A COMPUTATIONAL "GENOME WALK" TECHNIQUE TO IDENTIFY REGULATORY INTERACTIONS IN GENE NETWORKS.

ANDREAS WAGNER
The Santa Fe Institute
1399 Hyde Park Road, Santa Fe, NM 87501, U.S.A.
Phone: +1-505-984-8800 Ext. 231; E-mail: aw@santafe.edu

To delineate the astronomical number of possible interactions of all genes in a genome is a task for which conventional experimental techniques are ill-suited. Sorely needed are rapid and inexpensive methods that identify candidates for interacting genes, candidates that can be further investigated by experiment. The subject of this paper is the application of a novel method to the genome of the yeast *Saccharomyces cerevisiae*. The method applies to an important class of gene interactions, that is, transcriptional regulation via transcription factors (TFs) that bind to specific enhancer or silencer sites on DNA. The method addresses the question: which of the genes in a genome are likely to be regulated by one or more TFs with known DNA binding specificity? It takes advantage of the fact that many TFs show cooperativity in transcriptional activation which manifests itself in closely spaced TF binding sites. Such "clusters" of binding sites are very unlikely to occur by chance alone, as opposed to individual sites, which are often abundant both in the genome and in promoter regions. Statistical information about binding site clusters in the genome, can be complemented by information about (i) known biochemical functions of the TF, (ii) the structure of its binding site, and (iii) function of the genes near the cluster, to identify genes likely to be regulated by a given transcription factor. Previously, binding sites of well characterized transcription factors in *Saccharomyces cerevisiae* were analyzed. Here, the method is applied to a somewhat different situation: the yeast DNA binding activity yE2F, similar to the mammalian transcription factor E2F. yE2F has a DNA binding specificity identical to E2F, and its binding site shows UAS activity in a GAL1-based promoter construct. However, despite its high conservation, the *in vivo* function of yE2F is unknown. The analysis carried out here suggests candidate genes for regulation by yE2F.

1 Introduction

Our ability to extract biologically important information about gene interactions from genome sequences is still quite limited. Most of the biological interpretation of genome sequences pertains to the number and types of genes in an organism. Sorely needed are novel approaches that permit the formulation of experimentally testable hypotheses about gene interactions from sequence data alone. Such approaches could vastly improve efficacy of experiments by pointing out likely candidates for interacting genes. In devising such tools, the fundamental question is: what types of gene interactions leave traces on

the DNA, traces that could lead to the identification of interacting gene products. Maybe the prime candidate for such interactions is the transcriptional regulation of protein coding genes in eukaryotes. Here, transcription factors (TFs) bind enhancer sequences near the coding region of a gene, recruit a basal transcription machinery to the transcription initiation site, and activate the transcription of the gene (1). Alternatively, TFs can repress transcription of a gene by interfering with the basal transcription apparatus in various ways (2). The common theme is that the binding of TFs to specific, often short sequences on the DNA is necessary for transcriptional regulation. Undoubtedly the predominant mechanism regulating gene expression in eukaryotes, transcriptional regulation accounts for an enormous number of gene interactions. The availability of an efficient tool for the analysis of genes that are regulated by a given TF would thus permit analysis of a significant part of the global network of gene interactions.

To simply look for binding sites of specific TFs near a gene to identify candidate genes for regulation by a TF is problematic. For example, the minimally functional binding site of the heat shock transcription factor (4,5) occurs more than 10^6 times in the genome of *S. cerevisiae* (unpubl. obs.). The promoters of most genes would contain one or more such binding sites, making any biological conclusions based on binding site occurrence meaningless. Is there a modification of this simple approach that would render it useful? It has long been recognized that most transcriptional regulators display (homotypic or heterotypic) cooperative interactions, either when binding to DNA, or when activating transcription. Cooperativity is often reflected in the occurrence of multiple closely spaced binding sites on the DNA (6). The approach introduced below takes advantage of the ubiquity of cooperative interactions to identify genes putatively regulated by given TFs. Its basic tenet is that groups ("clusters") of TF binding sites linked much more tightly than expected by chance alone, are probably relevant to the transcriptional regulation of a nearby gene. The central problem is to find a statistically sensible definition of a highly significant cluster of binding sites. In only accepting the statistically most significant groups of binding sites, it is attempted to minimize the method's false positive rate, that is, the rate of identifying candidate genes for regulation by a TF that turn out not to be regulated by the factor. However, the price paid for such conservativism is that many genes regulated by a TF may not be detected. It is a price well worth paying, given that a conservative approach will generate candidate genes that seriously merit further experimental investigation.

