Full text provided by www.sciencedirect.com

Transcriptional networks: reverse-engineering gene regulation on a global scale

Gordon Chua, Mark D Robinson, Quaid Morris and Timothy R Hughes*

A major objective in post-genome research is to fully understand the transcriptional control of each gene and the targets of each transcription factor. In yeast, large-scale experimental and computational approaches have been applied to identify co-regulated genes, *cis* regulatory elements, and transcription factor DNA binding sites *in vivo*. Methods for modeling and predicting system behavior, and for reconciling discrepancies among data types, are being explored. The results indicate that a complete and comprehensive yeast transcriptional network will ultimately be achieved.

Addresses
Banting and Best Department of Medical Research, University of Toronto, 112 College Street, Room 307, Toronto, Ontario M5G 1L6, Canada
*e-mail: t.hughes@utoronto.ca

Current Opinion in Microbiology 2004, **7**:638–646

This review comes from a themed issue on
Growth and development
Edited by Mike Tyers and Mark Buttner

Available online 27th October 2004

1369-5274/$ – see front matter

DOI 10.1016/j.mib.2004.10.009

Abbreviations

chIP–chip	chromatin immunoprecipitation/chip (microarray) analysis
PWM	position weight matrix
TF	transcription factor
TFBS	transcription factor binding site

Introduction

Microarray analysis has provided abundant evidence that sets of functionally related genes are coordinately induced or repressed in response to developmental or environmental cues, presumably via the action of sequence-specific DNA-binding transcription factors (TFs). This provides a mechanism to control specific aspects of physiology; it also enables the use of gene co-regulation to predict gene function, and underlies the fact that expression profiles can be used to classify samples. Creating a full network diagram of transcriptional control will reveal how far these concepts can be extended.

The intriguing problem of determining how each gene is transcriptionally regulated, and which genes are controlled by each TF, has been attacked over the past few years by a variety of experimental and computational approaches. Here, we review recent accomplishments and challenges in this endeavor, focusing on the budding yeast *Saccharomyces cerevisiae*, in which many of these approaches have been field-tested. We also refer readers to several other recent reviews on the same general topic [1–7].

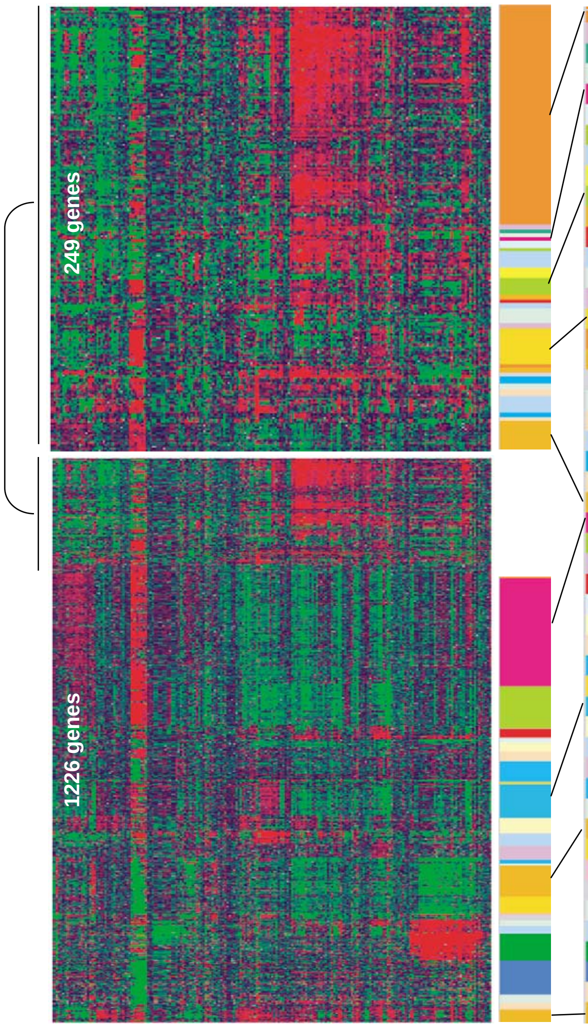
So that the significance of recent work can be appreciated, we begin by considering the overall problem.

Problem definition, expectations and success criteria

Deciphering a transcriptional network is a reverse-engineering problem. The goal is to understand how an existing system works, without being told. We are able to perturb the inputs (changes in growth conditions and/or genetic alterations) and monitor reaction of the outputs (measurements of relative abundance of transcripts in treatment versus control experiments, usually using microarrays). This is not a black box problem, because we have a working knowledge of the general architecture of the system. We have a relatively complete list of the TFs, and we expect that each of them will regulate a minority of genes (i.e. several to a few hundred) and that the genes controlled by any one factor will tend to be functionally related. We expect that TFs bind to regulatory sites in the promoters of genes, and that they may tend to work together with a limited number of other TFs. The regulatory sites are expected to be more conserved over evolution than background sequence. Moreover, the wiring of the network can be uncovered to considerable extent using biochemistry (e.g. by assessing the DNA-binding specificity of any specific TF *in vitro* and *in vivo*).

