function prediction. HPhash [40] uses a hierarchy preserving hashing technique to encode massive GO terms via compact binary codes and compresses GO annotations of proteins of different species into the low-dimensional hash space. It then performs PSI-BLAST based function prediction in the hash space, and finally maps the predictions back to the original space of terms.

Network alignment methods can also accomplish function transfers between proteins across species [43], [44]. These methods first align PPI networks of different species, and then transfer functional annotations between the aligned nodes based on the observation that the aligned protein pairs are generally annotated with the same GO terms. In contrast, AsyRW first uses PSI-BLAST tool and amino acids to establish the homology relationships between proteins from different species, and then transfers functional annotations of proteins based on these relationships. In addition, AsyRW uses the PPI networks and GO hierarchy to dynamically predict protein functions within and between species. For a protein without an aligned protein in another species, AsyRW can predict its functions based on the functions of its homology (or interacting) proteins, but network alignment based approaches may not. For a comprehensive comparison, PrimAlign [41], a recent network alignment approach which incorporates the principle of PageRank and Markov chain model to globally align nodes, is also included for experiments. The input parameters of these comparing methods are specified as the authors stated in the paper (code), or optimized in the suggested ranges. As to AsyRW, $\alpha = 0.1$, $l = 15$. We will investigate these parameters in the next subsections.

3.1.3 Evaluation Metrics

The performance of protein function prediction can be measured from different aspects [2]. To quantitatively and comprehensively compare the predictions made by these comparing methods, we adopt eight evaluation metrics, which include *MicroAvgF1*, *MacroAvgF1*, *RankingLoss*, *Coverage*, *AvgPrecision*, *AvgAUC*, *Fmax* and *Smin*. These metrics were extensively used in many studies [2], [22]. Since protein function prediction can be viewed as a multi-label classification problem [7], [45], [46], the first five metrics are representative multi-label learning evaluation metrics. Their formal definitions can be found in [45], [46]. *AvgAUC*, *Fmax* and *Smin* are suggested and used by CAFA [1], [2]. *AvgAUC* firstly computes receiver-operating characteristics (ROC) curve for each GO term under different thresholds, and then takes the average Area Under the ROC curve (AUC) of these GO terms. *Fmax* is the overall maximum harmonic mean of precision and recall across all possible thresholds on the predicted protein-term association matrix $\mathbf{F}$. *Smin* uses information theoretic analogs of precision and recall based on the GO hierarchy to measure the minimum semantic distance between the predictions and ground-truths across all possible thresholds. The formal definition of the last three evaluation metrics can be found

1545-5963 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2943342, IEEE/ACM Transactions on Computational Biology and Bioinformatics

in [2]. The higher the value of these evaluation metrics (except *RankingLoss*, *Coverage* and *Smin*), the better the performance is. Since these metrics evaluate the performance from different aspects, it is difficult for one approach to consistently outperform the other one across all the metrics.

3.2 Prediction on GO annotations archived in different years

We use the processed historical (archived in 2016) GO annotations of three model species to train these comparing methods, and then validate their predictions by referring to recent (archived in 2018) annotations. The experimental results of these comparing methods are reported in Table 3 (Homo sapiens), and Table S2 (Mus musculus), Table S3 (Rattus norvegicus) of the supplementary file. In these tables, data highlighted in **boldface** is the best result among all the comparing methods under each evaluation metric. We also count the win and loss of AsyRW with respect to each comparing method in 72 configurations (3 species $\times$ 3 branches $\times$ 8 metrics) and report the counts in Table 4.

TABLE 4
Statistics of the win/loss of AsyRW with comparing methods, out of 72 configurations.

		SW	MashUP	SimNet	NewGOA	
Single species	win	97.22%	90.28%	88.89%	91.67%	
	lose	2.78%	9.72%	11.11%	8.33%	
		PSI-BLAST	PrimAlign	InterGFP	HPhash	GOlabeler
Cross species	win	93.06%	98.61%	93.06%	73.61%	87.50%
	lose	6.94%	1.39%	6.94%	26.39%	12.50%

AsyRW achieves a better performance than these comparing methods across most evaluation metrics. In addition, cross-species solutions generally outperform single-species approaches (SW, MashUP, SimNet and NewGOA), supporting the GO annotations of proteins are complementary among species. AsyRW not always have the best results across all the evaluation metrics, since these metrics measure the predictions from different perspectives and have their own preferences.