A well known general problem in the analysis of DNA sequences is the enormous heterogeneity of sequence composition, which violates assumptions

needed for most conventional statistical techniques (7,8). Any statistical approach to the analysis of DNA sequences will thus provide only a crude assessment, of sequence properties. The method used here can not altogether avoid the problems of sequence heterogeneity, but it attempts to alleviate them by taking both global (genome-wide) and local sequence properties into account.

While the technique is applicable to any eukaryote, it is here illustrated with the genome of *S. cerevisiae.* The reasons for this choice are outlined in (9), a paper that also illustrates several applications of the method to known transcription factors. The application illustrated here regards a well characterized DNA binding activity whose *in vivo* function in *S. cerevisiae* is unknown. The reasons why this factor is interesting for the type of analysis carried out here are (i) its binding specificity is virtually identical to that of a mammalian transcription factor (E2F; ref. 10) involved in cell-cycle regulation, (ii) its binding site acts as a UAS sequence in a GAL1-reporter construct in *S. cerevisiae*, and (iii) its activity or that of a closely related factor is cell-cycle regulated (11). These findings suggest that a transcription factor similar to E2F may exist in *S. cerevisiae.* However, no genes regulated *in vivo* by this putative factor are known. Statistically highly significant clusters of yE2F binding sites in the promoter region of several yeast genes suggest candidate genes for regulation by yE2F. Needless to say, all these candidates have to be tested experimentally. However, while tentative, the results presented here provide a relatively inexpensive way to identify the most promising candidates among the enormous number of genes that yE2F might potentially regulate *in vivo.*

2 Statistical Methods

This section illustrates the statistical techniques used to identify highly significant clusters of transcription factor binding sites. The general approach has three steps. First, significant clusters of particular binding sites are detected by what is referred to as a "genome walk" analysis. Second, some of the clusters thus identified are eliminated from further consideration because of their location in the genome. Third, the statistical significance of the remaining clusters is reassessed on the basis of local sequence composition. Both the first and the third step critically depend on methods to estimate the probability of binding site occurrence on the DNA. These methods are therefore discussed first. Then, the three steps are explained in greater detail.

Estimates of the probability of site occurrence. What is the probability that a random oligonucleotide with compositional features similar to those of genomic DNA, and with the same length as the binding site of interest, matches that site? To ensure wide applicability of the technique, conventional consensus

sequences are used here instead of position weight matrices (PWMs, [12-13]) for binding sites, because very few transcription factors are sufficiently well characterized to allow construction of a PWM. When addressing the above question, one has to take into account that functional transcription factor binding sites S (i) may occur in either orientation on the DNA (the reverse complement of a site S will be denoted as $\overline{S}$), (ii) may have relaxed sequence requirements at some positions, as reflected by standard IUB nucleotide codes (14), (iii) in addition to such 'ambiguous' positions, may show a substantial number of mismatches to their consensus binding site.

The relative frequency of a binding site S of length l (an l-word) in a DNA sequence of N nucleotides is denoted by p_S, and determined by dividing the number of word occurrences N_S in that sequence by the maximally possible number $N - l + 1$, i.e.,

$$p_S = \frac{N_S}{N - l + 1} \qquad [1]$$

Special cases are the mono- and dinucleotide frequencies p_A, p_C, p_G, p_T, p_{AA}, ..., p_{TT}. The relative frequencies of a word with exactly k or at most k mismatches to a given word S of the same length are denoted as p_{S^k}, and $p_{S\leq k}$, respectively, where $p_S = p_{S^0}$. Obviously,

$$p_{S\leq k} = \sum_{i=0}^{k} p_{S^k}. \qquad [2]$$

Statistical estimators of the probabilities of word occurrence will be denoted as $\hat{p}_S$, $\hat{p}_{S^k}$, and $\hat{p}_{S\leq k}$.

Global estimator based on site counts. Here, the estimator $\hat{p}_{S\leq k}$ of site occurrence probability is the relative frequency $p_{S\leq k}$, as determined by [1] and [2], for an admissible number of mismatches, k. Under the Poisson model of site distribution, where the probability of observing k sites in a DNA sequence of length N is given by

$$Prob(k) = \exp(-\lambda)\frac{\lambda^k}{k!}, \qquad [3]$$

$\hat{p}_S = p_S$ is a maximum likelihood estimator of the distribution parameter λ. One has to count a large number of sites to ensure a narrow confidence interval for this estimator (15). To maximize site count, $\hat{p}_S$ was not estimated for each yeast chromosome separately, but for all 16 chromosomes together.

