Article No. mb981972 *J. Mol. Biol.* (1998) **281**, 663–673

JMB

DNA Structure in Human RNA Polymerase II Promoters

Anders Gorm Pedersen[1], Pierre Baldi[2], Yves Chauvin[2] and Søren Brunak[1]*

[1]*Center for Biological Sequence Analysis, The Technical University of Denmark Building 208, DK-2800 Lyngby Denmark*

[2]*Net-ID, Inc., 4225 Via Arbolada, Suite 500 Los Angeles, CA 90042, USA*

**Corresponding author*

The fact that DNA three-dimensional structure is important for transcriptional regulation begs the question of whether eukaryotic promoters contain general structural features independently of what genes they control. We present an analysis of a large set of human RNA polymerase II promoters with a very low level of sequence similarity. The sequences, which include both TATA-containing and TATA-less promoters, are aligned by hidden Markov models. Using three different models of sequence-derived DNA bendability, the aligned promoters display a common structural profile with bendability being low in a region upstream of the transcriptional start point and significantly higher downstream. Investigation of the sequence composition in the two regions shows that the bendability profile originates from the sequential structure of the DNA, rather than the general nucleotide composition. Several trinucleotides known to have high propensity for major groove compression are found much more frequently in the regions downstream of the transcriptional start point, while the upstream regions contain more low-bendability triplets. Within the region downstream of the start point, we observe a periodic pattern in sequence and bendability, which is in phase with the DNA helical pitch. The periodic bendability profile shows bending peaks roughly at every 10 bp with stronger bending at 20 bp intervals. These observations suggest that DNA in the region downstream of the transcriptional start point is able to wrap around protein in a manner reminiscent of DNA in a nucleosome. This notion is further supported by the finding that the periodic bendability is caused mainly by the complementary triplet pairs CAG/CTG and GGC/GCC, which previously have been found to correlate with nucleosome positioning. We present models where the high-bendability regions position nucleosomes at the downstream end of the transcriptional start point, and consider the possibility of interaction between histone-like TAFs and this area. We also propose the use of this structural signature in computational promoter-finding algorithms.

Keywords: RNA polymerase II promoter; DNA structure; nucleosome; TBP-associated factors (TAFs); hidden Markov model (HMM)

Introduction

Packaging of DNA into chromatin limits the accessibility of the DNA template for the transcriptional apparatus and has been found to inhibit transcriptional initiation (Paranjape *et al.*, 1994; Kornberg & Lorch, 1995; Gottesfeld & Forbes, 1997; Grunstein, 1997). Compared to naked DNA, chromatin is therefore in a state of transcriptional repression with a much lower basal activity. This is presumably important for the tight regulation of gene activity *in vivo*: while activator proteins typically increase transcription from a naked DNA template around tenfold, the activation seen *in vivo* can be a thousandfold or more. Thus, derepression of transcription by partial unfolding of chromatin is likely to constitute an important part of gene regulation, and in accordance with this notion several transcription factors and transcriptional co-activators have recently been shown to work by

Abbreviations used: TBP, TATA-box binding protein; TAF, TBP-associated factor; HMM, hidden Markov model; Inr, initiator.

E-mail address of the corresponding author: brunak@cbs.dtu.dk

0022–2836/98/340663–11 $30.00/0

disrupting or remodeling chromatin structure (Brownell & Allis, 1995; Brownell *et al.*, 1996; Mizzen *et al.*, 1996; Ogryzko *et al.*, 1996; Pazin & Kadonaga, 1997; Tsukiyama & Wu, 1997). For example, some transcription factors and co-activators have been found to be histone acetyltransferases. Most likely these work by masking positively charged residues on the highly basic histones, thereby lowering the affinity of the histones for DNA and consequently facilitating access of the transcriptional apparatus. The yeast factor SWI2/SNF2 and related proteins seem to function as ATP-driven machines that move along DNA and disrupt nucleosomes.

DNA three-dimensional structure can influence initiation of transcription in other ways. Several so-called architectural transcription factors that bend DNA have been described (van der Vliet & Verrijzer, 1993; Grosschedl, 1996; Werner & Burley, 1997). Their task is most likely to bring transcription factors bound at distant sites into close contact. The TATA-box binding protein (TBP) is known to bend DNA and thereby facilitate subsequent binding of the basal transcription factor TFIIB (Horikoshi *et al.*, 1992; Kim *et al.*, 1993a,b; Nikolov *et al.*, 1995). Interestingly, some transcription factors have been found to contain structural motifs that are present also in histones and other architectural chromatin proteins (Baxevanis *et al.*, 1995; van Holde & Zlatanova, 1996; Burley *et al.*, 1997). For instance, the TBP-associated factors $hTAF_{II}80$, $hTAF_{II}31$ and $hTAF_{II}20$ are structurally similar to core histones, and apparently form a histone octamer-like structure within the basal transcription factor TFIID (Burley & Roeder, 1996; Hoffmann *et al.*, 1996; Orphanides *et al.*, 1996; Xie *et al.*, 1996). Thus, there is an intricate interplay between protein-associated bending and folding of DNA, and transcriptional regulation.

The three-dimensional structure of DNA has been found to depend on the exact sequence of nucleotides, an effect that seems to be caused largely by interactions between neighboring base-pairs (Klug *et al.*, 1979; Dickerson & Drew, 1981; Hagerman, 1984; Satchwell *et al.*, 1986; Calladine *et al.*, 1988; Bolshoy *et al.*, 1991; Hunter, 1993, 1996; Goodsell & Dickerson, 1994; Brukner *et al.*, 1995a). Generally, periodic repetitions of bent DNA in phase with the helical pitch will cause the DNA to assume a macroscopically curved structure. Several models for estimating DNA structure from di- or trinucleotides have been devised, based on many different kinds of experimental data (e.g. see Goodsell & Dickerson, 1994; Sinden, 1994; Brukner *et al.*, 1995a; Hassan & Calladine, 1996). Sequence-dependent DNA structure is also important for protein-DNA interactions: e.g. it is energetically more favorable to wrap flexible or intrinsically curved DNA around histones compared to rigid and unbent DNA, and this has been shown to influence nucleosome positioning (Drew & Travers, 1985; Satchwell *et al.*, 1986; Richard-Foy & Hager, 1987; Simpson, 1991; Lu *et al.*, 1994; Bolshoy, 1995; Iyer & Struhl, 1995; Wolffe & Drew, 1995; Ioshikhes *et al.*, 1996; Widom, 1996; Zhu & Thiele, 1996; Liu & Stein, 1997). Similar mechanisms are also important for sequence-specific DNA-binding proteins (Parvin *et al.*, 1995; Starr *et al.*, 1995; Grove *et al.*, 1996).

Here, we investigate structural features in a set of RNA polymerase II promoters using sequence-derived models of DNA bendability. We find a general structural profile that is present in a majority of the promoters. The results show that a region downstream of the transcription start may have the ability to be wrapped around protein. We suggest that this is a signal for positioning a nucleosome and discuss the possibility of interactions between histone-like TAFs and this region.

