Querying Pathways in Protein Interaction Networks Based on Hidden Markov Models

Xiaoning Qian[1,2], Sing-Hoi Sze[3,4], and Byung-Jun Yoon[1,*]

[1] Department of Electrical & Computer Engineering, [2] Department of Statistics,

[3] Department of Computer Science, [4] Department of Biochemistry & Biophysics

Texas A&M University, College Station, TX 77843, USA

Abstract

High-throughput techniques for measuring protein interactions have enabled the systematic study of complex protein networks. Comparing the networks of different organisms and identifying their common substructures can lead to a better understanding of the regulatory mechanisms underlying various cellular functions. To facilitate such comparisons, we present an efficient framework based on hidden Markov models (HMMs) that can be used for finding homologous pathways in a network of interest. Given a query path, our method identifies the top k matching paths in the network, which may contain any number of consecutive insertions and deletions. We demonstrate that our method is able to identify biologically significant pathways in protein interaction networks obtained from the DIP database, and the retrieved paths are closer to the curated pathways in the KEGG database when compared to the results from previous approaches. Unlike most existing algorithms that suffer from exponential time complexity, our algorithm has a polynomial complexity that grows linearly with the query size. This enables the search for very long paths with more than 10 proteins within a few minutes on a desktop computer. A software program implementing the algorithm is available upon request from the authors.

Keywords: pathway alignment, protein interaction network, hidden Markov model (HMM).

[*] Corresponding author: Dept. of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843-3128, USA. Email: bjyoon@ece.tamu.edu.

1 Introduction

Recent advances in high-throughput experimental techniques for measuring protein interactions, such as the two-hybrid systems (Ito *et al.*, 2001) and co-immunoprecipitation assays (Mann *et al.*, 2001; Uetz *et al.*, 2004), have enabled the systematic study of biological interactions on a global scale for an increasing number of organisms (von Mering *et al.*, 2002). These complex interaction networks can be represented by graphs, in which the nodes represent biological entities in a given network and the edges indicate interactions between them. Comparing the networks of different organisms and identifying their common substructures, such as signaling pathways and protein complexes, can lead to a better understanding of the regulatory mechanisms underlying various cellular processes. For this reason, there have been growing efforts to find conserved interaction patterns in protein interaction networks (Kelley *et al.*, 2003; Koyutürk *et al.*, 2004; Sharan *et al.*, 2005; Scott *et al.*, 2006; Shlomi *et al.*, 2006; Yang and Sze, 2007; Dost *et al.*, 2008; Singh *et al.*, 2008), metabolic networks (Koyutürk *et al.*, 2004; Pinter *et al.*, 2005; Yang and Sze, 2007), gene regulatory networks (Akutsu *et al.*, 1998), and signal transduction networks (Steffen *et al.*, 2002). It has been demonstrated that searching for conserved interaction patterns can detect many well known pathways and can also make statistically significant predictions of novel interaction pathways.

Despite the initial success of existing methods, there still exist a number of problems that limit the applicability of these methods. For example, many of these methods suffer from high computational complexity that makes it difficult to use them to search for long query paths. The number of consecutive insertions and deletions are often restricted in these methods, which may prevent the detection of distant homologous pathways, and many of them rely on heuristics or randomized algorithms that may not necessarily yield optimal results.

In this paper, we focus on the problem of finding conserved pathways in a protein interaction network that are homologous to a known linear pathway. We formulate the problem as a pathway alignment problem, in which the goal is to find the path in a given protein network that is most similar to a given query. Based on hidden Markov models (HMMs), we propose a general probabilistic framework for scoring pathway alignments and present an efficient search algorithm that overcomes the aforementioned limitations of previous algorithms. Given a query path, our method identifies the top k matching paths in a network of interest, where the detected paths may contain any number of consecutive insertions and deletions. Since the search algorithm has a very low computational complexity that is linear in the query size as well as in the number of edges in the network, our method can be used to effectively search for long query paths in remotely related organisms. Although HMMs have been widely used in sequence homology search, we believe that this is their first application to homology search in pathways.

2 Methods

In this section, we present an algorithm for solving the following pathway alignment problem: Given a linear query path $\mathbf{p}$ and a protein interaction network $\mathcal{G}$, find the optimal path $\mathbf{q}$ in the network $\mathcal{G}$ that is closest to the query.

2.1 Pathway alignment

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph representing a protein interaction network. We assume that $\mathcal{G}$ has a set $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ of N nodes, representing the proteins in the network, and a set $\mathcal{E} = \{e_{ij}\}$ of M edges, representing the interactions between proteins v_i and v_j. As the network $\mathcal{G}$ is an undirected graph, we assume that if there exists an interaction between v_i and v_j, both e_{ij} and e_{ji} are present in the set $\mathcal{E}$. For a protein pair (v_i, v_j) with $e_{ij} \in \mathcal{E}$, we define their interaction reliability as $w(v_i, v_j)$. Given a query path $\mathbf{p} = p_1 p_2 \ldots p_L$ that consists of L proteins, we define the sequence similarity score between p_i and v_j as $h(p_i, v_j)$. Our goal is to find the best matching path $\mathbf{q} = q_1 q_2 \ldots q_{L'}$ ($q_i \in \mathcal{V}$) in the protein interaction network that maximizes a predefined pathway alignment score $S(\mathbf{p}, \mathbf{q})$.

To obtain meaningful pathway alignments, the alignment score $S(\mathbf{p}, \mathbf{q})$ should be defined in such a way that combines the similarity score $h(p_i, q_j)$ between the aligned proteins p_i and q_j, the interaction reliability score $w(q_j, q_{j+1})$ between the proteins q_j and q_{j+1} ($1 \leq j \leq L' - 1$), and the penalty for insertion (q_j that does not have a matching protein in the query path $\mathbf{p}$) and deletion (p_i that does not have a matching protein in the retrieved path $\mathbf{q}$). Figure 1C illustrates an example of a query path $\mathbf{p}$ (Fig. 1A) that is aligned to the best matching path $\mathbf{q}$ in the network $\mathcal{G}$ (Fig. 1B). A dashed line between p_i and v_j or between p_i and q_j indicates that they have significant sequence similarity. In this example, the optimal alignment that maximizes the alignment score $S(\mathbf{p}, \mathbf{q})$ has one insertion (node q_3 in the retrieved path) and one deletion (node p_4 in the query).

