BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 17 2006, pages 2175–2177
doi:10.1093/bioinformatics/btl287

Systems biology

NetAlign: a web-based tool for comparison of protein interaction networks

Zhi Liang[1,2,†], Meng Xu[1,2,†], Maikun Teng[1,2,*] and Liwen Niu[1,2,*]

[1]Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences, University of Science & Technology of China, 96 Jinzhai Road, Hefei, Anhui 230027, China and [2]Key Laboratory of Structural Biology, Chinese Academy of Sciences, 96 Jinzhai Road, Hefei, Anhui 230027, China

Received on April 12, 2006; accepted on May 16, 2006
Advance Access publication June 9, 2006
Associate Editor: Jonathan Wren

ABSTRACT

Summary: NetAlign is a web-based tool designed to enable comparative analysis of protein interaction networks (PINs). NetAlign compares a query PIN with a target PIN by combining interaction topology and sequence similarity to identify conserved network substructures (CoNSs), which may derive from a common ancestor and disclose conserved topological organization of interactions in evolution. To exemplify the application of NetAlign, we perform two genome-scale comparisons with (1) the *Escherichia coli* PIN against the *Helicobacter pylori* PIN and (2) the *Saccharomyces cerevisiae* PIN against the *Caenorrhabditis elegans* PIN. Many of the identified CoNSs correspond to known complexes; therefore, cross-species PIN comparison provides a way for discovery of conserved modules. In addition, based on the species-to-species differences in CoNSs, we reformulate the problems of protein–protein interaction (PPI) prediction and species divergence from a network perspective.

Availability: http://www1.ustc.edu.cn/lab/pcrystal/NetAlign

Supplementary Data: Supplementary data are available at *Bioinformatics* online.

Contact: mkteng@ustc.edu.cn, lwniu@ustc.edu.cn

1 INTRODUCTION

A key aim of contemporary biology is to characterize topology and dynamics of the extremely complex intracellular networks (Barabási and Oltvai, 2004). Recent progresses in proteomics have provided us with a first chance to characterize protein interaction networks (PINs), but also raised new challenges in analyzing and interpreting the accumulating data (Pellegrini *et al*., 2004). To meet the demand of the fast-growing field, new methods and tools need to be developed. Cytoscape (Shannon *et al*., 2003) and PathBLAST (Kelley *et al*., 2004) server as good examples.

In this paper, we present a new web server NetAlign for comparative analysis of PINs (Sharan *et al*., 2005; Sharan and Ideker, 2006), which compares a user-specified query PIN with a target PIN to identify conserved network substructures (CoNSs).

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

2 METHODS

Network comparison

In NetAlign, a PIN is modeled as an undirected graph with vertices representing proteins and edges representing PPIs. We formulate the identification of CoNSs as subgraph isomorphism and take network comparison as enumerating all the maximal common subgraphs (MCSs) between two PINs. The correspondence between a pair of vertices in two PINs is established, if they are putative orthologs as determined by BLAST search with a user specified *E*-value threshold. The correspondence between a pair of PPIs is defined, if the two pairs of interacting proteins correspond to each other simultaneously. To avoid meaningless combinations of components in disconnected MCSs, we only take connected MCSs into account and define them as s-CoNSs (single CoNSs; see Fig. 1a for examples). We implement a modified Bron–Kerbosch algorithm (Koch, 2001) to solve the problem.

Clustering

Each s-CoNS is an exact match between two subnetworks in two PINs. However, redundancy resulting from paralog interaction and inexact match due to evolutionary events and data incompleteness exist. To handle these, we introduce c-CoNSs (clustered CoNSs; see Fig. 1b for examples) by single-linkage mergence of s-CoNSs. Two s-CoNSs are clustered if their number of intersecting vertices is equal to or greater than 80% of the smaller one for either of the two PINs. c-CoNSs allow inexact match between orthologous regions in two PINs.

Scoring

Each connected component of a CoNS is independently scored as $n(n+1)/2$, where n is the number of conserved PPIs in it. The score of a CoNS is the sum of these individual scores. This strategy gives higher scores to CoNSs with larger size and better connectivity, since they are more likely to occur by conservation. As no simple analytic formula is available and the numerical alternative is time-consuming, the statistical evaluation of CoNSs is not implemented currently.

Web interface

The NetAlign web server has an intuitive user-interface. The query page prompts a user to specify a user-defined query PIN, a target PIN and a BLAST *E*-value threshold. The target PINs are regularly updated from the data released by the DIP (Xenarios *et al*., 2002). The result page reports the identified s-CoNSs and c-CoNSs as well as cross references to other databases. Besides, formatted reports are also available through hyperlinks, which facilitate storage and automatic analysis of the result.

