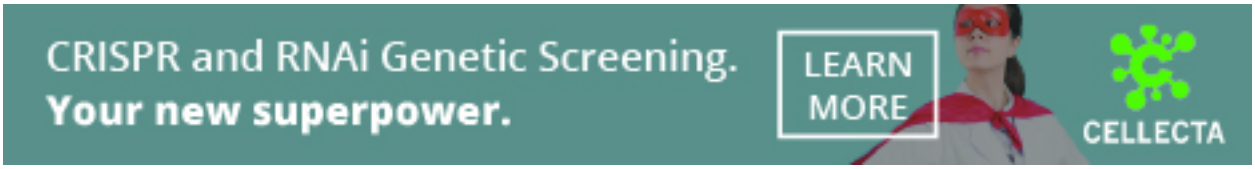
Candidate Regulatory Sequence Elements for Cell Cycle-Dependent Transcription in *Saccharomyces cerevisiae*

Tyra G. Wolfsberg, Andrei E. Gabrielian, Michael J. Campbell, et al.

Genome Res. 1999 9: 775-792
Access the most recent version at doi:10.1101/gr.9.8.775



Candidate Regulatory Sequence Elements for Cell Cycle-Dependent Transcription in *Saccharomyces cerevisiae*

Tyra G. Wolfsberg,[1,4] Andrei E. Gabrielian,[1,4] Michael J. Campbell,[2] Raymond J. Cho,[3] John L. Spouge,[1] and David Landsman[1,5]

[1]*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894 USA;* [2]*Molecular Applications Group, Palo Alto, California 94304 USA;* [3]*Department of Genetics, Stanford University School of Medicine, Stanford, California 94305 USA*

Recent developments in genome-wide transcript monitoring have led to a rapid accumulation of data from gene expression studies. Such projects highlight the need for methods to predict the molecular basis of transcriptional coregulation. A microarray project identified the 420 yeast transcripts whose synthesis displays cell cycle-dependent periodicity. We present here a statistical technique we developed to identify the sequence elements that may be responsible for this cell cycle regulation. Because most gene regulatory sites contain a short string of highly conserved nucleotides, any such strings that are involved in gene regulation will occur frequently in the upstream regions of the genes that they regulate, and rarely in the upstream regions of other genes. Our strategy therefore utilizes statistical procedures to identify short oligomers, five or six nucleotides in length, that are over-represented in upstream regions of genes whose expression peaks at the same phase of the cell cycle. We report, with a high level of confidence, that 9 hexamers and 12 pentamers are over-represented in the upstream regions of genes whose expression peaks at the early G_1, late G_1, S, G_2, or M phase of the cell cycle. Some of these sequence elements show a preference for a particular orientation, and others, through a separate statistical test, for a particular position upstream of the ATG start codon. The finding that the majority of the statistically significant sequence elements are located in late G_1 upstream regions correlates with other experiments that identified the late G_1/early S boundary as a vital cell cycle control point. Our results highlight the importance of MCB, an element implicated previously in late G_1/early S gene regulation, as most of the late G_1 oligomers contain the MCB sequence or variations thereof. It is striking that most MCB-like sequences localize to a specific region upstream of the ATG start codon. Additional sequences that we have identified may be important for regulation at other phases of the cell cycle.

[A companion website to this manuscript is available from http://www.ncbi.nlm.nih.gov/CBBresearch/Landsman/Cell_cycle_data]

The recent surge in the availability of complete genome sequences, as well as the development of technologies such as DNA microarrays, is ushering in a new era in the analysis of gene regulation. The yeast *Saccharomyces cerevisiae*, whose genome is the first to be fully sequenced from a eukaryote, is ideal for these studies. A number of groups have begun to merge the previously gathered wealth of biochemical information about yeast with its newly published complete genome sequence. One way to analyze gene expression is to scan the genome for the consensus-binding sequence of a known transcription factor to predict additional genes that may be regulated by that factor. This method has been used to look for Gcn4p-binding sites (Schuldiner et al. 1998) and stress-response elements (Moskvina et al. 1998; Treger et al. 1998). Another strategy focuses on the identification of upstream gene regulatory sites in groups of coregulated genes. Van Helden et al. (1998) and Brazma et al. (1998) looked at groups of coregulated genes to find over-represented oligonucleotide sequences. Both groups detected new candidate regulatory sites, as well as sites that had already been characterized. Brazma et al. (1998) also identified sequence patterns that occur more frequently in promoter regions than in other regions of the genome.

The availability of the complete yeast genome sequence has also facilitated whole genome expression analyses in which the expression levels of the ~6200 yeast genes are assayed in different conditions. Recent studies have identified the genes that are expressed during growth on rich and minimal medium (Wodicka et al. 1997), during the diauxic shift from anaerobic to aerobic metabolism (DeRisi et al. 1997), and when the transcriptional corepressor Tup1p is deleted or the

[4]*These authors contributed equally to this work.*
[5]Corresponding author.
E-MAIL landsman@ncbi.nlm.nih.gov; FAX (301)435-7794.

transcriptional activator Yap1p is overexpressed (DeRisi et al. 1997). The role of key components of the transcriptional machinery on the population of yeast genes was assessed by Holstege et al. (1998). Chu et al. (1998) assayed genes induced during sporulation in yeast. Roth et al. (1998) measured transcript abundance in the galactose response, heat shock, and mating type regulation systems, and took the further step of using a modified Gibbs sampling strategy to identify sequences that may be involved in the gene regulation. Cho et al. (1998) identified 420 genes whose mRNA expression exhibits cell cycle periodicity. They classified the genes by whether their expression peaked in the early G_1, late G_1, S, G_2, or M phase of the cell cycle. This analysis paves the way for a new type of experiment—a computational prediction of the sequence elements involved in cell cycle regulation. A preliminary analysis of these elements was described previously (Cho et al. 1998). In this manuscript, we present a more detailed interpretation.

We have developed a statistical technique to predict short sequence elements (i.e., oligomers) that may be involved in the expression of groups of coregulated genes. Our strategy is to look for pentamers and hexamers that are over-represented among the upstream regions of genes whose expression peaks at a particular phase of the cell cycle. We have identified 9 hexamers and 12 pentamers that may play a role in cell cycle regulation. Some of these oligomers may function in an orientation-dependent manner; the function of others may be dependent on their position within the promoter. The highest scoring hexamer to come out of this study, by all criteria, is the previously characterized sequence element MCB, which is known to control the expression of genes expressed in late G_1 (McIntosh 1993). Thus, this method is able to select biologically relevant sequence elements, and we predict that the other elements we identified also play roles in cell cycle regulation.

A companion website to this manuscript is available from http://www.ncbi.nlm.nih.gov/CBBresearch/Landsman/Cell_cycle_data. This site includes the upstream sequence of each of the cell cycle-regulated yeast genes, electronic versions of Tables 1–4 with links from the oligomer to the names of the genes that contain the oligomer, and the 50–100 top scoring oligomers for each analysis, including some sequences that are not statistically significant.

RESULTS

Strategy for Finding Novel Candidate Regulatory Elements

Our goal, in general, is to identify novel regulatory elements in upstream regions of coexpressed yeast genes. In this specific case, we apply our method to genes that may be involved in the cell cycle-dependent regulation of transcription. Cho et al. (1998) have identified those genes whose transcription exhibits cell cycle-dependent periodicity, and have furthermore classified the genes into five sets, those expressed during the early G_1 (63 genes), late G_1 (134 genes), S (74 genes), G_2 (56 genes), and M (56 genes) phases of the cell cycle. Our hypothesis is that sequence elements that are found more frequently (i.e., over-represented) in the upstream regions of genes expressed during one phase of the cell cycle, as compared with genes expressed during other phases of the cell cycle, may play a role in gene expression during that phase. As many transcription factors bind to short, highly conserved stretches of DNA, our analysis centers on short oligomers of length five or six, pentamers or hexamers. We limited our search to the sequence 600 nucleotides upstream of the translation start site of each gene, as most yeast regulatory elements are found within this region (Struhl 1995). Many yeast regulatory elements are analogous to mammalian enhancer sequences, and function in both orientations and at variable distances upstream of the transcription start site (Struhl 1995; Kunzler et al. 1996). Thus, we searched for oligomers whose representation is statistically significant, independent of their position and orientation. However, in higher eukaryotes, some regulatory elements act only when placed in certain locations or orientations with respect to the transcription start site (see, for example, Godambe et al. 1995; Nolan et al. 1996; Pfaff and Taylor 1998). To cover all biologically relevant possibilities, we also searched for potential orientation- and position-dependent elements, whose distribution is statistically significant on the basis of the strand or location in which the element is found.

Position-Independent Elements

The first analysis was to identify candidate elements important for cell cycle regulation in a position-independent manner. The statistical procedure is illustrated in Figure 1 with two oligomers, ACGCGT and GATGTA. Details are presented in Methods. The height of the white bar in each graph indicates the number of genes in each of the five cell cycle-regulated data sets. The number of genes whose upstream regions contain one or more copy of the sequence element is shown in purple. Pink squares indicate the number of upstream regions in that data set that would be expected to contain the sequence element, if the element were distributed evenly among the upstream regions of all of the cell cycle-regulated genes. Figure 1A illustrates that there are more ACGCGT elements in late G_1 than are expected, and fewer in G_2 and M. This observation is confirmed by the score for each data set (light blue), which is close to 0 in early G_1 and S, in which the observed number of elements is similar to the expected

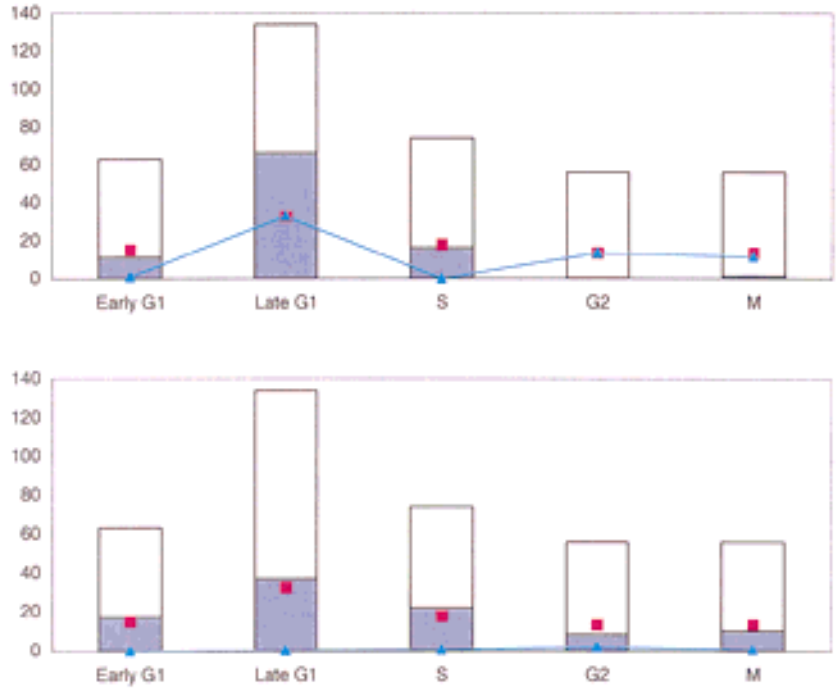


Figure 1 Test for hexamers over-represented in a position-independent manner. (*A*) Over-represented in late G_1: ACGCGT. The height of the white bar represents the number of genes in each of the five cell cycle-regulated data sets. The purple shading indicates the number of upstream regions in that data set that contain one or more copy of the oligomer. The number of upstream regions expected to contain that sequence element, if the element were evenly distributed among all five data sets, is marked with the pink box. The blue line indicates the contribution that each data set makes to the χ^2 score. (*B*) Not over-represented in any phase: GATGTA. Ninety-four cell cycle yeast upstream regions contain one or more copy of the sequence ACGCGT, and a different set of 94 contain one or more copy of the element GATGTA. However, these elements are distributed differently among the five data sets. The element ACGCGT, which has a χ^2 score of 60.4, is over-represented in late G_1—the observed number of elements (purple) is greater than the expected number (pink). It is also under-represented in G_2 and M, as the observed number is less than the expected number. The score in those three phases (blue) is thus higher than it is in early G_1 and S. The element GATGTA, which has a χ^2 score of 4.9, is not significantly over- or under-represented in any data set.

number, and higher in late G_1, G_2, and M, in which the observed and expected numbers differ. Conversely, the observed number of GATGTA elements in each of the five data sets is similar to the expected number, so the score for each data set is close to 0 (Fig. 1B). Therefore, this hexamer is not over- or under-represented in any of the five phases.

