

Computational dissection of regulatory networks using diverse high-throughput data

Ron Shamir
School of Computer Science
Sackler Faculty of Exact Sciences
Tel Aviv University
Tel Aviv 69978 Israel
rshamir@tau.ac.il

Abstract

The maturation of high-throughput technologies and the availability of whole genome sequences make it possible to apply holistic computational approaches to the study of biological systems. The use of high-throughput technologies requires the development of advanced computational methods and tools that would enable the elicitation of significant biological knowledge from the vast amounts of data generated by these methods. Our group has been developing a battery of such methodologies and incorporated some of them in several tools:

- **CLICK** (CLuster Identification via Connectivity Kernels): a clustering algorithm that combines graph-theoretic approaches and statistical considerations to yield solutions that balance intra-cluster homogeneity and inter-cluster separation.
- **PRIMA** (PRomoter Integration in Microarray Analysis): a promoter sequence analysis tool that aims at the identification of transcription factors whose binding sites are significantly over-represented in promoters of co-expressed genes. Using microarrays to compare the transcriptional response in wild-type and Atm-deficient mice, we used CLICK and PRIMA to identify, on a genomic scale, a DNA damage transcriptional response that is dependent on the ATM protein kinase, and dissected this response network into two major arms that are mediated by the p53 and NF- κ B transcriptional regulators.
- **SAMBA** (Statistical-Algorithmic Method for Bicluster Analysis): a method for finding subsets of genes that manifest a significant co-expression within particular subsets of the conditions. The method is graph-theoretic and based on a statistical model of the data generation. We demonstrated the utility of SAMBA in mining biological knowledge out of large and highly heterogeneous genome-wide yeast datasets. These included gene expression profiles, and data on protein-protein interactions, growth phenotypes, and transcription factor binding locations. Our approach analyzes such heterogeneous data set in an inherently integrative manner. SAMBA dissected the yeast system into *modules*, each comprising a set of genes that share common features over diverse data sources. Using these modules, we were able to predict the function of over 800 unknown genes, and validated some predictions experimentally. We were also able to obtain broad perspectives on the interaction of transcription factors and modules, and on the hierarchical organization of modules in yeast.
- **EXPANDER** (EXpression Analyzer and DisplayER): an integrative platform for the analysis of gene expression data, providing multiple analysis algorithms including CLICK, PRIMA and SAMBA, along with a variety of data normalization and visualization utilities.
- **SHARP** (SHowcase for ATM Related Pathways): an interactive software environment that displays graphically biological interaction networks, allows dynamic layout and navigation through these networks, and the superposition of DNA microarray data on interaction maps.
- **Binding Site Evolution**: a novel genome-wide analysis method for detecting binding sites in aligned promoters of related species, which is based primarily on identifying selection forces and not mere conservation. We demonstrate the method to the data of several yeast species, and the analysis reveals novel fascinating details on the evolution of transcription factor binding sites.
- **MetaReg**: a methodology for the representation and analysis of heterogeneous biological networks. The

network elements include mRNAs, proteins and metabolites. A discrete, synchronous model is assumed, and the network may contain cycles. We developed methods for the comparison of the model predictions to actual measurements, and for the generation of hypotheses where discrepancy between predictions and observations is large. We demonstrate the approach on the lysine biosynthesis pathway in yeast.

Our (more mature) tools are available at <http://www.cs.tau.ac.il/~rshamir>

Joint work, in parts, with Amos Tanay¹, Irit Gat-Viks¹, Ran Elkon², Roded Sharan^{1,4}, Chaim Linhart¹, Adi Maron-Katz², Nir Orlev², Giora Sternberg², Martin Kupiec³, Sharon Rashi-Elkeles², Yosef Shiloh²

References

R. Sharan, A. Maron-Katz and R. Shamir, "CLICK and EXPANDER: A System for Clustering and Visualizing Gene Expression Data" *Bioinformatics* Vol. 19 No. 14 pp. 1787--1799 (2003).

R. Elkon, C. Linhart, R. Sharan, R. Shamir, Y. Shiloh, "Genome-wide In-silico Determination of Transcriptional Regulation Modules Controlling Cell Cycle in Human Cells" *Genome Research* Vol. 13(5) pages 773-780, May 1 2003.

A. Tanay, R. Shamir, "Modeling Transcription Programs: Inferring Binding Site Activity and Dose - Response Model Optimization". *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology (RECOMB 03)*, pp. 301-310, ACM Press, 2003.

A. Tanay, R. Shamir, "Multi-level Modeling and Inference of Transcription Regulation". To appear in *Journal of Computational Biology*.

A. Tanay, I. Gat-Viks, and Ron Shamir "A Global View of the Selection Forces in the Evolution of Yeast Cis-Regulation". *Genome Research* 14: 829--834 (2004).

I. Gat-Viks, A. Tanay, R. Shamir. "Modeling and Analysis of Heterogeneous Regulation in Biological Networks" to appear in *Journal of Computational Biology*.

I. Gat-Viks and R. Shamir. "Chain functions and scoring functions in genetic networks". *Proc. 11th International Conference on Intelligent Systems for Molecular Biology (ISMB 03)*, Brisbane, Australia, July 2003. *Bioinformatics* Vol. 19 Supplement 1, pp. i108--i117 (2003).

A. Tanay, R. Sharan, M. Kupiec, R. Shamir "Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genome-wide data". *Proc. National Academy of Science USA* 101 (9) 2981--2986 (2004).

¹ School of Computer Science, Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel

² Department of Human Genetics, Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel

³ George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel

⁴Current address: International Computer Science Institute, Berkeley, CA.