BIOINFORMATICS

Vol. 17 no. 10 2001
Pages 878–889

Detection of cis*-element clusters in higher eukaryotic DNA*

Martin C. Frith[1], *Ulla Hansen*[1, 2] *and Zhiping Weng*[1, 3,*]

[1]*Bioinformatics Program, Boston University, 44 Cummington St,* [2]*Department of Biology, Boston University, 5 Cummington St and* [3]*Department of Biomedical Engineering, Boston University, 44 Cummington St, Boston, MA 02215, USA*

Received on April 20, 2001; revised and accepted on July 6, 2001

ABSTRACT

Motivation: Computational prediction and analysis of transcription regulatory regions in DNA sequences has the potential to accelerate greatly our understanding of how cellular processes are controlled. We present a hidden Markov model based method for detecting regulatory regions in DNA sequences, by searching for clusters of *cis*-elements.

Results: When applied to regulatory targets of the transcription factor LSF, this method achieves a sensitivity of 67%, while making one prediction per 33 kb of non-repetitive human genomic sequence. When applied to muscle specific regulatory regions, we obtain a sensitivity and prediction rate that compare favorably with one of the best alternative approaches. Our method, which we call Cister, can be used to predict different varieties of regulatory region by searching for clusters of *cis*-elements of any type chosen by the user. Cister is simple to use and is available on the web.

Availability: http://sullivan.bu.edu/~mfrith/cister.shtml

Contact: mfrith@bu.edu; zhiping@bu.edu

Introduction

The transcription of a eukaryotic gene is controlled by an intricate choreography of proteins binding to promoter, enhancer, and repressor sites on the DNA sequence (Ptashne, 1988; Ptashne and Gann, 1997). Currently we are extremely far from a predictive understanding of these interactions (Fickett and Hatzigeorgiou, 1997; Reese *et al.*, 2000), especially in more complex eukaryotes. However, an ability to predict which Transcription Factors (TFs) participate in regulating which regions of DNA would provide two enormous benefits. First, it would contribute significantly to the unsolved problem of finding and accurately delimiting the genes in genomes of complex eukaryotes such as human. More interestingly, such an ability would help us understand the detailed mechanisms of transcriptional control, telling us when, where, and under what conditions a gene is transcribed, and opening up the possibility of therapeutic intervention in these mechanisms.

The nucleotide binding specificities of many transcription factors have been studied, and there are programs that search for these signature sequence patterns, or motifs, in DNA sequence. Many of these methods, such as Matinspector (Quandt *et al.*, 1995), use Position Specific Scoring matrices (PSSMs) to represent motifs. However, these programs make far too many predictions, presumably false positives, to be used easily to understand transcriptional regulation. There is evidence that, while PSSMs can be used to predict accurately TF binding sites on DNA strands *in vitro*, they simply do not contain all the information for determining their sites of regulation *in vivo* (Tronche *et al.*, 1997). Thus, contextual effects must be important. Two obvious candidates for important contextual effects are the requirement for multiple factors to bind DNA and cooperatively regulate transcription, and the packaging of DNA into chromatin. Our method addresses the first of these effects, by searching for clusters of regulatory motifs in the DNA sequence.

There have been many attempts to predict regulatory regions in DNA sequences, some based, in one form or another, on searching for clusters of transcription factor binding sites. Of course, 'regulatory region' is a broad term encompassing many subcategories. The majority of these efforts have focused on detecting RNA polymerase II (Pol II) promoters, itself a large and apparently heterogeneous category. Some state-of-the-art Pol II promoter prediction tools, such as PromoterInspector (Scherf *et al.*, 2000), do not explicitly consider *cis*-element motifs at all, but perform no worse because of it. This may be because Pol II promoters are too heterogeneous to be detected efficiently using a fixed set of motifs, or because their binding site organization is poorly understood.

Of the previous methods that consider clusters of TF binding site motifs, two early attempts to detect Pol II

*To whom correspondence should be addressed.

promoters are PromoterScan (Prestridge, 1995) and FunSiteP (Kondrakhin *et al.*, 1995). These programs both use consensus sequences rather than PSSMs to detect motifs, so that motifs are treated in a binary fashion (present or absent), their relative strengths not being considered. The Logistic Regression Analysis (LRA) method of Wasserman and Fickett (Wasserman and Fickett, 1998) uses PSSMs to measure motif strengths, but it shares with PromoterScan and FunSiteP the necessity to introduce an *ad-hoc* window size. Motifs are considered together if they lie within a sequence segment, or window, of a certain length, and distances between motifs are not considered otherwise. A method by Wagner (1999) treats the distances between motifs in a more comprehensive way, using a Poisson model, but does not consider motif strengths. ModelInspector (Frech *et al.*, 1997) can consider diverse types of regulatory element, but they must occur in a specific order, and the distances between them are either specified to lie in a simple range (Klingenhoff *et al.*, 1999), or are described by a 'distance histogram' with many parameters. The intriguing hidden Markov model (HMM) method of Crowley *et al.* (1997) estimates all of its parameters, including which motif types to consider, from the sequence data. It also treats motifs in a binary fashion, ignoring weak motifs that do not match the consensus.

These methods illustrate a variety of advantages and disadvantages relative to one another. We set out to create a method that combines their best features, while avoiding their disadvantages as far as possible. Specifically, our requirements were that the method should consider both the strengths of motif hits and the distances between them, should avoid *ad-hoc* window sizes or motif score cutoffs, and should have a small number of parameters to avoid the danger of over-fitting. We also emphasized ease of use. The HMM algorithm that we present fulfills all of these requirements.

We applied our method to search for regulatory sites of LSF in DNA sequences, and also to detect muscle specific regulatory regions. The transcription factor LSF is a member of a small family of DNA-binding proteins, with strong sequence similarities from human to *Drosophila*. In mammals, LSF is uniformly expressed in all cell types (Swendeman *et al.*, 1994). It binds a repeated DNA sequence as an obligate tetramer (Murata *et al.*, 1998; Shirra and Hansen, 1998). In addition, in a tissue-specific manner, LSF can partner with other types of DNA-binding proteins to bind novel DNA-binding sites (Jane *et al.*, 1995; Romerio *et al.*, 1997; Murata *et al.*, 1998; Casolaro *et al.*, 2000).

