© 1999 Nature America Inc. • http://genetics.nature.com

Systematic determination of genetic network architecture

Saeed Tavazoie[1], Jason D. Hughes[1,2], Michael J. Campbell[3], Raymond J. Cho[4] & George M. Church[1]

Technologies to measure whole-genome mRNA abundances[1–3] and methods to organize and display such data[4–10] are emerging as valuable tools for systems-level exploration of transcriptional regulatory networks. For instance, it has been shown that mRNA data from 118 genes, measured at several time points in the developing hindbrain of mice, can be hierarchically clustered into various patterns (or 'waves') whose members tend to participate in common processes[5]. We have previously shown that hierarchical clustering can group together genes whose *cis*-regulatory elements are bound by the same proteins *in vivo*[6]. Hierarchical clustering has also been used to organize genes into hierarchical dendograms on the basis of their expression across multiple growth conditions[7]. The application of Fourier analysis to synchronized yeast mRNA expression data has identified cell-cycle periodic genes, many of which have expected *cis*-regulatory elements[8]. Here we apply a systematic set of statistical algorithms, based on whole-genome mRNA data, partitional clustering and motif discovery, to identify transcriptional regulatory sub-networks in yeast—without any a priori knowledge of their structure or any assumptions about their dynamics. This approach uncovered new regulons (sets of co-regulated genes) and their putative *cis*-regulatory elements. We used statistical characterization of known regulons and motifs to derive criteria by which we infer the biological significance of newly discovered regulons and motifs. Our approach holds promise for the rapid elucidation of genetic network architecture in sequenced organisms in which little biology is known.

We designed our approach to be systematic and minimally biased by previous knowledge of yeast biology. Our objective was to discover distinct expression patterns (clusters) in mRNA data sets and then identify upstream DNA sequence patterns specific to each expression cluster. A DNA sequence pattern that is specific to a single expression cluster constitutes the primary hypothesis for the *cis*-regulatory element through which co-regulation of the genes within the cluster is achieved.

We used data gathered by Cho *et al.*[11] who used Affymetrix oligonucleotide microarrays[12] to query the abundances of 6,220 mRNA species in synchronized *Saccharomyces cerevisiae* batch cultures. The data provided us with 15 time points, across two cell cycles. We variance-normalized the expression profile of each ORF and clustered the most variable 3,000 ORFs into 30 clusters of 49–186 ORFs per cluster. The clustering procedure groups together ORFs on the basis of their common expression patterns across the time points. We and others have previously used hierarchical algorithms[13] for clustering such data[4–8]. Here we use the *k*-means algorithm[14], a partitional method[13] that by iterative reallocation of cluster members minimizes the overall within-cluster dispersion.

We found the members of each cluster to be significantly enriched for genes with similar functions. We mapped the genes in each cluster to the 199 functional categories in the Martinsried Institute of Protein Sciences functional classification scheme (MIPS) database[15]. For each cluster, we calculated *P* values for observing the frequencies of genes in particular functional categories. There was significant grouping of genes within the same

Table 1 • Enrichment of clusters for ORFs within functional categories

Cluster	Periodicity index	Number of ORFs (*n*)	MIPS functional category (total ORFs)	ORFs within functional category (*k*)	*P* value $-\log_{10}$
1	0.07	164	ribosomal proteins (206)	64	54
			organization of cytoplasm (555)	79	39
			organization of chromosome structure (41)	7	4
2	0.38	186	DNA synthesis and replication (82)	23	16
			cell-cycle control and mitosis (312)	30	8
			recombination and DNA repair (84)	11	5
			nuclear organization (720)	40	4
4	0.14	170	mitochondrial organization (339)	32	10
			respiration (79)	10	5
7	0.35	101	cell-cycle control and mitosis (312)	17	5
			budding, cell polarity, filament formation (161)	10	4[a]
			DNA synthesis and replication (82)	7	4[a]
8	0.09	148	TCA pathway (22)	5	4[a]
			carbohydrate metabolism (411)	22	4[a]
14	0.45	74	organization of centrosome (28)	6	6
			nuclear biogenesis (5)	3	5
			organization of cytoskeleton (93)	7	4[a]
30	0.24	60	nitrogen and sulphur metabolism (75)	9	8
			amino acid metabolism (203)	12	7

Periodicity index is a quantitative measure of cell-cycle periodicity. The most highly enriched functional categories are given for each cluster. We calculated *P* values using the cumulative hypergeometric probability distribution for finding at least (*k*) ORFs from a particular functional category within a cluster of size (*n*). Because 199 MIPS functional categories were tested for each cluster, *P* values greater than 3×10^{-4} are not reported, as their total expectation within the cluster would be greater than 0.05. [a]Because all 30 clusters were tested independently, these *P* values may have marginal significance.

