

## COMMUNICATION

# A Statistical Model for Locating Regulatory Regions in Genomic DNA

Evelyn M. Crowley<sup>1</sup>, Kathryn Roeder<sup>1</sup> and Minou Bina<sup>2\*</sup>

<sup>1</sup>Department of Statistics  
Carnegie Mellon University  
Pittsburgh, PA 15213-3890  
USA

<sup>2</sup>Department of Chemistry  
Purdue University  
W. Lafayette, IN 47907, USA

In addition to genes, chromosomal DNA contains sequences that serve as signals for turning on and off gene expression. These signals are thought to be distributed as clusters in the regulatory regions of genes. We develop a Bayesian model that views locating regulatory regions in genomic DNA as a change-point problem, with the beginning of regulatory and non-regulatory regions corresponding to the change points. The model is based on a hidden Markov chain. The data consist of nucleotide positions of protein-binding elements in a genomic DNA sequence. These positions are identified using a reference catalogue containing elements that interact with transcription factors implicated in controlling the expression of protein-encoding genes. Among the protein-binding elements in a genomic DNA sequence, the statistical model automatically selects those that tend to predict regulatory regions. We test the model using viral sequences that include known regulatory regions and provide the results obtained for human genomic DNA corresponding to the  $\beta$  globin locus on chromosome 11.

© 1997 Academic Press Limited

**Keywords:** hidden Markov chains; Bayesian statistics; HIV-1 regulatory regions; adenovirus regulatory regions; LCR

\*Corresponding author

We propose a solution to the following problem: given the sequence of genomic DNA, how can one locate the regulatory regions in this DNA? This problem is complex because the DNA segments that control gene expression can be anywhere in the genome, including in non-coding sequence (introns) of genes and in areas that might be many kilobases upstream or downstream from the transcription initiation sites. In addition, because the control elements are relatively short, they occur not only in the regulatory regions but also elsewhere in the DNA sequence, probably by chance. See, for example Figure 1, which shows the complex distribution pattern of transcription factor binding elements in a long DNA segment (73,308 bp) that defines the human  $\beta$  globin locus on chromosome 11. Our hypothesis is that, despite its complexity,

the distribution of control elements or “words” in genomic DNA is non-random.

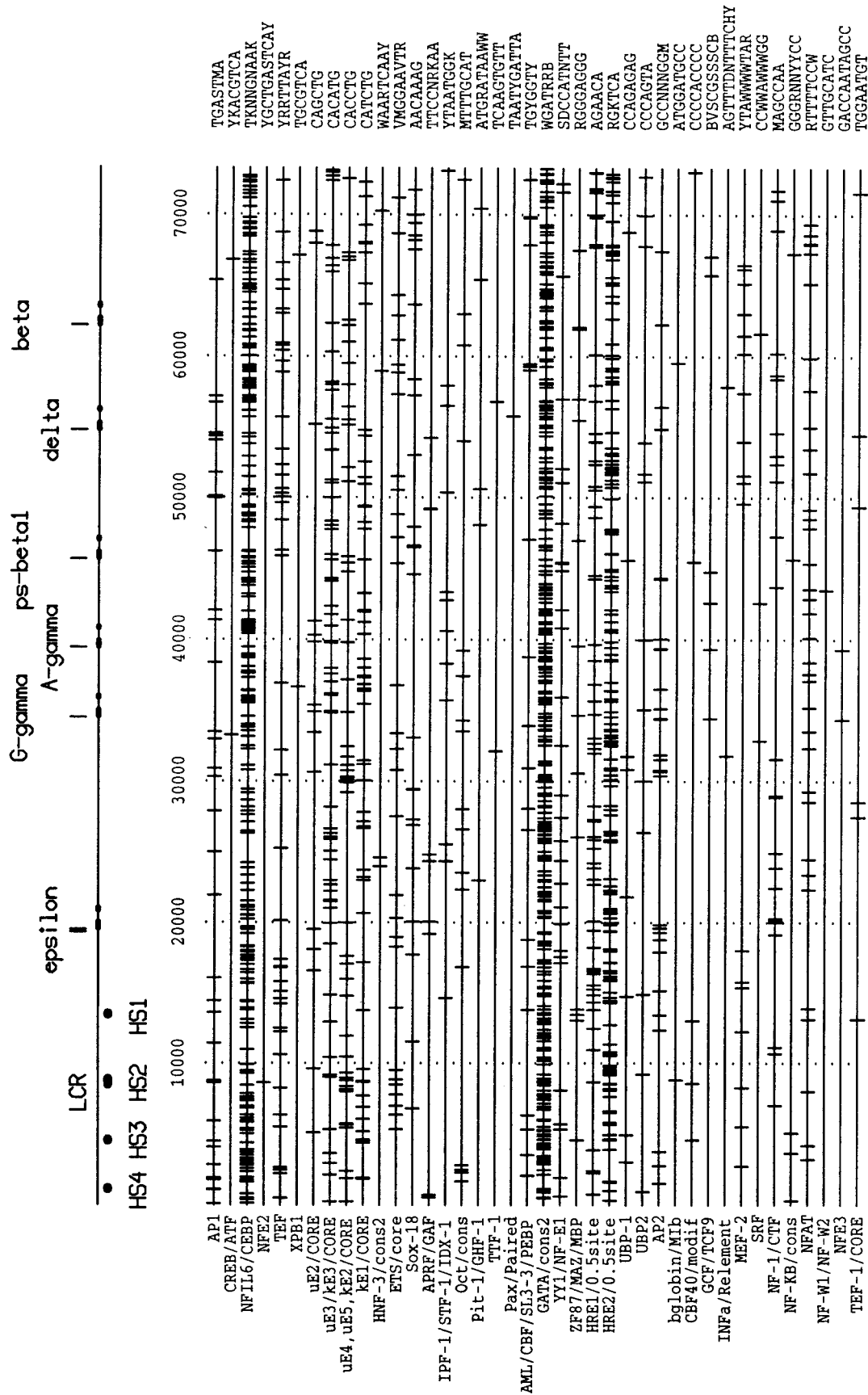
Here, we develop a Bayesian model for locating control regions in genomic DNA. The statistical model is known as a Hidden Markov Chain, a class of models that has been used with remarkable success in the genomic literature (e.g. see Churchill, 1989; Kruglyak *et al.*, 1996).