Results

A set of human RNA polymerase II promoters was extracted from the GenBank database. All sequences are 501 nt long and contain 250 nt up- and downstream of the transcriptional start point. After carefully reducing the redundancy of these data, which include both TATA-containing and TATA-less promoters, we used hidden Markov models (HMMs) to construct a multiple alignment. This alignment formed the basis for further investigations of DNA structural features. For the purpose of analyzing the sequence-dependent DNA structure, we have used three different models of DNA flexibility based on completely independent experimental measurements: a trinucleotide model based on DNase I cutting frequencies (Brukner *et al.*, 1995a); another trinucleotide model based on tabulation of preferred sequence locations on nucleosomes (Satchwell *et al.*, 1986; Goodsell & Dickerson, 1994); and a dinucleotide model based on X-ray crystallography of DNA oligomers (Hassan & Calladine, 1996).

Bendability profiles: structure downstream of the transcriptional start point

The trinucleotide model proposed by Brukner *et al.* (1995a) is based on the observation that DNase I preferably binds and cuts DNA that is bent (or bendable) towards the major groove (Lahm & Suck, 1991; Weston *et al.*, 1992; Suck, 1994). Thus, DNase I cutting frequencies on naked DNA can be interpreted as a quantitative measure of major groove compressibility or anisotropic bendability. Such data have been used to calculate bendability parameters for the 32 complementary trinucleotide pairs (Brukner *et al.*, 1995a).

Application of the bendability model to all overlapping triplets in a promoter gives the bendability values at each position in that particular sequence. By doing this for all sequences in the multiple alignment and averaging the resulting values at each position, we obtained an average bendability profile (Figure 1(a)). In the

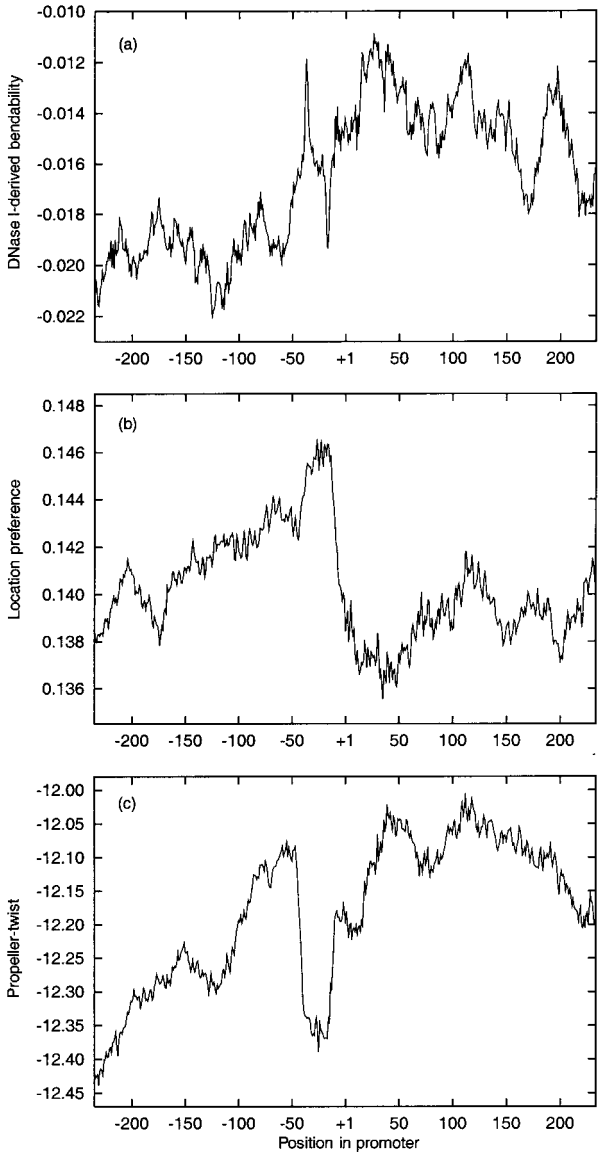

Figure 1. Average bendability profiles of the human promoter sequences. Position +1 corresponds to the transcriptional start point. (a) The non-redundant set of 624 human promoters was aligned using a hidden Markov model, and the average bendability for each position in the promoter calculated using DNase I-derived bendability parameters (Brukner *et al.*, 1995a). Higher parameters correspond to higher bendability (or propensity for major groove compressibility). The two peaks around position −30 are caused by TATA-box-containing promoters. The profile has been smoothed by calculating a running average with a window of size 20. (b) Average flexibility profile calculated from the aligned promoters using a trinucleotide model based on preferred sequence location on nucleosomes (Satchwell *et al.*, 1986). Lower values correspond to more flexible sequences, which have less preference for being positioned specifically. (c) Average flexibility profile based on propeller-twist values from X-ray crystallography of DNA oligomers (Hassan & Calladine, 1996). Higher (less negative) propeller-twist corresponds to higher flexibility. Profiles (b) and (c) have been smoothed by calculating a running average with a window of size 30. Note that all three profiles show a tendency for higher flexibility downstream of the transcriptional start point.

alignment, position +1 largely corresponds to the transcriptional start point (see Methods). This profile shows the general tendency for major groove compressibility in different parts of the aligned promoters. Around position −25, a high bendability spike just upstream of a low-bendability spike is clearly visible in the plot. This signal is caused mainly by TATA-box containing promoters (data not shown). High propensity for major groove compression in the TATA-box is consistent with the high-resolution crystal structure of the TATA-box/TBP complex: TBP binds to the minor groove of the TATA-box and bends the DNA about 80° towards the major groove (Kim *et al.*, 1993a,b). Thus, the DNase I bendability model captures an essential structural feature that is experimentally well established.

The most distinct feature of this graph, however, is the very clear difference in average bendability between the upstream and downstream regions: the bendability is significantly higher in a region downstream of the transcriptional start point compared to the region upstream of the TATA-box. The high-bendability region extends for roughly 150 bp beginning in between the TATA-box and the start point and ending around position +150. Since it is energetically more favorable to wrap flexible DNA around protein, this profile suggests that, on average, the region downstream of the TATA-box in a diverse set of human RNA polymerase II promoters has this ability. It is noteworthy that the length of the high-bendability region is similar to the length of DNA in a nucleosome, and it is tempting to suggest that the high-bendability region is a signal for positioning nucleosomes. Specific positioning of nucleosomes near the transcriptional start point in a set of widely different promoters may be related to chromatin-based repression of transcription (see Introduction and Discussion).

In order to further investigate the putative high-bendability region, we analyzed the promoter alignment using a completely independent approach. From experimental investigations of the positioning of DNA in nucleosomes, it has been found that certain trinucleotides have strong preferences for having minor grooves facing either towards or away from the nucleosome core (Satchwell *et al.*, 1986). Based on the premise that flexible sequences can occupy any rotational position on nucleosomal DNA, while rigid sequences will be restricted in rotational location, these preference values can be used as measures of DNA flexibility: all triplets having close to zero preference are assumed to be flexible, while triplets with preference for facing either in or out are taken to be more rigid.

We applied the preference values to the multiple alignment and produced the resulting flexibility profile (Figure 1(b)). It is obvious from this plot that a large transition takes place around the transcriptional start point: downstream of the start point, the average preference values are signifi-

cantly lower than upstream, meaning that downstream DNA is more flexible. Thus, the main result from the DNase I-based analysis is in agreement with the nucleosome-positioning model.

We also used a dinucleotide model that is based on the direct measurement of base-pair geometry in crystallized oligonucleotides (Hassan & Calladine, 1996). Dinucleotides with a large propeller-twist have a tendency to be more rigid than dinucleotides with low propeller-twist. Propeller-twist can therefore also be used as a measure of DNA flexibility.

