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

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

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

1 Introduction

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

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

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

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

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

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

2 Statistical Methods

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

denotes the distance between site X_j and X_i.

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

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

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

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

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

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

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