3594–3604 *Nucleic Acids Research, 1997, Vol. 25, No. 18*

A computational genomics approach to the identification of gene networks

Andreas Wagner*

The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

Received June 5, 1997; Revised and Accepted August 4, 1997

ABSTRACT

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. Such a method is introduced here for an important class of gene interactions, i.e., transcriptional regulation via transcription factors (TFs) that bind to specific enhancer or silencer sites. 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 in the genome. Here, statistical information about binding site clusters in the genome, is 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. Several applications are illustrated with the genome of *Saccharomyces cerevisiae*, and four different DNA binding activities, SBF, MBF, a sub-class of bHLH proteins and NBF. The technique may aid in the discovery of interactions between genes of known function, and the assignment of biological functions to putative open reading frames.

INTRODUCTION

The ultimate challenge to molecular biology is to identify and fully characterize the complete network of interactions among genes and their products in an organism. In facing this challenge, the wealth of information created by genome sequencing efforts will be an invaluable resource. However, 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. The advantage of such approaches are clear. They 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. It would put cell biology a large step closer to its ultimate goal.

Naively, one might assume that it is sufficient to look for binding sites of specific TFs near a gene to identify candidate genes for regulation by the TF. This approach is standard practice on a small scale, and its extension to entire genomes is straightforward (3). However, for many known enhancer sites, it is also deeply 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 *Saccharomyces cerevisiae* (unpublished observation). 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 approach that would render it useful? It has long been recognized that most transcriptional regulators display (homotypic or heterotypic) cooperative interactions, either when binding DNA, or when activating transcription. Cooperativity is usually 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

* To whom correspondence should be addressed. Tel: +1 505 984 8800; Fax: +1 505 982 0565; Email: aw@santafe.edu

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. It will be seen below that common plausibility arguments about the significance of binding site clusters can be quite misleading, if one takes the genome-wide distribution of binding sites into account. In only accepting the statistically most significant groups of binding sites, it is attempted to minimize the method's false positive rate. In addition, various sources of biological information are incorporated into the analysis, information that is likely to decrease this rate further. 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 cannot altogether avoid the problems of sequence heterogeneity, but 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 main reasons are that potential yeast promoter regions are in general short and located upstream of the coding region (9,10), and that the yeast genome does not contain many tandemly repeated sequences other than rDNA and CUP1 genes (11). Four different applications are illustrated with different yeast DNA binding proteins. They include, but are not limited to the identification of novel interactions among genes of known function, and the putative assignment of biological function (cell cycle regulation, etc.) to ORFs with unknown function. The particular choice of four DNA-binding proteins (out of the ~75 characterized to date) was motivated by (i) their well characterized DNA binding sites, (ii) the length of their binding sites (for methodological reasons discussed below), and (iii) the variety of applications that they can illustrate. Needless to say, all candidate gene interactions identified by the method have to be tested experimentally. However, while tentative, the results may aid in sifting through the astronomical number of possible gene interactions, and identify candidates worthy of experimental investigation.

STATISTICAL METHODS

This section illustrates the statistical techniques used to identify highly significant clusters of transcription factor binding sites which are then further analyzed using biological information about the respective transcription factors. 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. By taking both global and local sequence properties into account, it is attempted to alleviate problems caused by compositional heterogeneity of DNA. 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 (i) may occur in either orientation on the DNA, (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.,

$$p_S = \frac{N_S}{N-l+1}. \qquad \mathbf{1}$$

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_{Sk}, and $p_{S\le k}$, respectively, where $p_S = p_{S0}$. Obviously,

$$p_{S\le k} = \sum_{i=0}^{k} p_{Sk}. \qquad \mathbf{2}$$

The corresponding statistical predictors of the probabilities of word occurrence will be denoted as $\hat{p}_S$, $\hat{p}_{Sk}$ and $\hat{p}_{S\le k}$.

Global predictor based on site counts. Here, the predictor $\hat{p}_{S\le k}$ of site occurrence probability is the relative frequency $p_{S\le k}$, as determined by equations **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

$$Prob(k) = \exp(-\lambda)\frac{\lambda^k}{k!}. \qquad \mathbf{3}$$

$\hat{p}_S = p_S$ (given by equation **1**) 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 λ (15). Given that many transcription factor binding sites are longer than 10 bases (16), very large amounts of sequence may have to be analyzed to ensure a narrow confidence interval. To maximize site count, $\hat{p}_S$ was not determined for each yeast chromosome separately, but for all 16 chromosomes together.

Prediction based on mononucleotide frequencies. For an oligonucleotide generated by independently and randomly selecting successive letters from an underlying alphabet, the predicted probability $\hat{p}_S$ is simply the product of the letter frequencies, p_A,

. . . , p_T. $p_{S \le k}$ is calculated via equation **2**. To calculate individual $\hat{p}_{Si}$'s, one sums the respective probabilities over all *i*-tupels of positions where *i* mismatches can occur. For example, to calculate $\hat{p}_{S2}$ for the 8-word 5′-CACWANAA-3′, one has to sum over $\binom{8-1}{2} = 21$ configurations of sites at which two mismatches can occur. To predict the probability of finding a word with mismatches at positions, say, 1 and 4, one calculates $(1 - p_C)p_A p_C(1 - p_W)p_A\, p_A\, p_A$, where $p_W = p_A + p_T$.

Prediction based on dinucleotide frequencies. In this case a DNA sequence is viewed as a sequence of letters generated as a first-order Markov chain (17). The probability of finding a particular word *S*, say 5′-CACTAA-3′ is then predicted as

$$\hat{p}_S = \frac{p_{CA}p_{AC}p_{CT}p_{TA}p_{AA}}{p_A p_C p_T p_A}.$$

For words *S* containing positions with relaxed sequence requirements (*W*, *N* etc.), and *k* permissible mismatches to the consensus, all words were explicitly generated that fulfill the sequence requirements, and contain only letters *A* through *T*. Their respective probabilities were calculated using the above formula with observed mono- and dinucleotide frequencies, and added to obtain $p_{S \le k}$.