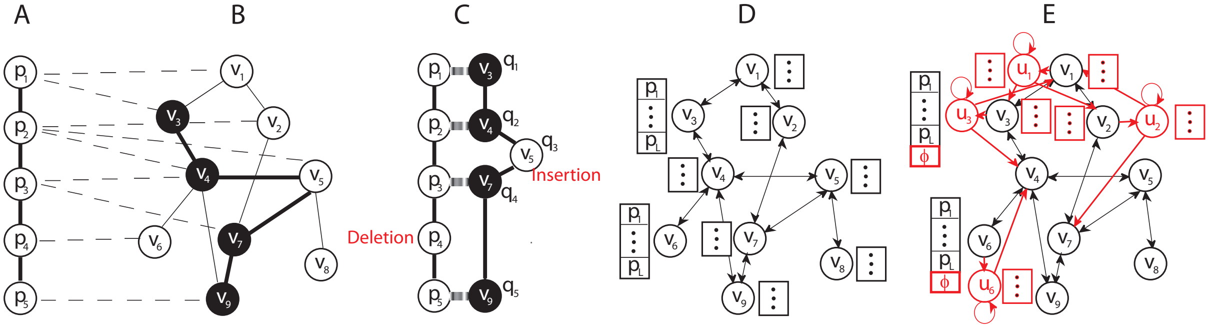

Figure 1: (A) Query path $\mathbf{p}$. (B) Protein interaction network $\mathcal{G}$. (C) An example of a pathway alignment with the best matching path $\mathbf{q}$ in the network $\mathcal{G}$. (D) Ungapped hidden Markov model (HMM) for finding the optimal pathway alignment. The dots next to the hidden states represent all possible symbols that can be emitted. (E) Modified HMM that allows insertions and deletions. The hidden states can emit a gap symbol ϕ in this modified model. For simplicity, changes to the HMM are shown only for the nodes v_1, v_2, v_3 and v_6.

2.2 Network representation using HMM

To define the alignment score $S(\mathbf{p}, \mathbf{q})$, we adopt the hidden Markov model (HMM) formalism. For simplicity, we start with an HMM that does not allow insertions or deletions in the pathway alignment. We construct the HMM based on the network graph $\mathcal{G}$ so that each node $v_i \in \mathcal{V}$ in the graph corresponds to a hidden state in the HMM. For convenience, we represent this hidden state using the same notation v_i and the HMM has an identical structure as the network graph $\mathcal{G}$. The resulting HMM can be viewed as a generative model that produces (or "emits") an interesting substructure of the original network, such as a signaling pathway. From this point of view, we regard the query path $\mathbf{p} = p_1 \dots p_L$ as an observation sequence generated by the constructed HMM. Figure 1D illustrates the HMM where the emittable symbols are shown next to the hidden states.

Based on this representation, the interaction reliability score and the sequence similarity score can be integrated naturally into the probabilistic framework. We define a mapping $\mathbf{f} : w(v_m, v_n) \mapsto t(v_n|v_m)$ that converts the interaction reliability $w(v_m, v_n)$ between two proteins to the following transition probability

$$P(q_i = v_n|q_{i-1} = v_m) = t(v_n|v_m) = \mathbf{f}(w(v_m, v_n)) \tag{1}$$

between the corresponding hidden states in the HMM. The mapping $\mathbf{f}$ is defined so that (i) $t(v_n|v_m) =$

0 for $e_{mn} \notin \mathcal{E}$, (ii) $\sum_n t(v_n|v_m) = 1$ for all m, and (iii) $t(v_{n1}|v_m) > t(v_{n2}|v_m)$ for $w(v_m, v_{n1}) > w(v_m, v_{n2})$. We define another mapping $\mathbf{g} : h(p_i, v_m) \mapsto e(p_i|v_m)$ that converts the sequence similarity score $h(p_i, v_m)$ to the following emission probability

$$P(p_i|q_j = v_m) = e(p_i|v_m) = \mathbf{g}(h(p_i, v_m)), \tag{2}$$

where $q_j = v_m$ is a hidden state in the HMM (representing a protein in the network $\mathcal{G}$) and p_i is an emitted symbol (representing a protein in the query $\mathbf{p}$ that is aligned to q_j). The mapping $\mathbf{g}$ is defined so that (i) $\sum_{p \in \mathcal{P}} e(p|v_m) = 1$ for all m, where $\mathcal{P}$ is the set of distinct proteins in the query path $\mathbf{p} = p_1 \dots p_L$, and (ii) $e(p_{i1}|v_m) > e(p_{i2}|v_m)$ for $h(p_{i1}, v_m) > h(p_{i2}, v_m)$.

2.3 Ungapped pathway alignment

Based on the HMM framework, the problem of finding the best matching path is transformed into the problem of finding the optimal state sequence in the HMM that maximizes the observation probability of the given query path. In an ungapped pathway alignment, the matching path $\mathbf{q} = q_1 q_2 \dots q_L$ has the same length as the query path $\mathbf{p} = p_1 p_2 \dots p_L$, hence q_i will be the underlying state for the "observed symbol" p_i.

To find the best matching path, a dynamic programming algorithm, called the Viterbi algorithm, can be used to solve the problem in polynomial time. We define $\gamma(t, j)$ as the log-probability of the most probable path for the sub-query $\widehat{\mathbf{p}} = p_1 \dots p_t$ of length $t\,(\leq L)$, where the underlying state of p_t is $q_t = v_j$. We compute $\gamma(t, j)$ recursively as follows:

$$\gamma(t, j) = \max_i \Big[\gamma(t-1, i) + \log t(v_j|v_i) + \log e(p_t|v_j)\Big]. \tag{3}$$

We repeat the above iterations until $t = L$. We then obtain the maximum log-probability of the query $\mathbf{p}$ as follows:

$$\log P(\mathbf{p}, \mathbf{q}^*) = \max_{\mathbf{q}} \Big[\log P(\mathbf{p}, \mathbf{q})\Big] = \max_j \gamma(L, j), \tag{4}$$

where $\mathbf{q}^* = \arg\max_{\mathbf{q}}[\log P(\mathbf{p}, \mathbf{q})]$ is the best matching path for $\mathbf{p}$ in the network. Once we have $\log P(\mathbf{p}, \mathbf{q}^*)$, it is straightforward to find $\mathbf{q}^*$ by tracing the recursive equations that led to the maximum log-probability $\log P(\mathbf{p}, \mathbf{q}^*)$. Although the above algorithm only finds the optimal path, we can extend the algorithm to find the top k paths simply by replacing the max operator by an operator that finds the k largest scores.

