© 2001 Nature Publishing Group http://genetics.nature.com

Regulatory element detection using correlation with expression

Harmen J. Bussemaker[1,2], Hao Li[1] & Eric D. Siggia[1]

We present here a new computational method for discovering *cis*-regulatory elements that circumvents the need to cluster genes based on their expression profiles. Based on a model in which upstream motifs contribute additively to the log-expression level of a gene, this method requires a single genome-wide set of expression ratios and the upstream sequence for each gene, and outputs statistically significant motifs. Analysis of publicly available expression data for *Saccharomyces cerevisiae* reveals several new putative regulatory elements, some of which plausibly control the early, transient induction of genes during sporulation. Known motifs generally have high statistical significance.

© 2001 Nature Publishing Group http://genetics.nature.com

Introduction

The availability of complete genome sequences has coincided with the development of technologies to monitor mRNA expression across the genome[1–4]. These expression data provide a global view of transcriptional regulation, but new methods of analysis are needed to extract biologically meaningful information. The DNA sequence elements that act as binding sites for transcription factors coordinate the expression of genes in whose regulatory region they appear, and are key to reducing the complexity of the observed expression patterns. One method for discovering them groups genes into disjoint clusters based on similarity in expression profile over a large number of different conditions[5,6]. The upstream regions of the genes in the cluster can then be analyzed for the presence of shared sequence motifs[7–9]. But the correlation between gene cluster and motifs is imprecise in both directions: there are genes in the cluster without the motif, and many genes with the motif do not respond. If gene control is multifactorial, groups of genes defined by a common motif will not be mutually disjointed, and partitioning the data into disjoint clusters will cause loss of information.

A natural and computationally efficient way to quantify the extent to which regulatory sequence elements can explain changes in genome-wide expression data is to fit the logarithm of the expression ratio to a sum of activating and inhibitory terms, each tied to a particular sequence motif (see Eq. (1) in Methods). Our algorithm selects the most statistically significant motifs from the set of all oligomers up to a specified length, dimers (two oligomers with a fixed spacing) and groups thereof based on sequence alignment. There are no adjustable parameters other than a probability cutoff on statistical significance, and plots of the fitting parameters as a function of time or conditions can suggest biological function. Because all genes are fit, our method is sensitive to the multifactorial nature of transcription control.

Results

We have applied our algorithm to publicly available data sets for yeast from microarray experiments on the diauxic shift[10], sporulation[11] and cell cycle[12,13]. We used a sampling of these data to illustrate four aspects of our algorithm: (i) the iterative selection of motifs to optimize their independence; (ii) the construction of weight matrices from groups of significant dimers; (iii) following the fitting parameters as a function of time to infer function; and (iv) analyzing the modulation of a consensus motif by variable positions and flanking bases (complete results are available at http://regulome.bio.uva.nl/REDUCE/).

Definitions

We define A_g as the logarithm base two of the expression ratio for gene g, and will refer to it as the expression level. For each motif, there is a parameter F, giving the contribution to the expression level for each occurrence in the regulatory region of the gene (see Eq. (1) in Methods). Motifs can be scanned for significance by fitting a single-motif model to the data for all genes, or all members of a group of significant motifs can be simultaneously fit.

The quality of the fit is measured by χ^2, the variance of the difference between the experimental and model values for A_g, normalized so that $\chi^2=1$ in the absence of any fitting parameters. The reduction in χ^2 when a motif is added as a model parameter is denoted by $\Delta\chi^2$ and represents a natural measure of motif significance to which a confidence or P value can be assigned (which we also verified by scrambling the expression data). For the zero time culture of Chu *et al.*[11] compared against itself, or the nonperiodic genes in Spellman *et al.*[13], the A_g are normally distributed with a variance $\mathrm{var}(A_g)\approx 0.06$ (or a 25% fluctuation in expression ratio around unity), which sets a lower bound on the reduction in χ^2.

