NetAlign and mNetAlign: Tools for Comparing Protein Interaction Networks

Zhi Liang, Meng Xu, Maikun Teng, and Liwen Niu
Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences
University of Science and Technology of China, China

1 Introduction

One key issue in postgenomic biology is reconstructing the complete molecular interaction networks within cells [Barabási & Oltvai, 2004]. Knowledge on these molecular networks will greatly increase our understanding of the extremely complex cellular machinery involved. With the advent of high-throughput techniques, data on molecular interactions, such as protein-protein interactions (PPIs) are accumulating rapidly. This fast-growing field requires the development of new methods and tools that can analyze and interpret large-scale datasets. As a powerful strategy, cross-species comparison provides a general framework to address the challenge. Recently, a series of works on comparative analysis of molecular interaction networks has been reported [Sharan & Ideker, 2006]. For instance, Ideker et al. performed cross-species analysis on protein interaction networks (PINs) using a dynamic programming-based network alignment tool named PathBLAST [Kelley et al., 2003, Kelley et al., 2004, Sharan et al., 2005]. Meanwhile, Pinter et al. adopted a fast subtree comparison algorithm to align metabolic pathways [Pinter et al., 2005], while Berg and Lässig devised a network alignment method grounded on statistical models of evolution and Bayesian parameter inference [Berg & Lässig, 2006]. Flannick et al. introduced Graælin, the first algorithm for scalable multiple network alignment [Flannick et al., 2006]. In their study, Koyutürk et al. modeled the pairwise local network alignment as a graph optimization problem [Koyutürk et al., 2006]. Dutkowski and Tiuryn proposed a framework for PIN alignment based on the reconstruction of an ancestral PIN [Dutkowski & Tiuryn, 2007]. Li et al. identified conserved subnetworks using integer quadratic programming [Li et al., 2007], while Singh et al. computed a global PIN alignment using IsoRank, a method analogous to Google’s PageRank [Singh et al., 2008]. Liao et al. implemented IsoRankN, a spectral clustering-based approach in their study [Liao et al., 2009]. In order to differentiate local and global alignment, Zaslavskiy formulated the PIN alignment as a graph matching problem [Zaslavskiy et al., 2009]. In his study, Klau used a Lagrangian relaxation approach to compute network alignment [Klau, 2009]. As a contribution to this field, NetAlign, a tool for pairwise PIN comparison based on subgraph isomorphism, has been reported by the authors [Liang et al., 2006a, Liang et al., 2006b]. Here, the authors present mNetAlign, a NetAlign-based database designed to enable comparison of PINs across multiple species. In the text below, the workflow of mNetAlign is described (Figure 1), and its application is demonstrated by performing a five-species PIN comparison among E. coli, H. pylori, S. cerevisiae, C. elegans, and H. sapiens.
Figure 1: Schematic of mNetAlign. mNetAlign accomplishes the comparison of multiple PINs in two steps. First, the comparison between any pair of PINs deposited in DIP are performed using NetAlign, after which the returned s-CoNSs are stored in a database. Second, when a user requests a multi-species PIN comparison, all the s-CoNSs derived from related species are pooled together and merged by a clustering rule to produce c-CoNSs that allow inexact matches among homologous interaction regions in multiple networks. The identified c-CoNSs are then scored on the basis of their interaction topologies.

2 Methods

2.1 Data source

The comparison among the PINs of several model organisms, including *E. coli*, *H. pylori*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, *M. musculus*, and *H. sapiens* is supported by the current implementation of NetAlign and mNetAlign. The PPI data were directly downloaded from the web site of Database of Interacting Proteins (DIP) [Xenarios et al., 2002]. The PPIs among different species and self-interactions were removed.

2.2 Graph model of PINs

A PIN was modeled as an undirected graph $G(V, E)$, where $V$ is a set of nodes representing proteins and $E$ is a set of edges representing PPIs. Prior to conducting a comparison between the two PINs $G_Q(V_Q, E_Q)$
and \( G_T(V_T, E_T) \), the correspondences of nodes and edges must be identified. Generally, the correspondence between the node \( A_Q \) in \( G_Q \) and the node \( A_T \) in \( G_T \) is established if they are putative orthologs as determined by a bi-directional BLAST search between the two species with significant E-values. The correspondence between the pair of PPIs \( A_Q-B_Q \) in \( G_Q \) and \( A_T-B_T \) in \( G_T \) is defined if the two pairs of interacting proteins correspond to each other simultaneously, namely, \( A_Q \) corresponds to \( A_T \), and \( B_Q \) corresponds to \( B_T \).

