BIOINFORMATICS

Vol. 15 no. 10 1999
Pages 776-784

Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes

Andreas Wagner

Department of Biology, University of New Mexico and The Santa Fe Institute, 167A Castetter Hall, Albuquerque, NM 87131-1091, USA

Received on January 13, 1999; revised on March 20, 1999; accepted on April 8, 1999

Abstract

Motivation: *The question addressed here is how cooperative interactions among transcription factors (TFs), a very frequent phenomenon in eukaryotic transcriptional regulation, can be used to identify genes that are regulated by one or more TFs with known DNA binding specificities. Cooperativity may be homotypic, involving binding of only one transcription factor to multiple sites in a gene's regulatory region. It may also be heterotypic, involving binding of more than one TF. Both types of cooperativity have in common that the binding sites for the respective TFs form tightly linked 'clusters', groups of binding sites often more closely associated than expected by chance alone.*

Results: *A statistical technique suitable for the identification of statistically significant homotypic or heterotypic TF binding site clusters in whole eukaryotic genomes is presented. It can be used to identify genes likely to be regulated by the TFs. Application of the technique is illustrated with two transcription factors involved in the cell cycle and mating control of the yeast Saccharomyces cerevisiae, indicating that the results obtained are biologically meaningful. This rapid and inexpensive computational method of generating hypotheses about gene regulation thus generates information that may be used to guide subsequent costly and laborious experimental approaches, and that may aid in the assignment of biological functions to putative open reading frames.*

Availability: *Software used for statistical analysis is available from the author upon request.*

Contact: *wagnera@unm.edu*

Introduction

The identification of regulatory regions in eukaryotic DNA has been the focus of great research interest in areas as diverse as microbial genetics and mammalian developmental biology. With several eukaryotic genome projects nearing completion, the unprecedented challenge of characterizing regulatory regions in entire genomes has arisen. The statistical techniques developed and applied here are concerned with a genome-wide characterization of regulatory regions mediating transcriptional regulation of protein coding genes.

All eukaryotes share a highly conserved mechanism of transcriptional regulation (Ptashne, 1988; Ptashne and Gann, 1997). Central to this mechanism are proteins called transcription factors (TFs), which bind to specific, short DNA sequence motifs in the *cis*-regulatory region (promoter, enhancer) of a gene and activate or repress its transcription. Because experimental characterization of enhancers is difficult, computational techniques leading to their tentative characterization have a long tradition. However, the success of these techniques is often very limited. It is common practice among molecular biologists to screen the DNA region near a gene of interest for the occurrence of specific DNA sequence motifs that are binding sites for known TFs, an approach that is easily extended to a genome-wide scale (Das *et al.*, 1997). The hope is that the encountered binding sites point towards TFs that play a role in the transcriptional regulation of that gene. This is often not the case, because several factors, such as chromatin structure, may influence the ability of a TF to bind a site or to regulate transcription. Also, encounters of non-functional TF binding sites are to be expected, given that many binding sites are abundant in genomic DNA.

A key feature of eukaryotic transcriptional regulation is that genes are often regulated by more than one TF. An illustrative example from higher eukaryotes is the developmental gene CyIIIa from the sea urchin *Strongylocentrotus purpuratus.* Its promoter comprises 2300 bp upstream of the coding region. At least nine different TFs regulate the expression of this gene via 23 binding sites contained in the 2.3 kb regulatory region (Kirchhamer *et al.*, 1996). Interactions among DNA-bound factors at such a promoter can be homotypic, in which case they involve interactions among multiple bound factors of the same kind; heterotypic, involving interactions among TFs of different kinds; or both (Figure 1). The method presented here will take advantage of the ubiquitous occurrence of homo- and heterotypic interactions at eukaryotic promoters, and the associated close spacing of TF binding sites. Groups ('clusters') of very closely spaced TF binding sites within the regulatory region of a gene are unlikely to have occurred 'by

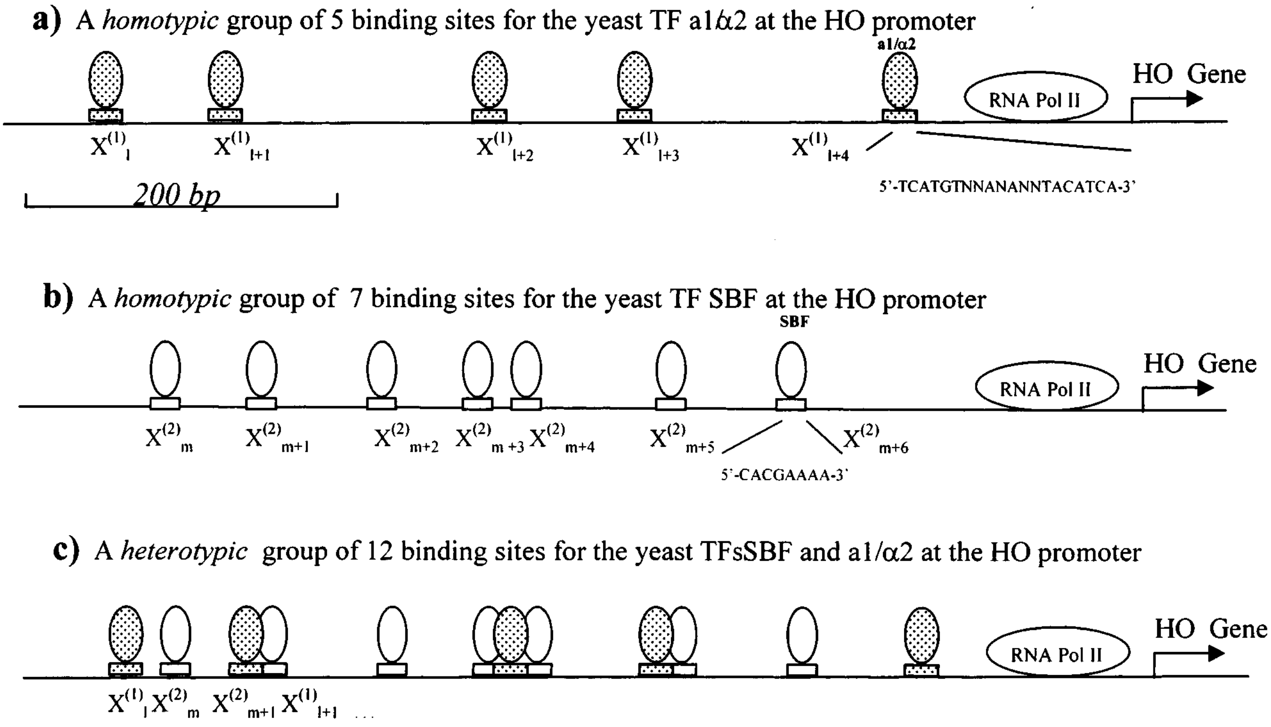
a) A *homotypic* group of 5 binding sites for the yeast TF a1/α2 at the HO promoter
a1/α2
RNA Pol II
HO Gene
$X^{(1)}_l$
$X^{(1)}_{l+1}$
$X^{(1)}_{l+2}$
$X^{(1)}_{l+3}$
$X^{(1)}_{l+4}$
200 bp
5'-TCATGTNNANANNTACATCA-3'
b) A *homotypic* group of 7 binding sites for the yeast TF SBF at the HO promoter
SBF
RNA Pol II
HO Gene
$X^{(2)}_m$
$X^{(2)}_{m+1}$
$X^{(2)}_{m+2}$
$X^{(2)}_{m+3}$
$X^{(2)}_{m+4}$
$X^{(2)}_{m+5}$
$X^{(2)}_{m+6}$
5'-CACGAAAA-3'
c) A *heterotypic* group of 12 binding sites for the yeast TFsSBF and a1/α2 at the HO promoter
RNA Pol II
HO Gene
$X^{(1)}_l X^{(2)}_m$ $X^{(2)}_{m+1} X^{(1)}_{l+1}$ ...

