Computational Applications of DNA Structural Scales

Pierre Baldi[*]
Net-ID, Inc.
Los Angeles, CA 90042
pfbaldi@netid.com

Yves Chauvin
Net-ID, Inc.
San Francisco, CA 94107
yves@netid.com

Søren Brunak   Jan Gorodkin   Anders Gorm Pedersen
Center for Biological Sequence Analysis
The Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark
brunak@cbs.dtu.dk   gorodkin@cbs.dtu.dk   gorm@cbs.dtu.dk

Abstract

We study from a computational standpoint several different physical scales associated with structural features of DNA sequences, including dinucleotide scales such as base stacking energy and propeller twist, and trinucleotide scales such as bendability and nucleosome positioning. We show that these scales provide an alternative or complementary compact representation of DNA sequences. As an example we construct a strand invariant representation of DNA sequences. The scales can also be used to analyze and discover new DNA structural patterns, especially in combinations with hidden Markov models (HMMs). The scales are applied to HMMs of human promoter sequences revealing a number of significant differences between regions upstream and downstream of the transcriptional start point. Finally we show, with some qualifications, that such scales are by and large independent, and therefore complement each other.

Keywords: promoters, strand invariance, DNA structure, bendability, nucleosomes

Introduction

DNA three-dimensional structure is essential to DNA function and depends 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; Nussinov 1985; Shapiro *et al.* 1986; Satchwell, Drew, & Travers 1986; Calladine, Drew, & McCall 1988; Bolshoy *et al.* 1991; Dickerson 1992; Hunter 1993; Goodsell & Dickerson 1994; Brukner *et al.* 1995; Hunter 1996).

Based on different kinds of experimental data, several models for estimating DNA structure from di- or trinucleotides have been devised (Sinden 1994). Notable examples of the resulting physical scales include the stacking

energy (Ornstein *et al.* 1978) and propeller twist (El Hassan & Calladine 1996) dinucleotide scales (Table 1), and the bendability (Brukner *et al.* 1995) and nucleosome positioning (Goodsell & Dickerson 1994) trinucleotide scales (not shown here for lack of space, but easily found in the references).

Table 1: Dinucleotide physical scales

Dinucleotide	Stacking Energy (kcal/mole)	Propeller Twist (degrees)
AA	-5.37	-18.66
AC	-10.51	-13.10
AG	-6.78	-14.00
AT	-6.57	-15.01
CA	-6.57	-9.45
CC	-8.26	-8.11
CG	-9.69	-10.03
CT	-6.78	-14.00
GA	-9.81	-13.48
GC	-14.59	-11.08
GG	-8.26	-8.11
GT	-10.51	-13.10
TA	-3.82	-11.85
TC	-9.81	-13.48
TG	-6.57	-9.45
TT	-5.37	-18.66

Here we apply and analyze such scales from three different standpoints that support their usefulness for a variety of computational tasks. First, we show how the scales can be used to provide invariant representations of DNA sequences. Second, we show how such scales can be applied to the analysis of specific DNA sequences, using human promoter sequences as an example. Third, we study the independence properties of such scales.

---

[*] Corresponding author

Invariant Representations of DNA

DNA Representation: Invariance and Compactness

Due to the complementarity of base pairs, the composition of a piece of DNA can be specified by giving the sequence of nucleotides in just one of the two strands. Although this notation is simple and functional it may pose problems in computational analysis of DNA sequences (Baldi & Brunak 1998). As an example consider promoter prediction. It is well known that some promoter elements are functional, independently of which orientation they have. These elements are usually binding sites for transcription factors, and such a lack of orientation dependence may occur if the transcription factor exerts its effect by unspecific protein-protein interactions with the basal transcriptional machinery. If the sequence of just one DNA strand in a set of promoters is presented to a prediction algorithm, the method will therefore essentially have to learn that the binding site and its complement are the same element. It seems reasonable to expect that recognition algorithms could benefit if this knowledge was available explicitly rather than being buried in the data sets. For this class of problems it would therefore be of interest to develop DNA sequence encodings that are directionally invariant in the sense that a sequence and its complement traversed in the same direction are encoded identically. However, some features in DNA sequences *do* depend on the orientation. *E.g.*, the orientation of a TATA-box has an influence on the position of the transcriptional start point and the direction of transcription (Wang, Jensen, & Stumph 1996). Thus, an encoding that is strand invariant but *not* directionally invariant (in the sense that a DNA sequence and its complement are represented by the same string of symbols in opposite orientations) might also be useful. Besides the matter of invariance, the compactness of sequence encoding is also problematic. Thus, DNA sequences are often presented to numerical information extraction algorithms—such as neural networks—using a sparse but wasteful binary encoding of the form A=(1,0,0,0), C=(0,1,0,0), *etc*. Sparse encoding has proven to be superior to some compact coding schemes, presumably because it does not introduce algebraic dependencies (Demeler & Zhou 1991). A less compact encoding, however, typically requires more parameters introducing a greater risk of overfitting. It seems, therefore, that although the sparse representation has proven to be very useful, it has certain shortcomings and it is natural to search for alternative or complementary analog representations.