Vertebrate LSF (also known as CP2 or LBP-1c) (Lim *et al.*, 1992; Shirra *et al.*, 1994; Yoon *et al.*, 1994) binds and regulates a variety of cellular and viral promoters, either as an activator or repressor of gene expression (Huang *et al.*, 1990; Kato *et al.*, 1991; Lim *et al.*, 1992; Sundseth and Hansen, 1992; Lim *et al.*, 1993; Parada *et al.*, 1995; Romerio *et al.*, 1997; Murata *et al.*, 1998; Casolaro *et al.*, 2000; Powell *et al.*, 2000). These include the thymidylate synthase genes in human and mouse fibroblasts (Powell *et al.*, 2000), the murine alpha-globin gene in erythroid cells (Lim *et al.*, 1992, 1993), the IL-4 gene in human T-cells (Casolaro *et al.*, 2000), the chicken alpha-crystallin gene (Murata et al., 1998), the SV40 major late promoter (Kim *et al.*, 1987; Huang *et al.*, 1990), and the HIV long terminal repeat (Yoon *et al.*, 1994; Romerio *et al.*, 1997).

We suspected that the transcription factors Sp1 and Ets-1 may be involved in co-regulating transcription with LSF, since binding sites for these factors have been found in close proximity to LSF binding sites (Kim *et al.*, 1987; Dong *et al.*, 2000). The majority of experimentally determined LSF binding sites lie within 300 bases of the Transcription Start Site (TSS), although some occur at more distant locations (see Section **Methods**). We decided to focus on predicting LSF regulatory regions close to TSSs, for fear that *cis*-element clusters further from the promoter may have distinct properties. Therefore, the TATA box motif is likely to be clustered with these other motifs too. So we searched for regions regulated by LSF by considering clusters of the four *cis*-element types: LSF, Sp1, Ets-1, and the TATA box. LSF binding sites occur in both orientations relative to the direction of transcription (Kim *et al.*, 1990; Powell *et al.*, 2000), as do binding sites for Sp1 (Jones and Tjian, 1985; Ishii *et al.*, 1986).

We also applied our approach to the muscle specific regulatory regions that were studied using the LRA method of Wasserman and Fickett, in order to facilitate comparison of the two methods, and gain more data on our method's performance. Muscle specific regulatory regions were studied with the LRA approach using PSSMs representing five types of binding site: those for Mef-2, Myf, SRF, Tef, and Sp1 (Wasserman and Fickett, 1998). In that study, two sets of PSSMs were used for these five *cis*-elements: one set derived from experimentally determined binding sites required for gene expression in skeletal muscle, and the other derived from muscle independent sources, including *in vitro* binding studies (Mef-2, Myf and SRF), and sites from genes not specifically expressed in muscle (Sp1 and Tef). The muscle derived PSSMs provide more accurate representations of the binding site motifs, but since they are constructed from sites in the sequences that the method will be tested on, the non-muscle derived PSSMs allow a more strictly non-circular test. We also used both of these sets of binding site matrices to test our method.

METHODS

Data for *cis*-element nucleotide frequencies and test sequences

We obtained 19 LSF binding sites supported by experimental evidence, associated with ten genes (Table 1). Fifteen of these sites, on nine sequences, were within 300 bp of a TSS, with the remaining sites in introns or UTRs. The nine sequences containing LSF sites near the promoter were taken as a positive test set. Eleven Ets-1 sites were obtained from the ooTFD database (Ghosh, 2000), excluding artificial and exact duplicate sequences. The 19 LSF motifs were aligned manually, as were the 11 Ets-1 binding sites. The number of each of the four types of nucleotide A, C, G and T at each position of the alignments was counted, producing matrices that we refer to as *nucleotide count matrices* (Table 2). Bucher's nucleotide count matrix (Bucher, 1990) derived from 389 sites was used for the TATA motif, and a matrix for Sp1 based on 108 sites was obtained from the TRANSFAC database (release 5.0, accession M00196).

Nucleotide count matrices representing the five muscle related *cis*-elements, and 43 sequences containing experimentally determined muscle specific regulatory regions, were obtained from the website http://www.cbil.upenn.edu/MTIR/DATATOC.html (Wasserman and Fickett, 1998). Since the muscle derived matrices were constructed from orthologous pairs of sequences, we divided all the counts by two. We discovered that the set of 43 regulatory regions includes many orthologs that share high levels of sequence identity. Therefore they do not constitute a sample of independent sequences, and the results of applying our method to one sequence will be highly determined by the results obtained for its ortholog. To avert this problem, we constructed a non-redundant subset of 27 of the sequences, and this subset was used to test our method. These sequences comprise 12 from mouse, 2 from rat, 11 from human, and 2 from chicken.

Data for human genome analysis

We obtained the sequences and annotation from the 7 October 2000 freeze of the draft human genome at the University of California at Santa Cruz (http://genome.ucsc.edu/) (Lander *et al.*, 2001). RepeatMasker results were taken from the tables 'chrN_rmask', gene predictions from 'genieAlt', known genes from 'genieKnown', and the GenBank accessions for their protein products were taken from 'knownInfo'. The gene predictions were derived using the program Genie (Kulp *et al.*, 1996), constrained by EST and protein homology data. The known genes are a subset of the predicted genes with further annotation and links to GenBank. Nearly all of them are derived from the RefSeq portion of GenBank. The descriptions of the known genes' protein products were extracted from their GenPept records. The genome sequence comprises 2.69 billion sequenced bases, of which 1.4 billion are nonrepetitive according to RepeatMasker.

Eukaryotic promoter sequences

The entire set of 1391 eukaryotic promoter sequences was obtained from the Eukaryotic Promoter Database, release 66_1 (Perier *et al.*, 2000). The majority of these promoters (943) are vertebrate, and 613 are mammalian. For prediction of muscle specific regulatory regions, sequences of length 300 bases, from -249 to $+50$ relative to the TSS, were used. For analysis using the LSF associated motifs, 2000 bp long sequences from -1499 to $+500$ were used.

HIDDEN MARKOV MODEL

We use a hidden Markov model (HMM), represented in Figure 1, to detect *cis*-element clusters in a DNA sequence (the query sequence). The HMM architecture and transition probabilities represent our prior expectations concerning the distribution of *cis*-elements. The emission probabilities describe the probability of observing the sequence data given a particular distribution of *cis*-elements. Posterior decoding is used to estimate the locations of *cis*-element clusters given the sequence data.

HMM architecture and transition probabilities

The motivation underlying this model is that we expect to see occasional clusters of *cis*-elements within a large ocean of background sequence. The inter-cluster background state represents this background 'ocean'. The model assumes that the distance between clusters is geometrically distributed with mean g.

