

COMMUNICATION

Naturally Occurring Nucleosome Positioning Signals in Human Exons and Introns

Pierre Baldi¹, Søren Brunak^{2*}, Yves Chauvin³ and Anders Krogh⁴

¹*Division of Biology
California Institute of
Technology, Pasadena
CA 91125, USA*

²*Center for Biological
Sequence Analysis
The Technical University of
Denmark, DK-2800 Lyngby
Denmark*

³*Net-ID, Inc., 601 Minnesota
St., San Francisco
CA 94107, USA*

⁴*The Sanger Centre, Hinxton
Cambridge CB10 1RQ, UK*

We describe the structural implications of a periodic pattern found in human exons and introns by hidden Markov models. We show that exons (besides the reading frame) have a specific sequential structure in the form of a pattern with triplet consensus non-T(A/T)G, and a minimal periodicity of roughly ten nucleotides. The periodic pattern is also present in intron sequences, although the strength per nucleotide is weaker. Using two independent profile methods based on triplet bendability parameters from DNase I experiments and nucleosome positioning data, we show that the pattern in multiple alignments of internal exon and intron sequences corresponds to a periodic “in phase” bending potential towards the major groove of the DNA. The nucleosome positioning data show that the consensus triplets (and their complements) have a preference for locations on a bent double helix where the major groove faces inward and is compressed. The in-phase triplets are located adjacent to GCC/GGC triplets known to have the strongest bias in their positioning on the nucleosome. Analysis of mRNA sequences encoding proteins with known tertiary structure exclude the possibility that the pattern is a consequence of the previously well-known periodicity caused by the encoding of alpha-helices in proteins. Finally, we discuss the relation between the bending potential of coding and non-coding regions and its impact on the translational positioning of nucleosomes and the recognition of genes by the transcriptional machinery.

© 1996 Academic Press Limited

*Corresponding author

Keywords: nucleosome; bendability; periodicity; exon; intron

Besides specifying the choice and order of amino acids in proteins, genetic material holds 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 & Travers, 1985; Goodman & Nash, 1989; Crothers & Steitz, 1992; Sinden, 1994; Elgin, 1995). In turn, these structural and functional properties of DNA change as a function of its base sequence.

DNA associated with protein coding regions, or exons, is highly constrained by the information it must carry. In contrast, non-coding regions or introns allow for a much higher degree of base variability or randomness, especially at large linear

distances from the splice sites. The distinctive characteristics of exon and intron sequences are of particular interest for understanding the processing of pre-mRNA in the cell nucleus (Lamond, 1995). Such intrinsic features are also highly interesting from a computational view point, due to the need for reliably separating coding from non-coding regions, in unannotated DNA, generated by the large genome sequencing projects.

Here, we report on the structural implications of a new periodic pattern found in human exons and introns by novel machine learning approaches (Rumelhart *et al.*, 1994). The pattern has an average period of about ten nucleotides, and features a fairly strong consensus sequence [^T][AT]G.

A large set of human genes was extracted from GenBank (rel. 81), and subdivided and separated into classes according to their function as coding or non-coding regions, for statistical detail see Baldi *et al.* (1995). From this basic data set, different

Abbreviations used: HMM, hidden Markov model; DSSP, Dictionary of Secondary Structures of Proteins.

training sets of flanked splice sites, pure exon sequence starting in specific or non-specific reading frame position, flanked exons or intron sequence, were extracted. On a pure exon experiment, for instance, a training set typically contained 500 exon sequences (for the patterns reported below we did not notice any important differences with larger training sets).

The periodicity was first observed when investigating the sequential structure in the sequence context of donor and acceptor splice sites. When training several hidden Markov models (Levinson *et al.*, 1983; Rabiner, 1989; Baldi *et al.*, 1994; Krogh *et al.*, 1994, 1995; Baldi & Chauvin, 1994) on donor and acceptor splice sites flanked by 100 nucleotides upstream and downstream, we observed in the emission probabilities (Baldi *et al.*, 1995) that the hidden Markov models did learn all the known features of human splice sites perfectly: the donor site consensus ([CA][A][G][G][T][AG][A][G]), the acceptor site consensus including the polypyrimidine tract ([TC]...[TC][N][CT][A][G][G]), the guanine-rich region in the 5' intron end (Nussinov, 1989; Engelbrecht *et al.*, 1992), and even other weak signals such as a branch point signal for the lariat with a strong A, in the 3' end of the intron (Lukashin *et al.*, 1992). However, most striking was the presence of extended periodic patterns, especially on the exon side of the splice sites, characterized by a minimal period of ten nucleotides, with A and G in-phase, and C and T in anti-phase (data not shown). When adding the probabilities, the combined value for C and G clearly showed a period of three associated with the triplet reading frame. In human genes the third codon position is biased towards these particular nucleotides (Trifonov, 1987).

