Mathematical Biology & Bioinformatics. 2011. V. 6. № 2. P. t36-t52.
URL: http://www.matbio.org/2011/Panjukov2011(6_t36).pdf
Original Russian text
Panyukov V.V., Nazipova N.N., Ozoline O.N., 2011, published in Matematicheskaya biologiya i bioinformatika, 2011, V. 6. № 2. P. 211–227. URL: http://www.matbio.org/2011/Panjukov2011(6_211).pdf

========== BIOINFORMATICS ==========

UDC: (577.214.625+004.93):519.688

Software package as a tool to study the spatial conformation of bacterial genome sites

Panyukov V.V.*[1], Nazipova N.N.[1], Ozoline O.N.[2]

[1]*Institute of mathematical problems of biology, Russian academy of sciences, Pushchino, Moscow Region, 142290, Russian Federation*
[2] *Institute of cell biophysics, Russian academy of sciences Pushchino, Moscow Region, 142290, Russian Federation*

Abstract. The paper describes an interactive package aSHAPE capable of revealing the conformational peculiarities in the families of DNA double helices by means of studying their 3D models. A package ability to discriminate between two families, containing promoters of Escherichia coli and non-promoter DNA, is demonstrated.

Key words: *DNA conformation, conformation chain, discriminant analysis, metrical parameters, DNA double helix, promoter, modeled segment*.

INTRODUCTION

The ability of genomic DNA to initiate RNA synthesis is largely determined by the structural and conformational features of its regulatory regions (promoters) [1–8]. The various protein factors play a pivotal role in creating an optimal spatial configuration of the promoter DNA and, depending on needs; these either help or, conversely, inhibit the formation of the transcription complex. Importantly, however, that the ability of the promoter region of undergoing the conformational transitions is supported by the optimal distribution of structure-specific homonucleote tracks and flexible kink-forming dinucleotides which are inherent in the nucleotide sequence of promoter regions, [9-11]. This allows one to conclude that the three-dimensional structure of the free promoter DNA may have a particular spatial configuration. Indeed, the investigation of short promoter - specific DNA-fragments with X-ray [12] and nuclear magnetic resonance (NMR) [13] confirmed their ability to form stable bends. Unfortunately, such tools are so far inapplicable to full-length promoters. Therefore, in this paper, we apply *in silico* internet resource “DNA tools” [14] to study the architectural features of promoters using 3D modeled DNA fragments.

A number of quantitative conformational parameters called metrics were introduced to describe the spatial form of a DNA fragment. Following this way we undertook the task to develop a software package that for a given two families of 3D modeled DNA regions makes it possible to figure out, which metric parameters discriminate between these families. The stated problem can be rephrased as a one-way analysis of variance (ANOVA) that is aimed to find variables discriminating given data sets [15, 16]. So we meet a challenge of choosing the metric parameters reflecting conformational state of such spatial multi-atomic system as DNA fragment. Digest of methods characterizing structural peculiarities within the DNA double helix can be found in [17].

Widely used approximation of the DNA double helix in literature is an elastic rod. In this case the base pairs are in the interior of the rod, while the sugar-phosphate chains lie on its surface. Undeformed rod imitates an ideal Watson-Crick B-form. Conformational transition of the rod from the ideal state deforms the geometrical lines tracking the shape of this rod. Thus, in order to describe conformational state of the rod and, accordingly, of the DNA fragment we choose several broken lines and watch for their conformation. Usually, broken line is defined by its segments, which we called *links*. But we identify the broken line via the endpoints of links called in the paper *vertices*. The package exploits atomic or virtual vertices.

Base pairs of the given DNA fragment were numbered in the natural order from the 5'- to the 3'-end. Selected atoms, for example Y(C6) and R(C8), connected in order of increasing base pair number constitute the atomic *chain*. And the centers or midpoints of Y(C6) –R(C8) doublet in a DNA fragment are considered as vertices of the virtual *chain*.

The virtual *chain* of the centers of base pairs is used by number of programs and servers [18–25] for calculating the DNA curvature. Since each of the computer program uses its own approximations for virtual vertices, the curvature values of the modeled molecules appeared to be different (reported by Barbic and Crothers [26], who compared such popular programs as 3DDNA [20], Curves 5.1 [21] and Freehelix98 [24]). We can add to this list two programs CURVATURE [23] and ADN-Viewer [27]. While CURVATURE approximates the curved helix axis by plane curve and measures the radius of curve bending, ADN-Viewer [27] proposes to calculate the angles between two consecutive *links* of the chain.

