Aligning Multiple PPI Networks with Representation Learning on Networks

Bo Song[1], Jianliang Gao[2], Hongliang Du[2], Zheng Chen[1], Xiaohua Hu[1]
[1] College of Computing and Informatics, Drexel University, Philadelphia, PA, USA
[2] College of Information Science and Engenering, Central South University, Changsha, China

***Abstract*—Protein-protein interaction (PPI) network alignment has been motivating researches for the comprehension of the underlying crucial biological knowledge, such as conserved evolutionary pathways and functionally conserved proteins throughout different species. Existing PPI network alignment methods have tried to improve the coverage ratio by aligning all proteins from different species. However, there is a fundamental biological justification needed to be acknowledged, that not every protein in a species can, nor should, find homologous proteins in other species. In this paper, we propose a novel approach for multiple PPI network alignment that tries to align only those proteins with the most similarities. To provide more comprehensive supports in computing the similarity, we integrate structural features of the networks together with biological characteristics during the alignment. For the structural features, we apply on PPI networks a representation learning method, which creates a low-dimensional vector embedding with the surrounding topologies of each protein in the network. This approach quantifies the structural features, and provides a new way to determine the topological similarity of the networks by transferring which as calculations in vector similarities. We also propose a new metric for the topological evaluation which can better assess the topological quality of the alignment results across different networks. Both biological and topological evaluations demonstrate our approach is promising and preferable against previous multiple alignment methods.**

Keywords—protein representation, multiple network alignment, PPI networks, topological assessment

I. Introduction

A. PPI Network Alignment

Network alignment as one of the most effective comparative analysis method has been successfully applied in variety of fields such as computer vision [1], social network [2, 3], biological networks [4-7] etc. Especially in biological applications, protein-protein interaction (PPI) network alignment facilitates the constructive explorations of the complex biological processes across different species and provides many important outcomes in identification of functional modules, detection of evolutionary pathways, discovery of functionally conserved complexes etc [8, 9]. By mapping proteins with corresponding maximized similarities in different PPI networks, network alignment is able to find conserved motif representing the subnetworks that have patterns of orthologous proteins with conserved interactions and activities, and utilize which in the prediction of protein functionalities as well [6].

The alignment of PPI networks across different species bridges the biological knowledge by transferring which from well-studied species to poor-studied species [4]. This is very beneficial and vital especially when the experimental studies of the poor-studied species are very costly or even impractical, such as knowledge transferring from Saccharomyces cerevisiae (yeast) or Caenorhabditis elegans (worm) to Homo sapiens (human), leading to new discoveries in evolutionary biology, drug targets, disease causing genes etc. In addition to the transfer of substantial knowledge, similarities between the networks determined through the alignment can also be used to infer phylogenetic relationships of different species [9].

Existing network alignment approaches can also fall into a categorization as aligning either locally or globally. Local network alignment aims to find smaller subnetworks with high local similarity irrespective of the overall similarity among the participating networks [4]. Since the subnetwork can overlap, a protein from one network can be mapped to several proteins in other networks, and hence generate a many-to-many local mapping. However, network alignment methods focusing on local alignment are generally not capable of finding global one-to-one mappings that maximize the overall similarity of the entire networks in a global network alignment [10].

This paper focuses on aligning multiple PPI networks globally, with local topological information integrated, to improve the performances of network alignment.

B. Motivations

Although continuous progress has been made in the field of PPI network alignment, there are two important problems remain unsettled with no satisfactory solutions:

(1) How to better represent and quantify the topological property of a protein in different network of species?

Previous researches once focused only on biological characteristics of proteins, such as the amino acid sequences, to align proteins with similarity and relevancy biologically. After topological properties that exclusively extracted from the PPI network structure gradually showed its advantages over the biological information, current network alignment methods tend to combine both sources of information to promote their alignment processes [4, 11].

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

978-1-5386-5488-0/18/$31.00 ©2018 IEEE

However, the metric used to capture the structural topology of a network and its proteins varies, an always focus on single attribute of edge or node, such as degree, centrality, eccentricity, betweenness etc. Alignment results may be optimized or not by one of these topological extractions than the others, according to various situations. In this paper, we propose to describe the characteristic of a node from multiple topological perspectives and apply node embedding method to generate low-dimensional vectors in representing the proteins together with their connectivity patterns. The topological properties of one protein are preserved to the great extent from its network, and which is also in the preferable quantified format of vector for further computation.

(2) How many proteins should be aligned towards a better alignment?

