Computational and Structural Biotechnology Journal 18 (2020) 2647–2656

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

A review of protein–protein interaction network alignment: From pathway comparison to global alignment

Cheng-Yu Ma [a], Chung-Shou Liao [b,*]

[a] Chang Gung Memorial Hospital, No. 5, Fu-Hsing St., Kuei Shan Dist., Taoyuan City 33305, Taiwan, ROC
[b] National Tsing Hua University, No. 101, Section 2, Kuang-Fu Rd., Hsinchu City 30013, Taiwan, ROC

ARTICLE INFO

Article history:
Received 21 May 2020
Received in revised form 1 September 2020
Accepted 5 September 2020
Available online 18 September 2020

Keywords:
Network alignment
Biological network
Protein interaction network
Systems biology

ABSTRACT

Network alignment provides a comprehensive way to discover the similar parts between molecular systems of different species based on topological and biological similarity. With such a strong basis, one can do comparative studies at a systems level in the field of computational biology. In this survey paper, we focus on protein–protein interaction networks and review some representative algorithms for network alignment in the past two decades as well as the state-of-the-art aligners. We also introduce the most popular evaluation measures in the literature to benchmark the performance of these approaches. Finally, we address several future challenges and the possible ways to conquer the existing problems of biological network alignment.

© 2020 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Contents

* Corresponding author.
E-mail address: csliao@ie.nthu.edu.tw (C.-S. Liao).

https://doi.org/10.1016/j.csbj.2020.09.011
2001-0370/© 2020 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

In the post-genomic era, with more and more omics data being generated, networks become a more appropriate representation to describe complicated biological systems, such as protein–protein interaction networks, gene regulatory networks, and transcription factor networks. Similar to sequence alignment demonstrating the main approach for biological sequence analysis, network alignment presents a comprehensive way to compare two or more biological networks in systems biology. In particular, network alignment considers not only biological interactions but also topological similarity of the neighborhood of biological nodes, like genes or proteins across different networks, which undoubtedly reveals deeper insight into molecular behaviors. One of the most straightforward applications of network alignment across biological networks is to transfer known biological knowledge from well-studied species to unknown ones. Moreover, we can discover the relationships between different species from a systems perspective [40]. Not surprisingly, the systems-analytical approach provides a way to more completely discuss a variety of interactions between objects throughout biological networks, even including other applications to social networks in recent years.

1.1. Problem definition

The goal of the biological network alignment problem is to cluster nodes across different networks based on their biological (sequence) similarity and the interaction patterns of their neighboring communities (i.e. topology similarity). The formal definition is as follow: given K networks (or graphs) $G_n = (V_n, E_n), 1 \leqslant n \leqslant K$, where V_n and E_n represent the sets of nodes and edges of network G_n respectively. The objective of network alignment is to find a one-to-one (or many-to-many) correspondence M, which is a set of node pairs (hyper-node pairs) $(u, v) \in V_i \times V_j$ $((S_u, S_v) \subseteq V_i \times V_j)$, $i \neq j, 1 \leqslant i, j \leqslant K$, where u and v (S_u and S_v) are closely conserved in sequence and topology. In a perfect scenario, node pairs $(p, q), p \in N(u), q \in N(v)$ are usually contained in M (or in the cluster set of (u, v) in M) due to topology similarity.

However, the lack of sufficient network data leads to a computational challenge. For example, there are a large amount of false negatives/positives (sometimes near 20%) in protein–protein interaction networks discovered by the yeast-two–hybrid (Y2H) technique [5,11,19]. Similar limitations occur in other high-throughput techniques, such as TAP-MS(tandem affinity purification mass spectrometry) [22] and ChIP-Seq [24]. Obviously, the above techniques that cause false identification significantly increase the computational hardness of network alignment.

1.2. Local vs. global

Network alignment can be classified into two categories: local and global alignments. The difference between the two types is similar to the one made for sequence alignment. The target of local alignment is to identify the closely mapping subnetworks between different networks [62]. Typically, local network alignment reports multiple subnetworks across networks, which may be mutually inconsistent [63]. On the other hand, global network alignment tries to match different networks as a whole, and the output result is a single mapping between the nodes of the networks [63]. Furthermore, global network alignment targets on searching for the best consistent mapping between all nodes across the networks, which can reveal evolutionarily conserved functions at a systems level.

In contrast, local network alignment matches the partial subnetworks between networks which are difficult to show the evolutionary trace of whole systems as shown in Fig. 1.

1.3. Pairwise vs. multiple

It is straightforward to see the hardness of computational complexity for the problem, as the *tripartite matching* problem is NP-hard. Even though we consider the case of $K = 2$, i.e., the pairwise alignment of two networks, the problem is still challenging because the *subgraph isomorphism* problem is still NP-hard.

Briefly speaking, network alignment can be divided into *pairwise* or *multiple*, just analogous to biological sequence alignment. That is, pairwise network alignment compares two networks at once, while multiple network alignment considers more than two networks at the same time (See Fig. 2). Obviously, the computational complexity increases exponentially in the number of networks.

1.4. One-to-one, one-to-many, many-to-many

Network alignment also can be classified as one-to-one, one-to-many, and many-to-many by the type of node mapping of their output. One-to-one network alignment maps one node of a given

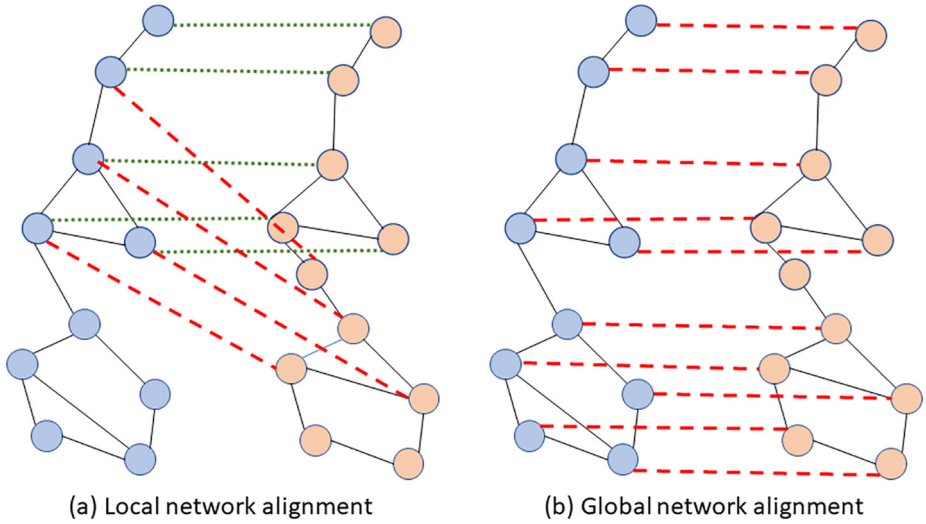

Fig. 1. Local network alignment vs. Global network alignment.

