Computational approaches to study transcriptional regulation

M. Madan Babu
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, U.K.

Abstract
In recent years, a number of technical and experimental advances have allowed us to obtain an unprecedented amount of information about living systems on a genomic scale. Although the complete genomes of many organisms are available due to the progress made in sequencing technology, the challenge to understand how the individual genes are regulated within the cell remains. Here, I provide an overview of current computational methods to investigate transcriptional regulation. I will first discuss how representing protein–DNA interactions as a network provides us with a conceptual framework to understand the organization of regulatory interactions in an organism. I will then describe methods to predict transcription factors and cis-regulatory elements using information such as sequence, structure and evolutionary conservation. Finally, I will discuss approaches to infer genome-scale transcriptional regulatory networks using experimentally characterized interactions from model organisms and by reverse-engineering regulatory interactions that makes use of gene expression data and genomewide location data. The methods summarized here can be exploited to discover previously uncharacterized transcriptional pathways in organisms whose genome sequence is known. In addition, such a framework and approach can be invaluable to investigate transcriptional regulation in complex microbial communities such as the human gut flora or populations of emerging pathogens. Apart from these medical applications, the concepts and methods discussed can be used to understand the combinatorial logic of transcriptional regulation and can be exploited in biotechnological applications, such as in synthetic biology experiments aimed at engineering regulatory circuits for various purposes.

Introduction
Regulation of gene expression at the transcriptional level is a fundamental mechanism that is evolutionarily conserved in all the cellular systems [1]. This form of regulation is typically mediated by TFs (transcription factors) that bind to DNA and either activate or repress the expression of nearby genes [2,3]. Extensive research into understanding this process at the biochemical and structural level has provided us with an understanding that now allows us to investigate this process on a genomic scale (see Table 1 for genome-scale experimental strategies to probe protein–DNA interactions). These experimental approaches have been applied to several prokaryotes and eukaryotes, resulting in a wealth of information that is stored in publicly available databases [4–7]. To take advantage of this deluge of information, computational approaches to visualize, analyse and prioritize genes for further experimental characterization, or to understand general principles on a genomic scale, have become indispensable. In the present mini-review, I will provide an overview of the key computational approaches to represent this information and investigate transcriptional regulation on a genomic scale.

Organization of the transcriptional regulatory network
Experiments performed over the previous years have resulted in a large amount of information on protein–DNA interactions and gene regulation in several model organisms [4–9]. In addition, advances in experimental techniques that detect protein–DNA interactions (see Table 1) have provided us with evidence for TF–DNA interactions on a genomic scale [10–13]. This deluge of information is best represented as the transcriptional regulatory network with nodes connected by edges (Figure 1) [13,14]. In such a network representation, the nodes represent either TFs or TGs (target genes) and the directed edges represent a regulatory interaction (protein–DNA interaction) between the TFs and their TGs [13,14].

Representing regulatory interactions as a network provides us with a conceptual framework and the abstraction to understand general principles of regulation on a genomic scale [15]. A number of recent studies on transcriptional networks of prokaryotes and eukaryotes have shown that the structure of such networks can be investigated at least at three distinct levels of organization [14]. At the most basic level, the network consists of a single regulatory interaction between a TF and its TG (Figure 1a). At the local level of organization, studies have uncovered that the basic unit is organized into fundamental units of transcriptional regulation, called network motifs (Figure 1b) [16]. Finally, at the global level of

Key words: cis-regulatory element, computational approach, protein–DNA interaction, reverse engineering, target gene, transcriptional regulation, transcription factor.

Abbreviations used: ChIP, chromatin immunoprecipitation; Dam, DNA adenine methyltransferase; DamID, Dam identification; HMM, hidden Markov model; PSSM, position-specific scoring matrix; TF, transcription factor; TG, target gene.