Finding motifs relevant to cell cycle

Our iterative scheme for selecting motifs is conveniently illustrated by the 14-minute time point data for the α-factor synchronized cultures of Spellman *et al.*[13], located near the M/G1 boundary of the cell cycle. When all oligomers up to 7 nt in length are tested for correlation with expression level, the strongest element found is the repressive AAAATTT with $\Delta\chi^2=3.4\%$, corresponding to a $P\leq 10^{-12}$ (Table 1). The 20 strongest motifs are shown, including several variants of AAAATTT. These motifs are compatible with motif M27 (ref. 14). When the model based on AAAATTT is subtracted from the experimental values, the signal for related motifs also disappears and the residuals are used to re-rank all oligomers up to heptamers. The new top-scoring motif is now ACGCGT, the well-known G1 element MCB (ref. 13), which is positively correlated with expression (that is, F(single) is positive). Continuing the iteration while $P<0.01$ yields a set of 11 motifs (Table 2). In addition to the two motifs already mentioned, the model contains the well-known

[1]*Center for Studies in Physics and Biology, The Rockefeller University, New York, New York, USA.* [2]*Present address: Swammerdam Institute for Life Sciences, Amsterdam Center for Computational Science, University of Amsterdam, Amsterdam, The Netherlands. Correspondence should be addressed to H.J.B. (e-mail: bussemaker@bio.uva.nl).*

© 2001 Nature Publishing Group http://genetics.nature.com

Table 1 • Significant regulatory motifs for cell cycle

Motif	$\Delta\chi^2$	*F*(single)	Matches	ORFs
AAAATTT	0.033534	–0.119555	1564	1331
AAATTTT	0.031324	–0.116968	1516	1291
ACGCGT	0.024535	0.209973	327	289
CGATGAG	0.022773	–0.249775	251	243
GATGAGC	0.019932	–0.275629	186	184
AAATTT	0.019839	–0.060756	3377	2426
AGGGG	0.019754	0.105028	1065	907
GATGAG	0.019091	–0.125441	756	669
AAAATT	0.018169	–0.057483	3663	2611
ACGCG	0.017018	0.103656	939	803
AAGGGG	0.017006	0.154128	476	447
AATTTT	0.015638	–0.051099	3644	2546
AATTTTT	0.014863	–0.075695	1633	1354
CTCATCG	0.014836	–0.206987	241	235
GGGG	0.014271	0.040594	2903	1742
AAAAATT	0.013716	–0.073302	1715	1445
CGCGT	0.013403	0.091417	948	803
TGACGCG	0.013336	0.328217	84	82
GACGCGT	0.012535	0.319782	79	75
CTCATC	0.012318	–0.094246	811	695

Detection of upstream sequence elements relevant to the genome-wide expression pattern for the 14-min time point in the α-synchronized cell-cycle experiment of Spellman *et al.*[13]. Motifs were chosen from the set of all oligonucleotides up to heptamers. Shown are motifs ranked by $\Delta\chi^2$, that is, the relative reduction in the error between the experimental data and a linear model based on a single motif. Only the first 20 motifs are shown out of a set of 21,844. For a definition of the listed quantities see the Methods section; for all motifs shown, $P<10^{-6}$; the number of upstream matches and the number of ORFs in which matches occur is also listed.

Table 2 • Final result of the iterative motif finding procedure

Motif	$\Delta\chi^2$	*F*(single)	*F*(multi)	Matches	ORFs
AAAATTT	0.033534	–0.119555	–0.080316	1564	1331
ACGCGT	0.024535	0.209973	0.211215	327	289
AGGGG	0.019754	0.105028	0.101450	1065	907
CGATGAG	0.022773	–0.249775	–0.200283	251	243
CTCATCG	0.014836	–0.206987	–0.179062	241	235
CCTCGAC	0.008866	0.350493	0.323390	49	48
CCCCT	0.007516	0.061382	0.060757	1146	954
TAAACAA	0.003290	–0.060218	–0.069649	610	565
ATTTTT	0.009661	–0.032125	–0.021880	5260	3167
TGACG	0.008097	0.068384	0.053070	1145	1012
TGAAAA	0.008472	–0.041577	–0.030628	3139	2325

Using a *P* value cutoff of 0.01, a model containing 11 motifs is constructed, using the same expression data as in Table 1.

stress response element (STRE) occurring in both directions (as CCCCT and AGGGG); the motifs CGATGAG and TGAAAA (M3a resp. M3b in ref. 14); CTCATCG, the reverse complement of CGATGAG; the motif TAAACAAA, similar to the SFF motif GTMAACAA (ref. 13); and two seemingly new motifs, CCTCGAC and TGACG. Comparable fits for all other data sets are on our web site (http://regulome.bio.uva.nl/REDUCE/).

We checked that our motifs contribute in an independent and linear way to expression in two ways: (i) the *F* value based on a single motif differs by at most 50% from the value obtained by simultaneously fitting multiple motifs in Table 2; and (ii) the single motif fits for the STRE element AGGGG using separately ORFs with exactly one, two or three copies all gave *F* values within 30% of each other.