### 2.3 Pairwise network comparison

NetAlign formulates the identification of conserved network substructures (CoNSs) between two PINs as a subgraph isomorphism problem. Concretely, network comparison was conducted to enumerate all the maximal common subgraphs (MCSs) in the two networks \( G_Q(V_Q, E_Q) \) and \( G_T(V_T, E_T) \). First, an edge compatibility graph \( G_C = (V_C, E_c) \) was built. Here, \( V_C \) is a set of corresponding edge pairs and defined as \( V_C = \{(e_{Qm}, e_{Tn}) | e_{Qm} \in E_Q, e_{Tn} \in E_T, \text{if } e_{Qm} \text{ corresponds to } e_{Tn}\}; e_C \) establishes the connection between the two edge pairs \( v_h = (e_{Qa}, e_{Ta}) \) and \( v_k = (e_{Qb}, e_{Tb}) \), where \( e_{Qa}, e_{Qb} \in E_Q, e_{Ta}, e_{Tb} \in E_T \), as follows: 

\[
E = \{(v_h, v_k) | v_h, v_k \in V_C; \text{ if } e_{Qa} \neq e_{Qb} \text{ and } e_{Ta} \neq e_{Tb}, \text{ and if either } e_{Qa} \text{ and } e_{Qb} \text{ in } G_Q \text{ are connected via a vertex corresponding to the vertex shared by } e_{Ta} \text{ and } e_{Tb} \text{ in } G_T \text{ or } e_{Qa} \text{ and } e_{Qb} \text{ as well as } e_{Ta} \text{ and } e_{Tb} \text{ are not adjacent in } G_Q \text{ and } G_T, \text{ respectively}\}.
\]

Each complete maximal subgraph in graph \( G_C \) is an MCS between \( G_Q \) and \( G_T \). The problem was then transformed into an all-maximal clique problem, which required enumerating all the complete maximal subgraphs. For this, the authors used Bron-Kerbosch algorithm, a fast and widely used algorithm for such a problem. The authors implemented a variant of this algorithm in order to detect all cliques representing connected MCSs [Koch, 2001]. Despite of the NP-hard nature of subgraph isomorphism, the actual constraints on PINs greatly reduce the search space of the problem such as the limited sizes of PINs and ortholog correspondence. Particularly in this study, only the connected MCSs were taken into account and defined as s-CoNSs (single CoNSs) in order to avoid meaningless and repetitive combinations of components in disconnected MCSs during the solution of the problem. Doing so reduced the recursion tree dramatically during the search.

### 2.4 Clustering CoNSs

Each identified s-CoNS is a solution of the network comparison and is an exact match between two subnetworks in the two PINs. However, redundancy exists in regions of interaction where paralogs interact and s-CoNSs can overlap with each other. In addition, there could be an inexact match between the conserved interaction regions in the two PINs due to loss, duplication, and divergence of genes and their associated interactions or data incompleteness; these regions could be disconnected. In order to handle these, c-CoNSs (clustered CoNSs) was introduced by merging similar s-CoNSs. Two s-CoNSs were clustered if the number of their intersecting vertices was equal to or greater than 80% of the smaller one for either of the two species. Three or more s-CoNSs were clustered by the rule of single linkage, that is, the clustering relation was transitive. If one s-CoNS could not be clustered with others, it would form a c-CoNS by itself.

### 2.5 Scoring strategy

Intuitively, the greater the number of conserved PPIs a CoNS contained and the better connectivity it had, the more likely it resulted from a conservation in evolution and not by chance. Based on this consideration, a simple scoring strategy was adopted, in which each connected component of a CoNS was considered independently and scored as \( n(n+1)/2 \). In this notation, \( n \) refers to the number of conserved PPIs, and the ultimate score of the CoNS is the sum of these individual scores. This strategy gave higher scores to CoNSs with larger size and better connectivity.
2.6 Multiple network comparison