So far, for all three predictors, only the probability of encountering the word *S*, and not that of its equally functional reverse complement $\overline{S}$ was given. For palindromic words, where $S = \overline{S}$, and for $k = 0$ allowed mismatches, the predicted probability of encountering the word or its reverse complement is simply $\hat{p}_S$ itself, because whenever *S* occurs, $\overline{S}$ will occur as well. For non-palindromic words, and for $k > 0$, the situation is more complicated because there may be non-palindromic words, e.g., 5′-GAWTTC-3′, that admit some palindromic matches, 5′-GAATTC-3′, and some non-palindromic matches, 5′-GATTTC-3′. In such cases, the quantity $\hat{p}_S + \hat{p}_{\overline{S}}$ will over-estimate word probability by as much as a factor of two, because it counts the palindromic word occurrences twice. However, because the binding sites to be analyzed below are either perfect palindromes, or contain features that prohibit palindromic matches, such as strong asymmetries, overestimation of site probability is not likely to be a problem here.

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 (equation **3**). It can be violated for two reasons, the first of which is the structure of the sites themselves. Very short sites, long sites in which a large number of mismatches is allowed, or sites with a repetitive structure (e.g., 5′-GGGGG-3′) will not follow a Poisson distribution even in random DNA with independently distributed nucleotides. However, this is not a problem for the sites studied here (see 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 $X_1, \ldots, X_n$ the positions at which a site *S* or its reverse $\overline{S}$ 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.

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

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 equation **3**, the distribution of the distance between successive words, $D_{i,i+1}$, is exponential with density

$$\lambda e^{-\lambda z}. \qquad \mathbf{5}$$

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

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

where $\Gamma(k) = (k-1)!$ is the gamma function. This is easily seen from the characteristic functions of equations **5** and **6** (18). The probability of observing a *k*-cluster of length less than *x* is

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

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

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

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 $N_S + N_{\overline{S}}$. This was done because in general only one of two overlapping sites can be functional, i.e., occupied by a TF at any given time. In terms of the statistical analysis, it leads to more conservative significance tests, because very short and thus highly significant 2-clusters are eliminated. 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 equation **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-10}, hence the name 'genome walk' analysis.

For all binding sites analyzed here, except those for the transcription factor MBF, a significance level of $\mathbf{P} = 0.001$ was chosen, because of the large number of site counts, and thus the large number of significance tests to be carried out. For example, for a TF with a genomic site count of $N_S + N_{\overline{S}} = 5000$, there are ~500 non-overlapping 10-clusters, and thus 500 independent significance tests for 10-clusters. A value of $\mathbf{P} = 0.05$ or $\mathbf{P} = 0.01$ would lead to a high type I error probability. The particular choice of **P** is motivated by the counts observed for the binding sites studied here (10^3–10^4 per genome), such that **P** is of the order of the number of independent tests carried out for a given cluster size *k*.

Table 1. Binding site counts and tests for Poisson distribution in *S.cerevisiae* genome and in random DNA

Site		Yeast genome				Random DNA	
		Mismatches allowed	No of sites	χ^2 (df)	G (df)	χ^2 (df)	G (df)
SBF	5′-CACGAAAA-3′	1	15331	26.40 (10)[a]	25.37 (10)[a]	3.78 (10)	4.32 (10)
MBF	5′-ACGCGT-3′	0	692	2.61 (6)	2.57 (6)	7.94 (7)	8.48 (7)
bHLH	5′-CACGTG-3′	0	953	7.28 (6)	7.29 (6)	3.01 (7)	3.21 (7)
NBF	5′-ATGTGAAAT-3′	1	5509	16.34 (9)	16.52 (9)	12.43 (9)	13.17 (9)

[a]Significant at $0.005 < \mathbf{P} < 0.001$

Step 2. Elimination of some statistically significant clusters

Yeast transcriptional regulators function in general only when bound upstream of the coding region (9,10), with the possible exceptions of the transcription of Ty retrotransposons (19). 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 composition

Estimating λ via actual site counts in step 1 is necessary because global sequence composition is a poor predictor of site occurrence (20). 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. Precisely those mono- and dinucleotides that occur in the binding sites will be overly frequent in small clusters. This is why a DNA segment larger than the actual cluster was used for small clusters. Two new estimates of λ, based on mono- and dinucleotide distributions in these regions were used to reassess the significance (equation **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 (21). Higher order correlations among nucleotides were not taken into account for reasons of computational feasibility.

R-scan analysis

This statistical technique (20,22,23) can be used to assess on a global level whether words show a clumped distribution in genomic DNA. It uses only the extreme values of the distribution of $D_{i,i+k}$ (a k-scan in Karlin's terminology). Denote as m_k^l the lth smallest k+1-cluster, $D_{i,i+k}$. R-scan analysis asks whether m_k^l is smaller than expected by chance alone under the Poisson null-hypothesis. The relevant formalism can be found in equation 5 of ref. 20.

Goodness of fit tests for exponential distribution

Likelihood ratio and χ^2 goodness of fit tests were carried out as described in ref. 24 (Ch. 17) to establish whether the lengths of $D_{i,i+1}$ followed an exponential distribution. Estimates of λ were based on global site counts. Williams' correction was applied to the likelihood ratio test (24, p704).

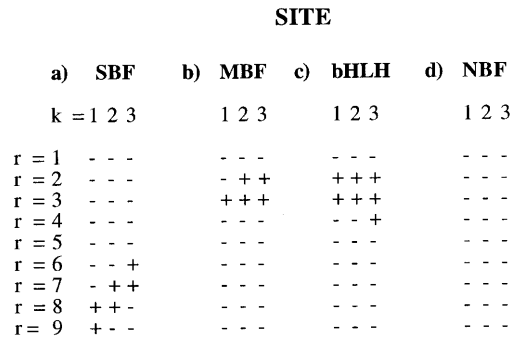

Figure 1. Tests for a clumped distribution of TF binding sites. Shown are the results of an r-scan analysis for clumped site distribution. An r-scan is defined as the length of DNA between r+1 consecutive binding sites. The test asks whether the kth smallest r-scan observed in the genome is significantly smaller than expected by chance alone under the null-hypothesis of exponentially distributed binding site distances. A '+' in the table indicates that the respective value is significantly ($\mathbf{P} = 0.01$) smaller than expected. The rather conservative significance level is chosen because of the large number of tests carried out. The figure shows that the binding sites for MBF, SBF, and the bHLH core motif 5′-CACGTG-3′ show a clumped distribution, whereas those of NBF do not.

