Comparative Analysis of Biological Networks Using Markov Chains and Hidden Markov Models

Byung-Jun Yoon, *Member, IEEE*, Xiaoning Qian, *Member, IEEE*, and Sayed Mohammad Ebrahim Sahraeian, *Student Member, IEEE*

I. Introduction to Comparative Network Analysis

The diverse cellular mechanisms that sustain the life of living organisms are carried out by numerous biomolecules, such as DNAs, RNAs, and proteins. During the past decades, significant research efforts have been made to sequence the genomes of various species and to search these genomes to track down genes that give rise to proteins and noncoding RNAs (ncRNAs) [1], [2]. As a result, the catalog of known functional molecules in cells has experienced a rapid expansion. Without question, identifying the basic entities that constitute cells and participate in various biological mechanisms within them is of great importance. However, cells are not mere collections of isolated parts. Biological functions are carried out by collaborative efforts of a large number of cellular constituents, and the diverse characteristics of biological systems emerge as a result of complicated interactions among many molecules [3], [4]. As a consequence, the traditional reductionistic approach, which focuses on studying the characteristics of individual molecules and their limited interactions with other molecules, fails to provide a comprehensive picture of living cells. In order to better understand biological systems and their intrinsic complexities, it is essential to study the structure and dynamics of the networks that arise from the complicated interactions among molecules within the cell.

In recent years, several high-throughput techniques for measuring protein-protein interactions, such as the two-hybrid screening [5] and co-immunoprecipitation followed by mass-spectrometry [6], have enabled the systematic study of protein interactions on a global scale. Since protein-protein interactions are fundamental to all biological processes, a comprehensive *protein-protein interaction (PPI) network* obtained by mapping the protein interactome (complete set of protein interactions) provides an invaluable framework for understanding the cell as an integrated system [7]. Furthermore, literature mining techniques have become increasingly popular to search through the vast amount of scientific literature to collect known biological interactions [8]. Nowadays, there exist a number of public databases, such as BioGRID [9] and DIP [10], that provide access to large collections of molecular interactions. In addition to these public databases, there also exist many commercial databases, such as the Yeast Proteome

B.-J. Yoon and S.M.E. Sahraeian are with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843.

X. Qian is with the Department of Computer Science and Engineering, University of South Florida, Tampa, FL 33620.

Database (YPD) [11], which provides a collection of manually curated PPIs, and Ingenuity [12] and Pathway Studio [13], which provide collections of known functional pathways.

Considering the rapidly growing number and size of biological networks, an important question is how we can utilize the available network data to gain novel biological insights. If we look back into the recent history of molecular biology research, there is no doubt that *comparative methods* have played central roles in analyzing the huge amount of genome sequencing data. In fact, comparative sequence analysis has been shown to be very useful for predicting novel genes and studying the organization of genomes, as well as in many other applications. Similarly, *comparative network analysis* can serve as a valuable tool for studying biological networks. Comparing the networks of different species provides an effective means of identifying functional modules (e.g., signaling pathways or protein complexes) that are conserved across multiple species, and it can lead to important insights into biological systems [14]. Based on the problem settings and goals, comparative network analysis methods

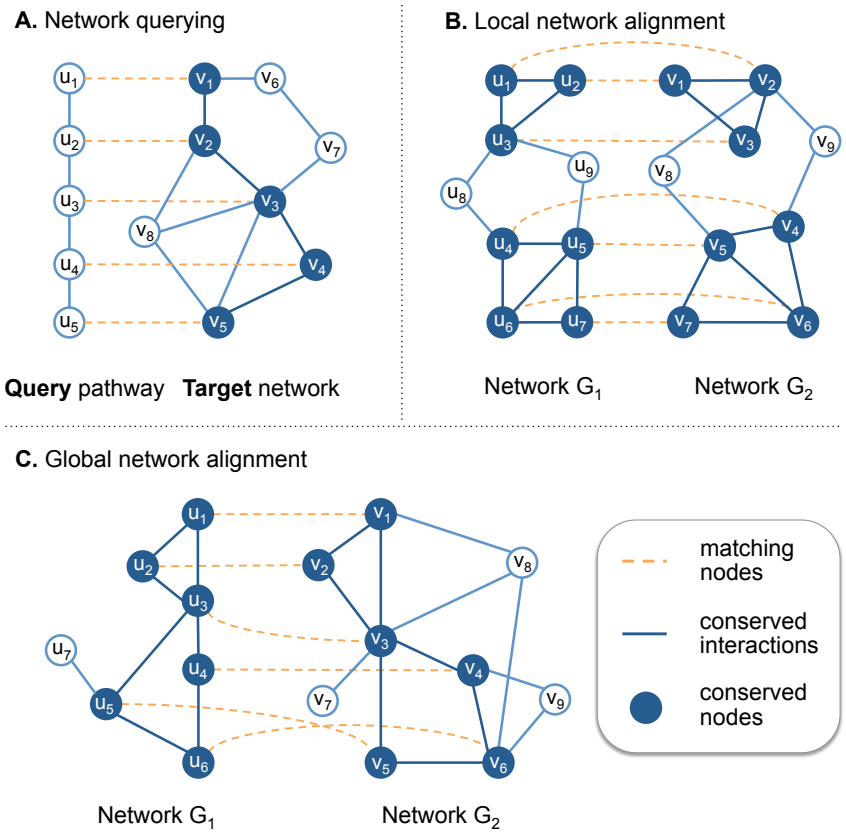

Fig. 1. Three different types of computational methods for comparative network analysis: (A) Network querying finds subnetwork regions in a target network that are similar to a given query pathway. (B) Local network alignment aims to identify similar subnetwork regions in different networks G_1 and G_2. (C) Global network alignment tries to find the overall coherent mapping between nodes in different networks G_1 and G_2.

can be broadly divided into three categories: (1) network querying, (2) local network alignment, and (3) global network alignment.

Network querying aims at finding the subnetworks in a "target network" that are similar to a given "query network." This can be used to search for a known functional module or pathway in the biological network of another species, thereby allowing us to transfer the existing knowledge of a well-studied species to other less-studied species. Figure 1A gives an illustration of network querying. The dashed lines connect the matching nodes in the query and the target networks, and the conserved nodes and interactions in the target network are shown in dark blue. *Local network alignment* tries to identify similar subnetwork regions that belong to different networks. This method can be useful for detecting *novel* functional modules that are conserved across different species. The local network alignment problem is illustrated in Figure 1B. As before, the matching nodes are connected by dashed lines and the nodes and interactions that are conserved in both networks are shown in dark blue. Finally, *global network alignment* aims to find the best overall alignment of two or more networks. This results in a consistent global mapping between nodes that belong to different networks, covering (nearly) all nodes in the given networks. An example of a pairwise global alignment is illustrated in Figure 1C. Finding the global network alignment can be especially useful for studying the cross-species variations in biological networks.

In this paper, we will give a brief review of existing computational methods and tools for comparative network analysis. Especially, we will focus on Markov model based methods and provide a tutorial overview of how Markov models can be used for comparative analysis of large-scale biological networks.

II. Review of Existing Comparative Network Analysis Methods

Recent research efforts to develop efficient tools for comparing biomolecular networks of different species have resulted in a number of promising network *querying* and *alignment* algorithms [15]–[41]. These algorithms compare two or multiple networks to identify regions of similarity. Mathematically, this is achieved by identifying a mapping between partial (or complete) sets of biomolecules in different networks that yields a high similarity score between the (sub)networks induced by these sets of biomolecules.

For simplicity, let us consider the alignment of two networks. Suppose we have two biomolecular networks, represented by two graphs $\mathsf{G}_1 = \{\mathcal{U}, \mathcal{D}\}$ and $\mathsf{G}_2 = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{U}$ and $\mathcal{V}$ are the sets of nodes that correspond to the biomolecules in these networks; $\mathcal{D}$ and $\mathcal{E}$ are the sets of edges that represent the molecular interactions. Both networks can be either directed (e.g., when modeling a regulatory network) or undirected (e.g., when modeling an interaction network). Let $\mathcal{U}' \subset \mathcal{U}$ be a subset of nodes in network G_1 and let $\mathsf{G}_1' = \mathsf{G}_1[\mathcal{U}']$ be the *induced*

network from G_1. Similarly, let $\mathcal{V}' \subset \mathcal{V}$ be a subset of nodes in network G_2 and let $G_2' = G_2[\mathcal{V}']$ be the induced network. The goal of pairwise alignment is to find $\mathcal{U}'$, $\mathcal{V}'$, and the best mapping between nodes $u \in \mathcal{U}'$ and $v \in \mathcal{V}'$ such that their alignment score $S(G_1', G_2')$ is maximized. To obtain biologically meaningful results, the predefined scoring scheme for computing the alignment score $S(G_1', G_2')$ must integrate both the similarity between the individual molecules (based on their compositions and/or functions) and the similarity between their interaction patterns. As shown by a reduction to the graph isomorphism problem [16], [32], [38], the optimal network alignment problem is NP-hard. Due to this reason, many comparative network analysis algorithms impose additional mathematical constraints or adopt various heuristics to make the problem computationally feasible.