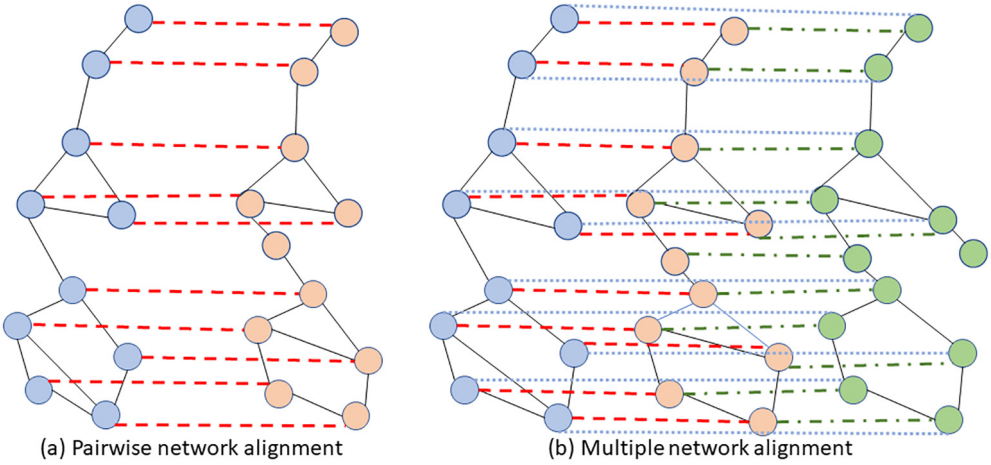

Fig. 2. Pairwise network alignment vs. Multiple network alignment.

network to at most one node of another network. Following the same logic, one-to-many network alignment maps one node of a given network to multiple nodes of another network. Many-to-many network alignment, on the other hand, maps a group of nodes of a given network to a group of nodes of another network, where the nodes in each of the groups are conserved in neighborhood topology and/or sequence similarity. The three types of node mappings fit different purposes.

As one use *edge correctness* or the number of conserved edges as the standard measure to evaluate network alignment outputs, one-to-one and one-to-many could have a better result. However, in biological networks, many protein/gene duplication, mutation and interaction rewiring events occurred during evolution.

Moreover, proteins/genes usually works as a complex or module which is represented as a community in the biological network. Thus, a many-to-many node mapping can be considered more reasonable and closer to the real world scenario because it can align functional similar complexes/modules between different networks (species).

That is, it is difficult to find perfectly matched neighborhood topology between nodes in different biological networks. On the other hand, it is hard to evaluate the topological quality of many-to-many node mappings, in comparison with one-to-one and one-to-many mappings, where the latter two have been more widely studied in the literature.

1.5. Biological similarity vs. topological similarity

Network alignment algorithms cluster nodes between networks mainly based on two measurements: biological similarity and topological similarity. Biological similarity represents the similarity between two biological objects themselves across networks, which is typically sequence similarity obtained from BLAST in most cases of biological networks. Topological similarity, on the other hand, describes how similar between the edge (interaction) patterns of two nodes' neighborhood. There have been many definitions to measure topological similarity in the literature, such as edge degree, edge density, eccentricity, edge coefficient, graphlet degree and so on. Later we will introduce some of them which have been often used in the field of network alignment.

1.6. Dataset

There are many databases which provide PPI data, such as DIP [60], HPRD [55], MIPS [45], IntAct [30], BioGRID [49], and STRING [66]. However, we particularly introduce two datasets: IsoBase

[51] and NAPAbench [58] because they are the most commonly used in many studies as evaluation datasets. IsoBase provides real PPI networks of five eukaryotes: yeast, worm, fly, mouse, and human that were collected from DIP, BioGRID and HPRD. Moreover, IsoBase identifies functionally related orthologs across the five organisms by IsoRankN based on sequence similarity and PPI data. In contrast with IsoBase, NAPAbench is a synthetic PPI dataset. Owing to the limitation of real PPI data and the lack of perfect alignment between real PPI networks, NAPAbench dataset offers a set of synthetic networks with no false positive/negative interactions. The authors of NAPAbench first observed intra-network properties of every individual network as well as cross-network properties between different networks in the real PPI data of the five species from IsoBase. Then, they generated families of synthetic PPI networks with three different *network growth models*, DMC (Duplication Mutation Completion) [69], DMR (Duplication with Random Mutation) [64,52], and CG (Crystal Growth) [31] based on an input phylogenetic relationships.

2. Evaluation methods

In the network alignment problem, unfortunately there is no gold standard for evaluation; that is, a best alignment is unknown from a biological perspective. Thus, it is important to find an approach for evaluating the quality of network alignment results from different perspectives. For the alignment between PPI networks, there are two main points of view to assess the results of network alignment: biological and topological evaluations. The former measures the consistency of biological functions between aligned proteins, if any, in an alignment, while the latter just measures several topological features of an alignment. Next we introduce several widely-used measurements in the field of network alignment. In order to explain the definitions of the following assessment methods, we recall the problem definition of network alignment. Given a network $G_n = (V_n, E_n)$, where n denotes the graph identity, V_n and E_n represent the set of nodes and edges in G_n respectively. We let $f : V_i \to V_j$ be a network alignment mapping function between two graphs G_i and G_j.

2.1. Biological evaluation

Most biological evaluation measures assess the functional similarity of aligned proteins based on Gene Ontology (GO) annotations [67,4]. GO is a hierarchical system for unifying the representation and annotation of gene and gene product attributes across all species. The ontology mainly have three domains, cellular component, molecular function, and biological process. Most GO-derived assessment methods are based on the GO terms of biological domains. Note that many GO terms are assigned to genes or gene products mainly based on sequence homology. The most trivial way to evaluate functional similarity between proteins is the ratio of common GO terms between proteins. However, there have been more elaborate approaches proposed in recent years.