RESULTS AND DISCUSSION

Several applications of the method introduced above are illustrated with different yeast transcriptional regulators. The first example concerns two transcriptional regulators, SBF and MBF (DSC1), known to regulate the expression of a large number of genes that are expressed in the late G1 phase of the cell cycle (25). Both factors are heterodimers that share a common subunit. However, their consensus DNA binding sequences differ (see Table 1), and they appear to regulate non-overlapping sets of genes (26–28). SBF regulates the transcription of the HO endonuclease, the cyclins CLN1 and CLN2, and the putative cyclin HCS26 (29). MBF regulates a large number of DNA synthesis genes, the cyclins CLB5 and CLB6, the kinase SPK1, and the transcription factor SWI4 (25,28,30).

Global analysis of genomic site distribution

Sites that would not follow a Poisson distribution in random DNA cannot be analyzed with this method, as discussed above. It was thus tested whether distances between SBF (MBF) binding sites follow an exponential distribution in a long (14 Mb) random DNA sequence with the same nucleotide composition as yeast. The

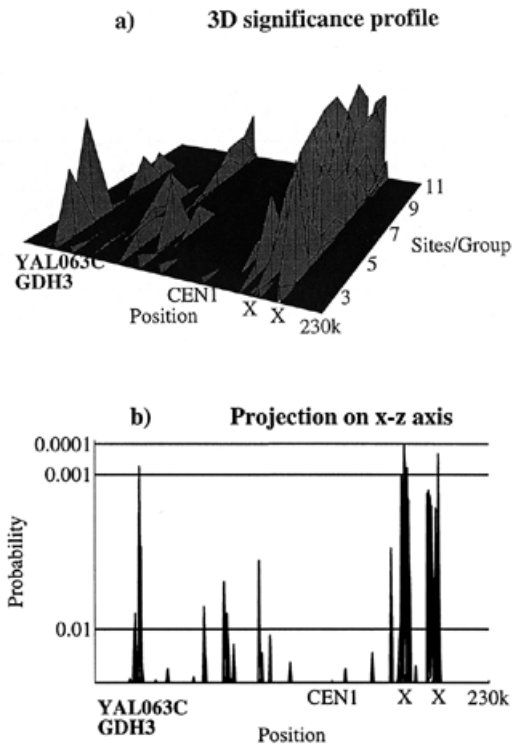

Figure 2. Significance profile of SBF binding site clusters on chromosome I. (**a**) The statistical significance of all groups of SBF binding sites on chromosome I. Each point in the $x-y$ plane corresponds to a group of binding sites comprising the number of sites indicated on the y-axis (2–11), whose 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 I. 'CEN1' indicates the approximate position of the centromere. Because there is a total of 292 binding sites on chromosome I, not all positions can be shown individually. 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. (**b**) 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 230209). The ordinate shows the **P**-values of clusters. Notice that there are three clusters with $\mathbf{P} < 0.001$, which are discussed in greater detail in the text.

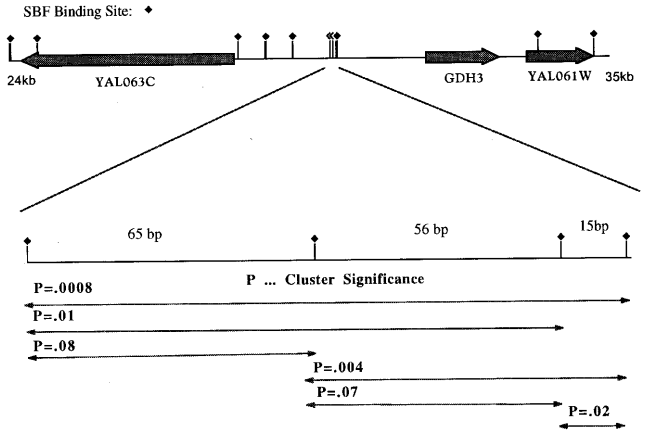

Figure 3. A significant cluster of SBF-binding sites on chromosome I between GDH3 and YAL063C. The displayed region corresponds to an 11 kb fragment starting at 24 kb counted from the left telomere of chromosome I. It includes the highly significant 4-cluster of SBF binding sites labelled in Figure 2 as YAL063C/GDH3. A detail of the cluster is shown in the lower part of the figure. It comprises four SBF binding sites spanning a total length of 144 bp to the last position of the fourth site. Also shown is the **P**-value of this 4-cluster, as well as the **P**-values of all sub-groups of binding sites, as indicated by the arrows. Notice that, despite their tight linkage, none of these sub-groups is significant at $\mathbf{P} < 0.001$. Because the neighboring ORFs are encoded on opposite strands, and transcribed in opposite directions, SBF is a candidate for transcriptional regulation of both genes.

distribution parameter λ was estimated via equations **1** and **2**. Results are consistent with a Poisson distribution in random DNA (Table 1). One mismatch to the SBF binding site was allowed, because the genes known to be regulated by SBF, such as HO, have several such near-matches to the SBF consensus in their promoter region (29). For *S.cerevisiae* genomic DNA, it would seem likely that site distribution would deviate from a Poisson, due to compositional heterogeneity. Perhaps surprisingly, only the SBF consensus site shows a deviation from the Poisson distribution (Table 1). However, a goodness of fit test to an exponential distribution provides only a very crude assessment of distribution properties. This is because (i) a large amount of distance information (see the site counts in Table 1) is pooled into a small number of bins, and (ii) no site distances other than those among nearest neighbors are included in the test. With these tests, a clumped distribution of binding sites, which may indicate the existence of biologically relevant clusters, could only be detected if a very large number of very closely spaced site-pairs occurred. A more sensitive test is provided by *r*-scan analysis (20,22,23). An *r*-scan is the cumulative length of DNA between $(r+1)$ consecutive binding sites. *R*-scan analysis for clumped distribution asks whether the *k*th smallest *r*-scan observed in the genome is smaller than expected by chance alone. Figure 1a and b shows the results of *r*-scan analysis for MBF and SBF, respectively. Both MBF and SBF show a clustered distribution, albeit for different *r*-values. A more fine grained analysis is encouraged by these findings.

