Published online July 27, 2005

Nucleic Acids Research, 2005, Vol. 33, No. 13 ***4255–4264***
doi:10.1093/nar/gki737

Large-scale structural analysis of the core promoter in mammalian and plant genomes

Kobe Florquin, Yvan Saeys, Sven Degroeve, Pierre Rouzé and Yves Van de Peer*

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

Received March 29, 2005; Revised June 10, 2005; Accepted July 10, 2005

ABSTRACT

DNA encodes at least two independent levels of functional information. The first level is for encoding proteins and sequence targets for DNA-binding factors, while the second one is contained in the physical and structural properties of the DNA molecule itself. Although the physical and structural properties are ultimately determined by the nucleotide sequence itself, the cell exploits these properties in a way in which the sequence itself plays no role other than to support or facilitate certain spatial structures. In this work, we focus on these structural properties, comparing them between different organisms and assessing their ability to describe the core promoter. We prove the existence of distinct types of core promoters, based on a clustering of their structural profiles. These results indicate that the structural profiles are much conserved within plants (*Arabidopsis* and rice) and animals (human and mouse), but differ considerably between plants and animals. Furthermore, we demonstrate that these structural profiles can be an alternative way of describing the core promoter, in addition to more classical motif or IUPAC-based approaches. Using the structural profiles as discriminatory elements to separate promoter regions from non-promoter regions, reliable models can be built to identify core-promoter regions using a strictly computational approach.

INTRODUCTION

During the last 20 years, the role of non-linear DNA structures in replication, recombination, DNA packaging within the nucleus and regulation of gene expression has become more and more appreciated (1–6). Non-linear DNA structures are not directly linked with the protein-coding part of sequences, but are exploited by the cell in a way in which the sequence itself plays no role other than to support or facilitate certain spatial structures. As a consequence, DNA actually encodes, in its sequence, at least two independent levels of functional information. The first level, which is well known, is used for encoding proteins and their regulatory elements. The DNA sequence is then the actual physical carrier for the genetic information that codes for the vast amount of proteins to be produced. In addition, the primary DNA sequence also holds the *cis*-regulatory elements responsible for directing spatial and temporal gene expression patterns in response to metabolic requirements, developmental programs and a plethora of external stimuli. Extensive research during the past decades has revealed that eukaryotic gene transcription is a remarkably intricate biochemical process that is tightly regulated at many levels (7–9). DNA-binding transcription factors (TFs) are one of the important components in this network and orchestrate gene expression by binding DNA through specific *cis*-acting regulatory elements, which are short conserved motifs of 5 up to 20 nt usually found in the vicinity of the 5′ end of genes, in what is called the promoter (see below).

The second level of information, and the focus of the current study, is contained in the physical and structural properties of the DNA molecule itself. Although DNA is often depicted as a uniformly straight and rigid double helix, it possesses inherent structural properties that play a role in many different biological processes (1–4,10). Because of the important role of these structural properties in different key biological processes, much research has focused on the biophysical understanding of the intrinsic curvature and bendability of DNA sequences (11–15). DNA does not behave as an isotropic rod, but depending on the sequence it might bend more easily in one plane than another, indicating that it possesses a degree of anisotropic flexibility. Such flexibility is crucial, since interactions between static structures are insufficient to explain the DNA–protein recognition events necessary for, e.g. gene regulation (14,16–18). In eukaryotes, RNA polymerase II (RNA Pol II) is responsible for transcribing nuclear genes encoding the mRNAs and several small nuclear RNAs. Similar to RNA Pol I and Pol III, RNA Pol II cannot recognize its target promoter directly and cannot initiate transcription without accessory proteins.

*To whom correspondence should be addressed. Tel: +32 9 331 3807; Fax: +32 9 331 3809; Email: Yves.VandePeer@psb.ugent.be

Instead, this large multisubunit enzyme relies on both general TFs and transcriptional activators and coactivators (8,19). The regulation imposed by the distal part of the promoter is mainly based on the binding by TFs on different binding sites. The regulation of promoters by distal enhancers, e.g. within complex genetic loci, has been the subject of intense investigation, and models of communication between distant protein–DNA complexes include DNA looping, protein tracking or changes in DNA topology, each of which is thought to activate transcription by increasing the local concentration of factors in the vicinity of the promoter (20–22). The general concept is that DNA topology, due to physical and structural properties of DNA, strongly influences transcription by immediately promoting the assembly of specialized structures, which are required for enhancers or inhibitors, to exert their regulatory effect. The proximal part, or the core promoter, then serves as the recognition site for the basal transcription apparatus, which comprises the multisubunit RNA Pol II protein complex and several auxiliary factors. Core promoters show specificity both in their interactions with enhancers and with sets of general TFs that control distinct subsets of genes. Although there are no known DNA sequence motifs that are shared by all core promoters, a number of motifs have been identified that are present in a substantial fraction, the most familiar of these motifs being the TATA-box.

The correct assembly of the stereo-specific RNA Pol II nucleoprotein complex requires proteins to bind to DNA in a sequence specific manner, and the correct assembly of the complex necessitates that the DNA at least facilitates the complex to be built (10,19). Because of the requirement of this physical support it is likely that specific higher-order structural elements are present within a core promoter and the question arises whether eukaryotic promoters contain general structural elements independently of the genes they control. Different studies have shown that, in general, eukaryotic core promoters indeed do have a distinct structural profile when compared with coding or non-regulatory sequences (2,23–26). However, many of these studies were conducted on specific groups of core promoters based on their mutual function (23,25) or the presence of specific core-promoter boxes (26).

Here, we investigate the higher-order structural topology of core promoters on a larger scale, for both mammal and plant sequences and report on a new approach that enables us to clearly differentiate between different core-promoters classes based on structural properties, such as intrinsic curvature, bendability, stacking energy and propeller twist.

MATERIALS AND METHODS

Core-promoter datasets

For human and mouse, we used the publicly available DBTSS database (http://tarawa.icm.edu.pl/dbtss/). This database contains information on the genomic positions of the transcriptional start sites (TSS) and the adjacent promoters for 8793 human and 6875 mouse genes, and was obtained by the mapping of 400 225 human and 580 209 mouse full-length cDNA sequences, respectively (27–30). For *Arabidopsis* and rice, we have constructed our own core-promoter datasets. The ARAPROM dataset (7088 promoter sequences) was constructed by aligning full-length cDNA sequences, generated by Seki *et al*. (31), with the genomic sequence (32). For the RICEPROM dataset, we adopted the same procedure, which resulted in 2195 putative promoter sequences. Because we were mainly interested in the regions flanking the putative TSS, i.e. the putative core-promoter regions, we extracted 100 bp upstream and 50 bp downstream of the TSS from each sequence in the dataset (33,34).

