Dissecting the transcription networks of a cell using computational genomics

Hao Li* and Wei Wang†

A great challenge in understanding biological complexity in the post-genome era is to reconstruct the regulatory networks governing the patterns of gene expression. In the past few years, the rapid accumulation of genomic sequence and functional data has led to the development of computational approaches to systematically dissect transcriptional regulatory networks. Effective algorithms have been developed to predict *cis*-regulatory elements in a genome, to identify the target genes of transcription factors, to infer the conditions under which each transcription factor is either activated or deactivated, and to analyze combinatorial regulation by multiple transcription factors. Genomic approaches have profoundly changed the way biologists investigate transcriptional regulation, and global pictures of the transcription networks for several model organisms are beginning to emerge.

Addresses
*Department of Biochemistry and Biophysics, University of California at San Francisco, 600 16th Street, San Francisco, CA 94143-2240, USA
e-mail: haoli@genome.ucsf.edu
†Department of Chemistry and Biochemistry, University of California at San Diego, La Jolla, CA 92093, USA
e-mail: wwang@chem.ucsd.edu
Correspondence: Hao Li

Current Opinion in Genetics & Development 2003, **13**:611–616

This review comes from a themed issue on
Genomes and evolution
Edited by Evan Eichler and Nipam Patel

0959-437X/$ – see front matter

DOI 10.1016/j.gde.2003.10.012

Abbreviations

ChIP	chromatin immunoprecipitation
TF	transcription factor
TFPE	TF perturbation experiment

Introduction

A great challenge in the post-genome era is to understand gene regulation on a genomic scale. Organisms devote a significant fraction of their DNA to encoding *cis*-regulatory programs that both control and coordinate gene expression at the transcript level. The outputs of the *cis*-regulatory program depend on the cellular state and extra-cellular inputs. Typically, an external stimulus activates a signal transduction pathway, which leads to the modification of the activities of several transcription factors. These transcription factors then target a subset of genes in the genome, effecting regulation that is often combinatorial in nature. Figure 1 depicts a simplified picture of transcription regulation at a genomic scale. Dissecting the complexities of transcriptional networks is essential for understanding development, cellular responses to environmental and genetic perturbations, and the molecular basis of many diseases.

To form a comprehensive picture of the transcription networks, one needs to address the following challenges: first, identification of *cis*-regulatory elements in the genome; second, accurate identification of the direct regulatory targets of transcription factors (TFs); third, identification of the cellular and environmental context in which these TFs are either activated or deactivated; and fourth, analysis of how gene expression is tailored to different conditions through combinatorial control by multiple TFs. Here we review recent progresses in developing computational approaches to meet these challenges, driven by the rapid accumulation of sequence and functional genomics data.

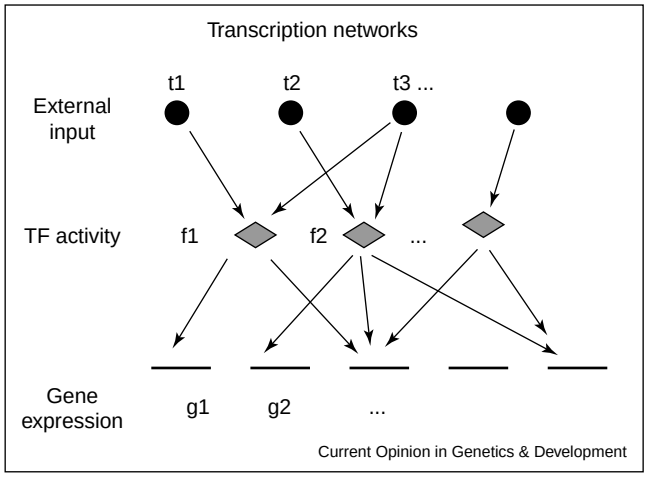
Identifying *cis*-regulatory elements in a genome

Until relatively recently, the identification of *cis*-regulatory elements in a genome has been difficult because these elements are typically short, degenerate, and obey few rules. The availability of large-scale gene expression data from DNA microarrays, complete genome sequences of many species for comparative analysis, and systematic ChIP–chip — chromatin immunoprecipitation followed by hybridization to DNA chip — experiments have led to the development of a large number of computational algorithms to identify *cis*-regulatory elements systematically. These algorithms generally fall into the following categories.