## Criteria for defining control elements in DNA

Our statistical model requires a catalogue of DNA sequences that act as signals in controlling gene expression. The entries in the catalogue are used to find the nucleotide positions of the protein-binding elements (words) in a given genomic DNA sequence in order to create a data file (e.g. see Figure 1). The focus of our study is on elements that control the expression of protein-encoding genes. Our initial approach involved constructing a catalogue containing the reported DNA control elements. However, our analyses indicate that this approach does not produce a suitable catalogue for testing statistical models, because of the problem of an underlying redundancy that is present in the reported experimental data.

A portion of this research was completed while E. M. C. was with the Department of Statistics, Purdue University, W. Lafayette, IN 47907, USA.

Abbreviations used: Ad-2, adenovirus-2; DNase I, deoxyribonuclease I; HIV-1, human immunodeficiency virus type 1; LCR, locus control region; LTR, long terminal repeat; pol II, polymerase II; SV40, simian virus 40.



**Figure 1.** Distribution of protein-binding elements in the human  $\beta$  globin locus on chromosome 11 (GenBank accession no. U01317). The top panel provides a map summarizing the location of the LCR and the genes in the locus;  $\epsilon$ ,  $G\gamma$ ,  $A\gamma$ ,  $\psi\beta 1$ ,  $\delta$  and  $\beta$ . Below the map, DNA-binding elements for transcription factors are shown on separate lines that define the same DNA sequence. The binding elements are from a catalogue (Pattern15.dat) constructed as described in the text. The names of the factors are shown on the left and their corresponding DNA-binding elements are provided on the right. The ambiguity codes in the sequences follow the standard notations: Y = T/C; R = A/G; W = A/T; S = G/C; K = G/T; M = A/C; B = C/G/T; D = A/G/T; H = A/C/T; V = A/C/G; N = A/G/C/T.

To tackle the problem of redundancy, when possible we define the control elements according to “core motifs” recognized by proteins that share a similar DNA-binding domain. For example, we define a DNA sequence as ETS/core to represent the binding site for ETS domain-containing proteins: i.e. ETS1, ETS2, ELK1/SAP1, E4TF1, E74, ELF-1, PU1/spi, GABP. Similarly, we designate a common site for proteins of the C/EBP family, including NF-IL6. In some cases, we construe a common site for related proteins: i.e. the Rel family of proteins including NF- $\kappa$ B. In other cases, we attempt to reconcile old and recent data by inferring a common site for proteins that appear to be related. For example, in the literature there are three different designations (ZF87, MAZ and MBP) for an identical protein; for brevity, we will refer to this protein as MAZ. We note that the reported binding site for MAZ (GGGAGGG) is similar to the reported site for H4TF-1. From inspection of experimental data (methylation interference analysis and sequence of protein-binding sites) we infer that H4TF-1 might be identical with or related to MAZ. Therefore, in this case, we modify the H4TF-1 and the MAZ site to RGGGAGGG. This modification provides a more restricted recognition sequence for MAZ and thus the modified site remains consistent with the reported DNA-binding properties of this protein. We classify the E-boxes (CANNTG) according to the dinucleotides that occupy the NN position; see for example Figure 1. In most instances we avoid “loose” consensus sequences. However, occasionally, loose sequences cannot be dismissed. This is the case for the GATA site and the half-site recognition sequence of nuclear receptors.