Since a di- or trinucleotide and its complement by definition have identical structural properties, the structural scales are inherently strand invariant and may therefore be good candidates for encoding schemes that meet the criterions mentioned above. It is likely that such encodings will be most successful when the quantity measured by the scale is directly relevant for the type of information being extracted. Most often, however, such relationships are unknown a priori and it may therefore be wise to test for them systematically. In other words, it may be useful to revisit some classical pattern recognition neural networks and retrain them using input representations based on physical scales.

Strand Invariance

To each nucleotide in a DNA sequence corresponds a unique stacking energy or propeller twist value, but given a stacking energy value we can only recover a dinucleotide up to strand symmetry. For instance, the propeller twist value of -13.10 degrees corresponds to to both AC and its reverse complement GT. Thus if we encode a nucleotide sequence by the corresponding sequence of propeller twist values we obtain a strand invariant encoding, except for the order of the numbers in the sequence[1]. With a small modification, however, this provides also a mean of constructing strand invariant representations for sequences of length greater than 2.

Consider the triplet ACT: the corresponding sequence of propeller twist values is $(-13.10, -14.00)$. The inverse complementary triplet on the other strand is AGT associated with the same sequence of propeller twist values but in reverse order $(-14.00, -13.10)$. We would like to build a representation where ACT and TGA are represented in the same way. This is easily achieved by using two symmetric functions, such as the sum $S = -27.10$ and the product $P = +183.4$. By solving a simple quadratic equation, from the value of S and P it is easy to recover uniquely the *unordered* pair $\{-13.10, -14.00\}$. Obviously any other pair of symmetric and reversible functions would do the job and we do not mean to imply here that S and P are necessarily the best choice. In fact, because of the particular discrete values of the propeller twist scale it is easy to check that a sum S is associated with a unique unordered pair of propeller twist values, although this requires reasonable numerical precision. As a result S alone suffice to provide a strand invariant encoding for any trinucleotide.

General Case

Consider a sequence of $n + 1$ nucleotides $A_1, \ldots, A_{n+1}$, and the associated sequence of propeller twist values $(a_1, \ldots, a_n)$. The inverse complementary sequence on the opposite strand is $\bar{A}_{n+1}, \ldots, \bar{A}_1$ with the scale sequence $(a_n, \ldots, a_1)$. We would like to find an encoding $E(A_1, \ldots, A_{n+1}) = E(a_1, \ldots, a_n)$ with two properties. First, it must be strand invariant in the sense that $E(A_1, \ldots, A_{n+1}) = E(\bar{A}_{n+1}, \ldots, \bar{A}_1) = E(a_n, \ldots, a_1)$. Second, we must be able to recover the sequence itself (and its inverse complementary form) from the encoding E. We have shown above that such encoding exists for sequences of length 2 and 3. To extend this process, we recursively construct the following encoding

$$E(A_1, \ldots, A_{n+1}) = \\ (a_1 + a_n, a_1 a_n, [a_1 + a_i][a_n + a_{n-i+1}], E(A_2, \ldots, A_n)) \quad (1)$$

where i (or $n - i + 1$) denotes the first position where $(a_2, \ldots, a_{n-1})$ differs from $(a_{n-1}, \ldots, a_2)$. When $n = 2$ or $n = 3$ the encoding is the one discussed in Section2. By induction $E(A_2, \ldots, A_n)$ is strand invariant. The functions appearing in the first 3 positions of the encoding are all symmetric with respect to the mirroring operation and therefore