Fig. 1. An example for homotypic and heterotypic groups of TF binding sites from the *S.cerevisiae* HO endonuclease promoter on chromosome IV (after Lewin, 1994, p. 1073). The five binding sites for TF$_1$, a1/α2, are located at positions $X^{(1)}_l$,...$X^{(1)}_{l+4}$ where $X^{(1)}_l$ would correspond to the a1/α2 binding site closest to the left telomere of chromosome IV. Similarly, the positions for the binding sites of TF$_2$, SBF, are indexed $X^{(2)}_m$,...,$X^{(2)}_{m+6}$. If binding sites for TF$_1$ and TF$_2$ are independently Poisson distributed with parameters (probabilities of site occurrence) $\lambda^{(1)}$ and $\lambda^{(2)}$, respectively,and if the positions $X^{(i)}_j$ of the two TFs are considered jointly [as indicated in (c)], then the joint site distribution is Poisson with parameter $\lambda^{(1)} + \lambda^{(2)}$

chance alone' (in contrast to individual sites). Rather, clustering of TF binding sites may indicate that the respective TFs are involved in the gene's regulation. The key idea underlying the technique developed below is that this observation, when formulated in a statistically rigorous way and applied to an entire genome, can be used to detect the 'best' candidate genes for regulation by one or more TFs. These are the genes whose *cis*-regulatory regions contain clusters of TF binding sites so closely linked that they are unlikely to have occurred by chance alone. The technique is not designed to identify as many genes as possible that are regulated by one or more TFs of interest, but only the best candidates. The price paid for such conservativism is that many genes regulated by a TF will not be detected. It is a price well worth paying, because a conservative approach will generate candidate genes that seriously merit further experimental investigation. Given the ubiquity of cooperativity in transcriptional regulation, it is perhaps surprising that no statistically rigorous techniques are currently available to ask questions about the combinatorial transcriptional regulation of specific genes.

The genome of the yeast *Saccharomyces cerevisiae*, currently the only eukaryote for which a complete, well-annotated genome sequence is available, will be used to illustrate two simple applications of the technique. Yeast has advantages as well as disadvantages for this type of analysis. Because of its fairly short (~600 bp on average) and less complex upstream regulatory regions compared to higher eukaryotes, yeast is probably not the best eukaryotic organism to screen for cooperative interactions among known TFs. On the other hand, the small size and organization of the yeast genome provide a number of advantages, such as that potential yeast promoter regions are in general located upstream of the coding region (Struhl, 1989, 1995), and that the yeast genome does not contain many tandemly repeated sequences other than rDNA and CUP genes (Olson, 1992).

Method and results

The statistical tests in this section are designed for the identification of clusters of TF binding sites. Tests are detailed for two TFs with different binding sites, TF$_1$ and TF$_2$. Generalization to three or more TFs is straightforward.

The point of departure is a null hypothesis, which is here that binding sites for TF$_i$ are Poisson distributed in genomic DNA with parameter $\lambda^{(i)}$ (Karlin and Taylor, 1975). This hypothesis is somewhat problematic, because it can be violated for reasons that have nothing to do with cooperative interactions in transcriptional regulation. However, the two most important confounding factors can be eliminated. The first has to do with the structure of the site itself. Very short sites, longer sites in which a large number of mismatches or nucleotide ambiguities are 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. Because such sequence features may not always be obvious, it is best to eliminate this possibility by testing whether site distribution is consistent with a Poisson distribution in a long stretch of computationally generated pseudorandom DNA. This is most conveniently done via a goodness-of-fit test that assesses whether the distances between adjacent sites are exponentially distributed (Sokal and Rohlf, 1981). The second confounding factor concerns the parameter λ, which is the probability of finding a binding site for a TF at an arbitrarily chosen position in the genome. The enormous compositional heterogeneity of genomic DNA implies that site occurrence probability may vary across regions of the genome, such that the assumption of a constant probability λ is unrealistic (for an example, see Figure 2). This issue is more difficult to address. Despite some advances in statistical modeling of DNA sequences (e.g. Almagor, 1983; Kleffe and Langbecker, 1990; Henderson *et al.*, 1997; for a review, see Li, 1997), there currently exists no satisfactory statistical model of DNA, that accounts for both the heterogeneity and non-stationary properties (Bernardi *et al.*, 1988; Karlin and Brendel, 1993) of genomic DNA. However, while there is currently no completely satisfactory solution to this problem, it will be alleviated here by incorporating information on both global (genome-wide) and local DNA composition into the statistical analysis.

