Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping

Albertha J. M. Walhout
University of Massachusetts Medical School

Let us know how access to this document benefits you.
Follow this and additional works at: https://escholarship.umassmed.edu/oapubs

Part of the Life Sciences Commons, and the Medicine and Health Sciences Commons

Repository Citation

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in Open Access Articles by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.
Unraveling transcription regulatory networks by protein–DNA and protein –protein interaction mapping

Albertha J.M. Walhout

Genome Res. 2006 16: 1445-1454 originally published online October 19, 2006
Access the most recent version at doi:10.1101/gr.5321506

References
This article cites 83 articles, 40 of which can be accessed free at:
http://genome.cshlp.org/content/16/12/1445.full.html#ref-list-1

Article cited in:
http://genome.cshlp.org/content/16/12/1445.full.html#related-urls

Email alerting service
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here

To subscribe to Genome Research go to:
http://genome.cshlp.org/subscriptions

Copyright © 2006, Cold Spring Harbor Laboratory Press
Unraveling transcription regulatory networks by protein–DNA and protein–protein interaction mapping

Albertha J.M. Walhout

Program in Gene Function and Expression and Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA

Metazoan genomes contain thousands of protein-coding and noncoding RNA genes, most of which are differentially expressed, i.e., at different locations or at different times during development, function, or pathology of the organism. Differential gene expression is achieved in part by the action of regulatory transcription factors (TFs) that bind to cis-regulatory elements that are often located in or near their target genes. Each TF likely regulates many targets in the context of intricate transcription regulatory networks. Up to 10% of a genome may encode TFs, but only a handful of these have been studied in detail. Here, I will discuss the different steps involved in the mapping and analysis of transcription regulatory networks, including the identification of network nodes (TFs and their target sequences) and edges (TF–TF dimers and TF–DNA target interactions), integration with other data types, and network properties and emerging principles that provide insights into differential gene expression.

E-mail marian.walhout@umassmed.edu; fax (508) 856-5460.

What is a regulatory network?

Network models are composed of nodes and edges that describe relationships between nodes. In biological networks, the nodes are bioactive macromolecules such as proteins, DNA, RNA, and metabolites (Barabasi and Oltvai 2004). Two types of regulatory networks can be distinguished: transcription regulatory networks and post-transcription regulatory networks (Fig. 2). Each of these types of networks can be subdivided into physical and functional networks. Physical networks contain protein–protein, protein–DNA, protein–RNA, and/or RNA–RNA interactions (Fig. 2A,C). Functional networks incorporate the consequences of these physical interactions, e.g., activation or repression of gene expression (Fig. 2B,D). Ultimately, transcription and post-transcriptional networks need to be combined to obtain a comprehensive picture of all aspects of the regulation of differential gene expression in complex metazoan systems (Fig. 2E).

In this review, I will focus on the mapping of transcription regulatory networks. I will discuss the identification of predicted TFs and cis-regulatory sequences, i.e., network nodes, and the protein–protein and protein–DNA interaction mapping approaches that are being used to identify physical interactions between these nodes, i.e., network edges. I will discuss several emerging insights and hypotheses that can be derived from such networks, and the future challenges that lie ahead in this rapidly evolving field.

Identifying network nodes

Transcription regulatory networks contain two types of nodes: regulatory TFs and their target DNA sequences. Many different strategies have been employed to identify both types of nodes, including computational and experimental methods.
Regulatory transcription factors

Regulatory TFs are composed of at least two types of domains: a DNA binding domain, which serves to interact with its cognate DNA target sequence, and a transcription regulation domain, which serves to activate or repress transcription (Fig. 1A). TFs are grouped into families based on their predicted DNA binding domains. To date, more than 100 different DNA binding domains have been found (Kummerfeld and Teichmann 2006). These domains have been used to computationally predict which genes in a genome of interest encode regulatory TFs. However, computational prediction alone is insufficient to obtain comprehensive and high-quality TF predictions. For instance, we recently obtained a high-quality compendium of Caenorhabditis elegans TFs by a combination of computational prediction and extensive manual curation (Reece-Hoyes et al. 2005). By doing so, the number of false positive and false negative predictions was drastically reduced. It should be feasible to obtain such comprehensive predictions for other organisms, including human, as well. However, even manually curated collections are likely incomplete as not all DNA binding domains have yet been uncovered. For example, both yeast and C. elegans proteins that bind DNA but that do not possess a known DNA binding domain have recently been retrieved (Hall et al. 2004; Deplancke et al. 2006).

