doi:10.1006/jmbi.2000.3519 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2000) 296, 1205–1214

JMB

Computational Identification of *Cis*-regulatory Elements Associated with Groups of Functionally Related Genes in *Saccharomyces cerevisiae*

Jason D. Hughes[1,2], Preston W. Estep[1], Saeed Tavazoie[1] and George M. Church[1*]

[1]*Department of Genetics Harvard Medical School, 200 Longwood Ave, Boston MA 02115, USA*

[2]*Graduate Program in Biophysics, Harvard University, 200 Longwood Ave Boston, MA 02115, USA*

**Corresponding author*

AlignACE is a Gibbs sampling algorithm for identifying motifs that are over-represented in a set of DNA sequences. When used to search upstream of apparently coregulated genes, AlignACE finds motifs that often correspond to the DNA binding preferences of transcription factors. We previously used AlignACE to analyze whole genome mRNA expression data. Here, we present a more detailed study of its effectiveness as applied to a variety of groups of genes in the *Saccharomyces cerevisiae* genome. Published functional catalogs of genes and sets of genes grouped by common name provided 248 groups, resulting in 3311 motifs. In conjunction with this analysis, we present measures for gauging the tendency of a motif to target a given set of genes relative to all other genes in the genome and for gauging the degree to which a motif is preferentially located in a certain distance range upstream of translational start sites. We demonstrate improved methods for comparing and clustering sequence motifs. Many previously identified *cis*-regulatory elements were found. We also describe previously unidentified motifs, one of which has been verified by experiments in our laboratory. An extensive set of AlignACE runs on randomly selected sets of genes and on sets of genes whose upstream regions contain known transcription factor binding sites serve as controls.

Keywords: bioinformatics; computational biology; genomics; DNA regulatory motifs; yeast

Introduction

The recent increase in the number of sequenced genomes and the amount of genome-scale experimental data allows the use of computational techniques to investigate *cis*-acting sequences controlling transcriptional regulation. Some methods seek to find new sites for a given transcription factor based on a set of known sites, often by using online search engines where one may submit sequences to be scanned for known motifs (Heinemeyer *et al.*, 1998; Zhu & Zhang, 1999). Others, such as AlignACE, seek to find unknown DNA binding motifs for unspecified transcription factors by searching the regions upstream of the translational start sites of a set of potentially coregulated genes (Spellman *et al.*, 1998; van Helden *et al.*, 1998; Brazma *et al.*, 1998; Roth *et al.*, 1998).

AlignACE is based on a Gibbs sampling algorithm and returns a series of motifs that are over-represented in the input set. It previously has been used to find transcriptional regulatory DNA motifs in *Saccharomyces cerevisiae* using groups of genes derived from genome-wide mRNA expression data (Roth *et al.*, 1998; Tavazoie *et al.*, 1999). While many known *cis*-acting elements were identified, AlignACE returned many more motifs about which no literature information was found. A distinguishing feature of most of the known motifs was that their corresponding highest scoring genomic sites tended to be strongly selective for the upstream regions of the genes used to find them.

Abbreviations used: ORF, open reading frame; SGD, *Saccharomyces* Genome Database; ORE, oleate response element; RRPE, ribosomal RNA processing element; PAC box, polymerase A and C box; STRE, stress response element; ECB, early cell-cycle box.

E-mail address of the corresponding author: church@arep.med.harvard.edu

0022-2836/00/051205–10 $35.00/0

One might expect this to be always true, since each motif is itself composed of sites in those regions, but we found that the vast majority of the unknown motifs were not very selective in this way. Also, a subset of the known motifs seemed to be preferentially positioned relative to the start of translation.

Here, we describe statistics to measure these two motif properties, which we call group specificity and positional bias. Furthermore, we present results from the systematic application of AlignACE to a sample set of functional groups of genes in *S. cerevisiae*, as well as positive and negative control sets. These data sets allow us to calibrate AlignACE and the associated motif measures so that empirical significance thresholds for these statistics may be determined. Many known *cis*-regulatory elements, as well as novel motifs, are identified by this method.

Results

The input sets of genes

A total of 248 groups were examined, including 135 groups from the database at the Munich Information Center for Protein Sequences (Heinemeyer *et al*., 1998), 17 groups from the Yeast Protein Database (Hodges *et al*., 1999), and 96 groups based on common name root as listed in the table of open reading frames (ORFs) from the *Saccharomyces* Genome Database (SGD) (ftp://genome-ftp.stanford.edu/pub/yeast/SacchDB; Cherry *et al*., 1998). We considered only groups of six or more genes. The number of genes in each of these groups range from this minimum of six to as many as 707, with an average of 42 genes per group. Runs of AlignACE on the upstream regions of these groups of genes produced 3311 motifs.

Since diverse sources of data were used to generate these groups of genes, no single mechanism of control is expected to exert an influence over all of the members of any of the groups. It is therefore important that each motif may consist of any number of sites, including zero, in each upstream region submitted to AlignACE. Furthermore, motifs found upstream of only a fraction of the submitted genes may still be considered very significant according to the measures developed here. For example, the motif corresponding to binding by the Leu3p transcription factor was found from an AlignACE run on the upstream regions of 116 amino acid residue biosynthetic genes as an alignment of 19 sites upstream of only 17 of the 116 genes. Nevertheless, according to the statistics discussed below, it was one of the strongest motifs found.

Positive and negative controls were also performed. A total of 29 known transcription factors with experimentally validated binding sites were used to create a test set to see how often AlignACE finds expected sequence motifs. To determine the false positive rate, a set of 250 control AlignACE runs were done, 50 each with 20, 40, 60, 80 and 100 randomly selected ORFs. This distribution of group sizes was chosen to be comparable to the functional categories studied here and to span the range of sizes of most gene sets to be analyzed by this method in future applications.