2.1.1. Functional Coherence (FC)

The FC was proposed by Singh et al. [63,9] and measures the functional consistency of the mapped proteins. The FC value of a mapping is computed as the average pairwise FC of the protein pairs that are aligned. A higher FC score indicates that the proteins in the mapping perform more similar functions. The method for computing FC values can be summarized as follows. First, the GO terms corresponding to each protein are collected. Notice that GO terms are a hierarchical description of protein functions. Then, each GO terms is mapped to a subset of the so-called *standardized* GO terms, which in this case are its ancestors lying within a fixed distance from the root of the GO tree. Finally, the similarity between each pair of aligned proteins is computed as the median of the fractional overlaps to their corresponding sets of standardized GO terms. The FC of each protein pair is defined to be:

$$FC(u, v) = \frac{|S_u \cap S_v|}{|S_u \cup S_v|}$$

where S_u and S_v are the GO term sets of protein u and protein v, respectively, for $u \in V_1$ and $v \in V_2$.

2.1.2. Resnik's semantic similarity

Resnik's semantic similarity is a semantic similarity metric in a hierarchical IS-A taxonomy based on the notion of information content [57]. This method considers the hierarchical structure of ontology and can be used in GO consistency measurement [54,61,43]. Given two networks G_1 and G_2, $|V_1| \leqslant |V_2|$, the Resnik's Semantic Similarity of an alignment between G_1 and G_2 is defined to be:

$$S_{Res}(G_1, G_2) = \frac{1}{|V_1|} \sum_{u \in V_1, f(u) \in V_2} Sim_{Res}(GO(u), GO(f(u)))$$

$$Sim_{Res}(t_1, t_2) = IC(t_{MICA})$$

$$IC(t) = -\log p(t)$$

where f is a network alignment mapping, $GO(u)$ is a set of GO terms of protein u, t_{MICA} is the Most Informative Common Ancestors between t_1 and t_2, t is a single GO term, IC is the information content of t, and p is the probability of occurrence of t in a specific corpus (in this case it is GO).

2.1.3. GO/KEGG entropy

Another GO-derived measurement is the entropy of the GO/KEGG annotations proposed by Liao et al. [38,59,2,23]. This measurement is designed to measure within-cluster consistency and GO/KEGG enrichment of the output clusters produced by many-to-many network alignment algorithms [67,28]. The entropy of a given cluster S_v^* is defined to be:

$$H(S_v^*) = H(p_1, p_2, \ldots, p_d) = -\sum_{i=1}^{d} p_i \log p_i$$

where p_i is the fraction of S_v with GO or KEGG group ID i. There is also a normalized version $\overline{H}(S_v^*) = \frac{1}{\log d} H(S_v^*)$ which normalizes the entropy by cluster size. Obviously, if a cluster has lower entropy, its GO/KEGG annotations are more within-cluster consistent.

2.1.4. Gene Ontology Consistency (GOC)

Gene Ontology Consistency (GOC) [1,10] is one of the most naive methods to assess the quality of an alignment by using the Jaccard index on sets of GO terms between two aligned proteins. The definition of GOC is:

$$GOC(G_1, G_2) = \sum_{(u_i, v_j) \in M} \frac{|GO(u_i) \cap GO(v_j)|}{|GO(u_i) \cup GO(v_j)|}$$

where (u_i, v_j) is a pair of matched nodes in the alignment M of G_1 and G_2, $u_i \in V_1$, $v_j \in V_2$. $GO(u)$ is a set of GO terms of protein u. Later, the normalized version of GOC was proposed by Elmsallati et al. [17]. The NGOC is defined to be:

$$NGOC(G_1, G_2) = \frac{1}{n} \left(\sum_{(u_i, v_j) \in M} \frac{|GO(u_i) \cap GO(v_j)|}{|GO(u_i) \cup GO(v_j)|} \right)$$

where n is the total number of aligned proteins.

2.2. Topological evaluation

2.2.1. Edge Correctness (EC)

Edge correctness [35,48,61,9,10] is one of the most straightforward methods to evaluate the quality of an alignment from the topological perspective. EC is the percentage of edges in the first network, which are aligned to edges in the second network. The definition of EC is:

$$EC(G_1, G_2) = \frac{|\{(u, v) \in E_1 \wedge (f(u), f(v)) \in E_2\}|}{|E_1|}$$

A slightly modified version of EC was proposed to consider the varied size of PPI networks [9]. The definition of the modified EC is:

$$EC(G_1, G_2) = \frac{1}{2}\left(\frac{|\{(u, v) \in E_1 \wedge (f(u), f(v)) \in E_2\}|}{|E_1|} + \frac{|\{(u, v) \in E_1 \wedge (f(u), f(v)) \in E_2\}|}{|E_2|}\right)$$

2.2.2. Induced Conserved Structure (ICS)

To overcome the EC's drawback that an alignment with a high EC value may not necessarily be the best alignment. For example, a sparse section of G_1 could map to a dense region of G_2 easily with a high EC score but actually, it is not a good mapping [53,35,48,36,10]. ICS penalizes this kind of circumstances by dividing the number of conserved edges by the edge number in the subnetwork of G_2 induced by the nodes that are aligned to the nodes of G_1. The definition of ICS is as follows:

$$ICS(G_1, G_2) = \frac{|\{(u, v) \in E_1 \wedge (f(u), f(v)) \in E_2\}|}{|E(G_2[f(V_1)])|}$$

2.2.3. Symmetric Substructure Score (S^3)

Symmetric substructure score, proposed in [61,70–72,43], further improves the weakness of ICS that ICS only considers the sparse-to-dense problem but does not punish dense-to-sparse condition. S^3 has been shown that it is a better predictor of correct alignment than EC and ICS for metaheuristic aligners [61]. The definition of S^3 is:

$$S^3(G_1, G_2) = \frac{|\{(u, v) \in E_1 \wedge (f(u), f(v)) \in E_2\}|}{|E_1| + |E(G_2[f(V_1)])| - |\{(u, v) \in E_1 \wedge (f(u), f(v)) \in E_2\}|}$$

2.2.4. Largest Common Connected Component (LCCS)

LCCS is the size of the largest common connected sub-graph induced by aligned nodes [63,61,10,43]. It reflects the connectivity of the aligned sub-graph between networks.