A. Network querying: Identifying pathways that are homologous to known pathways

Network querying algorithms (Figure 1A) are used to scan an unannotated biological network (the target network) and search for subnetworks that are similar to a known functional pathway (the query network). PathBLAST [16] is probably the first algorithm to address the querying problem in protein-protein interaction networks. It can identify simple paths containing up to five nodes in a so-called *alignment graph*, in which nodes correspond to pairs of orthologs from the query and target networks and edges correspond to interactions in the respective networks. The search algorithm is based on a greedy *"seed-and-extend"* approach, which finds the alignment of larger subgraphs by growing the alignment of small subgraphs of high similarity. The prediction results are limited to alignments of linear paths with restricted node insertions and deletions. To improve the computational efficiency and flexibility of network querying, several methods have used dynamic programming coupled with randomized techniques such as color coding [42]. Examples include MetaPathwayHunter [18], QPath [23], QNet [26], and Torque [39]. However, many of these algorithms can still handle only queries with specific structures (e.g., paths and trees) [18], [23], [26] or perform querying without explicitly using the topology of the query network [39]. Moreover, the complexity of many algorithms still increase exponentially with the query size, making them impractical for large queries. To reduce the time complexity, another algorithm called PathMatch [25] translated the query problem into that of finding the longest path in an acyclic graph. The computational complexity of PathMatch is polynomial in terms of the query size, where the reduction in complexity comes from allowing multiple occurrences of the same node in the retrieved paths (in the target network). However, this algorithm has more restrictions on insertions and deletions, as the penalty of indels and mismatches are treated in an identical manner. Its extension to queries with a general network structure, called GraphMatch, is highly complex and is only applicable in limited cases.

B. Local network alignment: Detecting conserved functional modules across networks

As mentioned earlier, the aim of local network alignment is to identify similar substructures in different biological networks (Figure 1B). As with network querying, finding the optimal local network alignment is NP-hard, and existing algorithms adopt diverse heuristic techniques to make the alignment problem computationally tractable. Many local alignment algorithms are implemented by extending the ideas used for network querying [16], [22], [23], [25], [26]. For example, NetworkBLAST [19], [28] generalizes the approach in PathBLAST, based on alignment graphs, for aligning two or more networks. There have been also research efforts to improve the scoring scheme by incorporating evolutionary [17] or functional relationships [31], [36] between molecules, with the goal of obtaining better alignment results that are biologically more significant. Essentially, the greedy "seed-and-extend" scheme still lies at the core of most of these algorithms. These methods can be effective for finding conserved subnetworks with relatively small sizes, but they typically suffer from high computational complexity that makes these methods not suitable for finding large subnetworks. Furthermore, these alignment methods have limited flexibility in handling node insertions and deletions and/or rely on randomized heuristics that yield suboptimal results. Some alignment methods adopt the *"divide-and-conquer"* strategy to reduce the overall computational complexity. These methods first partition the given networks into smaller network modules (e.g., using a "match-and-split" approach [43]) and subsequently align these modules to construct the network alignment [44], [45]. Although most network alignment algorithms focus on pairwise alignment, a few algorithms have been also proposed for local alignment of *multiple* networks [19], [28]. However, the alignment graphs, which lie at the core of these methods, do not provide a scalable framework for aligning a large number of networks. As a result, the aforementioned methods can be used for aligning only a small number of networks and they can handle only very limited types of network isomorphism due to the high complexity. For better scalability, another algorithm named Græmlin [21], [31] takes a "progressive alignment" approach, which has been widely adopted by many multiple sequence alignment algorithms. Basically, Græmlin performs multiple network alignment by successively aligning the closest pair of networks, where the networks are aligned by greedily extending a small high-scoring subnetwork alignment used as a seed [21]. The combination of seed-and-extend scheme and progressive alignment allows relatively efficient local alignment up to ten networks. However, due to its greedy nature, the optimality of the alignment results is not guaranteed.

C. Global network alignment: Macroscopic comparison between biological networks

Global network alignment aims to find a coherent *global mapping* between nodes in different networks (Figure 1C), instead of finding multiple independent local mappings that may not be

necessarily coherent with each other. Until now, several global network alignment algorithms have been proposed based on various strategies, including integer programming [24], [32], [37], network diffusion [29], [33], and message-passing [30], [37]. Heuristic techniques, such as the greedy extension of high scoring subnetwork alignments and the divide-and-conquer strategy [33], [37], have been also used for global network alignment to reduce complexity.

Recently, hidden Markov models (HMMs) and Markov chains (MCs)—two probabilistic models widely used in biological sequence analysis—have been also adopted for comparative network analysis. A number of studies have shown the effectiveness of HMMs and MCs in network querying and alignment [29], [30], [33]–[35], [40], [41], estimation of functional similarity between biomolecules that belong to different networks [29], and identification of functional orthologs in different species [46]. These methods possess several important advantages over other existing methods, clearly showing that HMMs and MCs provide promising mathematical frameworks for comparative analysis of biological networks. In the following sections, we provide a tutorial review of the various Markov model based techniques for comparing biological networks. Our main goal is to expose this new set of problems to researchers in the signal processing community, who are familiar with the theory and application of Markov models, which may provide them an exciting new venue for future research.

III. Optimal Network Querying Using HMMs

Let us consider the following network querying problem: given a linear query path $\mathbf{p}$ and a biological network G, how can we find the path $\mathbf{q}$ that is closest to the given query among all paths embedded in the network? In order to solve this problem, we first need to define a scoring scheme $S(\mathbf{p}, \mathbf{q})$ for comparing different paths and evaluating their similarity. This path similarity score $S(\mathbf{p}, \mathbf{q})$ should sensibly integrate the similarity between nodes in $\mathbf{p}$ and $\mathbf{q}$ (e.g., in terms of sequence and/or functional similarity between proteins) as well as the similarity between their interaction patterns. Given a scoring scheme, we next need an efficient way for finding the path $\mathbf{q}$ that maximizes this score $S(\mathbf{p}, \mathbf{q})$, without enumerating all possible paths in the network. Figure 2A illustrates the network querying problem. The solid lines show interactions between nodes within the query or the target network, while the dashed lines indicate the similarity between nodes that belong to $\mathbf{p}$ and G, respectively. The best matching path $\mathbf{q}$ in G is shown in dark blue. Figure 2B shows the alignment between the matching paths $\mathbf{p}$ and $\mathbf{q}$, where the matching nodes are connected by dashed lines. Note that u_4 in the query path $\mathbf{p}$ is deleted in the matching path $\mathbf{q}$, while a new node v_5 is inserted to $\mathbf{q}$. Despite the outward simplicity of this querying problem, it becomes fairly nontrivial once we consider all possible node insertions and deletions in $\mathbf{q}$.

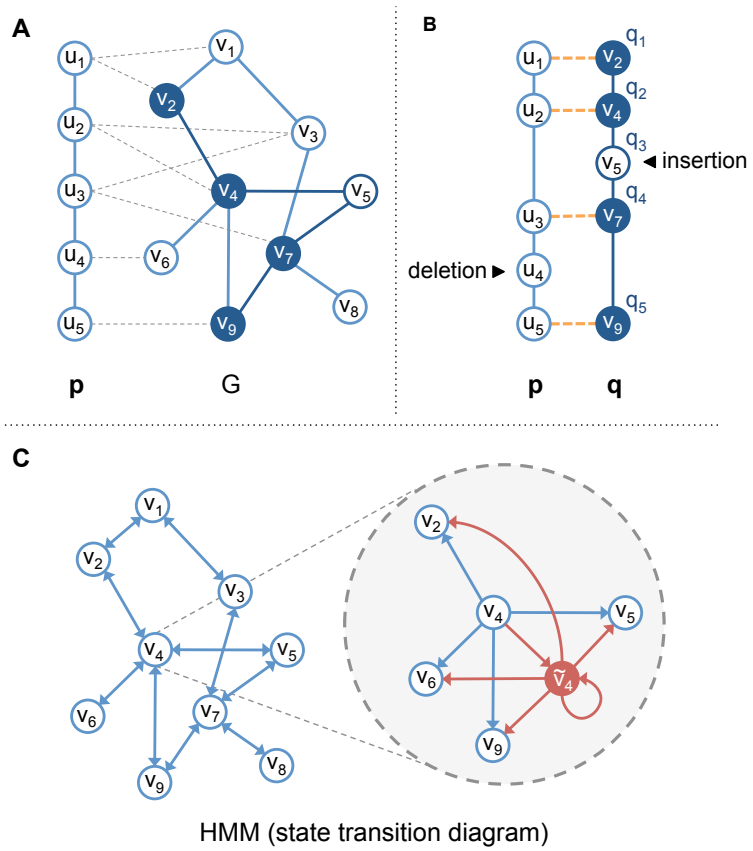

Fig. 2. Optimal network querying using HMMs. (A) Example of a query path $\mathbf{p}$ and a target network G. As shown by dashed lines, each node may have multiple similar nodes. (B) Alignment between the query $\mathbf{p}$ and the best matching path $\mathbf{q}$ in the target network. Matching nodes are connected by dashed lines. (C) State transition diagram of the HMM constructed from the target network G.

