

# Functional genomics as applied to mapping transcription regulatory networks

Nila Banerjee and Michael Q Zhang\*

The sequencing of the human genome and the entire genomes of many model organisms has resulted in the identification of many genes. Many large-scale experiments for generating gene disruptions and analyzing the phenotypes are underway to ascertain gene function. A future challenge will be to determine interaction and regulation of all the genes of an organism. Recent advances in functional genomic technology have begun to shine light on such gene network problems at both transcriptomic and proteomic levels. Functional genomics will not only elucidate what the genes do, but will also help determine when, where and how they are expressed as an orchestrated system. In this review, we discuss the functional genomics approaches to extract knowledge about transcription regulatory mechanisms from combinations of sequence data, microarray data and ChIP data. We focus in particular on the budding yeast *Saccharomyces cerevisiae*.

## Addresses

Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA  
\*e-mail: mzhang@cshl.org

**Current Opinion in Microbiology** 2002, **5**:313–317

1369-5274/02/\$ – see front matter  
© 2002 Elsevier Science Ltd. All rights reserved.

**Published online 7 May 2002**

## Abbreviations

**ChIP** chromatin immunoprecipitation  
**TF** transcription factor

## Introduction

The sequencing of the human genome and the entire genomes of many model organisms has resulted in the identification of most protein-coding genes. The functional genomics techniques for mapping transcription regulatory networks have evolved on the basis of advances in experimental approaches and the kinds of data generated. Various high-throughput expression techniques, such as serial analysis of gene expression (SAGE) and microarrays, have been developed that exploit the huge body of sequence data and provide rapid, parallel surveys of gene-expression patterns for hundreds of thousands of genes in a single assay. Several computational algorithms have been developed and applied to uncover co-regulated genes or causal relationships from the large-scale gene expression data. As transcription is mainly controlled and regulated by the binding of transcription factors (TFs) to the promoter DNA sequence, significant progress has also been made in identifying these *cis*-regulatory elements in the promoters, providing more insight to gene function and regulation pathways [1]. Recently, large-scale chromatin immunoprecipitation (ChIP), which allows genome-wide location

analysis of TF-binding *in vivo*, has offered an additional powerful tool in dissecting global regulatory networks. In this review, we discuss the functional genomics approaches to extract knowledge about transcription regulatory mechanisms from combinations of sequence data, microarray data and ChIP data. We focus on eukaryotes. In particular, much emphasis is placed on the budding yeast, *Saccharomyces cerevisiae*. With its small genome (predicted to encode roughly 6200 proteins) and tractable genetics, this yeast has played a prominent role in the development of many methodologies for functional genomics [2].

## Clustering expression profiles to identify co-regulated genes

Although mRNA is not the ultimate product of a gene, transcription is the first step in gene regulation and information about the transcript levels is needed in order to understand gene regulatory networks. To understand how a genetic system is regulated, one has to perturb the system and watch how the system responds. After perturbation, the simplest question one can ask is: which genes have been upregulated (or downregulated)? If the perturbation consists of a TF knockout or overexpression, one could in principle identify its target genes (activated or repressed) by sorting through expression levels (relative to the control). The difficulty lies in the fact that many of these targets may not be the direct ones; they may contain secondary targets unless the mRNA samples were collected fast enough or translation was blocked. However, identifying patterns of gene expression and grouping genes into expression classes may provide much greater insight into their biological function, because many genes belonging to the same complex (such as the ribosome) or to the same regulatory pathway tend to have similar or correlated expression profiles. For instance, if two or more genes have correlated (or anticorrelated) expression profiles in different experiments or at different time points, these genes may be co-regulated and, possibly, functionally related. Different metrics, like Euclidean distance, correlation coefficient, ranked correlation coefficient and mutual-information-based measure have been used to quantify the similarity (or distance) between the expression patterns. Once a similarity metric has been chosen, the expression matrix can be analyzed in either an unsupervised or supervised manner [3,4] Unlike unsupervised analysis, the supervised approach groups genes and is guided by previous knowledge of which genes belong together.