[1]*Department of Genetics, Harvard Medical School, 200 Longwood Ave, Boston, Massachusetts 02115, USA.* [2]*Graduate Program in Biophysics, 200 Longwood Ave, Harvard University, Boston, Massachusetts 02115, USA.* [3]*Molecular Applications Group, 607 Hansen Way, Building One, Palo Alto, California 94303-1110, USA.* [4]*Department of Genetics, B400 Beckman Center, 279 Campus Drive, Stanford Medical Center, Palo Alto, California 94304, USA. Correspondence should be addressed to G.M.C. (e-mail: church@salt2.med.harvard.edu).*

© 1999 Nature America Inc. • http://genetics.nature.com

© 1999 Nature America Inc. • http://genetics.nature.com

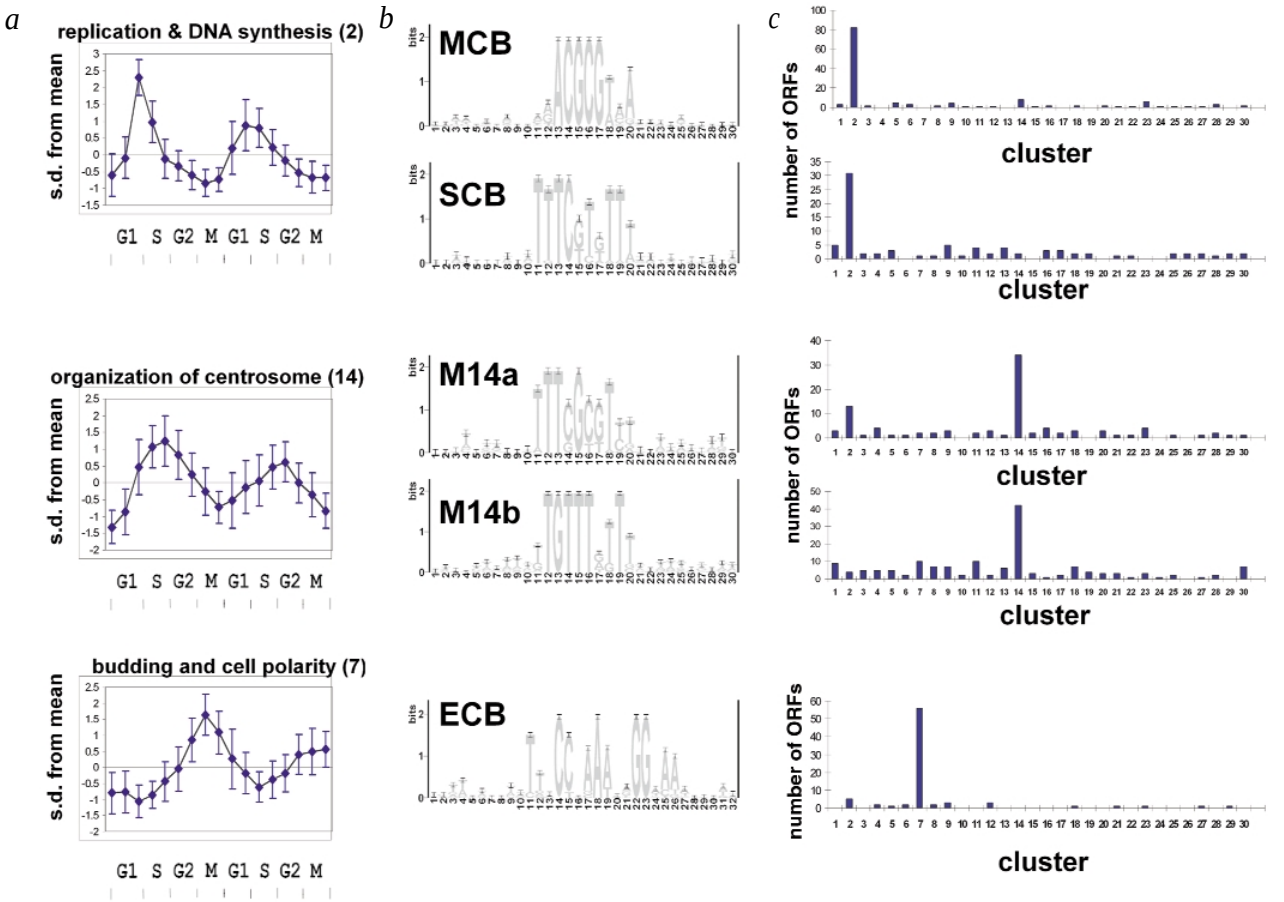

Fig. 1 Top periodic clusters, their motifs and overall distribution in all clusters. ***a***, Mean temporal profile of a cluster, named according to the biological functions for which it is most highly enriched (with the numerical designation of the cluster in parentheses). Error bars represent the standard deviation of the members of each cluster about the mean of the particular time point. ***b***, Sequence logo representation of the motif(s) discovered within the cluster. The height of each letter is proportional to its frequency. The letters are sorted with the most frequent one on top. The overall height of the stack signifies information content of the sequence at that position (0–2 bits). Motifs M14a and M14b were identified in this study. ***c***, The occurrence of the motif across all 30 clusters.

functional class (Table 1). The most notable functional grouping occurred for genes in cluster 1, where 64 of 164 genes encode ribosomal proteins (*P* value of 10^{-54}). Not all clusters showed significant enrichment for function. The members of such clusters may participate in multiple classically defined processes and therefore may not show significant enrichment in any one functional category. Alternatively, the number of clusters (30) may overestimate the underlying diversity of biological expression classes in the data set. We erred on the side of over-classification, however, to avoid missing significant expression classes. Subsequently, independent analyses, such as functional category enrichment and motif searching, aid in determining the biological significance of the clusters a posteriori. Note that the functional categories are only used to represent the enrichment of the clusters and were not used in any aspect of the analysis, including the motif discovery phase. The complete analysis is available (http://arep.med.harvard.edu/network_discovery).