2.2.5. Node Correctness (NC) and Interaction Correctness (IC)

Another two measures, NC [35,48,36,61,43] and IC [48], were discussed only on synthesized networks for benchmarking because the true (correct) alignment is required; however, it is impossible for real biological networks [58]. The definition of NC is:

$$NC(G_1, G_2) = \frac{|\{u_i | f(u_i) = g(u_i)\}|}{|V|},$$

and the definition of IC is:

$$IC = \frac{|\{(u, v) \in E_1 | (f(u), f(v)) \in E_2, f(v) = g(v)\}|}{|E_1|},$$

where $u_i \in V_1$, and $g(u_i)$ is the true mapping.

3. Methodology

In this section, we first discuss some landmark aligners which have received a considerable amount of attention in the literature in the past two decades. Next we talk about other recent tools and then make a summary.

3.1. Pathway based

In the early stage of the development of network alignment algorithms, only pairs of pathways between networks were considered. Kelly et al. [29] introduced the first local network alignment algorithm, PathBLAST in 2003. PathBLAST searches for pairs of pathways between protein–protein interaction networks based on the sequence similarity of proteins and the same order of the putative orthologs in each pathway. It considers only the simplest topology similarity of the aligned subnetwork. Later, Sharan, R. et al. [62,25,26] brought the NetworkBLAST and NetworkBLAST-M as an extension of PathBLAST. NetworkBLAST generates a network graph by integrating interactions with sequence similarity. In this graph, each node comprises a pair of sequence-conserved proteins, one from each network, and each edge represents conserved interactions between corresponding proteins from each network. NetworkBLAST seeks two types of conserved subnetworks over the graph, short linear paths and dense clusters. The first type models the signal transduction pathways and the second type models protein complexes. The search algorithm follows the seed-extension strategy in a greedy fashion by exhaustively identifying high-scoring subnetwork seeds based on the reliability of protein interactions.

3.2. IsoRank series

IsoRank [63] which uses the concept of the well-known Google PageRank algorithm [7] is the first global pairwise network alignment approach. IsoRank simultaneously considers both protein sequence and topology similarity in an eigenvalue-based framework to score protein pairs between networks, and then generates the match by high-scoring protein pairs. Also, IsoRank can tolerate gap alignment when a node cannot find a good mapping. In 2009, a multi-aligner version of IsoRank, IsoRankN [38], was introduced. It inherits the same scoring function of IsoRank, but further uses another spectral graph-theoretic concept, similar to the idea of PageRank-Nibble algorithm [3], to identify a many-to-many overall mapping. It uses iterative spectral clustering algorithm to replace the second phase of IsoRank and generates the final result. Later, Kollias et al. [33,34] used network similarity decomposition approach and half-approximation algorithm in the two phases of IsoRank, respectively, to speed up IsoRank significantly.

3.3. Graphlet based

GRAAL [35], a pairwise alginer proposed in 2010, is the first network alignment algorithm which purely relies on topological similarity. The idea of GRAAL uses a graphlet vector to describe topological features of local neighborhood for each node in both the networks, and then heuristically aligns the nodes that have the smallest distance between their graphlet vectors. Next, a series of extentions of GRAAL, such as H-GRAAL [48], Mi-GRAAL [36], C-GRAAL [44], and L-GRAAL [42] were presented later on.

H-GRAAL also relies on the graphlet degree similarity and employs the same cost function used by GRAAL to define the quality of network alignment, but H-GRAAL replaces the greedy seed-and-extension algorithm with the Hungarian algorithm to search the minimum cost of a bipartite matching. The change leads to a better result but costs more time complexity, i.e. $O(n^3)$ than GRAAL. Later, Mi-GRAAL, like the concept of cocktail, can integrate multiple types of similarity metrics, including sequence and topological similarity, between nodes of different networks. In addition

to the graphlet degree vector distance and sequence similarity, Mi-GRAAL integrates the other three topological similarities: degree difference, clustering coefficient difference, and eccentricity difference. Mi-GRAAL combines the five matrices into one confidence matrix, and again uses a seed-and-extension algorithm to find the final alignment. Regarding the last two variations of GRAAL, C-GRAAL employs not only graphlet degree signatures but also neighborhood density to illustrate the presences of similar neighborhoods. L-GRAAL integrates the seed-and-extend strategy with graphlet degree signatures by Lagrangian relaxation technique. Moreover, both of them have a flexibility to include sequence similarity to provide additional information to the algorithms.

3.4. Index-based series

One of the most common challenges for network alignment algorithms is the exponential time computational complexity. In order to reduce the running time of searching alignment, some research used indices to accelerate the execution process. In this section, we introduce two recent index-based approaches.

3.4.1. IBNAL

IBNAL, proposed by Elmsallati et al. [17] first divides the PPI network into two individual subsets and creates a *Clique Degree Signature* vector for each node in the subset of subordinate nodes to keep the number of cliques that the subordinate node connected. Then IBNAL indexes all subordinate nodes and cliques for accelerating the next step of alignment extraction. In the next step, a similarity matrix is maintained to describe the Euclidean distance between every two subordinate node vectors. The distance is zero if and only if two subordinate nodes connect the same number of cliques with the same sizes; that is, they are likely to be matched. Based on the similarity scores, an ordered priority queue is built for all pairs of subordinate nodes. Subordinate pairs are popped from the queue and matched along with all the cliques they connected until the queue is empty. Therefore, cliques with the same size are aligned to each other. The experimental result showed that IBNAL favors the S^3 metric to other state-of-the-art approach, and it has comparable but unstable GOC score.

3.4.2. SSAlign

Another index-based approach for indexing all maximal substructure of up to a specified size is called SSAlign, proposed by Elmsallati et al. [16]. It excludes larger substructures due to the expensive time cost of extracting and indexing. Note that when conducting its implementation, they actually considered substructures of size up to five only. The first step of SSAlign is to build a priority queue for each type of symmetric substructures based on their GOC score. Then SSAlign starts aligning the symmetric substructures with the highest GOC score from the queue. All the partially matched substructures are put into another queue for the next step in this phase. Next, all the partially matched substructure that were queued are popped, and all unaligned nodes in the partially matched substructure are assigned to their corresponding nodes. Finally, the aligning process matches all the neighbors of each node in the alignment set. SSAlign was shown to be the only aligner that can obtain S^3 and GOC scores at the same time [16], in comparison with other alignment approaches.

