

Characterization of Prokaryotic and Eukaryotic Promoters Using Hidden Markov Models

Anders Gorm Pedersen *, Pierre Baldi †, Søren Brunak ‡ and Yves Chauvin §

Abstract

In this paper we utilize hidden Markov models (HMMs) and information theory to analyze prokaryotic and eukaryotic promoters. We perform this analysis with special emphasis on the fact that promoters are divided into a number of different classes, depending on which polymerase-associated factors that bind to them. We find that HMMs trained on such subclasses of *Escherichia coli* promoters (specifically, the so-called σ^{70} and σ^{54} classes) give an excellent classification of unknown promoters with respect to sigma-class. HMMs trained on eukaryotic sequences from human genes also model nicely all the essential well known signals, in addition to a potentially new signal upstream of the TATA-box. We furthermore employ a novel technique for automatically discovering different classes in the input data (the promoters) using a system of selforganizing parallel HMMs. These selforganizing HMMs have at the same time the ability to find clusters and the ability to model the sequential structure in the input data. This is highly relevant in situations where the variance in the data is high, as is the case for the subclass structure in for example promoter sequences.

Key words: hidden Markov models (HMMs), information theory, DNA sequence analysis, *Escherichia coli*, *Homo sapiens*, promoters.

* Center for Biological Sequence Analysis, building 206, The Technical University of Denmark, DK-2800, Denmark, gorm@cbs.dtu.dk, (+45) 4525-2484, (+45) 4593-4808 (fax).

† Division of Biology, California Institute of Technology, Pasadena, CA 91125, pfbaldi@cco.caltech.edu, (213) 222-6007, (213) 222-7742 (fax).

‡ Center for Biological Sequence Analysis, building 206, The Technical University of Denmark, DK-2800, Denmark, brunak@cbs.dtu.dk, (+45) 4525-2477, (+45) 4593-4808 (fax).

§ Net-ID, Inc., San Francisco, CA 94107, yves@netid.com, (415) 647-9402 (415) 642-9265 (fax).

Introduction

Initiation of transcription is the first step in gene expression, and constitutes an important point of control in prokaryotes as well as in eukaryotes (Reznikoff *et al.* 1985). Transcription initiates when RNA-polymerase recognizes and binds to certain DNA-sequences termed promoters. Subsequent to binding, a short stretch of the DNA double helix is disrupted, and the polymerase starts to synthesize RNA by the process of complementary basepairing. The sequence of the promoter determines the position of the transcriptional start point, and is furthermore important for the frequency with which the gene is transcribed (the strength of the promoter).

Escherichia coli Promoters

In the prokaryote *E.coli*, the form of the RNA-polymerase that is responsible for recognizing promoter sequences, has the protein subunit composition $\alpha_2\beta\beta'\sigma$. This so-called holo-enzyme can be divided into two functional components: the core enzyme ($\alpha_2\beta\beta'$, also designated E) and the sigma factor (σ). The sigma factor plays an important role in recognizing promoter sequences, and after successful initiation it is released from the holoenzyme (Gross & Lonetto 1992; Loewen & Hengge-Aronis 1994). Several different sigma factors exist, each recognizing a specific subset of promoters. These subsets have different nucleotide sequences. The biological significance of this is that each promoter group controls genes that are needed under physiologically similar conditions, and that therefore need to be expressed simultaneously. *E.g.*, all promoters recognized by the holo enzyme E σ^{32} control genes which are important for helping the bacterium survive prolonged exposure to higher-than-normal temperatures (heat shock). Sigma factors derive their names from the molecular weight of the proteins (thus, σ^{32} has a M_w of 32 kDa). *E.coli* is known to contain the sigma factors σ^{70} , σ^{54} , σ^{32} , σ^F , and σ^S . Briefly the genes controlled by the different factors are:

- σ^{70} Primary sigma, majority of all *E.coli* genes.
- σ^{54} Nitrogen assimilation.
- σ^{32} Heat shock response.
- σ^F Flagellum genes.
- σ^S Starvation stress response.

Comparison of *E.coli* σ^{70} promoters has led to the identification of three major conserved features: the “-10 box”, the “-35 box”, and a pyrimidine (C or T) followed by a purine (A or G) at the initiation site (Rosenberg & Court 1979; Hawley & McClure 1983). The -10 and -35 boxes are conserved hexanucleotide elements that are named according to the approximate position of their central nucleotides relative to the transcriptional start point. The well known consensus sequences are TTGACA for the -35 box, and TATAAT for the -10 box.

The sigma factors σ^{32} , σ^F , and σ^S are homologous to σ^{70} and all bind to promoters which have the same overall architecture (signals at -10 and -35), but which differ at one or both sites (Lonetto, Gribskov, & Gross 1992). However, σ^{54} is not homologous to the σ^{70} -family, and promoters recognized by *E* σ^{54} , have been found to contain two consensus boxes located at positions -12 and -24 (Merrick 1993; Morett & Segovia 1993).

