BIOINFORMATICS

Vol. 15 nos 7/8 1999
Pages 644–653

Investigating extended regulatory regions of genomic DNA sequences

V. N. Babenko, P. S. Kosarev, O. V. Vishnevsky, V. G. Levitsky, V. V. Basin and A. S. Frolov

Laboratory of Theoretical Genetics, Institute of Cytology and Genetics, Lavrentyev Avenue, 10, Novosibirsk, 630090, Russia

Received on December 4, 1998; revised and accepted on April 20, 1999

Abstract

Motivation: *Despite the growing volume of data on primary nucleotide sequences, the regulatory regions remain a major puzzle with regard to their function. Numerous recognising programs considering a diversity of properties of regulatory regions have been developed. The system proposed here allows the specific contextual, conformational and physico-chemical properties to be revealed based on analysis of extended DNA regions.*

Results: *The Internet-accessible computer system RegScan, designed to analyse the extended regulatory regions of eukaryotic genes, has been developed. The computer system comprises the following software: (i) programs for classification dividing a set of promoters into TATA-containing and TATA-less promoters and promoters with and without CpG islands; (ii) programs for constructing (a) nucleotide frequency profiles, (b) sequence complexity profiles and (c) profiles of conformational and physico-chemical properties; (iii) the program for constructing the sets of degenerate oligonucleotide motifs of a specified length; and (iv) the program searching for and visualising repeats in nucleotide sequences. The system has allowed us to demonstrate the following characteristic patterns of vertebrate promoter regions: the TATA box region is flanked by regions with an increased G+C content and increased bending stiffness, the TATA box content is asymmetric and promoter regions are saturated with both direct and inverted repeats.*

Availability: *The computer system RegScan is available via the Internet at http://wwwmgs.bionet.nsc.ru/Systems/RegScan, http://www.cbil.upenn.edu/mgs/systems/regscan/*

Contact: *bob@bionet.nsc.ru*

Introduction

Eukaryotic promoters are DNA sequences providing gene expression regulation at the stage of transcription initiation. According to the modern concept, the eukaryotic promoters are modularly organised (Wingender, 1993; Kel *et al*., 1997). The elementary unit of transcription regulation is represented by *cis*-regulatory elements, which correspond to the transcription factor binding sites, and have a small length of ~5–20 bp (Wingender, 1993). Pairs of transcription factor binding sites brought into proximity form composite elements. The protein–protein interaction of transcription factors within composite elements can provide either synergistic or antagonistic effects on transcription regulation. The typical size of the composite elements is 50–60 bp (Kel *et al*., 1995). Core promoters, which are ~100 bp long (Roeder, 1996), provide for the assembly of basal transcription complexes and are the regulatory elements absolutely necessary for transcription initiation. The set of the regulatory elements listed above, located within the 5′-flanking regions of the genes, together with enhancers and silencers provide for gene-specific transcription regulation depending on cell cycle stage, ontogenetic stage, tissue type and the effect of external factors (Wingender, 1997). The typical size of 5′-flanking regions, significant for transcription regulation of various genes, falls into the range of 500 to several thousand base pairs.

A number of methods and programs for analysing and recognising transcription factor binding sites are currently available (Frech *et al*., 1993; Prestridge *et al*., 1993; Chen *et al*., 1995; Gelfand, 1995; Quandt *et al*., 1995; Wolfertstetter *et al*., 1996; Quandt *et al*., 1997). Prediction of certain sites and their combinations are used to recognise promoters (Kondrahin *et al*., 1995; Prestridge, 1995; Hutchinson, 1996); however, the accuracy of such recognition is low (Fickett and Hatzigeorgiou, 1997).

A considerable increase in the accuracy of recognition can be achieved through taking into account the properties of extended promoter regions, such as the patterns of oligonucleotide distribution in different regions of promoter (Zhang, 1998) or other specific properties of these regions (Baldi *et. al*., 1998).

Here we describe the Internet-accessible computer system RegScan, designed to analyse the extended regulatory regions of eukaryotic genes. This system allows us: (i) to divide the sample of promoters into subsamples of TATA-containing and TATA-less promoters and promoters with and without CpG islands; (ii) to construct nucleotide frequency profiles, sequence

complexity profiles, and profiles of conformational and physico-chemical properties; (iii) to create sets of degenerate oligonucleotide motifs; and (iv) to scan the nucleotide sequences for the occurrence of repeats of various types.

In comparison with previous works analysing DNA structural peculiarities (Baldi *et. al*., 1998), the system presented is supplemented with the possibility of preliminary classification of the initial sample relative to a set of properties significant for regulatory regions (occurrence of a TATA box or CpG island).

The set of programs developed was applied to the analysis of the promoter regions of eukaryotic genes. In the general case, this set of programs can be used for analysing any genomic sequences.

System and methods

The system RegScan (http://wwwmgs.bionet.nsc.ru/Systems/RegScan, http://www.cbil.upenn.edu/mgs/systems/regscan/) consists of six units that are logically divided into two main modules: the classification module contains one unit and the analysis module contains five units.

The system units are implemented by using the C++ computer language equipped with the Web interface written in HTML. The repeat visualisation unit is implemented as a Java applet.

Classification of the source sample

This unit is designed to sort out the source sample on the basis of certain criteria. So far, sorting of TATA-containing and TATA-less promoter sequences and DNA sequences with and without CpG islands is implemented in the classification unit.

Sorting of TATA-containing (TATA+) and TATA-less (TATA–) promoters. The sorting of a source sample into TATA-containing and TATA-less promoters is performed basing on weight matrix method using the TATA-box weight matrix, obtained by Bucher through analysing 502 sequences of vertebrate and invertebrate promoters (Bucher, 1990). Here, the TATA box is represented as a sequence of 15 nucleotides with the core consensus TATAa/tAa/t.

A given sequence is considered as TATA-containing, if the score of any of its motifs within the region from –39 to –9 relative to the transcription start site (TSS) exceeds a given cut-off value for the TATA-box matrix.

Sorting promoters into CpG+ and CpG- subsamples. Cytosine residues in CG dinucleotides are methylated in the major part of vertebrate DNA. However, there are specific DNA regions, the so-called CpG islands, where the CG dinucleotides are non-methylated in all tissues. Characteristics of these DNA islands are (Gardiner-Garden and Frommer, 1987): (i) the length is of over 200 nucleotides; (ii) the content is over 50% G+C; and (iii) ratio of observed/expected CG dinucleotides is more than 0.6.

