Cell, Vol. 117, 185–198, April 16, 2004, Copyright ©2004 by Cell Press

Predicting Gene Expression from Sequence

Michael A. Beer and Saeed Tavazoie*
Lewis-Sigler Institute for Integrative Genomics
and Department of Molecular Biology
Princeton University
Princeton, New Jersey 08544

Summary

We describe a systematic genome-wide approach for learning the complex combinatorial code underlying gene expression. Our probabilistic approach identifies local DNA-sequence elements and the positional and combinatorial constraints that determine their context-dependent role in transcriptional regulation. The inferred regulatory rules correctly predict expression patterns for 73% of genes in *Saccharomyces cerevisiae,* utilizing microarray expression data and sequences in the 800 bp upstream of genes. Application to *Caenorhabditis elegans* identifies predictive regulatory elements and combinatorial rules that control the phased temporal expression of transcription factors, histones, and germline specific genes. Successful prediction requires diverse and complex rules utilizing AND, OR, and NOT logic, with significant constraints on motif strength, orientation, and relative position. This system generates a large number of mechanistic hypotheses for focused experimental validation, and establishes a predictive dynamical framework for understanding cellular behavior from genomic sequence.

Introduction

At the heart of the complexity of multicellular life is the proper context-dependent expression of genes. To achieve this, cells have evolved a highly interconnected transcriptional network composed of signaling molecules, transcription factors (TFs), and their DNA targets (Levine and Tjian, 2003). The mRNA expression level of a gene is typically determined by several input signals, through the *cis*-regulatory logic encoded in its noncoding regulatory DNA sequences (Davidson et al., 2003). This *cis*-regulatory logic is fundamental to many processes, including physiological adaptation, generation of cell diversity, and morphological development.

With the arrival of whole-genome approaches for measuring the expression of genes, and computational methods for extracting biological insights from such data, there is an emerging movement to learn the structural and dynamical properties of transcriptional networks on a genomic scale (Eisen et al., 1998; Gardner et al., 2003; Hughes et al., 2000; Lee et al., 2002; Tavazoie et al., 1999). An important and fortuitous aspect of transcriptional network organization is that large sets of genes tend to be coexpressed at the mRNA level (Eisen et al., 1998; Tamayo et al., 1999; Tavazoie et al., 1999), consistent with the notion that many cellular processes require the simultaneous participation of many gene products. Comparative analysis has shown that the coexpression of many of these genes are conserved across diverse species (Stuart et al., 2003), but little is known about the underlying mechanisms by which these genes are regulated. Pattern recognition algorithms can be used to identify overrepresented DNA sequence elements, or motifs, in the presumptive regulatory regions of these groups of coexpressed genes (Tavazoie et al., 1999). In the yeast *S. cerevisiae*, many of these motifs correspond to previously known transcription factor binding sites (Tavazoie et al., 1999). However, the presence of a single motif is only marginally predictive of a gene's expression pattern. This reflects the extent of combinatorial regulation even in this simple eukaryote, where the expression level of a gene can depend on the occupancy states of multiple TF binding sites. Consistent with this, many *S. cerevisiae* genes have been experimentally shown to bind multiple TFs within their regulatory regions (Lee et al., 2002).

In this article we describe a computational approach for inferring the *cis*-regulatory logic of transcriptional networks from genome-wide mRNA expression data and DNA sequence. We use a probabilistic framework that is complementary to classical genetic techniques: after finding sets of coexpressed genes, our approach identifies the common, but potentially complex, DNA sequence features which are responsible for their regulation. We begin with a set of microarray expression data, and use a clustering algorithm (Hartigan, 1975) to find diverse sets of genes that are coexpressed across a set of conditions. Each of these sets of genes defines a distinct expression pattern across the experimental conditions (others have used the term module, or regulon). We then find a large set of putative regulatory DNA elements, or motifs, which are overrepresented in each expression pattern (Lawrence et al., 1993; Neuwald et al., 1995; Roth et al., 1998). A Bayesian network (Friedman et al., 2000; Pearl, 1988) is then used to infer the mapping between these sequence elements and the expression patterns. The network uses each gene's 5′ upstream sequence elements and their related properties as input variables, and outputs the probability of having a particular expression pattern. Thus, the inferred network describes the set of sequence elements and the positional and combinatorial constraints required for a gene to be expressed in a particular expression pattern. The network encodes that part of the *cis*-regulatory code which is active under the experimental conditions explored in the expression data.

Results and Discussion

Predicting Gene Expression

Systematic experimentation has acquired a detailed understanding of the mechanisms of transcriptional regulation for a handful of well-studied genes, but we lack the tools to achieve this level of understanding on a whole-genome scale. Here, as a formal step in this direc-

*Correspondence: tavazoie@molbio.princeton.edu

tion, we quantify the degree to which we can predict a gene's expression pattern by looking only at its regulatory sequences. We separate the genes into two sets, a training set where we will learn the regulatory DNA elements and combinatorial rules, and a test set which we will reserve only for prediction, or evaluation of our model.

The results of this approach have several fundamental biological implications. First, we are able to measure the degree to which gene expression is determined by local sequence, and we find that it is, perhaps surprisingly, high. Second, we can globally evaluate the degree to which different types of combinatorial regulation are utilized across the experimental conditions explored in the dataset. Third, we generate a set of high confidence predictions for regulatory DNA sequence elements, and the positional and combinatorial constraints that determine their function. Thus for thousands of genes, simultaneously and systematically, our approach finds the set of DNA sequence elements most likely to be responsible for each gene's proper context dependent expression.

While our automated approach is generally applicable to any microarray expression dataset, here we combine environmental stresses (Gasch et al., 2000) and cell cycle (Spellman et al., 1998), for 255 total conditions. This dataset explores a diverse set of experimental conditions, and the significant redundancy improves signal to noise. Noise in the expression data may present the greatest limitation on our ability to correctly predict gene expression, and imposes certain constraints on our approach. We must deal with the fact that under each condition, the measured gene expression level may be significantly different than the gene's actual expression. The degree to which coregulated genes are actually coexpressed in the expression data is demonstrated in Figure 1. For purposes of visualization, in Figure 1A, we have used a force-directed placement algorithm (Davidson et al., 2001; Kim et al., 2001) which places highly correlated genes near each other. This visualization emphasizes the fact that gene expression is continuous, not discrete, and that groups of coregulated genes are not generally distinct, but overlap. Figure 1A shows two large responses, a stress induced response in the mid-lower left, and a stress-repressed response in the top of the figure. But within these large sets are smaller groups of genes, with tighter coexpression, which participate in common biological processes. We find these expression profiles using a modification of the standard *k*-means algorithm (see Experimental Procedures). The number of expression patterns is determined automatically, and for what follows, we focus on a clustering which assigned 2587 genes to 49 expression patterns. These expression patterns are significantly enriched for genes of similar function, as shown in Table 1, using Bonferoni corrected P-values from the hypergeometric distribution. The mean of each of these expression patterns is shown in Supplemental Data, Supplemental Figure S1 available at http://www.cell.com/cgi/content/full/117/2/185/DC1. The genes in nine of these expression patterns are emphasized in color in Figure 1A.

While many of these expression patterns are similar over subsets of the data, using all conditions to distinguish between subtly different expression patterns allows us to learn their distinct modes of regulation. Figure 1B shows the expression of each gene in four of these expression patterns, as well as the mean of the expression pattern. For example, 138 genes participate in expression pattern (1), 122 of which are ribosomal proteins ($P < 8.5 \times 10^{-175}$). Expression pattern (4) has 114 genes, 21 of which are known to be involved in rRNA transcription ($P < 3.5 \times 10^{-14}$). While these sets of genes are very similar in expression, the subtle differences are potentially biologically significant: the rRNA transcription genes are more repressed under heat shock and turn off more rapidly under diamide treatment than the ribosomal proteins (see bottom of Figure 1B). We correctly predict 94% and 92% of the genes in these expression patterns, recapitulating the subtle difference in expression. Our predictions for these two sets of genes involve completely different programs of regulation: the ribosomal proteins are predicted to be regulated by the DNA element known to be bound by RAP1 (with significant constraints on its orientation and the presence of an appropriate regulatory partner), while the rRNA transcription genes are predicted to be regulated by the PAC and RRPE DNA elements (with constraints on their position relative to ATG), as discussed below. These results indicate that having two separate regulatory mechanisms for the production of the RNA and protein components of the ribosome may be important in the biology of yeast. Similar distinctions separate the stress-induced expression pattern (3) and the proteolytic degradation (proteasome) expression pattern (28).

Probabilistic Model

A Bayesian network (Pearl, 1988) describes relationships of probabilistic dependency between variables. In our case, we are interested in learning how a given gene will be expressed, under certain experimental conditions, given its 5′ upstream DNA sequences. While many methods could be applied to this task, we chose the Bayesian framework because of its natural way of dealing with incomplete information, and its ability to encode arbitrary dependencies between variables. The Bayesian approach to gene regulatory networks has been pioneered by Friedman and coworkers (Friedman et al., 2000; Segal et al., 2003), and our work is motivated by their progress. Their work (Friedman et al., 2000; Segal et al., 2003) builds regulatory networks by finding correlations between the mRNA levels of a regulatory gene (e.g., a known TF) and a regulated gene. However, many TFs are regulated by posttranscriptional mechanisms (e.g., nuclear import/export, phosphorylation, proteolytic degradation, interaction with small ligands, or at the level of translation). Ideally, we need to correlate the nuclear concentration of a transcription factor protein in its active state with a regulated gene's mRNA transcript abundance. However, we do not in general have this information.

