BIOINFORMATICS

Vol. 17 Suppl. 1 2001
Pages S49–S55

Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles

Derek Y. Chiang[1], Patrick O. Brown[2] and Michael B. Eisen[1,3]

[1]Dept. of Molecular and Cell Biology, U. of California, Berkeley, CA 94720, U.S.A., [2]Dept. of Biochemistry and Howard Hughes Medical Institute, Stanford U. School of Medicine, Stanford, CA 94305, U.S.A. and [3]Life Sciences Division, Ernest Orlando Lawrence Berkeley National Lab, Berkeley, CA 94720, U.S.A.

Received on February 6, 2001; revised and accepted on April 3, 2001

ABSTRACT

The combination of genome-wide expression patterns and full genome sequences offers a great opportunity to further our understanding of the mechanisms and logic of transcriptional regulation. Many methods have been described that identify sequence motifs enriched in transcription control regions of genes that share similar gene expression patterns. Here we present an alternative approach that evaluates the transcriptional information contained by specific sequence motifs by computing for each motif the mean expression profile of all genes that contain the motif in their transcription control regions. These genome-mean expression profiles (GMEP's) are valuable for visualizing the relationship between genome sequences and gene expression data, and for characterizing the transcriptional importance of specific sequence motifs.

Analysis of GMEP's calculated from a dataset of 519 whole-genome microarray experiments in *Saccharomyces cerevisiae* show a significant correlation between GMEP's of motifs that are reverse complements, a result that supports the relationship between GMEP's and transcriptional regulation. Hierarchical clustering of GMEP's identifies clusters of motifs that correspond to binding sites of well-characterized transcription factors. The GMEP's of these clustered motifs have patterns of variation across conditions that reflect the known activities of these transcription factors.

Software that computed GMEP's from sequence and gene expression data is available under the terms of the Gnu Public License from http://rana.lbl.gov/

Contact: mbeisen@lbl.gov

INTRODUCTION

As genome sequencing projects move forward at a rapid pace, and as the use of DNA microarrays and related techniques becomes more widespread, there is a growing number of organisms for which both complete genome sequences and large volumes of genome-wide transcript abundance measurements are available. An obvious challenge in the analysis of these data is to understand the cellular mechanisms used to orchestrate genomic expression programs. As complex models of transcriptional networks have yet to reach maturity, most recent research has focused on the more modest goal of using genome-wide expression patterns and genome sequences to identifying likely (and ideally previously unidentified) transcription factor binding sites.

Most common strategies adopt a "group-by-expression" approach, in which genes with similar expression are identified, and then their transcription control regions are analyzed for the presence of shared sequence motifs (reviewed in Ohler and Niemann, 2001). These approaches postulate that genes with similar patterns of expression are likely to be regulated by common factors, and thus should share binding sites for these factors in their non-coding regions. Co-expressed genes are identified by cluster analysis of gene expression data (c.f. DeRisi et al., 1997, Spellman *et al.,* 1998, Cho *et al.,* 1998, Tavazoie *et al.,* 1999, Gasch *et al.,* 2000). Sequences upstream of co-expressed genes are analyzed for statistically over-represented sequence motifs using a variety of algorithms, including: expectation maximization (Bailey and Elkan 1995), over-represented oligomers (van Helden *et al.,* 1998, Wolfsberg *et al.,* 1999), weight matrices (Hertz and Stormo, 1999), Gibbs sampling (Hughes *et al.,* 2000), enumerative statistics (Sinha and Tompa, 2000), probabilistic segmentation (Bussemaker *et al.,* 2000), and sequence pattern discovery (Vilo *et al.,* 2000). As has as been previously noted (Holmes and Bruno, 2000; Wagner 1999), a problem with this approach is that it does not take into account the multiple, independent mechanisms by which most genes are regulated. For example, two genes can be co-regulated under one set of conditions,

but differentially regulated under others. Although these genes would not be easily identified as co-expressed, they nonetheless share important regulatory information.