In Table 1 we present a list of predicted position-independent regulatory elements. The *P* value associated with each element provides an estimate of the probability with which a score of that magnitude occurs by chance alone. Although we list all elements with $P \leq 0.2$, (i.e., those whose score has a ≤20% chance of occurring by chance), we are most confident in those with $P \leq 0.05$. Twelve elements that could function on either strand, assayed by the presence of a hexamer and/or its reverse complement on one strand, have $P \leq 0.2$ (Table 1A). The distribution of an additional 10 hexamers on one strand also has $P \leq 0.2$ (Table 1B). However, only one of these hexamers, CCCTTT, represents a new element; the other nine were also identified in the search for hexamers and/or their reverse complements.

The statistical significance is made on the basis of the sum of the scores across all five data sets, and provides a measure of whether the hexamer is randomly distributed among the five sets of upstream regions. However, it does not identify in which data set the hexamer is over- or under-represented. For each sequence in Table 1, we present the number of upstream regions in each phase in which it occurs, as well as the contribution that the data set makes to the final score. We have also included the number of upstream regions of non-cell cycle-regulated genes in which each oligomer occurs. This number is supplied for reference purposes only; it was not included in any calculations. We classify a hexamer as being over-represented in a particular phase on the basis of the data set that provides the highest contribution to the final score (bold in Table 1). For example, the element defined by the hexamer CCCCGC and its reverse complement GCGGGG, has a total score of 31.37 (Table 1). Its over-representation in early G_1 contributes 25.60 units to the score. Some scores are also inflated by under-representation. For example, the sequence ACGCGT has a total score of 60.44. Late G_1, in which it is over-represented, contributes 33.34 units to the score, whereas G_2 and M, in which it is under-represented, contribute 13.74 and 11.82 units to the score, respectively. The distribution of some hexamers suggests that they may be important in more than one phase. For example, the M and early G_1 upstream regions contribute equally to the final score of the hexamer CCCTTT.

This method provides us with nine oligomers in which we have high confidence ($P \leq 0.05$), one in early G_1, seven from late G_1, and one from G_2. With lower confidence ($P \leq 0.2$) we present an additional four oligomers, one in early G_1, two in S, and one in both early G_1 and M. In late G_1, the highest scoring hexamer is the well-characterized palindrome ACGCGT, also known as the *Mlu*I cell cycle box (MCB),

Table 1. Position-Independent Hexamers (Over-Represented Hexamers)

A. Hexamers and/or reverse complements

		Early G1		Late G1		S		G2		M				Non-cell cycle regulated
		No.	χ^2	No.	χ^2	No.	χ^2	No.	χ^2	No.	χ^2	Total		No.
		63	score	134	score	74	score	56	score	56	score	χ^2 score	p-value ≤	5553
Early G1	CCCCGC/GCGGGG	14	**25.60**	6	0.68	2	1.50	1	1.79	1	1.79	31.37	0.005	378
	ACGCGG/CCGCGT	11	**13.77**	4	2.04	6	0.54	0	3.36	2	0.55	20.26	0.200	342
Late G1	ACGCGT/ACGCGT	11	1.29	66	**33.34**	16	0.26	0	13.74	1	11.82	60.44	0.001	270
	AACGCG/CGCGTT	14	1.13	72	**26.66**	22	0.00	3	11.07	2	12.76	51.62	0.001	641
	CGCGTC/GACGCG	12	0.71	63	**28.52**	13	1.37	5	5.44	0	13.60	49.64	0.001	380
	ACGCGA/TCGCGT	8	4.59	65	**24.08**	22	0.27	6	5.33	1	12.98	47.25	0.001	516
	CGCGTA/TACGCG	7	2.65	46	**12.83**	21	2.33	1	9.49	3	6.19	33.50	0.001	416
	CGCGAA/TTCGCG	9	3.26	58	**15.76**	20	0.04	8	2.90	4	7.58	29.53	0.005	654
	CGACGC/GCGTCG	7	0.28	33	**12.05**	11	0.09	1	5.73	0	7.60	25.76	0.050	421
S	CCGTGC/GCACGG	2	2.89	4	6.50	15	**7.99**	7	0.38	10	3.55	21.31	0.150	486
	AGTCAG/CTGACT	4	1.80	15	0.13	19	**10.83**	9	0.66	0	6.87	20.29	0.200	816
G2	GAGTCA/TGACTC	3	1.95	5	5.78	12	2.36	15	**14.32**	5	0.12	24.53	0.050	895

B. Hexamers

		Early G1		Late G1		S		G2		M				Non-cell cycle regulated
		No.	χ^2	No.	χ^2	No.	χ^2	No.	χ^2	No.	χ^2	Total		No.
		63	score	134	score	74	score	56	score	56	score	χ^2 score	p-value ≤	5553
Early G1	CCCCGC	10	**18.56**	5	0.15	1	1.59	0	2.49	1	0.89	23.67	0.200	201
Late G1	ACGCGT	11	1.29	66	**33.34**	16	0.26	0	13.74	1	11.82	60.44	0.001	270
	AACGCG	9	0.61	48	**21.59**	11	0.54	2	6.77	1	8.48	37.99	0.005	362
	CGCGTT	6	2.17	42	**15.48**	16	0.83	1	7.75	1	7.75	33.99	0.005	352
	GACGCG	7	0.75	41	**20.08**	7	1.70	4	2.48	0	8.63	33.64	0.005	207
	ACGCGA	3	4.19	38	**17.29**	10	0.06	5	1.24	0	8.19	30.97	0.010	256
	CGCGTC	9	0.03	38	**15.45**	9	0.43	2	4.95	0	8.48	29.35	0.050	197
	CGCGAA	4	3.63	38	**13.00**	12	0.00	7	0.41	0	8.92	25.97	0.100	348
	TCGCGT	5	1.48	32	**10.48**	13	0.87	1	5.73	1	5.73	24.29	0.200	277
Early G1, M	CCCTTT	23	**8.06**	13	7.48	10	1.71	11	0.01	21	**8.07**	25.34	0.150	1180

The number of upstream regions in each of the five sets of cell cycle-regulated genes containing one or more copies of a hexamer was counted. The counts are indicated in the column labeled "No." The contribution to the χ^2 score made by each of the data sets is shown, along with the total value of the χ^2 score. The data set(s) that make the greatest contribution to χ^2 are indicated in boldface type. *P* values for each χ^2 score were calculated from a Monte Carlo simulation. The final column shows the number of upstream regions of the non-cell cycle-regulated genes containing one or more copies of each hexamer. These counts are shown for illustrative purposes only; the numbers were not included in the χ^2 calculation.

which is involved in the transcriptional regulation of genes at the late G_1/S boundary (McIntosh 1993). It is notable that all of the hexamers that are over-represented in late G_1 contain some variation of the sequence ACGCGT. Thus, the seven late G_1 hexamers listed in Table 1 may not be independent, but rather, variants or parts of the same regulatory site.

Position-Dependent Elements I: Elements Over-Represented in Defined Intervals

Preliminary analysis of the top scoring hexamer from the previous analysis, ACGCGT, the MCB element, showed that most copies of this sequence reside within the interval 100–200 nucleotides upstream of the ATG start codon. Thus, our next step was to calculate whether any other oligomers are over-represented in a particular interval within one of the data sets. For statistical reasons discussed in Methods, we were unable to search for hexamers in this analysis; we instead turned to pentamers. The method is similar to that which we used to identify position-independent elements, except that in this experiment, we identified sequences that are over-represented in a single 99-nucleotide interval within a data set, not in the data set as a whole. Figure 2 illustrates the method, with the pentamer ACGCG as an example. The number of genes in each data set is indicated by the height of the white bar. Purple shading denotes the number of genes in each data set whose upstream region contains one or more copy of the sequence element in the 99-nucleotide interval. Pink squares indicate the number of upstream regions in that interval that would be expected to contain the sequence element, if the element were distributed evenly among all intervals in all data sets. The pentamer ACGCG occurs more frequently than expected in the interval −104 to −202 of late G_1,

