Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw402
Advance Access Publication Date: 4 July 2016
Applications Note

OXFORD

Genome analysis

The SMAL web server: global multiple network alignment from pairwise alignments

Jakob Dohrmann[1] and Rahul Singh[1,2,*]

[1]Department of Computer Science, San Francisco State University, San Francisco, CA 94132, USA and [2]Center for Discovery and Innovation in Parasitic Diseases, University of California, San Diego, CA 92093, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on February 2, 2016; revised on May 11, 2016; accepted on June 20, 2016

Abstract

Motivation: Alignments of protein-protein interaction networks (PPIN) can be used to predict protein function, study conserved aspects of the interactome, and to establish evolutionary correspondences. Within this problem context, determining multiple network alignments (MNA) is a significant challenge that involves high computational complexity. A limited number of public MNA implementations are available currently and the majority of the pairwise network alignment (PNA) algorithms do not have MNA counterparts. Furthermore, current MNA algorithms do not allow choosing a specific PPIN relative to which an MNA could be constructed. Also, once an MNA is obtained, it cannot easily be modified, such as through addition of a new network, without expensive re-computation of the entire MNA.

Results: SMAL (Scaffold-Based Multiple Network Aligner) is a public, open-source, web-based application for determining MNAs from existing PNAs that addresses all the aforementioned challenges. With SMAL, PNAs can be combined rapidly to obtain an MNA. The software also supports visualization and user-data interactions to facilitate exploratory analysis and sensemaking. SMAL is especially useful when multiple alignments relative to a particular PPIN are required; furthermore, SMAL alignments are persistent in that existing correspondences between networks (obtained during PNA or MNA) are not lost as new networks are added. In comparative studies alongside existent MNA techniques, SMAL MNAs were found to be superior per a number of measures, such as the total number of identified homologs and interologs as well as the fraction of all identified correspondences that are functionally similar or homologous to the scaffold. While directed primarily at PPIN-alignment, SMAL is a generic network aligner and may be applied to arbitrary networks.

Availability information: The SMAL web server and source code is available at: http://haddock6.sfsu.edu/smal/

Contact: rahul@sfsu.edu

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

1 Introduction

Network alignment, in both the pairwise alignment and multiple network alignment (MNA) formulations, is an NP-hard problem. Of these formulations, in practice, MNA requires significantly greater time to compute. Understandably therefore, at the state-of-the-art, the number of pairwise network alignment (PNA) algorithms and software significantly surpasses those for MNA. However, with increasing availability of interactomes, the ability to compare PPINs across species through MNA becomes increasingly critical for a spectrum of problems including identification of conserved functional components, function prediction and phylogenetics. SMAL (Scaffold-Based Multiple Network Aligner) incorporates a number of operational and theoretical novelties: (i) unlike current MNA methods, it allows MNA construction with respect to

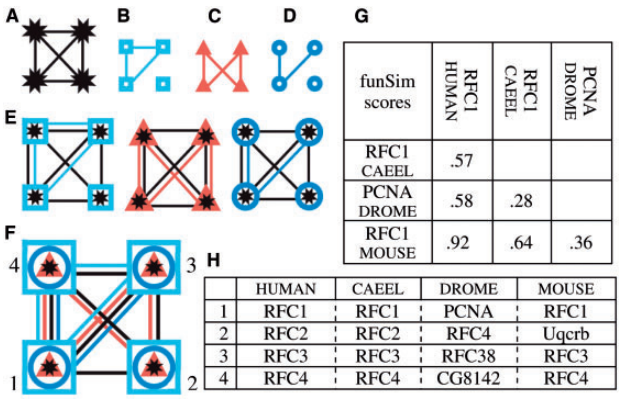
a specific network of interest (called the scaffold), which can be selected based on domain knowledge or other characteristics, such as representativeness or phylogenetic significance. (ii) In determining the MNA, PNAs with respect to the scaffold can be computed using arbitrary PNA algorithms. This property allows SMAL to utilize the large number of specialized PNA methods that currently exist. (iii) SMAL MNAs are invariant to the order in which the PNAs are incorporated. (iv) SMAL MNAs are progressively persistent, in that, network correspondences already part of the MNA are not altered as new networks are added (Supplementary Section S1). (v) SMAL has linear-time complexity with respect to the number of networks being aligned (Supplementary Section S2), leading to extremely fast MNA computations in practice, and (vi) the MNA obtained with SMAL can be related to the constituent pairwise alignments in a straightforward manner, aiding interpretability.