Combining sequence and expression data

A common approach for combining sequence and expression data is to first define groups of co-regulated genes on the basis of similarity in their expression profiles using clustering algorithms [1,2], then to search for enriched sequence patterns in the upstream regulatory regions of genes in a group. The underlying assumption is that genes with similar expression profiles are likely to be regulated by the same TFs. The search algorithms range from enumerating over-represented substrings or regular expression patterns [3–5] to local multiple sequence alignments [6–12]. (Some of these algorithms have been discussed in previous reviews [13,14].)

Figure 1



A diagram of transcription networks of a cell. The transcriptional response of the cell is determined by the cellular state and external input, as represented by the conditions t1, t2 (etc.) Elements f1, f2 (etc.) are transcription factors that are activated under specific conditions. Typically, transcription factors work together in a combinatorial fashion to control the expressions of genes g1, g2 (etc.).

The clustering-based approach has been quite successful in identifying regulatory elements but has its limitations. Clustering is far from an exact and objective process. Genes sharing the same motif may or may not cluster together depending on the expression measurement conditions. Partitioning genes into disjointed clusters may cause loss of information because groups of genes defined by a common motif may not be mutually exclusive, as a result of combinatorial regulation. In addition, clustering is not applicable in situations where only a single microarray measurement is available (e.g. a mutant/wild type comparison, or a ChIP–chip measurement). Several algorithms have been developed to extract regulatory elements without the need for clustering. Bussemaker, Li and Siggia developed the REDUCE (Regulatory Element Detection Using Correlation with Expression) algorithm that can identify combinatorial regulatory elements from a single microarray measurement, based on a linear regression model in which regulatory motifs contribute additively to the log of gene expression [15]. Liu *et al.* developed the MDscan algorithm which combines gene expression data with local multiple sequence alignment to identify TF binding sites from ChIP–chip data [16•]. Recently, Conlon *et al.* generalized the linear regression scheme used by the REDUCE algorithm to evaluate motifs described by position-specific weight matrices (which specify the probability of occurrence of the 4 nucleotides at each single base position) generated from the MDscan algorithm [17].

Single genome statistical analysis

Regulatory elements in a genome may be found on the basis of intra-genome statistics [18–20,21•]. A TF, in general, regulates more than one target and its binding site appears in many places in the genome — thus the binding site motif will be over-represented. Algorithms have been developed to identify putative regulatory elements using genome sequence information only. These algorithms search for over-represented motifs on the basis of certain 'background' models. One example is the Moby Dick algorithm developed by Bussemaker, Li and Siggia. This algorithm treats the genome as if it were a scrambled novel with 'words' representing putative regulatory elements. The algorithm reconstructs the lexicon by finding recurrent words using a probabilistic segmentation model [19,20]. When supplemented by specific knowledge of binding-site motifs, searches for over-represented motifs on the basis of genome-wide statistics can be very effective in finding regulatory elements. For example, using the observation that many DNA-binding proteins in bacteria bind to a bipartite motif with two short segments more conserved than the intervening region, Li *et al.* developed an algorithm that successfully identified about one-third of known regulatory motifs in the *Escherichia coli* genome and predicted many new ones [21•].

Comparative genome analysis

The availability of completely sequenced genomes of closely related species provides a great opportunity for delineating conserved regulatory elements. These elements are more conserved than general noncoding sequences because of functional constraints. Choosing species separated by appropriate evolutionary distances is essential for the success of this approach. The species have to be close enough to achieve sensible alignment of noncoding sequences, but sufficiently diverged such that conserved regulatory elements will stand out from the background.

Comparative genome analysis of regulatory sequences involves the identification of orthologous noncoding regions across species, followed by the search for conserved DNA segments. Some publicly available data sources and analysis tools are reviewed in [22]. The computational algorithms range from BLAST-like [23–26], to Hidden Markov model based [27,28], to local multiple sequence alignment [29]. Algorithms have also been developed where the statistical significance of alignment is evaluated under an appropriate background mutation model that takes into account the relatedness of the species [30]. The resolution at which the regulatory elements can be delineated depends on the type of sequence data available. Pair-wise alignment usually identifies highly conserved segments that are much longer (hundreds to thousands of bases) than the typical length of a TF binding site. This approach has been used by various groups, for example by Loots *et al.* [31] to identify regulatory sequences for interleukins in the human genome by comparison to mouse; by Waterston *et al.* [32•] to systematically analyze conserved noncoding