The boundaries of CpG islands are determined as follows. Obs./Exp. ratio is calculated as:

$$\frac{\text{Number of } CG}{\text{Number of } C \times \text{Number of } G} \times L,$$

where L is the length of the sequence in question. Obs./Exp. CG ratio and G+C percentage are determined within a window of 100 nucleotides in length (L=100) sliding along the sequence with a 1 nucleotide shift. The overlapping windows with (G+C)% over 50 and Obs./Exp. CG ratio over 0.6 are merged; if the resulting CG-rich fragment exceeds 200 nucleotides, it is considered a CpG island.

Methods and algorithms

Construction of position-specific nucleotide frequency profiles over the sample. The construction of nucleotide frequency profiles is performed as follows. A set of phased sequences with a fixed length represents a nucleotide matrix. This matrix is decomposed into a series of successive overlapping windows (columns) of a given size. The number of nucleotides of a definite type N_{ia} (i, the number of a window and a, designation of nucleotides in IUPAC 15 single letter-based code according to Cornish-Bowden (1985) is calculated within each window. The relative nucleotide frequency N_{ia}/N [N is the total number of nucleotides in a window (column) of the nucleotide matrix] is plotted at the position corresponding to the window's centre.

Estimation of nucleotide sequence complexity. Nucleotide sequence complexity is determined in terms of the model of Kolmogorov (1965) and Lempel and Ziv (1976). This model considers each finite string as a mixture of random symbols and repeated fragments. Kolmogorov was first to describe the complexity of an object as a finite automation-generated model. Lempel and Ziv applied this theory to describe finite sequences. Gusev *et al*. expanded the model through introducing the concepts of symmetry and complementarity (Gusev *et al*., 1993). Currently, variants of complexity measures are widely used for analysing DNA sequences (Allison and Yee, 1990; Milosavlievich and Jurka, 1993; Grumbach and Tahi, 1994; Allison *et al*., 1998).

In our case the complexity measure may be defined as the least number of events required to generate a given sequence algorithmically.

Events are as follows: (i) generation of a new symbol; and (ii) copying of a longest fragment from the already generated sequence portion in the following orientations: direct (D), symmetrical (S), inverted (I) or direct complementary (C).

A subset of orientations for copying can be specified from the set {D, S, I, C}. Then, different variants of the complexity measure reflect particular properties of a nucleotide sequences. The DSIC complexity, allowing any copying operations, is the most general.

For example, the DSIC complexity of the sequence ATGCATCGTACATC amounts to 6, and this sequence can be generated through the following steps:

(i) generation of the symbol A: Ø + A → A

(ii) generation of the symbol T: A + T → AT

(iii) generation of the symbol G: AT + G → ATG

(iv) copying of the fragment ATG in inverted orientation: ATG + I(ATG) = ATG + CAT → ATGCAT

(v) copying of the fragment ATGC in symmetrical orientation: ATGCAT + S(ATGC) = ATGCAT + CGTA → ATGCATCGTA

(vi) copying of the fragment CATC in direct orientation: ATGCATCGTA + D(CATC) = ATGCATCGTA + CATC → ATGCATCGTACATC

Nucleotide sequence of length L is decomposed into a series of successive overlapping windows of a given size w; complexity of the corresponding sequence region is calculated within each window. The series of values $[C_1, C_2, \ldots, C_{L-w+1}]$ obtained represents the complexity profile.

The complexity profile for sequence sample is a series

$$\left[\frac{1}{N}\sum_{i=1}^{N} C_1^{(i)}, \frac{1}{N}\sum_{i=1}^{N} C_2^{(i)}, \ldots, \frac{1}{N}\sum_{i=1}^{N} C_{L-w+1}^{(i)}\right]$$

where $C_j^{(i)}$is the calculated measure of complexity of the corresponding region of the ith sequence within the jth window as it was demonstrated in the example above (Gusev *et al.*, 1993), and N is the sample size.

DNA dinucleotide conformational and physico-chemical properties. At present, 38 dinucleotide sets of parameters of conformational and physico-chemical DNA properties have been accumulated in the database PROPERTY (Ponomarenko *et al.*, 1999). These parameters are applied by the unit for constructing the feature-specific profiles of DNA sequences considered as described by Levitsky *et al.* (1999).

Searching for repeats. This unit is designed to search for the direct, inverse, direct complementary and symmetrical substring copies (repeats) in a target DNA sequence and to visualise the results obtained with the help of a Java viewer applet.

Finding repeats in a non-coding sequence is usually a multistage problem, consisting of: (i) finding and marking the simple repeats, such as mono-, di- and trinucleotide tracts; (ii) finding and marking the moderate repeats, such as SINE and LINE elements and other annotated repeats, usually through comparing the target DNA sequence with CENSOR (Jurka *et al.*, 1996); and (iii) considering other regularities, such as a specific repeated structures attributed to the sequence analysed. This is the step we are aiming our system at.

We assess the statistical significance of the repeat abundance in terms of pairwise matches (Solovyer *et al.*, 1994), sequence assumed i.i.d. position-wise.

The advantage of this method is that it accounts for the total number of repeated regions. The well-known problem of pairs dependency, unless using extreme statistics (Arratia, Waterman,1989; Karlin *et al.*, 1990), is not completely overcome. Thus, the model proposed is a first order approximation and is considered as a characteristic of overall abundance of repeats.

The probability of two randomly selected non-overlapping DNA regions with length l to differ by k nucleotides [denoted (l,k) repeats] can be assessed in terms of binomial distribution:

$$P(l,k) = C_l^k p^{l-k}(1-p)^k,$$

where p is the probability of two randomly selected nucleotides to be identical:

$$p = \sum_{i=1}^{4} p_i^2,$$

where p_1, p_2, p_3, p_4 are the nucleotide frequencies.

The number of all the possible locations of two nonoverlapping segments, each l nucleotides long, in a DNA sequence of length L can be assessed as:

$$\varphi_l = C^2_{L-2l+2} = \frac{(L-2*l+1)(L-2*l+2)}{2}.$$

So the average number of repeats is calculated as follows:

$$En(l,k) = \varphi_l * P(l,k).$$

Let us consider the (l,k) repeats whose average number in a random sequence with the length L is close to 1. In this case, the binomial distribution is applicable to estimate the probability that n (l,k) repeats occur in this sequence:

$$P(n) = C^n_{\varphi_l} * (P(l,k))^n * (1-P(l,k))^{\varphi_l - n}$$

To determine the upper boundary of the confidence interval with the significance level q (0.05, 0.01, etc.), we find for the number of repeats expected to occur by chance such n_o that:

$$\sum_{n=0}^{n_0-1} P(n) < 1-q \text{ and } \sum_{n=0}^{n_0} P(n) \geq 1-q.$$

If the number of (l,k) repeats is equal to or exceeds the upper boundary of the confidence interval, i.e. $n(l,k) \geq n_o(l,k)$, such a number of repeats is considered to differ significantly from their expected number $En(l,k)$, and the repeats themselves are regarded as non-random.