3 RESULTS

PIN comparison between *Escherichia coli* and *Helibobacter pylori*

As the first example, we NetAlign the query PIN of *E.coli* (913 proteins and 2072 PPIs) against the target PIN of *H.pylori* (710 proteins and 1420 PPIs), both of which are derived from the DIP 20051016 release. The default *E*-value threshold 10^{-7} is used. A total of 7 s-CoNSs and 5 c-CoNSs are found, and they all correspond to known conserved complexes (Fig. 1).

Species divergence at the network level can be inferred based on CoNS differences between species. For example, s-CoNS 1, s-CoNS 2 and their mergence c-CoNS 1 suggest that a duplication of RPOB or RPOC results in the symmetric topology of the *E.coli* c-CoNS 1, while the *H. pylori* c-CoNS 1 lacks the duplication and provides a prototype of this molecular machine.

To facilitate the discovery of difference between two PINs, PPIs that exist in only one of them are also reported (see red and green edges in Fig. 1). Based on the query-to-target CoNS difference, we predict two PPIs (rpoA with rpoD, rpoA with fliA) for *H. pylori* and one PPI (uvrA with uvrC) for *E.coli*.

PIN comparison between *Sacchoromyces cerevisiae* and *Caenorhabditis elegans*

As the second example, we NetAlign the *S.cerevisiae* PIN (2635 proteins and 6574 PPIs) against the *C.elegans* PIN (2638 proteins and 4030 PPIs). The yeast PIN is from the DIP 20041003 core subset, and the worm PIN is from the DIP 20051016 release. A total of 167 s-CoNSs and 33 c-CoNSs are identified. We compare the yeast c-CoNSs with the MIPS yeast complex database, and if the proportion of the intersecting proteins between a yeast c-CoNS and a MIPS complex exceeds 80%, the c-CoNS is accepted as a hit. In our analysis, only those MIPS complexes that are manually annotated independently from the DIP data are considered. As a result 12 hits concerning 11 yeast c-CoNSs are found (Supplementary Table 1). This demonstrates discovery of conserved structures in PINs by cross-species comparison. Based on the same notion as above, we predict five PPIs (T13H5.4 with F11A10.2, rfc-2 with rfc-3, rfc-2 with rfc-4, rfc-3 with Q8ST15, rfc-4 with Q8ST15) for *C.elegans*. The predicted PPI between T13H5.4 and F11A10.2 is also present in the Interolog database (Yu *et al*., 2004; Fig. 1b c-CoNS 8).

Conjectures on species divergence between *S.cerevisiae* and *C.elegans* can also be made. For example, c-CoNS 6 is related to the cAMP-dependent protein kinases (Fig. 1b c-CoNS 6). In *S.cerevisiae*, a regulatory subunit exists as BCY1 and it interacts with three types of catalytic subunits TPK1, TPK2 and TPK3. However, in *C.elegans*, the regulatory subunit kin-2 interacts with only one catalytic subunit kin-1. This may reflect the difference of these two species in their cAMP-dependent protein kinases. Theoretically, each pair of matched but topologically non-identical CoNSs can reflect species difference in some aspect.

Currently due to the incompleteness and the unreliability of the available data, our results are fairly limited, for instance, few CoNSs are found and some of the identified species differences may be false. However, with the fast growth of data, our method offers a way to explore species conservation and divergences at the network level.