It should be clear that particular *chain* reflects only some features of the modeled DNA fragment; so more or less complete picture can be obtained by using several different *chains*. Indeed, when one end of the ideal DNA fragment – an elastic rod is fixed, while another end is rotated around the helix axis, each *chain* of vertices associated with the helix axis remains unchanged and one needs another chain to reveal the conformational transition. The package aSHAPE uses several chains to study the shape of DNA fragments. Having been the methodological feature of the package aSHAPE, this approach allows revealing important conformational details of DNA fragment.

Whereas a given chain reflects the shape of the fragment, its metric parameters show the amplitudes of the observed deformations. The package provides a list of metric parameters. Not all of them are equally informative. Therefore, the description of the DNA molecule in terms of its conformation involves two steps: the choice of a representative *chain* and the numerical estimate of its shape with metric parameters. Currently we are not ready to offer the universal recipes for the two steps but give the opportunity to choose the optimal way within the range of available options.

DESCRIPTION OF THE aSHAPE

Source of data

Two families (Prm and Cont), each one composed of 300bp modeled DNA regions, were used for developing, debugging and testing the software package aSHAPE. Prm contains the models of 180 bacterial promoters recognized by the σ^{70}-subunit of the *E. coli* RNA polymerase. These promoters have been used previously [11] as examples of single transcription signals. The sequences of Prm were aligned so that the transcription start was placed in position 151 and was flanked by 149 bp of the transcribed region. Previously [11] was also characterized a set that contains 272 DNA regions chosen from the nonpromoter genomic regions of *E.coli*. Fifty-one randomly selected members of this set of length 300bp were 3D modeled and placed into *Cont*.

The atomic coordinates in PDB format of modeled DNA regions were calculated with DNA Tools [14] and were used to compute the coordinates of the chain vertices.

Designations and agreements

The package is designed to analyze and visualize the deformity of 300 bp DNA modeled regions combined into a family. Since biological properties of DNA are dependent on its geometric form, we searched the family of DNA regions with definite biological function for its spatial peculiarities, which have been localized by means of fragments of the modeled region. It is clear that the different spatial peculiarities require fragments of different size for their localization. Thus, for an adequate analysis one has to assign the type of chain, metric parameter *f* and the fragment size *FragSz*.

A *RegSz* bp long region contains $RegSz - FragSz + 1$ fragments, where each next fragment occupies the position shifted by one base pair compared to the previous one (Fig. 1).

A fragment is identified completely by the position *p* of its first base pair relative to the beginning of modeled DNA region, hence the value of parameter *f* for the fragment may be written as *f*(*p*). We assume that *FamSz*(*RegSz* – *FragSz* + 1) ensemble of *f*(*p*) values pooled over all the family fragments reflects the spatial properties of this family with regard to considered metric parameter *f*. Such ensembles obtained for different DNA families can reveal their structural peculiarities especially if the corresponding histograms do not completely overlap each other.

In order to implement the histogram analysis each modeled DNA region of all DNA families was analyzed in the reference frame so as its first vertex coincided with the frame origin, where *x*-axis enumerated positions of base pairs and *y*-axis was assigned for the metric parameter values. It allows using a common plane for the plotting and comparing *f*-histograms of different families.

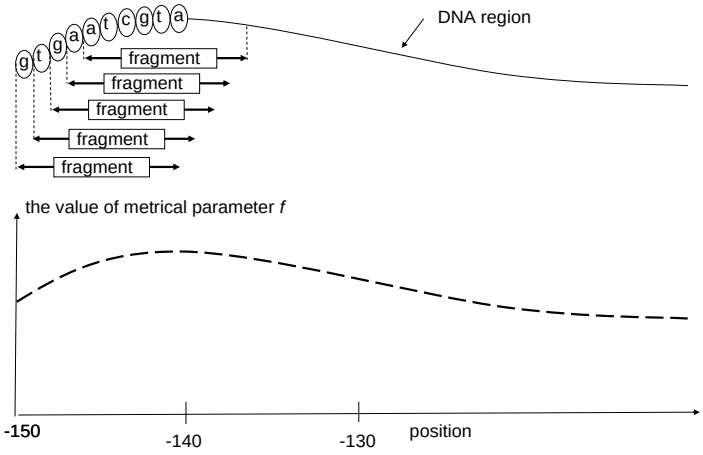

Fig. 1. The upper scheme illustrates fragments covering the region of interest. The dashed line on the lower plot presents the values of a metric parameter *f* for all fragments within the modeled region.

If *Fm* is a family of modeled DNA regions and *Ens*(*Fm*) is the corresponding ensemble of metric data then the histogram denoted by *h*(*Fm*) can be plotted as a portrait of analyzed feature. For the bin Δ sized by user the term *h*(*x*) by default means the normalized frequencies of *f* within $\delta = [x - \Delta/2,\ x + \Delta/2]$. Another optional case calls normalizing factor *RegNu* and visualizes *h*(*x*) as the density of $f \in \delta$ per family.