In Figure 1, the circles with the label 'm' correspond to positions in the *cis*-elements, and we refer to them as 'motif states'. For purposes of illustration, they are shown in the figure as a grid with two rows, each row having three states. In actual fact the number of rows is equal to twice the number of *cis*-element types selected by the user, one for each element on each DNA strand. The number of states in each row is the number of bases in the *cis*-element motif. The transition probabilities to the start of each row are all set equal to one other: within a cluster the model expects to see any *cis*-element on either strand with equal probability. This equal occurrence model is partly justified by noting that many *cis*-elements, such as Sp1 and LSF, are known to occur on both strands relative to the transcribed strand. In addition, there is insufficient experimental data to estimate reliably the relative frequencies of different *cis*-element types, and hence the simplest prior assumption is that their frequencies are equal. There are some cases, such as the orientation of the TATA box, where a more complex

Table 1. Experimentally determined LSF binding sites used to estimate position-specific nucleotide preferences

Sequence	GenBank accession	Length/bases	Number of LSF sites	Reference
SV40	NC_001669	5 243	1	Huang *et al.* (1990)
HIV type 1	NC_001802	9 181	1	Yoon *et al.* (1994)
Human fos	K00650	6 210	1	Volker *et al.* (1997)
Human ornithine decarboxylase	M81740	9 373	1	Volker *et al.* (1997)
Human beta-polymerase	J04201	1 866	1	Weis and Reinberg (1992)
Mouse thymidylate synthase, 5′ region	J02617	730	2	Powell *et al.* (2000)
Mouse alpha-globin	V00714	1 441	6	Kim *et al.* (1990)
Rat fibrinogen gamma-chain	X05860	3 299	1	Lim *et al.* (1993)
Chicken αA-crystallin	M17627	6 311	1	Murata *et al.* (1998)
Not in test set				
Mouse thymidylate synthase, 3′ region	–	–	3	Powell *et al.* (2000)
Mouse Ea	–	–	1	Bellorini *et al.* (1996)
Totals		43 654	19	

model might be more appropriate. Since we do not know *a priori* which strand is transcribed, the TATA box could be accommodated by having two alternative sets of motif and intra-cluster background states, with the TATA box orientation fixed in the opposite sense in each. Further details could then be added, such as an initiator motif downstream from the TATA box. However, our aim was not detailed prediction of basal Pol II promoters, and so we kept to the simpler and more general model of figure 1.

The *cis*-elements in a cluster are not required to be immediately adjacent to one another, and so the intra-cluster background state represents stretches of background sequence that may appear between motifs within a cluster. The distance between motifs in a cluster is modeled as a geometric distribution with mean a, and the number of *cis*-elements in a cluster is supposed geometrically distributed with mean b. The following equations relate the means a, b and g to the HMM transition probabilities α, β and γ:

$$\alpha = \frac{1}{a+1}$$
$$\beta = \frac{1}{b} \qquad (1)$$
$$\gamma = \frac{1}{g+1}.$$

Motif state emission probabilities

The emission probabilities of the motif states correspond to the nucleotide preferences of each position in the *cis*-element. The program takes nucleotide count matrices, such as those in Table 2, as input. It then follows Laplace's rule of succession and adds 1 to each of the counts, and these are normalized to obtain the emission probabilities for the motif states.

Table 2. Position specific nucleotide frequency matrices for binding sites of LSF and Ets-1

Position	LSF				Ets-1			
	A	C	G	T	A	C	G	T
1	5	0	11	3	6	0	5	0
2	0	17	2	0	0	8	3	0
3	2	0	0	17	7	4	0	0
4	0	0	17	2	0	0	11	0
5	0	1	18	0	0	0	11	0
6	2	3	6	8	11	0	0	0
7	1	2	1	15	8	0	0	3
8	4	6	2	7	0	0	11	0
9	2	3	10	4	0	4	0	7
10	6	6	4	3	3	1	7	0
11	5	1	9	4				
12	0	17	2	0				
13	1	6	2	10				
14	5	1	8	5				
15	0	5	14	0				

Background emission probabilities

Since DNA sequences can differ significantly in abundance of A, C, G and T from location to location, it is not obvious how to obtain suitable emission frequencies for the background states in our model. Options that we considered include: set them all to 0.25, set them to the genomic average, set them to the average for the entire query sequence, and obtain them from a local window around the position currently being scanned in the query sequence. We found that the last option gave us the best results. With the second-to-last option the results could change significantly just by including more flanking sequence at the edges of the query.

As the query sequence is scanned base by base, the

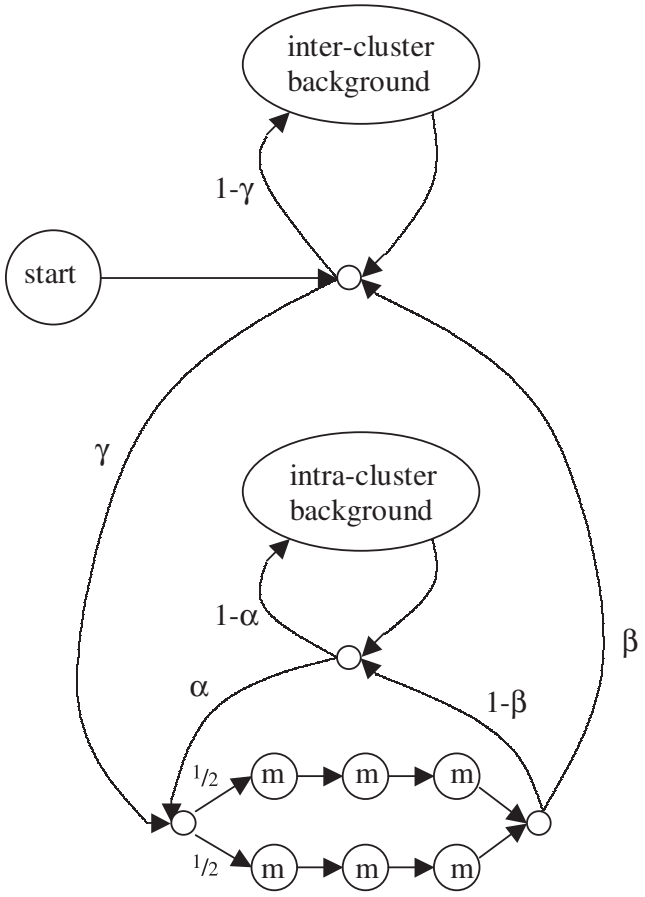

Fig. 1. The HMM used by Cister. The circles and ovals represent states, and the arrows represent possible transitions between the states. The labels by the arrows indicate transition probabilities. The small, unlabeled circles are silent states, as defined in Durbin *et al.* (1998).

nucleotide frequencies are counted in a window of length $2W + 1$ centered on the current base, where W is a user-defined parameter. The counts for each nucleotide are incremented by 1 (Laplace's rule), and the background emission probabilities for the current base are obtained from these counts.

