Periodic sequence patterns in human exons

Pierre Baldi[*]
Division of Biology, California Institute of Technology
Pasadena, CA 91125.
Tel: +1-818-3549038, fax:+1-818-3935013, email: pfbaldi@juliet.caltech.edu

Søren Brunak
Center for Biological Sequence Analysis, The Technical University of Denmark
DK-2800 Lyngby, Denmark.
Tel: +45-45252477, fax: +45-45934808, email: brunak@cbs.dtu.dk

Yves Chauvin[†]
Net-ID, Inc.
San Francisco, CA 94107.
Email: yves@netid.com

Jacob Engelbrecht
Center for Biological Sequence Analysis, The Technical University of Denmark
DK-2800 Lyngby, Denmark.
Tel: +45-45252477, fax: +45-45934808, email: engel@cbs.dtu.dk

Anders Krogh
NORDITA, Blegdamsvej 19
DK-2100 Copenhagen Ø, Denmark.
Tel: +45-35325503, fax:+45-31421016, email:krogh@norsci0.nordita.dk

Abstract

We analyse the sequential structure of human exons and their flanking introns by hidden Markov models. Together, models of donor site regions, acceptor site regions and flanked internal exons, show that exons — besides the reading frame — hold a specific periodic pattern. The pattern, which has the consensus: non-T(A/T)G and a minimal periodicity of roughly 10 nucleotides, is not a consequence of the nucleotide statistics in the three codon positions, nor of the well known nucleosome positioning signal. We discuss the relation between the pattern and other known sequence elements responsible for the intrinsic bending or curvature of DNA.

Keywords: DNA, sequential structure, periodicity, exon, intron, hidden Markov models.

[*] and Jet Propulsion Laboratory, Caltech.
[†] and Department of Psychology, Stanford University.

Introduction

Besides specifying the choice and order of amino acids in proteins genetic material hold a multitude of additional signals playing an important role in a variety of DNA transactions (Brendel et al. 1986; Trifonov 1989; Haran et al. 1994). Packaging, recombination and transcription of DNA are highly influenced by the bending and flexibility of the double helix (Drew and Travers 1985; Goodman and Nash 1989; Crothers and Steitz 1992). In turn these structural and functional properties of DNA change as a function of its base sequence.

The part of DNA represented by protein coding regions, or exons, belongs obviously to a class of sequence being highly constrained by the information capacity it needs to have. In contrast, non-coding regions or introns, especially in sequence parts at large linear distances from the splice sites, allow for a much higher degree of base variability or randomness. The sequential structure of coding and non-coding regions is of particular interest from a biological view point in revealing essential details necessary for understanding the assembly of the spliceosome and the splicing process in general.

In Proc. of Third Int. Conf. on Intelligent Systems for Molecular Biology
Eds. C. Rawlings et al., AAAI Press 1995, Pages 30-38

However, due to the need for reliably separating coding regions from non-coding regions in unannotated DNA generated by the large genome sequencing projects, such intrinsic features are also highly interesting from a computational view point. Gene parsing requires the statistical integration of several weak signals, some of which are poorly known, over length scales of at least several hundred nucleotides. In addition to consensus sequences at the splice sites, there seem to exist a number of other weak signals (Senapathy 1989; Engelbrecht et al. 1992) embedded in the 100 intron nucleotides upstream and downstream of an exon.

Due to the superposition of many signals in the same DNA sequence periodicities are hard to separate (Trifonov 1989). In particular, a specific *oscillatory* pattern easily detectable by one method, may be more or less invisible to others (Drew and Travers 1985). Two of the most well known periodic codes carried by DNA are the ribosome reading frame (Trifonov 1987, and below) and the chromatin code, which provides instructions on the proper placement of nucleosomes along the DNA molecule (Trifonov and Sussman 1980; Uberbacher et al. 1988; Muyldermans and Travers 1994).

In revealing these patterns a different alternative to conventional algorithms is the use of machine learning approaches. Adaptive algorithms are ideally suited for domains characterised by the presence of large amounts of data and the absence of a comprehensive underlying theory (Rumelhart et al. 1994). The fundamental idea behind adaptive algorithms is to learn the theory from the data, through a process of model fitting. Models can be selected from a number of different classes, such as Neural Networks (NNs), Hidden Markov Models (HMMs) or Stochastic Context Free Grammars (SCFGs). Such models are usually characterised by a large number of parameters.

