Bioinformatics Advance Access published September 16, 2011

An Efficient Network Querying Method Based on Conditional Random Fields

Qiang Huang [1], Ling-Yun Wu [1,*] and Xiang-Sun Zhang [1]

[1]National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

Associate Editor: Dr. Trey Ideker

ABSTRACT

Motivation: A large amount of biomolecular network data for multiple species have been generated by high-throughput experimental techniques, including undirected and directed networks such as protein-protein interaction networks, gene regulatory networks and metabolic networks. There are many conserved functionally similar modules and pathways among multiple biomolecular networks in different species, therefore, it is important to analyze the similarity between the biomolecular networks. Network querying approaches aim at efficiently discovering the similar subnetworks among different species. However, many existing methods only partially solve this problem.

Results: In this paper, a novel approach for network querying problem based on conditional random fields (CRF) model is presented, which can handle both undirected and directed networks, acyclic and cyclic networks, and any number of insertions/deletions. The CRF method is fast and can query pathways in a large network in seconds using a PC. To evaluate the CRF method, extensive computational experiments are conducted on the simulated and real data, and the results are compared with the existing network querying methods. All results show that the CRF method is very useful and efficient to find the conserved functionally similar modules and pathways in multiple biomolecular networks.

Availability: Code and data are available at `http://doc.aporc.org/wiki/CNetQ`

Contact: lywu@amt.ac.cn

Supplementary information: Supplementary materials are available at Bioinformatics online.

1 INTRODUCTION

The high-throughput experimental techniques have been dramatically improved in past decades and a large amount of biomolecular network data have been produced. For example, protein-protein interaction (PPI) networks are available in many databases such as DIP (Xenarios *et al.*, 2000) which catalogs experimentally determined protein interactions for many species including *Saccharomyces cerevisiae, Drosophila melanogaster, Escherichia coli, Caenorhabditis elegans*. Since the biomolecular networks contain many conserved functionally similar modules and pathways, it is important to analyze the similarity among multiple biomolecular networks to discover signaling pathways, find conserved modules, and reveal new biological function subnetworks. Network querying approaches attempt to address this problem via optimally mapping the queried network into a subnetwork of the given biomolecular network.

There already exist many methods for network querying in literature. PathBLAST (Kelley *et al.*, 2004) is a network alignment and search tool for comparing PPI networks, but it only could be used to query a short protein path. Pinter *et al.* (2005) devised an algorithm called MetaPathwayHunter for querying metabolic networks. MetaPathwayHunter could find queries for more general pathways, but it only could be applied to small networks without cycles. QPath (Shlomi *et al.*, 2006) is a comprehensive framework for querying a linear pathway in a network of interest, while QNet (Dost *et al.*, 2008) extends the type of queried network from the linear pathways to trees and networks of bounded tree width. Qian *et al.* (2009) proposed an efficient framework for finding pathways based on Hidden Markov Models (HMM) and it only could deal with the simple path.

The efficiency, accuracy and capability to handle general network structures are three major criteria for devising network querying approaches. Many previous methods are either limited by the size of networks, the types of networks, or the maximum allowable number of insertions/deletions. For example, PathBLAST, QPath and HMM method only consider the simple path (i.e. linear chain), while MetaPathwayHunter and QNet can not query the network with cycles.

Some recent approaches advanced the step to the network querying problem with general network structure. SAGA (Tian *et al.*, 2007) is an approximate graph matching technique that minimizes a subgraph distance function with a parameter λ over all possible matchings to get the similar subgraph. PathMatch/GraphMatch (Yang and Sze, 2007) formally define the path matching and graph matching problems and give algorithms for these problems respectively. They reduced the path matching problem to find a longest weighted path in a directed acyclic graph, and the graph matching problem to find the highest scoring subgraphs in a large graph. MNAligner (Li *et al.*, 2007) is a network alignment method based on the integer quadratic programming. It only works well for the small network alignment problem since solving the large scale integer quadratic programming is time-consuming. TORQUE (Bruckner *et al.*, 2009) is a topology-free querying algorithm. Given a query, TORQUE finds a matching set of proteins that are sequence similar to the query and then span a connected region of the network to find the matching subnetwork. PADA1 (Blin *et al.*, 2010) is an exact algorithm

*To whom correspondence should be addressed. Email: lywu@amt.ac.cn

for querying networks based on dynamic programming and color-coding technique, which transforms a network with cycles to a tree by using a brute-force method. However, due to the intrinsic contradiction between the computational complexity and the capability of handling general networks, developing an efficient general network querying method for analyzing large scale biomolecular networks is still a challenge.