How will we know when we have succeeded in (correctly) modeling the global transcriptional network? Certainly, the basic observations must be explained. Estimates regarding the number of yeast TFs that are likely to directly regulate specific groups of genes vary from 141 to 209, depending on selection criteria [8••,9]. We obtained a list of 173 by taking those that contain a DNA-binding domain typical of TFs, and removing the few candidates that are known to have other functions. Searching through Medline and other databases reveals that something is known about the physiological function of most of these 173 (around 97). Among these 97, the majority (around 61) are involved in small-molecule metabolism or transport, processes easily

Figure 1



	Cluster label	Sample genes	Candidate regulator(s)
🟧	Amino acid metabolism	TRP4, HIS3	GCN4
🟪	Arginine biosynthesis	ARG1, ARG3	ARG80/81
🟩	Arginine catabolism	CAR1, CAR2	ARG80/81/UME6/RPD3
⬜	Aromatic AA metabolism	ARO9, ARO10	ARO80
🟥	Asparagine biosynthesis	ASN1, ASN2	GCN4/HAP1/HAP2
⬜	Branched chain AA synth	ILV1,2,3,6	LEU3, GCN4
🟩	Lysine biosynthesism	LYS2, LYS9	LYS14
🟦	Methionine biosynthesis	MET3,16,28	CBF1, MET28, MET32
🟨	Sulfur AA tnsprt, metab	MUP1, MHT1	MET31,MET32
🟩	Adenine biosynthesis	ADE1,4,8	BAS1, BAS 2, GCN4
🟧	Aldehyde metabolism	AAD4,14,16	
🟥	Biotin biosynthesis	BIO3,4	
🟦	Citrate metabolism	CIT1,2	RTG3
⬜	Ergosterol biosynthesis	ERG1,5,11	ECM22/UPC2
🟪	Fatty acid biosynthesis	FAS1,FAS2	INO4
🟨	Gluconeogenesis	PGK1, TDH1,2,3	GCR1
🟧	NAD biosynthesis	BNA4,6	
🟧	One-carbon metabolism	GCV1,2,3	
⬜	Pyridoxine metabolismt	SNO1, SNZ1	
⬜	Thiamin biosynthesis 1	THI5,12	THI2/THI3
⬜	Thiamin biosynthesis 2	THI2,20	THI2/THI3
🟦	Hexose transport	HXT4,GSY1	GCR1
🟦	Sodium ion transport	ENA1,2,5	NRG1,MIG1
⬜	Polyamine transport	TPO2,3	HAA1
🟧	Nucleocytoplasmic transport	KAP123,NUP100	RRPE-binding factor
🟥	Ribosome/RNA biogenesis	MAK16, CBF5	PAC/RRPE-binding factors
🟩	Ribosomal proteins	RPS1A, RPL28	RAP1, FHL1
🟪	Translational elongation	TEF1,2	
🟥	Protein folding	SSA1,HSP60	HAC1,ROX1
⬜	Secretion	VTH1,KRE11	RLM1
🟨	Protein glycosylation	ALG6,CAX4	XBP1
⬜	Vesicle-mediated transport	VPS5,IMH1	
🟦	Proteasome	RPN6,RPT5	RPN4
🟨	Vacuole fusion	VTC1,3,4,PHO84	PHO4
🟦	Mitoribosome/respiration	MRPL1,MRPS5	
🟨	Mitochond. electron trans.	ATP1,COX4	HAP2/3/4/5
🟦	Iron transport/TCA cycle	FRE1,FET3	MAC1/RCS1/AFT1/PDR1/3
🟪	Chromatin/transcription	SNF2,CHD1,DOT6	
🟦	Histones	HTA1,HHF1	HIR1,HIR2
⬜	MCM2/3/6/CDC47	MCM2,3,6	ECB
🟧	DNA replication	RFA1,POL12	MCB
🟨	Mitotic cell cycle	SPC110, CIN8	HCM1
🟪	CLB1/CLB6/BBP1	CLB1,6	FKH1
⬜	Cytokinesis	CTS1,EGT2	ACE2,SWI4
⬜	Development	PAM1,GIC2	
🟦	Pheromone response	FUS3,FAR1	MATALPHA2,STE12
🟩	Conjugation	CIK1,KAR3	KAR4
🟦	Sporulation/meiosis	SPO11, SPO19	NDT80
⬜	Oxidative stress response	GDH3,HYR1	ROX1,MSN2, MSN4
🟧	Stress/heat shock	HSP104, SSA4	MSN2,MSN4

Current Opinion in Microbiology

A selection of non-overlapping, biologically meaningful clusters in yeast. This diagram is not comprehensive, but was constructed to ask whether observed gene-expression measurements can be explained on the basis of current literature and databases. Two hundred clusters were obtained from 424 microarray experiments compiled from the literature [11] using average linkage hierarchical clustering with Pearson correlation. Each cluster was analyzed for biological 'significance' using the hypergeometric P value against gene ontology annotations using FunSpec (http://funspec.med.utoronto.ca/). Fifty clusters encompassing 1226 genes were retained and each was analyzed manually using TRANSFAC [10], YRSA [36], Medline, and the *Saccharomyces* Genome Database (http://www.yeastgenome.org/) to identify known or putative transcriptional regulators.

manipulated experimentally by changing growth conditions. However, an additional 76 of the 173 appear to be poorly characterized. Presently, TRANSFAC [10], a reference database of established TF-binding sites, contains a total of 319 targets that bind directly (demonstrated by gel shifts or DNaseI footprinting) to 88 of these 173 TFs. Thus it appears that much is known in general, but much is uncertain in detail.

These observations are consistent with results of clustering microarray expression data. From a collection of 424 microarray experiments [11] we compiled 50 non-overlapping, biologically meaningful clusters encompassing 1226 genes (Figure 1). Most of these clusters also relate to small-molecule metabolism and transport, although these in fact account for only a small fraction of the target genes. The expression patterns of some of the clusters are complicated and resemble patterns of other clusters, suggesting multiple regulators. It is heartening that for most of the clusters (40/50) at least one putative DNA-binding regulator can be identified. However, most of the individual TF–target inferences are unproven, and some of the largest clusters cannot yet be associated with specific TFs. Two of the most frequently observed clusters in yeast microarray data are the mitochondrial ribosome, for which no promoter elements or binding factors have yet been identified, and the ribosome biogenesis/ RNA processing regulon, for which two promoter elements (PAC and RRPE) have been identified by many studies (e.g. [12]) but for which no binding factors have yet been identified. Hence, there is still much experimental work to be done before all of the observations can be explained.