As shown in [34], hidden Markov models (HMMs) can provide an elegant mathematical framework for solving this optimal querying problem. The basic idea of the HMM-based querying approach is to construct a HMM based on the target network G and view this HMM as a *generative model* that gives rise to a series of biomolecules, forming a biological pathway with a linear structure. According to this model, we can view the query path $\mathbf{p}$ as an observation sequence generated by the given HMM, and the problem of finding the best matching path $\mathbf{q}$ in the network G is translated into that of finding the optimal state sequence in the HMM that maximizes the observation probability of the query path $\mathbf{p}$.

To elaborate on the HMM-based querying approach in more detail, let us consider a query path $\mathbf{p} = p_1 p_2 \cdots p_L$ that consists of L biomolecules (e.g., proteins). Let $\mathsf{G} = (\mathcal{V}, \mathcal{E})$ be a graph that represents the biological network at hand, with a set $\mathcal{V} = \{v_1, v_2, \cdots, v_N\}$ of N nodes and a set $\mathcal{E} = \{e_{ij}\}$ of M edges. Our goal is to find the path $\mathbf{q} = q_1 q_2 \cdots q_{L'}$ embedded in G (i.e., $q_k \in \mathcal{V}$) that is most similar to the query $\mathbf{p}$. For a pair of nodes (v_i, v_j), the presence

of an edge $e_{ij} \in \mathcal{E}$ indicates that there exists a biological interaction (e.g., protein binding or transcriptional regulation) between the corresponding biomolecules. Depending on the type of biological network being modeled, G can be either a directed or an undirected graph. For every (v_i, v_j) such that $e_{ij} \in \mathcal{E}$, we denote the interaction reliability as $w(v_i, v_j)$. In addition to this, we denote the similarity between two nodes u_i (in the query $\mathbf{p}$) and v_j (in the target network G) as $h(u_i, v_j)$. Sequence alignment scores are typically used for measuring the similarity between biomolecules, although functional similarity can be used for this purpose as well [29], [36].

In order to construct the HMM to be used in network querying, we first determine its state-transition diagram based on the structure of the network G. More precisely, every node $v_i \in \mathcal{V}$ in the network G will have a corresponding hidden state in the HMM. For notational simplicity, we represent this hidden state using the same notation v_i. Figure 2C illustrates the HMM that is constructed according to the network G shown in Fig. 2A. Each state in this HMM is allowed to emit any of the "symbols" $\{u_1, u_2, \cdots, u_L\}$ that constitute the query path $\mathbf{p}$. The next step is to determine the parameters of the constructed HMM based on the available information, namely the node similarity score $h(u_i, v_j)$ and the interaction reliability score $w(v_i, v_j)$. Basically, we have to define two mappings $\mathbf{f} : w(v_i, v_j) \mapsto t(v_j|v_i)$ and $\mathbf{g} : h(u_i, v_j) \mapsto e(u_i|v_j)$, where $t(v_j|v_i) = P(q_n = v_j|q_{n-1} = v_i)$ is the transition probability from state v_i to state v_j and $e(u_i|v_j) = P(u_i|q_n = v_j)$ is the emission probability of symbol u_i at state v_j. Although there can be numerous ways to define these mappings, the general idea is to define $\mathbf{f}$ and $\mathbf{g}$, such that higher $t(v_j|v_i)$ is assigned to v_i and v_j with a stronger interaction, and higher $e(u_i|v_j)$ is assigned to u_i and v_j that share a larger similarity [34].

Based on the above HMM-based framework, we can compute the joint probability $P(\mathbf{p}, \mathbf{q})$ for the query path $\mathbf{p} = u_1 \cdots u_L$ (viewed as an observation sequence) and a matching path $\mathbf{q} = q_1 \cdots q_L$ in G (viewed as the underlying hidden state sequence), which provides an effective way for evaluating the similarity between the two paths. We can now find the best matching path in G for the query $\mathbf{p}$ by identifying the state sequence that maximizes this probability:

$$\mathbf{q}^* = \arg\max_{\mathbf{q}} P(\mathbf{p}, \mathbf{q}). \tag{1}$$

The above problem can be efficiently solved in polynomial time using the Viterbi algorithm [47].

It should be noted that the above framework does not yet allow gaps in the matching paths and requires that $\mathbf{p}$ and $\mathbf{q}$ have the same length. This limitation can be easily overcome by extending the HMM as follows. To model insertions in a matching path $\mathbf{q}$, we allow the hidden states $v_1, \cdots, v_N$ in the HMM to emit a gap symbol ϕ in addition to the "symbols" $\{u_1, \cdots, u_L\}$ in the original query path $\mathbf{p}$. The gap emission probability $e(\phi|v_m)$ can be specified to control the penalty for inserting v_m in $\mathbf{q}$. Next, to deal with deletions of nodes in the original query $\mathbf{p}$, we add an *accompanying state* $\tilde{v}_j$ for every $v_j \in \mathcal{V}$. We also add an outgoing edge from v_j to $\tilde{v}_j$

and add outgoing edges from $\tilde{v}_j$ to all the neighboring states v_k such that $e_{jk} \in \mathcal{E}$. In addition to this, we allow self-transitions from $\tilde{v}_j$ to itself, in order to model consecutive deletions. This is illustrated in Fig. 2C for the accompanying state $\tilde{v}_4$. For simplicity, other accompanying states are not shown in the figure. Emission of u_i at one of the accompanying states $\tilde{v}_j$ implies that u_i in the query $\mathbf{p}$ does not have a corresponding node in the matching path $\mathbf{q}$.

Figure 2B shows an example of a query path $\mathbf{p}$ along with its best matching path $\mathbf{q}$, where the matching nodes are connected by dashed lines. In this example, v_5 does not have a counterpart in $\mathbf{p}$, corresponding to an insertion, which implies that a gap symbol ϕ is emitted at state v_5 in the HMM. We can also see that u_4 does not have a matching node in $\mathbf{q}$, which corresponds to a deletion and implies that u_4 is emitted at one of the accompanying states (in this example, either at $\tilde{v}_7$ or $\tilde{v}_9$).

As reported in [34], the HMM-based querying algorithm outperforms other querying algorithms, in terms of computational efficiency as well as the accuracy and biological significance of the querying results. Table I compares the computational complexity of the HMM-based querying algorithm along with three other querying methods, where L is the length of the query, M is the number of interactions in the target network, N is the number of nodes in the target network, and D is the maximum number of allowed insertions. As shown in the table, the HMM-based querying algorithm has a very low computational complexity, which is linear with respect to the length of the query (L) and the number of interactions in the network (M), while other algorithms suffer from either exponential or polynomial computational complexity. Thanks to the low complexity, the HMM-based algorithm can search for very long query paths in large networks while many existing methods are limited to short queries with three to ten nodes, as their complexity grows exponentially with the query size. As recently reviewed in [48], queries that take minutes to hours for other methods, including PathBLAST [16] and QPath [23], can be executed within a few seconds on a personal computer using the HMM-based algorithm.

In [34], the HMM-based querying algorithm was used to search the *Drosophila melanogaster* network for pathways that are similar to the *Homo sapiens* hedgehog signaling pathway (Figure 3A) and the mitogen-activated protein (MAP) kinase pathway (Figure 3B). In both cases, the top matching pathways agreed well with the query pathways, according to the functional

TABLE I
COMPUTATIONAL COMPLEXITY OF NETWORK QUERYING ALGORITHMS.

Algorithm	Computational complexity
PathBLAST [16]	$O(L!M)$
QPath [23]	$Q(2^L M)$
PathMatch [25]	$O(M + N \log N + L \log L)$
HMM [34]	$O(LDM)$

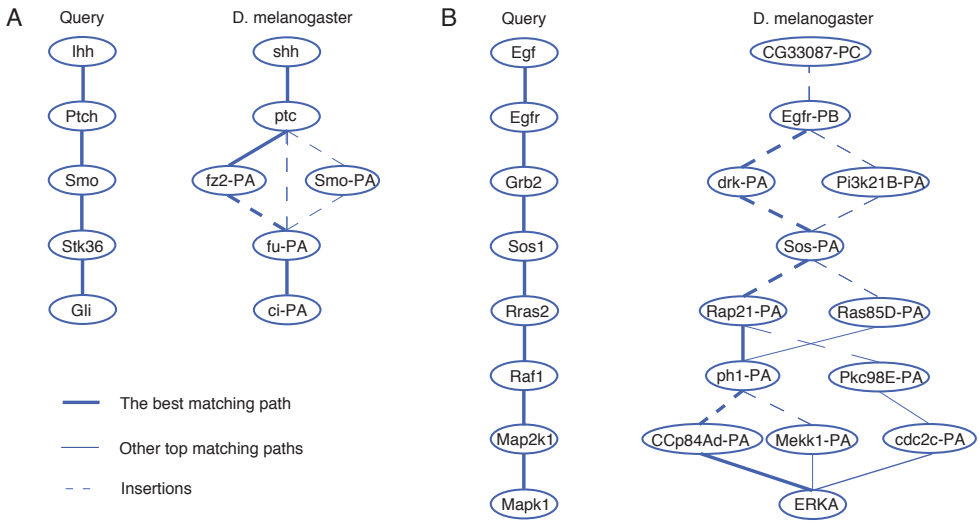

Fig. 3. Results for querying *H. sapiens* pathways in *D. melanogaster* PPI network (modified from Fig. 4 in [34]): (A) Human hedgehog pathway and the matching paths in the fly network. (B) Human MAP kinase pathway and the matching paths.

annotations of *D. melanogaster* [49], [50]. Furthermore, the predicted pathways in Figures 3A and 3B significantly overlapped with the putative homologous pathways in *D. melanogaster* reported in the KEGG database [51], [52]. For example, one of the top matching pathways in Figure 3A (shh–ptc–Smo–fu–ci) is the core of the *D. melanogaster* hedgehog signaling pathway given in KEGG (`http://www.genome.jp/kegg/pathway/dme/dme04340.html`), and Egfr–drk–Sos–Ras85D–ph1–Mekk1–ERKA (Figure 3B) is part of the putative MAP kinase pathway for *D. melanogaster* in KEGG (`http://www.genome.jp/kegg-bin/show_pathway?org_name=map&mapno=04010&mapscale=1.0&show_description=show`). The accuracy of the predictions compares favorably to the previously reported results [23], where the identified pathways had lower agreement with the putative pathways in KEGG [34], [48]. Furthermore, analysis of the querying results in the *S. cerevisiae*, *D. melanogaster*, *C. elegans*, and *E. coli* networks also showed that the HMM-based network querying algorithm can yield biologically meaningful predictions [34].

