Systems Biology

SAMNetWeb: identifying condition-specific networks linking signaling and transcription

Sara JC Gosline[1], Coyin Oh[1] and Ernest Fraenkel[1*]

[1]Department of Biological Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139.

*Corresponding author

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation*:* High throughput datasets such as genetic screens, mRNA expression assays, and global phospho-proteomic experiments are often difficult to interpret due to inherent noise in each experimental system. Computational tools have improved interpretation of these datasets by enabling the identification of biological processes and pathways that are most likely to explain the measured results. These tools are primarily designed to analyze data from a single experiment (e.g. drug treatment vs. control), creating a need for computational algorithms that can handle heterogeneous datasets across multiple experimental conditions at once.

Summary: We introduce SAMNetWeb, a web-based tool that enables functional enrichment analysis and visualization of high throughput datasets. SAMNetWeb can analyze two distinct data types (e.g. mRNA expression and global proteomics) simultaneously across multiple experimental systems to identify pathways activated in these experiments and then visualize the pathways in a single interaction network. Through the use of a multi-commodity flow based algorithm that requires each experiment 'share' underlying protein interactions, SAMNetWeb can identify distinct and common pathways across experiments.

Availability and Implementation: SAMNetWeb is freely available at http://fraenkel.mit.edu/samnetweb.

Contact*:* fraenkel-admin@mit.edu

1 INTRODUCTION

Due to large consortium-based data collection efforts such as ENCODE (Birney et al., 2007) and TCGA(TCGA Network, 2012), the quantity of biological signaling data is growing at an astounding rate. These data repositories release mRNA expression data, chromatin accessibility data, DNA sequencing data and proteomics data for hundreds of tissues and thousands of patients. As a result, there is a growing need for algorithms to meaningfully interpret diverse types of data across experimental conditions such as different diseases and tissues.

There are numerous network-based analysis tools (Tuncbag, Braunstein, et al., 2012; Yeger-Lotem et al., 2009), including their web-server counterparts (Lan et al., 2011; Tuncbag, McCallum, Huang, & Fraenkel, 2012) that perform visualization and functional enrichment of high throughput datasets within a single experiment. These algorithms improve the functional enrichment of basic statistical approaches (Eden, Navon, Steinfeld, Lipson, & Yakhini, 2009; Jiao et al., 2012) by identifying "hidden" nodes that are likely involved in the experiment but not measured as changing. However, these tools only analyze one experimental condition at a time, making visualization of multiple networks cumbersome. Furthermore, when applied to different perturbations in the same experimental system, these algorithms select similar networks (Gosline, Spencer, Ursu, & Fraenkel, 2012).

SAMNetWeb is able to analyze multiple experiments by mapping both signaling and transcriptomic datasets to the same underlying protein-protein interaction network. Representing each experiment as a 'commodity', the algorithm identifies common and distinct paths between the two datasets for each experiment using the SAMNet multi-commodity flow-based algorithm (Gosline et al., 2012). This approach makes it easier to visualize multiple experiments in the same network as well as enhances the ability to perform pathway enrichment to identify biological processes perturbed in each pathway. SAMNetWeb is implemented in an easy-to-use web server interface that enables the experimentalist to analyze multiple high-throughput experiments with both transcriptional and signaling data from each experiment. The end result is the identification of pathways shared between experiments and those unique to each experiment.

2 DESCRIPTION

SAMNetWeb analyzes signaling and transcriptomic data from multiple experiments by mapping the data to a weighted graph and then running a constrained optimization algorithm that selects the combination of nodes and edges that best explains the observed data. Protein-level signaling changes are mapped to a *protein interaction network* which connects to mRNA expression changes via a *transcriptional regulatory network*. Each edge in the network is weighted by confidence, enabling the SAMNet algorithm to create a reduced network that best explains the connections among the inputs. Usage is described in Figure 1.

2.1 Required input

SAMNetWeb requires two sets of data, representing protein-level

*To whom correspondence should be addressed.

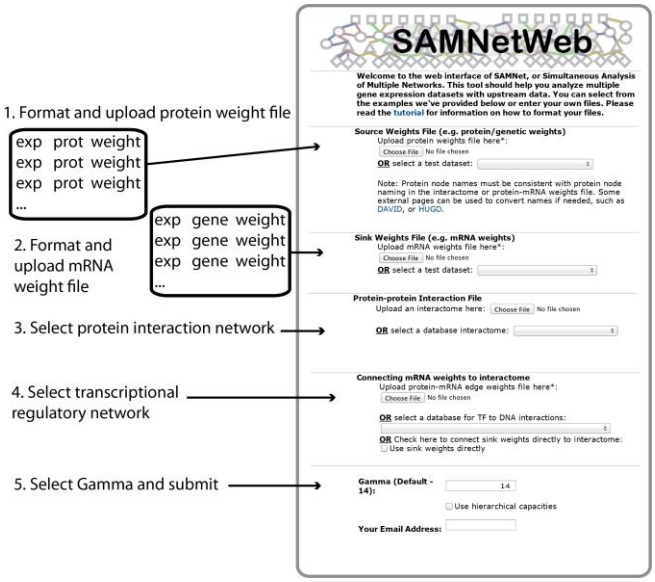

Figure 1: The five required steps to SAMNetWeb data analysis (left) with the respective parts of the SAMNetWeb submission page (right).

signaling changes (Step 1) and mRNA expression changes (Step 2). These data are uploaded to the web page in tab-delimited format. Multiple experiments are accommodated by including an experiment description in the first column ("exp"), next to the protein or gene identifier ("prot" or "gene", second column) and the weight (third column) representing their measured change. The weight allows the algorithm to prioritize inclusion of particular edges, and can be set to the log-fold change across conditions or the significance of the change (e.g. –log(p-value)). SAMNet processes the inputs by taking the absolute value and normalizing the weights.