In tracing the sequences of protein-binding elements, we made extensive use of the survey reported by Boulikas (1994). To make functionally relevant predictions, we impose the criterion that the entries in our catalogue must represent elements that have experimentally been shown to be functional in transcription and/or in DNA-binding assays. In some genomic sequences, we note relatively dense clustering of the recognition sequence for the factor known as GCF/TCF9 (Kageyama & Pastan, 1989; Boulikas, 1994). Since extensive clustering of binding sites for a single transcription factor might bias the results the GCF/TCF9 sites are not included in the analysis.

### Statistical model

The Bayesian model that we develop for locating control regions specifically reflects the research hypothesis: a clustering of “words” in a given region of genomic DNA reflects the presence of a regulatory region (Ambrose & Bina, 1990). For data analysis, the genomic DNA is divided into  $I$  evenly spaced, short intervals, with length  $\Delta$  equal to the longest of the words in the sequence. To determine if an unusual number of words occurred in a region of the sequence, the presence or absence of

each type of word within each interval is noted. Thus for  $j = 1, \dots, J$  words, let  $X_{ij} = 1$  if a binding element of  $j$ th type occurs in the  $i$ th interval, and 0 otherwise.

Positions on the DNA are either in a regulatory region ( $r = 1$ ) or a non-regulatory region ( $r = 0$ ). Let  $Y_i$ ,  $i = 1, \dots, I$  represent the state of the DNA, by position. In general, these states are unknown; they are the quantities to be estimated. Based on the distribution of words observed within the intervals, we estimate the probability that the sequence is in a regulatory region for each of these intervals. The statistical model assumes that if the sequence is currently in a non-regulatory state, then it is considerably more likely to remain in this state for the next interval than to move to a regulatory state. Alternatively, if the sequence is currently in a regulatory state, then it is somewhat more likely to remain in that state for the following interval. This feature is modeled using a Markov chain with transition probabilities  $Pr(Y_{i+1} = 1|Y_i = 0) = \lambda$  and  $Pr(Y_{i+1} = 0|Y_i = 1) = \tau$ .

A key feature of our model is the flexibility it allows for the number of regulatory regions observed in a sequence. A prior is specified for  $\lambda$  and  $\tau$  that influences the number of regulatory regions likely to be found in an interval (say, zero to four), but does not force a particular number of regions. For instance, the prior may favor two regulatory regions, but if the data do not suggest the presence of any regulatory regions, then the model is not likely to predict any. In our model, a high prior probability is given to regulatory regions occurring zero to four times every 5000 base-pairs. Also, we anticipate regulatory regions of length approximately 200 to 600 base-pairs. These features are incorporated into the model with beta priors for  $\lambda$  and  $\tau$ . For example, when  $\Delta$  is equal to 13 bp, we use a beta (1.3,100) prior for  $\lambda$  and a beta (5.5,100) for  $\tau$ .

Some words are more common than others, and hence we expect to observe them more often whether or not the process indicates a regulatory region. We reflect the relative density of different words in genomic DNA by using weights that are a function of the relative frequency of each type of word. For each  $j$ , let  $n_j = \sum_i X_{ij}$  and let  $w_j$  be the relative frequency of words of type  $j$ ,  $n_j / \sum_j n_j$ . Then we define a weight function,  $c_{jr}$ , that depends upon  $w_j$ , but varies depending upon whether the sequence is in a regulatory region or not. The fundamental assumption is that words occur with greater frequency in regulatory regions. Let  $\theta_r$  denote the probability of observing any of the words in the catalogue, in a given interval, when the process is in state  $r$ . Given that  $Y_i = r$ , we assume that a word of the  $j$ th type occurs in the  $i$ th interval with probability  $\theta_r c_{jr}$ . We assume that  $X_{ij}$  given  $Y_i = r$  is distributed Bernoulli ( $\theta_r c_{jr}$ ),  $i = 1, \dots, I$ ;  $j = 1, \dots, J$ ; and  $r = 0$  or 1. The observations are assumed to be independent. Because we know  $\theta_0 < \theta_1$  we use a uniform prior with an order re-

striction. In a non-regulatory region we choose  $c_{j0} = w_j$  because it describes the probability of observing a word in a non-regulatory interval times the probability that it is of the  $j$ th type. The same reasoning does not hold in regulatory regions, where rare words might be slightly more probable than expected, based on the relative frequency of rare words in the datafile. Therefore, we choose  $c_{j1}$  so that  $c_{j1}/c_{j0}$  is a decreasing function of  $w_j$ . Experience in predicting known regulatory regions indicates that  $c_{j1} = w_j \log(1/w_j)$  is a good choice.