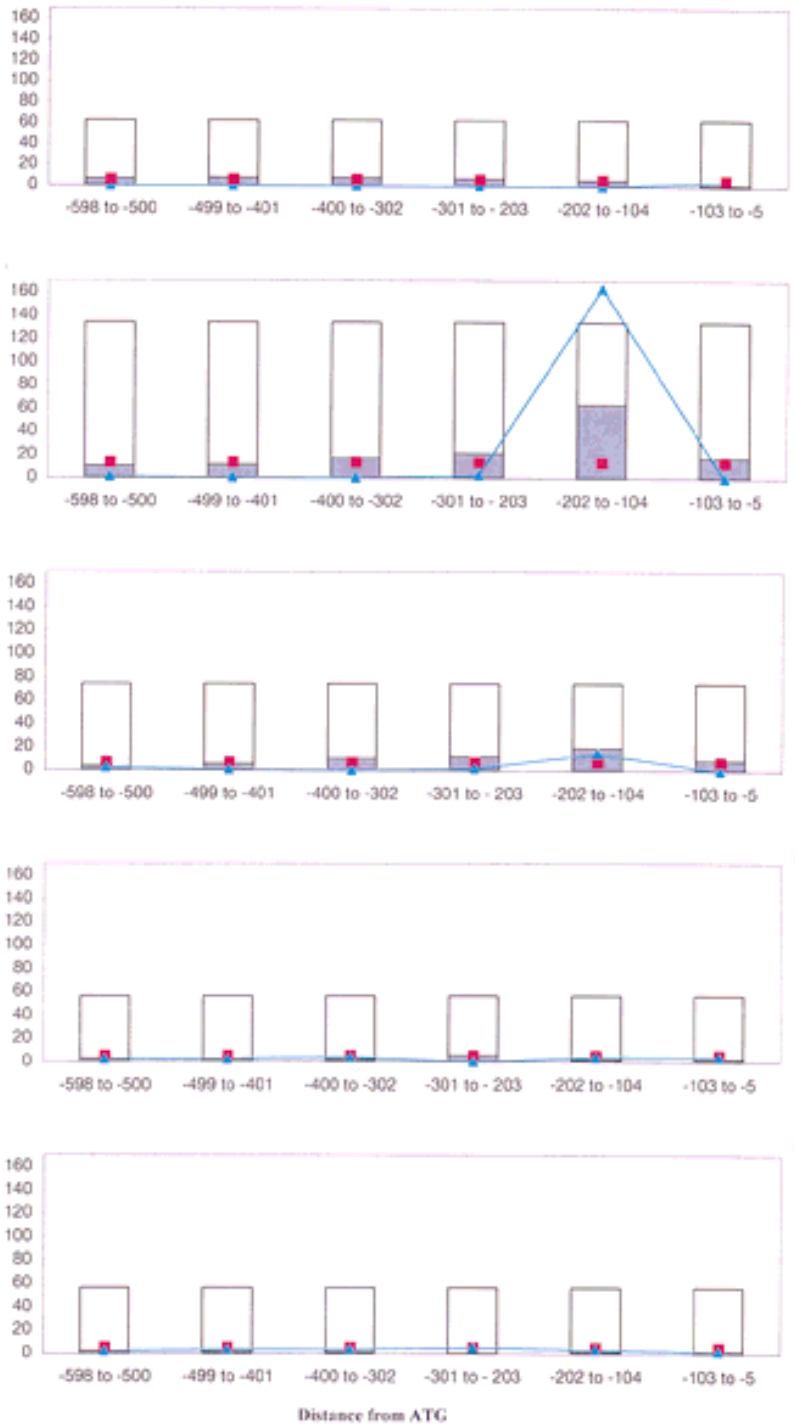


Figure 2 Test for pentamers over-represented in a position-dependent manner. (*A*) Early G_1: ACGCG. The height of the white bar represents the number of genes in each interval in each data set. The purple shading indicates the number of upstream regions in that interval that contain one or more copy of the sequence element. The number of upstream regions expected to contain that sequence element, if the element were evenly distributed among all intervals in all data sets, is marked with the pink box. The blue line indicates the contribution that each interval makes to the χ^2 score. (*B*) The pentamer ACGCG is over-represented in late G_1 upstream regions in the interval −104 to −202 nucleotides upstream of the ATG start codon—the observed number of elements (purple) is greater than the expected number (pink). (*C*) It is somewhat over-represented in the same interval in S. The total χ^2 score for ACGCG is 239.8. (*D*) G_1: ACGCG; (*E*) M: ACGCG.

and, to a lesser extent, in the interval −104 to −202 of S. Thus, its score (light blue) in those intervals is higher than it is in the other intervals.

Table 2 is a list of all statistically significant position-dependent pentamers. With $P \leq 0.05$, we find five pentamers and/or their reverse complements over-represented in late G_1 in the interval −104 to −202 from the ATG start codon, one in M in the interval −203 to −301, and one in G_2 from −401 to −499 and late G_1 from −5 to −103 (Table 2A). We also find 13 individual pentamers with a $P \leq 0.05$ (Table 2B); 8 of these pentamers represent one or both strands of the elements listed in Table 2A, whereas one is a new oligomer for early G_1, two for late G_1, one for S, and one for multiple sets. If we raise the threshold to $P \leq 0.2$, we add in four additional pentamers in early G_1 and late G_1, three in S, and one in many data sets. Eight of the pentamers identified by this method are contained within hexamers identified above and may not represent novel regulatory sites. Once again, the highest scoring pentamers are contained within the MCB element (ACGCGT) and variations of that sequence (Table 2A).

Position-Dependent Elements II: Clustered Within a Set of Upstream Regions

The experiment described above identifies oligomers that are over-represented in one or more 99-nucleotide intervals within the upstream regions of one or more of the five sets of cell cycle-regulated genes. However, it does not identify those elements that exhibit a preference for a specific location within one set of upstream regions. The details of this procedure are presented in Methods, and the basic concept is illustrated in Figure 3. The hexamers ACGCGT and CGACGC are both over-represented in late G_1 (Table 1). Their

Table 2. Position-Dependent Pentamers (Over-Represented in 99 Nucleotide Intervals)

A. χ^2 scores of pentamers and/or reverse complements

Distance from ATG		-598 to -500					-499 to -401					-400 to -302					-301 to -203					-202 to -104					-103 to -5					χ^2	p-value
Phase		EG	LG	S	G2	M	EG	LG	S	G2	M	EG	LG	S	G2	M	EG	LG	S	G2	M	EG	LG	S	G2	M	EG	LG	S	G2	M	score	≤
Early G1	ATATA/TATAT	6.02	6.38	1.77	0.17	0.09	0.19	4.26	5.59	0.17	0.00	0.79	2.10	1.83	0.03	1.27	0.21	0.30	0.01	0.09	0.00	5.05	4.32	2.44	0.03	0.29	**8.52**	1.34	2.44	0.09	1.56	57.35	0.1
	GGCCA/TGGCC	0.00	4.32	0.30	1.24	0.40	**10.53**	0.01	1.91	0.02	0.11	3.54	2.12	3.24	4.31	1.24	0.15	1.31	0.30	1.24	1.24	4.85	0.01	1.91	0.40	2.55	4.85	2.74	0.51	0.40	0.40	56.17	0.15
	CCGCG/CGCGG	0.04	1.25	3.12	0.06	0.06	2.06	1.25	0.40	2.94	0.79	2.06	0.97	0.25	0.79	1.13	**7.09**	0.02	0.25	1.13	2.36	2.06	0.48	4.81	5.59	2.36	2.66	5.66	0.25	2.36	0.06	54.31	0.2
Late G1	ACGCG/CGCGT	0.09	1.35	3.09	2.68	2.68	0.01	0.80	1.10	2.68	4.19	0.01	0.47	0.53	4.19	4.19	0.09	3.02	2.06	0.17	6.02	0.46	**163.97**	15.34	4.19	4.19	4.92	0.47	0.00	4.19	2.68	239.80	0.001
	CGCGA/TCGCG	2.38	2.13	0.07	3.24	3.24	0.49	1.37	0.42	0.22	1.84	1.26	11.79	0.82	3.24	0.22	0.95	2.01	0.02	0.00	0.83	2.38	**79.25**	8.07	0.18	1.84	5.67	0.36	0.02	3.24	3.24	140.79	0.001
	GACGC/GCGTC	1.64	2.38	0.53	0.96	5.24	0.21	2.38	2.22	5.24	0.29	0.21	0.96	0.12	0.96	3.43	0.21	7.14	0.00	1.45	3.43	0.61	**47.73**	3.72	0.96	3.43	4.06	0.51	5.07	3.43	5.24	113.77	0.001
	AACGC/GCGTT	0.19	0.17	4.41	1.30	1.30	0.01	0.17	0.38	3.37	0.20	0.05	0.88	0.35	3.37	1.30	0.58	0.88	3.34	1.30	1.30	0.30	**55.51**	7.48	2.22	1.30	4.29	1.35	0.79	1.30	0.20	99.61	0.001
	GCGAA/TTCGC	3.38	0.35	1.03	0.01	0.01	2.30	0.00	0.03	0.81	1.54	2.30	0.08	0.22	3.69	0.32	1.87	0.08	1.03	0.64	0.81	0.31	**30.92**	2.71	6.24	1.54	7.84	1.92	1.83	0.81	1.54	76.17	0.001
	GCGTA/TACGC	0.11	2.97	2.04	0.72	0.00	0.11	0.01	1.16	0.60	0.72	0.11	0.40	1.94	1.57	2.75	0.50	0.40	0.47	0.20	0.14	3.44	**23.28**	3.04	2.75	2.75	1.44	1.34	0.47	2.75	0.14	58.29	0.1
	CGAAA/TTTCG	4.16	0.18	0.34	2.05	3.93	0.00	0.67	0.14	3.93	5.10	0.00	0.61	0.14	0.37	0.11	0.66	2.23	1.14	0.59	0.00	0.00	**13.16**	5.22	2.54	2.92	0.23	0.08	2.46	0.00	2.92	55.90	0.15
S	GCCAC/GTGGC	0.61	0.17	2.40	0.59	0.96	0.75	0.02	1.37	2.70	1.45	0.00	0.17	3.72	0.96	0.59	0.00	1.00	**21.07**	0.96	0.01	4.06	3.41	0.62	0.96	0.01	0.14	2.45	0.53	2.00	2.00	55.69	0.15
M	AACCC/GGGTT	5.51	6.09	4.40	0.00	0.14	0.11	5.05	0.00	0.72	2.51	1.19	1.34	1.16	0.20	1.39	0.67	2.13	0.00	3.95	**10.26**	0.11	5.05	3.17	0.14	0.00	0.00	3.94	0.52	0.14	0.72	60.58	0.05
G2, Late G1	CAGCC/GGCTG	0.00	5.51	0.15	0.55	0.99	1.56	0.14	5.17	**8.52**	0.99	4.29	0.22	0.57	0.10	5.29	6.15	0.01	0.58	3.48	0.31	0.00	1.05	2.28	0.31	0.31	4.12	**8.97**	0.14	4.20	0.31	66.27	0.005
Many	GCCGC/GCGGC	1.52	0.61	4.47	2.41	0.22	0.08	0.61	0.64	4.24	0.22	4.77	0.21	0.13	4.24	1.10	2.91	3.09	0.92	4.24	3.92	0.45	2.05	4.47	0.00	0.94	2.64	4.35	0.27	0.22	0.94	56.86	0.15

