Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation

Frederick P. Roth[1-3†], Jason D. Hughes[1,2†], Preston W. Estep[2], and George M. Church[1,2*]

*[1]Harvard University Graduate Biophysics Program and [2]Harvard Medical School Department of Genetics, Boston, MA 02115. [3]Current address: Millennium Information, Cambridge, MA 02139. *Corresponding author (e-mail: church@salt2.med.harvard.edu). [†]These authors contributed equally to this work.*

Received 2 April 1998; accepted 10 August 1998

Whole-genome mRNA quantitation can be used to identify the genes that are most responsive to environmental or genotypic change. By searching for mutually similar DNA elements among the upstream noncoding DNA sequences of these genes, we can identify candidate regulatory motifs and corresponding candidate sets of coregulated genes. We have tested this strategy by applying it to three extensively studied regulatory systems in the yeast *Saccharomyces cerevisiae*: galactose response, heat shock, and mating type. Galactose-response data yielded the known binding site of Gal4, and six of nine genes known to be induced by galactose. Heat shock data yielded the cell-cycle activation motif, which is known to mediate cell-cycle dependent activation, and a set of genes coding for all four nucleosomal proteins. Mating type α and a data yielded all of the four relevant DNA motifs and most of the known a- and α-specific genes.

Keywords: functional genomics, gene expression

Complete DNA sequence is now known for more than 10 different organisms[1]. For even the most intensely studied of these organisms, a large fraction of genes is completely uncharacterized—about 40% and 50% for *Escherichia coli* and *Saccharomyces cerevisiae*, respectively[2,3]. Furthermore, annotation of noncoding regions has typically lagged behind discovery and prediction of gene function. Given that sequence elements in noncoding regions often control gene expression, and that knowing a gene's place in the larger regulatory network of a cell is essential to understanding its function, it is critical that we develop methods for rapidly characterizing noncoding regions.

A common approach to the discovery of regulatory elements entails the construction of a series of deletions or replacements in the upstream intergenic region of a gene, followed by an assay for altered regulation. An efficient method for predicting the most likely locations of regulatory sequences could guide these experiments more quickly to the sought-after elements. Given a set of genes "enriched" for coregulated members (obtained, for example, by genetic evidence), a search for conserved upstream sequence elements can predict the location of gene regulatory sequences[2].

Recently, it has become possible to measure the abundance of mRNA transcripts on a whole-genome scale[5-9]. By comparing transcript levels between different conditions or different strains we can find the set of genes whose transcript levels respond to a difference in environment or genotype. With this set of genes in hand, a number of questions naturally arise: Which of these changes in expression constitutes a primary response to an environmental change and which are indirect effects? Which are most critical for adaptation to a new condition? By what mechanisms are changes in transcript abundance achieved? What DNA or RNA sequence elements mediate the regulation of transcript abundance? It is the last question that we seek to address here.

Given a set of induced (or repressed) genes, one can search the regions upstream of translation start for short DNA sequence motifs, i.e., aligned sets of short, conserved DNA elements that are candidate DNA regulatory motifs. This search for regulatory motifs does not depend on prior information about gene regulatory mechanism and can be contrasted with approaches that search in upstream regions of induced (or repressed) genes for new examples of known DNA regulatory motifs[8].

Although they have been less extensively studied, conserved sequence elements in a gene's upstream region may also be determinants of mRNA stability[10,11] or even sites for regulation by antisense transcripts[12]. Regardless of mechanism, a highly conserved DNA sequence upstream of genes with similar expression responses is of interest. Similarly expressed genes that share a conserved upstream DNA motif constitute a candidate set of coregulated genes.

To test our strategy of combined expression analysis and upstream sequence alignment, we examined three extensively studied, transcriptionally regulated systems in the yeast *S. cerevisiae*: galactose utilization, heat shock response, and mating type regulation. To examine these three systems, we measured mRNA transcript abundance in *S. cerevisiae* on a whole-genome scale in each of four different cultures, which allowed three comparisons to be made: (1) growth on galactose vs. glucose, (2) strains of mating type α vs. mating type **a**, and (3) continuous growth at 30°C vs. 30°C growth followed by a 39°C heat shock. Expression was measured for each of these comparisons using photolithographically synthesized oligonucleotide microarrays ("chips")[7,9]. Change in transcript abundance for each open reading frame (ORF) was calculated for each comparison (Fig. 1).