The temporal profile of each cluster is represented by a plot of the mean, variance-normalized expression level of all the genes within the cluster (Figs 1*a*, 2*a*). Dispersion bars represent the standard deviation of the points along a particular dimension (in this case, time point). We used an index of cell-cycle periodicity to quantitate the extent of periodicity at the cell-cycle period of 80 minutes (Table 1). Of the top periodic clusters, three are profiled (Fig. 1*a*). Many of the genes in these clusters encode proteins which function in cell-cycle phase-specific processes such as replication (cluster 2), organization of centrosome (cluster 14), and budding and cell polarity (cluster 7). Note that the timing of maximum expression for the genes in these clusters agrees with the phase during which their product is required (G1-S for replication, S-G2 for organization of centrosome and M phase for budding and cell polarity).

Most clusters have non-periodic temporal profiles (Fig. 2*a*), with some showing complex behaviour. Members of cluster 1 show a relatively steady expression level, except for the peak during M-G1, but as can be seen from the relatively small dispersion bars, the members of this cluster are tightly co-regulated—a fact recapitulated in its 10-fold enrichment for ribosomal proteins.

We next conducted a blind and systematic search for upstream DNA sequence motifs that were common to members of each cluster. We did this to identify known or novel *cis*-regulatory elements that may contribute to the co-regulation of genes in a cluster. We used the program AlignACE (ref. 16), which finds globally optimal alignments within unaligned input sequences. We found that 18 motifs from 12 different clusters passed our criteria for biological significance; their average MAP score was 35 (range 12–82). Of these motifs, seven had been identified experimentally and are known to regulate the expression of many genes in their respective clusters. Multiple factors may account for why we did not find significant motifs in all clusters. First, our criteria for calling a motif 'significant' may be too stringent. Second, the co-regulation of the members of some clusters may be achieved through post-transcriptional mechanisms (such as those controlling