Structural models

In a first step, we converted the core-promoter sequences to a string of numerical values using the di- or trinucleotide values, coming from different structural models (Figure 1). To this end, we used a sliding window approach with a step of 1 and a

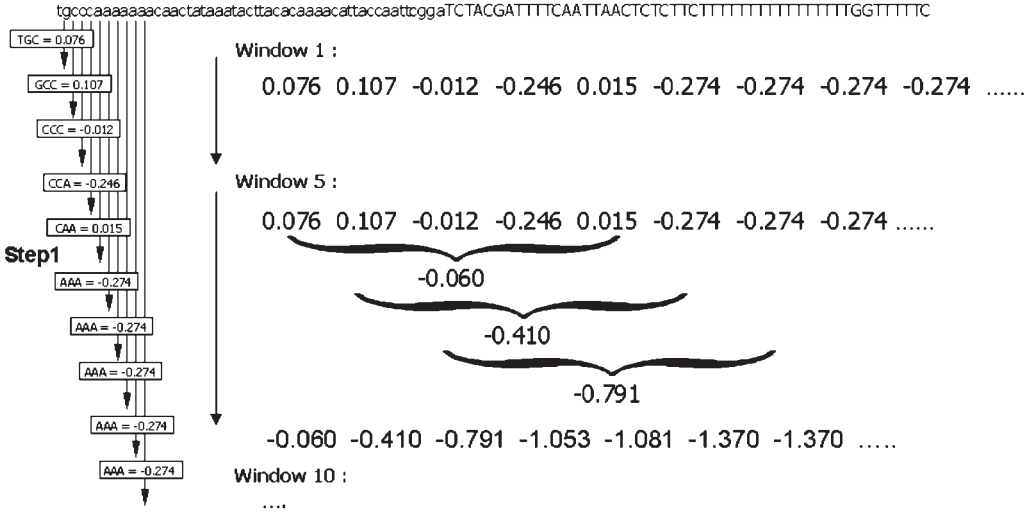

Figure 1. Sequence information is converted to numerical profiles. In this example, the trinucleotide bendability model by Brukner *et al*. (11) is used, based on Dnase I cutting frequencies. The enzyme Dnase I preferably binds to the minor groove and cuts DNA that is bent, or bendable toward the major groove. Therefore, Dnase I cutting frequencies on naked DNA can be interpreted as a quantitative measure of major groove compressibility or bendability. These frequencies allow for the derivation of bendability parameters for 32 complementary trinucleotide pairs and range from −0.280 to +0.194. To evaluate different smoothings of the raw profile signal (see text), a sliding window approach was used with steps of 1 and a window size of 1–10, respectively.

Table 1. The different structural models that were considered in our analysis

Structural feature	Property measured	Min	Max	Reference
Stacking energy	Dinucleotide base-stacking energy scale expressed in kilocalories per mol, derived from approximate quantum mechanical calculations on crystal structures. High peaks in base stacking reflect regions of the helix that de-stack or melt more easily; conversely a minimal peak would represent more stable regions	−14.59 kcal	−3.82 kcal	(65)
Propeller twist	The dinucleotide propeller twist angle scale is measured in degrees and is based on X-ray crystallography of DNA oligomers. A region with high propeller twist would mean that the helix is quite rigid in this area. Correspondingly, regions that are quite flexible would have low propeller twist values	−18.66°	−8.11°	(66)
Nucleosome position preference	NPP is a trinucleotide model based on the preferential location of sequences within a nucleosomal core. The study was performed on sequences wrapped around nucleosome cores and in closed circles of DNA. They calculated the fractional preference of each base pair triplet for a position facing out. High value peaks represent more rigid regions where nucleosomes are less likely to appear	−36%	+45%	(67)
Bendability	The trinucleotide bendability model is based on Dnase I cutting frequencies. The enzyme Dnase I preferably binds to the minor groove and cuts DNA that is bent, or bendable toward the major groove. Thus Dnase I cutting frequencies on naked DNA can be interpreted as a quantitative measure of major groove compressibility or bendability. DNA regions with a high peak correspond to regions that are more flexible than regions with a low peak value	−0.280	+0.194	(11)
A-philicity	The free energy dinucleotide base pair scale, for the ethanol-induced B- to A-DNA conformational transitions in solution, was determined for a series of carefully designed synthetic duplexes. A region in the DNA with a high A-philicity value is more easily converted to the A-form than a low value region, which is more resistant to transition			(68)
Protein-induced deformability	The dinucleotide protein deformability scale is derived from empirical energy functions extracted from the fluctuations and correlations of structural parameters as determined by the examination of more than a hundred crystal structures of DNA–protein complexes. On this scale, a larger value reflects a more deformable sequence while a smaller value indicates a region where the DNA helix is less likely to be changed dramatically by proteins	1.6	12.1	(69)
Duplex disrupt energy	The DNA disrupt energy was calculated using calorimetric calculation on 19 DNA oligomers and 9 DNA polymers. It has been shown that the stability of a DNA duplex depends on its base sequence and that it is not the base composition that determines the stability of the duplex. Regions with a high disrupt energy value will be more stable than a region with a lower energy value	0.9 kcal	3.1 kcal	(70)
Duplex free energy	For 50 DNA/DNA duplexes the thermodynamic parameters of the DNA free energy were calculated. The melting behavior of these duplexes was observed and the transition enthalpy was calculated giving dinucleotide values. Regions with a low free energy content will be more stable than regions with high thermodynamic energy content	−2.1 kcal/mol	−0.9 kcal/mol	(71)
DNA denaturation	The denaturation equilibrium is calculated by UV electronic spectroscopy at 270 nm of high-resolution melting experiments on 42 plasmids, containing synthetic repeated inserts. DNA regions with a low peak value are more likely to denaturate than regions with a higher peak value	64.35 cal/mol	135.38 cal/mol	(72, 73)
DNA-bending stiffness	The bending stiffness is regarded as the translational positioning of nucleosomes and more precisely the string correlation with the anisotropic flexibility of the DNA. In the analysis, a simple algorithm is used that accounts for nucleosome translational positions in terms of bending free energy. The values are given in nm, which stand for the persistence length value that is derived from experimental data. High peak values correspond to DNA regions that are more rigid, while low peak values correspond to regions that will bend more easily	20 nm	130 nm	(74)
B-DNA twist	The study focuses on the mean twist angles in B-DNA and was calculated on 38 B-DNA crystal structures. Structures with a low twist region appear to unwind in response to steric clashes of large exocylic groups in the major and minor grooves, and those with high twist values are subject to lesser contact	30.6°	43.2°	(75)
Protein–DNA twist	Olson *et al*. (69) looked at the behaviour of dimer steps from 92 protein–DNA crystals complexes and calculated the average distributions of the conformational parameters that can describe the DNA variability. High peak values are more likely to be deformed by proteins than regions with a lower peak value	31.5°	37.8°	(69)
Stabilizing energy of Z-DNA	To search for particular DNA segments, which can adopt a left-handed Z-conformation, empirically determined energetic parameters are used. The dinucleotide parameters represent the free energy values for a transition from B- to Z-DNA. Stretches of DNA with low energy minima are more likely to form Z-DNA than a high-energy region	5.9 kcal/mol	0.7 kcal/mol	(76)

window size of 1–10 nt, respectively, in order to evaluate different smoothings of the raw profile signal. The different structural models that we considered are listed in Table 1.

Next to these structural models, we have also used the presence of CpG dinucleotides and CpNpG trinucleotides as additional models to describe a promoter region, although they are not linked to specific structural elements (35–38). In our study, we will not use the standard definition of Gardiner-Garden and Frommer (39), because this definition was based on a study on human sequences solely and as a consequence is highly correlated with the specific animal (32). In this regard, we opted for a simpler and more straightforward method where we associated a positive value with each CpG dinucleotide and each NpCpG, CpGpN or CpNpG trinucleotide and a zero value for all other di- or trinucleotides.