Both AsyRW and NewGOA utilize protein interactions and GO structure to construct a heterogenous network, and then apply random walks on it for function prediction. But AsyRW frequently outperforms NewGOA. The causes are two fold: (i) AsyRW performs asynchronous random walk with each node having its own walk-lengths, whereas NewGOA can only specify two fixed walk-lengths; (ii) AsyRW can bi-directionally transfer annotations between species, whereas NewGOA only transfers annotations within species. Because of the two causes, AsyRW also achieves better results than single species data fusion solutions (SW, MashUP and SimNet). Although InterGFP can transfer annotations between species using the semantic similarity between proteins, it often loses to single-species comparing methods, and says nothing of cross-species ones. That is because the incomplete and imbalanced annotations of proteins can not support a semantic similarity as reliable as the similarity derived from other biological data. Furthermore, InterGFP can only transfer annotations from one species to another one, whereas AsyRW can dynamically and bidirectionally transfer annotations between species. Both GOLabeler and PSI-BLAST only make use of sequence data to accomplish cross-species function prediction, and they do not concretely account for the mutually complementary annotations of proteins across species. For these reasons, they both frequently lose to AsyRW. AsyRW also outperforms the network alignment based method PrimAlign. In fact, we found that the number of aligned protein pairs is smaller than that of inter-species connected proteins determined by PSI-BLAST. In addition, AsyRW uses the GO hierarchy. As a result, AsyRW makes a better use of topology information of the heterogeneous network than PrimAlign, and gives a better prediction performance.

HPhash outperforms other comparing methods under several evaluation metrics, such as *Fmax* and *Smin*. That is because GO annotations are quite sparse, which make accurately predicting the associations between proteins and a large number of GO terms a quite hard job. HPhash performs label (GO term) compression and is less affected by sparse annotations. The contribution of compressing multiple terms is also supported by the improved results of HPhash to PSI-BLAST under *MacroAvgF1*, *Fmax* and *Smin*.

Since the experiments are conducted in a history to recent protocol, no standard deviation is reported. To further check the statistical significance between these comparing methods, we adopt the widely-used signed-rank test (p-value threshold 0.05) [47] to assess the significance between AsyRW and these comparing methods on multiple datasets and evaluation metrics, and find the p-values are all smaller than 10^{-3}, which indicates that AsyRW outperforms these comparing methods by chance is close to zero. Given these results and statistics, we can conclude that AsyRW is more capable to predict the GO annotations of proteins. The superiority of AsyRW is mainly contributed by asynchronous random walk for dynamic and bidirectional function transfer between species.

We also conducted experiments to study how different species contribute to the prediction task, and how changing the number of species helps or hurts the accuracy. We reported the results in Section 1 and Table S4 of the supplementary file. Overall, these results indicate that functional annotations aggregated from more homologous species contribute to a more significant prediction performance. In addition, we also tried other ensemble strategies to combine the obtained predictions from case 1, 2 and 3 in Section 2.2.2. These results are provided in Section 6 of the supplementary file.

1545-5963 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2943342, IEEE/ACM Transactions on Computational Biology and Bioinformatics

TABLE 3
Results on archived GO annotations of *Homo sapiens*. ↓ means the lower the value, the better the performance is.

