Integrating regulatory motif discovery and genome-wide expression analysis

Erin M. Conlon*†, X. Shirley Liu†‡, Jason D. Lieb§, and Jun S. Liu*¶

*Department of Statistics, Harvard University, Cambridge, MA 02138; ‡Department of Biostatistics, Harvard School of Public Health, Dana Farber Cancer Institute, Boston, MA 02115; and §Department of Biology and Carolina Center for the Genome Sciences, University of North Carolina, Chapel Hill, NC 27599

Communicated by Richard M. Losick, Harvard University, Cambridge, MA, January 30, 2003 (received for review December 11, 2002)

We propose MOTIF REGRESSOR for discovering sequence motifs upstream of genes that undergo expression changes in a given condition. The method combines the advantages of matrix-based motif finding and oligomer motif-expression regression analysis, resulting in high sensitivity and specificity. MOTIF REGRESSOR is particularly effective in discovering expression-mediating motifs of medium to long width with multiple degenerate positions. When applied to *Saccharomyces cerevisiae*, MOTIF REGRESSOR identified the ROX1 and YAP1 motifs from Rox1p and Yap1p overexpression experiments, respectively; predicted that Gcn4p may have increased activity in *YAP1* deletion mutants; reported a group of motifs (including GCN4, PHO4, MET4, STRE, USR1, RAP1, M3A, and M3B) that may mediate the transcriptional response to amino acid starvation; and found all of the known cell-cycle regulation motifs from 18 expression microarrays over two cell cycles.

sequence motif discovery | microarray data | correlation | transcription regulation

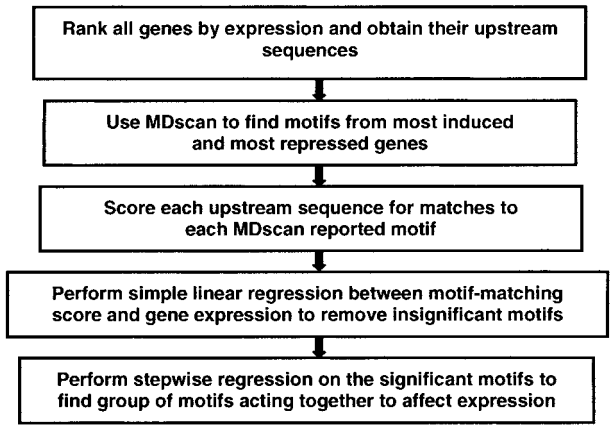

Fig. 1. MOTIF REGRESSOR strategy diagram.

Direct experimental determination of transcription factor DNA-binding motifs (TFBM) is not practical or efficient in many biological systems. Therefore, computational algorithms such as the word-enumeration (1–4), the position-specific matrix update (5–7), and the dictionary (8) methods have been developed to identify putative motifs and guide experimentation. One of the most successful computational tactics for TFBM discovery is to cluster genes based on their expression profiles, and then search for motifs in the sequences upstream of tightly clustered genes (9). When noise is introduced into the cluster through spurious correlations, however, such an approach may result in false positives. A filtering method (10) based on the specificity of the motif occurrences has been shown to effectively eliminate false positives, but the sensitivity of the algorithm is still low in some cases. An iterative procedure for simultaneous clustering and motif finding has been suggested (11), but no effective algorithm has been implemented to demonstrate its advantage in biological data. Two novel methods for TFBM discovery via the association of gene expression values with oligomer motif abundances have been proposed (12, 13). They first conduct word enumeration and then use regression to check whether the genes whose upstream sequences contain a set of words have significant changes in their expression. These methods are effective for discovering conserved short motifs and sometimes interactions among them, but are not effective with longer motifs and may lose sensitivity in cases where TFBMs have multiple degenerate positions.‖

We present an alternative approach operating under the explicit assumption that, in response to a given biological condition, the effect of a TFBM is strongest among genes with the most dramatic increase or decrease in mRNA expression. We first use a fast and sensitive motif-finding method, MDSCAN (14), to generate a large set of motif candidates that are enriched in the DNA sequence upstream of genes with the highest fold change in mRNA level relative to a control condition. Then we verify each candidate motif by associating every gene's upstream sequence motif-matching score with its downstream expression measure (Fig. 1). In the motif-matching score, both the number of sites and the strength of matching are incorporated by using a third-order Markov background model and a position-specific motif matrix, thereby increasing both the sensitivity of the method and the specificity of the reported motifs.

Methods

Microarray and Sequence Data. *Saccharomyces cerevisiae* microarray experiments in our study include Rox1p (15) and Yap1p (16) overexpression, *YAP1* deletion (17), amino acid starvation response (16), α-factor synchronized cell-cycle progression (18), and chromatin immunoprecipitation (ChIP)-array experiments on Gcn4p (19), Rpd3p, Ume1p, and Ume6p (20). Yeast ORFs were ranked according to their relative change in mRNA level under each condition. We extracted up to 800 bp upstream of each ORF that was nonoverlapping with the adjacent ORF. Single repeats (e.g., AAAA. . .) of >10 bases and double repeats (e.g., ACACAC. . .) of >16 bases were removed.