2 Algorithmic underpinnings of SMAL

In the following, we summarize the key algorithmic steps in SMAL; an illustrative example is presented in Figure 1. A detailed algorithmic treatment can be found in Dohrmann *et al.* (2015). Given m networks, let $G_s = (E_s, V_s)$, denote a (designated) scaffold. Let also m-1 pairwise alignments between the scaffold and each of the remaining networks be computed using arbitrary PNA method(s). In SMAL, the resulting m-1 PNAs are related to each other as follows: for each node $p \in V_s$, all nodes corresponding to it in any of the PNAs are combined into a single group $P(p)$, to represent the entire set of protein correspondences. Forming the union over pairwise alignments induces a relationship between nodes aligned to the same scaffold protein, which is termed weak correspondence transitivity (WCT). The biological relevance of WCT was investigated in Dohrmann *et al.* (2015) and additional results are summarized in Supplementary Section S3 and Tables S1 and S2. To determine the induced edges (conserved interactions), for each interaction i in G_s, $i = (u, v) \in E_s$, the Cartesian product of the nodes aligned with the interacting partners u and v is computed and tuples that form an interaction in a PPIN participating in the MNA are retained. The set of conserved interactions for a given scaffold interaction $i = (u, v)$ is therefore expressed as:

$$I(i = (u, v)) = \{(k, l) \in P(u) \times P(v) \exists t : (k, l) \in E_t\} \quad (1)$$

It may be noted that a subset of conserved interactions can be computed using Equation (1) even if certain interactions are absent in specific PPINs, making SMAL robust to missing data. During network alignment, any measure of node-node similarity (e.g. BLAST bit scores) can be provided to create biologically more relevant correspondences. In absence of such information, SMAL determines the alignments purely using network topology. Determining MNAs in SMAL can be accomplished extremely rapidly, taking only few seconds even for large networks. Given m networks, the speedup in SMAL compared to native MNA implementations can be attributed to the lower operational complexity of computing m-1 pairwise alignments and then combining them as opposed to directly determining a single MNA. Furthermore, given the mutual independence of PNAs used in SMAL, their computation can be easily parallelized. Actual computation times depend largely on the employed pairwise alignment algorithm and can range from minutes to several hours with observed speedups ranging between 1.5 up to 50 for SMAL versus native MNA implementations (Supplementary Section S4 and Table S3).



G

funSim scores	RFC1 HUMAN	RFC1 CAEEL	PCNA DROME
RFC1 CAEEL	.57		
PCNA DROME	.58	.28	
RFC1 MOUSE	.92	.64	.36

H

	HUMAN	CAEEL	DROME	MOUSE
1	RFC1	RFC1	PCNA	RFC1
2	RFC2	RFC2	RFC4	Uqcrb
3	RFC3	RFC3	RFC38	RFC3
4	RFC4	RFC4	CG8142	RFC4

Fig. 1. Illustration of the working of SMAL. Subnetworks involved in DNA replication in (**A**) *Homo sapiens* (stars) used as the scaffold, (**B**) *Caenorhabditis elegans* (squares), (**C**) *Drosophila melanogaster* (triangles) and (**D**) *Mus musculus* (circles). (**E**) PNA with Human using PINALOG (Phan and Sternberg, 2012). (**F**) The SMAL MNA. Note that the PPI subnetwork of *D. melanogaster* does not have an interaction between node 3 (RFC38) and node 4 (CG8142). The *H. sapiens* and *C. elegans* subnetworks however, contain an interaction between the corresponding homologs. Consequently, the MNA can be used to hypothesize an analogous PPI in *D. melanogaster*. Supporting this hypothesis, STRING v10 (Szklarczyk *et al.*, 2015) lists the interaction between RFC38 and CG8142 with high confidence (0.999). (**G**) The functional similarity (Schlicker and Albrecht, 2008) of a set of the aligned nodes. (**H**) The SMAL node correspondences. Each row represents the correspondence for the vertices numbered as shown in (**F**)

3 The SMAL web server

The SMAL web server is implemented using python-CGI. Its key modules and workflow are depicted in Figure 2. Examples of input and output file formats are summarized in Table 1 and expanded upon in Supplementary Section S5 and Fig. S1. Briefly, network and alignment files are simple space or tab delimited text. In a network file each line describes one interaction, while in an alignment file (created by an arbitrary PNA algorithm), each line contains a cluster with two or more corresponding nodes. Typically, nodes are encoded using protein names or accession numbers.