Homo sapiens Promoters

Eukaryotes have three RNA-polymerases that are responsible for transcribing different subsets of genes: RNA-polI transcribes ribosomal RNA, RNA-polII (which we will focus on in this paper) transcribes mRNA, while RNA-polIII transcribes tRNA and other small RNAs. RNA-polII consists of more than 10 subunits, some of which are partly homologous to the bacterial subunits α , β , and β' . As with bacterial RNA-polymerase, the eukaryotic RNA-polII is dependent on additional factors for initiation. However, in the case of RNA-polII the factors are more numerous and play a larger role in the determination of the startpoint (Gill & Tjian 1992; Pugh & Tjian 1992; Eick, Wedel, & Heumann 1994). The factors that assist RNA-polII in initiating transcription, can be divided into three groups:

- The so-called **basal factors** are required for successful initiation at all promoters. Together with RNA-polII they form a complex surrounding the startpoint, and they determine the transcriptional startpoint. The basal factors include the so-called TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIH, and TFIJ, many of which are multi-subunit factors.

- **Upstream factors** are DNA-binding proteins that recognize short sequence elements upstream of the startpoint. They enhance the efficiency of initiation, presumably by protein-protein interactions with the basal transcriptional apparatus. Different promoters may contain different combinations of binding sites for these factors, in various distances from the startpoint.

- **Inducible factors** are synthesized or activated under certain conditions or at certain times, but otherwise work like the upstream factors. They are responsible for the control of transcription with regard to time and space.

Eukaryotic promoters are less similar to each other than bacterial promoters. Only two sequence elements are reasonably conserved with respect to composition and location: the TATA-box and the initiator element (Smale & Baltimore 1989; Guarente & Bermingham-McDonogh 1992; O'Shea-Greenfield & Smale 1992). The initiator is a sequence that is located at the startpoint in some promoters. It has the consensus Py₂CAPy₅, where Py is a pyrimidine (C or T). Most promoters have a TATA-box, which is a short sequence with the consensus TATAAAA usually centered approximately 25 bp upstream of the startpoint. The TATA-box plays a crucial role when RNA-polII recognizes TATA-box containing promoters: the transcriptional apparatus is assembled factor by factor, starting with the TFIID-subunit TBP (the TATA Binding Protein). TBP arrives at the promoter and binds to the TATA-box. During the subsequent steps, the transcriptional apparatus is assembled on the promoter by a process involving numerous protein-protein interactions. The so-called TAFs (TBP-associated factors) are important in this respect (Gill & Tjian 1992; Pugh & Tjian 1992). In addition to the two elements mentioned above, most promoters contain additional upstream elements which bind activating factors, but the sequence, number, orientation, and position of these is highly variable. However, there is an overall preference for G's and C's in promoter regions.

Purpose

In this paper we have analyzed prokaryotic and eukaryotic promoters with statistical methods, and by using powerful hidden Markov models (HMMs). HMMs are excellent for investigating conserved sequence signals that have variable spacing, and indeed we easily find the known sequence signals in *E.coli*-genes, and the most conserved sequences in human genes. We find that combinations of HMMs which have been trained on sequences belonging to specific sigma-subclasses,

are able to classify unknown sequence with great success. Furthermore, we have developed a novel method, involving selforganizing parallel HMMs, for automatic classification of the promoter sequences.

Data

Prokaryotic Promoters

The *E.coli* promoter sequences were taken from the compilation by Lisser and Margalit (Lisser & Margalit 1993). This database, which contains 300 sequences, is superior to most other available *E.coli* promoter databases on two accounts:

- Each sequence has been compared to the original paper, minimizing the chance of database entry errors.
- For each sequence, the assignment of transcriptional start point(s) has been verified with the relevant papers, and the most reliable have been chosen.

We processed the data in the following ways: first, we concatenated the sequences that are partially overlapping (*e.g.*, dnaK-P1 and dnaK-P2). This removed a number of contradictions, since the nucleotide that is marked as a transcriptional start point in one sequence is not labeled as such in the partially overlapping sequence, and *vice versa*. Concatenation resulted in a subset consisting of 248 sequences. Second, we discarded all the sequences that contain multiple start points, and all sequences not including at least 75 bp upstream and 25 bp downstream of the transcriptional startpoint. Then we cut out sequence surrounding the startpoint, so that all sequences contain exactly 75 bp upstream and 25 bp downstream of the transcriptional startpoint. The resulting set, which we use in this study, contains 166 sequences. For the purpose of training HMMs that were able to analyze un-annotated sequences with respect to sigma-class, we divided the data into three sets: those sequences known to be recognized by σ^{70} (38 sequences), those known to be recognized by σ^{54} (3 sequences), and those where it was not known which sigma-factor is responsible for transcription (the remaining 125 sequences).