3.5. Swap-based series

Next, we introduces three methods based on swap strategy to optimize the network alignment results. Moreover, PISwap and MAGNA can be used as a booster to refine the alignment generated by other approaches.

3.5.1. PISwap

PISwap is a pairwise global alignment booster that uses 3-Opt heuristic swapping strategy to refine an existing alignment efficiently [9]. Also, PISwap can produce an alignment by itself and improve it through swapping. The aligner first generates an alignment by using the famous Hungarian algorithm to derive a maximum weight bipartite match between two networks. Then PISwap swaps the edges of the bipartite graph by interchanging the mapping of three nodes at a time. The swapping step does not affect the biological similarity matching but improve the number of conserved edges.

3.5.2. MAGNA

MAGNA is a Genetic Algorithm (GA) based pairwise global network alignment approach [61]. It uses GA to optimize biological and topological similarities of match nodes. The GA needs to maintain a population pool of candidate solutions (i.e. a solution is an alignment). In the pool, the best solution remains unchanged but randomly recombined with each other to generate new candidates for each generation. Note that there is no mutation operation in MAGNA. The initial pool (i.e. alignments) can be generated by randomness or other approaches. The solutions (i.e. matches) are ranked by their scores (based on biological and topological similarities). The evolution process repeats until there is no further improvement or given criteria are reached. Later MAGNA++ [70] and multiMAGNA++ [71], which are the parallel and multiple version of MAGNA respectively, were proposed. In recent years, the DynaMAGNA++ [72] further extends to align dynamic networks and adds a time-series factor to describe an edge as a time event. For every pair of mapped edges, conserved event time (CET) is defined to be the amount of time during which the two edges exists at the same time. Besides, the entire amount of time during which the two edges are non-conserved is called non-conserved event time (NCET) [72]. The goal of DynaMAGNA++ is to find an alignment by maximizing the CET while minimizing the NCET.

3.5.3. Optnetalign

Optnetalign [10] also uses a branch of GA, i.e. a multi-objective memetic algorithm, to discover the optimal solution between both the goals of sequence and topological similarity by the concept of Pareto dominance. It outputs a wide variety of representative solutions from the Pareto optimal set to cover a number of alignments with different tradeoffs between sequence and topology.

3.6. Multiple aligner

3.6.1. Graemlin 1.0 and 2.0

Graemlin 1.0 [20] is a multiple local network alignment approach based on *progressive alignment* and seed-extension strategies. Graemlin 1.0 uses phylogenetic relationship to relate the species while aligning networks. It first successively aligns the closest networks pairs with a seed-extension strategy based on the phylogenetic tree. Next, it transforms the alignment result with the unaligned nodes into generalized networks for the next phase of progressive alignment with the next closest network. Graemlin 2.0 [21] which was later introduced automatically learns for a scoring function to search the approximate matches based on the training set of known network alignments and phylogenetic information.

3.6.2. SMETANA

SMETANA is a many-to-many global alignment algorithm across multiple networks [59]. It uses a semi-Markov random walk model to compute the node correspondence scores and effectively output the maximum expected alignment for large networks. Precisely, SMETANA works through two steps. In the first step, it

calculates a similarity matrix, which describes the likelihood between any pair of nodes from different networks based on the semi-Markov random walk model. Then, the algorithm incorporates the matrix with biological information and uses both intra-network and cross-network probabilistic consistency transformation to improve the estimated probabilities of the alignment of all node pairs. In the second step, SMETANA uses a greedy seed-and-extension strategy to construct the final alignment based on the scoring matrices obtained in the first step.

3.6.3. BEAMS

BEAMS (Backbone Extraction And Merge Strategy) is also a many-to-many global alignment algorithm between multiple networks. It first constructs a weighted k-partite similarity graph from the k input networks and extracts all possible cliques greedily from the k-partite similarity graph as seeds (backbones) based on their pairwise sequence similarity. In order to maximize the total alignment score of the whole clustering output, these seeds are iteratively clustered and merged by the seed-and-extension strategy.

3.6.4. NetCoffee

NetCoffee is another many-to-many global aligner across multiple networks. It is an extension of T-Coffee, which is a multiple sequence alignment approach [23]. The key idea of NetCoffee uses simulated annealing, a metaheuristic search technique, on a set of weighted bipartite graphes to maximize a target function for seeking a global alignment match. The algorithm first builds a bipartite graph library, i.e., a set of bipartite graphs representing each pair of networks. Then it uses an integrated strategy, just similar to T-Coffee, to assign weights to edges in the bipartite graphs. Next, NetCoffee selects candidate edges from all bipartite graphs and reduces the search space to speed up the final alignment step. In the last step, NetCoffee outputs the final alignment by using simulated annealing to optimize a target function.

3.7. Other aligners

3.7.1. PrimAlign

PrimAlign [27] is a global pairwise and many-to-many network aligner. It models networks as a Markov chain in which each node in one network is related to nodes in the other network based on their protein sequence similarity scores. The Markov chain is transited iteratively state by state until convergence, and re-distributes the scores of relations between nodes across networks by a PageRank-inspired algorithm. Finally, all the relations are filtered out by a certain threshold, and those nodes that involve remaining relations are matched to form an alignment. Kalecky and Cho [27] showed that the time complexity of PrimAlign is $O(n)$, which is the theoretically minimum for this problem and thus it guarantees high scalability.

3.7.2. SANA

SANA (Simulated Annealing Network Aligner) is another pairwise global network alignment algorithm [43] using simulated annealing. SANA considers the network alignment problem into two parts: an objective function which measures the quality of an alignment, and a search algorithm which investigates a good solution according to the objective function. It also uses the simulated annealing technique to find an alignment that can maximize the objective function which is a linear combination of sequence and topological similarities. The algorithm generates a new alignment by randomly changing one or two mappings of individual pairs of aligned nodes and chooses the better one according to the objective function in each iteration. Besides, the algorithm uses a temperature schedule function $T()$ to determine the probability to accept the worse solution in each iteration. They claimed that if given a perfect objective function, the search algorithm can quickly converges to a perfect solution while many other approaches falter.