Our approach circumvents this difficulty by building a network, which is not gene-to-gene, but is sequence-to-gene. The determinants of gene expression levels in our model are short DNA sequence elements, not transcription factor mRNA levels. These sequence elements serve as a proxy for the active nuclear concentration of a TF: if an active TF recognizes a particular DNA

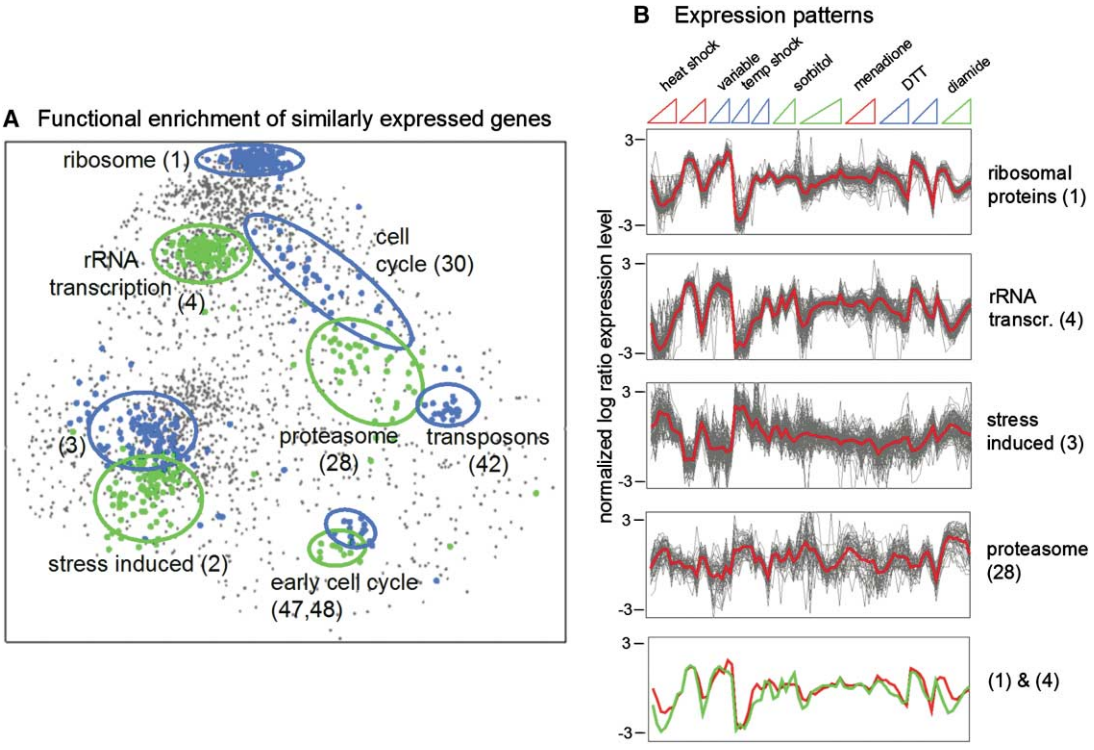

Figure 1. Examples of Expression Profiles and Their Functional Enrichment for Common Biological Processes

(A) Two-dimensional visualization of the expression data, where coexpressed genes are placed close to each other. Genes in nine of our 49 expression patterns are highlighted, showing that large sets of genes that are coexpressed share common functions.

(B) Four of our 49 expression patterns over the first 77 conditions in the dataset, and their average expression (red). Expression patterns (1) and (4) (compared at bottom) are similar, but the subtle differences select for different functions and distinct regulatory mechanisms.

sequence element, those genes with the sequence element will respond in a particular way—those genes without the element will not. Figure 2A shows a graphical representation of our framework. This approach can also describe regulation by chromatin-modifying complexes to the extent that they are targeted by a sequence specific factor (Kurdistani et al., 2002). Our approach may also describe the possibility that DNA structural elements (e.g., a particularly rigid or flexible region) may be involved in regulation, if such elements allow or impede access by other factors.

Several features of our approach turn out to be essential for predictive accuracy. (1) We learn DNA sequence motifs from expression patterns found in the expression dataset, so our set of motifs are those which are functional over the set of conditions we wish to predict. (2) We represent the motifs with position weight matrices (PWMs; Stormo and Fields, 1998), rather than consensus words or *k*-mers. (3) We learn the functional depth of each motif from the expression data, instead of using a fixed number of sites for each motif, as has been standard. (4) Our description of the sequence constraints is as general as possible, and potentially includes: the position of the motif relative to translation start (ATG), the orientation of the motif, the order and spacing between particular motifs, combinations of motifs, or the absence of motifs. (5) The mapping from sequence to expression is probabilistic and nonlinear, i.e., the expression level of each gene is not modeled as a linear superposition of effects of individual transcription factors. A nonlinear model allows flexibility for cooperativity between various transcription factors.

RRPE and PAC, a Case Study in Combinatorial Regulation

Two computationally discovered sequence elements, PAC and RRPE (Hughes et al., 2000; Tavazoie et al., 1999), exemplify the type of combinatorial regulation our network must describe in order to achieve predictive accuracy. These motifs were found in an expression pattern enriched for ribosomal RNA transcription and processing genes. If we take the top 404 genes with an upstream PAC element, and the top 403 genes with RRPE, 167 of these genes have both elements (by random chance we expect only 27 to have both). The degree of coregulation of any set of genes selected by a regulatory sequence constraint can be quantified by finding the distribution of pair-wise Pearson correlation coefficients, C_{ij}, for all genes in the set, as shown in the insert in Figure 2C. This probability distribution is a histogram of the observed correlation coefficients. Figure 2C shows this distribution for genes in three sets: those genes with only PAC, only RRPE, or both PAC and RRPE, compared to the distribution for all genes. Those genes with both elements are significantly more correlated than genes with just one element, reflecting their involvement in coregulation.

While genes with both PAC and RRPE are highly corre-

Table 1. MIPS Functional Category Enrichment for Each Expression Pattern, and the Number of Correctly Predicted Genes in the Training and Test Sets

Expression Pattern	Number of Genes	MIPS Functional Enrichment	P-Value	Number of Genes		Fraction	Number of Genes		Fraction
				In Training Sets	Correctly Predicted	Correctly Predicted	In Test Sets	Correctly Predicted	Correctly Predicted
1	138	122 ribosome biogenesis (215 ORFs)	8.5E-175	124	115	0.93	124	117	0.94
2	123	65 UNCLASSIFIED (2399 ORFs)	>0.1	113	86	0.76	113	75	0.66
3	115			107	92	0.86	107	84	0.79
4	114	21 rRNA transcription (109 ORFs)	3.5E-14	105	98	0.94	105	97	0.92
5	89	23 C-compound metabolism (415 ORFs)	1.6E-06	84	71	0.85	84	65	0.77
6	86	12 aminoacyl-tRNA-synthetases (37 ORFs)	5.3E-12	84	69	0.82	84	62	0.74
7	84	8 stress response (175 ORFs)	>0.1	82	75	0.92	82	66	0.80
8	82	50 UNCLASSIFIED (2399 ORFs)	5.1E-03	80	60	0.75	80	53	0.66
9	81			76	58	0.77	76	54	0.71
10	77	8 rRNA transcription (109 ORFs)	9.0E-03	68	61	0.90	68	61	0.90
11	76	8 peroxisome (39 ORFs)	2.6E-06	74	58	0.78	74	51	0.69
12	74	12 endoplasmic reticulum (157 ORFs)	4.2E-05	67	53	0.80	67	49	0.73
13	73	18 CELLULAR TRANSPORT (495 ORFs)	1.8E-03	69	58	0.85	69	55	0.80
14	72	9 TRANSPORT FACILITATION (313 ORFs)	>0.1	69	58	0.84	69	50	0.72
15	68	21 C-compound metabolism (415 ORFs)	2.1E-07	64	61	0.96	64	53	0.83
16	68	32 METABOLISM (1066 ORFs)	1.6E-06	67	51	0.77	67	49	0.73
17	68	12 rRNA transcription (109 ORFs)	2.5E-07	64	57	0.90	64	54	0.84
18	64	5 aminoacyl-tRNA-synthetases (37 ORFs)	5.7E-03	57	51	0.90	57	53	0.93
19	61			58	42	0.73	58	29	0.50
20	58	29 mitochondrion (366 ORFs)	7.6E-19	56	49	0.89	56	47	0.84
21	57	8 endoplasmic reticulum (157 ORFs)	1.4E-02	55	38	0.70	55	27	0.49
22	57	5 rRNA transcription (109 ORFs)	>0.1	53	37	0.71	53	34	0.64
23	55	18 CELLULAR TRANSPORT (495 ORFs)	1.8E-05	54	29	0.54	54	26	0.48
24	54	12 nitrogen and sulfur metabolism (67 ORFs)	3.6E-11	51	45	0.89	51	31	0.61
25	54	5 rRNA transcription (109 ORFs)	>0.1	50	37	0.75	50	40	0.80
26	53	10 rRNA transcription (109 ORFs)	3.2E-06	53	50	0.94	53	49	0.92
27	53	15 CELLULAR TRANSPORT (495 ORFs)	1.6E-03	53	39	0.75	53	28	0.53
28	53	25 proteolytic degradation (160 ORFs)	2.2E-24	52	39	0.75	52	30	0.58
29	52	7 nuclear transport (59 ORFs)	7.8E-05	49	41	0.84	49	39	0.80
30	52	31 CELL CYCLE AND DNA PROC. (628 ORFs)	2.8E-16	49	41	0.85	49	41	0.84
31	49	7 nuclear transport (59 ORFs)	5.1E-05	48	32	0.68	48	30	0.62
32	44	18 METABOLISM (1066 ORFs)	2.8E-02	42	34	0.81	42	30	0.71
33	42	11 C-compound metabolism (415 ORFs)	1.1E-02	39	31	0.81	39	27	0.69
34	42	4 cell wall (38 ORFs)	1.9E-02	40	21	0.54	40	18	0.45
35	39	8 cytoskeleton (108 ORFs)	4.3E-05	39	29	0.76	39	17	0.44
36	39	6 translation (64 ORFs)	4.0E-04	37	28	0.77	37	24	0.65
37	36	2 peroxisomal transport (16 ORFs)	>0.1	31	25	0.81	31	21	0.68
38	33	6 fermentation (33 ORFs)	2.3E-06	31	28	0.91	31	24	0.77
39	32			32	23	0.73	32	22	0.69
40	30	14 METABOLISM (1066 ORFs)	2.8E-02	29	24	0.83	29	21	0.72
41	29	20 mitochondrion (366 ORFs)	1.6E-16	27	20	0.74	27	17	0.63
42	23	transposons	NA	1	0	0.75	1	0	0.00
43	20	4 cytoskeleton (108 ORFs)	5.7E-02	18	14	0.78	18	10	0.56
44	20	8 stress response (175 ORFs)	5.4E-06	18	18	1.00	18	16	0.89
45	18	7 cell cycle (451 ORFs)	2.7E-02	18	14	0.79	18	12	0.67
46	17	5 protein folding(59 ORFs)	6.2E-05	15	13	0.88	15	12	0.80
47	17	14 UNCLASSIFIED (2399 ORFs)	4.8E-02	16	16	1.00	16	14	0.88
48	15	13 UNCLASSIFIED (2399 ORFs)	3.0E-02	8	7	0.94	8	7	0.88
49	14	9 glycolysis (35 ORFs)	6.2E-16	11	8	0.73	11	7	0.64
				2587	2120	0.82	2587	1898	0.73

We correctly predict the expression patterns of 73%, or 1898 of 2587 genes in the five test sets.

