BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 4 2008, pages 594–596
doi:10.1093/bioinformatics/btm630

Systems biology

NetworkBLAST: comparative analysis of protein networks

Maxim Kalaev[1,*], Mike Smoot[2], Trey Ideker[2] and Roded Sharan[1,*]

[1]School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel and [2]Department of Bioengineering, UC San Diego, La Jolla, CA 92093, USA

Received on July 22, 2007; revised on October 29, 2007; accepted on December 18, 2007

Advance Access publication January 2, 2008

Associate Editor: Limsoon Wong

ABSTRACT

Summary: The identification of protein complexes is a fundamental challenge in interpreting protein–protein interaction data. Cross-species analysis allows coping with the high levels of noise that are typical to these data. The NetworkBLAST web-server provides a platform for identifying protein complexes in protein-protein interaction networks. It can analyze a single network or two networks from different species. In the latter case, NetworkBLAST outputs a set of putative complexes that are evolutionarily conserved across the two networks.

Availability: NetworkBLAST is available as web-server at: www.cs.tau.ac.il/~roded/networkblast.htm

Contact: kalaevma@post.tau.ac.il; roded@post.tau.ac.il

1 INTRODUCTION

Recent progress in high-throughput technologies such as the two-hybrid system (Fields, 2005) and co-immunoprecipitation assays (Aebersold and Mann, 2003) have generated large protein–protein interaction (PPI) networks for multiple species. The increasing availability of such data underscores the importance of organizing it into models of cellular signaling and regulatory machinery.

Similar to other applications in biology, an approach based on cross-species comparison may provide a valuable framework for addressing this challenge. By comparing networks drawn from different species it is possible to reduce measurement noise and to reinforce the common signal present in the networks (Sharan *et al.*, 2005).

We have recently devised a method, called NetworkBLAST (Sharan *et al.*, 2005), for the identification of protein complexes within and across species. Here, we report on the development of a web-server allowing users to upload PPI network data and analyze them to obtain a visualized list of putative protein complexes over the input networks via a simple and intuitive web-interface.

2 THE WEB-SERVER

The NetworkBLAST web-server implements the algorithm in (Sharan *et al.*, 2005) for analyzing PPI networks across species to identify protein complexes that are conserved in evolution. Briefly, the algorithm constructs an alignment of the analyzed networks, which is searched for conserved protein complexes. Each candidate complex is scored by its fit to a protein complex model, which assumes a certain density of interactions within a complex, versus the likelihood that it arises at random. The server currently supports the analysis of one or two networks and the output consists of a set of putative protein complexes along with their visualization.

In contrast to previous tools for protein network comparison (Flannick *et al.*, 2006; Kelley *et al.*, 2004), which support only single queries of a given pathway/complex in a network of interest, NetworkBLAST allows the uncovering of all conserved complexes across two networks. This all-versus-all computation is much more costly and requires an efficient implementation.

2.1 Input

NetworkBLAST can operate in two-species or single-species modes (Fig. 1). For two species, the input consists of the two respective PPI network files and a sequence similarity file containing BLASTP E-values between pairs of proteins from each of the species. In the single-species mode, only PPI data for one species are needed. A PPI file may be in text or XML format. In the former case, each row represents an interaction and contains the IDs of the interacting proteins pair and a confidence value for it. Alternatively, the user can upload a PSI MI 2.5 file, as e.g. available in the Database of Interacting Proteins [DIP; (Xenarios *et al.*, 2002)]. In this case, the server uses information on the types of experiments in which each interaction was detected to automatically infer a confidence value for it; the computation is based on the logistic regression scheme of (Sharan *et al.*, 2005). The sequence similarity file is a text file, in which each row contains a pair of proteins from different species and their BLASTP E-value.

The user can control several parameters of the algorithm, including the density of the sought protein complexes, the estimated rate of false negatives in each network, the sequence similarity threshold for potential orthology, and the way experiments are categorized for interaction reliability computation (see web-site documentation). The first two parameters affect the scoring of the candidate complexes, see Sharan *et al.* (2005) for details. The third parameter provides a trade-off between speed and sensitivity (see Running time section below).

*To whom correspondence should be addressed.

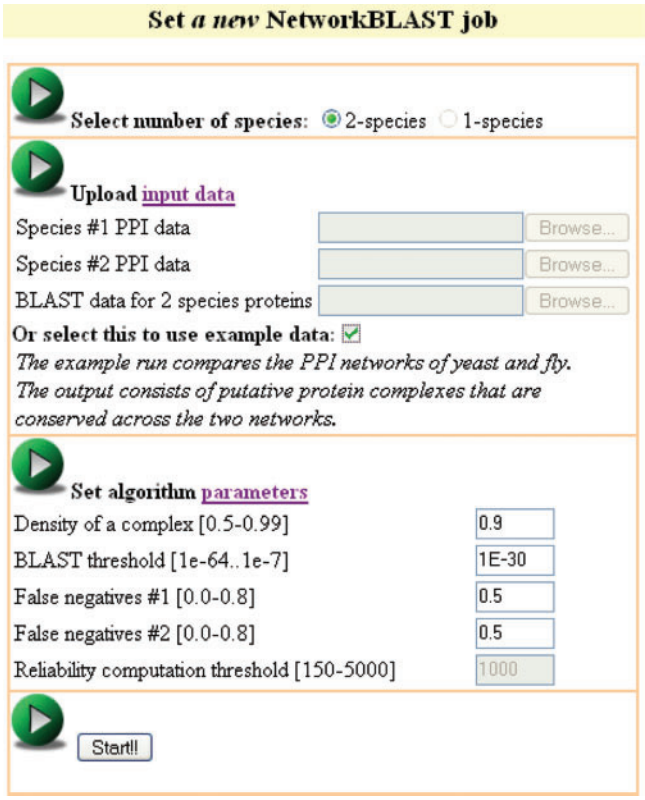

Fig. 1. NetworkBLAST's main page.

