W106–W108 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* *Published online 2 June 2009*
doi:10.1093/nar/gkp474

Torque: topology-free querying of protein interaction networks

Sharon Bruckner[1,*], Falk Hüffner[1], Richard M. Karp[2], Ron Shamir[1] and Roded Sharan[1,*]

[1]The Blavatnik School of Computer Science, Tel Aviv University, 69978 Tel Aviv, Israel and [2]International Computer Science Institute, 1947 Center St., Berkeley, CA 94704, USA

Received January 27, 2009; Revised April 24, 2009; Accepted May 17, 2009

ABSTRACT

Torque is a tool for cross-species querying of protein–protein interaction networks. It aims to answer the following question: given a set of proteins constituting a known complex or a pathway in one species, can a similar complex or pathway be found in the protein network of another species? To this end, Torque seeks a matching set of proteins that are sequence similar to the query proteins and span a connected region of the target network, while allowing for both insertions and deletions. Unlike existing approaches, Torque does not require knowledge of the interconnections among the query proteins. It can handle large queries of up to 25 proteins. The Torque web server is freely available for use at http://www.cs.tau.ac.il/~bnet/torque.html.

INTRODUCTION

In a *network querying* problem, one is given a small network, corresponding to a known pathway or a complex of interest. The goal is to identify similar instances in a large network, where similarity is measured in terms of sequence or interaction patterns. The resulting matches constitute possible protein complexes in queried species.

Previous approaches to the query problem required precise information on the interaction pattern of the query and were usually limited to small queries (2–4 proteins). PathBLAST—a server for querying linear pathways within a protein–protein interaction (PPI) network (1)—was subsequently extended to allow searching for more general structures (2). A general framework for subnetwork querying was developed (3), but it is applicable only to very small queries due to its complexity. Two other methods are NetMatch, a Cytoscape (4) plugin implementing the work of Ferro *et al*. (5) that utilizes fast heuristics for subgraph isomorphism to identify approximate matches of queries within a collection of networks, and NetGrep (6), a system for searching networks for patterns corresponding to small sets of proteins with specified attributes and topology.

The Torque server implements a novel method for querying protein networks that does not require information on the interconnections (topology) among the query proteins (7) (Figure 1). This makes Torque applicable in broader scenarios, such as querying complexes or pathways whose topologies are not completely known, or even when querying from species for which PPI information is not available. Lacroix *et al*. (8) also studied queries with no topology information, but since their method is enumerative it was applied to very small queries (2–4 proteins). In contrast, Torque can currently support queries of up to 25 proteins. It was tested extensively on hundreds of queries of known complexes from a variety of species (7). It was shown to yield far more matches than the QNet topology-based approach (2), while providing results that are highly functionally coherent.

IMPLEMENTATION AND FEATURES

The Torque web server implements the algorithms in (7) for querying protein sets across species. It combines three approaches: a dynamic programming method utilizing color coding, integer linear programming and a fast heuristic based on shortest paths. Torque automatically selects the best method to apply at each stage and outputs the highest scoring match. Scores are based on the underlying network structure, on interaction confidence values and on sequence similarities between matching proteins. The matching process is flexible, allowing a few insertions and deletions if needed. The server currently supports queries of size 4–25 (Figure 1).

Input

The input for Torque consists of:

(i) a query set of proteins in species A;
(ii) their protein sequences;

*To whom correspondence should be addressed. Tel: +972 3 640 7139; Fax: +972 3 640 9357; Email: bruckner@tau.ac.il
Correspondence may also be addressed to Roded Sharan. Email: roded@tau.ac.il

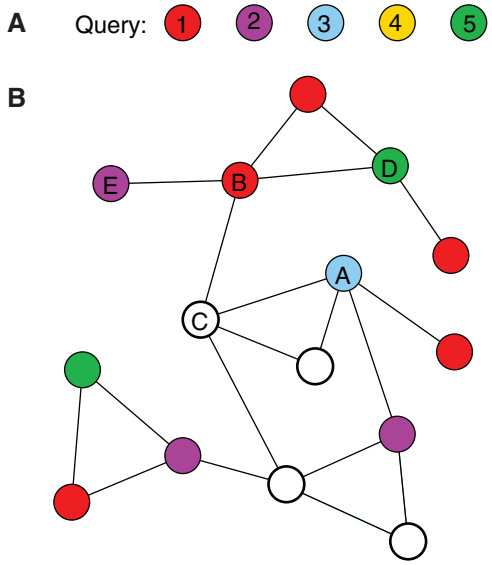

Figure 1. An example of a Torque query. (**A**) The query proteins; (**B**) the queried network. Colored vertices in the network match nodes of the same color in the query. Non-colored vertices do not match any query elements. The network induced by the vertices labeled *A*,*B*,*C*,*D* and *E* is a match to the query, where vertex C is an insertion and query element 4 is deleted.

