Deciphering gene expression regulatory networks
John J Wyrick* and Richard A Young†

In the past year, great strides have been made in our understanding of the regulatory networks that control gene expression in the model eukaryote Saccharomyces cerevisiae. The development and use of a number of genomic tools, including genome-wide location and expression analysis, has fueled this progress. In addition, a variety of computational algorithms have been devised to mine genomic sequence for conserved regulatory motifs in co-regulated genes. The recent description of the genetic network controlling the cell cycle illustrates the tremendous potential of these approaches for deciphering gene expression regulatory networks in eukaryotic cells.

Addresses
Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA; Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
* e-mail: jwyrick@mit.edu
† e-mail: young@wi.mit.edu

0959-437X/02/$ – see front matter © 2002 Elsevier Science Ltd. All rights reserved.

Abbreviations
GMEP genome-mean expression profile
SVD singular value decomposition

Introduction
Cells respond to environmental changes by reprogramming expression of specific genes throughout the genome. The transcription rate of a particular gene is determined by the interaction of diverse regulatory proteins — transcriptional activators and repressors — with specific DNA sequences in the gene’s promoter. How a collection of regulatory proteins accomplishes the task of regulating a set of genes can be described as a regulatory network (Figure 1). Some investigators are beginning to map the regulatory networks that govern gene expression throughout living cells. For example, a mathematical model of the regulatory network that controls the lysis/lysogeny decision in bacteriophage-λ has been constructed [1,2]. Regulatory networks in eukaryotic cells are much more complex than networks in bacteriophage: even the simple eukaryote Saccharomyces cerevisiae has >200 proteins that regulate transcription of its ~6200 genes.

Our ability to map gene-regulatory networks in eukaryotic cells has been enhanced greatly by the sequencing of genomes and the development of new tools to study genome expression. The yeast S. cerevisiae was the first eukaryote to have its genome sequenced [3], and has proven to be a workhorse for functional genomics. The past few years have seen an explosion of genome-wide expression data from yeast cells exposed to dozens of different environmental stimuli (e.g. see [4,5]) or deleted for one of many hundred different genes (e.g. see [6]). More recently, DNA microarrays have been used to profile the genomic binding sites of transcription factors and other DNA-binding proteins [7••–9••,10]. Computational algorithms have been developed that identify potential regulatory sequences in promoter regions throughout the genome (reviewed in [11]). In this review, we discuss the contributions of expression analysis, genome-location studies, and computational analysis of promoter sequence elements to our understanding of gene-expression regulatory networks in S. cerevisiae.

Genome-wide expression analysis
Genome-wide expression analysis involves the use of oligonucleotide or cDNA microarrays to measure, in a massively parallel fashion, the mRNA levels of many or all genes in a genome [12–15] (see Figure 2a). Genome-wide expression analysis has been used to investigate the regulatory networks controlling a variety of cellular processes in yeast, including the cell cycle [16–18], phosphate metabolism [19], galactose metabolism [20•], zinc metabolism [21], copper ion homeostasis [22], amino acid biosynthesis [23], sporulation [24,25], glucose repression [26], response to pheromone [27], and the general stress response [4,5]. This trove of data has been analyzed with a variety of clustering and pattern-finding algorithms to group together genes with similar patterns of expression [28–31].

As mRNA levels are the output of gene-expression regulatory networks, it is theoretically possible to use expression data to reverse engineer the architecture of the controlling regulatory networks. A number of groups have tackled this problem using singular value decomposition (SVD) analysis [32–35]. In these studies, SVD analysis was used to find underlying patterns or ‘modes’ in expression data, with the intention of linking these modes to the action of transcriptional regulators. An alternative approach is to use prior knowledge of the regulatory network’s architecture to design competing models, and then use Bayesian belief networks to pick the model that best fits the expression data [36]. Gifford and co-workers have used this approach to distinguish between two competing models for galactose regulation [37]. Friedman and co-workers have used Bayesian networks to analyze genome-wide expression data in order to identify significant interactions between genes in a variety of metabolic and regulatory pathways [38,39].