Multipartite *cis*-elements

Positions in a *cis*-element can be specified as 'background/blank', in which case they are assigned the same emission probabilities as the background states. This option allows multipartite elements to be specified. If bases flanking the *cis*-elements are specified as 'background/blank', there will be a motif-dependent minimum spacing between *cis*-elements on any one path through the HMM. This restriction models the occlusion of flanking bases by an overhanging DNA binding factor.

Treatment of masked bases

The program RepeatMasker was used to mask repetitive regions of DNA by replacing them with strings of the letter 'N'. This procedure raises the question of how Cister should treat these regions. We addressed this problem by modifying the HMM emission probabilities, so that all states have an equal probability, p_N, of emitting an 'N', and the emission probabilities for the other four letters were multiplied by $1 - p_N$. There is no need to estimate the parameter p_N since it cancels entirely from the posterior probability calculations (equation (2) in the next section). An alternative solution, giving the motif states zero probability of emitting 'N', appeared to perform worse (data not shown).

ALGORITHM

The query sequence is analyzed using our HMM with posterior decoding (Durbin *et al.*, 1998). The probability (Pr) that any path through the HMM is taken and emits the observed sequence can be calculated by multiplying the transition and emission probabilities. All paths are considered, and a posterior probability of each base being in any state is assigned according to the following equation:

$$\Pr(\text{base } i \text{ is in state } s) = \frac{\sum_{\text{paths } p \text{ with base } i \text{ in state } s} \Pr(p)}{\sum_{\text{all paths } p} \Pr(p)}. \quad (2)$$

This calculation can be performed efficiently using the forward and backward dynamic programming algorithms (Durbin *et al.*, 1998).

The technique of posterior decoding reflects our intuition of how transcription factors interact with DNA. A single path through the HMM represents one possible arrangement of transcription factors binding to the sequence at specific locations. If there are two overlapping motifs in the sequence, they cannot both be selected by the same path, just as overlapping *cis*-elements cannot be bound by their factors simultaneously. However, alternative paths can pass through both motifs, and the probabilities of all possible paths are summed to give the final posterior probability. Therefore, our algorithm calculates the overall probability that a segment of DNA interacts with competing transcription factors.

IMPLEMENTATION

This algorithm was implemented as a C program called 'Cister' (for *cis*-element clusTER finder). The time and memory requirements are linearly dependent upon the length of the query sequence. The memory requirement can of course be reduced by splitting a long sequence into shorter segments. Analysis of the draft human genome,

1.4 billion bases after masking repetitive regions, required 22 h on a single processor of a DEC alpha ES20.

Web interface

We attached great importance to making this method easy to use. We therefore constructed a website with a straightforward form to enter queries, and an intuitive graphical display of the results. Three pieces of information are requested on the input form: a query sequence, a choice of *cis*-element types to search for, and algorithmic parameters. The query sequence can be specified with a GenBank accession number, or it can be pasted directly into the form. If GenBank format is used, annotated coding regions ('CDS') will be displayed along with the *cis*-element results. *Cis*-elements can be chosen from a set of options, or specified by the user as nucleotide count matrices. These tables are unavoidably nontrivial to construct, since they require alignment of a set of trustworthy examples of the *cis*-element type. In order to minimize this difficulty, we provide a link to the 356 matrices in the TRANSFAC database (Wingender *et al.*, 2000), and the matrix information can be directly cut-and-pasted from there into the Cister form. Sensible defaults are provided for the algorithmic parameters a, b, g and W.

Cister displays an intuitive graphical output, illustrated in Figure 2a. The black curve indicates $1-\Pr$ (inter-cluster background). It represents the posterior probability that each base is within a cluster of *cis*-elements. The colored lines indicate Pr (position 1 of a *cis*-element). Motifs on the + strand are displayed in the upper half of the plot, and motifs on the − strand are displayed in the lower half. Figure 2 shows the output for the genome of SV40, correctly identifying the regulatory region and the LSF binding site.

RESULTS

Anecdotal example

We were interested to see whether Cister could be used to find sites of transcriptional regulation by LSF. By examining the graphical output produced by Cister, it is clear that the program can make specific and useful predictions. For example, Figure 2a shows the output for the genome of the SV40 virus. The regulatory portion at the beginning of the sequence is correctly identified, and there is an impressive lack of false positive predictions anywhere else in the sequence. For comparison, we also searched this sequence for matches to individual PSSMs for each of the four *cis*-elements. For these PSSM searches, the background nucleotide frequencies were counted using a sliding window of width 2001, in the same manner as Cister. Figure 2b shows that strong PSSM matches can be found at many locations in the sequence. So Cister reveals the regulatory region much more clearly than the simple PSSM search.

Systematic evaluation of performance

It is desirable to have a more quantitative measure of the method's ability to detect regulatory regions. Performance is traditionally measured using sensitivity: the percentage of true cases detected, and specificity: the percentage of false cases correctly rejected. Of course there is a tradeoff between these measures. The sensitivity can be made arbitrarily close to 100% by altering parameters, at the cost of lowering the specificity. Our parameter choices gave sensitivities of around 60%, and the method's performance is judged by its specificity (or prediction rate) at this sensitivity level. We measured sensitivity by examining the percentages of the nine LSF regulatory regions and the 27 muscle specific sequences that were identified by Cister. It is harder to get a handle on the method's specificity, since we do not have a set of sequences that are known not to contain regulatory regions associated with LSF or muscle specific expression. In this situation it is common to measure a method's background prediction rate on a large sequence set, and argue that a lower prediction rate is more useful, because it implies a smaller number of hypotheses that need to be tested experimentally. This sequence set should be chosen to match the kinds of sequence that the method will actually be applied to. We take this approach, measuring Cister's prediction rate for the draft human genome, and for 1391 eukaryotic promoters from the Eukaryotic Promoter Database (EPD). We argue in the discussion that the method is more appropriately applied to genomic sequence, but the performance on EPD sequences allows comparison with the LRA method.

Procedure for defining and evaluating predictions from cister output

Cister is not appropriate for detecting individual *cis*-elements, since it tends to over-predict motifs within a strong cluster. Moreover, the data on individual *cis*-elements within studied regulatory regions is incomplete, and we were interested in Cister's ability to find the overall cluster. We faced the difficulty that Cister's output consists of posterior probability curves, which are very informative to the human eye, but difficult to assess rigorously in terms of true positives and false positives. It was necessary to define an *ad-hoc* way of extracting discrete 'predictions' of regulatory regions from the Cister output, and then of declaring them 'true' or 'false'.