In this paper, we will present a conditional random field (CRF) model for the network querying problem by utilizing the similarity of molecular nodes and the interaction topology. CRF is a probabilistic framework which firstly is used for labeling and segmenting sequential data (Lafferty *et al.*, 2001). As an extension and generalization of HMM and maximum entropy Markov models, CRF has many advantages. It allows the strong independence assumptions of HMM to be relaxed, as well as overcomes the label-bias problem exhibited by maximum entropy Markov models. CRF has been successfully applied to labeling problem in many fields including natural language processing (Lafferty *et al.*, 2001; Sha and Pereira, 2003) and bioinformatics (DeCaprio *et al.*, 2007; Wu *et al.*, 2009).

The rest of this paper is organized as follows. In Section 2, we will describe the network querying problem formally and introduce a novel CRF model for network querying problem. To evaluate the proposed method, we conduct computational experiments on several simulated and real data including undirected and directed networks. Compared with some popular network querying methods such as PathMatch/GraphMatch (Yang and Sze, 2007), QNet (Dost *et al.*, 2008), HMM (Qian *et al.*, 2009) and MNAligner (Li *et al.*, 2007), the results show that the new method is accurate and efficient. The results of computational experiments will be reported in Section 3. Finally, conclusion and discussion will be given in Section 4.

2 METHODS

Let $G = (V_G, E_G)$ be a biomolecular network (e.g. protein-protein interactions networks, gene regulatory networks), where each node $v \in V_G$ denotes a protein/gene, and each edge $e = (v_i, v_j) \in E_G$ represents the interaction between nodes v_i and v_j. Given a target network G and a query network X, the task of network querying is to find the best matching subnetwork Y for X from the target network G.

2.1 Conditional random fields model

If we consider G as a label set, that is, each node $v \in V_G$ as a label and each edge $e = (v_i, v_j) \in E_G$ as the relation between labels v_i and v_j, the network querying problem can be treated as a network labeling problem: given a label set G and a query network X, find the best matching labels Y for X from the label set G. Therefore, in the rest of this paper, the word 'label' denotes the node in the target network G. Figure 1(a) illustrates the general process of network querying and Figure 1(b) shows an example of transforming the network querying problem into labeling problem.

In the CRF model for network querying problem, the conditional probability of a set of labels $Y = (V_Y, E_Y) \subseteq G$ for a given query network $X = (V_X, E_X)$ is defined as:

$$\Pr(Y|X) = \frac{1}{Z(X)} \prod_{x_i \in V_X} f_N(y_i, X, i) \prod_{(x_i,x_j) \in E_X} f_E(y_i, y_j, X, i, j)$$

where $Z(X) = \sum_Y \prod f_N(y_i, X, i) \prod f_E(y_i, y_j, X, i, j)$ is the normalization factor, $y_i \in V_Y$ is the corresponding label for the node $x_i \in V_X$, and two feature functions f_N and f_E are used to model the matching of nodes and edges respectively. The first product is over all the nodes $x_i \in V_X$, and the second product is over all the edges $(x_i, x_j) \in E_X$. The underlying

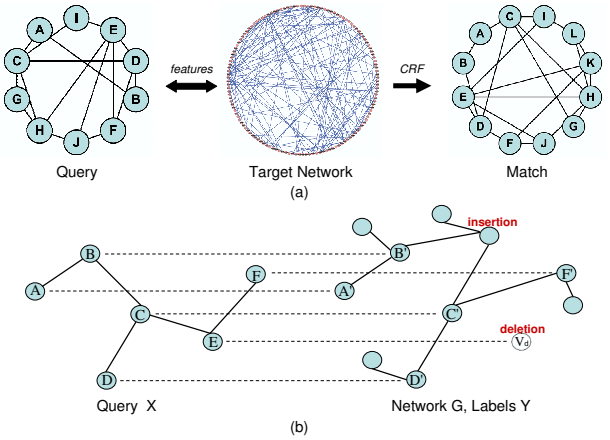


Fig. 1. (a) An illustration of network querying process. Given the query network (left) and target network (middle), the feature functions are constructed based on the node similarity and the topology similarity between two networks. Then the CRF model is applied to find the best matching subnetwork (right) in the target network. (b) An example of transforming the network querying problem to labeling problem. The left is the query network X (nodes $ABCDEF$), and the right is the target network G. The matching result Y (nodes $A'B'C'D'F'$) with an insertion and a deletion is shown. The correspondences between the nodes of X and Y are shown by dashed lines.

assumption is that the probability of the labels only depends on the similarity of nodes as well as edges.