Indeed, in recent years, the parsing problem has also been tackled using Neural Networks (Lapedes 1988; Brunak 1991; Uberbacher and Mural 1991; Xu et al. 1994) with encouraging results. Conventional neural networks typically use a fixed window size input, and perhaps are not ideally suited to handle the sort of elastic deformations introduced by evolutionary tinkering in genetic sequences. Another trend in recent years, has been the casting of DNA and protein sequences problems in terms of formal languages using stochastic context free grammars (Searls 1992; Sakakibara et al. 1993), probabilistic automata and HMMs (see also Churchill 1989). HMMs in particular have been used to model protein families and address a number of task such as multiple alignments, classification and database searches (Baldi et al. 1993 and 1994a-d; Haussler et al. 1993; Krogh et al. 1994a; and Baldi and Chauvin 1994a). It is the success obtained with this method on protein sequences, and the ease with which it can handle insertions and deletions, that naturally suggests its application to human genes.

Thus, the main thrust of this effort is towards the development and application of HMMs and other related adaptive techniques for modeling and parsing human genes and splice sites, and specifically for the detection of new statistical regularities. In Krogh et al. (1994b), HMMs are applied to the problem of detecting coding/non-coding regions in bacterial DNA (E. coli), which is characterized by the absence of true introns (like other prokaryotes). Their approach leads to an HMM that integrates both genic and intergenic regions, and can be used to locate genes fairly reliably. A similar approach for human DNA, that is not based on HMMs, but uses dynamic programming and neural networks to combine various gene finding techniques, is described in (Snyder and Stormo 1993). Here, we focus on detecting novel features of human exons by HMMs.

In this paper we report on a new periodic pattern found in human exons. The pattern, which is descibed in statistical terms, has an average period of about 10, and features a fairly strong consensus pattern [ˆ T][AT]G[1]. This pattern was found using several different types of HMMs, and it was checked that the pattern is not commensurate with a period of three, i.e., it seems not to stem from the exon reading frame. Because the period is close to the period of the double helix in its B-form we suggest that it may be related to structural properties of the DNA.

In section 2, we briefly review HMMs, and the learning algorithms and models used in the experiments. In section 3, we describe our main results using, in particular, HMMs to model acceptor sites, donor sites, exons, flanked exons and introns. The main new result is the detection of particular periodic patterns, with a period of roughly 10 nucleotides that may have significant biological implications. The results and their potential implications are discussed in section 4.

Materials and methods

Hidden Markov models

HMMs are a class of statistical models that have been used in a number of applications, especially speech recognition (Levinson et al. 1983; Rabiner 1989) but also for other problems, such as single ion channel recordings (Ball and Rice 1992).

A first order discrete HMM is completely defined by a set of states **S**, an alphabet of m symbols, a probability transition matrix $\mathbf{T} = (t_{ij})$, and a probability emission matrix $\mathbf{E} = (e_{iX})$. The model is intended to describe a stochastic system that evolves from state to state, while randomly 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,

[1] In the language of regular expressions [ˆ T] means 'non T' and [AT] means 'A or T'. That is, [ˆ T][AT]G means a string of 3 characters, the first is A, G, or C, the second is A or T and the third is always G.

and a probability e_{iX} of emitting symbol X. The model is called hidden because what is observed is the output string of symbols from the system and one of the goals is to gather information about the hidden set of transitions that may have led to its production.

As in the application of HMMs to speech recognition, a family of related primary sequences can be seen as a set of different utterances of the same word, generated by a common underlying HMM with a left-right architecture, i.e. once the system leaves a given state it can never return to it. An example of a standard architecture used in some of our experiments can be seen in Fig. 1. For the corresponding alphabets, $m = 4$ in the case of DNA or RNA sequences, one symbol per nucleotide, and $m = 20$ in the case of proteins sequences, one symbol per amino acid. Common knowledge about evolutionary mechanisms suggests to introduce three classes of states (in addition to the start and end states): the main states, the delete states and the insert states with **S** $= \{start, m_1, ..., m_N, i_1, ..., i_{N+1}, d_1, ..., d_N, end\}$, N is the length of the model. Usually, it is set equal to the average length of the sequences in the family being modeled. Alternatively, N can be iteratively adjusted during learning, as in (Krogh et al. 1994a). Prior to any learning, the transition and emission parameters of a model can be initialized uniformly, at random or according to any other desirable distribution. The main and insert states always emit a letter of the alphabet, whereas the delete states are mute. The linear sequence of state transitions $start \rightarrow m_1 \rightarrow m_2 .... \rightarrow m_N \rightarrow end$ we call the backbone of the model. Corresponding to each main state, insert and delete states are needed to model insertions and deletions, with respect to the backbone. Self loops on the insert states allow for multiple insertions. Architectural variations — including more complex loop structures — are possible and may be taylored to particular problems when additional information is available, see (Baldi et al. 1994d).