Adding words incrementally to the model is efficient: in the first pass there are 192 motifs in which $P<0.01$, but there are only 11 motifs in the final model, which achieves a combined χ^2 reduction of 13.6% when fit to the original data. Based on the values of $\mathrm{var}(A_g)\approx 0.12$ for the data set we analyzed and the experimental noise level of $\mathrm{var}(A_g)\approx 0.06$, we estimate the maximum achievable χ^2 reduction to be 50%. The fit does not improve when motifs are allowed to contain up to two IUPAC symbols.

Table 3 • Dimer alignment for MCM1 binding site

```
.ACC.....AGGA.
.ACC......GGAA
..CCTA...AGGA.
.ACCT...AAGG..
..CCT.....GGAA
..CCTA....GGAA
TACC....AAGG..
.ACCT.....GGA.
.ACCT....AGGA.
TACC......GGA.
TACC.....AGGA.
.ACCT.....GGAA
TACC......GGAA
```

Automatically generated alignment of significant dimers compatible with MCM1 site. Data used are same as in Table 1.

We have also searched for 'dimer' patterns of the form oligo–gap–oligo, in which each oligomer can have a length up to 4 nt, and the gap varies from one to ten. For the same 14-minute data used above, we obtained a cluster of significant dimers (Table 3), which is compatible with the MCM1 site discussed in Spellman *et al.*[13]. When a single *F* is used for all members of the group, $\Delta\chi^2=0.010$ ($P\sim 10^{-5}$). Generalizing to a weight matrix[15] W (see Methods and Table 4), replacing the integer copy number of the motif in Eq. (1) by the information score of the best match of W to the upstream region and using an optimal information score threshold $S_{\min}=9$, we obtained $\Delta\chi^2=0.018$, an 80% improvement over the dimer cluster. After a model fit based on all ten Spellman motifs and our MCM1 matrix is subtracted from the data, a few dozen highly significant dimer groups remain (including those similar to TGAA-4-TTTT, GAT-4-TGA, CTG-5-CCT and AGG-10-AAC).

In the paper by Spellman *et al.*[13], ten motifs were found that are over-represented in one or more expression clusters. We performed single-motif fits for all of these for each time point in the data set of Spellman *et al.*[13] and took the time with the most significant *P* value (which was discounted by the number of time points selected from). For five motifs, ACGCGT (MCB), CRCGAAA (SCB), RRCCAGCR (Swi5p), ACRMSAAA and GTMAACAA (SFF), the correlation is significant, $P\leq 10^{-3}$. Our algorithm identified motifs identical or very similar to all of these (ACGCGT, CGCGAAA, CACGAAA, AACCAGC and GTAAACA). We also verified that fitting our motifs first greatly reduced the $\Delta\chi^2$ associated with the homologous motif of Spellman *et al.*[13] and conversely. The correlation for the remaining five motifs was less significant. Nevertheless, our algorithm identified the motif CCACAGT, which closely matches CCACAK, whose reverse complement is similar to the site AAACTGTGG associated with the Met31/32p cluster. Our algorithm did not find any motifs similar to the 'histone' motif ATGC GAAR, and AGAAGAAA and GCSCRGC.

Tavazoie *et al.*[14] analysed the cell-cycle data of Cho *et al.*[12] and obtained 17 motifs (many nonperiodic), of which we identified nine (MCB, SCB, STRE, Met31/32p, M3a, M3b, M27, M4 and M14a). We failed to identify Cbf1, Rap1p, M5 and ECB, the latter two because their weight matrices have more than one gap, so go beyond our dimer class. The four remaining motifs (M1a, M14b, M13 and M26) show no significant correlation with expression for any time point, and we found no similar motifs.

Time courses for cell cycle and sporulation motifs

We fit each expression pattern separately; thus it is informative to plot the value of *F*(single) for a given motif as a function of time, as illustrated for four cell-cycle motifs: MCB, SCB, MCM1 and SFF (Fig. 1). A small but significant lag between the peaks in the MCB and SCB motifs, previously known by collating the expression of individual genes[13], is visible, as is a new result that the MCM1 and SFF factors are antagonistic (out of phase). Spellman *et al.*[13] gave a combined consensus pattern for these two motifs which should not be construed to imply synergism. The MCM1 and SFF motifs (Fig. 1*b*) occur sufficiently often in the genome that it is meaningful to

© 2001 Nature Publishing Group http://genetics.nature.com

Table 4 • Weight matrix for MCM1 binding site

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
A	111	135	135	121	83	266	0	0	69	148	151	132	167	192	0	0	354	274	191	132	131
C	86	64	61	58	70	31	389	389	48	67	59	80	63	42	0	0	13	30	48	60	59
G	70	73	71	56	70	37	0	0	44	83	75	51	54	62	389	389	10	27	76	76	93
T	122	117	122	154	166	55	0	0	228	91	104	126	105	93	0	0	12	58	74	121	106

Automatically generated weight matrix for the MCM1 binding site obtained by tallying base counts for all matches to the alignment in Table 3. Data used are same as in Table 1.