---

[1]There is a small exception in the base stacking energy scale because -6.57 corresponds to CA and TG, but also to AT (which is self-complementary)

the proposed encoding is strand invariant. To recover the original sequence up to its strand position, suppose we are given an encoding of the form (x, y, z, E). By induction, from E we can recover the sequence $A_2, \ldots, A_n$ and its inverse complement $\bar{A}_n, \ldots, \bar{A}_2$. From x and y we can recover the unordered pair $\{a_1, a_n\}$. This yields two alternative solutions

$$\begin{cases} (a_1, A_2, \ldots, A_n, a_n) & \text{and} & (a_n, \bar{A}_n, \ldots, \bar{A}_2, a_1) \\ \text{or} \\ (a_n, A_2, \ldots, A_n, a_1) & \text{and} & (a_1, \bar{A}_n, \ldots, \bar{A}_2, a_n). \end{cases} \quad (2)$$

The notation uses the fact that for any nucleotide, there exist a unique left nucleotide and unique right nucleotide with which it can form a dinucleotide with a given propeller twist value. We can now use $[a_1 + a_n] + [a_i + a_{n-i+1}]$ and z to recover the unordered pair $\{a_1 + a_i, a_n + a_{n-i+1}\}$ and use it to break the tie in Equation 2 in favor of the first solution. The length of this encoding grows like $3n$. It is not unique nor optimal in any sense. The usefulness of a strand invariant encoding as an input to a neural network algorithm depends also on other factors, such as how local, symmetric and complex are the functions it uses. An example of more compact (growing like 1.5n), local and symmetric encoding is provided by:

$$E(a_1, \ldots, a_{2p}) = (S_1, P_1, \ldots, S_p, P_p, P_1^2, \ldots, P_{p-1}^2) \quad (3)$$

where $S_i = a_i + a_{2p-i+1}$, $P_i = a_i a_{2p-i+1}$, and $P_i^2 = [a_i + a_{i+1}][a_{2p-i+1} + a_{2p-i}]$ are all strand symmetric functions. We leave as an exercise for the reader to determine the corresponding form for sequences of even length, and to prove strand invariance. The same ideas apply of course to other strand invariant scales, as well as both strand and direction invariance, and to other types of sequences with similar invariances. Experiments are in progress to test whether such invariant representations can be used to improve neural network performance.

Promoter Applications
Compact Encoding: TATA Boxes

We have recently investigated the structure of a large set of human promoters (A. G. Pedersen, P. Baldi, Y. Chauvin, and S. Brunak, submitted) and found what appears to be a general structural profile. Therefore, promoter prediction algorithms may be one example where coding schemes based on structural measures are relevant. Here we test whether the bendability scale can be used to provide compact representations of TATA box regions in conjunction with neural network algorithms.

Data was extracted from the GenBank nucleotide database, release 95 (Benson *et al.* 1997). Specifically, all human sequences that contained at least 250 nucleotides upstream and downstream from an experimentally determined transcriptional start point were extracted. Sequences containing non-nucleotide symbols were excluded. Redundancy was reduced using algorithm 2 from (Hobohm *et al.* 1992) and a novel method for finding a similarity cut-off (A. G. Pedersen, H. Nielsen and S. Brunak, in preparation). Briefly,

this method is based on performing all pairwise alignments for a data set, fitting the resulting Smith-Waterman scores to an extreme value distribution (Altschul *et al.* 1994; Waterman 1995), and choosing a value above which there are more observations than expected from the distribution.

In one experiment, we extracted 127 sequences with an annotated TATA box of length between 5 and 8 nucleotides and constructed two sets of examples consisting of windows that are 37 nucleotides long. In one set, a window is considered positive if it has a TATA-box starting 4 nucleotides from the left border, while in the other set a windows is considered positive if it has a TATA-box starting in the middle, 18 nucleotides from each border. The first set provides examples containing very little upstream context but enough downstream context to include the transcriptional start point. Negative examples were generated by keeping every fifth of all other windows of length 37. Thus each data set consists of 11881 examples, 127 of which are positive.