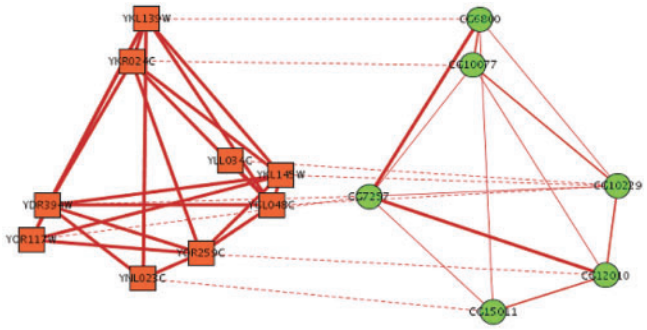

Fig. 2. A representative yeast–fly conserved complex from NetworkBLAST's output. Yeast proteins appear in orange; fly proteins appear in green. Sequence-similar proteins are connected with dashed lines. Solid lines represent PPIs, with the line width corresponding to the reliability of the corresponding interaction.

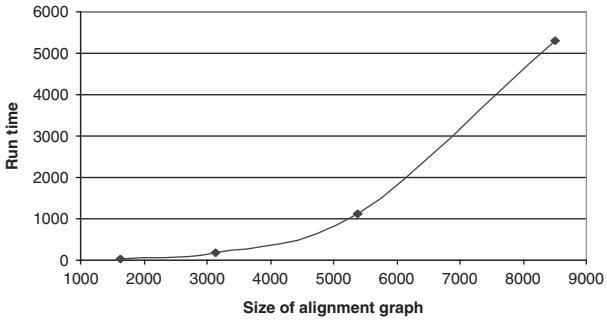

Fig. 3. NetworkBLAST's running time (in seconds, measured on AMD XP2500+, 1 GB memory) versus the alignment graph's size.

2.2 Output

The web-server generates an html page with links to images of the discovered complexes, as well as textual graph data files that can be viewed using the Cytoscape software (Shannon *et al*., 2003). The output putative complexes are sorted by their score. Their images are generated automatically using the Dual Layout plugin of the Cytoscape software. For a single species, the images show the member proteins of each putative complex with edges connecting interacting proteins and edge width corresponding to the interaction's reliability. In two-species mode, the images show aligned putative complexes across two species (drawn side by side), where dashed lines connect orthologous proteins and solid lines denote PPIs (Fig. 2).

2.3 An example run

We demonstrate the utility of the algorithm by presenting an example run on the PPI networks of *Saccharomyces Cerevisiae* and *Drosophila Melanogaster*. These networks are also available to the user as default inputs. Protein–protein interaction data for yeast and fly were downloaded from DIP (Xenarios *et al*., 2002) and contained 15 147 interactions among 4738 proteins in yeast and 23 484 interactions among 7165 proteins in fly. To assign confidence scores to these interactions we used the logistic—regression-based scheme employed in Sharan *et al*. (2005). The false negative rates were estimated at 50% (Sharan *et al*., 2005). The BLAST threshold was set to 1E-30 to allow fast running time. The analysis revealed 49 putative conserved complexes. We tested for functional coherency of these complexes using the GoTermFinder tool (Boyle *et al*., 2004). Yeast complexes of 78% and 63% of the fly complexes were functionally enriched (after correcting for multiple testing using the false discovery rate procedure), serving as a validation of these predictions.

2.4 Running time

NetworkBLAST's performance depends on the sizes of the input networks and the similarity between their proteins. One of the main factors influencing the running time is the size of the constructed network alignment graph (Fig. 3). This in turn depends on the sequence similarity threshold required for two proteins to be considered potentially orthologs. For efficiency reasons, the server runs are currently limited to alignment graphs with up to 5000 nodes. Larger graphs can be handled using an offline version of the program, available for download at the website.

3 CONCLUSIONS

NetworkBLAST is a web-server for protein complex detection within one or two PPI networks. The server allows uploading PPI network data, analyze the networks and visualize

the results. The identified complexes can be utilized to predict protein function and interaction (Sharan *et al*., 2005).

ACKNOWLEDGEMENTS

This research was supported by the Israel Science Foundation (grant no. 385/06).

Conflict of Interest: none declared.

REFERENCES

Aebersold,R. and Mann,M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.

Boyle,E.I. *et al*. (2004) GO::TermFinder–open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.

Fields,S. (2005) High-throughput two-hybrid analysis. The promise and the peril. *FEBS J*., **272**, 5391–5399.

Flannick,J. *et al*. (2006) Graemlin: general and robust alignment of multiple large interaction networks. *Genome Res*., **16**, 1169–1181.

Kelley,B.P. *et al*. (2004) PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Res*., **32**, W83–W88.

Shannon,P. *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*., **13**, 2498–2504.

Sharan,R. *et al*. (2005) Conserved patterns of protein interaction in multiple species. *Proc. Natl Acad. Sci. USA*, **102**, 1974–1979.

Xenarios,I. *et al*. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*., **30**, 303–305.