A plot of the average propeller-twist at all positions in the alignment is shown in Figure 1(c). Here, higher values correspond to smaller (less negative) propeller-twist equivalent to higher flexibility. Although the transition is less marked than that seen in the other two profiles, there is, nevertheless, a clear difference between the region upstream of the TATA-box and the region downstream of the transcriptional start point. The average propeller-twist in the downstream region is smaller than upstream, indicating that DNA in this region is more flexible. This is again in agreement with the conclusion derived from the other two structural profiles.

Since the putative high-bendability region is downstream of the transcriptional start point, we were concerned that the signal might reflect trivial codon usage bias in the coding portions present in a subset of the sequences. In order to investigate this, we divided the promoters into a subset that does contain some coding sequence in the 250 nucleotides downstream of the transcriptional start point (261 sequences), and a subset that does not (363 sequences). Analysis of these two sets showed that the structure was not caused by codon bias: bendability profiles calculated from a data set that does not contain any coding DNA are qualitatively very similar to profiles obtained from a set with coding DNA (data not shown).

The structural profiles described above are averaged over a set of 624 different promoters. It is highly unlikely that every single promoter in this set has the same overall profile. In order to investigate this point further we subdivided the data set based on relative average bendabilities in different regions. When calculating the bendability in an upstream region located between −200 and −50, and a downstream region extending from +1 to +150, we find that 398 promoter sequences (corresponding to 65%) have higher bendability in the downstream region. Furthermore, analysis of a large number of individual sequences shows that many promoters have bendability profiles that are very similar to the average profiles shown above. Another smaller group have high-bendability regions that are positioned further upstream (data not shown). We therefore conclude that a majority of the promoter sequences possess the average structural feature described above.

Upstream and downstream regions have different sequence composition

We analyzed the nucleotide composition in two 150 bp regions that appeared strongly different from the DNase I profile: an upstream region extending from position −200 to position −50, and a downstream region starting at +1 and extending downstream to position +150. By selecting the regions in this way we assured that the results are not greatly influenced by TATA-box containing promoters, since most TATA boxes are positioned around −30. Results were, however, very similar when the analysis was performed for the entire 250 nt regions up- and downstream of the transcriptional start point (data not shown).

We analyzed both nucleotide, dinucleotide and trinucleotide compositions. For all three classes, we observe varying degrees of non-uniform distribution between the up- and downstream regions. Careful analysis of the trinucleotide distribution (Figure 2(a)) shows that the bendability difference seen between up- and downstream regions in the DNase I profile can be explained by the preferential usage of a number of high-bendability triplets downstream and low-bendability triplets upstream. Thus, out of the 13 most over-represented triplets in the downstream region, ten have high DNase I-derived bendability parameters (TCT, GCA, GCC, CAG, CTG, GAG, CTC, AGC, GCT and TGC). Conversely, the six most over-represented trinucleotides upstream (GGG, CCC, AAA, TTT, AAT and ATT) all have low bendability parameters.

An additional interesting observation that can be made from the plot of triplet frequencies is that complementary trinucleotides show a tendency for being located adjacently: e.g. observe the positions of the complementary pairs CCC/GGG, AAA/TTT, AAT/ATT, CAG/CTG, GAG/CTC and AGC/GCT in Figure 2(a). Thus, within each region, trinucleotides and their complementary trinucleotides are present with almost equal frequencies. Comparison of all 32 complementary trinucleotide pairs showed a near-perfect correlation between the frequency of a triplet and its complementary triplet (data not shown). Since the analysis of triplet frequencies is based on only one strand of the promoter DNA, we find it unlikely that this phenomenon is caused by the complementarity of the two strands. The observation is, however, consistent with the relation between DNA structure and the trinucleotide distribution, since a triplet and its complementary triplet by definition have the same bendability.

To further examine the distribution of trinucleotides, we calculated triplet frequencies at different positions in the promoters from the HMM-alignment. This analysis shows that practically all non-uniformly distributed triplets have highly distinct usage profiles with sharp transitions near the transcriptional start point (see Figure 3 for two examples). Triplets that generally are more frequent

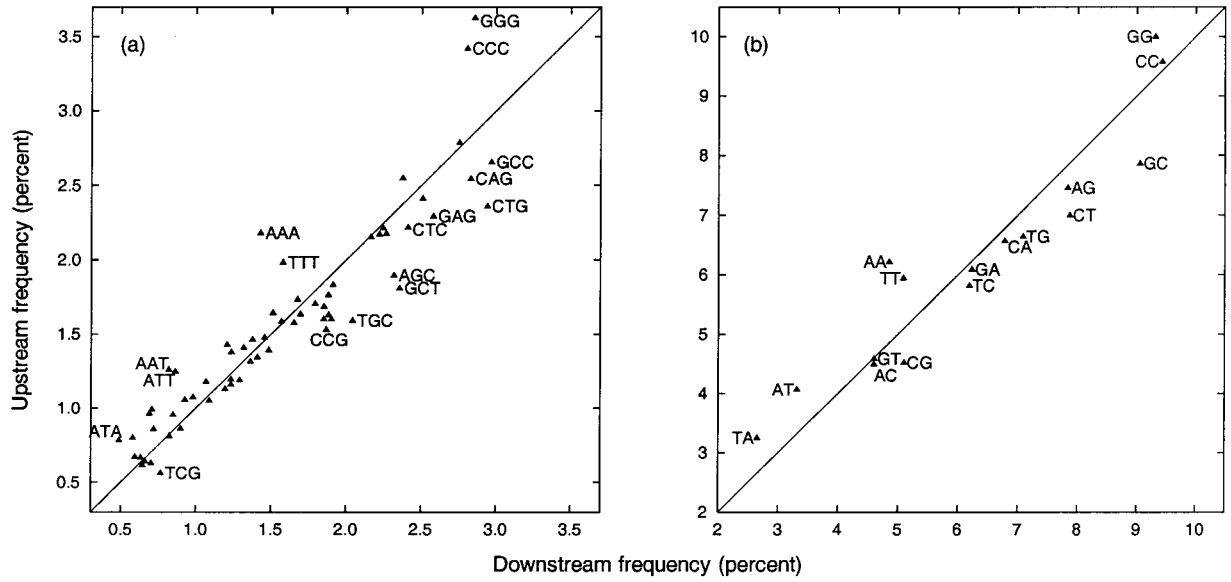

Figure 2. Comparison of di- and trinucleotide frequencies in regions upstream and downstream of the transcriptional start point. The frequency of (a) all trinucleotides and (b) dinucleotides was calculated in two regions: an upstream region extending from position −200 to position −50 (relative to the transcriptional start point) and a downstream region starting at +1 and ending at position +150. For any given di- or trinucleotide, frequencies in the two regions are compared by plotting a point at a position determined by the downstream (x-axis) and upstream (y-axis) frequencies of that sequence element. Di- or trinucleotides that are used more often downstream than upstream will be plotted beneath the diagonal and *vice versa*. In the trinucleotide plot, only some points are marked for clarity. Most of the triplets that are preferentially found in the downstream region have high DNase I-derived bendability parameters, while most of the triplets that are more frequent in the upstream region have low values.

upstream show a sharp decline in usage close to the start point, while triplets that are more frequent downstream have opposite frequency profiles that rise steeply around the transcription start. Thus, the clear transition between low and high-bendability near the start point is mirrored in the frequency profiles of several individual trinucleotides rather than being merely an average phenomenon.