The following section first reviews a significance test that asks whether k consecutive binding sites for one TF, i.e. a homotypic group of binding sites, are more tightly linked than expected by chance alone (Wagner, 1997). It then extends the test to groups of binding sites for two TFs, i.e. heterotypic groups of binding sites. The section after that is concerned with variation in site occurrence probability due to compositional variation along a chromosome.

Significance measures for groups of binding sites

Denote as $(X_1, \ldots, X_K)$ the positions at which a transcription factor binding site S or its reverse complement are encountered on the DNA. Further, define as X_0 the beginning (5′ end of the top strand) of the DNA sequence (genome) to be analyzed. The quantity $D_{i,j} = X_j - X_i$ denotes the distance between site X_j and X_i. $D_{i,i+k-1} = \sum_{j=0}^{k-2} D_{i+j,i+j+1}$ $(k > 1)$is the length of a stretch of DNA spanning exactly k sites. It will be referred to as a homotypic k-cluster. Under the Poisson null hypothesis, the distribution of the distance between successive sites (2-clusters), $D_{i,i+1}$, is exponential with parameter λ, where λ is the probability of finding a binding site at a specific position on the DNA. It follows that the length z of k-clusters follows a Pearson Type III distribution with density:

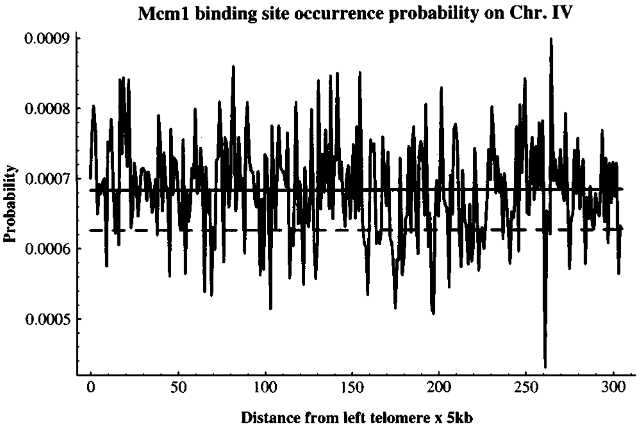

Fig. 2. Three estimators for the probability of occurrence of the Mcm1 binding site with consensus sequence 5-DCCYWWWNNRG-3 on chromosome IV of *S.cerevisiae*. The solid fluctuating graph is the estimated probability $\lambda_l(y)$ based on the dinucleotide composition of consecutive 5 kb windows spanning the chromosome. To obtain this estimator, all DNA words matching the consensus and containing only letters ACTG were explicitly evaluated, and their probabilities of occurrence, calculated as described in Wagner (1997), were added. The solid horizontal line is the average of this estimator over all windows. The dashed horizontal line is the estimator $\lambda_g = K/N$ which is based on the actual number, K = 959, of MCM1 binding sites on this chromosome of length $N \approx 1.53 \times 10^6$ nucleotides.

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

where $\Gamma(k) = (k-1)!$ is the gamma function. This can be seen from the characteristic function of the exponential distribution, $\phi(t) = (1 - it/\lambda)^{-1}$ (Abramowitz and Stegun, 1972, 26.1.31).For a given significance level P, an observed k-cluster of length x is called significantly shorter than expected by chance alone (i.e. the null hypothesis is rejected) if

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

The extension to two TFs, TF_1 and TF_2, is straightforward. In complete analogy to homotypic groups of binding sites, the null hypothesis here will be that the binding sites for these TFs, $S^{(1)}$ and $S^{(2)}$,are independently distributed in the genome, and that each follows the Poisson model with parameters $\lambda^{(1)}$ and $\lambda^{(2)}$, respectively. More specifically, denote as $(X_1^{(i)}, \ldots, X_{K_i}^{(i)})$ the $K_i(N)$ positions at which a binding site for

TF$_i$ occurs along a stretch of DNA comprising N nucleotides. $K_i(N)$ can be viewed as a Poisson process with parameter $\lambda^{(i)}$. It follows from the characteristic function of $K_i(N)$, $\phi_i(t) = \exp[\lambda^{(i)}N(e^{it} - 1)]$ that the sum $K_1(N) + K_2(N)$ is again a Poisson process with characteristic function $\phi_1(t) \times \phi_2(t) = \exp[(\lambda^{(1)} + \lambda^{(2)})N(e^{it} - 1)]$; hence, a Poisson process with parameter $\lambda = \lambda^{(1)} + \lambda^{(2)}$ (see also Karlin and Taylor, 1975). This is convenient, since it permits 'pooling' of information from occurrences of sites for each of the two TFs. The resulting new Poisson process can be analyzed with the techniques already introduced for homotypic groups of binding sites. Specifically, given a significance level P, a group of k consecutive binding sites for two transcription factors (a heterotypic k-cluster), and the number of nucleotides x from the first to the last site in the group, significant clustering is observed if

$$\frac{\lambda^{(1)} + \lambda^{(2)}}{\Gamma(k-1)} \int_0^x \left[(\lambda^{(1)} + \lambda^{(2)})z\right]^{k-2} e^{-(\lambda^{(1)}+\lambda^{(2)})z} dz < P \qquad (3)$$

As in the case of homotypic interactions, TF binding sites can violate the null hypothesis for reasons inherent in the sites themselves. The hypothetical TF binding sites 5-CATGGC-3 [$S^{(1)}$] and 5-ATAGCCA-3 [$S^{(2)}$] serve as examples. $S^{(2)}$ overlaps the reverse complement of $S^{(1)}$ by 4 bp, thus site occurrences will not be statistically independent. While $K_1(N)$ and $K_2(N)$ may behave as Poisson processes when taken individually, the sum $K_1(N) + K_2(N)$ is no longer a Poisson process. It is thus advisable to test, via a long stretch of pseudorandom DNA and a goodness-of-fit test, whether $K_1(N)$, $K_2(N)$ and $K_1(N) + K_2(N)$ are Poisson processes with parameters $\lambda^{(1)}$, $\lambda^{(2)}$ and $[\lambda^{(1)} + \lambda^{(2)}]$, respectively. If they are not, the respective sites are not amenable to this type of analysis.