Another classical test of whether a system is understood is to ask whether one can predict how the system behaves in response to any new condition or modification. Herrgard *et al.* [7] pointed out that models must ultimately consist of equations describing dynamics, as is already possible in metabolic analysis. However, we are still short of drawing even a complete and accurate ball-and-arrow diagram. We focus here on how completion of the ball-and-arrow diagram is progressing, beginning with the experimental data, and progressing through computational approaches.

Perturbing transcription factors for microarray expression analysis

Perhaps the most obvious experiment to define transcriptional targets would be microarray expression analysis of deletion mutants or overexpressors in all yeast TF-encoding genes. Indeed, at least 61 yeast TFs have been perturbed experimentally and analyzed using microarrays [13]. There are two major difficulties with this approach. The first is that one cannot rule out that any expression changes observed are secondary effects. However, the major obstacle appears to be ensuring that the TF is active under the chosen growth condition. Of the 173 TFs on our list, only six (3.4%) are essential for cell viability in rich medium, in comparison to around 19% of yeast genes on the whole. This suggests that most of these TFs are either redundant with other TFs, that they regulate non-essential processes, or that they are inactive in yeast grown under standard laboratory conditions. This latter possibility is supported by analysis of both expression data and promoter binding assays ([14•,15••]; see below).

A strategy to identify the conditions that activate a TF of interest (and to identify physiological functions of the TF) is to screen for conditions that cause hypersensitivity for the TF mutant, because this may indicate that the TF is required for the proper response to this perturbation (for recent examples, see [16,17•]). In these conditions, comparison of the wild-type and TF mutant strains often results in differential gene expression of targets [18–22].

Identifying the activating conditions might be difficult for many TFs. Three approaches to artificial activation of TFs circumvent this problem, and these have been successful in discovering TF targets. First, Wilcox *et al.* [23] identified potential targets of Upc2p by expression profiling an activated allele of the TF. Second, the overexpression of native *HAP4* was sufficient to induce genes enriched in mitochondrial function consistent with its role in respiration [24]. Third, activated forms of several Zn_2Cys_6 zinc-finger TFs have been engineered by fusion of the DNA-binding domain of the TF and a constitutive transcription-activating domain [13,25,26•].

ChIP–chip: microarray analysis of promoter binding *in vivo*

The 'chIP–chip' methodology involves the chromatin immunoprecipitation of an epitope-tagged TF bound to DNA fragments containing target promoters, followed by the hybridization of those amplified DNA fragments to an intergenic microarray [27,28]. Two major studies have been conducted under nutrient rich conditions; one consists of an extensive chIP–chip investigation on 106 TFs [8••], while the other [29] focuses on 11 TFs that function at the G1/S transition.

ChIP–chip data do not prove physiological relevance of the bound sites, but do have the advantage that they provide a direct biochemical link between TFs and promoters, and should be devoid of secondary effects. ChIP–chip data also have the potential to identify targets without knowing the activating conditions, if the TFs are promoter-bound in an inactive state. However, it appears as if this may not generally be the case. Manual examination of the Lee *et al.* [8••] chIP–chip data indicates that known targets were detected for only around 50% of the TFs. Many of these TFs are involved in basic aspects of cell growth, such as cell cycle regulation (e.g. *ABF1*, *ACE2*, *FKH1*, *FKH2*, *MCM1*, *NDD1*, *STB1*, *SWI4*, *SWI5* and *SWI6*) as well as nitrogen (e.g. *DAL81*, *DAL82* and *GLN3*), glucose (e.g. *GCR1*, *GCR2* and *MIG1*) and fatty acid metabolism (e.g. *INO2* and *INO4*). By contrast, TFs that play roles in stress responses, and thus are not required for growth in rich medium, did not bind to their known targets (*HAL9*, *HSF1*, *MSN1*, *MSN2*, *MSN4*, *PDR1*, *RIM101*, *ROX1*, *SKN7*, *YAP5*, and *YAP7*). This observation is supported by other more sophisticated analyses (e.g. [15••]; see below).

Consequently, the chIP–chip technology has begun to be extended to include perturbations as a means to increase the range of their targets and to uncover new biological roles not detected in the previous studies [30•,31]. In a few cases, the comparison of chIP–chip binding profiles of a TF under non-stimulating and stimulating conditions has resulted in identification of additional potential targets. For example, Ste12p bound to the promoters of 30 and 106 genes when cells were grown in the absence and presence of pheromone, respectively [31]. Harbison *et al.* [32] have recently extended the chIP-chip technique to a wide variety of yeast transcription factors under different growth conditions (see also Update).

Computational identification of *cis* regulatory sites

A major mechanism by which TFs recognize genes to be regulated is the presence of cognate DNA-binding sequence of TFs in the promoter. Two general computational approaches have emerged for finding TF binding sites (TFBSs) in promoters *de novo*: analysis of co-regulated genes; and phylogenetic footprinting. Both methods aim to distinguish small (6–15 bp) conserved (or enriched) elements against the remainder of intergenic sequence (around 500 bp in yeast). These elements can then be compared with the binding specificity of known TFs, and/or subjected to more traditional biochemical and genetic analyses, such as reporter and gel-shift assays, to prove that they are functional and to determine binding factors.