Motif Discovery by MDSCAN. Using every occurring w-mer ($5 \leq w \leq 15$ in our studies) in the top k (default as 100) sequences as a seed, MDSCAN (14) finds all w-mers in the top sequences that are similar to the seed and constructs from them a motif matrix. A pair of w-mers is said to be "similar" if they share at least m matched positions, where the probability for two random w-mers to share $\geq m$ matched positions is ε (default as 0.0015). All of the motif matrices are evaluated by a semi-Bayesian scoring func-

Abbreviations: TFBM, transcription factor DNA-binding motif; ChIP, chromatin immunoprecipitation.

†E.M.C. and X.S.L. contributed equally to this work.

¶To whom correspondence should be addressed at: Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138. E-mail: jliu@stat.harvard.edu.

‖Throughout the article, we use italicized, all capital designations (*YAP1*) for genes, first-letter capitalized designations ending with "p" (Yap1p) to represent proteins, and all capital, nonitalic designations (YAP1) to represent DNA binding motifs and sites bound by the respective protein.

GENETICS

Table 1. Motifs discovered from Rox1p and Yap1p overexpression experiments

Method	Rox1p overexpression	Yap1p overexpression
ALIGNACE	AAAAAAAAAAAAAAAAAAAAAAA	AAGAAGAAAA
Best 10 motifs (10 input sequences)	AAAAAAAAAAG	GAAGAGAAGAA
	AAGGAAAAAAAGAAAAAAAAA	AAAGAAGAAA
	AAAAAAAAGAAAAGAAAAAAA	CA**TTTCTAA**TCT
	AAGGAAAAAAGAAA	GAAAAGCG
	AAGAAAAAAA	AAGAGGAG
	GCGCCCCGGA	AAAAAAAGAAG
	GAGCGCTCATGCCGCTGTTTT	GAAAAAAAAG
	AAAATAAAAAAAAAAAA	AAAAAAAGAAA
	CTGCGGAAAA	AAAATGAAAAATG
MEME	TTTTTTTCTTTT	CA**TTACTAA**TCA
All motifs (10 input sequences)	TTCCGCGGA	AAAGAGGTG
MDSCAN	**TTGTT** (w = 5)	**TTACT** (w = 5)
Best motif at each width	**TATTGT** (w = 6)	GAAGAA (w = 6)
(10 input sequences)	**CTATTGT** (w = 7)	**TTACTAA** (w = 7)
	CTATTGTT (w = 8)	**TTACTAA**T (w = 8)
	A**TCTATTGT** (w = 9)	GA**TTACTAA** (w = 9)
	TATGTACGTA (w = 10)	GA**TTACTAA**T (w = 10)
	GTACGTATGTA (w = 11)	GA**TTACTAA**TC (w = 11)
	TCTCTTGCCTTT (w = 12)	TCA**TTACTAA**GC (w = 12)
	AGGACAAAAGGAA (w = 13)	GA**TTACTAA**TCAC (w = 13)
	TATATACATATATA (w = 14)	ATTATTAATCAAAT (w = 14)
	ATATATACGTATATA (w = 15)	GA**TTACTAA**TCACAT (w = 15)
MDSCAN	GTTTG (w = 5)	AAGGG (w = 5)
Best motif at each width	**CTATTG** (w = 6)	CCGCGG (w = 6)
(100 input sequences)	**TATTGTT** (w = 7)	AAGGGGA (w = 7)
	CTATTGTT (w = 8)	G**TTAC**CCG (w = 8)
	GCGATGAGC (w = 9)	CACACACCA (w = 9)
	GCGATGAGCT (w = 10)	C**TTACTAA**TCCA (w = 10)
	TGCGATGAGCT (w = 11)	ACACATATATATA (w = 11)
	ATGCGATGAGCT (w = 12)	CACACACACACACA (w = 12)
	GCGATGAGCTGAG (w = 13)	ATATATATATATATA (w = 13)
	ACACACACACACAC (w = 14)	CACACACACACACACAC (w = 14)
	CACACACACACACAC (w = 15)	TATATATATATATATATA (w = 15)
MOTIF REGRESSOR	**TCTATTGTT** (0)	ATGA**TTACTAA**TCA (5.58e-11)
Top 10 motifs (regression P value)	TT**TCTATTGT** (0)	TGC**TTACTAA**TC (1.39e-10)
(100 input sequences)	**CTATTGTTT**TC (0)	GA**TTACTAA**TC (3.30e-10)
	ACT**TCTATTGT** (0)	GTGA**TTACTAA**TC (3.26e-9)
	TT**TCTATTGTTT**T (0)	C**TTACTAA**TC (1.48e-7)
	TT**TCTATTGTTT**TT (0)	**TTACTAA**TC (4.23e-6)
	CTATTGTT (1.11e-16)	TCCTGCCCCTT (5.55e-6)
	ATTGT (1.20e-14)	TCCCA**TTAC** (4.80e-5)
	GGTGGC (1.38e-11)	GCGCCCC**TTAC** (1.68e-4)
	TATTGTT (1.04e-10)	TCCCTCTCCCTT (1.92e-4)

