Protein2Vec: Aligning Multiple PPI Networks with Representation Learning

Journal:	*IEEE/ACM Transactions on Computational Biology and Bioinformatics*
Manuscript ID	TCBBSI-2019-04-0181.R1
Manuscript Type:	SI - BIBM 2018
Keywords:	protein representation, multiple network alignment, PPI networks, topological assessment

Protein2Vec: Aligning Multiple PPI Networks with Representation Learning

Jianliang Gao, Ling Tian, Tengfei Lv, Jianxin Wang, Bo Song, and Xiaohua Hu

Abstract—Research of Protein-Protein Interaction (PPI) Network Alignment is playing an important role in understanding the crucial underlying biological knowledge such as functionally homologous proteins and conserved evolutionary pathways across different species. Existing methods of PPI network alignment often try to improve the coverage ratio of the alignment result by aligning all proteins from different species. However, there is a fundamental biological premise that needs to be considered carefully: not every protein in a species can, nor should, find its homologous proteins in other species. In this work, we propose a novel alignment method to map only those proteins with the most similarity throughout the PPI networks of multiple species. For the similarity features of the protein in the networks, we integrate both topological features with biological characteristics to provide enhanced supports for the alignment procedures. For topological features, we apply a representation learning method on the networks that can generate a low dimensional vector embedding with its surrounding structural features for each protein. The topological similarity of proteins from different PPI networks can thus be transferred as the similarity of their corresponding vector representations, which provides a new way to comprehensively quantify the topological similarities between proteins. We also propose a new measure for the topological evaluation of the alignment results which better uncover the structural quality of the alignment across multiple networks. Both biological and topological evaluations on the alignment results of real datasets demonstrate our approach is promising and preferable against previous multiple alignment methods.

Index Terms—Protein representation, multiple network alignment, PPI networks, topological assessment

1 Introduction

1.1 PPI Network Alignment

The comparative analysis of protein-protein interaction (PPI) networks across different species by network alignment is very effective in discovering functional orthologs of proteins among diverse species and identifying conserved subnetworks or motifs in the PPI network [1]. PPI network alignment can be implemented as either one-to-one or many-to-many node mapping by comparing networks based upon various supportive information such as sequence similarity and topology conservation. By aligning PPI networks of multiple species, knowledge such as conserved proteins and complexes can be transferred from well-studied species to poor-studied species.

Network alignment has already been successfully applied in many applications. (1) While plenty of crucial biological and disease processes in a species of interest are experimentally expensive to study, network alignment is capable to serve as a bridge and transfer knowledge from well-studied species, such as yeast Saccharomyces cerevisiae or worm Caenorhabditis elegans, to high-valued but less well-studied species such as human, and consequently lead to new discoveries in system biology. (2) In addition to the knowledge transferring across species, network alignment is also utilized for inferring phylogenetic relationships of different species based on the similarities between their biological networks [2].

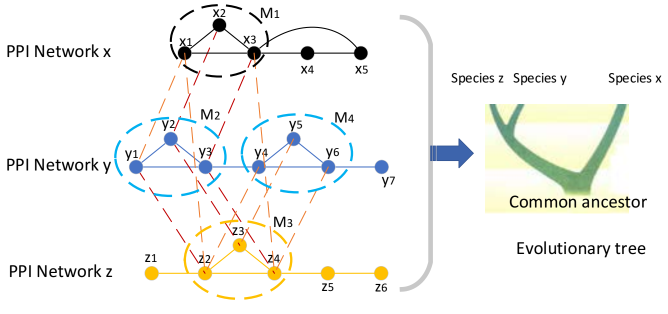


Fig. 1 PPI Network Alignment

According to the number of networks being aligned, the network alignment approaches can be categorized as either pairwise or multiple. Pairwise network alignment aligns only two networks at a time, whereas multiple network alignment aims to align more than two networks at the same time. Aligning multiple networks other than pairwise networks promises additional insights into the complex conservation as well as the knowledge transfer across multiple species. Figure 1 shows an example of three PPI network alignment. The substructure $M_1 = (x_1, x_2, x_3)$ in the network of species x is aligned with $M_2 = (y_1, y_2, y_3)$ in the network of species y; Similarly, M_2 is also aligned with $M_3 = (z_2, z_3, z_4)$ in the the network of species z. Then M_1 being further aligned with M_3 from the consistent perspective can only made possible during a multiple network

- *J. Gao, L. Tian, L Teng, J Wang are with the School of Computer Science and Engineering, Central South Uinversity, Changsha, P.R. China, 410083. E-mail: {gaojianliang,tianling,lvtengfei}@csu.edu.cn.*
- *B. Song, X. Hu.are with the College of Computing & Informatics, Drexel University, Philadelphis, PA, USA, 19104. E-mail: {bosong, xh29}@drexel.edu.*

alignment. To the contrast, $M_4 = (y_4, y_5, y_6)$ might be the best alignment if it was a pairwise alignment between Network y and Network z. In addition, species y and z could share more closed relationship than with species x in this example. Evolutionary tree can hence be further drawn from the results of multiple network alignment.

Network alignment methods can also be categorized as either local alignment or global alignment. Many existing network alignment methods focus on the Local Network Alignment (LNA), which aims to find smaller subnetworks with high similarities, such as protein complexes, that are irrespective of the overall similarity among compared networks [3]. Since the subnetworks can overlap, a node in one network can be mapped to multiple nodes in another network [4],[5]. Consequently, LNA is generally not capable of finding the global best mapping between the input networks [6]. Therefore, most of the recent efforts have been attracted to the Global Network Alignment (GNA), which typically attempts to map the entire network as a whole to other networks and maximize overall similarity of the participating networks [7].

This study mainly focuses on how to improve the alignment of multiple PPI networks from the global perspective.

1.2 Motivations

Although significant progresses have been made in the alignment research of PPI networks, there exist two crucial problems that still need to be solved:

(1) How to better quantify the topological similarity of proteins from different species?

Network alignment provides an effective way to identify conserved protein complexes across multiple PPI networks [8]. The conserved functional and topological features are the two focuses during the alignment, where functional module represents a collection of molecular interactions that work together to achieve a particular functional objective in a biological process, and topological module represents the locally dense neighborhoods in a PPI network [9].

Prior methods attempted with different metrics to obtain the topological similarity and generated various performances without a unified agreement and constancy. Some methods consider only the topological similarity in degree-based measures that capture the graph structure partially instead of retaining more comprehensive structural characteristics of the network. In this study, we try to improve this problem by proposing to apply representation learning on the networks to obtain vector representation of each protein with topological features embedded comprehensively. The topological similarity of proteins across the networks can be consequently converted and easily quantified as unified similarity of vectors.