Correlation of promoter sequence with expression

A variety of statistical algorithms have been developed for identifying sequence elements common to promoters of co-regulated genes, typically expression clusters (reviewed in [5]). The most popular of these algorithms use stochastic search methods, such as Gibbs sampling, to find sequences that are over-represented among the promoter sequences in the cluster (it is rare to find a sequence motif that is unique to a set of co-regulated genes and common to all of them; this is what makes the problem challenging). The algorithms differ in the way that they represent the binding profile of TFs and in the assumptions they make about the presence and location of TFBSs in the promoter sequence. The most popular model for TF-binding profiles is the positional weight matrix (PWM); although the assumption that the positions are independent is incorrect [33,34], it is not disastrously inaccurate [35].

Yeast regulatory sequence analysis (YRSA) [36] is one recent extension of these approaches. It consists of a web implementation that automatically couples Gibbs sampling of the promoters of yeast genes with a comparison of the sequences identified to the sites of known yeast TFs. An advanced algorithm, stochastic dictionary data augmentation (SDDA [37•]), finds binding sites for multiple TFs, does not require the length of the binding site to be pre-specified, and allows for TFBSs that contain large, variable-sized gaps between conserved elements. An alternative to the cluster-based approach has also been developed in which regression is used to identify sequences that correlate with expression levels derived from a single experiment: Roven and Bussemaker [38] have created a web implementation called REDUCE (regulatory element detection using correlation with expression).

A common outcome of all of these approaches is that one is faced with a large number of candidate regulatory sequences. A promising approach for downstream analysis is to find 'discriminative' sequence elements (i.e. combinations of TFBSs that are represented *only* in the co-regulated cluster; for example, [39••], discussed below).

Sequence conservation of regulatory sequences

It is also possible to predict TFBSs using multiple-sequence alignments to identify conserved features. The compared species need to be similar enough to reliably align their intergenic regions, but distinct enough that there is a discernible difference in the conservation of functional sequence elements (i.e. TFBSs) compared with that of the non-functional flanking sequence. On the basis of a preliminary sequencing study of 13 yeast species, Cliften *et al.* [40] argued that the complete genomes of three other *sensu stricto Saccharomyces* species would be sufficient to reliably detect functional elements in *S. cerevisiae*. This work prompted two groups — Cliften *et al.* [41••] and Kellis *et al.* [42••] — independently to sequence the required three genomes. Kellis *et al.* used *S. mikatae*, *S. paradoxus* and *S. bayanus*; whereas Cliften *et al.* replaced *S. paradoxus* with *S. kudriavzevii*, and added two *sensu lato* species for additional reference.

The two groups also used different algorithms for detecting conserved regulatory elements. Cliften *et al.* [44••] enumerated all 6- to 30-mers perfectly conserved in either four- or six-way alignments of intergenic regions. They found more than 16 000 conserved elements of which around 25% matched a characterized TF-binding profile, matching, in total, 53 of 62 characterized, ungapped profiles. Of the remaining conserved elements, Cliften *et al.* identified 79 candidate TFBSs by finding those whose associated gene set was significantly enriched for either co-expression, functional annotation or TF binding as measured by chIP–chip [8••].

By contrast, Kellis *et al.* [42••] used an algorithm that allows gaps in TFBSs, and whose output is a degenerate sequence motif. Over 2400 mini-motifs — high-scoring intergenic sequences with two conserved 3-mers flanking a gap of length 0–21 bp — were initially identified, and then iteratively extended and clustered to obtain 72 motifs that satisfy a 'motif conservation score' threshold, taken as the ratio of conservation of a given motif to that of a random motif having the same length and degeneracy. Among the 72 emerging motifs were strong matches to 35 known motifs. Of the novel putative TFBSs, 25 had associated gene sets that were significantly enriched for either co-expression, specific Gene Ontology annotations, or chIP–chip binding.

Variations on phylogenetic footprinting can also be applied even if the positions of regulatory sites have re-arranged over evolution, complicating multiple alignment. A phylogeny can be used to distinguish between conserved non-functional sequences which have not yet diverged in closely related species and less conserved functional sequence whose variation across the species pool recapitulates its phylogeny. Blanchette and Tompa [43] described an algorithm, Footprinter, that finds sequence elements which have less than a user-specified parsimony with respect to a supplied phylogeny and allows the absence of these elements in certain parts of the phylogeny, to model changes in the regulatory structure of distantly related species. In a different approach, Pritsker *et al.* [44] generated a large set of putative TFBSs by running Gibbs sampling on sets of orthologous promoters from 13 hemiascomycetous yeast species. These were then filtered by 'network-level conservation' (i.e. the genes with strong matches to a TFBS in one species [*S. cerevisiae*] should be the orthologs of those with strong matches in another, closely related species [in this case, *S. bayanus*]). Many known motifs (e.g. binding sites for Rap1p, Ume6p, Abf1p) received high network-level conservation scores, and the algorithm had an overall sensitivity of 82% in identifying 48 known TFBSs conserved between *S. cerevisiae* and *S. bayanus*. Specificity was not reported, although the authors estimated that they detected around 400 different binding sites, which is at least twice the estimated number of yeast DNA-binding transcription factors. Nonetheless, high-scoring putative TFBSs showed high levels of conservation of position and orientation.

How do the different data types compare?

The types of data described thus far (gene expression, TF DNA-binding, promoter elements) all represent aspects of the same regulatory networks, and can presumably be combined to formulate global or more accurate models, or both. It is logical to ask first how well they correspond to one another. Figure 2 shows that there is, apparently, not a high degree of agreement between some of the most commonly cited datasets. The highest correspondence to TRANSFAC [10] appears to be the expression clusters from a compilation of published microarray data [11].

This suggests that some effort will be required to reconcile all types of information. Indeed, much of the literature over the past two years has been aimed at coupling one or more data types together with other information or laboratory assays, or both, to more rigorously identify and verify 'modules' of the network (i.e. groups of genes controlled by the same TF or group of TFs). To illustrate this, Table 1 displays a (non-exhaustive) list of recent papers describing methods to identify yeast transcriptional networks, and the data types each utilized to do so. We discuss several that introduce or address key points.

