Bioinformatics Advance Access published November 22, 2007

NET-SYNTHESIS: A software for synthesis, inference and simplification of signal transduction networks

Sema Kachalo [1], Ranran Zhang [2], Eduardo Sontag [3], Réka Albert [4] and Bhaskar DasGupta [5*]

[1] Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607
[2] Penn State Cancer Institute and Integrative Biosciences Graduate Program, Pennsylvania State University, Hershey, PA 17033
[3] Department of Mathematics, Rutgers University, New Brunswick, NJ 08903
[4] Departments of Physics and Biology, Pennsylvania State University, University Park, PA 16802
[5] Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607

Associate Editor: Prof. Thomas Lengauer

ABSTRACT

Summary: We present a software for combined synthesis, inference and simplification of signal transduction networks. The main idea of our method lies in representing observed indirect causal relationships as network paths and using techniques from combinatorial optimization to find the sparsest graph consistent with all experimental observations. We illustrate the biological usability of our software by applying it to a previously published signal transduction network (Li *et al*., 2006) and by using it to synthesize and simplify a novel network corresponding to activation induced cell death in large granular lymphocyte leukemia.

Availability: NET-SYNTHESIS is freely downloadable from `http://www.cs.uic.edu/~dasgupta/network-synthesis/`

Supplementary Information: attached separately.

Contact: dasgupta@cs.uic.edu

1 INTRODUCTION

Identification of every reaction and regulatory interaction participating even in a relatively simple function of a single-celled organism requires a concerted and decades-long effort. Consequently, the state of the art understanding of many signaling processes is limited to the knowledge of key mediators and of their positive or negative effects on the whole process. For example, evidence of differential responses to a stimulus in wild-type organisms versus a mutant organism implicates the product of the mutated gene in the signal transduction process. The resulting causal inference relates three components (the signal, the mutated gene and the response) and only in a minority of cases corresponds to a single reaction (namely, when the stimulus is the reactant of the reaction, the mutated gene encodes the enzyme catalysing the reaction and the studied output is the product of the reaction). We previously introduced (Albert *et al*., 2007) a method of synthesizing interactions and causal inferences into a parsimonious network by incorporating positive (activating) or negative (inhibitory) causal relationships as signed network paths with known starting and end vertices (nodes) and putative intermediary pseudonodes. Here we describe an automated version of the method available for use by the community.

*to whom correspondence should be addressed

2 SOFTWARE OVERVIEW

Our software uses as input a text file whose lines represent causal relationships such as "$A \rightarrow B$" (representing activation), "$A \dashv B$" (representing inhibition), or "$A \rightarrow (B \dashv C)$" (indicating a double causal inference). Relationships that correspond to direct interactions are specified by the label "Y", e.g., "$A \rightarrow B$ Y". In addition, the relationship between the enzyme (E) and product (P) of a chemical reaction (i.e., "$E \rightarrow P$") is labeled both "Y" and "E" (for enzymatic edge). The entire network synthesis procedure is given in the Supplementary Information; here we briefly describe some key steps. Double causal relationships of the form $A\,x\,(B\,y\,C)$ with $x, y \in \{\rightarrow, \dashv\}$ are represented by adding a new "pseudo-vertex" P and three new edges, $A\,x\,P$, $B\,a\,P$ and $P\,b\,C$, where a and b are determined by y. Two graph-theoretic procedures, the pseudo-vertex collapse (PVC) and binary transitive reduction (BTR), are used as key steps in the algorithm. Intuitively, the PVC problem is useful for reducing the pseudo-vertex set to the the minimal set that maintains the graph consistent with all indirect experimental observations and the BTR problem is useful for determining a sparsest graph consistent with all experimental observations. Although the initial motivation for introducing pseudonodes is to represent the intersection of the two paths corresponding to three-node inferences, PVC can be used in the broader context of network simplification. In many large-scale regulatory networks only a subset of the nodes are of inherent interest, e.g., because they are differentially expressed in different exogenous conditions, and the rest serve as background or mediators. Our software enables users to designate vertices of less interest or confidence as pseudo-vertices and then collapse them, thereby making the network among high-interest/confidence nodes easier to interpret. To allow gradual simplification we also provide the choice to collapse degree two pseudonodes only or only collapse one pair of equivalent pseudo-vertices. A detailed manual of the software is available from the software's website. The software should run on any machine with MS Windows (Win32). The source files for a non-graphic version of the program for LINUX/UNIX systems can be obtained by sending an email to the authors.

2.1 Data sources

Large-scale repositories such as Many Microbe Microarrays (`http://m3d.bu.edu/cgi-bin/web/array/index.pl?read=aboutM3D`), NASCArrays (`http:// affymetrix.arabidopsis.info/narrays/experimentbrowse.pl`) and Gene Expression Omnibus (`http://www.ncbi.nlm.nih.gov/geo/`) contain expression information for thousands of genes under tens to hundreds of experimental conditions. Network inference algorithms applied to gene expression data based on e.g. mutual information, regression or Bayesian analysis lead to indirect causal relationships among genes. NET-SYNTHESIS can be used to filter redundant inferred relationships by binary transitive reduction. In addition, information about differentially expressed genes responding to a combination of two experimental perturbations, e.g. the presence of a signal in normal versus mutant organisms, can be expressed as double causal inferences. NET-SYNTHESIS can be used to interpret these inferences by pseudo-vertex collapse. Signal transduction pathway repositories such as TRANSPATH (`http://www.gene-regulation.com/pub/databases.html#transpath`) and protein interaction databases such as the Search Tool for the Retrieval of Interacting Proteins (`http://string.embl.de/`) contain up to thousands of interactions, a large number of which are not supported by direct physical evidence. NET-SYNTHESIS can be used to filter redundant information while keeping all direct interactions.