			MicroAvgF1	MacroAvgF1	AvgAUC	AvgPrecision	Fmax	RankLoss↓	Smin↓	Coverage↓
BPO	Single species	SW	0.8452	0.8352	0.9141	0.8333	0.7386	0.0186	14.3895	1136.6593
		MashUP	0.8425	0.8348	0.9000	0.8301	0.7709	0.0192	7.3375	1142.7047
		SimNet	0.0717	0.1398	0.9181	0.1133	0.8622	0.0166	5.5782	1151.5466
		NewGOA	0.8388	0.7772	0.9068	0.8348	0.8256	0.0175	6.2734	933.2567
	Cross species	PSI-BLAST	0.8518	0.8241	0.9155	0.8652	0.7759	0.0451	14.4123	1211.2308
		PrimAlign	0.8092	0.7098	0.9028	0.7813	0.8047	0.1305	12.1881	1766.1570
		InterGFP	0.8135	0.8181	0.8791	0.7818	0.7728	0.1707	7.2445	1730.6965
		HPhash	0.8506	**0.8537**	0.9076	**0.8898**	**0.9398**	0.0611	**4.2905**	1738.9219
		GOLabeler	0.8438	0.8234	0.8837	0.8386	0.7061	0.1200	12.0762	2028.0257
		AsyRW	**0.8655**	0.7893	**0.9491**	0.8719	0.8444	**0.0113**	7.5351	**776.0617**
MFO	Single species	SW	0.8437	0.8210	0.9147	0.8562	0.8349	0.0153	2.6768	163.3105
		MashUP	0.8526	0.8409	0.9069	0.8889	0.8639	0.0156	2.6363	165.8412
		SimNet	0.0078	0.0102	0.9391	0.1059	0.8802	0.0126	2.6108	170.4164
		NewGOA	0.8518	0.7867	0.9110	0.8933	0.8892	0.0116	2.4949	111.8337
	Cross species	PSI-BLAST	0.8676	0.8325	0.9486	0.9018	0.8630	0.0321	3.0622	145.9895
		PrimAlign	0.8093	0.7364	0.9103	0.8220	0.8661	0.0862	3.0841	260.7224
		InterGFP	0.7984	0.8226	0.8961	0.8083	0.8498	0.1471	2.6591	335.5220
		HPhash	0.8576	**0.8687**	0.9185	0.8944	**0.9416**	0.0503	2.4475	274.6675
		GOLabeler	0.8725	0.8586	0.9375	0.9087	0.8559	0.0314	2.9266	147.1112
		AsyRW	**0.8907**	0.8217	**0.9711**	**0.9279**	0.9014	**0.0060**	**2.2946**	**72.5635**
CCO	Single species	SW	0.8723	0.7640	0.8854	0.8211	0.8077	0.0223	3.1272	99.0686
		MashUP	0.9047	0.7886	0.8763	0.9306	0.8465	0.0091	2.3189	90.2753
		SimNet	0.0161	0.0437	0.9014	0.1688	0.8822	0.0079	2.1919	121.1863
		NewGOA	0.9081	0.7541	0.8974	0.9343	0.9029	0.0093	2.2380	88.5321
	Cross species	PSI-BLAST	0.9054	0.7716	0.8979	0.9312	0.8664	0.0219	2.5901	117.1352
		PrimAlign	0.8555	0.6649	0.8867	0.8465	0.8797	0.0971	2.5056	218.8075
		InterGFP	0.8473	0.7588	0.8568	0.8355	0.8400	0.1409	2.2967	188.2302
		HPhash	0.8883	**0.7964**	0.8830	0.9080	**0.9469**	0.0427	**1.8632**	218.3693
		GOLabeler	0.9108	0.7851	0.8663	0.9326	0.8409	0.0348	2.4961	177.5298
		AsyRW	**0.9157**	0.7693	**0.9423**	**0.9447**	0.9048	**0.0062**	2.2210	**72.8292**

3.3 Case study

To further clarify the strength of AsyRW, we selected 10 proteins with more than 35 new annotations (between 2016 and 2018) in CCO of Homo sapiens for case study. We counted how many annotations are added for these selected proteins, and how many of these annotations are correctly predicted by AsyRW and by comparing methods. We report these counts in Table 5. The numbers in (*) represent the overlap between the annotations predicted by AsyRW and by another method.

AsyRW can more accurately predict annotations of proteins than other methods in most cases. For example, 'ATPB' is additionally annotated with 52 terms from 2016-05 to 2018-11, AsyRW correctly predicts 40 of them, while all the other comparing methods correctly predict fewer than 35 of them. We also observe that different methods have different predictive performance, probably due to the various preferences of them. Among them, AsyRW not only predicts more correct functional annotations, but also almost always has larger overlaps with them than any pair of these comparing methods, which slightly complement the predictions of AsyRW. In summary, these examples again prove the effectiveness of AsyRW.

3.4 Contribution of different components

In this subsection, we conduct experiments to investigate the contribution of three subnetworks of the constructed heterogeneous network. For this investigation, we introduce three variants of AsyRW: (1) AsyRW-noGO disregards the GO hierarchy; (2) AsyRW-noIntra discards the intra-relations between proteins of the same species; (3) AsyRW-noInter discards the inter-relations between proteins of different species.