Predictions of regulatory regions were defined to occur at local maxima of the cluster posterior probability curve, where its value is greater than that at all locations up to 1200 bases distant, and greater than 0.5. For LSF regulatory regions, we additionally required an LSF *cis*-

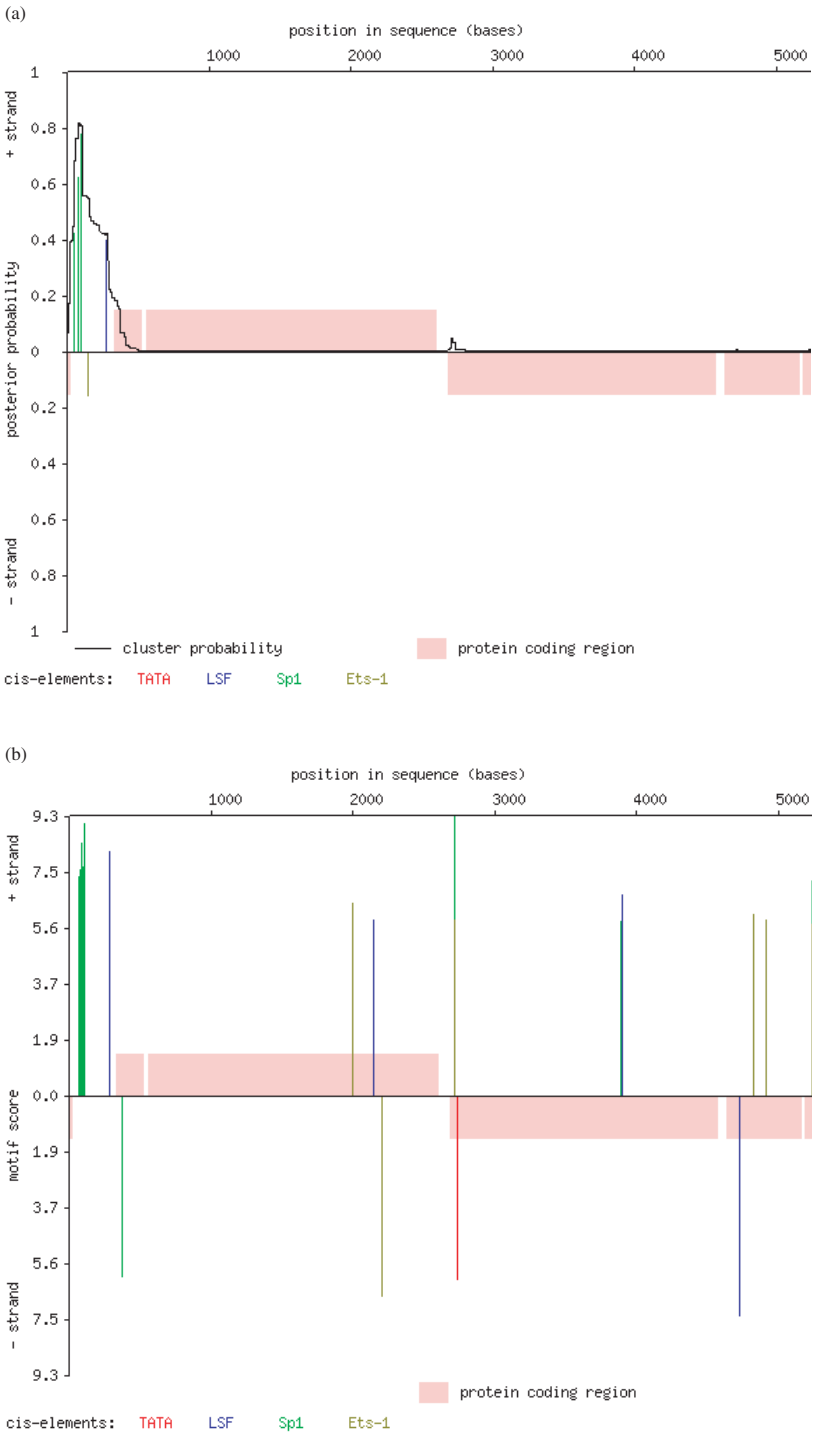

Fig. 2. The results of analyzing the SV40 genome sequence with Cister (a), and using individual PSSM searches for the four *cis*-elements (b); Cister parameter settings: $a = 35$, $b = 6$, $g = 20\,000$, $W = 1000$.

element prediction with posterior probability greater than 0.1 to occur within 1200 bases of the local maximum. For tests of sensitivity using the known LSF regulatory regions, a prediction was declared a true positive if a known LSF site occurred within 1200 bases from it. All other predictions were false positives. No more than one

true positive was counted per sequence. The 27 muscle specific regulatory regions are all shorter than 1200 bases and so false positives are impossible for this test set. The number 1200 was chosen because some regulatory regions are known to extend over several kilobases, such as that of the sea urchin gene Endo16, which covers about 2300 bases upstream of the gene (Arnone and Davidson, 1997).

Parameter choices

Cister has four tunable parameters: a, b, g, and W. Values for a, b and g should be chosen according to what typical *cis*-element clusters are like, i.e. how many nucleotides they span on average, how many *cis*-elements they contain, and the typical distance between clusters. Unfortunately, the experimental data on known *cis*-element clusters is too incomplete to estimate these parameters, reducing us to intelligent guesswork. For LSF associated clusters, we chose $a = 35$, $b = 6$, $g = 50\,000$, and $W = 1000$. The values for a and b are sensible choices and seem to give good results. We reasoned that the number of LSF regulatory regions in the genome is unlikely to be greater than the number of genes, hence the value for g has the same order of magnitude as the distance between human genes.

We discovered that the values of g and W exert a strong influence on the prediction rate and sensitivity. The prediction rate and sensitivity both rise if g decreases or if W increases. For g this effect is not surprising, since high values of g make *cis*-element clusters less probable. It appears that when W increases, the HMM background states capture local variations in nucleotide frequencies less precisely, so that some regions become less well modeled by the background states compared to motif states.

Since most of the 27 muscle regulatory regions are only 300 bases long, W is effectively constrained to values below 150, and lower values of g were used to maintain sensitivity. Alternatively, the posterior probability threshold of 0.5 could have been lowered for the muscle regions, but we did not do this. Because many of the muscle regulatory regions have been localized to within 200 bp, and since they are detected using five rather than four motifs, we chose values of a and b that reflect a higher concentration of *cis*-elements. For the muscle derived set of motifs, we used $a = 10$, $b = 10$, $g = 20\,000$, and $W = 150$. For the nonmuscle derived motifs, we kept the same values for a, b and W, but it was necessary to reduce g to 1000 to maintain the same sensitivity.

Measurements of sensitivity

The nine LSF test sequences and the 27 muscle sequences were first processed with RepeatMasker in order to remove repetitive sequences (Smit and Green, personal communication), using the 'slow' speed/sensitivity setting, and the 'DNA source' setting appropriate to each sequence. This procedure was used for consistency with the human genome analysis described below.