(iii) a PPI network for species B;
(iv) the sequences of the network proteins.

All inputs are in simple text format:

- the query set can be entered directly as a comma-delimited or whitespace-delimited list.
- Protein sequences are given in the standard FASTA format.
- The PPI network is given as a text file, where each row represents an interaction and contains the IDs of the interacting pair and a confidence value for it in the range [0, 1].

It is possible to use a single FASTA file (input 2) for many queries, if it contains the sequences for all proteins in all queries. When the query field is left blank, TORQUE will use all the proteins in input 2 as the query. If input 1 contains Uniprot protein IDs (www.uniprot.org), their sequences need not be entered in input 2; instead, TORQUE automatically retrieves them from the Uniprot database. For several target species, the user need not provide inputs 3 and 4. Currently, the server supports this option for the three target species *Saccharomyces cerevisiae*, *Drosophila melanogaster* and *Homo sapiens*. The user can indicate one of these target species instead of providing inputs 3 and 4. Details on how these networks were constructed can be found in (7).

The user can control two parameters of the algorithm, setting a trade-off between speed and sensitivity. First, the user can control the threshold for sequence similarities. TORQUE applies BLAST to find putative matches between query and target proteins. The user can set the threshold for BLAST similarity (*E*-value). By setting a lower threshold, less homologs will be identified for the query proteins, making the algorithm faster but less sensitive. The second parameter is a threshold for the confidence values of PPI network edges provided as part of input 3. Edges whose confidence value is lower than the threshold are discarded; hence, this parameter determines the sparsity of the target network and affects the number of possible matches and the running time.

Processing

The running time of TORQUE is typically a few minutes, but may be up to an hour, depending on the size and other properties of the query (for more details, see (7)). If several queries are submitted to the server at the same time, they are queued and executed sequentially. Rather than waiting for the results online, they can be accessed later, in two ways:

- when a TORQUE job is started, it is assigned a nine-digit *job ID*. This ID can be used to access the results later from the main TORQUE page.
- Before submitting a query, the user may enter an email address. Once TORQUE has finished processing the query, the results will be sent to the email address provided. This process is done automatically and the email address is then discarded.

Output

The web server generates a web page (Figure 2) with the image of the top-scoring match for the query in the target network, as well as an auxiliary file in .sif format that can be viewed using the Cytoscape software (4). The content of the .sif file includes, for each edge, a numerical value representing the confidence in the interaction it represents, as provided in the input. This value determines the thickness of the edge in the Cytoscape visualization. The image shows the subgraph induced by the top-scoring match in the PPI network. Each vertex is labeled with its protein name in the PPI network and its matching query protein, if such exists. Insertion vertices are shown in gray, and proteins from the query for which there was no match in the solution (deletions) are listed separately.

A sample run

The following example uses as query the mouse DNA synthesome complex, downloaded from the CORUM website (http://mips.gsf.de/genre/proj/corum/index.html). This 13-member complex was queried in the yeast network (5430 proteins, 39 936 interactions). The result of the TORQUE run is shown in Figure 2. Examining the subnetwork identified by TORQUE we find that it is functionally coherent (nucleotidyltransferase activity, $P < 1.4E-10$) and significantly intersects the yeast alpha DNA polymerase: primase complex, supporting its biological plausibility. This example can be run from the TORQUE main page by checking 'Use example data'.