Based on NetAlign, a tool called mNetAlign was implemented in order to achieve the comparison of multi-
species PINs. All pairwise NetAlign comparisons were first performed among the PINs using a $10^{-7}$ BLAST
E-value cut-off. The resulting s-CoNSs were then stored in a database. On request, mNetAlign would
pool together the precomputed s-CoNSs derived from the user-specified species, cluster them, and return
the resulting c-CoNSs. mNetAlign has a very simple and intuitive user-interface. Before initiating the
computation, a user must simply specify the PINs to compare and a GO level for GO enrichment analysis. In
the report page, the identified c-CoNSs were sorted by their scores. Aside from cross-referencing with other
databases, mNetAlign provided a report of GO enrichment within c-CoNSs, the aim of which is to facilitate
the analysis of the identified c-CoNSs. Briefly, for each c-CoNS, the enrichment of each GO term associated
with it would be computed for the respective species involved, if the GO term was at least at the user-specified
level from the root of the GO hierarchy. All the information was also formatted into flat files that can be
downloaded through hyperlinks. mNetAlign is accessible at http://netalign.ustc.edu.cn/mNetAlign.

3 Applications

As an example, a five-species comparison was performed across the PINs of E. coli, H. pylori, S. cerevisiae
C.elegans, and H. sapiens. Among the 114 c-CoNSs returned, 104 were identified from two species, 8
from three species, 2 from four species, and none from all the five species (Figure 2). For example, E. coli,
S. cerevisiae, C. elegans, and H. sapiens harbored a conserved interaction region identified as c-CoNS 14,
involved in the mismatch repair (MMR) in DNA and were found to be of paramount importance. Meanwhile,
S. cerevisiae, C. elegans, and H. sapiens shared c-CoNS 2 corresponding to the complex of replication factor
C (RFC). Finally, E. coli and H. pylori were found to have c-CoNS 37, which is the UvrABC repair system
catalyzing the recognition and processing of DNA lesions.

The identified CoNSs were shown to have a variety of applications [Liang et al., 2006b]. In the follow-
ing, the authors will focus the discussion on an evolutionary theme. Species divergence can be inferred at
the PIN level based on c-CoNS differences between species. For example, c-CoNS 6 is the RNA polymerase
(RNAP) controlling the transcription of RNA in prokaryotes. Its topology suggests that the duplication of
rpoB, rpoC, or rpoD results in the symmetric organization of the E. coli RNAP, while the H. pylori counter-
part lacks the duplication and may provide a prototype of this molecular machine. For the second example,
in c-CoNS 7 (an ATP synthase), E. coli only has one component, whereas S. cerevisiae has two discon-
nected components, one of which is mitochondrial ATP synthase and the other vacuolar ATP synthase. The
topological difference correctly reflects divergence of the two species in ATP synthase.

In addition, divergence between evolutionary branches can be identified through the absence or presence
of c-CoNSs. For example, c-CoNS 6 and c-CoNS 54 exist only in E. coli and H. pylori. In contrast, c-CoNS
9 and c-CoNS 22 are exclusively harbored by S. cerevisiae, C. elegans, and H. sapiens. The observation
is consistent with further analyses of these c-CoNSs. Meanwhile, c-CoNS 6 is the prokaryotic RNAP, and
c-CoNS 54 is involved in the sensor and transmission of chemical signals in bacteria, both of which are
specific to prokaryotes. Similarly, c-CoNS 9 and c-CoNS 22 are eukaryote-specific; the former functions
as the general transcription and DNA repair factor IIH (TFIIH) complex, while the latter functions in the
nuclear protein import.

Currently, due to the incompleteness of available data, our results are fairly limited. However, with
the rapid growth of data, the proposed method of identifying CoNSs across multiple species is a powerful
method with which to analyze PINs. For instance, given a number of species from a certain evolution
Figure 2: Representative c-CoNSs identified from a five-specie PIN comparison across *E. coli*, *H. pylori*, *S. cerevisiae*, *C. elegans*, and *H. sapiens*. The corresponding c-CoNSs of different species are shown in separate panels; putative homologs are shown in the same horizontal level in each panel. c-CoNS 2, RFC; c-CoNS 6, prokaryotic RNAP; c-CoNS 7, ATP synthase; c-CoNS 9, TFIH; c-CoNS 14, involved in DNA MMR; c-CoNS 22, related to nuclear protein import; c-CoNS 37, the UvrABC repair system; and c-CoNS 54, participating in the sensor and transmission of chemical signals in bacteria.
branch, this method can identify branch-specific topological organizations of interactions and shed light on the conservation of species and their divergence from the network level.

References