IV. Local Network Alignment Using HMMs

Let us consider a more general problem, where we want to compare two biological networks and identify the common pathways that are conserved in both networks. Suppose we have two biological networks $\mathsf{G}_1 = (\mathcal{U}, \mathcal{D})$ and $\mathsf{G}_2 = (\mathcal{V}, \mathcal{E})$. We assume that G_1 has a set $\mathcal{U} = \{u_1, u_2, \cdots, u_{N_1}\}$ of N_1 nodes and a set $\mathcal{D} = \{d_{ij}\}$ of M_1 edges, where d_{ij} indicates the presence of interaction between the two nodes u_i and u_j. Similarly, we assume that G_2 has a set $\mathcal{V} = \{v_1, v_2, \cdots, v_{N_2}\}$ of N_2 nodes and a set $\mathcal{E} = \{e_{ij}\}$ of M_2 edges. We denote the interaction reliability score between u_i and u_j as $w_1(u_i, u_j)$ and the score between v_i and v_j as $w_2(v_i, v_j)$. The node similarity between $u_i \in \mathcal{U}$ and $v_j \in \mathcal{V}$ is denoted as $h(u_i, v_j)$. Based on this setting,

our goal is to identify the most similar pair of linear paths $(\mathbf{p}, \mathbf{q})$, where the path $\mathbf{p} = p_1 \cdots p_L$ is embedded in the network G_1 and $\mathbf{q} = q_1 \cdots q_L$ in G_2. Now the question is how we can efficiently find such similar paths in different networks. Fortunately, the HMM-based framework that was discussed in the previous section can be extended in a straightforward manner to address this question [35].

In order to use HMMs for a local alignment of G_1 and G_2, we first construct two HMMs based on the given networks. Let us first focus on the HMM for G_1. As before, we design the initial state transition diagram of the HMM based on the network G_1. The resulting HMM contains a hidden state for each node $u_i \in \mathcal{U}$, which we also denote as u_i for convenience. State transition is allowed from u_i to u_j for (u_i, u_j) such that $d_{ij} \in \mathcal{D}$. The HMM for G_2 can be constructed in a similar way. Now that we have constructed the HMMs, how can we use these models to find the most similar paths in G_1 and G_2? In the network querying problem, the query path $\mathbf{p}$ was viewed as the observation sequence and the matching path $\mathbf{q}$ in the network was viewed as the state sequence of the constructed HMM that gives rise to this observation. In the local network alignment problem, we do not have a specific query path that can be viewed as the observation sequence, and both $\mathbf{p}$ and $\mathbf{q}$ will correspond to hidden state sequences in the respective HMMs. To compare $\mathbf{p}$ and $\mathbf{q}$, we can adopt the concept of a "virtual observation sequence" $\mathbf{s} = s_1 \cdots s_L$ that is *jointly* emitted by the two HMMs. Based on this model, the problem of finding the most similar pair of paths becomes that of finding the optimal pair of state sequences in the two HMMs that jointly maximize the probability $P(\mathbf{s}, \mathbf{p}, \mathbf{q})$ of the virtual sequence $\mathbf{s}$:

$$(\mathbf{p}^*, \mathbf{q}^*) = \underset{(\mathbf{p},\mathbf{q})}{\arg\max}\, P(\mathbf{s}, \mathbf{p}, \mathbf{q}). \tag{2}$$

As before, we can define two mappings $\mathbf{f}_1 : w_1(u_i, u_j) \mapsto t_1(u_j|u_i)$ and $\mathbf{f}_2 : w_2(v_i, v_j) \mapsto t_2(v_j|v_i)$ to transform the interaction reliability scores into state transition probabilities in the HMMs. In addition to this, we define another mapping $\mathbf{g} : h(u_i, v_j) \mapsto e(u_i, v_j)$ to obtain the joint emission probability $e(u_i, v_j)$ of a "virtual symbol" (in $\mathbf{s}$) at the state-pair (u_i, v_j) from the node similarity score $h(u_i, v_j)$. Further discussion on these transformations can be found in [35]. Given these HMMs, the optimal pair of hidden state sequences—and equivalently, the pair of the most similar paths—can be efficiently found through dynamic programming [35].

Finally, in order to handle insertions and deletions in matching paths, we again add accompanying states to both HMMs: $\tilde{u}_i$ for each $u_i \in \mathcal{U}$ (in the HMM for G_1) and $\tilde{v}_j$ for each $v_j \in \mathcal{V}$ (in the HMM for G_2). Note that "insertions" and "deletions" are relative terms. An insertion in $\mathbf{p}$ can be viewed as a deletion in $\mathbf{q}$, and similarly, a deletion in $\mathbf{p}$ can be viewed as an insertion in $\mathbf{q}$.

Figure 4 illustrates the overall idea of the HMM-based approach for comparing networks and identifying conserved paths. Two example networks G_1 and G_2 are shown in Fig. 4A, where

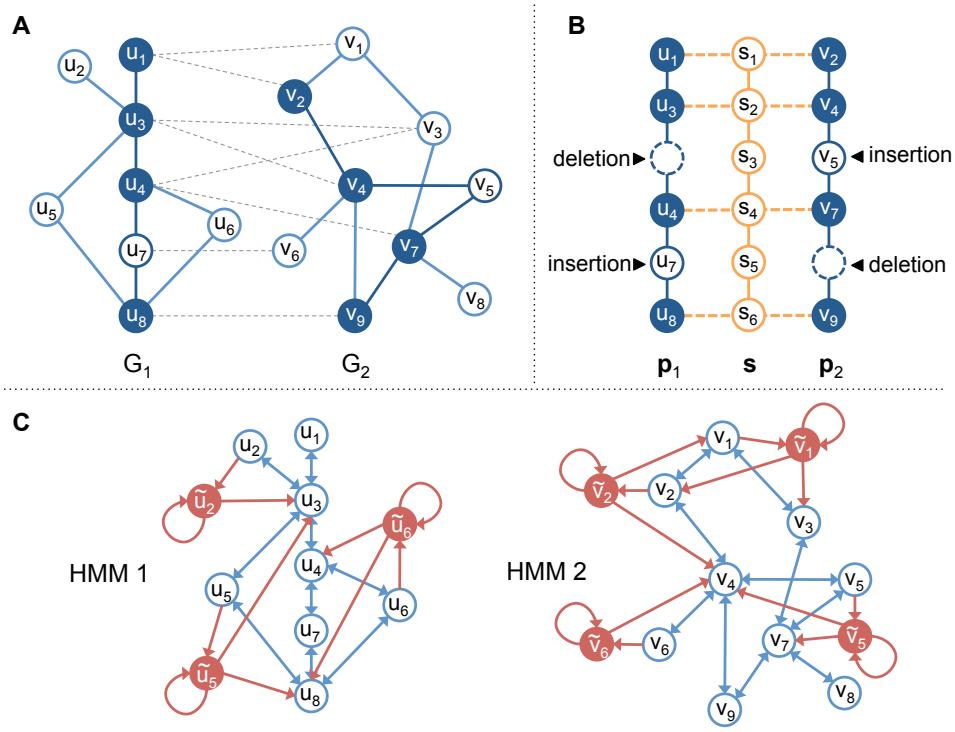

Fig. 4. Local network alignment using HMMs. (A) Two networks to be aligned. Conserved nodes and interactions are shown in dark blue, and the dashed lines indicate nodes with high similarity. (B) Optimal alignment of the conserved paths ($\mathbf{p}_1$ and $\mathbf{p}_2$) and the virtual sequence $\mathbf{s}$. (C) HMMs constructed from the networks G_1 and G_2.

similar nodes are connected by dashed lines. The nodes and interactions that are conserved in both networks are shown in dark blue. Figure 4B shows the alignment between the conserved paths $\mathbf{p}$ (in G_1) and $\mathbf{q}$ (in G_2). The virtual observation sequence $\mathbf{s}$ is also shown in Fig. 4B. The HMMs that are constructed from G_1 and G_2, respectively, are shown in Fig. 4C. For simplicity, only a few accompanying states are shown in the figure. According to these HMMs, the optimal pair $(\mathbf{p}, \mathbf{q})$ of hidden state sequences that results in the alignment in Fig. 4B would be:

$$\mathbf{p} = p_1 \cdots p_6 = u_1 u_3 \tilde{u}_3 u_4 u_7 u_8 \quad \text{and} \quad \mathbf{q} = q_1 \cdots q_6 = v_2 v_4 v_5 v_7 \tilde{v}_7 v_9. \tag{3}$$

The emission of s_3 at the state pair $(p_3, q_3) = (\tilde{u}_3, v_5)$ implies that the node v_5 is only included in the path $\mathbf{q}$ in network G_2 and the matching path $\mathbf{p}$ in G_1 does not contain a corresponding node. Similarly, the emission of s_5 at $(p_5, q_5) = (u_7, \tilde{v}_7)$ implies that u_7 is inserted in $\mathbf{p}$ but a corresponding node is not present in $\mathbf{q}$.

