



Detection of eukaryotic promoters using Markov transition matrices

Stéphane Audic* and Jean-Michel Claverie

Structural and Genetic Information Laboratory, C.N.R.S.-E.P. 91, Institute of Structural Biology and Microbiology, 31 Chemin Joseph Aiguier, Marseille 13402, France

(Received 25 July 1996; Accepted 2 December 1996)

Abstract—Eukaryotic promoters are among the most important functional domains yet to be characterized in a satisfactory manner in genomic sequences. Most current detection methods rely on the recognition of individual transcription elements using position-weight matrices (PWM) or consensus sequences. Here, we study a simple promoter detection algorithm based on Markov transition matrices built from sequences upward from proven transcription initiation sites. The performances have been evaluated on the training set and on a test set of promoter-containing sequences. The results on the training set are surprisingly good, given that the algorithm does not incorporate any specific knowledge about promoters. Yet, the program exhibits the pathological behaviour typical of all training set-based methods: a significant decline in performance when confronted with previously unseen sequences. Thus, the Markov algorithm, like the others presently available, does not truly capture the essence of eukaryotic promoters. A detection program based on a Markov model is likely to be blind to categories of promoters without close representatives in the training set. © 1997 Elsevier Science Ltd

1. INTRODUCTION

Two main approaches can be used when trying to recognize functional domains, such as promoters, in anonymous DNA sequences. Signal-based methods look for combinations of biologically determined subsequences known to be present in promoter regions (i.e. individual promoter elements). Content-based methods use overall statistical criteria (or artificial neural networks) to look for a local resemblance of the anonymous sequence with previously identified promoter regions. The first approach is the most satisfactory as it involves a detailed understanding of the biological function under study, here the initiation of mRNA transcription, and attempts to mimic the process by which the transcription machinery recognizes its target. However, pending the complete understanding of a biological process, content-based methods offer provisional solutions of practical use.

A number of programs for recognizing prokaryotic promoters have been published (Alexandrov and Mironov, 1990; Demeler and Zhou, 1991; Hirst and Sternberg, 1992; Horton and Kanehisa, 1992; Hertz and Stormo, 1996). Locating prokaryotic (mainly

E. coli) promoters is a relatively easy task, thanks to the conservation of two well-defined elements, a TATAAT consensus sequence (the -10 box), a TTGACA consensus sequence (the -35 box), as well as the nearby presence of a putative coding region downstream.

The situation is far less ideal for eukaryotic Polymerase II promoters, the structure of which is both more complex and variable, and where “consensus” sequence signals are facultative, and/or of low statistical significance (Claverie *et al.*, 1997; Claverie and Audic, 1996) when present. As an example, Table 1 lists, for PWMs corresponding to three of the major transcriptional elements (TATA-box, CCAAT-box and GC-box; see Bucher, 1990), the number of promoter regions exhibiting a significant ($p < 0.1$) match. The most ubiquitous signal is the TATA-box, but a method based on detecting this signal only (at this stringency) would miss more than half of the promoters and would find one putative promoter every 2500 bp (assuming a 250-bp promoter window). No single signal can serve as a reliable anchor in delineating a promoter region. The simple strategy consisting of looking (at $p < 0.1$) for a TATA-box, or a CAAT-box, or a GC-box would obviously have a better success rate (roughly 68%), but would also increase the rate of false positive (approximately one every 1000 bp).

The recognition of eukaryotic promoters has been attempted using a variety of approaches, from simple

* Author to whom correspondence should be addressed.
E-mail: jmc@igs.cnrs-mrs.fr; Fax: +33 (0) 491164549.
Software inquiries should be addressed to audic@igs.cnrs-mrs.fr.