A jackknife procedure was applied to the LSF sequences. For each sequence, the LSF sites present in its regulatory region were excluded from the LSF nucleotide count matrix given to Cister. We confirmed that none of the sites used to construct the Ets-1 matrix were present in these regulatory regions. Although we could not make the same confirmation for the Sp1 and TATA box matrices, these matrices are constructed from so many sites that the presence or absence of one or two would make little difference to the nucleotide emission probabilities.

For the nine LSF sequences, Cister produced six true positives, finding six out of the nine regulatory regions, and zero false positives. For our parameter choices, it failed to make any predictions for chicken αA-crystallin, mouse alpha globin, and mouse thymidylate synthase. To pick up these sequences while retaining a reasonable specificity, it may be necessary to consider additional types of *cis*-element. For example, the LSF site in alpha globin is clustered with sites for NF-E1, NF-5, IRP, and CP1 (Kim *et al.*, 1990), and the αA-crystallin promoter contains binding sites for AP-1, Pax6, USF, and CREB/CREM (Ilagan *et al.*, 1999). It is quite possible that LSF *cis*-elements appear in more than one type of regulatory region, and that different *cis*-element combinations and parameters are appropriate to these different types. In addition, while W was set to 1000, so that the base frequencies were counted in windows of size 2001, the thymidylate synthase and alpha globin sequences are considerably shorter than 2001 bases. So these two sequences are penalized by having a lower effective value for W, and it is quite possible that Cister would detect these regulatory regions if longer sequences containing them were available.

Using the muscle derived nucleotide count matrices, Cister made predictions for 16 out of the 27 muscle specific regulatory regions. With the nonmuscle derived matrices, we obtained the same figure of 16 out of 27, or 59%. These sensitivities are comparable to those reported for the LRA method (see Tables 3a and b).

Human genome analysis

In tests on a portion of the draft human genome sequence, we found that masking repetitive elements with RepeatMasker reduced the number of Cister predictions by about a factor of 6, even though only 47% of the sequence was masked. Cister predictions were over-represented in all three subtypes of repetitive sequence examined: Alu elements, other interspersed repeats, and especially low complexity/simple repeats (data not shown). We therefore decided to mask all repetitive elements before proceeding with our analysis.

Table 3a. Performance of Cister

Motif set	Parameter settings	Sensitivity	Prediction rate for human genomic sequence	Prediction rate for EPD sequences
LSF	$a = 35$ $b = 6$ $g = 50\,000$ $W = 1000$	6/9 (66%)	1 per 63 kb	15%
Muscle derived	$a = 10$ $b = 10$ $g = 20\,000$ $W = 150$	16/27 (59%)	1 per 68 kb	2.9%
Nonmuscle derived	$a = 10$ $b = 10$ $g = 1000$ $W = 150$	16/27 (59%)	1 per 32 kb	5.2%

Table 3b. Performance of the LRA method. Data from Wasserman and Fickett (1998)

PSSM set	Sensitivity	Prediction rate for human genomic sequences	Prediction rate for EPD sequence
Muscle derived	60%	1 per 32 kb	4%
Nonmuscle derived	60%	Not reported	13%

The masked human genome sequence was analyzed with Cister, using the four LSF associated nucleotide count matrices, the five muscle derived matrices, and the five nonmuscle derived matrices. In each case, the same parameter settings were used as for the tests of sensitivity. With the LSF associated motifs, Cister made a total of 42 995 predictions: a rate of one prediction per 33 kb of nonrepetitive sequence, or one per 63 kb of total sequence. This figure gives an upper limit to our method's false positive rate. For the muscle derived and nonmuscle derived matrices, Cister made 39 770 and 84 185 predictions, respectively. The prediction rate for the muscle derived matrices is thus one per 35 kb of non repetitive sequence, or one per 68 kb of total sequence. For the nonmuscle derived matrices these figures are one per 17 kb of non-repetitive sequence, or one per 32 kb of total sequence (Table 3a). By way of comparison, the reported prediction rate for the LRA method using the muscle-derived PSSMs is one per 32 kb of human genomic sequence, after masking repetitive elements (Table 3b; Wasserman and Fickett, 1998). It is not quite clear whether this rate was measured per unmasked sequence or per total sequence.

We investigated the predicted LSF regulatory regions in the human genome in more detail. It seems unlikely that there are really 43 000 sites of regulation by LSF, since this figure is greater than recent predictions of the total number of human genes (Lander *et al.*, 2001; Venter *et al.*, 2001). Reasoning that true LSF regulatory regions are likely to be over-represented near gene start locations, we investigated whether our predictions show a bias to the vicinity of TSSs. Of the 42 995 predictions, 3028 are less than or equal to 1200 bases distant from a predicted TSS. We estimated the probability of obtaining this figure by chance alone. There are 21 968 predicted TSSs in total. We counted the number of unmasked bases within a distance of 1200 from a TSS, which comes to about 2.4% of the total 1.4 billion nucleotides. If 42 995 sites are chosen at random, the mean number falling within 1200 bases of a TSS is 1053, with a standard deviation of 32. The probability that 3028 or more sites lie this close to a TSS is less than 10^{-12}. So the over-representation of Cister predictions near to TSSs is highly significant, suggesting that the program has some success in detecting transcription regulatory signals. Moreover, the TSS locations are based on tentative gene structure predictions, which are especially unreliable at the 5′ ends, owing to low EST coverage, and the difficulty of predicting the extent of the 5′ untranslated region *ab initio*. Unfortunately, it is hard to judge what fraction of these predictions are sites of LSF regulation rather than general Pol II promoter sites, since we include the ubiquitous TATA box and Sp1 motifs in our search.

Descriptions from GenPept (Benson *et al.*, 2000) records were available for 1889 of the genes whose annotated TSSs are within 1200 bases of a predicted LSF regulatory region. Ornithine decarboxylase does not appear in this list, owing to mis-assembly of the draft genome sequence in the region of the ODC promoter. We wanted to examine whether this list of genes shows any bias towards particular functional categories. From the LocusLink database (Pruitt and Maglott, 2001), we obtained assignments of 1172 of these genes to functional categories according to the Gene Ontology scheme (Ashburner *et al.*, 2000). Unfortunately, after accounting for multiple testing, we found no significant bias among our genes towards any one GO category.