Results and discussion

Examining upstream noncoding DNA sequence. We examined sets of upstream DNA sequences corresponding to (1) the top 10 ORFs as ranked by ratio of first-condition to second-condition abundance (e.g., ratio of galactose to glucose expression), (2) the top 10 ORFs as ranked by ratio of second-condition to first-condition abundance (e.g., ratio of glucose to galactose expression), and (3) the combination of the two preceding ORF sets. The rationale for examining the combined set of upstream regions is that a single regulatory motif

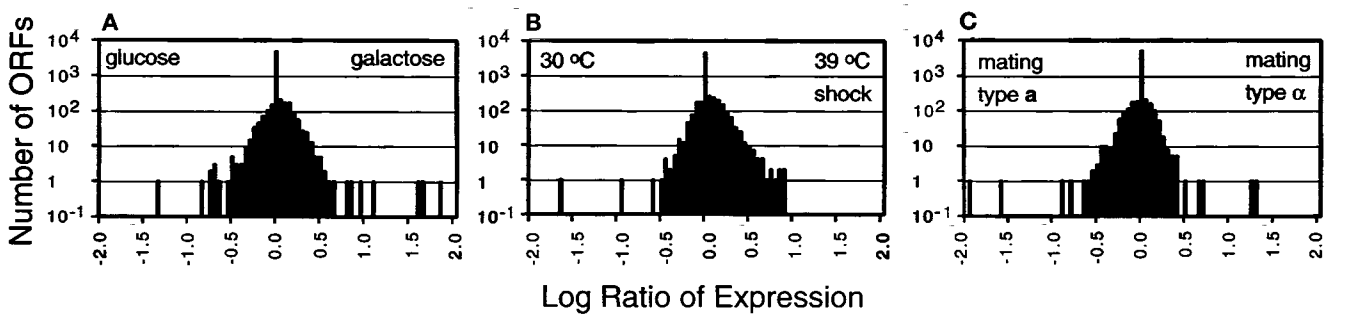

Figure 1. Histograms of change in expression level for each of three whole-genome expression comparisons. (A) Growth in galactose vs. glucose. (B) Growth after heat shock vs. 30°C. (C) Growth of mating type α vs. a. Transcripts with undetectable abundance in both conditions were assigned a log ratio value of 1.

may act as either a negative or a positive regulator depending on its sequence context. The upstream DNA sequence corresponding to each ORF was bounded at the 3′ (or downstream) end by the ORF's translation start. The 5′ end of the upstream region was bounded by the translation start or stop of the nearest upstream ORF, except that this boundary was never >600 DNA bp or <300 bp from translation start. The choice of upstream region boundaries is justified by an examination of those *S. cerevisiae* regulatory sites listed in the TRANSFAC database for which locations are given relative to translation start. Eighty-five percent (94 of 110) of these sites lie between 0 and 600 bases upstream of translation start.

Several algorithms for discovering recurring motifs in unaligned sequences have been developed. Those that are capable of automatically producing alignments containing multiple sites from a single input sequence include Gibbs Motif Sampling (GMS), CoreSearch, three CONSENSUS variants[13], and MEME[14]. While each method has its advantages and limitations with respect to our intended application, we chose GMS[15,16] to serve as our starting point for further development as we believed it to have the most flexible and exhaustive search methodology. There are several major distinctions between the application used here (called "AlignACE," for *Align*s *N*ucleic *A*cid *C*onserved *E*lements) and GMS, as implemented by Neuwald et al.[15] AlignACE has been optimized for finding multiple motifs (via an iterative masking procedure) and for alignment of DNA sequences (by automatic consideration of both strands). It also scores alignments by frequency of occurrence in the intergenic DNA sequence of a given genome.

AlignACE was applied to each of the sets of upstream DNA sequence described above. Of the many resulting DNA site alignments, we considered only those motifs that: (1) exceeded a threshold alignment score (a measure of "goodness" of sequence alignment), and (2) had an occurrence score (a measure of the fraction of ORFs in the *S. cerevisiae* genome with matching upstream sites) below 1%. The latter criterion requires that motifs be selective—that is, occur infrequently among upstream regions.

Those motifs that passed both alignment and occurrence score criteria were then compared with motifs one might have expected to find, with varying levels of confidence, given the relevant literature. For this purpose, we developed an objective measure of similarity between sequence motifs.

Galactose vs. glucose comparison. Using the set of upstream regions corresponding to the 10 ORFs with transcripts ranked most increased in galactose relative to glucose, we identified a motif, which we called gal-1, that matched the galactose upstream activation sequence (UAS_G) motif. UAS_G is known to regulate galactose-utilization genes via the Gal4/Gal80 activation complex[17]. No motif passing both alignment and occurrence criteria was obtained when upstream regions corresponding to the 10 ORFs with transcripts ranked most increased in glucose relative to galactose were used. Another UAS_G-like motif, gal-glu-1, was obtained when the preceding two upstream region sets (corresponding to a total of 20 ORFs) were combined (Table 1 and Fig. 2).

Table 1. DNA motifs found and expected.