© 1999 Nature America Inc. • http://genetics.nature.com

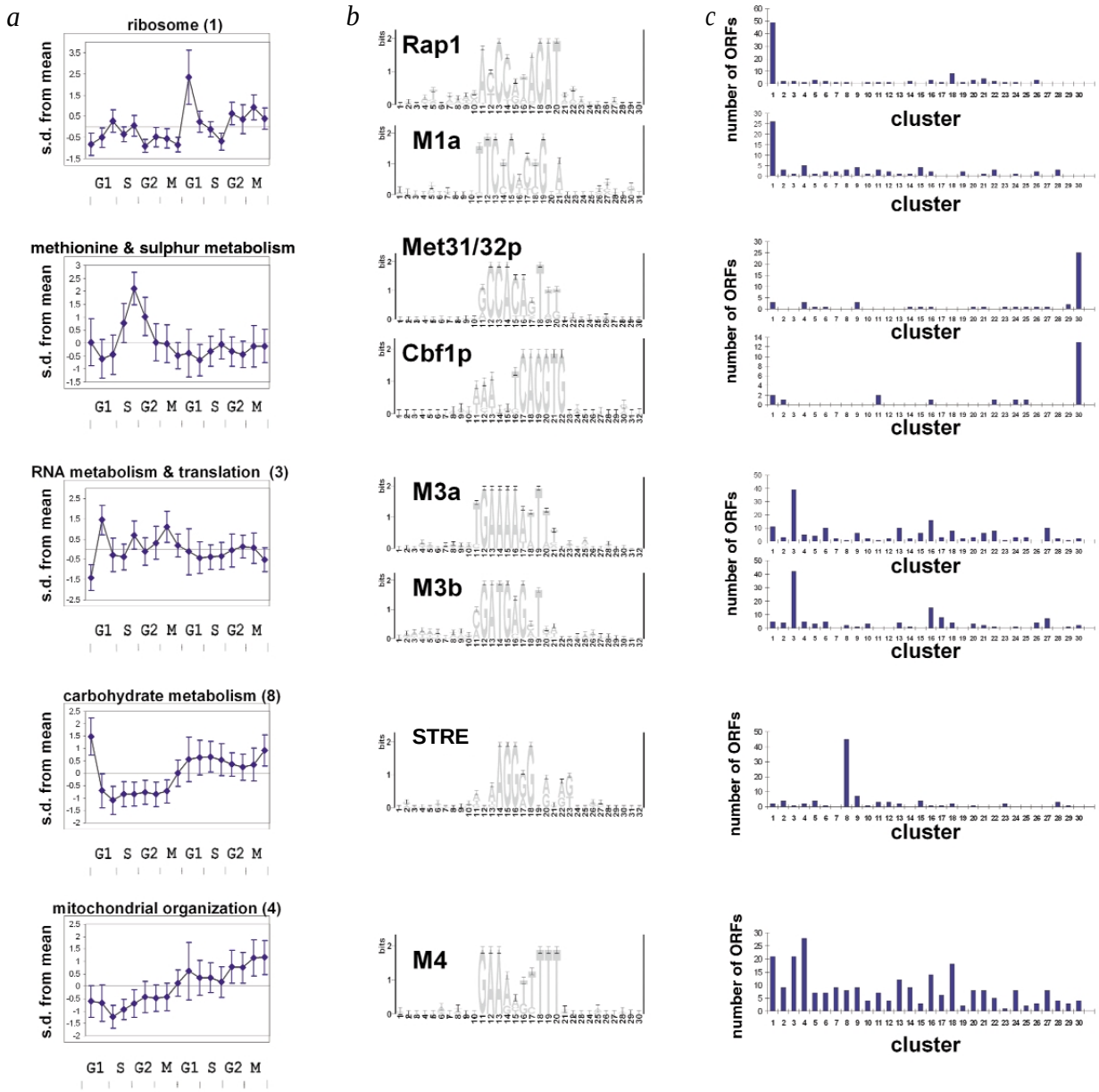

Fig. 2 'Non-periodic' clusters, their motifs and overall distribution in all clusters. ***a***, Mean temporal profile of a cluster. ***b***, Sequence logo representation of the motif(s). Motifs M1a, M3a, M3b and M4 were identified in this study. ***c***, The occurrence of the motif across all 30 clusters.

mRNA stability). Finally, some of the clusters may represent noise in the data set, with little underlying biological coherence.