Genome walk

As an example of the results obtained with the method, Figure 2 shows a significance profile of all binding site clusters of chromosome I of *S.cerevisiae* (see Fig. 2 legend). Peaks of the plot correspond to highly significant clusters, clusters that are very unlikely to have occurred by chance alone. Three clusters significant at $\mathbf{P} < 0.001$ are evident. They are located at positions 29760, 188430 and 198837 (not shown in Figure), counted from the left telomere of chromosome I, and are labelled YAL063C/GDH3, X and X, respectively. Despite their high significance, two of these clusters (labelled X) have to be eliminated from further consideration. The cluster starting at position 188430 shows a large overlap with the open reading frame YAR033W. Although it would be possible to include such clusters under the assumption that some putative ORFs will turn out to be non-coding regions (31), the conservative approach of considering only clusters located in bona-fide non-coding regions is taken here. The second cluster, starting at position 198837, is eliminated because it occurs downstream of both neighboring ORFs, OSH1 and YAR047C (not shown). The remaining cluster at position 29760 occurs in the 5′ non-coding region of the ORF YAL063C, encoded on the bottom strand, and the gene GDH3, encoded on the top strand. A detail of this region is shown in

Figure 3. The figure shows a tightly packed group of four SBF binding sites spanning 144 bp, significant at $\mathbf{P} = 7.55 \times 10^{-4}$. The figure also shows the significance values for all sub-groups of binding sites, none of which is below the threshold of $\mathbf{P} = 0.001$. This 4-cluster of binding sites makes both GDH3 and YAL063C candidate genes for regulation by SBF.

Table 2 summarizes the results of an analogous analysis for all 16 chromosomes. It shows all clusters of SBF binding sites significant at $\mathbf{P} < 0.001$ that are also candidates for the regulation of some ORF. There is experimental evidence that two of the genes are regulated by SBF (29). Clusters in the 5′ regions of two divergently transcribed genes might be involved in the regulation of one or both of the genes (e.g., the first pair in the table, GDH3/YAL063C, also shown in Figure 3)

Table 2. Candidate genes for regulation by SBF

Chr.	ORF	Cluster statistics		Estimated significance			Gene structure/function[c]
		Sites/bp[a]	Position[b]	Global	Mono	Di	
1	YAL063C		–1791				unknown
		4/144		7.55×10^{-4}	5.37×10^{-4}	2.9×10^{-3}	
1	GDH3		–1677				possibly NADP-linked glutamate dehydrogenase
2	YBR162c		–3[e]				unknown
		7/551		8.31×10^{-5}	2.31×10^{-5}	2.43×10^{-3}	
2	YSY6		–115				component of secretory pathway
4	HO[d]	10/619	–242	1.40×10^{-7}	2.32×10^{-8}	5.10×10^{-7}	mating type switch
4	UME6		–269				TF involved in meiosis and nitrogen repression
		3/28		3.17×10^{-4}	1.36×10^{-4}	5.78×10^{-4}	
4	MSS4		–413				required for cell growth
7	PDE1	4/113	–110	3.58×10^{-4}	1.68×10^{-4}	2.39×10^{-4}	3′5′-cyclic-nucleotide-phosphodiesterase
7	SNG1		–344				involved in nitrosoguanidine resistance
		4/114		3.68×10^{-4}	1.45×10^{-4}	6.82×10^{-4}	
7	YGR198W		–103				unknown
7	YGR033C		–187				unknown
		6/252		1.84×10^{-5}	5.68×10^{-6}	5.92×10^{-5}	
7	YGR034W		–531				unknown
9	YIL169C		–1043				unknown
		5/253		3.05×10^{-4}	1.93×10^{-4}	5.32×10^{-4}	
9	SDL1		–1638				L-serine dehydratase
12	YLR179C		–240				unknown
		5/308		6.49×10^{-4}	2.48×10^{-4}	6.46×10^{-3}	
12	SAM1		–9[f]				*S*-adenosylmethionine synthetase
12	YLR308W	6/323	–544	6.13×10^{-5}	4.45×10^{-5}	5.10×10^{-5}	unknown
12	MID2	5/202	–651	1.26×10^{-4}	4.32×10^{-5}	3.72×10^{-4}	required for mating
14	YNR051C	8/718	–114	4.40×10^{-5}	1.34×10^{-5}	3.24×10^{-4}	unknown
15	YOL157C		–380				unknown
		4/144		7.55×10^{-4}	3.23×10^{-4}	1.30×10^{-3}	
15	HXT11		–463				high-affinity hexose transporter
15	YOL104C		–6314				unknown
		6/227		1.10×10^{-5}	8.10×10^{-6}	9.60×10^{-5}	
15	ITR2		–5[g]				myo-inositol transporter
15	YOL007C	4/138	–259	6.63×10^{-4}	2.94×10^{-4}	9.90×10^{-4}	unknown
16	CLN2[d]		–531				G_1 cyclin
		6/536		6.50×10^{-4}	2.59×10^{-4}	1.34×10^{-2}	
16	BBP1		–52				deletion mutants defective in cell division

[a]Number of binding sites in the cluster/length of cluster in base pairs. Only one value is given for two genes if the genes share a promoter region, i.e., if they are divergently transcribed.
[b]Distance of the 3′-most site in the cluster from the start codon.
[c]From the *S.cerevisiae* genome database (http://genome-www.stanford.edu/Saccharomyces); see also references in the text.
[d]Gene is known to be regulated by SBF.
[e]A significant sub-cluster with a higher **P**-value exists which ends at position –56.
[f]No significant sub-cluster exists whose 3′-most site lies upstream of –9.
[g]A significant sub-cluster with a higher **P**-value exists which ends at position –21.

The clusters listed in the table were identified on the basis of their **P**-values (given in column 5), which are calculated from genomic binding site counts. Local **P**-values (columns 6 and 7 of Table 2) based on local mono- and dinucleotide composition in the respective promoter region are included here to account for compositional heterogeneity in genomic DNA (32). Any cluster with a local **P**-value vastly higher than the global **P**-value indicates that local base-composition may have favored occurrence of the cluster. 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 only be considered further if other evidence argues for its biological relevance.

Twenty six ORFs emerge as candidates for further investigation, based on global **P**-values <0.001. Fourteen of these are genes with known function, two of which, HO and CLN2, are known to be regulated by SBF (29). Indeed, the regulatory region of HO contains the cluster of SBF binding sites with the highest significance of all, $\mathbf{P} = 2.32 \times 10^{-8}$. Two other genes known to be regulated by SBF, CLN1 and HCS26, were not detected by this analysis, because the significance of the respective binding site clusters is well above **P** = 0.001 (not shown). This illustrates the price paid for trying to minimize the false positive rate, i.e., a high false negative rate of not detecting genes regulated by a TF. Given only four genes known to be regulated by SBF, a statistically reliable estimate of this rate is clearly impossible, but it may well be of the order of 50% or higher. Of the 24 ORFs that are not known to be regulated by SBF, some are suspicious based on features of the site clusters. The cluster associated with the gene pair YBR162C/YSY6 has a suspiciously high local **P**-value of 2.43×10^{-3}, and its 3′ most site lies only 3 bp upstream of the start codon of YSY6. Such a site would lie downstream of the TATA-box (9), and would thus probably be irrelevant to transcriptional regulation. YLR179C/SAM1 and YOL104C/ITR2 might be excluded on similar grounds. Nine of the remaining 18 strong candidate ORFs are functional genes, and biological criteria can be applied to identify good candidate genes for further investigation among them. For example, four of these nine ORFs, UME6, MSS4, MID2, and BBP1, are thought to have a function in the cell-cycle, although not necessarily in the G_1/S-transition (Table 2). No such criteria can be applied to ORFs of unknown function, and one can only consider **P**-values as rough guides to identify promising candidates for further investigation (e.g., YGR033C/YGR034W with a 6-cluster of $\mathbf{P} < 5.92 \times 10^{-5}$).

