BIOINFORMATICS

Vol. 17 no. 11 2001
Pages 998–1010

Nucleosome formation potential of eukaryotic DNA: calculation and promoters analysis

Victor G. Levitsky, Olga A. Podkolodnaya, Nikolay A. Kolchanov and Nikolay L. Podkolodny*

Laboratory of Theoretical Genetics, Institute of Cytology & Genetics, 630090, Lavrentiev Ave. 10, Novosibirsk, Russia

Received on December 10, 2000; revised on April 14 and June 28, 2001; accepted on July 27, 2001

ABSTRACT

Motivation: A rapid growth in the number of genes with known sequences calls for developing automated tools for their classification and analysis. It became clear that nucleosome packaging of eukaryotic DNA is very important for gene functioning. Automated computer tools for characterization of nucleosome packaging density could be useful for studying of gene regulation and genome annotation.

Results: A program for constructing nucleosome formation potential profiles of eukaryotic DNA sequences was developed. Nucleosome packaging density was analyzed for different functional types of human promoters. It was found that in promoters of tissue-specific genes, the nucleosome formation potential was essentially higher than in genes expressed in many tissues, or housekeeping genes. Hence, capability of nucleosome positioning in the promoter region may serve as a factor regulating gene expression.

Availability: The program for nucleosome sites recognition is included into the GeneExpress system; section 'DNA Nucleosomal Organization', http://wwwmgs.bionet.nsc.ru/mgs/programs/recon/.

Contact: levitsky@bionet.nsc.ru

INTRODUCTION

The nucleosome is the basic unit of chromatin packaging (Kornberg and Lorch, 1999). Nucleosome structure is similar in different eukaryotic taxa, as well as the proteins constituting the nucleosome which are extremely conservative (van Holde, 1989). The nucleosomal level of chromatin packaging is a winding of DNA sequence of about 146 bp long on the protein globule (octamer or histone core) formed of eight histones (H2A, H2B, H3, and H4, two molecules of each). The recent x-ray analysis data at 2.8 Å resolution suggest that nucleosomal DNA is only slightly bent at positions located in 1–1.5 and 4 turns off from the center of the nucleosome site, each end 10 bp segment of nucleosomal DNA is essentially straight, so an effective number of superhelical DNA turn equals to 1.65 (Luger *et al.*, 1997).

Several approaches to computer analysis of nucleosome formation sites have been proposed (Trifonov and Sussman, 1980; Mengeritsky and Trifonov, 1983; Calladine and Drew, 1986; Satchwell *et al.*, 1986; Uberbacher *et al.*, 1988; Staffelbach *et al.*, 1994; Fitzgerald *et al.*, 1994; Ulyanov and Stormo, 1995; Ioshikhes *et al.*, 1996; Sivolob and Kharpunov, 1995; Stein and Bina, 1999; Levitsky *et al.*, 1999). The problem of contextual specificity of nucleosomal DNA (Trifonov, 1997) is of particular interest. First, a periodic occurrence of certain dinucleotides rendering DNA the ability to bend was discovered (Trifonov and Sussman, 1980). Fourier analysis demonstrated that the sequences of nucleosome sites differ from random sequences (Satchwell *et al.*, 1986). Ulyanov and Stormo (1995) proposed a method for detecting weak consensuses in nucleosome sites in a 15 single letter-based degenerate code. The degeneracy means that no stringent conditions are imposed on a nucleotide sequence, and many similar DNA sequences are capable of nucleosome positioning. Baldi *et al.* (1996) found in human exons and introns a pattern with triplet consensus non-T(A/T)G (abbreviated to VWG), with periodicity of roughly 10 nucleotides. The presence of this pattern was related with phased bending potential and nucleosome positioning. Based on this observation it was demonstrated (Stein and Bina, 1999) that the experimentally determined preferences for nucleosome positioning data could be predicted by counting the occurrences of the period-10 VWG consensus.

Among the characteristics of nucleosomal organization of the chromatin (Trifonov, 1997) are its imperfection and degeneracy; therefore, classical methods of computer analysis (alignment and search for consensus) are poorly applicable here. In certain cases, the nonhistone proteins localized to the sites adjacent to nucleosome positioning

*To whom correspondence should be addressed.

sites have a pronounced effect on nonrandom nucleosome positioning (Thoma, 1992). A complex and degenerate code of the nucleosome sites positioning makes their computer recognition difficult. Unlike transcription factor binding sites which are characterized by more or less pronounced consensuses and weight matrices (Staden, 1984), the intricate coding typical of nucleosome sites is responsible for a certain DNA conformation, rather generalized and acceptable for many DNA sequences, that provides the interaction with the histone octamer (Trifonov, 1997).

Numerous experimental data accumulated so far suggest an important role of nucleosome positioning around promoter regions in the regulation of gene transcription (Wolffe, 1994; Fragoso *et al.*, 1995; Weinmann *et al.*, 1999). Regulation of eukaryotic gene transcription is tightly connected with the changes in nucleosome structure of the chromatin (Steger and Workman, 1996). A nucleosome positioned in the promoter region is capable of inhibiting the transcription initiation (Li *et al.*, 1998), whereas its displacement is capable of surmounting the repressive effect. One of the underlying mechanisms involves formation of a triple complex activator–nucleosome–DNA (Adams and Workman, 1993; Moreira and Holmberg, 1998).

On the other hand, a precise nucleosome positioning provides sometimes for drawing spatially distant DNA regions together, thereby facilitating the interactions of transcription factors bound to them and finally promoting the transcription activation (Montecino *et al.*, 1996; Herrera *et al.*, 1997). Thus, the ability of DNA to interact with the histone core appears essential for the gene function. It is also demonstrated that the accuracy of promoter recognition increases if the patterns of nucleosome positioning sites are analyzed additionally (Levitsky *et al.*, 2000).

Thus, it becomes apparent that the nucleosome packaging of DNA is one of the key factors underlying the specific functions of genomic sequences. To gain the insight into the characteristic features of genomic DNA nucleosome packaging, we have developed a program for calculating nucleosome formation potential profiles in eukaryotic DNA sequences. For calculating nucleosome potential, we have used the discriminant analysis, which was not applied previously for this purpose, but only for recognition of the coding gene sequences (Solovyev *et al.*, 1994; Zhang, 1997) and promoters (Solovyev and Salamov, 1997; Zhang, 1998). By the program developed, we have analyzed the nucleosome packaging density of different functional types of human promoters. Each type of promoter was shown to exhibit a specific pattern of nucleosome formation potential.

SYSTEM AND METHODS

Basic scheme of the system NucleoMeter

In order to analyze nucleosomal DNA, we have developed a computer system NucleoMeter (Figure 1). At the first step, we have designed method of partitioning the nucleosome site into separate regions with a more homogeneous nucleotide context than that of the entire site. This task is implemented by the block PartitionSearch of the system NucleoMeter, by using Monte Carlo methods and discriminant analysis of dinucleotide frequencies of nucleosome site sequences. PartitionSearch searches for such a partition of the nucleosome site sequence into nonoverlapping regions that provides the maximal value of the Mahalanobis distance R^2 functional while discriminating between nucleosome sites and the rest sequences.

The nucleotide sequences of nucleosome sites used as the input information for PartitionSearch performance are stored in the database SiteSequences.

The partition of a nucleosome site is used by the block PotentialBuilder to calculate the nucleosome formation potential $\varphi(X)$. This potential is constructed so that its mean value over the initial set of nucleosome site sequences equals $+1$; over the set of non-site (random) sequences, -1. This means that the $\varphi(X)$ values close to $+1$ correspond to a higher probability of nucleosome positioning. The interface Recon http://wwwmgs.bionet.nsc.ru/mgs/programs/recon/ of the system NucleoMeter allows the user to construct the nucleosome formation potential $\varphi(X)$ profile for the sequence of interest.

The database GenomeSequences comprises the nucleotide sequences of exons, introns, splice sites, promoters, and repetitive sequences used to calculate their nucleosome formation potentials.