Motifs reported by ALIGNACE, MEME, MDSCAN, and MOTIF REGRESSOR for two microarray experiments. Both ALIGNACE and MEME used only the top 10 genes, and incorporating more genes led to worse results. MDSCAN ranked the correct motifs as the best in several widths when the top 10 sequences were used, but failed when 100 sequences were used. MOTIF REGRESSOR ranked the correct motifs as number 1 for all settings. The most difficult case, using the top 100 genes to find motifs and the top 500 genes to refine them, is reported. Boldface indicates agreement between a discovered motif and the known motif consensus.

tion, and the 50 highest-scoring motifs are saved. Each retained motif matrix is refined by adding or removing w-mers in the top K (default as 500, this includes the original k) sequences to increase the motif score. Motifs with average frequency of the consensus bases <0.7 are eliminated. If several motifs of the same width share similar consensus, the one with the highest score is retained. Here, two consensuses are said to be "similar" if they share $\geq m$ matched positions, allowing frame shifts and considering both forward and reverse-complements. MDSCAN reports up to 30 distinct motifs.

Sequence Motif-Matching Scoring. We determine how well the upstream sequence of a gene g matches a motif m, in terms of both degree of matching and number of sites, by the following function:

$$S_{mg} = \log_2\left[\sum_{x \in X_{wg}} \Pr(x \text{ from } \theta_m)/\Pr(x \text{ from } \theta_0)\right] \quad [1]$$

where θ_m is the probability matrix of motif m of width w, θ_0 is the third-order Markov model estimated from all of the yeast

intergenic sequences, and X_{wg} is the set of all w-mers in the upstream sequence of gene g. Using a motif matrix and a background with Markov dependency greatly improves the sensitivity and specificity of the scoring function (14).

Linear Regression. For each motif reported by MDSCAN, we first fit the simple linear regression:

$$Y_g = \alpha + \beta_m S_{mg} + \varepsilon_g \quad [2]$$

where Y_g is the $\log_2$-expression value of gene g, S_{mg} is defined in (1), and ε_g is the gene-specific error term. The baseline expression α and the regression coefficient β_m will be estimated from the data. A significantly positive or negative β_m indicates that upstream sequences containing motif m are correlated with enhanced or inhibited downstream gene expression, respectively.

The candidate motifs with a significant P value ($P \leq 0.01$) for the simple linear regression coefficient β_m are retained and used by the stepwise regression procedure to fit a multiple regression model:

$$Y_g = \alpha + \sum_{m=1}^{M} \beta_m S_{mg} + \varepsilon_g. \quad [3]$$

Stepwise regression begins with Model (3) with only the intercept term, and adds at each step the motif that gives the largest reduction in residual error. After adding each new motif m_i, the model is checked to remove the ones whose effects have been sufficiently explained by m_i. The final model is reached when no motif can be added with a statistically significant coefficient.

Results

The Discovery of Single Motifs That Influence Gene Expression. To test the validity of MOTIF REGRESSOR, we sought to identify the TFBMs of Rox1p and Yap1p by examining the upstream regions of genes with highest fold change in expression upon the overexpression of Rox1p (15) and Yap1p (16) (Table 1). Rox1p is a heme-induced TF that recognizes YYNATTGTTY (21) and represses genes normally expressed in hypoxic conditions. We used MDSCAN to search the upstream sequences of the 10, 25, 50, and 100 most repressed genes ranked by fold change, respectively, to generate up to 30 candidate motifs for each width from 5 to 15 bases. We used five times the number of input sequences for refinement. The correct motif, as defined by previous studies (21), was the top-ranked motif discovered for input sequence sizes 10, 25, and 50, but was not top-ranked when using 100 sequences. In contrast, the top-ranked motifs found by MOTIF REGRESSOR for all input sequence sizes (10, 25, 50, 100) had consensuses that matched the known ROX1-binding consensus. For all input sequence sizes, at least 8 of the top 10 motifs had a motif consensus that matched perfectly with the known ROX1 consensus, all with very low ($<10^{-9}$) regression P values (unadjusted for multiple testing). These results (Table 1) were then compared with those obtained by other motif-finding algorithms. We applied ALIGNACE (9) to search the upstream sequences of 10, 25, 50, and 100 most repressed genes ranked by fold change, respectively, specifying the correct motif width and using all other default parameters. We failed to find any motifs resembling the known ROX1-binding consensus for any input sequence size. Similarly, we ran MEME (6) for input of 10, 25, 50, and 100 sequences, respectively, allowing motif width to vary between 5 and 15 and each sequence to contain 0-n repetitions of a motif. Again, MEME did not report any motifs with consensus similar to the known ROX1-binding consensus.