Local estimators based on mono- and dinucleotide frequencies. These estimators (detailed in ref. 9) assume that the statistical structure of DNA in a local region of interest can be described by a first order Markov chain, whose transition probabilities are estimated from the base composition in that region.

The next three sections list the principal steps of the statistical analysis carried out here.

Step 1: Identification of binding site clusters by genome walk analysis. The most simple, albeit problematic, null-hypothesis of binding site distribution is the Poisson approximation [3]. Very short sites or sites with a repetitive structure (e.g., 5'-GGGGG-3') will not follow a Poisson distribution (9) but, this is not a problem for the site studied here (see the next section). The second reason for deviations from the Poisson approximation is compositional heterogeneity and the complex statistical structure of DNA. It is addressed in step 3 below. In step 1, however, statistically significant clusters of transcription factor binding sites are identified by testing site spacing against the null-hypothesis of a Poisson distribution.

Denote as $X_i, \ldots, X_n$ the positions at which a site S or its reverse $\overline{\text{S}}$ complement are encountered on the DNA. Further, define as X_0 the beginning (5' end of the top strand) of the DNA sequence. The quantity

$$D_{i,j} = X_j - X_i$$

denotes the distance between site X_j and X_i.

$$D_{i,i+k-1} = \sum_{j=0}^{k-2} D_{i+j,i+j+1} \quad k > 1, \qquad [4]$$

is the length of a stretch of DNA spanning exactly k words. It will be referred to as a k-cluster. Under the Poisson null-hypothesis [3], the distribution of the distance between successive words, $D_{i,i+1}$, is exponential with density

$$\lambda e^{-\lambda z} \qquad [5]$$

This is the probability distribution of the length of 2-clusters. More generally, the length of k-clusters follows a Pearson Type III distribution with density

$$\frac{\lambda}{\Gamma(k-1)}(\lambda z)^{k-2}e^{-\lambda z} \quad k > 1, \qquad [6]$$

where $\Gamma(k) = (k-1)!$ is the gamma function. This is easily seen from the characteristic functions of [5] and [6] (16). The probability of observing a k-cluster of length less than x is

$$Prob(D_{i,i+k-1} < x) = \frac{\lambda}{\Gamma(k-1)} \int_0^x (\lambda z)^{k-2}e^{-\lambda z}dz. \qquad [7]$$

To assess whether the length, x, of an observed k-clusters, $D_{i,i+k-1}$, is shorter than would be expected "by chance alone" under the null-hypothesis, and for a given significance level **P**, [7] is used to determine whether

$$Prob(D_{i,i+k-1} < x) < \mathbf{P} \qquad [8]$$

The appropriate choice of **P** is discussed below.

The parameter λ needed in the above statistical tests was estimated here via relative site frequencies in the genome. However, from each pair of overlapping sites only one site was (randomly) chosen, and included in the absolute site count $N_S + N_{\overline{S}}$, for reasons explained in (9). Starting at X_0, the lengths of all k-clusters up to $k = 11$, i.e., $D_{0,1}, D_{0,2}, \ldots, D_{0,10}$, was determined. If for any of these k-clusters [8] was true, the cluster was retained for further analysis. This procedure was repeated for clusters starting at X_1 ($D_{1,2}, D_{1,3}, \ldots, D_{1,11}$), X_2, through X_{n-11}, hence the name "genome walk" analysis.

What is the appropriate choice of a significance threshold **P** for this method? Here, it is important to take into account the often large number of significance tests carried out. For example, for a TF with a genomic site count of $N_S + N_{\overline{S}} = 5000$, there are approximately 500 non-overlapping 10-clusters, and thus 500 independent significance tests for 10-clusters. A value of $\mathbf{P} = .05$ or $\mathbf{P} = .01$ would lead to high type I error probability. The problem of choosing an appropriate significance level is aggravated by the fact that many of the significance tests are carried out for overlapping clusters of binding sites, and are therefore not independent. The approach chosen here is to make **P** dependent on the specific size k of a k-cluster. More specifically $\mathbf{P} = (k-1)/(N_S + N_{\overline{S}})$ will be used as a significance threshold for any k-cluster. This value is chosen because it makes **P** approximately equal to the number of non-overlapping k-clusters, i.e., approximately equal to the number of independent statistical tests carried out. In other word, with this value of **P** one would expect, for any given k, of the order of one false positive k-cluster, i.e., a cluster for which the null-hypothesis is falsely rejected (a type I error, 17).

