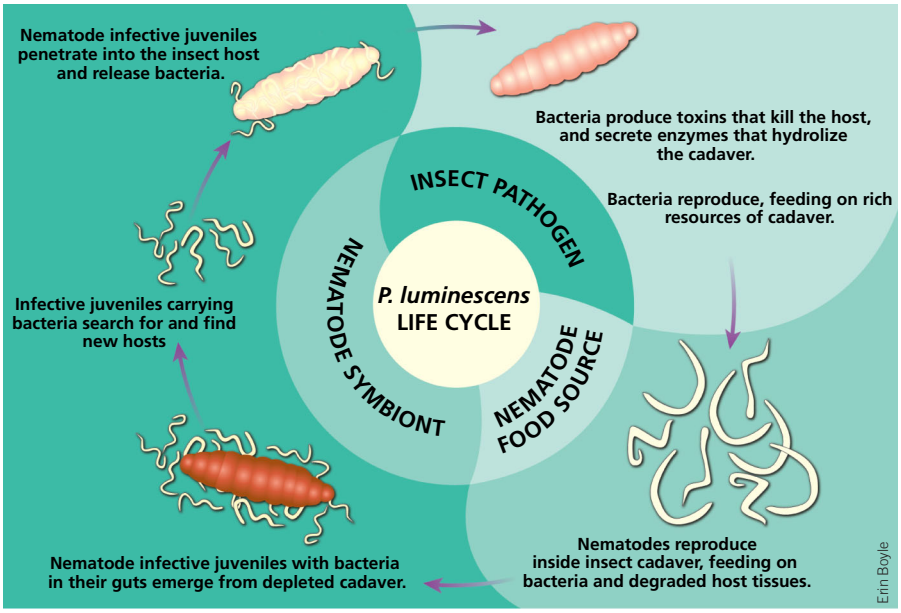

Figure 1 Symbiont or pathogen? During its complex life cycle, *P. luminescens* rotates between being a nematode symbiont, an insect pathogen and a nematode food source.

bacterium forms an intimate association with the gut of the nematode, and a large number of genes with potential roles in this symbiotic interaction have been identified, including those encoding proteins that probably function in surface adhesion. The availability of the complete sequence of the *P. luminescens* genome thus represents an important milestone toward exploiting the potential of this bacterium as a biocontrol agent.

Access to new gene sequences encoding potential protein toxins has further implications for bioengineering insect resistance in plants. Transgenic plants expressing *B. thuringiensis* genes are among the most successful and widely applied biotechnological products[7]. In some instances, transgenic *B. thuringiensis* crops have had a large impact on yield and have resulted in less pesticide use. But there is concern that insect resistance to *B. thuringiensis* toxins in transgenic plants, arising from changes in insect populations, will reduce the effectiveness of this toxin and its transgenic products. In addition, *B. thuringiensis* toxin proteins are generally effective against a narrow range of insects, and toxins have not been identified or developed against some insect pests. Several of the predicted toxin proteins in *P. luminescens* have been shown to have oral toxicity, but toxicity upon expression in transgenic plants has not yet been reported. The best-characterized toxic proteins from *P. luminescens* are large, and their expression in plants may be problematic as was the case initially with *B. thuringiensis* peptides[8].

Perhaps the most fascinating story yet to be told from the analysis of the *P. luminescence* genome is how this organism came to acquire the genes that allow it to fill its specialized niche so successfully. Comparison with the genomes of related bacteria indicates that extensive horizontal gene transfer has occurred. For example, *Yersinia pestis*, a flea-colonizing bacterium and the causal agent of plague, is a close relative. Other clues to the evolution of the *P. luminescens* genome are provided by the multitude of pathogenicity islands, phage remains and abundant transposable elements found in the *P. luminescens* genome.

Further genomic analyses will provide answers to such questions as how and why homologs of an insect juvenile hormone esterase gene were incorporated into the genome of *P. luminescens* and how *P. luminescens* implements and regulates all its different insecticidal capabilities. The answers will provide the tools to exploit this organism's capabilities to fight insect pests in new, untested ways.