The most important aspect of HMMs is that they are adaptive: given a set of training sequences, the parameters of a model can be iteratively modified to optimize the fit of the model to the data according to some measure, usually the product of the likelihoods of the sequences. Different algorithms are available for HMM training, such as the classical Baum-Welch algorithm (Baum 1972; Rabiner 1989), which is a special case of the more general EM algorithm in statistical estimation (Dempster et al. 1977), and different forms of gradient descent and their approximations (for instance Baldi and Chauvin 1994a). In order to avoid over-fitting, the models can be regularized. In this work we used the method introduced in (Krogh et al. 1994b), which is derived from a Dirichlet prior distribution.

Data sets

To train HMMs on human DNA we prepared several data sets of training and testing sequences from GenBank, release 81.0. The aim was to make a large unique set of internal exons. Entries were excluded if: (1) the Feature Table was missing, (2) the ORIGIN Label was missing, (3) the CDS Feature Key was missing, (4) the CDS Feature Key did contain a complement operator, (5) the CDS Feature Key had no operator and no intron Feature Key (assumed to be cDNA), (6) they had alternative splicing, (7) the CDS Feature Keys had overlapping, multiple reading frames. From the remaining set of entries the internal exons only were kept in the set. Exons with no information about acceptor and donor sites were also not included.

The main data set contains 2,019 non-redundant human internal exon sequences and their flanking regions. From this basic set, we extracted different training sets for pure exons, in open or closed reading frame, as well as for flanked exons, or flanked splice sites. On a pure exon experiments, for instance, a typical training set typically contains 500 exon sequences (for the patterns reported we did not notice any important differences with larger training sets). The bulk of the data set contained exons with a length from 100 to 200 nucleotides; most of the experiments were done using exons from this subset only. For full statistical detail on the data set, see (Baldi et al. 1994d).

Results

Unlike the case of protein families, it is essential to remark that, all exons are not directly, nor closely related by evolution. However, they still form a “family” in the sense of sharing certain general characteristics. For instance, in a multiple alignment of a set of flanked exons, the consensus sequences associated with the splice sites should stand out as highly conserved regions in the model, exactly like a protein motif in the case of a protein family. As a result, insertions and deletions in the HMM model should be interpreted here in terms of formal operations on the strings rather than evolutionary events. The main point is to apply a novel technique (HMMs) to an old problem, and see whether any new patterns emerge.

Below we first briefly report results from experiments where HMM’s have been trained on *splice sites*, either as paired sites linked by exons, or separate acceptor and donor sites flanked by intron or exon nucleotides. Secondly, we report results from a large number of experiments on *exons*, where these have been mixed or in one particular reading frame only.

Acceptor and donor sites linked by exons

To see whether an HMM would pick up easily known features of human acceptor and donor sites, a model with the architecture of Fig. 1, was trained on 500 randomly selected

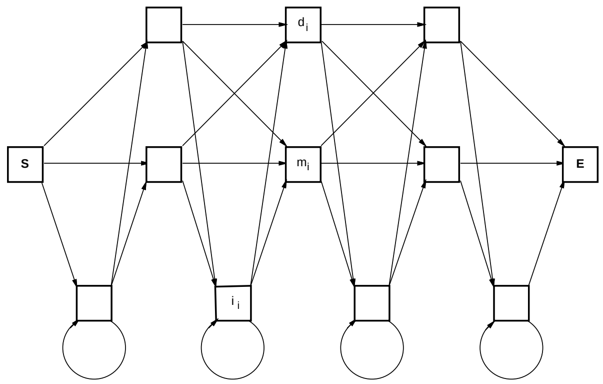

Figure 1: Typical left-right HMM architecture used in the current experiments. Notice that all states have fan-in 3 and fan-out 3. N is the length of the model, usually equal to average length of the sequences being modeled.

flanked internal exons, with the length of the exons restricted to being between 100 and 200 nucleotides only.

The probability of emitting each one of the four nucleotides, across the main states of the model, is plotted in Fig. 2. We see striking periodic patterns, especially present in the exon region, characterized by a minimal period of 10 nucleotides, with A and G in phase, and C and T in antiphase. Additional interesting patterns can be detected by close inspection of the parameters of the model.

By close inspection of the parameters of an HMM trained specifically on flanked acceptor sites we observed that the model learns the acceptor consensus sequence perfectly: ([TC]...[TC][N][CT][A][G][G]). The pyrimidine tract is clearly visible, as were a number of other known weak signals such as a branching (lariat) signal with a high A, in the 3' end of the intron.