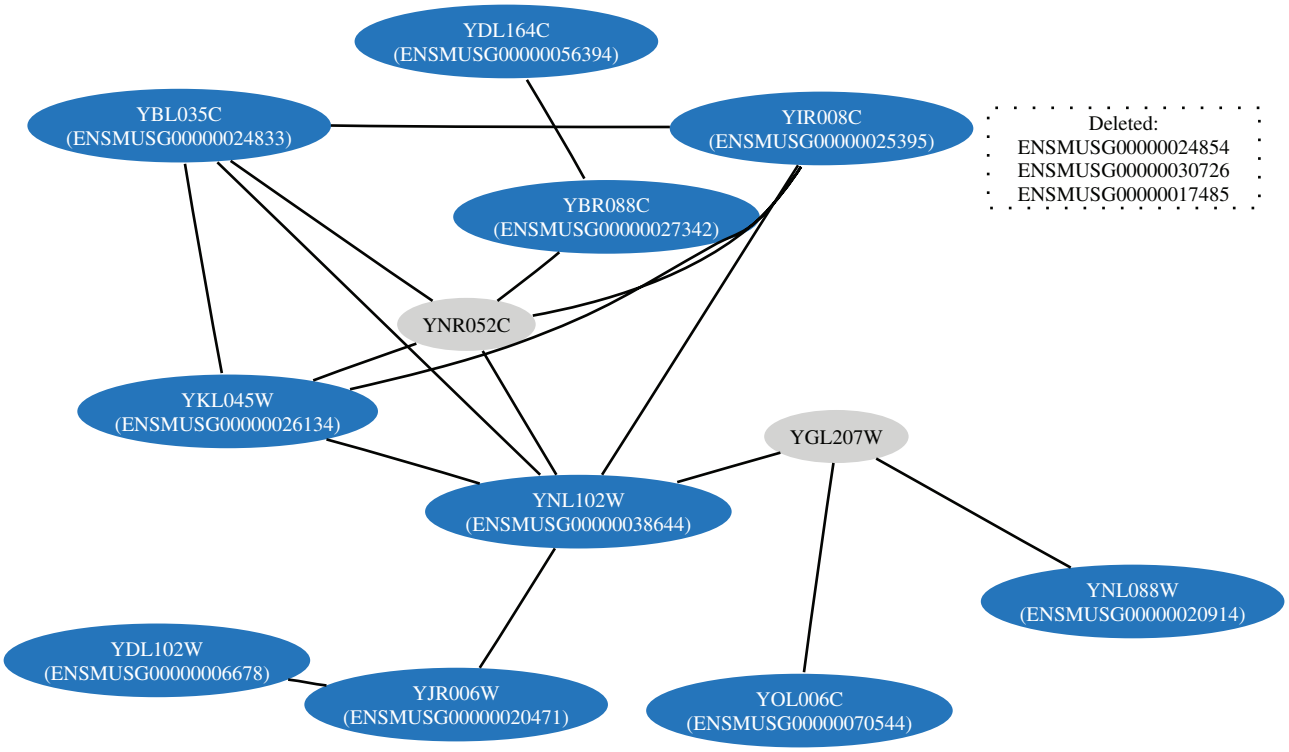

Figure 2. An example of the output of a Torque run. The query consists of 13 proteins. The match has 12 proteins with two insertions and three deletions.

SUMMARY

The Torque web server allows users to run topology-free queries on predefined or user-provided target networks. The result is a subnetwork of the target network most similar to the query, and is presented both graphically and as a downloadable text file.

ACKNOWLEDGEMENTS

We thank Maxim Kalaev for providing us with infrastructure code for the web server.

FUNDING

Israel Science Foundation (Grant no. 385/06 to R.S. and R.S.); German-Israel Foundation grant (to R.S.); postdoctoral fellowship from the Edmond J. Safra Bioinformatics Program at Tel Aviv University (to F.H.). Funding for open access charge: Israel Science Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. 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**, W83–W88.
2. Sharan,R., Dost,B., Shlomi,T., Gupta,N., Ruppin,E. and Bafna,V. (2008) QNet: a tool for querying protein interaction networks. *J. Comput. Biol.*, **15**, 913–925.
3. Sohler,F. and Zimmer,R. (2005) Identifying active transcription factors and kinases from expression data using pathway queries. *Bioinformatics*, **21(Suppl. 2)**, ii115–ii122.
4. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
5. Ferro,A., Giugno,R., Mongiovì,M., Pulvirenti,A., Skripin,D. and Shasha,D. (2008) GraphFind: enhancing graph searching by low support data mining techniques. *BMC Bioinformatics*, **9(Suppl. 4)**, S10.
6. Banks,E., Nabieva,E., Peterson,R. and Singh,M. (2008) NetGrep: fast network schema searches in interactomes. *Genome Biol.*, **9**, R138.
7. Bruckner,S., Hüffner,F., Karp,Richard M., Shamir,R. and Sharan,R. (2009) Topology-free querying of protein interaction networks. In *Proceedings of 13th RECOMB*. Vol. 5541 of *Lecture Notes in Bioinformatics*. Springer-Verlag, Berlin Heidelberg.
8. Lacroix,V., Fernandes,C.G. and Sagot,M (2006) Motif search in graphs: application to metabolic networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **3**, 360–368.