A critical issue here is to choose an appropriate significance level P. When analyzing entire genomes for clusters of TF binding sites, a large number of significance tests are carried out, not all of which are mutually independent. For example, for $K = 10^3$ encountered sites and for the analysis of heterotypic 5-clusters ($k = 5$),there are of the order of 10^3 groups of heterotypic 5-clusters. These can be grouped into approximately $K/(k - 1) = 250$ non-overlapping 5-clusters. Significance tests for non-overlapping 5-clusters are independent. $1/P$ should be greater than the number of independent significance tests to avoid high Type I error probabilities (Sokal and Rohlf, 1981). For a given k, and a total number of K binding sites in the genome, the significance threshold $P_k = (k - 1)/K$ is used here.

A minor complication occurs if the binding sites for one TF are more abundant than those for the other TF, e.g. if $\lambda^{(2)} >> \lambda^{(1)}$. It can be shown that the mean number of sites $S^{(2)}$ between two consecutive occurrences of site $S^{(1)}$ is given by $\lambda^{(2)}/\lambda^{(1)}$. Thus, if $\lambda^{(2)} >> \lambda^{(1)}$, many sites $S^{(2)}$ may lie between two consecutive occurrences of $S^{(1)}$. In this case, the test introduced thus far will not be very sensitive to heterotypic associations, but may largely measure homotypic site interactions for the more frequent site. Various remedies for this situation, which will be explored in a forthcoming contribution, are conceivable. They revolve around the analysis of clusters that include heterotypic site pairs, i.e. adjacent sites of different types.

Estimation of site occurrence probability λ

It has so far been assumed that site occurrence probability $\lambda^{(i)}$ is constant along a chromosome. In view of compositional heterogeneity of genomic DNA, this assumption has to be relaxed, which changes the statistical model to that of an inhomogeneous Poisson process (Parzen, 1962). To model variation of $\lambda^{(i)}$ appropriately, two complementary estimates of $\lambda^{(i)}$are used. For simplicity of notation, consider for the moment only TF$_1$ and set $\lambda = \lambda^{(1)}$. The first estimate is a global estimate λ_g, which is the number of sites K found per N nucleotides, i.e. $\lambda_g = K/N$. This is a maximum-likelihood estimator, whose sampling standard deviation scales as $1/\sqrt{K}$ (Kendall, 1952, p. 22); hence, the need to base the estimate on large stretches of DNA.

The second estimate of site occurrence probability is a local estimate $\lambda_l(y)$ whose value is based on the dinucleotide composition in a (short) region of interest around a location y in the genome. It is currently limited to dinucleotide composition and assumes that the underlying DNA sequence has a Markov property (Karlin and Taylor, 1975; Karlin and Macken, 1991a,b; Wagner, 1997). While certainly not the only way to model compositional heterogeneity (for a review, see Li, 1997), this approach is chosen here because of its computational tractability for the large genomic DNA regions to be analyzed.

λ_g and $\lambda_l(y)$, taken individually, are not adequate. $\lambda_l(y)$ alone, applied to each location y in the genome, has the undesirable property that its average, $\overline{\lambda_l(y)}$, over entire chromosomes does often not agree with the observed quantity λ_g (for an example, see Figure 2). The reason for this discrepancy, which has also been observed for the distribution of restriction sites in *Escherichia coli* (Karlin and Macken, 1991a), may have to do with higher order correlations among nucleotides as well as with selective pressures that affect the abundance of specific sequence motifs because they have some unknown function in the cell. λ_g, on the other hand, completely ignores the enormous variability in site occurrence probabilities along a chromosome (Figure 2). In light of these observations, the following compound estimator for site occurrence probability is proposed.

$$\lambda(y) = \lambda_g + \lambda_l(y) - \overline{\lambda_l(y)} \qquad (4)$$

$\overline{\lambda_l(y)}$ is a site occurrence probability estimate based on the average dinucleotide composition of a chromosome. $\lambda(y)$ takes both local sequence composition and observed genomic site counts into account, and it compensates for differences between $\overline{\lambda_l(y)}$ and λ_g. Following the last section:

$$\lambda^{(1)}(y) + \lambda^{(2)}(y) = \lambda_g^{(1)} + \lambda_g^{(2)} + \lambda_l(y)^{(1)} + \lambda_l(y)^{(2)} - \overline{\lambda_l(y)}^{(1)} - \overline{\lambda_l(y)}^{(2)}$$

is the appropriate measure for the compound process comprising two different TFs.

A location y on a chromosome is represented here by a window over which dinucleotide composition is evaluated. Since statistical significance of groups of TF binding sites is at issue here, nucleotide composition is calculated for any analyzed group of k consecutive sites from the DNA sequence between the first and last site in the group. The implicit assumption here, made for reasons of computational feasibility, is that nucleotide composition is constant within this window. For very tightly linked groups of binding sites, an important source of statistical bias are those mono- and dinucleotides that are contained in individual binding sites. They will be overly frequent. To avoid this potential problem, base composition is calculated in a 500 bp window centered around the group for groups of binding sites spanning <500 bp.

To ensure wide applicability of the technique, conventional consensus sequences are used instead of position weight matrices (PWMs; Stormo, 1990; Fickett, 1996) in determining the number K of matches to a site. When available, a good PWM is vastly superior to a simple consensus sequence, because it makes use of a much larger amount of binding data. Indeed, it has been shown in particular cases (Fickett, 1996) that PWM-based models provide accurate estimates of the binding affinity of a TF at its site. However, because this information is not easily obtained, the number of TFs for which well-supported PWMs are available is small.