Table 1. Common transcription elements detected in the promoter regions of EPD sequences in the vertebrate training set

Signal	Number of matches	Significance
TATA	111/245 (45%)	< 0.1
CAAT	45/245 (18%)	< 0.1
GC	76/245 (31%)	< 0.1

combinatorial pattern matching (Claverie and Sauvaget, 1985), promoter element consensus or position-weight matrix (PWM) matching (Bucher, 1990; Wingender, 1994; Prestridge, 1995), to hexamer statistics (Hutchinson, 1996), and neural networks (Matis *et al.*, 1996). Most studies have made use of the Eukaryotic Promoter Database (EPD) (Bucher, 1996), an extensive collection of sequences in proven promoter regions, as an information ("knowledge") source for training algorithms and testing them. According to the same principles on which programs for detecting coding regions have been designed (reviewed in Fickett and Tung, 1992), various measures can be defined and optimized to distinguish between promoter and non-promoter regions. Those measures are then applied to score successive sliding windows of anonymous DNA sequences, and a threshold is used to classify them in the promoter-like or non-promoter-like categories.

Here, we report on our effort to detect promoters using a Markov transition matrix built on EPD sequences to classify anonymous vertebrate genomic sequences into promoter and non-promoter regions. Given that no prior biological knowledge of promoter structure is incorporated into the algorithm, it is not expected to achieve better results than those of previously published methods. Rather, it is used as an objective way to assess the information content of the 250-nucleotide window immediately upstream of the transcription start sites, and as a benchmark against which the contribution of signal-specific recognition can be evaluated. Despite its simplicity and speed, we were surprised to find that the Markov algorithm could achieve, on the training set, performances similar to that previously reported for more sophisticated programs. However, the method exhibits the pathological behaviour seen in its predecessors, and all training-based methods, which is a significant decrease in performance when confronted with previously unseen promoter sequences (test sets).

2. METHODS

Markov chains, essentially a rigorous implementation of older oligomer counting methods (reviewed in Claverie *et al.*, 1990), have been extensively used for sequence analysis, and enjoyed their best success in the detection of protein coding regions in prokaryotes. They are at the core of the popular GenMark program (Borodovsky and McIninch, 1993). A Markov model of order k for a set of DNA sequences is designed as follows. A transition matrix T is computed by recording for each k -mer (i.e. each word of k nucleotides) its probability of being followed by A, C, G or T. Such probabilities are estimated simply by counting the occurrence of each $(k + 1)$ -mers in the data set.

Now, consider a window W of length L in a non-characterized DNA sequence. The probability of this sequence having been generated by a stochastic process according to the transition matrix T is:

$$P(W|T) = p(s_0) \cdot p(n_k|s_0) \cdot \dots \cdot p(n_{L-1}|s_{L-k-1}) \quad (1)$$

where s_i denotes the k -mer starting at the position i of the window, n_i is the nucleotide at position i , $p(s_i)$ the probability of the corresponding k -mer, and $p(n_k|s_i)$ the probability for nucleotide n_k to follow the k -mer s_i .

We now introduce several transition matrices T_i , built from the analysis of l mutually exclusive data sets of functionally homogeneous sequences (e.g. intron/exon/intergenic, or coding/non-coding, etc.), and use Baye's theorem to compute the probability for each of the transition matrices to be at the origin of a given sequence window W :

$$P(T_i|W) = \frac{P(W|T_i)P(T_i)}{\sum_{j=1}^l P(W|T_j)P(T_j)} \quad (2)$$

A promoter detection algorithm based on equation (2) requires two data sets, one for promoter sequences, and one for non-promoter sequences, from which 2 (i.e. $l = 2$) transition matrices T_{prom} (T_1) and T_{nprom} (T_2) are computed. A given anonymous sequence window W will be assigned to the promoter or non-promoter category depending on which of $P(T_{\text{prom}}|W)$ or $P(T_{\text{nprom}}|W)$ is the largest. The a priori probability $P(T_i)$ is one of the free parameters of the model. It would be quite unrealistic to assume an equal a priori probability for promoter and non-promoter regions when scanning anonymous genomic sequences. The a priori probability $P(T_i)$ is denoted $P_{(\text{prom})}$ in Table 2 and 3, and has been varied from 0.5 to 0.01.