Sequences used for analysis

The set of nucleosome sites used (totalling 141 nucleotide sequences) was stored in the SiteSequences database. It is composed from sequences extracted from the EMBL databank according to the codes and positions indicated in the database NUCLEOSOMAL DNA (Ioshikhes and Trifonov, 1993 http://www.embl-heidelberg.de/Services/index.html, section Molecular Biology Databases, catalogue nucleosomal_dna). The nucleosomal DNA sequences were aligned with respect to the centres of footprints. The sequences extracted were added to the database SAMPLES of the system GeneExpress. This database is freely available at the server of the Institute of Cytology and Genetics SB RAS, http://wwwmgs.bionet.nsc.ru/Dbases/NSamples/auto1.exe.
Since we have observed the internal central symmetry of the nucleosome site, we have analyzed the nucleosome site sequences in both orientations, with lengths of 160 bp,

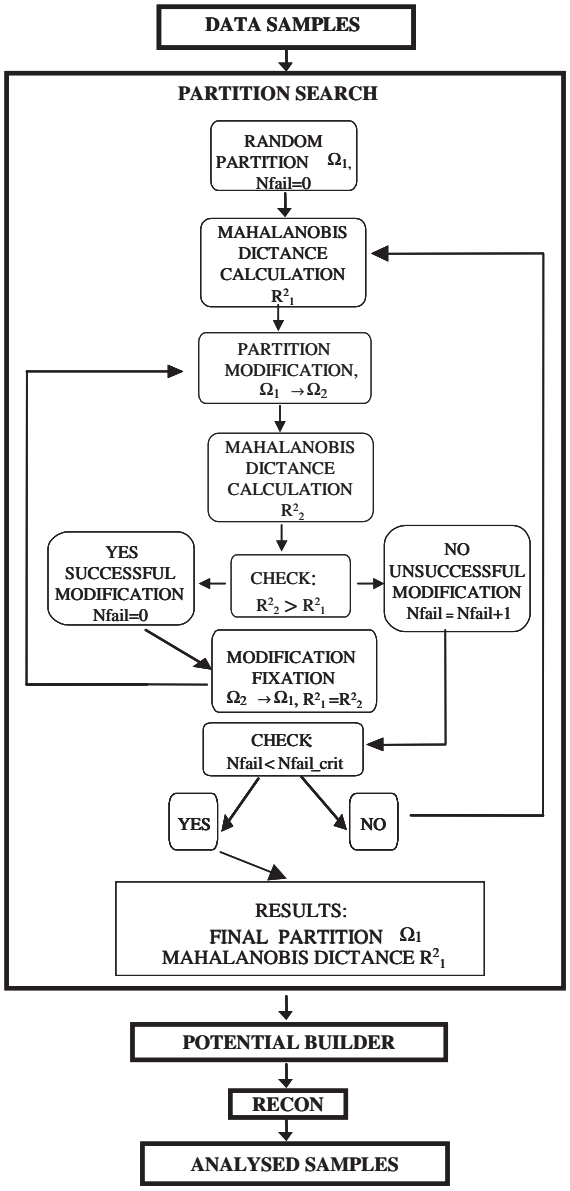

Fig. 1. Basic components of the system NucleoMeter (bold type) and scheme of the algorithm constructing the nucleosome formation potential profile (block PartitionSearch).

i.e. [−80; +80] relative to the footprint centres.

Nucleosomal DNA sequence database (Ioshikhes and Trifonov, 1993) accumulates experimentally verified nucleosome positioning sequences. The nucleosome center locations were deduced from different mapping techniques with the mapping accuracy ranging from 1 to 70 bp. In case several overlapping positions were observed, the most prominent ones were accepted. For construction of the potential for nucleosome formation, we have used the set applied by Ioshikhes and Trifonov (1993). Therefore, in this paper, we define nucleosome positioning in terms of rotational/translational parameters rather than DNA affinity for the histone octamer. This definition should be taken into account when interpreting the results of nucleosome potential calculation (see Section **Discussion**).

Table 1. Sets of DNA sequences used in the analysis

Set		Size
Sequences of nucleosome sites with lengths of 160 bp		141
Promoter regions of human genes [−300; +100] relative to the transcription start	Housekeeping	32
	Expressed in a wide range of tissues	30
	Tissue-specific	141
Stable nucleosome sequences		86
Anti-nucleosome sequences		40

Promoter sequences were extracted from the EPD database on promoters (Perier *et al.*, 2000). In total, 219 human promoters with a length of 400 bp ([−300; +100] relative to the transcription start) were analyzed. By classifying promoters in accordance with the gene expression patterns, we distinguish distinct promoter classes. These patterns were detected on the basis of information stored in the databases EPD (Perier *et al.*, 2000), TRRD (Kolchanov *et al.*, 2000), and relevant literature sources. As a result, promoters are classified into three classes respectively the types containing those genes: (i) housekeeping genes; (ii) genes expressed in a wide range of tissues; and (iii) tissue-specific genes (see Table 1 for gene names and numbers of promoters involved).

In addition, we have analyzed two sets of sequences that were experimentally obtained by the SELEX technique. The set 1 contains the mouse genome sequences, which possess by the maximal affinity to histone octamer and form the 'stable' nucleosomes (Widlund *et al.*, 1997). The set 2 is composed by synthetic sequences, or the so-called 'anti-nucleosomal' DNA fragments, characterized by the least affinity to histone octamer (Cao *et al.*, 1998). The lengths of the sequences from the sets 1 and 2 varied within the ranges 109–151 bp and 76–126 bp, respectively. So, the flanking regions of these sequences were symmetrically supplemented to 160 bp by random sequences, generated from the real sequences, with conservation of their dinucleotide content.

Dinucleotide relative abundance distance

For analysis of sequences generated by the SELEX technique, we have used the dinucleotide relative abundance

distance $\delta(S_1, S_2)$ (Karlin and Ladunga, 1994):

$$\delta(S_1, S_2) = \frac{1}{16} * \sum_{i=1}^{16} |g_i(S_1) - g_i(S_2)|,$$

where S_1, S_2 is a pair of sequences analyzed;

$$g_{XY} = \frac{f^*_{XY}}{f^*_X * f^*_Y};$$
$$f^*_{XX} = f^*_{YY} = 1/2 * (f_{XX} + f_{YY});$$

$f^*_X = f^*_Y = 1/2 * (f_X + f_Y)$, where X and Y are complementary nucleotides.

ALGORITHMS

Optimal partitioning of the nucleosome site into local regions

General definitions. The partition $\Omega(b_1, b_2, \ldots, b_{p-1})$ of a nucleosome site $[a, b]$ is defined as a set of P segments $[a_p, b_p]$, where $(p = 1, \ldots, P)$ meet the following conditions: (1) $a_1 = a$, (2) $a_{p+1} = b_p$ for $p = 1, \ldots, P-1$ and (3) $b_P = b$. The quality of partition was assessed according to the Mahalanobis distance R^2 (Fisher, 1936) between the positive (nucleosome sites) and negative (non-nucleosome sites) sets of nucleotide sequences. The Mahalanobis distance R^2 is defined as the difference between two distributions in the multidimensional space of $16 \times P$ variables (definition of a variable will be given below). An increase of R^2 means the rise in the distance between the centres of two distributions (positive and negative sets of sequences in our case) in a multidimensional space. The use of the Mahalanobis distance makes an allowance for relationships between the parameters used in analysis (Fisher, 1936). This allowance is extremely important for analysis of genome sequences, which are characterized by different types of inter-relationships between the elements of the context.

Thus, the optimization algorithm searching for the best partition of the nucleosome site into nonoverlapping regions solves the following problem: to find such a partition of the nucleosome site that provides the maximal value of the Mahalanobis distance R^2 while discriminating between the nucleosome sites and the rest sequences.

Let us consider the complete set of dinucleotides $\{D_i\} i = 1, \ldots, 16$ and two sets of sequences with equal lengths—the set of sites to be analyzed (nucleosome sites) and the set of random sequences. If the nucleosome site is dissected into P parts, the distance R^2 depends on $N = 16 \times P$ variables (16 is the number of dinucleotides), as each variable is the frequency of individual dinucleotide within a particular region.