In order to characterize the periodicity further, a wide range of different hidden Markov model architectures (Baldi *et al.*, 1995) were trained on non-flanked internal exons, in order to separate features from the special gradients in the nucleotide composition known to be present in initial and terminal exons (Engelbrecht *et al.*, 1992). When training on the bulk of the internal exons in the length interval between 100 and 200 nucleotides (internal exons have an average length of ≈ 150 nucleotides), a clear and consistent pattern emerged in the emission probabilities, no matter which architecture was applied. The architectural variation included conventional left-right HMM models, left-right models with identical segments "tied" together, and circular "wheel" models with better ability to reveal a periodicity in the presence of noise.

Figure 1(a) displays a wheel model architecture (in this case, ten nucleotides in length), where sequences can enter the wheel at any point. The thickness of the arrows from "outside" represents the probability of starting from the corresponding state. After training the emission parameters in the wheel model did show a periodic pattern

[[^]T][AT]G in a clearly recognizable form in states 8, 9 and 10.

By comparing the cumulative negative log-likelihood of the training set on wheel architectures with different numbers of states, we found that wheel models of length ten nucleotides yield the best fit. Implicitly, this is also confirmed by the fact that the skip probabilities are not strong in this model (Figure 1(a)). In other words, if the data were 9-periodic, a wheel model with a loop of length 10 should be able to fit the data, by heavy use of the possibility of skipping a state in the wheel. State repeating in a 9-state wheel is non-equivalent to state skipping in a 10-state wheel. These wheel models do not contain independent insert states (as the linear left-right HMM architectures do). A repeat of the same state does not give the same freedom in terms of likelihood as it would if independent inserts were allowed. Moreover, in analogy to gap penalties in conventional multiple alignments, the HMM training procedure uses a regularization term favoring main states over skip states.

All the experiments were repeated using several subsets of exons starting in one of the three codon positions in the reading frame, without any significant change in the observed patterns in the emission probabilities.

For comparison, Figure 1(b) shows the emission probabilities from a 9-state wheel model trained on the coding part of complete mRNA sequences of concatenated exons. This model clearly recognizes the triplet reading frame (see Figure legend).

Figure 2 shows a similar training of the HMM wheel architecture on large sets of complete intron sequences. Again a 10-periodic pattern with consensus [[^]T][AT]G appears in three subsequent states, now in states 7, 8 and 9. The fact that the pattern is present in intron sequences also provides additional evidence against a reading frame-associated origin for the pattern in the exons. Below we return to the question of the relative strength of the pattern in exons and introns.

It is well known that "bent DNA" requires a number of small individual bends that are in-phase (Sinden, 1994). Only when bends are phased at ≈ 10.5 bp (corresponding to one full turn of the double helix), can stable long-range curvature be obtained. It is therefore natural to assess profiles of the bending potential of exons and introns in relation to the triplet consensus signal described here.

Most of the experimental work on quantitative aspects of DNA bending has been reported as (nearest neighbor) dinucleotide propensities. Recently, parameters for trinucleotide sequence-dependent bendability deduced from DNase I digestion data have appeared (Brukner *et al.*, 1995a). DNase I interacts with the surface of the minor groove, and bends the DNA molecule away from the enzyme. The experiments (Brukner *et al.*, 1995b) therefore quantitatively reveal bendability parameters on a scale where low values indicate no