An alternate strategy is to adopt a "group-by-sequence" approach in which the transcriptional control content of sequence motifs is evaluated on the basis of the expression patterns of genes that contain the motif in their nominal transcription control regions (TCR; the adjacent cis-DNA that is believed to contain sequences that determine the transcriptional regulation of the gene). If a sequence motif carries transcriptional information - namely if it is bound by a transcription factor and this binding alters the transcription rate of adjacent genes - we expect the expression patterns of genes that contain this motif in their TCR's to have non-random features that reflect the activity of the corresponding factor. In contrast, if a motif does not encode transcriptional regulatory information, the genes that contain the motif in their TCR's should not have expression patterns that differ significantly from those of the entire population of genes.

To evaluate and exploit this expectation, we define the genome-mean expression profile (GMEP) of a sequence motif as the mean expression profile of all genes (regardless of the expression profiles) that contain this motif in their TCR's. To understand the reasons for using GMEP's, consider a set of genes whose transcription is increased by the activity of a given transcription factor in some set of conditions, but whose expression patterns are otherwise unrelated. Although the multi-factorial nature of transcription control could easily obscure the commonalities in these gene expression profiles, we nonetheless expect that, on average, these genes will have higher expression levels in the conditions where this transcriptional activator is active when compared to some randomly chosen subsets of genes, and we expect the magnitude of this elevation will reflect the activity level of the activator. Additional genes that contain this motif in their TCR's but which are not regulated by the particular activator should also have mean expression profiles that are close to the population mean profile. Thus, the GMEP of the sequence motif recognized by the activator should differ significantly from the population mean profile only when the activator is present and active and this difference should be greatest when the activator has its highest level of activity. Note that this should still be true even if transcription of the regulated genes is also independently and separately controlled by additional non-overlapping factors.

ALGORITHM

Data

To compute GMEP's, we begin with a data matrix D with r rows, each representing a single gene, and c columns, each representing a single experimental condition. Each cell D_{gj} represents the expression level of gene g in condition j. Missing values are allowed. In the data used here these values are log-transformed (base 2) relative expression ratios (compared to a suitable reference sample) and the columns are mean-centered. For each gene, we define a sequence $S(g)$ that is the genome sequence of the gene's nominal transcription control region. Note that for most organisms this are no well-defined rules for identifying TCR's; for analyses presented here using the yeast *Saccharomyces cerevisiae*, $S(g)$ is the 600 basepairs upstream of the translation start site for gene g.

Genome-mean expression profiles

For a DNA sequence motif m, let G be the set of genes that contain this motif in their TCR's:

$$g \in G \Leftrightarrow m \in S(g)$$

Define the **genome-mean expression profile** of motif m [denote GMEP(m)] to be the c-dimensional vector equal to the weighted mean of the c-dimensional vectors that represent the expression profiles of each gene in G:

$$\text{GMEP}(m)_j = \frac{\sum_{g}^{g \in G} w_{mg} \cdot D_{gj}}{\sum_{g}^{g \in G} w_{mg}} \qquad (1)$$

where w_{mg} is the number of occurrences of motif m in $S(g)$. A weighted mean was used since transcription factors may have a higher affinity to genes that contain multiple copies of their cognate sites (Wagner, 1999).

For simplicity, here we only enumerate motifs containing the symbols A, C, G, or T although this is not a necessary constraint. For a given data matrix D and a fixed motif length L, we compute the $(4^L \times c)$ matrix where each row is the GMEP of a single motif. To correspond with the data matrix D, the columns in the GMEP matrix are mean-centered.