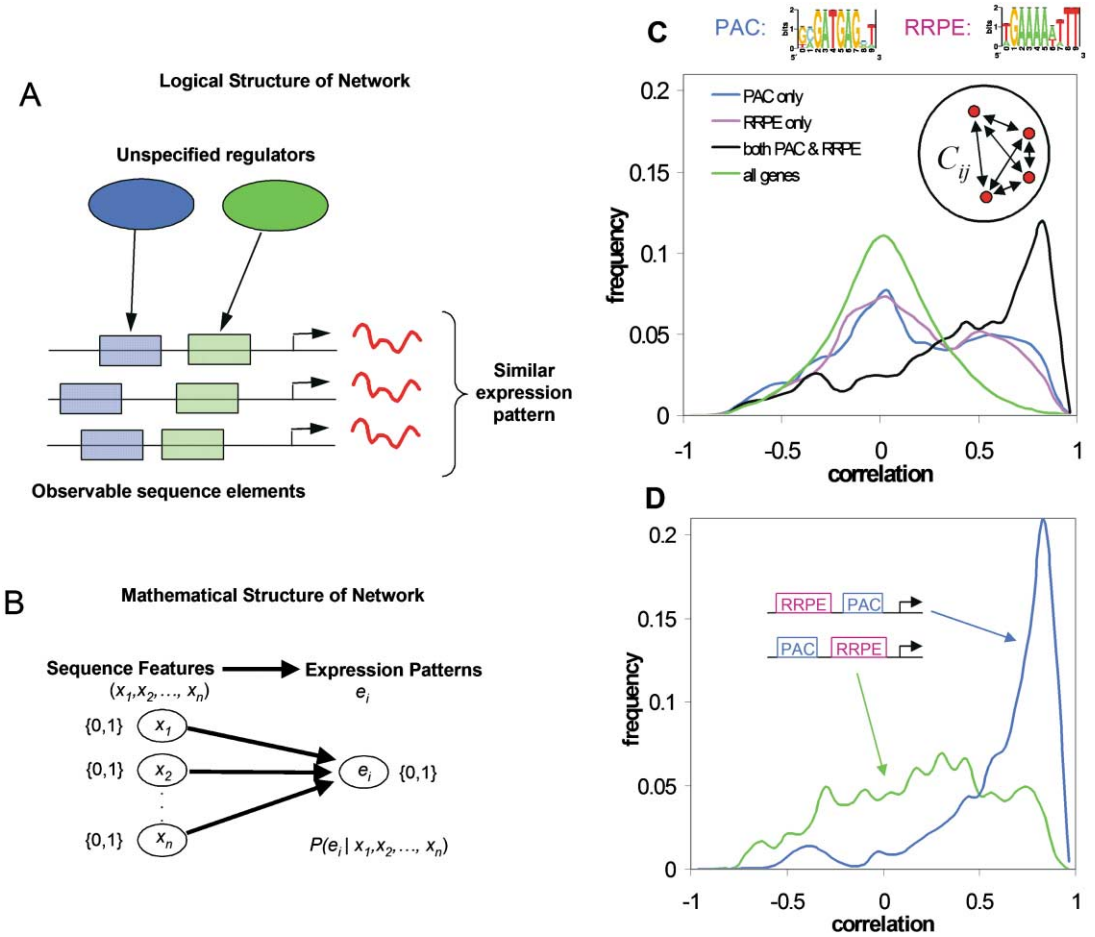

Figure 2. Sequence Elements that Determine the Regulation of a Set of Genes Involved in Ribosomal RNA Transcription and Processing

(A) Schematic of the logical structure of our approach. We use observable sequence elements to find complex promoter features, which are predictive of coexpression.

(B) The mathematical framework is a Bayesian network which maps general sequence features ($x_1 \ldots x_n$) to expression patterns (e_i) by encoding $P(ei|x1,x_2,\ldots,x_n)$, the probability that genes with these sequence features will participate in expression pattern i.

(C) Distribution of pair-wise correlation coefficients (insert) of all genes in the following sets: PAC only, RRPE only, both PAC and RRPE, and all genes. Genes with both PAC and RRPE are significantly more correlated.

(D) The effect of the order of the two elements relative to the promoter. Genes are much more tightly correlated if PAC is on the promoter side of RRPE.

lated, there are still many genes in that set which are not. By testing a large number of pair-wise sequence constraints, we find that the order of the two elements strongly affects the degree of correlation. If PAC is closer to the promoter than RRPE, the genes are much more correlated than if RRPE is closer to the promoter (Figure 2D). It has been previously noted that there are also statistically significant biases in the spacing between the two motifs, distance to ATG, and orientation, in addition to order (Hughes et al., 2000; Pilpel et al., 2001; Sudarsanam et al., 2002; Tavazoie et al., 1999). While we find that order is a significant determinant of coregulation, each of these alternative constraints also selects a more correlated subset of the genes that have both PAC and RRPE. We can find many rules that select sets of coregulated genes, but the Bayesian approach finds the most probable constraint, that which makes the observed data most likely.

Systematic and Genome-Wide Learning of Combinatorial Regulatory Rules

To incorporate specific combinatorial effects like PAC and RRPE, our description of motif interactions is as general as possible. We allow our probabilistic model to encode all the constraints on a motif discussed above: its presence in the 5′ upstream region of a gene, its orientation, its distance to ATG, its functional depth (PWM score cut-off for closeness to "consensus"), and the presence or absence of other motifs. If two or more motifs are present, we allow the interaction of any pair to be constrained by the distance between them, or by their order relative to the promoter. Because the Bayesian network encodes a joint probability distribution, any of these constraints may be satisfied individually or in particular combinations. These sequence constraints are represented by variables x_i, which are either satisfied for a particular gene ($x_i = 1$), or not ($x_i = 0$).

These sequence constraints are input variables, or possible parent nodes, in our Bayesian network. The final network encodes the distribution $P(ei|x1,x_2,\ldots,x_n)$, the probability of being expressed ($e_i = 1$) or not being expressed ($e_i = 0$) in expression pattern i, given the states of the sequence constraints x_i. This mathematical structure is represented in Figure 2B.

We learn the structure and probability distributions of our Bayesian network using modifications of standard techniques (Heckerman, 1998). We search through sequence space to find the network (N) which maximizes the probability that our network is correct, given the data (D), using Bayes' rule: $P(N|D) = P(N)P(D|N)/P(D)$ (see Experimental Procedures). We would like to find the most probable network, but because it is computationally infeasible to score all networks, we use a greedy search through network space. To avoid local optima, we learn several ($\sim$10) networks from independent bootstrap samples (a random selection of N samples, with replacement, from the N samples in the training dataset). Because of this sampling, each of these networks can potentially find different sequence constraints, and each gives a prediction for the probability of each gene being expressed in a particular expression pattern. We average these probabilities to give a final prediction (Breiman, 1996).

Gibbs sampling (AlignACE; Roth et al., 1998) is performed on the 5′ upstream sequences of the genes in each of the 49 expression patterns (described above and in Experimental Procedures) to find overrepresented sequences, which we represent by position weight matrices (Stormo and Fields, 1998). The predictive power of these motifs can be measured by a Bayesian score, which is further optimized using Monte-Carlo simulated annealing (see Experimental Procedures). We then score all sites in the genome for closeness to each of these motifs (ScanACE; Hughes et al., 2000), and normalize the score of each motif to the maximum possible score.