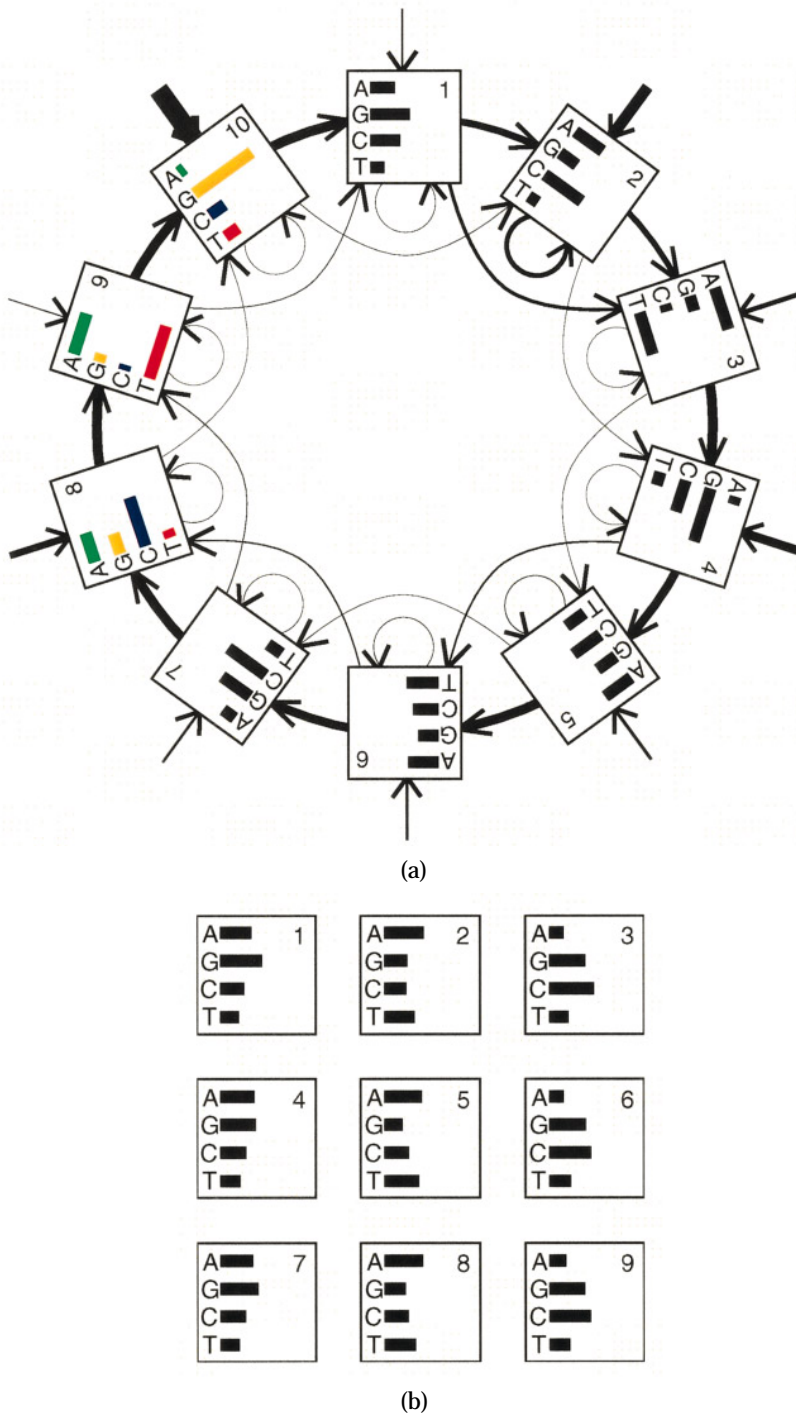


Figure 1. (a) A 10-state wheel hidden Markov model trained on 500 internal exons of length between 100 and 200 nucleotides. The HMM is completely defined by the set of ten states, the alphabet of four symbols, a probability transition matrix $T = (t_{ij})$, and a probability emission matrix $E = (e_{ix})$. 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 , and a probability e_{ix} of emitting symbol X . Transitions and emission probabilities can be iteratively modified to optimize the fit of the model to the data according to a given measure, usually the product of the likelihoods of the sequences. The model is called hidden because what is observed is the output string of symbols from the system. One of the goals is to gather information about the hidden set of transitions that may have led to its production.

In the wheel architecture the thickness of the external arrows shows the probability of starting in the corresponding state. Emission probabilities are represented by bars inside boxes. In this highly flexible unbiased model without any distinction between main and insert states, sequences can enter the wheel at any point. States may be skipped and repeated as well. The point of entry can be determined by dynamic programming. The pattern [^T][AT]G is clearly recognizable in states 8, 9 and 10, while non-perfect alignment and interference with the reading frame causes features of the pattern to appear in states 2, 3 and 3 as well. We also used a loop architecture (Baldi *et al.*, 1995), containing true insert and delete states, and obtained equivalent results. (b) The emission probabilities from a 9-state wheel model trained on complete mRNA sequences

shown without the skip and loop arrows, which were very thin in this case. The 3-periodic reading frame pattern is clearly visible, with higher frequencies of A and G, A and T, and C and G, on the first, second and third codon positions, respectively.

bending potential, and high values correspond to large bending or bendability towards the major groove, for the 32 double-stranded triplets: AAA/ATT, AAA/TTT, CCA/TGG, and so on (see Table 1).

Nucleotide, dinucleotide and trinucleotide statistics in exon and intron sequences differ markedly (Fickett, 1982; Brunak *et al.*, 1991; Fickett & Tung, 1992). Using the DNase I bendability parameters,

the overall bending potential of sequences from these functionally distinct classes can be compared easily, by multiplying the triplet frequencies and the bending values. Interestingly, the average value for human intron sequence, -0.0178 , is very low, compared to the average for human exon sequence, -0.0092 , indicating that non-coding human DNA (on the average) may be more inflexible. Not surprisingly, the corresponding value for sequences