The HMM-based framework for network querying and network comparison has a number of important advantages [34], [35]. First of all, the HMM framework can deal with a large class of path isomorphism, hence it allows us to find matching paths with any number of gaps at arbitrary locations. Furthermore, the given framework makes it very flexible to choose the scoring scheme $S(\mathbf{p}, \mathbf{q})$ for comparing paths, where different penalties can be assigned

to mismatches, insertions, and deletions. Despite its generality, the HMM-based framework enables us to use an efficient dynamic programming algorithm for identifying the closest paths. In fact, the computational complexity of finding the best matching paths of length L is only $O(LM_1M_2)$, where M_1 is the number of edges in G_1 and M_2 is the number of edges in G_2. The memory complexity of this dynamic programming algorithm is $O(LN_1N_2)$, where N_1 and N_2 are the number of nodes in G_1 and G_2, respectively. This makes it possible to compare biological networks with thousands of nodes and tens of thousands of interactions and predict the conserved paths with tens of nodes within a few minutes on a personal computer. Furthermore, the dynamic programming algorithm can find the mathematically optimal path alignment that maximizes the alignment score $S(\mathbf{p}, \mathbf{q})$. Of course, the mathematical optimality does not guarantee the biological significance of the obtained results, but it can certainly lead to more accurate predictions when coupled with a realistic scoring scheme for evaluating the similarity between biological pathways.

Network querying can be viewed as a special case of local network alignment, and the HMM-based local alignment algorithm inherits all the advantages of the HMM-based querying algorithm, such as the computational efficiency, flexibility, and the mathematical optimality of the solution. An important goal of local network alignment is to identify putative functional pathways that are conserved across different species. One standard way for evaluating the biological significance of the predicted alignment results is to verify whether aligned nodes and subnetworks share similar gene ontology (GO) annotations [53] or KEGG orthology (KO) group annotations [51], [52]. In [35], the HMM-based alignment algorithm was used to align microbial networks—including *E. coli*, *C. crescentus*, and *S. typhimurium*—which showed that the algorithm is capable of making accurate alignments that are highly consistent according to the KO group annotations. Figure 5 illustrates the trend of the cumulative specificity, which measures the cumulative ratio of the number of pairs of aligned proteins with consistent KO annotations. The cyan curve in the figure shows the cumulative specificity for the top $k = 200$ local alignments (for conserved pathways with length $L = 30$) from the pairwise alignment of the *E. coli* and the *C. crescentus* networks. For increasing k and L, the local alignment algorithm finds a larger number of aligned proteins, while the cumulative specificity generally decreases. This is expected since alignments with lower alignment scores correspond to less conserved pathways with larger variations. However, the cumulative specificity (for the top 200 alignments) of the HMM-based local alignment algorithm is above 90%, which is higher than the reported specificity of a popular alignment algorithm called Græmlin 2.0 [31], which indicates that the HMM-based local alignment can yield accurate network alignment results that are biologically meaningful. Further analysis of the predicted local alignments in [35], [40] showed that the aligned proteins share similar functional characteristics, which implies that the HMM-based local alignment

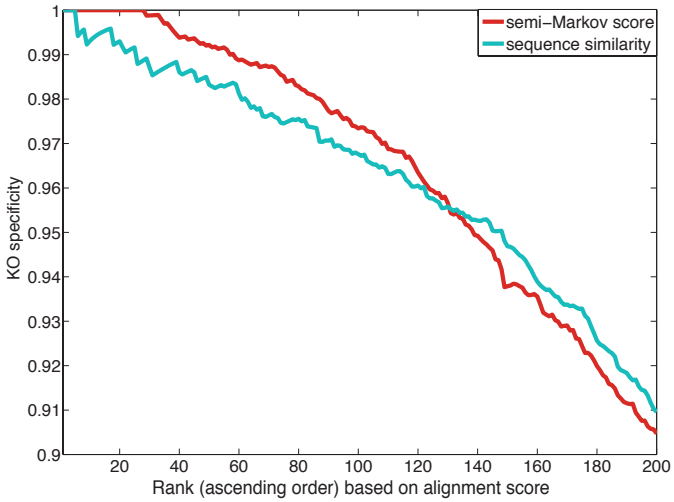

Fig. 5. Functional specificity for microbial network alignment (extracted from Fig. 4 in [40]): The cumulative specificity of the top 200 aligned pathways obtained from the pairwise alignment between the *E. coli* and *C. crescentus* networks.

method may be potentially used for automatic functional annotation of proteins, as well as other biomolecules.

V. Estimation of Functional Similarity Through Markov Random Walk

Suppose we want to compare two biological networks and predict the global correspondence between nodes in the respective networks. The basic goal is to map each node u_i in the network G_1 to one or more nodes v_j in the other network G_2 based on their overall *functional* similarity. When measuring the functional similarity between nodes, we should of course consider the similarity between the biomolecules themselves, in terms of their sequence and structure. However, considering that biomolecules carry out their functions through intertwined interactions with other molecules, it is important to consider these interaction patterns as well when evaluating the functional similarity between nodes. As originally proposed in [29] for the IsoRank algorithm, Markov random walk can provide an elegant framework for seamlessly integrating the node similarity and the interaction similarity and evaluating the global correspondence between nodes that belong to different networks.

Let us first focus on the problem of evaluating the interaction similarity between two nodes u_i in G_1 and v_j in G_2. Basically, our goal is to estimate the topological similarity between the subnetwork in G_1 around the node u_i and the subnetwork in G_2 around v_j. One possible way to estimate such topological similarity is to compare the neighboring nodes around u_i with those around v_j. For example, if many nodes around u_i share high similarity with many nodes around v_j, this would imply that the subnetworks around u_i and v_j are topologically similar. Of course, the similarity between the neighboring nodes would have to be estimated in a similar

way, which implies that the interaction pattern of every network node will affect the interaction similarity of all the other nodes in the network, unless they are disconnected. Therefore, we can view the resulting interaction similarity score as a *global* correspondence score that measures the overall similarity between nodes, in view of the entire networks to which they belong.

For a mathematical formulation of the above problem, let us denote the global similarity between u_i and v_j as $s(u_i, v_j)$. Let $\mathcal{U}(i) = \{u_k \in \mathcal{U} | d_{ik} \in \mathcal{D}\}$ be the set of neighboring nodes of u_i and $\mathcal{V}(j) = \{v_\ell \in \mathcal{V} | e_{j\ell} \in \mathcal{E}\}$ be the set of neighboring nodes of v_j. The global similarity score between u_i and v_j can then be estimated as:

$$s(u_i, v_j) = \sum_{u_k \in \mathcal{U}(i)} \sum_{v_\ell \in \mathcal{V}(j)} \bar{w}_1(i,k) \bar{w}_2(j,\ell) s(u_k, v_\ell), \tag{4}$$

where $\bar{w}_1(i,k)$ measures the normalized contribution from the neighboring node u_k to the node u_i based on the interaction reliability (or strength) $w_1(u_i, u_k)$ between u_i and u_k

$$\bar{w}_1(i,k) = \frac{w_1(u_i, u_k)}{\sum_{u_{k'} \in \mathcal{U}(i)} w_1(u_i, u_{k'})}, \tag{5}$$

and similarly, $\bar{w}_2(j,\ell)$ measures the normalized contribution from v_ℓ to v_j

$$\bar{w}_2(j,\ell) = \frac{w_2(v_j, v_\ell)}{\sum_{v_{\ell'} \in \mathcal{V}(j)} w_2(v_j, v_{\ell'})}. \tag{6}$$

We can conveniently rewrite (4) for all (u_i, v_j) in a matrix equation

$$\mathbf{s} = \mathbf{W}\mathbf{s}, \tag{7}$$

where $\mathbf{s}$ is a $N_1 N_2$-dimensional column vector such that $\mathbf{s}[(i-1)N_2 + j, 1] = s(u_i, v_j)$, and $\mathbf{W}$ is a $N_1 N_2 \times N_1 N_2$ matrix such that $\mathbf{W}[(i-1)N_2 + j, (k-1)N_2 + \ell] = \bar{w}_1(i,k)\bar{w}_2(j,\ell)$, for $1 \le i, k \le N_1$ and $1 \le j, \ell \le N_2$. From (7), we can compute the global similarity score by finding the eigenvector $\mathbf{s}$ of the matrix $\mathbf{W}$ with unit eigenvalue. To obtain a unique $\mathbf{s}$, we further normalize the vector such that $\mathbf{1}^T\mathbf{s} = 1$, where $\mathbf{1}$ is an all-one column vector.