Motif measures

To reduce the set of 3311 motifs under consideration, we devised two motif measures: one related to group specificity, the other to positional bias. The group specificity score gauges how well a given motif targets the upstream regions of the genes used to find it relative to the upstream regions of all of the genes in the genome. The positional bias score indicates the degree to which a motif tends to be preferentially positioned in a particular distance range upstream of the translational start (see Methods).

These measures are distinct from and supersede those used in the initial report using AlignACE (Roth *et al*., 1998). In that study, the two relevant measures were MAP score and a general specificity score, not to be confused with the new group specificity score. The MAP score measures the degree to which a motif is over-represented relative to the expectation for the random occurrence of such a motif in the sequence under consideration. It is the central score used by AlignACE to rate the different alignments it samples (see Methods). The main drawback to the MAP score is the fact that some motifs occurring ubiquitously in a genome (e.g. A-rich motifs in *S. cerevisiae*) are scored very highly, but are not likely to be relevant to the specific set of genes in question. The general specificity score was designed to give an indication of how frequently a motif occurs in the genome. Cutoffs based on this score were then used in an attempt to eliminate the ubiquitous motifs with high MAP scores. However, many real motifs occur frequently in the genome. In fact, the more important the motif is in terms of the number of genes it controls, the worse it scores by this measure.

The new measure, which we call group specificity, does not have this drawback. It serves as a powerful adjunct to the MAP score in that it takes into account the sequence of the entire genome and highlights those motifs that are found preferentially in association with the genes under consideration. Cutoffs based on group specificity serve to eliminate motifs that correspond to sequence features that are over-represented throughout the genome. It provides better balance between motifs with many genomic sites and motifs with fewer sites, since it is only a measure of the degree to which the distribution of sites is skewed toward the input set, the greater total number of sites is not as much of an advantage.

Motif clustering

Many examples of identical or very similar motifs were generated by AlignACE. This occurs when the same motif is found from AlignACE runs on overlapping or related groups of ORFs and also when multiple, similar examples of a very strong motif are returned from a single AlignACE run. The latter case is caused by the iterative masking procedure used to find multiple motifs (see Methods).

To automatically group very similar motifs together, we needed a computational measure for motif similarity. While many established tools exist for comparing one sequence with another sequence (Altschul *et al*., 1990) or one sequence with an alignment of many sequences (Berg & von Hippel, 1987), methods for comparing two sequence alignments in a way appropriate to the short DNA motifs here are considerably less well developed. We previously devised and used one algorithm for this purpose (Roth *et al*, 1998). A modified and simpler algorithm is used here, which we name CompareACE. A hierarchical clustering technique based on CompareACE was developed and used to group similar motifs (see Methods).

Highly group-specific motifs

In order to present only the strongest of the great number of motifs found, we chose a MAP score cutoff of 10.0, which reduced the set of motifs under consideration to 1234. While largely arbitrary, this threshold did not lead to the rejection of any of the best examples of known *cis*-regulatory elements. To focus on the most selective motifs, a cutoff of 10^{-10} for the group specificity score was chosen. A total of 54 highly specific motifs fulfilled both criteria and were grouped into 25 distinct motif clusters. Figure 1 lists a representative motif from each of the motif clusters along with its corresponding MAP, group specificity, and positional bias scores. If the motif has been identified with the binding preferences of a known transcription factor, that is also indicated. Otherwise, a short description is given of the group of genes upstream of which the motif was found.

Known motifs

Assignment of AlignACE motifs to known *cis*-regulatory elements from the literature is an ideal application for CompareACE. This algorithm was not used in this case, however, because databases of known transcription factor binding sites are still incomplete with respect to what is known in the literature. The main criterion used to identify an AlignACE motif as a known *cis*-regulatory element was that the AlignACE motif matched the literature consensus and was found upstream of an appropriate set of genes. For motifs with numerous annotated, well-defined binding sites, this criterion allowed us to easily make the assignment. In cases involving very few known sites, the criterion used was whether the top genomic sites for the AlignACE motif included a significant fraction of the sites verified in the literature. We were able to identify the following 16 known motifs from among the 25 highly specific motif clusters: Rap1p, Gcn4p, the heat shock element (HSE), the Cbf1p-Met4p-Met28p complex, the Hap2p-Hap3p-Hap4p complex, Lys14p, the MluI cell-cycle box (MCB), the stress response element (STRE), the Met21p-Met32p complex, Leu3p, Oaf1p, the carbon source responsive element (CSRE), Pho4p, Ste12p, and Pdr3p (Svetlov & Cooper, 1995; Warner, 1989; Lundin *et al*., 1994; Martinez-Pastor *et al*., 1996; Blaiseau *et al*., 1997; Fisher & Goding, 1992; Becker *et al*., 1998; McIntosh, 1993; Karpichev *et al*., 1997; Caspary *et al*., 1997; Baur *et al*., 1997; Delahodde *et al*., 1995). Known real motifs found with slightly lower group specificity scores include Aft1p, Ga14p, the early cell-cycle box (ECB), and the cell-cycle activation (CCA) (not shown in Figure 1) (Yamaguchi-Iwai *et al*., 1996; Lohr *et al*., 1995; McInerny *et al*., 1997; Freeman *et al*., 1992). The given names correspond either to the known transcription factor or to an acronym corresponding to the motif's function. Among those motif clusters that were not identified with known transcription factors, we were generally unable to find information to indicate a possible cellular function. In most assignments the motif found by AlignACE matched very closely the literature motif, but two exceptions are worth noting as they illustrate different interpretation issues with AlignACE motifs.