Several multivariate techniques have been used for unsupervised analysis of gene-expression data, including hierarchical clustering [5], *k*-means [6] and self-organizing maps [7,8]. DeRisi *et al.* [9] used a cDNA array containing

a complete set of yeast genes to study the time course of the diauxic shift. They selected small groups of genes with similar expression profiles and showed that these genes are functionally related and contain relevant TF-binding sites upstream of their open reading frames (ORFs). Later, more expression studies of yeast and other organisms under various conditions were carried out, including cell cycle [10,11] and sporulation [12]. The expression profiles of cell-cycle-dependent genes are periodic, and Fourier analysis has been successfully applied to discover these genes [11]. Clustering has been applied and groups of functionally related and co-regulated genes have been revealed. Given the numerous clustering algorithms, selection of an appropriate algorithm for exploration of the data at hand is a non-trivial task. Recently, Chen *et al.* [13] introduced algorithm-independent measures to evaluate the quality of clustering results from different algorithms.

Supervised analysis constructs classifiers, such as linear discriminants, decision trees or support vector machines (SVMs) [14], which assign predefined classes to a given expression profile. These classifiers are trained on a subset of data with a previous given classification and tested on another subset with known classification. They are used to assign an unknown member into a known group.

Some genes may only be co-regulated within a subset of experiments or during a subset of time intervals. These gene clusters may only be discovered by so-called two-way clustering methods [15–17].

Supervised or unsupervised approaches are, however, only the initial steps in gene expression analysis. To understand why a cluster of genes share similar expression patterns and how different clusters are related, it is necessary to further study transcription factors that regulate these genes through binding with their promoter *cis*-regulatory elements.

### Identification of *cis*-regulatory elements in promoters

Finding *cis*-regulatory elements in promoters has been a long-standing problem in computational molecular biology [18]. Traditionally, one starts by collecting known upstream sequences and employing motif-discovering algorithms (either based on over-represented words or on alignment) to identify the common *cis*-elements. One can then use the consensus or the position weight matrix (PWM) characterizing the motif to search novel promoter sequences for additional *cis*-elements. Clustering of microarray expression has provided a high-throughput technology for identifying co-regulated genes. The promoter sequences of each gene in a cluster can be immediately fed to a motif discovery algorithm. Motifs that are common to a set of apparently co-expressed genes are plausible candidates for binding sites implicated in transcriptional regulation. Van Helden *et al.* [19] and Brazma *et al.* [20] looked at groups of co-regulated genes to find over-represented oligonucleotide sequences. Both groups

detected new candidate regulatory sites, as well as sites that had already been characterized. Zhang [21] and Wolfsberg *et al.* [22] developed a statistical technique to predict short oligomers that may be involved in the expression of groups of co-regulated genes. Their strategy looked for pentamers and hexamers that are over-represented among the upstream regions of genes whose expression peaks at a particular phase of the cell cycle. Both Spellman *et al.* [11] and Tavazoie *et al.* [6] used their modified versions of the Gibbs sampler to look for longer motifs in the yeast cell-cycle clusters. Holmes and Bruno [23] suggested that the two stages — clustering of expression profiles followed by Gibbs sampling of sequences — may be combined and viewed as operating on the marginal distributions of a joint probabilistic model for both sequence and expression data. In this case, the presence or absence of a motif will have an influence on which cluster a gene is assigned to. The hope was that using an integrated approach and a better-formulated optimization problem will result in significantly improved discriminative power for regulatory signal identification. When positional information is also incorporated, TF-site prediction may be further improved [24].

Given that not all co-regulated gene promoters in a cluster share a common motif (because they may not all be the direct targets of the same TF) and that the same *cis*-element can be functional in genes of different clusters (the same TF can act at different times or under different conditions), there have been other attempts to locate the *cis*-elements. There are a few approaches to mining for regulatory *cis*-elements without clustering expression profiles. Using the cooperative property of certain *cis*-elements, Wagner [25] was able to detect some clustered TF sites in whole eukaryotic genomes using sequence information alone. Bussemaker *et al.* [26•] used an elegant regression model, in which upstream motifs contribute additively to the log-expression level of a gene, to identify statistically significant promoter motifs by a  $\chi^2$ -fit. In another algorithm [27•], *cis*-TF tries to match a TF to its binding sites by correlating the TF expression profile with the binding-site profile, which is defined by the composite expression patterns of all genes containing the site, even when those genes are not mutually correlated. As the method focuses on binding sites rather than genes, these target genes do not have to be co-regulated.