Based on the above CRF model, the network querying problem is solved by finding the optimal labels $Y \subseteq G$ that maximizes the conditional probability $\Pr(Y|X)$ for given query network X. The task to find optimal labels in CRF model is called decoding. There are already many decoding algorithms in literature (Wainwright and Jordan, 2008). When the network X is a tree (or forest), the decoding problem can be solved exactly by tree belief propagation algorithm. If X is a chain, the simplest case, Viterbi algorithm (Viterbi, 1967) is a faster exact algorithm. For the general networks, there exist many approximate algorithms, in which we choose loopy belief propagation algorithm in this study.

The proposed CRF model is a very general framework which is flexible to incorporate various information. In this paper, the node feature function is defined directly by the similarity between the corresponding nodes of X and Y as follows:

$$f_N(y_i, X, i) = S(x_i, y_i),$$

where $S(x_i, y_i)$ is the non-negative similarity score between nodes x_i and y_i. The edge feature function is a little more complicated. One edge in X is considered to be matched well only if its two end nodes are matched well by two closely connected labels in G. Therefore, the following feature function is used:

$$f_E(y_i, y_j, X, i, j) = \frac{S(x_i, y_i) + S(x_j, y_j)}{2} W(y_i, y_j).$$

where $W(y_i, y_j)$ is the non-negative connectivity score between nodes y_i and y_j. In the next section, we will discuss how to accommodate the gaps in the CRF model by properly defining the similarity score and the connectivity score.

2.2 Modeling gaps

In the network querying problem, there are two kinds of gaps need to handle: insertion and deletion. Insertion is the case that two neighboring nodes in X are labeled by two non-neighboring nodes in G, that is, some additional

nodes are needed so that the matching subnetwork Y can have similar network structure as the query X. Conversely, deletion is that some nodes in the query X do not have corresponding counterparts in the matching subnetwork Y. In other words, the counterparts of these nodes are deleted in the network G. Figure 1(b) shows an example of insertion and deletion.

The insertions are penalized by the length of the shortest path between two non-neighboring nodes y_i and y_j. To model deletions, a dummy node v_d is added to the network G as the label for deletions.

Formally, the similarity score is defined as follows:

$$S(x,y) = \begin{cases} \Delta_d & \text{if } y = v_d, \\ 0 & \text{if } R(x,y) \leq \Delta_d, \\ R(x,y) & \text{otherwise.} \end{cases}$$

where Δ_d ($\Delta_d > 0$) is a penalty parameter for deletions. The smaller Δ_d is, the heavier penalty for deletions. $R(x,y)$ is the raw similarity score of nodes x and y. For protein or gene networks, we set $R(x,y) = 1 - E_{x,y}$, where $E_{x,y}$ is the BLAST E-value of two proteins or genes x and y.

The connectivity score is defined as follows:

$$W(y_i,y_j) = \begin{cases} \Delta_c & \text{if } y_i = y_j = v_d, \\ \Delta_e & \text{if } y_i = v_d \text{ or } y_j = v_d, \\ \frac{1}{L(y_i,y_j)^{\Delta_s}} & \text{otherwise.} \end{cases}$$

where Δ_c ($0 < \Delta_c \leq 1$) is a penalty parameter for consecutive deletions, Δ_e ($0 < \Delta_e \leq 1$) is a penalty parameter for initial deletions, Δ_s ($\Delta_s \geq 0$) is a penalty parameter for insertions. Smaller Δ_e and Δ_c mean heavier penalty for initial and consecutive deletions respectively, while larger Δ_s indicates heavier penalty for insertions. $L(y_i,y_j)$ is the length of the shortest path between nodes y_i and y_j in the network G. If node y_j is not reachable from node y_i, $L(y_i,y_j)$ is set to infinity.

2.3 Network simplification

For large networks, the computation of CRF model can be greatly reduced by simplifying the network as follows. Given a query network X and a target network G. Firstly, the similarity scores $S(x,y)$ are calculated for all pairs of nodes x in X and y in G, and the connectivity scores $W(y_i,y_j)$ are calculated for all pairs of nodes y_i, y_j in G. A node y in G is called redundant if $S(x,y) \leq \Delta_d$ for all nodes x in X. All redundant nodes can be safely removed from G since they will never appear in the optimal solutions of CRF model.