In applying the above method to the genome of *S.cerevisiae*, the following steps were taken. First, for two different transcription factors of interest, the number and positions of all their binding sites in the genome are recorded. If two binding sites for the same factor overlap, one of the two sites (randomly chosen) is eliminated from further analysis, the rationale being that usually two overlapping sites can be occupied by only one TF. To increase the accuracy of estimates for $\lambda_g^{(i)}$, site counts were pooled from all 16 chromosomes. Then, for $k = 2$ to $k = 15$, the significance of all heterotypic k-clusters, i.e. groups of k consecutive binding sites, regardless of site type [$S^{(1)}$ or $S^{(2)}$], is evaluated. Only k-clusters with greater statistical significance than the threshold P_k located in the upstream region of an open reading frame (ORF) are listed below. Considering that there may be thousands of TF binding sites in a genome, and considering that local dinucleotide composition is evaluated for each k-cluster, it becomes evident that the computational requirements are considerable.

Application to the yeast TFs Mcm1 and Ste12

Mcm1 and Ste12 are two key regulators of cell cycle and mating response. MCM1, originally identified as a gene required for minichromosome maintenance (Maine *et al.*, 1984), encodes a transcription factor that is a close relative to the mammalian serum response factor, SRF (Wynne and Treisman, 1992). In cooperation with the TF α1, Mcm1 activates the transcription of α-cell type-specific genes; in cooperation with α2, it represses transcription of a-type-specific genes in α-cells (for a review, see Dolan and Fields, 1991); in cooperation with Sff, it regulates G_2-specific transcription (Althoefer *et al.*, 1995). Furthermore, Mcm1 has been implicated in the regulation of arginine metabolism, as well as in the synthesis of cell wall and cell membrane structures (Messenguy and Dubois, 1993; Kuo and Grayhack, 1994). Mcm1 can bind to DNA by itself *in vitro*, but its affinity is increased in the presence of the appropriate cofactor (Bender and Sprague, 1987). Two independent studies provide information on the range of DNA sequences bound by Mcm1. One of the studies used an *in vitro* selection scheme, starting from a library of yeast genomic DNA to identify promoter fragments strongly bound by Mcm1 (Kuo and Grayhack, 1994); the other study selected Mcm1 binding sites from a pool of random DNA sequences (Wynne and Treisman, 1992). The consensus binding sites derived from these studies are 5-TTTCCNAWWNNRGNAA-3 and 5-DCCYWWWNNRG-3, respectively, similar to the recognition site deduced earlier from mating-type genes regulated by Mcm1 (Dolan and Fields, 1991).

The transcription factor Ste12 regulates both the basal and mating pheromone-induced transcription of many genes involved in mating. Regulation is mediated by binding of Ste12 to at least one pheromone-responsive element (PREs; 5-TGAAACA-3; Sprague and Thorner, 1992). Cooperative binding at multiple PREs or with other TFs greatly enhances transcriptional activation via Ste12 (Yuan and Fields, 1991; Sprague and Thorner, 1992). The basal expression level of FAR1, a gene necessary for mating pheromone-induced cell cycle arrest, is cell cycle regulated with expression peaking in G_1 and in the G_2/M phase (Oehlen *et al.*, 1996). This regulation is functionally important. For example, elimination of the G_1 expression peak causes failure of cell cycle arrest in response to mating pheromone. Ste12 and Mcm1 jointly regulate the basal expression of FAR1. Given the importance of all three genes in cell cycle regulation, it is natural to ask what other important genes might be regulated jointly by Mcm1 and Ste12. To address this question with the approach pursued here, one first has to establish that the respective binding sites, when considered both separately and jointly, are Poisson distributed in random DNA. This is the case (Table 1). Moreover, on the coarse level of resolution provided by a conventional goodness-of-fit test, the sites appear Poisson distributed in yeast genomic DNA as well. Table 2 shows the number of significant clusters of binding sites ranked by their significance, i.e. the statistically most unlikely (most tightly linked) clusters appear on top of the table. Two ORFs in one entry of the table indicate that the respective cluster occurs in the upstream region of both adjacent ORFs, i.e. the ORFs are in a head-to-head orientation. Because information from both types of sites is pooled in the technique applied here, Table 2 includes clusters that contain only Mcm1 binding sites, only Ste12 binding sites, as well as heterotypic clusters.

Table 1. Binding site counts and tests for Poisson distribution of Mcm1 and Ste12 binding sites in the *S.cerevisiae* genome and in random DNA

Site	Yeast genome				Random DNA	
	Mismatches allowed	No. of sites	χ^2 (d.f., *P*)	G (d.f., *P*)	χ^2 (d.f., *P*)	G (d.f., *P*)
MCM1	0	7036	12.7 (9, 0.18)	12.5 (9, 0.19)	11.4 (9, 0.25)	12.9 (9, 0.17)
STE12	0	3400	3.5 (8, 0.9)	3.47 (8, 0.9)	2.6 (8, 0.96)	2.6 (8, 0.96)
MCM1/STE12		10 436	9.6 (10, 0.47)	9.7 (10, 0.47)	5.2 (9, 0.81)	6.0 (9, 0.74)

χ^2 and G (likelihood ratio) tests (Sokal and Rohlf, 1981) test for consistency with a null hypothesis of exponential inter-site distributions. None of the *P* values in the table suggest rejection of the null hypothesis. The Mcm1 consensus sequence 5-DCCYWWWNNRG-3 (no mismatches allowed; Wynne and Treisman, 1992) was used here. Analogous results (not shown) are obtained with an alternative Mcm1 consensus sequence, 5-TTTCCNAWWNNRGNAA-3 (one mismatch allowed; Kuo and Grayhack, 1994).