Genome-wide location analysis
The information provided by expression analysis is the product of all the regulatory events that impinge on gene expression. To understand how genes are controlled by transcriptional regulatory proteins, an additional, more direct measure is needed. To this end, a number of techniques have been developed to identify the genomic
binding sites of transcriptional regulators [7••,8••,40–42]. We believe that the most powerful of these methods is genome-wide location analysis [7••,8••]. This approach combines a chromatin immunoprecipitation protocol, which has been previously used to study protein–DNA interactions at a small number of specific DNA sites [43], with DNA microarray analysis (Figure 2b). DNA is enriched by immunoprecipitation with an antibody against the transcription factor of interest. Enriched and unenriched (control) DNA is labeled with Cy5 and Cy3 dyes, respectively, and then is hybridized to DNA microarrays containing the complete set of yeast intergenic regions. Intergenic regions enriched for Cy5 fluorescence relative to Cy3 correspond to transcription factor binding sites.

The binding of a transcriptional activator to the promoter region of a gene suggests that the activator has a regulatory effect on the gene, but it is also possible that the activator does not fully or even partially control the gene. Identifying the set of promoters where factor binding correlates with gene-expression increases the probability that the factor binding site is associated with adjacent gene expression [7••,9••]. This approach also helps resolve potential ambiguities that result from the fact that some of the intergenic regions contain the promoters for two divergently transcribed genes. Furthermore, the sets of promoters/genes found within the intersection of binding and expression datasets contains few false positives, because the noise that is inherent in each of the microarray datasets tends to be either reduced or eliminated within this intersection.

Initial studies using genome-wide location analysis focused on Gal4 [7••]. Gal4 is a well-characterized transcriptional activator that regulates genes required for galactose metabolism. Genome-wide location analysis of Gal4 identified a set of ten genes that were bound by Gal4 and whose expression was induced in galactose media. All seven of the known Gal4 target genes were identified, in addition to three novel target genes (PCL10, MTH1 and FUR4), which are functionally linked to galactose metabolism. The fact that a single genome-wide location experiment
Chromosomes and expression mechanisms

Several interesting themes have emerged from the location analysis experiments that have been reported to date. One striking observation is that the sequence which is bound by a specific factor occurs at many more sites in genomic DNA than are actually bound by the factor. For example, the Gal4 consensus binding motif (CGGN_{11}CCG), determined from structural studies [44], is found in the promoters of >200 genes (Saccharomyces cerevisiae Promoter Database: http://cgsigma.cshl.org/jian/; [45]), but Gal4 binds to and regulates only 10 of these promoters in vivo [7**]. Similar observations have been made for other transcriptional activators [8**,46**], suggesting that selective binding to a subset of potential target sequences in vivo is a general feature of transcriptional activators. Lieb et al. [46**] have investigated this issue using the transcriptional activator Rap1 as a model.

They surveyed the yeast genome for motifs that fit the Rap1 consensus binding sequence ACACCCRYACAYM. Although many of these matches were identified in intergenic sequences, a substantial number were located in open-reading frames (ORFs). However, genome-wide location analysis of Rap1 revealed that few of these ORFs (14%) were bound in vivo. Among intergenic regions, Rap1 showed a preference for binding to motifs found in promoters (46%) relative to motifs found in non-promoter intergenic regions (17%). This binding preference could not be accounted for by either the sequence or frequency of the motifs in promoter regions, indicating that the specificity of Rap1 binding is determined, in part, by some unknown genome-wide mechanism.

How do cells modify the binding specificity of transcriptional activators? Two general mechanisms [47] have been proposed: first, through cooperative binding of multiple...
transcriptional activators with different binding specificities, or second by inhibiting binding at selective sites through the formation of repressive chromatin structure. The first mechanism parallels studies in mammalian systems in which multiple transcription factors bind cooperatively to a promoter as an enhanceosome [48]. The second mechanism is particularly attractive in light of the recently described histone code, in which it has been proposed that covalent modifications of histone amino-terminal tails serve as a fundamental mechanism regulating protein binding [49]. Whether either (or both) of these mechanisms is used in yeast cells to modify transcription factor binding remains an open question.