Similarly, the donor sites are also clearly visible in a model trained on flanked donor sites, but much harder to learn than the acceptor sites. The consensus sequence of the donor site is learnt perfectly: ([CA][A][G][G][T][AG][A][G]), as was the G-rich region (Engelbrecht et al. 1992), extending roughly 75 bases downstream from the donor site. The fact that the acceptor site is easier to learn is most likely explained by the more extended nature of acceptor site regions as opposed to donor sites. However, it could also result from the fact that exons in the training sequences are always flanked by *exactly* 100 nucleotides upstream. To test this hypothesis, we trained a similar model (Fig. 2) using the same sequences, but in *reverse* order. Surprisingly, the model still learns the acceptor site (which is now downstream from the acceptor site) much better than the donor site. The periodic pattern in the reversed exon region is still present. The periods we observe could also be an artifact of the method: for instance, when presented with random training sequences, periodic HMM solutions could appear naturally as local optima of the training procedure. To test this hypothesis, we trained a model using random sequences of similar average composition as the exons and found no distinct oscillatory patterns in the emission distribution. We also tested that our database of exons does not correspond prevalently to the 3.6 amino acid period found in α-helical domains of proteins. This was done simply by computing from the reading frame assignments the amino acid composition and comparing it to the ranking of the helix forming potential of the twenty amino acids (Creighton 1993).

In summary, after a number of initial experiments, the main results were that: (1) donor sites are harder to learn than acceptor sites; (2) there seem to be some kind of statistical periodicity, at least in the exon regions, with a period of about 10 nucleotides. In the following, we shall try to elucidate (2), by training several architectures, either with off-line Baum-Welch with initialization favoring the backbone, or on-line gradient descent with uniform initialization and backbone regularization. In all cases we have tested, the two training algorithms have given very similar results. To test the periodic patterns, we also use tied and loop architectures, as discussed in the section on methods.

Exons

The HMMs were trained using a set of non-redundant internal exon sequences, typically 500, without any flanking nucleotides. To avoid any effects due to very short or very long exons, all exons had again length between 100 and 200 nucleotides. The average length (and therefore the length of the models) was typically 142 or 143. The experiments were repeated using several randomly selected sets without any change in the observed patterns in the emission probabilities.

A periodic pattern in the parameters of the models of the form [AT][CG], (or [AT]G) with a periodicity of roughly 10 base pairs, could be seen at positions: 10, 19, 28, 37, 46,

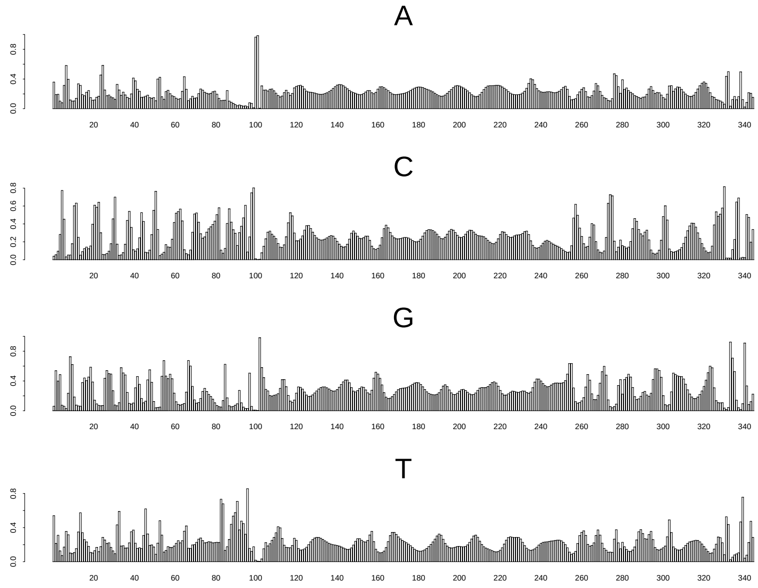

Figure 2: Emission distribution from main states of an HMM model trained on 500 flanked internal exons. The length of the training exons this time is constrained to be between 100 and 200 nucleotides, with average of 142, and fixed intron flanking of 100 on each side. The model was not fully regularized, with no bias favoring the main states backbone path. The donor site is not as clear as the acceptor site. Notice the oscillatory pattern in the exon region, and outside.

55, 72, 81, 90, 99, 105, 114, 123, 132, 141. Notice that this pattern is detected in the weights of the model, and not directly in the sequences themselves. There is also an apparent TGCA diagonal signal, starting at position 7, which emerges quite consistently across different experiments.

The emission profile of the backbone was compared to the cumulative distributions of two nucleotides jointly (data not shown). The plots of A+G and C+T are considerably smoother than those of A+T and C+G both in the intron and the exon side. The 10 periodicity is visible both in the smooth phase/antiphase pattern of A+G and C+T, and in the sharp contrast of high A+T followed by high C+G. There is also a rough 3 base pair periodicity, especially visible in C+G, where every third emission corresponds to a local minimum. This is consistent with the reading frame features of human genes (Trifonov 1987), which are strong especially on the third codon position (≈30% C and ≈26% G).