Yap1p is a transcriptional activator required for oxidative stress tolerance, and is known to recognize the DNA sequence TTACTAA (22). Genes induced by Yap1p overexpression presented another interesting case for comparing motif-finding methods. ALIGNACE reported a motif ranked fourth with a consensus that differs from the reported YAP1 consensus (22) by one base for input sequence sizes 10 and 25, but did not find any motifs resembling the YAP1 motif for input sequence sizes 50 and 100. MEME ranked the correct YAP1 motif first for input sequence size 10, and third for input sequence sizes 25 and 50, but failed to find the YAP1 motif for input sequence size 100 (Table 1). MDSCAN outperformed both ALIGNACE and MEME, finding and ranking the correct motif first over a range of motif widths when using 10, 25, and 50 input sequences, respectively. With 100 input sequences, MDSCAN found and ranked the correct motif first at motif width $w = 10$. MOTIF REGRESSOR also performed strongly, with at least the top 6 of the top 10 motifs containing motif consensuses that matched the known YAP1 consensus for all input sequence sizes.

We also analyzed the expression data from *YAP1* deletion mutants (17), expecting to find the YAP1 motif among the sequences upstream of down-regulated genes. To our surprise, MOTIF REGRESSOR found only one significant motif ($P = 0.0075$) partially matching the YAP1 motif, with a consensus CGTTACCCTCC. Using the sequences upstream of induced genes, how-

Motif #,group	Motif Sequence Logo	Known Motif	Motif coeff-icient	Motif p-value
1,1		MET4	0.107	8.3e-13
2,2		PHO4	0.088	1.2e-8
3,3		M3A	-0.09	2.0e-7
11,3			-0.77	3.9e-4
20,3			0.063	6.2e-3
4,4			-0.08	4.4e-6
5,5		RAP1	-0.1	5.8e-6
22,5			-0.06	6.6e-3
6,6		STRE	0.084	7.9e-5
13,6			0.063	9.5e-4
18,6			0.06	3.5e-3
7,7			0.064	8.0e-5
8,8			0.072	9.2e-5
21,8			0.046	6.4e-3
9,9			0.068	9.6e-5
19,9			0.051	4.9e-3
10,10			0.057	3.7e-4
12,11			0.045	4.8e-4
14,12		GCN4	0.059	1.1e-3
15,12			0.056	1.2e-3
23,12			0.05	6.9e-3
16,13			0.056	1.3e-3
17,14		URS1	0.057	1.4e-3
24,15		M3B	-0.08	8.5e-3
25,15			0.081	9.7e-3

Fig. 2. Motifs discovered from the amino acid starvation microarray experiment. The MOTIF REGRESSOR multiple regression model reported a total 25 significant motifs active in amino acid starvation response. The 25 motifs can be organized into 15 groups, 8 of which represent previously known TF motifs. The identity of the known motifs reflects how the cell responds to amino acid starvation: slowing cell growth (M3A, M3B, and RAP1), responding to general environmental stress (STRE and URS1), and initiating phosphate (PHO4), sulfur (MET4), and amino acid (GCN4) biogenesis and metabolism. Motif matrices are represented as sequence logos (23).