1. ffrench-Constant, R. *et al. FEMS Microbiol. Rev.* **26**, 433–456 (2003).
2. Forst, S. & Clarke, D. in *Entomopathogenic Nematology* (ed. Gaugler, R.) 35–56 (CABI Publishing, Oxon, UK, 2002).
3. Duchaud, E. *et al. Nat. Biotechnol.* **21**, 1307–1313 (2003).
4. Kaya, H.K. & Gaugler, R. *Annu. Rev. Entomol.* **38**,181–206 (1993).
5. Bowen, D. *et al. Science* **280**, 2129–2132 (1998).
6. Daborn, P.J. *et al. Proc. Natl. Acad. Sci. USA* **99**, 10742–10747 (2002).
7. Sheldon, A.M., Zhao, J.-Z. & Roush, R.T. *Annu. Rev. Entomol.* **47**, 845–881 (2002).
8. Estruch, J.J. *et al. Nat. Biotechnol.* **15**, 137–141 (1997).

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.

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. 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.

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

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

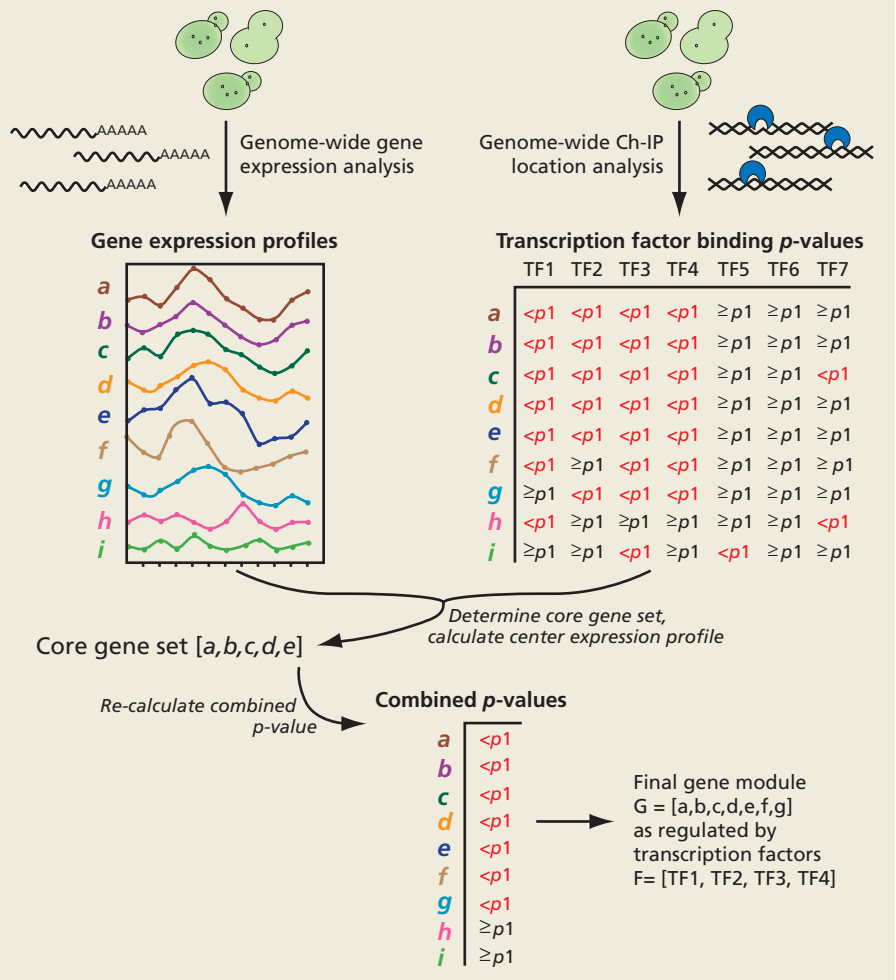

Figure 1 Schematic describing the GRAM algorithm. The data from gene expression and ChIP-chip experiments are presented on the top as stacked expression profiles and a *P* value table, respectively. In the *P* value table, the confidence values that are less than the strict threshold ($p1$) are colored 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.