These equations would hold true for direct, direct complementary, symmetrical and inverted repeats. In case of direct complementary and inverted repeats, the probability of two randomly selected nucleotides being complementary is used instead of the probability of them being identical:

$$p = 2 * (p_A p_T + p_G p_C)$$

Oligonucleotides unit

It is known that oligonucleotide composition is specific for various regions of genome (Karlin *et al*., 1994; Karlin and Burge, 1995; Karlin *et al*., 1997). Oligonucleotide composition analysis is widely used to reveal peculiarities of regulatory (Hertz and Stormo, 1996; Zhang, 1998) and coding sequences (Snyder and Stormo, 1995). We have previously developed a set of programs for revealing the oligonucleotides that are specific for the coding regions of the families of isofunctional genes (Kolchanov *et al*., 1995). Based on the method of search for degenerate oligonucleotide motifs, we have developed a new software package ARGO for analysis of functional sites and gene regulatory regions.

Assessing significance for the oligonucleotide motif. The aim of the algorithm described is to find the oligonucleotide motifs that are significantly presented in a set of RGS (regulatory genome sequences) and which, therefore, may play a specific biological role.

Let us consider an oligonucleotide motif $M=m_1,m_2,..,m_l$ of length l in the expanded 15 single letter-based code (Cornish-Bowden, 1985). The probability of this motif occurring at a particular position in the DNA sequence S_k of length L is:

$$P(M) = \prod_{i=1}^{l} P_i,$$

where P_i is a frequency of a letter m_i assessed from the nucleotide content of S_k.

If the expected number of occurrences of a particular oligonucleotide in a sequence calculated as $(L\text{-}l+1)*P(M)$ is less than 1, as is our case, the probability of the motif in question occurring at least once in the sequence S_k can be approximated by Poisson distribution:

$$P(Sk) = 1 - e^{-(L-l+1)*P(M)}$$

Consider the set of the sequences $S = \{S_1,\ldots,S_N\}$. The binomial probability $P(n,N)$ if observing the motif M in more or equal than $n(0\le n\le N)$ sequences is:

$$P(n, N) = \sum_{i=n}^{N} C_N^i P^i (1-P)^{N-i}.$$

*Description of the algorithm.*The method of the search for significant motifs is based on considering the complete oligonucleotide vocabulary for each RGS with subsequent clusterisation of similar oligonucleotides belonging to different RGS. If Hamming's distance R between oligonucleotides from different sequences is lower than the threshold value r_o, they are united into one class. The consensus is created for each class as follows. The significance of each of the 15 letters to occur at each position is evaluated by binomial criterion, and the signal with the minimal probability of appearing by chance is selected. The oligonucleotide motif obtained by this procedure is considered significant if it meets the following conditions: (i) the fraction f of the RGS containing the motif is higher than a certain given level f_o; and (ii) the binomial probability $P(n,N)$ of observing this motif by accident in n and more RGS of the N RGS considered is lower than a given significance level a.

Results of the analyses performed are accumulated in the knowledge base of the system ARGO in its inner standard format. This approach requires no preliminary alignment of the RGS analysed, representing an advantage compared to all available methods for constructing consensuses and weight matrices.

Implementation and results

We will illustrate operation of the system RegScan by computer analysis of vertebrate promoter sequences.

Data

The sample of promoter sequences was created as follows. The non-redundant promoter sequences of vertebrate chromosomal genes were extracted from EMBL based on the information contained in the EPD42 database (Bucher and Trifonov, 1986) using the software package MGL (Kolpakov and Babenko, 1997). In total, 310 sequences were extracted; each represented a completely sequenced [–300; +100] region relative to the transcription start site (TSS) at position +1 (the list of the corresponding identifiers is available upon request).

The same entries were used for extracting exons and introns. The resulting sequences were searched for coincidence with the TATA-box matrix. Only those motifs were selected whose score exceeded the cut-off value of –8.16 and which were located at least in 20 bp one from another. The motifs selected were supplemented with flanks, so that their total lengths amounted to 400 nucleotides and false TATA boxes were located in the 270 region. To avoid shifts in the nucleotide composition, not more than five TATA boxes were extracted from one sequence. Thus, 135 sequences of exons (with the mean score of false TATA boxes of –6.77) and 762 sequences of introns (with the mean score of false TATA boxes of –6.92) were obtained and constituted a control sample.

Classification

The initial sample of promoters was divided into TATA+, TATA–, CpG+ and CpG– subgroups and their intersections TATA+CpG+, TATA+CpG–, TATA–CpG+ and TATA–CpG– by applying the classification module of the RegScan system.

The weight matrix for TATA box with a cut-off value of –8.16 (Bucher, 1990) was used to separate the initial sample into TATA-containing (TATA+) and TATA-less (TATA–) subsamples. The mean score of the TATA boxes in the TATA+ subsample was –4.31. The promoters containing CpG islands

in the 5′ region relative the TSS formed the subsample CpG+. The rest of the promoters constituted the subsample CpG–.

The subsample CpG+ contains an approximately equal number of TATA-containing and TATA-less promoters (70 and 74, respectively), whereas TATA-containing promoters are predominant in the subsample CpG– (124 versus 33 TATA-less; Table 1). Nucleotide composition of the subsamples is listed in Table 2. The locations of CpG islands relative to the TSS in the 5′ region have been determined for promoters from subsample CpG+. The 5′ boundaries of CpG islands relative to the TSS are distributed evenly. Analysis of more extended regions has demonstrated that this trend is valid within the interval [–500; +100].

Table 1. The number and percentage of promoters in samples

	CpG+	CpG–	
TATA+	70 (23%)	124 (41%)	**194 (64%)**
TATA–	74 (25%)	33 (11%)	**107 (36%)**
	144 (48%)	**157 (52%)**	

Table 2. Nucleotide composition of promoters in samples (%)

	CpG+ TATA+	CpG+ TATA–	CpG– TATA+	CpG– TATA–
A	19.6	18.6	26.4	25.2
T	17.8	18.6	24.8	25.4
G	30.9	31.5	24.0	25.2
C	31.7	31.3	24.8	24.2

Analysis

Position-specific nucleotide frequency profiles. Position-specific nucleotide frequency profiles were constructed for the initial sample of 301 vertebrate promoters as well as for all the subsamples constructed. The regions corresponding to the core elements (TATA box and Inr) differed in their nucleotide contexts. The nucleotide frequency profile *W(A+T)* of the initial sample indicates an increased concentration of these nucleotides within the TATA box region (Figure 1). The transition from pyrimidine (–1) to purine (+1) on the background of a 'purine pit' is very evident (Figure 2).