TF predictions have led to the observation that, in increasingly complex metazoan organisms, a larger proportion of the genome encodes TFs, compared with relatively simple, unicellular eukaryotes. For instance, the genome of the unicellular yeast Saccharomyces cerevisiae encodes ~200 predicted TFs (Harbison et al. 2004), which is ~3% of all protein-coding genes; the relatively simple metazoan nematode C. elegans contains 934 predicted TFs, which is ~5% of all protein-coding genes (Reece-Hoyes et al. 2005); and more complex eukaryotes such as humans may devote up to 10% of their coding power to regulatory TFs (Levine and Tjian 2003).

TFs interact with different types of DNA sequences, including promoters and cis-regulatory modules, and, within such larger elements, bind to specific cis-regulatory elements or TF binding sites. Considerable efforts are underway to identify each of these elements in order to decipher the "regulatory code" that controls differential gene expression. For instance, the ENCODE (ENCyclopedia of DNA elements) Consortium aims to identify all functional elements in the human genome (ENCODE Project Consortium 2004). So far the efforts of this consortium have focused on 1% of the genome, or 30 Mb of sequence, which contains ~600 predicted protein-coding genes. In order to gain

---

**Figure 1.** Regulators of gene expression physically interact with their targets. (A) Transcriptional regulation. Regulatory TFs function by binding to proteins and to DNA. Black boxes, exons; blue squares, cis-regulatory DNA elements; red ellipses, TFs; arrow, transcription start site. Curved arrows indicate activation of gene expression and blunt "arrow" indicates transcriptional repression. AD, transcription activation domain; RD, transcription repression domain; DB, DNA binding domain. (B) Post-transcriptional regulation. RNA binding proteins and microRNAs function by directly interacting with their target mRNAs. Arrow, transcription start site; purple box, microRNA gene; green line, 3′ UTR of target microRNA (purple line); yellow circle, RNA binding protein. Upstream regulation of miRNA expression (by the pink TF binding to the light blue element) is indicated and connects transcriptional and post-transcriptional gene regulation.

**Figure 2.** Regulatory networks. (A) Protein–DNA and protein–protein interaction network involving regulatory TFs and their target genes. (B) Protein–RNA, protein–protein, and microRNA–RNA interaction network involving RNA binding proteins and their target mRNAs and microRNAs and their target mRNAs. (C) Transcription regulatory networks. The transcriptional consequences of the protein interactions shown in A are included. (D) Post-transcription regulatory networks. The effects of the protein interactions shown in B on target gene expression are included. (E) Combined transcription and post-transcription regulatory networks. Red nodes, TFs; blue nodes, target genes; green node, a gene can be a target gene and encode a TF; yellow nodes, RNA binding proteins; purple nodes, microRNAs. Black edges, protein–DNA interactions; dashed edges, protein–RNA interactions; red edges, microRNA–RNA interactions; blue edges, protein–protein interactions. Arrows, activation; blunt "arrow," repression.
insight into the regulatory code of a genome, one of the first steps is to comprehensively identify all gene promoters.

**Promoters**

A gene promoter is defined as the regulatory sequence (a few hundred base pairs) that is located immediately upstream of the transcription start site (for review, see Maston et al. 2006). Eukaryotic protein and microRNA-encoding gene promoters are composed of two parts: a proximal promoter that serves as a recognition sequence for the pre-initiation complex and RNA polymerase II, and a distal promoter that performs a regulatory function by interacting with regulatory TFs. The identification of promoters is relatively straightforward in unicellular eukaryotes such as *S. cerevisiae* as its genome is compact (Goffeau et al. 1996): It contains very few introns and short intergenic regions (median length shorter than 400 bp). Hence, the intergenic regions contain most cis-regulatory elements and can be used as a proxy for gene promoters when transcription start sites have not been precisely mapped. Since the genomes of higher eukaryotes contain longer intergenic regions with many repeat sequences, and because transcription start sites are often poorly defined, it is more difficult to accurately pinpoint metazoan gene promoters. Moreover, higher eukaryotic genes may frequently be regulated from multiple, alternative promoters.