We also analyzed nucleotide and dinucleotide frequencies (Table 1 and Figure 2(b)). Although the distribution of single nucleotides between the up- and downstream regions is slightly skewed, this far from accounts for the difference in the trinucleotide distributions. As a simple illustration, consider the trinucleotides TCG and CTG. If the single-nucleotide composition fully explained triplet frequencies, these triplets would occur with very similar frequencies. However, as it can be seen from Figure 2(a), these triplets are located at two opposite ends of the frequency spectrum (in the downstream region $f_{\mathrm{TCG}} = 0.77\%$ while $f_{\mathrm{CTG}} = 2.94\%$).

The average bendabilities expected in the up- and downstream regions were calculated from the nucleotide and dinucleotide distributions, and compared to the actual trinucleotide-based values (Table 2). This analysis shows that the single-nucleotide distribution far from accounts for the observed values. The dinucleotide distribution, on the other hand, gives an almost exact estimate of the observed trinucleotide-based bendability in the upstream region, but does less well in the downstream region.

Sequence periodicity and periodic bendability

Sequence-based DNA structure is related to sequence periodicity: periodic repetition of an intrinsically bent sequence element in phase with the DNA helical pitch causes the DNA to assume a macroscopically curved structure. In the case of sequence elements that are anisotropically bendable (as opposed to rigidly curved), periodic repetition will have a similar effect, in that the DNA will have the ability to be curved. In order to investigate whether the regions downstream of the transcriptional start point contain periodic sequence patterns, we used HMMs with circular architectures of varying length. HMM wheel-architectures have previously been used in an analysis of eukaryotic genes to find periodic patterns in human exons and introns (Baldi *et al.*, 1997). Such periodic patterns have recently been confirmed to occur in connection with *in vivo* positioned nucleosomes (Liu & Stein, 1997).

Interestingly, wheel architectures comprising ten main states trained on the downstream regions were found to do much better than the free linear architecture: they train much faster and achieve lower negative log-likelihood, indicating that there are indeed periodic patterns in the data. We also observed that wheels with 20 main states were able to model the data considerably better than wheels comprising 9, 10 or 11 states (data not shown).

Figure 4(a) shows the emission parameters of a circular model of length 20. In this plot, the periodic sequence pattern C[AT]G can be clearly seen

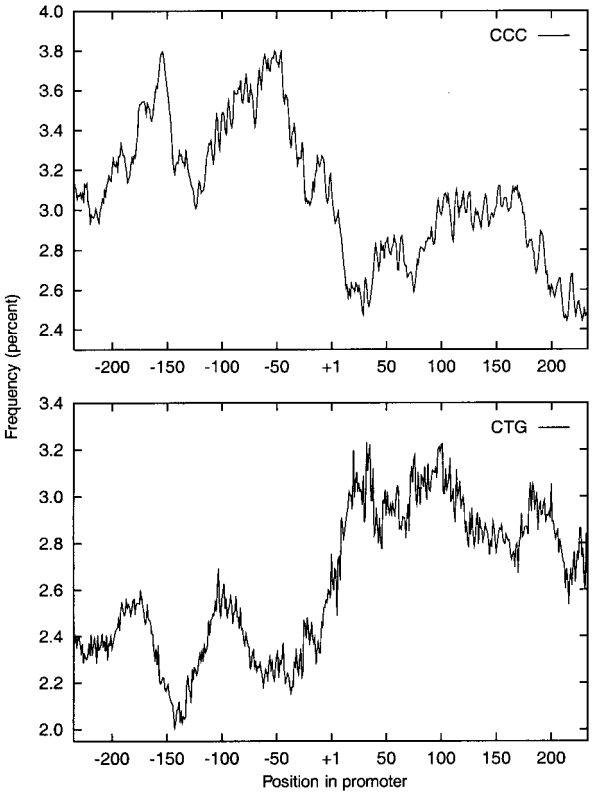

Figure 3. Frequency profiles for the trinucleotides CCC (upper panel) and CTG (lower panel). Average trinucleotide frequencies were calculated at all positions from the aligned promoters. Note that CCC, which is more frequent in the upstream region (see Figure 2(b)), displays a sharp drop in frequency around the start point. The opposite is true for CTG, which is more frequent downstream. All triplets that are over-represented in the upstream region have frequency profiles qualitatively similar to that of CCC, while triplets over-represented in the downstream region have profiles similar to that of CTG (data not shown).

in states 16, 17 and 18. From a structural viewpoint, the two triplets (CAG and CTG) that conform to this consensus are very interesting. First, they are complementary and therefore have identical structural properties. Second, they have very high bendability: the triplet pair CAG/CTG is number 3 on the DNase I scale, while it is one of the most flexible (only 2% rotational preference) on the nucleosome-positioning scale. Third, this triplet pair is included in the more general periodic nucleosome signal (non-T[AT]G) found earlier (Baldi *et al.*, 1997; Liu & Stein, 1997). CAG/CTG is also one of the high-bendability triplet pairs found to be over-represented in the downstream region.

Just upstream, and overlapping one nucleotide, the periodic pattern G[CG]C can be seen in states 14, 15 and 16. Interestingly, the two triplets contained within this consensus (GCC and GGC) are also complementary. Furthermore, these triplets have very high major groove compressibility according to the DNase I scale (fifth from the top).

Table 1. Nucleotide frequencies

	Upstream (%)	Downstream (%)
A	22.2	20.6
C	27.7	29.2
G	28.4	29.2
T	21.6	21.0

Table 2. Expected and measured DNase I-derived bendability

	Expected (nucleotide)	Expected (dinucleotide)	Measured (trinucleotide)
Upstream	−0.0164	−0.0195	−0.0194
Downstream	−0.0160	−0.0125	−0.0137

They are also the triplets that have the highest preference (45%) for being located on nucleosomes with the minor groove facing out (i.e. with a compressed major groove).

Thus, the consensus pattern seen in the model parameters of the loop HMM is consistent with previously identified periodic patterns that are known to be associated with nucleosome positioning.

In order to examine whether the sequence periodicity implies a periodic bendability, we estimated a bendability profile from the trained wheel-model with length 20 (Figure 4(b)). This profile has two distinct high-bendability peaks next to each other and a less strong peak offset 10 bp from the highest peak. The two high peaks correspond to the G[CG]C[AT]G pattern described above. Close analysis of the weaker peak shows that it is caused by the presence of a second, less pronounced G[GC]C pattern that is located exactly 10 bp from the C[AT]G pattern. The pattern showing strong bendability every 20 bp and somewhat weaker bendability every 10 bp, fits the crystal structure of the nucleosome core particle very well (Richmond *et al.*, 1984; Luger *et al.*, 1997): DNA wrapped around a histone octamer does not follow a uniformly circular path, but shows larger distortions separated by approximately 20 bp (corresponding to two helical turns). This further supports the notion that DNA in the region downstream of the transcriptional start point is able to assume a structure very similar to that seen in the nucleosome.

