RESQUE: Network Reduction Using Semi-Markov Random Walk Scores for Efficient Querying of Biological Networks (Extended Abstract)

Sayed Mohammad Ebrahim Sahraeian and Byung-Jun Yoon

Department of Electrical and Computer Engineering,
Texas A&M University, College Station, TX 77843, USA,
msahraeian@tamu.edu, bjyoon@ece.tamu.edu

Abstract. In this work, we present RESQUE, an efficient algorithm for querying large-scale biological networks. The algorithm uses a semi-Markov random walk model to estimate the correspondence scores between nodes across different networks. The target network is iteratively reduced based on the node correspondence scores, which are also iteratively re-estimated for improved accuracy, until the best matching subnetwork emerges. The proposed network querying scheme is computationally efficient, can handle any network query with arbitrary topology, and yields accurate querying results.

1 Introduction

The increasing availability of large-scale biological networks, including protein-protein interaction (PPI) networks, metabolic networks, and co-expression networks, asks for effective tools that can be used to analyze these network data and gain useful biological insights. Network querying is one such example, whose goal is to identify subnetwork regions in a large target network that are similar to a given query network [3]. For instance, a network querying algorithm can be used to search for novel potential pathways in a given biological network that are similar to known pathways, thereby enabling knowledge transfer from well-studied species to less-studied ones. The optimal network querying problem has been shown to be NP-complete, by reduction to the graph isomorphism problem [2], and various approaches have been proposed so far to make it computationally feasible [2, 4, 1]. Despite recent advances in the field, most of the currently existing algorithms still have a number of practical limitations. For example, most algorithms cannot handle network queries with arbitrary topology, and the structure of the query network is often limited to linear paths or trees [2]. Or alternatively, the query network is simply treated as a collection of nodes by ignoring the network topology [1]. Moreover, many existing algorithms suffer from high computational complexity, which often increases exponentially with the query size, making them unsuitable for performing large queries.

In this work, we propose a novel network querying algorithm called RESQUE (**RE**duction-based scheme using **S**emi-Markov scores for network **QUE**rying) that can effectively address the aforementioned issues. RESQUE takes a probabilistic approach for fast and accurate network querying, which iteratively reduces the target network based on the so-called node correspondence scores computed via a semi-Markov random walk model. These scores provide a probabilistic measure of the similarities between nodes that belong to different networks (i.e., the query and the target networks) and can be efficiently computed by a closed-form formula. At each iteration, the estimated node correspondence scores are used to shrink the search space in the target network by removing the nodes that have minimal correspondence to the query nodes. The node correspondence scores are then re-estimated based on the reduced network, and the network reduction process is repeated until the target network has been sufficiently reduced. The iterative re-estimation of the correspondence scores leads to better querying results with higher expected accuracy. After the iterative reduction process, the final querying result is obtained using two different strategies. The first strategy, called RESQUE-M, uses the maximum weighted matching algorithm to find the best matching subnetwork that maximizes the expected accuracy. The second strategy, called RESQUE-C, finds the largest connected subnetwork in the reduced target network. RESQUE does not restrict the topology of the query network, and it can handle linear paths, trees, and general structures with loops. Furthermore, the query can be connected, partially connected, or simply a collection of nodes with unknown network topology.

To assess the performance of RESQUE, we carried out network queries of 1,184 protein complexes of size $4 \sim 25$ in the PPI networks of fly, yeast, and human, which are the three largest PPI networks that are currently available. We compared our algorithm with Torque [1], a state-of-the-art network querying algorithm, which has been shown to outperform many existing algorithms. Our results show that both RESQUE-M and RESQUE-C yield significantly larger number of hits compared to Torque, with higher functional coherence rates at comparable specificity levels. The computational complexity of RESQUE is polynomial in terms of the query size, and it is able to complete each query within a few seconds for all complexes considered in our experiments, while Torque may need more than an hour for some of them. Further performance assessment based on simulated data shows that RESQUE is also highly robust against distortions in the node similarity score as well as random node deletions and insertions.

References

1. Bruckner, S., Huffner, F., Karp, R.M., Shamir, R., Sharan, R.: Topology-free querying of protein interaction networks. J. Comput. Biol. 17, 237–252 (Mar 2010)
2. Dost, B., Shlomi, T., Gupta, N., Ruppin, E., Bafna, V., Sharan, R.: QNet: A tool for querying protein interaction networks. J Comput Biol 15(7), 913–925 (2008)
3. Sharan, R., Ideker, T.: Modeling cellular machinery through biological network comparison. Nat. Biotechnol. 24, 427–433 (Apr 2006)
4. Yang, Q., Sze, S.: Path matching and graph matching in biological networks. J Comput Biol 14, 56–67 (2007)