regions between human and mouse and to estimate the fraction of noncoding regions under selection; and by Kent and Zahler [28] to compare *Caenorhabditis elegans* and *C. briggisae*. When multiple species data are available, finer resolution can be achieved. For example, McCue *et al.* [29] used Gibbs sampler to identify regulatory motifs in the orthologous noncoding regions from several bacterial species [29]. The power of comparing several closely related species with appropriate evolutionary distance is clearly demonstrated by the recent sequencing and comparative analysis of several yeast species [33•,34•], where many known regulatory elements were identified by simply searching for bipartite patterns or oligonucleoties that are more conserved than expected by chance. There are also recent works where targeted genomic regions in multiple mammalian (e.g. [35]) and vertebrate (e.g. [36]) species were sequenced and novel regulatory sequences identified by comparative analysis. Although comparative analysis is quite successful, it is known that many regulatory elements lie outside the conserved regions, and thus will escape detection (E Emberly, N Rajewsky, E Siggia, personal communication).

Predicting *cis*-regulatory modules on the basis of clustering of binding sites

Identification of regulatory elements in metazoans (e.g. fly, mouse and human) is more difficult than in unicellular organisms (e.g. yeast). In contrast to yeast, where *cis*-regulatory elements are typically located a few hundred base pairs away from the translation start site, *cis*-regulatory elements in metazoans can be tens or even hundreds of kilobases away from the genes they regulate. In addition, the binding sites are, in general, not as sharply defined as in yeast. Thus, false positives occur frequently. Recently, notable progress has been made on the basis of the following simple observation. Analysis of the transcriptional program governing early fly embryo development revealed that the *cis*-regulatory elements organize into well separable modules, each defining a specific aspect of the spatio-temporal pattern [37,38]. Such a modular structure has also been revealed, for instance, in the studies of sea urchin development [39–41]. In an early study, Fickett and Wasserman [42] used a combination of muscle-specific TF binding sites to search for muscle-specific genes in the human genome. Recently, several groups [43–45,46•] developed algorithms to search for *cis*-regulatory modules responsible for early fly embryo patterning. Most of the algorithms are based on counting the number of matches of a certain minimum similarity to known motifs in a sequence window. Rajewsky *et al.* [46•] used known motif profiles and a statistical segmentation algorithm (discussed in [19,20]) to compute the likelihood ratio of a given sequence being 'module' versus 'background'. This algorithm circumvents the arbitrary cut-off on motif matches and potentially permits multiple weak motifs to contribute. Frith *et al.* [47,48] have developed an algorithm based on hidden Markov model to analyze clusters in the human genome and have made the tool available free online.

Identifying target genes of TFs

It remains a significant challenge to link predicted *cis*-regulatory elements to the TFs that recognize them. Typically, the potential functions of the predicted elements are evaluated by comparison with known TF binding sites and targets, or by functional analysis of the genes that contain the element. This approach was used, for example, by Kellis *et al.* [33•] and Cliften *et al.* [34•] to assign putative functions for *cis*-regulatory elements identified by comparative analyses.

One exciting development in the past few years has been the invention [49,50] and large-scale application [51•] of the ChIP–chip technology to identify the direct targets of a TF. Recently Lee *et al.* [51•] applied the technology systematically to yeast and published a dataset for 106 TFs, the most comprehensive dataset for TF binding in the yeast genome to date. The ChIP–chip technology is now used to study TF binding in mammalian cells [52–54]. Using DNA microarray containing the proximal promoters of ~5,000 well annotated genes, Li *et al.* systematically identified the targets of c-Myc in Burkitt's lymphoma cells [53]. The amount of ChIP–chip data are rapidly accumulating as various laboratories are using similar approaches to analyze TFs under various conditions. However, these data cannot be used blindly to define the target genes of a TF. It is important to have the ChIP–chip experiment done under the right conditions where the TF is activated. Apart from identifying target genes, it is also non-trivial to accurately locate the binding site of a TF, because ChIP–chip data only allows the identification of TF binding loci with a resolution of ~1 kb. One approach is to first identify a set of potential target genes on the basis of ChIP–chip data and then to search for common sequence patterns in their promoters using local sequence alignment algorithms [51•]. Other algorithms have been developed to identify binding sites ([16•]; W Wang *et al.*, unpublished data) and target genes of a TF (W Wang *et al.*, unpublished data) more effectively by combining ChIP–chip data and sequence information.