A second theme emerging from genome-wide location studies is that transcriptional activators bind to a subset of their target promoters even in non-inducing conditions. This has been demonstrated most clearly for the transcriptional activator Ste12, which regulates pheromone-responsive genes. Genome-wide location analysis of Ste12 following pheromone treatment identified 29 Ste12 target genes, which function in a variety of cellular processes involved in mating [7••]. Ste12 binds to the promoters of a subset of its target genes (e.g. STE12, PCL2, FIG2 and FUS1) prior to pheromone treatment, and the expression of these genes is induced immediately following pheromone treatment [7••,27]. Gal4 [7••] and Ralp [46•] are also pre-bound to a subset of their target promoters, indicating that this may be a general feature of transcriptional activators. The pre-bound target genes are not activated in non-inducing conditions because the negative regulators Dig1/Dig2 and Ga80 act to inhibit the function of Ste12 and Gal4, respectively, prior to induction. Interestingly, Ste12 binds to its own promoter and induces its transcription immediately following pheromone treatment. This may explain the increase in the number of sites bound by Ste12 following pheromone treatment, and is an interesting example of a positive feedback loop in gene-expression networks.

**DNA motif-finding algorithms**

A gene’s expression pattern is largely determined by short promoter sequences (or motifs) that serve as transcription-factor binding sites. Hence, identifying and characterizing regulatory motifs is an important step in deciphering gene-expression regulatory networks. Two general approaches have been used to identify conserved regulatory motifs: first, comparing promoter DNA sequences of closely related species, and second, comparing promoter DNA sequences of co-regulated genes in the same species. The first of these methods takes advantage of the continuing flood of genomic sequence information to identify evolutionarily conserved promoter sequences, with the assumption that these sequences define potential transcription-factor binding sites. A study by Johnston and co-workers [50•] compared small segments of genomic sequence from several closely related Saccharomyces species. From this analysis, they identified a number of conserved DNA motifs, many of which had been previously identified as transcription factor binding sites. These encouraging results will hopefully spur efforts toward obtaining complete genome sequences for this family of *Saccharomyces* species, so this analysis can be extended genome-wide.

The second method takes advantage of a variety of computational algorithms that have been devised to identify conserved sequence motifs in the promoters of co-regulated genes (reviewed in [11]). In many of these studies, groups of co-regulated genes are defined by clustering analysis of genome-wide expression data sets, and then the promoter sequences of these gene clusters are analyzed for statistically over-represented DNA motifs [16,51–53]. Alternatively, potential DNA motifs can be classified by their genome-mean expression profile (GMEP), which is calculated by averaging the expression profiles of all genes that contain the motif in their promoters [54]. DNA motifs with coherent, non-random GMEPs are likely to represent functional regulatory motifs.

These motif-finding algorithms focus on detecting individual regulatory elements. Many eukaryotic genes, however, are bound by multiple transcription factors that act synergistically to regulate transcription. Two recent papers [55••,56•] describe computational algorithms designed to identify combinations of regulatory motifs. Bussemaker *et al.* [56•] used a linear model, in which regulatory motifs contribute additively to the expression level of a gene, to analyze the cell cycle and sporulation expression data sets. Pilpel *et al.* [55••] searched a database of putative regulatory motifs for motif pairs that had synergistic GMEPs. The identified motif pairs were used to build motif synergy maps and to analyze the causal relationship between individual motifs and expression patterns. Interestingly, they found that combinations of a small number of regulatory motifs could account for a complex set of expression patterns.