© 2001 Nature Publishing Group http://genetics.nature.com

investigate nonlinear interactions between them by fitting the model: $A = F_{MCM1}N_{MCM1} + F_{SFF}N_{SFF} + F_{MCM1+SFF}N_{MCM1}N_{SFF}$. The fit does not significantly improve.

We determined the expected time courses for the early (URS1) motif DSGGCGGCND and the middle sporulation motif DNCRCAAAW (MSE) for the data of Chu *et al.*[11] (Fig. 1*c*). The motif CWBYSCTTT, proposed as a potential inducer for mid-sporulation genes by Chu *et al.*[11], gave no signal. By contrast, the strong signal we obtained for the mitotic MCB element, ACGCGT, indicates a role in sporulation that was not detected by clustering genes[11]. We also observe that the induction profiles for the reverse complement of MSE are similar to those for MSE itself, indicating bidirectional binding of the associated factor (Ndt80). We plotted the complete time course for all motifs that are significant at early times, revealing several new early inducers and repressors (Fig. 1*d*). In the former category are GATAAG, which is known to have a role in nitrogen response[9,16], as well as CCACAGT and ATGACT, which seem new and whose reverse complements are also inducers. In Chu *et al.*[11], a 'metabolic' cluster of early and transiently induced genes was identified; 37 (versus 17 expected by chance) of the 49 genes in this cluster have a match to at least 1 of these 5 motifs in their upstream region. We also found that four cell-cycle regulatory motifs, AAATTTT, CGATGAG, TGAAAAA and CTCATC (Table 2), function as early time repressors in sporulation and reach their maximum negative values at 0.5 hours.

Modulation of the MSE motif in sporulation

The MSE element DNCRCAAAW is responsible for the induction of a large number of genes in the middle and late stages of sporulation[11]. Yet only 17% of the 700 genes in which it occurs respond by more than twofold at some point during sporulation. To locate other elements that might discriminate between the active and inactive genes, we fitted just these 700 genes. When motifs up to 8 nt and containing up to 2 IUPAC wild cards are used, 3 significant motifs are found, and $\Delta\chi^2$ is 26%, but they all overlap with either MSE or its reverse complement (Table 5).

To look specifically for flanking bases, we used a site-specific fitting scheme, where B_i stands for the presence of nucleotide B at position i of a wild-card–containing consensus motif when it is matched to the upstream region of a gene. To allow for an investigation of both the variable positions and the flanking regions of the MSE motif, we used the pattern N_5CRCAAAWN$_5$ to 'anchor' variable sites relative to the MSE core. Thus, A_7 corresponds to having an A at the R-position, and F(single) then measures the corresponding change in induction/repression relative to the average effect of having an upstream match to any oligonucleotide compatible with CRCAAAW. Having R=A or W=A leads to an increase in expression level, whereas having R=G or W=T has a suppressive effect (Table 6). The total reduction in variance achieved with this site-specific model fit is 14%, using 8 significant parameters from a total set of 2.2 + 10.4. The implications, however, are similar to the three-motif fit (Table 5) as regards the most significant bases.

Discussion

The approach of clustering genes on the basis of their expression profile across many experiments reduces experimental noise and is well suited for uncovering groups of genes (for example, ribosomal proteins or histones) that co-vary under most circumstances. *Cis*-regulatory elements often then can be found that 'explain' the clusters[5]. Most genes, however, are not part of such a 'synexpression group'[17]. Their regulation is combinatorial and results from