To assess the predictive performance of our network, before inference, we randomly partition the genes into 5 test sets for crossvalidation. We then infer five networks, using 80% of the genes as a training set, and 20% as a test set. Examples of the sequence constraints selected are shown in Figure 3. Figure 3A shows a network for the "ribosomal RNA transcription" expression pattern (4). For this bootstrap training sample, the network growth stopped after two parent nodes were added: PAC and RRPE, constrained by distance to ATG. By looking at all genes in the training set, the network finds that if the PAC element is not within 140 bp of ATG, and RRPE is not within 240 bp of ATG, the probability of being in expression pattern (4) is only 1%. If PAC is not within 140 bp of ATG, but RRPE is within 240 bp, the probability of being in expression pattern (4) is 22%. If PAC is within 140 bp of ATG, but RRPE is not within 240 bp, the probability of being in expression pattern (4) rises to 67%. Finally, having both constraints satisfied increases a gene's probability of being in the expression pattern to 100%. We refer to this as **AND** logic. To confirm this result from the network, the correlation distribution for all genes with PAC within 140 bp and RRPE within 240 bp, shown in Figure 4A, is as tight as the order constraint in Figure 2D, but applies to a larger set of genes and is thus better supported by the data. In addition, the set of genes which have PAC and RRPE with the learned functional depth, but which do not satisfy the distance to ATG constraint, have very low correlation, indistinguishable from random genes (Figure 4A). Since PAC is constrained to be closer to ATG (140 bp) than RRPE (240 bp), most of these genes have PAC on the promoter side of RRPE, but there exist coregulated genes with RRPE closer to the promoter than PAC, and by selecting the position constraint over order, the network has chosen the maximally predictive constraint. This clear delimitation of the set of coregulated genes would not have been obtained without simultaneously varying the thresholds for both the functional depth (closeness to consensus) and the distance to ATG. In other expression patterns, RRPE and PAC operate with other sequence elements (see Supplemental Data available on *Cell* website). Examples of genes which have PAC and RRPE and satisfy the positional constraint are shown in Figure 4B. *DRS1* is a putative ATP dependent RNA helicase, *RRB1* is involved in ribosome assembly, *RPA49* is the 49 kDa subunit of RNA polymerase A, and *DIM1* is a dimethyladenosine transferase involved in 35S primary transcript processing. These functions are consistent with their tight regulation, as shown on the right of Figure 4B. The genes in Figure 4C have strong PAC and RRPE elements, but do not satisfy the positional constraint. *ATP5* is an ATP synthase, *NMT1* is an N-myristoyl transferase, *ESBP6* is a putative monocarboxylate permease, and *PTP2* is a protein tyrosine phosphatase. These genes are completely uncorrelated, as shown on the right of Figure 4C.

Other common logical constraints inferred by the network are redundancy (**OR** logic), or the requirement for the absence of a particular motif (**NOT** logic). A network learned on a stress-induced expression pattern (Figure 3B) demonstrates both. Here, the STRE-like elements (x_1 and x_2) are similar, but select different sets of genes. These two motifs were chosen over the canonical STRE (AGGGG; Martinez-Pastor et al., 1996), and x_2 has a strict requirement for upstream Ts. Genes with x_1 are in expression pattern (2) 59% of the time, in the absence of the other two motifs, and genes with only x_2 are in expression pattern (2) 75% of the time, but the sets are largely distinct (only five genes have both x_1 and x_2). This is an example of **OR** logic, but it could also be due to an imperfect representation of the same underlying binding site. RRPE (x_3), on the other hand, is selected as a constraint because its presence guarantees that the gene will not be in this expression pattern (**NOT** logic), even if x_1 or x_2 is present. Any genes with x_3 have a zero probability of participating in this expression pattern (green boxes in Figure 3B).

The ribosomal protein expression pattern demonstrates another example of redundancy (Figure 3C). RAP1 is the main regulator of ribosomal proteins in *S. cerevisiae*, and 90% of the 137 ribosomal protein genes are reported to have a RAP1 binding site upstream (Lascaris et al., 1999). However, many of these RAP1 sites deviate from the consensus binding site. At a normalized depth (motif score) of 0.6, 81% of the ribosomal proteins have a RAP1 binding site, but so do 273 other genes, most of which are not coexpressed with the ribosomal protein genes. What determines which of these RAP1 binding sites are functional? In the first bootstrap sample, the network chooses RAP1 in

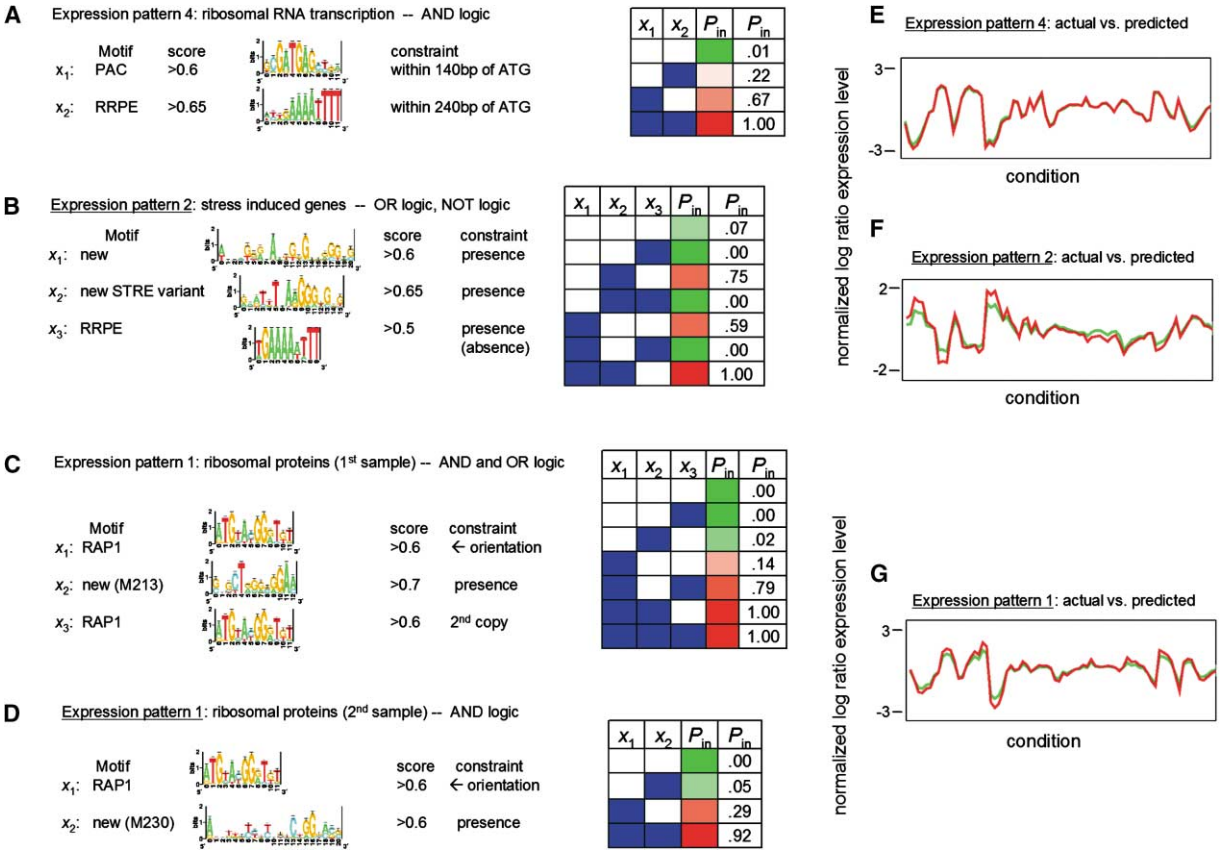

Figure 3. Examples of Network Structures Inferred from Expression Data and Regulatory Sequence

(A–D) Sequence constraints selected by each network and the combinations which are predictive of a particular expression pattern. For each network, all realized states of the sequence variables, x_i, and the fraction (probability) of genes participating in the expression pattern for each state of the sequence variables are shown: filled (blue) for $x_i = 1$, and empty for $x_i = 0$. Red indicates high, and green indicates low probability of being in the expression pattern.

(E–G) Actual mean expression pattern (red) and predicted mean expression pattern (green) for all genes predicted to participate in patterns (1), (2), and (4). For prediction, each gene is assigned to the most probable pattern found from all networks.

the ← orientation with the regulatory partner M213, and finds that a second copy of RAP1 can substitute for this regulatory partner. It was previously observed that two RAP1 binding sites can activate transcription synergistically (Woudt et al., 1986). If a gene has RAP1 in the ← orientation and either M213 or a second copy of RAP1, the gene will be expressed in pattern (1) 100% of the time, while if both M213 and a second copy of RAP1 are absent, the chances of being expressed in pattern (1) drop to 14%. Without one RAP1 in the ← orientation, the presence of either M213 or a second copy of RAP1 → is insufficient to produce this expression pattern (2% and 0%, respectively).

A second resampling (Figure 3D) finds a second regulatory partner, M230. If a gene has a RAP1 in the ← orientation and M230, there is a 92% chance that it will be regulated like a ribosomal protein, in expression pattern (1). If there is only RAP1 ←, the chances of being in expression pattern (1) drop to 29%. These results indicate that these motifs can substitute for the second copy of RAP1, given the proper orientation of the first RAP1 binding site.

The distribution of correlation coefficients for the genes selected by the predictive rule [RAP1← and (M213 or M230 or a second copy of RAP1)] is shown in Figure 4D. Also shown is the distribution for an equally strong RAP1 →. These rules are clearly able to distinguish a set of genes that are tightly expressed in pattern (1). Examples of genes that are selected by these rules are shown in Figure 4E: all are ribosomal proteins. Examples of genes which have an equally strong RAP1 binding site, but in the → orientation, are shown in Figure 4F. These genes are known to be involved in other processes, and do not participate in expression pattern (1).

Many of the motifs chosen by the network closely match one of roughly 20 known regulatory elements, and citations to the experimental support are included in the Supplemental Data (available on *Cell* website). But the subtle differences appear to be significant for prediction, since our motifs learned by Gibbs sampling and optimization were selected over the known motif set by a substantial margin. More examples of learned regulatory rules for other expression patterns are described in Supplemental Data, Supplemental Figure S2 available on *Cell* website. A list of the most frequently learned motifs is shown in Supplemental Data, Supplemental Table S1 available on *Cell* website.