(2) How many proteins really need to be aligned?

The coverage and consistency are the two most considered metrics to evaluate the effectiveness of a network alignment method. Achieving both high coverage and high consistency at the same time is one important goal when trying to involve all proteins of a network in the alignment procedure. However, other than these overall measures, it is more important to achieve precised alignment among multiple network alignment. Unfortunately, existing alignment methods mostly concern with only the overall coverage. In reality, many proteins from different species should never be aligned as they are not homologous at the first place. To eliminate these commonly adopted limitations, we propose an algorithm that only partially aligns those most homologous proteins from diverse species, which is robust and scalable for multiple networks and can achieve balanced high consistency.

1.3 Contributions

In our research, we propose a new approach for aligning multiple PPI networks. The main contributions are as follows:

- To solve the existing limitation of quantifying the topological features of proteins, we propose to learn the representation of proteins in PPI networks, which creates a vector for each protein. The topological similarity of proteins is then easily transferred by computing the similarity of vector representations of the proteins.
- We propose an effective method to partially align multiple PPI networks with representation learning. With this method, only those proteins that really need to be aligned are considered instead of involving all protiens for the overall coverage. It is also more efficient to find homologous proteins or protein complexes across various species.
- A more comprehensive topological evaluation called mean neighbor similarity (MNS) is introduced. It measures topological quality of alignment result in replacing the conventional measure of overall coverage.
- Extensive experiments are conducted on the real datasets consist of PPI networks from five species. The results for different evaluation measures illustrate the outstanding performances of the proposed method when compared with four widely adopted network alignment methods.

2 Related Works

The general objective of PPI network alignment is to obtain the similarities among proteins from different networks through the graph mapping. The alignment result usually reaches to the highest score of similarity. To determine the protein similarities with the best score, many current network alignment algorithms adopt a cost function combining with the biological and structural properties[10]. For the network structural properties, representation learning is a recently raising approach that could reflect more comprehensive network topology than those conventional measures like degree related approaches. In this section, we review the related research in the PPI network alignment and representation learning.

2.1 Alignment of PPI Networks

Network alignment methods from previous research could be categorized into either or combined categories of local or global, pairwise or multiple [4, 6, 10, 12], while each of them has its own way in attempting to achieve an optimal alignment result.

For pairwise PPI networks, IsoRank as a classic pairwise global alignment algorithm for biological networks is one of the most referred alignment methods in the field [12]. In IsoRank, a version of Google's PageRank algorithm is adopted to estimate the similarity of proteins base on the topology of their respective neighbors. Intuitively, if the neighbors of two proteins from different networks are similar, the two proteins are considered as similar too. Based upon which, IsoRank associates similarity scores for protein pairs and screens out the candidate pairs to construct similarity matrix for the search of final global alignment in a greedy algorithm. MAGNA is another recently proposed pairwise and global network alignment method. It relies on a genetic algorithm that optimizes edge conservation directly and chooses high scoring alignment results according to an objective function that combines both topological and biological factors [4].

There exist more research interests of network alignment that shift toward multiple networks, including IsoRankN[12], SMETANA[13], NetCoffee[18], BEAMS[14], GMAlign[26] and MPGM[27]. Evolved from IsoRank, IsoRankN [12] applies spectral graph theorectic approach that is similar to the PageRank-Nibble algorithm. It can generate aligned clusters globally from multiple networks, where each cluster can contain several proteins from the same network. SMETANA [13] tries to effectively find among multiple networks a maximum global alignment in two stages. It first applies a node cost function bases on a semi-Markov random walk model in order to calculate similarities between nodes that serves as a probabilistic index; then, the aligned node clusters with the maximum expected accuracy can be obtained by a greedy approach. NetCoffee is another global multiple aligner that combines in its objective function both the sequence and topological similarities [6]. It is the first multiple network aligner that measures weights on the edge according to not only bipartite sequence similarity, but also a triplet extension over all networks. This topological approach is also similar to the multiple sequence aligner T-Coffee [25]. Alkan et al., propose a global heuristic algorithm BEAMS that applies a backbone extraction and merge strategy. It maximizes the number of conserved edges between disjoint cliques to repeatedly merge which into identified backbones and form the final aligned protein clusters from multiple PPI networks in a greedy manner [14]. Zhu Y et al present a two stage global network aligner GMAlign[26]. In the first stage, it selects several pairs of important proteins as seeds to obtain an initial protein mapping by expanding the seeds; In the second stage, it refines the initial result to obtain an optimal alignment result iteratively based on the vertex cover. MPGM[27] is a global alignment algorithm recently proposed that generates one-to-one mappings through multiple networks. It firstly maps few proteins and produces an initial set of seed tuples using only protein sequence similarities, and combined which with the structure of networks later to align the rest of unmapped proteins across all networks.

2.2 Representation Learning

Current literature of node embedding technique mainly define the nodes similarity in terms of different types of proximity between the local neighborhood structures of the nodes.

Representation learning of node is the approach that try to obtain similar embeddings for similar nodes while preserving the node features in the embedded space. Deepwalk [15] is the first representation learning algorithms inspired by the word2vec algorithm to learn a language model from a network. It is an unsupervised method which learns adaptable latent representations for the nodes in a network. The representative sequences of nodes are learnt by sampling from a stream of unbiased and truncated random walks over the network, which effectively maps local features into a low-dimensional vector space. Deepwalk has attracted considerable interests in network analysis field as it conveyed the idea of representation learning from language modeling to the realm of networks, spurring extended and fruitful outcomes. Node2vec [16] is proposed as an algorithmic framework to learn the representation of continuous node features and capture in network the observed diverse patterns of connectivity. It produces high quality and informative embeddings through biased second order random walk model to maximize the likelihood of preserving neighborhood features of nodes. Node2vec provides the flexibility in capturing the context of nodes with both homophily and structural equivalence, andcan explore neighborhood diversity efficiently. Struc2vec [17] is another rising framework of representation learning with great novelty and flexibility, which is able to preserve the structural similarity of nodes in the network at different scales and regardless of their proximity. It attempts to learn latent representations of nodes that have similar role in the network with a hierarchy measures by constructing a multilayer graph for topological similarity encoding and structural context generation.

Many applications and downstream tasks related to node embedding become very promising and largely promoted by these recent advances in the representation learning research, but which has not been widely applied in the biological networks analysis nor extended beyond a single network, let alone for multiple biological networks studies.