Significance testing

To analyze the likelihood that specific values in our GMEP matrixes are expected to have occurred by chance we compute approximate Z-scores. Consider the calculation of a GMEP as the mean of a sample $(X_1, \ldots, X_n)$ of n gene expression levels drawn randomly (with replacement) from a population. This population comprises all relative gene expression measurements from a single microarray experiment. If a motif does not contain transcriptional information, the expression levels of genes that contain this motif in their TCR's represent a randomly drawn sample, and the GMEP for this motif should not differ significantly from the the population mean. Alternatively,

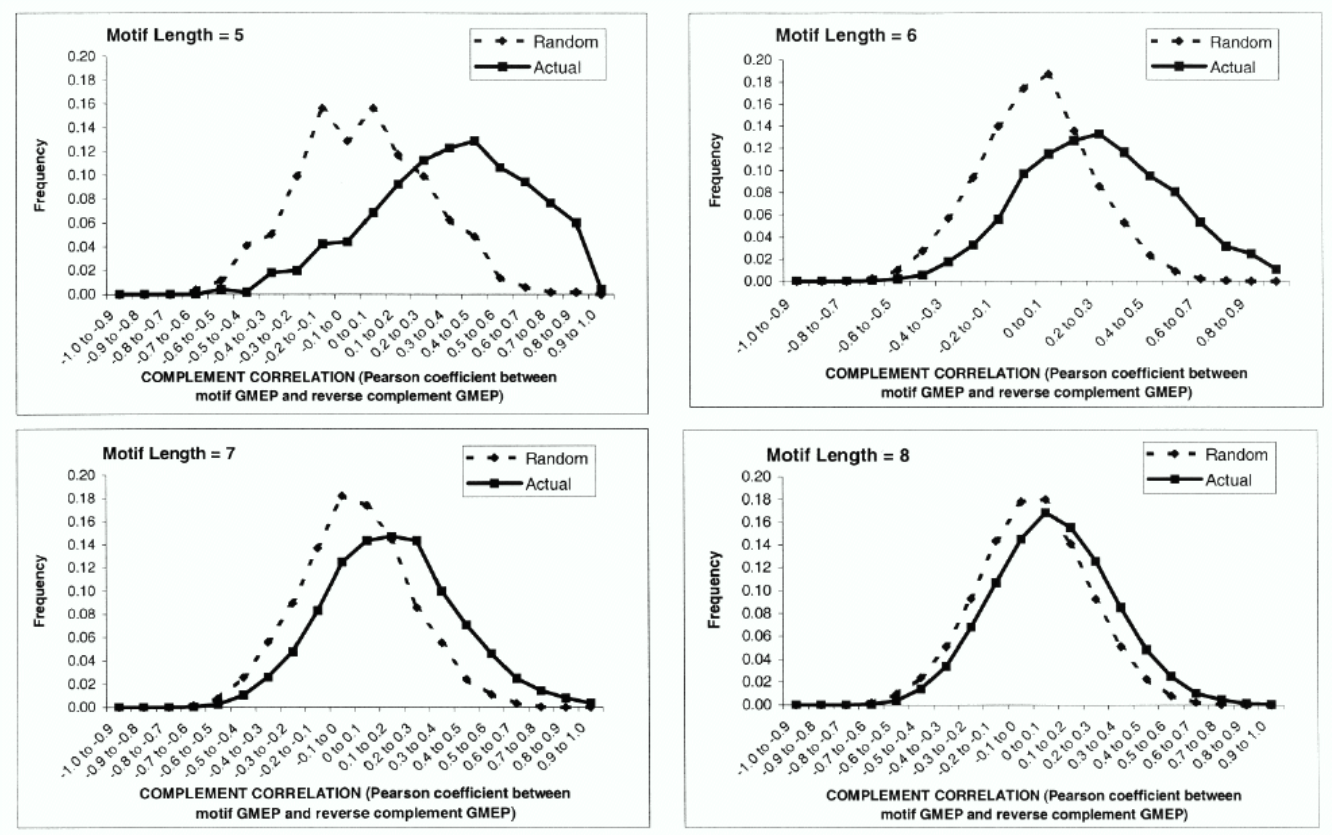

Fig. 1. Distributions of complement correlations for all motif/reverse complement pairs. Correlations between the GMEP for a motif and the GMEP of its reverse complement were calculated as described. Dashed lines indicate the distribution of Pearson coefficients for randomly permuted associations between each TCR sequence and a gene expression profile, whereas solid lines indicate the distributions for actual data. Mean values of the actual vs. randomized distributions: 0.357 vs. 0.009 (length 5); 0.245 vs. −0.006 (length 6); 0.148 vs. 0.003 (length 7); 0.080 vs. −0.001 (length 8).

if a motif does contain transcriptional information, and the corresponding transcription factor(s) are active, then we expect the GMEP to be different from the population mean.