Several experimental approaches have been developed for transcription start site and, thus, promoter annotation. First, full-length cDNA sequencing has led to the annotation of thousands of transcripts for both the murine and human genome (Imanishi et al. 2004; Carninci et al. 2005). Second, the use of genome-wide tiling arrays has enabled the identification of 5' and 3' boundaries of transcripts (Carninci et al. 2005). Third, cap analysis of gene expression (CAGE) has been used to more precisely define transcription start sites in mammalian promoters (Carninci et al. 2006). Fourth, chromatin-immunoprecipitations (see below) with anti-TFIID and anti-RNA polymerase II antibodies have been used to identify many active human promoters (Kim et al. 2005a,b). Finally, by transient transfection assays, 387 gene promoters from the ENCODE regions that drive gene expression in more than 16 different cell-lines were identified (Cooper et al. 2006). Although a lot of progress has been made, it is likely that highly sensitive experimental methods need to be developed to identify promoters that are rarely active.

**Cis-regulatory modules**

Many gene promoters have been identified to date. However, the genome-wide identification of enhancers and silencers in higher eukaryotes has been relatively slow. This is because they can be located at a great genomic distance from the target's transcription start site(s) and can be found upstream, downstream, or within introns (for review, see Maston et al. 2006).

It has been postulated that functional TF binding sites often occur in clusters and form cis-regulatory modules (Davidson 2001). Recently, this hypothesis has been utilized by several groups for the computational prediction of cis-regulatory modules that may constitute enhancers or silencers (Aerts et al. 2003; Sharan et al. 2003; Gupta and Liu 2005; Blanchette et al. 2006; Hallikas et al. 2006). The methods used by these groups provide powerful tools to search for cis-regulatory modules containing consensus binding motifs for TFs for which the recognition sequence has been mapped. However, to date such information is only available for a limited number of TFs.

In addition to using computational tools to predict regulatory elements, experimental methods can be used for the discovery of cis-regulatory modules (for review, see Elnitski et al. 2006). For instance, the observation that the genome is more accessible to DNasel when TFs are bound, leading to DNasel hypersensitive sites, can be used to identify cis-regulatory modules. Until recently, the unbiased, genome-wide mapping of such sites has been hampered by a lack of high-throughput "readout" methods that can be used to map such sites onto genome sequences. Several groups have already made significant progress toward this goal, for instance by combining DNasel treatment with microarrays or massive parallel sequencing (Dorschner et al. 2004; Crawford et al. 2006a,b; Sabo et al. 2006). Integration with other types of data will be necessary to delineate the function of each DNasel hypersensitive site and to find the transacting factors that bind to these sites.

**TF binding sites and cis-regulatory elements**

For a thorough understanding of transcription regulatory networks, it is not only important to find promoters and cis-regulatory modules, but also to precisely map the cis-regulatory elements located within these longer sequences. Individual cis-regulatory elements are short (usually <20 bp) DNA sequences that interact directly with regulatory TFs. Such TF binding sites have traditionally been mapped using a combination of deletion analyses and reporter gene expression (see, e.g., Davidson et al. 2002). However, such methods are not readily adaptable to high-throughput settings.

Recently, several methods have been employed to computationally identify putative cis-regulatory elements (Fig. 3) (for review, see Elnitski et al. 2006). The first method is based on the hypothesis that genes that are coexpressed under a particular condition are subject to control by the same TF(s). The advent of gene expression analysis by microarrays greatly facilitated the identification of coexpressed genes (DeRisi et al. 1997). Using a variety of computational algorithms, the regulatory regions of coexpressed genes can be interrogated for the occurrence of over-represented DNA sequences that may constitute binding sites for the TF responsible for the coexpression (for information and performance on such algorithms, see Tompa et al. 2005).