We have represented the motifs in a graphical format, which shows the information content in bits[17] at each base (Figs 1*b*, 2*b*). Once a motif was deemed biologically significant, we searched the yeast genome to identify additional sites that scored greater than one standard deviation below the mean Berg-von-Hippel score[18] for that motif. We constructed histograms for the distribution of a motif upstream of all the ORFs within all 30 clusters (Figs 1*c*, 2*c*). Most motifs are highly selective for the cluster in which they were found. For example, 55 of 101 members of cluster 7 (budding and cell polarity) had the early cell cycle box (ECB) motif, whereas no other cluster had more than 4% of its ORFs containing this motif.

For the clusters whose members belong to known regulons, the expected *cis*-regulatory element(s) emerged as the highest scoring motif(s) in every case. For example, the highest scoring motifs that emerged from periodic clusters 2 and 7 have well-established roles in the periodic transcription of the genes in these clusters (MCB box (*Mlu*I cell-cycle box) and SCB (Swi 4/6 cell-cycle box; ref. 19) for cluster 2 and ECB (refs 20,21) for cluster 7). The same was true for cluster 1 (ribosomal proteins, Rap1p binding site[22]), cluster 8 (carbohydrate metabolism, STRE binding site[23]) and cluster 30 (methionine and sulphur metabolism, Cbf1p and Met31/32p binding sites[24]).

We also found a cohesive cluster of 73 ORFs (cluster 14) that achieves maximal expression during the S-G2 phase. The members of this cluster function in the organization of centrosome and nuclear biogenesis (Table 1). Cluster 14 has the highest periodicity index (0.45) and its profile peaks slightly later than that of cluster 2 (replication and DNA synthesis). Our search in the upstream region of cluster 14 ORFs identified 2 new motifs (M14a and M14b) that are highly specific for these genes (Fig. 1*b*,*c*). The discovery of this large and tightly co-regulated class of periodic genes, together with their putative *cis*-regulatory motifs, extends the

© 1999 Nature America Inc. • http://genetics.nature.com

© 1999 Nature America Inc. • http://genetics.nature.com

number of known periodic classes, adding S-G2 to the well-known G1-S and M-phase induced regulons.

We found 2 of the highest scoring motifs upstream of cluster 3 genes (Fig. 2*b*,*c*). The MIPS classification scheme did not adequately capture the functional enrichment of this cluster, as its members spanned multiple categories with the common theme of RNA metabolism and translation. These genes encode Pol I and Pol III subunits, RNA/tRNA splicing factors, translation initiation factors, RNA helicases and various other proteins involved in RNA metabolism. The cluster 3 motifs (M3a and M3b) have not been previously described, but their strong specificity for genes within cluster 3 and for RNA and translation-related genes outside of cluster 3 suggests that they have a role in the global regulation of protein synthesis. Furthermore, motifs M3a and M3b had tight upstream distributions (Fig. 3*f*). The distances between the 25th and 75th percentile from the ATG were 144 (m3a) and 111 (m3b) bp, whereas the average for all the known motifs was 229 bp.

Only half of the 30 clusters were significantly enriched for functional categories or had significant motifs. What are the statistical characteristics of clusters that correlate with these independent measures of cluster coherence? An important characteristic of a cluster is its 'tightness', or roughly speaking, how close its members are to the mean of the cluster. We defined a mean Euclidean distance (MUD) for every cluster (the average Euclidean distance of all the members of a cluster from the cluster mean). Based on this metric, the clusters with significant functional enrichment tend to be tighter (MUD=0.60 versus 0.66; *P* value=0.02). We saw a stronger correlation between tightness of clusters and presence of significant motifs (MUD=0.58 versus 0.66; *P* value=0.006). Furthermore, genes containing significant motifs within a cluster tend to be closer to the centre of their clusters.

We should stress that we designed our approach with minimal biases. Information about yeast biology did not influence the formation of clusters or evaluation of motifs. These are important criteria for the validation of emerging methodologies, as they must correctly identify the structure of known networks without any a priori knowledge of their structure or any assumptions about their dynamics. In this context, our identification of known and expected *cis*-regulatory elements as the top-scoring motif, in every case, is a significant outcome. In terms of novel regulons and their motifs, we have introduced new post-clustering analyses that characterize and validate the biological significance of expression clusters and their motifs. These analyses also provide quantitative means by which it is possible to compare alternative clustering approaches.