The first exception involves the assignment to the STRE of motif cluster S13, which contains motifs from carbohydrate utilization categories. The consensus binding site for the STRE is AGGGG (Martinez-Pastor *et al*., 1996). The motifs in this cluster are very G-rich, but include more columns of information than are in this simple consensus. This may indicate that the literature consensus has ignored information in the flanking regions of the motif, or it may indicate that AlignACE has chosen an alignment in which the motif has been overspecified.

Another difficult assignment was motif cluster S16, which was derived from a group of genes involved in peroxisomal organization. The motif identified by AlignACE is a superposition of a half-site of the oleate response element (ORE) and the multifunctional URS1 consensus CGGCGGC (Karpichev *et al*., 1997; Gailus-Durner *et al*., 1997). It therefore demonstrates two possible ways in which AlignACE can fail to find the appropriate motif. The ORE is defined in the literature as two near palindromic half-sites separated by a 17-bp or 18 bp spacer. The AlignACE alignment, however, only matched one half-site. AlignACE is designed to look for compact sites and so penalizes sites that are diffuse. This penalty is not so great as to preclude its finding the Ga14p binding site, a CGG inverted repeat separated by 11 bp. It is possible

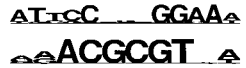
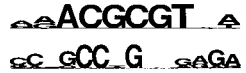
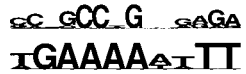
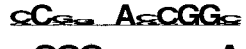
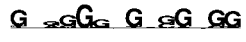
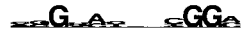
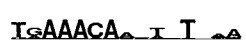
Cluster	MAP	Spec	PosBias	PrefPos	Logo	Notes
S1	123	8.6×10^{-48}	4.03×10^{-6}	279		Rap1
S2	128	1.27×10^{-32}	3.02×10^{-10}	125		Rpn4
S3	31.1	5.29×10^{-23}	0.0163	270		Gcn4
S4	55	8.73×10^{-21}	0.0015	169		HSE
S5	26.8	3.3×10^{-18}	0.00305	557		Mig1
S6	30.9	5.36×10^{-18}	0.00159	246		Cbf1
S7	12.4	9.98×10^{-14}	0.000546	241		RPS genes
S8	15.9	2.32×10^{-13}	0.0113	558		Hap2,3,4
S9	25.3	6.13×10^{-13}	1.49×10^{-5}	191		Lys14
S10	39.3	1.43×10^{-12}	4.85×10^{-8}	124		MCB
S11	11.1	1.44×10^{-12}	0.00527	168		RPL genes
S12	102	2.34×10^{-12}	2.04×10^{-43}	123		RRPE
S13	18.7	2.37×10^{-12}	0.0447	316		STRE
S14	20.6	2.6×10^{-12}	0.00213	171		Met31,32
S15	19.4	3.08×10^{-12}	0.00269	153		Leu3
S16	17.2	3.08×10^{-12}	0.0204	105		Oaf1
S17	14.1	4.17×10^{-12}	0.000374	201		mito. transp.
S18	21	1.14×10^{-11}	0.000413	344		stress
S19	24.6	1.57×10^{-11}	9.29×10^{-6}	308		CSRE
S20	16	2.07×10^{-11}	8.66×10^{-5}	388		TCA cycle
S21	21.2	2.2×10^{-11}	0.000205	172		cytoskel. transp.
S22	20.3	2.27×10^{-11}	0.00241	374		TCA cycle
S23	10.5	2.49×10^{-11}	0.0221	349		Pho4
S24	10.7	4.27×10^{-11}	0.000847	161		Ste12
S25	44.8	8.64×10^{-11}	0.00647	382		Pdr3

Figure 1. Motifs ranked by specificity score. For each cluster, the statistics for the motif member with the best (lowest) specificity score are listed. The second, third, and fourth columns correspond to the MAP, group specificity, and positional bias scores, respectively. The fifth column is the number of base-pairs upstream of the translational start site that the center of the most enriched 50 bp window is found (see Methods). The sixth column is a sequence logo representation of the motif (Schneider & Stephens, 1990). An algorithm for determining a unique orientation for each motif was developed and applied (see Methods). The last column lists the common name or the binding factor for the motif, if known. Otherwise a short description is given of the group of genes upstream of which the motif was found.

that in this case the 17 bp spacer incurred too great a penalty for the full site to be considered significant, that the variability in the length of the spacer prevented alignment of the full site, or that the sampling was not sufficiently extensive for even a strong motif to be found from among the large sample space of motifs with 17 bp spacers. The alignment was further complicated by the presence of a few sites that were perfect matches to the URS1 consensus, which is very similar to the ORE half-site. It is not known whether these are functional URS1 sites. Nevertheless, although this motif does not directly correspond to the binding preferences for any one transcription factor, and the aligned sites seem to match the binding preferences for either of two different factors, it is encouraging that the sites aligned by AlignACE largely correspond to functional control elements in the cases where the upstream regions in question have been studied.

Unknown motifs

Three highly specific motifs were found to be associated with ribosomal proteins. One of these, the Rap1p motif, is well known. The other two are motifs S7 and S11, which are primarily associated with small and large ribosomal subunits, respectively. These findings are especially interesting since the transcriptional regulation of a number of ribosomal proteins has been studied in detail, and the known Rap1p and Abf1p sites, along with a T-rich region, are generally found to be sufficient

to explain their transcriptional control (Warner, 1989; Goncalves *et al*., 1995).