The second method, referred to as phylogenetic footprinting,
ing, is based on the conservation of functional cis-regulatory elements in closely related organisms. This method has been used to identify putative regulatory elements in yeast (Cliffen et al. 2001, 2003; Kellis et al. 2003), Drosophila melanogaster (Glazov et al. 2005), and human genomes (Bejerano et al. 2004; Siepel et al. 2005; Woolfe et al. 2005; Xie et al. 2005). In addition to using computational tools to find putative cis-regulatory elements in complete genome sequences, experimentally mapped consensus TF binding motifs (see below) can also be used to interrogate a genome sequence of interest. However, depending on the length of the motif, many functional and nonfunctional sequences will be identified. Phylogenetic footprinting and coexpression can then be used to determine which motifs have a higher likelihood of being functional in vivo (Fig. 3; Elnitski et al. 2006).

Only a small portion of all TF binding sites that occur in a genome of interest have been identified to date. For instance, by comparative genomics, Xie and colleagues found 174 candidate DNA motifs that likely correspond to numerous TF binding sites in human promoters (Xie et al. 2005). However, these different elements may only represent ~10% of all TF binding motifs as the human genome may encode more than 2000 TFs (Levine and Tjian 2003), each of which likely binds DNA with different specificity and affinity. On the other hand, it is likely that some TFs from one family may have overlapping binding specificities, and that therefore the number of different TF binding motifs may be considerably less than 2000. In addition, certain TFs may exclusively bind to regulatory elements that are located in transcriptional enhancers or silencers. These TF binding motifs will be missed in studies that focus solely on promoter sequences.

The computational prediction of cis-regulatory modules has relied on the observation that TF binding sites are often clustered. However, the generality of this phenomenon has not been investigated and, thus, it is not clear how many functional, non-clustered TF binding sites occur in the genome. In addition, most researchers have focused on elements that are conserved between related species. Such phylogenetic footprinting likely increases the specificity of motif finding. However, the sensitivity will suffer from only interrogating conserved sequences because many important, species-specific elements are not conserved. The success of phylogenetic profiling for the identification of functional regulatory elements also depends on the evolutionary distance between the organisms used in the analysis: The use of closely related species may result in relatively low specificity and the use of distantly related species may result in high specificity, but relatively low sensitivity (Ruvinsky and Ruvkun 2003).

Identifying network edges

**TF–TF dimers**

Many TFs bind their target genes as dimers. For instance, bZIP, bHLH, and nuclear hormone receptor TFs all dimerize. The comprehensive identification of TF dimers requires the use of protein–protein interaction detection methods that can be used in (semi) high-throughput settings. One assay that is particularly suited to identify binary protein–protein interactions is the yeast two-hybrid system (Fields and Song 1989), and multiple putative TF homo- and heterodimers have already been found using this system (Li et al. 2004; Reece-Hoyes et al. 2005; Rual et al. 2005; Stelzl et al. 2005). Putative TF dimers have also been identified by protein arrays. For instance, Newman and Keating tested >2400 combinations of human bZIP protein–protein interactions and found multiple putative dimers (Newman and Keating 2003). Finally, 15 putative yeast TF–TF heterodimers have been identified by large-scale TAP-TAG purification methods (Gavin et al. 2006; Krogan et al. 2006).

Only a small portion of all TF dimers has been identified to date. For instance, <10% of all predicted TFs are present in the current C. elegans protein–protein interaction network (Li et al. 2004), even though at least 30% of all TFs belong to the bZIP, bHLH, or nuclear hormone receptor families (Reece-Hoyes et al. 2005). Since TF dimerization may be condition-dependent, many TF dimers have also likely been missed by TAP-TAG assays in yeast. In the future, it will be important to comprehensively map all dimerization interactions between TFs and to incorporate this information into network models (Fig. 2).

**Interactions between TFs and their target genes/sequences**

Protein–DNA interactions between TFs and their target DNA sequences can be mapped using two conceptually different strategies. First, one can identify for a TF or set of TFs of interest, the target genes, and/or cis-regulatory elements these TFs bind to. Alternatively, one can take a DNA sequence as a starting point and aim to identify the TFs that can interact with this sequence. We refer to these strategies as “TF-centered” and “gene-centered” methods, respectively (Fig. 4A; Deplancke et al. 2006).