3 RESULTS

The proposed CRF method is implemented by using Matlab. The Matlab toolbox UGM[1] is used to solve the CRF model, and the lengths of shortest paths for all pairs of nodes in the network G are calculated by the Matlab toolbox MatlabBGL[2].

Several simulated and real data are used to evaluate the CRF method on both undirected and directed network querying problems. In this study, we set $\Delta_d = 10^{-10}$, $\Delta_c = 0.5$, $\Delta_e = 1$, and $\Delta_s = 1$.

3.1 Querying undirected networks

Firstly, the CRF method is applied to query known pathways in PPI networks. Pathways and PPI networks are both undirected. For the sake of comparison, the PPI networks data of *Saccharomyces cerevisiae* and *Drosophila melanogaster* are obtained from Dost *et al.* (2008). The *S. cerevisiae* PPI network contains 15147 interactions among 4738 proteins, while the *D. melanogaster* PPI network contains 26201 interactions among 7481 proteins.

[1] http://www.di.ens.fr/~mschmidt/Software/UGM.html

[2] http://www.stanford.edu/~dgleich/programs/matlab_bgl/

3.1.1 Simulated data We randomly generate three types of query networks: simple paths, trees, and loopy subnetworks. For each type of query networks, we extract 30 instances, with sizes ranged from 7 to 12, from the *S. cerevisiae* PPI network. Each extracted subnetwork is perturbed with 2 node insertions and deletions. Meanwhile, we simulate 3 levels of mutations to every protein sequence, that is, randomly changing 50%, 60% and 70% residues respectively. More detail of the simulation is provided in Supplementary Materials.

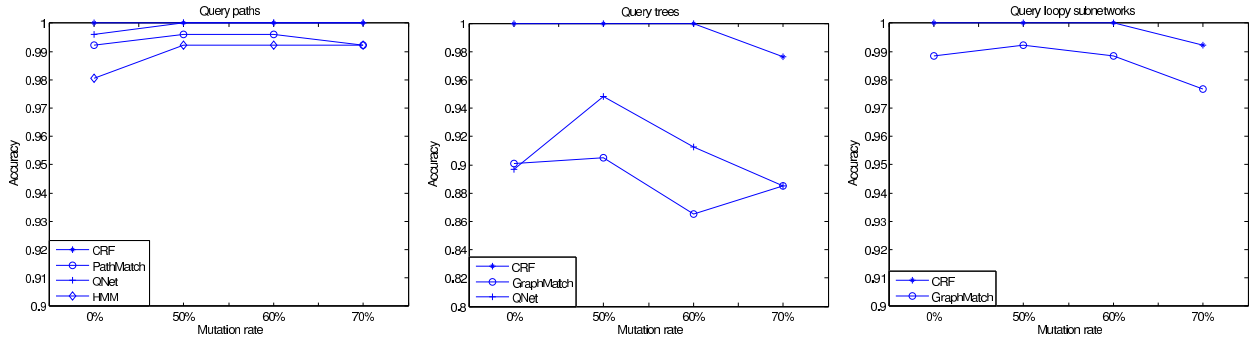
The simulated subnetworks are queried in the *S. cerevisiae* PPI network. To evaluate the querying methods, we compute the accuracy of nodes by comparing the querying results with the original extracted subnetworks except the insertion and deletion nodes. The results of CRF method are compared with HMM (Qian *et al.*, 2009), QNet (Dost *et al.*, 2008), PathMatch/GraphMatch (Yang and Sze, 2007), as shown in Figure 2. In these methods, HMM and PathMatch only can query the simple path, while QNet is designed for tree querying. As the extension of PathMatch, GraphMatch can query the general subnetworks, but needs longer computation time than CRF method. Specially, CRF method can find the exact solution for all instances with mutation level equal or smaller than 60% in seconds. The comparison of running time on the simulated paths are shown in Supplementary Materials. In a word, CRF method outperforms all other methods in terms of both the accuracy and the applicable query network types.

In order to evaluate the influence of node similarity to the results, we conduct additional experiments on the simulated data (The procedure and results are provided in Supplementary Materials). The results show that the CRF method well balances the node similarity information and the topological information, compared with the existing methods. The effects of parameters are also analyzed in Supplementary Materials.

3.1.2 Real data Three real pathways of different types are selected from KEGG database (Kanehisa and Goto, 2000) to evaluate the CRF method: simple path, tree, and loopy subnetwork.