ChIP–chip experiments map the genomic location of a TF's binding site, but do not provide direct evidence for the regulation of the genes bound by the TF. A functional assay is a TF perturbation experiment (TFPE). In a TFPE, the expression profile of the wild type is compared to a mutant in which the TF has been perturbed (e.g. either deleted or overexpressed) under conditions where the TF plays a regulatory role. Identification of the binding sites and the direct targets of a TF using TFPE has received less attention because of concerns over the difficulty of distinguishing direct and indirect targets. However, Wang *et al.* recently demonstrated that the

binding site and target genes of a TF can be identified with high specificity by combining promoter sequence analysis with TFPE data ([55•]; W Wang *et al.*, unpublished data). Their work suggests that TFPEs for all the TFs in the genome may be a comprehensive and efficient way to map transcriptional networks on a genomic scale.

Identifying the cellular and environmental context in which a transcription factor is active

Although significant progress has been made in identifying *cis*-regulatory elements and mapping the links from TFs to their targets (the bottom portion of the network diagram in Figure 1), the development of tools to map the links from conditions to transcription factors (the top portion of the network) is still in its infancy. Identification of the cellular and environmental contexts in which each TF is either activated or deactivated is crucial for translating the static information encoded in the DNA sequence into an understanding of the dynamic regulatory network. At present, there is no high-throughput method to measure the activities of all the TFs in a genome directly. mRNA expression level, for example, is insufficient because the activity of TFs is often regulated by post-translational modifications. Several computational approaches have been developed to infer the activities of TFs from microarray expression data indirectly. Wang *et al.* [55•] have developed an inference scheme on the basis of 'local similarity' between the expression data from a TFPE experiment and that from a condition of interest, under the assumption that if the TF is activated under that condition, genes regulated by the TF should have responses similar to those in the TFPE. Barkai *et al.* [56] developed an algorithm to identify groups of genes that are coherently expressed under a subset of conditions. If genes in a group are known to be regulated by a TF, then the TF can be inferred to be active under those conditions. Algorithms have also been developed to search for TFs regulating a gene cluster on the basis of similarity between the expression profile of a TF and that of the cluster [57]. Segal *et al.* used a similar idea to infer potential condition specific regulators [58•]. This approach is limited, for instance by the fact that many TFs are not regulated at the transcript level, and by the difficulty of inferring causality from correlations.

Combinatorial regulation

Combinatorial regulation is known to be an essential feature of transcriptional regulation. Examples include combinatorial control for spatial temporal patterning during development [37–41], and the stress response in yeast [59]. An understanding of combinatorial regulation at a genomic scale is a major challenge, as the number of possible combinations is huge and the cooperation between TFs is context-dependent. With the rapid accumulation of data on gene expression, TFs, and their target genes, it is possible now to systematically analyze genes regulated by multiple TFs and to relate the complex transcriptional response of a gene to the combinations of TF binding sites. We expect that this will become one of the focuses in computational analysis of transcriptional regulation in the next few years.

One straightforward approach to identifying combinatorial regulation is to examine the overlaps between the target genes of different TFs ([51•]; W Wang *et al.*, unpublished data). This approach can be very powerful if TFPE or ChIP–chip data under the right activation condition is available for TFs involved in the regulation. Using ChIP–chip data in conjunction with expression data, Lee *et al.* identified genes bound by a common set of regulators as well as co-expressed throughout the cell cycle, and built a model of a transcriptional network for cell-cycle regulation [51•]. Wang *et al.* integrated TFPE, ChIP–chip and gene expression data to derive a mechanistic model for combinatorial regulation during sporulation (W Wang *et al.*, unpublished data). In a different approach, Pilpel *et al.* [60] screened for pairs of regulatory motifs which may function together on the basis of the assumption that genes sharing both motifs should be more tightly co-regulated. Segal *et al.* developed a scheme to infer a binary decision tree suggesting potential combinatorial regulation [58•]. Taking advantage of multiple yeast species sequence data, Chiang *et al.* [61] searched for potential combinatorial motifs by enumerating pairs of hexameric sequences that are jointly conserved and exhibit non-random spacing.