Three encoding strategies were used: (1) standard sparse encoding [A=(1,0,0,0), C=(0,1,0,0), G=(0,0,1,0), T=(0,0,0,1)], (2) bendability encoding, where the bendability of a triplet is used to encode the middle nucleotide, and (3) a combination of the two: each nucleotide is encoded in a sparse fashion as above, but instead of 1 the bendability value of the corresponding triplet is used, as a way of presenting the network simultaneously with both types of information. For instance, A may be encoded as (0.127,0,0,0) or (−0.024,0,0,0) etc. Notice that the bendability encoding used here is compact but not strand invariant.

The two data sets, encoded using either of the three methods described above, were used as input to a standard feed-forward neural network, containing two hidden units and one output unit representing the probability of the input window being a member of the positive class. Networks were trained by gradient descent with a learning rate of 0.1, and using the cross-entropy error function. Choices of architecture and learning rate were determined from a number of preliminary pilot experiments. The networks trained using the sparse and combination encoding had 148 input units and therefore a total of 301 parameters, including 3 thresholds. The network with the bendability encoding had 37 input units, and a total of 79 parameters only, including 3 thresholds. All networks were trained on-line for a thousand epochs. In all cases, the training performance converged before the end of the 1000 epochs, while test performance generally peaked early and then deteriorated by overfitting.

In each case, we used 6-fold cross validation to address the overfitting problem. Specifically, examples were partitioned into 6 equal subsets. For each of 6 different permutations, the networks are trained on 4/6 examples (7921), 1/6 examples (1980) are used for early stopping, and performance is assessed on the remaining 1/6. Thus in a typical stop or test set there are about 21 positive examples. Overall performance is then the average of the 6 test results. We calculate also the average of the best performance in all 12 test sets (6 stop and 6 evaluation). Performance for all three types of encoding and both classes of positive examples was evaluated using the Mathews correlation coefficient

(Table 2, "stop" denotes performance assessed using distinct stop and test sets).

Sparse encoding consistently performs a little better than the other two representations (Table 2). The bendability encoding, however, in spite of its lack of independence actually performs almost as well *with about one fourth the number of parameters* (Table 2).

Table 2: Performance of TATA-box recognizing neural nets.

	TATA at 5		TATA at 19	
	stop	non-stop	stop	non-stop
Sparse	0.6719	0.7487	0.6884	0.7336
Bend	0.4887	0.5824	0.5651	0.6070
Comb	0.5328	0.6426	0.5631	0.6560

Pattern Detection: Profiles

DNA or protein scales can be effectively combined with hidden Markov models (HMMs) (Baldi & Brunak 1998). In particular, any DNA scale can be convolved with the parameters of an HMM to produce an expected profile for the corresponding property. Alternatively, the scales can be applied directly to the corresponding HMM-derived multiple alignment with indels.

In (Baldi *et al.* 1997), a new weak periodic pattern was detected in human exon and intron DNA sequences using a number of different HMMs. This statistical pattern is characterized by the consensus pattern [non-T][A or T][G] and a periodicity of roughly 10 nucleotides. From a structural point of view, this periodicity is interesting since it is well known that "bent DNA" requires a number of small individual bends that are in phase. Only when bends are phased at approximately 10.5 bp (corresponding to one full turn of the double helix) can stable long-range curvature be obtained. The pattern found is related to the DNase I-derived bendability trinucleotide scale (Brukner *et al.* 1995). (DNase I interacts with the surface of the minor groove, and bends the DNA molecule away from the enzyme.) In fact, five (ATG, CAG, CTG, GAG, GTG) of the six triplets associated with the consensus pattern are found in the high end of the bendability scale. These results are consistent with the sequence signal having a role in nucleosome positioning, and it is possible that the differences that are observed between the strength of signals in coding and non-coding regions have implications for the recognition of genes by the transcriptional machinery.

