Data Fusion in Functional Genomics

Chen-Hsiang Yeang & Tommi Jaakkola

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: The goal of this research project is to infer the underlying mechanism and pattern of gene regulation in yeast on the basis of combined constraints arising from multiple information sources. The available data sources include gene expression measurements, location analysis data, as well as upstream DNA binding motifs. We need a flexible computational framework that fully exploits the partial constraints that can be inferred from each data source.

Motivation: Inferring the principles and mechanism of gene expression is one of the grand challenges in functional genomics. A single source of information such as gene expression data is, however, rather limited in its ability to restrict possible interpretations of the underlying regulatory processes. Other sources of information such as (tentative) factor-gene pairs revealed by location analysis (Ren et al. [1]) provide complementary (physical level) constraints on the models of regulatory processes. By integrating limited but complementary data sources we can realize a mutually consistent hypothesis bearing stronger similarity to the underlying causal structure.

Previous Work: A typical approach for exploiting two or more data sources in functional genomics uses one type of data to validate the results generated independently from the other (i.e., no data fusion). For example, Tavazoie et al. ([6]), Spellman ([5]), and Zhang ([8]) perform clustering on the basis of gene expression data and identify consensus motifs in the promoters within each cluster. The underlying assumption here is that genes co-expressed under varying experimental conditions are likely to be co-regulated by the same (possibly unknown) transcription factors. For example, a specific transcription factor may bind to the region of DNA that possesses its consensus sequence. Holmes et al. ([4]) construct a joint likelihood score based on consensus motif and gene expression and use this score to perform clustering (thus performing a type of data fusion). Segal et al. ([3]) build relational probabilistic models by incorporating gene expression and functional categories as input variables. Pe'er et al. ([2]) infer causal models from knock-out data by exploiting the distinction between *observational* and *interventional* data in causal inference ([7]).

Approach: We develop a computational framework which integrates gene expression data, sequence data, location analysis and knock-out experiments to identify regulatory networks. Expression data captures the *statistical dependencies* of genes. We build graphical models (Gaussian-Markov models, Bayesian networks, dynamic Bayesian networks) from expression data and extract local dependencies of gene pairs. Two genes g_1 and g_2 are directly dependent if there exist no other genes which can explain away their interactions, i.e., there exists no S s.t. $g_1 \perp g_2 | S$. Each dependency must be explained by some causal relations. In functional genomics, this corresponds to finding the pathway(s) linking the dependent genes.

Expression data alone is rather limited in providing information about the underlying causal relations. However, sequence data, location analysis and knock-out experiments, albeit limited in quantity, reflect physical processes of gene regulation. We build a physical models consistent with all the available data. Location analysis and consensus motifs impose constraints on links between specific (or unknown) transcription factors and genes that they potentially regulate. Knock-out experiments, on the other hand, provide partial ordering constraints for the regulatory pathways. Using a physical level model to explain gene expression data may not yield a high likelihood score since many intermediate steps are not carried out (directly) via transcriptional changes. The goal here is to determine what part of the network is exercised in a specific expression dataset or, in another words, what part of the physical model can explain the statistical dependencies observed in the expression dataset.

We use physical level models and explain statistical dependencies in gene expression data by hypothesizing potential paths mediating the observed dependence. There may be multiple paths which explain the same relation.

We identify paths in the physical model that participate in explaining a large number of statistical dependency relations. Such paths are likely to reflect the underlying causal processes. We can determine the direction and, in principle, also the delay in the effect of regulation when time series data is available.

Impact: Our method allows us to infer the regulatory network from the genomic data rather than merely depicting the phenomenon of gene expression. Such models correspond more closely to biologists' understanding of the regulatory system and naturally lend themselves to optimum experiment design methods. The information fusion methodology developed in this project is more widely applicable to tasks of inferring underlying causal processes from multiple limited data sources.

Future Work: An immediate extension of this work is to cast the physical level constraints on a sound probabilistic framework. Instead of evolving (or realizing) a single underlying physical model we can generate multiple models and assign probability scores to them on the basis of statistics derived from individual data items. The inference task consequently changes to identifying pathways with the ability to explain expression dependencies and which occur with high probability in the physical level. We are also in the process of adapting and further developing appropriate experiment design methods for models of this type.

Research Support: This project is in part supported by Nippon Telegraph and Telephone Corporation and by ARO MURI grant DAAD19-00-1-0466.

References:

[1] B. Ren et al. Genome-wide location and function of dna-binding proteins. *Science*, 5500:2306–2309, 2000.

[2] D. Pe'er et al. Inferring subnetworks from perturbed expression profiles. In *Proc. ISMB*, 2001.

[3] E. Segal et al. Rich probabilistic models for gene expression. In *Proc. ISMB*, 2001.

[4] I. Homles et al. Finding regulatory elements using joint likelihoods for sequence and expression profile data. In *Proc. ISMB*, 2000.

[5] P.T. Spellman et al. Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular biology of the cell*, 9:3273–3297, 1998.

[6] S. Tavazoie et al. Systematic determination of genetic network architecture. *Nature genetics*, 22:281–285, 1999.

[7] J. Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, 2000.

[8] M.Q. Zhang. Large-scale gene expression data analysis: a new challenge to computational biologists. *Genomic research*, 9:681–688, 1999.