Following the same experimental protocol used in previous subsections, we record the results of AsyRW and these variants, and plot them in Fig. 2 and Fig S1 of the supplementary file. We can observe that AsyRW generally holds the best performance among its variants. An interesting observation is that AsyRW-noIntra has the lowest *AvgAUC* and *Fmax* on Homo sapiens and Mus musculus, whereas AsyRW-noInter gets the lowest *AvgAUC* and *Fmax* on Rattus norvegicus. That is because the functional annotations of Rattus norvegicus are more sparse and imbalanced than the other two species, and the statistics in Table 2 also suggest the sparseness and imbalance. Therefore, the predicted functional annotations of Rattus norvegicus are dependent on the orthologous relationships between proteins, which serve the function transfer bridges between species. On the other hand, the prediction on Homo sapiens and Mus musculus are more dependent on intra-species subnetwork. That is because these two species have high-homology and their annotations are often bi-referenced. AsyRW-noInter always has the highest *Smin*. That is because *Smin* evaluates the performance from the perspective of GO hierarchy, and inter-species relationships between proteins generally contributes much more than intra-species relationships. From these analysis, we can conclude that the inter- and intra-species relationships, and the GO hierarchy should be leveraged for protein function prediction. These investigations also suggest that our constructed heterogeneous network can effec-

1545-5963 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2943342, IEEE/ACM Transactions on Computational Biology and Bioinformatics

TABLE 5
Statistics of new annotations of ten proteins in CCO of *Homo sapiens* during 2016-05 and 2018-11, and the number of annotations correctly predicted by each comparing method. The numbers in (*) count the number of annotations predicted by AsyRW and also by another method.

	ATPB	ATPA	AP4AT	NSG2	ADCY8	HIP1R	ATP5E	ATPD	ATPG	DYN1
New annotations	52	49	46	45	43	42	38	38	38	36
SW	34(34)	31(31)	26(26)	18(18)	14(13)	13(12)	18(18)	18(18)	18(17)	17(17)
MashUP	34(34)	31(31)	26(26)	18(18)	14(13)	13(12)	18(18)	18(18)	18(17)	17(17)
SimNet	13(11)	10(8)	6(5)	4(4)	2(2)	3(1)	2(0)	2(0)	2(0)	2(0)
NewGOA	1(0)	1(0)	1(0)	2(0)	15(13)	13(12)	1(0)	1(0)	1(0)	18(18)
PSI-BLAST	30(29)	28(27)	24(24)	21(16)	19(18)	15(11)	20(20)	17(17)	15(15)	22(22)
PrimAlign	1(0)	1(0)	27(22)	21(16)	14(10)	11(9)	28(26)	13(13)	1(0)	18(17)
InterGFP	1(0)	1(0)	1(0)	2(0)	14(13)	13(12)	1(0)	1(0)	1(0)	17(17)
HPhash	2(1)	1(1)	3(3)	4(3)	3(2)	2(1)	1(1)	1(1)	1(1)	1(1)
GOLabeler	31(31)	30(30)	26(24)	19(16)	20(20)	12(11)	27(27)	27(26)	20(19)	19(19)
AsyRW	40	39	26	18	20	12	27	27	26	26

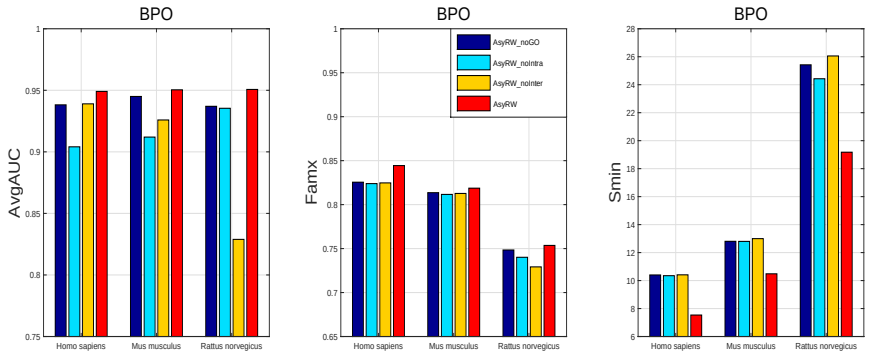


Fig. 2. The *AvgAUC*, *Fmax* and *Smin* values of AsyRW and its variants on BPO. Different from *AvgAUC* and *Fmax*, the lower the value of *Smin*, the better the performance is.

tive integrate different types of biological data.