Eukaryotic Promoters

The human data was extracted from Genbank release 90 (Benson *et al.* 1994). Specifically, all human sequences containing the feature key "prim_transcript", were selected. This feature key indicates that the sequence is an unprocessed transcript, and that it may therefore contain one or more transcriptional startpoints. From these sequences, those which had at least 250 bp upstream and 250 bp downstream of the first transcriptional startpoint were selected, and the 501

bp that symmetrically surrounded the startpoint were cut out, and kept for training. This resulted in a set of 340 sequences, of which 37 contained more than one transcriptional startpoint.

Methods

Measures of Information Content

The Kullback Leibler distance (or relative entropy) for each position in sequences aligned by the transcriptional startpoint, was calculated by the formula:

$$D(p, q) = \sum_i p_i \log_2 \frac{p_i}{q_i}$$

where p_i and q_i are the probabilities of occurrence for a particular nucleotide i (A, C, G, T) at the position (Kullback & Leibler 1951). Specifically, we took the probability distribution from one sigma-subset (*e.g.*, σ^{70}) and compared it to the distribution from another subset. $D(p, q)$ has values that range from 0 to ∞ . $D(p, q) = 0$ indicates that the two distributions are identical at the given position, while larger values of $D(p, q)$ means that the distributions differ at that position.

The traditional Shannon measure was also used and calculated by the formula:

$$I(p) = H_{max} - \sum_i p_i \log_2 p_i$$

where $H_{max} = \log_2(\text{length of alphabet}) = 2$, since the nucleotide alphabet contains 4 letters (Shannon 1948).

The results from both kinds of analysis were depicted by the use of sequence logos replacing the conventional numeric curves. The sequence logos were constructed according to Schneider and Stephens (Schneider & Stephens 1990). Briefly, sequence logos combine the information contained in consensus sequences with a quantitative measure of information, by representing each position in an alignment by a stack of letters. The height of the stack is a measure of the non-randomness at the position (here the Kullback Leibler distance or the Shannon measure), while the height of a letter corresponds to its frequency.

HMMs of Promoter Sequences

A first order discrete HMM can be viewed as a stochastic generative model defined by a set of states S , an alphabet A of m symbols, a probability transition matrix $T = (t_{ij})$, and a probability emission matrix $E = (e_{ix})$. The system randomly evolves from state to state, while emitting symbols from the alphabet. When the system is in a given state i , it has

a probability t_{ij} of moving to state j , and a probability e_{jX} of emitting symbol X . As in the application of HMMs to speech recognition, a family of DNA sequences can be seen as a set of different utterances of the same word, generated by a common underlying HMM. One of the standard HMM architectures for molecular biology applications, first introduced in (Krogh *et al.* 1994), is the left-right architecture. The alphabet has $m = 4$ symbols, one for each nucleotide ($m = 20$ for protein models, one symbol per amino acids). In addition to the start and end states, there are three classes of states: the main states, the delete states and the insert states with $S = \{start, m_1, \dots, m_N, i_1, \dots, i_{N+1}, d_1, \dots, d_N, end\}$. N is the length of the model, typically equal to the average length of the sequences in the family. The main and insert states always emit a nucleotide, whereas the delete states are mute. The linear sequence of state transitions $start \rightarrow m_1 \rightarrow m_2 \dots \rightarrow m_N \rightarrow end$ is the backbone of the model. For each main state, corresponding insert and delete states are needed to model insertions and deletions. The self-loop on the insert states allows for multiple insertions at a given site.

Given a sample of K training sequences O_1, \dots, O_k , the parameters of an HMM can be iteratively modified, in an unsupervised way, to optimize the data fit according to some measure, usually based on the likelihood of the data. Since the sequences can be considered as independent, the overall likelihood is equal to the product of the individual likelihoods. Two target functions, commonly used for training, are the negative log-likelihood:

$$Q = - \sum_{k=1}^K Q_k = - \sum_{k=1}^K \ln P(O_k) \quad (2.1)$$

and the negative log-likelihood based on the optimal paths:

$$Q = - \sum_{k=1}^K Q_k = - \sum_{k=1}^K \ln P(\pi(O_k)) \quad (2.2)$$

where $\pi(O)$ is the most likely HMM production path for sequence O . $\pi(O)$ can be computed efficiently by dynamic programming (the Viterbi algorithm). When priors on the parameters are included, one can also add regulariser terms to the objective functions for MAP (Maximum A Posteriori) estimation. Different algorithms are available for HMM training, including the Baum-Welch or EM (Expectation-Maximization) algorithm, and different forms of gradient descent and other GEM (Generalized EM) algorithms (Dempster, Laird, & Rubin 1977; Rabiner 1989; Baldi & Chauvin 1994b). Regardless of the training method, once an

HMM has been successfully trained on a family of sequences, it can be used in a number of different tasks. First, for any given sequence, one can compute its likelihood according to the model, and also its most likely path. A multiple alignment results immediately from aligning all the optimal paths of the sequences in the family. The model can also be used for discrimination tests, and data base searches (Krogh *et al.* 1994; Baldi & Chauvin 1994a), by comparing the likelihood of any sequence to the likelihoods of the sequences in the family. Finally the parameters of a model, such as the emission distributions of the backbone states and their entropies, can be used to detect consensus patterns and other signals (see (Baldi *et al.* 1995) for an example).