Clustering and classification

In a second step, we have used the converted sequence information to classify the core-promoter sequences into different groups. To this end, we used a novel approach, which is based on clustering the converted sequences. Because the clustering algorithm looks for similarity in the given profiles, sequences that exhibit comparable structural elements will be grouped together.

Here, we used the adaptive quality-based clustering (AQBC) tool because of its advantages over more classical clustering methods (40–42). For example, AQBC does not require the number of clusters to be defined in advance and will not force every sequence into a specific cluster. The AQBC was developed specifically to find large clusters that have a good quality guarantee. In short, the algorithm uses a two-step heuristic approach, where in a first step, i.e. the quality-based approach, a sphere is defined within the high-dimensional representation of the data where the density of the elements is maximal. This maximum is based on a preliminary estimation of the radius of the sphere or cluster. In a following step, called the adaptive approach, the optimal radius of the sphere is adapted in such a way that only the elements with the highest significance are included. The size of the radius is determined for different spheres separately by a trade-off between the probability of false-positive results and negative results. In this way, we can guarantee that a certain gene is assigned to the right sphere or cluster with a certain probability. In our analyses we used a probability threshold of 99% (42).

To investigate the significance of the core-promoter classes obtained by our clustering approach, the acquired datasets were compared with non-core-promoter datasets. To test how well the core-promoter sets could be differentiated from non-core-promoter or 'negative' datasets, a benchmark procedure was set up as follows. As negative datasets, both randomized datasets and non-promoters sequences from the corresponding coding (containing exon and intron information) and intergenic sequences were used (Figure 2).

The randomized datasets were constructed based on different shuffling methods applied to the real core-promoter sequences: (i) simple shuffling: shuffling of the nucleotides at random, (ii) dinucleotide shuffling: the order of the nucleotides in the sequence is randomized but the dinucleotide frequencies are preserved (43) and (iii) Markov shuffling: a Markov chain of degree 1 is derived from the original sequence and this probability model is used to generate new sequences, starting from a random nucleotide (44).

Two different approaches were then used to classify core-promoter and non-promoter sequences, namely AQBC (42) and a support vector machine (SVM)-based approach. The SVM is a data-driven method for solving two-class classification tasks (45,46). In our experiments, we used the SVM with a linear kernel (LSVM). To evaluate the performance of our classification, i.e. the ability to discriminate between core-promoter and non-core-promoter sequences, both the recall, defined as the proportion of positives that are correctly predicted, and the precision, defined as the proportion of predicted positives that are indeed correct, were evaluated:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

TP stands for true positives, TN for true negatives, FN for false negatives and FP for false positives. We used the *F*-measure

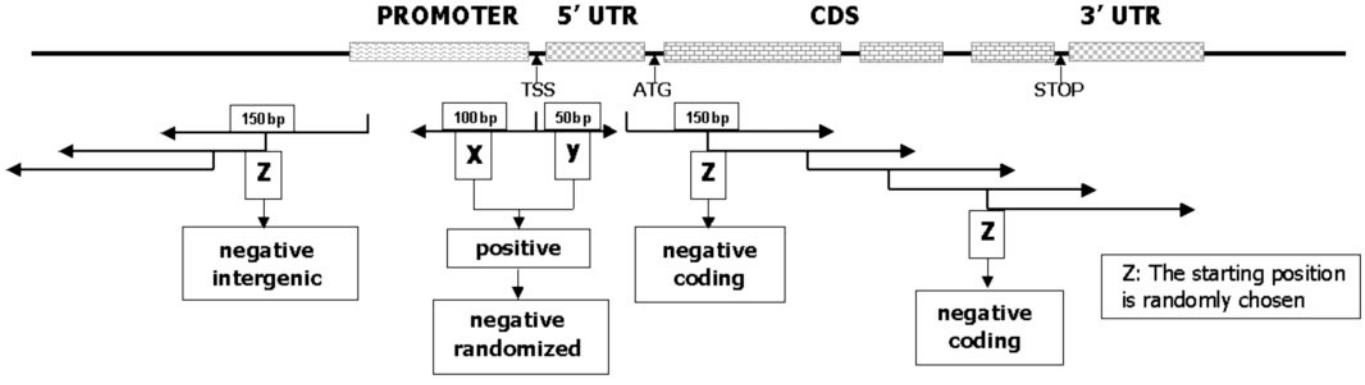

Figure 2. For *Arabidopsis* and rice, in-house core-promoter datasets were constructed. The ARAPROM dataset (7088 promoter sequences) was constructed by aligning full-length cDNA sequences, generated by Seki *et al*. (31), with the genomic sequence. The RICEPROM dataset consists of 2195 putative promoter sequences. From each original promoter sequence, 100 bp upstream and 50 bp downstream of the TSS were selected. As negative datasets, we extracted 150 bp from the non-promoter sequence part, including intron, exon and intergenic sequences. In addition, three randomized datasets were constructed, based on randomizing the core-promoter sequences.

that combines precision and recall, and is a measure for the overall classification performance. The *F*-measure takes the harmonic mean of recall and the precision:

$$F = 2 \times \frac{\text{Precision} \times \text{recall}}{\text{Precision} + \text{recall}}$$

Known core-promoter elements

To analyze the different core-promoter classes for the presence of known motifs, we used two different scanning approaches. The major difference between these two scanning approaches is the representation of the known motifs. On the one hand, we used the IUPAC consensus representation derived from a set of aligned sequences of the known binding sites. The consensus sequences were derived from the literature (34,47,48) and we used the TATA-box (TATAAA: −25/−30 nt TSS), the Inr-box (YYANWYY: +1 nt TSS), the DPE-box (RGWYV: +30 nt TSS), the BRE-box (SSRCGCC: −30 nt TSS), the CCAAT-box (GGNCAATCN: −75 nt TSS) and the GC-box (GGGCGG: −90/−100 nt TSS). A Perl-script was used to search for the occurrence of these motifs in the different core-promoter classes. Thereby, two different strategies were adopted. First, we scanned without position restrictions for the occurrence of a specific motif. Next, we limited the region where a motif can be found, based on the description of that specific known motif in the literature.

On the other hand, we used a quantitative representation of the motifs using a position weight matrix (PWM). These matrices are calculated by simply counting the occurrences of the nucleotides at each position of the motif. For the PWM matrices, we chose the matrices described by Bucher (49), such as the Initiation-, the TATA-, the CCAAT and the GC-box, which are available in the Transfac database (http://www.gene-regulation.com/). To scan with the PWMs, we used the PoSSum software (50).

RESULTS AND DISCUSSION

When core-promoter sequences are converted to numerical values on the basis of the structural models listed above (see Materials and Methods), some first general conclusions can be drawn. For example, the number of inferred clusters for the different structural profiles is pretty similar and seems to be correlated with the window size that was used to evaluate different smoothings of the raw profile signal (see Materials and Methods). As can be expected, using a large window size gives fewer but larger promoter classes, whereas a small window size generates more, yet smaller promoter classes (see Supplementary Material at http://www.psb.ugent.be/bioinformatics/). The average number of clusters formed by the AQBC method (42) is 17, for the different structural models, using a window of 5 nt, while a window size of 10 nt results in 12 clusters on average. In general, most core promoters belong to one of the clusters discerned by the AQBC algorithm (see Supplementary Material). For example, regarding the propeller twist structural model, 95% of the human and 92% of the *Arabidopsis* core-promoter sequences belong to one of the clusters using a window size of 10 bp, while for a window size of 5 bp, these percentages decrease to 88 and 85, respectively. These values are very similar for all structural models and show that the majority of the core promoters can be ascribed to a specific core-promoter class (see Supplementary Material). If we use very small window sizes this percentage will drop. As a consequence, using smaller window sizes will ascribe fewer genes to specific clusters. It should also be noted that sequences that belong to the same cluster do not show any, or very little, sequence similarity (data not shown).