Step 2: Elimination of some statistically significant clusters. Yeast transcriptional regulators function in general only when bound upstream of the coding region (9), with the possible exceptions of the transcription of Ty retrotransposons (18). Moreover, regulatory regions that lie interspersed among various genes and in enormous distances from the gene they regulate seem to be absent or infrequent in *S. cerevisiae* (9). Thus, statistically significant clusters were not considered further, if they (i) overlapped or were located inside exons, and (ii) if they occurred downstream of both adjacent open reading frames (ORFs).

Step 3: Analysis of remaining clusters based on local sequence com-

position. Estimating λ via actual site counts in step 1 is necessary because global sequence composition is a poor predictor of site occurrence (19). However, local biases in sequence composition may affect the local probabilities of site occurrence, and thus the actual significance of the detected clusters. Thus, in the last step of the analysis, DNA mono- and dinucleotide composition was analyzed in each of the remaining clusters, or in a 500 bp window centered around the cluster, whichever was longer. Two new estimates of λ, based on mono- and dinucleotide distributions in these regions were used to reassess the significance [8] of the clusters remaining after step 2. In statistical terms, the underlying null hypothesis is that site distribution in the genome follows an inhomogeneous Poisson process, i.e., a Poisson process whose parameter $\lambda = \lambda(y)$ is a function of the location y in the genome (20). Higher order correlations among nucleotides were not taken into account for reasons of computational feasibility.

3 Results and Discussion

The "genome walk" technique presented in the previous section can be used to detect statistically highly significant clusters, that is, groups of very tightly linked binding sites of a transcription factor. Because the cooperativity of many eukaryotic transcription factors is reflected in the occurrence of such binding site clusters, genes in the vicinity of a cluster are good candidates for regulation by the respective factor.

Mammalian E2F is a transcription factor which regulates a number of genes implicated in DNA replication and cell-cycle control. It interacts with members of the retinoblastoma protein family (e.g., 21), and its activity is regulated by cyclin-dependent kinases (22). Because central features of the cell-cycle are conserved across eukaryotes, it would seem natural to search for similar transcription factors in organisms where E2F-like activities have not yet been found. Such a search was carried out successfully in *S. cerevisiae* (10, 11) which appears to encode at least one transcription factor with a DNA binding activity identical to that of mammalian E2F at the AdE2-promoter (5'-GCGCGAAA-3'), a binding activity that has been named yE2F (10) and SCELA (11). The binding site acts as an UAS element for a GAL1-reporter gene. Multiple binding sites in the upstream region of the reporter gene lead to a massive increase in transcriptional activation, suggesting that the factor shows cooperativity in DNA binding and/or transcriptional activation. DNA binding of the factor or of a closely related activity is cell-cycle dependent, reaching its peak at the G_1/S transition (11). These findings suggest that an E2F-like transcription factor exists in *S. cerevisiae*, and that it may play a

role in the regulation of the cell-cycle. However, no genes regulated *in vivo* by yE2F have been characterized.

The ability of yE2F to activate transcription cooperatively makes it an ideal object for the genome walk analysis. However, it is useful to first consider some global statistics of binding site distribution. It was discussed above that sites which do not follow a Poisson distribution in random DNA can not be analyzed with this method. To exclude this possibility, it was tested whether distances between yE2F binding sites (5'-GCGCGAAA-3', allowing one mismatch to the consensus) follow an exponential distribution in a long (14Mb) random DNA sequence with the same nucleotide composition as yeast. The distribution parameter λ was estimated via [**1**] and [**2**]. Both the result of a likelihood ratio test (G=11.91, 9 df; $0.1 < \mathbf{P} < 0.5$) and a chi square test ($\chi^2 = 13.03$, 9 df; $0.1 < \mathbf{P} < 0.5$) for exponential distribution of inter-site distances (9) are consistent with a Poisson distribution in random DNA.

The nuclear genome of *S. cerevisiae* contains 4328 non-overlapping sites with no more than one mismatch to the consensus 5'-GCGCGAAA-3', with a mean site distance of 2839 base pairs. 1674 of these sites are located in non-coding regions. This number illustrates the very limited use of identifying candidate genes regulated by a given transcription factor on the basis of the occurrence of individual binding sites in promoter regions. Hundreds of (mostly spurious) candidate genes for regulation by yE2F would have been identified.