2.1. Selection of the Data Sets

Sequences containing proven promoters were extracted from the Genbank or EMBL databases as referenced in EPD (Bucher, 1996). Only sequences from vertebrates were used, including 180 primate sequences, 57 rodent sequences, 24 other mammal sequences and 46 other vertebrate sequences, for a total of 307 promoter sequences and 330 transcription initiation sites. Care was taken to retain a single sequence from each homologous group, as annotated in EPD. The training set was constituted by three-quarters (randomly selected) of the available promoter-containing sequences, while the remainder was used to constitute the test set. Following the commonly accepted definition of a promoter region (Bucher, 1990; Prestridge, 1995), the 250 nucleotides upstream of the transcription initiation site were selected from the training set sequences and used to build the promoter region transition matrix T_{prom} . The remaining nucleotides of the same sequences were used to build the non-promoter region transition matrix T_{nprom} .

2.2. Scoring Successful Predictions

Unknown sequence regions were assigned to the promoter or non-promoter categories according the following simple algorithm. Successive 250-bp

Table 2. Influence of the order of the Markov model on the recognition of promoters. The performances are evaluated on the whole *vertebrate* training set and on a test set of *vertebrate* sequences. Markov transition matrices were computed from the promoter and non-promoter sequence parts of a *vertebrate* training set

$P(\text{prom})$	Vertebrate training set		Vertebrate test set	
	Success	False positive	Success	False positive
Order 4				
0.5	188/245 (77%)	2090 bp	48/84 (57%)	1745 bp
0.1	146/245 (60%)	3110 bp	39/84 (46%)	2675 bp
0.05	136/245 (55%)	3530 bp	38/84 (45%)	2850 bp
0.01	126/245 (51%)	4735 bp	31/84 (37%)	4070 bp
Order 5				
0.5	217/245 (88%)	3800 bp	40/84 (47%)	2515 bp
0.1	209/245 (85%)	5615 bp	31/84 (37%)	3290 bp
0.05	201/245 (82%)	6250 bp	25/84 (30%)	3640 bp
0.01	189/245 (77%)	9125 bp	19/84 (23%)	5900 bp
Order 6				
0.5	229/245 (93%)	22 500 bp	20/84 (24%)	6580 bp
0.1	228/245 (93%)	36 310 bp	17/84 (20%)	8550 bp
0.05	228/245 (93%)	53 570 bp	16/84 (19%)	10 060 bp
0.01	224/245 (91%)	72 620 bp	12/84 (14%)	17 100 bp

sequence windows, sliding by steps of 50, were used to compute the $P(T_{\text{prom}}|W)$ or $P(T_{\text{nprom}}|W)$ conditional probabilities, i.e. the probabilities for the window to be or not to be of the promoter type. A promoter region was considered successfully "detected" when a window centred inside the proven promoter region (defined as the 250 nucleotides before the transcription initiation site) was assigned to the promoter category. The false positive rate of a promoter detection algorithm is an important property. Considering that the 3×10^9 -bp human genome contains about 100 000 genes, we expect one promoter every 30 000 bp (when analysing both strands). Current programs approximately "detect" one promoter every 5000 bp (single strand), a false positive rate that makes their predictions unusable without experimental validation. Here, we evaluated the false positive rate by running our program against the sequences from the non-promoter data set, for which, by definition, each detected "promoter" is very likely to represent a false positive identification.

2.3. Implementation

The promoter detection method was implemented using a general purpose package (Audic, unpublished) for the Markov chain analysis of DNA sequences. This package allows the number of different transition matrices/categories, their a priori probability, the length of the sequence window, etc., to be varied. Using a typical workstation, a promoter detection scan is performed in about 1 s for the single strand analysis of a 50 000-bp sequence. The output

is simply a list of sequence positions (window centres) for which $P(\text{prom}|W) > 0.5$. The sequence scanning program, as well as companion modules to generate transition matrices from sequence data sets, are available.