GENETICS

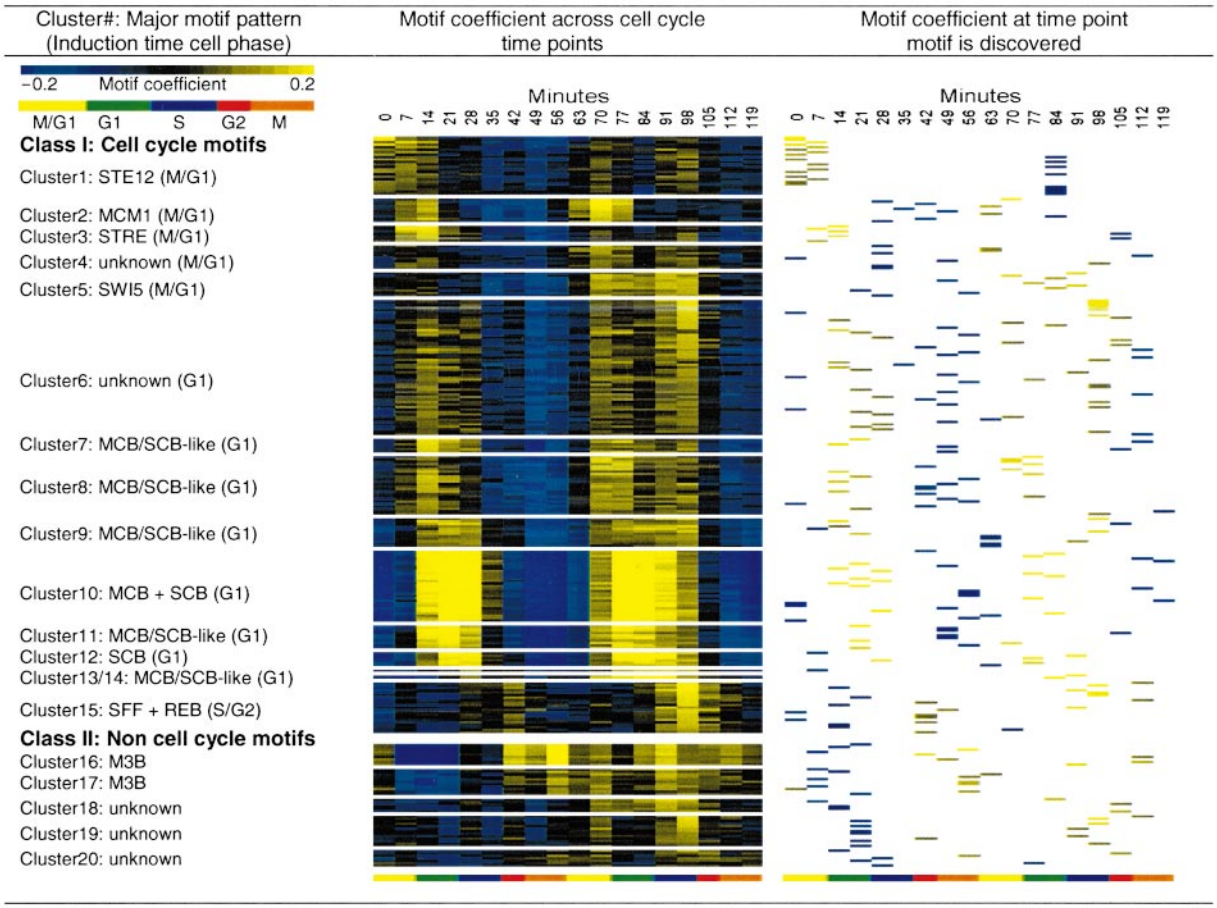

Fig. 3. Motif clusters from cell cycle expression time series experiments. The 273 significant motifs reported by MOTIF REGRESSOR over two cell cycles are clustered by motif coefficients over the 18 time points. Motif coefficients can be interpreted as the putative influence a particular motif has on the expression of downstream genes. The 20 resulting clusters include the known cell cycle-related TF motifs MCB, SCB, SFF, MCM1, and SWI5. Other motif clusters also have coefficients that fluctuate with the cell cycle, such as STE12, STRE, groups of motifs that resemble MCB and SCB, and some novel G_1 motifs. Five motif clusters have coefficients that do not fluctuate with the cell cycle, including M3B and some motifs of unknown function. The clusters were ordered by first appearance of their cell-cycle influence. This figure was produced using TREEVIEW software (30).

ever, MOTIF REGRESSOR found many significant motifs matching the GCN4 consensus TGACTCA (21). It has been previously shown that Gcn4p can bind efficiently to YAP1 sites TTACTAA (22), so it is possible that the activity of Gcn4p is increased in *YAP1* mutants to partially compensate for the lack of Yap1p. Increased Gcn4p activity would then lead to the induced expression of many Gcn4p targets, and our discovery of the GCN4 motif. Among the 135 Gcn4p intergenic targets identified by Gcn4p ChIP-array experiments (19), 95 showed increased downstream expression in *YAP1* mutants ($P < 10^{-6}$). Furthermore, among the 118 up-regulated genes with >1.5-fold change in *YAP1* mutants, 20 were identified as Gcn4p target downstream genes by genome-wide ChIP assays (19) ($P < 10^{-12}$).

The Discovery of Multiple Motifs That Influence Gene Expression. MOTIF REGRESSOR was applied to the sequences upstream of yeast genes whose expression changed after 30 min of amino acid starvation (16). MDSCAN found 414 motifs of width 5–15 from the most induced genes and most repressed genes, respectively. The simple linear regression step screened out 179 insignificant motifs ($P > 0.01$). The stepwise regression on the remaining 235 motifs yielded 25 that were significant. The resulting model had an R-square of 19.8%, implying that together the 25 motifs might account for 19.8% of the variation in genomic expression. Stepwise regression overestimates the true R-square due to its selective use of explanatory variables (www.stata.com/support/faqs/stat/stepwise.html). However, we performed a simulation study by using 6,000 genes and 400 independent motifs, which found the overestimation of R-square under these conditions to be small, ≈1%.

The 25 motifs can be classified into 15 different DNA patterns (Fig. 2). Eight of these patterns (STRE, GCN4, M3A, M3B, MET4, PHO4, RAP1, and URS1) were previously known, and together they have an R-square of 17.6%. The stress response element STRE and the GCN4 motif, which regulates amino acid biosynthesis, are known to positively regulate transcription during amino acid starvation. Two patterns involved in RNA processing (M3A and M3B; ref. 24) have been previously found in genes repressed under environmental stress (web supplement to ref. 16; http://genome-www.stanford.edu/yeast_stress/images/sfigure0.html). Met4p, with its auxiliary factors Cbf1p, Met28p, Met31p, and Met32p, regulates the transcriptional activation of genes involved in sulfur metabolism, especially the sulfur amino acid pathway (25). Pho4p, together with Pho2p, are the master transcription regulators of the *PHO* genes responsible for the scavenging and uptake of inorganic phosphate under phosphate-limiting conditions (26). Rap1p is the primary regulator of the yeast ribosomal protein genes (RPGs), and the RAP1 motif occurs in most of the RPG promoters (27). Rap1p is required for the heavy transcription of RPGs under favorable growth conditions, and for