For all structural models evaluated, the profiles are very similar between human and mouse, and between *Arabidopsis* and rice, but are clearly different between mammals and plants. For example, Figure 3 shows the difference in profile between plants and mammals for the four clusters with the highest quality value obtained for duplex disrupt energy (see Materials and Methods). The corresponding profiles for *Arabidopsis* (Figure 3, 1.a) and rice (Figure 3, 2.a) and for human (Figure 3, 3.a) and mouse (Figure 3, 4.a) are very similar, whereas the plant profiles do not match any of the mammalian profiles. The same is true for profiles based on other structural models (see Supplementary Material).

Although the structural models described here are obtained through different experimental techniques, one could wonder to what extent some of those may be redundant, i.e. basically providing the same information. Both Liao *et al*. (4) and Baldi *et al*. (51) studied the computational correlation between different structural models and concluded that the models are by and large independent from each other. In this paper, we compared the gene content of the different core-promoter classes based on different structural models and found little overlap (see Supplementary Material), hereby confirming the results of previous studies that the structural models are complementary to one another and can provide independent supporting evidence (4,51). From this, we can conclude that the models used to describe DNA structure capture different aspects of it. Several models may agree on some structural elements but can also uncover divergent interpretations of other structural elements. While no final consensus regarding these models exists, it is likely that each one provides a slightly different and partially complementary view of the DNA structure.

To investigate the significance of each core-promoter class and to get an overall view of the performance and discriminative power of our clustering approach based on the different structural models, we evaluated how well our core-promoters could be separated from non-core-promoter sequences. To this end, sequences from the cluster with the highest quality value of each structural model, using a window size of 5, were grouped with non-core-promoter sequences obtained from the five different negative datasets. Each dataset that was generated contained 25% positive sequences and 75% negative sequences. An LSVM-based approach was then applied to separate core-promoter sequences from non-core-promoter sequences (see Materials and Methods). These classification results are generally comparable for the different organisms, but can differ considerably for different structural models (see Supplementary Material). When using a window size of 5, six structural models are able to discriminate core-promoter sequences from non-promoter sequences for all the organisms, such as A-philicity, DNA bending stiffness, DNA denaturation, duplex disrupt energy, nucleosome positioning preference and propeller twist, and the average classification performance for these five models is 82%. The choice of the negative datasets can lower or increase the classification

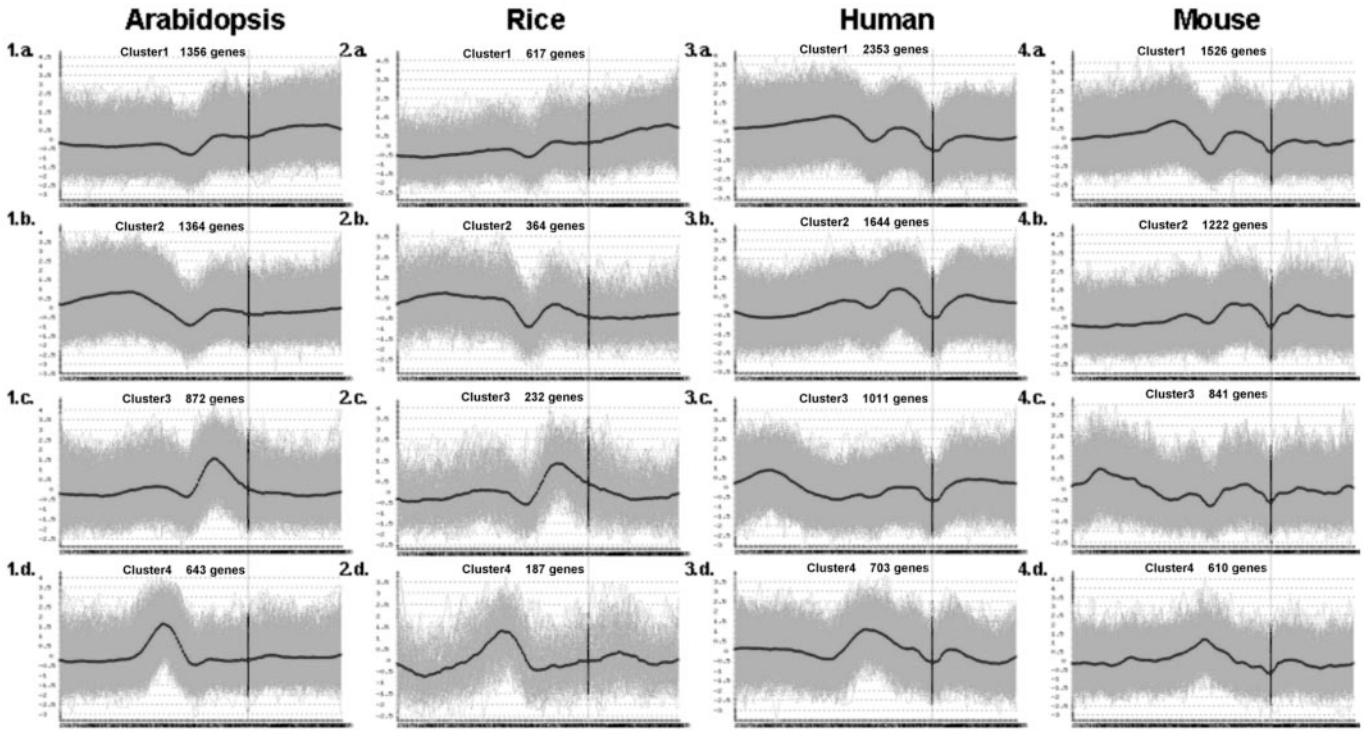

Figure 3. Profiles based on the structural model 'duplex disrupt energy' and window size 10 are shown for the four highest quality value clusters for *Arabidopsis*, rice, human and mouse (42). The position of the transcription start site is shown on the different structural profiles.

power, but this gain or loss is on average not significant (see Supplementary Material). In our study we only consider clusters, or core-promoter classes, that have an F-value of >65% as being significant.

The discriminatory power between core-promoter sequences and the negative instances also depends on the window size, or smoothing we perform, as can be seen in Figure 4. Using a small window size will give small but very specific promoter classes that can also be easily distinguished from non-promoters according to our classification benchmarks. On the contrary, a class of promoters delineated by using a larger window will lead to poorer classification results. For example, for the structural model nucleosome positioning, the classification performance of the *Arabidopsis* core promoters decreases from 97 to 87% and 79% when we enlarge the window size from 1 to 5 and 10, respectively (Figure 4). Thus, enlarging the window size increases, as expected, the smoothness of the structural profile, but eventually too much local information will be lost and only the global information content will be uncovered, rendering the outcome less productive. In the end, the difference between the promoter sequences and non-promoter sequences will become unclear due to an excessive smoothing of the raw data. We have to keep in mind that the classification results for the smaller window sizes are only applicable to a subset of the core promoters, whereas when using a larger window size these results are applicable for almost all core promoters (see above and Supplementary Material).