3.7.3. Network query and complex identification

In addition to the above network alignment approaches, there are also some approaches that use network alignment techniques to solve other tasks such as network query [13,6,74,75,14] and protein complex identification. For example, QNet [13] was the first algorithm developed for querying subgraphs in PPI networks. It defines the similarity between two networks based on node and edge similarity with the penalty for node deletion and insertion. Then QNet uses the color coding algorithm to perform tree queries and bounded-treewidth graph queries. On the other hand, regarding protein complex identification, NEOComplex [41] integrates functional similarity orthology information that can obtain from different types of multiple network alignment approaches to expand the search space of protein complex detection across different species. NEOComplex identifies candidate complexes for different networks based on new edge clustering coefficient (NECC) and expands the candidate complexes from one species to the others. To incorporate multiple network alignment into the protein complex identification task enables NEOComplex to tolerate edge loss in PPI networks and even to discover sparse protein complexes that have traditionally been a challenge to predict.

Aligner	Year	Local/global	Pairwise/multiple	One-to-one/many-to-many	Features
PathBLAST [29]	2004	Local	Pairwise	Many-to-many	Linear chain topology
NetworkBLAST [62]	2008	Local	Pairwise	Many-to-many	Complex detection
NetworkBLAST-M [26]	2008	Local	Multiple	Many-to-many	Layered alignment graph
Graemlin [20]	2006	Local	Pairwise	Many-to-many	Phylogenetic information, Progressive alignment
Graemlin 2.0 [21]	2009	Local	Multiple	Many-to-many	Phylogenetic information, Machine learning
IsoRank [63]	2008	Global	Pairwise	One-to-one	PageRank
IsoRankN [38]	2009	Global	Multiple	Many-to-many	PageRank-Nibble, Spectral graph theory
GRAAL [35]	2009	Global	Pairwise	One-to-one	Graphlet, Purely topological
H-GRAAL [48]	2010	Global	Pairwise	One-to-one	Graphlet, Purely topological, Hungarian algorithm

(continued on next page)

(*continued*)

Aligner	Year	Local/global	Pairwise/multiple	One-to-one/many-to-many	Features
MI-GRAAL [36]	2011	Global	Pairwise	One-to-one	Graphlet, Multi-types of similarity metrics
C-GRAAL [44]	2012	Global	Pairwise	One-to-one	Graphlet, Neighborhood density, Purely topological
L-GRAAL [42]	2015	Global	Pairwise	One-to-one	Graphlet, Lagrangian relaxation
SMETANA [59]	2013	Global	Multiple	Many-to-many	Semi-Markov random walk
BEAMS [2]	2014	Global	Multiple	Many-to-many	Backbone extraction and merge strategy
NetCoffee [23]	2014	Global	Multiple	Many-to-many	Metaheuristic search
PISwap [9]	2013	Global	Pairwise	One-to-one	3-Opt heuristic swapping, Alignment booster
MAGNA [61]	2014	Global	Pairwise	One-to-one	Metaheuristic search, Symmetric substructure score
MAGNA++ [70]	2015	Global	Pairwise	One-to-One	Parallel version of MAGNA
multi-MAGNA++ [71]	2016	Global	Multiple	One-to-one	Multiple version of MAGNA++
IBNAL [17]	2016	Global	Pairwise	One-to-one	Clique-based index
SSAlign [17]	2017	Global	Pairwise	One-to-one	Symmetric substructure
DynaMAGNA++ [72]	2017	Global	Pairwise	One-to-one	Metaheuristic search, Dynamic network alignment
Optnetalign [10]	2015	Global	Pairwise	One-to-one	Metaheuristic search, Pareto optimality
SANA [43]	2017	Global	Pairwise	One-to-one	Metaheuristic search, Temperature schedule function
PrimAlign [27]	2018	Global	Pairwise	Many-to-Many	Markovian representation, PageRank

4. Conclusion and discussion

In this survey paper, we have reviewed some well-known algorithms for network alignment between protein–protein interaction networks in the past two decades and compared them with recent aligners from a different perspective. We have also presented the most widely used measures in the literature, regarding biological functions and network topological properties, to evaluate the performance of these approaches. In particular, we have introduced several types of classification for network alignment algorithms, and addressed the characteristics of different types of aligners, which may benefit the understanding of evolutionary relationship across species, and even help identify conserved functions.

Generally speaking, there are still some flaws in current biological and topological assessments. For instance, concerning the widely-used GO-based evaluation, the core issue is that many GO terms are assigned mainly based on sequence homology [50]. This could lead to a bias because many algorithms generate alignment, based on or at least partially based on sequence similarity. Another issue is that most of GO-derived metrics did not consider the hierarchical structure of GO. This may mislead the evaluation results. On the other hand, regarding the topological assessments, we also have some observations. For example, NC and IC require the ground truth of the alignment, which is often unavailable in real biological networks [15]. Therefore, NC and IC are not that suitable for real PPI networks. Another example is that EC may not be able to distinguish better alignment from worse alignment since both of them may have the same EC. Hence, the performance evaluation cannot purely rely on topological assessments, even considering topology-based aligners such as graphlet-based algorithms.

All in all, the alignment algorithms that are mainly or partially based on sequence similarity, may favor biological assessments more. It also happens for topology-based aligners which may have better results for topological assessments. Even sometimes, biological and topological assessments might have trade-off relationship [15]. Here we particularly remark that duo to the nature of evolution, it is hard to have comparable studies between conducting a one-to-many or many-to-many multiple aligner and performing a one-to-one pairwise aligner many times across multiple networks. Precisely, many-to-many multiple aligners usually perform better in finding conserved complexes or functional modules than one-to-one pairwise aligners because they better tolerate the edge loss in PPI networks and they even directly or indirectly provide intra-species relationships between proteins. Moreover, many-to-many multiple aligners are more suitable for discovering phyletic relationships between different organisms from the systems-level perspective, analogous to the fact that multiple sequence alignment can help construct phylogenetic relationships. In contrast, one-to-one pairwise aligners are good at detecting similar topological substructures as well as functional orthologs.