The histograms *h*(*Fm1*) and *h*(*Fm2*) representing two families *Fm1* and *Fm2*, are plotted in Fig. 2. They allow one to know whether the ensembles *Ens*(*Fm1*) and *Ens*(*Fm2*) are separated in terms of metrical parameter *f* or not.

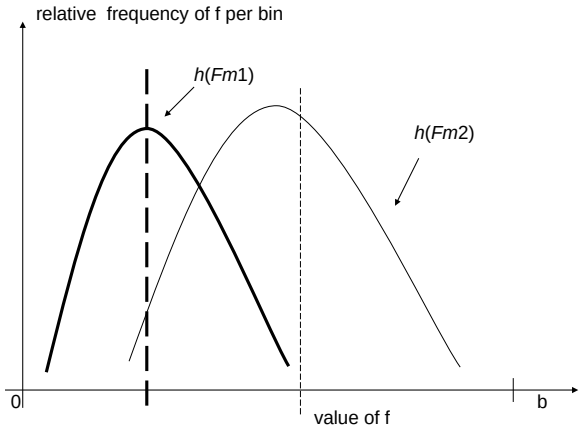

Fig. 2. The histograms of the ensembles *Ens*(*Fm1*)∪*Ens*(*Fm2*). Vertical dashed lines indicate the means each across the ensemble.

Let σ(*Ens*) denotes STD deviation for a quantitative ensemble *Ens*. Fig.2 shows that $\sigma(Ens(Fm1) \cup Ens(Fm2)) > max\{\sigma(Ens(Fm1)), \sigma(Ens(Fm2))\}$. In this case discriminant analysis testifies that the ensembles *Ens*(*Fm1*) and *Ens*(*Fm2*) are separated and we signify it as $h(Fm1) < h(Fm2)$.

If particular metric parameter *f* from the package list was used to compare *Ens*(*Fm1*)) and *Ens*(*Fm2*), then relative disposition of histograms *Fm1* and *Fm2* depends on *chain* type and *FragSz*. By our definition, the parameter *f* is discriminative if inequality $h(Fm1) < h(Fm2)$ (or $h(Fm2) < h(Fm1)$) is fulfilled despite of both *FragSz* and chain type variations.

Histograms *h*(*Fm*1) and *h*(*Fm*2) usually overlap depending on *FragSz*. The overlaping area *Sq* may be quantified. The *FragSz* giving the smallest *Sq* is considered as biologically relevant.

The choice of the spatial model of the DNA molecule

Most adequately the spatial configuration of DNA can be modeled using 3D structures obtained for short DNA fragments by X-ray diffraction or NMR, but these methods are not yet able to process long areas of the promoter DNA. That is why, we employed online software “DNA tools” [14], that models up to 700 base pairs, exploiting experimentally estimated distance and angular parameters of dinucleotides.

This instrument allows us to build a numerous conformational chains associated with each modeled DNA region and to use another conformational parameters instead of standard ones like *tilt*, *roll* and *twist* [28]. The point is that the standard parameters are not completely independent [29], and hence the width of histogram that shows, for example, the *tilt*-specific deformity may be influenced by variations of *roll*. The greater *roll* variations the wider histograms, and, as a consequence, the greater deal of histogram overlap. Wide histograms conceal the genuine relations between families.

Conformation and chains

The package aSHAPE estimates DNA conformation via *chains*. In this chapter we specify certain chains from the package list, which are aimed to reveal different conformation features of the modeled region. A chain is identified by its own vertices, whose coordinates are defined by 3D vector.

Vertex that is *doublet center* (phosphorous or carbon)**.** If v_i and w_i are the coordinates of the doublet atoms of the *i*-th base pair then the coordinates of *doublet center* is $(v_i + w_i)/2$. Fig. 3 illustrates the *chain* of phosphorus doublet centers.

Vertex that is *helix center*. Consider two neighboring phosphorus doublets. Let $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ be the coordinates of the phosphors which belong to the strand **A**, while $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ be the coordinates of phosphors on the strand B. We assume that this dinucleotide fragment forms a double helix with common axis if the coordinates satisfy the equations:

$$|\boldsymbol{a}_1 - \boldsymbol{b}_1| = |\boldsymbol{a}_2 - \boldsymbol{b}_2|$$

and

$$|\boldsymbol{a}_1 - \boldsymbol{a}_2| = |\boldsymbol{b}_1 - \boldsymbol{b}_2| \ (*).$$

The plane passing through $\boldsymbol{a}_1$, $\boldsymbol{b}_1$ perpendicular to the helix axis intersects it at the point which we call *helix center*. Ideal B-form DNA strongly satisfies (*). As to real *Prm* and *Cont* families, the carried out calculations showed:

– the variance of phosphorus doublet size is 0.14 Å with a mean equal to17.8Å;

– the variance of the distance between the ends of neighboring doublets is 0.18 Å with a mean value 6.5Å.