1 email madanm@mrc-lmb.cam.ac.uk
Table 1 | Genome-scale experimental methods to probe protein–DNA interactions

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chip-chip and Chip-seq</td>
<td>The DNA-binding protein is tagged with an epitope and is expressed in a cell. The bound protein is covalently linked to DNA by using an in vivo cross-linking agent such as formaldehyde. After cross-linking, DNA is sheared and the protein–DNA complex is pulled down using an antibody for the tag. Reversal of the cross-link releases the bound DNA, allowing the sequence of the fragments to be determined by hybridization to a microarray (Chip-chip) or by sequencing (Chip-seq). In Chip-chip experiments, intergenic regions are spotted on to a microarray chip. Following a ChIP step, the bound fragments are reverse cross-linked and hybridized on to the chip. Complementary sequences will bind to specific spots on the chip, thereby providing the exact intergenic region to which the protein was bound [57–59]. In Chip-seq experiments, the bound fragments are directly sequenced using 454/Solexa/Illumina Sequencing Technology. The sequences are then computationally mapped back to the genome sequence. Fragments that were bound by the protein will be sequenced several times providing a direct measure of enrichment [60–62].</td>
</tr>
<tr>
<td>DamID (DNA adenine methyltransferase identification)</td>
<td>To overcome any potential non-specific cross-linking of protein to DNA as could happen with Chip-chip experiments, the DamID technique was introduced. The protein of interest is fused to an E. coli protein, Dam. Dam methylates the N6 position of the adenine in the sequence GATC, which occurs at reasonably high frequency in any genome (~1 site in 256 bases). Upon binding DNA, the Dam protein preferentially methylates adenine in the vicinity of binding. Subsequently, the genomic DNA is digested by the DpnI and DpnII restriction enzymes that cleave within the non-methylated GATC sequence, and remove fragments that are not methylated. The remaining methylated fragments are amplified by selective PCR and quantified using a microarray [63].</td>
</tr>
<tr>
<td>PBMs (protein-binding universal DNA microarrays)</td>
<td>In contrast with the methods described above, this is an in vitro method to probe protein–DNA interaction. A DNA-binding protein of interest is epitope-tagged, purified and bound directly to a double-stranded DNA microarray spotted with a large number of potential binding sites. Labelling with fluorophore-conjugated antibody for the tag allows detection of binding sites from the significantly bound spots [64].</td>
</tr>
</tbody>
</table>

Figure 1 | Organization of the transcriptional regulatory network

(a) The basic unit consists of a regulatory interaction (grey arrow) between a TF (black circle) and a TG (grey circle). (b) The basic unit forms small patterns of regulatory interactions called network motifs. The three most commonly occurring motifs are shown here: feed-forward motif (FFM, top), single input motif (SIM, middle) and multiple input motif (MIM, bottom). (c) The set of all transcriptional regulatory interactions is referred to as the transcriptional regulatory network.

Local network structure

A network motif is defined as a small pattern of interconnections that recur at many different parts of the network at frequencies much higher than what is expected by chance when compared with random networks of similar size [17]. Analysis of the transcriptional networks of *E. coli* and yeast has revealed the presence of three commonly occurring motifs, each of which has distinct kinetic properties in the control of gene expression [16]. These are (i) feed-forward motif (FFM; Figure 1b, top), where a top-level TF regulates both the intermediate-level TF and the TGs, and the intermediate-level TF regulates the TG. If both TFs are activators, such a connectivity pattern might ensure that the TG is expressed only when persistent signal is received by the top-level TF. Since the concentration of the intermediate TF should be built up for the regulation of the final TG, random fluctuations and noise in activation of the top-level TF are filtered and do not get propagated. (ii) Single input motif (SIM; Figure 1b, middle), where a single TF regulates the expression of several TGs simultaneously. Depending on the promoter strength of the regulated genes, it may respond to different concentration levels of the active TF [16]. Therefore, if the concentration of the active TF changes with time, such a motif could set a temporal pattern in the expression of the individual targets. (iii) Multiple input motif (MIM; Figure 1b, bottom), where multiple TFs regulate the expression of...
multiple TGs. Since the TFs could potentially respond to different signals, such motifs could therefore integrate diverse signals and bring about differential expression of the relevant targets. Thus, regulation of genes via network motifs provides distinct ways of regulating gene expression. Since maintaining the right levels of TGs can affect fitness [18], it is very much possible that controlling gene expression via distinct motifs is advantageous under different conditions. Local network properties such as motif identification can be carried out using programs such as Mfinder, FanMod and Cytoscape (Table 2).