Prediction rates for eukaryotic promoters

To make further performance comparisons with the LRA method, Cister was applied to the 1391 promoter sequences from the EPD. The three sets of matrices were used (LSF associated, muscle derived and nonmuscle derived), with the same parameter settings as for the previous two tests. For analysis with the muscle derived and nonmuscle derived matrices, we used EPD sequences of length 300 bp, since most of the 27 muscle sequences are

300 bp long, and the parameter W is 150 in these cases. For the LSF associated matrices, on the other hand, we used EPD sequences of length 2000 bp. The parameter W is 1000 in this case, and so shorter EPD sequences would not provide a fair test, owing to the strong dependence of Cister's prediction rate on the size of the window for counting background nucleotide frequencies.

The results are summarized in Table 3a. With the muscle derived matrices, Cister made positive predictions for 40 promoters, or 2.9%, compared to 4% reported for the LRA method. Using the nonmuscle derived matrices, we obtained 73 predictions, or 5.2%, representing a dramatic improvement over 13% for the LRA method. The LSF matrices, on the other hand, gave positive predictions for 213 promoters, or 15%. This much higher prediction rate is not surprising given that TATA box and Sp1 motifs are included among the LSF matrices.

DISCUSSION

The results summarized in Table 3 demonstrate that Cister performs at least as well as, and arguably better than the LRA method, a well established technique for detecting clusters of *cis*-elements. We believe that our algorithm has two main advantages over the LRA approach: it is not constrained by an *ad-hoc* window size of 200 bp for the regulatory region, and it is able to integrate all possible motif matches over a region of sequence. A possible advantage of the LRA method is that its parameters are trained from positive and negative training sets. It would be possible to train our HMM transition probabilities from a positive training set using the Baum–Welch algorithm (Durbin *et al.*, 1998).

Compared to previous methods that search for clusters of *cis*-elements, our method presents an important alternative. We do not preprocess the query sequence using arbitrary cutoffs or segmentation. Our method is able to consider very weak motifs if they are clustered tightly enough, and widely separated motifs if they are strong enough, having no cutoff at either of these extremes. Our approach is the only one we know of with an explicit theory to deal with overlapping motifs, and to estimate the overall probability of TFs occupying alternative sets of binding sites (Fickett and Wasserman, 2000). It also allows for varying nucleotide composition along a sequence.

Further advantages of Cister are that it is freely available on the web, it has a user-friendly input form and graphical output, and it is fast enough to analyze large genomes in a reasonable amount of time (a day for the draft human genome on a single DEC alpha processor).

Our algorithm demonstrates a reasonably low prediction rate for muscle specific regulatory regions on both human genomic sequence and promoter sequences. Therefore it is a practical tool for predicting these regulatory regions, thereby generating hypotheses to guide experiments, on sequence sets up to perhaps a few megabases in length. However, it is not yet accurate enough for straightforward use on a whole genome scale. Cister demonstrates a similar performance for discriminating LSF regulatory regions against a background of genomic sequence. Our use of the TATA box and Sp1 motifs makes the method less appropriate for discrimination of LSF regulatory regions from a set of promoter sequences. Realistically, promoters are hard to identify and the locations of most human promoters are unknown, and so the real challenge is to find regulatory regions in genomic sequence, rather than to discriminate them from a set of promoters.

We did not achieve as high a sensitivity and specificity as we would like. There are three likely reasons. First, the regulatory regions that we studied (LSF associated and muscle specific) might actually consist of several subtypes with distinct *cis*-element types and spacings. Second, related to the first reason, our knowledge of which other transcription factors co-regulate transcription with LSF and with the muscle related factors may be insufficient. To address these problems, Cister can be used to explore properties of regulatory regions, by trying several different *cis*-element combinations, and settings for the model parameters a, b, and g. Such investigation may lead to hypotheses about which other factors are involved in transcriptional regulation by LSF. Finally, chromatin structural properties may distinguish true regulatory regions from false positives.

Possible future directions

The HMM approach allows endless flexibility by varying the model architecture and the transition probabilities. For example, the *cis*-elements could be constrained or made more likely to appear in a certain order.

Phylogenetic footprinting could greatly magnify the power of this method, by allowing us to consider sequence conservation between species separated by a suitable evolutionary distance, such as human and mouse. A recent study successfully combined phylogenetic footprinting with Gibbs sampling to detect regulatory motifs (Wasserman *et al.*, 2000). Our HMM could easily be adapted to analyze a pair-wise sequence alignment, with the states emitting pairs of bases instead of single bases. The motif states would then be given high probabilities of emitting conserved bases.

ACKNOWLEDGEMENTS

We are especially grateful to Temple Smith for suggesting use of the forward algorithm to account for overlapping *cis*-elements, and to Weiwei Tong for supplying an algorithm to find the peaks in the cluster probability curve. We thank the research groups of David Landsman and John Spouge for their helpful suggestions and feedback, and the NCBI for their hospitality and the use of their

resources. Martin Frith is a Howard Hughes Medical Institute Predoctoral Fellow. Martin Frith and Zhiping Weng are partially supported by NSF grant DBI#0078194. Ulla Hansen is funded by NIH grant #CA81157.

REFERENCES

Arnone,M.I. and Davidson,E.H. (1997) The hardwiring of development: organization and function of genomic regulatory systems. *Development*, **124**, 1851–1864.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nature Genet.*, **25**, 25–29.

Bellorini,M., Dantonel,J.C., Yoon,J.B., Roeder,R.G., Tora,L. and Mantovani,R. (1996) The major histocompatibility complex class II Ea promoter requires TFIID binding to an initiator sequence. *Mol. Cell. Biol.*, **16**, 503–512.

Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2000) GenBank. *Nucleic Acids Res.*, **28**, 15–18.

Bucher,P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.*, **212**, 563–578.

Casolaro,V., Keane-Myers,A.M., Swendeman,S.L., Steindler,C., Zhong,F., Sheffery,M., Georas,S.N. and Ono,S.J. (2000) Identification and characterization of a critical CP2-binding element in the human interleukin-4 promoter. *J. Biol. Chem.*, **275**, 36 605–36 611.

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.

Dong,S., Lester,L. and Johnson,L.F. (2000) Transcriptional control elements and complex initiation pattern of the TATA-less bidirectional human thymidylate synthase promoter. *J. Cell. Biochem.*, **77**, 50–64.

Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.

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

Fickett,J.W. and Wasserman,W.W. (2000) Discovery and modeling of transcriptional regulatory regions. *Curr. Opin. Biotechnol.*, **11**, 19–24.

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.

Ghosh,D. (2000) Object-oriented transcription factors database (ooTFD). *Nucleic Acids Res.*, **28**, 308–310.

Huang,H.C., Sundseth,R. and Hansen,U. (1990) Transcription factor LSF binds two variant bipartite sites within the SV40 late promoter. *Genes Dev.*, **4**, 287–298.