B. χ^2 scores of pentamers

Distance from ATG		-598 to -500					-499 to -401					-400 to -302					-301 to -203					-202 to -104					-103 to -5					χ^2	
Phase		EG	LG	S	G2	M	EG	LG	S	G2	M	EG	LG	S	G2	M	EG	LG	S	G2	M	EG	LG	S	G2	M	EG	LG	S	G2	M	score	p-value ≤
Early G1	CCCTT	1.18	1.98	0.36	0.85	1.76	2.12	0.11	1.91	0.22	0.18	7.44	1.24	1.91	0.85	10.06	1.07	2.88	0.33	0.85	7.41	**15.99**	2.88	0.03	0.77	0.00	0.06	0.06	3.15	1.76	0.22	69.63	0.05
	ATATA	5.63	8.36	1.96	2.29	0.02	0.00	4.79	4.45	0.94	0.02	0.03	0.62	0.92	0.01	0.18	1.19	0.20	0.00	0.43	0.01	**10.61**	2.79	2.05	0.15	0.15	9.01	1.74	2.78	0.15	1.44	62.97	0.15
	TATAA	3.94	3.63	2.74	0.01	0.26	0.00	3.63	2.74	1.19	0.64	0.00	0.04	0.41	3.33	3.87	0.00	0.13	0.80	1.54	0.13	8.80	1.69	1.93	0.43	0.05	**16.54**	0.54	0.78	0.64	2.35	62.82	0.15
Late G1	ACGCG	0.17	3.97	3.94	2.59	2.59	0.24	0.20	1.33	1.28	4.36	0.17	0.23	0.10	2.59	2.59	0.17	2.96	0.87	0.09	4.36	0.74	**207.71**	6.75	4.36	4.36	3.11	1.22	0.54	4.36	2.59	270.55	0.001
	CGCGT	0.00	0.59	2.49	2.61	2.61	0.00	0.59	1.35	2.61	2.61	0.18	1.93	1.77	4.39	4.39	0.18	2.89	0.84	2.61	4.39	0.76	**141.25**	17.96	2.61	2.61	3.14	0.21	0.11	2.61	1.30	211.61	0.001
	GACGC	2.05	0.19	0.00	0.00	3.00	2.05	0.10	3.96	3.00	0.33	0.56	1.11	0.23	1.33	1.33	0.79	6.50	0.23	0.34	3.00	0.56	**54.81**	0.27	1.33	1.33	1.67	0.00	2.21	3.00	3.00	98.28	0.005
	CGCGA	3.34	3.68	0.22	1.31	1.31	0.13	0.63	0.22	0.35	0.32	0.54	11.10	1.09	1.31	0.32	0.04	0.50	0.00	0.32	0.32	1.64	**55.59**	1.09	0.00	2.97	3.34	0.00	0.29	2.97	1.31	96.26	0.005
	GCGTT	0.53	0.74	2.09	1.03	2.29	0.46	0.29	0.02	4.05	2.29	0.07	0.56	1.32	1.03	1.03	0.53	2.92	5.98	1.03	0.27	0.07	**46.96**	4.05	1.03	0.27	4.55	0.01	1.03	0.27	0.27	87.06	0.005
	TCGCG	0.71	0.35	0.14	3.19	1.51	1.87	1.73	0.01	1.51	1.51	1.87	7.09	1.83	1.51	0.20	1.62	1.48	0.14	0.20	0.01	0.71	**27.00**	7.92	1.02	0.45	3.59	2.82	0.01	1.51	3.19	76.70	0.005
	AACGC	0.03	0.03	3.00	0.69	0.14	1.09	0.02	1.08	0.69	0.29	0.46	0.52	3.00	3.03	0.69	0.38	0.03	0.06	0.69	1.65	0.46	**47.65**	3.35	1.65	1.65	1.09	1.03	0.30	1.65	0.14	76.54	0.005
	ATAAA	5.72	5.29	3.87	0.58	0.28	1.37	0.92	0.15	5.35	0.63	3.63	0.09	0.00	1.10	1.04	0.86	0.39	1.14	1.10	0.06	2.28	0.77	0.08	6.78	0.63	0.61	**20.78**	2.67	1.72	0.08	69.97	0.05
	AATAA	3.76	1.05	3.18	4.35	0.08	2.12	1.05	1.23	0.05	0.05	0.53	4.79	0.05	1.58	0.64	0.06	0.28	0.00	1.73	0.00	0.06	0.65	0.46	0.05	0.24	8.48	**24.64**	3.23	3.14	0.05	67.58	0.05
	GCGTC	2.36	3.81	2.06	2.88	2.88	1.54	3.81	0.17	2.88	0.27	0.96	0.65	0.01	0.27	2.88	0.47	5.44	0.01	1.57	1.22	0.02	**12.09**	4.64	0.27	2.88	3.23	0.11	3.80	1.22	2.88	67.27	0.05
	TAACA	0.72	0.51	3.56	1.65	0.31	0.03	1.38	0.87	0.85	0.31	0.03	0.06	0.38	0.04	1.65	0.03	8.02	1.57	0.81	1.60	0.03	0.06	0.09	0.04	0.31	1.47	**16.00**	3.68	2.65	14.56	63.25	0.15
	TAATA	2.88	5.90	0.06	1.18	4.05	1.98	2.05	0.06	0.21	0.58	0.11	1.52	0.06	0.58	1.18	0.11	0.06	0.31	0.22	0.02	0.39	1.34	1.10	0.03	0.21	7.08	**24.43**	3.61	0.61	0.02	61.95	0.2
S	AACAA	0.21	4.72	4.40	0.92	1.48	2.54	0.00	1.93	1.51	0.17	0.04	0.35	0.00	0.47	1.51	0.32	0.06	0.20	0.16	1.51	2.54	0.59	0.20	8.11	2.22	0.94	6.62	**12.81**	6.64	5.31	68.49	0.05
	GTTGG	0.15	6.15	1.52	4.02	2.12	7.32	1.14	0.04	0.54	0.83	0.15	1.99	**9.84**	0.13	1.64	0.82	1.12	7.08	1.64	1.64	2.02	0.00	0.45	3.34	1.64	0.15	4.49	0.08	1.64	0.54	64.21	0.1
	AAACA	0.01	3.61	3.65	2.59	0.12	3.76	0.01	0.88	0.40	0.12	0.01	0.66	0.28	0.83	1.77	3.76	0.20	0.20	0.01	0.04	2.80	0.01	0.00	8.25	2.17	0.94	0.66	**15.50**	9.94	0.83	64.01	0.1
Many	ACAAA	0.05	8.03	0.14	0.04	2.40	0.76	0.32	5.79	1.58	1.58	0.35	0.57	0.15	0.04	1.17	0.00	0.57	3.60	2.40	1.17	2.02	3.57	1.29	**10.92**	0.45	0.00	**9.16**	5.89	**10.92**	3.40	78.34	0.005

The number of upstream regions containing one or more copies of a pentamer was counted over intervals of 99 nucleotides in each of the five sets of cell cycle-regulated genes. The interval, measured in nucleotides from the ATG start codon, is indicated at the *top*. The χ^2 contribution made by each of the 30 populations is shown, along with the total value of χ^2. The population(s) that make the greatest contribution to χ^2 are indicated in boldface type.
(EG) Early G_1; (LG) late G_1; (S) S; (G_2) G_2; (M) M.
P-values for each χ^2 value were calculated from a Monte Carlo simulation.

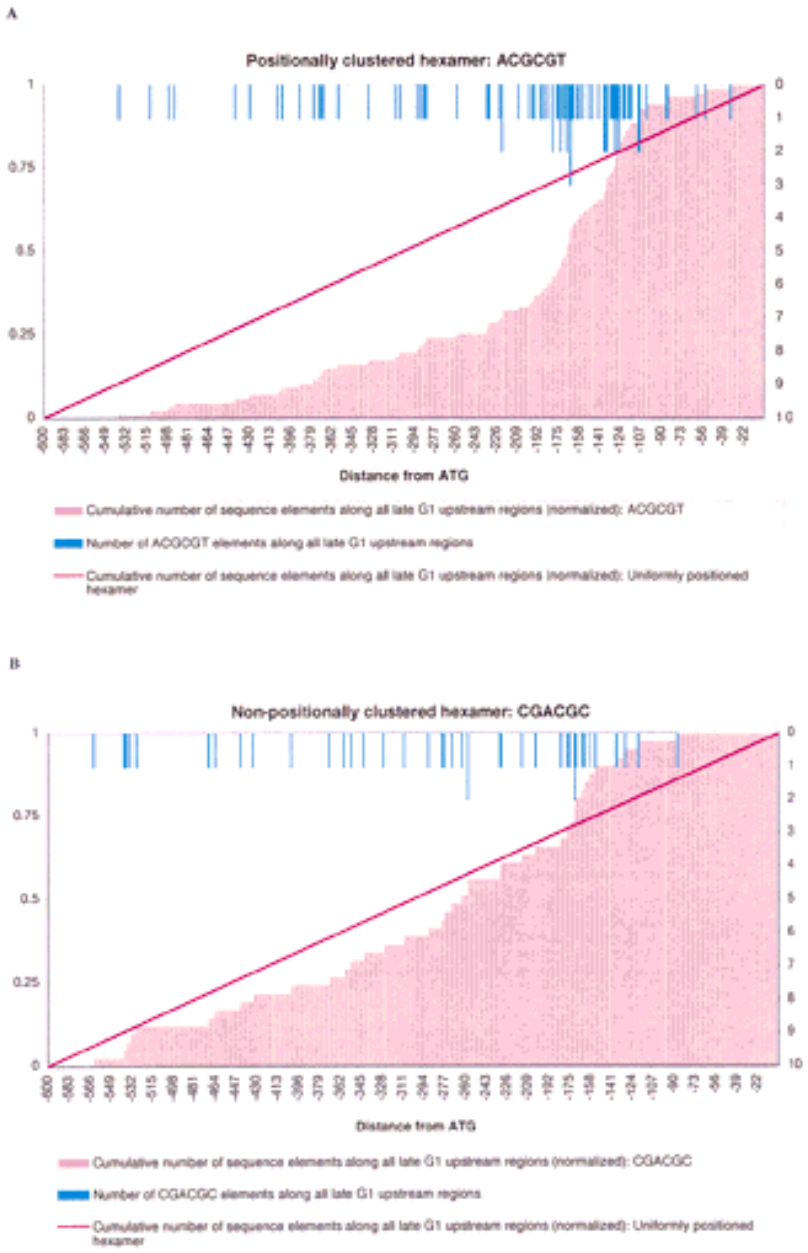


Figure 3 Test for clustered pentamers or hexamers. The light blue lines at the *top* represents the position of the sequence element along all upstream regions in a given data set—in this case, late G_1. The scale for these blue lines is along the *right* axis; the scale for all other lines is along the *left* axis. The pink histogram is a cumulative representation of the light blue lines, i.e., it shows the (normalized) cumulative number of oligomers present at that position. The dark pink line is the expected (normalized) cumulative number of sequence elements at that position, if the element is not clustered along the upstream region. Both elements ACGCGT and CGACGC are over-represented in late G_1 (Table 1). The two overlapping pentamers that make up ACGCGT, ACGCG, and CGCGT are both over-represented in late G_1 in the interval from −104 to −202 nucleotides upstream of the ATG start codon (Table 2). One of the two pentamers that makes up CGACGC, GACGC, is also over-represented in late G_1 in the interval from −104 to −202 nucleotides upstream of the ATG start codon (Table 2). However, although the 87 ACGCGT hexamers are clustered along late G_1 upstream regions according to the statistic used here, the 41 CGACGC hexamers are not clustered. Empirically, the ACGCGT hexamers are clustered between ~−100 and −200 nucleotides upstream of the ATG start codon.

positioning along all late G_1 upstream regions is shown in the figure. The light blue lines indicate the number of times that the element is found at each position upstream of the ATG start codon across all late G_1 genes. We calculated the normalized, cumulative number of occurrences of each oligomer along the late G_1 upstream regions (light pink histogram), and compared it with the cumulative number of occurrences of a hypothetical, completely uniformly positioned sequence (dark pink line). The hexamer ACGCGT is observed less frequently than the hypothetical sequence from position ~−600 to −125, whereas the hexamer CGACGC behaves similarly to the hypothetical sequence. The statistic (not illustrated) confirms that ACGCGT is clustered within late G_1 upstream regions, whereas CGACGC is not. We observe empirically that ACGCGT is clustered between ~−100 and −200 from the ATG start codon, whereas the element CGACGC is fairly evenly positioned (light blue lines).

We repeated this test to determine whether any of the other hexamers and pentamers identified above (Tables 1 and 2) are clustered in a particular location within the group of upstream regions in which they were shown to be over-represented. Tables 3 and 4 show the hexamers and pentamers, respectively, whose positioning within a set of upstream regions is statistically significant. Because of space constraints, we do not list the number of elements present at each position in the 600-nucleotide upstream region. Rather, for illustrative purposes, we have tabulated the total number of elements found in 50-nucleotide intervals in a particular data set. The total number of elements found in this method is different from the that presented in Tables 1 and 2; in Tables 1 and 2, the number of upstream regions containing one or more copy of an oligomer are counted, whereas in Tables 3 and 4, each occurrence of an element is counted separately (see Methods).