The simple path instance is querying the human hedgehog signaling pathway (KEGG:hsa04340) in the *D. melanogaster* PPI network. The best matching result is shown in Figure 3(b). Compared with the *D. melanogaster* hedgehog signaling pathway (KEGG:dme04340), the human Fu is mismatched with PKA instead of Fu in *D. melanogaster*, although PKA is also in the *D. melanogaster* hedgehog signaling pathway. We found the reason is that the edge between Smo and Fu, which exists in the *D. melanogaster* hedgehog signaling pathway in the KEGG database, does not exist in the *D. melanogaster* PPI network used in this study. If we add the edge between Smo and Fu, the best matching result becomes the same as the *D. melanogaster* hedgehog signaling pathway as shown in Figure 3(c). The results of HMM (Qian *et al.*, 2009) and QPath (Shlomi *et al.*, 2006) are extracted from their papers respectively and shown in Supplementary Materials for comparison.

The tree instance is shown in Figure 4. The query is a classical human MAPK pathway involved in cell proliferation and differentiation (KEGG:hsa04010). In the best matching result found in the *D. melanogaster* PPI network, the proteins EGFR, Sev, Ras85D, Ph1, Dsor1, Rolled are the members of the known *D. melanogaster* MAPK pathway (KEGG:dme04013). The results of HMM (Qian *et al.*, 2009) and QNet (Dost *et al.*, 2008) are extracted from their papers respectively and shown in Supplementary Materials for comparison.



(a) paths (b) trees (c) loopy subnetworks

Fig. 2. Comparison of the querying results on simulated data. The y axis is the accuracy, i.e. the percentage of correctly matched nodes, and the x axis is the mutation level.

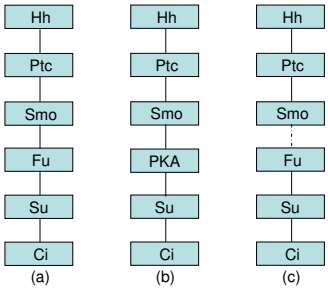


Fig. 3. (a) The human hedgehog signaling pathway and its best match in *D. melanogaster* (b) without and (c) with artificial edge between Smo and Fu.

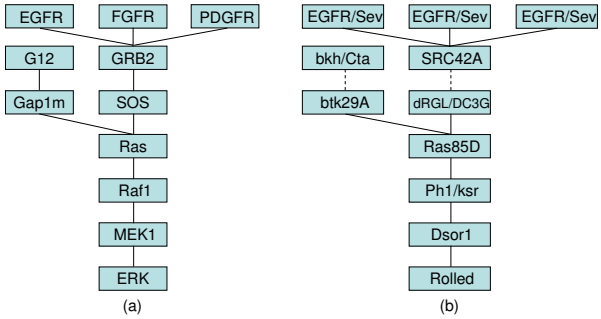


Fig. 4. Human MAPK signaling pathway and its best match in *D. melanogaster*. Solid lines indicate there is no insertion/deletion between the nodes, and dashed lines on the contrary. All gap nodes are omitted.

The loopy subnetwork instance is querying a part of the *D. melanogaster* MAPK signaling pathway (KEGG:dme04013) in the *S. cerevisiae* PPI network, as shown in Figure 5. The proteins in the best matching result, SLT2, FUS3, STE11, STE7, HOG1, KSS1 are all in the *S. cerevisiae* MAPK signaling pathway (KEGG:sce04011). The ABP1 is an actin-binding protein and associated with the cortical cytoskeleton of *S. cerevisiae*. Its SH3 region indicates that it might serve to bring together signal transduction proteins and their targets or regulators, or both, in the membrane cytoskeleton (Drubin *et al.*, 1990). Cdc25 is a membrane bound guanine nucleotide exchange factor. PTP3 is phosphotyrosine-specific protein phosphatase involved in the inactivation of MAPK during osmolarity sensing. RAS1 is GTPase involved in G-protein signaling in the adenylate cyclase activating pathway and has similar function with Ras85D and Ras.

Several additional experimental results on loopy subnetworks are provided in Supplementary Materials. All these results show that CRF method is very efficient to find the functionally similar pathways of different structures in real PPI networks.

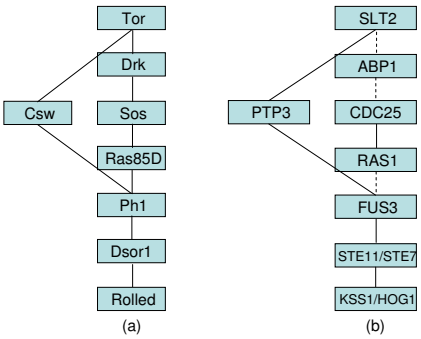


Fig. 5. *D. melanogaster* MAPK pathway and its best match in *S. cerevisiae*. Solid lines indicate there is no insertion and deletion between the nodes while dashed lines on the contrary. All gap nodes are omitted.