While no consensus exists on which evaluation measures should be used for different situation, consistency and coverage are the two most considered in evaluating the quality of network alignment results. Functional consistency is determined as measuring the common biological functionality shared by aligned proteins. Coverage on the other hand, serves as the topological measure in inspecting the total amount of proteins being aligned across networks by an alignment method.

The pursue of aligning more proteins in order to increase the overall coverage is unfortunately the most concern for many network alignment methods. Besides achieving high consistency, the topological measure for the quality of alignment should not simply focuses on the coverage for improvement. It is rooted on an intuitive reason: not all proteins from different networks should be attempted to get aligned, which is precisely due to they are different species, and quite many proteins are supposed not to be homologous. To break the conventional limitation, we propose a partial alignment approach that only aligns those proteins with the most similarity across species while achieving balanced high consistency at the same time.

C. Contributions

In this paper, we propose a new approach to align multiple PPI networks. The main contributions are as follows:

- To improve the description of topological properties during the alignment, we propose to adopt representation learning method to embed protein node as vector while capturing enriched structural features such as triangle motifs. The topological similarity can consequently be quantified and transferred directly as similarity of vectors for computation.
- We propose a partial alignment approach which focus on only mapping proteins with the most similarities and those supposed to be aligned, instead of aiming at increasing overall coverage by attempting to align all the proteins.
- 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.

II. Related works

The general idea behind PPI network alignment is to obtain the similarities between proteins through the mapping of different networks and determine among them the alignment with the highest score of similarity. To decide protein similarity, many current network alignment algorithms adopt a node cost function that combines together the biological information and structural information [10]. For network structural properties, representation learning is a recent popular and more comprehensive approach in reflecting the topology than conventional parameters such as degree. In this section, we review the related researches in PPI network alignment and representation learning.

A. Alignmet of PPI Networks

Previous network alignment methods could fall in either or combined categories of local or global, pairwise or multiple [4, 6, 10, 12], and each of them has its own features in attempting to achieve an optimal alignment result.

For pairwise PPI networks, IsoRank is one of the most classic and referred alignment method in the field. It is a typical pairwise global alignment algorithm for biological networks [12]. The idea of PageRank algorithm is used for reference in computing the similarity of protein pairs according to their neighboring topology. Intuitively, if neighbors of two nodes from different networks are similar, the two nodes are also considered as similar. Based upon which, IsoRank assigns pairwise functional similarity scores for node pairs and screens out the candidate pairs to construct similarity matrix for the search of global alignment results with greedy strategy. MAGNA is another method recently proposed for pairwise and global network alignment. It relies on genetic algorithm in choosing alignment results with high scores according to an objective function that combines topological and biological factors [4].

More recent research interests shift to the alignment on multiple networks, such as IsoRankN, SMETANA, NetCoffee, and BEAMS. Extended from IsoRank, IsoRankN [12] applies spectral clustering on multiple networks, which improves the global network alignment and produces aligned clusters as the result. Each of the clusters could consist of multiple proteins from a same network. SMETANA [13] tries to effectively find among large networks the maximum global alignment. It aligns multiple networks in two stages. It first applies a semi-Markov random walk model with its cost function in order to calculate similarities between nodes, which serves as a probabilistic index; Then, a greedy approach is used in producing alignment results with the maximum expected accuracy. NetCoffee is another global multiple aligner proposed recently that combines sequence and topological similarity together in its scoring function [6]. It is the first multiple network aligner that weights score of protein pairs according to not only pairwise sequence similarity, but also a triplet extension across multiple networks. This topological approach is similar to the multiple sequence aligner T-Coffee. Alken et al., proposed a heuristic approach based on the strategy of backbone extraction and merge in the alignment algorithm of BEAMS [14] to globally align multiple PPI networks. They break down the alignment process into two phases. In the first phase, a partite node similarity graph is

constructed from the given networks to determine backbones by identifying a set of disjoint cliques that maximizes the number of conserved edges between each pair of cliques. Once all the backbones are determined, the cliques are repeatedly merged during the second phase of backbone merging to form aligned node clusters until the alignment score reaches to its maximum.

B. Representation Learning

Current literatures of node embedding technique mainly define the nodes similarity in terms of proximity or neighborhood structure. Representation learning is the approach works by making similar nodes have more similar embedding.

