Reconstructing genetic networks in yeast

Zhaolei Zhang & Mark Gerstein

By combining data from gene expression and DNA-binding experiments, a computational algorithm identifies the genetic regulatory network in yeast.

[AU: OK?]

A central challenge in genomic biology is to determine how cells coordinate the expression of thousands of genes throughout their life cycle or in response to external stimuli, such as nutrients or pheromones. [AU: OK?] In eukaryotes, gene expression is modulated by various transcription factors that bind to the promoter regions, and different combinations of transcription factors may alternatively activate or repress gene expression. This is analogous to an electronic circuit, in which components are switched on and off by a network of transistors. In this issue, Bar-Joseph and colleagues[1] report a computational approach to show that in yeast, genes are indeed regulated in networks that are controlled by groups of transcription factors. Furthermore, they show that these regulatory networks also have a modular structure in which groups of genes under the control of the same regulators tend to behave similarly. [AU: OK?]

Genetic regulation and its mechanisms have been investigated since the days of Jacob and Monod and the discovery of the *lac* operon. Traditionally, such studies are labor-intensive and gene-specific and often require years of bench work. More recently, and with the complete genome sequences of a number of eukaryotic organisms becoming available, several high-throughput genomic technologies have been developed, which allow biologists to study gene expression and gene regulation on a whole-genome scale. The concept of determining gene expression at the whole-genome level was first introduced by Brown and colleagues[2], who developed DNA microarrays to measure the expression level for every gene in yeast simultaneously.

In recent years, several groups have also implemented chromatin immunoprecipitation (ChIP) DNA chips for directly mapping the *in vivo* physical interactions between transcription factors and their DNA binding sites[3–6]. Briefly, a cell line expressing a specific tagged transcription factor is constructed. After growth under experimental conditions, DNA fragments bound to the tagged transcription factors are recovered by a ChIP assay and hybridized to DNA microarrays containing the complete set of the yeast intergenic regions. Strong hybridization in a region proximal to a gene would indicate transcription factor binding to that gene's promoter site.

Many researchers have attempted to apply statistical or computational approaches to reconstruct genetic regulatory networks based on data sets derived from these whole-genome methodologies. Most of the approaches have consisted of applying clustering algorithms to gene expression data to identify coexpressed genes, which are surmised to be coregulated by shared transcription factors[7]. Such approaches have also been expanded to incorporate previous knowledge about the genes, such as cellular functions or promoter sequence motifs[8,9]. These methods have achieved various levels of success, but an intrinsic limitation is their over-reliance on expression data, which represent the result rather than the cause of genetic regulation. In addition, some of these methods assume that expression levels are correlated between the transcription factors and the genes that they regulate. This has been proven not always to be true[10].

Other computational methods have also been developed to extract regulatory information from whole-genome DNA-binding data sets[5,6]. The rationale behind these approaches is that if two genes share a common set of transcription factors, then they are probably coregulated and belong to the same gene module. Using this location-based approach, researchers have successfully identified some basic regulatory motifs in the yeast network. But this approach has its own limitations. First, location information does not indicate whether the nature of the regulation is in the positive or negative direction; second, DNA-protein interaction data are noisy owing to much nonspecific binding.

As reported in the present paper, Bar-

Zhaolei Zhang and Mark Gerstein are at the Department of Molecular Biophysics and Biochemistry, Yale University, 266 Whitney Avenue, New Haven, Connecticut 06520-8114, USA. e-mail: Mark.Gerstein@yale.edu