Assume that $X_1, \ldots, X_n$ are independent and identically distributed. From the central limit theorem (CLT), we expect that the sample means of many samples chosen randomly from a given (microarray) population with mean μ and standard deviation s will fall on a normal distribution with mean μ and standard deviation $\frac{s}{\sqrt{n}}$. Using the above sampling distribution as a null distribution, we can approximate a Z-score for each value in the GMEP data matrix:

$$Z_{mj} = \frac{\text{GMEP}(m)_j - \mu_j}{s_j/\sqrt{n_m}} \qquad (2)$$

where n_m is the number of observations for motif m that went into the mean, and μ_j and s_j are the mean and standard deviation, respectively, of all relative expression measurements for microarray experiment j.

However the above assumption may not be valid because: (1) relative gene expression levels for individual genes may be correlated; (2) the distribution of w_{mg} is not uniform across the genome. In this case, the mean GMEP value is still expected to equal μ, but the standard deviation of the GMEP value will vary. We investigated the variability of the standard deviation from the predicted value of $\frac{s}{\sqrt{n}}$ by permutation tests. For each of 51 hexameric motifs, a distribution of randomized GMEP values was obtained from 5000 random permutations of the d_{gj} values for a single experiment. Since the mean and standard deviations for these permutation distributions varied by less than 5% from the values given by the CLT, we assume that equation (2) provides a good approximation to the Z-score.

Our GMEP software has an option of filtering entries in

the $(4^L \times c)$ data matrix D according to a user-determined threshold T. Values of D_{gj} that have Z-scores between $-T$ and $+T$ are deemed not significant and are set to zero.

RESULTS

We computed GMEP's for all motifs of length 5, 6, 7, or 8 nucleotides in length using an input gene expression dataset of 517 different DNA microarrays, each containing ~5300 yeast genes (overlapping genes and duplicated genes were not included in this analysis). This dataset comes from the Stanford Microarray Database (Sherlock *et al.*, 2001) and includes the published results of DeRisi *et al.*, 1997; Spellman *et al.*, 1998; Chu *et al.*, 1998; Gasch *et al.*, 2000; Ogawa *et al.*, 2000 and some unpublished results that will be described in a forthcoming publication.

Many asymmetric transcription factor binding sites confer similar regulation irrespective of their orientation relative to the target gene. If GMEP's reflect transcriptional regulation associated with a sequence motif, then we would expect GMEP's for many *bona fide* regulatory motifs to be highly correlated with the GMEP's of their reverse complements (note that in computing GMEP's we only use motifs found on the positive strand of adjacent non-coding DNA, so there is no *a priori* expectation that GMEP's of reverse complementary motifs should be correlated). We computed the correlation between all motifs and their reverse complements (excluding motifs that are self-reverse complements) using the GMEP matrixes described above, and compared the results to a negative control in which the associations between TCR-sequences and gene expression vectors were randomly permuted. Since this control maintained the same set of expression profiles and regulatory sequences (only the assignment of the expression profiles to each motif were permuted) effects due either to the expression patterns themselves or to the distribution of motifs in non-coding sequences would be found in both the real and permuted data.

Figure 1 shows the histograms of reverse-complement correlations for motifs of lengths 5, 6, 7, and 8. As expected, for all of these motif lengths, the distributions of reverse-complement correlations for the randomly permuted datasets resembled normal distributions with mean values close to zero. In contrast, when the correct associations of gene expression data were used, there was a striking shift in the distribution of complement correlations towards positive correlations, with a distribution mean ranging from 0.357 (motif length 5) to 0.080 (motif length 8). The positive correlations in the GMEP's between many of the motifs and their reverse complements support the assumption that many regulatory motifs encode information when present on either of the DNA strands and validates the biological relevance of GMEP's.