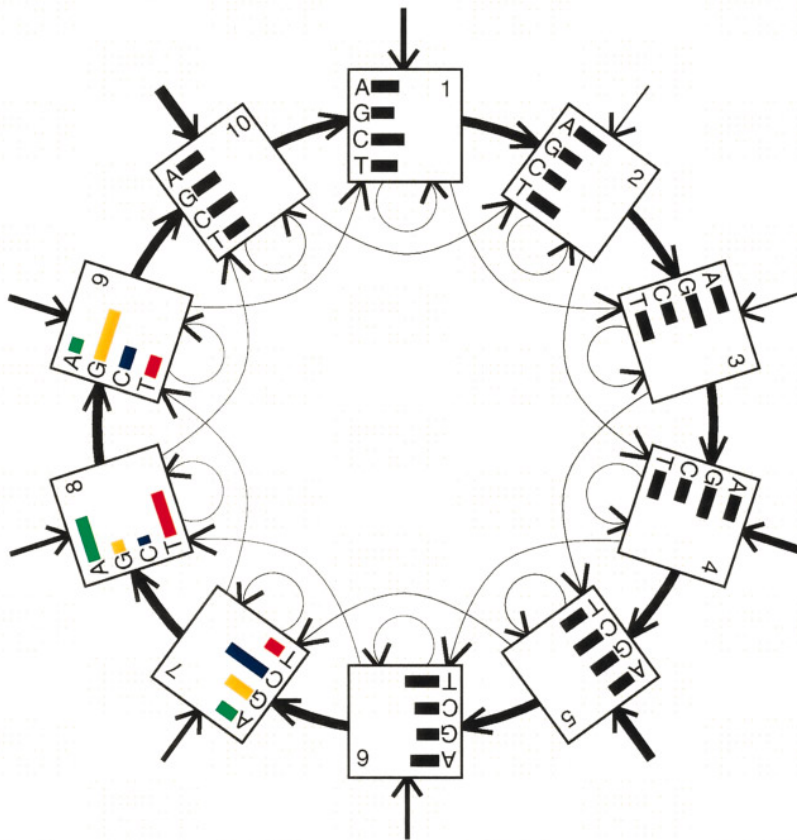


Figure 2. A 10-state wheel hidden Markov model trained on 2000 human introns; 25 nucleotides were removed at the 5' and 3' ends in order to avoid effects of the conserved sequence patterns at the splice sites. The pattern [[^]T][AT]G is clearly recognizable in states 7, 8 and 9.

with a uniform distribution of triplets, -0.0184 , is closer to the value for introns.

However, unless the positioning of the relevant sequence patterns is in-phase, these overall values do not have any long-range bending implications. The periodic pattern found [[^]T][AT]G covers six triplets: AAG (-0.081), GAG (0.031), GTG (0.040), ATG (0.134), CAG (0.175), CTG (0.175). The bendability parameters indicated for each triplet are quite striking: they belong almost exclusively to the high end of the bendability spectrum, and none of them is very negative. This means that the in-phase triplets have a relatively high bending potential towards the major groove.

We have computed the bendability profile for an alignment of the entire set of exons and, for comparison, a set of aligned deep intron segments, and for random sequences as well. Unlike the case of protein families (Baldi *et al.*, 1994; Baldi & Chauvin, 1994; Krogh *et al.*, 1994, 1995), it is essential to state that arbitrary exons are not directly, or closely, related by evolution. However, they may still form a "family" in the sense of sharing certain general characteristics (Trifonov, 1987).

We have used the wheel model (Figure 1) and dynamic programming to produce multiple alignments of the internal exons and, for comparison, intron segments and random sequences as well. For each nucleotide in a particular sequence, we list the state number in the wheel architecture (from 1 to 10) associated with its production, thus aligning the

sequential structures in the sequences to each other. From the alignments, a bendability profile displaying the sequential pattern in the bendability of the sequences can be computed from the triplet frequencies and the associated states.

The bendability profiles of the alignments made by this procedure of "rolling" the wheel over the actual sequences are shown in Figure 3. They show clearly that the in-phase triplets in coding regions (and less strongly in non-coding regions) correspond to a periodic bending potential, with a period of roughly 10. The other states are associated with strong negative values, or values close to zero.

The consensus sequence for the periodic pattern does not describe in full detail the underlying distribution of the in-phase triplets. When the actual distribution of exon triplets (with the middle nucleotide assigned to state 9 by the wheel in Figure 1(a)) was sorted according to frequency, it appeared that all the six triplets covered by the consensus had high frequencies (6 to 16%), but in addition complementary partners for four of the triplets did follow just below. Three of the four partner triplets have overabundant frequencies around 3 to 4% (CTC, CTT, CAC), while one (CAT) has a level of around 2% (Table 1). Except for CAT, the nine other triplets fill the top nine positions in the frequency list. Two of the six consensus triplets (CTG/CAG) are complementary already, meaning that the bending-associated periodic pattern centered at state 9, is created by the frequent occurrence of five complementary triplet pairs.