Note that $S(\mathbf{p},\mathbf{q}) = \log P(\mathbf{p},\mathbf{q})$ can serve as a good alignment score for the paths $\mathbf{p}$ and $\mathbf{q}$ that effectively combines sequence similarity and interaction reliability. In principle, we can also use non-stochastic emission scores $s_{em}(p_t|v_j)$ and transition scores $s_{tr}(v_j|v_i)$ in place of the log-probabilities $\log e(p_t|v_j)$ and $\log t(v_j|v_i)$, respectively, in the recursive equation (3). This will yield a non-stochastic alignment score instead of an observation probability.

2.4 Pathway alignment with gaps

To accommodate insertions and deletions, we modify the HMM as follows. To model deletions, we add an accompanying state u_m for every state v_m in the HMM. We add an outgoing edge from v_m to u_m and add outgoing edges from u_m to all of the neighboring states of v_m in the network $\mathcal{G}$. To be more precise, u_m will have an outgoing edge to every $v_n \in \mathcal{V}(m) = \{v_n | e_{mn} \in \mathcal{E}\}$. By varying the transition probability $t(u_m|v_m)$, we can control the probability (hence, the penalty) of having deletions. We adjust the outgoing transition probabilities from v_m so that $t(u_m|v_m) + \sum_{v_n} t(v_n|v_m) = 1$. We control the probability of having consecutive deletions by adjusting the probability $t(u_m|u_m)$ for making self-transitions at u_m. The outgoing transition probabilities $t(v_n|u_m)$ from an accompanying state u_m are chosen so that they are proportional to $t(v_n|v_m)$ and satisfy $t(u_m|u_m) + \sum_{v_n} t(v_n|u_m) = 1$. To model insertions, we allow the original states $v_1, \ldots, v_N$ in the HMM to emit a gap symbol ϕ in addition to the proteins in the query path $\mathbf{p}$. The gap emission probability $e(\phi|v_m)$ can be used to control the probability (hence, the penalty) of having insertions. The structure of the modified HMM is depicted in Fig. 1E. Note that the matching path $\mathbf{q} = q_1 q_2 \ldots q_{L'}$ might have different length L' from the length L of the query path $\mathbf{p} = p_1 p_2 \ldots p_L$ when there are insertions or deletions in the pathway alignment. Hence we may have $j \neq i$ for the underlying state q_j of the "observed symbol" p_i.

To find the optimal path that may include one or more gaps, we modify our dynamic programming algorithm correspondingly. We define $\gamma(t,d,j)$ as the log-probability of the most probable path for $\widehat{\mathbf{p}} = p_1 \ldots p_t$ that contains d insertions and ends at v_j $(1 \leq j \leq 2N)$, where we use $v_{m+N} = u_m$ for convenience. The value of $\gamma(t,d,j)$ is computed recursively as follows:

$$\gamma(t,d,j) = \max_i \begin{cases} \gamma(t-1,d,i) + \log t(v_j|v_i) + \log e(p_t|v_j), & \text{emits } p_t \\ \gamma(t,d-1,i) + \log t(v_j|v_i) + \log e(\phi|v_j), & \text{emits } \phi. \end{cases} \tag{5}$$

We repeat the above iterations until we reach $t = L$ and $d = D$, where D is the maximum number of allowed insertions. Note that there is no explicit limit for the number of deletions. We then compute the log-probability of the optimal path as follows:

$$\log P(\mathbf{p}, \mathbf{q}^*) = \max_{\mathbf{q}} \Big[\log P(\mathbf{p}, \mathbf{q}) \Big] = \max_{d,j} \gamma(L, d, j). \tag{6}$$

The path $\mathbf{q}^* = \arg\max_{\mathbf{q}}[\log P(\mathbf{p}, \mathbf{q})]$ is the closest match to the query $\mathbf{p}$, and it may contain some number of insertions and deletions. As before, we can replace the max operator by an operator that finds the k largest scores if we want to find the top k matching paths instead of a single top-scoring path.

The computational complexity of the above algorithm is $O(kLDM)$ for finding the top k matching paths, where L is the length of the query, D is the maximum number of allowed insertions, and M is the number of edges in the network $\mathcal{G}$. Note that the complexity is linear with respect to the query size, the number of edges in the network, the maximum number of insertions, and the number of best matching paths we want to retrieve.

2.5 Statistical significance

We estimate the statistical significance of a retrieved path using a similar approach as the one proposed in Kelley *et al.* (2003). We first generate a large number of random graphs by permuting the protein locations in the original network $\mathcal{G}$. Therefore, all random graphs will be comprised of the same set of proteins and retain a similar structure as the original network. For each random graph, we construct an HMM and compute the best pathway alignment score for the query $\mathbf{p}$. It is well known that the Gumbel distribution provides a good approximation for the extreme value distribution (EVD) of various random variables, and it has been widely used in sequence homology search to assess the significance of predicted results (Karlin and Altschul, 1990; Durbin *et al.*, 1998). As we are interested in evaluating the statistical significance of the maximum pathway alignment score obtained from the original network, the Gumbel distribution provides a better approximation when compared to the widely used Gaussian distribution. The two unknown parameters α and β in the Gumbel distribution function $D(x) = 1 - \exp(-\exp(\frac{x-\alpha}{\beta}))$ can be estimated using simple least squares regression based on the alignment scores obtained from the random graphs. Once we have estimated α and β, we can compute the p-value of the best pathway alignment score in the original network based on the estimated distribution.

3 Results

3.1 HMM parameterization

We tested our algorithm using several protein interaction networks in the Database of Interacting Proteins (DIP) (Xenarios *et al.*, 2002). We adopted a simple non-stochastic scoring scheme for parameterizing the HMMs.

We set the transition scores $s_{tr}(v_n|v_m)$ based on the presence of interaction between the corresponding proteins. If there exists an interaction between the two nodes v_m and v_n in the network $\mathcal{G}$, we set the transition score to $s_{tr}(v_n|v_m) = 0$ (and also $s_{tr}(v_m|v_n) = 0$). Otherwise, we set the score to $s_{tr}(v_n|v_m) = -\infty$ (and also $s_{tr}(v_m|v_n) = -\infty$). This keeps the state v_m in the HMM from making a direct transition to a non-relevant state v_n, and vice versa, thereby preventing the inclusion of any irrelevant protein interactions with no biological support in the retrieved path $\mathbf{q}$. The score for making a transition into an accompanying state u_m was set to $s_{tr}(u_m|v_m) = 0$, and we set the self-transition score to $s_{tr}(u_m|u_m) = 0$ to allow consecutive deletions. The score for making a transition from u_m back to a regular state v_n was set to $s_{tr}(v_n|u_m) = 0$ for $v_n \in \mathcal{V}(m) = \{v_n|e_{mn} \in \mathcal{E}\}$ and $s_{tr}(v_n|u_m) = -\infty$ for $v_n \notin \mathcal{V}(m)$.