Here, a standard linear HMM architecture with length $N = 500$ was trained using the redundancy-reduced promoter data set described above containing 625 sequences, all with length 501, *i.e.*, 250 nucleotides up- and downstream of the transcriptional start point. The training was facilitated by initializing the main state emissions associated with the TATA-box using consensus probabilities from promoters with experimentally verified TATA-boxes. We then computed the profile of the HMM backbone for each of the DNA scales (Figure 1), by multiplying the HMM probabilities with the scale values and averaging over a sliding window of length 21, as in (Baldi & Brunak 1998). We have checked that the main results are robust over sliding window sizes in the range of 3 to 31. All profiles consistently show a large signal around the transcriptional start point (position +1) with differences between the upstream and downstream regions. The most striking feature is a significant increase in bendability in the region immediately *downstream* of the transcriptional start point. As promoters most often have been characterized by a number of upstream patterns and compositional tendencies, it is interesting that the HMM alignment corresponds to structural similarity in the downstream region of these otherwise unrelated promoter sequences. The signature around the transcriptional start point is not the result of conservation since the sequences are highly variable. We also checked that randomly generated scales do not yield a consistent signal around the transcription start point. From a careful analysis of the sequence periodicities and composition, we conjecture that the increase in downstream bendability is related to nucleosome positioning and/or facilitation of interaction with other factors involved in transcriptional initiation (A. G. Pedersen, P. Baldi, S. Brunak, Y. Chauvin, submitted). Additional experiments, including the obvious application to promoter prediction, are in progress.

Computational Independence

All the physical DNA scales considered here (bendability, nucleosome positioning, stacking energies, propeller twist) show a signal around the TATA box and the start point with differences between the upstream and the downstream regions. Although these scales are obtained via *completely different* experimental techniques, it is then natural to ask whether there are any computational relationships between the scales, and whether each one of them provides an independent element of supporting evidence or not. Consider, for instance, two scales such as bendability and nucleosome positioning. It is clear from their tables that the value of one of them determines the value of the other one, and the value of the corresponding triplet up to strand invariance. Thus in general a function exists that relates one scale to another. The real question is what is the complexity of the function. Clearly if the function is linear, the promoter signals observed would be a scaled version of each other and the second scale would not bring any new evidence to the results obtained with the first. From inspection of the plots in Figure 1, it is obvious that the relationship is *not* linear. But could it have some other relatively simple form? And how to deal with scales that belong to different dimensions, such as dinucleotides (base stacking) versus trinucleotides scales (bendability)? Such questions can be addressed in a number of ways such as correlation coefficients, and polynomial and/or neural network regression techniques (i.e. looking at the complexity of the neural network required to learn the transformation of one scale into another). Because of the large number of possible comparisons, we present here a sample of our analysis. But in all cases tested so far, we find that the physical scales are largely uncorrelated to each other.
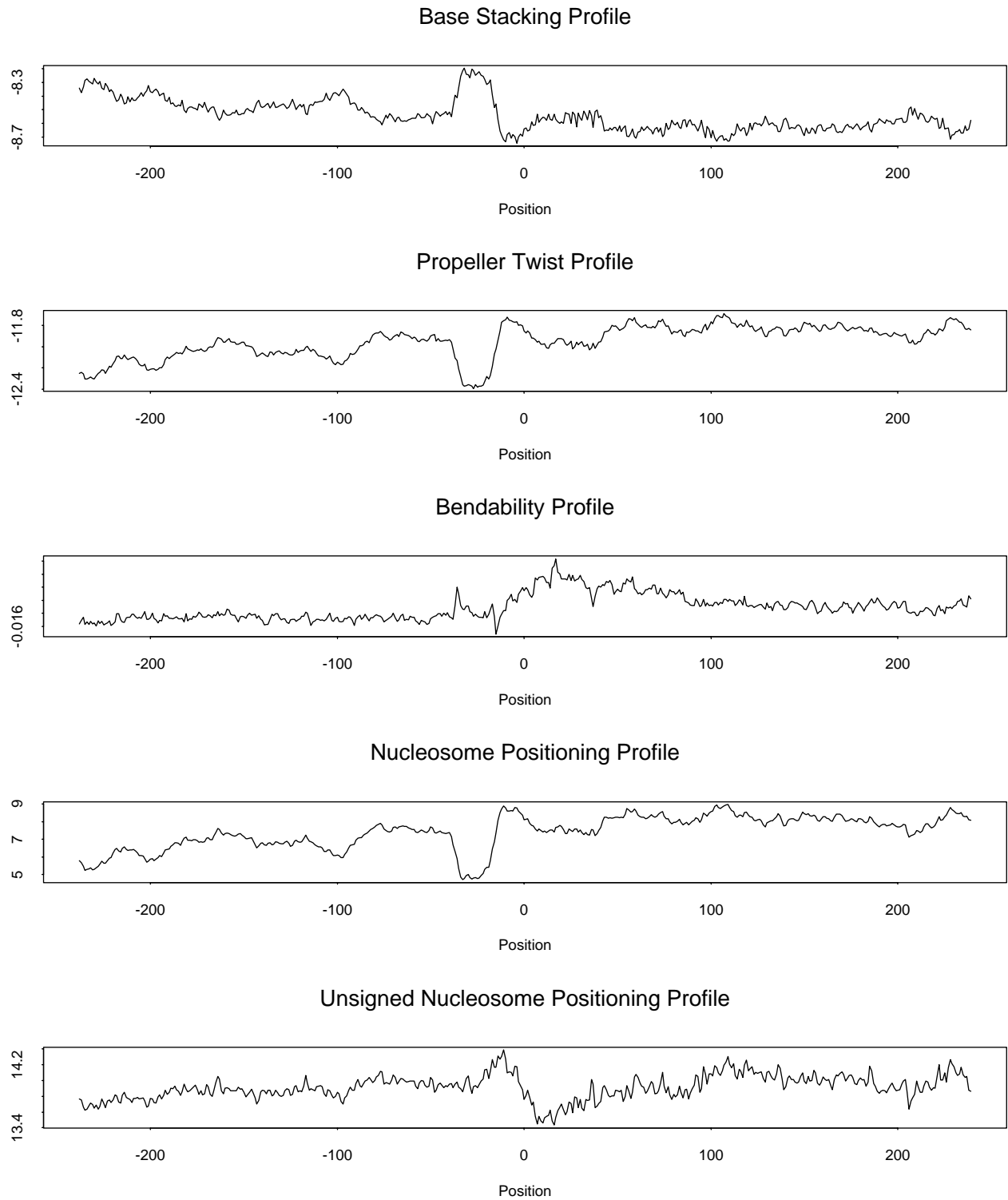
Figure 1: Profile of several scales around human promoter regions derived using an HMM of length 500, with 250 positions upstream and downstream form the transcription start site.