Let us consider the pth region of the nucleosome site $(p = 1, \ldots, P)$. Frequency of the ith dinucleotide $(i = 1, \ldots, 16)$ within the pth region of mth sequence X_m of the nucleosome site equals $f^{(1)}_{i,p}(X_m)$. Let us calculate the average frequency of the ith dinucleotide in the pth region over all M sequences of nucleosome sites:

$$f^{(1)}_{i,p} = \frac{1}{M} \sum_{m=1}^{M} f^{(1)}_{i,p}(X_m),$$

where $m = 1, \ldots, M$ and M is the number of sequences in the set of nucleosome sites.

If a dinucleotide is localized on the border between two neighbouring regions, it is considered in both regions, but its frequency is calculated with a coefficient of 1/2 (the frequencies of dinucleotides within a region are calculated with the coefficient 1).

Having completed these calculations for all the dinucleotides, we obtain an N-dimensional vector of average frequencies of all the dinucleotides and all the dissected regions of the nucleosome site:

$$f^{(1)}_n = f^{(1)}_{i,p}, n = 1, \ldots, N,$$

where $N = 16 \times P$. For the pth partition region and ith dinucleotide, $n = (p-1) \times 16 + i$; $i = 1, \ldots, 16$; $p = 1, \ldots, P$. Let us construct the same vector of average frequencies for all the dinucleotides and partition regions over the set of random sequences $f^{(2)}_n$, $n = 1, \ldots, N$, where $N = 16 \times P$.

In an N-dimensional space ($N = 16 \times P$, where P is the number of partition regions composing the site; 16, number of dinucleotides), the Mahalanobis distance R^2 between the set of nucleosome sites (vector $f^{(1)}_n$) and the set of random sequences (vector $f^{(2)}_n$) is found according to the following equation:

$$R^2 = \sum_{k=1}^{N} \sum_{n=1}^{N} \{[f^{(2)}_n - f^{(1)}_n] * S^{-1}_{n,k} * [f^{(2)}_k - f^{(1)}_k]\}, \quad (1)$$

where $n = 1, \ldots, N$; $k = 1, \ldots, N$; and $S^{-1}_{n,k}$ is an element of the matrix $|S^{-1}|$, inverse to the matrix $|S| = |S^{(1)}| + |S^{(2)}|$. In turn, $|S^{(1)}|$ and $|S^{(2)}|$ are the covariance matrices of the vectors of dinucleotide frequencies over the sets 1 and 2, respectively. Elements of the covariance matrices $|S^{(1)}|$ and $|S^{(2)}|$ are calculated by equation:

$$S^{(\gamma)}_{n,k} = \frac{1}{M-1} * \sum_{m=1}^{M} \{[f^{(\gamma)}_n(X_m) - f^{(\gamma)}_n] * [f^{(\gamma)}_k(X_m) - f^{(\gamma)}_k]\},$$

where $m = 1, \ldots, M$; M is the number of sequences in the set of nucleosome sites; $\gamma = 1, 2$ is the number of a set; and $f^{(\gamma)}_n(X_m)$, the frequency of ith dinucleotide in the pth region of mth sequence with the set γ, so that $n = (p-1) \times 16 + i$.

Partition symmetry is a natural condition for choosing the optimal partition. By accounting this parameter, the effective number P of partition regions composing the site equals to 7, instead of 13. All frequency variables averaged over symmetrically located regions. The central symmetry of the nucleosome site is determined by the structure of the nucleosome: the DNA positions equidistant from the centre are under similar conditions.

The sum of dinucleotide frequencies over each local region is equal to 1, thus, decreasing the number of independent dinucleotide frequency variables in each region by 1. By using complementary sequences in the training set and by symmetrical averaging, we decrease the number of independent dinucleotide frequency variables in each region from 15 to 9. Finally, the number of independent variables is equal to $7 \times 9 = 63$.

A Monte Carlo method for detection of the optimal partition. The Monte Carlo method underlies the algorithm for seeking the optimal partition of the nucleosome site, outlined in Figure 1. This algorithm is based on searching for such a partition $\Omega(b_1, b_2, \ldots, b_{p-1})$ that provides for the maximal Mahalanobis distance $R^2(\Omega)$ (1). This actually means that the partition of the nucleosome site into such local regions is looked for wherein the mutual correlations between dinucleotide frequencies provide a more pronounced distinction between nucleosome sites and random sequences.

The algorithm for seeking the optimal partition comprises the following stages (Figure 1):

(1) an initial arbitrary partition (Ω_1 in Figure 1) is specified in a random manner; the example of initial arbitrary partition is represented in the Figure 2a;

(2) the Mahalanobis distance $R^2(\Omega_1)$ between the nucleosome sites and random sequences is calculated for this partition;

(3) the partition Ω_1 is modified into Ω_2 (Figure 1). Modification of the initial partition (Figure 2a) changes the boundary positions of certain individual regions, as illustrated in Figures 2b–e. The modifications require preservation of the total number of local regions (equalling the initial number). In addition, it is assumed that the minimal size of any local region $[a_p, b_p]$ is fixed;

(4) the Mahalanobis distance $R^2(\Omega_2)$ is calculated for the modified partition Ω_2;

(5) compliance with the condition $R^2(\Omega_2) > R^2(\Omega_1)$ is tested; otherwise, the modification performed is considered unsuccessful;

(6) The number N_{fail} of consecutive unsuccessful modifications of Ω_1 into Ω_2, that is, such modifications that fail to meet the condition $R^2(\Omega_2) > R^2(\Omega_1)$ is calculated ($N_{\text{fail}} = 0$ at the beginning of the algorithm performance);

(7) in case the modification is successful, that is, $R^2(\Omega_2) > R^2(\Omega_1)$, the partition Ω_2 is renamed into Ω_1;

(8) here, the successful modification is fixed (Figure 1), and the initial zero value is assigned to the parameter N_{fail};

(9) if the number of consecutive unsuccessful modifications exceeds certain specified threshold level $N_{\text{fail\_crit}}$, the algorithm completes the optimization cycle with the specified initial random partition. The current partition Ω_1 and the corresponding value $R^2(\Omega_1)$ are stored and considered as the intermediate result of the algorithm operation for a given initial arbitrary partition of the site;

(10) a new initial random partition of the site into local regions is specified, and the algorithm proceeds from item (2);

(11) the algorithm operation is completed when K initial partitions are processed;

(12) the partition Ω displaying the maximal $R^2(\Omega)$ value among all the partitions obtained under different initial arbitrary partition is considered the final result of the algorithm performance.

Nucleosome formation potential $\varphi(X)$

A 160-bp sliding window containing the nucleotide fragment X is considered while analyzing an arbitrary nucleotide sequence. At each window position, the value of nucleosome formation potential $\varphi(X)$ is calculated according to the following equation of discriminant analysis:

$$\varphi(X) = \frac{1}{R^2} * \sum_{n=1}^{N} \sum_{k=1}^{N} \times \{[f_n(X) - (1/2) * [f_n^{(2)} + f_n^{(1)}] * S_{n,k}^{-1} * [f_k^{(2)} - f_k^{(1)}]\}.$$

Here R^2 is the Mahalanobis distance calculated according to (1); $f_n(X)$ ($n = 1, \ldots, N$), vector of dinucleotides frequencies in the sliding window of the sequence under study considering the optimal partition of the nucleosome site; and the rest designations are as in equation (1).

The function $\varphi(X)$ determined according to the sets of dinucleotides frequencies $\{f(X)\}$ for the sliding window of the DNA sequence considered may be called the DNA potential of nucleosome positioning (DNA potential of interaction with the nucleosome core).