The second motif listed in Figure 1 has an extremely well-conserved consensus and is very specific for genes coding for proteasome subunits. The motif also shows a great deal of positional bias, with the most significant enrichment occurring between approximately −50 and −200 bp relative to the translational start. The top 100 genes ranked by the strength of their best site in this upstream window include 31 proteasome subunits, five ubiquitin-related genes, five chaperonin genes, two mitochrondrial proteases, and three nuclear transport genes. The corresponding binding factor for this motif has recently been identified as Rpn4p (Mannhaupt *et al*., 1999). This result has been independently verified in our lab using a one-hybrid selection with confirmation by mRNA expression analysis of Rpn4p knockout and overexpressing strains.

Cluster S12 contains a motif that to our knowledge has not been noted in the literature. It is very specific for genes involved in rRNA processing, and it demonstrates strong positional bias, preferring sites between approximately −50 and −200 bp upstream of the translational start. We refer to this motif as RRPE, which stands for ribosomal RNA processing element.

Many other unknown motifs were found. See the web site (http://arep.med.harvard.edu/) for a complete current list of all known and unknown motifs found by AlignACE.

Positionally biased motifs

To focus on the most positionally biased motifs, a MAP score cutoff of 10.0 was again applied, followed by a positional bias score cutoff of 10^{-8}. The 448 motifs passing these criteria demonstrated great redundancy and were separated into only 17 distinct clusters (see Figure 2).

The vast majority of these motifs are homopolymeric A-rich sequences, which are commonly found despite the fact that AlignACE corrects for the 62 % A + T content of the yeast genome. These are generally the strongest motifs found in a search of any selection of upstream regions in yeast, and they demonstrate strong positional bias toward locations between about −50 and −150 relative to the translational start. The only known transcription factor that binds such sequences is Datin, which has been observed to act both as an activator and as a repressor (Moreira *et al*., 1998). Such sites have also been observed to exert transcriptional effects that are consistent with their sequence-specific structural properties (Iyer & Struhl, 1995).

Other motifs found here include AT-repeat and GT-repeat motifs, Rap1p, Reb1p, Abf1p, Rpn4p, the MCB element and two unknown motifs that were also ranked highly in terms of specificity (Svetlov & Cooper, 1995). These two distinct unknown motifs were found to be positionally biased and specific for rRNA and tRNA synthesis and processing. One is the RRPE motif discussed

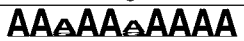
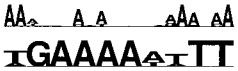
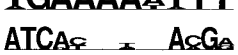
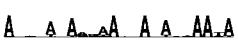
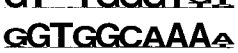
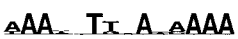
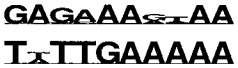
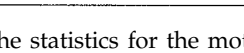
Cluster	MAP	Spec	PosBias	PrefPos	Logo	Notes
P1	637	0.000114	1.55×10^{-166}	140		
P2	15.6	0.0559	2.8×10^{-48}	114		AT repeat
P3	16.4	0.109	8.09×10^{-48}	133		
P4	102	2.34×10^{-12}	2.04×10^{-43}	123		RRPE
P5	14.3	0.025	2.29×10^{-41}	148		Abf1
P6	23.8	0.0906	1.58×10^{-35}	130		
P7	29.6	0.00888	2.79×10^{-33}	101		PAC
P8	17.6	0.00102	3.53×10^{-31}	148		Reb1
P9	10.2	0.0111	5.66×10^{-21}	35		GT repeat
P10	32.6	1.98×10^{-14}	5.66×10^{-21}	35		Rap1
P11	125	4.08×10^{-28}	1.13×10^{-14}	112		Rpn4
P12	12.5	0.00818	6.6×10^{-11}	123		
P13	13.3	7.5×10^{-6}	7.01×10^{-10}	249		
P14	19.2	0.028	7.45×10^{-10}	114		
P15	10.5	9.72×10^{-5}	5×10^{-9}	141		MCB
P16	11.8	0.000216	5×10^{-9}	37		
P17	20.2	0.00495	9.18×10^{-9}	127		

Figure 2. Motifs ranked by positional bias score. For each cluster, the statistics for the motif member with the best (lowest) positional bias score are listed. See the legend to Figure 1 for details.

above, and the other is a motif with consensus GATGAG that has been noted before. It was named the polymerase A and C box (PAC box) because of its association with polymerase A and C subunits (Dequard-Chablat *et al*., 1991), but as yet, neither a function nor a *trans*-acting factor for this motif has been identified.

Negative controls: motifs found from searches upstream of randomly chosen sets of ORFs

We ran AlignACE on 50 each of randomly chosen sets of 20, 40, 60, 80 and 100 ORFs. Varying MAP score and group specificity cutoffs were applied as in the functional categories, and motif clustering was performed (see Table 1 for a comparison of results). Runs of AlignACE on the upstream regions of genes in functional categories did result in higher-scoring motifs overall. There were a number of motifs from the randomly chosen sets of ORFs, however, that scored well within the range of some real motifs from the functional category runs. Inspection of the best of these motifs showed no indication that any of them might correspond to motifs noted in the literature.

If one considers these motifs to represent the background noise inherent in this method, then Table 1 may be used to choose cutoffs with prescribed false positive rates. Accordingly, four-fifths of the motifs listed in Figure 1 should correspond to a real signal above that background. In cases in which one searches upstream of genes that are known to be controlled by a common transcription factor, the false positive rates estimated in this manner are likely too high. Greater credence would be given to any motif specific to a coregulated group of genes as opposed to a motif specific to a randomly chosen group of genes.