Another use of HMMs is in the classification of sequences within a family, and the discovery of sub-classes. This may be particularly relevant for promoter sequences, especially in eukaryotes, where a diverse array of sub-classes may exist, with different signals. Different classification algorithms can be considered depending on the amount of prior knowledge available. One basic approach (Krogh *et al.* 1994) is to use a “super-HMM”, consisting of several basic sub-HMMs in parallel, one for each sub-class, see Figure 1. The super-HMM can be trained using some form of competitive learning. For example, when a training sequence is presented to the super-HMM, its Viterbi path is first computed. Such a path goes through only one of the sub-HMMs, and only the parameters of this HMM are updated during training, to increase the likelihood of the corresponding sequence. Thus, these selforganizing HMMs have at the same time the ability to find clusters and the ability to model the sequential structure in the input data. Another more general approach to classification, based on hybrid HMM/NN (Neural Network) architectures, is briefly described in (Baldi & Chauvin 1995). In hybrid HMM/NN architectures, a NN is used to calculate and modulate the parameters of an HMM, so that a slightly different HMM is generated for each sub-class.

In practice, however, it is well known that selforganizing processes of this sort are far from trivial, especially in the absence of any prior knowledge, and can be plagued by local minima problems. The number of classes, their definition and separation, the degree to which each is represented in the available training set, are all crucial issues that impact the performance of the algorithms. As an example, in one experiment, we tried a super-HMM consisting of two similar sub-HMMs, initialized randomly, against the entire subset of *E.coli* promoter sequences with unique transcriptional starting point. Numerous training cycles always

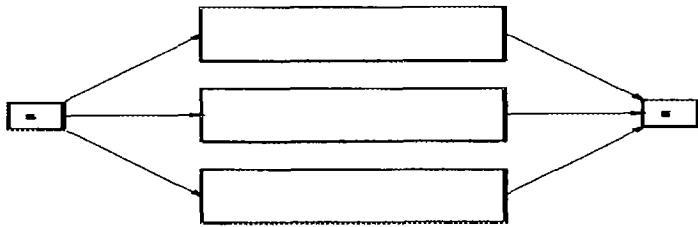


Figure 1: A “super-HMM”, consisting of several basic sub-HMMs in parallel.

resulted in a disappointing result: essentially all the sequences were classified as a single sub-class, associated with one of the two sub-HMMs. The underlying reason is as follows: with random initialisation, the initial sub-HMMs are far away, in sequence space, from the cloud of promoter sequences. The first training sequence selects whichever sub-HMM happens to be closer to the promoter cloud, and pulls it towards the cloud during the corresponding parameter update. This phenomenon is only repeated and reinforced by the presentation of the following training sequences, so that only one model is selected. Thus to produce a bifurcation between the sub-models one must introduce some additional elements in the algorithm. One possibility we have used is to take advantage of the prior knowledge gathered during the training of single HMM models, to initialise the sub-HMMs parameters close to the promoter cloud, instead of randomly. More generally, a bootstrap procedure of this sort can be used anytime to go from n to $n + 1$ classification.

Results: *Escherichia coli*

Statistical analysis

The Shannon information measure was calculated for three different subsets of the 166 sequences in the *E.coli* database: sequences known to be recognized by sigma-70, sequences known to be recognized by sigma-54, and the rest (Figure 2). Not surprisingly, no strong and clear picture emerges from the analysis of the three sequences in the σ^{54} -set. However, the well-known -10 box signal, and the CA-signal can be seen in the subsets recognized by σ^{70} , and the larger subset of un-annotated sequences (Figure 2). No

clear -35 signal can be seen in the σ^{70} -subset, and only a weak signal is visible in the larger subset of un-annotated sequences. This can probably be explained in part by the fact that the position of the -35 box, relative to the transcriptional start point, is somewhat flexible (Galas, Eggert, & Waterman 1985; Harley & Reynolds 1987). Consequently, the sequence signal will not be clearly recognized without multiple alignment.

In order to learn more about the differences between the sequences in the different subsets, we used the Kullback Leibler measure also in the analysis. Specifically, we calculated the Kullback Leibler distance between sequences belonging to the σ^{70} -subset and the entire set of sequences (data not shown). This analysis demonstrated that sequences from the σ^{70} -set were quite similar to the average sequence in the entire set. This is to be expected, since most of the sequences in the un-annotated set probably do belong to the σ^{70} -class. Nevertheless, a small overrepresentation of T's in the -10 box area was apparent, in good agreement with the fact that σ^{70} -sequences have the consensus TATAAT at this position. Furthermore, the σ^{70} -sequences displayed some under-representation of A's and T's in the area upstream of the -10 box.