We find that a total of six hexamers and/or their reverse comple-

Table 3. Position-Dependent Hexamers (Clustered Hexamers)

A. Hexamers and/or reverse complements

Phase	Hexamer	Non-uniform distribution *Kolmogorov-Smirnov test p-value ≤*	Number of sequence elements in the interval												
			-600 to -551	*-550 to -501*	*-500 to -451*	*-450 to -401*	*-400 to -351*	*-350 to -301*	*-300 to -251*	*-250 to -201*	*-200 to -151*	*-150 to -101*	*-100 to -51*	*-50 to -1*	*totals*
Late G1	ACGCGT/ACGCGT	0.001	0	2	2	4	6	3	5	7	**25**	**28**	4	1	87
	AACGCG/CGCGTT	0.001	2	3	4	6	8	3	4	8	**25**	**38**	6	2	109
	CGCGTC/GACGCG	0.001	2	5	0	5	7	7	11	7	**25**	**20**	4	2	95
	ACGCGA/TCGCGT	0.005	4	3	2	5	5	5	6	6	**14**	**24**	6	3	83
	CGCGTA/TACGCG	0.001	0	0	2	4	3	2	5	3	**15**	**17**	4	1	56
	CGCGAA/TTCGCG	0.050	2	1	2	2	**12**	8	5	7	**10**	**19**	5	3	76

B. Hexamers

Phase	Hexamer	Non-uniform distribution *Kolmogorov-Smirnov test p-value ≤*	Number of sequence elements in the interval												
			-600 to -551	*-550 to -501*	*-500 to -451*	*-450 to -401*	*-400 to -351*	*-350 to -301*	*-300 to -251*	*-250 to -201*	*-200 to -151*	*-150 to -101*	*-100 to -51*	*-50 to -1*	*totals*
Late G1	ACGCGT	0.001	0	2	2	4	6	3	5	7	**25**	**28**	4	1	87
	AACGCG	0.001	0	3	1	4	3	1	1	5	**14**	**20**	5	2	59
	CGCGTT	0.050	2	0	3	2	5	2	3	3	**11**	**18**	1	0	50
	GACGCG	0.010	1	0	0	3	5	2	6	3	**14**	**14**	2	1	51
	ACGCGA	0.001	1	0	0	4	2	1	3	4	**9**	**13**	5	2	44

We performed a Kolmogorov–Smirnov test on each hexamer listed in Table 1. Shown here are those hexamers whose *P*-value from this test is ≤0.05. The data set on which we performed the test is listed in the *left* column. We also counted the number of upstream regions in that data set that contain one or more copies of the hexamer over the 50-nucleotide intervals shown. The interval(s) with the highest numbers is in bold. Intervals are measured by the distance from the ATG start codon.

Table 4. Position-Dependent Pentamers (Clustered Pentamers)

A. Pentamers and/or reverse complements

Phase	Pentamer	Non-uniform distribution *Kolmogorov-Smirnov test* *p-value* ≤	Number of sequence elements in the interval															
			-600 to -551	*-550 to -501*	*-500 to -451*	*-450 to -401*	*-400 to -351*	*-350 to -301*	*-300 to -251*	*-250 to -201*	*-200 to -151*	*-150 to -101*	*-100 to -51*	*-50 to -1*	*totals*
Early G1	ATATA/TATAT	0.001	16	22	22	22	29	32	26	17	**50**	**44**	**47**	23	350
	GGCCA/TGGCC	0.001	2	5	7	8	4	6	3	1	0	0	0	0	36
Late G1	ACGCG/CGCGT	0.001	5	9	7	16	20	14	20	21	**70**	**84**	15	7	288
	CGCGA/TCGCG	0.010	6	5	3	8	**20**	16	10	13	**24**	**28**	6	3	142
	AACGC/GCGTT	0.001	8	11	8	15	13	15	12	17	30	**50**	19	5	203
	GCGAA/TTCGC	0.010	12	7	13	11	19	9	14	15	**24**	**24**	16	10	174
G2, Late G1	CAGCC/GGCTG	0.005	13	15	**20**	9	12	9	10	6	7	5	5	3	114

B. Pentamers

Phase	Pentamer	Non-uniform distribution *Kolmogorov-Smirnov test* *p-value* ≤	Number of sequence elements in the interval												
			-600 to -551	*-550 to -501*	*-500 to -451*	*-450 to -401*	*-400 to -351*	*-350 to -301*	*-300 to -251*	*-250 to -201*	*-200 to -151*	*-150 to -101*	*-100 to -51*	*-50 to -1*	*totals*
Early G1	ATATA	0.001	7	11	11	12	16	15	13	5	**27**	**25**	**24**	16	182
	TATAA	0.001	3	2	7	7	8	5	8	9	12	14	**18**	**18**	111
Late G1	ACGCG	0.001	1	3	2	9	10	5	9	11	**37**	**44**	10	5	146
	CGCGT	0.001	4	6	5	7	10	9	11	10	**33**	**40**	5	2	142
	CGCGA	0.050	2	0	0	5	10	7	5	6	12	**15**	5	2	69
	GCGTT	0.050	3	4	5	4	8	5	8	9	13	**24**	8	1	92
	AACGC	0.050	5	7	3	11	5	10	4	8	17	**26**	11	4	111
	ATAAA	0.001	13	14	22	13	15	25	20	16	24	31	**44**	31	268
	AATAA	0.001	21	15	22	14	11	16	20	22	21	27	**50**	42	281
	TAATA	0.001	9	5	9	11	15	8	15	16	18	**26**	**25**	**33**	190
S	AACAA	0.001	3	6	3	10	9	12	11	10	7	19	**34**	17	141
	AAACA	0.001	6	5	7	9	10	12	6	12	7	18	**36**	23	151

We performed a Kolmogorov–Smirnov test on each pentamer listed in Table 2. Shown here are those pentamers whose *P* value from this test is ≤0.05. The data set on which we performed the test is listed in the *left* column. We also counted the number of upstream regions in that data set that contain one or more copies of the pentamer over the 50-nucleotide intervals shown. The interval(s) with the highest numbers is in boldface type. Intervals are measured by the distance upstream of the ATG start codon.

ments are clustered among a set of upstream regions (Table 3A). An additional five individual hexamers are also clustered, but these do not represent new binding sites (Table 3B). All clustered hexamers are from late G_1, and all show a preference for the interval −101 to −200 from the ATG start codon. Seven pentamers and/or their reverse complements are clustered, as are twelve individual pentamers (Table 4). Four of the late G_1 pentamers also tend to be located between −101 and −200. Pentamers assessed in other data sets do not show such an overwhelming preference for one location.

DISCUSSION

Using high-density oligonucleotide arrays, Cho et al. (1998) identified 420 transcripts, of the 6220 ORFs in the *S. cerevisiae* genome, whose expression is cell cycle regulated. They further divided these transcripts into those whose expression peaks during the early G_1, late G_1, S, G_2, or M phase of the cell cycle. We have developed a statistical method to scan upstream regions for short oligomers that may be involved in the regulation of coexpressed genes. These elements can be position- and orientation-independent or dependent. We have applied this technique to the transcripts identified by Cho et al. (1998) and have produced a list of elements, some new, some identified previously, which are likely responsible for the coordinate regulation of genes during the yeast cell cycle. Although the oligomers themselves may not represent the complete protein-binding sites, they will serve as excellent starting points for future studies of longer, more complex, regulatory sites.

Table 5 presents a composite list of the elements that may be important for cell cycle-dependent transcriptional regulation in yeast. The statistically significant elements with $P \leq 0.05$ are marked with an asterisk; unmarked sequences have $P \leq 0.2$. To make it easier to identify the potential regulatory sites, we have grouped together pairs of sequences that are reverse complements. We list the pentamers and hexamers separately, even though many may be part of the same binding site, because some pentamers are contained within two or more different hexamers. A total of nine hexamers have $P \leq 0.05$, and another four have $P \leq 0.2$. Twelve pentamers have $P \leq 0.05$; six of these elements are contained within a hexamer. Another 12 pentamers have $P \leq 0.2$.

The hexamers listed in Table 5 have a statistically significant distribution independent of their position within an upstream region. Thus, these elements could act at any location within the upstream region. However, we have found that six of them are clustered in a particular region within the set of upstream regions in which they are over-represented (Table 3). This positioning may be important for their function. The pentamers, which were identified as being over-represented in a particular 99-nucleotide upstream interval, are likely to be part of binding sites that function in a position-dependent manner. Furthermore, within a group of coregulated upstream regions, some of the pentamers are clustered in a specific region (Table 4). It is not surprising that only a subset of the position-dependent pentamers is also clustered. The method identifies position-dependent pentamers by counting the number of upstream regions that contain one or more copies of the element; the clustering portion of the method counts all occurrences of the element.

One result of our study is that many of the sequence elements likely function in an orientation-independent manner, that is, they could act when placed on either DNA strand. The over-representation of some orientation-independent elements is due to the number of upstream regions that contain either the oligomer or its reverse complement on one strand, such as GAGTCA/TGACTC in G_2 (Table 1A). In other orientation-independent elements, the sequence and its reverse complement are both individually over-represented, such as the pair AACGCG and CGCGTT in late G_1 (Table 1B). In contrast, elements in which only one of the two elements is statistically significant, such as CCCTT in early G_1, may function in an orientation-dependent fashion, with the sequence CCCTT being on the coding strand (Table 1B).

An important validation of our method is the fact that one of the oligomers on our list has been identified previously as being involved in cell cycle-dependent gene regulation. The transcription factor complex MBF, a heterodimeric complex composed of Mbp1p and Swi6p, binds to the consensus sequence ACGCGT, the MCB (for review, see McIntosh 1993; Koch and Nasmyth 1994). It has been shown that this sequence is present upstream of many genes involved in DNA synthesis and repair that are regulated at the late G_1/early S transition (for review, see McIntosh 1993). This palindromic sequence is the highest scoring hexamer by our methods, and our analysis supports the idea that it functions during late G_1 (Table 1). The element occurs in 49% of late G_1-upstream regions, and, in 65% of these upstream regions, it is located between −101 and −200 from the start codon. It is also present in 22% of the upstream regions from S phase. Furthermore, 29% of the late G_1 upstream regions that contain the MCB element contain two or more copies of it (not shown). This result correlates with the experimental observations that increasing the number of MCB sequences in a gene upstream region leads to a higher rate of transcription of the gene (McIntosh 1993).

It is striking that the highest scoring elements are all variants of the MCB sequence, ACGCGT, are all

Table 5. Summary

A. Hexamers

Phase	Hexamer	Aligns with MCB at positions 4/6	Aligns with MCB at positions 5/6	Contains pentamer	Clustered
Early G_1	*C C C C G C / G C G G G G				
	C C C C G C				
	A C G C G G / C C G C G T	Y	Y	Y	
Late G_1	*A C G C G T / A C G C G T	Y	Y	Y	Y
	*A C G C G T	Y	Y	Y	Y
	*A A C G C G / C G C G T T	Y	Y	Y	Y
	*A A C G C G	Y	Y	Y	Y
	*C G C G T T	Y	Y	Y	Y
	*C G C G T C / G A C G C G	Y	Y	Y	Y
	*G A C G C G	Y	Y	Y	Y
	*C G C G T C	Y	Y	Y	
	*A C G C G A / T C G C G T	Y	Y	Y	Y
	*A C G C G A	Y	Y	Y	Y
	T C G C G T	Y	Y	Y	
	*C G C G T A / T A C G C G	Y	Y	Y	Y
	*C G C G A A / T T C G C G	Y		Y	Y
	C G C G A A	Y		Y	
	*C G A C G C / G C G T C G	Y		Y	
S	C C G T G C / G C A C G G				
	A G T C A G / C T G A C T				
G_2	*G A G T C A / T G A C T C				
Early G_1, M	C C C T T T			Y	

B. Pentamers

Phase	Pentamer	Aligns with MCB at 4/5 positions	Contained within hexamer	Clustered
Early G_1	A T A T A / T A T A T			Y
	A T A T A			Y
	G G C C A / T G G C C			Y
	C C G C G / C G C G G	Y	Y	
	*C C C T T		Y	
	T A T A A			Y
Late G_1	*A C G C G / C G C G T	Y	Y	Y
	*A C G C G	Y	Y	Y
	*C G C G T	Y	Y	Y
	*C G C G A / T C G C G	Y	Y	Y
	*C G C G A	Y	Y	Y
	*T C G C G	Y	Y	
	*G A C G C / G C G T C	Y	Y	
	*G A C G C	Y	Y	
	*G C G T C	Y	Y	
	*A A C G C / G C G T T	Y	Y	Y
	*G C G T T	Y	Y	Y
	*A A C G C	Y	Y	Y
	*G C G A A / T T C G C		Y	Y
	G C G T A / T A C G C	Y	Y	
	C G A A A / T T T C G			
	*A T A A A			Y
	*A A T A A			Y
	T A A C A			
	T A A T A			Y
S	G C C A C / G T G G C			
	*A A C A A			Y
	G T T G G			
	A A A C A			Y
M	*A A C C C / G G G T T			
G_2, Late G_1	*C A G C C / G G C T G			Y
Many	G C C G C / G C G G C			
	*A C A A A			

Summary of the data presented in Tables 1–4. (*) Hexamers and pentamers with $P \leq 0.05$ from Table 1 or 2.

over-represented in late G_1, and, to some extent, in S, and most are located between 100 and 200 nucleotides upstream of the ATG start codon. Table 5 documents the amount of similarity that the elements share with MCB. All seven of the late G_1 hexamers with a $P \leq 0.05$ are identical to MCB at four of six positions, and five are identical at five of six positions. Of the seven late G_1 pentamers with $P \leq 0.05$, four overlap with MCB at four of five positions. Furthermore, all of the MCB-like pentamers were identified because they are over-represented in the interval between -202 and -104 (Table 2), and all of the MCB-like pentamers and hexamers are clustered in the same region (Tables 3 and 4). MCB and the related sequences are thus different from other yeast upstream-activating sequences, as most elements function at variable distances from the

mRNA start site (Struhl 1995; Kunzler et al. 1996). The positioning of the MCB-like sequences is likely important for their function.