The types of constraints learned by the network indicate the prevalence of various modes of combinatorial regulation in *S. cerevisiae* (see Supplemental Data avail-

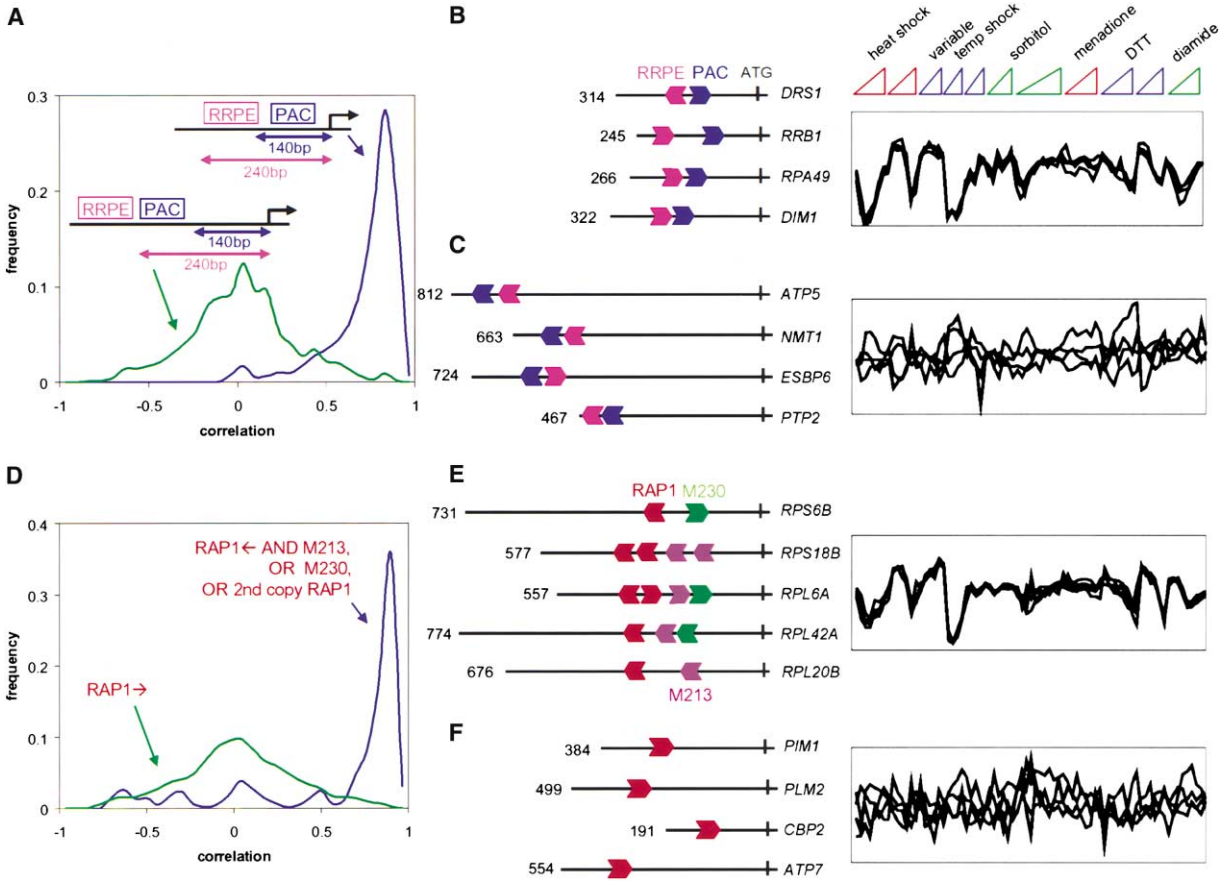

Figure 4. Sets of Genes Selected by the Inferred Constraints for Expression Patterns (4) and (1) (rRNA transcription and ribosomal proteins).

(A) When PAC is within 140 bp of ATG and RRPE is within 240 bp of ATG, the genes are tightly coregulated (blue). When PAC and RRPE are further from ATG, coregulation is lost, and the distribution of correlations is close to random, (compare to Figure 2C).

(B) Examples of genes that satisfy the positional constraint and their expression pattern (right).

(C) Examples of genes that do not satisfy the positional constraint and their expression (right).

(D) When RAP1 is present in the ← orientation with the motifs M213, M230, or a second copy of RAP1, the genes are tightly coregulated (blue). When an equally strong RAP1 ← is present alone (data not shown), or in the → orientation, the distribution is close to random (green).

(E) Examples of genes that satisfy the orientation and partner constraints, and their expression (right).

(F) Examples of genes that have an equally strong RAP1 →, and their expression (right).

able on *Cell* website). It is important to note that the degree of combinatorial regulation uncovered here represents a lower limit, and a broader sampling of physiological conditions may yield a higher average number of regulators per gene and perhaps more complex rules.

Predicting Gene Expression Patterns from Sequence

Postgenome biology is largely defined by the challenge of mapping genetic information to phenotype. Because of its direct physical coupling to the gene, mRNA expression dynamics provides the most proximal "phenotype" for addressing this challenge. In this context, predictive accuracy is the objective arbiter of how well we understand this process. The more accurate our predictions are, the more likely our model is capturing the essential underlying mechanisms. To this end, we assess the network's ability to predict expression patterns of genes from sequences which it had not seen before (the 20% test set). We infer rules using the 80% training set genes, and then predict the expression pattern of the reserved test set genes by only looking at their promoter sequences. We repeat this for each of the five test sets.

Even small noise levels in the expression data would make it impossible to predict a gene's expression pattern exactly, so we must incorporate a reasonable amount of flexibility in our assessment of what qualifies as a correct prediction. We do so by predicting a gene's participation in one of the 49 expression patterns described above. Thus to be correctly predicted, a gene must be predicted to be participating in the expression pattern to which it is closest. The correlation coefficient cut-off ($C > 0.65$) used to define these expression patterns seems appropriate given the measurement noise, and the strong functional enrichments we found in Table 1. A complication is that some of these expression patterns are very similar or overlapping, e.g., expression patterns (2) and (3) or (47) and (48) shown in Figure 1A. Thus to avoid penalizing for prediction in a very close expression pattern, we consider a prediction in an overlapping expression pattern correct. Overlapping expression patterns are defined to have correlation between

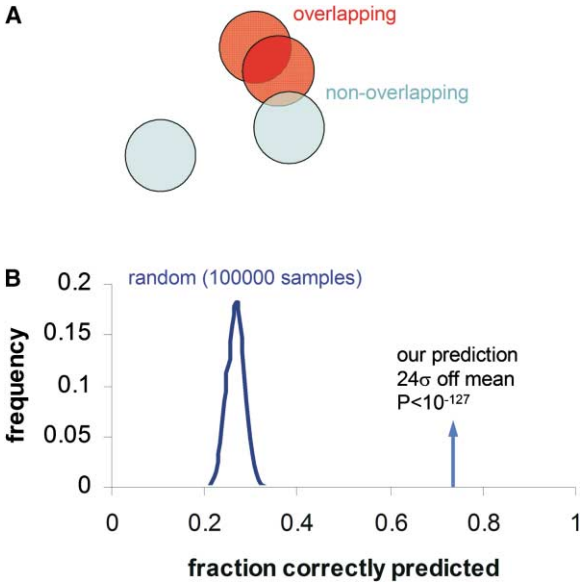

Figure 5. Statistical Significance of 73% Prediction Accuracy

(A) Because some of our 49 expression patterns are similar, we consider prediction in an overlapping pattern (red) correct, but prediction in a nonoverlapping pattern (blue) is incorrect. Overlapping patterns must have mean expression patterns which are correlated better than $C > 0.65$, the radius of each pattern.

(B) 100,000 independent random assignments of all genes to expression patterns always get near 26.6% correct. The distribution is close to normal and has a standard deviation of 1.9%, so the P-value for our prediction of 73%, 24 standard deviations from the mean, is $P < 10^{-127}$.

their mean expression greater than 0.65, the same as the cut-off defining the radius of an expression pattern, as shown schematically in Figure 5A. Prediction in an overlapping expression pattern most frequently occurs when (perhaps because of noise) a gene is assigned to a neighboring expression pattern but is actually successfully predicted using rules learned in a different, but very similar, expression pattern.

We correctly predict the expression patterns of 73%, or 1898 of the 2587 clustered genes in the five test sets, as shown in Table 1. Each gene is in four crossfold validation training sets (these are averaged) but only one test set (these are combined). This degree of accuracy is highly significant. If no expression patterns overlapped we would expect to predict correctly 1/49 or approximately 2% of the time. Because there is significant overlap of some expression patterns, randomly assigning genes to expression patterns gives 26.6% $\pm$ 1.9% correct on average. The distribution of 100,000 independent random assignments for all genes is shown in Figure 5B. Since these independent random samples are normally distributed to a good approximation, the P-value for our prediction of 73% is $<10^{-127}$.

Another measure of our ability to predict gene expression, which does not have complications due to overlapping expression patterns, is the distribution of correlation coefficients of each gene to its predicted expression pattern, as shown in Figure 6A. Three distributions are shown, comparing full networks, networks that are constrained to use only single motifs, and random assignment to expression patterns. Using only single motifs, with the optimal depth cut-off, the mean correlation of each gene to its predicted expression pattern is 0.36. With full networks the average number of parent nodes (selected motif elements or constraints) is 2.8, and the mean correlation of each gene to its predicted expression pattern is 0.51. While the percent of genes correctly assigned to an expression pattern increases from 26.6% to 73% using our full network, the average correlation of the genes to their predicted expression pattern increases dramatically, from 0.02 to 0.51. This increased correlation is a global measure of the degree to which we can predict gene expression, and the dramatic shift of the curve to higher correlation demonstrates that our global predictive accuracy approaches that detailed above for the specific examples shown for PAC/RRPE and RAP1 in Figure 4. The increase from single parents to full networks also indicates the significant degree of combinatorial regulation. Without resampling, the mean correlation is only 0.42. That resampling improves the results is primarily indicative of redundant modes of regulation: each resampling can learn different ways of regulating a particular expression pattern. For comparison, we have also tried multiple linear regression (Bussemaker et al., 2001) with our full motif set, and find that it performs somewhat worse than single motifs (mean correlation of 0.25), likely due to overfitting.

To provide a visual demonstration of our predictive accuracy, we compare the actual expression profiles of genes in one test set to the model's prediction based solely on sequence (Figure 6B). The actual expression profile of the test set was hierarchically clustered (Eisen et al., 1998), and our predicted expression profile for each gene is displayed in the same order. As can be seen, except for the $\sim$27% of genes which we fail to predict correctly, there is impressive global concordance between the two sets of profiles. These predictions are somewhat "coarse-grained," presenting us with the challenge of extracting increasingly subtle features from future refinements of our model.