3 Methods and Algorithms

In order to achieve the optimal alignment result with enhanced supports, we establish a similarity scoring function that could reflect comprehensive information from both functional and structural aspect of the participating species and their networks. Biological characteristics and topological features are well quantified and integrated in our overall scoring function to guide the aligning process and promote the final alignment result. All match connections between proteins across different networks that have high scores form a candidate pool for the later heuristic searching procedure to generate the final optimized alignment result, which only consists of proportional proteins with the most similarities.

By quantifying and denoting the biological similarity between two proteins u and v as $B(u,v)$, and the topological similarity of which as $T(u,v)$, our scoring function below that integrates both features can be formulated as:

$$S(u,v) = \alpha * T(u,v) + (1-\alpha) * B(u,v) \qquad (1)$$

where α is a controllable parameter to weight and balance the contribution of topological and biological similarity towards the overall similarity score of $S(u,v)$.

3.1 Protein Node to Vector

The conserved proteins across species often share similar structural patterns of interactions and have similar functions [18]. Conventional approaches describe the structural feature of proteins mainly with metrics of topology such as the degree. We apply in our research an alternative approach to represent proteins in the PPI network as vectors, utilizing more comprehensive structural features. As struc2vec builds its algorithm based on the intuitive assumption that two proteins should be deemed structurally similar if their neighbors also share same degrees, we propose to further consider with proteinaceous pattern that is more meaningful in PPI networks, where the over-represented triangle motifs (fully connected 3-node subgraph) often act as the basic building block and essential functional units of biological processes [19].

Denote $G = (V, E)$ as a PPI network with node set V and edge set E. We compute in the first step a hierarchic variance H as follows:

$$H_k(u,v) = d\big(t(U_k), t(V_k)\big) + d\big(s(U_k), s(V_k)\big) + H_{k-1}(u,v) \quad (2)$$

where $U_k(\cdot)$ or $V_k(\cdot)$ denotes a node set at k hop away from u or v in G, $s(\cdot)$ denotes the ordered sequence of degree of a node set. $t(\cdot)$ denotes the sequence of number of triangle motif composed with node set k-1 hop away. The function $d(\cdot)$ measures the distance between two sequences. The design of this hierarchy is able to capture structural characteristics of nodes with both neighborhood degree chain and motif features for every two nodes.

In the second step, a weighted k-layer complete graph is constructed for a biased random walk to generate context sequence for each node. The weight on the edge of two nodes in the k^{th} layer is assigned as its normalized hierarchic variance on the total variances of that layer:

$$e_k(u,v) = H_k(u,v) / \textstyle\sum_{v\in V, v\neq u} H_k(u,v) \qquad (3)$$

The weights on the connection of a node u to its upper and lower layers are assigned as $c_{k+1}(u)$ and $c_{k-1}(u)$ separately by:

$$c_{k+1}(u) = \frac{\log(1+\sum_{v\in V, u\neq v} |e_k(u,v) > Q_1(e_k)|\,)}{1+\log(1+\sum_{v\in V, u\neq v} |e_k(u,v) > Q_1(e_k)|\,)} \qquad (4)$$

$$c_{k-1}(u) = 1 - c_{k+1}(u) \qquad (5)$$

where $Q_1(e_k)$ is the lower quartile of all edge weights of the complete graph in the k^{th} layer. Then the biased random walk similar to node2vec is applied on the k-layer graph instead of one, with the in-layer moving probability as $e_k(u,v)$ and cross-layer moving probability as $c_{k+1}(u)$ and $c_{k-1}(u)$, to create neighbors in sequences as its context.

Once the context sequences are generated, we apply word2vec model to effectively learn from the sequences a node embedding and get its latent representation as a low-dimensional vector for each of the proteins.

With the structural property of a node quantified and embedded in a vector, the topological similarity $T(u,v)$ can be readily transformed by calculating the vector similarity with various choices of coefficient. We apply the cosine measure for the vector similarity calculation:

$$T(u,v) = \cos(\mathrm{u,v}) = \frac{emb(u)\cdot emb(v)}{|emb(u)| * |emb(v)|} \qquad (6)$$

where emb(.) is the low-dimensional vector representation of a protein node embedded with its topological features. We further conduct normalization on the results to have topological similarity scores of all proteins fall in the same scope of [0,1] for fair comparison. The closer the $T(u,v)$ of two proteins is to the value 1 the more similar their topological features are in their own networks.

3.2 Protein Sequence to Biological Similarity

Besides topological features of interactions in a network, a protein also has its biological identity, such as the amino acid sequence, which can be used to assess from biological aspect its homology relationships with other proteins. Higher similarity between protein sequences indicate greater likelihood of them having similar molecular functions [8].

We take biological similarity into consideration to support and complement our scoring function in guiding the alignment process to a more compelling result. We determine the biological similarity $B(u,v)$ between proteins as our previous work [20] by comparing their biological significance of homology, which is quantified as a statistical index called Expect values (E-value). The all-against-all sequence comparison of Protein to Protein Basic Local Alignment Search Tool (BLASTP) [21] is applied to calculate the E-value, which describes the number of hits that can be expected to get by chance in a pairwise comparison.

The lower the E-value the more the similarity of the two proteins is statistically significant. We utilize such index of significance to quantify biological similarity of each protein pair (u,v), which is to be assigned a score s_e if its E-value is within a cutoff threshold :

$$B(u,\ v) = \begin{cases} s_e, & BLASTP(u,v) \leq threshold \\ 0, & otherwise. \end{cases} \qquad (7)$$

The $B(u,\ v)$ is also further normalized to fall into the scope of [0, 1] in order to keep the consistency of dimensionality with that of $T(u,v)$ for the integration of our scoring function.

3.3 Heuristic Searching to Optimum Matchset

When previous research attempted to align every protein from one PPI network to others, we propose against which to only focus on just proportion of the proteins that deserve to be aligned to their homologues in other species. Under this guiding principle that we deem is more natural and rational, the new strategy is applied accordingly in our heuristic alignment procedure.

A candidate pool can be firstly constructed from protein pairs possessing high overall similarity score S determined from our integrated scoring function. Maximum weighted bipartite matching method is then applied on all pairs to search for a maximum number of pairs whose sum of score S is as large as possible. The outcomes form the candidate pool where each protein pair is aligned by a virtual link called match connection with its similarity score. The candidate pool contains less number of pairs while prior quality of quantified similarities of the networks is well preserved. During the alignment process, matchsets will be created and updated from the candidate pool, where each matchset contains aligned proteins with their match connections from multiple networks.