Evidence supporting biological relevance of significant clusters

In addition to (i) the detection of genes known to be regulated by SBF, and (ii) the detection of genes with a likely role in the cell cycle, two pieces of evidence suggest that this type of statistical analysis yields biologically meaningful results. First, consider all clusters of binding sites significant at $\mathbf{P} < 0.001$, 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 ~72% of the individual sites to occur in coding regions, because coding regions account for ~72% of the yeast genome (33). However, SBF binding sites belonging to significant clusters occur with vastly higher frequency in non-coding regions (Table 3, $\chi^2 = 109.53$, $\mathbf{P} << 10^{-3}$). Could this simply be due to differences in the base composition of coding and non-coding regions that favor site occurrence in non-coding regions? The predicted probabilities of site occurrence (Table 4) based on the base composition in non-coding and coding regions do not support this possibility. Predicted site probabilities either differ by <2% for non-coding and coding regions, or even suggest that SBF binding sites should occur more frequently in coding regions, in stark contrast to the observation. It is tempting to speculate that this biased distribution has to do with transcriptional regulation. For example, it might be the result of (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. If this is true, the distribution of site clusters among coding/non-coding regions might aid in assessing whether the binding site of a DNA-binding protein with unknown function has a role in transcriptional regulation.

Table 3. Binding sites belonging to significant clusters occur preferably in non-coding regions

Site	Total	Coding		Non-Coding		χ^2 (1 df)
		Obs.	Exp.	Obs.	Exp.	
SBF	461	231	331.9	230	129.1	109.53
MBF	114	37	82.1	77	31.9	88.54
bHLH	76	31	54.7	45	21.3	36.64
NBF	195	114	140.4	81	54.6	17.73

Expected values in non-coding and coding regions are based on the fact that 72% of the *S.cerevisae* genome encodes for proteins. All χ^2 values significant at $\mathbf{P} < 0.001$.

Table 4. Estimated probabilities[a] of binding site occurrence in coding and non-coding regions of *S.cerevisiae*

Site	Non-coding		Coding	
	Mono	Di	Mono	Di
SBF	9.88×10^{-4}	1.11×10^{-3}	9.73×10^{-4}	1.22×10^{-3}
MBF	1.13×10^{-4}	7.46×10^{-5}	1.31×10^{-4}	5.78×10^{-5}
bHLH	1.13×10^{-4}	1.01×10^{-4}	1.31×10^{-4}	9.87×10^{-5}
NBF	5.39×10^{-4}	6.34×10^{-4}	4.80×10^{-4}	6.19×10^{-4}

[a]Estimates are based on 1000 randomly chosen 1 kb DNA segments from coding or non-coding regions, i.e., on 1 Mb of genomic DNA.

The second piece of evidence concerns the distribution of observed mismatches to the consensus. If one considers SBF binding sites in the regulatory regions of the four genes known to be regulated by SBF, it appears that some positions are more variable than others (29). A statistically sound argument is difficult to make, partly because the number of binding sites is small (29). If the sites observed in the clusters shown in Table 2 were irrelevant to SBF-binding and transcriptional regulation, one would expect the mismatches to the consensus to be evenly distributed across the sites. This is not what is observed. The listed clusters consist of 70 individual sites, 69 of which show one mismatch to the consensus, 5′-CACGAAAA-3′. The number of sites with mismatches at each position is

C	*A*	*C*	*G*	*A*	*A*	*A*	*A*
18	10	18	8	3	4	3	5

Table 5. Candidate genes for regulation by MBF

Chr.	ORF	Cluster statistics		Estimated significance			Gene structure/function[c]
		Sites/bp[a]	Position[b]	Global	Mono	Di	
1	RFA1[d]		–130				replication factor A, 69 kDa subunit
		2/36		1.72×10^{-3}	3.83×10^{-3}	5.54×10^{-3}	
1	YAR008W		–186				unknown
2	POL12[d]	2/29	–194	1.32×10^{-3}	2.30×10^{-3}	1.94×10^{-3}	DNA polymerase I, β subunit
3	YCL060C	2/35	–838	1.66×10^{-3}	3.05×10^{-3}	2.01×10^{-3}	unknown
3	YCR064C		–487				unknown
		2/54		2.75×10^{-3}	5.10×10^{-3}	1.01×10^{-2}	
3	HCM1		–269				isolated as suppressor of a calmodulin (CMD1) mutant
4	YDL018C		–122				unknown
		2/44		2.18×10^{-3}	4.59×10^{-3}	3.00×10^{-3}	
4	CDC7		–539				protein kinase required for initiation of mit. DNA synthesis
4	MCD1	2/86	–292	4.58×10^{-3}	5.99×10^{-3}	8.13×10^{-3}	mitotic chromosome determinant; similar to *Schizosaccharomyces pombe* RAD21
4	YDR097C	2/26	–171	1.15×10^{-3}	2.2×10^{-3}	3.42×10^{-4}	unknown
4	YDR134C	2/54	–344	2.75×10^{-3}	9.35×10^{-3}	1.15×10^{-3}	unknown
5	RNR1[d]	4/192	–306	2.01×10^{-7}	1.37×10^{-6}	1.64×10^{-6}	ribonucleotide reductase regulatory subunit 1
5	PUP3		–433				putative proteasome subunit
		2/47		2.35×10^{-3}	6.91×10^{-3}	5.99×10^{-3}	
5	RAD51[d]		–160				recombinational DNA repair
7	CLB6[d]		–372				cyclin
		2/32		1.49×10^{-3}	3.26×10^{-3}	4.80×10^{-3}	
7	YGR110W		–6810				unknown
9	YIL026C	2/13	–123	4.01×10^{-4}	9.98×10^{-4}	1.65×10^{-3}	unknown
10	NCA3		–1197				mutation affects mitochondrial ATP synthase
		2/42		2.06×10^{-3}	4.13×10^{-3}	3.80×10^{-3}	
10	ASF1		–180				causes expression of silent loci when overexpressed
10	YJR030C	2/16	–216	5.73×10^{-3}	9.16×10^{-4}	1.40×10^{-3}	unknown
11	RAD27		–123				exonuclease required for processing of Okazaki fragments
		2/58		2.98×10^{-3}	6.04×10^{-3}	8.97×10^{-3}	
11	ABFI		–520				transcription factor and ARS binding protein
12	CDC45		–145				DNA replication initiation protein
		2/36		1.72×10^{-3}	2.73×10^{-3}	5.28×10^{-3}	
12	YLR104W		–469				unknown
12	FKS1	2/60	–583	3.09×10^{-3}	7.20×10^{-3}	1.15×10^{-2}	1,3-β-D-glucan synthase
14	YNL313C		–145				unknown
		2/18		6.88×10^{-4}	1.14×10^{-3}	9.27×10^{-4}	
14	RFA2[d]		–108				replication factor A, 32 kDa subunit
14	YNL274C		–558				unknown
		2/25		1.09×10^{-3}	3.26×10^{-3}	2.99×10^{-3}	
14	YNL273W		–136				unknown
14	POL1[d]	2/41	–173	2.00×10^{-3}	2.26×10^{-3}	3.13×10^{-3}	DNA polymerase I
15	YOL018C		–271				unknown
		2/19		7.45×10^{-4}	1.08×10^{-3}	9.34×10^{-4}	
15	YOL017W		–171				unknown
15	CDC21[d]		–116				thymidilate synthase
		2/43		2.12×10^{-3}	4.05×10^{-3}	2.12×10^{-3}	
15	UFE1		–470				null-mutant defective in spore germination and veg. growth