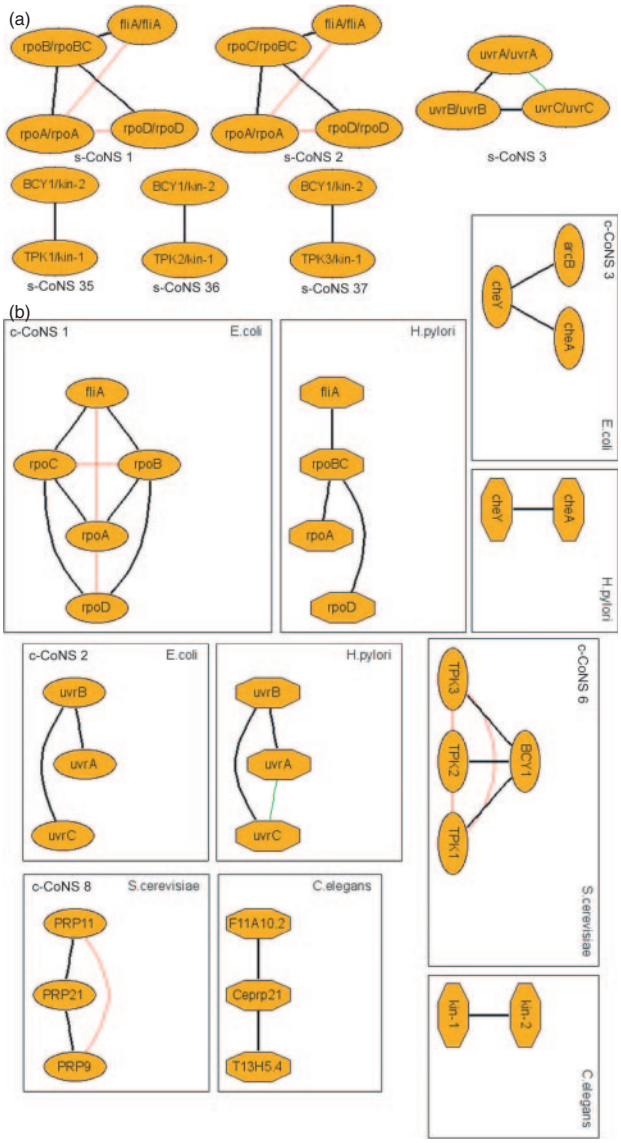

Fig. 1. Examples of identified CoNSs. **(a)** s-CoNSs. Each pair of matched query and target proteins is shown in the same node and delimited by a slash; black edges are conserved PPIs observed in both PINs, red and green edges are interactions observed only in the query and the target, respectively. s-CoNS 1 ~ s-CoNS 3 derived from the PIN comparison between *E.coli* and *H.pylori*, and s-CoNS 35 ~ s-CoNS 37 from the comparison between *S.cerevisiae* and *C.elegans* are shown. s-CoNS 1 and s-CoNS 2 correspond to the RNA polymerase (RNAP) that controls the transcription of RNA in prokaryotes; s-CoNS 3 is the UvrABC repair system catalyzing the recognition and processing of DNA lesions; s-CoNS 35 ~ s-CoNS 37 are the cAMP-dependent protein kinases. **(b)** c-CoNSs. Query and corresponding target c-CoNSs are shown in two separate panels; putative orthologs are shown in the same horizontal level in each panel. c-CoNS 1 ~ c-CoNS 3 identified from the PIN comparison between *E.coli* and *H.pylori*, c-CoNS 6 and c-CoNS 8 derived from the NetAlign analysis between *S.cerevisiae* and *C.elegans* are shown. c-CoNS 1 (obtained by merging s-CoNS 1and s-CoNS 2 in Fig. 1a) is the RNAP; c-CoNS 2 (s-CoNS 3 in fig. 1a) is the UvrABC repair system; c-CoNS 3 is related to chemical sensor systems of bacteria; c-CoNS 6 (generated by merging s-CoNS 35 ~ s-CoNS 37 in Fig. 1a) corresponds to the cAMP-dependent protein kinases; c-CoNS 8 participates the formation of -pre-mRNA splicing factor.

ACKNOWLEDGEMENTS

The authors would like to thank the two anonymous reviewers and Prof. Haiyan Liu for suggestive comments on the manuscript and all members of our lab for discussion. Financial support for this project was provided by research grants from Chinese National Natural Science Foundation (grant nos 30121001, 30025012, 30130080), the '973' and '863' Plans of the Chinese Ministry of Science and Technology (grant nos G1999075603, 2004CB520801 and 2002BA711A13) and the Chinese Academy of Sciences (grant no. KSCX1-SW-17).

Conflict of Interest: none declared.

REFERENCES

Barabási,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nature Rev. Genet.*, **5**, 101–113.

Kelley,B., Yuan,B., Lewitter,F., Sharan,R., Stockwell,B. and Ideker,T. (2004) PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Res.*, **32**, 83–88.

Koch,I. (2001) Enumerating all connected maximal common subgraphs in two graphs. *Theor. Comput. Sci.*, **250**, 1–30.

Pellegrini,M., Haynor,D. and Johnson,J.M. (2004) Protein Interaction Networks. *Expert Rev Proteomics*, **1**, 239–249.

Shannon,P., Markiel,A., Ozier,O., Baliga,N., Wang,J., 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.

Sharan,R., Suthram,S., Kelley,R., Kuhn,T., McCuine,S., Uetz,P., Sittler,T., Karp,R. and Ideker,T. (2005) Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. USA*, **102**, 1974–1979.

Sharan,R. and Ideker,T. (2006) Modeling cellular machinery through biological network comparison. *Nature Biotech.*, **24**, 427–433.

Xenarios,I., SalwÃ-nski,L., Duan,X., Higney,P., Kim,S. and Eisenberg,D. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–305.

Yu,H., Luscombe,N., Lu,H., Zhu,X., Xia,Y., Han,J., Bertin,N., Chung,S., Vidal,M. and Gerstein,M. (2004) Annotation transfer between genomes: protein–protein interologs and protein–DNA regulogs. *Genome Res.*, **14**, 1107–1118.