Table 2. Candidate genes for regulation by Mcm1, Ste12 or both factors

Chromosome	ORF	Sites/ bp[a]	Mcm1 (consensus)/ Ste12[b]	Position[c]	Statistical significance *P* [d]	Structure or function[e]
12	YLR037C	8/741	7 (sh)/1	–350	9.3×10^{-6}	Unknown
7	SCW4	6/432	5 (sh)/1	–274/–257	4.6×10^{-5}	Cell wall protein
14	ERG24/YNL279W	3/29	0 (lo)/3	–325/–159	4.9×10^{-5}	Sterol C-14 reductase/Unknown
6	YFL027C/STE2+	5/306	3 (sh)/2	–239/–134	6.7×10^{-5}	Unknown/Mating pheromone receptor
1	CLN3+	3/84	3 (lo)/0	–890	7.6×10^{-5}	G1 cyclin
8	STE12	4/166	1 (sh)/3	–319	8.0×10^{-5}	Transcription factor
14	YGP1	6/484	6 (sh)/0	–368	1.2×10^{-4}	Glycoprotein synthesized in response to nutrient limitation
10	ELO1/CDC6+	3/74	3 (lo)/0	–321/–177	1.2×10^{-4}	Fatty acid biosynthesis/Initiation of DNA replication
14	SUI1/SLA2	4/178	3 (sh)/1	–316/–68	1.3×10^{-4}	Translation factor/Membrane cytoskeleton assembly
14	CLA4	3/93	2 (lo)/1	–330	1.8×10^{-4}	Protein kinase, septin ring formation in cytokinesis
10	PRY3	4/510	2 (lo)/2	–155	1.9×10^{-4}	Similar to plant pathogen related protein
6	YFR044C/YFR045W	5/357	4 (sh)/1	–366/–313	2.6×10^{-4}	Unknown/Unknown
16	YPL059W	9/2507	5 (sh)/4	–422	2.9×10^{-4}	Unknown
2	UBC4/TEC1	4/422	0 (lo)/4	–1491/–94	3.5×10^{-4}	Ubiquitin-conjugating enzyme/TF regulating Ty1 expression

[a] Number of TF binding sites in the significant cluster/cluster length, from first to last site, in base pairs.

[b] Number of Mcm1 and Ste12 binding sites in the cluster, respectively. 'sh' and 'lo' in parentheses indicate whether the short or the long MCM1 consensus sequence was used (see below).

[c] Binding site closest to first codon of ORF.

[d] Statistical significance based on global site count in the genome, and on local dinucleotide distribution, as explained in Method and results.

[e] From the *S.cerevisiae* genome database (http://genome-www.stanford.edu/Saccharomyces); see also references in the text.

+, Gene is known to be regulated by one or both TFs. Clusters shown are based on separate searches for two Mcm1 consensus sequences, 5-DCCYWWWNNRG-3 ('sh'; no mismatch allowed; Wynne and Treisman, 1992) and 5-TTTCCNAWWNNRGNAA-3 ('lo'; one mismatch allowed; Kuo and Grayhack, 1994).

Notably, two of the three most tightly linked clusters are associated with ORFs of unknown function. One of them contains only binding sites for Ste12, making it a candidate for regulation by Ste12 alone or in cooperation with factors other than Mcm1. The fourth cluster is associated with STE2, which is the only gene other than FAR1 whose regulation by

Ste12 and Mcm1 has been demonstrated experimentally (Hwang-Shum *et al.*, 1991). The genes CLN3 and CDC6 are known experimentally to be regulated by Mcm1. Consistent with this, they are associated with significant clusters of Mcm1 sites. These three examples show that the method identifies genes known to be regulated by the respective factors. The STE12 gene itself is associated with a significant cluster of four sites, three of which are perfect matches to the PRE, suggesting the possibility of Ste12 autoregulation in response to pheromone. Consistent with this is the observation that STE12 transcription is induced moderately in response to pheromone (Sprague and Thorner, 1992). The remaining ORFs in the table include ORFs of unknown function, as well as known genes, most of which are poorly characterized. Among them, the most promising candidates for further investigation may be SLA2 and CLA4, based on their and MCM1's implication in cell wall synthesis and maintenance (Table 2 and Kuo and Grayhack, 1994). For example, CLA4, a protein kinase, is involved in formation of the septin ring required for cytokinesis (Cvrckova *et al.*, 1995). Its activity is cell cycle regulated, peaking near mitosis (Benton *et al.*, 1997). Another such candidate is the otherwise poorly characterized cell wall protein SCW4 (see also below).

Among their other functions, Mcm1 and Ste12 regulate the expression of cell cycle regulated genes, both separately and jointly (Althoefer *et al.*, 1995; Oehlen *et al.*, 1996). Thus, if Table 2 contains genes other than those discussed above that are cell cycle regulated in a pattern consistent with that found for other Mcm1 and Ste12 regulated genes, such genes are prime candidates for further study. Circumstantial evidence of this sort is provided by Spellman *et al.* (1998) in a paper using micro array technology to determine which of all yeast genes are cell cycle regulated. According to these results, five additional genes in Table 2 show cell cycle-dependent expression. These are SCW4 (peak expression in M), YGP1 (M/G_1), ELO1 (M/G_1), PRY3 (G_1) and TEC1 (M/G_1; Spellman *et al.*, 1998). While a consistent expression pattern is certainly not proof of a regulatory interaction, it certainly serves to identify candidates for further study.

Discussion

Statistical significance of TF binding site clusters does not, of course, imply biological significance. However, the observation that the method detects genes known to be regulated by the studied TFs suggests that its results are biologically meaningful. This is further supported when genes are found whose regulation by the studied TFs has not been demonstrated, but for which circumstantial experimental evidence points towards such regulation. By pursuing a statistically conservative approach of forming hypotheses only based on clusters of TF binding sites, and not on individual sites alone, it is hoped that the enormously high false-positive rate normally associated with *in silico* promoter characterization can be lowered. Precise estimates of this false-positive rate would require knowledge of the genes regulated by a TF of interest. This knowledge is likely to become available soon, as large-scale expression studies in suitable mutants become possible in *S.cerevisiae* (DeRisi *et al.*, 1997).