The group specificity statistic can be modified to compare a motif's targets against a set of genes other than that used to find the motif. We refer to a statistic measuring the specificity of a motif for some different group of genes as cross-specificity. Using this measure, 82 motifs from the AlignACE runs on randomly selected groups of ORFs are found to have cross-specificities of less than 10^{-8} for one or more functional categories. This is despite the fact that no randomly chosen set of ORFs included more than five members of any of the smaller functional categories (those having less than 50 ORFs) or 10 % of the members of any of the larger functional categories (those having 50 or more ORFs). All of these motifs correspond to one of the following: Rap1p, Rpn4p, PAC box, RRPE and ECB. There were numerous matches to each of the first four motifs, but only one match to the ECB motif. By comparison, no motif found from any functional category had a cross-specificity of less than 10^{-5} to any of the randomly chosen ORF sets.

The most positionally-biased motifs found in the negative controls are very similar to those found analogously in the functional groups analysis (Figure 2). These include Abf1p, Reb1p, Rap1p, the PAC box and RRPE. The only novel motif was derived from amino acid repeats in the coding regions of two proteins found very nearby or overlapping one of the randomly chosen ORFs. Many of the motifs found by this method correspond to known real motifs, though no biological information beyond the genome sequence and predicted translation start sites was used to find them.

Positive controls: motifs from searches upstream of genes with known transcription factor binding sites

Groups of genes controlled by known transcription factors were used for the positive controls (Zhu & Zhang, 1999). Only 29 factors having five or more unique reported binding sites were considered. AlignACE was used as above to search for motifs upstream of the reportedly controlled genes, and the resulting alignments were checked for the presence of the sites cited in the literature. A motif was considered a match if it contained half or more of the literature sites in its alignment or if half or more of the aligned sites were cited in the literature. An alignment corresponding to the lit-

Table 1. Comparison between AlignACE runs on upstream regions of ORFs in functional categories and randomly chosen sets of ORFs

	Random groupings motifs		Functional category motifs	
Spec. score cutoff	MAP > 0	MAP > 10	MAP > 0	MAP > 10
1	3692(1063)	1792(205)	3311(1324)	1234(208)
10^{-1}	2766(1038)	1047(202)	2713(1284)	815(194)
10^{-2}	2026(978)	553(181)	2198(1201)	530(179)
10^{-3}	1416(831)	285(149)	1622(1016)	337(153)
10^{-4}	935(641)	151(104)	1109(753)	226(121)
10^{-5}	554(425)	72(56)	750(543)	160(90)
10^{-6}	329(290)	31(29)	446(329)	122(67)
10^{-7}	151(143)	15(15)	270(199)	91(47)
10^{-8}	60(59)	9(9)	164(118)	73(35)
10^{-9}	37(36)	6(6)	97(62)	60(28)
10^{-10}	14(14)	5(5)	69(38)	54(25)

Columns 2-5 list the numbers of motifs found from random groups and functional categories having group specificity scores less than the cutoff listed in the first column and MAP scores greater than that listed in the column headings. The number of independent motif clusters is listed in parentheses.

erature motif was found in 21 of the 29 test cases. Of the eight that were not found here, five were found in appropriate functional category runs. Therefore it is likely that many of the false negatives are the result of the limited number of true sites in the small input sequence sets. In any case, the false negative rate is no more than 30%, and with appropriate input data might be much lower.

Discussion

We present a set of analytical tools for the computational discovery and validation of *cis*-acting regulatory elements in a sequenced and annotated genome.

The group specificity score is a useful statistic for gauging whether a given motif is real in the sense that it describes a sequence feature that is functionally relevant for the genes under consideration. This measure is independent of the method being used to find motifs. It works as long as there is a method of ranking potentially regulated gene targets and could therefore serve as an independent measure by which to judge the performance of different motif-finding algorithms. Alternatively, different methods of grouping genes could be rated by the ability of those groupings to lead to the discovery of very self-specific motifs. The group specificity score also might serve as a new basis on which to design improved motif-finding algorithms.

The observation that some real motifs are preferentially located in certain distance ranges upstream of translational start sites is intriguing. The most positionally biased motifs tend to have sites centered around positions between −100 and −150 relative to translational start. Since the 5′ UTRs in *S. cerevisiae* are very short, this could indicate that a precise positioning relative to the transcriptional start site is necessary for the function of these motifs. Alternatively, since some of these motifs regulate the transcription of multiprotein complexes, one possible explanation for their precise positioning is that nearly identical modes of transcriptional induction and translational efficiency are required for the stoichiometric production of the protein subunits. The fact that there are many motifs that do not demonstrate this property implies that the reason for this positional bias, whatever it may be, is not a property of all transcription factor binding sites.

The method presented here is applicable to groups of genes other than functional categories. Possibilities include clusters of genes sharing common expression profiles across different conditions, sets of genes sharing a common phenotype, and genes coding for interacting proteins. With the advent of high-throughput technologies, in many cases it is becoming possible to obtain these types of information on a whole-genome basis with only one or a few experiments. Furthermore, although the *S. cerevisiae* genome with its compact upstream regions and independently transcribed genes seems ideal for the approach used here to find motifs, it may prove applicable to many other organisms. AlignACE has already proven useful in bacterial genomes (McGuire *et al*., unpublished results), though some distinct challenges will be encountered as it is applied to larger eukaryotic organisms. As new technologies generate great quantities of data concerning organisms about which little is known, the methods presented here, perhaps in the context of a more general model of genetic networks, could help to piece together much of the functionality of these organisms.

The analysis performed here will also form the starting point for a database of information about known and hypothetical sequence control features in *S. cerevisiae* and other organisms. Not only will researchers be able to use these tools to determine the most likely potential regulatory sequences for the genes they are studying, but they will be able to quickly determine whether the resulting hypothetical motifs are similar to any known or already suspected motifs.