Some elements, such as AACGCG and CGCGTC clearly overlap with MCB, but the sequence ACGCGA may represent a novel binding site. Although it is tempting to create a longer MCB-like consensus sequence by combining overlapping sequences, this procedure is not sound. The heptamers AACGCGA/TCGCGTT, AACGCGT/ACGCGTT, and ACGCGAA/TTCGCGT can be formed by combining two hexamers listed in Table 1. The over-representation of these heptamers in late G_1 upstream regions is statistically significant (not shown). However, the distribution of the heptamers, ACGCGTA/TACGCGT, GACGCGA/TCGCGTC, and TACGCGA/TCGCGTA, also created by merging hexamers from Table 1, is not statistically significant. Thus, it is likely that the sequence context of the pentamers and hexamers plays a role in protein binding. Two sequence elements that activate transcription at the G_1/S boundary are SCB (CACGAAA), which is recognized by the Swi4p–Swi6p complex, and MCB, recognized by a Mbp1p–Swi6p complex. The yeast CLN1 upstream region contains three MCB-like sequence elements, ACGCGT, TCGCGA, and CCGCGT. Surprisingly, these elements are bound by Swi4p–Swi6p, not the typical MCB-binding complex of Mbp1p–Swi6p (Partridge et al. 1997). Thus, the MCB-like elements may be part of longer sequences that interact with Mbp1p–Swi6p, Swi4p–Swi6p, or even a novel combination of factors. We plan to analyze the sequences flanking the MCB and MCB-like elements in the future to better classify particular subclasses of binding sites.

One of the major cell cycle control checkpoints in yeast, called Start, occurs during late G_1 phase (for review, see Lew et al. 1997). Thus, it is interesting that we find a larger collection of statistically significant sequences upstream of late G_1 genes than upstream of other cell cycle-regulated genes. Two-thirds of the elements that may be important for cell cycle control are over-represented in late G_1, and all but two of the late G_1 elements share sequence similarity, and are over-represented or clustered in the region 100 to 200 nucleotides upstream of the ATG start codon. This finding leads us to speculate that control of transcription during late G_1, around the time of Start, is especially important to the cell. However, transcription during other phases may be controlled by a wide range of activities, including the regulation of mRNA stability, which do not necessarily require upstream sequence elements.

Some of the pentamers and hexamers that we have identified are also contained within sequences shown previously to be important for transcription in yeast. As a reference for transcriptionally active yeast sequences, we used the 312 yeast protein-binding sites listed in the TRANSFAC database, version 3.5 (Heinemeyer et al. 1999). Of the 21 hexamers and pentamers with $P \leq 0.05$ (Table 5), 17 are contained within one or more TRANSFAC-binding sites. These 119 TRANSFAC entries represent the binding sites for 23 binding factors (Table 6). Two of these binding factor entries, T01094 and T01120, are for MCB, the late G_1 transcription factor described above. Another entry, T00096, is for the binding factor SBF, a late G_1-specific transcription factor (Koch and Nasmyth 1994). However, this factor binds at least two similar sites, and the one that contains the pentamers and hexamers is not the one cited as the consensus in the literature. A third factor, T00798, is TBP, the TATA-binding protein (Hahn et al. 1989). The pentamers that overlap with the TBP-binding site, ATAAA, AATAA, and ACAAA are over-represented in the interval from -5 to -103 (Table 2), the region in which TBP is active (Struhl 1995; Kunzler et al. 1996). These pentamers could be a part either of the TBP-binding site, or of a different site.

This method clearly identifies sequence elements that are over-represented in one of the five sets of cell cycle-regulated upstream regions. However, elements that are evenly distributed among the five data sets will not be highlighted. The SCB element, which has the consensus sequence CACGAAA, was identified previously as a regulatory sequence located upstream of genes transcribed in late G_1 and early S. SCB is bound by the SBF transcription factor, a complex of Swi4p and Swi6p (for review, see Koch and Nasmyth 1994). Because the two hexamers that compose SCB, CACGAA and ACGAAA, are evenly distributed among all cell cycle-regulated upstream regions, neither is statistically significant by our analysis. A pentamer, CGAAA, that is part of SCB is over-represented in late G_1 in the interval -104 to -202, but with $P \leq 0.15$. However, CACGAAA may be part of a longer sequence that serves as the SBF-binding site. This longer, yet unidentified sequence may be over-represented in late G_1 upstream regions. Some of the late G_1 oligomers that we have identified in this study may also be contained within this longer binding sequence.

A second analysis of cell cycle-regulated yeast genes was recently performed by Spellman et al. (1998). These researchers used different technology and methods from those described in Cho et al. (1998), and identified ~800 genes that are cell cycle regulated. Although this set of 800 genes includes only 304 of the 420 identified by Cho et al. (1998), the cell cycle periodicity described by Cho et al. (1998) has been confirmed in separate clustering experiments (Tamayo et al. 1999). Spellman et al. (1998) searched for potential regulatory elements among groups of coregulated genes by a Gibbs sampling strategy, a different statis-

tical procedure from that used here. They describe ~15 motifs, some known, some novel, that are found upstream of clusters of coregulated genes. They also identified the MCB site, ACGCGT, upstream of a set of genes whose expression peaks in G_1. In the same group of genes, their strategy identified SCB, another late G_1 element. One of our late G_1 pentamers, GCGAA, is part of a motif they identified in the histone cluster. And the pentamer AACAA that we identified in S phase is part of a regulatory site they predict for their MCM cluster. Other sites do not overlap with those we identified. The discrepancies between our results and those of Spellman et al. (1998) are not surprising. Most important, because of the techniques used to analyze the gene expression data, the lists of coregulated genes are not the same. Furthermore, it appears that many of the motifs they describe are extended consensus sequences for previously identified regulatory binding sites, whereas our analysis is performed with no pre-existing knowledge of the sites. We are currently applying our statistical methods to the coregulated genes described in Spellman et al. (1998), and plan a detailed comparison with the results presented here.

The sequence elements that we have identified are responsible for much of the cell cycle regulation described by Cho et al. (1998). When the elements are analyzed independently, it is clear that none is solely responsible for transcription at a specific time during the cell cycle. The highest scoring element, ACGCGT (MCB), is present in 49% of late G_1 upstream regions; the hexamers with $P \leq 0.05$ over-represented in early G_1 and G_2 are found in only 22% and 27% of their respective upstream regions. These frequencies are in agreement with those found by Chu et al. (1998), who looked for sequence elements controlling sporulation in yeast. However, previous work on cell cycle regulation in yeast suggests that more than one sequence element may be responsible for transcription at the same phase (e.g., SCB and MCB both regulate late G_1 mRNA expression). Thus, we counted the number of upstream regions that had either one of the statistically significant hexamers ($P \leq 0.05$), at any position, or one of the significant pentamers, in the interval specified in Table 2. The results are as follows: 35% of early G_1 upstream regions contain one or more of the predicted elements, as do 98% of late G_1 upstream regions, 45% of S upstream regions, 27% of G_2 upstream regions, and 25% of M upstream regions. Thus, our work predicts that a variety of elements can be responsible for transcription at each phase of the cell cycle. However, other mechanisms, such as differential mRNA stability, may also play a role. We anticipate that the oligomers described here will be used by researchers studying yeast cell cycle gene regulation as a basis for further investigation.

METHODS

Creating a Database of Cell Cycle-Regulated Yeast Upstream Regions

We have created a database of all yeast upstream region sequences (T.G. Wolfsberg and D. Landsman, unpubl.) after extracting information present in YPD (Hodges et al. 1999) and MIPS (Mewes et al. 1999) as of February, 1997. This database contains the sequence 600 nucleotides upstream of each ATG start codon. We identified all ORFs that were annotated in both yeast databases. We then extracted the sequence 600 nucleotides upstream of each of these ORFs from the MIPS version of the yeast genome sequence. All upstream region sequences are oriented in the direction in which transcription occurs, with the start of the sequence at -600 and the end of the sequence at -1 with respect to the ATG start codon.

Cho et al. (1998) have carried out a comprehensive analysis of cell cycle-regulated transcription in yeast. We used the results of their work to create five data sets that contain the upstream region sequences of genes whose transcription is induced in one phase of the cell cycle. The data sets are defined as follows: early G_1; 63 upstream regions; late G_1; 134 upstream regions; S; 74 upstream regions; G_2; 56 upstream regions; M; 56 upstream regions.

Assessing Nonrandom Representation of Short Sequences Within the Cell Cycle-Regulated Yeast Upstream Regions

Our search for oligomers responsible for cell cycle-dependent regulation was made on the basis of the following hypothesis. If a short sequence is responsible for gene regulation in a specific phase of the cell cycle, it should be over-represented among the corresponding upstream regions, relative to other cell cycle-regulated upstream regions. Moreover, any short subsequences contained in it should also be over-represented. As a hypothetical example, if the sequence element AGGCATTC were responsible for gene regulation specifically in late G_1, it should be over-represented in late G_1 upstream regions, relative to upstream regions specific to the other cell cycle phases. Moreover, AGGCATTC is composed of three shorter hexamers, AGGCAT, GGCATT, and GCATTC, they also should be similarly over-represented in late G_1 upstream regions.

Thus, we were interested in identifying short DNA oligomers that are over-represented in upstream regions specific to a particular phase of the cell cycle, relative to other cell cycle-regulated upstream regions. We therefore wished to apply statistical tests for this over-representation. Most available statistical tests, for example, Gibbs sampling, representation ratios, *z*-scores, etc., assume a statistical parent population in which random DNA bases are generated by a Markov model, or perhaps independently. The model becomes part of what is being tested statistically, which is undesirable. We therefore developed a statistical test whose parent population represents real DNA, so that in principle, before the experiments were performed, samples from the parent population were possible experimental outcomes.

