Paper Title

A. B. AUTHOR* and C. D. AUTHOR
University Department, University Name,
City, State ZIP/Zone, Country
**E-mail: ab_author@university.com*
www.university_name.edu

A. N. AUTHOR
Group, Laboratory, Street,
City, State ZIP/Zone, Country
E-mail: an_author@laboratory.com

Here should come the abstract.

Keywords: keyword 1; keyword 2; keyword 3;

1. Introduction

Analysis of protein-protein interactions has a growing potential of providing a better understanding about collective protein function and cellular machinery. Such interactions can be modeled as a protein interaction network; that is a weighted graph where each node represents a protein and each edge represents an interaction between a pair of proteins, with a weight representing the level of confidence that this interaction truly exists. Such a structure allows easier extraction of hidden information due to the analogy with well-known graph problems.

An important aspect of the analysis of protein interaction networks is detecting signaling pathways. A signaling pathway is a series of proteins in which each protein signals its successor to transmit some biological information through their interaction. Signaling pathways can be viewed as simple paths in protein interaction networks.[1]

Given a confidence value for each interaction, we can assign an overall confidence value for the existence of a pathway simply by multiplying the confidence values of its constituting interactions. An optimal pathway is one with the highest confidence value. To work in an additive framework instead of the multiplicative one, edge weights can be assigned logarithmic values of the original interaction confidence.

Problem Consider a set V of nodes denoting proteins, and a probability value $p(u, v)$ denoting interaction confidence for each $u, v \in V$. A set of undirected weighted edges E can be obtained by adding an edge for each pair (u, v) with $p(u, v) > 0$, and assigning it a weight $w(p, v) = -\log[p(u, v)]$. Now consider the undirected weighted graph $G = (V, E, w)$ representing the protein interaction network, a set of start nodes $S \subset V$ and a set of end nodes $T \subset V$. We want to find the minimum-weight simple path of length m, starting at any node $s \in S$ and ending at any node $t \in T$.

Scott et al.[2] mentioned that the traveling-salesman problem is polynomial-time reducible to this problem; therefore it is NP-hard. They presented a method for detecting pathways

using a generic technique devised by Alon et al.[3] called *color-coding*. The basic idea of this method is to randomly assign each node in the graph one of m different colors, and search for an optimal pathway in the restricted domain of colorful pathways. A pathway is considered colorful if and only if all of its nodes are different in color from each other. This process is repeated for several iterations until reaching a given level of confidence that the unknown optimal pathway was among the colorful ones at some instance. This confidence level builds up with each iteration by calculating a probability value that an optimal path is indeed colorful in this iteration. This *success probability* value depends solely on the pathway length m and doesnt capitalize on available information like the network topology and color assignment. As a result, the method provides a theoretically correct but very conservative success probability value, and hence loses a very good potential for decreasing the number of iterations needed to accomplish a given confidence level.

Gülsoy et al.[4] presented an enhanced color-coding technique called *k-hop coloring*. In essence, k-hop coloring makes use of knowing the network topology and the node colors to assign the network a maximal value k such that the network is k-hop colorable. A colored network is k-hop colorable if the shortest path between any pair of same-color nodes is more than k hops in length. This additional piece of information allows for higher success probability in each iteration, with a higher k value resulting in a higher success probability. However, an obvious limitation is that the subnetworks with higher level of connectivity diminishe the k value assigned to the whole network. For example, a network containing a clique of size m cannot be colored with $(m-1)$-hop coloring using m colors.[4]

Contribution Our motivation comes from the need for a deeper understanding of the relation between network topology, random color assignment and success probability. We study the possibility of assigning k values to nodes on an individual basis instead of a single k value for the whole network. We also study how this reflects on the resulting success probability for each iteration. We examine the idea that a pathway whose nodes are assigned different k values should result in a higher success probability than if we only consider the minimum of these k values for all nodes. Given different k values for each node on a pathway, we show how to obtain a better bound on success probability.