Expression data alone: beyond clustering

To use microarray expression data alone to identify regulatory modules, Segal *et al.* [45] presented a probabilistic graphical model algorithm to infer regulation 'programs'. Given a list of probable regulatory factors, and assuming the regulator has a similar expression profile to the target genes (at least on a subset of conditions), the algorithm groups genes into modules that show coordinate expression with the factor (or factors). This facilitates testing hypotheses of the form 'regulator X regulates module Y under conditions W'. Segal *et al.* [45] confirmed some well-known mechanisms, and shed light on potential new programs.

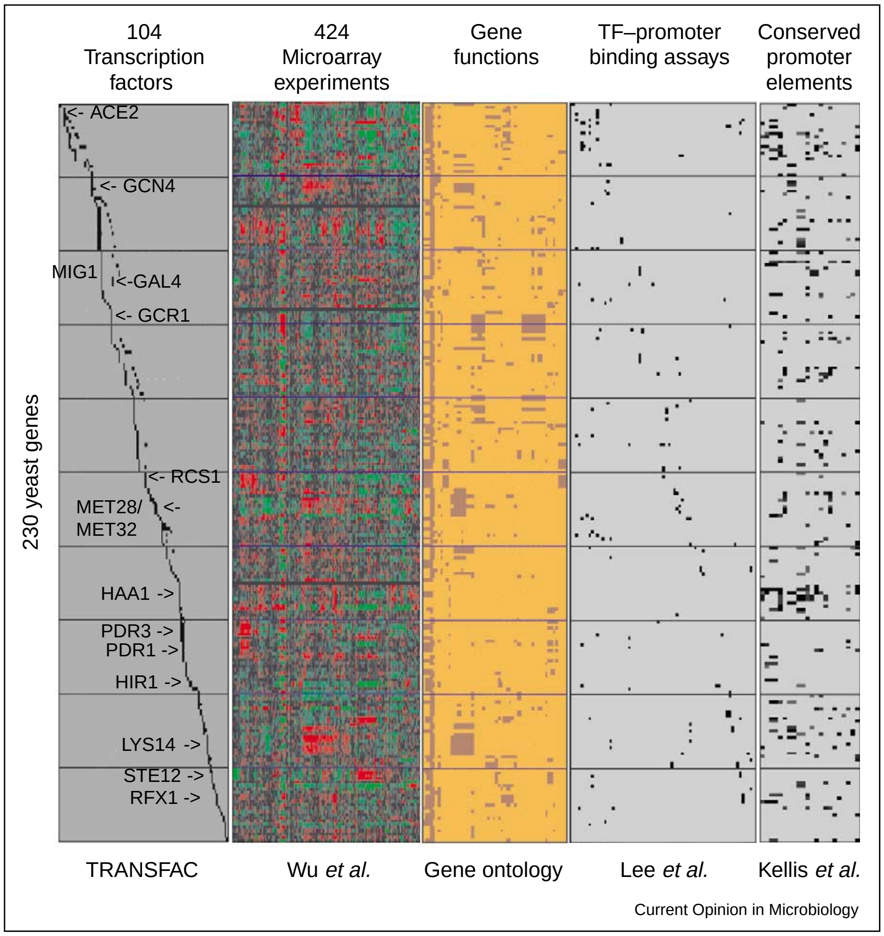
One difficulty with this and related approaches (Table 1) is that the expression of the regulator itself might not correlate well with expression of its targets (e.g. if the TF is regulated post-transcriptionally). In fact, Herrgard *et al.* [46••] showed that significant correlations between known TF–target pairs are infrequent, occurring in less than 20% of 925 'known' regulator–target pairs tested. This is also supported by the results of Qian *et al.* [14•], who took a supervised learning approach to identify relationships in gene expression measurements among known TF–target pairs.

Expression data combined with chIP–chip

A tried-and-true approach to reduce false-positives with noisy biological data is to use multiple data sources. Bar-Joseph *et al.* [30•] described the GRAM (genetic regulatory modules) algorithm, which creates initial gene regulatory modules with strict criteria for both chIP–chip binding confidence and expression correlation. Having identified a putative regulatory module, the algorithm revisits less significantly bound genes and appends them to the module if they are sufficiently co-expressed. The approach recapitulated many known regulatory modules, and identified new TF–target pairs that seem functionally relevant. Although the authors commented on the large number of connected gene modules, an equally significant result is that they obtained many unconnected or low-connectivity regulatory modules (i.e. groups of functionally related genes regulated by only one or a few TFs), suggesting that the structure of the yeast transcriptional network may become simpler as noise in the data is reduced.

Gao *et al.* [15••] employed a multiple linear regression approach to infer the 'activity profile' of a TF from both gene expression data over 750 diverse expression patterns and ChIP occupancy data of 113 TFs from Lee *et al.* [8••]. (Although Lee *et al.* [8••] described analysis of 106 TFs, the downloadable data includes 113 TFs.) They defined the notion of 'coupling' as the co-expression of a gene with its activity profile. Genes that are both bound and coupled (i.e. the genes bound by the same TF that also display coordinate expression) are likely to be functional direct targets. This analysis produced several important quantitative results. Among the 113 TFs analyzed by Lee *et al.* [8••], only 37 were significant predictors of mRNA