HMMs

A hidden Markov model was trained on the entire set of 166 *E.coli*-sequences. The main state emission probabilities of the resulting model are shown in figure 3. As it can be seen, the HMM is very successful at modelling the known sequence signals: around -10 the TATAAT consensus is very clear. Further upstream the most

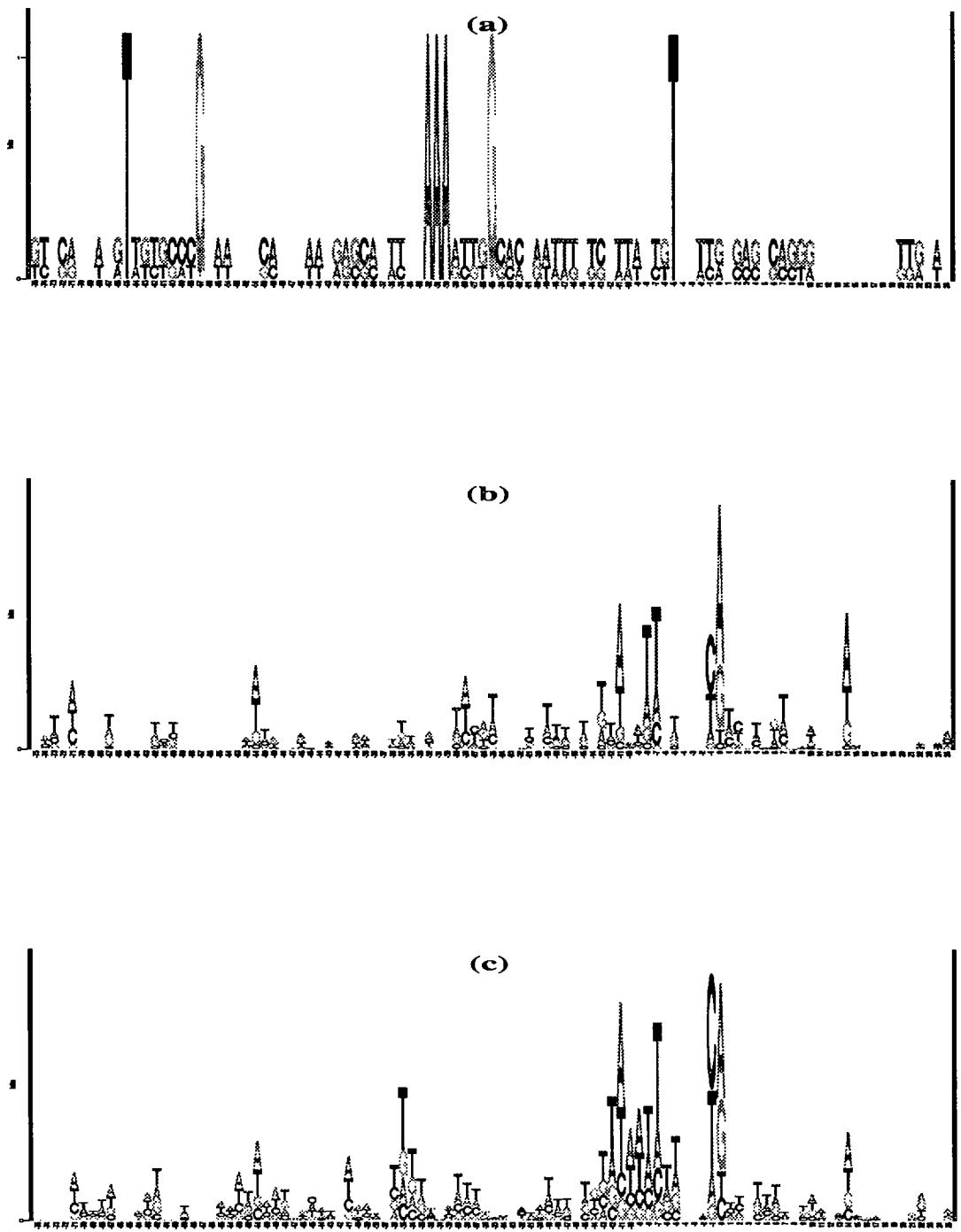


Figure 2: Shannon information content in the various subsets of the data depicted as sequence logos. a) Sequences known to be recognized by σ^{70} (3 sequences). b) sequences known to be recognized by σ^{70} (38 sequences). c) the remaining sequences in our *E. coli* database (125 sequences). The sequences are aligned by their transcriptional initiation site (position 0).