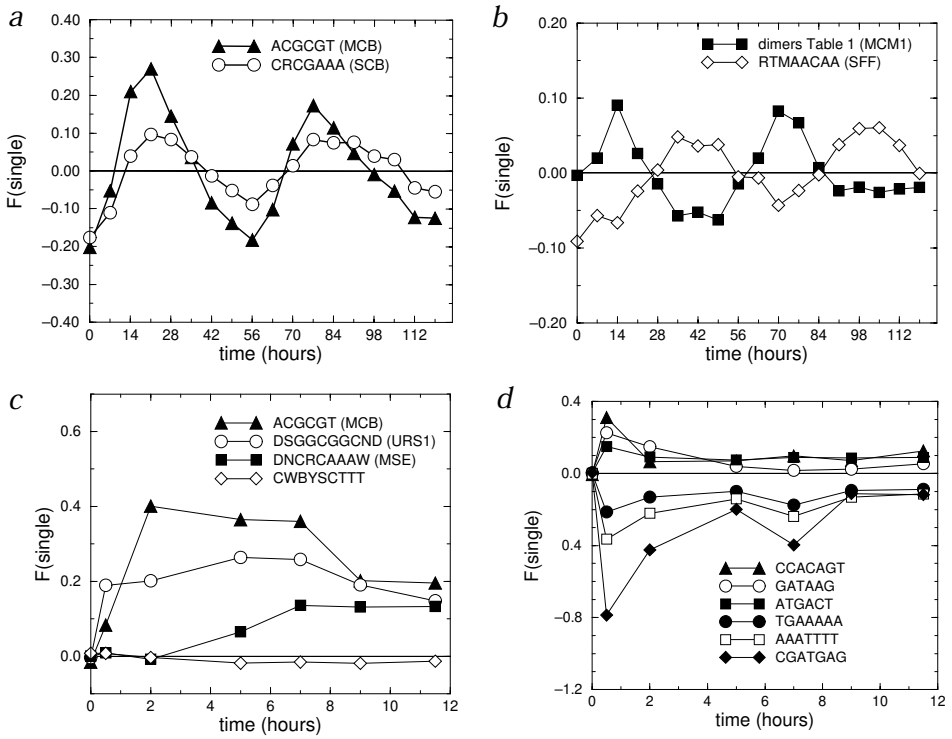

Fig. 1 Time courses for cell cycle and sporulation. Parameter F (Eq. (1)) for a single-motif fit to genome wide expression data, plotted versus time. ***a***, α-factor-synchronized cell-cycle data[13] showing the relative phasing of the MCB and SCB motifs. ***b***, Same for the MCM1 motif as defined in Table 3 and the SFF motif. ***c***, Several motifs discussed in ref. 11 for sporulation: URS1, a regulator of early genes; MSE, the mid-sporulation element; and a putative element CWBYSCTTT, which we find not to be significant within our fit. In addition, we find the known cell-cycle motif ACGCGT to be significant in this context. ***d***, Several new motifs including inhibitory ones for which F(single) is negative.

© 2001 Nature Publishing Group http://genetics.nature.com

Table 5 • Modulation of the MSE consensus motif DNCRCAAAW

Motif	$\Delta\chi^2$	*F*(single)	*F*(multi)	Matches
GHCACAWA	0.1425	1.46	1.36	118
KWTTGTG	0.0694	0.88	0.92	130
ACAAAWTC	0.0559	1.35	1.20	49

Expression data used are from the 11.5-h time point in the sporulation experiment of ref. 11. Outcome of fitting procedure selecting from the set of all motifs of 8 nt or less with at most two IUPAC symbols. Only the 700 genes with an upstream match to DNCRCAAAW are used. An error reduction of 26% is achieved with 3 significant fitting parameters, and C = – 0.14.

Table 6 • Position-specific fit using N_5CRCAAAWN$_5$ as anchor

Base	$\Delta\chi^2$	*P*	*F*(single)	Matches
G_4	0.0604	0.000001	0.85	187
A_7	0.0321	0.000192	0.47	585
G_7	0.0242	0.001269	–0.45	340
G_2	0.0211	0.002889	0.51	182
G_5	0.0192	0.004931	–0.48	188
A_{12}	0.0181	0.006752	0.36	550
C_{14}	0.0111	0.070088	0.41	140
C_4	0.0103	0.094045	–0.36	177
T_{12}	0.0099	0.106903	–0.28	375
T_5	0.0077	0.251208	0.27	267
G_{15}	0.0071	0.320387	0.29	179
C_5	0.0063	0.441155	0.29	160

Positions are counted starting at the left-most position in the anchor, so that G_7 denotes the presence of an G at the R position, etc. Data used are same as in Table 5.

the integration of various signals through the cooperative or competitive binding to multiple sequence elements in the promoter region. There are inherent limitations to clustering methods that do not take the DNA sequence into account. It is also hard to assign a confidence value to the output of a clustering algorithm. On the other hand, expression clustering would link genes controlled by a regulatory cascade even if it involves many different transcription factors and binding sites.