Deepwalk [15] is one of the representation learning algorithms inspired by the word2vec algorithm from language modeling. It aims at learning adaptable social representations for the nodes in a network and generates sequences from a stream of truncated random walks on the network, which effectively maps local features into a lower dimensional embedding. Deepwalk has drawn many interests in the machine learning community as it conveyed the idea of representation learning from word2vec to the realm of networks, spurring extended and fruitful discussions. In order to capture the diverse patterns of connectivity observed in networks, node2vec [16] is proposed as an algorithmic framework to learn for nodes the representation of continuous features. It generates node mapping in a low-dimensional space that maximizes the likelihood of preserving neighborhood features of nodes. Node2vec defines a flexible notion of network neighborhoods and designs a biased random walk procedure that can explores neighborhoods diversity efficiently. Struc2vec [17] is another rising representation learning framework with great novelty and flexibility, which learns latent representations of nodes for their structural identities with a hierarchy measures of nodes at different scales for similarity. It constructs a multilayer graph to encode topological similarities and generate structural context for the nodes in the network.

Recent advances in representation learning promote the node embedding which is very promising and could be used in many downstream tasks (e.g., link prediction), but typically has not been extended beyond a single network, especially for multiple biological networks.

III. Methods and Algorithms

To achieve optimal alignment result with enhanced supports, we establish a 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 similarity scoring function to guide the aligning process. All match connections with high scores between proteins across networks form a candidate pool for a heuristic searching procedure to be later conducted and generate the final optimized alignment result which only consists of proportional proteins with the most overall similarities.

If we can quantify and denote the biological similarity between two proteins u and v as $B(u,v)$, and the topological similarity of which as $T(u,v)$. Then our scoring function integrating both features can be formulated as following:

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

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

A. Protein Node to Vector

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

Denote $G = (V, E)$ as a considered 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 characteristic of node with both neighbor degrees chain and motif features for every two nodes.

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

$$e_k(u,v) = \frac{H_k(u,v)}{\sum_{v \in V, v \neq u} H_k(u,v)} \quad (3)$$

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

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

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

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 the latent representations as a low-dimensional vector for each of the protein nodes. With the structural property quantified in vector, topological similarity $T(u, v)$ can be readily transformed by calculating vector similarity with various choice of coefficient such as cosine measure.

B. Protein Sequence to Index

Besides structural features of interactions in a network, protein has its biological identity, such as the amino acid sequences, that can be used to assess its homology relationships with others. High 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 works [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 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.

For a pair of proteins, the lower the E-value the more their similarity is statistical 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 set threshold cutoff accordingly:

$$B(u,\ v) = \begin{cases} s_e, & BLASTP(u, v) \le threshold \\ 0, & otherwise. \end{cases} \quad (6)$$

C. Heuristic Searching to Optimum Matchset

Unlike previous works that attempted to align every proteins from one PPI network to others, we propose to against which by only focusing on just proportion of those proteins that are deserved to be aligned to their homologues in other species. Under this guiding principle that we deem is naturally more rational, new strategy is applied accordingly in our heuristic alignment procedure.

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

Instead of considering all proteins, we start the alignment by randomly select in a source network from the participating multiple networks a percentage of proteins to create the initial matchsets, where each protein 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 be linked with the proteins in one of the existed matchsets. The effected matchset will be updated according to a merging rules, where the proteins of the match connection could either:

1) Being fully merged into an existing matchset and expand the size of the matchset, if the one or both of the proteins in the selected match connection can be found in only one of the existed matchsets;

2) Replacing one or more proteins in one or both of the matchsets and adjust the protein composition of affected matchsets accordingly, either change the matchset size or delete the matchset when no alignment hold on remaining proteins, if proteins in the selected match connection exist in different matchsets;

3) Creating a new matchset on its own and substitute for one of the existed matchsets as a whole, if whose current alignment score ranks the last;

The alignment score S(M) for the current alignment results consist 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:

$$h(m) = \sum_{i=1}^{\mathrm{N}_m} S(u, v) \quad (7)$$

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

$$H(M) = \sum_{i=1}^{\mathrm{N}_M} h(\mathrm{m}_i) \quad (8)$$

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

To solve the computationally intractable (NP-hard) issue of network alignment, we apply the approach of 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 $H(\mathrm{M})$, which is then the best alignment of multiple networks.

IV. Experiments and Results

A. Dataset Preparations

To evaluate the quality of alignment results, we collected real PPI networks from five species to test our proposed alignment approach in experiments. Five eukaryotic species were included: *Homo sapiens* (human), *Mus musculus* (mouse), *Dorsophila melanogaster* (fruit fly), *Caenorhabditis elegans* (worm) and *Saccharomyces cerevisiae* (yeast). They were retrieved from public molecular interaction database IntAct [22].