We set the emission score $s_{em}(p_t|v_m)$ based on the sequence similarity of the proteins p_t and v_m. We assume that every state in the HMM (including the accompanying states) can emit any protein p_t in the query path $\mathbf{p}$. The emission score was made larger for more similar proteins so that it is more likely that a hidden state v_m will emit a protein p_t that is closer to its corresponding protein. For all protein pairs (p_t, v_m) between a protein p_t in the query and a protein v_m in the network $\mathcal{G}$, we computed their E-values using the PRSS routine in the FASTA package (Pearson and Lipman, 1988). PRSS (Pearson, 1996) computes accurate E-values using the Smith-Waterman algorithm with sophisticated shuffling methods, and it is believed to be better than BLASTP in detecting significant matches (Pagni and Jongeneel, 2001). We regarded a protein pair (p_t, v_m) as a “match” if its E-value $E_v(p_t, v_m)$ was below some predefined threshold λ_{th}. Otherwise, we viewed (p_t, v_m) as a “mismatch”, which implies that the two proteins do not contain significant similarity. Based on this criterion, we set the emission score $s_{em}(p_t|v_m)$ as follows:

$$s_{em}(p_t|v_m) = \begin{cases} -\log_{10} E_v(p_t, v_m), & \text{if } E_v(p_t, v_m) \le \lambda_{th} \\ -\Delta, & \text{otherwise.} \end{cases} \tag{7}$$

The value Δ can be viewed as the mismatch penalty, and is selected so that $-\Delta \ll -\log_{10} \lambda_{th}$. We set

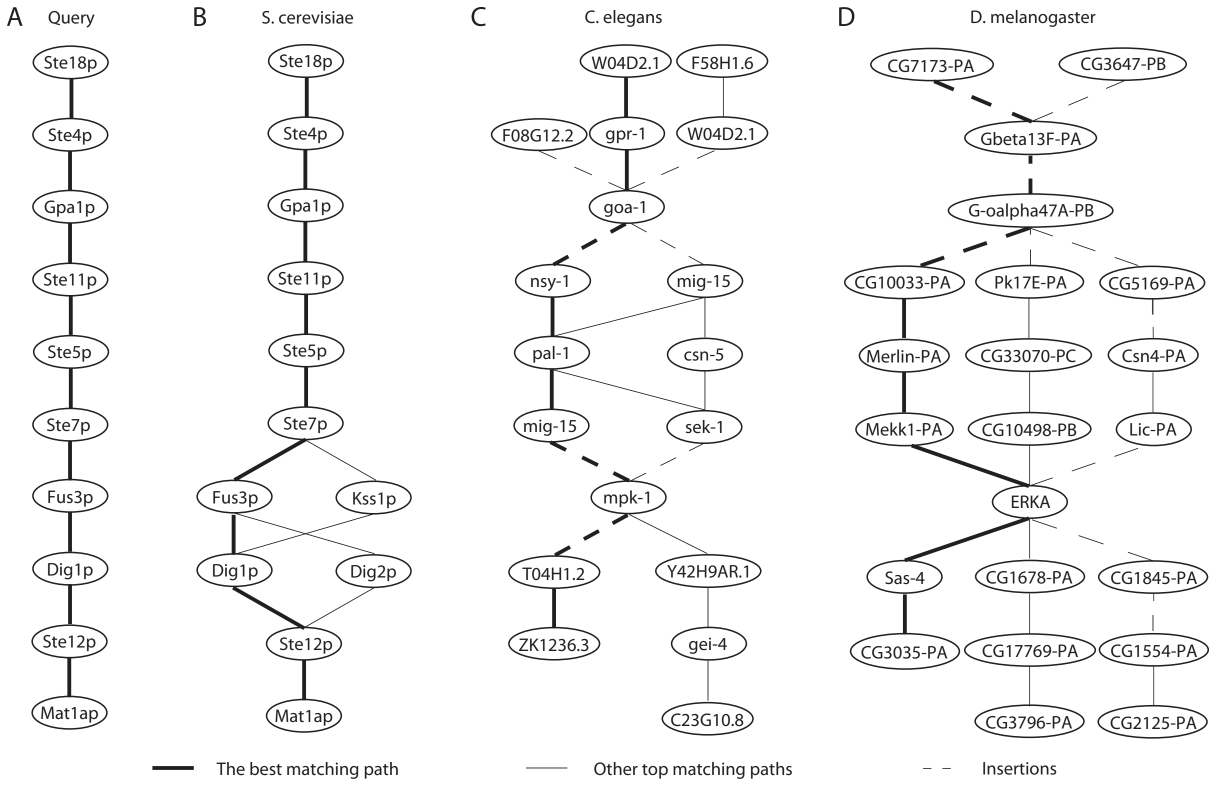

Figure 2: Query results in different networks. (A) The *S. cerevisiae* query path. (B) Matching paths in the *S. cerevisiae* network. (C) Matching paths in the *C. elegans* network. (D) Matching paths in the *D. melanogaster* network.

the insertion penalty to $s_{em}(\phi|v_m) = -\Delta_i$ and the deletion penalty to $s_{em}(p_t|u_m) = -\Delta_d$. Note that the accompanying state u_m cannot emit a gap, hence $s_{em}(\phi|u_m) = -\infty$. Finally, we set the maximum number of insertions to be the same as the length of the query ($D = L$).

3.2 Querying yeast pathways in various organisms

To verify the capability of our method for identifying relevant pathways in different organisms, we obtained the protein interaction networks of *S. cerevisiae*, *C. elegans*, *D. melanogaster*, and *E. coli* from DIP. We took a mating-pheromone response pathway of *S. cerevisiae* with 10 proteins as our query path, which contains the mitogen-activated protein (MAP) kinase cascade Ste11p–Ste7p–Fus3p (Fig. 2A). The same pathway has been used by other existing algorithms for performance evaluation (Scott *et al.*, 2006; Yang and Sze, 2007). We searched for similar paths in the *S. cerevisiae* network with