**TF–centered protein–DNA interaction mapping**

The most widely used protein–DNA interaction mapping methods are TF-centered, and most are based on chromatin-
immunoprecipitation (ChIP) (for review, see Elnitski et al. 2006). In ChIP assays an anti-TF antibody is used to precipitate DNA bound by the TF in vivo (Fig. 4B). These DNA fragments can subsequently be identified and quantified using a variety of readouts, including PCR, microarrays (referred to as ChIP-on-chip), and cloning/sequencing as in SAGE-like methods (serial analysis of gene expression, Fig. 4B; for review, see Blais and Dynlacht 2005; Elnitski et al. 2006). For yeast ChIP-on-chip assays, endogenous TFs were replaced by hybrid proteins in which the TFs were fused to a universal protein tag (Lee et al. 2002). Thus, almost 200 individual yeast strains were created, each carrying a different TF-TAG fusion protein. The advantage of this strategy is that the same antibody can be used for each TF. ChIP-on-chip has been used to identify target sequences for most yeast TFs under standard laboratory growth conditions (Lee et al. 2002). In addition, target binding has been examined under multiple experimental conditions for a subset of these TFs (Harbison et al. 2004; Workman et al. 2006). In addition to yeast, ChIP-on-chip has been used for a variety of mammalian TFs by using tissue culture cells (Cawley et al. 2004; Carroll et al. 2005; Bieda et al. 2006). So far, only a few studies focused on the DNA binding of metazoan TFs in their natural environment. For instance, in a pioneering study, endogenous promoters bound by HNF1α, HNF4α, and HNF6 within human liver and pancreas were identified (Odom et al. 2004). Similarly, by a combination of computational target prediction and ChIP-on-chip, many target promoters bound by CREB were identified in both tissue culture cells and primary hepatocytes (Zhang et al. 2005). ChIP-on-chip was also used to identify promoters bound by the TFs OCT4, SOX2, and NANOG in hepatocytes (Zhang et al. 2005). ChIP-on-chip was also used to identify target sequences for most yeast TFs under standard laboratory growth conditions (Lee et al. 2002). In addition, target binding has been examined under multiple experimental conditions for a subset of these TFs (Harbison et al. 2004; Workman et al. 2006). In addition to yeast, ChIP-on-chip has been used for a variety of mammalian TFs by using tissue culture cells (Cawley et al. 2004; Carroll et al. 2005; Bieda et al. 2006). So far, only a few studies focused on the DNA binding of metazoan TFs in their natural environment. For instance, in a pioneering study, endogenous promoters bound by HNF1α, HNF4α, and HNF6 within human liver and pancreas were identified (Odom et al. 2004). Similarly, by a combination of computational target prediction and ChIP-on-chip, many target promoters bound by CREB were identified in both tissue culture cells and primary hepatocytes (Zhang et al. 2005). ChIP-on-chip was also used to identify promoters bound by the TFs OCT4, SOX2, and NANOG in human embryonic stem cells (Boyer et al. 2005). Finally, ChIP-cloning (i.e., the cloning and sequencing of precipitated DNA) was used to identify in vivo target genes for the C. elegans FOXO TF, DAF-16 (Oh et al. 2005).

In DamID, a TF is fused to Escherichia coli DNA adenine methyltransferase (Dam) and expressed in tissue culture cells or intact model organisms (van Steensel and Henikoff 2000). Upon binding of the TF to DNA, the surrounding nucleotides are methylated. This methylation can be detected by PCR or microarrays after immunoprecipitation of methylated DNA. DamID has mainly been used to identify the DNA targets of general chromatin-binding proteins, but has also been used to dissect the Drosophila Myc TF network (Orian et al. 2003).

In protein-binding microarrays, a TF is fused to GST, expressed in bacteria or yeast, purified, and hybridized to a double-stranded DNA array that contains DNA sequences of interest (Mukherjee et al. 2004). To date, this method has been used to find targets for the yeast TFs Abd1, Rap1, and Mig1. The target sequences were then used to identify the consensus TF binding sites for each of these factors. DIP-ChIP can also be used to identify consensus TF binding sites. This method uses naked genomic DNA and a purified TF. Briefly, after incubation of the DNA with the factor, an immunoprecipitation is performed and TF-associated DNA is identified by microarray analysis (Liu et al. 2005). Although both DIP-chip and PBM are carried out in vitro, the TF binding sites obtained were in very good agreement with data obtained using in vivo methods, suggesting that they are effective and rapid methods to identify TF binding specificities, and, perhaps, affinities (Mukherjee et al. 2004; Liu et al. 2005).