Most of the motif-searching algorithms identify individual protein-binding-site patterns in DNA sequences. But it is known that transcriptional activation in eukaryotic organisms normally requires combinatorial interactions of multiple transcription factors. A few methods [28,29] have been designed for discovery of binding-site patterns for cooperatively acting factors. Recently, GuhaThakurta and Stormo [30] introduced an algorithm, Co-Bind (for *co*-perative *binding*), which utilizes a Gibbs sampling strategy to model the co-operativity between two transcription factors in a training set discriminated against the genome background.

With the genome-wide expression data, synergistic motif combinations that can produce more correlated expression profiles have been screened [31\*\*]. To get a global view of the connections between regulators of the transcriptional networks within the cell in different conditions, motif synergy maps were generated that displayed the motif associations discovered in the study.

In addition to *in silico* identification of putative *cis*-regulatory elements, genome-wide mapping of TF binding sites *in vivo* has recently become possible. Working in yeast, two groups have pioneered the so-called 'ChIP-chip' method that can combine chromatin immunoprecipitation (ChIP) and microarray (DNA chip) technology to efficiently identify all the crosslinked regions in chromatin DNA bound by a given TF. Using this approach, Ren *et al.* [32\*] identified the sites bound by Gal4 and Ste1, and Iyer *et al.* [33\*] analyzed the genomic binding sites of the transcription factors SBF and MBF. Similar technology has also recently been applied to mammalian systems [34,35].

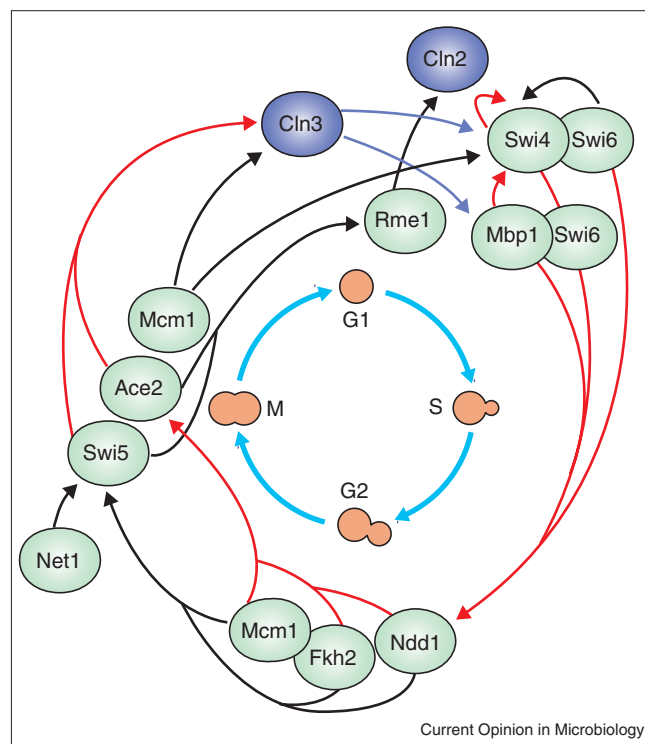
### Mapping regulatory networks

Although substantial work has been done on clustering algorithms and motif-discovery algorithms, a more ambitious goal for functional genomics is to delineate the gene regulation networks. Construction of gene network architectures from expression profiles is often referred to as 'reverse engineering'. One of the first large-scale attempts at reverse engineering was done with real time (RT)-PCR [36]. Later, a random Boolean network model [37] and continuous network models [38] were developed to infer network architecture. As the possible number of networks grows exponentially with the number of genes, it is not possible to derive a unique network with only limited data. To deal with the inherent complexity of network inference, Friedman [39] examined local statistical properties of network components using Bayesian network approaches. With a large set of gene knockout expression data, they were able to extract a finer structure of interactions between genes, such as causality, mediation, activation and inhibition and uncovered some robust regulatory pathways [40].

Even though there has been much progress in developing network models, it is important to note that the current experimental data from which networks are inferred is extremely noisy. Moreover, mRNA expression data alone only gives a partial picture that does not include key post-transcriptional events. Additionally, the amount of samples, even in the largest experiments in the foreseeable future, does not provide enough information to construct a full, detailed model with high statistical confidence. Compounded by these issues, there is a great need to integrate diverse data types and construct tools that will assimilate them into biological models [41].