The monotonic increase in the C+G frequencies over the entire region considered in the subsample CpG+TATA+ (Figure 3) results from an even distribution of 5′ boundaries of CpG islands relative to the TSS (data not shown). However, the regions with the increased G+C content flank locally the TATA-box region, as is demonstrated by the analysis of CpG–TATA+ (Figure 3). The G+C-rich flanks are lacking in a sample with false TATA boxes (Figure 3).

Search for oligonucleotide motifs. The [–50; +1] region of the core promoter of both TATA+ and TATA– subsamples

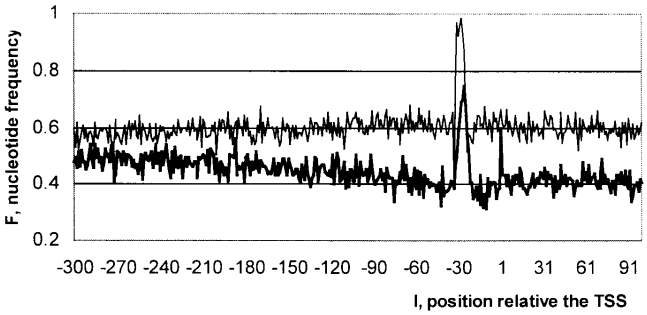

Fig. 1. The frequency profiles of W(A+T) nucleotides for initial set of 301 promoters (bold line) and 301 false TATA-boxes (thin line). An increased concentration of these nucleotides is observed within the region with the center at –30 corresponding to the TATA-box.

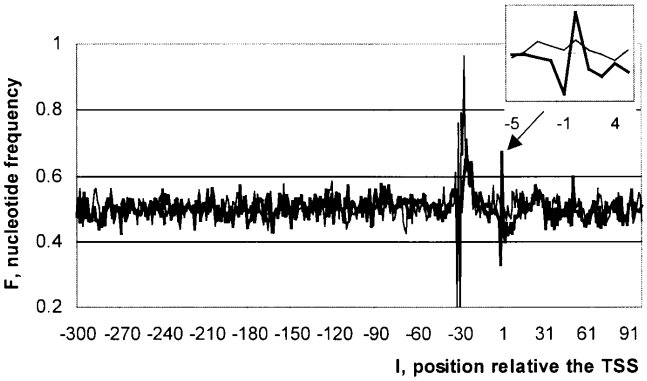

Fig. 2. The frequency profiles of purines (A+G) for initial set of 301 promoters (bold line) and 301 false TATA-boxes (thin line). The transition from pyrimidine (–1) to purine (+1) on the background of a 'purine pit' characterising transcription start site is very evident for the sample of promoters in contrast with the TATA-false sample.

was used to generate samples of characteristic oligonucleotide motifs denoted as M_{TATA+} and M_{TATA-}, respectively, by the program ARGO (Table 3).

Table 3. The significant oligonucleotide motifs obtained from the TATA+ and TATA– subsamples of vertebrate promoters

M_{TATA+}	M_{TATA-}
TWWWWD	GGMRGVV, RGSHGGV,
TNTWWW	VRGVAGNV, BCCYDCHV,
WWWWAD	DGGVDSDG, RRGSHRG,
HTWWWW	CCDSCYB, CYKCCBB,
TWWWWW	VVRGVAG, CYBNBCCY,
	RGNNDGRG, GGMRGVV,
	VVRGVAG, GGVDSDGS,
	VCTBCYB

Each motif occurs in at least 60% of the RGS within the [–50; +1] region, while the binomial probability for these motifs to occur in this number of sequences by chance is $<10^{-10}$.

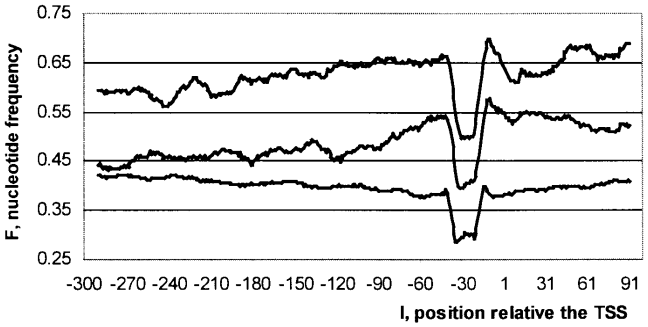

Fig. 3. The frequency profiles of S(G+C) nucleotides for TATA+CpG+, TATA+CpG–and false TATA-boxes subsamples of promoters (upper, intermediate and lower profiles, respectively). The window size is 20 bp.

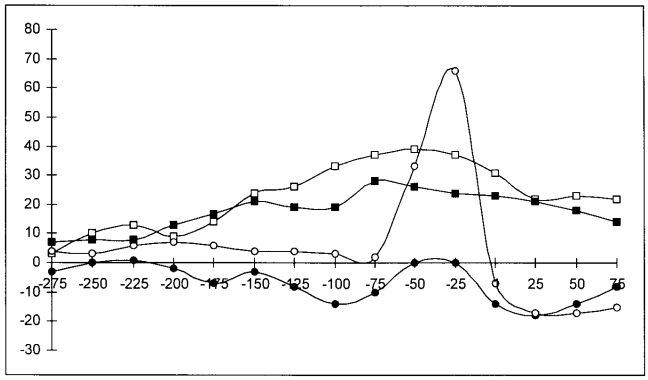

Fig. 4. The profiles of occurrence of the oligonucleotide motifs M_{TATA+} and M_{TATA-} along the sequences of the TATA+ and TATA– promoter subsamples: ○, distribution of HTWWWW (typical M_{TATA+} motif) along the sequences of TATA+ promoters; ●, distribution of HTWWWW along the sequences of TATA– promoters; □, distribution of DGGVDSDG (typical M_{TATA-} motif) along the sequences of TATA– promoters; and ■, distribution of DGGVDSDG along the sequences of TATA+ promoters. Position relative the TSS is on the abscissa; occurrence of the motif at this position over the sample of sequences, on the ordinate.