In addition, we conducted another set of experiments to investigate the prediction results using only GO hierarchy, intra-relations between proteins, and inter-relations between proteins. In other words, case 1, 2 and 3 in Section 2.2.2 are separately used for function prediction. We present the results in Figure S2 and analysis of the supplementary file. We find that only using GO hierarchy, intra-relations or inter-relations between proteins give lower results than AsyRW.

3.5 Individual walk-length is better than fixed walk-length

To study the contribution of our proposed individual walk-lengths, we also test the performance of AsyRW with fixed walk-lengths for all the nodes by varying walk-length in the three subnetwork from 0 to 5, respectively. Fig. 3 reveals the *AvgAUC*, *Fmax* and *Smin* of AsyRW under different combined configurations of $\mathbf{L}_p^k$, $\mathbf{L}^o$ and $\mathbf{L}_g^k$. For comparison, we also plot the results of AsyRW with $l = 5$ (highlighted by red star) in Fig. 3. From this figure, we can clearly see that the *AvgAUC* and *Fmax* stop increasing and *Smin* stops decreasing, once $\mathbf{L}_p^k$, $\mathbf{L}^o$ and $\mathbf{L}_g^k$ are larger than 2. AsyRW with individual walk-lengths generally obtains better results than the counterpart with fixed walk-length. This study further corroborates the effectiveness of individual walk-lengths.

3.6 Runtime and parameter sensitivity analysis

We also record the runtime cost of AsyRW and other comparing methods, and report them in Table S9 of the supplementary file. Because of the multiple networks integration, specification of each subnetworks, the individual walk-length of each node, AsyRW does not show an advantage in runtime, but it almost always achieves higher accuracy than these comparing methods. In addition, we conduct experiment to investigate the sensitivity of AsyRW to random walk restart probability α and the scale of walk-length l. From Figure S3-S5 of the supplementary file, we can find that AsyRW can easily select input parameters with effectiveness.

4 Conclusions

Research interests of biologists and the experimental ethics cause the functional annotations of proteins across species are rather imbalanced and incomplete. We introduce an asynchronous random walk based solution called AsyRW to alleviate the above issues. AsyRW firstly leverages complementary annotations of different species, GO hierarchy and multi-level biological data related to proteins to construct a composite heterogeneous

1545-5963 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2943342, IEEE/ACM Transactions on Computational Biology and Bioinformatics

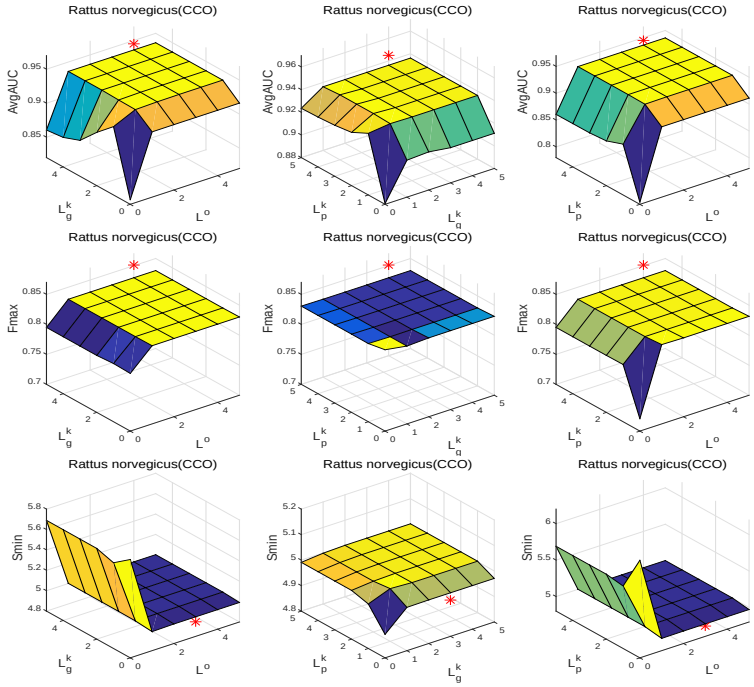


Fig. 3. The *AvgAUC*, *Fmax* and *Smin* values of AsyRW with different fixed walk-lengths. The red star is the value of AsyRW with individual walk-lengths.