More recently, Simon *et al.* [42\*\*] have used genome-wide location (ChIP-chip) analysis to determine how the yeast cell-cycle gene-expression program is regulated by each of

Figure 1



Mapping of transcriptional regulatory network in yeast cell cycle. The relationship between TFs (light green) and cyclin genes (purple), as collated from experimental approaches (black arrows) and functional genomics approaches with genome-wide binding (red arrows), are shown. The red arrows and black arrows represent binding of transcriptional activators to the promoter of another regulatory factor. Purple arrows indicate post-transcriptional regulation. The data from genome-wide binding analysis helped establish the sequence in which TFs regulate activators acting in the next cell cycle stage. The figure is adapted from [42\*\*].

the nine major transcriptional activators. By combining all (sequence, expression and TF-binding) data, they found that the cell-cycle transcriptional activators that function during one stage of the cell cycle regulate those that function during the next stage (Figure 1). This study has led to a more coherent picture of the transcriptional regulatory network in yeast mitotic cell-cycle progression.

Taking advantage of genome-wide location and gene expression data, Hartemink *et al.* [43] developed a method to automatically discover regulatory network models. Their Bayesian network models are led into a model induction process that capitalizes on the location data to generate models which would not have resulted from using gene expression data in isolation.

### Conclusions

As the computational approaches to analyzing functional genomics data are further developed and refined, extracting and integrating orthogonal information will become increasingly important. A combination of sequence data,

global expression profiling and binding-site mapping has already produced a more complete picture of the genetic circuitry that is responsible for transcription regulation. Different types of large-scale data can be interrelated to reveal potentially important but not apparent relationships, such as between gene expression and the position of genes on chromosomes [44], or between gene expression and the subcellular localization of proteins [45], or between gene expression and the protein interaction [46]. Ideker *et al.* [47] were able to build, test and refine a model of the galactose utilization pathway in *S. cerevisiae* by integrating both genomic and proteomic approaches. The systematic combination of diverse data types along with new functional genomics approaches will provide a rigorous platform to map transcription regulatory networks more efficiently. But to efficiently dissect large amount of functional genomics data for transcription regulatory network studies, more promoter prediction tools [48], more promoter extraction tools [49] and more specialized promoter databases, such as SCPD (the promoter database of *S. cerevisiae*) [50], are clearly going to be needed.

## Update

Recently, Davidson *et al.* [51] have delineated a gene regulatory network that controls the specification of endoderm and mesoderm in the sea urchin embryo. Large-scale perturbation analyses, in combination with computational methodologies, genomic data, *cis*-regulatory analysis and molecular embryology, were used to generate their network.