Then the profiles of both M_{TATA+} and M_{TATA-} motif occurrences along the promoter were constructed as follows. The promoter region [–300; +100] was divided into 15 overlapping windows 50 bp long with a 25-bp shift. Occurrence of each motif was considered independently in each window. The occurrence of each motif analysed in each window can be introduced via 'representation coefficient' $F = f_+ - f_-$, where f_+ is the RGS fraction containing a given motif in a given window and f_- is the fraction of random sequences housing this motif. The graphs for distribution of F coefficient were constructed for the M_{TATA+} and M_{TATA-} motifs (Figure 4).

We compared the method proposed with the standard Gibbs free energy algorithm (Lawrence *et al.*, 1993). Since the TATA+ promoter subsample contains *a priori* the TATA signal, we used the TATA– subsample for this comparison. The Gibbs algorithm is a probabilistic method; therefore, we carried out 50 individual alignment events for the [–50; +1] regions of TATA– promoters and obtained 31 different consensus nuclei and the values F were calculated for them (Table 4). Then, we searched for the motifs in four letter-based code using the program ARGO. As a result, oligonucleotides GNGCNGG (F = 32%) and GGGNGGNG (F = 24%) were revealed. Note that the motif GNGCNGG contains the oligonucleotide GNAGG, found by the Gibbs method, which was less represented (F = 22%), while the motif GGGNGGNG contains both GNAGG and GGAG (F = 19%), found by the Gibbs method too. In turn, these signals are constituents of the oligonucleotide motifs found by ARGO for 15 single letter-based code (Table 3) and displaying F~40%.

Thus, the estimations performed have demonstrated the compliance of our method for searching oligonucleotide motifs with the other available methods used for analysing low-homology DNA sequence samples..

Table 4. Consensus nuclei obtained through Gibbs multiple alignment

Consensus nucleus	Percent of aligning events, D (%)	Information content value, G	Maximal representation, F (%)
gnagg	2	464	22
ggag	30	512	19
ctct	2	466	14
gggagg	2	510	14
ggga	2	420	14
ggaag	2	487	12
gnag	4	420	10
angag	2	471	10
gngg	2	418	10
others	48 (each <2%)		<10

Each consensus nucleus is characterised by the weight matrix and its information content value

$$G = - \sum_{k=\{A,T,G,C\}} \sum_{i=0}^{w} p_{ki} \log_2 p_{ki},$$

where w is the length of consensus nucleus and p_{ki}, frequency of nucleotide k in the i-th position of consensus nucleus. In addition, each consensus nucleus is characterised by its maximal representation F and the percent of aligning events D performed to obtain this consensus nucleus

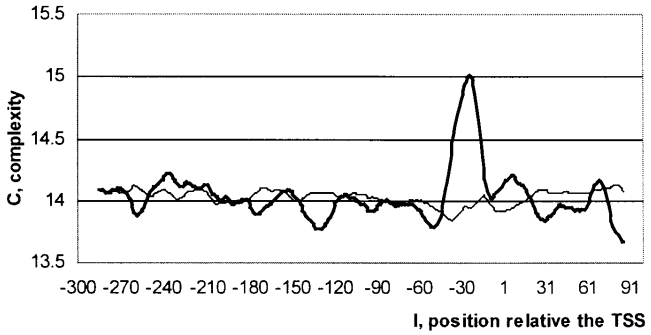

Fig. 5. The S-complexity profiles constructed over the TATA+ promoter subsample (bold line) and false TATA-boxes (thin line). The window size is 20 bp. The TATA-box region in TATA+ promoter subsample is characterised by increased complexity values on the contrary to the respective profile for the false TATA-box subsample.

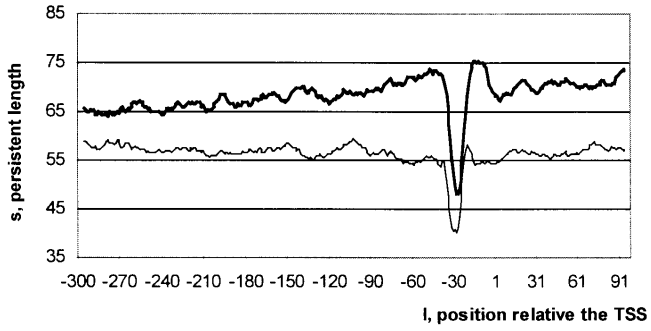

Fig. 6. The bending stiffness profiles for the TATA+ promoter subsample (bold line) and for 195 TATA-false sequences (thin line). Persistent length (nm) is on the ordinate. Window size is 10 bp.

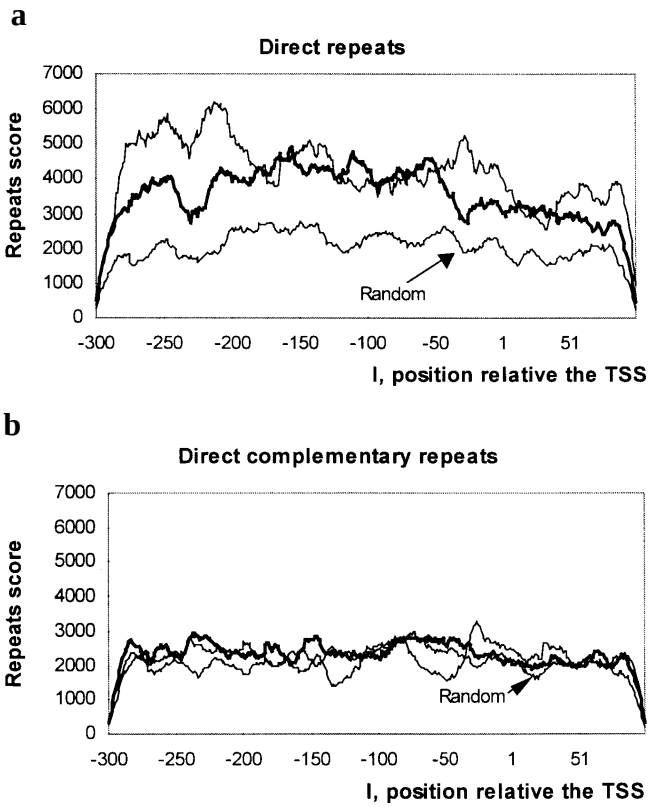

Fig. 7. The profiles of saturation by (**a**) direct repeats; (**b**) direct complementary repeats in the sample of 301 promoters (bold line) and the sample of 301 non-promoter sequences (thin line). An arrow indicates profile for random sequences of the same local nucleotide content. Random sequences were produced as follows: each promoter sequence was divided into non-overlapping 10 bp windows, and the nucleotides were jumbled within each of them. The system for repeat searching was used to analyse the repeats. The search parameters are as follows: *min length* = 7 bp, *max length* = 20 bp, *max mismatch number* = 5 bp, *limit* = 0.0001, *limit type* = P_LIMIT. The repeat score is calculated as follows: every time an (l,k)-repeat is found in a sequence, the score for each position within the repeat is increased by l-k. The profile is the score averaged over all sample sequences.