One possibility is to look for possible reading frame effects on the patterns we observe. Therefore we also trained models using 500 exons with identical reading frame. The exon length was again filtered in the [100,200] interval. The average length was 143. So a model of length 143 was trained as above. Interestingly, we obtain very similar results including the TGCA signal (this time starting at position 8) and 10 periodicity. Therefore the models do not seem to be affected by reading frame effects.

To further test our findings, we trained a "tied" exon model with a hard-wired periodicity of 10, see (Baldi et al. 1994b). The tied model consists of 14 identical segments of length 10, and 5 additional positions in the beginning and end of the model, making a total length of 150. During training the segments are kept identical by *tying* of the parameters, i.e. the parameters are constrained to be exactly the same throughout learning, as in the weight sharing procedure for neural networks. The model was trained on 800 internal exon sequences of length between 100 and 200, and it was tested on 262 different sequences. The parameters of the repeated segment after training, are shown in Fig. 3. Emission probabilities are represented by horizontal bars of corresponding proportional length. There is a lot of structure in this segment. The most prominent feature is the regular expression [^T][AT]G at position 12–14. The same pattern was often found at positions with very low entropy in the "standard models" described above. In order to test the significance, the tied model was compared to a standard model of the same length. By comparing the average negative log-likelihood they both assign to the exon sequences and to random sequences of similar composition, it was clear that the tied model achieves a level of performance comparable to the standard model, but with significantly less free parameters. Therefore a period of around 10 in the exons seems to be a strong hypothesis.

However, the type of left-right architecture we have used

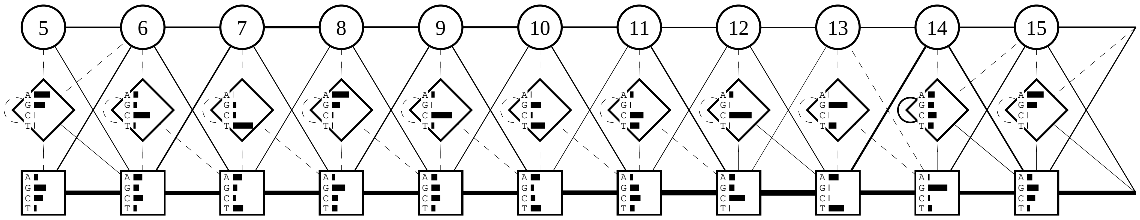

Figure 3: The repeated segment of the tied model. Rectangles represent main state and circles represent delete states. Histograms represent emission distributions from main and insert states. Thickness of connections is proportional to corresponding transition distribution. Note that position 15 is identical to position 5.

is not the ideal model of an exon, because of the large length variations. It would be desirable to have a model with a loop structure such that the segment can be entered as many times as necessary for any given exon, see (Krogh et al. 1994b) for a loop structure used for E. coli DNA.

So we finally trained a different sort of loop models, using a data set of 500 exons. The model was a "wheel" model of length 10, without flanking, without any distinction between main and insert states, and without delete states. Thus there are no problems associated with potential silent loops. Sequences can enter the wheel at any point. The point of entry can of course be determined by dynamic programming. The structure of the model obtained after training with the EM algorithm is shown in Fig. 4. The thickness of the arrows from "outside" represents the probability of starting from the corresponding state. Remarkably, the emission parameters in the wheel have a structure very similar to those found in the repeated segment of the tied model. In particular the pattern `[^T][AT]G` is clearly recognizable.

One obvious question one can ask about the 10 periodicity is how likely is it to arise by pure chance? This question itself is not well defined because the periodicity itself is not well defined. Suppose, for the sake of the argument, that we observe something like [AT][GC] every 10 base pair or so, that is with a positional variability of +1 or -1. Suppose also for simplicity that each nucleotide occurs with probability 0.25. If currently we observe [AT] as a starting point of the pattern, there is a 0.5 chance of immediately seeing a [GC] right after. There is a 0.25 chance of seeing the pattern [AT][GC] 9 positions downstream, a 0.25 chance of seeing it 10 positions downstream, and a 0.25 chance of seeing it 11 positions downstream, in a randomly generated sequence. So the total chance of observing a first period, knowing that the starting point is a [AT], is 0.5×0.75. Similarly the chance of seeing n such "periods" is 0.5×0.75^n. In the case of a typical exon $n \approx 13$ or so. This gives a probability of observing the oscillatory pattern in random uniform sequences of approximately $0.5 \times 0.75^{13} \approx 0.01$. Even if we allow for 10 possible different starting positions, we get a probability of 0.1. In other words, the pattern would occur in at most one in ten training sequences, and there is no reason why the HMM should pick it up (given that in a random sequences there are many other "periodic" patterns with the same likelihood, such as the reverse pattern [GC][AT]). These probabilities become even smaller if we use any skewed nucleotide distribution, such as the one found in real exons.