Methods

AlignACE

AlignACE is an algorithm implemented in C++ for finding multiple motifs in any given set of DNA input sequences. We define a motif as the characteristic base-frequency patterns of the most information-rich columns of a set of aligned sites. AlignACE is based on a Gibbs sampling algorithm previously used to find motifs in protein sequences (Neuwald *et al*., 1995; Lawrence *et al*., 1993; Liu *et al*., 1995). It differs from this method in the following ways: (1) the motif model was changed so that the base frequencies for non-site sequence was fixed according to the source genome (62% A + T in the case of *S. cerevisiae*). (2) Both strands of the input sequence are simultaneously considered at each step of the algorithm. Overlapping sites are not allowed even if the sites are on opposite strands. (3) Simultaneous multiple motif searching was replaced by an approach in which single motifs were found and iteratively masked. The masking is done by determining the most information-rich column in each motif, mapping that column back to the input sequences, and placing a marker at each of those positions. The sampler is then reinitialized to find another motif with the stipulation that no sites that contain a masking marker may be resampled. Such sites may, however, be added to any found motif at the end of sampling so that the AlignACE output includes all relevant sites for each output motif. In the case of a very strong motif, it is possible for the motif to have one of its positions masked and yet still retain enough information in its other positions for a variant of the original motif to be found. We refer to these as mask variants. (4) The near-optimum sampling method used by AlignACE is different from that used by Neuwald *et al*. (1995). The MAP score is now the criterion on which the final output motif is based (see below).

AlignACE accepts input either as a FASTA-formatted sequence file, or as a list of ORF names, along with an SGD ORF table and a FASTA-formatted genome sequence. In the latter case, AlignACE will take sequence upstream of the translational starts of the listed ORFs for

motif searching. The translational start site is used as a proxy for the transcriptional start site, since the latter is difficult to determine computationally. The amount of sequence to be taken is specified by parameters such that at least a minimum amount of sequence is taken (default 300 bp) and as much as a maximum amount is taken (default 600 bp) so long as sequence belonging to other ORFs, transfer RNAs (tRNAs), small nuclear (snRNAs) and transposons are not included. In the case that some ORF overlaps an ORF of interest including part of its upstream region, the presence of that ORF is ignored. It is assumed that only one of such pairs of overlapping ORFs in *S. cerevisiae* is real.

Since a number of upstream regions in *S. cerevisiae* are nearly identical, AlignACE also includes the option to purge very similar input sequence before sampling. A Smith-Waterman algorithm (Smith & Waterman, 1981) is used to find such sets of repeated input sequences, all but one of which are then removed from consideration. The cutoffs used for this are such that at least 60% sequence identity is required for a sequence to be purged.

All results generated for this work used version 2.1 of AlignACE. The only non-default options used were $-y$ (automatic selection of upstream regions) and $-e$ (purging of input sequences based on Smith-Waterman comparisons).

MAP score

The MAP (maximum *a priori* log likelihood) score is used by AlignACE to judge different motifs sampled during the course of the algorithm. A crude, but useful approximation is given by the formula $N \log R$, where N is the number of aligned sites and R is the degree of over-representation of the motif in the input sequence. In other words, if a site matching a given motif is expected to occur once every kilobase according to background genomic mononucleotide frequencies, and ten sites are observed in 2 kb of input sequence, then $R = 5$. A detailed development of the formula is given by Liu *et al*. (1995).

The general properties of MAP score can be summarized by stating that all of the following lead to higher scores for otherwise similar motifs: (1) greater numbers of aligned sites; (2) more tightly conserved motifs; (3) less total input sequence; (4) more tightly packed information-rich positions; and (5) enrichment of the motif with nucleotides that are less prevalent in the genome.

ScanACE

ScanACE is a program written in C++ that searches a genome for close matches to a motif found by AlignACE. The scoring method used is identical with that used by AlignACE to sample sites. Specifically, the score S for a site Q whose sequence as a function of position is given by $q(p)$ is:

$$S(Q) = \sum_p M_{p,q(p)}$$

where the matrix M is calculated as:

$$M_{p,b} = \log \frac{F_{p,b} + p_b}{N + 1} - \log p_b$$

Here $F_{p,b}$ is the number of bases of type b aligned at position p, N is the number of aligned sites, and p_b is the genomic background nucleotide frequency for base b. The first term in the second equation above corresponds to the log of the frequency of a given base at a particular position in the motif alignment, estimated with a Bayesian prior distribution corresponding to the genomic mononucleotide frequencies and a total pseudocount of 1, as is the default for AlignACE.

ScanACE can be set to return all genomic sites scoring better than a cutoff based on the mean and standard deviation of the scores of the aligned sites, or it can return a given number of best sites. The positions of the sites are returned along with information concerning neighboring genomic features according to the tables of ORFs and other features it is given. This information may then be used to generate the necessary data for calculating group specificity and positional bias scores.

Group specificity

The group specificity score is a measure of how well a given motif targets the genes whose upstream regions were used to find it. For each motif, ScanACE output is used to rank all ORFs according to the strength of the site best matching the scoring matrix in each ORFs 5′ upstream non-coding region between −100 and −500 bp relative to translation start. The top 100 ORFs in this list are compared to the genes in the group used to find the motif. More than 100 ORFs are included in the target list if there are ties according to the ranking criteria. The probability that these sets would have the observed intersection or greater is calculated. This probability is what we refer to as the group specificity score. It is given by the formula:

$$S = \sum_{i=x}^{\min(s_1,s_2)} \frac{\binom{s_1}{i}\binom{N-s_1}{s_2-i}}{\binom{N}{s_2}}$$

where N is the total number of ORFs (6226 at the time of these calculations), s_1 and s_2 are the numbers of ORFs in the group used to find the motif and in the list of target genes, respectively, and x is the number of ORFs in the intersection of the two lists. Each term of this sum represents the probability of having obtained an intersection of i ORFs assuming a random sampling of the two sets of ORFs being compared. The sum S is then the probability of observing this intersection or greater. This is the statistic we use to quantify the degree to which a motif is specific to the ORFs from which it was found. If the assignment of sites to ORFs was not as straightforward or if it was believed that the occurrence of multiple sites for a given ORF was very significant, it would be possible to modify this statistic to instead consider specificity between the genomic sites in the input sequence set and the top target sites in the genome. This is the method used by McGuire *et al*. (unpublished results). Other variations are also possible.