The classification results, i.e. the success in discriminating between core- promoter and non-core-promoter sequences, are similar for all core-promoter classes and are not only good for those with the highest quality values. A slight decrease in classification power for the clusters with a lower quality label can be observed for some structural models (see Supplementary Material). This can be due to the fact that these classes contain fewer sequences than the classes with a higher quality label, so that the signal of the distinctive structural elements is less pronounced and as a consequence the classification performance is lower. However, the overall classification power remains very high (close to 80%) and shows that even the smaller core-promoters classes contain sufficient structural information to differentiate them from the negative sequences.

Apart from the LSVM for discriminating core-promoters from non-core-promoter sequences, the AQBC method was also used for this purpose. On average, the AQBC method performs ~10% better than the LSVM (data not shown). However, we believe this to be mainly due to the fact that the same technique was also used for the initial delineation of the different core-promoter classes. Owing to this potential bias, we preferred to use the LSVM in evaluating the discriminative power for different structural model characteristics.

Specific structural profiles

The different profiles that correspond to specific core-promoter classes based on certain structural models can give us a more detailed view of the structure of a core promoter and can provide us with information about the various structural elements that describe or delineate the given core-promoter class. Here, we will limit our discussion to one of the structural models (bendability) while other models will be discussed in more detail elsewhere.

Regarding the cluster with the highest quality value, our observations from the bendability profile are in very good

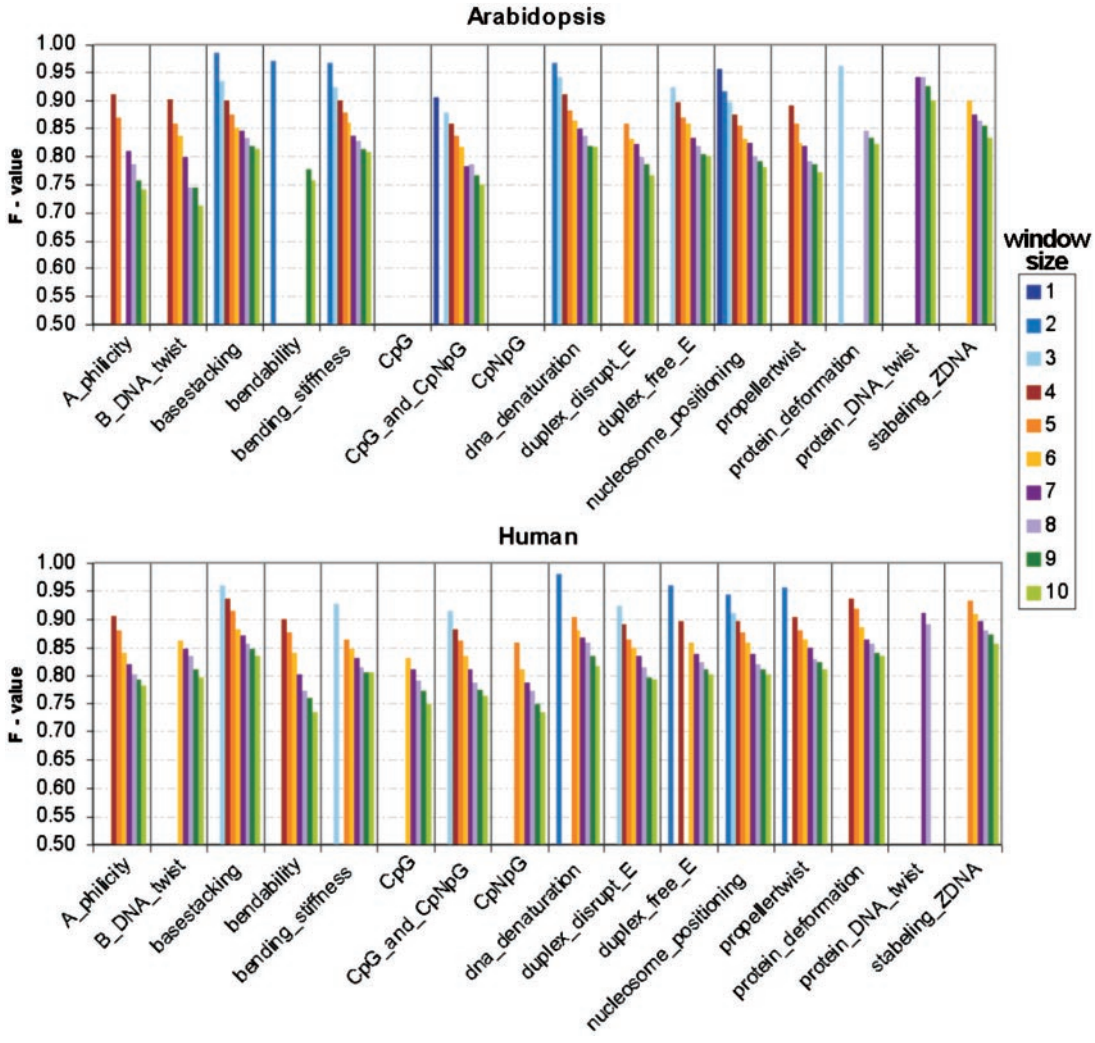

Figure 4. Influence of the window size on the classification results. This shows the discriminative power to distinguish core-promoter sequences from non core-promoter sequences—for all structural models and for window sizes 1–10. For each structural model, all core-promoter sequences from the clusters with the highest quality value were mixed with 75% sequences coming from the dinucleotide-randomized dataset. The *F*-value, which combines sensitivity and specificity, is a measure for the overall performance of discriminating between core-promoter sequences and non-core-promoter sequences. Classification results were based on applying the LSVM classification method.

agreement with those of Pedersen *et al*. (2), who performed a large-scale investigation of three different structural models describing DNA bendability. These authors described a common structural profile for bendability as being low in a region upstream of the TSS and significantly higher downstream of the TSS. As can be observed in Figure 5a, the region upstream of the TSS has indeed a higher bendability value than the region downstream, at least for cluster 1 (highest quality value cluster). More recently, Fukue *et al*. (26) confirmed these observations using the bendability parameters from Brukner *et al*. (11) as a structural model. They investigated in greater detail human promoter sequences that contained only a TATA-box or an Inr-box and concluded that, in addition to the profile observed by Pedersen *et al*. (2), TATA-box-only core promoters have a region between the TATA-box and the transcription start site that is also more flexible.

In contrast to these previous observations, we can clearly distinguish new core-promoter classes that exhibit alternative structural profiles (Figure 5). If we look at these structural profiles, we can clearly see that the different maxima or minima often coincide with regions where known core-promoter elements are believed to be present (Figure 5). For example, a core-promoter class that shows two peaks at positions +1 and −30 could represent a set of promoters that has an Inr-box and a TATA-box at those positions. It is well known that the structure of the TATA-box (at position −30 of the TSS) is very important for the initiation of transcription, because the TATA-binding protein introduces two sharp kinks through the intercalation of phenylalanine side chains on the DNA-binding domain (52–54).