network (composed with proteins and GO terms). Next, it executes asynchronous random walk by restricting each node having its own walk-length to predict the associations between proteins and terms. Extensive experimental results show that AsyRW often outperforms other related competitive solutions across various evaluation metrics. Our study suggests that the annotations of proteins of different species are mutually complementary to each other, and they can be collaboratively used to improve protein function prediction. Both the cross-species relationships between proteins and individual walk-lengths in asynchronous random walk can boost the prediction performance.

5 Acknowledgements

This work is supported by Natural Science Foundation of China (61741217, 61872300, 61873214, 61871020, 61571163 and 61532014), Fundamental Research Funds for the Central Universities (XDJK2019B024), the National Key Research and Development Plan Task of China (Grant No. 2016YFC0901902), Natural Science Foundation of CQ CSTC (cstc2018jcyjAX0228).

References

[1] P. Radivojac, W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, and A. Ben-Hur, "A large-scale evaluation of computational protein function prediction," *Nature Methods*, vol. 10, no. 3, pp. 221–227, 2013.

[2] Y. Jiang, T. R. Oron, W. T. Clark, A. R. Bankapur, D. D'Andrea, R. Lepore, C. S. Funk, I. Kahanda, K. M. Verspoor, and A. Benhur, "An expanded evaluation of protein function prediction methods shows an improvement in accuracy:," *Genome Biology*, vol. 17, no. 1, p. 184, 2016.

[3] A. Shehu, D. Barbará, and K. Molloy, *A survey of computational methods for protein function prediction*. Switzerland: Springer International Publishing, 2016.

[4] M. S. Alexandra, C. R. David, W. T. Alexander, C. B. Patricia, and F. Iddo, "Biases in the experimental annotations of protein function and their effect on our understanding of protein function space," *PLoS Computational Biology*, vol. 9, no. 5, p. e1003063, 2013.

[5] T. G. O. Consortium, M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, and S. S. Dwight, "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–9, 2000.

[6] J. A. Blake, "Ten quick tips for using the gene ontology," *PLoS Computational Biology*, vol. 9, no. 11, p. e1003343, 2013.

[7] G. Yu, C. Domeniconi, H. Rangwala, G. Zhang, and Z. Yu, "Transductive multi-label ensemble classification for protein function prediction," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 1077–1085.

[8] L. Lan, N. Djuric, Y. Guo, and S. Vucetic, "Msknn: protein function prediction by integrating multiple data sources," *BMC Bioinformatics*, vol. 14, no. 3, p. S8, 2013.

[9] G. Yu, G. Fu, J. Wang, and H. Zhu, "Predicting protein function via semantic integration of multiple networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 2, pp. 220–232, 2016.

[10] H. Cho, B. Berger, and J. Peng, "Compact integration of multi-network topology for functional analysis of genes," *Cell Systems*, vol. 3, no. 6, p. 540, 2016.

[11] G. Yu, H. Rangwala, C. Domeniconi, G. Zhang, and Z. Zhang, "Predicting protein function using multiple kernels," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 1, pp. 219–233, 2015.

[12] M. Frasca, "Automated gene function prediction through gene multifunctionality in biological networks," *Neurocomputing*, vol. 162, pp. 48–56, 2015.

[13] R. Cao and J. Cheng, "Integrated protein function prediction by mining function associations, sequences, and protein-protein and gene-gene interaction networks," *Methods*, vol. 93, pp. 84–91, 2016.