The distributions of reverse-complement correlations displayed motif-position dependence. We computed the distribution of reverse-complement correlation values for hexameric motifs found in 100 bp windows between −1000 and +1000 bp relative to the translation start site. Figure 2 shows the reverse-complement correlations associated with motifs found at different positions. The highest reverse-complement correlations occur for motifs found between −100 and −200 of the translation start site, while the reverse-complement correlation decays to near-background as the distance from the start site increases. This result agrees with other data on the positional distribution of transcriptionally active transcription factor binding sites (Wolfsberg *et al.*, 1999).

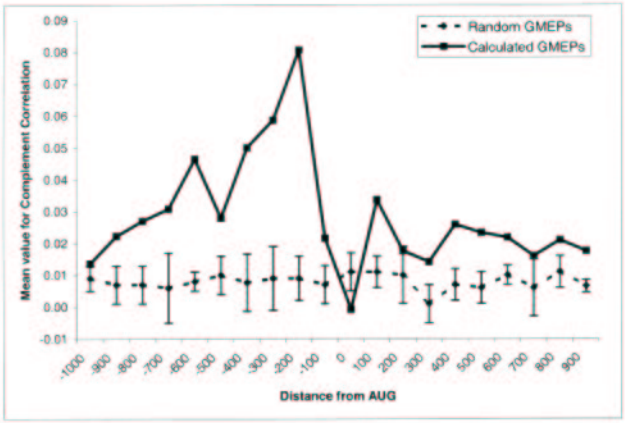

Fig. 2. Position-dependent effects of shifts in the complement correlation distribution. Mean values were calculated for reverse-complement correlation distributions in 100 bp windows at varying distances away from the translation start site. The dashed line represents the mean of mean values obtained from five different trials in which the association between each non-coding sequence and gene expression level was randomly permuted. The error bars indicate the standard deviation for these five trials. The solid line represents the mean values of the reverse-complement correlation distribution for actual data.

We chose to examine in more detail the data for all 4096 possible hexameric motifs. After calculating the GMEP associated with each motif, we then organized this data using hierarchical clustering (Eisen *et al.*, 1998). The logic of applying clustering to GMEP's was that motifs that encode similar regulatory information would display similar GMEP's and would thus be clustered together, and that motifs within a cluster might comprise different submotifs of a single consensus binding site. The clustered GMEP matrix for motifs of length six is available at http://rana.lbl.gov/.

We found that examining the clustered GMEP data in TreeView (Eisen *et al.*, 1998), provided an efficient way to visually identify clusters of motifs associated with

Table 1. Examples of sequences that regulate gene expression in many conditions. Consensus sequences were assembled from the individual motifs comprising clusters that were selected using the criteria described in the text. Cluster correlation refers to the Pearson correlation among normalized GMEP's for all motifs found in the cluster. The mean complement correlation refers to the mean value for the Pearson correlation between the GMEP of each motif found in the cluster with the GMEP of its reverse complement. The number of motifs in each cluster is indicated in parentheses.

Consensus Sequence R = [A/G]; S = [C/G];W = [A/T]	Transcription factor	Cluster correlation	Mean complement correlation	Characteristics of GMEP
TGAAAATTTT	RRPE[1]	0.974	0.977 ($n = 4$)	Generally repressed
AWTTTTCWTTT	RRPE[1]	0.963	0.809 ($n = 14$)	Generally repressed
SCACGTG	Pho4[2]	0.775	0.628 ($n = 6$)	Induced in Δpho80, Δpho85 mutants
TGASTCA	Gcn4[2]	0.751	0.698 ($n = 3$)	Induced during amino acid starvation
AGGGG	STRE	0.897	0.880 ($n = 26$)	Induced during stress
ARGGGAWA	STRE	0.840	0.794 ($n = 15$)	Induced during stress
CAG[C/A]GATGAG[C/A]T	Unknown[3]	0.834	0.880 ($n = 20$)	Repressed during stress
WCGCGW	MCB[4]	0.814	0.757 ($n = 11$)	Cell cycle periodicity
GATAAG	MET[2]	0.810	0.854 ($n = 2$)	Induced during amino acid starvation