There are various reasons why site distribution in genomic DNA might deviate from a Poisson, such as compositional heterogeneity, or excessive clumping of sites on the DNA. However, the distribution of inter-site distances in the genome is remarkably consistent with an exponential distribution (G=4.85, 8 df; $0.5 < \mathbf{P} < 0.9$; $\chi^2 = 4.98$, 8 df; $0.5 < \mathbf{P} < 0.9$). However, notice that a goodness-of-fit test to an exponential distribution provides only a very crude assessment of site distribution properties. This is because (i) a large amount of distance information (cf. the number of sites above) is pooled into a small number of bins, and (ii) no site distances other than those among nearest neighbors are included in the test. A small number of clusters of closely spaced sites would probably go undetected by this test. The " genome walk" analysis is a fine grained assay more suitable to detect such clusters.

Genome walk. Figure 1 shows a representative example of the results obtained by the genome walk analysis for the yE2F binding site. Shown is a significance profile of all binding site clusters on chromosome 13 of *S. cerevisiae* (see also the figure legend). Peaks in the plot correspond to highly significant clusters, clusters whose constituent binding sites are very tightly linked. There is only one cluster, comprising $k = 4$ binding sites, on chromosome 13 whose significance ($\mathbf{P} = 6.72 \times 10^{-4}$) is higher than the threshold value of $\mathbf{P} =$