Positional bias

A statistic to measure positional bias was constructed as follows. The 200 best sites in the genome for a given motif are found and their positions relative to the translational start sites of the nearest ORFs are extracted from

the ScanACE output. More than 200 genomic sites are considered if many equally good sites are tied in the ranking for the 200th best. Among these sites, t are found within 600 bp upstream of some ORF. The 50 bp window containing the greatest number, m, of these t sites is considered further. The probability of observing m or more sites out of a possible t in a 50 bp window of a 600 bp region is determined by the formula:

$$P = \sum_{i=m}^{t} \binom{t}{i} \left(\frac{w}{s}\right)^{i} \left(1 - \frac{w}{s}\right)^{t-i}$$

where $w = 50$ and $s = 600$. To make the expected distribution of randomly chosen sites as flat as possible, the presence of sites inside coding regions is ignored. That is, if a site is inside some ORF and yet is also 450 bp upstream of some other ORF, it is counted as occurring at −450 relative to a start site. The only deviations from this presumed flat background occur when a complete ORF is contained within the 600 bp upstream region of another ORF. This happens for 261 ORFs (4 %).

Since a sliding window of 50 bp is being considered, the expected distribution of scores for a randomly chosen site distribution is itself not flat. To determine what threshold score should be considered significant, sample distributions of sites were randomly generated. Out of 100 sets of 200 randomly selected sites in a 600 bp range, only two scored better than 10^{-3} and one better than 10^{-4} by this statistic. Over one-third of the motifs considered in this study passed a cutoff of 10^{-8}, indicating a very significant degree of positional bias.

CompareACE

To compare motifs, we chose a scoring method based on the Pearson correlation coefficient between the nucleotide base frequencies of two motif alignments (Pietrokovski, 1996). We decided to consider only alignments that contained at least the most informative six positions of each motif. This precludes the possibility of high scores resulting from alignments involving only the weak regions of motifs. The region of alignment is allowed to be as wide as necessary to accommodate these positions for each motif, but is made no wider. Positions of unknown sequence are modeled as being 25 % each A, C, G and T. The final score is the maximum value of the correlation coefficient over the space of all allowable alignments. This score varies between −1 and 1, and approaches 1 for a perfect match between motifs. We have named this algorithm CompareACE by analogy to the related tools ScanACE and AlignACE.

Motif clustering

The specific method used was the simple joining algorithm (Hartigan, 1975) in which comparisons between groups of motifs are done by averaging all of the CompareACE scores between relevant pairs of motifs. The purpose of the clustering in this case was not to highlight distant relationships, but rather to automatically group identical motifs. A cutoff score of 0.7 was used to define the cluster boundaries. The clusters were largely insensitive to cutoffs in the range of 0.6 to 0.9.

Orienting motifs

Since DNA sequences can be read in either of two ways, for consistency we designed a method of orienting motifs. The information-weighted nucleotide base content of the motif is calculated, with values I_A, I_C, I_G, and I_T. The function $1.5(I_G) + I_A - I_T - 1.5(I_C)$ is evaluated, and the motifs are orientated so that this value is positive. As a result, purine residues are preferentially displayed, and G + T-rich motifs are displayed instead of A + C-rich motifs.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.

Baur, M., Esch, R. K. & Errede, B. (1997). Cooperative binding interactions required for function of the Ty1 sterile response element. *Mol. Cell. Biol.* **17**, 4330-4337.

Becker, B., Feller, A., el Alami, M., Dubios, E. & Pierard, A. (1998). A nonameric core sequence is required upstream of the LYS genes of *Saccharomyces cerevisiae* for Lvsl4p-mediated activation and apparent repression by lysine. *Mol. Microbiol.* **29**, 151-163.

Berg, O. G. & von Hippel, P. H. (1987). Selection of DNA binding sites by regulatory proteins: statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* **193**, 723-750.

Blaiseau, P.-L., Isnard, A.-D., Surdin-Kerjan, Y. & Thomas, D. (1997). Met31p and Met32p, two related zinc-finger proteins, are involved in transcriptional regulation of yeast sulfur amino acid metabolism. *Mol. Cell. Biol.* **17**, 3640-3648.

Brazma, A., Jonassen, I., Vilo, J. & Ukkonen, E. (1998). Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.* **8**, 1202-1215.

Caspary, F., Hartig, A. & Schuller, H.-J. (1997). Constitutive and carbon source-responsive promoter elements are involved in the regulated expression of the *Saccharomyces cerevisiae* malate synthase gene MLS1. *Mol. Gen. Genet.* **255**, 619-627.

Delahodde., A., Delaveau, T. & Jacq, C. (1995). Positive auroregulation of the yeast transcription factor Pdr3p, which is involved in control of drug resistance. *Mol. Cell. Biol.* **15**, 4043-4051.

Dequard-Chablat, M., Riva, M., Carles, C. & Sentenac, A. (1991). RPC19, the gene for a subunit common to yeast RNA polymerases A (I) and C (III). *J. Biol. Chem.* **266**, 15300-15307.