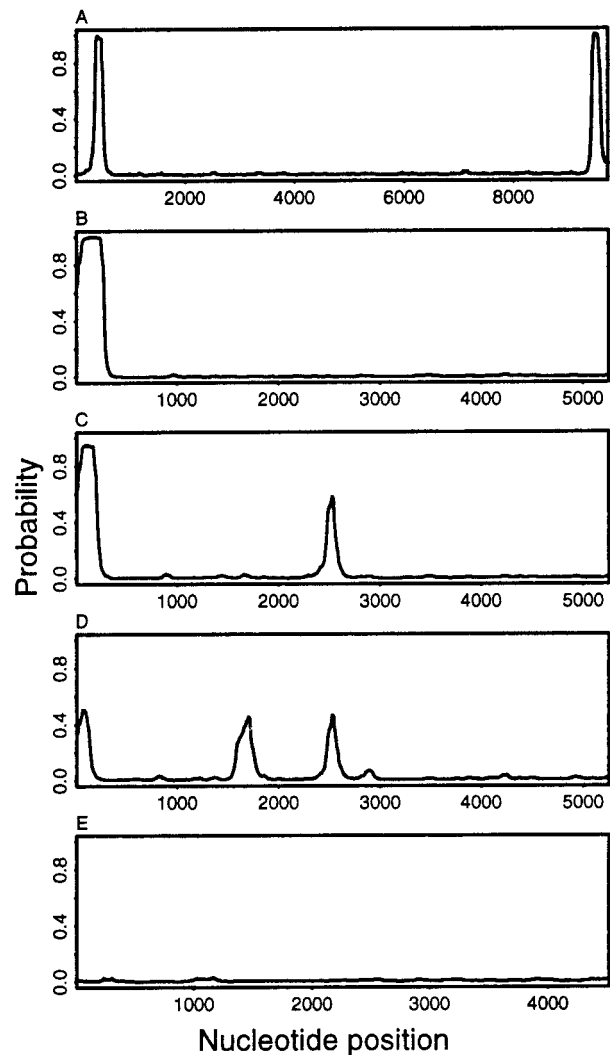
It is realistic to assume that some words might have little predictive power, at least for some types of regulatory regions. To compensate for this inherent uncertainty, we introduce a binary variable  $Z_j$  that equals 1 if the  $j$ th word is a "correct word" and zero if it is an "incorrect word". Then we use a Bayesian technique to estimate the probability that  $Z_j = 1$  in the particular sequence at hand. We modify our model to allow for incorrect words in the catalogue by treating regulatory regions and non-regulatory regions alike when  $Z_j = 0$ . Thus  $X_{ij}|Y_i = r, Z_j = 0 \sim \text{Bernoulli}(\theta_0 c_{j0})$  and  $X_{ij}|Y_i = r, Z_j = 1 \sim \text{Bernoulli}(\theta_r c_{jr})$ . Because we do not know which words belong in the catalogue *a priori*, we model  $Z_j$  with a Bernoulli (0.5). This procedure is also helpful in identifying potential redundant entries in the catalogue of words.

We use Gibbs sampling with Metropolis steps to estimate the parameters of our model (for reviews of these techniques, see Gilks *et al.*, 1993; Smith & Roberts, 1993). The result of the analysis is an estimate of the posterior probability that each interval is in a regulatory state. The posterior probabilities are plotted as a function of the location in the sequence to give a picture of the estimated process. Strong peaks in the plot suggest regulatory regions.

### Representative examples

As examples, we examine the predictions of the model for genomic sequences of three viruses (HIV-1, SV40 and Ad-2), and subsequently analyze the human  $\beta$  globin locus on chromosome 11. We selected viruses for testing the model because their genome includes known regulatory regions (see Jones *et al.*, 1988; Gaynor, 1992). The control of HIV-1 gene transcription involves regulatory signals located in the viral long terminal repeats (Gaynor, 1992). In the analysis of a prototype HIV-1 proviral DNA, the model yields two well-defined peaks (Figure 2A). There is an exact correspondence between the location of these peaks and the two long terminal repeats in the HIV-1 DNA.