Higher eukaryotes have regulatory regions that are incomparably more complex than most yeast promoters. They are thus better candidates for fruitful applications of this method. However, they also pose unique problems 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. Given these complexities, one method alone will not be sufficient to characterize regulatory regions, and information from several complementary techniques may have to be combined. Many such techniques have been developed in the recent past (for reviews, see Duret and Bucher, 1997; Fickett and Hatzigeorgiou, 1997; Lavorgna *et al.*, 1998). They fall into two categories, the first of which attempts to distinguish between promoter and non-promoter sequences, based on (i) distinct oligonucleotide distribution profiles (Chen *et al.*, 1997; van Helden *et al.*, 1998), (ii) detection of complex regulatory modules, such as retroviral long terminal repeats (Frech *et al.*, 1997), (iii) differential distribution of individual known TF binding sites and TATA boxes (Kel *et al.*, 1995; Kondrakhin *et al.*, 1995; Prestridge, 1995, 1996; Kolchanov *et al.*, 1998), (iv) pattern recognition algorithms based on neural networks (Lukashin *et al.*, 1989; O'Neill, 1991; Matis *et al.*, 1996), (v) Bayesian statistics (Crowley *et al.*, 1997) and (vi) phylogenetic footprinting (Shelton *et al.*, 1997). A second group of methods, such as the one presented here, attempts to identify more specific regulatory interactions (e.g. Fickett, 1996; Brazma *et al.*, 1997, 1998, Kolchanov *et al.*, 1998; Wasserman and Fickett, 1998). Each of these has its unique strengths and weaknesses. In combination, they will aid in sifting through an astronomical number of possible gene interactions, and identify candidates worthy of further experimental investigation, at a cost incomparably lower than any experimental approach.

Acknowledgements

The author wishes 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 and of the Albuquerque High Performance ComputingCenter (AHPCC) is gratefully acknowledged.

References

Abramowitz,M. and Stegun,I.A. (1972) *Handbook of Mathematical Functions*. Dover, New York.

Almagor,H. (1983) A Markov analysis of DNA sequences. *J. Theor. Biol.*, **104**, 633–645.

Althoefer,H., Schleiffer,A., Wassmann,K., Nordheim,A. and Ammerer,G. (1995) Mcm1 is required to coordinate G2-specific transcription in *Saccharomyces cerevisiae. Mol. Cell. Biol.*, **15**, 5917–5928.

Bender,A. and Sprague,G.F.,Jr (1987) MAT alpha 1 protein, a yeast transcription activator, binds synergistically with a second protein to a set of cell-type-specific genes. *Cell*, **50**, 681–691.

Benton,B.K., Tinkelenberg,A., Gonzalez,I. and Cross,F.R. (1997) Cla4p, a *Saccharomyces cerevisiae* Cdc42p-activated kinase involved in cytokinesis, is activated at mitosis. *Mol. Cell. Biol.*, **17**, 5067–5076.

Bernardi,G., Mouchiroud,D., Gautier,C. and Bernardi,G. (1988) Compositional patterns in vertebrate genomes: conservation and change in evolution. *J. Mol. Evol.*, **28**, 7–18.

Brazma,A., Vilo,J., Ukkonen,E. and Valtonen,K. (1997) Data mining for regulatory elements in the yeast genome. *Intell. Syst. Mol. Biol.*, **5**, 65–74.

Brazma,A., Jonassen,I., Vilo,J. and Ukkonen,E. (1998) Prediciting gene regulatory elements on a genomic scale. *Genome Res.*, **8**, 1202–1215.

Chen,Q.K., Hertz,G.Z. and Stormo,G.D. (1997) PromFD 1.0: a computer program that predicts eukaryotic pol II promoters using strings and IMD matrices. *Comput. Applic. Biosci.*, **13**, 29–35.

Crowley,E.M., Roeder,K. and Bina,M. (1997) A statistical model for locating regulatory regions in genomic DNA. *J. Mol. Biol.*, **268**, 8–14.

Cvrckova,F., De Virgilio,C., Manser,E., Pringle,J.R. and Nasmyth,K. (1995) Ste20-like protein kinases are required for normal localization of cell growth and for cytokinesis in budding yeast. *Genes Dev.*, **9**, 1817–1830.

Das,S., Yu,L., Gaitatzes,C., Rogers,R., Freeman,J., Blenkowska,J., Adams,R.M., Smith,T.F. and Lindellen,J. (1997) Biology's new Rosetta stone. *Nature*, **385**, 29–30.

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

Dolan,J.W. and Fields,S. (1991) Cell-type-specific transcription in yeast. *Biochim. Biophys. Acta*, **1088**, 155–169.

Duret,L. and Bucher,P. (1997) Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.*, **7**, 399–406.

Fickett,J.W. (1996) Quantitative discrimination of MEF2 sites. *Mol. Cell. Biol.*, **16**, 437–441.

Fickett,J.W. and Hatzigeorgiou,A.G. (1997) Eukaryotic promoter recognition. *Genome Res.*, **7**, 861–878.

Frech,K., Danescu-Mayer,J. and Werner,T. (1997) A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. *J. Mol. Biol.*, **270**, 674–687.

Henderson,J., Salzberg,S. and Fasman,K.H. (1997) Finding genes in DNA with a Hidden Markov Model. *J. Comput. Biol.*, **4**, 127–141.

Hwang-Shum,J.J., Hagen,D.C., Jarvis,E.E., Westby,C.A. and Sprague,G.F.,Jr (1991) Relative contributions of MCM1 and STE12 to transcriptional activation of a- and alpha-specific genes from *Saccharomyces cerevisiae. Mol. Gen. Genet.*, **227**, 197–204.

Karlin,S. and Brendel,V. (1993) Patchiness and correlations in DNA sequences. *Science*, **259**, 677–680.

Karlin,S. and Macken,C. (1991a) Assessment of inhomogeneities in an *E. coli* physical map. *Nucleic Acids Res.*, **19**, 4241–4246.

Karlin,S. and Macken,C. (1991b) Some statistical problems in the assessment of inhomogeneities of DNA sequence data. *J. Am. Stat. Assoc.*, **86**, 27–35.