Instead of considering all proteins, we start the alignment by randomly selecting a source network from the participating multiple networks and a percentage of proteins in the source network to create the initial matchsets, where each of the proteins form one matchset. In each of the repeated step of the alignment procedure, a candidate match connection from the candidate pool will be randomly selected with replacement. It is attempted to link with the proteins in one of the existed matchsets. The effected matchset will be updated according to a merging rules.

For the protein nodes u and v in a selected match connection c=(u,v), $node_i$ or $node_j$ belongs to a considered matchset, and the source nerwork is G, the merging rule can be described as follows:

1) If both u and v do not belong to any existed matchset, but one of the proteins comes from the source network G, the u, v and c will become a new matchset to replace the existed matchset with the lowest alignment score;

2) If one of the proteins belongs to an existed matchset:

2.1) when u belongs to the source network G or neither of the proteins belong to the G, add v and the c to the matchset;

2.2) if v belongs to the source network, then add u, v, and $\{node_i | (node_i, node_2) \in \text{pool} \,\&\, node_i \notin \text{G}\}$ into the matchset, and take (u, v) together wit $\{(node_j, node_2) | (node_j, node_2) \in \text{the matchset}\}$ to form a new matchset;

3) If u and v belong to the same matchset:

3.1) leave without updating if $(\text{u}, \text{v}) \in$ matchset;

3.2) add (u, v) into the matchset if$(\text{u}, \text{v}) \notin$ matchset;

4) If u and v belong to deifferent matchsets $m_i and\ m_j$:

4.1) if either of u and v belongs to G (e.g. the u),

a) add (u,v) into the matchset m_i and the rest of $\text{m}_j$compose a new matchset;

b) add the (u, v) and $\{(node_i, node_j) \mid (node_i, node_j) \in \text{m}_j \,\&\, node_i, node_j \notin \text{G}\}$ into the m_j and then randomly select one match connection from the pool to replace the m_i;

c) choose from a) or b) whichever could obtain higher alignment score.

4.2) if none of the u and v belongs to G:

a) add (u,v) into the matchset m_i and the rest of $\text{m}_j$compose a new matchset;

b) add the(u, v) into the m_j and then randomly select one match connection from the pool (where u or v $\in$ G) to replace the m_i;

c) choose from a) or b) whichever could obtain higher alignment score.

The alignment score $S(M)$ for the current alignment result that consists of matchsets M, are calculated along with each update step. To obtain $S(M)$, the score of each matchset m will first be calculated with function h:

$$h(m) = \sum_{i=1}^{\text{N}_m} S(u, v) \tag{8}$$

where N_m is the number of match connections in that matchset m. Then the alignment score function H for the alignment result with all the matchsets can be formulated as:

$$H(M) = \sum_{i=1}^{\text{N}_M} h(\text{m}_i) \tag{9}$$

where m_i is a matchset in the matchsets M, and N_M is the number of matchsets in matchset M.

To solve the computationally intractable (NP-hard) issue of network alignment, we apply Simulated Annealing (SA) [11] to heuristically search for an alignment result whose matchsets hold the global maximum alignment score. Match connection in the candidate pool is incrementally selected in the update procedure until the alignment result reaches to its highest possible score $H(M)$, which is then the best alignment of multiple networks. The detail of the algorithm is shown in the following Algorithm 1.

```
Algorithm 1 Heuristic Selection
Input: maximum temperature $T_{max}$,
    minimum temperature $T_{min}$,
    candidate match connection pool P,t,K
Output: Set of matched protein complexes
        $M = \{m_1, m_2, \dots, m_i\}$
while $T_{max} > T_{min}$ do
  for i ← 1 to K do
    link ← GetMatch(P);
      $m_i$ ←merge(link, $m_{current}$);
    if Sum($m_{current}$)< Sum($m_i$)
      $m_{current} \leftarrow m_i$;
    elseif rand(0,1)<exp($\frac{\text{Sum}(m_{current} - \text{Sum}(m_i))}{\text{t} * T_{max}}$)
      $m_{current} \leftarrow m_i$;
    end if
  end for
end while
return $M = \{m_1, m_2, \dots, m_i\}$
```

4 Experiment

4.1 Dataset preparations

We use real PPI networks from five eukaryotic species to conduct experiments for the evaluation of our proposed alignment method. The five adopted species in our alignment experiments include: Homo sapiens (human), Mus musculus (mouse), Dorsophila melanogaster (fruit fly), Caenorhabditis elegans (worm) and Saccharomyces cerevisiae (yeast). They are pulled from public molecular interaction database IntAct [22]. Through cleaning and filtering the raw data from these five eukaryotic species, we eventually obtained a total of 21,472 proteins and 87,310 interactions in constructing five PPI networks. The descriptive details for each network are listed in Table 1. The sequence of each protein from all five PPI networks is further retrieved from UniProtKB/Swiss-Prot database [23].

With the variousness of five PPI networks from diverse species, we examine our proposed method and compare our outcome with other four previous widely acknowledged global multiple PPI network alignment methods on the same three testing datasets. The three Datasets named as A, B, C are each composed of three different PPI networks, and their specific compositions are also shown in Table 1.

TABLE 1
Datasets

Species	#Proteins	#Interactions	Dataset		
			A	*B*	*C*
H.sapiens	8828	37956	√		
M.musculus	1569	3129	√	√	
D.melanogaster	1547	3292	√	√	√
C.elegans	784	1493		√	√
S.cerevisiae	5744	41440			√

Proteins and interactions of five species and the composition of three datasets A, B, and C. Blue blocks in each dataset indicate the species included in the according dataset, e.g. dataset-A contains PPI networks of H.sapiens, H.musculus, and D.melanogaster.

Besides the novel topological measure proposed in this study, we also evaluated the quality of our alignment outcomes with commonly applied biological criteria on all datasets. For the purpose of biological evaluations, the Gene Ontology (GO) annotations for each protein were retrieved accordingly from Uniprot-GOA database [24].

4.2 Experiment setups

To integrate biological information into our score assignment, we calculate the E-values of pairwise proteins by implementing BLASTP. The cutoff value was set to be 1e-7 as a filter to keep only those pairs with more potential homologous from each bipartite network. The filtered pairs are all assigned a biological similarity score B with their normalized E-values.

The integrated score of each protein pair is then obtained by combining both $B$and topological similarity score T together on a customizable coefficient α. We also test α by assigning different values and discuss its corresponding influences on the alignment results. All the pairs with integrated scores are further computed with maximum weighted matching algorithm to form a candidate pool for the heuristic update procedure of alignment. For the alignment procedure, we also discuss the effect on the alignment results with different choices of aligning percentages.