Table 1. Table with bendability parameters from the DNase I experiments (Brukner *et al.*, 1995a), fractional variation of occurrence on the outside (or inside) of the DNA wound around nucleosomes (Satchwell *et al.*, 1986), and roll angles, for the 32 triplet pairs

	DNase I	% Out	Roll	Exon	Intron
AAT/ATT	-0.280	-30	0.7	1.63/2.14	2.68/5.07
AAA/TTT	-0.274	-36	0.0	0.83/1.13	1.99/3.68
CCA/TGG	-0.246	8	5.4	0.05/0.07	0.10/0.06
AAC/GTT	-0.205	-6	3.7	1.73/1.06	1.58/1.53
ACT/AGT	-0.183	11	5.8	0.07/0.11	0.04/0.21
CCG/CGG	-0.136	2	4.7	1.02/2.09	0.60/0.86
ATC/GAT	-0.110	7	5.3	2.58/0.80	2.01/0.62
AAG/CTT	-0.081	6	5.1	8.50/3.30	7.13/4.81
CGC/GCC	-0.077	25	7.5	0.38/0.38	0.22/0.18
AGG/CCT	-0.057	8	5.4	0.73/0.37	1.46/0.24
GAA/TTC	-0.037	-12	2.9	0.15/1.09	0.27/0.74
ACG/CGT	-0.033	8	5.4	0.33/0.27	0.16/0.07
ACC/GGT	-0.032	8	5.4	0.12/0.08	0.10/0.01
GAC/GTC	-0.013	8	5.4	1.03/0.97	0.51/0.83
CCC/GGG	-0.012	13	6.0	0.53/1.16	0.27/0.93
ACA/TGT	-0.006	6	5.1	0.00/0.00	0.01/0.00
CGA/TCG	-0.003	31	8.3	0.08/0.05	0.00/0.09
GGA/TCC	0.013	-5	3.8	0.04/0.00	0.00/0.01
CAA/TTG	0.015	-9	3.3	1.14/2.75	1.14/4.02
AGC/GCT	0.017	25	7.5	0.01/0.10	0.01/0.01
GTA/TAC	0.025	-6	3.7	0.30/0.68	0.44/0.48
AGA/TCT	0.027	-9	3.3	0.00/0.00	0.00/0.00
CTC/GAG	0.031	8	5.4	3.93/6.02	5.29/5.22
CAC/GTG	0.040	17	6.5	2.91/6.73	2.59/5.58
TAA/TTA	0.068	-20	1.9	0.05/0.02	0.22/0.28
GCA/TGC	0.076	13	6.0	0.00/0.00	0.00/0.00
CTA/TAG	0.090	-18	2.2	1.13/0.48	1.60/2.11
GCC/GGC	0.107	45	10.0	0.08/0.17	0.03/0.15
ATG/CAT	0.134	18	6.6	7.87/1.88	5.97/2.08
CAG/CTG	0.175	-2	4.2	11.40/16.30	9.25/11.83
ATA/TAT	0.182	-13	2.8	0.47/0.41	1.72/0.62
TCA/TGA	0.194	8	5.4	0.00/0.00	0.00/0.00

The roll angles are computed from the percentages using the simple formula: $\text{roll} = 10^\circ \times (\% + 36) / (45 + 36)$, see Goodsell & Dickenson (1994). The roll angles are normalized using an arbitrary maximum roll of 10° for the GGC/GCC triplet pair. The two last columns list the frequencies of triplets where the middle nucleotide is assigned to state 9 in exons and introns, respectively. The periodic triplets and their complements account for close to 70% and 60% of the triplets in state 9.

In the wheel alignment of the introns the five triplet pairs also account for most of the periodic pattern, yet the distribution is not as skew as in the case of the exons. Of the exon and intron triplets with the middle nucleotide in state 9, 68.8% and 59.8%, respectively, are among the five complementary triplets. In the case of random assignments these percentages would equal $1/64 \times 10 = 15.625\%$. This means, for example, that at approximately every ten exon nucleotides we will find one of the triplets in seven out of ten cases. Another way of quantifying the relative strength of the periodicity in exons and introns is to compute from the probabilities p_i of the four nucleotides assigned to state 9 and the probabilities at the two adjacent positions in the actual sequences, the Shannon information content: $-\sum_{i=1}^4 p_i \log_2 p_i$. The accumulated deviation from the random case (corresponding to two bits per position) amounts to 1.43 and 1.19 bits in exons and introns, respectively.

Our experiments indicate that the periodicity is

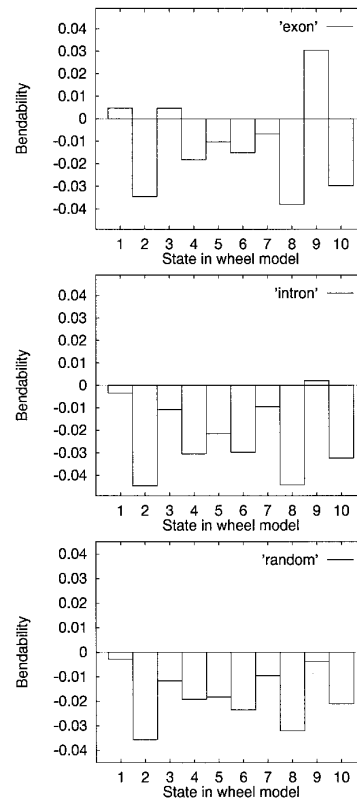


Figure 3. Bendability profiles of alignments of internal exons, deep intron segments and random sequences, made by a hidden Markov model trained on internal exons from human genes (Figure 1). The bendability potential of internal exons shows a distinct periodicity of 10, which is weaker in introns. No states have positive values in the alignments of 1000 randomly generated sequences of 200 nucleotides in length. The period in the alignments (average distance between state 9 nucleotides) is of the order of 10.1 to 10.2 nucleotides.