Based on these findings, We present a new method for detecting signaling pathways in protein interaction networks using an enhanced k-hop coloring technique. For a required optimal pathway of length m, we start by assigning one of m colors to each node in the graph, then we extract the optimal colorful pathway. We then calculate our new bound on success probability. We repeat this process until the cumulative success probability is at least equal to a given confidence level. Although our theoretical findings are based on assuming the knowledge of the k values assigned to the unknown global optimal pathway, we empirically show that the local optimal pathway extracted from the domain of colorful pathways actually fits the purpose.

We provide validation experiments to test the biological significance of our results. We use *weight p-value* and *functional enrichment* as validation measures. We also compare the performance of our method against the one presented by Scott et al.[2] with respect to how fast our method reaches a given confidence level as opposed to theirs.

The rest of the paper is organized as follows. Section 2 discusses the background and related work. Section 3 explains how to obtain a tighter bound on success probability and describes our enhanced k-hop coloring method. Section 4 shows the experiments performed and their results. Section 5 is the conclusion of the paper.

2. Background

Some different but closely related problems have been studied in the literature. Zhao et al.[5] used integer linear programming to find signaling networks in protein interaction networks. They formulated the problem as a linear optimization problem of finding maximum weight subnetwork with a given size. This approach is concerned with finding signaling subnetworks in their general form rather than linear pathways. Kelley et al.[1] detected conserved signaling pathways between related organisms by performing global alignment between their protein interaction networks. They scored each pathway in terms of probability of true homology between aligned pair of proteins, as well as probability of true interactions between pairs of proteins along the pathway. This approach detects conserved pathways only and requires coupling between two datasets. Shlomi et al.[6] introduced QPath, a method for querying protein interaction networks for pathways using known homologous pathways as queries. They scored results based on their similarity to the query, number of insertions and deletions used, as well as the reliability of their interactions. This method only detects pathways that are similar to a given one.

On the other hand, the problem we address has also been studied in the literature. Lu et al.[7] presented a divide-and-conquer algorithm for finding generic pathway structures in protein interaction networks. They recursively partitioned the network into two sets of vertices, enumerated substructures present in each set, and then built larger structures from them. They assumed that all edges have the same weight and scored the resulting pathway structures based on the biological function relatedness of their nodes to the given source and destination nodes. We are more interested in detecting pathways based on confidence in interactions rather than similarity of proteins.

Steffen et al.[8] studied detecting signaling pathways in protein interaction networks as guided by expression data. They listed all pathway candidates in a protein interaction network using exhaustive search. They scored each candidate based on how similar the expression profiles of its genes are. Bebek et al.[9] presented a method called PathFinder for finding new signaling pathways using association rules of known ones. They started with mining association rules for known pathways, guided by the knowledge of functional annotations of their proteins. They then performed an exhaustive search for candidate pathways. From these candidates, they selected the ones having at least a certain number of the known association rules and an average interaction weight above a given threshold. The drawback of both of these methods is that the time complexity of exhaustive graph search is exponential in terms of the network size, and hence is very inefficient.

Gitter et al.[10] presented a method for discovering signaling pathways by adding edge orientation to protein interaction networks. They selected an optimal orientation of all edges in the network that maximizes the weights of all satisfied length-bound paths. A path is satisfied

if it follows the same direction along its edges from a source node to a destination node. They proved that this problem is NP-hard. They provided two approximation algorithms for it based on available solution methods for weighted Boolean satisfiability, and a third algorithm based on probabilistic selection. As shown in their results, these methods do not scale well with increasing the number of source and destination nodes and the required path length.

The present work builds on the method presented by Scott et al.[2] for detecting signaling pathways in protein interaction networks using color coding. It also enhances the model presented by Gülsoy et al.[4] for topology-aware color coding. We indroduced both methods in section 1.

3. Methods

3.1. *Term Definition and Problem Formulation*

Here we define the terms we commonly use like: coloring instance, k-value configuration, ..etc. Then we accurately define the problem of finding signaling pathways in PPI networks.

3.2. *Success Probability: a Tighter Bound*

Here we should explain the calculation of success probability, how the number of possible colorings for the optimal path is its key factor, and how obtaining a tighter (smaller) bound on it results in a tighter bound on success probability, and hence a better result. We then explain our notion of k-value configuration and how we calculate this bound from it. We explain the lattice structure and the subset relation between k-value configurations.