Discussion

The findings reported here indicate that human promoters may have general structural features independently of what type of genes they control. Specifically, computational analysis using three independent models of DNA flexibility shows that a set of promoters with a low level of sequence similarity displays an average tendency for low bendability upstream of the transcriptional start point and high bendability downstream. Closer examination reveals that this is caused by the pre-

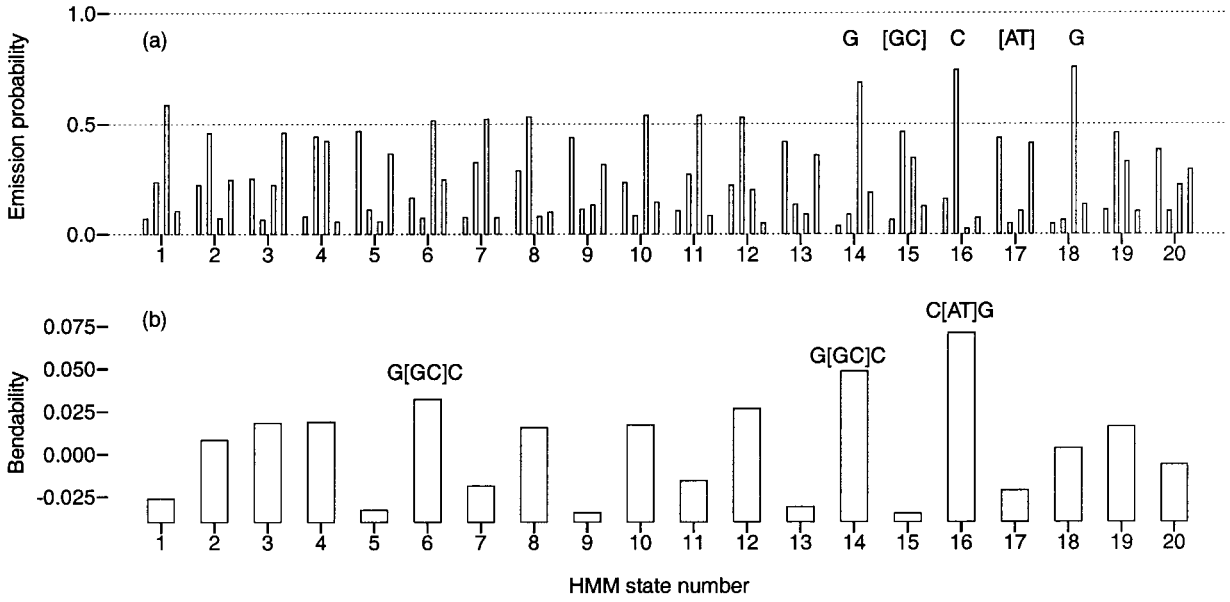

Figure 4. Analysis of periodic sequence patterns downstream of the transcriptional start point using circular HMMs. (a) Main state emission probabilities from a circular HMM of length 20 trained on the region downstream of the start point in all 624 promoter sequences. For each main state the emission probabilities of the four nucleotides are indicated in the order A, C, G, T. The periodic consensus pattern **G**[GC]**C**[AT]**G** emerges from the high emissions in states 14, 15, 16, 17 and 18. Two triplet pairs contained within this consensus (GGC/GCC and CAG/CTG) are complementary and have high DNase I-derived bendability. (b) Bendability profile estimated from the circular HMM of length 20. The bendability profile was calculated from the nucleotide emission probabilities in the main states of the trained HMM using DNase I-derived bendability parameters. Note the pattern with a strong peak every 20 bp and a weaker peak offset about 10 bp. This pattern is reminiscent of the path of DNA in a nucleosome. The complementary triplet pairs that are mainly responsible for the peaks are indicated.

ferred use of a number of high-bendability triplets in the downstream region and a corresponding over-representation of low-bendability triplets in the upstream region. Within the downstream region there are strong indications of periodic sequence and bendability patterns in phase with the DNA helical pitch. This periodic pattern is very similar to that known from X-ray structures of the nucleosome core particle and tabulations of preferred sequence locations on nucleosomes. The presence of structure in the downstream region is in agreement with results we obtained in a previous HMM-based analysis of human promoters (Pedersen *et al.*, 1996). HMMs trained on a large set of RNA polymerase II promoters displayed low main state entropies in a region downstream of the transcriptional start point, indicating the presence of non-random sequence patterns. The fact that trinucleotides and their complementary trinucleotides occur with almost identical frequencies and at the same periodic positions further supports the notion that there is a structural basis for the regularities in the findings reported here.

Our results indicate that the DNA in the region downstream of the start point is able to assume a macroscopically curved structure (e.g. to be wrapped around protein) very similar to that of DNA in a nucleosome. Since the length of the high-bendability region is approximately the same as the length of DNA wrapped around a histone octamer, it is tempting to suggest that this is a signal for positioning nucleosomes right at the transcriptional start point. This notion is supported by several observations: (1) the periodicity of the bendability pattern (20/10 bp) corresponds to the physical structure of the histone octamer (Richmond *et al.*, 1984; Luger *et al.*, 1997); (2) the triplet pair GGC/GCC, which constitutes an important part of the periodic signal, has previously been found to have a large preference for being positioned in phase with the helical pitch on the outside of nucleosomes (Satchwell *et al.*, 1986); and (3) the complementary triplet pair CAG/CTG, which is over-represented in the downstream region and makes up the most important part of the periodic signal, has also been found to be correlated with nucleosome positioning (Baldi *et al.*, 1997; Liu & Stein, 1997). Positioning of nucleosomes near the transcriptional start point could be related to the tight regulation of gene expression that is often observed *in vivo*: although promoters are usually thought of in terms of their function in promoting initiation of transcription, another important characteristic of many genes is a low basal or uninduced activity. Positioning of a nucleosome at the transcriptional start point would seem to be in full agreement with this idea. An alternative but not contradictory model is that it is the upstream low-bendability region that is important. A number of promoters in yeast and higher eukaryotes contain homopolymeric dA:dT elements. Such homopolymeric tracts are known

from X-ray crystallography to be straight and rigid (Nelson *et al*., 1987). Recent studies in two different yeast species have shown that homopolymeric elements destabilize nucleosomes and thereby facilitate the access of transcription factors bound nearby (Iyer & Struhl, 1995; Zhu & Thiele, 1996). In this way, rapid response to activation is possible. Conceivably, the low-bendability region we observe upstream of the transcriptional start point may have this function. Activation mechanisms similar to those seen in the yeast promoters may therefore be active in a wide range of different eukaryotic promoters.

The TBP-associated factors $hTAF_{II}80$, $hTAF_{II}31$ and $hTAF_{II}20$ have been found to be histone-like in their structure (Burley & Roeder, 1996; Xie *et al*., 1996; Hoffmann *et al*., 1996). Thus, it may be imagined that these TAFs can bind to the high-bendability region in a manner reminiscent of a nucleosome. In support of this, histone-like TAFs have indeed been shown to directly interact with the region downstream of the transcriptional start point in some promoters (Oelgeschläger *et al*., 1996; Burke & Kadonaga, 1997; Smale, 1997; Hoffmann *et al*., 1997). It is, furthermore, intriguing that $hTAF_{II}250$ has been found to have histone acetyltransferase activity and that TFIID may therefore in itself be able to displace or destabilize nucleosomes (Mizzen *et al*., 1996). Taken together, these facts suggest an interesting scenario where promoters exist in two functionally different but structurally similar complexes: a transcriptionally repressed state complexed with histones, and a transcriptionally potentiated state in a similar complex with TFIID (Hoffmann *et al*., 1997).