Table continued

Table 5. *continued*

Chr.	ORF	Cluster statistics		Estimated significance			Gene structure/function[c]
		Sites/bp[a]	Position[b]	Global	Mono	Di	
16	SPK1[d]		–254				S-phase specific kinase
		2/31		1.43×10^{-3}	2.47×10^{-3}	2.35×10^{-3}	
16	YPL152W		–557				unknown
16	DSS4		–327				GDP dissociation factor for Sec4p
		2/16		5.73×10^{-4}	1.32×10^{-4}	1.21×10^{-3}	
16	RLF2		–222				involved in DNA-replication-linked nucleosome assembly
16	YPR075C	2/29	–159	1.32×10^{-3}	3.37×10^{-3}	2.20×10^{-3}	unknown

[a]Number of binding sites in the cluster/length of cluster in base pairs. Only one value is given for two genes if the genes share a promoter region, i.e., if they are divergently transcribed.
[b]Distance of the 3′-most site in the cluster from the start codon.
[c]From the *S.cerevisiae* genome database (http://genome-www.stanford.edu/Saccharomyces); see also references in the text.
[d]Gene is known to be regulated by MBF.

This highly significant deviation from the expected uniform distribution [$\chi^2 = 31.99(7df)$, $\mathbf{P} < 0.001$] further suggests that the clusters in Table 2 are not only statistically significant, but also biologically relevant. Moreover, for most of the positions, the pattern of mismatches is similar to that for the sites in the four genes known to be regulated by SBF (29).

The analysis of binding site distribution for MBF proceeds analogously. MBF binding sites show a clumped distribution in the genome (Fig. 1). Sites belonging to significant clusters occur preferentially in non-coding regions (Table 3), an observation that cannot be explained by differences in base composition (Table 4). Table 5 shows 39 genes identified through the genome walk analysis. Because of the small number of MBF binding sites in the genome (Table 1), a somewhat higher significance level of $\mathbf{P} = 0.005$ was used here. Because of this small number of sites, a group of only two closely spaced sites can be significant. In fact, all candidate genes except RNR1 (which is known to be regulated by MBF; 30) have only two binding sites in their non-coding region. Significance estimates based on mono- and dinucleotide distributions are to be taken with caution here, because global genome composition considerably overestimates the probability of site occurrence (not shown). If this holds for local composition as well, then the values shown in column 6 and 7 of Table 5 will considerably underestimate cluster significance. For nine of the 39 identified candidate genes, regulation by MBF has already been shown or proposed (25,30). Of the remaining 30 candidates, 17 are ORFs of unknown function. Among the 13 genes with known function are some good candidates for regulation by MBF, based on their role in the cell-cycle, and based on the fact that MBF is known to regulate the expression of genes involved in DNA replication. One example is RLF2, involved in the assembly of nucleosomes on replicating DNA (34). Another example is RAD27 (RTH1), a 5′–3′ exonuclease required for the processing of Okazaki fragments during replication (35). Notably, 28 out of the 39 candidate genes are members of gene pairs that are transcribed divergently on opposite strands.

Families of DNA binding activities

Families of transcription factors with widely overlapping binding specificities are common in eukaryotes, and a one-to-one relation between distinct transcription factors and different binding sites does not always exist (36,37). Where this is the case, one may only be able to analyze binding sites common to a group of factors (13), but the genome walk approach may still be useful in identifying genes regulated by one or more factors in such a group.

DNA binding proteins belonging to the basic helix–loop–helix (bHLH) family of transcription factors (38) bind the core motif 5′-CANNTG-3′. In budding yeast, at least six genes encoding members of this family exist. PHO4, a transcriptional regulator of genes needed for phosphate utilization (39), CBFI, necessary for centromere binding and methionine prototrophy (40), INO2 and INO4, which form a transcriptional regulator of the expression of phospholipid biosynthetic genes (41,42), SGC1, required for the expression of the yeast enolase genes (43), and RTG1, a protein involved in the communication between nucleus, mitochondria and peroxisomes (44). A sub-group of bHLH proteins binds the palindromic motif 5′-CACGTG-3′, and CBFI and PHO4 are members of this sub-group in budding yeast (45). INO2/INO4 seem to bind DNA with a slightly different specificity (46,47), and the binding activities of both SGC1 and RTG1 are not well characterized. Because the bHLH core binding motif is too short to be analyzed with the method used here, a search for groups of the 5′-CACGTG-3′ motif was carried out. Genes whose promoters contain such groups are candidates for regulation by all characterized bHLH proteins except Ino2p/Ino4p, plus potentially unknown bHLH factors. As in the above cases, the bHLH motif shows a clumped distribution (Fig. 1c), and a strong bias for cluster occurrence in non-coding regions (Tables 3 and 4). Table 6 shows genes associated with highly significant clusters. There are between 8 and 15 candidate genes, depending on whether one or both members of the gene pairs in promoter–promoter orientation are counted. The most significant cluster ($\mathbf{P} = 2.9 \times 10^{-6}$) is associated with two ORFs of unknown function. Notably, three of the nine candidate genes with known function, ATP7, NDI1 and IDH1, encode mitochondrial proteins involved in energy metabolism. It is tempting to speculate that RTG1 may be involved in their regulation, given that it may have a role in regulating mitochondrial metabolism (44).