Our actual data set consisted of 383 cell cycle-regulated upstream regions (nucleotides -1 to -600), divided into five classes corresponding to early G_1 (63 genes), late G_1 (134 genes), S (74 genes), G_2 (56 genes), and M (56 genes). If only these numbers are known, before the experimental results classify these 383-upstream regions, every classification into

Table 6. Overlap of Hexamers and Pentamers with Yeast TRANSFAC Entries

TRANSFAC binding factor	TRANSFAC binding site	Hexamer or pentamer	Over-represented phase
T00056: BAF1.	R00716: ATCAATACATCATAAAATACGAACGATC	ATAAA	Early G1
	R03820: AATCACCGCGAACGGA	CGCGA/TCGCG	Late G1
		CGCGAA/TTCGCG	Late G1
		GCGAA/TTCGC	Late G1
T00096: CCBF. T00775: SWI4. T01013: SWI6.	R03735: CGCGAAA	CGCGA/TCGCG	Late G1
		CGCGAA/TTCGCG	Late G1
		GCGAA/TTCGC	Late G1
T00179: CUP2.	R01846: gatGCGTCTTTTCCGCTGAACCGttc	GACGC/GCGTC	Late G1
	R01043: CGATGCGTCTTTTCCGCTGAACCGTT	GACGC/GCGTC	Late G1
T00302: GAL4.	R00491: CGCGCCGCACTGCTCCGAACAAT	AACAA	S
	R01791: CGGGTGACAGCCCTCCGaCGGGTGACAGCCCTCCGaCGGGTGACAGCCCTCCG	CAGCC/GGCTG	G2, Late G1
T00321: GCN4.	R04022: TTGACTCTT	GAGTCA/TGACTC	G2
	R04023: ATGAATAAT	AATAA	Late G1
	R00645: GAGTCA	GAGTCA/TGACTC	G2
	R00648: TGACTC	GAGTCA/TGACTC	G2
	R00651: TGACTC	GAGTCA/TGACTC	G2
	R00655: TGACTC	GAGTCA/TGACTC	G2
	R00829: GAGTCA	GAGTCA/TGACTC	G2
	R02022: TTGACTCTCtaaaaaATGATTCAT	GAGTCA/TGACTC	G2
T00346: HAP1.	R00257: TGGCCGGGGTTTACGGACGATGA	AACCC/GGGTT	M
T00480: MAL63.	R01633: GATTTATCCGGAAATTTTCGCGGAC	CGCGA/TCGCG	Late G1
		CGCGAA/TTCGCG	Late G1
		GCGAA/TTCGC	Late G1
T00487: MATalpha2. T00715: RAP1. T00488: MATa1.	R01019: GATGTCTGGGTTTT	AACCC/GGGTT	M
T00500: MCM1.	R01476: TGACTTTCCAAATTGGGTTAAAA	AACCC/GGGTT	M
T00509: MIG1.	R02024: CCCCGCaTTTTT	CCCCGC/GCGGGG	Early G1
	R04156: TTAAAAGCGGGG	CCCCGC/GCGGGG	Early G1
	R04154: CCCCGCgTTTAT	CCCCGC/GCGGGG	Early G1
T00689: PHO2. T01027: BAS1.	R00652: TAATAGTGACTCCGGTAAATTAGTTAATTAA	GAGTCA/TGACTC	G2
T00690: PHO4.	R01207: AATTAGCACGTTTTCGCATA	GCGAA/TTCGC	Late G1
T00715: RAP1.	R01210: GAACCCATACACT	AACCC/GGGTT	M
	R03124: ACATCCGTACAACGaGAACCCATACATTA	AACCC/GGGTT	M
	R04074: ttacacctggACACCCCTTTTCTggcatccagtt	CCCTT	Early G1
	R00714: ATGGGTTTTTG	AACCC/GGGTT	M
	R00717: tattgCAAAAACCCATcaaccttg	AACCC/GGGTT	M
T00723: RC1. T00724: RC2.	R00259: GCCGGGGTTTA	AACCC/GGGTT	M
T00724: RC2.	R00258: GGCCGGGGTTTAC	AACCC/GGGTT	M
T00725: REB1.	R00489: CGGGTGACAGCCCTCCGA	CAGCC/GGCTG	G2, Late G1
	R03753: actGGGTTACCCGGggcacctg	AACCC/GGGTT	M
T00798: TBP.	R03162: gtatgTATATAAAac	ATAAA	Early G1
	R03172: TATAAA	ATAAA	Early G1
	R03811: TTTAAATAAGT	AATAA	Late G1
	R03812: CAATTTAATACCTAAATATAAAAAATGTTATTATA	ATAAA	Early G1
	R03813: TATAAATAGTATCAATATATATATATATATATATATTTATTTG	ATAAA	Early G1
	R03814: ATACAAAACATAAAAATAAAT	ATAAA	Early G1
		AATAA	Late G1
		ACAAA	Many
T01027: BAS1.	R00650: TAATAGTGACTCCGGTAAATT	GAGTCA/TGACTC	G2

Table 6. (*Continued*)

TRANSFAC binding factor	TRANSFAC binding site	Hexamer or pentamer	Over-represented phase
T01094: DSC1.	R03656: ACGCGAaaACGCGT	ACGCG/CGCGT	Late G1
		ACGCGA/TCGCGT	Late G1
		ACGCGT/ACGCGT	Late G1
		CGCGA/TCGCG	Late G1
T01120: MCB.	R03724: gcgACGCGTttta	ACGCG/CGCGT	Late G1
		ACGCGT/ACGCGT	Late G1
	R03725: aACGCGTttg	ACGCG/CGCGT	Late G1
		ACGCGT/ACGCGT	Late G1
	R03726: ACGCGT	ACGCG/CGCGT	Late G1
		ACGCGT/ACGCGT	Late G1
	R03727: tattACGCGTaac	ACGCG/CGCGT	Late G1
		ACGCGT/ACGCGT	Late G1
	R03728: gcgACGCGAggcctcACGCGTcgg	ACGCG/CGCGT	Late G1
		ACGCGA/TCGCGT	Late G1
		ACGCGT/ACGCGT	Late G1
		CGCGA/TCGCG	Late G1
T01146: QBP.	R02071: TTAGAAGCCGCCGAGCGGGTGACAGCCCT	CAGCC/GGCTG	G2, Late G1
T01291: DAL82.	R03864: GGTGCGATAGAATACCGCGGATTTTGGAAAATTGCGTTTGCTTTTCTTATCACA	AACGC/GCGTT	Late G1
T01306: SKO1, T00321: GCN4.	R00649: ATGACTCAT	GAGTCA/TGACTC	G2
None listed	R03750: TGTCTTTTCTCACCCCTTATGGGGAC	CCCTT	Early G1
	R03790: GTTCGGATAAATTTTT	ATAAA	Early G1
	R03819: TATATAAA	ATAAA	Early G1
	R03832: CACGAAAAACGAGACAAACGAA	ACAAA	Many

Hexamers and pentamers from Tables 1 and 2 with $P \leq 0.05$ were compared to the yeast binding site entries in TRANSFAC v. 3.5. All binding sites that contain one or more of the hexamers or pentamers are listed. The binding sites are grouped together by the factors that bind to them, according to the TRANSFAC database.

early G_1 (63 genes), late G_1 (134 genes), S (74 genes), G_2 (56 genes), and M (56 genes) is equally likely. The random classifications are the natural parent population for our problem. To sample the parent population, we used Monte Carlo random sampling to create 1000 mock data sets by classifying the 383-upstream regions into the five classes at random. The 1000 mock data sets were an adequate sample for the over-representation tests described next.

Statistical Tests for Orientation-Dependent, Position-Independent Over-Representation

If an oligomer within an upstream region can regulate regardless of its specific position but only in a particular orientation, we call it orientation dependent, position independent. We can test all DNA oligomers of a fixed length for orientation-dependent, position-independent over-representation as follows. (For orientation independence, etc., see below.) Because the over-representation is orientation dependent, each oligomer is considered different from its reverse complement from the other DNA strand (e.g., ACGCGA separately from TCGCGT).

First, in the real data set, count the number of upstream regions in each class that contained the oligomer. (Note: not all upstream regions contain the oligomer, and multiple appearances in the same upstream region are counted only once.) To rank oligomers by non random representation in the five classes, calculate a χ^2 score for each oligomer. The χ^2 score is defined as

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

where o_i, the observed value, is the number of upstream regions in class i that contain the oligomer, and e_i, the expected value, is the number of upstream regions in class i expected to contain the oligomer if the upstream regions had been classified at random.

The χ^2 score for each oligomer can then be used to rank the oligomers for over-representation. On one hand, oligomers that are represented in each class according to the expectation have a small χ^2 score (i.e., they are randomly represented). On the other hand, oligomers that are over-represented in a class should have a large χ^2 score. Over-representation in one class is logically equivalent to under-representation in some other classes. Accordingly, the i^{th} term in the χ^2 score for a particular oligomer indicates whether class i deviates from the oligomer's expected representation; it does not distinguish directly whether the oligomer was over-represented or under-represented. Once statistical significance has been established, however, over-represented classes for an oligomer are easily determined by selecting the largest terms with $o_i > e_i$ in the χ^2 score.

Each oligomer has now been assigned a χ^2 score corresponding to the real data set, so a statistical significance from

the 1000 mock data sets can then be calculated as follows. In each mock data set, calculate the χ^2 score for each oligomer as above. Find the largest χ^2 score from each of the 1000 mock data sets, and then sort these 1000 numbers. For any particular P value, find the $1000 \times P$th largest χ^2 score among these 1000 numbers. For the standard $P = 0.05$, for example, find the fiftieth largest (to avoid missing biologically interesting oligomers, we also used $P = 0.2$, corresponding to the two-hundredth largest). This cutoff has a direct interpretation: For any oligomer, a χ^2 score larger than the cutoff occurs only 5% of the time in a random data set. The cutoff therefore provides a statistical test at the $P = 0.05$ level of significance. This statistical test automatically accounts for multiple testing, as it must, because all oligomers of a fixed length have been tested.

Under the random classification of the mock data sets, the χ^2 score for each oligomer approximately follows a χ^2 distribution with $(D = 4 = 5 - 1)$ degrees of freedom, as long as the expected number in each class satisfies $e_i \geq 5$. Thus, of two oligomers, if one is overall more frequent within cell cycle-regulated upstream regions (i.e., in the aggregate of all five classes), the two χ^2 scores have the same distribution in the absence of any systematic over-representation. Thus, in examining over-representation, the χ^2 score has a very desirable feature: It is insensitive to the overall frequency of the corresponding oligomer.

As mentioned above, under the assumption that the expected number in each class satisfies $e_i \geq 5$, the χ^2 score for individual oligomers follows a χ^2 distribution with $D = 4$ degrees of freedom. By the Bonferroni inequality, because there are $N = 4l$ oligomers of length l, each with a χ^2 score, the maximum χ^2 score satisfies $P(\max \chi^2 \geq x) \leq N\, P(\chi^2 \geq x)$. The right side $P(\chi^2 \geq x)$ is a standard χ^2 tail probability and is available from a function built into Microsoft Excel. Our results indicate that the simple approximation $P(\max \chi^2 \geq x) \approx N\, P(\chi^2 \geq x)$ holds for probabilities up to $\sim P = 0.05$. Above $P = 0.05$, however, $P(\max \chi^2 \geq x)$ and $N\, P(\chi^2 \geq x)$ diverge as they become larger, making the P values from the χ^2 statistical test too conservative. Thus, in the absence of any theoretical reassurance, the stringent estimation of $P(\max \chi^2 \geq x)$ does require the mock data sets.