the repression of these genes under unfavorable conditions (28). Finally, the URS1 site is known to be present in the promoters of many yeast genes induced under stress conditions (29). From the eight discovered known motifs, we gain insight to the cell's amino acid starvation response: slowing cell growth (M3A, M3B, and RAP1), responding to general environmental stress (STRE and URS1), initiating nutrient scavenging and production as evidenced by the phosphate and sulfur regulatory pathways (PHO4 and MET4), and promoting amino acid biogenesis (GCN4). Although the remaining seven unknown patterns (9 discovered motifs) accounted for only 2.2% additional variation, the two most significant motifs from simple linear regression (motif 19 and 10) still have simple regression *R*-square of 5% and 1.33%, respectively, suggesting their biological relevance to the amino acid starvation response.

The M3B motif has been discovered by many computational algorithms (4, 10, 12), and it was recently suggested to be bound by the histone deacetylase and repressor Rpd3p and its associated proteins Ume1p and Ume6p (20). We applied MOTIF REGRESSOR to targets separately selected by ChIP-array experiments on Rpd3p, Ume1p, and Ume6p, respectively (20). MOTIF REGRESSOR found URS1 to be the most significant motif for Ume6p, RAP1 for Ume1p, and M3B for Rpd3p. Because targets of Rpd3p and Ume1p overlap by >50%, M3B and RAP1 motifs are found significant for both Ume1p and Rpd3p. This finding suggests that Rpd3p and Ume1p might have cooperative binding interactions, and that Rpd3p, Ume1p, and Ume6p might be involved in regulating transcription during amino acid starvation.

To determine the false-positive rate of MOTIF REGRESSOR, we assigned randomly the observed expression fold changes to yeast genes, ran MDSCAN to find motifs of width 5–15, and screened the results by simple linear regressions. This process was repeated 100 times, giving a total of 40,324 motifs, among which 1,398 (3.5%) had regression *P* values <0.01. This number is slightly higher than the expected 1% because the genes used to find candidate motifs are also used to fit the regression. By comparison, 235 of the 414 candidate motifs (56.8%) from the amino acid starvation had regression *P* values <0.01, among which we expect <15 (3.5% of 414) to occur by chance.

The Discovery of Multiple Motifs That Influence Gene Expression over Many Time Points. We ran MOTIF REGRESSOR on expression data from 18 time points over two complete yeast cell cycles, starting from release from alpha-factor arrest in M/G_1 (18), and obtained a total of 273 significant motifs (including overlapping motifs). Many motifs are found at the time points at which they are known to have the strongest effects. For example, MBF motifs are found during G_1 with a strong inducing effect, and during G_2 with a strong repressing effect, but are not found in M or S phases. To examine these motifs' effects over all time points, we regressed for each motif at each time point the gene expression values against the upstream sequence motif-matching scores. With each motif represented by a vector of 18 simple regression coefficients, we hierarchically clustered the 273 motifs into 20 groups based on their Euclidean distances (Fig. 3).

Fifteen of the 20 motif clusters have putative influences that fluctuate with the cell cycle. Among them are the binding motifs of well-known cell-cycle regulators MCM1, SWI5, MCB, SCB, and SFF (Fig. 4*a*). Our findings that SWI5 promotes expression at M/G_1, SCB promotes expression at G_1, and MCB promotes expression at G_1 agree well with established data (29). MCM1 was known to promote expression in both G_2/M and M/G_1, and our findings are consistent with the latter. SFF is thought to function throughout cell cycle, but has been demonstrated to have an inductive influence in S/G_2, which our observations support. A number of clusters contain motifs such as ACGCGC and GACGC that resemble both MCB (consensus ACGCGT) and SCB (consensus CGCGAAA). One possibility is that these variant sites match the real MCB or SBF targets well enough to score highly, causing these motifs counted as significant in regression by the MCB/SCB targets. Another explanation is that Mbp1p and Swi4p, which recognize MCB and SCB respectively, are partially redundant regulators and the two proteins recognize each other's binding sites (31). Mbp1p and Swi4p could indeed bind sequences resembling MCB and SCB (although perhaps not optimally) and induce transcription enough to make regression significant.