Our approach to the discovery of regulatory elements from expression data is a quantitative expression of the widespread notion[18] that transcription initiation occurs through the recruitment of the polymerase by reversible binding to transcription factors and hence to the regulatory sequences. We fit the logarithm of the expression ratio, a surrogate for the binding free energy, to a sum of contributions from the available motifs. The response is not binary. We find from a large pool of potential motifs those that best correlate with the data. Each motif contributes a fixed increment to the expression, *F*, which can be of either sign, corresponding to enhancement or inhibition. All genes are fit; inactive genes that contain known functional motifs are particularly informative for inhibitory motifs, co-activators for the original site or the influence of position on the activity of a factor.

Using the cell-cycle and sporulation experiments as examples, we reconfirmed almost all motifs found by clustering methods, at least to the extent of finding a related sequence motif that captures the same experimental signal. We have numerous examples that point to combinatorial effects in transcription regulation, or groups of genes that co-vary in one circumstance but vary differently in another, for which expression-based clustering would be poorly suited. For instance, the MCB element is important for mitosis, but also has a role in sporulation, in a way that was not picked up by clustering[11]. The *F* value for MCB is in fact larger than the canonical MSE and URS1 motifs, presumably because the latter also occur in many inactive genes. We found several new factors implicated in early stages of sporulation that together accounted for 16% of the experimental variance, and are candidates for combinatorial control (for example, one is a known nitrogen regulator). The antagonistic effects of MCM1 and SFF are another example, for which clustering analysis only revealed co-occurrence in one gene cluster[13].

We have uncovered another plausible instance of combinatorial control by analysing the absolute mRNA levels in the wild-type reference cultures of Holstege *et al.*[19], for which the logarithm of the mRNA level is close to normally distributed. Our analysis picked up several motifs, among them a group of dimers (and associated weight matrix) corresponding to the binding motif TC-7-ACG for the general factor Abf1p (refs. 20,21), whose correlation with expression had the same significance level as did the MCM1 motif with the cell-cycle data. Matches are found in 1,796 genes and thus occur in combination with a variety of other, more specific factors. Holstege *et al.*[19] searched for regulatory effects associated with components of the general transcription apparatus by comparing wild-type cells with temperature-sensitive mutants, but partial information is implicit already in their reference cultures.

Our linear model, corresponding to independent *cis*-regulatory elements, accounts for at best 30% of the total signal present in genome-wide expression patterns, and there are many directions for improvement. We assumed any motif 600 bp upstream of the translation start site was equally effective in regulating transcription, but location relative to the TATA box should certainly matter. For higher organisms with clusters of regulatory sites far upstream, interspecies comparisons will be essential to filter out the junk DNA and leave a manageable set of loci to be fit to expression[22,23]. Our model is more plausible for transcription rate than transcript abundance, but measurements of mRNA lifetime are possible[19] and allow one to convert between the two types of data. Finally, we stress the importance of simultaneously fitting as many genes as possible. Non-responding genes can be as informative as active ones regarding *cis*-regulatory elements.

Methods

DNA sequence motifs. Motifs were chosen from the set of all oligomers up to a given length (given the paucity of many octamers in our data set, we stopped at 7 nt). In addition, dimers consisting of a pair of oligos with a fixed spacing were used. Optionally, oligos were improved by iteratively adding one degenerate IUPAC symbol at a time in all possible ways to the top 100 motifs. A sequence similarity score was computed between all pairs of statistically significant dimers that were then grouped using the 'cast' routine[24]. A weight matrix was computed by aligning the actual DNA sequence elements that matched any member of the cluster (with no double counting) and then counting the bases at each position, adding a pseudocount of one. For the fit of the MCM1 weight matrix in Table 4, only sequence matches with an information score greater than a given threshold were used.

We count only those motifs that occur in the 600 bp upstream of the translation start site for each ORF, because most of the known transcription factor binding sites fall in that range[13]. We shorten the upstream region to eliminate any overlap with a coding region on either strand. The chromosome sequence and ORF coordinates for *Saccharomyces cerevisiae* were obtained from the Saccharomyces Genome Database[1].

Definition of the model. The model we use to fit the expression data assumes additivity of the contributions from different regulatory factors and is defined as:

$$A_g = C + \sum_{\mu \in M} F_\mu N_{\mu g} \qquad (1)$$

Here A_g is the logarithm base two of the ratio of mRNA abundances between two cell populations for gene *g*. The integer $N_{\mu g}$ equals the number of occurrences of motif μ in the regulatory region, and M denotes the set of significant motifs. For weight matrices, $N_{\mu g}$ is the information score of the best match of the matrix to the upstream region provided it exceeds a threshold based on the score of the matrix against itself and its variance. The model parameters C and $\{F_\mu\}$ are the same for each gene: C represents a baseline expression level

© 2001 Nature Publishing Group http://genetics.nature.com

© 2001 Nature Publishing Group http://genetics.nature.com

when no motifs are present in the upstream sequence, whereas F_μ is the increase/decrease in expression level caused by the presence of motif μ. The sign of F_μ determines whether the putative protein factor that binds to sequence element μ acts as an activator or as an inhibitor.