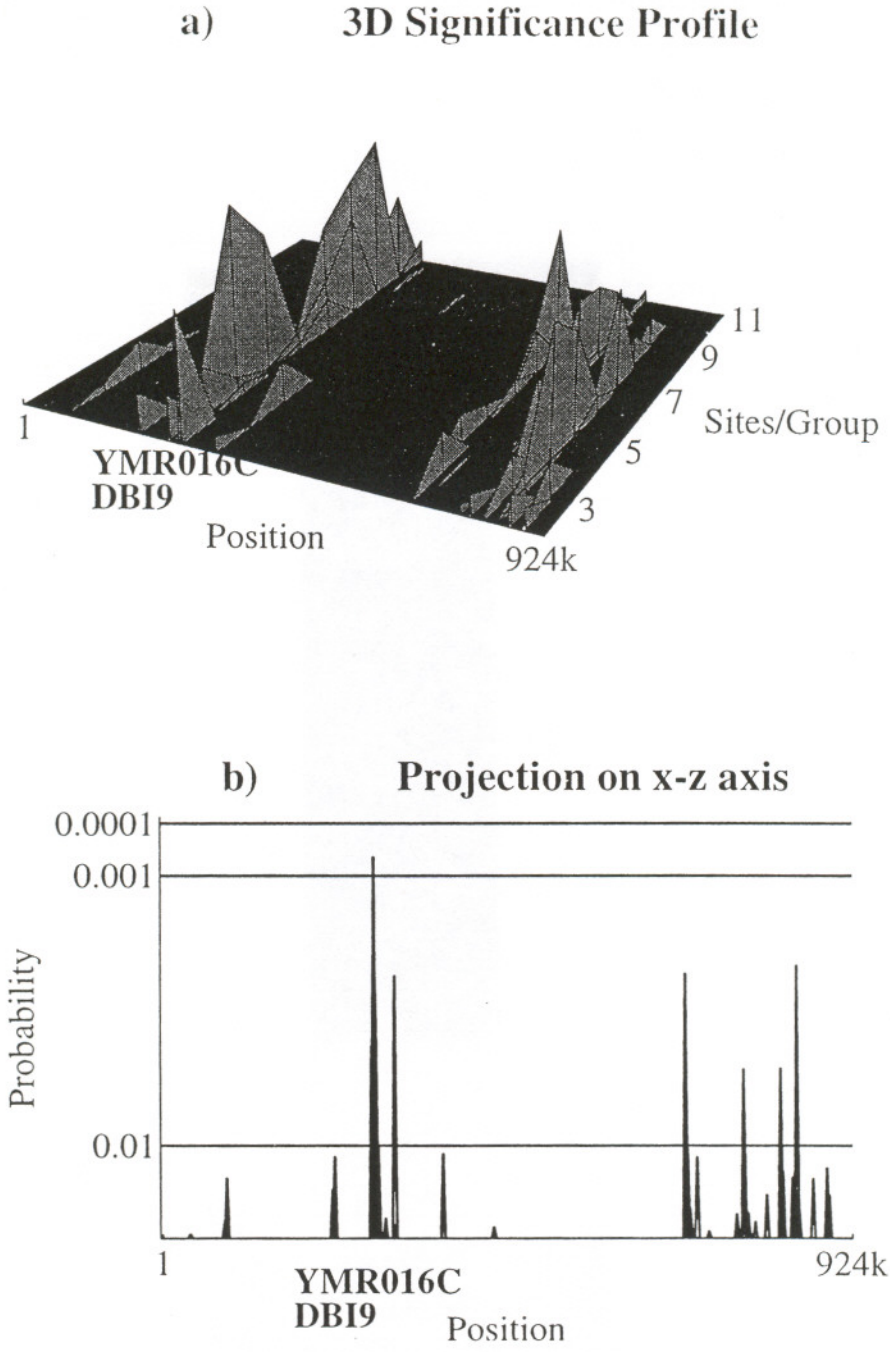

Fig. 1: Significance profile of yE2F binding site clusters on chromosome XIII. Panel **a)** shows the statistical significance of all groups of yE2F binding sites on chromosome XIII. Each point in the $x-y$ plane corresponds to a group of binding sites comprising the number of sites indicated on the y-axis (2-11), whose 5' most site starts at the position indicated on the x-axis. The origin (lower left corner) corresponds to the first group of two binding sites starting at the site closest to the left telomere of chromosome XIII. The z-axis shows a measure of the probability **P** of finding a group of sites spaced at the observed or a smaller distance under the assumption of the null-hypothesis. More precisely, the plotted values are $(1-\mathbf{P})^{150}$. Because of this transformation, (i) peaks on the plot correspond to highly significant clusters, and (ii) all but the most significant values will be effectively zero. Panel **b)** shows the same plot, but projected onto the $x-z$ plane. The abscissa indicates the position along the chromosome from left telomere (position 1) to right telomere (position 924430). The ordinate shows the **P**-values of clusters. Notice that there is only one cluster that exceeds the significance threshold used here. See text for details.

7.06×10^{-4} for 4-clusters. It lies between the ORF YMR016C, encoded on the bottom strand, and the gene DBI9, encoded on the top strand. yE2F could thus be involved in the regulation of either gene, or of both genes.

Table 1 summarizes the results of the genome walk analysis carried out for all 16 chromosomes. The first and second column of the table show chromosome number and gene (ORF) name, respectively. The third column of the table indicates the number of sites in the cluster, and the length of the cluster in base pairs. Two neighboring genes in the table are transcribed in opposite, divergent directions if only one value is given in this column, i.e., they share their 5' non-coding region. The fourth column shows the distance of the yE2F binding site closest to the start-codon. Only clusters of k binding sites located upstream of a protein coding gene that are significant at $\mathbf{P} \geq (k-1)/(N_S+N_{\overline{S}}) = (k-1)/4250$ are listed. It is obvious from the table, that, despite the thousands of potential yE2F binding sites in the genome, there is only a small number of significant binding site clusters in the promoter region of one or more genes. The clusters listed in the Table are included on the basis of their *global* **P**-values (given in column 5), which are calculated from genomic binding site counts. To take local nucleotide composition into account, cluster significance was reassessed based on local mono- und dinucleotide composition in the respective promoter region (columns 6 and 7 of Table 2). Dinucleotide composition is included here, because it is known to be an important factor in compositional heterogeneity (21). To avoid assigning a cut-off point to significance, all local **P**-values are listed. However, any cluster that shows a *local* **P**-value vastly higher than its *global* **P**-value should be approached with caution, and only included into further consideration if other evidence argues for its biological relevance.

When analyzing a well characterized yeast transcription factor, one can use information about its biological function to identify genes likely to be regulated by the factor (9). Because the *in vivo* role of yE2F is not known, this is not possible for the candidate genes listed in Table 1. However, the biological function of mammalian E2F may provide hints as to the nature of such genes, given that E2F and yE2F may have similar biological roles. For example, E2F activates the transcription of a human H2A histone gene during S-phase (22). It also regulates the expression of a mouse H2A gene (23). Remarkably, a highly significant group of yE2F binding sites is found on chromosome 4 of *S. cerevisiae*, in the vicinity of two genes encoding the yeast H2A and H2B genes. These genes are divergently transcribed, and yE2f might be involved in the regulation of one or both of the genes. Another example concerns the regulation of the murine gene Htf9-a, encoding a protein that interacts with the Ran GTPase, a member of the ras-superfamily. Transcription of Htf9-a appears to be regulated by E2F in a cell-cycle specific manner, reaching a peak

Table 1: Candidate genes for regulation by yE2F.