The regulatory region of SV40 genome is compact: it includes the SV40 replication origin, elements required for bidirectional transcription of the viral "early" and "late" genes and two copies of a segment (ENH1 and ENH2) that act in upregulation of transcription (Jones *et al.*, 1988). The model identifies the regulatory region in SV40 DNA as a single peak with probability 1



**Figure 2.** Posterior probability distribution for the location of regulatory regions within several DNA sequences: A, HIV-1 DNA (GenBank accession no. K03455); B, SV40 DNA (GenBank accession no. J02400); C, an artificial construct in which one SV40 enhancer is at the *Bam*HI site; D, an artificial construct in which one enhancer repeat is at the *Bam*HI site and the other is at the *Eco*RI site; E, an artificial construct (pSV0CAT) made by deleting the regulatory region of SV40 from the pSV2CAT sequence (GenBank accession no. VB0065).

(Figure 2B). To further evaluate the model, we examine two artificial constructs using the *Bam*HI and *Eco*RI sites in SV40 DNA as landmarks. In the first construct, ENH1 is moved to the *Bam*HI site. As expected, for this artificial construct the model predicts a major peak at the modified regulatory region and another peak at the *Bam*HI site (Figure 2C). In the second construct, ENH1 is moved to the *Bam*HI site, as above, and ENH2 to the *Eco*RI site. In this case, the model also makes correct predictions: a peak at the modified regulatory region and peaks at the *Eco*RI and *Bam*HI sites (Figure 2D). As a potential control, we analyze

another artificial construct (pSV0CAT) that does not include regulatory sequences (Gorman *et al.*, 1982). As expected, no region with substantial probability is observed for pSV0CAT (Figure 2E). The minor bumps in the profile provide an example of the level of background noise.

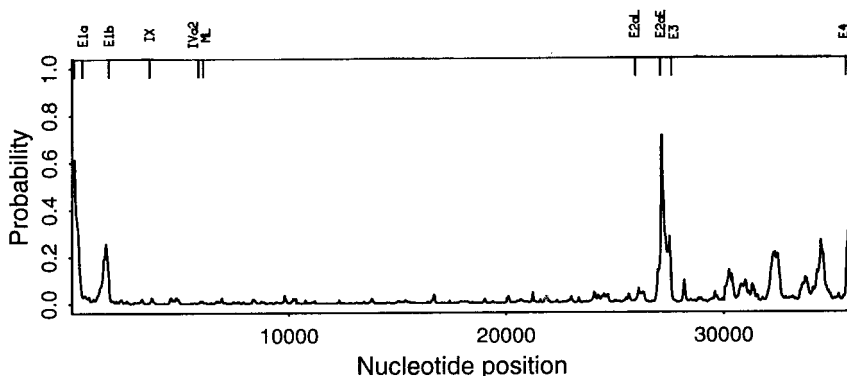
To test a longer sequence, we examined a genome (Ad-2) containing 35,937 base-pairs. Even though the organization of genes in Ad-2 DNA is relatively complex (Roberts *et al.*, 1986), our statistical model makes predictions that correlate closely with several regions that are expected to function in the regulation of Ad-2 gene expression (Jones *et al.*, 1988). For example, the model predicts a regulatory region between positions 1 and 320 in Ad-2 DNA (Figure 3). This region (0.61 probability) includes the 5' inverted repeat and sequences that are upstream of the transcription initiation site of the E1a gene. A second predicted regulatory region (0.25 probability, at position 1575) is upstream of the E1b gene. A third predicted region (0.71 probability, at 27,170) is upstream of the gene (E2a-early) that encodes the 72 K early mRNAs. A fourth region (0.28 probability, at 27,520) is upstream of the E3 gene. A fifth predicted regulatory region (0.26 probability at 34,450) is within the E4 gene. Lastly, a sixth predicted region (0.36 probability, between 35,690 and 35,940) is upstream of the E4 gene and includes the 3' inverted repeat in the adenovirus genome.