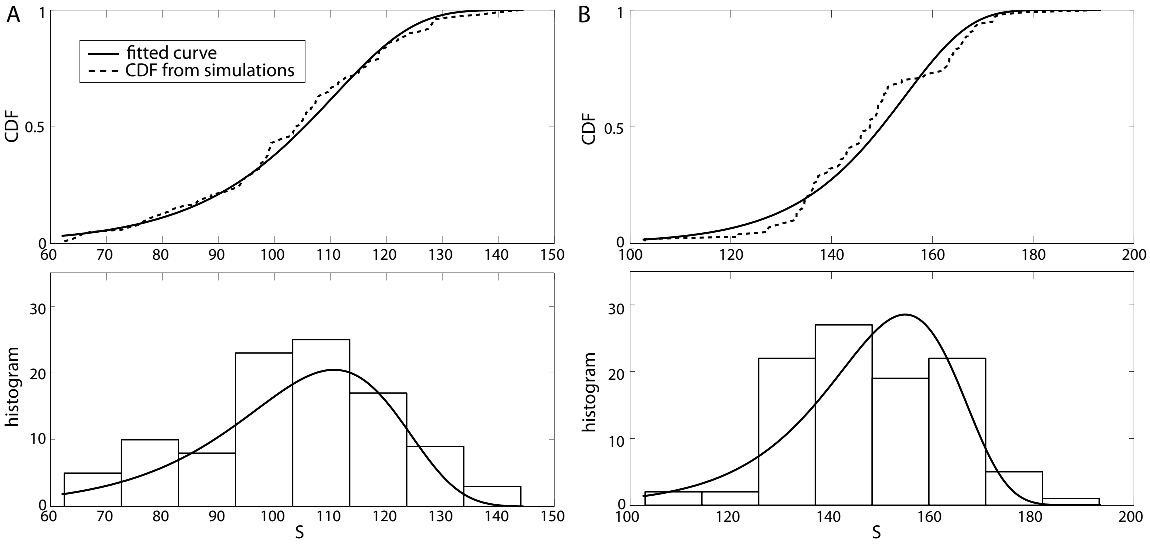

Figure 3: The cumulative distribution functions (CDFs) and histograms of the maximum alignment scores S computed from the random networks. The estimated extreme value distribution (EVD) curves are shown together. (A) *C. elegans.* (B) *D. melanogaster.*

$17,579$ interactions among $4,969$ proteins, the *C. elegans* network with $4,037$ interactions among $2,647$ proteins, and the *D. melanogaster* network with $22,840$ interactions among $7,476$ proteins. We set $\lambda_{th} = 0.5$ and $\Delta = \Delta_i = \Delta_d = 12$.

As we would expect, the best matching path in *S. cerevisiae* network was identical to the query path. This is shown in Fig. 2B, where other top matches are shown with the best matching path. The retrieved paths in the *C. elegans* network and the *D. melanogaster* network are shown in Fig. 2C and Fig. 2D, respectively. It is interesting to note that many proteins in the retrieved paths share similar functions with the corresponding proteins in the query path. The proteins Ste11p, nsy-1, mig-15, CG10033-PA, Pk17E-PA, and CG5169-PA, which are aligned to each other, all belong to the serine/threonine protein kinase family. Similarly, Ste7p, mig-15, sek-1, Mekk1-PA, CG10498-PB, and Lic-PA, also belong to the serine/threonine protein kinase family, with sek-1, Mekk1-PA, and Lic-PA being MAPK kinases. All the proteins Fus3p, Kss1p, mpk-1, and ERKA, which are aligned to each other in Fig. 2, are MAP kinases (The Flybase Consortium, 1996; Gustin *et al.*, 1998; Stein *et al.*, 2001). These results clearly indicate that our method is able to effectively identify similar pathways that are biologically meaningful.

In order to estimate the statistical significance of the predicted results, we computed the p-values for the best matching paths in the respective networks. The p-values have been computed as described in Sec. 2.5 using 100 random networks. Figures 3A and 3B show the resulting cumulative distribution functions (CDFs) and histograms of the maximum alignment scores in the random networks of *C. elegans* and *D. melanogaster*, respectively. The fitted Gumbel distributions are also shown in the figure. The p-value of the optimal path in the original *C. elegans* network was $p = 0.014$, while the p-value of the optimal path in the *D. melanogaster* network was $p = 0.069$. For both organisms, the optimal alignment scores in the original networks ranked among the top scores, indicating that the retrieved results are statistically significant. One point to note is that using a shorter query path typically leads to a smaller p-value, as it becomes more likely to detect good matches that contain less insertions and deletions. For example, if we take a small portion (Ste11p–Ste5p–Ste7p–Fus3p) of the original query shown in Fig. 2A and search the *C. elegans* network, the algorithm retrieves the corresponding part in Fig. 2C, with a smaller p-value $p = 2.6 \times 10^{-3}$. Similarly, if we use the same short query to search the *D. melanogaster* network, the corresponding portion in Fig. 2D is retrieved, where the p-value of the optimal path is only $p = 3.6 \times 10^{-9}$.

In addition to the *D. melanogaster*, *C. elegans*, and *S. cerevisiae* networks, we used the same mating-pheromone response pathway in Fig. 2A to find relevant paths in the *E. coli* protein interaction network, which contains $6,976$ interactions among $1,850$ proteins. We reduced the mismatch and indel penalties to $\Delta = \Delta_i = \Delta_d = 2$ so that the retrieved paths may contain more indels and mismatches. The p-value of the optimal path was $p = 0.66$ and its alignment score was among the lowest when compared to the alignment scores obtained from the random networks. Similarly, the p-value of the retrieved path was also high ($p = 0.59$) for the short query. This implies that the search results are statistically insignificant, which is consistent with the fact that there are no known MAP kinase pathways in bacteria (Chang and Stewart, 1998). This is a good indication that our method is very useful in identifying conserved pathways that are biologically meaningful.

3.3 Querying human pathways in fly

We further applied our algorithm to search the *D. melanogaster* network for matching paths that are similar to known human signaling pathways. We used the same parameters as before: $\lambda_{th} = 0.5$ and $\Delta = \Delta_i = \Delta_d = 12$. Figure 4A shows the retrieved paths for the human hedgehog signaling pathway and Fig. 4B shows the retrieved paths for the human MAP kinase pathway. In both cases, the top matching paths agreed well with the query paths, according to the known functional annotations