Comparative study of biological networks has the potential to explore the mechanism of life from a comprehensive, systems point of view. For the purpose of revealing a deeper insight, network alignment is clearly a powerful and reasonable way. However, there are still open problems that have not yet been solved through all the efforts made in the past decades. The fist problem is that there is no gold standard ground truth for network alignment across protein–protein interaction networks because the nature of biological research is basically a reverse-engineering process due to the unknown mechanisms of evolutional events. Moreover, there remains noise and information loss in protein–protein interaction data caused by the limitation of existing molecular operating techniques. The second issue needed to be addressed is that there have been getting more and more large and dense biological networks, and even different types of networks discovered. It is obviously a challenge for algorithm design and even for physical limits of computing power. Nevertheless, network alignment could be applied to not only systems biology, but also many other fields, such as neural science, social network analysis and knowledge management [46,47,37,39,65].

In recent years, deep neural networks (DNN) have demonstrated a tremendous success in many applications especially convolution networks. In addition, graph convolution neural networks have been exploited in graph clustering and classification

[32,76,73,18,8] while network alignment tasks can be seen as a clustering problem between the nodes from different networks. It would be worthwhile to transform network alignment problems into a multivariate clustering optimization problem, where the latter one could be solved by using graph deep learning approaches. On the other hand, concerning no gold standard ground truth for biological network alignment, it may rely on semi-supervised or unsupervised deep learning approaches [12,56,68], which can overcome similar challenges in the area of data mining.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by MOST Taiwan under Grants MOST108-2634-F-007-010 and MOST109-2634-F-007-018, and the Maintenance Project of the Center for Artificial Intelligence in Medicine (Grant CLRPG3H0012, CIRPG3H0012) at CGMH.

References

[1] Aladag AE, Erten C. SPINAL: scalable protein interaction network alignment. Bioinformatics 2013;29(7):917–24.
[2] Alkan F, Erten C. BEAMS: backbone extraction and merge strategy for the global many-to-many alignment of multiple PPI networks. Bioinformatics 2014;30(4):531–9.
[3] Andersen R, Chung F, Lang K. Local graph partitioning using pagerank vectors. In: Proceedings of the 47th annual IEEE symposium on foundations of computer science. Berkeley, California, USA; 2006. p. 475–86..
[4] Ashburner M et al. Gene Ontology: tool for the unification of biology. Nat Genet 2000;25(1):25–9.
[5] Bader JS et al. Gaining confidence in high-throughput protein interaction networks. Nat Biotechnol 2004;22:78–85.
[6] Blin G, Sikora F, Vialette S. Querying protein-protein interaction networks. In: Măndoiu I, Narasimhan G, Zhang Y., editors. Bioinformatics research and applications. ISBRA 2009. Lecture notes in computer science, 5542; 2009..
[7] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Comput Net ISDN Syst 1998;30:107–17.
[8] Chiang, W-L. et al. Cluster-GCN: an efficient algorithm for training deep and large graph convolutional networks KDD 2019; 2019. arXiv:1905.07953..
[9] Chindelevitch L et al. Optimizing a global alignment of protein interaction networks. Bioinformatics 2013;29(21):2765–73.
[10] Clark C, Kalita J. A multiobjective memetic algorithm for PPI network alignment. Bioinformatics 2015;31(12):1988–98.
[11] Deane CM et al. Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol Cell Proteom 2002;1(5):349–56.
[12] Derr T. et al. Deep adversarial network alignment; 2019. arXiv, 1902.10307.
[13] Dost B et al. QNet: A tool for querying protein interaction networks. J Comput Biol 2008;15(7):913–25.
[14] El-Kebir M et al. NatalieQ: A web server for protein-protein interaction network querying. BMC Syst Biol 2014;8:40.
[15] Elmsallati A, Clark C, Kalita J. Global alignment of protein-protein interaction networks: a survey. IEEE/ACM Trans Comput Biol Bioinform 2015;13(4):689–705.
[16] Elmsallati A, Roy S, Kalita JK. Exploring symmetric substructures in protein interaction networks for pairwise alignment. Int Conf Bioinf Biomed Eng 2017:173–84.
[17] Elmsallati A, Msalati A, Kalita JK. Index-based network aligner of protein-protein interaction networks. IEEE/ACM Trans Comput Biol Bioinf 2018;15(1):330–6.
[18] Fey M. Just jump: dynamic neighborhood aggregation in graph neural networks. ICLR 2019; 2019. arXiv:1904.04849..
[19] Fields S, Song O. A novel genetic system to detect protein-protein interactions. Nature 1989;340(6230):245–6.
[20] Flannik J et al. Graemlin: General and robust alignment of multiple large interaction networks. Genome Res 2006;16(9):1169–81.
[21] Flannik J et al. Automatic parameter learning for multiple local network alignment. J Comput Biol 2009;16(8):1001–22.
[22] Gavin AC et al. Proteome survey reveals modularity of the yeast cell machinery. Nature 2006;440:631–6.
[23] Hu J, Kehr B, Reinert K. NetCoffee: a fast and accurate global alignment approach to identify functionally conserved proteins in multiple networks. Bioinformatics 2014;30(4):540–8.
[24] Johnson DS et al. Genome-wide mapping of in vivo protein-DNA interactions. Science 2007;316:1497–502.
[25] Kalaev M et al. NetworkBLAST: comparative analysis of protein networks. Bioinformatics 2008;24(4):594–6.
[26] Kalaev M, Bafna V, Sharan R. Fast and accurate alignment of multiple protein networks. In: Proceedings of the 12th annual international conference on research in computational molecular biology. p. 246–56.
[27] Kalecky K, Cho YR. PrimAlign: PageRank-inspired Markovian alignment for large biological networks. Bioinformatics 2018;34:i537–46.
[28] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucl Acids Res 2000;28:27–30.
[29] Kelley BP et al. Pathblast: a tool for alignment of protein interaction networks. Nucl Acids Res 2004;32(suppl 2):W83–8.
[30] Kerrien S et al. The intact molecular interaction database in 2012. Nucl Acids Res 2012;40:841–6.
[31] Kim WK, Marcotte EM. Age-dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence. PLoS Comput Biol 2008;4(11):e1000232.
[32] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. ICLR 2017; 2016. arXiv:1609.02907..
[33] Kollias G, Mohammadi S, Grama A. Network similarity decomposition (nsd): A fast and scalable approach to network alignment. IEEE TransKnowl Data Eng 2012;24(12):2232–43.
[34] Kollias G, Sathe M, Mohammadi S, Grama A. A fast approach to global alignment of protein-protein interaction networks. BMC Res Notes 2013;6(1):35.
[35] Kuchaiev O et al. Topological network alignment uncovers biological function and phylogeny. J R Soc Interf 2010;7(50):1341–54.
[36] Kuchaiev O, Pržulj N. Integrative network alignment reveals large regions of global network similarity in yeast and human. Bioinformatics 2011;27(10):1390–6.
[37] Li C. et al. Adversarial learning for weakly-supervised social network alignment. In: Thirty-Third AAAI Conference on Artificial Intelligence; 2019..
[38] Liao CS et al. IsoRankN: spectral methods for global alignment of multiple protein networks. Bioinformatics 2009;25(12):253–8.
[39] Liu L et al. Aligning users across social networks using network embedding. Int Joint Conf Artif Intell 2016:1774–80.
[40] Ma CY et al. Reconstruction of phyletic trees by global alignment of multiple metabolic networks. BMC Bioinf 2013;14(Suppl 2):S12.
[41] Ma CY et al. Identification of protein complexes by integrating multiple alignment of protein interaction networks. Bioinformatics 2017;33(11):1681–8.
[42] Malod-Dognin N, Pržulj N. L-GRAAL: Lagrangian graphlet-based network aligner. Bioinformatics 2015;31(13):2182–9.
[43] Mamano N, Hayes WB. SANA: simulated annealing far outperforms many other search algorithms for biological network alignment. Bioinformatics 2017;33(14):2156–64.
[44] Memišević V, Pržulj N. C-GRAAL: common–neighbors–based global graph alignment of biological networks. Integr Biol 2012;4:734–43.
[45] Mewes H-W et al. Mips: Analysis and annotation of proteins from whole genomes. Nucl Acids Res 2004;32(suppl 1):D41–4.
[46] Milano M, Guzzi PH, Cannataro M. Using multi network alignment for analysis of connectomes. Int Conf Comput Sci, ICCS; 2017..
[47] Milano M et al. An extensive assessment of network alignment algorithms for comparison of brain connectomes. BMC Bioinf 2017;18(Suppl 6):235.
[48] Milenković T et al. Optimal network alignment with graphlet degree vectors. Cancer Inf 2010;9:121.
[49] Oughtred R. et al. The BioGRID interaction database: 2019 update. Nucl Acids Res 2019;47(Database issue):D529–D541..
[50] Pal D. On gene ontology and function annotation. Bioinformation 2006;1(3):97.
[51] Park, D. et al. IsoBase: a database of functionally related proteins across PPI networks. Nucl Acids Res 2011;39(Database issue):D295–D300..
[52] Pastor-Satorras R, Smith E, Solé RV. Evolving protein interaction networks through gene duplication. J Theoretical Biol 2003;222(2):199–210.
[53] Patro R, Kingsford C. Global Network Alignment Using Multiscale Spectral Signatures. Bioinformatics 2012;28(23):3105–14.
[54] Pesquita C et al. Semantic similarity in biomedical ontologies. PLoS Comput Biol 2009;5(7):e1000443.
[55] Prasad TK et al. Human protein reference database–2009 update. Nucl Acids Res 2009;37(suppl 1):D767–72.
[56] Qu M, Tang J, Bengio Y. Weakly-supervised knowledge graph alignment with adversarial learning; 2019. arXiv arXiv:1907.03179.
[57] Resnik P. Using information content to evaluate semantic similarity in a taxonomy; 1995. arXiv preprint cmp-lg/9511007..
[58] Sahraeian SME, Yoon BJ. A network synthesis model for generating protein interaction network families. PLoS ONE 2012;7(8):e41474.
[59] Sahraeian SME, Yoon B-J. SMETANA: Accurate and scalable algorithm for probabilistic alignment of large-scale biological networks. PLoS ONE 2013;8(7):e67995.