In bacterial one-hybrid assays, a plasmid encoding a TF of interest is transformed into bacteria containing a library of random DNA elements (Meng et al. 2005). Binding of the TF to a specific element is transformed into bacteria containing a library of random DNA elements (Meng et al. 2005). Binding of the TF to a
evant. Such criteria have previously been used for large-scale protein–protein interaction maps, generated by high-throughput yeast two-hybrid assays (Li et al. 2004; Rual et al. 2005; Stelzl et al. 2005), and we have recently devised stringent criteria to filter Y1H data (Deplancke et al. 2006). In summary, both sensitivity and specificity are important issues to consider when choosing a protein–protein or protein–DNA interaction identification method, and the choice depends on the question being addressed. As the various methods are highly complementary, it is desired, in the long term, to use a multitude of techniques for comprehensive, high-quality protein–protein and protein–DNA interaction mapping.

Emerging concepts and future challenges

Large sets of protein–protein and protein–DNA interactions can be visualized as network models using various freely available software packages, including Cytoscape (Shannon et al. 2003) and N-browse (Lall et al. 2006).

Network models serve multiple purposes. For instance, they provide a great tool for the visualization and navigation of large interaction data sets. In addition, networks can be analyzed at different levels, i.e., at the level of the network as a whole, the level of subgraphs and network motifs, and the level of individual nodes or edges. By doing so, they enable the derivation of hypotheses regarding different levels of gene expression.

Network analysis

Once visualized, networks can be analyzed topologically using different network parameters such as connectivity, path length, clustering coefficient, etc. (For review, see Barabasi and Oltvai 2004). As has been observed for other networks, transcription regulatory networks are highly connected and display a scale-free degree distribution (Albert et al. 2000), i.e., they contain a small number of disproportionately highly-connected nodes, or hubs, and many less-well-connected nodes (Guelzim et al. 2002; Lee et al. 2002; Luscombe et al. 2004; Deplancke et al. 2006). Transcription regulatory networks potentially contain two types of hubs: TF hubs (TFs that bind many promoters) and promoter hubs (promoters that interact with many TFs). Interestingly, transcription regulatory networks predominantly contain TF hubs, rather than promoter hubs (Guelzim et al. 2002; Deplancke et al. 2006). As in other networks, such hubs provide integrity to the network: When nodes are randomly removed, the network stays connected. However, when hubs are sequentially removed the network disintegrates rapidly (for review, see Barabasi and Oltvai 2004). The biological implication of this became apparent when it was demonstrated that TF hubs have a higher tendency to be essential for the organism (Jeong et al. 2001; Yu et al. 2004; Deplancke et al. 2006).

We recently mapped a protein–DNA interaction network of genes expressed or involved in the C. elegans digestive tract (Deplancke et al. 2006). By visualizing and analyzing this network, we can derive hypotheses at different levels of gene regulation. For instance, we observed that the network is highly connected, contains several TF hubs, and is enriched for TFs expressed in the digestive tract (Fig. 5A; Deplancke et al. 2006). In addition, we found that most promoters are bound by a combination of TF hubs and less well-connected TFs, some of which may be master regulators. This led to a model in which we propose that C. elegans transcription is regulated by a layered organization of TF function (Fig. 5A; Deplancke et al. 2006). The digestive tract is predominantly composed of the pharynx and intestine, each of which is derived from distinct germ layers. We found that TF hubs interact with both pharyngeal and intestinal genes. This suggests that these TFs function as global regulators of gene expression and leads to the prediction that they interact with promoters of large numbers of genes that are expressed in other tissues as well.

In addition to hypotheses regarding gene regulation at the level of an entire network and system, one can also derive hypotheses by zooming into network subgraphs.