The context-dependent nature of combinatorial regulation poses a great challenge for reconstructing transcription networks. Because a TF can work together with different TFs to regulate different sets of genes depending on the conditions, context-dependent methods such as TFPE or ChIP–chip experiments (TF binding is also condition-dependent) are essential. On the other hand, because enumerating all different contexts is a daunting task, one needs to develop computational tools to assemble all the partial information into an integrated picture of the network. Context-independent approaches, such as those identifying all TF-binding sites and combinations of sites in the genome on the basis of sequence analysis only, will be indispensable for extending knowledge gained in specific contexts and for suggesting new contexts to be explored.

Conclusions

In the past few years, the availability of genomic sequence and functional data has led to the development of computational approaches to dissecting transcription networks at the system level. For simple model organisms such as yeast, global pictures of the network are beginning to emerge. In the future, there will be continuing efforts to collect increasing amounts of sequence and functional data and develop better theoretical models and computational algorithms to obtain a comprehensive picture of the network, both in uni-cellular and multi-cellular organisms.

We believe one step beyond reconstructing the network is to have a mechanistic understanding of how the network performs its regulatory function. In the long run, analyzing transcriptional networks by combining bioinformatic analysis with physical modeling is likely to yield insights into the basic constraints and underlying principles for how the transcription network and the *cis*-regulatory system of a genome is designed.

Acknowledgements

We thank Erin O'shea and Eric Siggia for helpful comments. H Li acknowledges support from a Sandler's startup fund, a Sandler's opportunity grant, and a David and Lucile Packard Science and Engineering Fellowship. W Wang acknowledges supercomputer time at NCSA through a small allocation grant.

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. 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.

2. 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.

3. 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.

4. Rigoutsos I, Floratos A: **Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm**. *Bioinformatics* 1998, **14**:55-67.

5. Sinha S, Tompa M: **Discovery of novel transcription factor binding sites by statistical overrepresentation**. *Nucleic Acids Res* 2002, **30**:5549-5560.

6. Stormo GD, Hartzell GW III: **Identifying protein-binding sites from unaligned DNA fragments**. *Proc Natl Acad Sci USA* 1989, **86**:1183-1187.

7. Lawrence CE, Reilly AA: **An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences**. *Proteins* 1990, **7**:41-51.

8. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment**. *Science* 1993, **262**:208-214.

9. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers**. *Proc Int Conf Intell Syst Mol Biol* 1994, **2**:28-36.

10. Hughes JD, Estep PW, Tavazoie S, Church GM: **Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae**. *J Mol Biol* 2000, **296**:1205-1214.

11. Liu X, Brutlag DL, Liu JS: **BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes**. *Pac Symp Biocomput* 2001:127-138.

12. Roth FP, Hughes JD, Estep PW, Church GM: **Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation**. *Nat Biotechnol* 1998, **16**:939-945.

13. Zhang MQ: **Large-scale gene expression data analysis: a new challenge to computational biologists**. *Genome Res* 1999, **9**:681-688.

14. Banerjee N, Zhang MQ: **Functional genomics as applied to mapping transcription regulatory networks**. *Curr Opin Microbiol* 2002, **5**:313-317.

15. Bussemaker HJ, Li H, Siggia ED: **Regulatory element detection using correlation with expression**. *Nat Genet* 2001, **27**:167-171.

16. • Liu XS, Brutlag DL, Liu JS: **An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments**. *Nat Biotechnol* 2002, **20**:835-839.
This paper presents the MDscan algorithm and its application to identifying the binding sites of a TF from ChIP–chip data. The algorithm first selects a small set of genes with highest fluorescence ratios and identifies candidate motifs by enumeration. These motifs are then refined by including potential sites in other genes that increase the statistical significance.

17. Conlon EM, Liu XS, Lieb JD, Liu JS: **Integrating regulatory motif discovery and genome-wide expression analysis**. *Proc Natl Acad Sci USA* 2003, **100**:3339-3344.

18. Brazma A, Jonassen I, Vilo J, Ukkonen E: **Predicting gene regulatory elements *in silico* on a genomic scale**. *Genome Res* 1998, **8**:1202-1215.

19. Bussemaker HJ, Li H, Siggia ED: **Regulatory element detection using a probabilistic segmentation model**. *Proc Int Conf Intell Syst Mol Biol* 2000, **8**:67-74.