Application to *Caenorhabditis elegans*

S. cerevisiae provides a suitable test of our algorithm in an organism where much is known about gene regulation and where there is an abundance of high quality expression data. However, we are also interested in the applicability of our approach to multicellular organisms. As a preliminary test, we applied our algorithm to Affymetrix expression data collected during embryonic and later development in *C. elegans* (Baugh et al., 2003; Hill et al., 2000). The combined dataset contains 20 points in a time course from staged embryos: 1 oocyte sample, 13 embryonic samples, 5 larval samples, and 1 adult sample. We used 2000 bp of 5′ upstream regulatory sequence for each gene that was expressed significantly in this dataset (5547 genes in 30 expression patterns). Given the larger regulatory DNA sequences, the potentially more complex regulation, and the tissue dependent expression of many genes, we were surprised to find that we could predict the expression patterns of roughly half of these genes. In this initial study, we have not looked for regulatory elements in introns or downstream regions. We have also ignored the effect of oper-

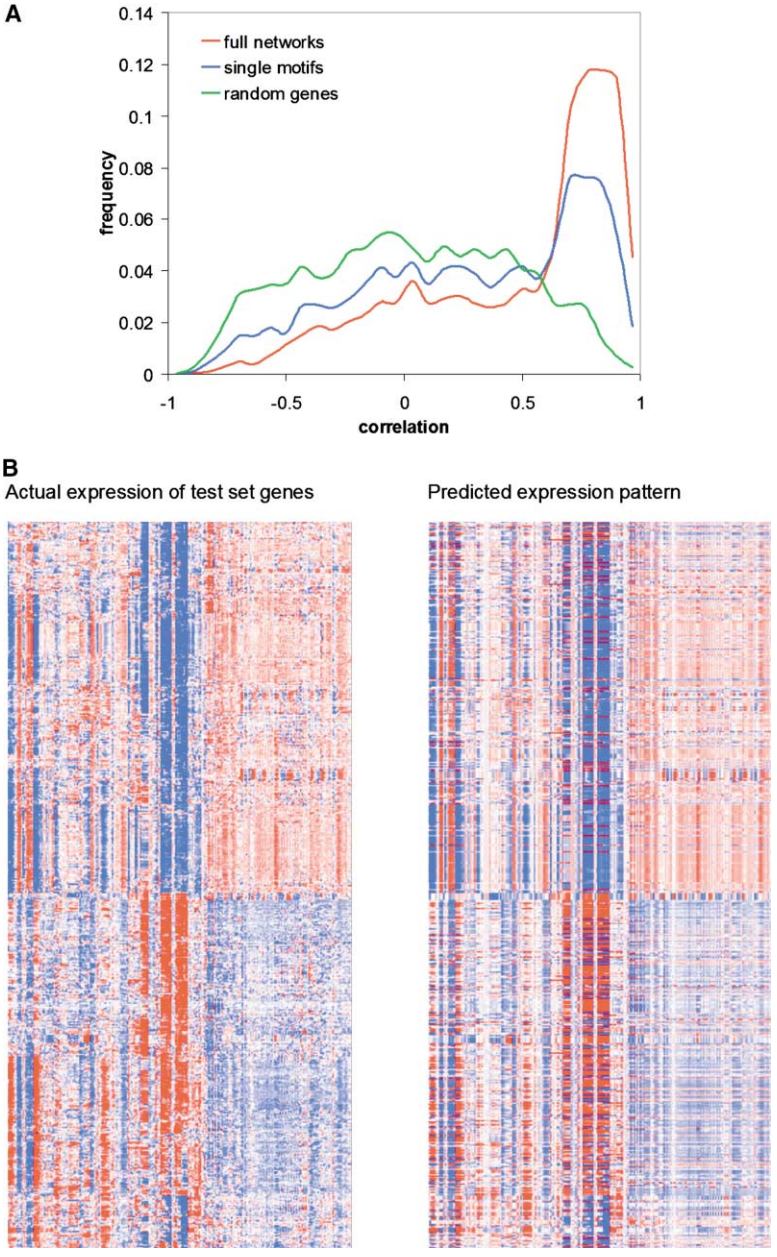

Figure 6. Global Predictive Accuracy

(A) The distribution of correlation coefficients of each gene to its predicted expression pattern, using full networks, single motifs, and random assignment. The mean correlation for full networks is 0.51, compared to 0.02 for random assignment.

(B) Hierarchical clustering of the 518 genes in one of the five test sets (left) and our prediction (right) for each of these genes, over all 255 conditions of the dataset.

ons and alternative splicing. In addition, gene architecture predictions are less well established in *C. elegans* compared to *S. cerevisiae*. All of these effects could change the relevant regulatory regions for some genes and including them should improve our predictive accuracy.

Comparatively little is known about transcription factor binding sites in *C. elegans*. Figure 7 shows the expression profiles for four of these expression patterns and the regulatory programs responsible for their expression. Expression patterns (4), (15), and (28), are strongly enriched for TFs (genes in the gene ontology classification "DNA dependent regulation of transcription") with Bonferroni corrected P-values of 9.3×10^{-18}, 3.6×10^{-17}, and 3.9×10^{-6}, respectively. Expression patterns (4), (15), and (28) peak in the late, middle, and early phases of embryonic development, respectively. Expression pattern (23) is enriched for genes involved in cell motility ($P < 3.2 \times 10^{-25}$), and peaks during larval phases of development. Genes whose promoters satisfy the inferred combinatorial rules recapitulate the phased temporal expression of the original patterns, as shown by the dashed lines in Figure 7G. Expression pattern (28) is found to be regulated by three motifs: M320, M324, and M323. The constraints governing the function of these motifs include AND and OR logic, and spatial constraints, as shown in Figure 7A. Motif M324 selects coregulated genes by itself if it is within 180 bp of ATG. M320 is functional only in combination with either M323, or M324, with the positional constraint that M320 and M324 must be within 160 bp of each other. The 83 genes which satisfy the combined rule [(M324 within 180 bp

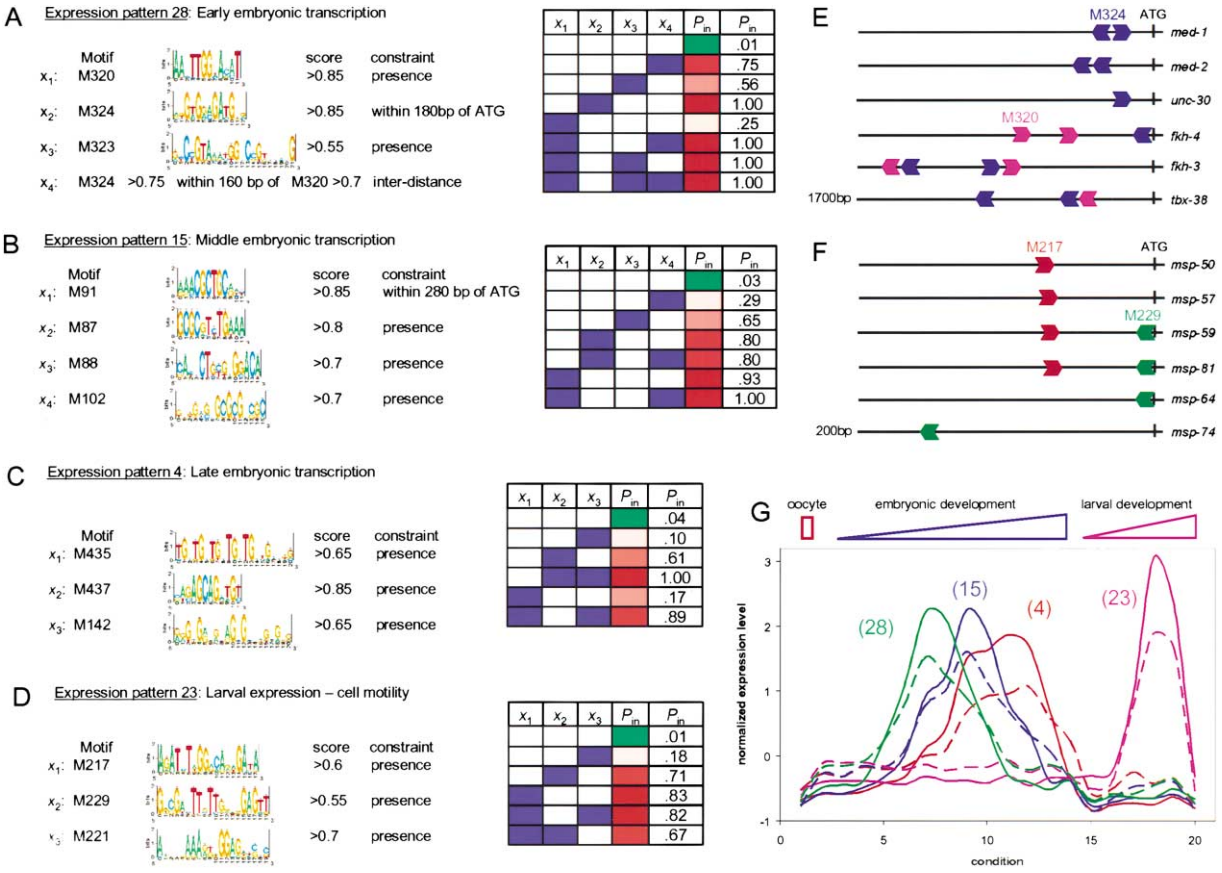

Figure 7. Examples of Regulatory Networks Learned for Embryonic and Larval Development in *C. elegans*

(A–D) Sequence constraints selected by each network and the combinations which are predictive of a particular expression pattern, as in Figure 3. Expression patterns (28), (15), and (4), peak during early, middle, and late embryonic phase, and are enriched for genes involved in transcriptional regulation. The three patterns are predicted to be turned on by distinct sets of motifs. Expression pattern (23) peaks during larval development and is enriched for genes involved in cell motility.