**Global network structure**

At the global level of organization, analysis of transcriptional networks has revealed that they display a scale-free topology [14]. In other words, such a network is characterized by the presence of a few highly influential TFs that regulate several genes and a large number of TFs that regulate only a few genes. The highly influential TFs are referred to as global regulators, or regulatory hubs, and their presence contributes to the inherent robustness of such a topology, where robustness is defined as the ability of complex systems to function even when the structure of the system is perturbed significantly [19,20]. A scale-free topology is robust because random inactivation of genes will probably affect the TFs that regulate a few genes as these occur in very high numbers. This would leave a central, highly connected subnetwork that may still be functional. However, the downside of such a network structure is that they are vulnerable to targeted attacks of hubs, i.e. targeted removal of the very highly connected nodes will result in the collapse of the system into small sets of isolated fragments that no longer interact with each other. Therefore the global regulators are believed to be crucial for the robustness and functioning of the regulatory network [21].

Global network properties such as connectivity, modularity and clustering can be investigated using packages such as Cytoscape, Pajek and Topnet (Table 2).

**Computational approaches to identify TFs and cis-regulatory elements**

**Identification of TFs**

The simplest unit of a transcriptional network involves a TF that senses changes in the environment and binds to a cis-regulatory element in the promoter region of the relevant TGs to regulate their expression. Investigations of proteins that function as TFs through functional, biochemical and structural methods have revealed that these proteins are modular and contain at least two domains, where a domain is defined as a structural, functional and evolutionary unit of proteins [22]. Of the two domains, one recognizes and binds to specific DNA sequences called the DNA-binding domain, and the other is an effector domain that senses changes in the internal or external environment [23,24]. Such an effector domain can detect the changes in the local environment by directly binding to a small molecule (e.g. one-component systems) or by being post-translationally modified by other proteins that sense the external milieu (e.g. two-component signal transduction systems) [25–27]. Therefore one can use the information about the domain encoded to identify novel TFs in completely sequenced genomes using remote homology detection methods.

Proteins containing specific domains can be identified by several ways, which includes the profile-based methods such as PSI-BLAST and HMM (hidden Markov model)-based methods such as HMMER and SAM. Domain databases such as PFAM database (Protein Families Database of Alignments and Hidden Markov Models; http://www.sanger.ac.uk/Software/Pfam/) use HMMER and protein sequence information for building its library of HMMs, whereas databases such as Superfamily use SAM-T99 and structural information to build a library of profiles (see Table 2). Several studies have used these domain-based approaches to predict TFs from completely sequenced genomes and have discovered novel TF families in key pathogens [28–32]. The information from such efforts has been stored in publicly available databases, which include BacTregulator, DBD and ArchaeTF (Table 2).

**Identification of cis-regulatory elements**

TFs bind to short DNA sequence motifs, also called binding sites, in promoter regions of transcriptional units. All the different binding sites recognized by the same TF can be conveniently represented as a consensus sequence. Alternatively, they can be represented as PSSMs (position-specific scoring matrices). Such matrices represent the probability of finding a particular nucleotide in a specific position and can be visualized using a logo representation (see Table 2, weblogo and enologos). The methods to detect cis-regulatory elements that could function as possible binding sites are normally referred to as pattern discovery algorithms. These algorithms either rely on probabilistic description of the cis-regulatory element [33,34] or exact words that are statistically over-represented in a set of sequences [35,36]. Using these, computational search procedures that scan promoter regions can detect sequence motifs, which potentially correspond to TF-binding sites. For these methods to work well, it is important that the background frequency of nucleotides and oligonucleotides is corrected accordingly, e.g. the GC content of the organism of interest.

Using these *in silico* methods in combination with external information [e.g. differential expression upon knockout of a TF or regions that are enriched in ChIP (chromatin immunoprecipitation)-chip experiments] could help in identifying motifs with high confidence for a particular TF. In addition, several evolutionary and biological principles can be applied to identify cis-regulatory elements. For instance, since cis-elements may be subjected to evolutionary selection, they evolve less rapidly than the surrounding non-coding regions within closely related organisms. Therefore conserved motifs upstream of orthologous genes in related species are more likely to be true cis-regulatory elements [37]. Similarly, since most TFs in prokaryotes function as dimeric units, they are more likely to recognize palindromic sequences. Therefore
motifs that are closely spaced and complementary are more likely to be functional cis-regulatory elements in bacteria [38]. Several algorithms that use PSSMs or over-represented oligonucleotide sequences to identify regulatory elements exist in the literature. A systematic comparison of some of the key DNA sequence motif-detection algorithms has been carried out recently on co-regulated genes [39]. Although many approaches aim to predict cis-regulatory elements in a given set of sequences, combined approaches that employ different principles but arrive at the same motif are likely to identify true motifs with the highest confidence. Some of the commonly used motif discovery platforms include RSAT, seqVISTA and web-MOTIFS (Table 2).