Table 6. Candidate genes for regulation by 5′-CACGTG-3′ bHLH transcription factors

Chr.	ORF	Cluster statistics		Estimated significance			Gene structure/function[c]
		Sites/bp[a]	Position[b]	Global	Mono	Di	
4	CDC34		–348				ubiquitin conjugating enzyme E2
		3/313		2.89×10^{-4}	8.48×10^{-4}	3.32×10^{-4}	
4	YDR055W		–545				unknown
5	SWI4		–1121				transcription factor
		2/13		5.53×10^{-4}	5.57×10^{-4}	5.86×10^{-4}	
5	USS1		–228				U6 snRNA-associated protein
6	LPD1		–263				dihydrolipoamide dehydrogenase precursor
		2/17		8.68×10^{-4}	1.10×10^{-3}	1.71×10^{-3}	
6	SNP2		–298				snRNP E protein
8	YHR136C		–26				unknown
		4/336		2.89×10^{-6}	1.79×10^{-5}	4.86×10^{-5}	
8	YHR137W		–253				unknown
10	YJL012C	2/18	–108	9.47×10^{-4}	1.46×10^{-3}	1.29×10^{-3}	unknown
11	ATP7		–262				ATP synthase subunit d
		3/35		2.62×10^{-6}	4.33×10^{-6}	8.83×10^{-6}	
11	PUT3		–264				transcriptional activator of proline utilization genes
13	NDI1		–553				mitochondrial NADH ubiquinone 6 oxidoreductase
		2/17		8.68×10^{-4}	1.15×10^{-3}	1.00×10^{-3}	
13	YML119W		–240				unknown
14	IDH1		–392				SU of mitochondrial isocitrate dehydrogenase 1
		3/89		2.14×10^{-5}	8.31×10^{-5}	1.78×10^{-4}	
14	NCE3		–337				involved in protein secretion

[a]Number of binding sites in the cluster/length of cluster in base pairs.
[b]Distance of the 3′-most site in the cluster from the start codon.
[c]From the *S.cerevisiae* genome database (http://genome-www.stanford.edu/Saccharomyces); see also references in the text.

Homotypic cooperativity of DNA binding activities with unknown function

The example used here is NBF, an activity binding to four sites in the promoter of the INO1 gene which is involved in the biosynthesis of membrane phospholipids (46). The products of at least three genes, OPI1, INO2 and INO4, contribute to the transcriptional regulation of INO1 (41,48). NBF appears to be distinct from their products (42,46). NBF binds specifically to a sequence with consensus 5′-ATGTGAAAT-3′, which is very similar to an octamer motif, 5′-ATGCAAAT-3′, known to be involved in the transcriptional regulation of immunoglobulin genes (36). Although promoter fragments that confer INO1 specific regulation contain at least one NBF binding site, an NBF site alone in front of a heterologous reporter gene cannot activate transcription (41). Thus, NBF may not be part of a transcriptional regulator, or base pairs flanking its binding site may be necessary for full UAS function (46,49).

5509 sites with at most one mismatch to the NBF consensus occur in the genome of *S.cerevisiae* (Table 1). Sites with a larger number of mismatches are too frequent, and would not allow a meaningful analysis. When analyzing the distribution of these 5509 sites, a pattern emerges that is fundamentally different from that of the three binding sites just discussed. First, NBF binding sites do not show a clumped distribution (Fig. 1d). Second, while still significant, the distribution among coding and non-coding regions of sites inside clusters (Table 3) shows a much less biased pattern than that of the other sites. Whereas the ratio s of binding sites in non-coding regions to those in coding regions is at least one for these sites (calculated from Table 3), $s = 0.71$ for NBF. The mononucleotide distribution of non-coding regions might account for a part of the remaining bias (Table 3). Third, despite the large number of NBF binding sites, only three significant clusters (not shown) occur that lie entirely in the 5′ non-coding regions of some ORF. Contrast this with the 15 candidate genes for regulation by bHLH proteins, despite the fact that their total number of binding sites is almost 6-fold lower. Thus, NBF binding sites show a pattern of site and cluster distribution vastly different from that of the transcriptional regulators analyzed thus far. If NBF is a transcriptional regulator at all, homotypic cooperative interactions are not a dominant mode of action for NBF.

CONCLUSION AND OUTLOOK

The transcription factors studied here illustrate that despite a large number of transcription factor binding sites in a genome, the

number of significant clusters of sites can 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 presented here detects genes whose regulation by a given transcription factor was shown experimentally, indicate its usefulness. However, the critical question regarding the false positive rate of the method can only be decided by experimentally testing its predictions. This may be a challenging task, especially because presence or absence of a transcription factor alone may not be sufficient for regulation of a target gene. The availability of necessary cofactors may critically depend on the environment, or on the physiological state of a cell.

Many further applications of the method are conceivable, other than analyzing all characterized transcription factor binding sites in yeast. For example, the usefulness of the method can be considerably enhanced by not only considering homotypic cooperativity, but also heterotypic interactions at a promoter. That is, consider not only clusters of binding sites for one transcription factor, but also clusters of binding sites for different transcription factors. This extension of the method 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. Such an application will raise new challenges because of (i) the vastly larger genomes involved, (ii) the abundance of tandem repeats, (iii) the existence of regulatory regions interspersed between genes, and (iv) the often ill-defined location of coding regions. In these cases, existing complementary techniques (50–52), e.g., techniques suitable to determine the location of likely promoter regions, will have to be used in conjunction with the method introduced here.