To see whether local maxima or minima in the profiles indeed correspond to previously described boxes, we scanned our core-promoter sequences for known core-promoter elements that are over-represented with respect to the entire promoter set or the negative datasets. In general, we could not find any clear case where a specific core-promoter class could be linked to the presence of known core-promoter motifs. On the contrary, for instance, when screening the

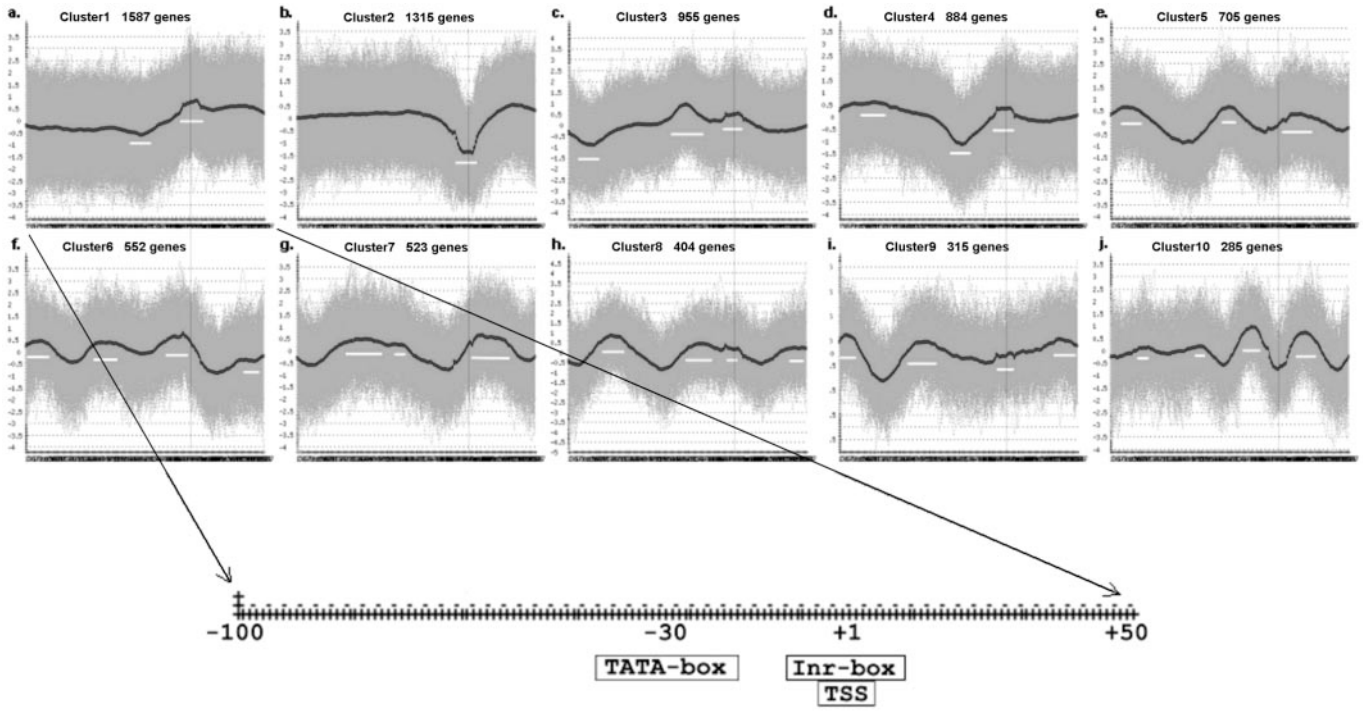

Figure 5. (**a–j**) The first 10 clusters, as inferred by the AQBC method, of human structural profiles obtained using bendability as a structural model with window size 10 are shown. All the core promoters are aligned based on the TSS and each profile corresponds to 100 bp downstream of the TSS and 50 bp upstream.

highest quality cluster of human core promoters (see Figure 5, cluster1), only ~5% of the sequences had a clear TATA-motif. Comparable results are obtained when we look at all the different core-promoter classes and the core-promoter datasets from the other organisms. These results are thus in strong contradiction to what has been previously reported by Bucher (49). However, a similar observation, namely that most human core promoters do not have a clear TATA-motif, was also made recently by Fukue *et al*. (26).

There might be two possible explanations for the apparent discrepancy between the lack of a clear TATA-motif in most of the sequences on the one hand but a clear signal of the profile at the position where one expects the TATA-box on the other hand. One reason could be that the consensus sequence or the PWMs we applied are not very accurate representations of the core-promoter motifs, due to the fact that they were constructed on the basis of a small and very specific group of promoters (47–49). As a result, they may not be able to capture the core-promoter boxes, or structural elements, present at those specific positions in most of the sequences. Alternatively, the 3D structure of the core promoter may be even more important for the initiation of transcription than the presence of the TATA-motif, or another motif, as a binding site (55). This would suggest that a TATA-motif might indeed not be present in the majority of sequences but that the signal in the profile is the result of higher-order structure information that is not being captured by using IUPAC words or a PWM.

In general, we believe that two main conclusions can be drawn from our analyses. The first one is that, based on the overall structure of a promoter, there exist a wide variety of different promoter classes. Nevertheless, these different promoter classes do show some conservation within larger phylogenetic groups (in our case plants and mammals). A wide diversity in core promoters seems to suggest an intricate level of regulation, which has been suggested before. For example, Smale (56) suggested that core promoters could be active contributors to combinatorial regulation. The second striking observation is that the different core-promoter classes cannot be simply depicted or characterized by the use of nucleotide representations. Different classes of core promoters have been described before based on the combination of known core-promoter motifs (34). However, as we have shown here, these core-promoter motifs are generally not sufficient to explain the different core-promoter classes that emerge when we look at the larger structure of a core promoter. Previous findings of over-representation of core-promoter motifs might have been biased because in general these studies were performed on very small and specific datasets (47–49). This is supported by recent genome-wide studies reporting that, for instance, the TATA-box is not a common core-promoter element [(26,57,58) and this study], contrary to what has been assumed previously (49,59).

Our results suggest that the binding of proteins to the DNA is not just a simple interaction between static structures but that it is a mixture of different higher-order interactions between the surfaces of the protein and the DNA helix, which are not necessarily linked to a particular nucleotide sequence.

Relation to computational promoter identification

Recently, there have been attempts to identify promoters on a large scale using strictly computational methods (58,60,61). It was shown that approximately only 50% of promoters could

be correctly predicted from the genomic sequence. On the other hand, it should be noted that these methods suffer from a high false-positive rate or only predict a very specific subset of core promoters, e.g. promoters linked to CpG islands (60). Fickett and Hatzigeorgiou (62) concluded that additional structural features describing a promoter region should be considered when building tools to predict promoter sequences. This conclusion was based on the concept that, although polymerase II promoters are quite different in terms of individual organization, they are embedded into a common genomic content. This concept has been applied in some prediction tools, where, for example, Markov modeling is used to capture similarity between promoters based on certain structural features (63,64). However, in all these modeling approaches, all promoters are expected to behave similarly. We believe that the performance of these tools can be greatly improved by considering that there exist different core-promoter classes, which all have a distinct structural organization. In this respect, the current study hopes to stimulate further research in computational promoter identification, based on the existence of various sub-types of core promoters.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by The Flanders Interuniversity Institute for Biotechnology (VIB).