Network subgraphs

Network subgraphs can be network modules, motif clusters, or other network neighborhoods. A network module can be defined as a subgraph consisting of highly interconnected nodes that may fulfill a particular biological function. Network modularity has been observed in yeast regulatory networks (Ihmels et al. 2002; Bar-Joseph et al. 2003; Segal et al. 2003; Luscombe et al. 2004), although there are few modules that can be clearly separated from the main network component (Babu et al. 2004). This may be because individual yeast TFs may function in multiple, apparently unrelated pathways. These observations suggest that regulatory networks of higher eukaryotes such as C. elegans may be organized in modules as well but that these modules share multiple TFs. This hypothesis is in agreement with our observation that many C. elegans TFs are expressed in multiple tissues (Deplancke et al. 2006).

Figure 5B shows an example of a subgraph of the C. elegans digestive tract protein–DNA interaction network that can be used to derive specific biological hypotheses. This subgraph is composed of multiple bifan motifs (see below for network motifs) of the TF hubs DIE-1 and ZTF-1 and their target promoters. Interestingly, these TFs share 22 promoters, which is 73% of their combined targets (Fig. 5B). This leads to the prediction that DIE-1 and ZTF-1 have a similar biochemical function: For instance, they may have similar DNA binding motifs. The observation that they do not share all of their targets suggests that the motifs are not completely identical. Interestingly, these proteins share no homology in their primary amino acid sequence but both proteins do contain two pairs of C2H2 zinc fingers that are separated by a long amino acid sequence. The observation of shared targets also leads to the prediction that these two TFs may share biological functions. However, while knockdown of DIE-1 is lethal, ZTF-1 is dispensable for the function of the organism. In contrast, we could not create stable transgenic lines expressing ZTF-1, suggesting that overexpression of this protein may be lethal (Deplancke et al. 2006). Whereas DIE-1 can activate gene expression (Deplancke et al. 2006), the transcriptional function of ZTF-1 remains to be elucidated.

Network motifs

Network motifs are the building blocks of networks (Milo et al. 2002). Several motifs are overrepresented in experimentally derived transcription regulatory networks compared with random networks (Milo et al. 2002; Shen-Orr et al. 2002). Such motifs provide insights into the properties of networks and the propagation of regulatory signals. As such, the analysis of network motifs may help to uncover the biochemical functions of both TFs and their target genes. For instance, feed forward loops are overrepresented in transcription regulatory networks of various organisms (Milo et al. 2002; Shen-Orr et al. 2002). This may not
be surprising as such loops offer a rapid gene expression output, for example in response to outside signals. In contrast, feed-back or autoregulatory loops can either reinforce or diminish a transcriptional output, whereas single input motifs can confer strong coexpression of downstream target genes (Shen-Orr et al. 2002). The analysis of each TF and the network context in which it functions will be important to unravel how each factor contributes to differential gene expression.

Network motifs can also be used to derive specific biological hypotheses, either for individual promoters, or TFs. For instance, we found a single input motif in which ZTF-2 interacts specifically with the promoters of five pharyngeal genes (Deplancke et al. 2006). This led to the prediction that ZTF-2 is a regulator of pharyngeal gene expression and that these promoters share a pharyngeal gene element to which ZTF-2 binds. We tested these hypotheses experimentally and found that ZTF-2 is itself expressed in the pharynx (and elsewhere), and that a knockdown of ztf-2 results in a pharyngeal phenotype. Furthermore, we used the five promoter sequences to define a ZTF-2 binding motif and found that it is highly similar to a previously described pharyngeal gene element. Finally, we demonstrated that ZTF-2 represses expression of its pharyngeal targets and that it can bind the pharyngeal element in vivo. Taken together, the mapping, analysis and deconvolution of a protein–DNA interaction network into subgraphs and motifs can be used to derive biological hypotheses regarding differential gene expression at different levels.