Discussion

We have developed a system for analysing extended regulatory regions of genomic DNA and applied it to study the vertebrate gene promoters transcribed by RNA polymerase II.

Application and comparison of the methods described is detailed above. Here we shall proceed with biological results obtained.

(i) The nucleotide sequence profile indicates that C+G-rich regions are flanking the TATA-box region (Figure 3). They are likely to play certain role in binding of TATA-binding protein (TBP) to TATA box. Since the TBP interacts with DNA minor groove due to the van der Waals interactions, TBP exhibits an increased affinity for A+T-rich DNA regions with a smooth minor groove along with the decreased affinity for G+C-rich regions due to the NH_2 group of guanine, which projects into the minor groove and prevents the close contact of the protein and DNA surfaces. Thus, the decreased TBP affinity for G+C-rich regions does not preclude its diffusion towards the TATA box. The profile of G+C frequencies may also reflect the fact that the binding sites of several transcription factors are G+C-rich.

On the contrary, a decreased G+C content in the TATA-box flanking regions can be seen on the corresponding profile of the sequence sample with the false TATA boxes.

(ii) The region of TATA box displays an increased S complexity (Figure 5); i.e., it is asymmetrical. It is likely that the TATA box asymmetry is necessary for the correct orientation of TBP

Note that no increased complexity is observed in the complexity profile of the sample with the false TATA boxes. This is likely to suggest the effect of the flanking regions or

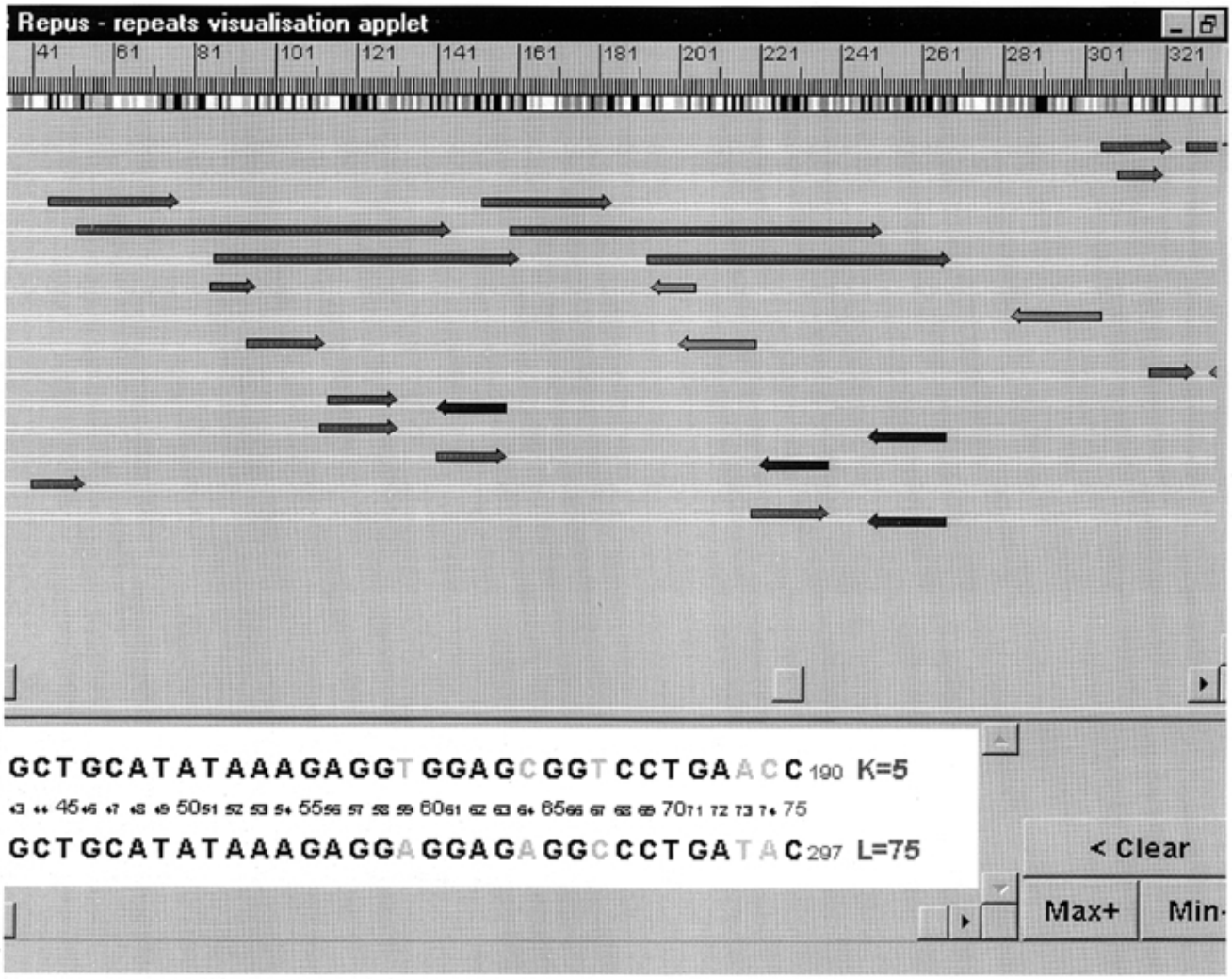

Fig. 8. The graphic output of the program is shown here. Direct and inverted repeats as well as symmetrical repeats in the sequence of *Xenopus laevis* histone H4 gene (EPD: 07051, EMBL: X00224, XLHIS4) are indicated. The double-stranded DNA molecule is displayed by two straight lines. The repeats are indicated by arrows with respect to their position and type. A scale indicating nucleotide positions relative to the sequence start is shown at the bottom. Clicking a repeat image opens an auxiliary window containing the following characteristics of a repeat: the length, nucleotide sequences of its regions with indication of coinciding nucleotides, number of mismatches, and the regions coordinates.

mutual dependence of nucleotides within the TATA box. Note also that the mean score of TATA boxes is –4.31, while that of the false TATA boxes is –6.9.