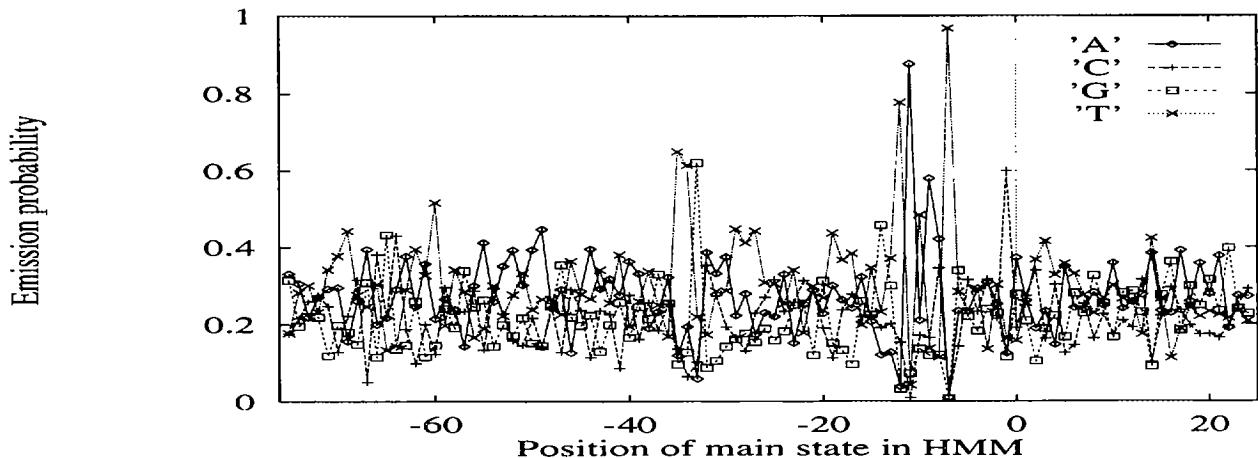


Figure 3: Emission probabilities of the main states in a hidden Markov model trained on the 166 *E.coli*-sequences. Notice the clear pattern of the well-known consensus sequences at -35 (TTG), -10 (TATAAT), and at the transcriptional startpoint (CA).

highly conserved part of the -35 box (TTG) can also be seen easily. Finally, there's a reasonably clear CA-signal at the transcriptional startpoint.

Thus it can be seen that the HMM is very good at handling the variably spaced *E.coli* promoter signals without a need for prior alignment, and certainly much better than the simple statistical methods employed above. This feature makes HMMs very strong tools for promoter analysis, since promoters are by nature modular.

The HMM was also trained on random sequences constructed by shuffling the nucleotides within each of the sequences in the dataset. In this way the nucleotide composition is the same, but the sequential structure of the sequences is different. When one calculates the negative log-likelihood of the sequences based on the model trained on the *E.coli*-data, it is found that the shuffled sequences have much higher values than the native sequences. Specifically, the average value of the negative log-likelihood of the random sequences was 144.0, while that of the *E.coli*-sequences was 137.6. This confirms that the *E.coli*-sequences have a lower Shannon information content than the randomly shuffled sequences (*i.e.*, they contain conserved signals).

A super HMM was constructed by combining two simple linear HMMs in parallel. One of the HMMs had been trained on the σ^{70} -sequences (38 sequences), while the other HMM had been trained on the three σ^{54} -sequences. The super-model was subsequently trained using the entire set of 166 sequences. During all cycles of training (including cycle 0, *i.e.* prior to training of the super-model), the classification induced on

the 166 sequences was completely constant: all the 166 sequences are classified as belonging to the first sub-HMM (the one trained on σ^{70}), with the exception of 4 sequences. These 4 sequences are: glnA-P2, glnH-P2, fdhF, and lacI. The first three of these sequences indeed belong to the σ^{54} class, while lacI belongs to the σ^{70} -class. The lacI promoter is associated with a weakly expressed gene (repressor for the lac-operon), and is known to be a 'non-consensus' σ^{70} representative. It is therefore not critical that the HMM doesn't characterize it as belonging to the σ^{70} -class. Furthermore, considering that the majority of *E.coli*-genes are transcribed by the holoenzyme $E\sigma^{70}$, it is a very reasonable result that most sequences end up being stably classified as belonging to the σ^{70} -class. In conclusion, the HMM is able convincingly to characterize unknown sequences based on a relatively small number of training sequences.