[60] Salwinski L et al. The database of interacting proteins: 2004 update. Nucl Acids Res 2004;32(suppl 1):D449–51.
[61] Saraph V, Milenković T. MAGNA: maximizing accuracy in global network alignment. Bioinformatics 2014;30(20):2931–40.
[62] Sharan R et al. Conserved patterns of protein interaction in multiple species. Proc Natl Acad Sci USA 2005;102(6):1974–9.
[63] Singh R, Xu J, Berger B. Global alignment of multiple protein interaction networks with application to functional orthology detection. Proc Natl Acad Sci USA 2008;105(35):12763–8.
[64] Solé RV et al. A model of large-scale proteome evolution. Adv Complex Syst 2002;5(01):43–54.
[65] Sun Z. et al. Knowledge graph alignment network with gated multi-hop neighborhood aggregation. In Thirty-fourth AAAI conference on artificial intelligence; 2019..
[66] Szklarczyk D et al. The string database in 2011: Functional interaction networks of proteins, globally integrated and scored. Nucl Acids Res 2011;39 (suppl 1):D561–8.
[67] The gene ontology consortium. The gene ontology project in 2008. Nucl Acids Res 2008;36(Database issue):D440–D444..
[68] Toan NT, Cong PT, Tho QT. Weakly-supervised network alignment with adversarial. Learning 2019. https://doi.org/10.5121/CSIT.2019.90809.
[69] Vázquez A et al. Modeling of protein interaction networks. Complexus 2002;1 (1):38–44.
[70] Vijayan V, Saraph V, Milenković T. MAGNA11: maximizing accuracy in global network alignment via both node and edge conservation. Bioinformatics 2015;31(14):2409–11.
[71] Vijayan V, Milenković T. Multiple network alignment via multiMAGNA++. In: Proceed-ings of the 15th international workshop on data mining in bioinformatics (BIOKDD) at the 22nd ACM SIGKDD 2016 conference on knowledge discovery & data mining (KDD), San Francisco, CA, USA, August 13–17; 2016..
[72] Vijayan V, Critchlow D, Milenković T. Alignment of dynamic networks. Bioinformatics 2017;33:i180–9.
[73] Wu Z, et al. A comprehensive survey on graph neural networks; 2019. arXiv:1901.00596..
[74] Zhang S, Li S, Yang J. GADDI: Distance index based subgraph matching in biological networks. EDBT 2009:192–203.
[75] Zhang S, Yang J, Jin W. SAPPER: Subgraph Indexing and approximate matching in large graphs. VLDB 2010;3(1):1185–94.
[76] Zhang S et al. Graph convolutional networks: a comprehensive review. Comput Soc Netw 2019;6:11.