For the purpose of alignment result comparisons, four most widely accepted multiple alignment methods are applied, including: IsoRank-N[12], SMETANA[13], NetCoffee[18], and BEAMS[14]. They are all executed with their recommended parameters from the original papers to compare with our proposed method on the same datasets in the experiment.

4.3 Evaluations

The topological evaluation of MNS and the biological evaluations of ME and MNE for all the alignment results generated by our proposed method as well as by the other four methods are compared in Table 2.

TABLE 2
Evaluation of the Alignment Results

Dataset	Evaluations	Alignment Methods				
		Protein2Vec	*BEAMS*	*SMETANA*	*IsoRankN*	*NetCoffee*
Dataset A	ME	0.001969	0.001506	0.002778	0.002402	0.002487
	MNE	0.000855	0.000408	0.000843	0.000631	0.000546
	MNS	250.7044	359.3844	307.7648	287.4407	300.2356
Dataset B	ME	0.003868	0.005123	0.006794	0.005994	0.002894
	MNE	0.001202	0.001210	0.001407	0.001614	0.001487
	MNS	18.60203	25.90799	28.81806	26.63474	27.34610
Dataset C	ME	0.001904	0.002501	0.003306	0.004856	0.002597
	MNE	0.000978	0.000633	0.001144	0.001595	0.001182
	MNS	2915.752	5346.524	5580.384	4399.162	5356.012

4.3.1 Biological measures

To evaluate the biological significance of the alignment results, we applied the commonly adopted measures of Mean Entropy (ME) and the Mean Normalized Entropy (MNE) to assess their functional homogeneity. The idea is based on an intuitive assumption that the more the proteins of a matchset in the alignment results have their GO annotations corresponding to a set of genes with the same function, the higher consistency that matchset possesses to a certain degree in terms of alignment. The higher the consistencies possessed in all the matchsets generated from an alignment, the better is the alignment method.

One of the measuares for assessing the consistency of aligned protein nodes in the same matchset is the ME. Given each protein corresponding to one or more GO annotations, ME is computed by finding all corresponding GO annotations of each protein in a matchset and obtaining the percentage of proteins containing the same GO annotation for each different GO annotation in that matchset. In order to compute the mean entropy of an alignment result with all matchsets, we can first calculate the entropy E(m) of each matchset as follows:

$$\mathrm{E(m)} = E(v_1, v_2 \cdots v_n) = -\sum_{i=1}^{d} p_i \times logp_i \qquad (10)$$

where v_i is the protein nodes in the matchset m; p_i represents the percentage of proteins containing the ith GO annotation, and d is the total number of different GO annotations in this matchset. The lower entropy a matchset hold, the more within-cluster consistency it possesses. The ME is then the evaluation on all matchsets from an alignment by calculating the average of their entropies, which can be formulated as follows:

$$\mathrm{ME} = \frac{1}{N}\sum_{i=1}^{N} E(m_i) \qquad (11)$$

where N is the number of all matchsets in the alignment result and m_i is the $\mathrm{i^{th}}$ matchset. Accordingly, the lower the ME of an alignment, the higher consistency it could obtained, which indicates a better biological quality.

For the purpose of comparison, the ME evaluations of all methods are illustrated in Figure 2.

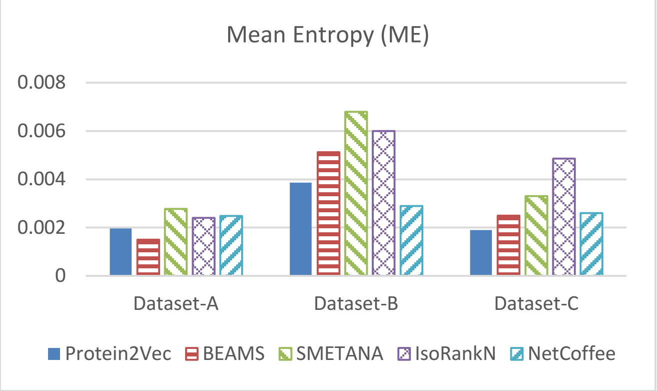


Fig. 2. Illustration and comparisons of the biological evaluation ME on the alignment results of all five alignment methods.

As shown in Figure 2, for the biological evaluation measure ME on Dataset-A, our method obtains slightly higher value than that of the BEAMS method, but performs better alignment when being compared to the other three methods of SMETANA, IsoRankN, and NetCoffee; On Dataset-B, NetCoffee performs slightly better than our method , while ME value of ours is much lower than the other three methods of BEAMS, IsoRankN, and SMETANA; Our proposed method outperforms all four other methods on Dataset-C with lower ME across all aligned networks and achieves better alignment in terms of the consistency of biological functionality.

Another measure of the consistency evaluation for a matchset is the MNE. MNE is a biological measure similar to the ME and more from a normalized aspect. It normalizes the entropy E(m) in each matchset with the following definition of NE(m):

$$\mathrm{NE(m)} = NE(v_1, v_2, \dots v_n) = -\frac{1}{\log d}\sum_{i=1}^{d} p_i \times \log p_i \qquad (12)$$

where the parameters in the formula share the same meaning of Equation 10. The MNE is then defined to be the average value of the normalized entropies of all matchsets generated from an alignment method, which is formulated as follows:

$$\mathrm{MNE} = \frac{1}{N}\sum_{i=1}^{N} NE(m_i) \qquad (13)$$

Similarly, the evaluation of MNE obtaining a lower value would indicate a better consistency and biological quality of an alignment method.

The comparison regarding the MNE evaluation measure of all five methods on the three datasets are shown in Figure 3.

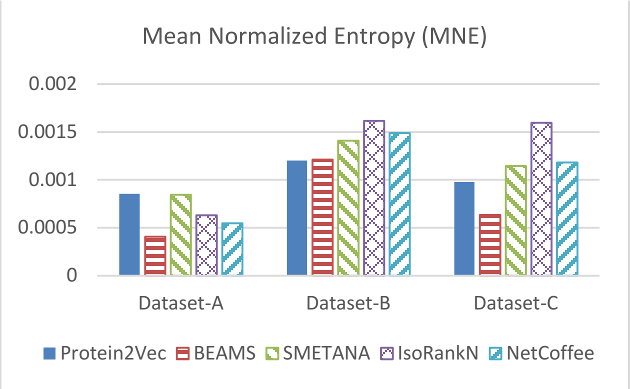


Fig. 3. Illustration and comparisons of the biological evaluation MNE on the alignment results of all five alignment methods.

From Figure 3 we can see that our method could not overtake all other methods on Dataset-A for the MNE evaluation. However, our method outperforms all other four methods on Dataset-B. On Dataset-C, except for the BEAMS, our method obtains better MNE than the other three methods.