(iii) The profile of bending stiffness demonstrates that the TATA box-flanking sequences possess an increased bending stiffness, whereas the region of TATA box itself displays an increased bending flexibility (Figure 6). The promoter DNA architecture may be changed, for example, by activators or HMG-box proteins, which bend DNA, or superhelical coils produced by topoisomerases, etc. It is likely that under the stresses causing DNA bending, the stiff CG blocks flanking the TATA box provide the precise location of the bend in AT-rich DNA region located between them and, consequently, the location of TATA box. It was demonstrated experimentally that TBP bound to the DNA pre-bent (17–20°) toward the major groove with a manifold increased affinity compared with the unbent DNA (Parvin *et al.*, 1995).

(iv) The distribution of M_{TATA+} oligonucleotide motifs along the sequences of TATA+ and TATA– promoters demonstrates that the frequency in the region [+1; +100] is drastically decreased in both subsamples (Figure 4). This suggests a possibility of a negative evolutionary pressure that forbids the presence of M_{TATA+} motifs in the [+1; +100] region of promoters. The competitive inhibition of the correct TATA box by the false TATA binding sites, randomly occurring in unexpected positions, may provide the explanation of this fact (Kolchanov and Lim, 1994).

(v) The system for repeat searching was used to analyse the repeats in the initial sample of 301 promoters. On the whole, the direct repeats were more abundant compared with the random sequences of the same nucleotide composition (Figure 7a). However, the abundance of direct complementary repeats, known with no biological function, was similar to those in random sequences (Figure 7b). Inverted repeats were not so abundant (data not shown).

Similar results were obtained on 301 nucleotide sequences randomly selected from a total of an 897-strong sample of TATA-false sequences, all of which are either exon or intron fragments (Figure 7a,b).

By means of the visualising unit (Figure 8) we also manually selected some promoters with unusual distinct structural features:

(a) EPD:07051. This contains a duplicated fragment of regulatory region, resulting in origination of an alternative transcription start site (Clerc *et al*., 1983);

(b) EPD:27009. Contains an extended tandem repeat (Sazer and Schimke, 1986);

(c) EPD:24032. Four tandem fragments of 48 nucleotides each (Kelly *et al*., 1986);

(d) EPD:41005. An extended tandem repeat (Nakasone *et al*., 1993);

(e) EPD:15032. Potential cruciform structure (Valerio *et al*., 1985).

RegScan-assisted analysis suggests that the distributions of the nucleotide content, the complexity profiles and the physical and chemical properties of the TATA-box region are different with those of the TATA-false control sample. The promoters have an abundance of direct and, to a lesser extent, inverted repeats. There are no significant differences between the promoters and other regions in this abundance.

The proposed approach is aimed at acquiring the knowledge on various peculiarities of regulatory regions. We are currently extending the classification module to provide additional classification schemes based on the local alignment by using Gibbs Sampling Strategy (Lawrence *et al*., 1993) combined with data from TRRD (Kolchanov *et al*., 1999) and TRANSFAC (Heinemeyer *et al*., 1998) databases on the transcription factor binding sites. Also classification based on specific genes involved in the same gene network will be taken into account. We believe that finding the best classification of eukaryotic gene promoters is a cornerstone for the task of revealing significant features of regulatory regions.

Acknowledgements

The authors are grateful to N.A.Kolchanov, M.P.Ponomarenko, F.A.Kolpakov and A.E.Kel for helpful discussions and anonymous referees for valuable suggestions. The work was supported by the Russian Foundation for Basic Research, the Russian Human Genome Program, Russian State Committee on Science and Technology, Integrated Program of Siberian Department of Russian Academy of Sciences, and by the Grant No. 5-R01-RR04026-09 NIH USA. The authors are grateful to Dr G.Chirikova for translating the paper from Russian into English.

References

Allison,L., Edgoose,T. and Dix,T.I. (1998) Compression of Strings with Approximate Repeats. In *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, Menlo Park, CA.

Allison,L.and Yee,C.N. (1990) Minimum message length encoding and the comparison of macromolecules. *Bull. Math. Biol*., **52**, 431–453.

Arratia,R. and Waterman,M.S. (1989) The Erdos–Renyi strong law for pattern matching with a given proportion of mismatches. *Ann. Prob*., **17**, 1152–1169.

Baldi,P., Chauvin,Y., Brunak,S., Gorodkin,J. and Pedersen,A.G. (1998) Computational applications of DNA structural scales. In *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, Menlo Park, CA.

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.

Bucher,P. and Trifonov,E.N. (1986) Compilation and analysis of eukaryotic POL II promoter sequences. *Nucleic Acids Res*., **14**, 10009–100026.

Chen,Q.K., Hertz,G.Z. and Stormo,G.D. (1995) MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput. Appl. Biosci*., **11**, 563–566.

Clerc,R.G., Bucher,P., Strub,K. and Birnstiel,M.L. (1983) Transcription of a cloned *Xenopus laevis* H4 histone gene in the homologous frog oocyte system depends on an evolutionary conserved sequence motif in the –50 region. *Nucleic Acids Res.*, **11**, 8641–8657.

Cornish-Bowden,A. (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res*., **13**, 3021–3030.

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

Frech,K., Hermann,G. and Werner,T. (1993) Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids. *Nucleic Acids Res*., **21**, 1655–1664.

Gardiner-Garden,M. and Frommer,M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol*., **196**, 261–282.

Gelfand,M.S. (1995) Prediction of function in DNA sequence analysis. *J. Comp. Biol*., **2**, 87–115.

Grumbach,S, Tahi,F. (1994) A new challenge for compression algorithms: genetic sequences. *Inf. Proc. and Management*, **30**, 875–886.

Gusev,V.D., Kulichkov,V.A. and Chupakhina,O.M. (1993) The Lempel–Ziv complexity and local structure analysis of genomes. *BioSystems*, **30**, 183–200.

Hutchinson, G.B. (1996) The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *Comput. Appl. Biosci*., **12**, 391–398.

Heinemeyer,T., Wingender,E., Reuter,I., Hermjakob,H., Kel,A.E., Kel,O.V., Ignatieva,E.V., Ananko,E.A., Podkolodnaya,O.A., Kolpakov,F.A., Podkolodny,N.L. and Kolchanov,N.A. (1998) Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res.*, **26**, 362–367.

Hertz,G. and Stormo,G. (1996) *Escherichia coli* promoter sequences: analysis and prediction. *Methods Enzymol.*, **273**, 30–42.

Jurka,J., Klonowski,P., Dagman,V. and Pelton,P. (1996) CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. *Comput. Chem.*, **20**, 119–122.