Fitting to expression data. It is convenient to transform both the logarithm of the expression ratio, A_g, and the number of occurrences of motif μ in gene g, $N_{\mu g}$, in Eq. (1) by subtracting their mean and applying a rescaling. To this end define $a_g = \delta A_g/(G\langle\delta A^2\rangle)^{1/2}$, where G denotes the total number of genes, $\langle X\rangle \equiv (1/G)\Sigma_g X_g$ defines an average of quantity X over all genes, $\delta A_g = A_g - \langle A\rangle$ is the deviation from the mean, and $\langle\delta A^2\rangle = var(A_g)$ the variance of A_g. Similarly we define $n_{\mu g} = \delta N_{\mu g}/(G\langle\delta N_\mu^2\rangle)^{1/2}$, with $\delta N_{\mu g} = N_{\mu g} - \langle\delta N_\mu\rangle$ for each motif μ. It is helpful to think of a_g and of $n_{\mu g}$ as vectors $\mathbf{a}$ and $\mathbf{n}_\mu$ in the G-dimensional space of all genes, and define the dot product $\mathbf{a}\cdot\mathbf{b} = \Sigma_g a_g b_g$ and norm $|\mathbf{a}| = (\mathbf{a}\cdot\mathbf{a})^{1/2}$. It follows from their definition that $\mathbf{a}$ and $\mathbf{n}_\mu$ have unit length and are perpendicular to the vector $\mathbf{1} = (1, 1, \ldots, 1)$. With this notation Eq. (1) takes the following form:

$$\boldsymbol{a} = \sum_{\mu\in M} f_\mu \boldsymbol{n}_\mu \tag{2}$$

By varying f_μ, we minimize the error between the model and the experimental data,

$$\chi^2 = \Big|\mathbf{a} - \sum_{\mu\in M} f_\mu \mathbf{n}_\mu\Big|^2 \tag{3}$$

It is straightforward to show that the optimal solution is obtained by solving f_μ from the linear equation

$$\sum_{\mu'} (\mathbf{n}_\mu \bullet \mathbf{n}_{\mu'}) f_{\mu'} = \mathbf{a}\bullet\mathbf{n}_\mu \tag{4}$$

Note that the dimensionality of this linear equation—equal to the number of fitting parameters—is very different from that of the space in which the vectors $\mathbf{a}$ and $\mathbf{n}_\mu$ reside. The corresponding solution to Eq. (1) can be recovered as follows: $F_\mu = f_\mu(\langle\delta A^2\rangle/\langle\delta N_\mu^2\rangle)^{1/2}$ and $C = \langle A\rangle - \Sigma_{\mu\in M} f_\mu\langle N_\mu\rangle$.

Iterative procedure for finding significant motifs. When constructing a model that achieves a significant reduction of χ^2, the number of fitting parameters can be kept to a minimum by adding motifs to the set M one at a time. When the model is based on a single motif, we have $f_\mu = \mathbf{a}\cdot\mathbf{n}_\mu$, and the error is given by:

$$\chi^2 = 1 - (\mathbf{a}\bullet\mathbf{n}_\mu)^2 \equiv 1 - \Delta\chi^2_\mu \tag{5}$$

One can thus rank all possible motifs by the reduction in the error, $\Delta\chi^2_\mu = (\mathbf{a}\bullet\mathbf{n})^2$, and select the largest one. We proceed inductively: after fitting a set of parameters M compute the residual $\mathbf{a}' = \mathbf{a} - \Sigma_\mu f_\mu\mathbf{n}_\mu$, rank all motifs by $\Delta'\chi^2_\mu = (\mathbf{a}'\cdot\mathbf{n}_\mu)^2$, and again select the μ giving the largest reduction in variance. Note that Eq. (4) is equivalent to the statement that $\mathbf{a}'$ is orthogonal to the space spanned by the $\mathbf{n}_\mu$ in the set M.