Note that the global similarity score obtained from (7) estimates only the interaction similarity between nodes. However, as shown in [29], we can easily extend this scheme to incorporate node similarity by modifying the previous equation as follows

$$\mathbf{s} = \lambda \mathbf{W}\mathbf{s} + (1-\lambda)\mathbf{h} = \left(\lambda \mathbf{W} + (1-\lambda)\mathbf{h}\mathbf{1}^T\right)\mathbf{s}, \tag{8}$$

where $\mathbf{h}$ is a $N_1 N_2$-dimensional column vector that contains the node similarity score $\mathbf{h}[(i-1)N_2 + j] = h(u_i, v_j)$, and $\lambda \in [0,1]$ is a parameter that controls the balance between the interaction similarity and the node similarity in evaluating the global functional similarity between nodes. We assume that $\mathbf{h}$ is normalized such that $\mathbf{1}^T\mathbf{h} = 1$. As before, we can compute the global similarity score between nodes simply by finding the normalized eigenvector $\mathbf{s}$ that satisfies (8) and $\mathbf{1}^T\mathbf{s} = 1$.

Conceptually, we can interpret the process of computing the global similarity score from the viewpoint of Markov random walk. Suppose we want to perform a *simultaneous* random walk on the two networks G_1 and G_2, as shown in Fig. 6A. At each time step, the walker randomly moves to one of the neighboring nodes in each network, where a neighbor with a higher interaction reliability score is more likely to be chosen. For example, let us assume that the random walker is currently located at the node u_i in G_1 and at v_j in G_2. At the next time step, the walker randomly moves from u_i to one of its neighboring nodes $u_k \in \mathcal{U}(i)$ with probability $\bar{w}_1(i,k)$ in the network G_1. Similarly, in the other network G_2, the walker moves from v_j to one of the neighbors $v_\ell \in \mathcal{V}(j)$ with probability $\bar{w}_2(j,\ell)$. This is equivalent to a random walk on a *product graph* $\mathsf{G}_\times$ of G_1 and G_2, where every node in $\mathsf{G}_\times$ corresponds to a node pair (u_i, v_j) and there exists an edge between two node pairs (u_i, v_j) and (u_k, v_ℓ) if and only if there exists an edge between u_i and u_k in G_1 (i.e., $d_{ik} \in \mathcal{D}$) and an edge between v_j and v_ℓ in G_2 (i.e., $e_{j\ell} \in \mathcal{E}$). This is illustrated in Fig. 6B.

We can measure the interaction similarity between u_i and v_j based on the long-run proportion of time that the random walker simultaneously stays at u_i and v_j. This measure is quite intuitive if we consider the following. For example, suppose u_i and v_j are surrounded by similar neighbors, which are likely to be simultaneously visited during the random walk. This will increase the probability that u_i and v_j will be simultaneously visited by the random walker, thereby increasing the long-run proportion of time spent at (u_i, v_j). This random walk can be modeled as a Markov chain with the same transition probability matrix $\mathbf{W}$ given in (7), and as before, the long-run proportion of time (or the *stationary probability* of the Markov chain) can be computed by finding the eigenvector of $\mathbf{W}$ with unit eigenvalue.

Similarly, we can also interpret the global similarity score obtained from (8) as the stationary probability of a *random walk with restart*. In this case, the random walker randomly decides

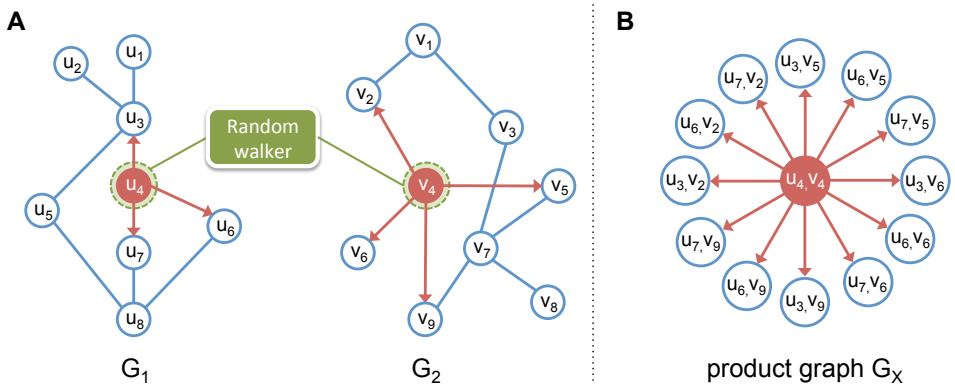

Fig. 6. Random walk for estimating the similarity between nodes. (A) Simultaneous random walk on two graphs. (B) The simultaneous random walk on G_1 and G_2 is equivalent to a random walk on their product graph $\mathsf{G}_\times$.

at each time step whether to *restart* the walk from a new position (with probability $1-\lambda$) or to continue the walk from the current position (with probability λ). In case the random walker chooses to restart, the new position (u_k, v_ℓ) is selected according to $\mathbf{h}$ with probability $\mathbf{h}[(k-1)N_2+\ell] = h(u_k, v_\ell)$. This implies that the random walk is more likely to restart at node pairs with higher node similarity. In case the random walker decides to continue the walk from the current position (u_i, v_j), the next position (u_k, v_ℓ) is randomly selected according to the weight matrix $\mathbf{W}$, as in a regular random walk. As we can see, the *random walk* is governed by the network topology and the reliability (or strength) of the interactions between nodes, while the *restart* operation is governed by the similarity between nodes in the two networks. As a consequence, the long-run proportion of time spent at a node pair (u_i, v_j) provides an effective measure of the global similarity between the two nodes u_i and v_j that combines node similarity and interaction similarity.

Instead of taking the "random walk with restart" approach discussed above, we can also adopt the concept of *semi-Markov random walk* to obtain an effective global node similarity score [41], [54]. In an ordinary Markov random walk, the random walker always spends a fixed amount of time at a given position before making the next move. On the contrary, in a semi-Markov random walk, the random walker may spend a different (and possibly random) amount of time at each position, before moving to the next position. Suppose we model the semi-Markov random walk such that the (expected) amount of time that the random walker spends at a node pair (u_i, v_j) is proportional to the node similarity $h(u_i, v_j)$. According to this model, both higher interaction similarity and higher node similarity will increase the long-run proportion of time that the random walker spends at a node pair, making it a good measure for estimating the global similarity between nodes. As shown in [54], based on this semi-Markov random walk model, the global similarity score can be computed as:

$$\mathbf{s} = \frac{\mathbf{r} \circ \mathbf{h}}{\mathbf{h}^T \mathbf{r}}, \tag{9}$$

where $\mathbf{r}$ is the stationary probability of the ordinary Markov random walk on the product graph that satisfies $\mathbf{r} = \mathbf{W}\mathbf{r}$ and $\circ$ denotes the Hadamard (or element-wise) product.

Markov random walk scores have been used in IsoRank [29] and in IsoRankN [33] for global multiple network alignment. IsoRank is widely known as the first effective global network alignment algorithm that aims to construct a single consistent alignment of multiple networks, instead of identifying a set of unrelated local subnetwork alignments. Given K networks $\{\mathsf{G}_1, \cdots, \mathsf{G}_K\}$, IsoRank first computes the global similarity scores between nodes of every pair of networks based on (8), and it uses these scores to construct a global alignment of the given networks through greedy K-partite matching [29]. IsoRank has been used to construct a global alignment of the protein-protein interaction (PPI) networks of five different species—

Saccharomyces cerevisiae, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Mus musculus*, and *Homo sapiens*—and the algorithm has been shown to be very effective in detecting conserved modules across networks and predicting functional orthologs [29]. IsoRankN [33], an improved version of the original IsoRank, also utilizes the global similarity scores obtained from (8) but takes a different spectral clustering approach to build the global network alignment. It was shown that IsoRankN yields improved alignment results, in terms of coverage and consistency, which may also lead to more accurate ortholog prediction [33].