Results: *Homo sapiens* HMMs

A hidden Markov model was trained on the 340 *H.sapiens*-sequences (see Fig. 4 for the resulting main state entropies, and emission probabilities). Also in this case the HMM can be seen to successfully model the most well-conserved part of the eukaryotic promoters, *i.e.*, the TATA-box. As it can be observed from the emission probability profiles, there is a clearly visible overabundance of A's and T's in the -25 region. Interestingly, what appears to be an additional signal can be seen around position -200 . When analysing sequence alignments made using the trained HMM, a

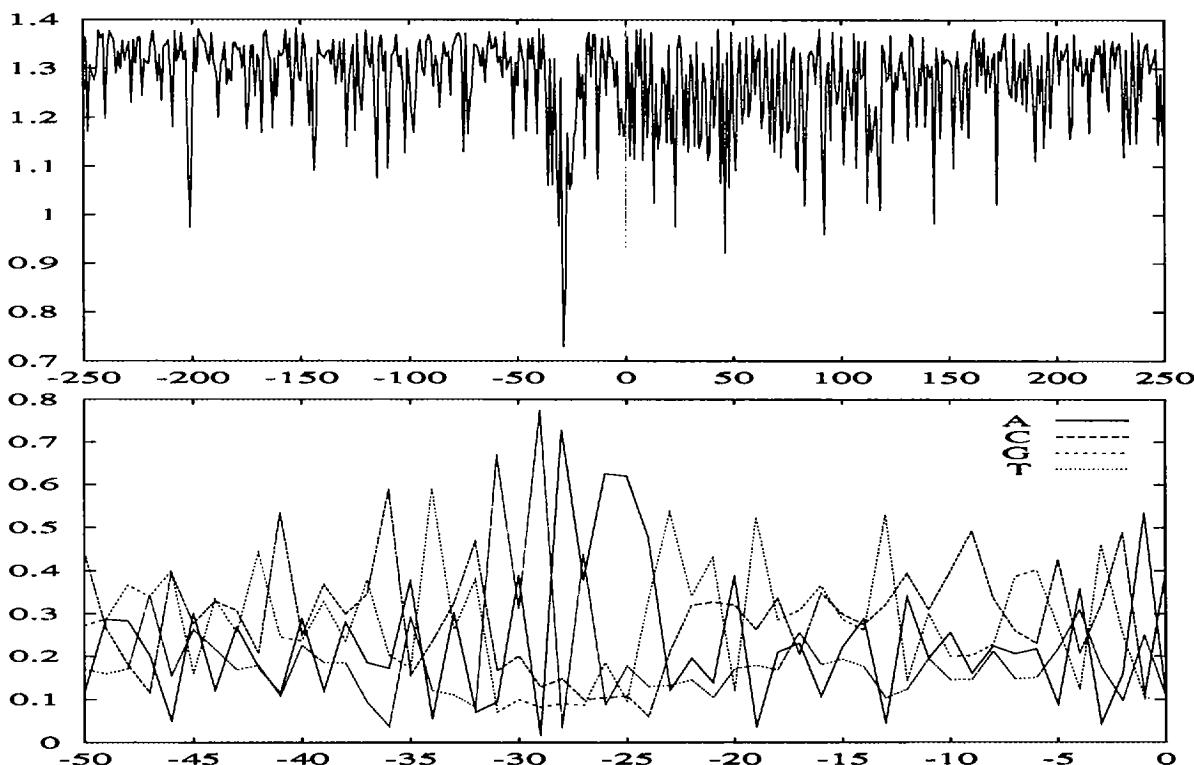


Figure 4: Entropies and emission probabilities of the main states in a hidden Markov model trained on the 340 human promoter sequences. The entropies and emission probabilities are shown as a function of the position on the HMM backbone (0=transcriptional startpoint). The HMM was trained on the 340 *H.sapiens*-sequences. Notice the TATA-box signal around -25. There is also a distinct low-entropy signal around -200, as well as several low-entropy peaks downstream of the start site.

clear pattern emerges from the nucleotides assigned by the mainstates in the -200 region. The conserved pattern covers four nucleotides, and has the consensus (C/T)(C/G)(T/A)(G/T). In this region the most abundant nucleotides are C and G, and it is the overabundance of A and T nucleotides that makes the pattern visible in the main state emission entropies and the sequence alignment. A common signal located this far upstream from the transcriptional startpoint is very interesting, further analysis of this potential enhancer/silencer-signal is clearly needed.

Discussion

We have shown that hidden Markov models are able to learn the sequential structure present in both prokaryotic and eukaryotic promoter sequences. They clearly enhance features which are being blurred by a rigid gap-free alignment of the sequences by the transcriptional start point.

We further introduce a new way of using the HMM technique for performing clustering experiments along with the need for modelling sequential structure. This is of importance in a large number of biosequence anal-

ysis situations, here exemplified by the investigation of promoter sequences known for their strong diversity related to the recognition by individual RNA-polymerase associated factors.

The results presented here need to be studied further, and several other experiments aimed at discovering sub-class structure in a larger eukaryotic dataset are in progress. We are currently extending the selforganizing bifurcation principle to more than two classes, and are also using these jointly with neural network learning techniques. We are developing hybrid HMM/NN architectures well suited for handling unbounded dependencies effectively equipped with larger windows around the transcriptional start point. These architectures may recognize correctly enhancer signals which are located at great distances from the well known local features of the minimal promoter.