4.3.2 Topological measure

There are several existing topological measures usually being applied in the previous PPI network alignment research, such as the Edge Correctness (EC), the Induced Conserved Structure (ICS), or the Symmetric Substructure Score (S3). Despite of their effectiveness in evaluating an alignment results from different aspects, they are designed to measure the alignment of pairwise networks instead of multiple networks.

To solve the above limitation, we propose a novel measure for topological evaluation of the result from multiple network alignment, called Mean Neighbor Similarity (MNS). Our idea is based on a very natural assumption that if two proteins from different networks are very similar in functional homogeneity, they should also share a very similar topological characteristics in terms of the protein interactions in their respective network of species. In another word, two well aligned proteins from different networks should share a very similar pattern of the neighboring structure and the structure of their neighbors. With such assumed guidance, we design to use the degree sequence of all neighbors of a protein in a network to represent its topological feature.

When evaluating the topological feature of a protein node in the matchsets of an alignment result, we obtain the degree sequence of its all neighbors in the network. The two similar protein nodes should share similar topological structures while their neighbor nodes are also having similar structures in their respective PPI networks, and such structural patterns can be represented by the degree sequences of their neighbors; If the degree sequences of different proteins from different networks are similar, the protein nodes are considered to have a more similar topological structure hence being more homogeneous.

To determine the topological similarity of aligned proteins, we need to calculate the similarity of their degree sequences which are usually unequal in length. Traditional sequence similarity calculations usually apply the algorithm such as Euclidean distance, but which cannot solve the complex case of sequences with different lengths. Dynamic Time Warping (DTW) is a very effective algorithm that can measure similarity between two sequences with different lengths. We calculate the similarity of the degree sequences of proteins for the MNS drawing on the same idea of DTW.

Given two finite degree sequences $\mathrm{Q} = (\mathrm{q}_1, \dots, \mathrm{q}_\mathrm{n})$ and $\mathrm{C} = (\mathrm{c}_1, \dots, \mathrm{c}_\mathrm{m})$, without loss generality, we assume both $\mathrm{q}_1, \dots, \mathrm{q}_\mathrm{n}$ and $\mathrm{c}_1, \dots, \mathrm{c}_\mathrm{m}$ are in ascending order, and n, m denote the length of sequence Q and sequence C. The algorithm we applied for calculating the similarity of degree sequences are as follows:

If n=m, the distance between the degree sequences can be directly calculated, such as using Euclidean distance;

if $\mathrm{n} \neq \mathrm{m}$, the two sequences need to be aligned before calculating the distance. Here we use the dynamic programming (DP) algorithm to align different sequences. We need to first create a matrix of n x m, where the matrix element (i, j) represents the distance $\mathrm{d}(q_i, c_j)$ between the two sequence element q_i and c_j, that is, the i$^{\mathrm{th}}$ element of the sequence Q and the j$^{\mathrm{th}}$ element of the sequence C. Calculating the similarity of sequences with different length by the DP algorithm can be then summarized as finding a path, where its passing elements are those that need to be aligned. The optimal distance is the path with the smallest accumulated distance; The shorter the distance, the more similar the two sequences are.

The specific steps of MNS calculation are: Firstly, the degree sequences of two protein nodes are aligned with DP algorithm to obtain their similarity distance. Secondly, the pairwise similarity distances of the degree sequences of all nodes in a matchset is calculated and accumulated. Thirdly, the average of the degree sequence similarity of all the matchsets is calculated as the value of the evaluation index MNS.

Figure 4 shows the MNS comparison of the alignment results of five methods on three datasets. Since the orders of magnitude are not at the same level for the numbers of nodes in the networks of different datasets, in order to show the comparison in one figure, the MNS values are processed and scaled for the results on each dataset separately. For example, on the Dataset-B, the MNS values of the five methods are all increased by 10 times; On the Dataset-C, the MNS values are all reduced by a factor of 10. The scaling processing has no effect on the comparison between the different methods on the same dataset.

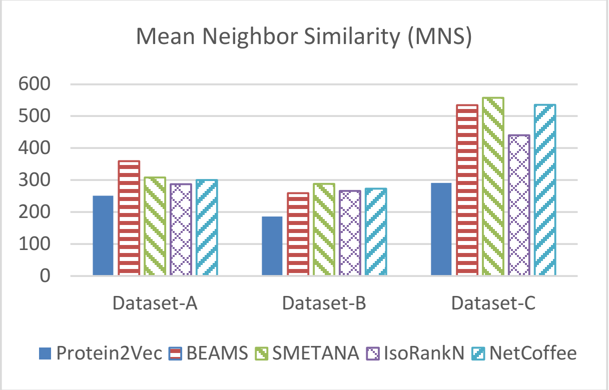


Fig. 4. Illustration and comparisons of the topological evaluation (MNS) on the alignment results of five alignment.

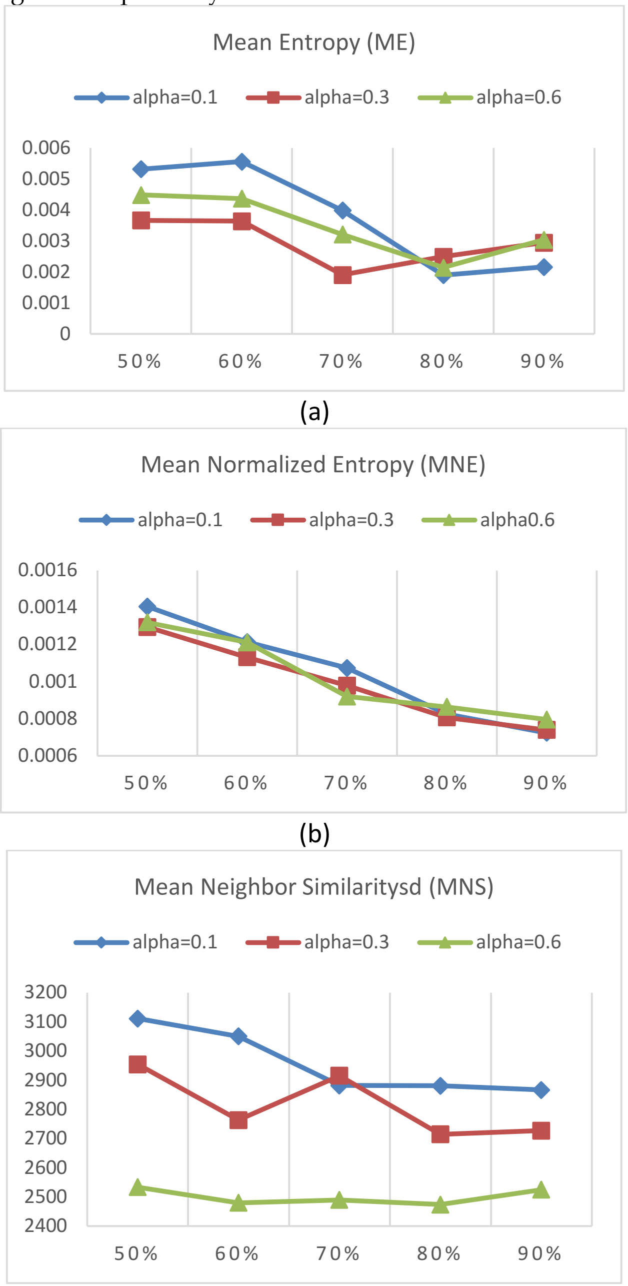
Since MNS calculates the distance between sequences, the lower the value of MNS, the more similar the sequences are. From Figure 4 we can see that for the evaluation of topological features, our method outperforms all other four methods on all three evaluation datasets.