**Deciphering the cell cycle regulatory network**

A recent study of the yeast cell cycle illustrates the great potential of using genomic tools to decipher gene-expression regulatory networks [9••]. In this study, a combination of genome-wide location and expression analysis was used to investigate how the nine known cell cycle transcriptional activators (Mbp1, Swi4, Swi6, Mcm1, Fkh1, Fkh2, Ndd1, Swi5, and Ace2) regulate the expression of ~800 cell-cycle genes. The data reveal that distinct sets of these nine transcriptional activators regulate genes expressed in different stages of the cell cycle. Mbp1, Swi4, and Swi6 bind predominately to the promoters of late *G1* genes, Mcm1, Fkh2, and Ndd1 to *G2/M* genes, and Swi5 and Ace2 to *M/G1* genes. Fkh1, on the other hand, binds to genes expressed in the *G1*, *S*, and *G2/M* phases of the cell cycle. Strikingly, the data also revealed that each of these sets of stage-specific transcriptional activators also regulates the expression of one or more activators involved in the next stage of the cell cycle, forming a fully connected regulatory network (see Figure 3a). In addition, the
expression of a variety of protein regulators (e.g., cyclins) of the cell cycle is also regulated by these sets of stage-specific transcriptional activators (see Figure 3b). For example, the Mcm1/Fkh2/Ndd1 complex regulates the M/G1 transcriptional activators Swi5 and Ace2, and the cyclins Clb1 and Clb2, which promote entry into mitosis [57]. Hence, a key insight from this study is that the cell cycle is regulated by a connected circular network of transcriptional activators.

An interesting feature of the yeast cell cycle transcriptional regulatory network is that pairs of activators exhibit partially redundant functions. These transcriptional activator pairs (e.g., Swi4 and Mbp1, Fkh1 and Fkh2, Ace2 and Swi5) have similar DNA-binding domains, similar binding sequences, but only partially overlapping target genes in vivo [9••]. Previous models of cell-cycle regulation suggested that each of these transcriptional activator pairs shares its function with its partner (i.e., they are functionally redundant), as each of these six transcriptional activator genes are not essential for cell-cycle progression. A striking example of this is seen for the FKH1 and FKH2 genes. Deletion of either gene individually has little effect, but deletion of both results in striking changes in the expression of a variety of protein regulators (e.g., cyclins) of the cell cycle.

Conclusions

It seems likely that, given the recent flurry of genome-wide location and expression papers this past year, the functional target genes of all yeast transcriptional activators might be identified in the very near future. Such data would provide a foundation for a complete map of the regulatory networks controlling gene expression in a eukaryotic cell. Analyzing these data, however, will be a considerable challenge, and new and more powerful computational tools are needed to build a gene-expression network using these data. Hence, the interface between mathematical modeling and DNA microarray experiments should prove to be fertile ground in the years ahead.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

* of special interest
** of outstanding interest


8. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO: ** Genomic binding sites of the yeast cell-cycle transcription factors SBF and MCB. Nature 2001, 409:533-538. These papers [7*,8*] provide the first descriptions of the genome-wide location analysis technique. Ren et al. [7*] use genome-wide location analysis to define the target genes of the yeast transcriptional activators Gal4 and Ste12. Iyer et al. [8*] use this technique to identify the genes regulated by the cell cycle transcription factors SBF and MCB.


19. Ogawa N, DeRisi J, Brown PO: ** Genomic location analysis technique. Ren et al. [7*] use genome-wide location analysis to define the target genes of the yeast transcriptional activators Gal4 and Ste12. Iyer et al. [8*] use this technique to identify the genes regulated by the cell cycle transcription factors SBF and MCB.


This paper outlines a novel computational algorithm to find synergistic motif combinations in the promoters of co-regulated genes. The authors apply this algorithm to yeast expression data sets from the cell cycle, sporulation, and stress response, in order to map the connections between the transcriptional regulators controlling these biological processes. They find that these maps are highly interconnected, indicating that combinatorial transcriptional regulation is a feature of yeast regulatory networks.


The authors describe a computational method for identifying new regulatory motifs. In this method, genome-wide expression data is fit with a linear model in which multiple DNA motifs can contribute additively to a gene's expression level. They test this method on the diauxic shift, cell cycle, and sporulation expression data sets, and identify several novel regulatory motifs.