Comparison	Found by AlignACE	Alignment score	% occurence	Similar motifs
galactose vs. glucose	gal-1	33.1	0.16	UAS_G
	gal-glu-1	24.9	0.20	UAS_G
heat shock vs. 30°C	39C-1	5.1	0.04	
	30C-1	40.1	0.26	CCA
	30C-2	5.5	0.44	-
	39C-30C-1	30.1	0.20	CCA
	39C-30C-2	8.5	0.10	
mating type α vs. mating type a	mtα-1	8.9	0.22	P Box
	mta-1	8.5	0.10	-
	mta-2	5.0	0.20	-
	mta-3	28.1	0.62	α2-binding
	mtα-mta-1	20.7	0.68	α2-binding
	mtα-mta-2	5.3	0.26	PRE
	mtα-mta-3	8.6	0.54	-
	mtα-mta-4	5.3	0.62	Q Box

Comparison	Expected motif	DNABP	Reference
galactose vs. glucose	UAS_G	Gal4p/Gal80p	40
	URS_G	Mig1p	40
	Rap1p-binding	Rap1p	40
	Gcr1p-binding	Gcr1p	40
heat shock vs. 30°C	HSE	HSF	42
	STRE	Msn2p/Msn4p	43, 44
	CCA	?/Hir1p/Hir2p	19
	NEG	?	20
	MCB	Mbp1p	21
	SCB	Swi4p/Swi6p	21
	ECB	Mcm1p	22
mating type α vs. mating type a	P Box	Mcm1p	23
	Q Box	Matα1p	23
	α2-binding	Matα2p	23
	PRE	Ste12p	23

Similar motifs: those expected motifs found to be similar by objective criteria; DNABP: the protein that binds an element where known.

Heat shock. Upstream regions of the 10 ORFs whose transcript levels increased the most in the heat-shocked culture relative to the 30°C culture yielded a single motif, 39C-1. Upstream regions of the 10 ORFs with transcripts ranked most increased in the 30°C relative to the heat-shocked culture yielded two motifs. The first of these, 30C-1, matched the cell cycle activation (CCA) motif, a known activator of histone genes[19,20]. The second motif, 30C-2, has not been described. When the combined set of upstream regions was used, another CCA-like motif—39C-30C-1—was identified along with 39C-30C-2,which was similar to 39C-1.

Mating type. When upstream regions of the 10 ORFs whose transcript levels increased the most in mating type α relative to

type a were examined, the motif mtα-1 was found. mtα-1 matched the P Box and the early cell-cycle box (ECB), as well as the Gcr1p-binding site and the heat shock element (HSE). ECB mediates M/G1-specific activation[21,22] and the P Box regulates mating type-specific genes[23]. The P Box and ECB both bind Mcm1p[25]. Using upstream regions of the 10 ORFs whose transcript levels increased the most in mating type **a**, relative to type α, three motifs emerged. The first two of these, mta-1 and mta-2 have not been described. The third motif, mta-3, matched the binding site of Matα2p[21].

When upstream regions from the combined set of 20 ORFs with most altered transcript abundance between mating type α and a were examined, four motifs emerged. The first of these, mtα-mta-1, matched the known Matα2p-binding site. The second, mtα-mta-2, matched the pheromone-response element (PRE)—the known binding site of Ste12p[23], an activator of mating type-specific genes. A weak second match between mtα-mta-2 to the PRE motif suggested a conserved spacing of 7 bp between PRE elements. The third motif, mtα-mta-3, bore some resemblance to the PRE motif, but had a similarity score below the threshold applied. The fourth motif, mtα-mta-4, corresponded to the Q Box element, which binds Matα1p and mediates activation of α-specific genes[23]. Motifs similar to the PRE and Q Box motifs were not found by examining only upstream regions of the combined set of 20 ORFs or either set of 10 ORFs. The PRE element confers inducibility by either α or a mating pheromone, and was previously known to be present upstream of both mating type α- and **a**-specific genes. The Q Box, on the other hand, was expected only upstream of α-specific genes so that finding a matching site upstream of *STE2* was unexpected. A closer examination of this *STE2* site revealed a T instead of the highly conserved A at the eighth position of the Q Box motif, so that the match was not a strong one.