Statistical significance measure. When the expression levels A_g are random variables drawn from a normal distribution, the correlation between A_g and $N_{\mu g}$, which can be written as $\mathbf{a}\cdot\mathbf{n}_\mu$, is a random variable with zero mean and standard deviation $G^{-1/2}$. When the significance of a specific motif is considered, it is therefore convenient to define a unit-variance Z-score as:

$$Z_\mu = (G)^{1/2}(\mathbf{a}\bullet\mathbf{n}_\mu) \tag{6}$$

Note that $\Delta\chi^2_\mu = (Z_\mu)^2/G$. When the motif that gives the largest error reduction $\Delta\chi^2_\mu$ is selected from a set of M possible motifs, the significance of $\Delta\chi^2_\mu$ is given by the extreme value distribution, describing the probability that the largest (in absolute value) of M samples from a normal distribution, $\text{prob}(Z) = (2\pi)^{-1/2}\exp(-Z^2/2)$, equals $|Z|_{\max} = (G\,\Delta\chi^2)^{1/2}$.

$$P = 1 - \left[\frac{2}{(2\pi)^{1/2}}\int_0^{|Z|\max} dZ\, e^{-Z^2/2}\right]^M \tag{7}$$

When using oligonucleotides up to length l, we have $M = (4/3)(4^l - 1)$. We verified our error estimates (for example, the assumption that the residuals are independent gaussians with a common variance) by randomizing the association between expression and gene. When a site-specific fitting scheme is used where $\mu = B_i$ stands for the presence of nucleotide B at position i of a wild-card–containing consensus motif, M equals the total number of possible parameters B_i, so that each N in the consensus motif will add four, an R or W will add two, and A, C, G or T will not add to M at all.

Acknowledgments

We thank B. Shraiman for suggesting linear multivariate fits to expression data, and L. Grivell, R. Lascaris and H. de Nobel for discussions and critical reading of the manuscript. Support was received from the NSF under grant number DMR 9732083 and from the Keck foundation to H.L.

Received 23 February 2000; accepted 3 January 2001.

1. Cherry, J.M. *et al.* Genetic and physical maps of *Saccharomyces cerevisiae*. *Nature* **387**, 67–73 (1997).
2. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
3. Lockhart, D.J. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.* **14**, 1675–1680 (1996).
4. Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
5. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
6. Roth, F.R., Hughes, J.D., Estep, P.W. & Church, G.M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnol.* **16**, 939–945 (1998).
7. Lawrence, C.E. *et al.* Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208–214 (1993).
8. Neuwald, A.F., Liu, J.S. & Lawrence, C.E. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.* **4**, 1618–1632 (1995).
9. Van Helden, J., Andre, B. & Collado-Vides, J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* **281**, 827–842 (1998).
10. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **287**, 680–686 (1997).
11. Chu, S. *et al.* The transcriptional program of sporulation in budding yeast. *Science* **282**, 699–705 (1998).
12. Cho, R.J. *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2**, 65–73 (1998).
13. Spellman, P.T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell.* **9**, 3273–3297 (1998).
14. Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. & Church, G.M. Systematic determination of genetic network architecture. *Nature Genet.* **22**, 281–285 (1999).
15. Berg, O.G. & Von Hippel, P.H. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* **193**, 723–750 (1987).
16. Magasanik, B. Regulation of nitrogen utilisation. in *The Molecular and Cellular Biology of the Yeast Saccharomyces: Gene Expression* (eds. Jones, E.W., Pringle, J.R. & Broach, J.R.) 283–318 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1992).
17. Niehrs, C. & Pollet, N. Synexpression groups in eukaryotes. *Nature* **402**, 483–487 (1999).
18. Ptashne, M. & Gann, A. Imposing specificity by localization: mechanism and evolvability. *Curr. Biol.* **8**, R897 (1998).
19. Holstege, F.C.P. *et al.* Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717–728 (1998).
20. Halfter, H., Kavety, B., Vandekerckhove, J., Kiefer, F. & Gallwitz, D. Sequence, expression and mutational analysis of BAF1, a transcriptional activator and ARS1-binding protein of the yeast *Saccharomyces cerevisiae*. *EMBO J.* **8**, 4265–4272 (1989).
21. Della Seta, F. *et al.* The ABF1 factor is the transcriptional activator of the L2 ribosomal 15 protein genes in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **10**, 2437–2441 (1990).
22. Loots, G.G. *et al.* Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**, 136–140 (2000).
23. Hardison, R.C., Oeltjen, J. & Miller, W. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.* **7**, 959–966 (1997).
24. Ben-Dor, A., Shamir, R. & Yakhini, Z. Clustering gene expression patterns. *J. Comput. Biol.* **6**, 281–297 (1999).