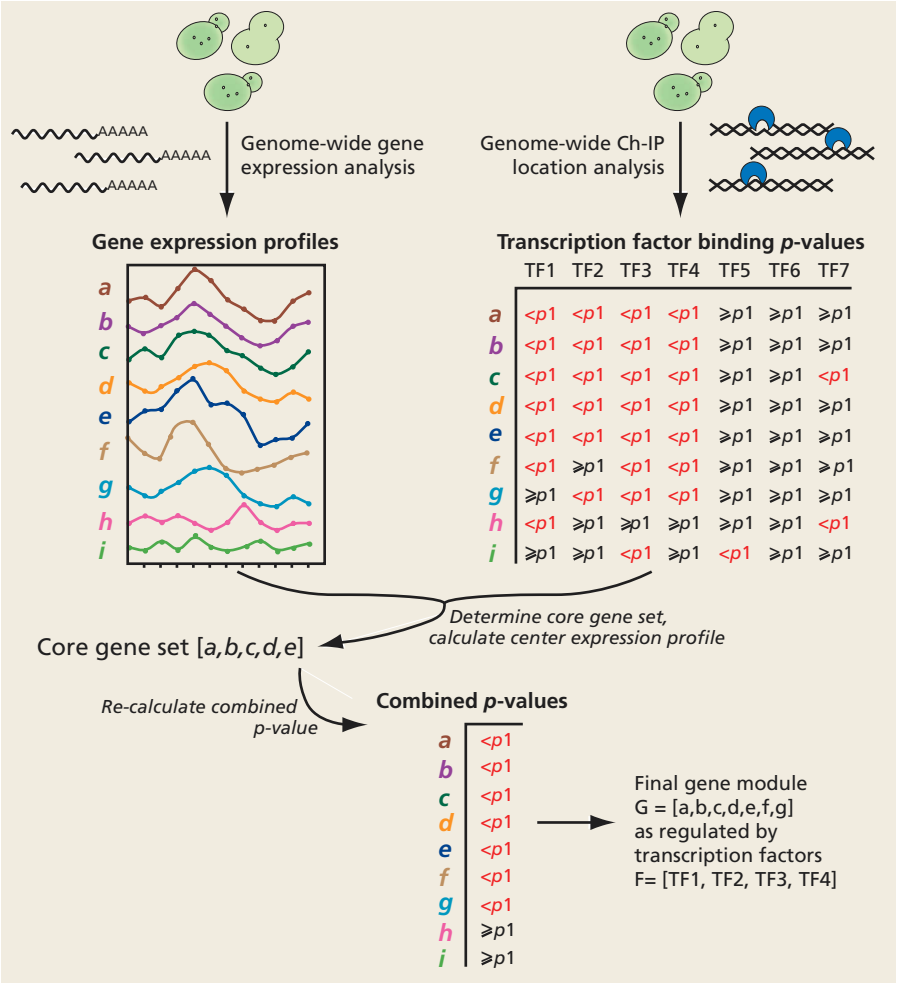

Figure 1 Schematic describing the GRAM algorithm. The data from gene expression and ChIP-chip experiments are presented on the top as stacked expression profiles and a *P* value table, respectively. In the *P* value table, the confidence values that are less than the strict threshold ($p1$) are colored red. In the ChIP-chip experiments, *P* values were calculated for each spot on the microarray to represent the confidence value (the smaller the *P* value, the more likely the observed DNA binding is real)[5]. The GRAM algorithm first selects a 'core set' of genes that share a common group of transcription factors and also have similar expression profiles. In this example, the core set consists of genes *a*, *b*, *c*, *d* and *e* but not *f* and *g* because only the first five genes have *P* values strictly less than $p1$ for the subset of regulators TF1, TF2, TF3 and TF4. A center expression profile is then computed from this core set of genes. The algorithm then revisits the *P* value table to recompute a combined *P* value for every gene with respect to the subset of regulators. A gene is added to the selected set if its expression profile is close to the center expression profile and the combined *P* value is less than $p1$. The final selected set of genes is exported as a gene module. The above procedures are repeated for every possible combination of transcription factors in yeast to derive the complete regulatory network. **[AU: Edits OK?]**

Joseph *et al.* improved previous algorithms incorporating both DNA-binding data and gene expression data. Their new algorithm, called GRAM (genetic regulatory modules), works in three steps, as shown in **Figure 1**. As described in their paper[1], the authors reconstructed a yeast rich media regulatory network using DNA-binding data from 106 transcription factors and over 500 gene expression data sets. The final regulatory network contains 655 distinct genes partitioned into 106 modules, and 68 transcription factors are placed in the network representing regulatory hubs (see Figure 1 in original paper; ref. 1). They carried out gene-specific ChIP experiments to verify a number of selected regulatory interactions predicted by GRAM.

The power of GRAM is evident in the fact that 40% of the 1,560 unique regulatory interactions it identifies in yeast would not have been detected using only the DNA-binding data. Another advantage of the combined approach is that it can also predict directionality of the edges in the network; that is, it can be inferred whether the genes in a module are upregulated or downregulated by examining their expression correlations. An important benefit of having a complete genetic network of an organism is its potential to provide clues on a gene's role in, for example, signal transduction pathways and thereby identify its interaction partners.

It is accepted that genes in the same network module generally have similar cellular functions. This has also been observed among network modules generated by GRAM. Notably, the authors found that in most cases in which a gene module is regulated by more than one transcription factor, previous evidence could always be found suggesting potential physical or functional interactions between these transcription factors. **[AU: Clarify – from the data sets used here or from previous published research?]** All these observations prove that the regulatory networks produced by GRAM are biologically relevant and promise to serve as a blueprint to direct future experiments.

Like microarrays in the late 1990s, it is almost certain that the new ChIP-chip technology will quickly catch on with researchers worldwide, and before long, hundreds of genome-wide DNA-binding data sets will be available. Powerful and sophisticated computer algorithms, such as GRAM, will be needed to analyze these data.

Finally, many other research avenues can be pursued. For example, these tools can be applied to determine the degree of conservation of modular network structures or regulatory interactions among closely related species, such as *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. This type of comparative analysis can potentially shed light on the evolution of regulatory networks. Also, the current knowledge on genetic networks does not paint a truly dynamic picture of the processes taking place inside a cell. Existing technologies and algorithms, such as GRAM, are the first steps toward the development of tools capable of capturing the dynamics of genetic regulatory networks. **[AU: Edits throughout paragraph OK?]**

1. Bar-Joseph, Z. *et al. Nat. Biotechnol.* 21, '–' (2003). [add pages]
2. Chu, S. *et al. Science* 282, 699–705 (1998).
3. Ren, B. *et al. Science* 290, 2306–2309 (2000).
4. Iyer, V.R. *et al. Nature* 409, 533–538 (2001).
5. Horak, C.E. *et al. Genes Dev.* 16, 3017–3033 (2002).
6. Lee, T.I. *et al. Science* 298, 799–804 (2002).
7. Qian, J., Dolled-Filhart, M., Lin, J., Yu, H. & Gerstein, M. *J. Mol. Biol.* 314, 1053–1066 (2001).
8. Ihmels, J. *et al. Nat. Genet.* 31, 370–377 (2002).
9. Pilpel, Y., Sudarsanam, P. & Church, G.M. *Nat. Genet.* 29, 153–159 (2001).
10. Yu, H., Luscombe, N.M., Qian, J. & Gerstein, M. *Trends Genet.* 19, 422–427 (2003).