Computational approaches to investigate transcriptional regulatory networks

While there has been significant progress in unravelling the transcriptional regulatory networks of various model organisms such as E. coli and Bacillus subtilis, much less information is available on the transcriptional networks of other prokaryotes. To gain a better understanding of the transcriptional regulatory network in other organisms, computational methods to extrapolate this information from model organisms to poorly studied organisms have been developed. The two major approaches to infer regulatory networks are described below.

Template-based methods

The template-based approach exploits the principle that orthologous TFs regulate orthologous TGs [32,40–44]. Thus, in this method, one starts with a known regulatory network and transfers information about interactions to orthologous genes in a target genome of interest. Such an approach may or may not explicitly involve the use of binding site information for the TF. Methods that do not use binding site data require the complete genome sequence and the transcriptional regulatory network for a reference organism. The protein sequence of TFs and TGs in the reference network is used to identify orthologous genes in the target organism to infer the conserved regulatory network. The method, which uses binding site data, requires reliable information on the cis-regulatory element for a TF. Such methods exploit the fact that the presence of a similar binding site upstream of different genes in a closely related species would imply that the orthologous TF regulates the TGs through similar binding sites. Both these methods have their advantages and disadvantages; the former method allows prediction of conserved interactions in distantly related organisms, but does not facilitate discovery of novel targets for a given TF. In contrast, the latter method allows detection of novel targets for a TF, but is not applicable to distantly related genomes because cis-regulatory elements are shorter and evolve relatively faster than the protein-coding regions, hence making detection of new interactions difficult. See Table 2 for databases that contain information about experimentally characterized transcriptional interactions and binding site data that can be used for network reconstruction.

Reverse engineering using gene expression data

In this approach, one scans for patterns in gene expression data from time-series experiments and from experiments conducted across different conditions [45–50]. If a gene is consistently differentially expressed (up- or down-regulated) after overexpression or knockout of a TF across several time points or different conditions, a regulatory interaction between the two is inferred. In the case of expression analysis over different experimental conditions, one infers sets of genes with a similar expression profile across many conditions to be co-regulated by the same set of TFs. Such inferences become more accurate as the number of measurements over a certain period of time (the time resolution of the data) increases, since this allows direct regulatory interactions to be distinguished from indirect (multistep) regulation. Variants of this approach make use of information about experimentally well-characterized TF-binding sites to make inferences about regulatory interactions. In this method, promoter regions in the genome of interest are scanned using known binding site profiles of characterized TFs. The set of genes that are predicted to have a binding site and are differentially regulated upon up-regulation or knockout of the TF are inferred to be regulated by the corresponding TF [51].

While the methods mentioned above exploit different principles, there have been considerable efforts to develop combined approaches to predict regulatory interactions with a higher degree of confidence. For instance, while analysing microarray expression data, the initially determined sets of co-regulated genes can be refined by investigating whether the same TF actually binds to all of them by predicting the presence or absence of a binding site in the promoter region of these genes. This approach could be extremely powerful if one explicitly uses genomewide binding location data (e.g. ChIP-chip or ChIP-seq). Such an integrated approach can directly link a detected binding event, with a predicted cis-regulatory element to a change in gene expression of a relevant TG. In this way, one can distinguish directly regulated genes from the indirectly regulated ones or even genes that just randomly happen to show a similar expression profile [8,52–54].

For a current evaluation of reverse-engineering methods, the reader is suggested to visit the DREAM project website (http://wiki.c2b2.columbia.edu/dream/). DREAM is a Dialogue for Reverse Engineering Assessments and Methods with its main objective to catalyse the interaction between experiment and theory in the area of cellular network inference [55].