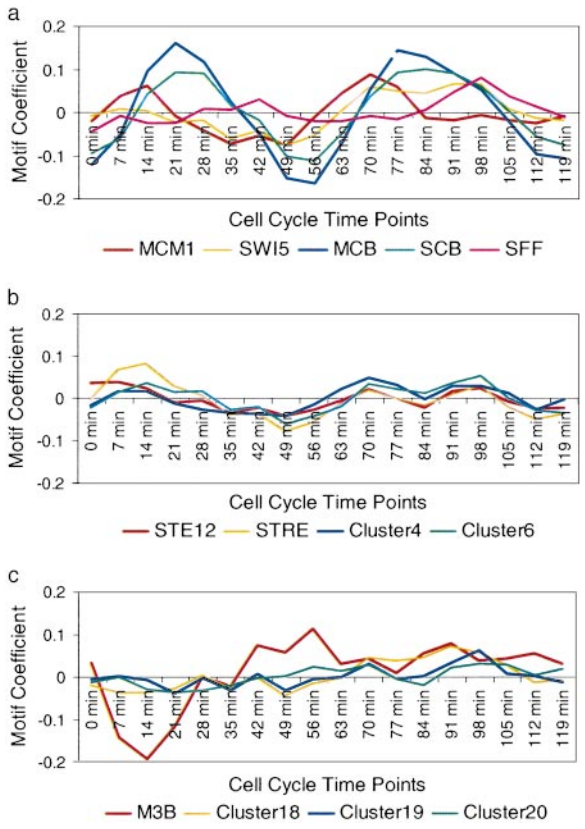

Fig. 4. Motif effects (coefficients) during the cell cycle. (*a*) Known cell-cycle-related motifs MCM1, SWI5, MCB, SCB, and SFF have coefficients that fluctuate with the cell cycle. (*b*) Other cell cycle motifs (STE12, STRE, and motifs in Cluster4 and Cluster6) influence expression through the cell cycle, but to a lesser extent than the known cell cycle regulators. (*c*) Non-cell-cycle motifs, M3B, and motifs in Cluster18, Cluster19, and Cluster20 showed sharp, low-amplitude fluctuations that correlate to a known experimental artifact that resulted from differential processing of odd- and even-numbered time points (G. Sherlock, personal communication).

We found the STE12 motif, which was correlated with an inductive effect, in the earliest time points. This result is expected because release from mating pheromone was used to synchronize the cells in this experiment and Ste12p is the key transcriptional activator for pheromone-induced transcription. We also found that STE12 retained its putative inductive effect in the second G_1 cycle, which is not likely to be explained by the synchronization method, but is consistent with its reported joint role with MCM1 to activate a subset of G_1-induced genes (32). The putative strong inductive effect of STRE observed immediately after cell-cycle arrest can be explained as a stress response to the centrifugation and handling required for the release from cell-cycle arrest. A milder inductive effect of STRE is also seen in the following cell cycle throughout G_1 and S phases (Fig. 4*b*). Although ≈50% of yeast genes contain the STRE motif in their promoters (16) and many STRE genes are also regulated by a host of other TFs, this result suggests further experiments to

GENETICS

determine whether STRE indeed induces gene expression during G_1/S.

Influences of five motif clusters do not vary with the cell cycle. They include two clusters for the M3B motif, which was also found in the amino acid starvation experiment and is present in upstream of genes involved in RNA processing. The variation in the coefficient of the other three motif clusters, especially cluster 19, seemed to correlate with a known experimental artifact (G. Sherlock, personal communication; Fig. 4*c*). The biological significance of these motifs is thus difficult to interpret.

Discussion

Previously described approaches for regulatory motif discovery either extract the motif features from the upstream sequences with little help from microarray-derived expression values (9, 10), or conduct feature selection based solely on the correlation between short oligomer motif occurrences and expression values (12, 13). MOTIF REGRESSOR uses MDSCAN as a feature extraction tool to find candidate motif matrices and then uses correlation analysis to select motifs relevant to changes in gene expression. Although our feature selection step resembles REDUCER (12) and the filtering method (10), our adoption of matrix-based motif finding and linear regression enhances both the sensitivity and the specificity. Microarray data from a single experiment can be rapidly analyzed to determine what motifs might be influencing the changes observed. For experiments with multiple time point measurements, the clustering of motif regression coefficients over time course provides a method to quickly recognize the effects of different TFs on the process being studied. With minor modifications, MOTIF REGRESSOR can also be applied to ChIP-array experiments.

For experimental biologists, MOTIF REGRESSOR is a useful catalyst for planning insightful and directed biological experiments. For example, an experiment is suggested by the analysis of the *YAP1* deletion experiment, in which the GCN4 motif was found upstream of up-regulated genes. One possible explanation for this result is that Gcn4p activity is increased in the absence of Yap1p, and the natural targets of Gcn4p are expressed at relatively higher levels. Alternatively, a subset of Gcn4p targets could be induced indirectly through a secondary pathway initiated by *YAP1* deletion not dependent on Gcn4p. A ChIP array experiment may determine whether in *YAP1* mutants Gcn4 protein actually binds to targets normally bound by Yap1p. Indeed, results obtained through computational approaches must always be tested for biological and mechanistic relevance *in vivo*.

The method described here assumes that the most interesting motifs are those that cause the most dramatic changes in gene expression under a given condition. Therefore, it may not be ideal for discovering motifs that specify consistent, but subtle, changes in expression. Another assumption is that a given motif can specify only one type of regulation at a given time point, either induction or repression. To detect motifs that may facilitate the binding of both activators and repressors, we may need to use the absolute value of the log-expression values as the dependent variable in the regression, or conduct regression for induced or repressed genes separately.