(E) Examples of known transcription factor genes that are selected by the network for expression pattern (28).

(F) Examples of major sperm proteins that are selected by the network for expression pattern (23).

(G) Actual expression patterns (solid), and predicted expression patterns (dashed) using the sequence constraints (A–D) which confer high probability of participating in each pattern.

of ATG) or (M320 and M323) or (M320 and M324, within 160 bp of each other)] are highly correlated, and their expression peaks in early embryonic development, as shown in Figure 7G. Among the genes selected by this constraint are the known transcription factors *med-1, med-2, unc-30, fkh-3, fhk-4*, and *tbx-38*, and their regulatory regions are shown in Figure 7E. The genes *med-1* and *med-2* have been shown (Maduro et al., 2001) to be regulated by the maternal transcription factor SKN-1 (Bowerman et al., 1992), which spatially restricts their expression to EMS. Motif M324 selects two sites 15 bp and 43 bp upstream of the SKN-1 consensus sites (An and Blackwell, 2003; Blackwell et al., 1994) in *med-1* and *med-2*. These sites are within the 180 bp fragment shown to be sufficient for proper expression of *med-1* and *med-2* (Maduro et al., 2001), and may contribute to the proper temporal regulation of these genes.

Expression pattern (15) is regulated by the rule [(M91 within 280 bp of ATG) or (M87) or (M88)], as shown in Figure 7B, which selects 202 genes. At the learned depth, M88 selects 64% of the 73 histone genes in *C. elegans*, and has strong similarity to a motif found to be overrepresented in the 5′ regulatory region of four histone genes (Roberts et al., 1989), but which has not previously been shown to be predictive of their expression. Examples of histone genes which are selected by this motif are shown in Supplemental Data, Supplemental Figure S3 available on *Cell* website. Expression pattern (4) is regulated by the rule [(M437) or (M435 and M142)], as shown in Figure 7C, which selects 226 genes. Expression pattern (23) is regulated by [M217 or M229], as shown Figure 7D, which selects 162 genes (M221 provides mild improvement). This rule selects 55% of the 47 major sperm protein genes (*msp*), responsible for the amoeboid locomotion of sperm cells (L'Hernault, 1997). M217 is close to a sequence found to be overrepresented in 10 *msp* genes (Klass et al., 1988), but which again has not been tested for its effect on expression. Examples of genes selected by these motifs are shown in Figure 7F. Figure 7G shows the actual expression patterns (solid) and the predicted expression patterns for genes that satisfy these constraints found by the

network (dashed), demonstrating that our approach is able to find sequence constraints that are predictive of proper expression during development.

Prevalent Themes and Biological Insights

The examples detailed above highlight the key discoveries of our approach. First, we find a great deal of redundancy in the modes of transcriptional regulation (**OR** logic). Second, many factors require at least one partner to be functional (**AND** logic). Third, one mode of combinatorial regulation is the absence of a factor that would cause a different mode of regulation (**NOT** logic). Finally, we can now account for a large fraction of the information required for the proper expression of genes in response to relevant physiological perturbations and developmental dynamics in the two model organisms. The fact that this information resides within their 5′ upstream regions provides a global statistical proof for this important dogma in molecular biology.

However, whether all the requisite information is resident in the local DNA, is an open question. Because of the statistical nature of our approach, we cannot correctly predict all genes. Higher-order combinatorial interactions may be difficult to learn, because they have few, or unique instances in the genome. Also, we may not be finding all of the relevant sequence features. Some relevant features may be downstream or within coding regions, or may be undetectable by standard motif finding algorithms. The proper description of some DNA regulatory elements may require nonadditive effects not included in our present position weight matrix description. Another potential limitation is that our heuristic learning algorithm may not be finding the optimal network. Finally, noise in the expression data may set a hard limit on our ability to learn the relevant sequence features and network structure.

But other failures may imply the existence of alternative regulatory mechanisms, e.g., because we learn the regulatory programs from local sequence, our failures may indicate genes where longer range interactions are important. A prominent cause of this type of failure may be silencing due to large scale chromatin modification near telomeres (Gottschling et al., 1990) and mating loci (Aparicio et al., 1991), boundary elements which inhibit local DNA sequences from signaling nearby genes (Kellum and Schedl, 1991), or similar mechanisms which set up chromosomal domains of gene expression (Cohen et al., 2000). The fact that our failures are not spatially clustered more than would be randomly expected indicates that such chromatin domains, if responsible for our failures, appear to be of intermediate scale. What is the role of local chromatin modifications? Are all such modifications subservient to the local sequence features that recruit transcription factors, which in turn recruit chromatin modifying machinery? These are important questions to address in future work, and we are currently in the process of exploring these possibilities.

Unlike the genetic code, the *cis*-regulatory code is not universal, requiring for individual genes, heroic experimental efforts to elucidate (Davidson et al., 2003). We have developed a whole-genome computational framework for the systematic extraction of this combinatorial code and prediction of gene expression patterns from DNA sequence alone. The large number of combinatorial rules which pass our predictive validation criterion, provide the community with a rich source of high-yield hypotheses for experimental analysis. Our success with *C. elegans* indicates that our general approach is applicable to multicellular eukaryotes, but the larger regulatory regions in these genomes still present a significant challenge. Also, combinatorial regulation is likely to be much more elaborate. In this setting, successful motif detection and predictive modeling will undoubtedly benefit from cross-species comparisons of regulatory regions.

The results presented here clearly demonstrate that a sufficiently general and systematic whole-genome approach is able to infer predictive regulatory constraints from mRNA expression data and DNA sequence alone. Our ability to decipher more complex regulatory programs is currently limited by the availability of gene expression data. From physiological perturbations and temporal expression responses at the organismal level, we have identified the regulatory information in many previously uncharacterized genes in *S. cerevisiae* and *C. elegans*. With the increasing availability of high quality tissue specific expression data in model organisms (Kim et al., 2001) and humans, our method presents a framework for rapidly elucidating the transcriptional regulatory mechanisms that orchestrate diverse spatiotemporal processes in multicellular organisms.

Experimental Procedures

Clustering

Our modified k-means clustering uses the standard algorithm (Hartigan, 1975), except that we constrain expression patterns to only include genes within some cut-off Pearson correlation coefficient, *C*, and we require each expression pattern to have 10 or more genes. Any expression pattern that does not satisfy the size constraint is reseeded from a random gene. With these constraints we choose the maximum number of expression patterns for which this algorithm converges. We have tried correlation cut-offs of $C = 0.6$, 0.65, and 0.7. At lower values of *C* more genes participate, but the expression patterns are less coherent. At $C = 0.65$, we find 49 expression patterns of 2587 genes, and we focus on this value. Expression patterns are considered overlapping if the correlation coefficient of their mean expression patterns is greater than this cut-off.

Motif Finding

We use AlignACE (Hughes et al., 2000; Roth et al., 1998; Tavazoie et al., 1999) with 12 bp motifs, and search up to 800 bp upstream of each gene. For *S. cerevisiae*, this yields a large set of motifs ($\sim$2000) with significant redundancy. We reduce the overlap in this set by allowing no two motifs to score 50% of the same sites with a normalized score >0.5, which reduces the set to 615 motifs. We augment this set with 51 known and experimentally documented TF binding sites (Hughes et al., 2000; Lee et al., 2002). Each motif is represented by a position weight matrix (PWM) and is graphically represented using sequence logos (Schneider and Stephens, 1990). The ability of each of these motifs to distinguish genes in the expression pattern from those not in the expression pattern can be measured with a Bayesian score, $P(N|D)$, given below, where in this case the network consists of single motifs as parent nodes. Using Monte-Carlo simulated annealing (Kirkpatrick et al., 1983), we further optimize the AlignACE motifs by perturbing columns of each PWM to maximize $P(N|D)$. Text files with the motifs, PWMs, their occurrences in genes, the expression data, and all expression patterns can be found in the Supplemental Data available on *Cell* website and at http://genomics.princeton.edu/tavazoie/Supplementary%20Data.htm.

Learning the Structure and Probability Distributions of the Bayesian Network

We incrementally add sequence constraints (parent nodes) which maximize the probability that our network is correct, given the data, using: $P(N|D) = P(N)P(D|N)/P(D)$. The probability of the data, D, given our network, N, assuming a Dirichlet prior is given by Cooper and Herskovits (1992), leading to:

$$P(N \mid D) = P(N) \prod_{i=1}^{n} \prod_{j=1}^{q} \frac{\Gamma(a_{ij})}{\Gamma(a_{ij} + N_{ij})} \prod_{k=0}^{r_i} \frac{\Gamma(a_{ijk} + N_{ijk})}{\Gamma(a_{ijk})} \quad (1)$$

where $\Gamma(.)$ is the *Gamma* function, $a_{ij} = \Sigma a_{ijk}$, $N_{ij} = \Sigma N_{ijk}$, n is the number of nodes with inputs in the network, $r_i + 1$ is the number of possible discrete states of node x_i, and q is the number of possible states of the parents of node x_i. N_{ijk} is simply the number of samples in the database where node x_i was in state k, when the parents of x_i were in state j. Each expression pattern is represented by a binary one layer network, where $n = 1$ and $r_i = 1$. Because our sequence variables are clearly distinguishable from the expression pattern variables e_i, we use the uninformative prior (Cooper and Herskovits, 1992) $a_{ijk} = 1$, rather than Heckerman's prescription from network equivalence (Heckerman et al., 1995).

For the probability of a given network $P(N)$, Cooper and Herskovits assumed that all networks were equally likely. We prefer to penalize dense networks using the method of Heckerman et al. (1995), which gives asymptotically the same penalty as MDL (Risannen, 1989) and other approaches (Friedman and Goldszmidt, 1998; Friedman and Koller, 2003). Heckerman penalizes each edge of the graph which changes relative to a reference prior network. We choose an empty prior network, so $P(N) \propto K^{-N_p}$, where N_p is the number of parent nodes in the network. A parameter of our learning procedure, $\log(K)$, specifies how strongly complexity is penalized.