The rapid sequencing of many organisms of biological and clinical importance has made urgent the task of identifying the function of many thousands of novel genes. The methodology presented here has expanded the membership of known regulons by placing hundreds of unknown ORFs into regulation and motif classes. The association of these unknown ORFs with well-characterized genes and motifs generates many hypotheses for their biological roles. Furthermore, the systematic approach presented here is ideally suited for determining the transcriptional regulatory networks of newly sequenced organisms in which little biology is known. The combination of experimental and computational approaches presented here, together with experimental verification of novel motifs and the discovery of their *trans*-acting factors, should allow the construction of the circuit diagram for the genetic network, allowing us to both understand and manipulate complex cellular processes on a systems level.

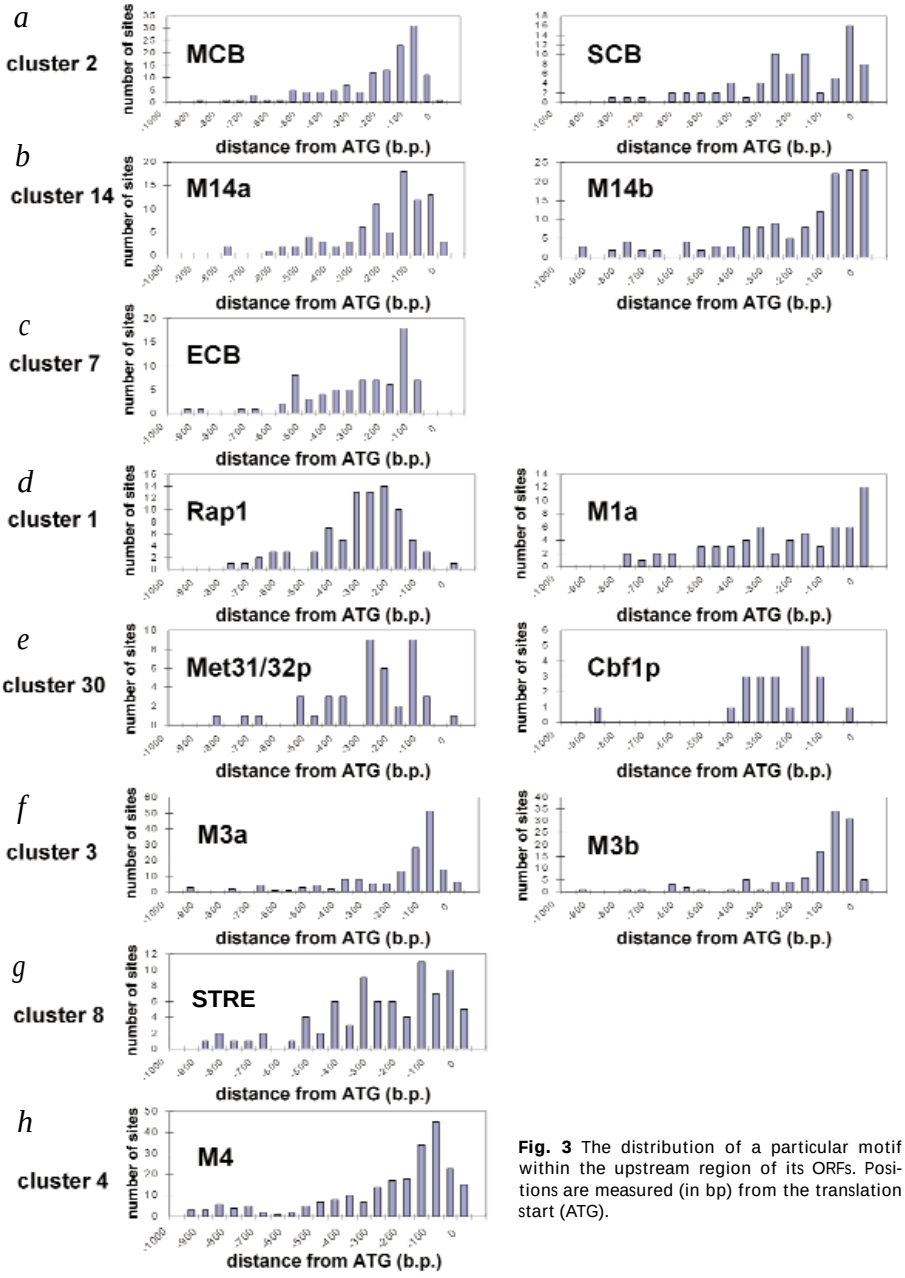

Fig. 3 The distribution of a particular motif within the upstream region of its ORFs. Positions are measured (in bp) from the translation start (ATG).