As an example of the human genomic sequences that we have analyzed, we provide the results obtained for the human  $\beta$  globin locus. We have chosen this example because the expression of the genes in this locus is regulated by a DNA segment that is far upstream of the genes in the locus (e.g. see Tuan *et al.*, 1985; Forrester *et al.*, 1987; Grosveld *et al.*, 1987; Pruzina *et al.*, 1991; Ellis *et al.*, 1993; Phillipson *et al.*, 1993; Bungert *et al.*, 1995; Stamatoyannopoulos *et al.*, 1995). This locus control region (LCR) acts during the developmental times when the  $\beta$  globin locus is transcriptionally active; the LCR includes four subregions; HS1, HS2, HS3 and HS4 (Tuan *et al.*, 1985; Forrester *et al.*, 1987; Grosveld *et al.*, 1987; Pruzina *et al.*, 1991; Ellis *et al.*, 1993; Phillipson *et al.*, 1993; Bungert *et al.*, 1995; Stamatoyannopoulos *et al.*, 1995). Our statistical

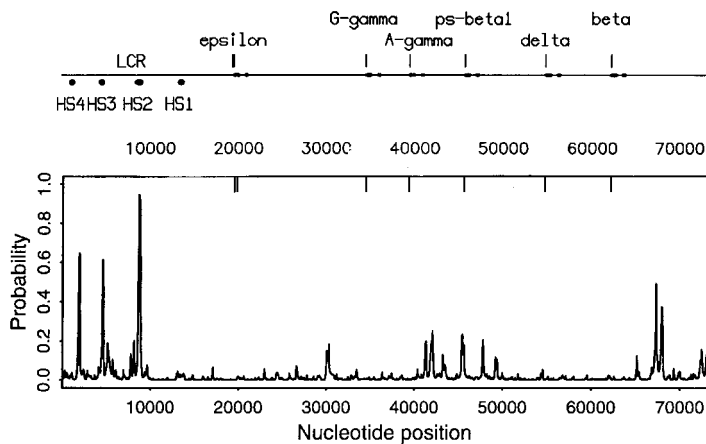
model predicts three regulatory regions that fall within the LCR (Figure 4). Experimental data suggest that HS4 is between 1120 and 1190; the statistical model predicts a regulatory segment between 1815 and 1990, with 0.65 probability. Experimental data indicate that HS3 is between 4581 and 4650; the statistical model predicts a regulatory region between 4510 and 4680, with 0.61 probability. Experimental data have localized HS2 between 8640 and 8870; the model predicts a regulatory region between 8560 and 8890, with 0.95 probability. The model does not detect a regulatory region corresponding to HS1 (experimental data 13,410 to 13,490). However, the model predicts two regulatory regions (0.49 probability at 67,275 to 67,370 and 0.37 probability at 67,895 to 68,070) downstream of the  $\beta$  globin gene (Figure 4). It is possible that these predicted regions represent the 3' HS1, localized downstream of the gene cluster (Tuan *et al.*, 1985; Forrester *et al.*, 1987; Grosveld *et al.*, 1987; Pruzina *et al.*, 1991; Ellis *et al.*, 1993; Phillipson *et al.*, 1993; Bungert *et al.*, 1995; Stamatoyannopoulos *et al.*, 1995).

### Comparison to other models

The majority of the algorithms developed to date have focused on predicting the regulatory regions that encompass the promoter elements and sequences that are upstream of transcription initiation sites (e.g. see Bucher & Trifonov, 1986; Bucher, 1990; Prestridge, 1995; Kondrakhin *et al.*, 1995; Chen *et al.*, 1995). In contrast, our statistical model does not consider the sequence elements that direct the initiation of transcription but rather attempts to identify regulatory regions regardless of their location. Clustering is the paradigm on which we have based our statistical model. Related work includes the method developed by Kondrakhin *et al.* (1995). In their analysis of sequences with "regulatory potential" Kondrakhin *et al.* (1995) have noted that the binding sites for a number of transcription factors are distributed unevenly in the DNA sequence and invoked the idea of looking for clusters of binding sites to improve the accuracy of finding regulatory regions in DNA. This idea agrees with a previous prediction that



**Figure 3.** Posterior probability distribution for the location of regulatory regions in adenovirus type 2 (Ad-2) DNA (GenBank accession no. J01917). The vertical bars shown above the theoretical plot mark the transcription start sites for the Ad-2 genes.



**Figure 4.** Posterior probability distribution for the location of regulatory regions in the human  $\beta$  globin locus on chromosome 11. The sequence (GenBank accession no. U01317) is the same as that shown in Figure 1. The top panel provides the organization of the genes in the locus. The locus is depicted as a line; the exons as thicker lines. The locations of the DNase I hypersensitive regions are shown below the LCR as dots marked by HS4, HS3, HS2 and HS1. The vertical bars in the map and in the theoretical plot mark the transcription start sites.

clustering of sites appears to be a hallmark of regions that can exert long-range effects on the regulation of transcription (Ambrose & Bina, 1990).