Fisher, F. & Goding, C. R. (1992). Single amino acid substitutions alter helix-loop-helix protein specificity for bases flanking the core CANNTG motif. *EMBO J.* **11**, 4103-4109.

Freeman, K. B., Karns, L. R., Lutz, K. A. & Smith, M. M. (1992). Histone H3 transcription in *Saccharomyces cerevisiae* is controlled by multiple cell cycle activation sites and a constitutive negative regulatory element. *Mol. Cell. Biol.* **12**, 5455-5463.

Gailus-Durner, V., Chintamaneni, C., Wilson, R., Brill, S. J. & Vershon, A. K. (1997). Analysis of a meiosis-specific URS1 site: sequence requirements and involvement of replication protein a. *Mol. Cell. Biol.* **17**, 3536-3546.

Goncalves, P. M., Griffioen, G., Minnee, R., Bosma, M., Kraakman, L. S., Mager, W. H. & Planta, R. J. (1995). Transcription activation of yeast ribosomal protein genes requires additional elements apart

from binding sites for Abf1p and Rap1p. *Nucl. Acids Res.* **23**, 1475-1480.
Hartigan, J. A. (1975). *Clustering Algorithms*, John Wiley and Sons, Inc., New York.
Heinemeyer, T., Wingender, E., Reuter, I., Hermjakob, H., Kel, A. E., Kel, O. V., Ignatieva, E. V., Ananko, E. A., Podkolodnaya, O. A., Kolpakov, F. A. L. P. N. & Kolchanov, N. A. (1998). Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucl. Acids Res.* **26**, 364-370.
Hodges, P., McKee, A., Davis, B., Payne, W. & Garrels, J. (1999). Yeast protein database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucl. Acids Res.* **27**, 69-73.
Iyer, V. & Struhl, K. (1995). Poly(dA:dT), a ubiquitous promoter element that stimulates transcription *via* its intrinsic DNA structure. *EMBO J.* **14**, 2570-2579.
Karpichev, I. V., Luo, Y., Marians, R. C. & Small, G. M. (1997). A complex containing two transcription factors regulates peroxisome proliferation and the coordinate induction of *B*-oxidation enzymes in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **17**, 69-80.
Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. & Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science,* **262**, 208-214.
Liu, J. S., Neuwald, A. F. & Lawrence, C. E. (1995). Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.* **90**, 1156-1170.
Lohr, D., Venkov, P. & Zlatanova, J. (1995). Transcriptional regulation in the yeast GAL gene family: a complex genetic network. *FASEB J.* **9**, 777-787.
Lundin, M., Nehlin, J. O. & Ronne, H. (1994). Importance of a flanking AT-rich region in target site recognition by the GC box-binding zinc finger protein MIG1. *Mol. Cell. Biol.* **14**, 1979-1985.
Mannhaupt, G., Schnall, R., Karpov, V., Vetter, I. & Feldmann, H. (1999). Rpn4p acts as a transcription factor by binding to PACE, a nonamer box found upstream of 26 S proteasomal and other genes in yeast. *FEBS Letters,* **450**, 27-34.
Martinez-Pastor, M. T., Marchler, G., Schuller, C., Marchler-Bauer, A., Ruis, H. & Estruch, F. (1996). The *Saccharomyces cerevisiae* zinc-finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress-response element (STRE). *EMBO J.* **15**, 2227-2235.
McInerny, C. J., Partridge, J. F., Mikesell, G. E., Creemer, D. P. & Breeden, L. L. (1997). A novel Mcm1-dependent element in the SWI4, CLN3, CDC6, and CDC47 promoters activates M/G1-specific transcription. *Genes Dev.* **11**, 1277-1288.
McIntosh, E. M. (1993). MCB elements and the regulation of DNA replication genes in yeast. *Curr. Genet.* **24**, 185-192.
Moreira , J. M., Remacle, J. E., Kielland-Brandt, M. C. & Holmberg, S. (1998). Datin, a yeast poly(dA:dT)-binding protein, behaves as an activator of the wild-type ILV1 promoter and interacts synergistically with Reb1p. *Mol. Gen. Genet.* **258**, 95-103.
Neuwald, A. F., Liu, J. S. & Lawrence, C. E. (1995). Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.* **4**, 1618-1632.
Pietrokovski, S. (1996). Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucl. Acids Res.* **24**, 3836-3845.
Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. (1998). Finding DNA regulatory motifs within unaligned non-coding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnol.* **16**, 939-945.
Schneider, T. D. & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequence. *Nucl. Acids Res.* **18**, 6097-6100.
Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195-197.
Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Cell. Biol.* **9**, 3273-3297.
Svetlov, V. A. & Cooper, T. G. (1995). Review: compilation and characteristics of dedicated transcription factors in *Saccharomyces cerevisiae*. *Yeast,* **11**, 1439-1484.
Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. (1999). Systematic determination of genetic network architecture. *Nature Genet.* **22**, 1-5.
van Helden, J., Andre, B. & Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* **281**, 827-842.
Warner, J. (1989). Synthesis of ribosomes in *Saccharomyces cerevisiae*. *Microbiol. Rev.* **53**, 256-271.
Yamaguchi-Iwai, Y., Stearman, R., Dancis, A. & Klausner, R. D. (1996). Iron-regulated DNA binding by the AFT1 protein controls the iron regulon in yeast. *EMBO J.* **15**, 3377-3384.
Zhu, J. & Zhang, M. Q. (1999). SCPD: a promoter database of yeast *Saccharomyces cerevisiae*. *Bioinformatics,* **15**, 607-611.

Edited by F. E. Cohen

(Received 23 August 1999; received in revised form 4 January 2000; accepted 4 January 2000)