Introns, intragenic regions and other experiments

A number of other experiments have been tried which are not reported here at the present time. For instance, it is known that surviving isolated insertions and deletions in exons are very rare, since they entirely disrupt the local reading frame. Accordingly we have trained architectures where insertions and deletions could only occur, while respecting the triplet reading frame structure. The result are consistent with the ones reported here. Likewise several alignment experiments have been considered.

As far as the periodic pattern of period 10, it is natural to wonder whether it is confined to exons or exons with their immediate flanking, or also in the middle of large intron regions and in intragenic regions. We are in the process of constructing data sets to check these possibilities. A preliminary experiment was run, starting with a data base of introns with length at least 800. From these sequences, we removed 400 base pair on each side, to remove any proximity effects due to splice sites. We were left with 447 "deep" intron sequences, of length greater than 100. 69 deep intron sequences had length above 200. 400 sequences were selected at random and further cut to match the length distribution of the exon data base to avoid possible length effects. Finally, an HMM as in Fig. 1 was trained by gradient descent with regularisation. No oscillations, or other particular patterns, seem to be clearly present. After 6 training cycles, the cumulative probabilities A+G and C+T are smooth, as for exons, but so is also A+T. Overall these curves are less smooth than in the exons and their proximal flanking. After 12 training cycles, all smoothness seem to disappear.

Using the wheel model to estimate the average negative log-likelihood per nucleotide we obtain the values given in Table 1. The figures are computed specifically for various types of sequence, different types of exons, introns and

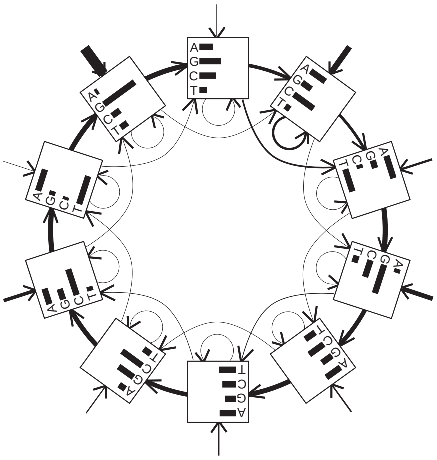

Figure 4: Wheel model trained on 500 exons of length between 100 and 200. Thickness of external arrows show the probability of starting in the corresponding state. Emission probabilities are represented by bars inside boxes.

intragenic regions. They strongly indicate that the above described periodic pattern belongs to exons, rather than non-coding deep intron sequence.

Sequence type	Negative log-likelihood
Exons	1.355150
Last coding exon	1.351415
First coding exon	1.357349
First exon coding/non-coding	1.361875
Last exon coding/non-coding	1.374327
Introns	1.397193
Intragenic regions	1.400396
Deep introns	1.402820
Randomized exons	1.402836

Table 1: Average negative log-likelihood per nucleotide in the wheel model. Non-coding exon is transcribed, but not translated.

Discussion

With HMMs we have been able to rapidly recognize the well known pattern and statistics related to exons and splice sites. Examples include the splice site consensus sequences, or the 3 periodicity inside exons. In addition, we are able to detect a new pattern which is a sort of periodicity, with a period of roughly 10. Our experiments indicate that this periodicity exists in the exons, but possibly also in the immediate flanking regions, but not in the deep introns. The period 10 signal is stronger than the 3 periodicity in the sense that models constrained to period 9 are harder to train.

The pattern is best seen in the weights of the model, and is also associated with the smoothness of the cumulative distributions of purines `A+G` (in phase) and pyrimidines `C+T` (in antiphase). Plots of `A+T` and `C+G` are much more jagged, with a greater tendency towards 3 periodicity. Exon regions seem to be characterised also by larger oscillation amplitudes than the immediately adjacent intron regions. Such patterns would be very difficult to detect with other methods, in part because of exon length variability.

All the tests we have conducted so far, have led to results that are consistent with these patterns. In particular, testing the 10 periodicity has forced us to expand the HMM method, for instance by developing new architectures. These may be useful for other problems also, where periodic effects are important.

It is intriguing that the new periodicity we observe is closely related to the periodicity of the DNA helix. If confirmed, the periodic patterns could have significant biological and algorithmic implications. They could be related to the superimposition of several signals, and/or to the way DNA bends and wraps around the histone octamer.