Our statistical model differs from approaches that directly utilize the density of transcription factor binding sites for finding potential regulatory regions in DNA. Our model draws on the methods of Churchill (1989, 1992), which model the DNA sequence as a hidden Markov process. His approach allows estimation of the local properties of a DNA sequence without the need for arbitrary specification such as a window size and further allows graphical representation of results as smooth curves.

Like a number of other theoretical approaches, our model relies on experimental data for defining the DNA sequence elements that bind factors that regulate genes transcribed by RNA polymerase II. Over the past decade, several listings have appeared in the literature (e.g. see Wingender, 1988; Ghosh, 1990; Faisst & Meyer, 1992; Boulikas, 1994). In general, compiling catalogues of DNA control elements is a challenging problem. For example, as noted previously (Chen *et al.*, 1995), a major drawback is the problem of redundancy. Another drawback is the nagging question of accuracy. In order to keep these problems somewhat under control, we have chosen to construct our own catalogue of pol II regulatory elements for analyzing genomic DNA. In most cases, we prefer naturally occurring sequences to those deduced by selection approaches based on polymerase chain reactions. We note that sequences in our catalogue primarily represent protein-binding elements in primates. Lacking are sequences representing binding sites for homeotic proteins, since these sites are not clearly defined. Many of the problems associated with constructing databases of regulatory sequences might be resolved as more refined experimental data emerge in the literature.

Chen *et al.* (1995) have developed a database of weight matrices of transcription factor binding sites to reduce the number of false positives if the length of a database entry is shorter than the actual pattern utilized by the regulators of gene ex-

pression. Our model incorporates weights that are a function of the relative abundance of a given site in a DNA sequence. However, in our model the weights vary depending upon whether or not a site appears in a regulatory region. Kondrakhin *et al.* (1995) have noted that some protein-binding elements appear more often than others in the regulatory regions of genes. We find that the occurrence of a site in a regulatory region appears to be context-dependent. That is, a site might be a good predictor of a regulatory region in one sequence but a poor predictor in another. We have therefore developed a statistical method to identify the words that make a significant contribution to the regulatory regions predicted by the model.

We note that the reported methods analyze primarily sequences that are less than 10,000 bp. Our goal is to analyze relatively large genomic DNA segments. Thus far, our program works effectively for sequences that are less than 30 kb, but problems with computation time are encountered for longer sequences. In some instances, the computer time becomes prohibitive and the analysis does not converge. We hope to eventually remedy this problem. Nonetheless, the statistical model appears to be relatively robust. So far, we have not detected examples of "over prediction", as observed for some of the competing algorithms. "Under prediction" appears common and is probably a consequence of incompleteness of our catalogue of binding sites. This problem is likely to disappear as sites for novel factors are discovered and as refined methods improve the accuracy of experimental data.

The Fortran code to perform the statistical analysis is available from the authors upon request.

## Acknowledgements

This research was partially supported by National Science Foundation grants DMS-9303556 (E.M.C.), DMS-9496219 and DMS-9508427 (K.R.), and by National Institutes of Health grant RO1AI29121 (M.B.).