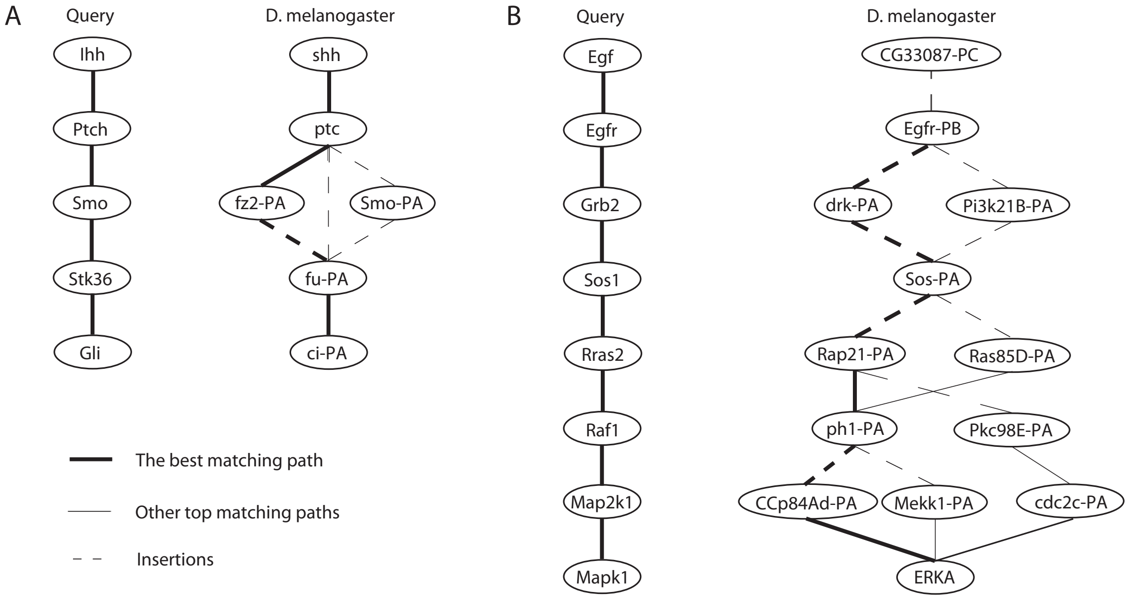

Figure 4: Human queries and their best matching paths in the *D. melanogaster* network. (A) Human hedgehog pathway and the matching paths. (B) Human MAP kinase pathway and the matching paths.

of *D. melanogaster*. In Fig. 4A, ptc is a receptor of the hedgehog pathway located at the plasma membrane (Lum and Beachy, 2004), and it has been shown that ci-PA plays an important role in the hedgehog pathway that regulates cell growth in many tissues (Lum and Beachy, 2004). For the MAP kinase query (Fig. 4B), Egfr-PB is a putative growth factor receptor; drk-PA is downstream of the receptor kinase; Rap21-PA and Ras85D-PA have putative GTPase activity; ph1-PA, Pkc98E-PA, Mekk1-PA, and cdc2c-PA all belong to the serine/threonine protein kinase family; and ERKA is an annotated nuclear MAP kinase which likely activates specific transcription factors (The Flybase Consortium, 1996; Gustin *et al.*, 1998).

Both Figs. 4A and 4B contain the putative homologous pathways in *D. melanogaster* reported in the KEGG database (Kanehisa and Goto, 2000): one of the top matching paths in Fig. 4A (shh–ptc–Smo–fu–ci) is the core of the *D. melanogaster* hedgehog signaling pathway given in Kanehisa and Goto (2000), and Egfr–drk–Sos–Ras85D–ph1–Mekk1–ERKA (Fig. 4B) is part of the putative MAP kinase pathway for *D. melanogaster* in Kanehisa and Goto (2000). By comparing the retrieved pathways with the corresponding putative pathways of *D. melanogaster* in the KEGG database (Kanehisa and Goto, 2000), we found that our algorithm was able to retrieve the identical core segment with five

Table 1: Running time of our algorithm with different parameters (time measured in seconds).

max # of	# of	Query path length							
insertions	paths	6	8	10	12	14	16	18	20
$D = 0$	$k = 1$	47.9	67.5	85.8	105.4	122.6	144.0	163.6	179.2
$D = 5$	$k = 5$	1373.4	1878.3	2382.4	2881.5	3384.4	3894.5	4387.5	4817.5

proteins in the putative hedgehog signaling pathway[1]. For the MAP kinase query, the retrieved pathway included seven proteins that exactly matched the proteins in the putative *D. melanogaster* MAP kinase pathway[2]. These results compare favorably to the previously reported results (Shlomi *et al.*, 2006), which found two and five matched proteins for the respective pathways, indicating that our algorithm can make biologically meaningful predictions with better accuracy. As before, we also computed the p-values for the top matching paths using 100 random networks. The p-value of the optimal path for the human hedgehog signaling pathway was $p = 7.0 \times 10^{-23}$ and the p-value of the top retrieved path for the human MAP kinase pathway was $p = 2.6 \times 10^{-4}$, which show the statistical significance of the predictions.

3.4 Running time

Our method has a very low computational complexity that is linear with respect to the length of the query as well as the number of interactions in the network. Unlike most existing methods whose utility is limited to relatively short queries with 3 to 10 proteins, our method can search for very long query paths in large protein interaction networks. Table 1 summarizes the running time of our method with different parameters. We searched for queries with 6 to 20 proteins in the *S. cerevisiae* network that consists of 4,969 proteins and 17,579 interactions. The running time has been measured on a desktop computer with 2.13GHz CPU and 2GB memory. From Table 1, we see that it takes only about three minutes to find the top path for a query of length 20, where the retrieved path may contain any number of deletions and mismatches but no insertions. We see clearly in Table 1 that the running time grows linearly with the query length. If we search for the top five paths using the same query (of length 20), and allow up to five insertions and any number of deletions and mismatches, it still takes only

[1] $http://www.genome.jp/dbget-bin/get\_pathway?org\_name=dme\&mapno=04340$

[2] $http://www.genome.jp/dbget-bin/get\_pathway?org\_name=map\&mapno=04010$

about 80 minutes. Note that this is about 25 ($= D \times k$) times larger compared to the previous running time (three minutes), which confirms that the computational complexity is also linear in the maximum number of insertions D and the number of paths k we want to retrieve.

3.5 Accuracy and robustness

To evaluate the accuracy and the robustness of our HMM-based algorithm, we performed the following experiments.