## 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

1. Zhang MQ: **Large-scale gene expression data analysis: a new challenge to computational biologists.** *Genome Res* 1999, **9**:681-688.
2. Winzler EA, Davis RW: **Functional analysis of the yeast genome.** *Curr Opin Genet Dev* 1997, **7**:771-776.
3. Brazma A, Vilo J: **Gene expression data analysis.** *FEBS Letters* 2000, **480**:17-24.
4. Quackenbush J: **Computational analysis of microarray data.** *Nat Rev Genet* 2001, **2**:418-427.
5. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95**:14863-14868.
6. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture.** *Nat Genet* 1999, **22**:281-285.
7. Toronen P, Kolehmainen M, Wong G, Castren E: **Analysis of gene expression data using SOM.** *FEBS Lett* 1999, **451**:142-146.
8. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.** *Proc Natl Acad Sci USA* 1999, **96**:2907-2912.
9. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale.** *Science* 1997, **278**:680-686.
10. Cho RJ, Campbell MJ, Winzler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabriellian AE, Landsman D, Lockhart DJ *et al.*: **A genome-wide transcriptional analysis of the mitotic cell cycle.** *Mol Cell* 1998, **2**:65-73.
11. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization.** *Mol Biol Cell* 1998, **9**:3273-3297.
12. Chen G, Jaradat SA, Banerjee N, Tanaka TS, Ko MSH, Zhang MQ: **Comparison of clustering algorithms in analyzing ES cell gene expression data.** *Statistica Sinica* 2002, **12**:241-262.
13. Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I: **The transcriptional program of sporulation in budding yeast.** *Science* 1998, **282**:699-705.
14. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D: **Knowledge-based analysis of microarray gene expression data by using support vector machines.** *Proc Natl Acad Sci USA* 2000, **97**:262-267.
15. Getz G, Levine E, Domany E: **Coupled two-way clustering analysis of microarray data.** *Proc Natl Acad Sci USA* 2000, **97**:12079-12084.
16. Cheng Y, Church GM: **Biclustering of expression data.** *Proc Int Conf Intell Syst Mol Biol* 2000, **8**:93-103.
17. Califano A, Stolovitzky G, Tu Y: **Analysis of gene expression microarrays for phenotype classification.** *Proc Int Conf Intell Syst Mol Biol* 2000, **8**:75-85.
18. Zhang MQ: **Computational methods for promoter recognition.** In *Current Topics in Computational Molecular Biology*. Edited by Jiang T, Xu Y, Zhang MQ. Cambridge: MIT Press; 2002:249-267.
19. van Helden J, Andre B, Collado-Vides J: **Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies.** *J Mol Biol* 1998, **281**:827-842.
20. Brazma A, Jonassen I, Vilo J, Ukkonen E: **Predicting gene regulatory elements *in silico* on a genomic scale.** *Genome Res* 1998, **8**:1202-1215.
21. Zhang MQ: **Promoter analysis of co-regulated genes in the yeast genome.** *Comput Chem* 1999, **23**:233-250.
22. Wolfsberg T, Gabriellian A, Campbell M, Cho RJ, Spouge J, Landsman D: **Candidate regulatory sequence elements for cell-cycle-dependent transcription in *Saccharomyces cerevisiae*.** *Genome Res* 1999, **9**:775-792.
23. Holmes I, Bruno WJ: **Finding regulatory elements using joint likelihoods for sequence and expression profile data.** *Proc Int Conf Intell Syst Mol Biol* 2000, **8**:202-210.
24. Kielbasa SM, Korbel JO, Beule D, Schuchhardt J, Herzog H: **Combining frequency and positional information to predict transcription factor binding sites.** *Bioinformatics* 2001, **17**:1019-1026.
25. Wagner A: **Genes regulated cooperatively by one and more transcription factors and their identification in whole eukaryotic genomes.** *Bioinformatics* 1999, **15**:776-784.
26. Bussemaker HJ, Li H, Siggia ED: **Regulatory element detection using correlation with expression.** *Nat Genet* 2000, **27**:167-171. This paper presents a novel method for discovering *cis*-regulatory elements that bypasses the need to cluster genes based on their expression profiles. This method is based on a model in which upstream motifs contribute additively to the log-expression level of a gene. It requires a single genome-wide set of expression ratios and the upstream sequence for each gene, and outputs statistically significant motifs.
27. Birnbaum K, Benfey PN, Shasha DE: ***Cis* element/transcription factor analysis (*cis*/TF): a method of discovering transcription factor/*cis* element relationships.** *Genome Res* 2001, **11**:1567-1573. This paper uses a binding-site approach to identify *cis*-elements and the transcription factors that are likely to bind them by finding a high correlation between the expression of a given TF and the composite expression patterns of all genes sharing a motif in their regulatory regions.
28. Klingenhoff A, Frech K, Quandt K, Werner T: **Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity.** *Bioinformatics* 1999, **15**:180-186.
29. Kel A, Kel-Margoulis O, Babenko V, Wingender E: **Recognition of NFATp/AP-1 composite elements within genes induced upon activation of immune cells.** *J Mol Biol* 1999, **288**:353-376.
30. GuhaThakurta D, Stormo GD: **Identifying target sites for cooperatively binding factors.** *Bioinformatics* 2001, **17**:608-621.