		Cluster Statistics		*Estimated Significance*			
Chr.	**ORF**	Sites/bp	Position	Glob.	Mono	Di	Gene Function or Structure[1]
4	**NHP10***	3/61	-221	1.72×10^{-4}	1.85×10^{-4}	3.9×10^{-4}	high mobility group like protein
4	**YDL001W**		-132				unknown
4	**HTB1**	3/58	-255	1.53×10^{-4}	2.72×10^{-4}	1.77×10^{-4}	histone H2B
4	**HTA1**		-513				histone H2A
4	**SAC7**	5/559	-83	5.06×10^{-5}	1.47×10^{-4}	1.1×10^{-3}	GTPase activating protein (GAP) for Rho1p, a GTP binding protein in the ras superfamily
7	**GOG5**	6/2062	-10	9.07×10^{-4}	3.32×10^{-4}	9.48×10^{-3}	implicated in glycosylation in Golgi apparatus
7	**YGL179C**	5/733	-451	1.44×10^{-4}	1.15×10^{-3}	7.84×10^{-3}	unknown
7	**MPT5**		-1085				interacts with Sst2p (a GAP), implicated in pheromone-induced growth arrest.
7	**YGL096W**	5/1025	-778	5.16×10^{-4}	1.61×10^{-3}	1.05×10^{-2}	unknown
11	**YKL102C**	5/351	-54	8.05×10^{-6}	2.15×10^{-5}	4.85×10^{-4}	unknown
11	**HSL1**		-156				negative regulator of Swe1p kinase (which regulates Cdc28p)
13	**YMR016C**	4/479	-606	6.72×10^{-4}	4.28×10^{-3}	3.82×10^{-2}	unknown
13	**DBI9****		-819				interacts with product of DBF2, a kinase required for late nuclear division

[1] from the *S. cerevisiae* genome database (http://genome-www.stanford.edu/Saccharomyces), see main text for further references
* non-standard name: HMO2, ** non-standard name: SPO20

during S-phase, and being down-regulated during growth arrest (24). The two *S. cerevisiae* genes SAC7 and MPT5 both are associated with highly significant clusters of E2F binding sites. Sac7p is a GTPase activating protein (GAP) of the GTP-binding protein Rho1p, itself a member of the ras-family. Mpt5p interacts with a GAP protein, and plays a role in pheromone induced growth arrest (25). Another interesting candidate gene may be HSL1, which is involved in the regulation of the SWE1 gene (26). The product of SWE1 inhibits the kinase Cdc28p, a protein central to cell cycle control in *S. cerevisiae*. Thus, HSL1 itself encodes probably a cell-cycle regulator. The likely involvement of yE2F in cell-cycle control makes HSL1 a good candidate gene for regulation by yE2F.

Obviously, no such functional criteria can be used to identify which of the five candidate ORFs with unknown function might actually be regulated by yE2F. However, it may sometimes be possible to exclude such ORFs on the basis of other criteria. For example, while the ORF YGL069W on chromosome 7 is associated with a 5-cluster highly significant ($\mathbf{P} = 5.16 \times 10^{-4}$) on the basis of global site distribution, its **P**-value based on local dinucleotide distribution is $\mathbf{P} \approx 10^{-2}$. Local base composition seems to favor binding site occurrence in this case, so that the high cluster significance apparent on the global level may be spurious. Another criterion that can sometimes be used to exclude genes based on features of the associated clusters is the distance of the cluster to the start-codon. For example, the downstream-most binding site of the cluster associated with the gene GOG5 on chromosome 7 ends only 2 base pairs upstream of the start-codon, and no statistically significant sub-group of binding sites exists further upstream of GOG5.

Obviously, any results obtained with this method are tentative and have to be validated experimentally. However, there is additional statistical evidence consistent with the hypothesis that significant clusters of yE2F binding sites may be associated with transcriptional regulation. Consider all clusters of significant binding sites, including clusters known to be overlapping with, or contained in ORFs. If the individual sites belonging to such clusters were randomly distributed among coding and non-coding regions, one would expect approximately 72 percent of the individual sites to occur in coding regions, because coding regions account for approximately 72 percent of the yeast genome (27). There are 215 sites belonging to significant clusters, 129 (86) of which are located in coding (non-coding) regions. A χ^2-test for the expected 72:28 distribution shows that these sites are much more likely to be found in non-coding regions ($\chi^2 = 15.36, 1df; \mathbf{P} = 8.9 \times 10^{-5}$). Could this simply be due to differences in the base composition of coding and non-coding regions, favoring site occurrence in non-coding regions? This seems