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 overreliance 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-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. 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.

1. Bar-Joseph, Z. *et al. Nat. Biotechnol.* 21, 1337–1342 (2003).
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).

Playing tag with the yeast proteome

Brenda J Andrews, Gary D Bader & Charles Boone

Two tagged proteome studies offer the most intimate and detailed view into the inner works of yeast cells to date.

Proteomics—the study of the complement of expressed cellular proteins (or proteome)—has catapulted to the forefront of biological research. This advance is due to the development of enabling technologies for producing large-scale data sets of protein activities and to the increasing number of annotated genome sequences that can serve as prerequisite proteome 'blueprints'. Pioneering methods for analysis of the proteome have been developed in yeast and have relied on the systematic cloning of open reading frames (ORFs) for subsequent expression or generation of genomic sets of strains expressing tagged proteins suitable for a variety of array-based manipulations. In two recent *Nature* papers, the Weissman and O'Shea groups[1,2] report two notable additions to the arsenal of tools available for the comprehensive analysis of gene and protein function in yeast. The authors describe two collections of yeast strains in which each ORF is fused with affinity or fluorescence tags, thereby providing the most comprehensive and sensitive view yet of the expressed proteome and its subcellular location in a eukaryotic cell.

In the past few years, myriad genetic and biochemical methods have been used to query genomic sets of proteins for biochemical activity and protein-protein interactions. Notable landmarks on the road to the functional description of the yeast proteome include large-scale two-hybrid screens, immunoprecipitation–mass spectrometric analysis of protein complexes and the generation of tagged sets of proteins for production of functional protein chips (reviewed in ref. 3). The generation of protein complex interaction maps and functional surveys of proteins for DNA binding and other activities are providing a rich, but relatively static, view of the yeast 'interactome'. A more complete 'cell biological' view of the proteome will emerge from integration of proteomics information with functional genomics data derived from transcriptional profiling and gene disruption projects, as well as a picture of the subcellular distribution of proteins and their relative abundance.

In a tour-de-force of strain construction, Ghaemmaghami *et al.*[1] used a PCR-based homologous recombination strategy to insert a tandem affinity purification (TAP) tag at the C termini of all predicted yeast ORFs. They reasoned that an explanation of the biological properties of the proteome would require not only a description of macromolecular complexes and their subcellular location, but also an experimental description of the expressed proteome and a reasonable measure of the absolute levels of proteins in the cell. Two features of the strain collection allow both a survey of expressed proteins in a particular physiological circumstance and a measure of their cellular abundance. First, the tagged proteins are expressed from their native promoters in their endogenous chromosomal location and should be responsive to normal regulatory circuitry. Second, each ORF is marked with a common tag allowing measurement of the absolute abundance of each protein using quantitative western-blot analyses (see http://yeastgfp.ucsf.edu/). A sensible set of test cases suggests that the regulation and activity of most yeast proteins is unperturbed by the C-terminal tag, which bodes well for the utility of the strain set in future genetic and cell biological studies and is good news for the many other projects that have used convenient tags to study gene and protein function.

The authors were able to successfully TAP-tag 6,109 of the 6,243 predicted ORFs and observed a protein product for 4,251 or 70% of the tagged proteome in log-phase yeast cells grown in optimal laboratory conditions[1]. A

Brenda J. Andrews is at the Department of Medical Genetics & Microbiology, and Charles Boone is at the Department of Medical Genetics & Microbiology and the Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario M5S 1A8, Canada. Gary D. Bader is at the Computational Biology Center, Memorial Sloan-Kettering Cancer Center, Box 460, 1275 York Ave., New York, New York 10021, USA.
e-mail: charlie.boone@utoronto.ca; brenda.andrews@utoronto.ca