Conflict of interest statement. None declared.

REFERENCES

1. Sinden,R.R. (1994) *DNA: Structure and Function.* Academic press, San Diego, CA.
2. Pedersen,A.G., Baldi,P., Chauvin,Y. and Brunak,S. (1998) DNA structure in human RNA polymerase II promoters. *J. Mol. Biol.*, **281**, 663–673.
3. Perez-Martin,J. and de Lorenzo,V. (1997) Clues and consequences of DNA bending in transcription. *Annu. Rev. Microbiol.*, **51**, 593–628.
4. Liao,G.C., Rehm,E.J. and Rubin,G.M. (2000) Insertion site preferences of the P transposable element in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA*, **97**, 3347–3351.
5. Lu,X.J., Shakked,Z. and Olson,W.K. (2000) A-form conformational motifs in ligand-bound DNA structures. *J. Mol. Biol.*, **300**, 819–840.
6. Pedersen,A.G., Jensen,L.J., Brunak,S., Staerfeldt,H.H. and Ussery,D.W. (2000) A DNA structural atlas for *Escherichia coli*. *J. Mol. Biol.*, **299**, 907–930.
7. Lemon,B. and Tjian,R. (2000) Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev.*, **14**, 2551–2569.
8. Novina,C.D. and Roy,A.L. (1996) Core promoters and transcriptional control. *Trends Genet.*, **12**, 351–355.
9. Svejstrup,J.Q. (2004) The RNA polymerase II transcription cycle: cycling through chromatin. *Biochim. Biophys. Acta*, **1677**, 64–73.
10. Coulombe,B. and Burton,Z.F. (1999) DNA bending and wrapping around RNA polymerase: a revolutionary model describing transcriptional mechanisms. *Microbiol. Mol. Biol. Rev.*, **63**, 457–478.
11. Brukner,I., Sanchez,R., Suck,D. and Pongor,S. (1995) Trinucleotide models for DNA bending propensity: comparison of models based on DNaseI digestion and nucleosome packaging data. *J. Biomol. Struct. Dyn.*, **13**, 309–317.
12. Dickerson,R.E. and Chiu,T.K. (1997) Helix bending as a factor in protein/DNA recognition. *Biopolymers*, **44**, 361–403.
13. Dickerson,R.E. (1998) DNA bending: the prevalence of kinkiness and the virtues of normality. *Nucleic Acids Res.*, **26**, 1906–1926.
14. Munteanu,M.G., Vlahovicek,K., Parthasarathy,S., Simon,I. and Pongor,S. (1998) Rod models of DNA: sequence-dependent anisotropic elastic modelling of local bending phenomena. *Trends Biochem. Sci.*, **23**, 341–347.
15. Strahs,D. and Schlick,T. (2000) A-tract bending: insights into experimental structures by computational models. *J. Mol. Biol.*, **301**, 643–663.
16. Gerland,U., Moroz,J.D. and Hwa,T. (2002) Physical constraints and functional characteristics of transcription factor–DNA interaction. *Proc. Natl Acad. Sci. USA*, **99**, 12015–12020.
17. Steffen,N.R., Murphy,S.D., Lathrop,R.H., Opel,M.L., Tolleri,L. and Hatfield,G.W. (2002) The role of DNA deformation energy at individual base steps for the identification of DNA–protein binding sites. *Genome Inform. Ser. Workshop Genome Inform.*, **13**, 153–162.
18. Steffen,N.R., Murphy,S.D., Tolleri,L., Hatfield,G.W. and Lathrop,R.H. (2002) DNA sequence and structure: direct and indirect recognition in protein–DNA binding. *Bioinformatics*, **18** (Suppl. 1), S22–S30.
19. Borukhov,S. and Nudler,E. (2003) RNA polymerase holoenzyme: structure, function and biological implications. *Curr. Opin. Microbiol.*, **6**, 93–100.
20. Barton,M.C., Madani,N. and Emerson,B.M. (1997) Distal enhancer regulation by promoter derepression in topologically constrained DNA *in vitro*. *Proc. Natl Acad. Sci. USA*, **94**, 7257–7262.
21. Bagga,R., Michalowski,S., Sabnis,R., Griffith,J.D. and Emerson,B.M. (2000) HMG I/Y regulates long-range enhancer-dependent transcription on DNA and chromatin by changes in DNA topology. *Nucleic Acids Res.*, **28**, 2541–2550.
22. Audit,B., Vaillant,C., Arneodo,A., d'Aubenton-Carafa,Y. and Thermes,C. (2002) Long-range correlations between DNA bending sites: relation to the structure and dynamics of nucleosomes. *J. Mol. Biol.*, **316**, 903–918.
23. Marilley,M. and Pasero,P. (1996) Common DNA structural features exhibited by eukaryotic ribosomal gene promoters. *Nucleic Acids Res.*, **24**, 2204–2211.
24. Gabrielian,A.E., Landsman,D. and Bolshoy,A. (1999) Curved DNA in promoter sequences. *In Silico Biol.*, **1**, 183–196.
25. Marilley,M., Radebaugh,C.A., Geiss,G.K., Laybourn,P.J. and Paule,M.R. (2002) DNA structural variation affects complex formation and promoter melting in ribosomal RNA transcription. *Mol. Genet. Genomics*, **267**, 781–791.
26. Fukue,Y., Sumida,N., Nishikawa,J. and Ohyama,T. (2004) Core promoter elements of eukaryotic genes have a highly distinctive mechanical property. *Nucleic Acids Res.*, **32**, 5834–5840.
27. Suzuki,Y., Tsunoda,T., Sese,J., Taira,H., Mizushima-Sugano,J., Hata,H., Ota,T., Isogai,T., Tanaka,T., Nakamura,Y. *et al.* (2001) Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res.*, **11**, 677–684.
28. Suzuki,Y., Taira,H., Tsunoda,T., Mizushima-Sugano,J., Sese,J., Hata,H., Ota,T., Isogai,T., Tanaka,T., Morishita,S. *et al.* (2001) Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Rep.*, **2**, 388–393.
29. Suzuki,Y., Yamashita,R., Nakai,K. and Sugano,S. (2002) DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res.*, **30**, 328–331.
30. Suzuki,Y., Yamashita,R., Sugano,S. and Nakai,K. (2004) DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. *Nucleic Acids Res.*, **32**, D78–D81.
31. Seki,M., Narusaka,M., Kamiya,A., Ishida,J., Satou,M., Sakurai,T., Nakajima,M., Enju,A., Akiyama,K., Oono,Y. *et al.* (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science*, **296**, 141–145.
32. Rombauts,S., Florquin,K., Lescot,M., Marchal,K., Rouze,P. and van de Peer,Y. (2003) Computational approaches to identify promoters and *cis*-regulatory elements in plant genomes. *Plant Physiol.*, **132**, 1162–1176.
33. Fickett,J.W. and Wasserman,W.W. (2000) Discovery and modeling of transcriptional regulatory regions. *Curr. Opin. Biotechnol.*, **11**, 19–24.
34. Butler,J.E. and Kadonaga,J.T. (2002) The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev.*, **16**, 2583–2592.
35. Panstruga,R., Buschges,R., Piffanelli,P. and Schulze-Lefert,P. (1998) A contiguous 60 kb genomic stretch from barley reveals molecular evidence for gene islands in a monocot genome. *Nucleic Acids Res.*, **26**, 1056–1062.