We thank Wei Wang for advice during algorithm development, and Douglas Brutlag, Audrey Gasch, Arkady Khodursky, Nobuo Ogawa, Gavin Sherlock, and the two anonymous referees for insight and help during the preparation of this manuscript. This research was supported by National Institutes of Health Grant 1F37LM07626-01 (to E.M.C.), the Helen Hay Whitney Foundation (to J.D.L.), National Institutes of Health Grant R01 HG02518-01 (to J.S.L.), and National Science Foundation Grant DMS-0204674 (to J.S.L.).

1. van Helden, J., Andre, B., Collado-Vides, J. (1998) *J. Mol. Biol.* **281,** 827–842.
2. Vilo, J., Brazma, A., Jonassen, I., Robinson, A. & Ukkonen, E. (2000) *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8,** 384–394.
3. Sinha, S. & Tompa, M. (2000) *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8,** 344–354.
4. Hampson, S., Kibler, D. & Baldi, P. (2002) *Bioinformatics* **18,** 513–528.
5. Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. & Wootton, J. C. (1993) *Science* **262,** 208–214.
6. Grundy, W. N., Bailey, T. L. & Elkan, C. P. (1996) *Comput. Appl. Biosci.* **12,** 303–310.
7. Stormo, G. D. & Hartzell, G. W., III (1989) *Proc. Natl. Acad. Sci. USA* **86,** 1183–1187.
8. Bussemaker, H. J., Li, H. & Siggia, E. D. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 10096–10100.
9. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. (1998) *Nat. Biotechnol.* **16,** 939–945.
10. Hughes, J. D., Estep, P. W., Tavazoie, S. & Church G. M. (2000) *J. Mol. Biol.* **296,** 1205–1214.
11. Holmes, I. & Bruno, W. J. (2000) *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8,** 202–210.
12. Bussemaker, H. J., Li, H. & Siggia, E. D. (2001) *Nat. Genet.* **27,** 167–171.
13. Keles, S., Van Der Laan, M. & Eisen, M. B. (2002) *Bioinformatics* **18,** 1167–1175.
14. Liu, X. S., Brutlag, D. L. & Liu, J. S. (2002) *Nat. Biotechnol.* **20,** 835–839.
15. Gasch, A. P., Huang, M., Metzner, S., Botstein, D., Elledge, S. J. & Brown, P. O. (2001) *Mol. Biol. Cell* **12,** 2987–3003.
16. Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D. & Brown, P. O. (2000) *Mol. Biol. Cell* **11,** 4241–4257.
17. Hughes, T. R., Roberts, C. J., Dai, H., Jones, A. R., Meyer, M. R., Slade, D., Burchard, J., Dow, S., Ward, T. R., Kidd, M. J., *et al*. (2000) *Nat. Genet.* **25,** 333–337.
18. Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, B. (1998) *Mol. Biol. Cell* **9,** 3273–3297.
19. Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. R., Thompson, C. M., Simon I., *et al*. (2002) *Science* **298,** 799–804.
20. Kurdistani, S. K., Robyr, D., Tavazoie, S. & Grunstein, M. (2002) *Nat. Genet.* **31,** 248–254.
21. Zhu, J. & Zhang, M. Q. (1999) *Bioinformatics* **15,** 607–611.
22. Fernandes, L., Rodrigues-Pousada, C. & Struhl, K. (1997) *Mol. Cell. Biol.* **17,** 6982–6993.
23. Schneider, T. D. & Stephens, R. M. (1990) *Nucleic Acids Res.* **18,** 6097–6100.
24. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. (1999) *Nat. Genet.* **22,** 281–285.
25. Blaiseau, P. L. & Thomas, D. (1998) *EMBO J.* **17,** 6327–6336.
26. Ogawa, N., DeRisi, J. & Brown, P. O. (2000) *Mol. Biol. Cell* **11,** 4309–4321.
27. Lieb, J. D., Liu, X., Botstein, D. & Brown, P. O. (2001) *Nat. Genet.* **28,** 327–334.
28. Warner, J. R. (1999) *Trends Biochem. Sci.* **24,** 437–440.
29. Gailus-Durner, V., Chintamaneni, C., Wilson, R., Brill, S. J. & Vershon, A. K. (1997) *Mol. Cell. Biol.* **17,** 3536–3546.
30. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998) *Proc. Natl. Acad. Sci. USA* **25,** 14863–14868.
31. Simon, I., Barnett, J., Hannett, N., Harbison, C. T., Rinaldi, N. J., Volkert, T. L., Wyrick, J. J., Zeitlinger, J., Gifford, D. K., Jaakkola, T. S. & Young, R. A. (2001) *Cell* **106,** 697–708.
32. Oehlen, L. J., McKinney, J. D. & Cross, F. R. (1996) *Mol. Cell. Biol.* **16,** 2830–2837.