3.3. *Method: Enhanced k-hop Coloring*

Here we detail our method. We start by asserting that we have no knowledge about the optimal path, but we use the local optimal as a replacement and experimentally test the correctness of this approach. Then we explain the algorithm in detail.

4. Experiments

4.1. *Datasets*

Here we list the datasets used in experiments. I think we can use the MINT datasets for multiple organisms.

4.2. *Validation Experiments*

Here we list the validation experiments we did and their results. I think we should do some validation experiments similar to those done in Sharan's paper, using weight p-value and functional enrichment as validitiy measures.

4.2.1. *Validation using Weight p-value*

For each dataset we use, we should obtain the 99 percent confidence optimal pathway and compare it with optimal pathways obtained we obtain from random networks. We generate

random networks by shuffling edges of the original network. The weight p-value is the percentage of cases where the algorithm produces a more optimal pathway when run on one of these random networks.

4.2.2. *Validation using Functional Enrichment*

For each dataset, we we obtain the 99 percent confidence optimal pathway and test its functional enrichment. For each GO term appearing on the dataset proteins, we count the total number of proteins annotated by it and the number of proteins in the resulting pathway annotated by it. We use these numbers, along with the total number of proteins and the number of proteins in the pathway, as parameters for a hypergeometric test (I still have to develop further understanding about the details of this test). The maximum enrichment value for any of the tested GO terms gives us the final functional enrichment p-value.

4.3. *Comparison with Sharan*

This is just a temporary title for this subsection, I'm not very sure what to name it.

We measure the time and number of iterations needed by our method to obtain 70%, 90% and 99% confidence pathways of lengths 6, 7, 8 and 9 nodes. We compare these numbers against the ones by Sharan's method for the same cases.

We run our method for 500 iterations and measure the incremental success probabilty against iteration number. We do this experiment many times take the average curve. We do the same experiment using Sharan's method and obtain a second curve. We also measure the average practical success probability, which is the observed probability that the DP algorithm finds the optimal solution in a certain iteration or before it. We compare the three curves targetting two conclusions: (1) our method is experimentally solid because our calculated success probabiilities are lower than the observed ones; and (2) our method outperforms Sharan's method.

5. Conclusion

Here goes the conclusion.

References

1. B. P. Kelley, R. Sharan, R. M. Karp, T. Sittler, D. E. Root, B. R. Stockwell and T. Ideker, *Proceedings of the National Academy of Sciences* **100**, 11394 (September 2003).
2. J. Scott, T. Ideker, R. M. Karp and R. Sharan, Efficient algorithms for detecting signaling pathways in protein interaction networks, in *Proceedings of the 9th Annual international conference on Research in Computational Molecular Biology*, RECOMB'05 (Springer-Verlag, Berlin, Heidelberg, 2005).
3. N. Alon, R. Yuster and U. Zwick, *J. ACM* , 844 (1995).
4. G. Gülsoy, B. Gandhi and T. Kahveci, Topology aware coloring of gene regulatory networks, in *Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, BCB '11 (ACM, New York, NY, USA, 2011).

5. X.-M. Zhao, R.-S. Wang, L. Chen and K. Aihara, *Nucleic Acids Research* **36**, p. e48 (2008).
6. T. Shlomi, D. Segal, E. Ruppin and R. Sharan, *BMC Bioinformatics* **7**, p. 199 (2006).
7. S. Lu, F. Zhang, J. Chen and S.-H. Sze, *Algorithmica* **48**, 363 (August 2007).
8. M. Steffen, A. Petti, J. Aach, P. D'haeseleer and G. Church, *BMC Bioinformatics* **3**, p. 34 (2002).
9. G. Bebek and J. Yang, *BMC Bioinformatics* **8**, p. 335 (2007).
10. A. Gitter, J. Klein-Seetharaman, A. Gupta and Z. Bar-Joseph, *Nucleic Acids Research* **39**, p. e22 (2011).