[14] Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[15] A. Mitrofanova, V. Pavlovic, and B. Mishra, "Prediction of protein functions with gene ontology and interspecies protein homology

1545-5963 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

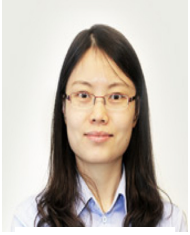
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2943342, IEEE/ACM Transactions on Computational Biology and Bioinformatics

data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 3, pp. 775–784, 2011.
[16] C. Y. Park, A. K. Wong, C. S. Greene, J. Rowland, Y. Guan, L. A. Bongo, R. D. Burdine, and O. G. Troyanskaya, "Functional knowledge transfer for high-accuracy prediction of under-studied biological processes," *PLoS Computational Biology*, vol. 9, no. 3, p. e1002957, 2013.
[17] V. Vidulin, T. Šmuc, and F. Supek, "Extensive complementarity between gene function prediction methods," *Bioinformatics*, vol. 32, no. 23, pp. 3645–3653, 2016.
[18] M. Kulmanov, M. A. Khan, and R. Hoehndorf, "Deepgo: predicting protein functions from sequence and interactions using a deep ontology-aware classifier," *Bioinformatics*, vol. 34, no. 4, pp. 660–668, 2017.
[19] R. You, Z. Zhang, Y. Xiong, F. Sun, H. Mamitsuka, and S. Zhu, "Golabeler: Improving sequence-based large-scale protein function prediction by learning to rank," *Bioinformatics*, vol. 34, no. 14, 2018.
[20] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," in *24th International Conference on Machine Learning*, 2007, pp. 129–136.
[21] Y. Jiang, W. T. Clark, I. Friedberg, and P. Radivojac, "The impact of incomplete knowledge on the evaluation of protein function prediction: a structured-output learning perspective," *Bioinformatics*, vol. 30, no. 17, p. i609, 2014.
[22] G. Yu, W. Luo, G. Fu, and J. Wang, "Interspecies gene function prediction using semantic similarity," *BMC Systems Biology*, vol. 10, no. 4, p. 121, 2016.
[23] S. Wang, M. Qu, and J. Peng, "Prosnet: Integrating homology with molecular networks for protein function prediction," in *Pacific Symposium on Biocomputing*, 2017, pp. 27–38.
[24] G. Yu, G. Fu, J. Wang, and Y. Zhao, "Newgoa: Predicting new go annotations of proteins by bi-random walks on a hybrid graph," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 4, pp. 1390–1402, 2018.
[25] G. Yu, H. Zhu, C. Domeniconi, and J. Liu, "Predicting protein function via downward random walks on a gene ontology," *BMC Bioinformatics*, vol. 16, no. 1, p. 271, 2015.
[26] W. Peng, M. Li, L. Chen, and L. Wang, "Predicting protein functions by using unbalanced random walk algorithm on three biological networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 2, p. 360, 2017.
[27] G. Valentini, "True path rule hierarchical ensembles for genome-wide gene function prediction." *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 3, pp. 832–847, 2011.
[28] G. Yu, H. Zhu, and C. Domeniconi, "Predicting protein functions using incomplete hierarchical labels," *BMC Bioinformatics*, vol. 16, no. 1, p. 1, 2015.
[29] G. Fu, J. Wang, B. Yang, and G. Yu, "Neggoa: Negative go annotations selection using ontology structure." *Bioinformatics*, vol. 32, no. 19, p. 2996, 2016.
[30] O. D. King, R. E. Foulger, S. S. Dwight, J. V. White, and F. P. Roth, "Predicting gene function from patterns of annotation," *Genome Research*, vol. 13, no. 5, p. 896.
[31] Y. Tao, L. Sam, J. Li, C. Friedman, and Y. A. Lussier, "Information theory applied to the sparse gene ontology annotation network to predict novel gene function," *Bioinformatics*, vol. 23, no. 13, pp. i529–i538, 2007.
[32] E. A. Codling, M. J. Plank, and S. Benhamou, "Random walk models in biology," *Journal of the Royal Society Interface*, vol. 5, no. 25, pp. 813–834, 2008.
[33] J. Zhang, L. Deng *et al.*, "Gene ontology-based function prediction of long non-coding rnas using bi-random walk," *BMC Medical Genomics*, vol. 11, no. 5, p. 99, 2018.
[34] E. Ernesto, "Generalized walks-based centrality measures for complex biological networks," *Journal of Theoretical Biology*, vol. 263, no. 4, pp. 556–565, 2010.
[35] S. F, G. MC, M. A, and B. AL, "A universal model for mobility and migration patterns," *Nature*, vol. 484, no. 7392, p. 96, 2012.
[36] F. Cheng, C. Liu, C.-C. Lin, J. Zhao, P. Jia, W.-H. Li, and Z. Zhao, "A gene gravity model for the evolution of cancer genomes: a study of 3,000 cancer genomes across 9 cancer types," *PLoS Computational Bbiology*, vol. 11, no. 9, p. e1004497, 2015.
[37] L. Lin, T. Yang, L. Fang, J. Yang, F. Yang, and J. Zhao, "Gene gravity-like algorithm for disease gene prediction based on phenotype-specific network," *BMC Systems Biology*, vol. 11, no. 1, p. 121, 2017.
[38] C. L. Myers, D. R. Barrett, M. A. Hibbs, C. Huttenhower, and O. G. Troyanskaya, "Finding function: evaluation methods for functional genomic data," *BMC Genomics*, vol. 7, no. 1, p. 187, 2006.
[39] S. Mostafavi and Q. Morris, "Fast integration of heterogeneous data sources for predicting gene function with limited annotation," *Bioinformatics*, vol. 26, no. 14, pp. 1759–1765, 2010.
[40] Y. Zhao, G. Fu, J. Wang, M. Guo, and G. Yu, "Gene function prediction based on gene ontology hierarchy preserving hashing," *Genomics*, vol. 111, no. 3, pp. 334–342, 2019.
[41] K. Kalecky and Y. R. Cho, "Primalign: Pagerank-inspired markovian alignment for large biological networks," *Bioinformatics*, vol. 34, no. 13, pp. i537–i546, 2018.
[42] C. Pesquita, D. Faria, H. Bastos, A. E. Ferreira, A. O. Falco, and F. M. Couto, "Metrics for go based protein semantic similarity: a systematic evaluation," *BMC Bioinformatics*, vol. 9, no. Suppl 5, pp. S4–S4, 2008.
[43] C. S. Liao, K. Lu, M. Baym, R. Singh, and B. Berger, "Isorankn: spectral methods for global alignment of multiple protein networks," *Bioinformatics*, vol. 25, no. 12, pp. 253–8, 2009.
[44] L. Meng, A. Striegel, and T. Milenkovi?, "Local versus global biological network alignment," *Bioinformatics*, vol. 32, no. 20, pp. 3155–3164, 2015.
[45] M. L. Zhang and Z. H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.
[46] N. Cesa-Bianchi, "Synergy of multi-label hierarchical ensembles, data fusion, and cost-sensitive methods for gene functional inference," *Machine Learning*, vol. 88, no. 1-2, pp. 209–241, 2012.
[47] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