Figure 2



Correspondence between information in the TRANSFAC database of TF-binding sites, and other datasets reflective of gene regulation mechanisms. The 104 TFs do not all contain a DNA-binding domain, and the target genes shown include only those that can be associated with a single yeast gene. TRANSFAC-derived targets were ordered to place genes regulated by the same factor(s) adjacent on the vertical axis. This order of genes was held fixed in the other four datasets. For microarray data and gene functions, the horizontal axis was ordered by clustering (i.e. if the data agreed with TRANSFAC then blocks of color would be formed). For the Lee *et al*. and Kellis *et al*. datasets, the horizontal axis is the same as that in the TRANSFAC panel, and only the TFs shared with TRANSFAC are shown (i.e. if the data agreed perfectly with TRANSFAC, it would form the same diagonal line as is shown left).

expression. Among these 37, on average 58% of significantly bound genes were coupled. For most of the 37, the coupled targets displayed a strong tendency to be functionally related, whereas the non-coupled targets did not. This suggests that chIP–chip results which do not correlate with expression data might be less physiologically relevant. The corresponding TFs could be prioritized for analysis under other growth conditions, for example those under which the TF is required for normal growth or behavior (e.g. [17•,31]).

Expression data combined with promoter sequence

Pilpel *et al.* [47] introduced the combinatorial analysis of promoter elements from gene expression data. Their study illustrated several examples of 'expression coherence' among the genes having the promoter elements of two TFs compared with those genes whose promoters have binding sites of two TFs compared with those genes whose promoters have the binding site of only one of the TFs. Their analysis also showed that, by virtue of the fact that some genes share multiple regulatory factors, an interconnected network diagram can be constructed.

More recently, Beer and Tavazoie [39••] made a major conceptual advance by using a Bayesian network learning algorithm to predict the positional, orientational and combinatorial constraints of upstream sequence elements which are predictive of expression patterns. An initial set of putative TFBSs, discovered using Gibbs sampling, was culled for those predictive of whether each of 2587 genes belonged to each of 49 partially overlapping gene

Table 1

Recently reported methods and types of data used for identifying regulatory modules.

Analysis paper	Gene expression data	Binding data	Intergenic sequence	Functional annotation	Method
Segal *et al.* [45]	●				Module networks
Qian *et al.* [14•]	●				Support vector machine
Tringe *et al.* [49]	●				Graph reconstruction
Lee *et al.* [8••]		●			ChIP–chip
Kellis *et al.* [42]			●		Comparative genomics
Chiang *et al.* [50]			●		Conserved word pairs
Gao *et al.* [15••]	●	●			Multiple regression-coupling
Banerjee *et al.* [51]	●	●			Cooperativity analysis
Bar-Joseph *et al.* [30•]	●	●			Genetic regulatory modules
Cliften *et al.* [41••]	●		●		Phylogenetic fingerprinting
Pilpel *et al.* [47]	●		●		Expression coherence
Haverty *et al.* [52]	●		●		Computational ascertainment of regulatory relationships inferred from expression (CARRIE)
Beer and Tavazoie [39••]	●		●		Bayesian networks with constraints
Chen *et al.* [53]	●		●	●	Functional clustering
Wang *et al.* [48]	●		●	●	Expression-weighted profiles

expression clusters. The algorithm used a 'greedy' iterative search to find the combination (and other constraints) of TFBSs that best distinguished between genes present and absent in the cluster. The cluster assignments of 73% of the genes included in the analysis were correctly classified from upstream sequence alone. The fact that 26% of genes could be classified correctly at random suggests that the initial gene and cluster selection substantially simplified the problem. Nonetheless, this is a highly significant result, and both the analytical framework and the evaluation criteria applied raise the bar from correlation or overlap between datasets, to using one data type to predict the results of the other in a blind test.

Conclusions and future prospects

Reverse-engineering the transcriptional regulatory network architecture in yeast has now been attacked by several approaches on a large scale. Computational analysis of extensive network structures is now possible. However, models constructed from current data sources are likely to contain many errors, because different approaches produce different lists of regulator–target associations. They will certainly be incomplete, because roughly one third of all TFs appear to be uncharacterized with respect to physiological function, and the observed regulation of many genes cannot yet be associated with a direct regulator, known or putative.

The realization that current datasets have not yielded a harmonized view of yeast transcriptional regulation is, in our view, one of the most significant results in the field in the past few years. Before progressing to more quantitative and dynamic models of whole-cell transcriptional regulation, considerably more work will be required to simply ensure that all of the connections are drawn properly. Consequently, one of the current challenges in the field remains to ascertain and confirm the individual regulatory modules (i.e. identify the physiological targets of each TF, and the mechanism underlying each gene regulation, including the upstream cues and signaling pathway that control each TF).

Although the wealth of data now available is often extolled, there is little doubt that much more data are needed. It will be invaluable to obtain mutant compared with wild-type expression profiles (and chIP–chip binding profiles) under conditions that activate each TF. Several network motif analysis approaches automatically predict the appropriate conditions for these experiments [45,48]. Microarray data and chIP–chip data resulting from artificial activation of TFs will also be extremely valuable. To take maximum advantage of promoter sequences, it would also be advantageous to know the sequence specificity of all of the known and predicted TFs, and to ask what protein(s) can bind to each putative *cis* element.

Despite these caveats, it is encouraging that virtually all approaches discussed here can yield coherent data, and they do relate to one another somewhat. This indicates that, although achievement of a comprehensive network might be more difficult than expected, the overall goal of reverse-engineering transcriptional networks is tractable with existing techniques and following current tenets.

Update

In a follow up paper to Lee *et al.* [8••], Harbison *et al.* [32] have extended chIP-chip analysis to 203 yeast TFs in rich media conditions and 84 of these regulators in at least one environmental perturbation. Many of the binding profiles appear to differ considerably with growth condition, generally with more promoters bound by the TFs under perturbed conditions.