Dinucleotide Scale Correlations

In table 3 it is shown that the correlation between base stacking and propeller twist is fairly small.

Table 3: Correlation between dinucleotide scales

Scale	BS	PT
Base Stacking	1	-0.293
Propeller Twist		1

Trinucleotide Scale Correlations

In the case of trinucleotide scales, we have computed all possible correlations between 8 scales (Table 4, next page). In addition to the bendability and nucleosome positioning scales, we have used the positive nucleosome positioning scale (suggested by Travers and obtained by taking the absolute value of the nucleosome positioning scales), the sum and product of the stacking energies of the two dinucleotides associated with a given triplet, and the similar sum and product of propeller twist values. The use of the sum and product is of course dictated by the considerations discussed in the first section and the need to build a bridge between dinucleotide and trinucleotide scales (see also below).

Most correlation coefficients are small, with two notable exceptions (Table 4). The sum and product of base stacking or propeller twist values are very highly correlated (-0.985 and -0.989). This is easily explained below. There is also a non-trivial correlation between the nucleosome positioning scale and the sum or products associated with the dinucleotide scales (-0.766, 0.753, 0.649, -0.661). We are currently investigating the origin of such mild correlation. Remarkably, such correlation is absent in the positive nucleosome positioning scale. This suggests that the positive version of the nucleosome positioning is a better tool for providing new independent evidence of a signal. It must also be noted that such correlations were computed using a *uniform* distribution across dinucleotides or trinucleotides and can be recomputed using any compositional bias.

Correlations Between Dinucleotide and Trinucleotides Scales

As an example, let us consider bendability and base stacking energy. Without any other information, it may be reasonable to make the hypothesis that they share a close relationship. For instance the higher the energy the greater the stiffness (or vice versa). As we shall see this is not the case.