[1]Hughes et al., 2000, [2]van Helden et al., 1998, [3]Gasch et al., 2000, [4]Spellman et al., 1998.

biologically interesting expression patterns. The software and processed GMEP data are available at http://rana.lbl.gov/. We used two stringent heuristic criteria for identifying "interesting" motif clusters: (1) the GMEP's within the clusters had correlations with each other of greater than 0.75; and (2) the motifs within each cluster were orientation-independent (*i.e.*, each cluster containing at least one reverse complement pair with a correlation greater than 0.7). Table 1 lists nine separate motif clusters that met these criteria. Each of these clusters contains previously-identified promoter motifs, including the MCB element bound by the MBF transcription factors, the STRE element recognized by the Msn2p and Msn4p transcription factors, and a site involved in environment stress response that has been previously identified but whose putative binding factor remains unknown. Figure 3 shows the GMEP clusters associated with these motifs. These GMEP profiles reflect conditions in which these transcription factors are known or believed to be active. Further analysis of numerous other apparently significant GMEP's is underway.

DISCUSSION

Genome-mean expression profiles represent one of several alternatives to "group-by-expression" approaches for analyzing gene expression data. Rather than look for statistical over-representation of sequences in a fixed subset of genes, these alternative methods introduce conceptual models that underlie microarray data. Holmes and Bruno (2000) have developed a likelihood framework to consider similarities in both sequences and gene expression profiles at the same time. The clustering of genes can thus by guided by choosing the most likely sequence-expression model that yields the observed gene sequences and gene expression levels. Bussemaker *et al.* (2001) use a regression method to fit gene expression data to a multivariate linear model. Significant motifs are defined to be those that yield the largest reduction in the χ^2 statistic.

The genome-mean expression profile introduced here is a simple and straightforward tool for assessing the information content of sequence motifs. The underlying model is a simple one. However, the observed correlation between reverse-complement pairs, the striking position-dependence of this correlation, and the success in identifying many known transcription factor binding sites strongly support continued analysis of the current data and the development of more sophisticated derivatives.

ACKNOWLEDGEMENTS

We thank Audrey Gasch for her advice and for a critical reading of this manuscript, as well as Terry Speed, Mark van der Laan and Ben Berman for helpful discussions. This work was supported by the Director, Office of Science, of the U.S. Department of Energy under contract no. DE-AC03-76SF00098. DYC was supported by a National Sciences and Engineering Research Council of Canada postgraduate scholarship. POB is an associate investigator of the Howard Hughes Medical Institute.

REFERENCES

Bailey, T. L., and Elkan, C. (1995). Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning,* **21,**51-80.

Bussemaker, H. J.; Li, H.; Siggia, E. D. (2001). Regulatory element detection using correlation with expression. *Nature Genetics,* **27,** 167-171.

Bussemaker, H. J.; Li, H.; Siggia, E. D. (2000). Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proceedings of the National Academy of Sciences of the U. S. A.,* **97,** 10096-10100.