TABLE 3
Evaluation of the Alignment Results on Different Settings of the Parameter percentage and α

		α for topological score		
Percentage	**Evaluations**	*0.1*	*0.3*	*0.6*
50%	ME	0.005318	0.003665	0.004487
	MNE	0.001405	0.001294	0.001319
	MNS	3111.352	2954.371	2533.671
60%	ME	0.005562	0.003642	0.004372
	MNE	0.001213	0.001131	0.001211
	MNS	3050.323	2763.615	2480.121
70%	ME	0.003985	0.001904	0.003213
	MNE	0.001073	0.000978	0.000919
	MNS	2881.996	2915.752	2489.956
80%	ME	0.001901	0.002493	0.002145
	MNE	0.000825	0.000806	0.000863
	MNS	2880.966	2714.952	2474.034
90 %	ME	0.002166	0.002950	0.003043
	MNE	0.000723	0.000739	0.000795
	MNS	2866.939	2727.614	2525.304

In order to figure the influences on the alignment results generated by our proposed method with different parameter assignments to α and the percentage of best aligning proteins in a target network, we additionally conduct a large number of experiments on the three datasets. Taking the experimental results of the parameter examination on the Dataset C as an example, we test different α values and different numbers of aligned matchsets as the percentage of target network. We experiment with α values spanning as 0.1, 0.3, and 0.6, and set the percentage of matchsets spanning as 50%, 60%, 70%, and 80%, and 90%. The detailed results are shown in Table 3.

In order to see the effects of setting different α values and different number of matchsets on the results of multiple network alignments with same dataset, we show different values of α across different aligning percentages of the target network with the three evaluation measures in Figure 5 respectively.



(a)

(b)

(c)

Fig. 5. Evaluation of the alignment results on different settings of the parameter of percentage and α

From Fig. 5, we can see that: Firstly, in the Figure 5(a) , with the same value of α, there is a trend that the ME value decreases at the beginning and then increases when the number of matchsets changes from 50% of the source network protein node number towards 90% of it. This trend remains when we test with the three different α values; Secondly, in the case of the evaluation measure of MNE showed in Figure 5(b), with the same α value, there is a tendency of MNE values decrease when the percentages increase. The overall decreasing rate however changes from large to small, and such tendency also occurs with the condition of three different α values. Thirdly, in the Figure 5(c), when the value of α remains and the number of matchsets increases, there is a trend that values of MNS decrease first and then increase, which has also appeared for three different α values.

These further large number of experiments and comparisons indicate that, initially the more the aligning matchsets involved, the more protein nodes with homogeneity from multiple networks can be aligned with our alignment method; There is a turning point of saturation however, if the allowed number of matchsets keeps increasing after reaching that point, the alignment performance and quality start to decrease. The findings also justify our alignment approach and guidance of only aiming at aligning part of the proteins that should be aligned from multiple PPI networks of different species, instead of making effort to involve all proteins in every network for the network alignment task.

7 Conclusion

In this study, we propose a new PPI network alignment method with representation learning on the networks. It transforms and quantifies the structural features of proteins into low-dimensional vectors. Topological similarity can thus be computed through the corresponding vectors. Along with the biological similarity, the proposed method aligns multiple PPI networks without requiring all proteins to be aligned, which is more efficient to find only most homologous proteins across multiple species. Besides biological evaluation measures, we also propose a new measure to better evaluate topological quality of the alignment results.

Acknowledgment

This work was supported in part by NSF III 1815256, NSF III 1744661, NSF CNS 1650431, and NSFC 61873288.

References

[1] M. Milano, P. Guzzi, M. Cannataro, "GLAlign: A Novel Algorithm for Local Network Alignment," IEEE/ACM Transactions on Computational Biology and Bioinformatics, April 2018. DOI: 10.1109/TCBB.2018.2830323

[2] F. Alkan and C. Erten, "Sipan: simultaneous prediction and alignment of protein–protein interaction networks," Bioinformatics, vol. 31, no. 14, pp. 2356–2363, 2015.

[3] L. Meng, A. Striegel, and T. Milenković, "Local versus global biological network alignment," Bioinformatics, vol. 32, no. 20, pp.

3155–3164, 2016.

[4] J. Hu and K. Reinert, “Localali: an evolutionary-based local alignment approach to identify functionally conserved modules in multiple networks,” Bioinformatics, vol 31, no. 3, pp. 363-372, 2014.

[5] J. Gao, P. Liu, X. Kang, L. Zhang, and J. Wang, “PRS: Parallel relaxation simulation for massive graphs,” Computer Journal, vol. 59, no. 6, pp. 848–860, 2016.

[6] J. Crawford, Y. Sun, and T. Milenković, “Fair evaluation of global network aligners,” Proceddings of IEEE International Conference on Bioinformatics and Biomedicine, pp. 1768-1770, 2015.

[7] A. Elmsallati, C. Clark, and J. Kalita, “Global alignment of protein-protein interaction networks: A survey,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 13, no. 4, pp. 689–705, 2016.

[8] FE Faisal, L. Meng, J. Crawford, T. Milenkovic, “The post-genomic era of biological network alignment,” EURASIP Journal on Bioinformatics and Systems Biology. 2015(1):1-19.

[9] Bhowmick SS, Seah BS. “Clustering and summarizing protein-protein interaction networks: a survey,” IEEE Transactions on Knowledge and Data Engineering. 2016;28(3):638-658.