For any triplet A_1, A_2, A_3 there is a unique bendability B shared by the complementary triplet on the other strand. There is also a unique pair of stacking energies a_1 and a_2 up to permutation. Using the strand invariant encoding of the first section, we see that for any one of the 32 possible values of B, there is a unique pair (S, P), with $S = a_1 + a_2$ and $P = a_1 a_2$. In fact, there is a unique S and a unique P. So now we can focus on the relationship between B and S, and B and P (Figure 2). One can also plot B as a function of a_1 and a_2 as a surface, with a symmetry around the $a_1 = a_2$ plane. By looking at the plots one notices that the (S, B) curve is almost exactly symmetric with respect to the (P, B) curve (Figure 2). This is easy to explain: the stacking energies are all negative and below -1. So a very small S (i.e. a_1 and a_2 very negative) correspond to a very large positive P. So we can focus on the (S, B) curve (Figure 2).
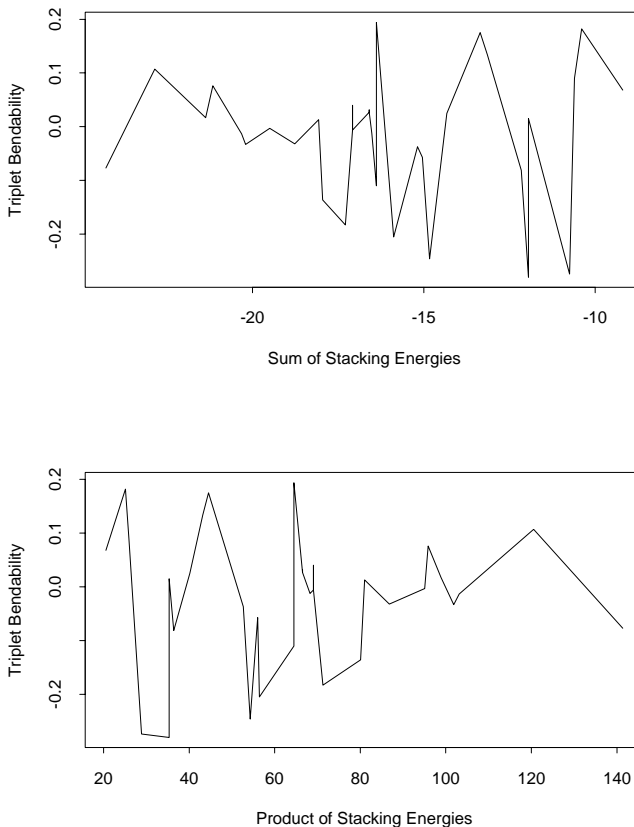

Figure 2: Bendability of 32 single strand triplets as a function of the sum and product of the base stacking energy of the corresponding pair of dinucleotides. Notice the position of the consensus triplets in the sum plot: AAG = $(-12.15, -0.081)$, ATG = $(-13.14, 0.134)$, CAG = CTG = $(-13.35, 0.175)$, GAG = $(-16.59, 0.031)$, GTG = $(-17.08, 0.040)$, and the TATA box triplets: ATA = TAT = $(-10.39, 0.182)$. The narrow maximum in the center corresponds to TCA = TGA = $(-16.38, 0.194)$.

By inspection, it is clear that the relationship is far from trivial. It is certainly not linear or quadratic, but multimodal. In connection with the promoter results, it is then useful to look at where the triplets associated with high bendability and with the TATA box are located. This is easily done taking the Brukner bendability scale (Brukner *et al.* 1995) and tracking back the corresponding points. The highest narrow peak in the center correspond to TCA/TGA. The rightmost peak is associated with ATA/TAT the TATA box triplet (so this triplet has very high bendability and very high stacking energy, which in part explains the TATA signal). The other peaks are associated with the 6 (rather 5) high bendability triplets conforming to [non-T][A or T][G] and these are scattered all over the spectrum of stacking energies. In other words, the high-bendability triplets can have low, medium,

Table 4: Correlation between trinucleotide scales

Scale	B	NP	PNP	SBS	PBS	SPT	PPT
Bendability	1	0.272	-0.0079	-0.025	0.022	0.316	-0.393
Nuc. Pos.		1	0.161	-0.766	0.753	0.649	-0.661
Pos. Nuc. Pos.			1	-0.123	0.186	-0.157	0.203
Sum Base Stack.				1	-0.985	-0.550	0.538
Product Base Stack.					1	0.541	-0.526
Sum Prop. Twist						1	-0.989
Product Prop. Twist							1

or high cumulative stacking energy.