unlikely, based on the following calculation. Average mono- and dinucleotide distribution was determined for 1000 randomly chosen 1kb fragments located entirely in coding or non-coding regions. Based on the nucleotide distributions thus obtained, the probability of occurrence of yE2F sites in coding regions (CR) and non-coding regions (NCR) was estimated. The estimated values are $\hat{p}^{CR}_{\mathrm{yE2f}} = 4.47 \times 10^{-4}$ and $\hat{p}^{NCR}_{\mathrm{yE2f}} = 3.89 \times 10^{-4}$ based on mononucleotide composition, and $\hat{p}^{CR}_{\mathrm{yE2f}} = 4.09 \times 10^{-4}$ and $\hat{p}^{NCR}_{\mathrm{yE2f}} = 3.80 \times 10^{-4}$ based on dinucleotide composition. Thus, the estimated probability of yE2F binding site occurrence is slightly higher in coding regions, yet sites belonging to significant clusters tend to accumulate in non-coding regions. This might reflect (i) positive selection for clusters in non-coding regions where they can play a role in regulating gene expression, or (ii) negative selection eliminating clusters in coding regions, because the binding of several copies of a transcription factor inside an ORF may interfere with transcription.

4 Conclusions

As opposed to the large number of transcription factor binding sites in a genome, the number of significant clusters of sites may be very small. Such clusters also show unexpected features, such as their preferred occurrence in non-coding regions. These features and the fact that the method used here (i) detects genes whose regulation by a given transcription factor was demonstrated experimentally (12), and (ii) detects genes that are functionally related to genes regulated by similar transcription factors in other organisms, indicate its usefulness. However, the critical question regarding the method's false positive rate, that is, the fraction of candidate genes for regulation by a factor that turn out not to be regulated by that factor, can only be answered by experimentally testing its predictions.

Many further applications of the method are conceivable, other than applying it to all characterized transcription factor binding sites in yeast. For example, combinatorial regulation of a gene by different transcription factors has not been explored yet. It would require only a slight modification to the statistical approach. The method can also be applied to higher eukaryotes, where genomic sequences are now rapidly accumulating. Hopefully, a sensible combination of statistical and biological information, taking sequence composition at all levels of genome organization into account, will permit genomic DNA sequences to be useful beyond the mere identification of genes.

Acknowledgments

I would like to thank Bill Bruno, Patrik D'haeseleer, Catherine Macken, and David Torney for invaluable discussions on the subject of this paper. The financial support of the Santa Fe Institute is gratefully acknowledged.

References

1. Ptashne, M., Gann, A. (1997) *Nature*, **386**, 569-577
2. Levine, M., Manley, J.L. (1989) *Cell*, **59**, 405-408
3. Das, S., Yu, L., Gaitatzes, C., Rogers, R., Freeman, J., Blenkowska, J., Adams, R.M., Smith, T.F., Lindellen, J. (1997) *Nature*, **385**, 29-30
4. Sorger, P.K. (1991) *Cell*, **65**, 363-366
5. Srinivas, U.K., Swamynathan, S.K. (1996) *Journal of Bioscience*, **21**, 103-121
6. Ptashne, M. (1988) *Nature*, **335**, 683-689
7. Karlin, S., Brendel, V. (1993) *Science*, **259**, 677-680
8. Bernardi, G., Mouchiroud, D., Gautier, C., Bernardi, G. (1988) *Journal of Molecular Evolution*, **28**, 7-18
9. Wagner, A. (1997) *Nucleic Acids Research* **25**, 3594-3604
10. Mai, B., Lipp, B. (1993) *FEBS Letters*, **321**, 153-158
11. Vemu, S., Reichel, R.R. (1995) *Journal of Biological Chemistry*, **270**, 20724-20729
12. Stormo, G.D. (1990) *Methods in Enzymology*, **183**, 211-220
13. Fickett, J.W. (1996) *Molecular and Cellular Biology*, **16**, 437-441
14. IUB Nomenclature Committee (1985) *European Journal of Biochemistry*, **150**, 1-5
15. Kendall, M.G. (1952) *The advanced theory of statistics. Vol. II. p22* Griffin, London.
16. Abramowitz, M., Stegun, I.A. (1972) *Handbook of mathematical functions. 26.1.28, 26.1.31* Dover, New York.
17. Sokal, R.R. and Rohlf, F.J. (1981) *Biometry.* Freeman, New York
18. Türkel, S., Farabaugh, P.J. (1993) *Molecular and Cellular Biology*, **13**, 2091-2103
19. Karlin, S., Macken, C. (1991) *Nucleic Acids Research*, **19**, 4241-4246
20. Parzen, G. (1962) *Stochastic Processes. Ch. 4.2.* Holden-Day, San Francisco.
21. Karlin, S., Cardon, L.R. (1994) *Annual Reviews of Microbiology*, **48**, 619-654