Karlin,S. and Taylor,H.M. (1975) *A First Course in Stochastic Processes*. Academic Press, New York.

Kel,A.E., Kondrakhin,Y.V., Kolpakov,Ph.A., Kel,O.V., Romashenko,A.G., Wingender,E., Milanesi,L. and Kolchanov,N.A. (1995) Computer tool FUNSITE for analysis of eukaryotic regulatory genomic sequences. *Intell. Syst. Mol. Biol.*, **3**, 197–205.

Kendall,M.G. (1952) *The Advanced Theory of Statistics*, Vol. II. Griffin, London.

Kirchhamer,K., Yuh,C.-H. and Davidson,E.H. (1996) Modular *cis*-regulatory organization of developmentally expressed genes: two genes transcribed territorially in the sea urchin embryo and additional examples. *Proc. Natl Acad. Sci. USA*, **93**, 9322–9328.

Kleffe,J. and Langbecker,U. (1990) Exact computation of pattern probabilities in random sequences generated by Markov chains. *Comput. Applic. Biosci.*, **6**, 347–353.

Kolchanov,N.A. *et al.* (1998) GeneExpress: a computer system for description, analysis and recognition of regulatory sequences in eukaryotic genome. *Intell. Syst. Mol. Biol.*, **6**, 95–104.

Kondrakhin,Y.V., Kel,A.E., Kolchanov,N.A., Romashchenko,A.G. and Milanesi,L. (1995) Eukaryotic promoter recognition by binding sites for transcription factors. *Comput. Applic. Biosci.*, **11**, 477–488.

Kuo,M.H. and Grayhack,E. (1994) A library of yeast genomic MCM1 binding sites contains genes involved in cell cycle control, cell wall and membrane structure and metabolism. *Mol. Cell. Biol.*, **14**, 348–359.

Lavorgna,G., Boncinelli,E., Wagner,A. and Werner,T. (1998) Detection of potential target genes *in silico*? *Trends Genet.*, **14**, 375–376.

Li,W.T. (1997) The study of correlation structures of DNA-sequences: a critical review. *Comput. Chem.*, **21**, 257–271.

Lukashin,A.V., Anshelevich,V.V., Amirikyan,B.R,. Gragerov,A.I. and Frank-Kamenetskii,M.D. (1989) Neural network models for promoter recognition. *J. Biomol. Struct. Dyn.*, **6**, 1123–1133.

Maine,G.T., Sinha,P. and Tye,B.K. (1984) Mutants of *S. cerevisiae* defective in the maintenance of minichromosomes. *Genetics*, **106**, 365–385.

Matis,S., Xu,Y., Shah,M., Guan,X., Einstein,J.R., Mural,R. and Uberbacher,E. (1996) Detection of RNA polymerase II promoters and polyadenylation sites in human DNA sequence. *Comput. Chem.*, **20**, 135–140.

Messenguy,F. and Dubois,E. (1993) Genetic evidence for a role for MCM1 in the regulation of arginine metabolism in *Saccharomyces cerevisiae. Mol. Cell. Biol.*, **13**, 2586–2592.

Oehlen,L.J., McKinney,J.D. and Cross,F.R. (1996) Ste12 and Mcm1 regulate cell cycle-dependent transcription of FAR1. *Mol. Cell. Biol.*, **16**, 2830–2837.

Olson,M.V. (1992) Genome structure and organization in *Saccharomyces cerevisiae*. In Jones,E.W., Pringle,J.R. and Broach,J.R. (eds), *The Molecular and Cellular Biology of the Yeast Saccharomyces*. Cold Spring Harbor Laboratory Press, New York, Vol. II.

O'Neill,M.C. (1991) Training back-propagation neural networks to define and detect DNA-binding sites. *Nucleic Acids Res.*, **19**, 313–318.

Parzen,G. (1962) *Stochastic Processes*. Holden-Day, San Francisco, CA.

Prestridge,D.S. (1995) Predicting Pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol.*, **249**, 923–932.

Prestridge,D.S. (1996) SIGNAL SCAN 4.0: additional databases and sequence formats. *Comput. Applic. Biosci.*, **12**, 157–160.

Ptashne,M. (1988) How transcriptional activators work. *Nature*, **335**, 683–689.

Ptashne,M. and Gann,A. (1997) Transcriptional activation by recruitment. *Nature*, **386**, 569–577.

Shelton,D.A., Stegman,L., Hardison,R., Miller,W., Bock,J.H., Slightom,J.L., Goodman,M. and Gumucio,D.L. (1997) Phylogenetic footprinting of hypersensitive site 3 of the beta-globin locus control region. *Blood*, **89**, 3457–3469.

Sokal,R.R. and Rohlf,F.J. (1981) *Biometry*. Freeman, New York.

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

Sprague,G.F.,J. and Thorner,J.W. (1992) Pheromone response and signal transduction during the mating response in *Saccharomyces cerevisiae*. In Jones,E.W., Pringle,J.R. and Broach,J.R. (eds), *The Molecular and Cellular Biology of the Yeast Saccharomyces*. Cold Spring Harbor Laboratory Press, New York, Vol. II.

Stormo,G.D. (1990) Consensus patterns in DNA. *Methods Enzymol.*, **183**, 211–220.

Struhl,K. (1989) Molecular mechanisms of transcriptional regulation in yeast. *Annu. Rev. Biochem.*, **58**, 1051–1077.

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

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

Wagner,A. (1997) A computational genomics approach to the identification of gene networks. *Nucleic Acids Res.*, **25**, 3594–3604.

Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, **278**, 167–181.

Wynne,J. and Treisman,R. (1992) SRF and MCM1 have related but distinct DNA binding specificities. *Nucleic Acids Res.*, **20**, 3297–3303.

Yuan,Y.L. and Fields,S. (1991) Properties of the DNA-binding domain of the *Saccharomyces cerevisiae* STE12 protein. *Mol. Cell. Biol.*, **11**, 5910–5918.