After data cleaning and filtering, we eventually obtained from five species a total of 21472 proteins and 87310 interactions in constructing five PPI networks. Details about the number and interactions of proteins for each network are listed in Table I. Amino acid sequence of each protein is further retrieved from UniProtKB/Swiss-Prot database [23].

Based upon these PPI networks from various species, we applied our approach on datasets of multiple networks containing number of networks scaling from 3 to 5, to examine the robustness of our proposed method. We also compare our alignment results against those from other three widely acknowledged global multiple PPI network alignment methods from previous researches on the same datasets. The three Datasets A,B,C are composed of increasing number of PPI networks in 3, 4, and 5 respectively, and they are described in Table I with more details.

TABLE I. DATASETS

PROTEINS AND INTERACTIONS OF FIVE SPECIES AND THE COMPOSITION OF THREE DATASETS A,B, AND C. "√" INDICATE THE SPECIES INCLUDED IN THE ACCORDING DATASET, E.G. DATASET-A CONTAINS PPI NETWORKS FROM H.SAPIENS, M.MUSCULUS, AND D.MELANOGASTER.

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			√

Besides a novel topological quality measure first proposed in this paper, we also evaluated the biological quality of alignment results under commonly applied criteria. For the purpose of biological evaluations, Gene Ontology (GO) of proteins were retrieved accordingly from Uniprot-GOA database [24].

B. Experiment setups

For the biological similarity in the score function, we calculate the E-values between proteins by BLASTP. The cutoff value was set as 1e-7 to filter all the E-values and keep only the match connections with more potential homologous in every bipartite network. The remaining match connections are all assigned a biological score of $B(u, v)$ with their normalized E-value.

The integrated score of each match connection are then obtained by combining both $B(u, v)$ and $T(u, v)$ on the customizable coefficient α, which we set as 0.5 for generality, to equally distribute the contributions from both biological and topological similarities from the participated networks. We also tested α assigned with different values and discussed the corresponding influences on the alignment results. For the alignment procedure, we set the percentage of aligning best matching proteins from the source network to the target networks as 30%. We also discuss the effect of choosing different percentages on the alignment results.

To compare our alignment results against others, we applied three widely accepted multiple alignment methods: IsoRankN [12], SMETANA [13], and BEAMS [14]. They are all executed with their recommended parameters from the original papers on the same datasets as we evaluate our proposed method with.

C. Evaluations

1) Biological measures

To evaluate the biological quality of the alignment results, we applied the commonly adopted measures of mean entropy (ME) to assess the functional homogeneity. The idea is based on an intuitive assumption that if all the proteins of a matchset from the alignment results have Gene Ontology (GO) annotations that correspond to a set of genes with the same function, then that matchset possesses a biological consistency to a certain degree. The higher the consistencies possessed in all matchsets generated by an alignment, the better the alignment method.

The consistency in a matchset can be measured by entropy E(M) defined as follows:

$$\mathrm{E(M)} = E(v_1, v_2, \ldots v_n) = -\sum_{i=1}^{d} p_i \times \log p_i \qquad (9)$$

where p_i is the percentage of all proteins with the annotation GO_i in a matchset, and d represents the total number of different GO annotations in that matchset. A matchset with more within-cluster consistency will hold lower entropy.

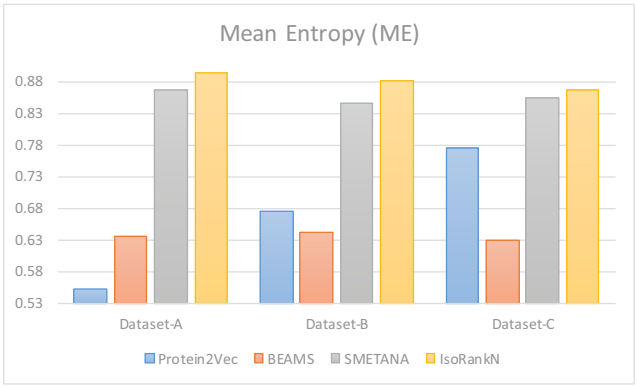

Fig. 1. Illustration and comparisons of the biological evaluation

The mean entropy (ME) is then the evolution on the whole alignment by calculating the average of the entropies of all generated matchsets. Accordingly, the lower the ME of an alignment, the higher consistency it could obtained, which indicates a better biological quality. We can see in Fig. 1 that in general our method obtains better alignment results in terms of biological evaluation on all datasets. The only exception is that BEAMS achieved lower ME than ours, but only on four and five networks alignments.