## References

- Ambrose, C. & Bina, M. (1990). Strategy for statistical-mapping of potential regulatory regions in the human genome. *J. Mol. Biol.* **216**, 485–490.
- Boulikas, T. (1994). A compilation and classification of DNA binding sites for protein transcription factors from vertebrates. *Crit. Rev. Euk. Gene Express.* **4**, 117–321.
- 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.
- Bucher, P. & Trifonov, E. N. (1986). Compilation and analysis of eukaryotic Pol II promoter sequences. *Nucl. Acids Res.* **14**, 10009–10026.
- Bungert, J., Davé, U., Lim, K., Lieu, K. H., Shavit, J. A., Liu, Q. & Engel, J. D. (1995). Synergistic regulation of human  $\beta$ -globin gene switching by locus control region elements HS3 and HS4. *Genes Dev.* **9**, 3083–3096.
- Chen, Q. K., Hertz, G. Z. & Stormo, G. D. (1995). MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *CABIOS*, **41**, 164–171.
- Churchill, G. A. (1989). Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.* **41**, 164–171.
- Churchill, G. A. (1992). Hidden Markov chains and the analysis of genome structure. *Comp. Chem.* **16**, 107–115.
- Crossley, M. & Orkin, S. H. (1993). Regulation of the  $\beta$ -globin locus. *Curr. Opin. Genet. Dev.* **3**, 232–237.
- Ellis, J., Talbot, D., Dillon, N. & Grosveld, F. (1993). Synthetic human  $\beta$ -globin 5'HS2 constructs function as locus control regions only in multicopy transgene concatamers. *EMBO J.* **12**, 127–134.
- Engel, J. D. (1993). Developmental regulation of human  $\beta$ -globin gene transcription: a switch of loyalties? *Trends Genet.* **9**, 304–309.
- Faisst, S. & Meyer, S. (1992). Compilation of vertebrate-encoded transcription factors. *Nucl. Acids Res.* **20**, 3–26.
- Forrester, W. C., Takegawa, S., Papayannopoulou, T., Stamatoyannopoulos, G. & Groudine, M. (1987). Evidence for a locus activation region: the formation of developmentally stable hypersensitive sites in globin-expressing hybrids. *Nucl. Acids Res.* **15**, 10159–10177.
- Fritsch, E. F., Lawn, R. M. & Maniatis, T. (1980). Molecular cloning and characterization of the human  $\beta$ -like globin gene cluster. *Cell*, **19**, 959–972.
- Gaynor, R. (1992). Cellular transcription factors involved in the regulation of HIV-1 gene expression. *AIDS*, **6**, 347–363.
- Ghosh, D. (1990). A relational database of transcription factors. *Nucl. Acids Res.* **18**, 1749–1756.
- Gilks, W. R., Clayton, D. G., Spiegelhalter, D. J., Best, N. G., McNeil, A. J., Sharples, L. D. & Kirby, A. J. (1993). Modelling complexity: applications of Gibbs sampling in medicine. *J. Roy. Stat. Soc. ser. B*, **55**, 39–52.
- Gorman, C. M., Moffat, L. F. & Howard, B. H. (1982). Recombinant genomes which express chloramphenicol acetyltransferase in mammalian cells. *Mol. Cell. Biol.* **2**, 1044–1051.
- Grosveld, F., van Assendelft, G. B., Greaves, D. R. & Kollias, G. (1987). Position-independent, high-level expression of the human  $\beta$ -globin gene in transgenic mice. *Cell*, **51**, 975–985.
- Jones, N. C., Rigby, P. W. J. & Ziff, E. B. (1988). Trans-acting protein factors and the regulation of eukaryotic transcription: lessons from studies on DNA tumor viruses. *Genes Dev.* **2**, 267–281.
- Kageyama, R. & Pastan, I. (1989). Molecular cloning and characterization of a human DNA binding factor that represses transcription. *Cell*, **59**, 815–825.
- Kondrakhin, Y. V., Kel, A. E., Kolchanov, N. A., Romashchenko, A. G. & Milanesi, L. (1995). Eukaryotic promoter recognition by binding sites for transcription factors. *CABIOS*, **11**, 477–488.
- Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. & Lander, E. S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.* **58**, 1347–1363.
- Phillipsen, S., Pruzina, S. & Grosveld, F. (1993). The minimal requirements for activity in transgenic mice of hypersensitive site 3 of the  $\beta$  globin locus control region. *EMBO J.* **12**, 1077–1085.
- Prestridge, D. S. (1995). Predicting pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol.* **249**, 923–932.
- Pruzina, S., Hanscombe, O., Whyatt, D., Grosveld, F. & Phillipsen, S. (1991). Hypersensitive site 4 of the human  $\beta$  globin locus control region. *Nucl. Acids Res.* **19**, 1413–1419.
- Roberts, R. J., Akusjarvi, G., Alestrom, P., Gelinas, R. E., Gingeras, T. R., Sciaky, D. & Pettersson, U. (1986). A consensus sequence for the adenovirus-2 genome. In *Developments in Molecular Virology* (Doerfler, W., ed.), pp. 1–51. Adenovirus DNA, Boston.
- Smith, A. F. M. & Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *J. Roy. Stat. Soc. ser. B*, **55**, 3–23.
- Stamatoyannopoulos, J. A., Goodwin, A., Joyce, T. & Lowrey, C. H. (1995). NF-E2 and GATA binding motifs are required for the formation of DNase I hypersensitive site 4 of the human  $\beta$ -globin locus control region. *EMBO J.* **14**, 106–116.
- Tuan, D., Solomon, W., Li, Q. & London, I. M. (1985). The “ $\beta$ -like-globin” gene domain in human erythroid cells. *Proc. Natl Acad. Sci. USA*, **82**, 6384–6388.
- Wingender, E. (1988). Compilation of transcription regulating proteins. *Nucl. Acids Res.* **16**, 1879–1902.

Edited by F. E. Cohen

(Received 17 September 1996; received in revised form 6 February 1997; accepted 6 February 1997)