strongest in exons, and possibly also in the immediate flanking intron sequence, but on the average somewhat weaker in arbitrarily selected deep intron segments (Van Wye *et al.*, 1991). In none of the experiments using simple linear left-right HMM architectures did we detect clear regular oscillations in the non-coding sequence. Using the wheel model to estimate the average negative log-likelihood per nucleotide, we also computed values specifically for various types of sequence, different types of exons, introns and intragenic regions. The ranking of these also strongly indicates that the above described periodic pattern is strongest in exons. The period in the alignments (average distance between state 9 nucleotides) is of the order of 10.1 to 10.2 nucleotides.

Alpha-helices in proteins have 3.6 amino acids per turn corresponding to 10.8 nucleotides in the DNA (Zhurkin, 1981), not far from the period in the internal exons described above. Almost all internal exons represent protein-encoding regions (exons in 3' and 5' untranslated regions are primarily of the initial or terminal type), and it is relevant to ask

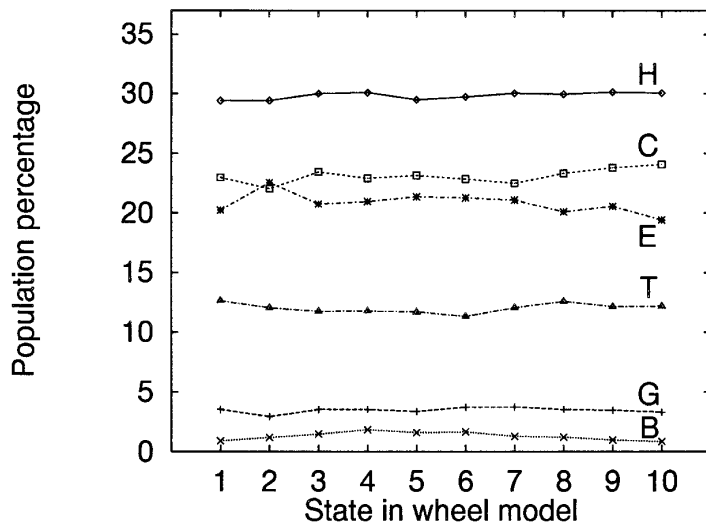


Figure 4. The distribution over the ten states of actual DSSP protein secondary structure, in alignments of human (mature) mRNA sequences coding for proteins with known three-dimensional structure. In the DSSP nomenclature, H corresponds to the conventional alpha-helix, C to random coil, E to beta-strand, B to beta-bridge, T to turn, and G to the 3-10 helix. The structural coordinates for the proteins were extracted from the Brookhaven Protein Data Bank, with the requirement that the resolution be better than 2.5 Å; the corresponding mRNA nucleotide sequences were extracted from GenBank. For details on the data extraction criteria, see Brunak & Engelbrecht (1996).

whether the different hidden Markov models are aligning regions encoding alpha-helices, instead of a DNA-related pattern. However, whereas alpha-helices only rarely exceed 15 amino acids in length (45 nucleotides), the periodic emission pattern in the left-right linear HMM architectures is clearly visible over an interval covering 140 nucleotides or so (Baldi *et al.*, 1995). It is unlikely that the HMM delete and insert states are used to produce the patterns from alignment of the structural features in the encoded proteins.

Using the wheel model to make alignments of mRNA sequences encoding 34 non-homologous human proteins with known three-dimensional structure (Brunak & Engelbrecht, 1996), we have shown directly that the periodic pattern is not related to the secondary structure of the associated proteins. Figure 3 displays the distribution of the DSSP protein secondary structure (Kabsch & Sander, 1983) over the ten states, nucleotides encoding alpha-helical residues do not cluster in states 8, 9 and 10, which represent the periodic pattern. In fact, the coil and turn categories are the only ones having a small increase in these states.

Eukaryotic DNA sequence patterns for rotational nucleosome positioning have been investigated in detail (see, for example: Trifonov & Sussman, 1980; Klug & Lutter, 1981; Zhurkin, 1983; Drew & Travers, 1985; Satchwell *et al.*, 1986; Trifonov, 1989; Pehrson, 1989; Haran *et al.*, 1994; Muylldermans & Travers, 1994; Bolshoy, 1995). From estimates of the fractional variation of occurrence of trinucleotides on the outside or inside of DNA wound around nucleosomes (Satchwell *et al.*, 1986; Goodsell & Dickerson, 1994), it follows that our in-phase triplets (and their complementary partners) have a tendency to appear on the outside, not on the inside (Table 1). This means that the triplets, in relation to nucleosome positioning, have a preference for locations on the bent double helix, where the major groove faces inward and is compressed (positive roll), completely in line with the bendability

profiles from the DNase I digestion data. From a frequency sorted list of exon triplets associated with states 10, 1 and 2 in the wheel architecture (see also Figure 1(a)), it appears that the in-phase triplets (states 8, 9 and 10) are located adjacent to highly frequent GCC/GGC triplets (states 10, 1 and 2) known to have the strongest preference in their positioning on the nucleosome (Satchwell *et al.*, 1986; Goodsell & Dickerson, 1994) at locations where the double helix has its major groove on the concave side of the curved helix.