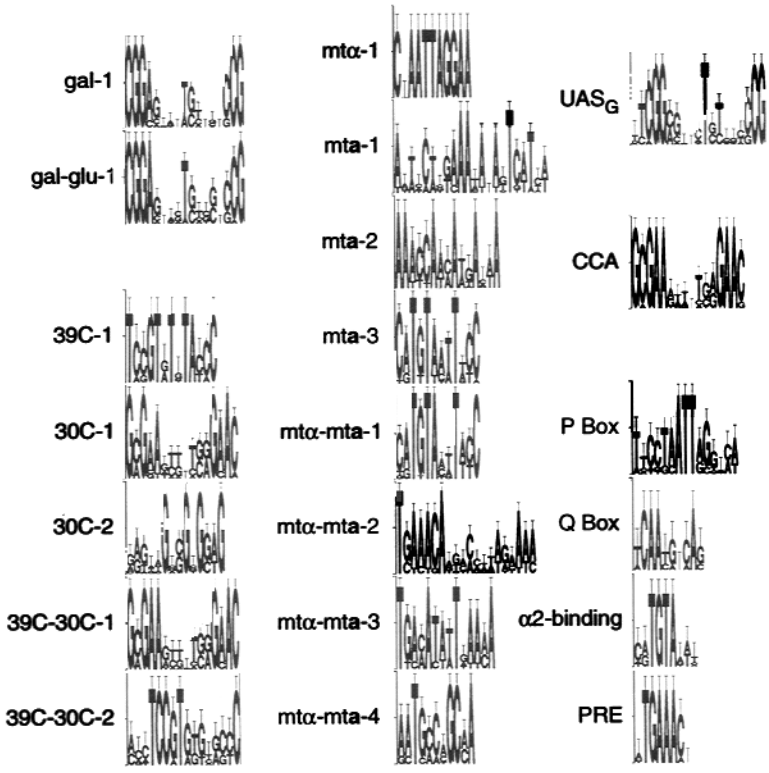

Figure 2. Sequence logos for DNA motifs found using AlignACE (first and second columns) and similar motifs that might have been expected a priori (third column). The height of each letter is proportional to its frequency, and the letters are sorted so the most common one is on top. The height of the entire stack signifies the information content of the sequences at that position, with information content at each position varying between 0 and 2 bits[18].

Comparison with other studies. Motifs common to many genes—for example, the TATA box—were neither found nor expected as these are excluded by the occurrence score constraint discussed above. In the case of transcripts less abundant in galactose than glucose, motifs corresponding to Rap1p- or Gcr1p-binding sites might be expected, but neither of these was found. Rap1p and Gcr1p are general transcription factors with diverse roles, including regulation of glycolytic enzymes and ribosomal proteins[24]. In the case of transcripts more abundant in galactose than glucose, we expected to find URS$_G$ (bound by Mig1p[24]), but did not. As expected, our procedure did find UAS$_G$, an essential regulatory element for galactose-utilization genes.

The HSE and stress response promoter elements (STREs), known to mediate heat shock response[25,26], were notably absent from the motifs found by AlignACE. Heat shock is known to have broad effects, including transient cell-cycle arrest in G1[27]. As a result, we might have expected to find genes with cell cycle–specific expression among genes affected by heat shock. Histone genes are strongly transcribed during S phase, and are regulated both by a negative regulatory sequence (NEG) and activated by the CCA motif[9,20]. Not found among heat shock data were the NEG motif, the Swi4/6p-dependent cell cycle box (SCB), the MluI cell cycle box (MCB) motifs (which regulate G1/S-specific transcription), or the ECB element[21,22]. AlignACE did find the CCA motif among the set of ORFs with transcripts decreased in heat shock, a set which contained several histone genes.

The α2 operator, P Box, PRE, and Q Box elements represent the complete set of DNA elements responsible for regulation of mating type-specific genes[23]. All four of these elements were found by AlignACE.

It is likely that the six motifs found by AlignACE that we did not identify with previously known motifs are false positives. Six motifs represent no more than the average number of false positives one might have expected, as determined below.

False-positive and false-negative motifs. The alignment and occurrence score criteria used here, 5 and 1%, respectively, were chosen permissively, so that few biologically relevant motifs would be excluded. To ensure that the alignment criterion was sufficiently permissive, we searched for conserved motifs among upstream sequences corresponding to 1000 randomly chosen sets of 10 ORFs. A maximum of three motifs was sought from each intergenic sequence set. Each randomly chosen sequence set returned at least one motif with a score greater than our threshold of five, and in 674 cases, all three motifs returned had scores greater than the permissive. An occurrence score of 1% seemed to be a permissive threshold, given our desire for motifs that do not match too frequently in upstream regions.

Because different users of this method will have varying tolerance for false-positive motifs, it is advantageous to know how the expected number of false-positive motifs depends on chosen alignment and occurrence score thresholds. We estimated the number of expected false positives by randomly generating 100 sets of 10 ORFs, obtaining the upstream DNA sequence corresponding to these ORFs, and plotting the expected number of false-positive motifs as a function of alignment and occurrence score thresholds (Fig. 3). For the permissive thresholds, an average of 1.7 motifs (number of false positives is Poisson-distributed with a coefficient of dispersion of 0.9) was obtained from randomly chosen upstream sequence sets. Eight motifs identified from 6 sets of 10 ORFs meet the permissive thresholds. Four of these match known regulatory motifs. The chance that a motif is a false positive decreases with increasing