The mean value of nucleosome formation potential is $+1$ for the set of nucleosome sites and -1 for the set of random sequences. Consequently, a higher probability

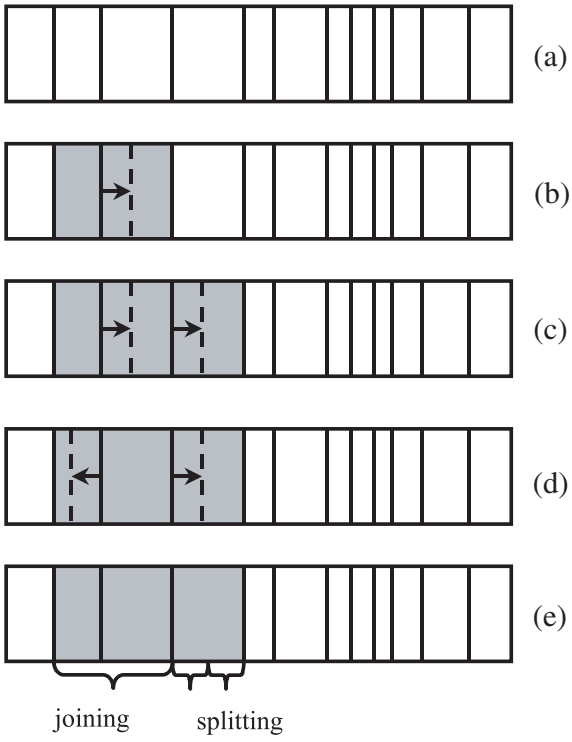

Fig. 2. Examples of modifications used by the algorithm searching for optimal partition: (a) arbitrary distribution; (b) shift of the border between adjacent regions; (c) shift of a region relative to the neighbouring regions; (d) symmetrical shift of the region's borders relative to its centre; and (e) joining and splitting.

of nucleosome positioning correlates with the nucleosome formation potential values close to +1.

IMPLEMENTATION AND RESULTS

Search for optimal partition of the nucleosome site

A minimal length W of the local region is an important parameter while seeking the optimal partition. It was selected equal to 8 bp, as this size was close to that of one helix turn (about 10 bp). In addition, this length allows characteristic patterns of nucleosomal DNA bending stiffness energy profile to be detected (Levitsky *et al.*, 1999).

Our calculations have demonstrated that 50 iterations of the algorithm ($K = 50$) is enough to obtain a stable partition. In this process, the threshold level for unsuccessful modifications was specified as $N_{\text{fail\_crit}} = 100$.

Figure 3a demonstrates the dynamics of increase in the distance R^2 while searching for the optimal partition. The resulting optimal partition (Figure 3b) comprises 13 regions. The axis of symmetry goes through the central region, while the remaining 12 regions are pairwise symmetrical. The sizes of individual regions vary from 8 to 31 bp.

Accuracy of recognition of a nucleosome site

Evaluation of recognition accuracy was made by the jack-knife procedure (Efron and Gong, 1983). The positive sample (nucleosome sites) were divided into the training set, which was used to determine the parameters of the method and included 80% of the sample sequences, and the control set used for evaluation of recognition accuracy. We have made multiple partitions of the positive sample into two parts. For the negative sample (non-sites), we set random sequences with the nucleotides frequencies of 0.25.

Let TP be True Positives (number of sites predicted as the sites), TN be True Negatives (number of non-sites predicted as the non-sites), FP be False Positives (number of non-sites predicted as the sites), FN be False Negatives (number of sites predicted as the non-sites). Then we determine the false negative estimate E_1 (underprediction) and the false positive estimate E_2 (overprediction) as:

$$E_1 = \frac{\text{FN}}{\text{FN} + \text{TN}}; \qquad E_2 = \frac{\text{FP}}{\text{TP} + \text{FP}}.$$

The curve of E_1 versus E_2 obtained by the jack-knife procedure is given in Figure 4. In order to determine optimal false positive and false negative estimations, we have made a maximization of the Correlation Coefficient CC, which evaluates the general accuracy of recognition:

$$\text{CC} = \frac{\text{TP} * \text{TN} - \text{FN} * \text{FP}}{\sqrt{(\text{TP} + \text{FN}) * (\text{TN} + \text{FP}) * (\text{TP} + \text{FP}) * (\text{TN} + \text{FN})}}.$$

In our case, $E_1 = 20.6\%$, $E_2 = 5.6\%$, CC = 0.74.

Nucleosome potential $\varphi(X)$ for randomly generated sequences

We have generated distribution $\varphi(X)$ for the three samples of random sequences. The first sample contained the sequences with nucleotide frequencies $f(X)$ of 0.25; the second and the third samples contained random sequences generated from the set of nucleosome sites with conservation of mono-or dinucleotide content, respectively (Markov models of zero or first order). These three distributions $\varphi(X)$ are illustrated in Figure 5 and compared to distribution $\varphi(X)$ of nucleosome sites. As seen, by viewing from the first to the third sample, there is a progressive shift of distribution $\varphi(X)$ to the right, that is, to distribution $\varphi(X)$ calculated for nucleosome sites.

Nucleosome potential $\varphi(X)$ for the sequences with the low and high affinity to histone octamer, generated by SELEX technique

By analyzing the anti-nucleosome set (see Table 1), it was shown that it contains the sequences with a poor dinucleotide content, which may be determined by synthetic origin of these sequences and their simple structure

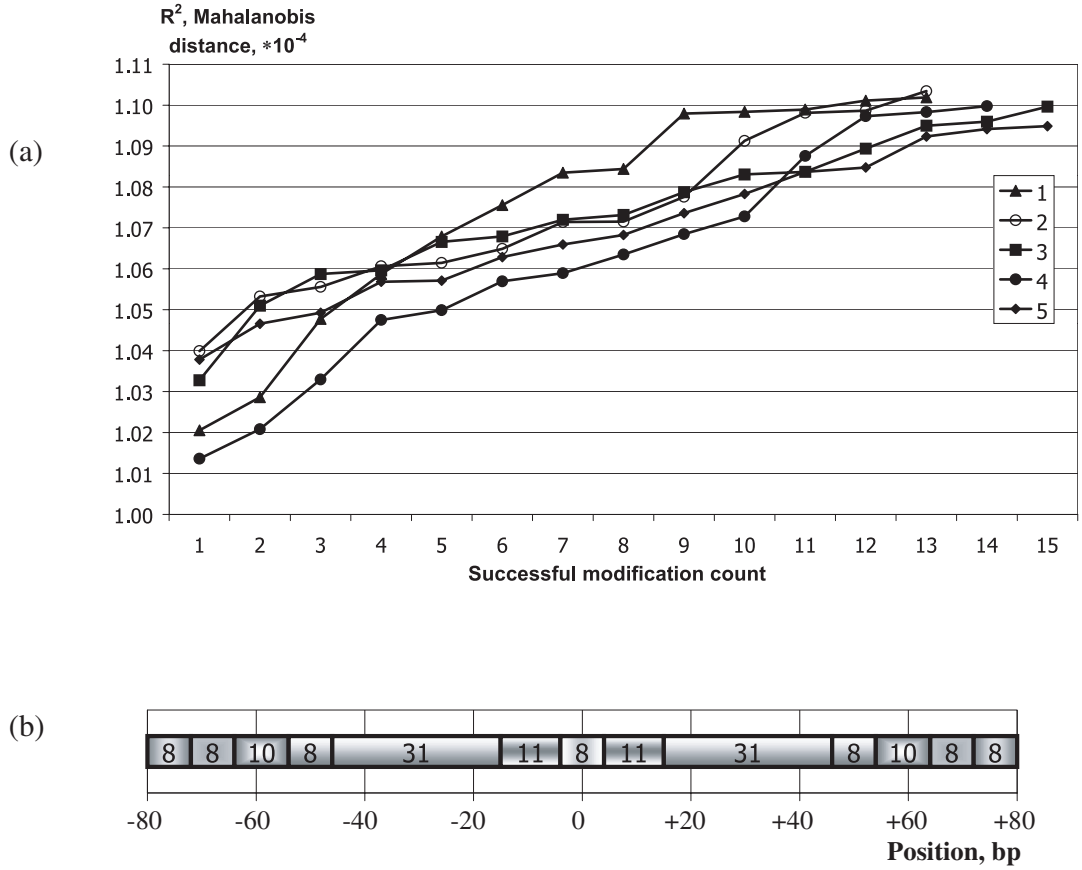

Fig. 3. (a) Growth dynamics of Mahalanobis distance R^2, five optimization cycles are presented; (b) partitions of the entire nucleosome site region [−80; +80] used for constructing nucleosome formation potential profile.