Ilagan,J.G., Cvekl,A., Kantorow,M., Piatigorsky,J. and Sax,C.M. (1999) Regulation of alphaA-crystallin gene expression. Lens specificity achieved through the differential placement of similar transcriptional control elements in mouse and chicken. *J. Biol. Chem.*, **274**, 19 973–19 978.

Ishii,S., Kadonaga,J.T., Tjian,R., Brady,J.N., Merlino,G.T. and Pastan,I. (1986) Binding of the Sp1 transcription factor by the human Harvey ras1 proto-oncogene promoter. *Science*, **232**, 1410–1413.

Jane,S.M., Nienhuis,A.W. and Cunningham,J.M. (1995) Hemoglobin switching in man and chicken is mediated by a heteromeric complex between the ubiquitous transcription factor CP2 and a developmentally specific protein. *EMBO J.*, **14**, 97–105.

Jones,K.A. and Tjian,R. (1985) Sp1 binds to promoter sequences and activates herpes simplex virus 'immediate-early' gene transcription *in vitro*. *Nature*, **317**, 179–182.

Kato,H., Horikoshi,M. and Roeder,R.G. (1991) Repression of HIV-1 transcription by a cellular protein. *Science*, **251**, 1476–1479.

Kim,C.G., Swendeman,S.L., Barnhart,K.M. and Sheffery,M. (1990) Promoter elements and erythroid cell nuclear factors that regulate alpha-globin gene transcription *in vitro*. *Mol. Cell. Biol.*, **10**, 5958–5966.

Kim,C.H., Heath,C., Bertuch,A. and Hansen,U. (1987) Specific stimulation of simian virus 40 late transcription *in vitro* by a cellular factor binding the simian virus 40 21-base-pair repeat promoter element. *Proc. Natl Acad. Sci. USA*, **84**, 6025–6029.

Klingenhoff,A., Frech,K., Quandt,K. and Werner,T. (1999) Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics*, **15**, 180–186.

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. Appl. Biosci.*, **11**, 477–488.

Kulp,D., Haussler,D., Reese,M.G. and Eeckman,F.H. (1996) A generalized hidden Markov model for the recognition of human genes in DNA. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **4**, 134–142.

Lander,E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lim,L.C., Swendeman,S.L. and Sheffery,M. (1992) Molecular cloning of the alpha-globin transcription factor CP2. *Mol. Cell. Biol.*, **12**, 828–835.

Lim,L.C., Fang,L., Swendeman,S.L. and Sheffery,M. (1993) Characterization of the molecularly cloned murine alpha-globin transcription factor CP2. *J. Biol. Chem.*, **268**, 18 008–18 017.

Murata,T., Nitta,M. and Yasuda,K. (1998) Transcription factor CP2 is essential for lens-specific expression of the chicken alphaA-crystallin gene. *Genes Cells*, **3**, 443–457.

Parada,C.A., Yoon,J.B. and Roeder,R.G. (1995) A novel LBP-1-mediated restriction of HIV-1 transcription at the level of elongation *in vitro*. *J. Biol. Chem.*, **270**, 2274–2283.

Perier,R.C., Praz,V., Junier,T., Bonnard,C. and Bucher,P. (2000) The Eukaryotic Promoter Database (EPD). *Nucleic Acids Res.*, **28**, 302–303.

Powell,C.M., Rudge,T.L., Zhu,Q., Johnson,L.F. and Hansen,U. (2000) Inhibition of the mammalian transcription factor LSF induces S-phase-dependent apoptosis by downregulating thymidylate synthase expression. *EMBO J.*, **19**, 4665–4675.

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

Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.

Ptashne,M. (1988) How eukaryotic transcriptional activators work.

Nature, **335**, 683–689.

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

Quandt,K., Frech,K., Karas,H., Wingender,E. and Werner,T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878–4884.

Reese,M.G., Hartzell,G., Harris,N.L., Ohler,U., Abril,J.F. and Lewis,S.E. (2000) Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.*, **10**, 483–501.

Romerio,F., Gabriel,M.N. and Margolis,D.M. (1997) Repression of human immunodeficiency virus type 1 through the novel cooperation of human factors YY1 and LSF. *J. Virol.*, **71**, 9375–9382.

Scherf,M., Klingenhoff,A. and Werner,T. (2000) Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol.*, **297**, 599–606.

Shirra,M.K. and Hansen,U. (1998) LSF and NTF-1 share a conserved DNA recognition motif yet require different oligomerization states to form a stable protein–DNA complex. *J. Biol. Chem.*, **273**, 19 260–19 268.

Shirra,M.K., Zhu,Q., Huang,H.C., Pallas,D. and Hansen,U. (1994) One exon of the human LSF gene includes conserved regions involved in novel DNA-binding and dimerization motifs. *Mol. Cell. Biol.*, **14**, 5076–5087.

Smit,A.F.A. and Green,P. (Personal communication) RepeatMasker: http://ftp.genome.washington.edu/RM/RepeatMasker.html.

Sundseth,R. and Hansen,U. (1992) Activation of RNA polymerase II transcription by the specific DNA- binding protein LSF. Increased rate of binding of the basal promoter factor TFIIB. *J. Biol. Chem.*, **267**, 7845–7855.

Swendeman,S.L., Spielholz,C., Jenkins,N.A., Gilbert,D.J., Copeland,N.G. and Sheffery,M. (1994) Characterization of the genomic structure, chromosomal location, promoter, and development expression of the alpha-globin transcription factor CP2. *J. Biol. Chem.*, **269**, 11 663–11 671.

Tronche,F., Ringeisen,F., Blumenfeld,M., Yaniv,M. and Pontoglio,M. (1997) Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. *J. Mol. Biol.*, **266**, 231–245.

Venter,J.C. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Volker,J.L., Rameh,L.E., Zhu,Q., DeCaprio,J. and Hansen,U. (1997) Mitogenic stimulation of resting T-cells causes rapid phosphorylation of the transcription factor LSF and increased DNA-binding activity. *Genes Dev.*, **11**, 1435–1446.

Wagner,A. (1999) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, **15**, 776–784.

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

Wasserman,W.W., Palumbo,M., Thompson,W., Fickett,J.W. and Lawrence,C.E. (2000) Human-mouse genome comparisons to locate regulatory sites. *Nature Genet.*, **26**, 225–228.

Weis,L. and Reinberg,D. (1992) Transcription by RNA polymerase II: initiator-directed formation of transcription-competent complexes. *Faseb J.*, **6**, 3300–3309.

Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Pruss,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.

Yoon,J.B., Li,G. and Roeder,R.G. (1994) Characterization of a family of related cellular transcription factors which can modulate human immunodeficiency virus type 1 transcription *in vitro*. *Mol. Cell. Biol.*, **14**, 1776–1785.