We learn network structure iteratively. First, each motif is scored at each depth, and the highest scoring motif is added with the optimal depth. Second, additional constraints are tested on each existing parent: a new depth, distance to ATG in 20 bp increments, orientation, or the presence of a second copy. Third, additional constraints between existing parents are tested: distance between any two motifs in 20 bp increments and their order relative to ATG. Because more complex networks are penalized, the deletion of each current parent constraint is also tested. Then the cycle repeats, and stops when no structural modification improves the score. The algorithm is relatively fast: overnight on eight 2.4 GHz processors of a Linux cluster for *S. cerevisiae*.

Network Complexity and Overfitting

We varied the network complexity penalty parameter $\log(K)$ to maximize the performance on the test dataset. As $\log(K)$ decreases, the networks grow larger, and the average number of parents per network increases. Initially this increase in complexity reflects actual combinatorial regulation, and the performance on the test set increases. But as $\log(K)$ is decreased further, the performance on the training set continues to improve, but the performance on the test set declines slightly. These more complex networks overfit the training data. Optimal performance was observed at $\log(K) = 15$, with an average network size of 2.8 parent nodes (motif constraints).

Linear Regression

An alternative to our approach is to model the expression data as a linear superposition (Bussemaker et al., 2001) where the log ratio expression level of gene i, under condition j, E_{ij}, is given by:

$$E_{ij} = \sum_{k} M_{ik} W_{kj}, \quad (2)$$

where M_{ik} is the number of times motif k occurs in gene i, and W_{kj} is the weight with which motif k contributes to the expression level under condition j. W_{kj} can be positive (for activation) or negative (for repression). The weights W_{kj} are found by least-squares regression on the 80% training set, and performance is evaluated on the 20% test set.

Acknowledgments

We thank the members of the Tavazoie laboratory and Sohail Tavazoie for helpful discussion and review of this work. M.B. is a Lewis Thomas Fellow of Princeton University. S.T. is supported in part by grants from NSF CAREER, DARPA, and DOE.

Received: August 20, 2003
Revised: February 13, 2004
Accepted: February 18, 2004
Published: April 15, 2004

References

An, J.H., and Blackwell, T.K. (2003). SKN-1 links *C. elegans* mesendodermal specification to a conserved oxidative stress response. Genes Dev. *17*, 1882–1893.

Aparicio, O.M., Billington, B.L., and Gottschling, D.E. (1991). Modifiers of position effect are shared between telomeric and silent mating-type loci in S. cerevisiae. Cell *66*, 1279–1287.

Baugh, L.R., Hill, A.A., Slonim, D.K., Brown, E.L., and Hunter, C.P. (2003). Composition and dynamics of the *Caenorhabditis elegans* early embronic transcriptome. Development *130*, 889–900.

Blackwell, T.K., Bowerman, B., Priess, J., and Weintraub, H. (1994). Formation of a monomeric DNA binding domain by Skn-1 bZIP and homeodomain elements. Science *266*, 621–628.

Bowerman, B., Eaton, B., and Priess, J. (1992). skn-1, a maternally expressed gene required to specify the fate of ventral blastomeres in the early *C. elegans* embryo. Cell *68*, 1061–1075.

Breiman, L. (1996). Bagging predictors. Mach. Learn. *24*, 123–140.

Bussemaker, H.J., Li, H., and Siggia, E.D. (2001). Regulatory element detection using correlation with expression. Nat. Genet. *27*, 167–171.

Cohen, B., Mitra, R., Hughes, J., and Church, G. (2000). A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat. Genet. *26*, 183–186.

Cooper, G., and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. Mach. Learn. *9*, 309–347.

Davidson, G.S., Wylie, B.N., and Boyack, K.W. (2001). Cluster stability and the use of noise in interpretation of clustering. Proc IEEE Information Visualization, 23–30.

Davidson, E.H., McClay, D.R., and Hood, L. (2003). Regulatory gene networks and the properties of the developmental process. Proc. Natl. Acad. Sci. USA *100*, 1475–1480.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA *95*, 14863–14868.

Friedman, N., and Goldszmidt, M. (1998). Learning Bayesian networks with local structure. In Learning in Graphical Models, M.I. Jordan, ed. (Cambridge, MA: MIT Press), pp. 421–459.

Friedman, N., and Koller, D. (2003). Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. Mach. Learn. *50*, 95–125.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. J. Comput. Biol. *7*, 601–620.

Gardner, T.S., di Bernardo, D., Lorenz, D., and Collins, J.J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. Science *301*, 102–105.

Gasch, A., Spellman, P., Kao, C., Carmel-Harel, O., Eisen, M., Storz, G., Botstein, D., and Brown, P.O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. Mol. Biol. Cell *11*, 4241–4257.

Gottschling, D.E., Aparicio, O.M., Billington, B.L., and Zakian, V.A. (1990). Position effect at S. cerevisiae telomeres: reversible repression of Pol II transcription. Cell *63*, 751–762.

Hartigan, J.A. (1975). Clustering Algorithms (New York, NY: Wiley).

Heckerman, D. (1998). A tutorial on learning with Bayesian networks.

In Learning in Graphical Models, M.I. Jordan, ed. (Cambridge, MA: MIT Press), pp. 301–354.

Heckerman, D., Geiger, D., and Chickering, M. (1995). Learning Bayesian networks: the combination of knowledge and statistical data. Mach. Learn. *20*, 197–243.

Hill, A.A., Hunter, C.P., Tsung, B.T., Tucker-Kellogg, G., and Brown, E.L. (2000). Genomic analysis of gene expression in *C. elegans*. Science *290*, 809–812.

Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G.M. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J. Mol. Biol. *296*, 1205–1214.

Kellum, R., and Schedl, P. (1991). A position-effect assay for boundaries of higher order chromosomal domains. Cell *64*, 941–950.

Kim, S., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J., Eizinger, A., Wylie, B., and Davidson, G. (2001). A gene expression map for Caenorhabditis elegans. Science *293*, 2087–2092.

Kirkpatrick, S., Gelatt, C.D., and Vecchi, M.P. (1983). Optimization by simulated annealing. Science *220*, 671–680.

Klass, M., Ammons, D., and Ward, S. (1988). Conservation in the 5′ flanking sequences of transcribed members of the *Caenorhabditis elegans* major sperm protein gene family. J. Mol. Biol. *199*, 15–22.

Kurdistani, S.K., Robyr, D., Tavazoie, S., and Grunstein, M. (2002). Genome-wide binding map of the histone deacetylase Rpd3 in yeast. Nat. Genet. *31*, 248–254.

Lascaris, R.F., Planta, W.H., and Mager, R.J. (1999). DNA-binding requirements of the yeast protein Rap1p as selected in silico from ribosomal protein promoters. Bioinformatics *15*, 267–277.

Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wootton, J.C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science *262*, 208–214.

Lee, T.I., Rinaldi, N., Roberts, F., Odom, D., Bar-Joseph, Z., Gerber, G., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., et al. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. Science *298*, 799–804.

Levine, M., and Tjian, R. (2003). Transcription regulation and animal diversity. Nature *424*, 147–151.

L'Hernault, S. (1997). Spermatogenesis. In C. elegans II, D. Riddle, T. Blumenthal, B. Meyer, and J. Priess, eds. (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press), pp. 271–294.

Maduro, M., Meneghini, M., Bowerman, B., Broitman-Maduro, G., and Rothman, J. (2001). Restriction of mesendoderm to a single blastomere by the combined action of SKN-1 and a GSK-3β homolog is mediated by MED-1 and -2 in *C. elegans*. Mol. Cell *7*, 475–485.

Martinez-Pastor, M.T., Marchler, G., Schuller, C., Marchler-Bauer, A., Ruis, H., and Estruch, F. (1996). The Saccharomyces cerevisiae zinc-finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress-response element (STRE). EMBO J. *15*, 2227–2235.

Neuwald, A.F., Liu, J.S., and Lawrence, C.E. (1995). Gibbs motif sampling: detection of bacterial outer membrane protein repeats. Protein Sci. *4*, 1618–1632.

Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems. (San Fransisco, CA: Morgan Kaufmann).

Pilpel, Y., Sudarsanam, P., and Church, G. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. Nat. Genet. *29*, 153–159.

Risannen, J. (1989). Stochastic Complexity in Statistical Inquiry. (River Edge, NJ: World Scientific Publishing Company).

Roberts, S.B., Emmons, S.W., and Childs, G. (1989). Nucleotide sequences of *Caenorhabditis elegans* core histone genes. J. Mol. Biol. *206*, 567–577.

Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. (1998). Finding DNA-regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat. Biotechnol. *16*, 939–945.

Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. *18*, 6097–6100.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat. Genet. *34*, 166–176.

Spellman, P.T., Sherlock, G., Zhang, M., Iver, V., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. (1998). Comprehensive indentification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiase by microarray hybridization. Mol. Biol. Cell *9*, 3273–3297.

Stormo, G., and Fields, D. (1998). Specificity, free energy and information content in protein-DNA interactions. Trends Biochem. Sci. *23*, 109–113.

Stuart, J.M., Segal, E., Koller, D., and Kim, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. Science *302*, 249–255.

Sudarsanam, P., Pilpel, Y., and Church, G.M. (2002). Genome-wide co-occurence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in Saccharomyces cerevisiae. Genome Res. *12*, 1723–1731.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovzky, E., Lander, E., and Golub, T. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. USA *96*, 2907–2912.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., and Church, G.M. (1999). Systematic determination of genetic network architecture. Nat. Genet. *22*, 281–285.

Woudt, L.P., Smit, A.B., Mager, W.H., and Planta, R.J. (1986). Conserved sequence elements upstream of the gene encoding yeast ribosomal protein L25 are involved in transcription activation. EMBO J. *5*, 1037–1040.