20. Bussemaker HJ, Li H, Siggia ED: **Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis**. *Proc Natl Acad Sci USA* 2000, **97**:10096-10100.

21. • Li H, Rhodius V, Gross C, Siggia ED: **Identification of the binding sites of regulatory proteins in bacterial genomes**. *Proc Natl Acad Sci USA* 2002, **99**:11772-11777.
This paper presents an algorithm capable of identifying a significant fraction of regulatory sites in a bacterial genome using single genome sequence information only. The algorithm searches for bipartite motifs by combining enumeration with local multiple sequence alignment.

22. Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC: **Cross-species sequence comparisons: a review of methods and available resources**. *Genome Res* 2003, **13**:1-12.

23. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W: **Human-mouse alignments with BLASTZ**. *Genome Res* 2003, **13**:103-107.

24. Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W: **PipMaker – a web server for aligning two genomic DNA sequences**. *Genome Res* 2000, **10**:577-586.

25. Kent WJ: **BLAT – the BLAST-like alignment tool**. *Genome Res* 2002, **12**:656-664.

26. Bray N, Dubchak I, Pachter L: **AVID: a global alignment program**. *Genome Res* 2003, **13**:97-102.

27. Jareborg N, Birney E, Durbin R: **Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs**. *Genome Res* 1999, **9**:815-824.

28. Kent WJ, Zahler AM: **Conservation, regulation, synteny, and introns in a large-scale C. briggsae-*C. elegans* genomic alignment**. *Genome Res* 2000, **10**:1115-1125.

29. McCue L, Thompson W, Carmack C, Ryan MP, Liu JS, Derbyshire V, Lawrence CE: **Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes**. *Nucleic Acids Res* 2001, **29**:774-782.

30. Rajewsky N, Socci ND, Zapotocky M, Siggia ED: **The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons**. *Genome Res* 2002, **12**:298-308.

31. Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA: **Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons**. *Science* 2000, **288**:136-140.

32. • Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P *et al.*: **Initial sequencing and comparative analysis of the mouse genome**. *Nature* 2002, **420**:520-562.

Published by the Mouse Genome Sequencing Consortium, this paper reports the initial sequencing of the mouse genome and human/mouse comparative analysis. The analysis revealed a large number of conserved noncoding sequences, many of them potential regulatory sequences. The availability of the complete mouse and human genome sequences for comparative analysis will change profoundly the way biologists study transcriptional regulation in mammalian systems.

33. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing**
• **and comparison of yeast species to identify genes and regulatory elements**. *Nature* 2003, **423**:241-254.
See annotation [34•].

34. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J,
• Waterston R, Cohen BA, Johnston M: **Finding functional features in Saccharomyces genomes by phylogenetic footprinting**. *Science* 2003, **301**:71-76.
Both [33•] and [34•] report the results of sequencing and comparative analysis of several yeast species. Many known as well as potential regulatory elements were identified by searching for bipartite motifs or oligonucleotides that are more conserved across species than expected by chance. The genome sequences for the seven yeast species are a very valuable resource for analyzing transcription regulation in yeast.

35. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM: **Phylogenetic shadowing of primate sequences to find functional regions of the human genome**. *Science* 2003, **299**:1391-1394.

36. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC *et al.*: **Comparative analyses of multi-species sequences from targeted genomic regions**. *Nature* 2003, **424**:788-793.

37. Stanojevic D, Small S, Levine M: **Regulation of a segmentation stripe by overlapping activators and repressors in the Drosophila embryo**. *Science* 1991, **254**:1385-1387.

38. Small S, Blair A, Levine M: **Regulation of even-skipped stripe 2 in the Drosophila embryo**. *EMBO J* 1992, **11**:4047-4057.

39. Arnone MI, Davidson EH: **The hardwiring of development: organization and function of genomic regulatory systems**. *Development* 1997, **124**:1851-1864.

40. Yuh CH, Bolouri H, Davidson EH: **Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene**. *Science* 1998, **279**:1896-1902.

41. 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.

42. Wasserman WW, Fickett JW: **Identification of regulatory regions which confer muscle-specific gene expression**. *J Mol Biol* 1998, **278**:167-181.

43. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB: **Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome**. *Proc Natl Acad Sci USA* 2002, **99**:757-762.