Using the program BEND (Goodsell & Dickerson, 1994), where the triplet positioning preferences have been converted into roll angles, implementing a method for calculating the magnitude of the local bending, we observed that this completely independent method also gave a phased bending profile for the exon alignments made by the hidden Markov model (Figure 1(a)). When the average of the local bends over the states was computed exactly as before, we got large values for two consecutive states, 10 and 1, and lower values in between. This means that the nucleosome positioning data also assign high values to transition points defined by the location of the periodic triplets. In both cases we find a periodic pattern in the tendency of the major groove compression (Brukner *et al.*, 1995a). The combined consensus signal for the nucleosome positioning signal reads [⁺T][AT]G[CG]C. In structural terms this sequence is a combination of a quite flexible part (the in-phase triplets), and a more rigid component with a preferred conformation displaying substantial hard-to-distort major groove compression. This could reflect different requirements for torsional compensation (which could be better accommodated by flexible sequences) and bending in a superhelical configuration.

The removal of introns in eukaryotic genes is known to affect the rate of transcription in negative direction, often quite drastically. If the pattern reported in this study is stronger in exons, and

weaker (as our work indicates) in introns in general, it could make nucleosomes wrapped by coding regions differ from nucleosomes wrapped by non-coding sequence (Widom, 1996). The translational positioning of the nucleosomes could be one such difference (Kornberg & Lorch, 1992). It is striking that the average length of internal exons, approximately 150 nucleotides, is almost equal to the length of the 1.75 turns of DNA wrapped around the nucleosome, 145-146 nucleotides. While prokaryotes lack both introns and the nucleosomal organization of the DNA, the structure of chromatin is highly conserved and unique for eukaryotic genomes. Under normal physiological conditions, DNA within the nucleus is coiled into a compact fiber about 30 nm in diameter, which is known to be a poor substrate for initiation and chain elongation by RNA polymerase (Clark *et al.*, 1993). If a difference in wrapping or positioning exists, genes could make themselves known to the transcriptional machinery on a scale different from the size of the promoter complex.