To create SMAL MNAs, first, the constituent networks and their pairwise alignments with the scaffold have to be uploaded. Subsequently, SMAL computes multiple alignment files for nodes where each line lists a scaffold node together with all nodes aligned

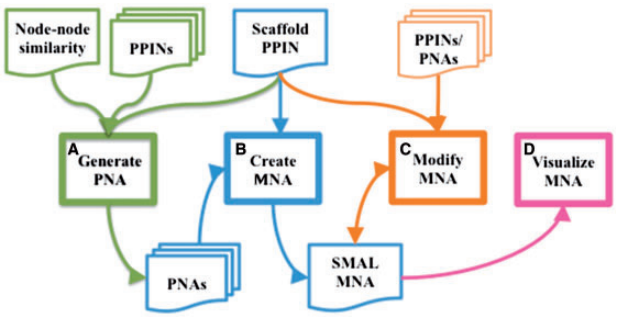


Fig. 2. Functional overview of SMAL. Boxes refer to primary functions supported by the SMAL web server. Document shapes above and below represent input and output files. Depending on the operational context, these functions may be invoked sequentially or independent of each other. (**A**) The SMAL web server links to existing PNA implementations and provides a frontend to generate PNAs from PPINs using SMETANA (Sahraeian and Yoon, 2013). (**B**) Using the designated scaffold, the PNAs are combined into a single MNA based on Equation (1). (**C**) The computed MNAs can be modified by adding or removing PNAs after which the MNA is seamlessly recomputed. (**D**) Resulting MNAs can be visualized and explored online

Table 1. Examples of SMAL input and output files

A	*interactorA*	*interactorB*		
	RFC3	RFC4		
B	*node*	*alignedNode*	[*alignedNodes*...]	
	RFC3	RFC38		
C	*scaffoldNode*	[*alignedNodes*...]		
	RFC3@hs	RFC3@ce	RFC38@dm	RFC3@mm
D	*scaffoldEdge*	[*inducedEdges*...]		
	RFC3@hs_RFC4@hs	RFC38@dm_CG8142@dm		

(A) network (PPIN), (B) PNA node alignment, (C) SMAL labeled MNA node alignment, (D) SMAL labeled MNA edge alignment.

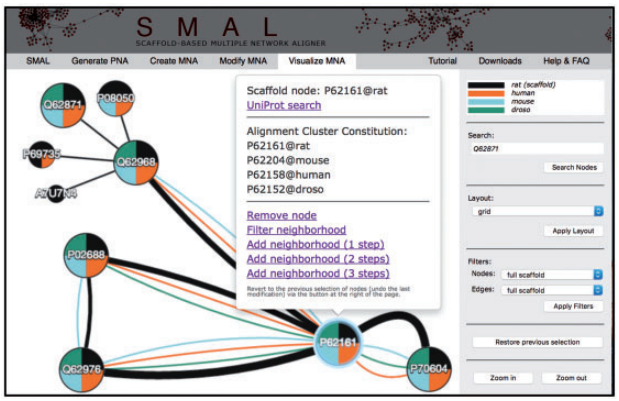


Fig. 3. Visualization of a 4-species MNA of *Rattus norvegicus* (scaffold), *Homo sapiens, Mus musculus* and *Drosophila melanogaster*. A well-conserved cluster of proteins related to ion-channel regulation around P62161 (Calmodulin in *R. norvegicus*) is shown. Each vertex represents an alignment cluster and is labeled with the accession number of the scaffold protein. Both vertex size and the weight of interactions increase with the level of conservation. Additionally, each species in the MNA is associated with an inlaid color in the vertex. Clicking a vertex displays details about the constitution of the corresponding alignment cluster and provides tools to modify and explore the alignment

in any of the PNAs. Based on this MNA node alignment, edge alignments listing scaffold interactions together with all induced interactions as identified by Equation (1) are generated. Both, node and edge alignment files are created as a labeled version where each node is appended with a label indicating the network of origin as well as a plain version without such extra mark-up. Additional information including basic statistics to quantify alignments, such as the percentage of scaffold nodes and edges that have correspondences at different levels of conservation, is provided.