We first estimated the accuracy of our algorithm in retrieving homologous pathways by using synthetic query paths. For this purpose, we followed a similar procedure used in Dost *et al.*, (2008). To obtain a reasonable set of query paths, we randomly extracted 10 paths from the *S. cerevisiae* network, whose lengths range from $L = 6$ to $L = 10$. Each of these paths was perturbed by inserting, deleting, and replacing one or more nodes. We also applied point mutations to the protein sequences in the query paths with different mutation rates of up to 80%. For each path, we used our algorithm to retrieve the top matching path in the *S. cerevisiae* network. As in the previous experiments, we used $\lambda_{th} = 0.5$ and $\Delta = \Delta_i = \Delta_d = 12$. The retrieved results have been compared to the original unperturbed paths that were used to obtain the query paths. We computed the edit distance between each retrieved path and the original path to evaluate the prediction accuracy. As a comparison, we also tried to find the best matching proteins based on sequence similarity alone. For each node in the query paths, we reported the node in the *S. cerevisiae* network with the highest alignment score using the PRSS routine (Pearson, 1996) as the "matching node" and counted the number of nodes that were correctly predicted. The prediction results are summarized in Table 2. We can see from this table that the predictions made by our HMM-based algorithm are far more accurate than those based on the best PRSS scores. For HMM-based predictions, the average distance between the retrieved optimal paths and the original paths was less than one for all types of perturbations with up to 70% point mutations. The advantage of our algorithm over the sequence-based approach becomes more pronounced for higher mutation rates. These results clearly show that the HMM-based algorithm can make accurate predictions by integrating sequence similarity and the interaction network in a sensible manner.

We also measured the sensitivity of our algorithm with respect to parameter changes. For this purpose, we used the human hedgehog signaling pathway shown in Fig. 4A to query the *D. melanogaster* network using different sets of parameters. For each parameter setting, we compared the proteins in

Table 2: The average edit distances of the retrieved paths and the original query paths for different types of node perturbations and different levels of point mutations. The results obtained by our algorithm are denoted as "HMM" and the results obtained based on the best PRSS hits are denoted as "PRSS".

Node perturbation	Point mutation rate									
	0%		50%		60%		70%		80%	
	PRSS	HMM	PRSS	HMM	PRSS	HMM	PRSS	HMM	PRSS	HMM
None	0	0	0	0	0.2	0	0.9	0.3	3.9	0.9
1 replacement	1	0.2	1	0.2	1.2	0.2	1.9	0.8	4.9	1.5
2 replacements	2	0.5	2	0.5	2.2	0.5	2.9	0.8	5.9	2.2
1 ins & 1 del	1.8	0.3	1.8	0.3	2.0	0.3	2.7	0.7	5.7	2.3
2 ins & 2 del	3.3	0.6	3.3	0.6	3.5	0.6	4.2	0.7	7.2	3.8

Table 3: The relative changes in the top 10 retrieved paths. We computed the relative number of proteins that were added to or removed from the predicted results due to the parameter changes. The changes have been measured by comparing the predicted results to the results obtained using the original parameters ($\lambda_{th} = 0.5$ and $\Delta = \Delta_i = \Delta_d = 12$) employed in Sec. 3.3. The relative number of added proteins is denoted by r_I, and the relative number of removed proteins is denoted by r_D.

λ_{th}	$\Delta = \Delta_i = \Delta_d$							
	6		12		18		26	
	r_I	r_D	r_I	r_D	r_I	r_D	r_I	r_D
0.05	0	0	0	0	0	0	0.07	0.03
0.25	0	0	0	0	0	0	0.07	0.03
0.5	0	0	0	0	0	0	0.07	0.03
1.0	0	0	0	0	0	0	0.07	0.03
5.0	0.02	0.02	0.02	0.02	0.02	0.02	0.09	0.05

the top 10 retrieved paths with the proteins included in the top 10 paths based on the original setting ($\lambda_{th} = 0.5$ and $\Delta = \Delta_i = \Delta_d = 12$) that was used to obtain the results in Sec. 3.3. To estimate the relative changes of the retrieved proteins, we computed the following values:

$$r_I = C_I/C \quad \text{and} \quad r_D = C_D/C,$$

where C is the number of proteins in the top retrieved paths based on the original setting, C_I is the number of proteins that were added as a result of the parameter change, and C_D is the number of proteins that were removed from the top paths due to the change.

Table 3 shows the relative number of added proteins and that of removed proteins for different sets of parameters. As we can see, the retrieved results are not significantly affected by parameter changes. For large ranges of λ_{th} (from 0.05 to 1.0) and $\Delta = \Delta_i = \Delta_d$ (from 6 to 18), the set of top 10 retrieved paths stayed the same. Even for larger values of $\lambda_{th} (= 5.0)$ and $\Delta = \Delta_i = \Delta_d (= 26)$, the changes in the retrieved results were relatively small. In addition to measuring the relative changes in the predicted results, we computed the p-values of the top paths based on 100 random networks. In all cases, the p-values ranged between 1.1×10^{-26} and 6.2×10^{-18}, showing that the predictions made by our HMM-based method are statistically significant. This explains why the retrieved paths were so close to each other. As we can see from these results, our HMM-based algorithm performs robustly for a considerably large range of parameter values, especially when the retrieved paths are biologically meaningful. The predicted results may have larger variations when the retrieved paths are not statistically significant, as in the case of querying the *E. coli* network using the yeast MAP kinase pathway, described in Sec. 3.2.

4 Discussion

There exist a number of algorithms that can be applied to pathway search (Kelley *et al.*, 2003; Pinter *et al.*, 2005; Scott *et al.*, 2006; Shlomi *et al.*, 2006; Yang and Sze, 2007). Among these algorithms, PathBLAST (Kelley *et al.*, 2003) focuses on finding conserved pathways by comparing two networks, and this reduces to a pathway search problem if one of the networks is a specific pathway. However, PathBLAST does not allow consecutive deletions of proteins in the query path nor consecutive insertions of proteins in the matching path, limiting its applicability to closely matching pathways. Furthermore, its computational complexity contains factorial factors in the query length, making it impractical for long paths that contain more than 6 to 7 proteins.

To solve these problems, two algorithms have been proposed based on the color coding technique (Scott *et al.*, 2006; Shlomi *et al.*, 2006) that can search for simple paths of length around 10. Although these algorithms have significantly reduced the running time, the algorithmic complexity is still exponential in the query length. Hence they quickly become infeasible for slightly longer queries. In addition, the use of randomized algorithms does not guarantee the optimality of the query result.

A more recent algorithm called PathMatch (Yang and Sze, 2007) reduces the pathway search problem to the problem of finding the longest weighted path in a directed acyclic graph by relaxing the constraint for finding simple paths. This allows top matching paths to be found in polynomial time and obviates the need for randomized algorithms. However, PathMatch has a limited flexibility in the choice of the scoring scheme since mismatches and indels are treated in an identical manner and it is difficult to use different penalties for different mismatches or indels.