We note that a bendability profile averaged over all possible heptamers conforming to the very loose consensus sequence of the initiator (Inr) element $PyPyA_{+1}N[TA]PyPy$ (Smale, 1997) displays a single distinct high-bendability peak at position +1 (data not shown). This is caused by the fact that all eight triplets described by the subconsensus $PyA_{+1}N$ have high bendability. The two trinucleotides contained within the corresponding region in the more limited *Drosophila* consensus ($CA_{+1}[GT]$) are two of the highest scoring on the DNase I scale (third and fourth from the top, respectively). We tentatively suggest that at least part of the sequence requirements for a functional Inr are of a structural nature. An average bendability profile of the downstream promoter element (DPE), consensus [AG]G[AT]CGTG) described in *Drosophila* and man also shows a single bendability peak caused by the 3′ GTG. The fact that this element has to be positioned precisely about 30 bp downstream of the Inr in order to have activity (Burke & Kadonaga, 1997) is also in agreement with part of its function being structural.

A protein bound in the region downstream of the transcriptional start point could have another effect on the initiation of transcription due to the topological connection between supercoiling and DNA strand twisting ($\Delta Lk = \Delta Tw + \Delta Wr$). Assuming no change in the linking number ($\Delta Lk = 0$), negative superhelical winding of DNA is equivalent to local untwisting of the DNA strands. Thus, it is possible that this region could be instrumental in strand separation at initiation of transcription.

We have described a structural feature that appears to be present in a large fraction of all human promoters. This opens the interesting possibility of using the structure as a signal in promoter-finding algorithms. All methods for localizing promoters in unannotated sequence, either explicitly or implicitly, work by searching for conserved promoter elements such as the TATA box or the initiator (Fickett & Hatzigeorgiou, 1997). Many algorithms, furthermore, mostly search for signals positioned upstream of the transcriptional start point. Thus, it might be imagined that the additional use of a hitherto unknown signal (of a structural nature) positioned downstream of the transcriptional start point could improve the performance of promoter-finding methods. We are in the process of developing a promoter-finding algorithm based on this premise.

Methods

Extraction of a non-redundant set of RNA polymerase II promoter sequences

A set of human RNA polymerase II promoters was extracted from the GenBank database release 95 (Benson *et al*., 1997). Specifically, we extracted sequences surrounding annotated, experimentally determined transcriptional start points of protein-encoding genes. This was done by selecting nuclear (i.e. not mitochondrial) sequences having an ''mRNA'' or ''5′ UTR'' feature key (indicating the presence of a transcriptional start point), as well as a ''CDS'' feature key (indicating that the gene is protein-encoding and hence transcribed by RNA polymerase II). The resulting set of sequences was subsequently controlled for consistency of annotation (i.e. no conflicting annotation of sequence elements), and manually reviewed for irregularities. Sequences whose annotation indicated that the transcription start had been assigned by non-experimental means were discarded. Since we wanted to avoid a bias in the analysis by assuming that only the upstream region is essential for promoter function, we extracted 250 nt from both up- and downstream of the start point. Sequences containing ambiguous nucleotide symbols were excluded. This gave a set of 820 promoter sequences of length 501 nt. It is, however, a well-known problem that sequence databases are redundant due to the presence of, for instance, genes belonging to gene families and identical sequences submitted to the database more than once. This has consequences for statistical analysis, since results will be biased for any over-represented sequences and it is therefore important to remove sequences that are too similar (Sander & Schneider, 1991; Hobohm *et al*., 1992; Nielsen *et al*., 1996; Baldi & Brunak, 1998). We performed very thorough reduction of the redundancy using algorithm 2 from Hobohm *et al*. (1992) and a novel method (unpublished) for finding a similarity cutoff. Briefly, this method is based on performing all pairwise local alignments for a data set, fitting the resulting Smith-Water-

man scores to an extreme value distribution (Altschul *et al.*, 1994), and choosing a cutoff value above which there are more observations than expected from the distribution. The size of the redundancy-reduced data set was 624 sequences. This set was further subdivided into a set containing coding sequence in the 250 nucleotides downstream of the transcriptional start point (261 sequences), and one without any coding sequence (363 sequences). The data sets are available from the authors upon request.

Hidden Markov models

Hidden Markov models (HMMs) are flexible probabilistic models that generalize the notion of a profile, and are applicable to a wide range of sequence analysis problems (Krogh *et al.*, 1994; Baldi *et al.*, 1994, 1997; Hughey & Krogh, 1996; Baldi & Brunak, 1998).

Briefly, a first-order HMM consists of a set of states, a transition probability matrix specifying the probability of moving from any state to any other state, and an emission probability matrix specifying the probability of producing any letter of the alphabet from any state. Sequences are generated stochastically by superimposing the corresponding emission and transition Markov processes.

The most widely used HMM architecture is the linear architecture, where the HMM states are arranged linearly from left to right into a sequence of positions that can be thought of as columns of a multiple alignment. The HMM states are partitioned into three classes: the main states, the delete states, and the insert states. The main states are ordered linearly from left to right and form the backbone of the model. For each main state, there is an associated insert and delete state to account for possible insertions and deletions in the multiple alignment. Delete states are mute, i.e. they produce a gap with probability 1. The transition of an insert state onto itself is used to model multiple insertions. In addition to HMMs with a linear architecture, we have developed and used HMMs with a circular structure to study periodic patterns (Baldi *et al.*, 1997).

There exist several efficient algorithms to fit an HMM to a family of sequences by maximum likelihood or maximum *a posteriori* optimization. This can be used to automatically estimate (or train) the model parameters from a set of unaligned sequences. Using a trained model it is conversely possible to produce a multiple alignment of a set of sequences. The trainable emission probabilities capture the local composition of the sequence family, while the trainable transition probabilities effectively correspond to position-dependent gap-penalties in a multiple alignment.

Preliminary investigations had demonstrated that HMMs are able to model human promoter sequences in a meaningful way (Pedersen *et al.*, 1996). We therefore trained several linear HMMs on the non-redundant data set. Examination of the resulting models show that they capture two well-known features common to eukaryotic promoters: a TATA-box (TATAAAA) approximately 30 bp upstream of the start point and a pronounced CA at the start point (data not shown). Approximately 25% of the sequences do not go through the states corresponding to the TATA-box. This number is in reasonable agreement with the expected number of TATA-less promoters in a human data set.

An HMM of length 500 has about 6000 free parameters (1000 emitting states $\times$ (4 nucleotides $-$ 1) + 1500 states $\times$ (3 transitions $-$ 1)), while the number of nucleotides in the training set is larger than 312,000 (624 sequences of length 501 nt). This ratio indicates that there is little danger of overfitting.

We have trained linear architectures of length 500, using the full promoter sequences, in addition to models of length 250 using either the upstream or downstream regions. We have trained a variety of wheel architectures on the same regions. In all cases, we have tested different training algorithms and parameters, as well as different training sets, including training sets devoid of downstream coding regions. The resulting multiple alignments were very similar across different training algorithms and parameters. In the multiple alignment of length 500, the large majority of transcription start sites were emitted by a main state very close to main state 251. Hence this position is referred to as position +1 throughout the text.

In the case of linear architectures of length 500, the quality of the models can be slightly improved by initializing the parameters in the region associated with the TATA-box by hand prior to training. The transition parameters of these architectures were regularized in a standard way using Dirichlet prior distributions in order to favor the use of main states over insert states (Krogh *et al.*, 1994; Hughey & Krogh, 1996; Baldi & Brunak, 1998).