Conclusions and outlook

With advancements in large-scale experimental methodologies that detect protein–DNA interactions on a genomic scale such as ChIP-seq and tiling arrays, the general methods discussed here will be useful for investigating transcriptional regulation for completely sequenced genomes and...
<table>
<thead>
<tr>
<th>Comment</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biocomputational databases, genome assignments and protein domain databases</td>
<td>Bioconductor, STRING, COG, Pfam, CDD, SCOP, HOMSTRAD, CATH, KEGG, SUPERFAMILY, Gene3D, Domain databases, genome assignments and TF databases</td>
</tr>
</tbody>
</table>
### Microbes Online
- **Domain assignment, expression data, evolutionary relationships and operon structure**

### Interpro
- **Sequence and structural domain assignments**
  - [http://www.ebi.ac.uk/interpro/](http://www.ebi.ac.uk/interpro/)

### BacTRegulators
- **Database of TFs in bacteria and archaea**
  - [http://www.bactregulators.org/](http://www.bactregulators.org/)

### DBD
- **A database of predicted TFs of over 700 completely sequenced genomes based on SCOP DNA-binding domains**
  - [http://dbd.mrc-lmb.cam.ac.uk/DBD/index.cgi?Home](http://dbd.mrc-lmb.cam.ac.uk/DBD/index.cgi?Home)

### Protein Lounge
- **A database of TFs (commercial)**

### Transfac
- **A TF database**

### ArchaeaTF
- **An archaeal TF database**
  - [http://bioinformatics.zj.cn/archaeatf/Homepage.php](http://bioinformatics.zj.cn/archaeatf/Homepage.php)

### cis-Regulatory element identification, visualization and transcriptional network databases
- **Tools for comparative analysis of genomic sequences**

- **A very powerful platform for regulatory sequence analysis**

- **Motif discovery, scoring, analysis, and visualization using different programs**
  - [http://fraenkel.mit.edu/webmotifs/finalout.html](http://fraenkel.mit.edu/webmotifs/finalout.html)

- **Platform for binding site discovery**
  - [http://zlab.bu.edu/SeqVISTA/index.htm](http://zlab.bu.edu/SeqVISTA/index.htm)

- **TF-binding sites and regulatory interactions**
  - [http://regtransbase.lbl.gov/cgi-bin/regtransbase?page=main](http://regtransbase.lbl.gov/cgi-bin/regtransbase?page=main)

- **Atlas of completely sequenced genomes**

- **Visualizing binding site information**
  - [http://weblogo.berkeley.edu/](http://weblogo.berkeley.edu/)

- **Logo visualization**
  - [http://biodev.hgen.pitt.edu/cgi-bin/enologos/enologos.cgi](http://biodev.hgen.pitt.edu/cgi-bin/enologos/enologos.cgi)

- **Database of TFs and binding sites for *E. coli***
  - [http://regulondb.ccg.unam.mx/](http://regulondb.ccg.unam.mx/)

- **Database of TFs and binding sites for *B. subtilis***
  - [http://dbtbs.hgc.jp/](http://dbtbs.hgc.jp/)

- **Database of regulatory network for several microbes**
  - [http://www.coryneregnet.de/](http://www.coryneregnet.de/)

- **Prokaryotic database of gene regulation**
  - [http://www.prodoric.de/](http://www.prodoric.de/)

- **Computationally predicted TF-binding sites in γ-proteobacterial genomes**
  - [http://www.tractor.lncc.br/](http://www.tractor.lncc.br/)
for complex microbial communities. These results and predictions from the computational approaches can serve as a scaffold for experimental studies on transcriptional control in poorly characterized genomes, and could be relevant for designing experiments to investigate regulation in medically important pathogens and for engineering regulatory interactions in organisms with biotechnological value. On a more general level, such methods can be used in synthetic biology experiments that aim to design circuits with specific kinetic and regulatory properties [56] and may identify TFs with new modes of regulation.

I acknowledge the Medical Research Council, Darwin College and Schlumberger for generous support. I thank Arthur Wuster, Rekin’s Janky and Nitish Mittal for critically reading this paper and for helpful suggestions.

References

Received 18 March 2008
doi:10.1042/BST0360758