ACKNOWLEDGEMENTS

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. and Gann, A. (1997) *Nature*, **386**, 569–577.
2 Levine, M. and 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. and Lindellen, J. (1997) *Nature*, **385**, 29–30.
4 Sorger, P.K. (1991) *Cell*, **65**, 363–366.
5 Srinivas, U.K. and Swamynathan, S.K. (1996) *J. Biosci.*, **21**, 103–121.
6 Ptashne, M. (1988) *Nature*, **335**, 683–689.
7 Karlin, S. and Brendel, V. (1993) *Science*, **259**, 677–680.
8 Bernardi, G., Mouchiroud, D., Gautier, C. and Bernardi, G. (1988) *J. Mol. Evolut.*, **28**, 7–18.
9 Struhl, K. (1989) *Annu. Rev. Biochem.*, **58**, 1051–1077.
10 Struhl, K. (1995) *Annu. Rev. Genet.*, **29**, 651–674.
11 Olson, M.V. (1992) In Jones, E.W., Pringle, J.R. and Broach, J.R. (eds) *The Molecular and Cellular Biology of the Yeast Saccharomyces. Vol. I*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
12 Stormo, G.D. (1990) *Methods Enzymol.*, **183**, 211–220.
13 Fickett, J.W. (1996) *Mol. Cell. Biol.*, **16**, 437–441.
14 IUB Nomenclature Committee (1985) *Eur. J. Biochem.*, **150**, 1–5.
15 Kendall, M.G. (1952) *The Advanced Theory of Statistics. Vol. II*. Griffin, London. p. 22.
16 Dhawale, S.S. and Lane, A.C. (1993) *Nucleic Acids Res.*, **24**, 5537–5546.
17 Karlin, S. and Taylor, H.. (1975) *A First Course in Stochastic Processes*. Academic Press, New York.
18 Abramowitz, M. and Stegun, I.A. (1972) *Handbook of Mathematical Functions*. 26.1.28, 26.1.31. Dover, New York.
19 Türkel, S. and Farabaugh, P.J. (1993) *Mol. Cell. Biol.*, **13**, 2091–2103.
20 Karlin, S. and Macken, C. (1991) *Nucleic Acids Res.*, **19**, 4241–4246.
21 Parzen, G. (1962) *Stochastic Processes*. Holden-Day, San Francisco. Ch. 4.2.
22 Karlin, S. and Macken, C. (1991) *J. Am. Stat. Assoc.*, **86**, 27–35.
23 Dembo, A. and Karlin, S. (1992) *Ann. Appl. Prob.*, **2**, 329–357.
24 Sokal, R.R. and Rohlf, F.J. (1981) *Biometry*. Freeman, New York.
25 Koch, C. and Nasmyth, K. (1994) *Curr. Opin. Cell Biol.*, **6**, 451–459.
26 Nasmith, K. and Dirick, L. (1991) *Cell*, **66**, 995–1013.
27 Lowndes, N.F., Johnson, A.L. and Johnston, L.H. (1991) *Nature*, **350**, 247–250.
28 Verma, R., Patapoutian, A., Gordon, C.B. and Campbell, J.L. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 7155–7159.
29 Ogas, J., Andrews, B.J. and Herskowitz, I. (1991) *Cell*, **66**,1015–1026.
30 McIntosh, E.M. (1993) *Curr. Genet.*, **24**, 185–192.
31 Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M. *et al.* (1996) *Science*, **274**, 546–567.
32 Karlin, S. and Cardon, L.R. (1994) *Annu. Rev. Microbiol.*, **48**, 619–654.
33 Dujon, B. (1996) *Trends Genet.*, **12**, 263–270.
34 Kaufman, P.D., Kobayashi, R. and Stillman, B. (1997) *Genes Dev.*, **11**, 345–357.
35 Sommers, C.H., Miller, E.J., Dujon, B., Prakash, S. and Prakash, L. (1995) *J. Biol. Chem.*, **270**, 4193–4196.
36 Johnson, P.F. and McKnight, S.L. (1989) *Annu. Rev. Biochem.*, **58**, 799–839.
37 Lamb, P. and McKnight, S.L. (1991) *Trends Biochem. Sci.*, **16**, 417–422.
38 Murre, C., McCaw, P.S. and Baltimore, D. (1989) *Cell*, **56**, 777–783.
39 Johnston, M. and Carlson, M. (1992) In Jones, E.W., Pringle, J.R. and Broach, J.R. (eds) *The Molecular and Cellular Biology of the Yeast Saccharomyces. Vol. II*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
40 Dowell, S.J., Tsang, J.S.H. and Mellor, J. (1992) *Nucleic Acids Res.*, **20**, 4229–4236.
41 Lopes, J.M., Hirsch, J.P., Chorgo, P.A., Schulze, K.L. and Henry, S.A. (1991) *Nucleic Acids Res.*, **19**, 1687–1693.
42 Nikoloff, D.M. and Henry, S.A. (1994) *J. Biol. Chem.*, **269**, 7402–7411.
43 Nishi, K., Park, C.S., Pepper, A.E., Eichinger, G., Innis, M.A. and Holland, M.J. (1995) *Mol. Cell. Biol.*, **15**, 2646–2653.
44 Rothermel, B.A., Shyjan, A.W., Etheredge, J.L. and Butow, R.A. (1995) *J. Biol. Chem.*, **270**, 29476–29482.
45 Fisher, F. and Goding, C.R. (1992) *EMBO J.*, **11**, 4103–4109.
46 Lopes, J.M. and Henry, S.A. (1991) *Nucleic Acids Res.*, **19**, 3987–3994.
47 Ashburner, B.P. and Lopes, J.M. (1995) *Mol. Cell. Biol.*, **15**, 1709–1715.
48 Paltauf, F., Kohlwein, S.D. and Henry, S.A. (1992) In Jones, E.W., Pringle, J.R. and Broach, J.R. (eds) *The Molecular and Cellular Biology of the Yeast Saccharomyces. Vol. II*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
49 Koipally, J., Ashburner, B.P., Bachhawat, N., Gill, T., Hung, G., Henry, S.A. and Lopes, J.M. (1996) *Yeast*, **12**, 653–665.
50 Kondrakin, Y.V., Kel, A.E., Kolchanov, A.E., Romashchenko, A.G. and Milanesi, L. (1995) *Comput. Appl. Biosci.*, **11**, 477–488.
51 Chen, Q.K., Hertz, G.Z. and Stormo, G.D. (1997) *Comput. Appl. Biosci.*, **13**, 29–35.
52 Frech, K., Quandt, K. and Werner, T. (1997) *Comput. Appl. Biosci.*, **13**, 89–97.