Yingwen Zhao is an MPhil student at the College of Computer and Information Sciences, Southwest University, Chongqing, China. She received B.Sc. degree in Internet of Things from Jingss University, Wuxi, China. Her research interests include data mining and bioinformatics.

Jun Wang is an Associate Professor in the College of Computer and Information Science, Southwest University, Chongqing, China. She received B.Sc. degree in Computer Science, M.Eng. degree in Computer Science and Ph.D. in Artificial Intelligence from Harbin Institute of Technology, Harbin, China in 2004, 2006 and 2010, respectively. Her current research interests include machine learning, data mining and their applications in bioinformatics.

Maozu Guo is a professor at the College of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China. He received the Ph.D. degree in Computer Science and Technology from Harbin Institute of Technology. His research interests include bioinformatics, machine learning, and data mining.



Xiangliang Zhang is an Assistant Professor and directs the Machine Intelligence and kNowledge Engineering (MINE) Laboratory in King Abdullahs University of Science and Technology (KAUST). She earned her Ph.D. degree in computer science with great honors from INRIA-University Paris-Sud 11, France, in 2010. Her main research interests and experiences are in diverse areas of machine learning and data mining.

Guoxian Yu is a Professor at the College of Computer and Information Science, Southwest University, Chongqing, China. He received the Ph.D. in Computer Science from South China University of Technology, Guangzhou, China in 2013. His current research interests include data mining and bioinformatics. He severs as reviewers for AAAI, KDD, ICDM, TNNLS, TCBB and other prestigious conferences and journals. He is a recipient of Best Poster Award of SDM2012 Doctral Forum and Best Paper Award of ICML-C2011.

1545-5963 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.