© 1999 Nature America Inc. • http://genetics.nature.com

Methods

Variance normalization and clustering of expression time series. We selected the top 6,000 genes according to average expression level. For subsequent analysis, we chose the 3,000 most variable ORFs using a metric of variation based on the normalized dispersion in expression level of each gene across the time points (s.d./mean). We used 15 time points to construct a 3,000 by 15 data matrix (time points at 90 and 100 min were excluded from the analysis due to the less efficient labelling of their mRNA during the original chip hybridizations). The data matrix was then transformed such that the variance of each gene was normalized across the 15 conditions[6]. This was done by subtracting its mean across the time points from the expression level of each gene, and dividing by the standard deviation across the time points:

$$Y_{ij} = \frac{X_{ij} - \langle X_i \rangle}{[\frac{1}{15}\sum_{j=1}^{15}(X_{ij} - \langle X_i \rangle)^2]^{\frac{1}{2}}}$$

The 3,000 members of this transformed data matrix occupy a 15-dimensional 'expression space. We then partitioned the 3,000 genes into clusters whose members share some measure of similarity in their expression pattern. The Euclidean distance metric was used to define distance between the coordinates of any two genes in the space[5]. Thus, the smaller the distance between any two genes, the more similar they are in expression pattern. Other metrics are also used in multivariate clustering[13], and our use of the Euclidean distance reflects our ignorance of a more biologically relevant measure of distance. We used the *k*-means algorithm[14] to cluster the 3,000 genes into different regulation classes. *k*-means is an unsupervised, iterative algorithm that minimizes the within-cluster sum of squared distances from the cluster mean. We used an implementation of *k*-means in the statistical software package SYSTAT 7.0 (SPSS). The first cluster centre was chosen as the centroid of the entire data set and subsequent centres were chosen by finding the data point farthest from the centres already chosen. We repeated the algorithm for 200–400 iterations and partitioned the 3,000 genes into 10, 30 and 60 clusters. By 200 iterations, the algorithm had converged because the cluster memberships did not change appreciably between 200 and 400 iterations. We chose the 30-cluster partitioning because it provided the best compromise between number of clusters and separation between them.

Searching for common upstream regulatory motifs. We used the program AlignACE (ref. 16) to conduct an unbiased search for common DNA-sequence motifs within 600 bp upstream of the ORFs within each cluster. We performed independent searches using three sets of ORF inputs from each cluster. The first set consisted of the 50 ORFs closest (in Euclidean distance) to the centre of each cluster, and the other 2 non-overlapping sets each contained approximately half of the 49–80 ORFs closest to the centre of each cluster. Our rationale for using multiple input sets was to increase the probability of finding rare motifs and to use the discovery of the same motif in multiple sets of ORFs to strengthen the case for its causal association with the cluster rather than by chance alone. We used the following AlignACE settings: the number of columns (expected number of conserved bases) was 10; the expected number of sites was 10; maximum number of initial sampling runs was 500; iterative masking was performed a maximum of 100 times; and near-optimum sampling commenced after 50 consecutive sampling runs without an increase in alignment score. AlignACE calculates a statistic called the MAP score[16]. This score is an internal metric used to determine the significance of an alignment. Our criteria for considering a motif 'biologically significant' consisted of two conditions: (i) that at least two of three searches in each cluster yielded the motif; and (ii) that the motif had a MAP score of ten or higher. The discovered motifs were displayed using a described method[17].

Determining the cell-cycle periodicity of clusters. The period of the cell cycle was determined using the cytological analysis of cell-cycle phase[11] and the Fourier spectrum of the temporal profiles of two well-studied periodic transcripts (*CLN1* and *CLN2*). These two independent approaches gave very similar results (80±10 min). We calculated Fourier amplitudes for the mean profiles of all clusters at eight equally spaced frequencies (0.00625–0.05 min^{-1}, corresponding to periods between 160 min and 2 min). The index of cell-cycle periodicity was defined as the ratio of Fourier component magnitude at 0.0125 min^{-1} (80 min) to the sum of all 8 Fourier components. The periodicity index ranged from 0.05 to 0.45 with a standard deviation of 0.11. The top four periodic clusters (14, 11, 2 and 7) had an average periodicity index greater than two standard deviations from the mean.