Recall that frequency insensitivity requires every class to have an expected number $e_i \geq 5$. This restriction places an upper limit on the choice of the length l of the oligomers. The oligomers should not be so long that they appear infrequently ($e_i < 5$) in five classes of 383 upstream regions, each of length 600. On the other hand, they should not be so short that they appear in almost every upstream region of length 600. Because $600 \leq 4^l \leq 383 \times 600/5$, giving $5 \leq l \leq 6$. We considered using pentamers ($l = 5$) instead of hexamers ($l = 6$), but upstream regions of length 600 contain too large a fraction of the $4^5 = 1024$ possible pentamers. Thus, theory indicates orientation-dependent, position-independent over-representation can be tested with hexamers, and we did so. There are $N = 4^6 = 4096$ hexamers.

The theoretical considerations above were borne out empirically by testing with pentamers and heptamers, which proved less effective than hexamers (data not shown). Additionally, it is interesting to note that many statistical surveys on DNA have been made on the basis of hexamers. Hutchinson (1996) used hexamer distributions to predict the location of upstream region and nonpromoter sequences in vertebrate DNA sequences. Fujibuchi and Kanehisa (1997) used a hexamer search as a basis for discriminating upstream regions of housekeeping genes from those of tissue-specific ones. Recently, van Helden et al. (1998) identified potential regulatory oligomers from groups of coregulated yeast genes using the theory of over-represented hexamers.

Many variants on our test are feasible. For example, we considered sampling our 1000 mock data sets from a sixth class, a control consisting of upstream regions that are not cell cycle regulated. We also considered adding an extra term in the χ^2 score to correspond to this sixth class. No feasible variant on the test described above, however, produced significant differences from the results presented.

Statistical Tests for Orientation-Independent, Position-Independent Over-Representation

If an oligomer within an upstream region can regulate, regardless of its position and orientation, we call it orientation-independent, position-independent. Because the over-representation is now orientation independent, each oligomer is considered the same as its reverse complement from the other DNA strand (e.g., ACGCGA is now the same as TCGCGT). For reasons similar to those given above, we tested orientation-independent, position-independent over-representation with hexamers ($l = 6$). Because of palindromes, there are $N = 1/2\ (4^6 + 4^3) = 2080$ distinct pairs of hexamers and reverse complements. For a given pair, we now count upstream regions that contain either member of the pair. Otherwise, all statistical calculations were carried out as described above for orientation-dependent, position-independent over-representation.

Statistical Tests for Orientation-Dependent, Position-Dependent Over-Representation

If an oligomer within an upstream region can regulate, but only in a particular position and particular orientation, we call it orientation dependent, position dependent. Because the over representation is now orientation dependent, each oligomer is considered different from its reverse complement from the other DNA strand (e.g., ACGCG is different from CGCGT). For the same theoretical reasons that led to hexamers ($l = 6$) above, this analysis was done with pentamers ($l = 5$). There are $N = 4^5 = 1024$ pentamers.

To detect orientation-dependent, position-dependent over-representation, we divided each upstream region into six nonoverlapping subsequences, numbered with respect to the ATG start codon: -1 to -103, -104 to -202, -203 to -301, -302 to -400, -401 to -499, and -500 to -598. This division should increase statistical power when searching for position-dependent upstream region elements, although it does suffer from inevitable difficulties with edge effects at the ends of the subsequences. Conceptually, for the purposes of statistical testing, each of the 6 subsequences is a set of 99 pentamers. The upstream region subsequences therefore give $383 \times 6 = 2298$ different sets of pentamers, falling into $5 \times 6 = 30$ different classes with respect to cell cycle regulation and position.

The statistical computations were carried out as described above for orientation-dependent, position-independent over-representation, with the following exceptions. To sample the parent population, we created the 1000 mock data sets by reclassifying the 2298 different sets of pentamers at random into the 30 different classes. The corresponding χ^2 score had 30 terms and $d = 29$ degrees of freedom. In the Bonferroni inequality, $N = 4^5 = 1024$.

Statistical Tests for Orientation-Independent, Position-Dependent Over-Representation

If an oligomer within an upstream region can regulate, regardless of its orientation but only in a particular position, we call it orientation independent, position dependent. Because the over-representation is now orientation independent, each oligomer is considered the same as its reverse complement from the other DNA strand (e.g., ACGCG is now the same as CGCGT). For reasons given above, we tested orientation-independent, position-dependent over-representation with pentamers ($l = 5$). Because pentamers cannot have reverse complement palindromes, there are $N = 1/2\ (4^5) = 512$ distinct pairs of pentamers and reverse complements. For a given pair, we now count upstream regions that contain either member of the pair.

Otherwise, to detect orientation-independent, position-dependent over-representation, we proceeded as described above for orientation-dependent, position-dependent over-representation, with upstream region subsequences representing $383 \times 6 = 2298$ different sets of pentamers, falling into $5 \times 6 = 30$ different classes with respect to cell cycle regulation and position. We created the 1000 mock data sets by reclassifying the 2298 different sets of pentamers at random into the 30 different classes. The corresponding χ^2 score had 30 terms and $d = 29$ degrees of freedom. In the Bonferroni inequality, $N = 4^5 = 512$.

The Kolmogorov-Smirnov Test for Assessing Nonuniform Position

We were also interested in determining and/or confirming the position dependence of any of the over-represented oligomers identified above, even if the position dependence was not apparent with the χ^2 score. The null hypothesis here is that if the oligomer has no position dependence, it will be uniformly positioned along the upstream region. There is a standard method to assess the uniformity of oligomer placement. We performed this analysis only if an oligomer had been statistically significant with $P \leq 0.2$ under the previous tests, and we tested its position dependence only in the upstream regions in which it was over-represented. To perform the Kolmogorov-Smirnov test, at each position from -1 to -600, we counted the number of occurrences of the oligomer. We then converted this to a cumulative score at each position giving the number of counts at the position or prior to it. The cumulative scores were then normalized by dividing by the total number of counts. The normalized cumulative score was then compared with the expected cumulative score corresponding to a uniform distribution of that oligomer across the range -1 to -600. Thus, we calculated

$$D = \max|F_0(X) - S_N(X)|$$

where $F_0(X)$ is the theoretical cumulative distribution, and $S_N(X)$ is the cumulative distribution of the observed values. The statistical significance of D was determined by first calculating it from a formula (Kendall and Stuart 1979), and then multiplying the resulting number by the number of independent Kolmogorov-Smirnov tests performed (55 in this analysis) to account for multiple testing. The test detects when an oligomer deviates from uniform random placement. However, it does not indicate which are the intervals of over- or under-representation.

ACKNOWLEDGMENTS

We thank Jo McEntyre and Francis Ouellette for a critical reading of the manuscript.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Brazma, A., I. Jonassen, J. Vilo, and E. Ukkonen. 1998. Predicting gene regulatory elements in silico on a genomic scale. *Genome Res*. **8:** 1202–1215.

Cho, R.J., M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart, and R.W. Davis. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2:** 65–73.

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

DeRisi, J.L., V.R. Iyer, and P.O. Brown. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278:** 680–686.

Fujibuchi, W. and M. Kanehisa. 1997. Prediction of gene expression specificity by promoter sequence patterns. *DNA Res*. **4:** 81–90.

Godambe, S.A., D.D. Chaplin, T. Takova, L.M. Read, and C.J. Bellone. 1995. A novel cis-acting element required for lipopolysaccharide-induced transcription of the murine interleukin-1 beta gene. *Mol. Cell. Biol*. **15:** 112–119.

Hahn, S., S. Buratowski, P.A. Sharp, and L. Guarente. 1989. Yeast TATA-binding protein TFIID binds to TATA elements with both consensus and nonconsensus DNA sequences. *Proc. Natl. Acad. Sci*. **86:** 5718–5722.

Heinemeyer, T., X. Chen, H. Karas, A.E. Kel, O.V. Kel, I. Liebich, T. Meinhardt, I. Reuter, F. Schacherer, and E. Wingender. 1999. Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. *Nucleic Acids Res*. **27:** 318–322.

Hodges, P.E., A.H.Z. McKee, B.P. Davis, W.E. Payne, and J.I. Garrels. 1999. The yeast proteome database (YPD): A model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res*. **27:** 69–73.

Holstege, F.C., E.G. Jennings, J.J. Wyrick, T.I. Lee, C.J. Hengartner, M.R. Green, T.R. Golub, E.S. Lander, and R.A. Young. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95:** 717–728.

Hutchinson, G.B. 1996. The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *Comput. Appl. Biosci*. **12:** 391–398.

Kendall, M. and A. Stuart. 1979. *The advanced theory of statistics*. Griffin, London, UK.

Koch, C. and K. Nasmyth. 1994. Cell cycle regulated transcription in yeast. *Curr. Opin. Cell. Biol*. **6:** 451–459.

Kunzler, M., C. Springer, and G.H. Braus. 1996. The transcriptional apparatus required for mRNA encoding genes in the yeast Saccharomyces cerevisiae emerges from a jigsaw puzzle of transcription factors. *FEMS Microbiol. Rev*. **19:** 117–136.

Lew, D.J., T. Weinert, and J.R. Pringle. 1997. Cell cycle control in Saccharomyces cerevisiae. In *The molecular and cellular biology of the yeast* Saccharomyces (ed. J.R. Pringle, J.R. Broach, and E.W. Jones), Vol. 3, pp. 607–695. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

McIntosh, E.M. 1993. MCB elements and the regulation of DNA replication genes in yeast. *Curr. Genet*. **24:** 185–192.

Mewes, H.W., K. Heumann, A. Kaps, K. Mayer, F. Pfeiffer, S. Stocker, and D. Frishman. 1999. MIPS: A database for genomes and protein sequences. *Nucleic Acids Res*. **27:** 44–48.

Moskvina, E., C. Schuller, C.T. Maurer, W.H. Mager, and H. Ruis. 1998. A search in the genome of Saccharomyces cerevisiae for

genes regulated via stress response elements. *Yeast* **14:** 1041–1050.

Nolan, E.M., T.C. Cheung, D.W. Burton, and L.J. Deftos. 1996. Transcriptional regulation of the human chromogranin A gene by its 5′ distal regulatory element: Novel effects of orientation, structure, flanking sequences, and position on expression. *Mol. Cell. Endocrinol.* **124:** 51–62.

Partridge, J.F., G.E. Mikesell, and L.L. Breeden. 1997. Cell cycle-dependent transcription of CLN1 involves swi4 binding to MCB- like elements. *J. Biol. Chem.* **272:** 9071–9077.

Pfaff, S.L. and W.L. Taylor. 1998. Xenopus TFIIIA gene transcription is dependent on cis-element positioning and chromatin structure. *Mol. Cell. Biol.* **18:** 3811–3818.

Roth, F.P., J.D. Hughes, P.W. Estep, and G.M. Church. 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* **16:** 939–945.

Schuldiner, O., C. Yanover, and N. Benvenisty. 1998. Computer analysis of the entire budding yeast genome for putative targets of the GCN4 transcription factor. *Curr. Genet.* **33:** 16–20.

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

Struhl, K. 1995. Yeast transcriptional regulatory mechanisms. *Annu. Rev. Genet.* **29:** 651–674.

Tamayo, P., D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub. 1999. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci.* **96:** 2907–2912.

Treger, J.M., A.P. Schmitt, J.R. Simon, and K. McEntee. 1998. Transcriptional factor mutations reveal regulatory complexities of heat shock and newly identified stress genes in Saccharomyces cerevisiae. *J. Biol. Chem.* **273:** 26875–26879.

van Helden, J., B. Andre, and J. Collado-Vides. 1998. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* **281:** 827–842.

Wodicka, L., H. Dong, M. Mittmann, M.H. Ho, and D.J. Lockhart. 1997. Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nat. Biotechnol.* **15:** 1359–1367.

Received March 9, 1999; accepted in revised form June 18, 1999.