Future challenges

The transcription regulatory networks that are currently available are likely to be a small representation of all the interactions that occur in vivo. Even in yeast, where binding of each TF has been examined under standard laboratory conditions and binding of a few under multiple conditions (Lee et al. 2002; Harbison et al. 2004; Workman et al. 2006), the regulatory information is likely far from complete. This is because many conditions remain to be tested and because TFs that bind DNA with low specificity or affinity may be difficult to analyze. The transcription regulatory networks that have been mapped in higher eukaryotes represent an even smaller sample of the entire network. This is because so
far (1) ChIP-on-chip assays mainly utilized arrays containing probes corresponding to promoter regions and, thus, TF binding to cis-regulatory modules located elsewhere in the genome will be missed; (2) only very few TFs have been examined by TF-centered methods; (3) $<1\%$ of all promoters in *C. elegans* have been examined by gene-centered methods (Deplancke et al. 2006), and, finally, not all cis-regulatory elements and TF binding sites have been identified either computationally (Xie et al. 2005) or experimentally (Mukherjee et al. 2004; Meng et al. 2005).

Cis-regulatory elements or TF binding sites are often found in intergenic regions. When intergenic regions are short (i.e., in yeast and *C. elegans*) and reside between genes that are transcribed from opposite strands, perhaps by bidirectional promoters, it is difficult to infer which of the two genes will be affected through such regulatory sequences. Similarly, in the genomes of higher eukaryotes, cis-regulatory modules can be located far from a transcribed unit and it may be difficult to infer the gene that is controlled by the cis-regulatory module. In the future, it will be important to scan larger genomic regions, both in cis and in trans for genes that can be affected by different individual cis-regulatory modules.

### From protein–DNA interaction networks to transcription regulatory networks

Networks that are solely based on protein–protein and protein–DNA interactions do not contain regulatory information because the protein–protein and protein–DNA interaction detection methods discussed above do not provide insight into the consequences of physical interactions (e.g., activation or repression of transcription). The transcriptional consequences of protein–DNA interactions need to be superimposed onto protein–DNA interaction networks by integrating interaction data with other data types (Deplancke et al. 2006). This can be done either at the level of individual interactions using detailed and often labor-intensive methods such as quantitative RT-PCR and RNAi (Baugh et al. 2005; Oh et al. 2005; Deplancke et al. 2006) or at the network level by integration with other large data sets, such as expression profiles (Lee et al. 2002; Segal et al. 2003; Yu et al. 2003; Luscombe et al. 2004). In the future, it will be important to develop methods that can be used to map, at a large scale, the transcriptional activity of each TF.

### Spatio–temporal network modeling

Protein–protein and protein–DNA interaction networks and transcription regulatory networks are static models of all the transcriptional events that can occur in a system of interest. To fully understand how such networks contribute to system development, function, and pathology, it is important to unravel where and when which parts of the network are active and what the biological consequences of this activity are (Davidson et al. 2002). Such analysis has again been pioneered in yeast. For instance, the binding of the cell cycle regulatory TF complexes SBF and MBF has been analyzed during different phases of the cell cycle (Horak et al. 2002). It was found that these TFs bind to and regulate many other TF-encoding genes that are involved in cell cycle progression and/or differentiation. In addition, more extensive networks that are active under different endogenous and exogenous experimental conditions were compiled (Luscombe et al. 2004). Surprisingly, it was found that these different subnetworks have different topological properties and motifs that may reflect their particular function. In the future, it will be important to extrapolate where and when which parts of transcription regulatory networks are active in higher eukaryotes as well.

Longer term, transcription regulatory networks need to be integrated to model more comprehensive regulatory networks in which transcription regulation of the expression of both protein-coding and microRNA-encoding genes is combined with gene regulation by both RNA binding proteins and microRNAs (Fig. 2E). Such networks need themselves to be integrated with spatiotemporal information about gene expression and TF/microRNA activity and with phenotypes conferred by TFs and microRNAs to obtain a comprehensive picture about regulatory networks and how they control the development, function, and pathology of complex metazoan systems.

### Acknowledgments

I thank members of the Walhout laboratory for discussions, J. Lieb for information on intergenic region length in yeast, S. Ryder for advice about RNA binding proteins, I. Barrasa for help with yeast TF–TF dimer retrieval, and B. Deplancke, V. Vermeirssen, N. Martinez, and J. Dekker for critical reading of the manuscript. The Walhout laboratory is supported by NIH grants CA097516, DK068429, and DK071713.

### References


Mapping transcription regulatory networks