3. RESULTS AND DISCUSSION

We computed the number of promoters successfully detected in both the training set and the test set for Markov transition matrices of order 4–6. Table 2 shows the results obtained for the various orders on the vertebrate training and test sets, computed from transition matrices built from promoter and non-promoter sequences of the vertebrate training set. Markov models of order 5 were found to offer the best compromise between sensitivity and specificity for the most relevant prior probability value [$P(\text{prom}) = 0.01$]. As expected, the lowest false positive rate was obtained when using the lowest prior probability $P_{(\text{prom})}$ (promoter regions should represent less than 1% of total genomic nucleotides). Unfortunately, this is associated with the lowest success rate (i.e. the specificity of detection cannot be increased without a loss in sensitivity). The performances computed on the whole vertebrate data set or on the primate-only data set were equivalent.

In Table 3, only primate sequences from the vertebrate training and test sets were scored, using the same transition matrices (order 5) as in Table 2. Similar but reciprocal calculations were performed using transition matrices from only primate

Table 3. Primate vs vertebrate promoter recognition. The performances are evaluated on the whole *primate* training set and on a test set of primate sequences. Markov transition matrices were computed from the promoter and non-promoter sequence parts in a *vertebrate* training set

$P(\text{prom})$	Primate training set		Primate test set	
	Success	False positive	Success	False positive
0.5	132/147 (90%)	3500 bp	24/50 (48%)	2080 bp
0.1	129/147 (88%)	5380 bp	19/50 (38%)	3090 bp
0.05	125/147 (85%)	6350 bp	17/50 (34%)	3420 bp
0.01	118/147 (80%)	9230 bp	12/50 (24%)	6380 bp

Table 4. Performances of the best Markov model [order 5, $P(\text{prom}) = 0.01$, see Table 2] detailed for TATA-box (TATA₊) or non-TATA-box (TATA₋) promoter recognition. The average percentages are computed from eight random partitions of the vertebrate promoter in training (2/3) and test (1/3) sets

	Training sets			Test sets		
	All	TATA ₊	TATA ₋	All	TATA ₊	TATA ₋
Success (%)	76 ± 2	86 ± 3	72 ± 2	26 ± 3	46 ± 8	19 ± 5

sequences with nearly identical results. Thus, by this test, primate promoters have nothing special compared with other vertebrate promoters.

Given its remarkable simplicity (and speed), and the fact that our algorithm does not incorporate any biological knowledge about the process of transcription (presence of specific signals, polarity, etc.), it achieves a very satisfactory recognition rate, and relatively low false positive rate, in the training set. However, recognition performances for the test set are below those obtained by current leading methods, in terms of sensitivity at a comparable level of specificity. For instance, PROMOTER Scan (Prestridge, 1995), apparently the best available program, detects 70% of its training set, 54% of its test set, and has a false positive rate of 1 every 5565 bp (per single strand). Our Markov model (Table 2, order 5) only achieves 23% success for a false positive rate of 1 every 5900 bp.

Prestridge's program (Prestridge, 1995) locates promoters by looking for the local accumulation, in a 250-bp window, of matches with previously recognized transcriptional elements. It is thus an explicit sequence similarity-based method. The difference in success rate between our Markov model detection and Prestridge's method, may be accounted for in part by a specific handling of TATA-box containing promoters. The TATA-box signal is separately detected using a PWM the score of which is combined in a somewhat *ad hoc* manner to the score obtained from the match with other transcription elements. The TATA-box signal is a dominant one. By itself, it allows 78% of all TATA-containing promoters to be located at the right location (using a PWM score threshold corresponding to a significance level of 1%, data not shown). This prompted us to analyse the behaviour of our Markov model in recognizing two classes of promoters: with or without TATA-boxes. For this, the set of vertebrate promoters was classified according to the presence or absence of a strong TATA-box signal. The selection was made from the vertebrate EPD sequences associated with a single transcription start site. Promoters were classified as TATA₊ when they exhibited a TATA-box PWM match with a significance level of 1% near the expected location (-30). This partition resulted in 66 TATA₊ promoters and 222 TATA₋ promoters. The recognition performances computed after repeating eight times the random partition in training and test sets are shown in Table 4. Clearly, much better performances were achieved in recognizing TATA₊ promoters, although we used a "general" promoter Markov model trained on a mixture of promoter types. This suggests that TATA₊ promoters constitute a more informative and, perhaps, a more homogeneous class of sequences.