2) Topological measure

We propose a novel measure of topological quality of the alignment result, called mean neighbor similarity (MNS). Our idea is based on a very nature assumption that if two proteins from different networks are very similar or functionally homologous, they should share a very similar topological structures of the protein interactions in their respective network of species. In another word, two well aligned proteins from different networks should have very similar neighbor structure pattern. With such guidance assumed, we design to use a degree

sequence of the neighbors of a protein in a network to represent that protein.

For each protein in a matchset of an alignment, we first obtain the degree sequence of all its neighbors, and then unify the sequence length to the maximum length in the matchset by making up with zeros and later sort the values in every sequence. Then the distances between each two degree sequences are calculated. The average value of all distances is hence the MNS. The lower the MNS means the better the similarity of topology in the alignment results. This measure with topological feature embedded does not have any number limit for the participating networks in the alignment, and also avoid the improper pursuing of the coverage as a whole.

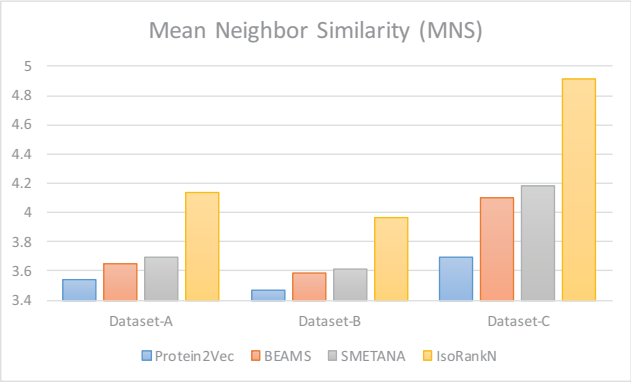

Fig. 2. Illustration and comparisons of the topological evaluation (MNS) on the alignment results from four alignment methods.

The Fig. 2 shows the comparison illustration of the alignment results on all datasets by all methods in terms of topological evaluation. It is obvious that our method outperform the other three methods and achieves much lower MNS.

V. Conclusions

In this paper, 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 proposed a new measure to better evaluate topological quality of the alignment results.

References

[1] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 532-539.

[2] Z. Chen, X. Yu, B. Song, J. Gao, X. Hu, and W.-S. Yang, "Community-Based Network Alignment for Large Attributed Network," in Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 587-596: ACM.

[3] J. Gao, B. Song, Z. Chen, W. Ke, W. Ding, and X. Hu, "Counter Deanonymization Query: H-index Based k-Anonymization Privacy Protection for Social Networks," in Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017, pp. 809-812: ACM.

[4] L. Meng, A. Striegel, and T. Milenković, "Local versus global biological network alignment," Bioinformatics, vol. 32, no. 20, pp. 3155-3164, 2016.

[5] C. Zhao, Y. Zang, W. Quan, X. Hu, and A. Sacan, "HIV1-human protein-protein interaction prediction based on interface architecture similarity," in Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on, 2017, pp. 97-100: IEEE.

[6] F. E. Faisal, L. Meng, J. Crawford, and T. Milenković, "The post-genomic era of biological network alignment," EURASIP Journal on Bioinformatics and Systems Biology, vol. 2015, no. 1, p. 3, 2015.

[7] J. Gao, B. Song, X. Hu, F. Yan, and J. Wang, "ConnectedAlign: a PPI network alignment method for identifying conserved protein complexes across multiple species," BMC bioinformatics, vol. 19, no. 9, p. 129, 2018.

[8] E. M. Marcotte, M. Pellegrini, H.-L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg, "Detecting protein function and protein-protein interactions from genome sequences," Science, vol. 285, no. 5428, pp. 751-753, 1999.

[9] 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.

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

[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. i253-i258, 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, p. 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, 2013.

[15] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014, pp. 701-710: ACM.

[16] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016, pp. 855-864: ACM.

[17] L. F. Ribeiro, P. H. Saverese, and D. R. Figueiredo, "struc2vec: Learning node representations from structural identity," in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 385-394: ACM.

[18] N. Malod-Dognin, K. Ban, and N. Pržulj, "Unified alignment of protein-protein interaction networks," Scientific Reports, vol. 7, no. 1, p. 953, 2017.

[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 et al., "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] S. Kerrien et al., "The IntAct molecular interaction database in 2012," Nucleic acids research, vol. 40, no. D1, pp. D841-D846, 2011.

[23] U. Consortium, "The universal protein resource (UniProt)," Nucleic acids research, vol. 36, no. suppl_1, pp. D190-D195, 2007.

[24] R. P. Huntley et al., "The GOA database: gene ontology annotation updates for 2015," Nucleic acids research, vol. 43, no. D1, pp. D1057-D1063, 2014.