In [40], it was shown that the semi-Markov random walk model can be used in combination with the HMM-based local network alignment method to improve the accuracy of the alignment results, especially when the individual node similarity scores (e.g., based on sequence similarity) are noisy or unreliable. In this work, the semi-Markov random walk framework was used to compute the global similarity score between nodes that belong to different networks. These scores were subsequently used by the HMM-based local network alignment algorithm to find the best matching linear pathways conserved in the given networks. Figure 5 shows the performance improvement, measured in terms of the cumulative specificity (using the KO annotations) of the top $k = 200$ conserved pathways (with length $L = 30$) by aligning the networks of *E. coli* and the *C. crescentus*. The red curve in Figure 5 shows the KO specificity for using the global similarity score based on the semi-Markov random walk and the cyan curve shows the specificity for directly using the sequence similarity score. As we can see from these results, the use of semi-Markov scores can significantly improve the specificity of the top predicted pathways, which indicates that semi-Markov random walk scores can be helpful in making predictions that are

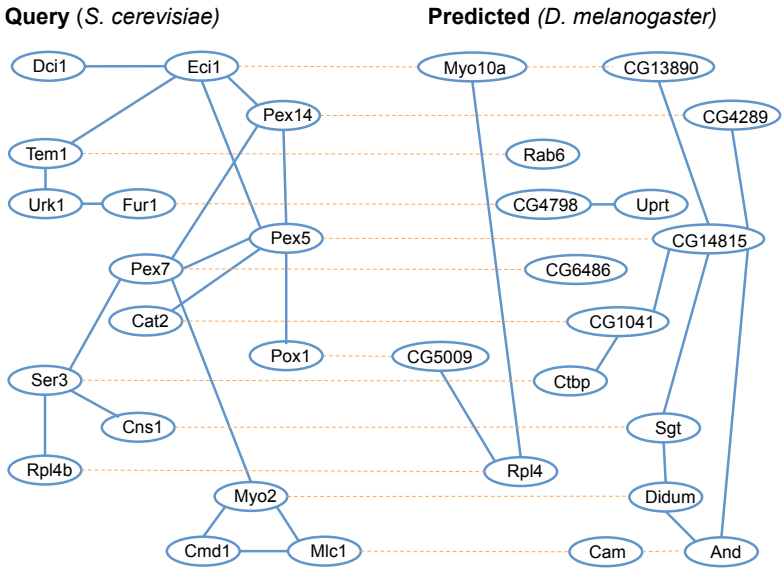

Fig. 7. Result for querying the peroxisomal pathway of *S. cerevisiae* in the *D. melanogaster* PPI network. The dashed lines indicate the matching proteins.

functionally more coherent. Recently, the semi-Markov random walk approach has also been applied to fast network querying [41], where the computed global similarity scores were used to identify the best matching region in the target network through an efficient network reduction technique. As an example, Figure 7 shows the result for querying the peroxisomal pathway of *S. cerevisiae* (obtained from [19]) in the PPI network of *D. melanogaster*. The querying result is in good agreement with the result previously reported in [19], and the identified subnetwork shows high functional coherence (p-value of 3.5e-4), measured based on GO annotations. The concept of semi-Markov random walk was also used to introduce an effective similarity measure for comparing two HMMs at a low computational cost [54], where the measure can be virtually used to compare any types of graphs and network models.

VI. Discussion and Conclusion

In this paper, we have reviewed the application of Markov models to various comparative network analysis problems. As we have seen, comparative network analysis provides a powerful means of gaining novel system-level understanding of diverse biological mechanisms within cells and the variegated functional roles of various cellular constituents. Comparative network analysis can help us take advantage of the available biological data and knowledge encoded in biological networks—which include the fast growing list of functional biomolecular entities within cells; their composition, structure, and annotated functions; and the interactions among these entities—in an integrative manner. As noted in [55], [56], this may expedite the genome-scale functional annotation of biomolecules at a relatively low cost. Furthermore, computational analysis of biomolecular interaction networks can help us better understand the functional organization of biological networks and elucidate the similarities and differences among networks that belong to different species. This may provide fundamental insights into biological systems that may ultimately lead to important advances in medical applications.

For example, comparative network analysis may be used to identify genes and pathways that are associated with a complex disease, such as cancer [57], [58]. Identification of disease-related subnetworks (or pathways) can help us better understand the detailed mechanisms of a disease and its development, thereby leading to the development of enhanced diagnostic techniques and novel drugs. A number of recent studies have shown that integrative analysis of gene expression data based on known pathways or disease-related subnetworks can significantly improve the accuracy and robustness of cancer diagnosis and prognosis [59]–[62]. As shown by PARADIGM [63], pathway-focused analyses that integrate multiple genomics data also can be very useful in identifying patient-specific molecular activities. These results imply that identifying novel disease-associated pathways through network analysis may contribute to improving such pathway or network based disease classification techniques. Comparative network analysis

can also provide a powerful tool for studying viral infection mechanisms [57], [64]. Recently, the HMM-based local network alignment method has been used to identify conserved pathways across species that are susceptible to lentiviruses, and to study the susceptibility of these conserved pathways to HIV-1 (human immunodeficiency virus type one) [65]. Such analyses can be useful for elucidating the infection mechanisms of HIV-1. Furthermore, similar approaches can be used to identify alternative pathways that can circumvent HIV infection [66], which may lead to the design of novel system-based therapeutics for acquired immunodeficiency syndrome (AIDS).

In this tutorial review, we focused on comparative network analysis methods that are based on Markov models. As shown throughout the paper, Markov model based methods provide a number of important advantages over existing methods. For example, the HMM-based querying and local alignment methods have significantly lower computational complexity compared to many existing methods, which allows us to efficiently compare large networks (with tens of thousands of nodes and hundreds of thousands of edges) and identify conserved pathways (with easily up to hundreds of nodes). Furthermore, the HMM-based framework does not impose any restrictions on the number of node deletions/insertions as well as their locations, hence it can handle a large class of path isomorphism. This allows us identify distantly related homologous pathways that may possess considerable differences. Another important advantage of the HMM-based framework is that it allows us to find the mathematically optimal alignment. Considering that many existing algorithms rely on various heuristics that cannot guarantee the optimality of the obtained solutions, the HMM-based method may lead to more accurate and biologically significant predictions, when combined with a realistic scoring scheme for assessing pathway homology. As also discussed in this paper, the Markov random walk and the semi-Markov random walk models can provide effective ways for measuring the functional similarity between biomolecules that belong to different networks, by sensibly integrating their molecular similarity (based on sequence and/or structure) as well as their interaction patterns. The resulting functional similarity scores can provide the basis for building accurate network querying and alignment tools that can make predictions that are functionally more coherent.

Unlike comparative sequence analysis, the field of comparative network analysis is still at an early stage. Current network analysis tools have a large room for further improvement and significant research efforts are needed to make these tools ready for everyday use in systems biology research. However, the preliminary results obtained through comparative analysis of biological networks are certainly very promising. We expect that probabilistic models, such as the Markov models discussed in this paper, will play essential roles in advancing the field of computational comparative network analysis and building practical tools that can effectively assist biomedical research with the identification and study of functional pathways.

REFERENCES

[1] J. H. Do and D. K. Choi, "Computational approaches to gene prediction," *J. Microbiol.*, vol. 44, pp. 137–144, Apr 2006.

[2] B.-J. Yoon and P. P. Vaidyanathan, "Computational identification and analysis of noncoding RNAs-Unearthing the buried treasures in the genome," *IEEE Signal Processing Magazine*, vol. 24, no. 1, pp. 64–74, 2007.

[3] H. Kitano, "Systems biology: a brief overview," *Science*, vol. 295, pp. 1662–1664, Mar 2002.

[4] U. Sauer, M. Heinemann, and N. Zamboni, "Getting closer to the whole picture," *Science*, vol. 316, pp. 550–551, Apr 2007.

[5] A. Osman, "Yeast two-hybrid assay for studying protein-protein interactions," *Methods Mol. Biol.*, vol. 270, pp. 403–422, 2004.

[6] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, pp. 198–207, Mar 2003.

[7] M. E. Cusick, N. Klitgord, M. Vidal, and D. E. Hill, "Interactome: gateway into systems biology," *Hum. Mol. Genet.*, vol. 14 Spec No. 2, pp. R171–181, Oct 2005.

[8] A. Skusa, A. Ruegg, and J. Kohler, "Extraction of biological interaction networks from scientific literature," *Brief. Bioinformatics*, vol. 6, pp. 263–276, Sep 2005.

[9] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Res.*, vol. 34, pp. D535–539, Jan 2006.

[10] I. Xenarios, L. Salwinski, X. Duan, P. Higney, S. Kim, and D. Eisenberg, "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions," *Nucleic acids research*, vol. 30, no. 1, p. 303, 2002.

[11] P. E. Hodges, A. H. McKee, B. P. Davis, W. E. Payne, and J. I. Garrels, "The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data," *Nucleic Acids Res.*, vol. 27, pp. 69–73, Jan 1999.

[12] "Ingenuity," http://www.ingenuity.com/.

[13] "Pathway Studio," http://www.ariadnegenomics.com/products/pathway-studio/.

[14] R. Sharan and T. Ideker, "Modeling cellular machinery through biological network comparison," *Nat. Biotechnol.*, vol. 24, pp. 427–433, Apr 2006.

[15] T. Akutsu, S. Kuhara, O. Maruyama, and S. Miyano, "Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions," in *Proc. 9th Annu. ACM-SIAM Symp. Discrete Alg.*, 1998, pp. 695–706.

[16] B. P. Kelley, R. Sharan, R. M. Karp, T. Sittler, D. E. Root, B. R. Stockwell, and T. Ideker, "Conserved pathways within bacteria and yeast as revealed by global protein network alignment," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 100, pp. 11 394–11 399, Sep 2003.

[17] M. Koyutürk, A. Grama, and W. Szpankowski, "An efficient algorithm for detecting frequent subgraphs in biological networks," *Bioinformatics*, vol. 20, pp. SI200–207, 2004.

[18] R. Pinter, O. Rokhlenko, E. Yeger-Lotem, and M. Ziv-Ukelson, "Alignment of metabolic pathways," *Bioinformatics*, vol. 21, no. 16, pp. 3401–3408, 2005.

[19] R. Sharan, S. Suthram, R. M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. M. Karp, and T. Ideker, "Conserved patterns of protein interaction in multiple species," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, pp. 1974–1979, Feb 2005.