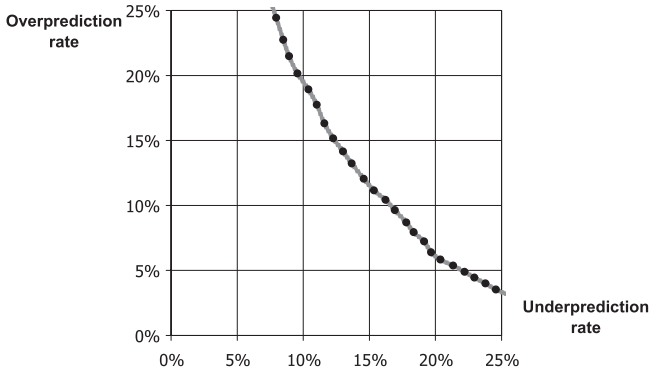

Fig. 4. The curve of overprediction rate (E_2) versus underprediction rate (E_1).

$(TGGA)_n$. This means that dinucleotide frequencies of some sequences from this set essentially differ from those of nucleosome sites. By analysis of a set of sequences forming the stable nucleosomes (see Table 1), we have also found a seria of sequences with dinucleotide frequencies markedly differing from those of nucleosome sites.

Since the nucleosome potential $\varphi(X)$ is based on dinucleotide frequencies that are typical for nucleosome sites, we have excluded from further consideration in both sets the sequences with abnormal dinucleotide frequencies. For this purpose, we have calculated the dinucleotide relative abundance distance $\delta(S_1, S_2)$ (Karlin and Ladunga, 1994).

In our case, S_1 is an arbitrary sequence, S_2, is an integrated sequence of 141 nucleosome sites. Comparison of each nucleosome site S_1 with the integrated sequence S_2 revealed that $\delta(S_1, S_2) < 0.5$. Hence, we use the value $\delta_0 = 0.5$ as a threshold for elimination of the SELEX-generated sequences with abnormal dinucleotide content. As a result, 35% of sequences with abnormal dinucleotide content were rejected from the anti-nucleosome set, and 30%—from the set of stable nucleosome sequences.

After refining of the sets of ‘stable nucleosomes’ and ‘anti-nucleosomes’ sequences, we have generated distri-

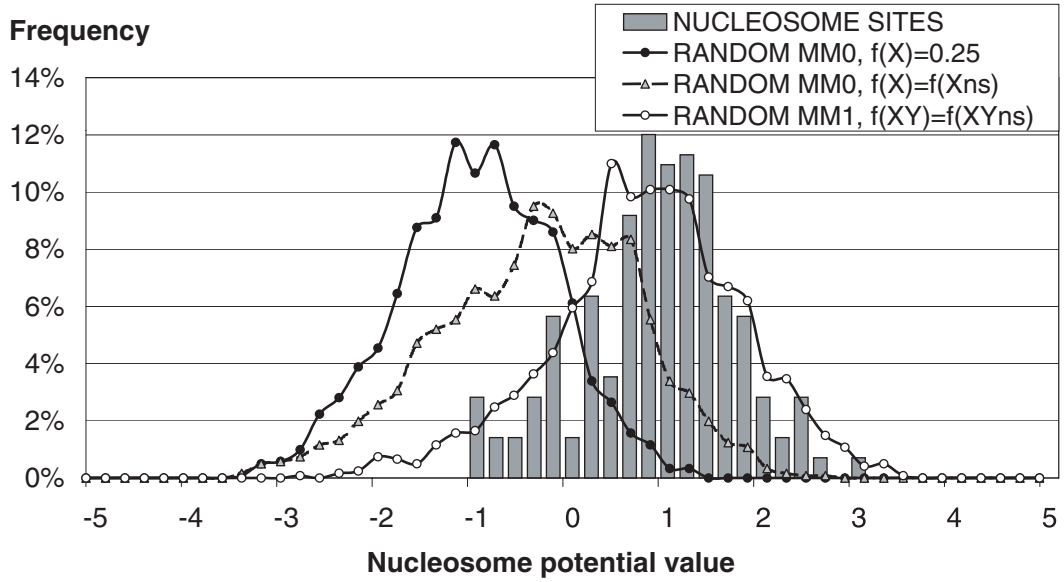

Fig. 5. Histogram of nucleosome formation potential $\varphi(X)$ distributions for randomly generated samples of sequences in comparison to distribution $\varphi(X)$ for the set of nucleosome sites: (1) nucleotide frequencies equal to 0.25; (2) and (3), the sequences are generated by the set of nucleosome sites with conservation of the mono- and dinucleotide content. MM0 is Markov model of order 0; MM1, of order 1. f(Xns) and f(XYns) denote mono- and dinucleotide frequencies of nucleosome site sequences.

butions of nucleosome potential $\varphi(X)$ (Figure 6), which were compared to distribution $\varphi(X)$ for the naturally occurring nucleosome sites sequences. As seen, distribution $\varphi(X)$ for the sets of stable nucleosomes and naturally occurring nucleosome sites are similar, whereas the distribution $\varphi(X)$ for anti-nucleosomes is shifted left-ward and differs significantly from two previously discussed distributions.

The results obtained give evidence that the method of calculation of nucleosome potential constructed on the basis of the set of nucleosome positioning sequences (Ioshikhes and Trifonov, 1993) gives adequate results also for nucleosome positioning sequences selected by the SELEX technique in terms of affinity to histone octamer.

This result enables estimation of the quantitative limitations on application of the approach based on estimation of nucleosome potential $\varphi(X)$ to nucleotide sequences. Factually, we may conclude that while calculating $\varphi(X)$ within the limits of the sliding window, it is necessary to control the mono-and dinucleotide content.

In our case, $A + T$ nucleotide content in the set of nucleosome sites is $f_A + f_T = 59.3 \pm 11.3\%$. By taking this estimation into account and following the Student's criterion, we may determine the limits of 99% confidence interval for nucleotide content of the sliding window, that is, such interval that nucleotide content with significance level $P < 0.01$ does not differ from the mean values along the set of nucleosome sites. This interval ranges from [29.9%; 88.6%].

Thus, by controlling nucleotide content, the application of our program is restricted by condition that $f_A + f_T$ for the sliding window fall within the interval indicated above:

$$29.9\% < f_A + f_T < 88.6\%. \quad (2)$$

The second control parameter is dinucleotide relative abundance distance, which should follow the condition within the limits of the sliding window:

$$\delta(S_1, S_2) < 0.5. \quad (3)$$

To account for this restriction, we have designed a filter into the program for calculation $\varphi(X)$. This filter controls $A + T$ content and dinucleotide content $\delta(S_1, S_2)$ according to conditions (2) and (3). In case even a single condition is not valid, the program refuses to calculate the nucleosome potential in a respective sliding window. Positions of the sliding window, which are ignored by the program, are marked by colour in the graphical representation of the program output and by symbol *, in numerical delivery.

Nucleosome potential $\varphi(X)$ for promoters

The results obtained by analysis of three gene classes—(1) housekeeping genes, (2) genes expressed in a wide range of tissues, (3) tissue-specific genes—are listed in Table 2.

