Discovering Functional Transcription Factor Binding from Superimposed Gene Networks

Matt Weirauch and Josh Stuart
Department of Biomolecular Engineering, University of California at Santa Cruz
{weirauch, jstuart}@soe.ucsc.edu

Abstract

The availability of entire genome sequences, coupled with genome-wide studies of gene expression, offers promise for discovering new pathways along with their regulatory programs. Clusters identified from gene co-expression networks (GCNs) reveal correlations between genes but say little about the mechanism behind their coregulation. We have constructed a co-binding network (CBN) to identify the potential combinations of transcription factors (TFs) that may regulate a set of genes. The CBN was built by connecting all pairs of genes bound by the same transcription factor observed in ChIP-Chip microarray experiments. We superimposed the CBN onto the GCN to identify clusters of genes in overlapping sub-networks. Applying our method to a GCN derived from four distantly related species, we identified transcription factor combinations for several conserved sub-networks in yeast.

1. Introduction

With the recent availability of datasets from large-scale studies on gene expression, we may finally begin to decipher the complex regulatory structure of the cell. One major challenge lies in deciphering the combinatorial and context-specific logic of gene regulation from these data. For example, we seek rules such as "transcription factors A and B up-regulate genes X, Y, and Z under condition C." As a first step in this direction, many current methods search for clusters of coregulated genes. While clustering gene expression profiles may help reveal co-regulation groups, the clusters themselves say little about the mechanism underlying their coregulation.

Recently, genomewide ChIP-Chip assays have started to provide us with more concrete examples of TF binding. These experiments provide evidence of where TFs bind in the genome, but do not offer any information on whether this binding is functional. Several methods have been proposed to combine coexpression data with transcription factor binding data in an attempt to address these issues. Our work extends these methods by using expression data from multiple species and locating coregulated groups of genes along with the TFs that regulate them.

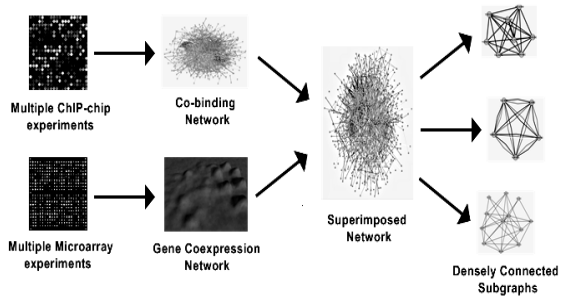

Figure 1: Method overview

2. Methods

Our method takes as input multiple DNA and ChIP-Chip microarray data and finds subsets of coregulated genes and the TFs that potentially regulate them (Fig. 1). First, we build a co-binding network (CBN) that describes which genes are bound by the same set of TFs. In the CBN, genes are represented as nodes and each TF is represented as a labeled edge. We compiled the TF binding data from several yeast ChIP-Chip microarray experiments [1-4]. Next, we create a gene coexpression network (GCN) by mapping the multispecies (human, fly, worm, yeast) network, as described in [5] to yeast. This network connects two genes if they are significantly coexpressed across several microarray experiments in multiple eukaryotic organisms. We then create a superimposed network (SN) by intersecting the CBN with the GCN. This network therefore connects all genes that are both coexpressed and share at least one common TF. Finally, we identify *kernels* as strongly connected

subgraphs in the SN. To identify these kernels, we use the MCODE algorithm [6], which is available as a plug-in for the Cytoscape software package [7].

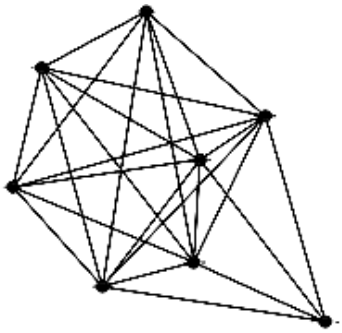

Figure 2: HSF1 kernel

3. Results and Discussion

We applied our method to the gene coexpression network determined in [5] and to the co-binding network created from [1-4]. We found 16 kernels having five or more genes and one or more significant TFs. Several of these kernels contain subkernels of genes regulated by different combinations of TFs. Future work will shed light on the complex interactions between the TFs in each kernel.

To assess our method's ability to find relevant TF-kernel pairings, we analyzed known targets of the yeast heat shock factor (HSF1). HSF1 is a member of heat shock proteins that is well conserved across eukaryotes. Hahn et al [8] created a 'gold standard' dataset by combining expression results of an HSF1 mutant with an HSF1 specific ChIP-Chip experiment. To create a gold standard to compare our method against, we intersected the gene list of Hahn et al with those contained in the CBN, yielding a total of 26 genes. 84 genes (out of 108) from the CBN alone were not contained in this gold standard set. Thus, many putative targets of HSF1 determined by ChIP-Chip data alone cannot be confirmed by specific knock-down of HSF1, supporting the idea that many of its binding sites may be non-functional.

Among the kernels identified by MCODE is a kernel of eight genes connected solely by HSF1 factors (Fig. 2). It is worth noting that the set of genes contained in the kernel are all connected strictly by HSF1, and do not share any other TFs. As is evidenced by the high connectivity of this kernel, the majority of these genes are coexpressed together in the GCN. Of the eight genes in this kernel, six of them are contained in the gold standard (0.75 specificity as opposed to the 0.24 specificity obtained from using ChIP-Chip data alone.) Of the remaining two, one is a known heat shock protein chaperonin. The final gene is not known to be involved in the heat shock response.

4. Future Directions

Eukaryotic genes are often regulated by more than one TF. In the future, we will extend our method to the automated discovery of TF combinations that regulate a set of genes. Using our superimposed network, we can test every combination of TFs within a kernel for its ability to distinguish kernel genes from non-kernel genes.

Finally, as more binding data is becoming available for multiple organisms, we will be able to construct a CBN representing conserved TF binding relationships. We expect that using an evolutionarily conserved CBN should increase our ability to identify functional combinations of TF binding. In this way, we hope to decipher a clearer picture of the evolution of gene regulation in eukaryotes.

5. References

[1] Harbison *et al*, "Transcriptional regulatory code of a eukaryotic genome", *Nature*, 2004 Sep 2, pp. 99-104.

[2] Horak *et al*, "Complex transcriptional circuitry at the G1/S transition in Saccharomyces cerevisiae", *Genes & Development*, 2002 Dec 1, pp. 3017-33.

[3] Iyer *et al*, "Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF", *Nature*, 2001 Jan 25, pp. 533-8.

[4] Lieb *et al*, "Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association", *Nature Genetics*, 2001 Aug, pp. 327-34.

[5] Stuart *et al*, "A gene-coexpression network for global discovery of conserved genetic modules", *Science*, 2003 Oct 10, pp. 249-55.

[6] Bayer *et al*, "An automated method for finding molecular complexes in large protein interaction networks", *BMC Bioinformatics*, 2003 Jan 13, 4:2.

[7] Shannon *et al*, "Cytoscape: a software environment for integrated models of biomolecular interaction networks", *Genome Research*, 2003 Nov, pp. 2498-504.

[8] Hahn *et al*, "Genome-wide analysis of the biology of stress responses through heat shock transcription factor", *Molecular Cell Biology*, 2004 Jun, pp. 5249-5