Eukaryotic DNA sequence patterns for nucleosome positioning have previously been investigated in detail, see for example (Klug and Lutter 1981; Zhurkin 1983; Drew and Travers 1985; Uberbacher et al. 1988; Trifonov 1989; Heran et al. 1994). These sequence patterns have many different features, the most predominant being runs af adenine (A-tracts) (and/or thymine), which allow the DNA axis to bend. The periodic pattern reported in this study, non-T(A/T)G, has absolutely no homopolymeric features. The bending properties of sequence with this kind of periodicity could be estimated theoretically, or even better be measured by observing its electrophoretic migration in gels (Heran et al. 1994), or directly by cryo-electron microscopy (Dubochet et al. 1994). If this pattern is specific for exons and the flanking intron sequence, and not, as our preliminary experiments indicate, of introns in general, it could make nucleosomes wrapped by coding regions differ from nucleosomes wrapped by non-coding sequence. If this is true

— a purely speculative proposition — genes could make themselves known to the transcription machinery on a scale different from the size of the promoter complex. It is known that under normal physiological conditions, DNA within the nucleous is packaged into a compact fiber about 30 nm in diameter, which is a poor substrate for initiation and chain elongation by the RNA polymerase (Clark et al. 1993).

Acknowledgements. We thank C. Kesmir and K. Rapacki for competent programming. This work (SB and JE) was supported by the Danish Natural Science Research Council and the Danish National Research Foundation. The work of PB was supported by grants from the ONR, the AFOSR and a Lew Allen Award at JPL. The work of YC was 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., Chauvin, Y., Hunkapiller, T. and McClure, M. A. 1993. Hidden Markov Models in Molecular Biology: New Algorithms and Applications. Advances in Neural Information Processing Systems 5:747-754, Morgan Kaufmann Pub.

Baldi, P., Chauvin, Y., Hunkapiller, T. and McClure, M. A. 1994. Hidden Markov Models of Biological Primary Sequence Information. PNAS USA, 91:1059-1063.

Baldi, P. and Chauvin, Y. 1994a. Smooth On-Line Learning Algorithms for Hidden Markov Models. Neural Comp., 6:305-316.

Baldi, P., Brunak, S., Chauvin, Y., Engelbrecht, J. and Krogh, A. 1994b. Hidden Markov Models of Human Genes. Advances in Neural Information Processing Systems 6:761-768, Morgan Kaufmann Pub.

Baldi, P. and Chauvin, Y. 1994c. Hidden Markov Models of the G-Protein Coupled Receptor Family. J. Comp. Biol., 1/4 in press.

Baldi, P., Brunak, S., Chauvin, Y., Engelbrecht, J. and Krogh, A. 1994d. Hidden Markov Models of Human Genes. CalTech Technical Report. Division of Biology, Caltech.

Ball, F. G. and Rice, J. A. 1992. Stochastic Models for Ion Channels: Introduction and Bibliography. Mathematical Bioscience.

Baum, L. E. 1972. An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes. Inequalities, 3:1-8.

Brendel, V., Beckmann, J.S. and Trifonov, E.N. 1986. Linguistics of Nucleotide Sequences: Morphology and Comparison of Vocabularies, J. Mol. Struct. Dyn. 4:11-21.

Brunak, S., Engelbrecht, J. and Knudsen, S. 1991. Prediction of Human mRNA Donor and Acceptor Sites from the DNA Sequence. J. Mol. Biol., 220:49-65.

Churchill, G. A. 1989. Stochastic Models for Heterogeneous DNA Sequences. Bull. Math. Biol., 51:79-94.

Clark, D., Reitman, M., Studitsky, V. and Chung, J. 1993. Chromatin Structure of Transcriptionally Active Genes, in Cold Spring Harbor Symp. Quant. Biol. 58:1-6.

Creighton, T.E. 1993. Proteins, Structures and Molecular Properties, W.H. Freeman, New York.

Crothers, D.M. and Steitz, T.A. in *Transcriptional Regulation eds. McKnight, S.L. and Yamamoto, K.R.*, 501-534 Cold Spring Harbor Laboratory Press, New York, 1992.

Dempster, A. P., Laird, N. M. and Rubin, D. B. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. J. Roy. Stat. Soc., B39:1-22.

Drew, H.R. and Travers, A.A. 1985. DNA Bending and its Relation to Nucleosome Positioning, J. Mol. Biol. 186:773-790.

Dubochet, J., Bednar, J., Furrer, P. Stasiak, A.Z., Stasiak, A. and Bolshoy, A.A. 1994. Determination of the DNA helical repreat by cryo-electron microscopy. Nature Struct. Biol. 1:361-363.

Engelbrecht, J., Knudsen, S. and Brunak S., 1992. G/C rich tract in 5' end of human introns, J. Mol. Biol., 227:108-113.