Initially, average nucleosome formation potential profiles $\varphi(X)$ within the interval of $[-300; +100]$ were

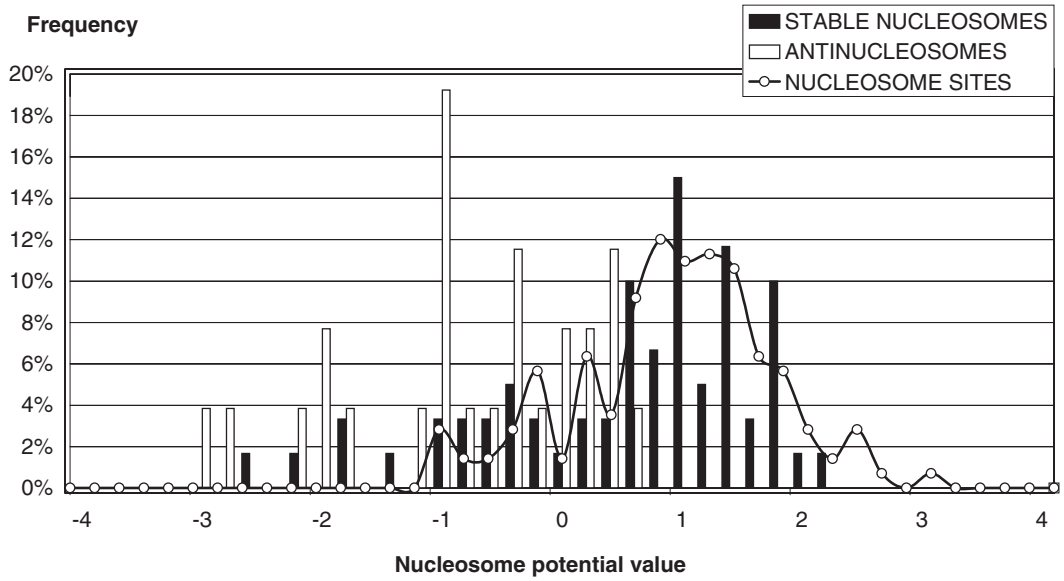

Fig. 6. Histogram of nucleosome formation potential $\varphi(X)$ distributions for the sequence sets of high and low affinity to histone octamer in comparison to distribution $\varphi(X)$ for the set of nucleosome sites.

Table 2. Distribution of nucleosome formation potential values according to the classes of promoters differing in their expression patterns (distribution values are calculated for the promoter region $[-50; +1]$)

Name of promoter class	Mean value*	Number of mean values, (%)	
		In the interval [0; 2]	Outside [0, 2]
Housekeeping genes	-1.48 ± 0.20	11.11	88.89
Genes expressed in a wide range of tissues	-0.66 ± 0.21	21.77	78.13
Tissue-specific genes	$+0.70 \pm 0.08$	80.15	19.85

*All paired differences are significant by Student's test, $P < 0.01$.

constructed for each class in question (Figure 7). All the three classes demonstrated that the nucleosome formation potential $\varphi(X)$ decreased with approaching the transcription start. This result is consistent with our previous data obtained analysing DNA conformational and physico-chemical characteristics on a decreased value of nucleosome site recognition function around the transcription start (Levitsky *et al.*, 1999).

However, the $\varphi(X)$ profiles of these three promoter groups are essentially different. The promoters of tissue-specific genes display the highest $\varphi(X)$ values; the promoters of genes expressing in many tissues show intermediate values; and those of housekeeping genes, exhibit the lowest $\varphi(X)$ values over the entire region considered.

Let us dwell on the promoter region $[-50; +1]$, which has a crucial significance for transcription initiation. The distributions of nucleosome formation potential $\varphi(X)$ for this region of the three promoter classes are shown in Figure 8. Note that the housekeeping gene promoters display the lowest values. The $\varphi(X)$ distribution of the widely expressed genes is appreciably shifted rightward compared with the housekeeping genes. In turn, that of the tissue-specific gene promoters is shifted further rightward.

Nucleosome formation potential values $\varphi(X)$ for various promoter classes are briefed in Table 2, listing the following data: the mean $\varphi(X)$ values within the region $[-50; +1]$ relative to the transcription start, its standard deviation, and the share of promoters in the class such that the $\varphi(X)$ values fall within and outside the interval $[0; +2]$. The mean $\varphi(X)$ value of tissue-specific gene promoters differs significantly from those of housekeeping and widely expressed gene promoters (Student's test, significance level of $P < 0.01$).

A comparison of typical nucleosome formation potential profiles for the genes with different expression patterns is demonstrated in Figure 9. Ubiquitin is a housekeeping gene, whereas prealbumin gene is tissue-specific. It is evident that the potential values in the promoter region of the gene with tissue-specific expression pattern are essentially closer to unity.

DISCUSSION AND CONCLUSIONS

We have developed a method for analyzing nucleosomal organization of genomic DNA sequences and applied it

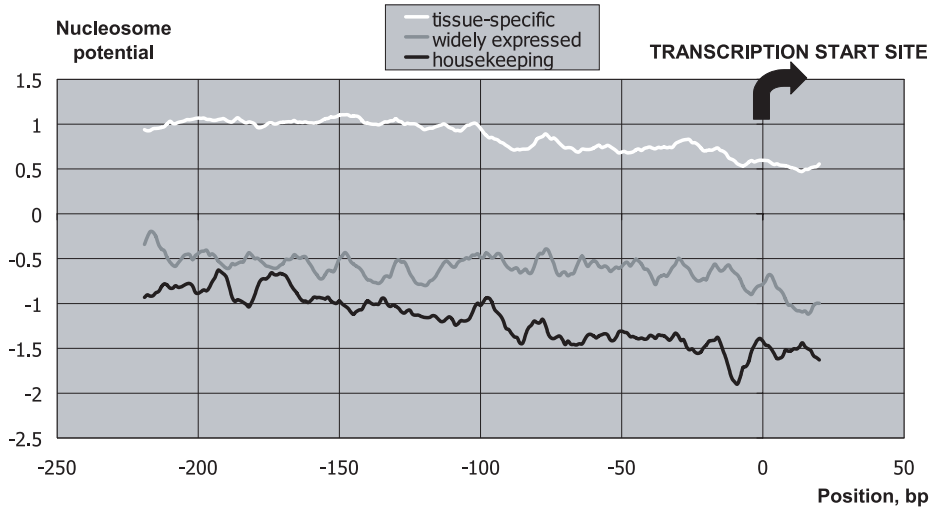

Fig. 7. Average nucleosome formation potential $\varphi(X)$ profiles of gene promoters of different expression classes.

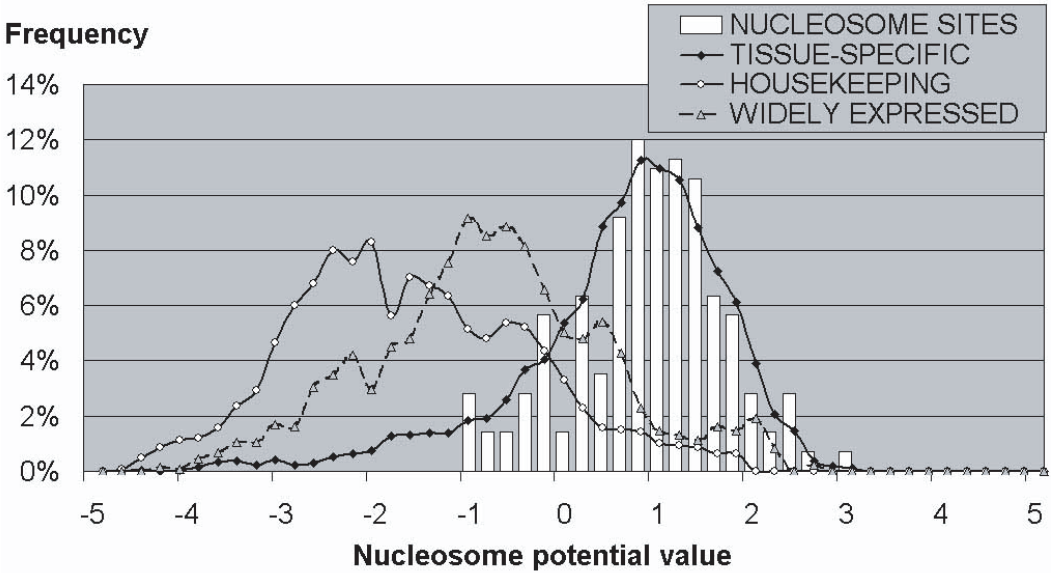

Fig. 8. Histogram of the nucleosome formation potential $\varphi(X)$ distributions for gene promoter regions (for comparison, see the distribution of the training set of nucleosome sites): promoters of housekeeping genes, promoters of genes expressed in various tissues, and promoters of tissue-specific genes.

to studying various functional types of human promoters. The analysis performed has demonstrated that each type of promoter has its own characteristic pattern of nucleosome formation potential profile.