All published promoter recognition algorithms are plagued by the same unpleasant feature: they all exhibit a significant decrease in performance when confronted with previously unseen sequences (the test set) rather than sequences on which they have been trained (training set). Despite the abstract and objective nature of its "education", our Markov chain-based program is no exception to this behaviour. The performances on the test set are divided by 2-3 in terms of sensitivity and divided by 2 in terms of specificity. A Markov model is thus good at capturing promoter sequence information, but does not show any progress in capturing the basic features of eukaryotic promoters in unrelated sequence contexts. The Markov model approach is plagued by the same conservatism that characterizes most sequence analysis programs available today (Claverie *et al.*, 1997).

Acknowledgements—S. Audic is supported by a grant from Incyte Pharmaceuticals to J.-M. Claverie.

REFERENCES

- Alexandrov, N. N. and Mironov, A. A. (1990) Application of a new method of pattern recognition in DNA sequence analysis: a study of *E. coli* promoters. *Nucleic Acids Research* **18**, 1847-1852.
- Bucher, P. (1990) Weight matrix description of four eukaryotic RNA Polymerase II promoter elements derived from 502 unrelated promoter sequences. *Journal of Molecular Biology* **212**, 563-578.
- Bucher, P. (1996) *The Eukaryotic Promoter Database EPD*, EMBL Nucleotide Sequence Data Library, Release 46. European Bioinformatics Institute, Hinxton Hall, Hinxton, Cambridge CB10 1RQ, U.K.
- Borodovsky, M. and McIninch, J. (1993) GENMARK: parallel gene recognition for both DNA strands. *Computers and Chemistry* **17**, 123-133.
- Claverie, J.-M. and Audic, S. (1996) The statistical significance of nucleotide position-weight matrix matches. *Computers Applications in the Biosciences* **12**, 431-439.
- Claverie, J.-M., Poirot, O. and Lopez, F. (1997) The difficulty of identifying genes in anonymous vertebrate sequences. *Computers and Chemistry* **21**, 203-214.
- Claverie, J.-M. and Sauvaget, I. (1985) Assessing the biological significance of primary structure consensus patterns using sequence databanks. I. Heat-shock and glucocorticoid control elements in eukaryotic promoters. *Computer Applications in the Biosciences* **1**, 95-104.
- Claverie, J.-M., Sauvaget, I. and Bougueleret, L. (1990) *k*-tuple frequency analysis: from intron/exon discrimination to T-cell epitope mapping. *Methods in Enzymology* **183**, 237-252.
- Demeler, B. and Zhou, G. W. (1991) Neural network optimization for *E. coli* promoter prediction. *Nucleic Acids Research* **19**, 1593-1599.
- Fickett, J. W. and Tung, C. (1992) Assessment of protein coding measures. *Nucleic Acids Research* **20**, 6441-6450.

- Hertz, G. Z. and Stormo, G. D. (1996) *E. coli* promoter sequences: analysis and prediction. *Methods in Enzymology* **273**, 30–42.
- Hirst, J. D. and Sternberg, M. J. (1992) Prediction of structural and functional features of protein and nucleic acids sequences by artificial neural networks. *Biochemistry* **31**, 7211–7218.
- Horton, P. B. and Kanehisa, M. (1992) An assessment of neural network and statistical approaches for prediction of *E. coli* promoter sites. *Nucleic Acids Research* **20**, 4331–4338.
- Hutchinson, G. B. (1996) The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *Computer Applications in the Biosciences* **12**, 391–398.
- Matis, S., Xu, Y., Shah, M., Guan, X., Einstein, J. R., Mural, R. and Uberbacher, E. (1996) Detection of RNA Polymerase II promoters and polyadenylation sites in human DNA sequences. *Computers and Chemistry* **20**, 135–140.
- Prestridge, D. S. (1995) Predicting Pol II promoter sequences using transcription factor binding sites. *Journal of Molecular Biology* **249**, 923–932.
- Wingender, E. (1994) Recognition of regulatory regions in genomic sequences. *Journal of Biotechnology* **35**, 273–280.