3 RESULTS AND DISCUSSIONS

3.1 Synthesizing a Network for T Cell Survival and Death in Large Granular Lymphocyte Leukemia

T-cell large granular lymphocyte leukemia (T-LGL) represents a spectrum of lympho-proliferative diseases in which cytotoxic T lymphocyte activation and elimination are uncoupled (Loughran, 1993). To date 33 proteins and small molecules related to cytotoxic T lymphocyte activation and activation-induced cell death have been shown to be deregulated in T-LGL and it is known that pro-survival signaling pathways are upregulated and that T-LGL cells are insensitive to Fas-induced apoptosis (Epling-Burnette *et al.*, 2004). However the interaction/regulatory network among these components remains largely unknown.

We synthesized a cell-survival/cell-death regulation-related signaling network from the TRANSPATH 6.0 database, with additional information manually curated from literature search. The 359 vertices of this network represent proteins/protein families and mRNAs participating in pro-survival and Fas-induced apoptosis pathways. The 1295 edges represent regulatory relationships between nodes, including protein interactions, catalytic reactions, transcriptional regulation (for a total of 766 direct interactions), and known indirect causal regulation. No double causal inferences (relationships among three nodes) were available for this network.

Performing BTR with NET-SYNTHESIS reduced the total edge-number to 873. To focus on pathways that involve the 33 known T-LGL deregulated proteins, we designated vertices that correspond to proteins with no evidence of being changed during T-LGL as pseudo-vertices and deleted the label "Y" for those edges whose both endpoints were pseudo-vertices. Recursively performing "Reduction (faster)" BTR and "Collapse degree-2 pseudonodes" of NET-SYNTHESIS until no edge/node could be further removed simplified the network to 267 nodes and 751 edges. Performing comprehensive PVC led to a drastic reduction to 38 vertices and 108 edges. The drawback of this dramatic simplification is that pairs of incoherent edges (two edges with opposite signs) can appear among pairs of nodes. While incoherent paths between pairs of nodes are often seen in biological regulatory networks, interpretation of incoherent edges is difficult without knowledge of the mediators of the two opposite regulatory mechanisms. The number of incoherent edge pairs ranged between 3 (when collapsing degree two pseudo-vertices only) and 19 (for comprehensive PVC). Thus optimal simplification may require several alternative applications of the various options of PVC algorithms.

3.2 Synthesizing a Network for Abscisic Acid(ABA)-induced Stomatal Closure

We have performed a comparison of the manually curated network for ABA-induced closure published in (Li *et al*., 2006) with the output of NET-SYNTHESIS as reported in (Albert *et al*., 2007). The input to NET-SYNTHESIS is a list of 140 interactions and causal inferences in ABA-induced closure published in Table S1 and Text S1 in (Li *et al*., 2006). The complete list of causal relationships is given in Table 1 in the Supplementary Information. A detailed comparison of the two networks is available in (Albert *et al*., 2007), here we briefly summarize the overall comparison of the two networks. The network of (Li *et al*., 2006) has 54 vertices and 92 edges; our network has 57 vertices (3 extra pseudo-vertices) but 84 edges. The two networks have 71 common edges and identical strongly connected components. All the paths present in the (Li *et al*., 2006) reconstruction are present in our network as well. Thus the two networks are highly similar and their divergence on a few edges is due not to algorithmic deficiencies but to human decisions. Finally, the entire network synthesis process was done within a few seconds by our software. A picture of our network is available as Figure 1 in the Supplementary Information.

4 CONCLUSION

The applications of NET-SYNTHESIS enable us to conclude that it can serve as a very important first step in formalizing the logical substrate of an inferred signal transduction network. We foresee its optimal application in conjunction with human expertise, as part of an interactive and iterative process. The NET-SYNTHESIS users would give the experimentally known information as input, then use the output network to augment the input information with additional facts or hypotheses, allowing them to simultaneously synthesize their knowledge and formalize their hypotheses regarding a signal transduction network.

5 FUNDING

This project was partially supported by NSF grants IIS-0346973 (to SK), EIA-0205116 and DMS-050455 (to EDS), MCB-0618402 and CCF-0643529 (to RA), IIS-0346973, IIS-0612044 and DBI-0543365 (to BD) and USDA grant 2006-35100-17254 (to RA).

REFERENCES

R. Albert, B. DasGupta, R. Dondi, et al. (2007). A Novel Method for Signal Transduction Network Inference from Indirect Experimental Evidence, Journal of Computational Biology, 14 (7), 927-949.

S. Li, S. M. Assmann and R. Albert (2006). Predicting Essential Components of Signal Transduction Networks: A Dynamic

Model of Guard Cell Abscisic Acid Signaling, PLoS Biology, 4 (10).

T. P. Loughran Jr (1993). Clonal diseases of Large Granular Lymphocytes, Blood, 82 (1), 1-14.

P. K Epling-Burnette, F. Bai, S. Wei, et al. (2004). ERK Couples Chronic Survival of NK Cells to Constitutively Activated Ras in Lymphoproliferative Disease of Granular Lymphocytes, Oncogene, 23, 9220-9229.