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

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].)
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 Siggsa 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].

**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.).

**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. briggsae. 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 com-
parative analysis of several yeast species [33*,34*], where many known regulatory elements were identified by simply
searching for bipartite patterns or oligonucleotides 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 transcrip-
tional program governing early fly embryo development
revealed that the cis-regulatory elements organize into
describable 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 combina-
tion 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 circum-
vents 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
genoae 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. system-
tically 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 condi-
tions. 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 algo-
rithms [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 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


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.


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.


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.


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 palindromic sequences 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.


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.


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.


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 regulatory by examining target genes shared by multiple TFs.


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 profiles of the group and the mRNA level of a precompiled set of regulatory TFs. The algorithm iteratively refines the partitioning of genes and the assignments of regulators.