Calculation of bendability profiles

The trinucleotide-based, DNase I-derived bendability model has been described elsewhere (Brukner *et al.*, 1990, 1995a,b). The correlation between propeller-twist and DNA flexibility is described by Hassan & Calladine (1996). The use of the nucleosome-positioning model for quantifying flexibility was suggested to us by Dr A. A. Travers and is based on data described by Satchwell *et al.* (1986). Specifically, we have used the trinucleotide location preferences without the sign indicating the direction of the preference.

Based on the di- or trinucleotide parameters in the models mentioned above, it is possible to calculate bendability profiles for any given sequence (Brukner *et al.*, 1995a; Baldi *et al.*, 1997; Baldi & Brunak, 1998). This can be done by looking up the bendability parameter for each consecutive overlapping di- or trinucleotide in the sequence. In this work, the aim was to investigate general structural features of human promoters, and we therefore calculated bendability profiles from the alignments rather than from single sequences. Thus, calculations were performed on di- or trinucleotides centered on HMM main states. Since, however, DNA structure is dependent on the exact non-gapped nucleotide sequence, we used the flanking nucleotides in the original sequence, and not the flanking main states, when looking up bendability parameters. This procedure was repeated for all main states in all sequences, and the average at each main state was subsequently calculated.

Acknowledgments

We thank Dr A. A. Travers for suggesting the use of nucleosome-positioning data as a measure of flexibility.

A.G.P. and S.B. are supported by a grant from the Danish National Research Foundation. The work of P.B. and Y.C. is, in part, supported by an NIH SBIR grant to NetID, Inc.

References

Altschul, S., Boguski, M. S., Gish, W. & Wootton, J. C. (1994). Issues in searching molecular sequence databases. *Nature Genet.* **6**, 119–129.

Baldi, P. & Brunak, S. (1998). *Bioinformatics: The Machine Learning Approach*, MIT Press, Cambridge, MA.

Baldi, P., Chauvin, Y., Hunkapillar, T. & McClure, M. (1994). Hidden Markov models of biological primary sequence information. *Proc. Natl Acad. Sci. USA,* **91**, 1059–1063.

Baldi, P., Brunak, S., Chauvin, Y. & Krogh, A. (1997). Naturally occurring nucleosome positioning signals in human exons and introns. *J. Mol. Biol.* **263**, 503–510.

Baxevanis, A. D., Arents, G., Moudrianakis, E. N. & Landsman, D. (1995). A variety of DNA-binding and multimeric proteins contain the histone fold motif. *Nucl. Acids Res.* **23**, 2685–2691.

Benson, D., Boguski, M., Lipman, D. & Ostell, J. (1997). GenBank. *Nucl. Acids Res.* **25**, 1–6.

Bolshoy, A. (1995). CC dinucleotides contribute to the bending of DNA in chromatin. *Nature Struct. Biol.* **2**, 447–448.

Bolshoy, A., McNamara, P., Harrington, R. E. & Trifonov, E. N. (1991). Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. *Proc. Natl Acad. Sci. USA,* **88**, 2312–2316.

Brownell, J. E. & Allis, D. D. (1995). An activity gel assay detects a single catalytically active histone acetyltransferase subunit in *Tetrahymena* macronuclei. *Proc. Natl Acad. Sci. USA,* **92**, 6364–6368.

Brownell, J. E., Zhou, J., Ranalli, T., Kobayashi, R., Edmondson, D. G., Roth, S. Y. & Allis, C. D. (1996). *Tetrahymena* histone acetyltransferase A: a homologue to yeast Gcn5p linking histone acetylation to gene activation. *Cell,* **84**, 843–851.

Brukner, I., Jurukovski, V. & Savic, A. (1990). Sequence-dependent structural variations of DNA revealed by DNase I. *Nucl. Acids Res.* **18**, 891–894.

Brukner, I., Sanchez, 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., Sanchez, R., Suck, D. & Pongor, S. (1995b). Trinucleotide models for DNA bending propensity: comparison of models based on DNase I digestion and nucleosome positioning data. *J. Biomol. Struct. Dynam.* **13**, 309–317.

Burke, T. W. & Kadonaga, J. T. (1997). The downstream promoter element, DPE, is conserved from *Drosophila* to humans and is recognized by $TAF_{II}60$ of *Drosophila*. *Genes Dev.* **11**, 3020–3031.

Burley, S. K. & Roeder, R. G. (1996). Biochemistry and structural biology of transcription factor IID (TFIID). *Annu. Rev. Biochem.* **65**, 769–799.

Burley, S. K., Xie, X., Clark, K. L. & Shu, F. (1997). Histone-like transcription factors in eukaryotes. *Curr. Opin. Struct. Biol.* **7**, 94–102.

Calladine, C. R., Drew, H. R. & McCall, M. J. (1988). The intrinsic structure of DNA in solution. *J. Mol. Biol.* **201**, 127–137.

Dickerson, R. E. & Drew, H. R. (1981). Structure of a *B*-DNA dodecamer. II. Influence of base sequence on helix structure. *J. Mol. Biol.* **149**, 761–786.

Drew, H. R. & Travers, A. A. (1985). DNA bending and its relation to nucleosome positioning. *J. Mol. Biol.* **186**, 773–790.

Fickett, J. W. & Hatzigeorgiou, A. G. (1997). Eukaryotic promoter recognition. *Genome Res.* **7**, 861–878.

Goodsell, D. S. & Dickerson, R. E. (1994). Bending and curvature calculations in *B*-DNA. *Nucl. Acids Res.* **22**, 5497–5503.

Gottesfeld, J. M. & Forbes, D. J. (1997). Mitotic repression of the transcriptional machinery. *Trends Biochem. Sci.* **22**, 197–202.

Grosschedl, R. (1996). Higher-order complexes in transcription: analogies with site-specific recombination. *Curr. Opin. Cell Biol.* **7**, 362–370.

Grove, A., Galeone, A., Mayol, L. & Geiduschek, E. P. (1996). Localized DNA flexibility contributes to target site selection by DNA-bending proteins. *J. Mol. Biol.* **260**, 120–125.

Grunstein, M. (1997). Histone acetylation in chromatin structure and transcription. *Nature,* **389**, 349–352.

Hagerman, P. J. (1984). Evidence for the existence of stable curvature of DNA in solution. *Proc. Natl Acad. Sci. USA,* **81**, 4632–4636.

El Hassan, M. A. & Calladine, C. R. (1996). Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. *J. Mol. Biol.* **259**, 95–103.

Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of representative data sets. *Protein Sci.* **1**, 409–417.

Hoffmann, A., Chiang, C. M., Oelgeschläger, C. M., Xie, X., Burley, S. K., Nakatani, Y. & Roeder, R. G. (1996). A histone octamer-like structure within TFIID. *Nature,* **380**, 356–359.

Hoffmann, A., Oelgeschläger, T. & Roeder, R. G. (1997). Considerations of transcriptional control mechanisms: do TFIID-core promoter complexes recapitulate nucleosome-like functions. *Proc. Natl Acad. Sci. USA,* **94**, 8928–8935.

Horikoshi, M., Bertuccioli, C., Takada, R., Wang, J., Yamamoto, T. & Roeder, R. G. (1992). Transcription factor TFIID induces DNA bending upon binding to the TATA element. *Proc. Natl Acad. Sci. USA,* **89**, 1060–1064.

Hughey, R. & Krogh, A. (1996). Hidden Markov models for sequence analysis: extension and analysis of the basic method. *CABIOS,* **12**, 95–107.

Hunter, C. A. (1993). Sequence-dependent DNA structure: the role of base stacking interactions. *J. Mol. Biol.* **230**, 1025–1054.