In addition to supporting the core algorithms for creating and modifying MNAs, the web server either directly integrates or links-to a number of PNA methods, including SMETANA (Sahraeian and Yoon, 2013), IsoRankN (Liao *et al.*, 2009), NETAL (Neyshabur *et al.*, 2013) and PINALOG (Phan and Sternberg, 2012). Finally, online visualization and user-data interactions with the MNAs are supported via the cytoscape.js library (Franz *et al.*, 2016). The computed MNA is displayed as a network where each vertex represents a scaffold node together with all the network vertices aligned to it. To assist in visual assimilation of the information, colors are inlaid inside the scaffold vertices to summarize the underlying alignment. Scaffold interactions are rendered together with color-coded edges to represent induced interactions. The visualization module supports searching for proteins, local filtering of aligned neighborhoods based on conservation, and also includes functionality to investigate the neighborhood of selected proteins (Fig. 3).

Table 2. Comparative analysis of SMAL alignments

	$NForH_S$	$EForH_S$	EA_S	EA-3
Native IR	2608.00	263.43	1688.71	137.14
SMAL IR	**4681.00**	**688.86**	**4224.71**	**160.86**
Native SM	6479.14	1446.71	5430.29	**616.29**
SMAL SM	**12 232.86**	**3032.86**	**11 100.43**	514.71

IR, IsoRankN; SM, SMETANA. Results are for an alignment of seven PPINs (*Homo sapiens*, *Drosophila melanogaster*, *Mus musculus*, *Rattus norvegicus*, *Caenorhabditis elegans*, *Escherichia coli* and *Saccharomyces cerevisiae*). $NForH_S$ counts aligned proteins that are homologous or functionally similar. $EForH_S$ counts conserved interactions between homologous or functionally similar proteins. EA_S represents the raw number of all conserved interactions. EA-3 counts the scaffold interactions that are conserved in at least three distinct PPINs. Results are averaged over all possible scaffold choices. Best values are bolded. SMAL outperforms IR and SM for most measures. The performance of SMAL degrades for measures that focus on the interrelation between non-scaffold networks. For such cases, existing MNA techniques may match or outperform SMAL as demonstrated by the EA-3 measure for SM.

3.1 Experiments and validation

SMAL alignments tend to be superior to alignments obtained with current MNA algorithms such as IsoRankN and SMETANA (Table 2). Generally, SMAL node alignment clusters contain a higher number of pairwise homologous or functionally similar proteins than those found by other methods. SMAL MNAs also retain more interactions and comprise a higher fraction of interologs as compared to other aligners. Detailed results examining alignment quality for different choices of the scaffold and a wider set of measures are included in Supplementary Section S6 and Tables S4 and S5. The robustness of SMAL to noisy data is analyzed in Supplementary Section S7 and Fig. S2. Further experimental investigation of SMAL can be found in Dohrmann *et al.*(2015).

Funding

This work was supported in part by the National Science Foundation [grant IIS-0644418].

Conflict of Interest: none declared.

References

Dohrmann,J. *et al.* (2015) Global multiple protein-protein interaction network alignment by combining pairwise network alignments. *BMC Bioinformatics*, **16(Suppl. 13)**, S11.

Franz,M. *et al.* (2016) Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics*, **32**, 309–311.

Liao,C. *et al.* (2009) Isorankn: spectral methods for global alignment of multiple protein networks. *Bioinformatics*, **25**, i253–i258.

Neyshabur,B. *et al.* (2013) Netal: a new graph-based method for global alignment of protein-protein interaction networks. *Bioinformatics*, **29**, 1654.

Phan,H. and Sternberg,M. (2012) Pinalog: a novel approach to align protein interaction networks—implications for complex detection and function prediction. *Bioinformatics*, **28**, 1239–1245.

Sahraeian,S. and Yoon,B. (2013) Smetana: accurate and scalable algorithm for probabilistic alignment of large-scale biological networks. *PLoS One*, **8**, e67995–e67911.

Schlicker,A. and Albrecht,M. (2008) FunSimMat: a comprehensive functional similarity database. *Nucleic Acids Res.*, **36(Database issue)**, D434–D439.

Szklarczyk,D. *et al.* (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43(Database issue)**, D447–D445.