36. Antequera,F. and Bird,A. (1999) CpG islands as genomic footprints of promoters that are associated with replication origins. *Curr. Biol.*, **9**, R661–R667.
37. Ashikawa,I. (2001) Gene-associated CpG islands in plants as revealed by analyses of genomic sequences. *Plant J.*, **26**, 617–625.
38. Takai,D. and Jones,P.A. (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl Acad. Sci. USA*, **99**, 3740–3745.
39. Gardiner-Garden,M. and Frommer,M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.*, **196**, 261–282.
40. Heyer,L.J., Kruglyak,S. and Yooseph,S. (1999) Exploring expression data: identification and analysis of coexpressed genes. *Genome Res.*, **9**, 1106–1115.
41. Sherlock,G. (2000) Analysis of large-scale gene expression data. *Curr. Opin. Immunol.*, **12**, 201–205.
42. De Smet,F., Mathys,J., Marchal,K., Thijs,G., De Moor,B. and Moreau,Y. (2002) Adaptive quality-based clustering of gene expression profiles. *Bioinformatics*, **18**, 735–746.
43. Workman,C. and Krogh,A. (1999) No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res.*, **27**, 4816–4822.
44. Katz,L. and Burge,C.B. (2003) Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res.*, **13**, 2042–2051.
45. Boser,B., Guyon,I. and Vapnik,V.N. (1992) A training algorithm for optimal margin classifiers. In Haussler,D. (ed.), *Proceedings of COLT*. ACM Press, New York, NY, USA, pp. 144–152.
46. Vapnik,V.N. (1995) *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin.
47. Lagrange,T., Kapanidis,A.N., Tang,H., Reinberg,D. and Ebright,R.H. (1998) New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes Dev.*, **12**, 34–44.
48. Kadonaga,J.T. (2002) The DPE, a core promoter element for transcription by RNA polymerase II. *Exp. Mol. Med.*, **34**, 259–264.
49. Bucher,P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.*, **212**, 563–578.
50. Beckstette,M., Strothmann,D., Homann,R., Giegerich,R. and Kurtz,S. (2004) PoSSuMsearch: fast and sensitive matching of position specific scoring matrices using enhanced suffix arrays. In *Proceedings of the German Conference on Bioinformatics 2004. GI Lecture Notes in Informatics*, **53**, 53–64.
51. Baldi,P., Chauvin,Y., Brunak,S., Gorodkin,J. and Pedersen,A.G. (1998) Computational applications of DNA structural scales. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 35–42.
52. Juo,Z.S., Chiu,T.K., Leiberman,P.M., Baikalov,I., Berk,A.J. and Dickerson,R.E. (1996) How proteins recognize the TATA box. *J. Mol. Biol.*, **261**, 239–254.
53. Patikoglou,G.A., Kim,J.L., Sun,L., Yang,S.H., Kodadek,T. and Burley,S.K. (1999) TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes Dev.*, **13**, 3217–3230.
54. Wu,J., Parkhurst,K.M., Powell,R.M., Brenowitz,M. and Parkhurst,L.J. (2001) DNA bends in TATA-binding protein–TATA complexes in solution are DNA sequence-dependent. *J. Biol. Chem.*, **276**, 14614–14622.
55. Leblanc,B.P., Benham,C.J. and Clark,D.J. (2000) An initiation element in the yeast CUP1 promoter is recognized by RNA polymerase II in the absence of TATA box-binding protein if the DNA is negatively supercoiled. *Proc. Natl Acad. Sci. USA*, **97**, 10745–10750.
56. Smale,S.T. (2001) Core promoters: active contributors to combinatorial gene regulation. *Genes Dev.*, **15**, 2503–2508.
57. Molina,C. and Grotewold,E.R. (2005) Genome wide analysis of *Arabidopsis* core promoters. *BMC Genomics*, **6**, 25.
58. Ohler,U., Liao,G.C., Niemann,H. and Rubin,G.M. (2002) Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol.*, **3**, RESEARCH0087.
59. Shahmuradov,I.A., Gammerman,A.J., Hancock,J.M., Bramley,P.M. and Solovyev,V.V. (2003) PlantProm: a database of plant promoter sequences. *Nucleic Acids Res.*, **31**, 114–117.
60. Down,T.A. and Hubbard,T.J. (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.*, **12**, 458–461.
61. Bajic,V.B., Seah,S.H., Chong,A., Krishnan,S.P., Koh,J.L. and Brusic,V. (2003) Computer model for recognition of functional transcription start sites in RNA polymerase II promoters of vertebrates. *J. Mol. Graph Model.*, **21**, 323–332.
62. Fickett,J.W. and Hatzigeorgiou,A.G. (1997) Eukaryotic promoter recognition. *Genome Res.*, **7**, 861–878.
63. Ohler,U., Niemann,H., Liao,G.C. and Rubin,G.M. (2001) Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics*, **17** (Suppl. 1), S199–S206.
64. Bajic,V.B., Choudhary,V. and Hock,C.K. (2004) Content analysis of the core promoter region of human genes. *In Silico Biol.*, **4**, 109–125.
65. Ornstein,R.L., Rein,R., Breen,D.L. and Macelroy,R.D. (1987) An optimized potential function for the calculation of nucleic acid interaction energies: base stacking. *Biopolymers*, **17**, 2341–2360.
66. el Hassan,M.A. and Calladine,C.R. (1996) Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. *J. Mol. Biol.*, **259**, 95–103.
67. Satchwell,S.C., Drew,H.R. and Travers,A.A. (1986) Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.*, **191**, 659–675.
68. Ivanov,V.I. and Minchenkova,L.E. (1995) The A-form of DNA: in search of the biological role. *Mol. Biol.*, **28,**, 780–788.
69. Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proc. Natl Acad. Sci. USA*, **95**, 11163–11168.
70. Breslauer,K.J., Frank,R., Blocker,H. and Marky,L.A. (1986) Predicting DNA duplex stability from the base sequence. *Proc. Natl Acad. Sci. USA*, **83**, 3746–3750.
71. Sugimoto,N., Nakano,S., Yoneyama,M. and Honda,K. (1996) Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res.*, **24**, 4501–4505.
72. Blake,R.D. and Delcourt,S.G. (1998) Thermal stability of DNA. *Nucleic Acids Res.*, **26**, 3323–3332.
73. Blake,R.D., Bizzaro,J.W., Blake,J.D., Day,G.R., Delcourt,S.G., Knowles,J., Marx,K.A. and SantaLucia,J.,Jr (1999) Statistical mechanical simulation of polymeric DNA melting with MELTSIM. *Bioinformatics*, **15**, 370–375.
74. Sivolob,A.V. and Khrapunov,S.N. (1995) Translational positioning of nucleosomes on DNA: the role of sequence-dependent isotropic DNA bending stiffness. *J. Mol. Biol.*, **247**, 918–931.
75. Gorin,A.A., Zhurkin,V.B. and Olson,W.K. (1995) B-DNA twisting correlates with base-pair morphology. *J. Mol. Biol.*, **247**, 34–48.
76. Ho,P.S., Zhou,G.W. and Clark,L.B. (1990) Polarized electronic spectra of Z-DNA single crystals. *Biopolymers*, **30**, 151–163.