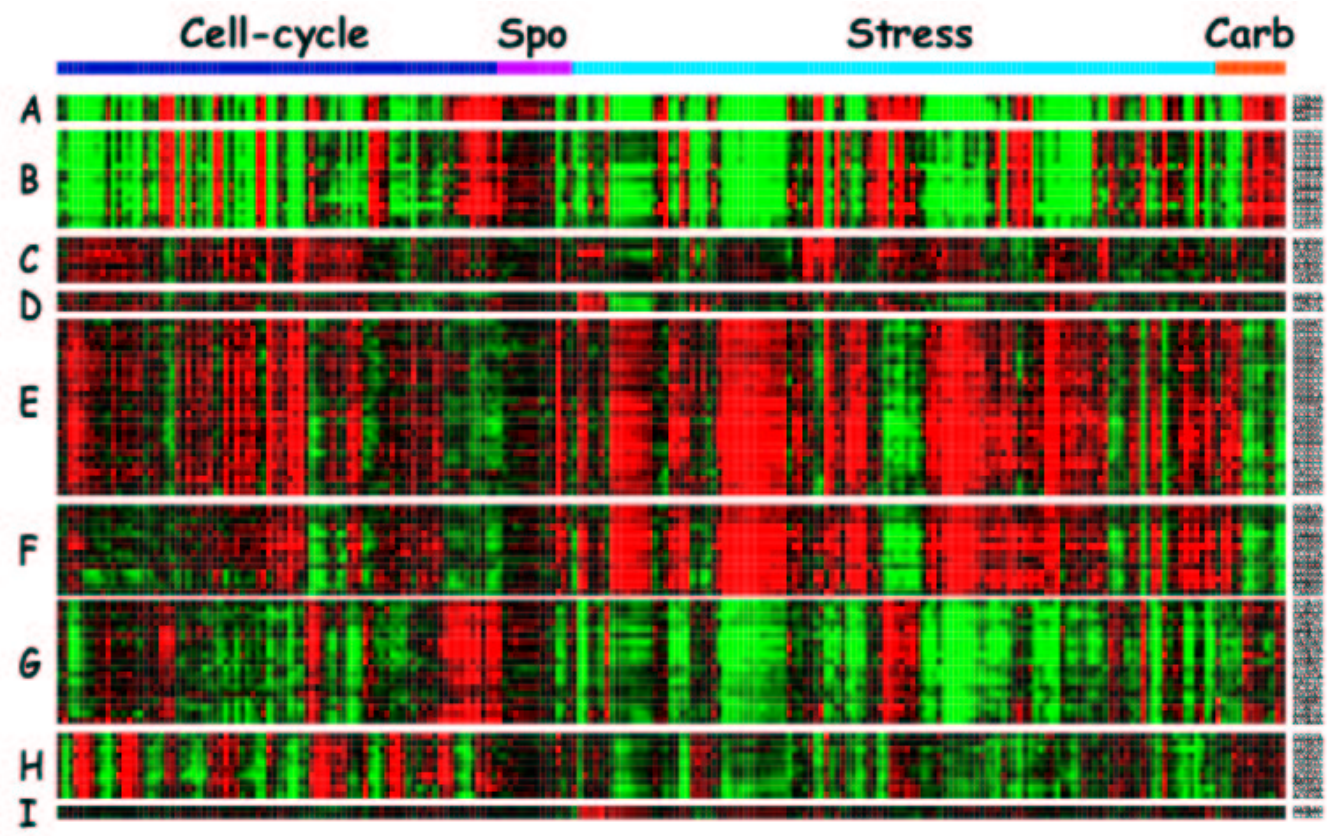

Fig. 3. Example GMEP Clusters. Each row represents the normalized GMEP for a single motif, calculated using equation (1). The GMEP's of all 4096 motifs were normalized for each column j by dividing GMEP$(m)_j$ by $\frac{s_j}{\sqrt{n_m}}$ and subtracting M_j, where n_m is the number of observations for motif m that went into the mean, s_j is the standard deviation of al relative expression measurements for column j, and M_j is the mean of all GMEP values for column j. Each column represents a single cDNA microarray experiment. The columns selected for display correspond to microarray experiments for the mitotic cell cycle conditions (Cell-cycle: Spellman et al., 1998; Zhu et al., 2000); sporulation conditions (Spo: Chu et al., 1998); environmental stress conditions (Stress: Gasch et al., 2000; Ogawa et al., 2000); and alternate carbon sources (Carb: Spellman PT, Brown PO and Botstein D, unpublished observations). Green pixels correspond to transcriptional repression, red pixels correspond to transcriptional induction, and the pixel intensity reflects the magnitude of the change in average gene expression. Nine clusters of GMEP's are displayed that meet our selection criteria as discussed in the text. Manual alignment of these motifs yields the following consensus sequences, which are also listed in Table 1: (A) TGAAAATTTT (RRPE); (B) AWTTTTCWTTT (RRPE-like); (C) SCACGTG (Pho4); (D) TGASTCA (Gcn4); (E) AGGGG (STRE); (F) ARGGGGAWA (STRE-like); (G) CAG[C/A]GATGAG[C/A]T (Repressed in stress - Gasch et al., 2000); (H) WCGCGW (MCB); (I) GATAAG (MET).