Goodman, S.D. and Nash, H.A. 1989. Nature, 341:251-254.

Haran, T.E., Kahn, J.D. and Crothers, D.M. 1994. Sequence Elements Responsible for DNA Curvature, J. Mol. Biol. 244:135-143.

Haussler, D., Krogh, A., Mian, I.S. and Sjölander, K. 1993. Protein Modeling using Hidden Markov Models: Analysis of Globins, Proceedings of the Hawaii International Conference on System Sciences, **1**, IEEE Computer Society Press, Los Alamitos, CA, 792-802.

Klug, A. and Lutter, L.C. 1981. The helical periodicity of DNA on the nucleosome, Nuc. Acids. Res. 9:4267-4283.

Krogh, A., Brown, M., Mian, I. S., Sjölander, K. and Haussler, D. 1994a. Hidden Markov Models in Computational Biology: Applications to Protein Modeling. J. Mol. Biol. 235:1501–1531.

Krogh, A., Mian, I. S. and Haussler, D. 1994b. A Hidden Markov Model that Finds Genes in *E. coli* DNA, Nuc. Acids Res., 22:4768–4778.

Lapedes, A., Barnes, C., Burks, C., Farber, R. and Sirotkin, K. Application of Neural Networks and Other Machine Learning Algorithms to DNA Sequence Analysis. In G.I. Bell and T.G. Marr, editors. The Proceedings of the Interface Between Computation Science and Nucleic Acid Sequencing Workshop. Proceedings of the Santa Fe Institute, volume VII, pages 157–182. Addison Wesley, Redwood City, CA, 1988.

Levinson, S. E., Rabiner, L. R. and Sondhi, M. M. 1983. . An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition. The Bell Syst. Tech. J., 62:1035-1074.

Muyldermans, S. and Travers, A.A. 1994. DNA Sequence Organization in Chromatosomes, J. Mol. Biol., 235:855-870.

Rabiner, L. R. 1989. A Tutorial on Hidden Markor Models and Selected Applications in Speech Recognition. Proc. IEEE, 77:257-286.

Rumelhart, D. E., Durbin, R., Golden, R. and Chauvin, Y. 1994. Back-propagation: the Theory. In: Back-propagation: Theory, Architectures and Applications. Y.E. Chauvin and D.E. Rumelhart Editors, Chapter 1, Lawrence Erlbaum Associates, in press.

Sakakibara, Y., Brown, M., Underwood, R.C., Mian, S.I. and Haussler, D. 1993. Stochastic Context-Free Grammars for Modeling RNA. Technical Report UCSC-CRL-93-16, University of California, Santa Cruz.

Searls, D. B. 1992. The Linguistics of DNA. American Scientist, 80:579-591.

Senapathy, P., Shapiro, M.B., and Harris, N.L. 1990. Splice Junctions, Branch Point Sites, and Exons: Sequence Statistics, Identification and Applications to Genome Project. Patterns in Nucleic Acid Sequences, Academic Press, 252–278.

Snyder, E.E. and Stormo, G.D. 1993. Identification of Coding Regions in Genomic DNA Sequences: an Application of Dynamic Programming and Neural Networks. Nuc. Acids Res., 21:607–613.

Trifonov, E.N. and Sussman, J.L. 1980. The pitch of chromatin DNA is reflected in its nucleotide sequence, PNAS USA 77:3816-3820.

Trifonov, E.N. 1987. Translation framing code and frame–monitoring mechanism as suggested by the analysis of mRNA and 16S rRNA nucleotide sequences, J. Mol. Biol., 194:643-652.

Trifonov, E.N. 1989. The Multiple Codes of Nucleotide Sequences, Bull. Math. Biol. 51:417-432.

Uberbacher, E. C., Harp, J.M. and Bunnick, G. J. 1988. DNA Sequence Patterns in Precisely Positioned Nucleosomes, J. Mol. Struct. Dyn. 6:105-120.

Uberbacher, E. C. and Mural, R. J. 1991. Locating Protein-Coding Regions in Human DNA Sequences by a Multiple Sensor-Neural Network Approach. PNAS USA, 88:11261-11265.

Xu, Y., Einstein, J. R., Mural, R. J., Shah, M. and Uberbacher, E. C. 1994. An Improved System for Exon Recognition and Gene Modeling in Human DNA Sequences. Proceedings of Second International Conference on Intelligent Systems for Molecular Biology Stanford University. , R. Altman and D. Brutlag and P. Karp and R. Lathrop and D. Searls Editors, AAAI Press, 376-383.

Zhurkin, V.B. 1983. Specific alignment of nucleosomes on DNA correlates with periodic distribution of purine-pyrimidine and pyrimidine-purine dimers, FEBS Lett. 158:293-297.