When compared to the existing methods, our HMM-based method has a significantly lower computational complexity that is linear with respect to the query length and the number of edges in the network, and it can search for query paths with more than 10 proteins in a network with thousands of nodes and tens of thousands of interactions within a few minutes. An important advantage of the HMM-based framework is that it is very flexible in the choice of the scoring scheme. We can use different penalties for mismatches, insertions, and deletions, and it is also possible to assign different penalties to different types of mismatches. Furthermore, our algorithm allows the matching paths to contain any number of consecutive deletions and insertions. We can also use an affine gap penalty model for scoring consecutive deletions, where gap openings and gap extensions are treated differently.

Similar to the PathMatch algorithm (Yang and Sze, 2007), we may in principle have a repeated occurrence of a network node v_j in the retrieved path. Such repeated occurrences are not frequently observed in practice unless all the proteins in the query $\mathbf{p}$ are very similar to each other. As mentioned in Yang and Sze (2007), these limited occurrences of repeats can be biologically useful in identifying proteins that have multiple roles in a signaling pathway, and there exist many known examples of such multi-functional proteins (Teichmann *et al.*, 2001). We did not observe many repeated occurrences of the same protein in our retrieved paths.

In our tests, we have used relatively simple non-stochastic scores for parameterizing the HMMs. We have shown that, even with this simple scoring scheme, our retrieved paths are closer to the putative homologous *D. melanogaster* pathways in KEGG than the paths reported in Shlomi *et al.* (2006). Considering the flexibility of the proposed framework, it would be beneficial to use a more elaborate scoring scheme in the future. For example, we may incorporate additional information, such as gene ontology (GO) annotations and gene expression data (Sharan *et al.*, 2005; Shlomi *et al.*, 2006), for more reliable estimations of the HMM parameters. We may also incorporate methods for evaluating the reliabilities of protein interactions (von Mering *et al.*, 2002; Bader *et al.*, 2004; Sharan *et al.*, 2005) to obtain more robust transition scores for the HMMs. Although the pathway alignment score $S(\mathbf{p}, \mathbf{q})$ used in this paper incorporates only the interaction reliability of the protein network, we can easily incorporate the interaction reliability of the query as well. This can be achieved by adding an additional term for the reliability of the interaction between p_{t-1} and p_t in the recursive equations (3) and (5). As a final remark, the HMM-based method proposed in this paper is currently limited to linear queries, and we are investigating the possibility of extending the framework to support more general queries such as trees.

Acknowledgments

This work was supported in part by the National Cancer Institute under 2 R25CA090301-06 (XQ) and the National Science Foundation under DBI-0624077 (SS).

Disclosure Statement

No conflicting financial interests exist.

References

[1] Akutsu T, Kuhara S, Maruyama O, Miyano S (1998) Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions. In *Proc 9th Annu ACM-SIAM Symp Discrete Alg (SODA 1998)*:695–702.

[2] Bader J, Chaudhuri A, Rothberg JM, Chant J (2004) Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnol* **22**:78–85.

[3] Chang C and Stewart RC (1998) The two-component system. Regulation of diverse signaling pathways in prokaryotes and eukaryotes. *Plant Physiol* **117**:723–731.

[4] Dost B, Shlomi T, Gupta N, Ruppin E, Bafna V, Sharan R (2008) QNet: a tool for querying protein interaction networks. *J Comput Biol* **15**(7):913–925.

[5] Durbin R, Eddy SR, Krogh A, Mitchison G (1998) *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* Cambridge University Press, Cambridge, UK.

[6] Gustin MC, Albertyn J, Alexander M, Davenport K (1998) MAP kinase pathways in the yeast *Saccharomyces cerevisiae. Microbiol Mol Biol Rev* **62**:1264–1300.

[7] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA* **98**(8):4569–4574.

[8] Kanehisa M and Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28**(1):27–30.

[9] Karlin S and Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA* **87**:2264–2268.

[10] Kelley BP, Sharan R, Karp RM, Sittler T, Root DE, Stockwell BR, Ideker T (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc Natl Acad Sci USA* **100**(20):11394–11399.

[11] Koyut*ü*rk M, Grama A, Szpankowski W (2004) An efficient algorithm for detecting frequent subgraphs in biological networks. *Bioinformatics* **20**:SI200–207.

[12] Lum L and Beachy PA (2004) The Hedgehog response network: sensors, switches, and routers. *Science* **304**(5678):1755–1759.

[13] Mann M, Hendrickson R, Pandey A (2001) Analysis of proteins and proteomes by mass spectrometry. *Annu Rev Biochem* **70**:437–473.

[14] Pagni M and Jongeneel CV (2001) Making sense of score statistics for sequence alignments. *Brief Bioinform* **2**(1):51–67.

[15] Pearson WR and Lipman DJ (1988) Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* **85**(8):2444–2448.

[16] Pearson WR (1996) Effective protein sequence comparison. *Methods Enzymol* **266**:227–258.

[17] Pinter RY, Rokhlenko O, Yeger-Lotem E, Ziv-Ukelson M (2005) Alignment of metabolic pathways. *Bioinformatics* **21**(16):3401–3408.

[18] Scott J, Ideker T, Karp RM, Sharan R (2006) Efficient algorithms for detecting signaling pathways in protein interaction networks. *J Comput Biol* **13**:133–144.

[19] Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T (2005) Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci USA* **102**(6):1974–1979.

[20] Shlomi T, Segal D, Ruppin E, Sharan R (2006) QPath: a method for querying pathways in a protein-protein interaction network. *BMC Bioinformatics* **7**:199.

[21] Singh R, Xu J, Berger B (2008) Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc Natl Acad Sci USA* **105**(35):12763-12768.

[22] Steffen M, Petti A, Aach J, D'haeseleer P, Church G (2002) Automated modelling of signal transduction networks. *BMC Bioinformatics* **3**:34.

[23] Stein L, Sternberg P, Durbin R, Thierry-Mieg J, Spieth J (2001) WormBase: network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Res* **29**:82–86.

[24] Teichmann SA, Rison SC, Thornton JM, Riley M, Gough J, Chothia C (2001) Small-molecule metabolism: an enzyme mosaic. *Trends Biotechnol* **19**:482–486.

[25] The Flybase Consortium (1996) FlyBase: the *Drosophila* database. *Nucleic Acids Res* **24**:53–56.

[26] Uetz P, Rajagopala SV, Dong YA, Haas J (2004) From ORFeomes to protein interaction maps in viruses. *Genome Res* **14**(10B):2029–2033.

[27] von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**:399–403.

[28] Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* **30**:303–305.

[29] Yang Q and Sze SH (2007) Path matching and graph matching in biological networks. *J Comput Biol* **14**:56–67.