Cho, R.; Campbell, M.; Winzeler, E.; Steinmetz, L.; Conway, A.; Wodicka, L.; Gabrielian, A.; Landsman, D.; Lockhart, D.; and Davis, R. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell,* **2,** 65-73.

Chu, S.; DeRisi, J.; Eisen, M.; Mulholland, J.; Botstein, D.; Brown, P. O.; and Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science,* **282,** 699-705.

Eisen, M. B.; Spellman, P. T.; Brown, P. O.; and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the USA,* **95,** 14863-14868.

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

Hertz, G. Z., and Stormo, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics,* **15,** 563-577.

Holmes, I. and Bruno, W. J. (2000). Finding regulatory elements using joint likelihoods for sequence and expression profile data. *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology,* 202-210. San Diego, CA: AAAI Press.

Hughes, J. D.; Estep, P. W.; Tavazoie, S.; Church, G. M. (2000). Computational identification of cis-regulatory elements associ-

ated with groups of functionally related genes in Saccharomyces cerevisiae. *Journal of Molecular Biology,* **296,** 1205-1214.

Ogawa, N.; DeRisi, J; and Brown, P.O. (2000). New components of a system for phosphate accumulation and polyphosphate metabolism in Saccharomyces cerevisiae revealed by genomic expression analysis. *Molecular Biology of the Cell,* **11,** 4309-4321.

Ohler, U., and Niemann, H. (2001). Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends in Genetics,* **17,** 56-60.

Sherlock, G.; Hernandez-Boussard, T.; Kasarskis, A.; Binkley, G.; Matese, J.C.; Dwight, S.S.; Kaloper, M.; Weng, S.; Jin, H.; Ball, C.A.; Eisen, M.B.; Spellman, P.T.; Brown, P.O.; Botstein, D.; Cherry, J.M. (2001). The Stanford microarray database. *Nucleic Acids Research,* **29,** 152-155.

Sinha, S., and Tompa, M. (2000). A statistical method for finding transcription factor binding sites. *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology,* 344-354. San Diego, CA: AAAI Press.

Spellman, P.T.; Sherlock, G.; Zhang, M. Q.; Iyer, V. R.; Anders, K.; Eisen, M. B.; Brown, P. O.; Botstein, D.; and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell,* **9,** 3273-3297.

Tavazoie, S.; Hughes, J. D.; Campbell, M. J.; Cho, R. J.; and Church G. M. (1999) Systematic determination of genetic network architecture. *Nature Genetics, 22,* 281-285.

van Helden, J.; Andr, B.; and Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology,* **281,** 827-842.

Vilo, J.; Brazma, A.; Jonassen, I.; Robinson, A.; and Ukkonen, E. (2000). Mining for putative regulatory elements in the yeast genome using gene expression data. *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology,* 384-394. San Diego, CA: AAAI Press.

Wagner, A. (1999). Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics,* **15,** 776-784.

Wolfsberg, T. G.; Gabrielian, A. E.; Campbell, M. J.; Cho, R. J.; Spouge, J. L.; and Landsman, D. (1999). Candidate regulatory sequence elements for cell-cycle transcription in Saccharomyces cerevisiae. *Genome Research,* **8,** 775-792.

Zhu, G.; Spellman, P.T.; Volpe, T; Brown, P.O.; Botstein, D.; Davis, T.N.; and Futcher, B. (2000). Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. *Nature,* **406,** 90-94.