Acknowledgements

AGP and SB are supported by a grant from the Danish National Research Foundation. The work of PB is supported by a grant from the ONR. The work of YC is

supported in part by grant number R43 LM05780 from the National Library of Medicine. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the National Library of Medicine.

References

- Baldi, P., and Chauvin, Y. 1994a. Hidden markov models of the G-protein-coupled receptor family. *Journal of Computational Biology* 1(4):311-335.
- Baldi, P., and Chauvin, Y. 1994b. Smooth on-line learning algorithms for hidden markov models. *Neural Computation* 6(2):305-316.
- Baldi, P., and Chauvin, Y. 1995. Protein modeling with hybrid hidden markov model/neural networks architectures. In *Proceedings of the 1995 Conference on Intelligent Systems for Molecular Biology (ISMB95)*, in Cambridge (UK). Menlo Park, CA: The AAI Press.
- Baldi, P.; Brunak, S.; Chauvin, Y.; Engelbrecht, J.; and Krogh, A. 1995. Periodic sequence patterns in human exons. In *Proceedings of the 1995 Conference on Intelligent Systems for Molecular Biology (ISMB95)*, in Cambridge (UK). Menlo Park, CA: The AAI Press.
- Benson, D.; Boguski, M.; Lipman, D.; and Ostell, J. 1994. Genbank. *Nucl. Acids Res.* 22:3441-3444.
- Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal Royal Statistical Society B* 39:1-22.
- Eick, D.; Wedel, A.; and Heumann, H. 1994. From initiation to elongation: Comparison of transcription by prokaryotic and eukaryotic RNA polymerases. *Trends in Gen.* 10:292-296.
- Galas, D. J.; Eggert, M.; and Waterman, M. S. 1985. Rigorous pattern-recognition methods for DNA sequences. analysis of promoter sequences from Escherichia coli. *J Mol Biol* 186:117-28.
- Gill, G., and Tjian, R. 1992. Eukaryotic coactivators associated with the TATA box binding protein. *Cur. Opin. Gen. Dev.* 2:236-242.
- Gross, C. A., and Lonetto, M. 1992. Bacterial sigma factors. In *Transcriptional regulation*. Cold Spring Harbor Laboratory Press.
- Guarente, L., and Bermingham-McDonogh, O. 1992. Conservation and evolution of transcriptional mechanisms in eukaryotes. *Trends in Gen.* 8:27-32.
- Harley, C. B., and Reynolds, R. P. 1987. Analysis of *E. coli* promoter sequences. *Nucleic Acids Res* 15:2343-61.
- Hawley, D. K., and McClure, W. R. 1983. Compilation and analysis of *Escherichia coli* promoter DNA sequences. *Nucleic Acids Res* 11:2237-55.
- Krogh, A.; Brown, M.; Mian, I. S.; Sjolander, K.; and Haussler, D. 1994. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology* 235:1501-1531.
- Kullback, S., and Leibler, R. A. 1951. On information and sufficiency. *Ann Math Stat* 22:79-86.
- Lisser, S., and Margalit, H. 1993. Compilation of *E. coli* mRNA promoter sequences. *Nucleic Acids Res* 21:1507-16.
- Loewen, P., and Hengge-Aronis, R. 1994. The role of the sigma factor σ^S (katf) in bacterial global regulation. *Annu. Rev. Microbiol.* 48:53-80.
- Lonetto, M.; Gribskov, M.; and Gross, C. 1992. The σ^{70} family: Sequence conservation and evolutionary relationships. *J. Bact.* 174:3843-3849.
- Merrick, M. 1993. In a class of its own — the RNA polymerase sigma factor σ^{54} (σ^N). *Mol. Microbiol.* 10:903-909.
- Morett, E., and Segovia, L. 1993. The σ^{54} bacterial enhancer-binding protein family: Mechanism of action and phylogenetic relationship of their functional domains. *J. Bact.* 175:6067-6074.
- O'Shea-Greenfield, A., and Smale, S. T. 1992. Roles of TATA and initiator elements in determining the start site location and direction of RNA polymerase II transcription. *J. Biol. Chem.* 267:1391-1402.
- Pugh, B., and Tjian, R. 1992. Diverse transcriptional functions of the multisubunit eukaryotic TFIID complex. *J. Biol. Chem.* 267:679-682.
- Rabiner, L. R. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2):257-286.
- Reznikoff, W. S.; Siegele, D. A.; Cowing, D. W.; and Gross, C. A. 1985. The regulation of transcription initiation in bacteria. *Annu Rev Genet* 19:355-87.
- Rosenberg, M., and Court, D. 1979. Regulatory sequences involved in the promotion and termination of RNA transcription. *Annu Rev Genet* 13:319-53.
- Schneider, T. D., and Stephens, R. M. 1990. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.* 18:6097-6100.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Tech. J.* 27:379–423, 623–656.

Smale, S. T., and Baltimore, D. 1989. The "initiator" as a transcription control element. *Cell* 57:103–13.