31. Pilpel Y, Sudarsanam P, Church GM: **Identifying regulatory networks by combinatorial analysis of promoter elements.** *Nat Genet* 2001, **29**:153-159.
- The authors of this paper introduce a method to identify synergistic motif combinations that control expression patterns during cell cycle, sporulation and various stress-response conditions. They map the associations of these motifs under different processes and find them to be highly connected, suggesting that a small number of TFs might be regulating expression patterns under various conditions.
32. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E *et al.*: **Genome-wide location and function of DNA binding proteins.** *Science* 2000, **290**:2306-2309.
- The authors of this paper and those of [33\*] developed a microarray method that reveals the genome-wide location of DNA-bound proteins and established that this method could be successfully used to monitor binding of gene-specific transcription activators in yeast to study gene regulatory networks, gene function and genome maintenance. A combination of location and expression profiles was used to identify genes whose expression is directly controlled by Gal4 and Ste12 as cells respond to changes in carbon source and mating pheromone, respectively. The results identified pathways that are co-ordinately regulated by each of the two activators and revealed previously unknown functions for Gal4 and Ste12.
33. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO: **Genomic binding sites of the yeast cell-cycle transcription factors-SBF and MBF.** *Nature* 2001, **409**:533-538.
- This study, along with [32\*], established that genome-wide binding data could be successfully exploited to analyze gene regulation.
34. Kel AE, Kel-Margoulis OV, Farnham PJ, Stephanie MB, Wingender E, Zhang MQ: **Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors.** *J Mol Biol* 2001, **309**:99-120.
35. Weinmann AS, Bartley SM, Zhang T, Zhang MQ, Farnham PJ: **The use of chromatin immunoprecipitation to clone novel E2F target promoters.** *Mol Cell Biol* 2001, **21**:6820-6832.
36. Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL, Somogyi R: **Large scale temporal gene expression mapping of central nervous system development.** *Proc Natl Acad Sci USA* 1998, **95**:334-339.
37. Liang S, Fuhrman S, Somogyi R: **REVEAL, a general reverse engineering algorithm for inference of genetic network architectures.** *Pac Symp Biocomput* 1998, **3**:18-29.
38. Kyoda KM, Morohashi M, Onami S, Kitano H: **A gene network inference method from continuous-value gene expression data of wild-type and mutants.** *Genome Inform Ser Workshop Genome Inform* 2000, **11**:196-204.
39. Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian networks to analyze expression data.** *J Comput Biol* 2000, **7**:601-620.
40. Pe'er D, Regev A, Elidan G, Friedman N: **Inferring subnetworks from perturbed expression profiles.** *Bioinformatics* 2001, **17**(Suppl 1):S215-S224.
41. Hasty J, McMillen D, Isaacs F, Collins JJ: **Computational studies of gene regulatory networks: in numero molecular biology.** *Nat Rev Genet* 2001, **2**:268-279.
42. Simon I, Barnett J, Hannett N, Harbison CT, Rinaldi NJ, Volkert TL, Wyrick JJ, Zeitlinger J, Gifford DK, Jaakkola TS, Young RA: **Serial regulation of transcriptional regulators in the yeast cell cycle.** *Cell* 2001, **106**:697-708.
- A comprehensive work in identifying genomic binding sites of nine known cell-cycle transcription factors and exploiting this data to extend understanding of cell cycle regulation.
43. Hartemink AJ, Gifford D, Jaakkola TS, Young RA: **Combining location and expression data for principled discovery of genetic regulatory network models.** *Pac Symp Biocomput* 2002, **7**:437-449.
44. Cohen BA, Mitra RD, Hughes JD, Church GM: **A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression.** *Nat Genet* 2000, **26**:183-186.
45. Drawid A, Jansen R, Gerstein M: **Genome-wide analysis relating expression level with protein subcellular localization.** *Trends Genet* 2000, **16**:426-430.
46. Ge H, Liu Z, Church GM, Vidal M: **Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*.** *Nat Genet* 2001, **29**:482-486.
47. Ideker V, Ranish J, Christmas R, Buhler J, Eng J, Bumgarner R, Goodlett D, Aebersold R, Hood L: **Integrated genomic and proteomic analysis of systematically perturbed metabolic network.** *Science* 2001, **292**:929-934.
48. Davuluri R, Grosse I, Zhang MQ: **Computational identification of promoters and first exons in the human genome.** *Nat Genet* 2001, **29**:412-417.
49. Zhang T, Zhang MQ: **Promoter extraction from GenBank (PEG): automatic extraction of eukaryotic promoter sequences in large sets of genes.** *Bioinformatics* 2001, **17**:1232-1233.
50. Zhu J, Zhang MQ: **SCPD: A promoter database of yeast *Saccharomyces cerevisiae*.** *Bioinformatics* 1999, **15**:607-611.
51. Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh CH, Minokawa T, Amore G, Hinman V, Arenas-Mena C *et al.*: **A genomic regulatory network for development.** *Science* 2002, **295**:1669-1678.