Acknowledgements

We thank Jack Greenblatt, Brenda Andrews, Ben Blencowe and Nevan Krogan for careful evaluation of this manuscript. The authors are funded by Natural Sciences and Engineering Research Council of Canada and Canadian Institutes of Health Research.We apologize to our many colleagues whose work could not be cited owing to space constraints.

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. Devaux F, Marc P, Jacq C: **Transcriptomes, transcription activators and microarrays**. *FEBS Lett* 2001, **498**:140-144.

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

3. Horak CE, Snyder M: **Global analysis of gene expression in yeast**. *Funct Integr Genomics* 2002, **2**:171-180.

4. Wyrick JJ, Young RA: **Deciphering gene expression regulatory networks**. *Curr Opin Genet Dev* 2002, **12**:130-136.

5. Qiu P: **Recent advances in computational promoter analysis in understanding the transcriptional regulatory network**. *Biochem Biophys Res Commun* 2003, **309**:495-501.

6. Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA: **Structure and evolution of transcriptional regulatory networks**. *Curr Opin Struct Biol* 2004, **14**:283-291.

7. Herrgard MJ, Covert MW, Palsson BO: **Reconstruction of microbial transcriptional regulatory networks**. *Curr Opin Biotechnol* 2004, **15**:70-77.

8. •• 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.

This landmark paper presents a comprehensive analysis of yeast transcription factors using chIP–chip technology.

9. Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR *et al.*: ***Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes**. *Science* 2000, **290**:2105-2110.

10. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV *et al.*: **TRANSFAC: transcriptional regulation, from patterns to profiles**. *Nucleic Acids Res* 2003, **31**:374-378.

11. Wu LF, Hughes TR, Davierwala AP, Robinson MD, Stoughton R, Altschuler SJ: **Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters**. *Nat Genet* 2002, **31**:255-265.

12. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture**. *Nat Genet* 1999, **22**:281-285.

13. Devaux F, Marc P, Bouchoux C, Delaveau T, Hikkel I, Potier MC, Jacq C: **An artificial transcription activator mimics the genome-wide properties of the yeast Pdr1 transcription factor**. *EMBO Rep* 2001, **2**:493-498.

14. • Qian J, Lin J, Luscombe NM, Yu H, Gerstein M: **Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data**. *Bioinformatics* 2003, **19**:1917-1926.

An analysis of gene expression data to predict TF–target pairs using a support vector machine. Instead of assuming expression levels of transcription factors are highly correlated with their targets, the algorithm discovers subtle relationships (e.g. lag, exaggerated profiles) between the expression levels of transcription factors and their targets.

15. •• Gao F, Foat BC, Bussemaker HJ: **Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data**. *BMC Bioinformatics* 2004, **5**:31.

This is an analysis of gene expression data in combination with chIP–chip data. Using a regression analysis approach, the authors are able to find when transcription factors are active, and to define direct targets.

16. Palkova Z, Devaux F, Icicova M, Minarikova L, Le Crom S, Jacq C: **Ammonia pulses and metabolic oscillations guide yeast colony development**. *Mol Biol Cell* 2002, **13**:3901-3914.

17. • Schuller C, Mamnun YM, Mollapour M, Krapf G, Schuster M, Bauer BE, Piper PW, Kuchler K: **Global phenotypic analysis and transcriptional profiling defines the weak acid stress response regulon in *Saccharomyces cerevisiae***. *Mol Biol Cell* 2004, **15**:706-720.

This paper elegantly shows an effective approach to identify direct targets by microarray profiling transcription factor deletion strains, in which emphasis is placed on finding the activating condition of the transcription factor by screening the transcription factor deletion strain for hypersensitivity to that condition. By comparing microarray profiles of single and double mutants in the presence of the activating condition, genes uniquely and combinatorially regulated by each transcription factor could be distinguished.

18. Cohen BA, Pilpel Y, Mitra RD, Church GM: **Discrimination between paralogs using microarray analysis: application to the Yap1p and Yap2p transcriptional networks**. *Mol Biol Cell* 2002, **13**:1608-1614.

19. Kwast KE, Lai LC, Menda N, James DT III, Aref S, Burke PV: **Genomic analyses of anaerobically induced genes in *Saccharomyces cerevisiae*: functional roles of Rox1 and other factors in mediating the anoxic response**. *J Bacteriol* 2002, **184**:250-265.

20. Santiago TC, Mamoun CB: **Genome expression analysis in yeast reveals novel transcriptional regulation by inositol and choline and new regulatory functions for Opi1p, Ino2p, and Ino4p**. *J Biol Chem* 2003, **278**:38723-38730.

21. Young ET, Dombek KM, Tachibana C, Ideker T: **Multiple pathways are co-regulated by the protein kinase Snf1 and the transcription factors Adr1 and Cat8**. *J Biol Chem* 2003, **278**:26146-26158.

22. Vachova L, Devaux F, Kucerova H, Ricicova M, Jacq C, Palkova Z: **Sok2p transcription factor is involved in adaptive program relevant for long-term survival of *Saccharomyces cerevisiae* colonies**. *J Biol Chem* 2004, in press.

23. Wilcox LJ, Balderes DA, Wharton B, Tinkelenberg AH, Rao G, Sturley SL: **Transcriptional profiling identifies two members of the ATP-binding cassette transporter superfamily required for sterol uptake in yeast**. *J Biol Chem* 2002, **277**:32466-32472.

24. Lascaris R, Bussemaker HJ, Boorsma A, Piper M, van der Spek H, Grivell L, Blom J: **Hap4p overexpression in glucose-grown *Saccharomyces cerevisiae* induces cells to enter a novel metabolic state**. *Genome Biol* 2003, **4**:R3.

25. Le Crom S, Devaux F, Marc P, Zhang X, Moye-Rowley WS, Jacq C: **New insights into the pleiotropic drug resistance network from genome-wide characterization of the YRR1 transcription factor regulation system**. *Mol Cell Biol* 2002, **22**:2642-2649.