SAMNetWeb provides interaction networks to facilitate analysis. The user can select from three *protein-protein interaction networks* from established repositories (Step 3). The provided *transcriptional regulatory networks* (Step 4) are derived from DNase I hypersensitive clusters from the ENCODE consortium to predict protein-DNA interactions using clusters of TRANSFAC motifs (MacIsaac et al., 2010). We also provide two cell-line specific networks, derived from DNase I hypersensitive regions from A549 cells (lung cancer cell line) and MCF7 cells (breast cancer cell line). For each of these, edge weights between transcription factors and mRNA are determined by motif scores within hypersensitive regions 2kb upstream of the gene transcription start site. When using only protein-level data in Step 2, the transcriptional regulatory network is not necessary so the user must check 'use sink weights directly' in Step 4.

Lastly, a user must select a value of *Gamma* to run the algorithm. *Gamma* limits the number of nodes that can connect to the source. *Gamma*=0 results in an empty network, and the network solution size will increase with *Gamma* until the network is saturated. We use a default *Gamma* value of 14.0, but we recommend the user experiment with a number of values to identify the best network.

2.2 SAMNetWeb results

Once the files are properly formatted they can be submitted (Step 5) tfor analysis. Results are queued and processed within ~15 minutes and displayed on a static web page. Users can download relevant Cytoscape (Shannon et al., 2003) files, resulting DAVID enrichment (Jiao et al., 2012) analysis and view the network using the Cytoscape plug-in (Lopes et al., 2010).

We demonstrate the results of SAMNetWeb at http://fraenkel.mit.edu/samnetweb/emt. In this analysis, we showcase the ability to study multiple conditions at once by combining phosphoproteomic measurements and mRNA expression measurements across four experimental models of epithelial-mesenchymal transition (EMT). The resulting network and DAVID enrichment demonstrate the individual interactions predicted in each EMT model as well as the biological processes that are enriched for each dataset.

3 DISCUSSION

With the publication of more high-throughput datasets there is a growing need for easy-to-use, integrative analysis tools. SAMNetWeb is able to integrate signaling and transcriptomic data across multiple experimental systems to facilitate hypothesis generation following initial high-throughput sequencing experiments. The intuitive web-server interface and pre-formatted interaction networks make the algorithm useable without prior programming knowledge. Furthermore, SAMNetWeb is very flexible in its inputs, allowing the user to map any type of data to the interactome. This flexibility will enable SAMNetWeb to remain useful as new technologies are developed to analyze genome-wide data.

ACKNOWLEDGEMENTS

The authors acknowledge Dr. Nurcan Tuncbag for her technical guidance and Scott Moskrin for technical support.
Funding: This work is supported by NIH grants U54CA112967 and R01GM089903 and used computing resources funded by the National Science Foundation under Award No. DB1-0821391.

REFERENCES

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., … de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, *447*(7146), 799–816. doi:10.1038/nature05874

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, *10*(1), 48. doi:10.1186/1471-2105-10-48

Gosline, S. J., Spencer, S. J., Ursu, O., & Fraenkel, E. (2012). SAMNet: a network-based approach to integrate multi-dimensional high throughput datasets. *Integrative Biology*. doi:10.1039/c2ib20072d

Jiao, X., Sherman, B. T., Huang, D. W., Stephens, R., Baseler, M. W., Lane, H. C., & Lempicki, R. A. (2012). DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics (Oxford, England)*, *28*(13), 1805–6. doi:10.1093/bioinformatics/bts251

Lan, A., Smoly, I. Y., Rapaport, G., Lindquist, S., Fraenkel, E., & Yeger-Lotem, E. (2011). ResponseNet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data. *Nucleic Acids Research*, *39*(Web Server issue), W424–9. doi:10.1093/nar/gkr359

Lopes, C. T., Franz, M., Kazi, F., Donaldson, S. L., Morris, Q., & Bader, G. D. (2010). Cytoscape Web: an interactive web-based network browser. *Bioinformatics (Oxford, England)*, *26*(18), 2347–8. doi:10.1093/bioinformatics/btq430

MacIsaac, K. D., Lo, K. A., Gordon, W., Motola, S., Mazor, T., & Fraenkel, E. (2010). A quantitative model of transcriptional regulation reveals the influence of binding location on expression. *PLoS Computational Biology*, *6*(4), e1000773. doi:10.1371/journal.pcbi.1000773

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–504. doi:10.1101/gr.1239303

TCGA Network, T. C. G. A. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), 61–70. doi:10.1038/nature11412

Tuncbag, N., Braunstein, A., Pagnani, A., Huang, S.-S. C., Chayes, J., Borgs, C., … Fraenkel, E. (2012). Simultaneous reconstruction of multiple signaling pathways via the prize-collecting Steiner forest problem. In *RECOMB* (pp. 127–1477).

Tuncbag, N., McCallum, S., Huang, S.-S. C., & Fraenkel, E. (2012). SteinerNet: a web server for integrating "omic" data to discover hidden components of response pathways. *Nucleic Acids Research*, *40*(Web Server issue), W505–9. doi:10.1093/nar/gks445

Yeger-Lotem, E., Riva, L., Su, L. J., Gitler, A. D., Cashikar, A. G., King, O. D., … Fraenkel, E. (2009). Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nature Genetics*, *41*(3), 316–23. doi:10.1038/ng.337