Hunter, C. A. (1996). Sequence-dependent DNA structure. *Bioessays,* **18**, 157–162.

Ioshikhes, I., Bolshoy, A., Derenshteyn, K., Borodovsky, M. & Trifonov, E. N. (1996). Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *J. Mol. Biol.* **262**, 129–139.

Iyer, V. & Struhl, K. (1995). Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J.* **14**, 2570–2579.

Kim, J. L., Nikolov, D. B. & Burley, S. K. (1993a). Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature,* **365**, 520–527.

Kim, Y., Geiger, J. H., Hahn, S. & Sigler, P. B. (1993b). Crystal structure of a yeast TBP/TATA-box complex. *Nature,* **365**, 512–520.

Klug, A., Jack, A., Viswamitra, M. A., Kennard, O., Shakked, A. & Steitz, T. A. (1979). A hypothesis on a specific sequence-dependent conformation of DNA and its relation to the binding of the *lac-repressor protein*. *J. Mol. Biol.* **131**, 669–680.

Kornberg, R. G. & Lorch, Y. (1995). Interplay between chromatin structure and transcription. *Curr. Opin. Cell. Biol.* **7**, 371–375.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. (1994). Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol.* **235**, 1501–1531.

Lahm, A. & Suck, D. (1991). DNase I-induced DNA conformation: 2 Å structure of a DNase I-octamer complex. *J. Mol. Biol.* **222**, 645–667.

Liu, K. & Stein, A. (1997). DNA sequence encodes information for nucleosome array formation. *J. Mol. Biol.* **270**, 559–573.

Lu, Q., Wallrath, L. L. & Elgin, S. C. R. (1994). Nucleosome positioning and gene regulation. *J. Cell. Biochem.* **55**, 83–92.

Luger, K., Maeder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature,* **389**, 251–260.

Mizzen, C. A., Yang, X.-J., Kokubo, T., Brownell, J. E., Bannister, A. J., Owen-Hughes, T., Workman, J. L., Berger, S. L., Kouzarides, T., Nakatani, Y. & Allis, C. D. (1996). The $TAF_{II}250$ subunit of TFIID has histone acetyltransferase activity. *Cell,* **87**, 1261–1270.

Nelson, H. C. M., Finch, J. T., Luisi, B. F. & Klug, A. (1987). The structure of an oligo(dA)-oligo(dT) tract and its biological implications. *Nature,* **330**, 221–226.

Nielsen, H., Engelbrecht, J., von Heijne, G. & Brunak, S. (1996). Defining a similarity threshold for a functional protein sequence pattern: the signal peptide cleavage site. *Proteins: Struct. Funct. Genet.* **26**, 165–177.

Nikolov, D. B., Chen, H., Halay, E. D., Usheva, A. A., Hisatake, K., Lee, D. K., Roeder, R. G. & Burley, S. K. (1995). Crystal structure of a TFIIB-TBP-TATA-element ternary complex. *Nature,* **377**, 119–128.

Oelgeschläger, T., Chiang, C.-M. & Roeder, R. G. (1996). Topology and reorganization of a human TFIID-promoter complex. *Nature,* **382**, 735–738.

Ogryzko, V., Schiltz, R. L., Russanova, V., Howard, B. H. & Nakatani, Y. (1996). The transcriptional coactivators p300 and CBP are histone acetyltransferases. *Cell,* **87**, 953–959.

Orphanides, G., Lagrange, T. & Reinberg, D. (1996). The general transcription factors of RNA polymerase II. *Genes Dev.* **10**, 2657–2683.

Paranjape, S. M., Kamakaka, R. T. & Kadonaga, J. T. (1994). Role of chromatin structure in the regulation of transcription by RNA polymerase II. *Annu. Rev. Biochem.* **63**, 265–297.

Parvin, J. D., McCormick, R. J., Sharp, P. A. & Fisher, D. E. (1995). Pre-bending of a promoter sequence enhances affinity for the TATA-binding factor. *Nature,* **373**, 724–727.

Pazin, M. J. & Kadonaga, J. T. (1997). SWI2/SNF2 and related proteins: ATP-driven motors that disrupt protein-DNA interactions? *Cell,* **88**, 737–740.

Pedersen, A. G., Baldi, P., Brunak, S. & Chauvin, Y. (1996). Characterization of prokaryotic and eukaryotic promoters using hidden Markov models. *Intell. Syst. Mol. Biol.,* **4**, 182–191.

Richard-Foy, H. & Hager, G. L. (1987). Sequence specific positioning of nucleosomes over the steroid-inducible MMTV promoter. *EMBO J.* **6**, 2321–2328.

Richmond, T. J., Finch, J. T., Rushton, B., Rhodes, D. & Klug, A. (1984). Structure of the nucleosome core particle at 7 Å resolution. *Nature,* **311**, 532–537.

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–58.

Satchwell, S. C., Drew, H. R. & Travers, A. A. (1986). Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191**, 659–675.

Simpson, R. T. (1991). Nucleosome positioning: occurrence, mechanisms, and functional consequences. *Prog. Nucl. Acids Res. Mol. Biol.* **40**, 143–184.

Sinden, R. R. (1994). *DNA Structure and Function*, Academic Press, San Diego, CA.

Smale, S. T. (1997). Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes. *Biochim. Biophys. Acta,* **1351**, 73–88.

Starr, D. B., Hoopes, B. C. & Hawley, D. K. (1995). DNA bending is an important component of site-specific recognition by the TATA binding protein. *J. Mol. Biol.* **250**, 434–446.

Suck, D. (1994). DNA recognition by Dnase I. *J. Mol. Recogn.* **7**, 65–70.

Tsukiyama, T. & Wu, C. (1997). Chromatin remodeling and transcription. *Curr. Opin. Genet. Dev.* **7**, 182–191.

van der Vliet, P. C. & Verrijzer, C. P. (1993). Bending of DNA by transcription factors. *BioEssays,* **15**, 25–32.

van Holde, K. & Zlatanova, J. (1996). Chromatin architectural proteins and transcription factors: a structural connection. *BioEssays,* **18**, 697–700.

Werner, M. H. & Burley, S. K. (1997). Architectural transcription factors: proteins that remodel DNA. *Cell,* **88**, 733–736.

Weston, S. A., Lahm, A. & Suck, D. (1992). X-ray structure of the Dnase I-dGGTATCC complex at 2.3 Å resolution. *J. Mol. Biol.* **226**, 1237–1256.

Widom, J. (1996). Short-range order in two eukaryotic genomes: relation to chromosome structure. *J. Mol. Biol.* **259**, 579–588.

Wolffe, A. P. & Drew, H. R. (1995). DNA structure: implications for chromatin structure and function. In *Chromatin Structure and Gene Expression* (Elgin, S. C. R., ed.), pp. 27–48, IRL Press, Oxford.

Xie, X., Kokubo, T., Cohen, S. L., Mirza, U. A., Hoffmann, A., Chait, B. T., Roeder, R. G., Nakatani, Y. & Burley, S. K. (1996). Structural similarity between TAFs and the heterotetrameric core of the histone octamer. *Nature,* **380**, 316–323.

Zhu, Z. & Thiele, D. J. (1996). A specialized nucleosome modulates transcription factor access to a *C. glabrata* metal responsive promoter. *Cell,* **87**, 459–470.

Edited by J. Karn

(Received 2 January 1998; received in revised form 2 June 1998; accepted 4 June 1998)