26. • Hikkel I, Lucau-Danila A, Delaveau T, Marc P, Devaux F, Jacq C: **A general strategy to uncover transcription factor properties identifies a new regulator of drug resistance in yeast**. *J Biol Chem* 2003, **278**:11427-11432.

The authors used an artificial activation-based strategy combined with microarray profiling to identify direct targets of the previously uncharacterized transcription factor Pdr8p. The artificially activated form of Pdr8p was constructed by the fusion of its DNA-binding domain to the GAL4 transcriptional activation domain, and the authenticity of its transcriptional targets was supported by their overlap to chromatin immunoprecipitation data and to the microarray expression profile of a strain ectopically expressing *PDR8*.

27. 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.

28. 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.

29. Horak CE, Luscombe NM, Qian J, Bertone P, Piccirrillo S, Gerstein M, Snyder M: **Complex transcriptional circuitry at the G1/S transition in Saccharomyces cerevisiae**. *Genes Dev* 2002, **16**:3017-3033.

30. • Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA *et al.*: **Computational discovery of gene modules and regulatory networks**. *Nat Biotechnol* 2003, **21**:1337-1342.

This is an analysis of gene expression data in combination with chIP–chip data. The method — 'GRAM' (genetic regulatory modules) — initializes modules based on strict binding and expression coherence and then relaxes the stringency to allow those genes showing either strong co-expression or binding.

31. Zeitlinger J, Simon I, Harbison CT, Hannett NM, Volkert TL, Fink GR, Young RA: **Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling**. *Cell* 2003, **113**:395-404.

32. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne J-B, Reynolds DB, Yoo J *et al.*: **Transcriptional regulatory code of a eukaryotic genome**. *Nature* 2004, **431**:99-104.

33. Man TK, Stormo G: **Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay**. *Nucleic Acids Res* 2001, **29**:2471-2478.

34. Bulyk M, Johnson P, Church G: **Nucleotides of transcription factor binding sites exert inter-dependent effects on the binding affinities of transcription factors**. *Nucleic Acids Res* 2002, **30**:1255-1261.

35. Benos P, Bulyk M, Stormo G: **Additivity in protein–DNA interactions: How good an approximation is it?** *Nucleic Acids Res* 2002, **30**:4442-4451.

36. Sandelin A, Hoglund A, Lenhard B, Wasserman WW: **Integrated analysis of yeast regulatory sequences for biologically linked clusters of genes**. *Funct Integr Genomics* 2003, **3**:125-134.

37. • Gupta M, Liu JS: **Discovery of conserved sequence patterns using a stochastic dictionary model**. *J Am Stat Assoc* 2003, **98**:55-66.

This paper describes the 'stochastic dictionary data augmentation' (SDDA) algorithm, the latest in a series of Gibbs sampling algorithms for fitting statistical models that incorporate positional weight matrices (PWMs) to sets of promoter sequences. Unlike previous models in the series, the described model uses a stochastic PWM dictionary (an extension of previous work by Bussemaker) that explicitly allows for multiple motifs per promoter sequence.

38. Roven C, Bussemaker HJ: **REDUCE: an online tool for inferring cis-regulatory elements and transcriptional module activities from microarray data**. *Nucleic Acids Res* 2003, **31**:3487-3490.

39. •• Beer MA, Tavazoie S: **Predicting gene expression from sequence**. *Cell* 2004, **117**:185-198.

This paper describes a two-stage algorithm for fitting a Bayesian network which describes the combinatorial transcription factor (TF) regulation of yeast gene expression. This algorithm is one of the first to incorporate successfully constraints on TF-binding-site position, TF-binding affinity, and relative location into its predictions. The algorithm is tested on yeast, rediscovering some well-known regulatory constraints, and is used to make predictions about transcriptional regulation in *Caenorhabditis elegans*.

40. Cliften PF, Hillier LW, Fulton L, Graves T, Miner T, Gish WR, Waterston RH, Johnston M: **Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis**. *Genome Res* 2001, **11**:1175-1186.

41. •• 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.

See also Kellis *et al.* [42••]. These two papers describe sequencing of genomes of several related yeast species. A major objective was identification of potential regulatory elements.

42. •• 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 Cliften *et al.* [41••].

43. Blanchette M, Tompa M: **FootPrinter: a program designed for phylogenetic footprinting**. *Nucleic Acids Res* 2003, **31**:3840-3842.

44. Pritsker M, Liu YC, Beer MA, Tavazoie S: **Whole-genome discovery of transcription factor binding sites by network-level conservation**. *Genome Res* 2004, **14**:99-108.

45. 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.

46. •• Herrgard MJ, Covert MW, Palsson BO: **Reconciling gene expression data with known genome-scale regulatory network structures**. *Genome Res* 2003, **13**:2423-2434.

This is a thorough and objective evaluation of whether different transcriptional network models and motifs are consistent with microarray expression data in *Escherichia coli* and *S. cerevisiae*.

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

48. 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.

49. Tringe SG, Wagner A, Ruby SW: **Enriching for direct regulatory targets in perturbed gene-expression profiles**. *Genome Biol* 2004, **5**:R29.

50. 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.

51. Banerjee N, Zhang MQ: **Identifying cooperativity among transcription factors controlling the cell cycle in yeast**. *Nucleic Acids Res* 2003, **31**:7024-7031.

52. Haverty PM, Hansen U, Weng Z: **Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification**. *Nucleic Acids Res* 2004, **32**:179-188.

53. Chen G, Hata N, Zhang MQ: **Transcription factor binding element detection using functional clustering of mutant expression data**. *Nucleic Acids Res* 2004, **32**:2362-2371.