[10] W. Eddine Djeddi, S. Yahia, E. Nguifo, “A Novel Computational Approach for Global Alignment for Multiple Biological Networks,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, Feb. 2018. DOI: 10.1109/TCBB.2018.2808529.

[11] N. Mamano and W. B. Hayes, "SANA: simulated annealing far outperforms many other search algorithms for biological network alignment," Bioinformatics, vol. 33, no. 14, pp. 2156-2164, 2017.

[12] 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. 253–258, 2009.

[13] S. M. E. Sahraeian and B.-J. Yoon, “Smetana: accurate and scalable algorithm for probabilistic alignment of large-scale biological networks,” PloS one, vol. 8, no. 7, pp. e67995, 2013

[14] F. Alkan and C. Erten, “Beams: backbone extraction and merge strategy for the global many-to-many alignment of multiple PPI networks,” Bioinformatics, vol. 30, no. 4, pp. 531–539, 2014.

[15] B. Perozzi, R. Al-Rfou, and S. Skiena. “Deepwalk: Online learning of social representations,” 2014, In KDD. ACM, 701–710.

[16] A. Grover and J. Leskovec. “Node2vec: Scalable feature learning for networks,” 2016, In KDD. ACM, 855–864.

[17] L. Ribeiro, P. Saverese, and D. Figueiredo. “Struc2vec: Learning Node Representations from Structural Identity,” 2017, In KDD. ACM, 385–394.

[18] J. Hu, B. Kehr, and K. Reinert, “Netcoffee: a fast and accurate global alignment approach to identify functionally conserved proteins in multiple networks,” Bioinformatics, vol.30, no. 4, pp. 540-548, 2014.

[19] J. Choi and D. Lee, "Topological motifs populate complex networks through grouped attachment," Scientific reports, vol. 8, no. 1, p. 12670, 2018.

[20] J. Gao, B. Song, W. Ke, and X. Hu, "Balanceali: multiple PPI network alignment with balanced high coverage and consistency," IEEE transactions on nanobioscience, vol. 16, no. 5, pp. 333-340, 2017.

[21] S. F. Altschul, T. L. Madden, A. A. Sch¨affer, J. Zhang, 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.

[22] Kerrien, Samuel, et al. "The IntAct molecular interaction database in 2012." Nucleic acids research, vol. 40, no. D1, pp. 841-846, 2011.

[23] UniProt Consortium. "The universal protein resource (UniProt) in 2010." Nucleic acids research 38.suppl 1 (2010): D142-D148.

[24] Huntley, Rachael P., et al. "The GOA database: gene ontology annotation updates for 2015." Nucleic acids research 43.D1 (2015): D1057-D1063.

[25] C. Notredame, D. G. Higgins, and J. Heringa, “T-coffee: A novel method for fast and accurate multiple sequence alignment,” Journal of molecular biology, vol. 302, no. 1, pp. 205–217, 2000.

[26] Zhu Y, Li Y, Liu J, et al. “Discovering large conserved functional components in global network alignment by graph matching,” BMC genomics, 2018, 19(7): 41-58.

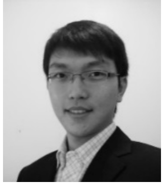
[27] Kazemi, Ehsan, and Matthias Grossglauser. "MPGM: Scalable and Accurate Multiple Network Alignment." IEEE/ACM transactions on computational biology and bioinformatics (2019). **DOI:** 10.1109/TCBB.2019.2914050

Jianliang Gao is received the PhD degree in computer science from Chinese Academy of Sciences, China. Currently, he is a professor in the the School of Computer Science and Engineering, Central South Uinversity, China. He was a visiting professor from 2015 to 2017 at Drexel University. His current research interests include big data processing, and bioinformatics. He is serving as the program member of several international conferences. He was the general chair of 2018 IEEE Conference on Big Data. He is a member of the Institute of Electrical and Electronics Engineers.

Ling Tian is a graduate student of the School of Computer Science and Engineering, Central South University, Changsha, China. She received the B.S. degree from ChangChun University of Technology, Changchun, Jilin, China, in 2017. Her current research focuses on network alignment and bioinformatics.

Tengfei Lv is a graduate student of the School of Computer Science and Engineering, Central South University，Changsha, China. He received the B.S. degree from Bohai University, Jinzhou, Liaoning, China, in 2019. His research interests are machine learning and bioinformatics.

Jianxin Wang received the PhD degree in computer science from Central South University, China, in 2001. Currently, he is a professor in the School of Computer Science and Engineering, Central South University, China. His current research interests include algorithm analysis and optimization, computer network, and bioinformatics. He has published more than 100 papers in various international journals and refereed conferences. He is serving as the program committee chair or member of several international conferences. He is a senior member of the Institute of Electrical and Electronics Engineers.



Bo Song received his B.S. degree in biomedical engineering from Northeastern University, Shenyang, China, in 2010 and his M.S. degree in biomedical engineering from Drexel University, Philadelphia, PA, in 2012. He is currently pursuing the Ph.D. degree in information science at Drexel University.Meanwhile, he is a Research Assistant and a Lecturer in the College of Computing & Informatics at Drexel University, and a Visiting Researcher at Children's Hospital of Philadelphia (CHOP). He had also worked as a Research Associate at Drexel Center for Integrated Bioinformatics. His research interests include Bioinformatics, Network Analysis, Data Mining, Bio/Medical image processing etc.

Xiaohua Hu is a full professor and the founding director of the data mining and bioinformatics lab at the College of Computing and Informatics, Drexel University. He is also serving as the founding co-director of the NSF Center (I/U CRC) on Visual and Decision Informatics (NSF CVDI), IEEE Computer Society Bioinformatics and Biomedicine Steering Committee Chair, and IEEE Computer Society Big Data Steering Committee Chair. He is a scientist, teacher, and entrepreneur. He joined Drexel University in 2002. He founded the International Journal of Data Mining and Bioinformatics (SCI indexed) in 2006, International Journal of Granular Computing, and Rough Sets and Intelligent Systems in 2008. Earlier, he was a research scientist in the world-leading R&D centers such as Nortel Research Center and Verizon Lab (the former GTE labs). In 2001, he founded the DMW Software in Silicon Valley, California. He has a lot of experience and expertise to convert original ideas into research prototypes, and eventually into commercial products, many of his research ideas have been integrated into commercial products and applications in data mining fraud detection and database marketing. His current research interests are in data/text/web mining, big data, bioinformatics, information retrieval and information extraction, social network analysis, healthcare informatics, rough set theory, and application. He has published more than 240 peer-reviewed research papers in various journals, conferences, and books.