First, we found that typical of promoters, essential for gene expression, is a specific nucleosome positioning pattern around the transcription start. Moreover, it is likely that the pattern of gene expression during the evolution favoured selection of the promoter nucleotide context that would provide for the nucleosome density optimal for their particular function. A trend to increase the nucleosome density might have occurred in the promoters of genes requiring fine tuning (tissue-specific promoters). However, when transcription inhibition under a variety of conditions is disadvantageous (for example, for the genes expressed in many tissues or housekeeping genes), the promoter nucleotide context providing a lesser nucleosome density or even its lack would be favourable.

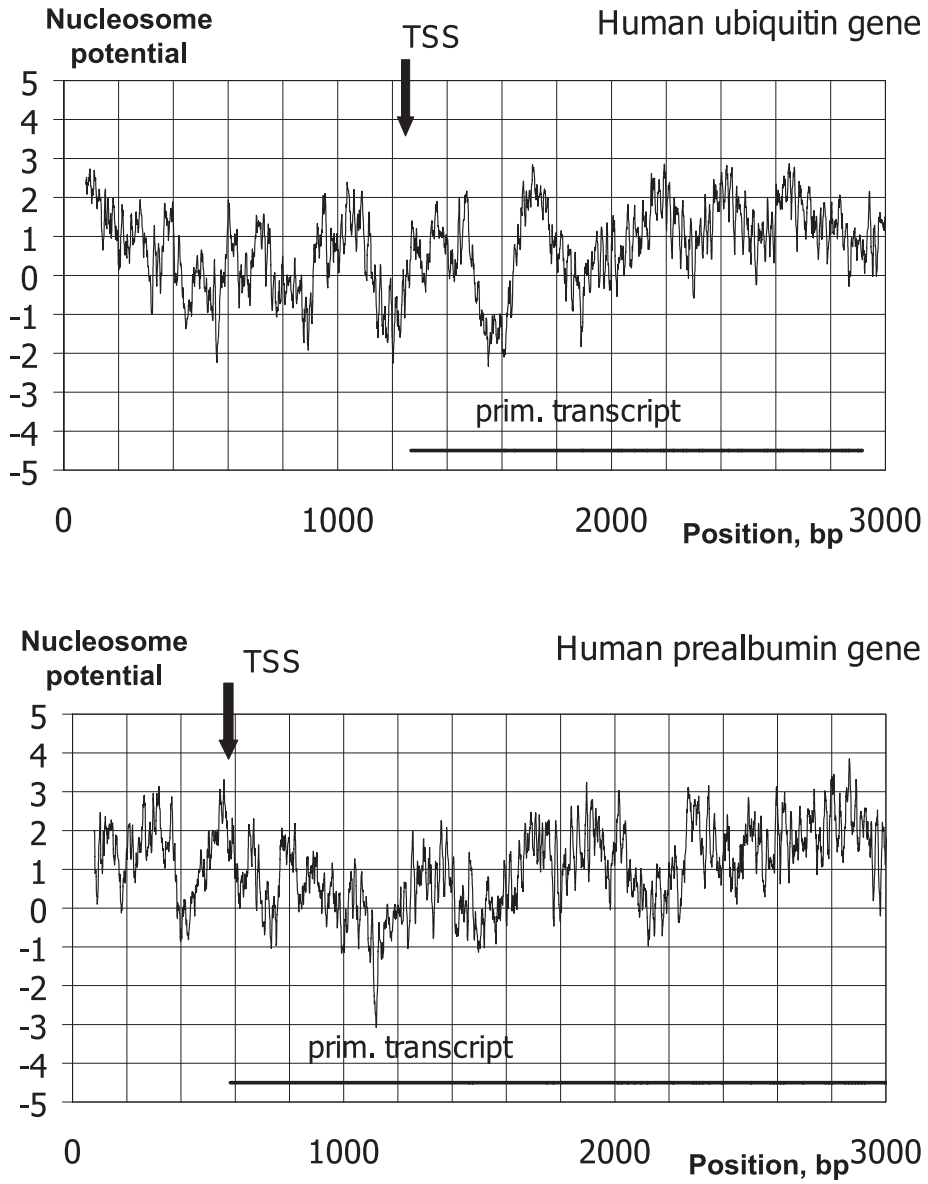

Fig. 9. Nucleosome formation potential $\varphi(X)$ profiles for housekeeping ubiquitin gene (AC U49869) and tissue-specific prealbumin gene (AC M11844). TSS—Transcription Start Site, primary transcript locations are shown below.

An insight into nucleosomal organization of the chromatin is especially important for understanding the mechanisms regulating the gene expression. The experimental data on occurrence of nucleosomes within transcriptionally active regions are miscellaneous. Nucleosome positioning with promoter region often plays an important functional role (Wolffe, 1994; Buttinelli *et al.*, 1993; Fragoso *et al.*, 1995; Weinmann *et al.*, 1999). Such is the case of nucleosome positioned at the site of TATA-Binding Protein (TBP), thereby preventing TBP from interaction with its binding site and repressing the corresponding gene. An alternative variant—the nucleosome precisely positioned in gene regulatory region forms the steric DNA structure necessary for activation of this gene. Enhancers and silencers (Scott *et al.*, 1999), present in eukaryotic genomes and extending their effects on DNA regions located several hundred base pairs away, suggest that nucleosomal organization of the chromatin may be involved in these remote interactions.

An increasing number of experiments report interactions of various transcription factors with nucleosomal DNA (Muro-Pastor *et al.*, 1999; Langst *et al.*, 1998; Shim *et al.*,

1998; Moreira and Holmberg, 1998). This phenomenon might be related to the recently discovered periodicity—a periodic (with a period of 10–10.5 bp) overrepresentation of transcription factor binding sites within regions [−50; +120] relative to the transcription start—assumed to be connected with nucleosome positioning (Ioshikhes *et al.*, 1999). This may be explained with a phased nucleosome location in promoter regions.

Thus, we are proposing a new method for studying the nucleosome organization of genomic DNA involving construction of nucleosome formation potential profiles based on discriminant analysis of dinucleotides frequencies of nucleosome site local regions. This method was used for systematic estimations of the nucleosome formation potential of eukaryotic gene promoters. It has been demonstrated that tissue-specific gene promoters display a higher nucleosome formation potential compared with the potentials of genes expressed in many tissues and housekeeping genes. Nucleosome formation potentials of exons, introns, splice sites, and repetitive sequences were calculated. Essential distinctions in the nucleosome formation potential profiles of donor and acceptor splice sites were discovered.

The method proposed may be useful for functional annotation of the newly determined genomic sequences. The possibility it provides—a computer search for DNA regions with high and low nucleosome formation potentials—is very important for clarifying the molecular mechanisms underlying the functions of genomic sequences.

The approach proposed will be further developed in the following directions:

improving the search for optimal nucleosome site partition into local regions using new optimization methods, in particular, genetic algorithm;

constructing nucleosome formation potential profiles involving the frequencies of trinucleotides, tetranucleotides, etc.; and

studying systematically the nucleosome organization of different classes of genomic sequences.

ACKNOWLEDGEMENTS

The work was supported by the Russian Foundation for Basic Research (grants no 98-07-90126, 98-07-91078, 99-07-90203, 00-07-90337), grant of the Russian National Human Genome Program, grants of the Integrated Program of Siberian Branch of the Russian Academy of Sciences, and the SB RAS grant for Young Scientists. The authors are grateful to D.A.Grigorovich and S.V.Lavryushev for interface design, A.V.Katokhin for useful discussions and to G.B.Chirikova for translating the paper from Russian into English.

REFERENCES

Adams,C.C. and Workman,J.L. (1993) Nucleosome displacement in transcription. *Cell*, **72**, 305–308.

Baldi,P., Brunak,S., Chauvin,Y. and Krogh,A. (1996) Naturally occurring nucleosome positioning signals in human exons and introns. *J. Mol. Biol.*, **263**, 503–510.

Buttinelli,M., Di Mauro,E. and Negri,R. (1993) Multiple nucleosome positioning with unique rotational setting for the *Saccharomyces cerevisiae* 5S rRNA gene in vitro and in vivo. *Proc. Natl Acad. Sci. USA*, **90**, 9315–9319.