3.2 Querying directed networks

In the real world, there are many directed biomolecular networks in which the interactions between biomolecules are directional,

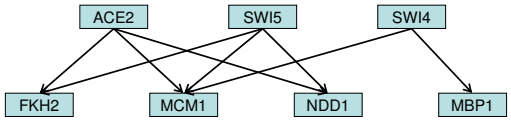


Fig. 6. A directed subnetwork query in the regulator-regulator interaction network. The best matching result is the same as query.

for example, gene regulatory networks and metabolic networks. However, only a few of existing network querying methods can successfully handle the directed networks. The proposed CRF method can deal with the directed network querying problem without modification. In this section, several simulated and real instances are shown to evaluate the capability of CRF method for directed networks.

3.2.1 Regulator-regulator interaction networks Figure 6 shows a simulated querying example of directed networks. The query is extracted from a regulator-regulator interaction network which contains 106 regulators (Lee *et al.*, 2002). For each node in the query, five nodes are randomly selected as fake sources. The similarity scores between the query node and its truth source as well as five fake sources are set as 1. The remaining similarity scores are random numbers in $[0\ 1]$. The best matching result of CRF method is identical with the query network.

3.2.2 Metabolic networks Figure 7 shows an instance that a tree-like network containing part of the α-aminoadipic pathway in *T. thermophilus* (Kobashi *et al.*, 1999) is queried in a combined *E. coli* metabolic network of glycolysis, gluconeogenesis, the citrate cycle and the glyoxylate metabolism. The raw similarity scores are defined following Yang and Sze (2007) but with simple normalization to the interval $[0\ 1]$. In detail, the raw similarity scores of enzymes are provided by the information content values based on the proximity of EC numbers, while the raw similarity scores of compounds are calculated by using SIMCOMP package from Hattori *et al.* (2003). The best matching result of CRF method is the same as the top result found by GraphMatch (Yang and Sze, 2007).

Another test of metabolic networks is following Pinter *et al.* (2005) and Li *et al.* (2007). The data of 113 *E. coli* pathways and 151 *S. cerevisiae* pathways are obtained from Pinter *et al.* (2005). In Pinter *et al.* (2005) and Li *et al.* (2007), they ran all-against-all network alignment and gave the results through the t-test and the statistical significance p-value. We transform the network alignment problem to the network querying problem as follows. Firstly, all the 151 *S. cerevisiae* pathways are combined as a whole metabolic network in which the same enzymes in different pathways are presented as different nodes, and the different pathways are not connected. There are totally 862 nodes in the constructed *S. cerevisiae* metabolic network. Secondly, each of the 113 *E. coli* pathways is queried in the constructed *S. cerevisiae* metabolic network. The node similarity is computed following Pinter *et al.* (2005) and Li *et al.* (2007). We filter the matching results which contain at most one mismatch. There are 44 *E. coli* pathways with at most one mismatch, in which 32 are completely matched. For comparison, by using the same procedure, GraphMatch found 42 matching results with at most one mismatch,

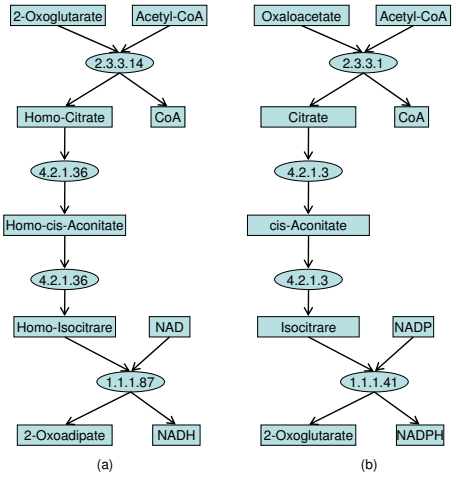


Fig. 7. (a) A tree-like pathway containing part of the α-aminoadipic pathway in *T. thermophilus* and (b) its best match in *E. coli*.