Karlin,S., Dembo,A. and Kawabata,T. (1990) Statistical composition of high-scoring segments from molecular sequences. *Ann. Stat.*, **18**, 571–581.

Karlin,S. and Burge,C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.*, **11**, 283–290.

Karlin,S., Ladunga,I. and Blaisdell,B. (1994) Heterogeneity of genomes: measure and values. *Proc. Natl Acad. Sci. USA*, **91**, 12837–12841.

Karlin,S., Mrazek,G. and Campbell,A. (1997) Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol.*, **179**, 3899–3913.

Kel,A.E., Kolchanov,N.A., Kel,O.V., Romashchenko,A.G., Anan'ko,E.A., Ignat'eva,E.V., Merkulova,T.I., Podkolodnaya,O.A., Stepanenko,I.L., Kochetov,A.V., Kolpakov,F.A., Podkolodnyi,N.L. and Naumochkin,A.A. (1997) TRRD: a database of transcription regulatory regions in eukaryotic genes. *Molek. Biol. (Mosk.)*, **31,** 626–636.

Kel,O.V., Romaschenko,A.G., Kel,A.E., Kolchanov,N.A. and Wingender,E. (1995) A compilation of composite regulatory elements affecting gene transcription in vertebrates. *Nucleic Acids Res.*, **23**, 4097–4103.

Kelly,J.M., Porter,A.C., Chernajovsky,Y., Gilbert,C.S., Stark,G.R. and Kerr,I.M. (1986) Characterization of a human gene inducible by alpha- and beta-interferons and its expression in mouse cells. *EMBO J.*, **5**, 1601–1606.

Kolchanov,N.A., Ananko,E.A., Podkolodnaya,O.A., Ignatieva,E.V., Stepanenko,I.L., Kel-Margoulis,O.V., Kel,A.E., Merkulova,T.I., Goryachkovskaya,T.N., Busigina,T.N., Kolpakov,F.A., Podkolodny,N.L., Naumochkin,A.N. and Romashchenko,A.G. (1999) Transcription Regulatory Regions Database (TRRD): its status in 1999. *Nucleic Acids Res.*, **27**, 303–306.

Kolchanov,N.A., Vishnevsky,O.V., Babenko,V.N. and Kel,A.E. (1995) Oligonucleotide Sets. Computer Tool and Application. In *Proceedings Third International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, Menlo Park, CA.

Kolchanov,N.A. and Lim,H.A. (1994) *Computer analysis of Genetic Macromolecules.: Structure, Function and Evolution*. World Scientific Pub. Co., Singapore, New Jersey, London, Hong Kong.

Kolmogorov,A.N. (1965) Three approaches to the quantitative definition of information. *Probl. Inf. Transmission*, **1**, 1–7.

Kolpakov,F.A. and Babenko,V.N. (1997) Computer system MGL: a tool for construction of samples, visualization and analysis of genomic regulatory sequences. *Molek. Biol. (Mosk.)*, **31**, 647–655.

Kondrahin,Y.V, Kel,A.E., Kolchanov,N.A., Romaschenko,A.G. and Milanesi,L. (1995) Eukaryotic promoter recognition by binding sites for transcription factors. *Comput. Appl. Biosci.*, **11**, 477–488.

Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Lempel,A. and Ziv,J. (1976) On the complexity of finite sequences. *IEEE Trans. Inf. Theory IT*, **22**, 783–795.

Levitsky,G.V., Ponomarenko,M.P., Ponomarenko,J.V., Frolov,A.S. and Kolchanov,N.A. (1999) Nucleosomal DNA property database. *Bioinformatics*, **15**, 582–592.

Milosavljevic,A. and Jurka,J. (1993) Discovering simple DNA sequences by the algorithmic significance method. *Comp. Appl. BioSci.*, **9**, 407–411.

Nakasone,K., Kenmochi,N., Toku,S. and Tanaka,T. (1993) The structure of the gene encoding chicken ribosomal protein L30. *Biochim. Biophys. Acta*, **1174**, 75–78.

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

Ponomarenko,J.V., Ponomarenko,M.P., Frolov,A.S., Vorobyev,D.G., Overton,G.C. and Kolchanov,N.A. (1999) Conformational and physico-chemical DNA features specific for transcription factor binding sites. *Bioinformatics*, **15**, 654–668.

Prestridge,D.S. and Stormo,G. (1993) SIGNAL SCAN 3.0: new database and program features. *Comput. Appl. Biosci.*, **9**, 113–115.

Prestridge,D.S. (1995) Predicting Pol II promoter sequences using transcription factor binding sites. *J Mol Biol*, **249**, 923–932.

Quandt,K., Frech,K., Karas,H., Wingender,E. and Werner,T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878–4884.

Quandt,K., Frech,K. and Werner,T. (1997) Analysis of transcription regulatory regions based on recognition of transcription factor binding sites and relative arrangement. *Molek. Biol. (Mosk.)*, **31**, 749–758.

Roeder,R.G. (1996) The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.*, **21**, 327–335.

Sazer,S. and Schimke,R.T. (1986) A re-examination of the 5′ termini of mouse dihydrofolate reductase RNA. *J. Biol. Chem.*, **261**, 4685–4690.

Snyder,E. and Stormo,G. (1995) Identification of protein coding regions in genomic DNA. *J. Mol. Biol.*, **248**, 1–18.

Solovyev,,V.V., Kolchanov,N.A. and Zhazkikh,A.A. (1994) Method for revealing imperfect repeats in nucleotide sequences. In Kolchanov, N.A. and Lim, H.A. (eds), *Computer Analysis of Genetic Macromolecules: Structure, Function and Evolution.* World Scientific, Singapore, New Jersey, London, Hong Kong, pp. 107–112.

Valerio,D., Duyvesteyn,M.G.C., Dekker,B.M.M., Weeda,G., Berkvens,Th.M., Van der Voorn,L. and Van der Eb,A.J. (1985) Adenosine deaminase: characterization and expression of a gene with remarkable promoter. *EMBO J.*, **4**, 437–443.

Wingender,E. (1993) *Gene Regulation in Eukaryotes*. VHC Verlagsgesellschaft mbH, D-69469 Weinheim, Germany.

Wingender,E. (1997) Classification of transcription factors in Eukaryotes. *Molek. Biol. (Mosk.)*, **31,** 584–600.

Wolfertstetter,F., Frech,K., Hermann,G. and Werner,T. (1996) Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. *Comput. Appl. Biosci.*, **12**, 71–80.

Zhang,M.Q. (1998) Identification of human gene core promoters in silico. *Genome Res.*, **8**, 319–326.