Calladine,C.R. and Drew,H.R. (1986) Principles of sequence-dependent flexure of DNA. *J. Mol. Biol.*, **192**, 907–918.

Cao,H., Widlund,H.R., Simonsson,T. and Kubista,M. (1998) TGGA repeats impair nucleosome formation. *J. Mol. Biol.*, **281**, 253–260.

Efron,B. and Gong,G. (1983) A leisurely look at the bootstrap, the jackknife, and cross-validation. *Am. Stat.*, **37**, 36–48.

Fisher,R.A. (1936) The use of multiple measurements in the taxonomic problems. *Ann. Eugen.*, **7**, 179–188.

Fitzgerald,D.J., Dryden,G.L., Bronson,E.C., Williams,J.S. and Anderson,J.N. (1994) Conserved patterns of bending in satellite and nucleosome positioning DNA. *J. Biol. Chem.*, **269**, 21 303–21 314.

Fragoso,G., John,S., Roberts,M.S. and Hager,G.L. (1995) Nucleosome positioning on the MMTV LTR results from the frequency-biased occupancy of multiple frames. *Genes Dev.*, **9**, 1933–1947.

Herrera,R.E., Nordheim,A. and Stewart,A.F. (1997) Chromatin structure analysis of the human c-fos promoter reveals a centrally positioned nucleosome. *Chromosoma*, **106**, 284–292.

Ioshikhes,I. and Trifonov,E.N. (1993) Nucleosomal DNA sequence database. *Nucleic Acids Res.*, **21**, 4857–4859.

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

Ioshikhes,I., Trifonov,E.N. and Zhang,M.Q. (1999) Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. *Proc. Natl Acad. Sci. USA*, **96**, 2891–2895.

Karlin,S. and Ladunga,I. (1994) Comparisons of eukaryotic genomic sequences. *Proc. Natl Acad. Sci. USA*, **91**, 12 832–12 836.

Kolchanov,N.A., Podkolodnaya,O.A., Ananko,E.A., Ignatieva,E.V., Stepanenko,I.L., Kel-Margoulis,O.V., Kel,A.E., Merkulova,T.I., Goryachkovskaya,T.N., Busygina,T.V., Kolpakov,F.A., Podkolodny,N.L., Naumochkin,A.N., Korostishevskaya,I.M., Romashchenko,A.G. and Overton,G.C. (2000) Transcription Regulatory Regions Database (TRRD): its status in 2000. *Nucleic Acids Res.*, **28**, 298–301.

Kornberg,R.D. and Lorch,Y. (1999) Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, **98**, 285–294.

Langst,G., Becker,P.B. and Grummt,I. (1998) TTF-I determines the chromatin architecture of the active rDNA promoter. *EMBO J.*, **17**, 3135–3145.

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

Levitsky,V.G., Katokhin,A.V. and Kolchanov,N.A. (2000) Inherent

modular promoter structure and its application for recognition tools development. *Comput. Technol. (Novosibirsk)*, **5**, 41–47.

Li,G., Chandler,S.P., Wolffe,A.P. and Hall,T.C. (1998) Architectural specificity in chromatin structure at the TATA box in vivo: nucleosome displacement upon beta-phaseolin gene activation. *Genes Dev.*, **12**, 5–10.

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

Mengeritsky,G. and Trifonov,E.N. (1983) Nucleotide sequence-directed mapping of the nucleosomes. *Nucleic Acids Res.*, **11**, 3833–3851.

Montecino,M., Lian,J., Stein,G. and Stein,J. (1996) Changes in chromatin structure support constitutive and developmentally regulated transcription of the bone-specific osteocalcin gene in osteoblastic cells. *Biochemistry*, **35**, 5093–6102.

Moreira,J.M. and Holmberg,S. (1998) Nucleosome structure of the yeast CHA1 promoter: analysis of activation-dependent chromatin remodeling of an RNA-polymerase-II-transcribed gene in TBP and RNA pol II mutants defective in vivo in response to acidic activators. *EMBO J.*, **17**, 6028–6038.

Muro-Pastor,M.I., Gonzalez,R., Strauss,J., Narendja,F. and Scazzocchio,C. (1999) The GATA factor AreA is essential for chromatin remodelling in a eukaryotic bidirectional promoter. *EMBO J.*, **18**, 1584–1597.

Pedersen,A.G., Baldi,P., Chauvin,Y. and Brunak,S. (1999) The biology of eukaryotic promoter prediction—a review. *Comput. Chem.*, **23**, 191–207.

Perier,R.C., Praz,V., Junier,T., Bonnard,C. and Bucher,P. (2000) The eukaryotic promoter database. *Nucleic Acids Res.*, **28**, 302–303.

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

Scott,K.C., Taubman,A.D. and Geyer,P.K. (1999) Enhancer blocking by the *Drosophila* gypsy insulator depends upon insulator anatomy and enhancer strength. *Genetics*, **153**, 787–798.

Shim,E.Y., Woodcock,C. and Zaret,K.S. (1998) Nucleosome positioning by the winged helix transcription factor HNF3. *Genes Dev.*, **12**, 5–10.

Sivolob,A.V. and Kharpunov,S.N. (1995) Translational positioning of nucleosomes on DNA: the role of sequence-dependent isotropic DNA bending stiffness. *J. Mol. Biol.*, **247**, 918–931.

Solovyev,V. and Salamov,A. (1997) The gene-finder computer tools for analysis of human and model organisms genome sequences. In *Proceedings, Fifth International Conference on Intelligent Systems for Molecular Biology(Ismb-97)*. pp. 294–302.

Solovyev,V.V., Salamov,A.A. and Lawrence,C.B. (1994) Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.*, **22**, 5153–5156.

Staden,R. (1984) Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.*, **12**, 505–519.

Staffelbach,H., Koller,T. and Burks,C. (1994) DNA structural patterns and nucleosome positioning. *J. Biomol. Struct. Dyn.*, **12**, 301–325.

Steger,D.J. and Workman,J.L. (1996) Remodeling chromatin structures for transcription: what happens with histones? *Bioessay*, **18**, 875–884.

Stein,A. and Bina,M. (1999) A signal encoded in vertebrate DNA that influences nucleosome positioning and alignment. *Nucleic Acids Res.*, **27**, 848–853.

Thoma,F. (1992) Nucleosome positioning (Review). *Biochim. Biophys. Acta*, **1130**, 1–19.

Trifonov,E.N. and Sussman,J.L. (1980) The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc. Natl Acad. Sci. USA*, **77**, 3816–3820.

Trifonov,E.N. (1997) Genetic level of DNA sequences is determined by superposition of many codes. *Mol. Biol. (Mosk.)*, **31**, 759–767.

Uberbacher,E.C., Harp,J.M. and Bunick,G.J. (1988) DNA sequence patterns in precisely positioned nucleosomes. *J. Biomol. Struct. Dyn.*, **6**, 105–120.

Ulyanov,A.V. and Stormo,G.D. (1995) Multi-alphabet consensus algorithm for identification of low specificity protein-DNA interactions. *Nucleic Acids Res.*, **23**, 1434–1440.

van Holde,K.E. (1989) *Chromatin*. Springer, Berlin.

Weinmann,A.S., Plevy,S.E. and Smale,S.T. (1999) Rapid and selective remodeling of a positioned nucleosome during the induction of IL-12 p40 transcription. *Immunity*, **11**, 665–675.

Widlund,H.R., Cao,H., Simonsson,S., Magnusson,E., Simonsson,T., Nielsen,P.E., Kahn,J.D., Crothers,D.M. and Kubista,M. (1997) Identification and characterization of genomic nucleosome-positioning sequences. *J. Mol. Biol.*, **267**, 807–817.

Wolffe,A.P. (1994) Nucleosome positioning and modification: chromatin structures that potentiate transcription. *Trends Biochem. Sci.*, **19**, 240–244.

Zhang,M.Q. (1997) Identification of protein coding region in the human genome by quadratic discriminant analysis. *Proc. Natl Acad. Sci. USA*, **94**, 565–568.

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