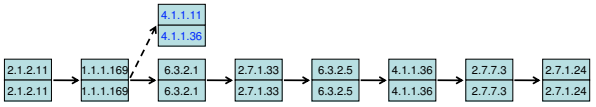


Fig. 8. A super pantothenate coenzymeA biosynth pathway in *E. coli* and the best matching pantothenate coenzymeA biosynth pathway in *S. ccerevisiae* with one mismatch. The solid edges are the matching interactions and the dashed edge is the mismatching edge.

in which 25 are completely matched and all included in the 32 completely matched results of CRF method. All the matching results of two methods can be found in Supplementary Materials. Figure 8 is an example of matching the *E. coli* super pantothenate coenzymeA biosynth pathway to *S. cerevisiae* pathway with a mismatch.

The running time of CRF method and GraphMatch on the metabolic network data from Pinter *et al.* (2005) are compared in Figure 9. CRF method gets the matching results in less than 20 seconds for all instances, while GraphMatch does not finish within 1000 seconds for the instances with query length larger than 20. The variance of running time of GraphMatch becomes very large when the size of network increases since the computational complexity of GraphMatch depends on the number of induced subgraphs. Additional comparison of the running time are also provided in Supplementary Materials.

4 CONCLUSION AND DISCUSSION

This paper presents a novel CRF method for network querying problem. A number of simulated and real experiments show that the CRF method is very efficient and helpful to find the conserved functionally similar modules and pathways in both undirected and directed biomolecular networks. Compared with the existing network querying approaches, the CRF method has several

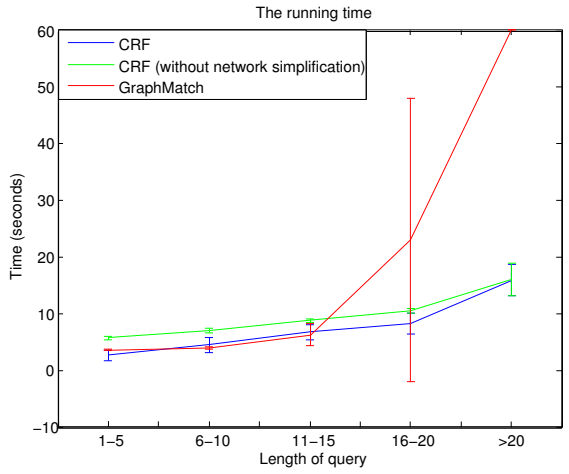


Fig. 9. Comparison on the running time of CRF and GraphMatch on the metabolic network data from Pinter *et al.* (2005).

advantages. Firstly, the CRF method allows unlimited insertions/deletions. Some existing methods can only handle a small number of insertions/deletions because the computational complexity of these methods depend on the maximum allowed number of insertions/deletions. In other words, if the maximum allowed number of insertions/deletions is too large, their models may become intractable within reasonable time. Conversely, the computation time of CRF method is independent of the maximum number of insertions/deletions. Therefore, the CRF method can discover better matching results.

Secondly, the CRF method can efficiently and accurately query networks with cycles. The existing methods can successfully deal with simple network structure such as linear chain and tree. However, most of them can not be straightforwardly extended to the networks with cycles. The CRF model can be exactly solved for linear chain and tree cases. For general networks, there are also many mature CRF algorithms in literature. Although these algorithms is not theoretically guaranteed optimal for the networks with cycles, extensive computational experiments on simulated and real data show the CRF method is practically efficient and accurate.

Thirdly, the CRF method addresses the directed network querying problem and the undirected network querying problem with the same model. Therefore, the results of directed networks are naturally consistent with that of undirected networks.

Finally, the CRF method can easily incorporate additional information as a result of the flexibility the CRF model possesses. New information can be integrated into the node and edge feature functions as well as new feature functions. For example, the BLAST E-value used for calculating the node similarity of protein or gene networks can be substituted by other measurement such as the functional similarity defined by common Gene Ontology terms (Ashburner *et al.*, 2000). The edge feature function can be defined by using other network properties instead of the shortest path. Additional feature functions also can be introduced to emphasize concrete biological implications, which will be one of the major research directions in the future.

The network querying problem is closely related to the graph matching problem. CRF and the related methods such as Markov random fields (MRF) have been used extensively in the non-bioinformatics graph matching literature (Caetano *et al.*, 2004; Caelli and Caetano, 2005; Torresani *et al.*, 2008; Bayati *et al.*, 2009). Banks *et al.* (2008) addressed another kind of network querying problem, in which a network schema (e.g. desired topology, types of nodes and interactions) instead of a concrete network is queried in the target networks. Some ideas in these works may be useful to improve the proposed method in this paper, which will be another major research work in the future.

ACKNOWLEDGEMENT

Funding: This work is supported by the National Natural Science Foundation of China (Grant No. 60970091 and 60873205), and Knowledge Innovation Program of the Chinese Academy of Sciences (Grant No. kjcx-yw-s7).