[20] F. Sohler and R. Zimmer, "Identifying active transcription factors and kinases from expression data using pathway queries," *Bioinformatics*, pp. ii115–ii122, 2005.

[21] J. Flannick, A. Novak, B. Srinivasan, H. McAdams, and S. Batzoglou, "Græmlin: general and robust alignment of multiple large interaction networks," *Genome Res*, vol. 16, no. 9, pp. 1169–1181, 2006.

[22] J. Scott, T. Ideker, R. Karp, and R. Sharan, "Efficient algorithms for detecting signaling pathways in protein interaction networks," *J Comput Biol*, vol. 13, pp. 133–144, 2006.

[23] T. Shlomi, D. Segal, E. Ruppin, and R. Sharan, "QPath: a method for querying pathways in a protein-protein interaction network," *BMC Bioinformatics*, vol. 7, no. 199, 2006.

[24] Z. Li, S. Zhang, Y. Wang, X. Zhang, and L. Chen, "Alignment of molecular networks by integer quadratic programming," *Bioinformatics*, vol. 23, no. 13, pp. 1631–1639, 2007.

[25] Q. Yang and S. Sze, "Path matching and graph matching in biological networks," *J Comput Biol*, vol. 14, pp. 56–67, 2007.

[26] B. Dost, T. Shlomi, N. Gupta, E. Ruppin, V. Bafna, and R. Sharan, "QNet: A tool for querying protein interaction networks," *J Comput Biol*, vol. 15, no. 7, pp. 913–925, 2008.

[27] A. Ferro, R. Giugno, M. Mongiovi, A. Pulvirenti, D. Skripin, and D. Shasha, "GraphFind: Enhancing graph searching by low support data mining techniques," *BMC Bioinformatics*, vol. 9(S4), p. S10, 2008.

[28] M. Kalaev, M. Smoot, T. Ideker, and R. Sharan, "NetworkBLAST: comparative analysis of protein networks," *Bioinformatics*, vol. 24, pp. 594–596, Feb 2008.

[29] R. Singh, J. Xu, and B. Berger, "Global alignment of multiple protein interaction networks with application to functional orthology detection," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 105, pp. 12 763–12 768, Sep 2008.

[30] M. Bayati, M. Gerritsen, D. Gleich, A. Saberi, and Y. Wang, "Algorithms for large, sparse network alignment problems," in *IEEE International Conference on Data Mining (ICDM)*, 2009, pp. 705–710.

[31] J. Flannick, A. Novak, C. B. Do, B. S. Srinivasan, and S. Batzoglou, "Automatic parameter learning for multiple local network alignment," *J. Comput. Biol.*, vol. 16, pp. 1001–1022, Aug 2009.
[32] G. Klau, "A new graph-based method for pairwise global network alignment," *BMC Bioinformatics*, vol. 10, no. Suppl 1, p. S59, 2009.
[33] 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, pp. i253–258, Jun 2009.
[34] X. Qian, S. H. Sze, and B.-J. Yoon, "Querying pathways in protein interaction networks based on hidden Markov models," *Journal of Computational Biology*, vol. 16, pp. 145–157, Feb 2009.
[35] X. Qian and B.-J. Yoon, "Effective identification of conserved pathways in biological networks using hidden Markov models," *PLoS ONE*, vol. 4, p. e8070, 2009.
[36] W. Tian and N. Samatova, "Pairwise alignment of interaction networks by fast identification of maximal conserved patterns," in *Pac Symp Biocomput*, vol. 14, 2009, pp. 99–110.
[37] M. Zaslavskiy, F. Bach, and J. Vert, "Global alignment of protein-protein interaction networks by graph matching methods," *Bioinformatics*, vol. 25, pp. 259–267, 2009.
[38] F. Ay, M. Kellis, and T. Kahveci, "SubMAP: Aligning metabolic pathways with subnetwork mappings," *Journal of Computational Biology*, vol. in press, 2011.
[39] S. Bruckner, F. Huffner, R. M. Karp, R. Shamir, and R. Sharan, "Topology-free querying of protein interaction networks," *J. Comput. Biol.*, vol. 17, pp. 237–252, Mar 2010.
[40] X. Qian, S. Sahraeian, and B.-J. Yoon, "Enhancing the accuracy of HMM-based conserved pathway prediction using global correspondence scores," *BMC Bioinformatics*, vol. 12 (Suppl 8): S6, 2011.
[41] S. M. E. Sahraeian and B.-J. Yoon, "Fast network querying algorithm for searching large-scale biological networks," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.
[42] N. Alon, R. Yuster, and U. Zwick, "Color-coding," *J ACM*, pp. 844–856, 1995.
[43] M. Narayanan and R. Karp, "Comparing protein interaction networks via a graph match-and-split algorithm," *Journal of Computational Biology*, vol. 14, no. 7, pp. 892–907, 2007.
[44] P. Jancura, J. Heringa, and E. Marchiori, "Divide, align and full-search for discovering conserved protein complexes," in *Proc. of the LNCS EvoBIO*, 2008, pp. 71–82.
[45] F. Towfic, M. H. W. Greenlee, and V. Honavar, "Aligning biomolecular networks using modular graph kernels," in *Proceedings of the 9th international conference on Algorithms in bioinformatics*, ser. WABI'09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 345–361.
[46] S. Bandyopadhyay, R. Sharan, and T. Ideker, "Systematic identification of functional orthologs based on protein network comparison," *Genome Res.*, vol. 16, pp. 428–435, Mar 2006.
[47] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
[48] V. Fionda and L. Palopoli, "Biological network querying techniques: Analysis and comparison," *J Comput Bio*, vol. 18, no. 4, pp. 595–625, 2011.
[49] The Flybase Consortium, "FlyBase: The Drosophila database." *Nucleic Acids Res*, vol. 24, pp. 53–56, Jan 1996.
[50] R. Drysdale, "FlyBase : A database for the Drosophila research community," *Methods Mol Biol*, vol. 420, pp. 45–59, 2008.
[51] M. Kanehisa and S. Goto, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Res*, vol. 28, pp. 27–30, Jan 2000.
[52] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi, "KEGG for linking genomes to life and the environment," *Nucleic Acids Res*, vol. 36, pp. D480–484, Jan 2008.
[53] M. Ashburner *et al.*, "Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet*, vol. 25, no. 1, pp. 25–29, 2000.
[54] S. Sahraeian and B.-J. Yoon, "A novel low-complexity HMM similarity measure," *IEEE Signal Processing Letters*, vol. 18, no. 2, pp. 87–90, 2011.
[55] E. Kolker, K. Makarova, S. Shabalina, A. Picone, S. Purvine, T. Holzman, T. Cherny, D. Armbruster, R. Munson, G. Kolesov, D. Frishman, and M. Galperin, "Identification and functional analysis of hypothetical genes expressed in Haemophilus influenzae," *Nucleic Acids Res*, vol. 32, no. 8, pp. 2353–2361, 2004.
[56] M. Campillos, C. von Mering, L. Jensen, and P. Bork, "Identification and analysis of evolutionarily cohesive functional modules in protein networks," *Genome Res*, vol. 16, pp. 374–382, 2006.
[57] T. Ideker and R. Sharan, "Protein networks in disease," *Genome Res.*, vol. 18, pp. 644–652, Apr 2008.
[58] S. Karni, H. Soreq, and R. Sharan, "A network-based method for predicting disease-causing genes," *Journal of Computational Biology*, vol. 16, no. 2, pp. 181–189, 2009.
[59] H. Chuang, E. Lee, Y. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Mol. Syst. Biol.*, vol. 3, p. 140, 2007.
[60] E. Lee, H. Chuang, J. Kim, T. Ideker, and D. Lee, "Inferring pathway activity toward precise disease classification," *PLoS Comput. Biol.*, vol. 4, p. e1000217, 2008.

[61] J. Su, B. J. Yoon, and E. R. Dougherty, "Accurate and reliable cancer classification based on probabilistic inference of pathway activity," *PLoS ONE*, vol. 4, p. e8161, 2009.
[62] J. Su, B.-J. Yoon, and E. Dougherty, "Identification of diagnostic subnetwork markers for cancer in human protein-protein interaction network," *BMC Bioinformatics*, vol. 11, no. Suppl 6, p. S8, 2010.
[63] C. Vaske, S. Benz, J. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, and J. Stuart, "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM," *Bioinformatics*, vol. 26, no. 12, pp. 237–245, 2010.
[64] R. Konig, Y. Zhou, D. Elleder, T. Diamond, G. Bonamy, J. Irelan, C. Chiang, B. Tu, P. D. Jesus, C. Lilley, and et al., "Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication," *Cell*, vol. 135, no. 1, pp. 49–60, 2008.
[65] X. Qian and B.-J. Yoon, "Comparative analysis of protein interaction networks reveals that conserved pathways are susceptible to HIV-1 interception," *BMC Bioinformatics*, vol. Suppl 1, no. S19, 2011.
[66] S. Balakrishnan, O. Tastan, J. Carbonell, and J. Klein-Seetharaman, "Alternative paths in HIV-1 targeted human signal transduction pathways," *BMC Genomics*, vol. 10, no. Suppl 3, p. S30, 2009.