Determination of statistical significance for functional category enrichment. The hypergeometric distribution was used to obtain the chance probability of observing the number of genes from a particular MIPS functional category within each cluster. More specifically, the probability of observing at least (k) ORFs from a functional category within a cluster of size (n) is given by:

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{f}{i}\binom{g-f}{n-i}}{\binom{g}{n}} \quad ,$$

where (f) is the total number of genes within a functional category and (g) is the total number of genes within the genome (6,220). As we tested 199 MIPS (ref. 15) functional categories for each cluster, P values greater than 3×10^{-4} are not reported, as their total expectation within the cluster would be higher than 0.05.

Note added in proof: As predicted motifs are experimentally verified, the data will be accessible from our web site.

Acknowledgements

We thank D. Lockhart and L. Wodicka for support, and B. Gewurz, V. Mootha, S. Tavazoie, M. Tavazoie and members of the Church lab, especially P. Estep, R. Mitra, B. Cohen, J. Johnson, M. Bulyk and J. Aach, for discussions and critical readings of the manuscript. This work was supported by the US Department of Energy (grant DE-FG02-87-ER60565), the office of Naval Research and DARPA (grant N00014-97-1-0865), the Lipper Foundation and Hoechst Marion Roussel.

Received 7 January; accepted 21 May 1999.

1. Velculescu, V.E. *et al*. Characterization of the yeast transcriptome. *Cell* **88**, 243–251 (1997).
2. Lockhart, D.J. *et al*. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.* **14**, 1675–1680 (1996).
3. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
4. Weinstein, J.N. *et al*. An information-intensive approach to the molecular pharmacology of cancer. *Science* **275**, 343–349 (1997).
5. Wen, X. *et al*. Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl Acad. Sci. USA* **95**, 334–339 (1998).
6. Tavazoie, S. & Church, G.M. Quantitative whole-genome analysis of DNA-protein interactions by *in vivo* methylase protection in *E. coli*. *Nature Biotechnol.* **16**, 566–571 (1998).
7. 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).
8. Spellman, P.T. *et al*. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiase* by microarray hybridization. *Mol. Biol. Cell* **9**, 3273–3297 (1998).
9. Holstege, F.C. *et al*. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717–728 (1998).
10. Tamayo, P. *et al*. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA* **96**, 2907–2912 (1999).
11. Cho, R.J. *et al*. A genome wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2**, 65–73 (1998).
12. Wodicka, L., Dong, H., Mittmann, M., Ho, M.H. & Lockhart, D.J. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nature Biotechnol.* **15**, 1359–1366 (1997).
13. Everitt, B. *Cluster Analysis* 122 (Heinemann, London, 1974).
14. Hartigan, J.A. *Clustering Algorithms* 351 (Wiley, New York, 1975).
15. Mewes, H.W. *et al*. Overview of the yeast genome. *Nature* **387**, 7–65 (1997).
16. Roth, F.P., 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**, 949–945 (1998).
17. Schneider, T.D. & Stephens, R.M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100 (1990).
18. 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).
19. Koch, C. & Nasmyth, K. Cell cycle regulated transcription in yeast. *Curr. Opin. Cell Biol.* **6**, 451–459 (1994).
20. McInerny, C.J., Partridge, J.F., Mikesell, G.E., Creemer, D.P. & Breeden, L.L. A novel Mcm1-dependent element in the SWI4, CLN3, CDC6, and CDC47 promoters activates M/G1-specific transcription. *Genes Dev.* **11**, 1277–1288 (1997).
21. Kuo, M. & Grayhack, E. A library of yeast genomic MCM1 binding sites contains genes involved in cell cycle control, cell wall and membrane structure, and metabolism. *Mol. Cell. Biol.* **14**, 348–359 (1994).
22. Planta, R.J., Goncalves, M. & Mager, W.H. Global regulators of ribosome biosynthesis in yeast. *Biochem. Cell Biol.* **73**, 825–834 (1995).
23. Moskovina, E. *et al.* A search in the genome of Saccharomyces cerevisiae for genes regulated via stress response elements. *Yeast* **14**, 1041–1050 (1998).
24. Thomas, D. & Surdin-Kerjan, Y. Metabolism of sulfur amino acids in *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.* **61**, 503–532 (1997).

© 1999 Nature America Inc. • http://genetics.nature.com