Conflict of interest statement. None declared.

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, **25**(1), 25–29.

Banks, E., Nabieva, E., Peterson, R., and Singh, M. (2008). Netgrep: fast network schema searches in interactomes. *Genome Biol*, **9**(9), R138.

Bayati, M., Gerritsen, M., Gleich, D. F., Saberi, A., and Wang, Y. (2009). Algorithms for large, sparse network alignment problems. In *Proc. Ninth IEEE Int. Conf. Data Mining ICDM '09*, pages 705–710.

Blin, G., Sikora, F., and Vialette, S. (2010). Querying graphs in protein-protein interactions networks using feedback vertex set. *IEEE/ACM Trans Comput Biol Bioinform*, **7**(4), 628–635.

Bruckner, S., Hüffner, F., Karp, R. M., Shamir, R., and Sharan, R. (2009). TORQUE: topology-free querying of protein interaction networks. *Nucleic Acids Res*, **37**(Web Server issue), W106–W108.

Caelli, T. and Caetano, T. (2005). Graphical models for graph matching: Approximate models and optimal algorithms. *Pattern Recognition Letters*, **26**(3), 339–346.

Caetano, T. S., Caelli, T., and Barone, D. A. C. (2004). Graphical models for graph matching. In *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition CVPR 2004*, volume 2.

DeCaprio, D., Vinson, J. P., Pearson, M. D., Montgomery, P., Doherty, M., and Galagan, J. E. (2007). Conrad: gene prediction using conditional random fields. *Genome Res*, **17**(9), 1389–1398.

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

Drubin, D. G., Mulholland, J., Zhu, Z. M., and Botstein, D. (1990). Homology of a yeast actin-binding protein to signal transduction

proteins and myosin-i. *Nature*, **343**(6255), 288–290.

Hattori, M., Okuno, Y., Goto, S., and Kanehisa, M. (2003). Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J Am Chem Soc*, **125**(39), 11853–11865.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, **28**(1), 27–30.

Kelley, B. P., Yuan, B., Lewitter, F., Sharan, R., Stockwell, B. R., and Ideker, T. (2004). PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Res*, **32**(Web Server issue), W83–W88.

Kobashi, N., Nishiyama, M., and Tanokura, M. (1999). Aspartate kinase-independent lysine synthesis in an extremely thermophilic bacterium, thermus thermophilus: lysine is synthesized via alpha-aminoadipic acid not via diaminopimelic acid. *J Bacteriol*, **181**(6), 1713–1718.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 282–289.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J.-B., Volkert, T. L., Fraenkel, E., Gifford, D. K., and Young, R. A. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**(5594), 799–804.

Li, Z., Zhang, S., Wang, Y., Zhang, X.-S., and Chen, L. (2007). Alignment of molecular networks by integer quadratic programming. *Bioinformatics*, **23**(13), 1631–1639.

Pinter, R. Y., Rokhlenko, O., Yeger-Lotem, E., and Ziv-Ukelson, M. (2005). Alignment of metabolic pathways. *Bioinformatics*, **21**(16), 3401–3408.

Qian, X., Sze, S.-H., and Yoon, B.-J. (2009). Querying pathways in protein interaction networks based on hidden Markov models. *J Comput Biol*, **16**(2), 145–157.

Sha, F. and Pereira, F. (2003). Shallow parsing with conditional random fields. In *Proceedings of Human Language Technology/North American chapter of the Association for Computational Linguistics annual meeting (HLT/NAACL)*.

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

Tian, Y., McEachin, R. C., Santos, C., States, D. J., and Patel, J. M. (2007). SAGA: a subgraph matching tool for biological graphs. *Bioinformatics*, **23**(2), 232–239.

Torresani, L., Kolmogorov, V., and Rother, C. (2008). Feature correspondence via graph matching: Models and global optimization. *Computer Vision–ECCV 2008*, pages 596–609.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, **13**(2), 260–269.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, **1**(1–2), 1–305.

Wu, L.-Y., Zhou, X., Li, F., Yang, X., Chang, C.-C., and Wong, S. T. C. (2009). Conditional random pattern algorithm for LOH inference and segmentation. *Bioinformatics*, **25**(1), 61–67.

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. (2000). DIP: the database of interacting proteins. *Nucleic Acids Res*, **28**(1), 289–291.

Yang, Q. and Sze, S.-H. (2007). Path matching and graph matching in biological networks. *J Comput Biol*, **14**(1), 56–67.