44. Markstein M, Markstein P, Markstein V, Levine MS: **Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo**. *Proc Natl Acad Sci USA* 2002, **99**:763-768.

45. Halfon MS, Grad Y, Church GM, Michelson AM: **Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model**. *Genome Res* 2002, **12**:1019-1028.

46. Rajewsky N, Vergassola M, Gaul U, Siggia ED: **Computational**
• **detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo**. *BMC Bioinformatics* 2002, **3**:30.
These authors developed a new algorithm for detecting *cis*-regulatory modules responsible for early fly embryo patterning. Modules are assumed to contain multiple binding sites of several characterized TFs in close proximity. For a given sequence segment, the algorithm evaluates the likelihood ratio of the sequence being a module versus background. The algorithm is based on a statistical segmentation model that avoids arbitrary cutoff on motif matches and is flexible in modeling the background.

47. Frith MC, Hansen U, Weng Z: **Detection of *cis*-element clusters in higher eukaryotic DNA**. *Bioinformatics* 2001, **17**:878-889.

48. Frith MC, Li MC, Weng Z: **Cluster-Buster: finding dense clusters of motifs in DNA sequences**. *Nucleic Acids Res* 2003, **31**:3666-3668.

49. 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.

50. 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.

51. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK,
• Hannett NM, Harbison CT, Thompson CM, Simon I *et al.*: **Transcriptional regulatory networks in Saccharomyces cerevisiae**. *Science* 2002, **298**:799-804.
ChIP–chip experiments were performed on 106 TFs in yeast. This is, to date, the most comprehensive data available on TF binding in yeast. The authors also analyzed statistical properties of the transcription networks and constructed a model for the cell cycle by combining ChIP–chip data with gene-expression data. One should keep in mind that TF binding can be condition-dependent, and that a TF may not always bind to its targets under the conditions in which a ChIP–chip experiment is performed.

52. Ren B, Cam H, Takahashi Y, Volkert T, Terragni J, Young RA, Dynlacht BD: **E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints**. *Genes Dev* 2002, **16**:245-256.

53. Li Z, Van Calcar S, Qu C, Cavenee WK, Zhang MQ, Ren B: **A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells**. *Proc Natl Acad Sci USA* 2003, **100**:8164-8169.

54. Weinmann AS, Yan PS, Oberley MJ, Huang TH, Farnham PJ: **Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis**. *Genes Dev* 2002, **16**:235-244.

55. Wang W, Cherry JM, Botstein D, Li H: **A systematic approach to**
• **reconstructing transcription networks in Saccharomyces cerevisiae**. *Proc Natl Acad Sci USA* 2002, **99**:16893-16898.
In this paper, we proposed a systematic approach to reconstructing transcription networks: identifying the binding site and target genes of a TF by modeling promoter sequence and gene-expression data jointly, inferring the activity of a TF based on a 'local similarity' measure, and analyzing combinatorial regulation by examining target genes shared by multiple TFs.

56. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N: **Revealing modular organization in the yeast transcriptional network**. *Nat Genet* 2002, **31**:370-377.

57. Birnbaum K, Benfey PN, Shasha DE: **cis element/transcription factor analysis (cis/TF): a method for discovering transcription factor/cis element relationships**. *Genome Res* 2001, **11**:1567-1573.

58. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D,
• Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data**. *Nat Genet* 2003, **34**:166-176.
These authors of this paper have developed a statistical inference scheme to identify groups of genes responding similarly to a set of conditions and to identify their potential regulators. Starting from a given partitioning of genes into groups, the algorithm searches for potential regulators for each group on the basis of the correlation between the expression profile of the group and the mRNA level of a precompiled set of regulators. The algorithm iteratively refines the partitioning of genes and the assignments of regulators.

59. O'Rourke SM, Herskowitz I, O'Shea EK: **Yeast go the whole HOG for the hyperosmotic response**. *Trends Genet* 2002, **18**:405-412.

60. Pilpel Y, Sudarsanam P, Church GM: **Identifying regulatory networks by combinatorial analysis of promoter elements**. *Nat Genet* 2001, **29**:153-159.

61. Chiang DY, Moses AM, Kellis M, Lander ES, Eisen MB: **Phylogenetically and spatially conserved word pairs associated with gene-expression changes in yeasts**. *Genome Biol* 2003, **4**:R43.