Conclusion

We have studied from a computational standpoint several different physical scales associated with DNA sequences, including dinucleotide scales such as base stacking energy and propeller twist, and trinucleotide scales such as bendability and nucleosome positioning. We have shown that these scales are useful as an alternative or complementary representation of DNA sequences. As an example we have constructed a strand invariant representation of DNA sequences and demonstrated the feasibility of a compact encoding for promoter TATA box regions. The scales can be used as well to analyze and discover new DNA patterns, especially in combinations with hidden Markov models (HMMs). We have applied the scales to HMMs of human promoters revealing a number of significant differences between regions upstream and downstream of the transcriptional start point. Finally we have shown with some qualifications, that such scales are by and large uncorrelated, and therefore complement each other. Because multiple codes (triplet, nucleosome positioning, etc.) are embedded in DNA, understanding the flexibility of each one with respect to the constraints posed by the others is important. Our results provide also further evidence of the importance of the bendability code and its flexibility with respect to other measures, such as base stacking energy.

Finally, it must not be forgotten that the DNA scales we have used are only a first order approximation. Evidence reviewed in (Dickerson 1992), for instance, suggests that the twist angle between bases probably depends on more than just the two adjacent bases. The exact range of the dependence in fact is not really known. A better approximation may be derived using the tetranucleotide formed by the two bases before and after the twist angle. Unfortunately, the structure of all possible 256 tetranucleotides is not known. But the methods we have developed are independent of any particular scale, approximation, or oligonucleotide length. They are readily applicable to the new scales, tetranucleotide and other, that will undoubtedly become available with future progress in experimental techniques and as more structural data becomes available. Furthermore, the methods are also applicable in conjunction with computational scales that are parameterised and fitted to the data using neural network representations and machine learning techniques.

References

Altschul, S.; Boguski, M. S.; Gish, W.; and Wootton, J. C. 1994. Issues in searching molecular sequence databases. *Nature Genetics* 6:119–129.

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

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

Benson, D.; Boguski, M.; Lipman, D.; and Ostell, J. 1997. Genbank. *Nucleic Acids Res.* 25:1–6.

Bolshoy, A.; McNamara, P.; Harrington, R. E.; and 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.

Brukner, I.; Sanchez, R.; Suck, D.; and Pongor, S. 1995. Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *EMBO J.* 14:1812–1818.

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

Demeler, B., and Zhou, G. W. 1991. Neural network optimization for E. coli promoter prediction. *Nucleic Acids Res.* 19:1593–9.

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

Dickerson, R. E. 1992. DNA structure from A to Z. *Meth. Enz.* 211:67–111.

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

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

Hassan, M. A. E., 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.

Hobohm, U.; Scharf, M.; Schneider, R.; and Sander, C. 1992. Selection of representative data sets. *Prot. Sci.* 1:409–417.

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.

Klug, A.; Jack, A.; Viswamitra, M. A.; Kennard, O.; Shakked, Z.; and 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.

Nussinov, R. 1985. Large helical conformational deviations from ideal B-DNA and prokaryotic regulatory sites. *J. Theor. Biol.* 115:179–189.

Ornstein, R. L.; Rein, R.; Breen, D. L.; and MacElroy., R. 1978. An optimised potential function for the calculation of nucleic acid interaction energies. I. Base stacking. *Biopolymers* 17:2341–2360.

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

Shapiro, B. A.; Nussinov, R.; Lipkin, L. E.; and Maizel, J. V. 1986. A sequence analysis system encompassing rules for DNA helical distortion. *Nucleic Acids Res.* 14:75–86.

Sinden, R. R. 1994. *DNA structure and function.* San Diego, CA: Academic Press.

Wang, Y.; Jensen, R. C.; and Stumph, W. E. 1996. Role of TATA box sequence and orientation in determining RNA polymerase II/III transcription specificity. *Nucleic Acids Res.* 24:3100–3106.

Waterman, M. S. 1995. *Introduction to Computational Biology.* London, UK: Chapman and Hall.