Acknowledgements

We thank an anonymous referee and Dr A. A. Travers for critical comments and suggestions about the manuscript. The work of P.B. was supported by a grant from the ONR. S.B. was supported by the Danish National Research Foundation. A.K. was supported by the Wellcome Trust. The work of Y.C. 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. (1994). Hidden Markov models of the G-Protein coupled receptor family. *J. Comp. Biol.* **1**, 311-335.
- Baldi, P., Chauvin, Y., Hunkapiller, T. & McClure, M. A. (1994). Hidden Markov models of biological primary sequence information. *Proc. Natl Acad. Sci. USA*, **91**, 1059-1063.
- Baldi, P., Brunak, S., Chauvin, Y., Engelbrecht, J. & Krogh, A. (1995). Periodic sequence patterns in human exons. In *Proc. Third Int. Conf. on Intelligent Systems for Mol. Biol.* (Rawlings, C., Clark, D., Altman, R., Hunter, L., Lengauer, T. & Wodak, S., eds), pp. 30-38. AAAI Press, Menlo Park.
- Bolshoy, A. (1995). CC dinucleotides contribute to the bending of DNA in chromatin. *Nature Struct. Biol.* **2**, 447-448.
- Brendel, V., Beckmann, J. S. & Trifonov, E. N. (1986). Linguistics of nucleotide sequences: morphology and comparison of vocabularies. *J. Mol. Struct. Dyn.* **4**, 11-21.
- Brukner, I., Sánchez, R., Suck, D. & Pongor, S. (1995a). Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *EMBO J.* **14**, 1812-1818.
- Brukner, I., Sánchez, R., Suck, D. & Pongor, S. (1995b). Trinucleotide models for DNA bending propensity: comparison of models based on DNase I digestion and nucleosome packaging data. *J. Biomol. Struct. Dynam.* **13**, 309-317.
- Brunak, S. & Engelbrecht, J. (1996). Protein structure and the sequential structure of mRNA: α -helix and β -sheet signals at the nucleotide level. *Proteins: Struct. Funct. Genet.* **25**, 237-252.
- Brunak, S., Engelbrecht, J. & Knudsen, S. (1991). Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.* **220**, 49-65.
- Clark, D., Reitman, M., Studitsky, V. & Chung, J. (1993). Chromatin structure of transcriptionally active genes. *Cold Spring Harbor Symp. Quant. Biol.* **58**, 1-6.
- Crothers, D. M. & Steitz, T. A. (1992). Transcriptional activation by *Escherichia coli* CAP protein. In *Transcriptional Regulation* (McKnight, S. L. & Yamamoto, K. R., eds.), pp. 501-534, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Drew, H. R. & Travers, A. A. (1985). DNA bending and its relation to nucleosome positioning. *J. Mol. Biol.* **186**, 773-790.
- Elgin, S. C. R., editor, (1995). *Chromatin Structure and Gene Expression*, IRL Press, Oxford.
- Engelbrecht, J., Knudsen, S. & Brunak S. (1992). G + C-rich tract in 5' end of human introns. *J. Mol. Biol.* **227**, 108-113.
- Fickett, J. W. (1982). Recognition of protein coding regions in DNA sequences. *Nucl. Acids Res.* **10**, 5303-5318.
- Fickett, J. W. & Tung, C. S. (1992). Assessment of protein coding measures. *Nucl. Acids Res.* **20**, 6441-6450.
- Goodman, S. D. & Nash, H. A. (1989). Functional replacement of a protein-induced bend in a DNA recombination site. *Nature*, **341**, 251-254.
- Goodsell, D. S. & Dickerson, R. E. (1994). Bending and curvature calculations in B-DNA. *Nucl. Acids Res.* **22**, 5497-5503.
- Haran, T. E., Kahn J. D & Crothers, D. M. (1994). Sequence elements responsible for DNA curvature. *J. Mol. Biol.* **244**, 135-143.
- Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical feature. *Biopolymers*, **22**, 2577-2637.
- Klug, A. & Lutter, L. C. (1981). The helical periodicity of DNA on the nucleosome. *Nucl. Acids Res.* **9**, 4267-4283.
- Kornberg, R. D. & Lorch, Y. (1992). Chromatin structure and transcription. *Annu. Rev. Cell Biol.* **8**, 563-587.
- Krogh, A., Brown, M., Mian, I. S., Sjölander, K. & Haussler, D. (1994). Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol.* **235**, 1501-1531.
- Krogh, A., Mian, I. S. & Haussler, D. (1995). A hidden Markov model that finds genes in *E. coli* DNA. *Nucl. Acids Res.* **22**, 4768-4778.
- Lamond, A. I. (1995). *Pre-mRNA Processing*, R. G. Landes Company, Austin.
- Levinson, S. E., Rabiner, L. R. & 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.
- Lukashin, A. V., Engelbrecht, J. & Brunak, S. (1992). Multiple alignment using simulated annealing:

- branch point definition in human mRNA splicing. *Nucl. Acids Res.* **20**, 2511–2516.
- Muyldermans, S. & Travers, A. A. (1994). DNA sequence organization in chromatosomes. *J. Mol. Biol.* **235**, 855–870.
- Nussinov, R. (1989). Strong patterns in homooligomer tracts occurrences in non-coding and in potential regulatory sites in eukaryotic genomes. *J. Biomol. Struct. Dynam.* **6**, 985–1000.
- Pehrson, J. R. (1989). Thymine dimer formation as a probe of the path of DNA in and between nucleosomes in intact chromatin. *Proc. Natl Acad. Sci. USA*, **86**, 9149–9153.
- Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.
- Rumelhart, D. E., Durbin, R., Golden, R. & Chauvin, Y. (1994). Back-propagation: the theory. In *Back-propagation: Theory, Architectures and Applications* (Chauvin, Y. E. & Rumelhart, D. E., eds), Lawrence Erlbaum Associates, Hillsdale, New Jersey and Hove, UK.
- Satchwell, S. C., Drew, H. R. & Travers, A. A. (1986). Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191**, 659–675.
- Sinden, R. R. (1994). *DNA Structure and Function*, Academic Press, San Diego.
- Trifonov, E. N. (1987). Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16 S 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.
- Trifonov, E. N. & Sussman, J. L. (1980). The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc. Natl Acad. Sci. USA*, **77**, 3816–3820.
- VanWye, J. D., Bronson, E. C. & Anderson, J. N. (1991). Species-specific patterns of DNA bending and sequence. *Nucl. Acids Res.* **19**, 5253–5261.
- Widom, J. (1996). Short-range order in two eukaryotic genomes: relation to chromosome structure. *J. Mol. Biol.* **259**, 579–588.
- Zhurkin, V. B. (1981). Periodicity in DNA primary structure is defined by secondary structure of the coded protein. *Nucl. Acids Res.* **9**, 1963–1971.
- Zhurkin, V. B. (1983). Specific alignment of nucleosomes on DNA correlates with periodic distribution of purine-pyrimidine and pyrimidine-purine dimers. *FEBS Letters*, **158**, 293–297.

Edited by T. Richmond

(Received 4 April 1996; received in revised form 2 September 1996; accepted 2 September 1996)