We see that modeled regions of families *Prm* and *Cont* satisfy equations (*) with high accuracy, which makes utilization of helix centers reasonable.

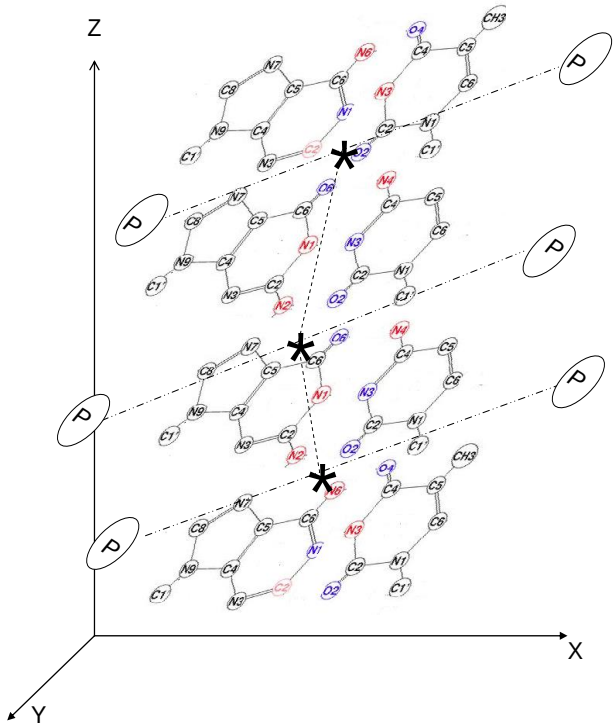

Fig. 3. The phosphorus doublet centers of 4 bp DNA fragment are marked by asterisks. Phosphorus atoms are rounded by large ovals. A pair of phosphorus belonging to the complementary base pairs are combined into a doublet (dashed segment line). Drawings of complementary base pairs are taken from [30] and adapted.

In summary, the package operates with four types of chains, namely, the chain of phosphorus doublet centers, the chain of carbon (C6−C8) doublet centers, the chain of phosphorus helix centers and the chain of carbon (C6−C8) helix centers. All they are virtual; however we chose them because the atomic chains of even ideal DNA have rather complex coiled structure, and deformation complicates it, making the structural analysis very difficult.

Local and integral conformation of chains

Conformational *chains* characterize the conformational state of the DNA fragment. Local *bending* angle θ, torsion angle φ and *stretching* are introduced for estimating chain conformation.

Consider the vertices of a chain identified by coordinates $\boldsymbol{v}_i,\boldsymbol{v}_{i+1},\boldsymbol{v}_{i+2},\boldsymbol{v}_{i+3}$, where *i* runs over all the *chain* vertices, and vector $\boldsymbol{v}_{i+1} - \boldsymbol{v}_i$ linking vertices $\boldsymbol{v}_i$ and $\boldsymbol{v}_{i+1}$.

The *bending* at the vertex $\boldsymbol{v}_i$, denoted by θ, is equal to the angle between adjacent links $\boldsymbol{v}_{i+1} - \boldsymbol{v}_i$ and $\boldsymbol{v}_{i+2} - \boldsymbol{v}_{i+1}$.

To measure the *torsion angle* at the vertex $\boldsymbol{v}_i$, denoted by ϕ, we need four neighboring vertices, $\boldsymbol{v}_i, \boldsymbol{v}_{i+1}, \boldsymbol{v}_{i+2}, \boldsymbol{v}_{i+3}$. By definition, ϕ equals to the angle between the link $\boldsymbol{v}_{i+3} - \boldsymbol{v}_{i+2}$ and the plane, passing through $\boldsymbol{v}_i$, $\boldsymbol{v}_{i+1}$, and $\boldsymbol{v}_{i+2}$.

The *stretching* at the vertex $\boldsymbol{v}_i$ equals to the difference in length of two adjacent links, namely $||\boldsymbol{v}_{i+1} - \boldsymbol{v}_i| - |\boldsymbol{v}_{i+2} - \boldsymbol{v}_{i+1}||$.

It is clear that evaluating these three parameters at every vertex of a chain provides the full information on the chain form, see Figure 4.

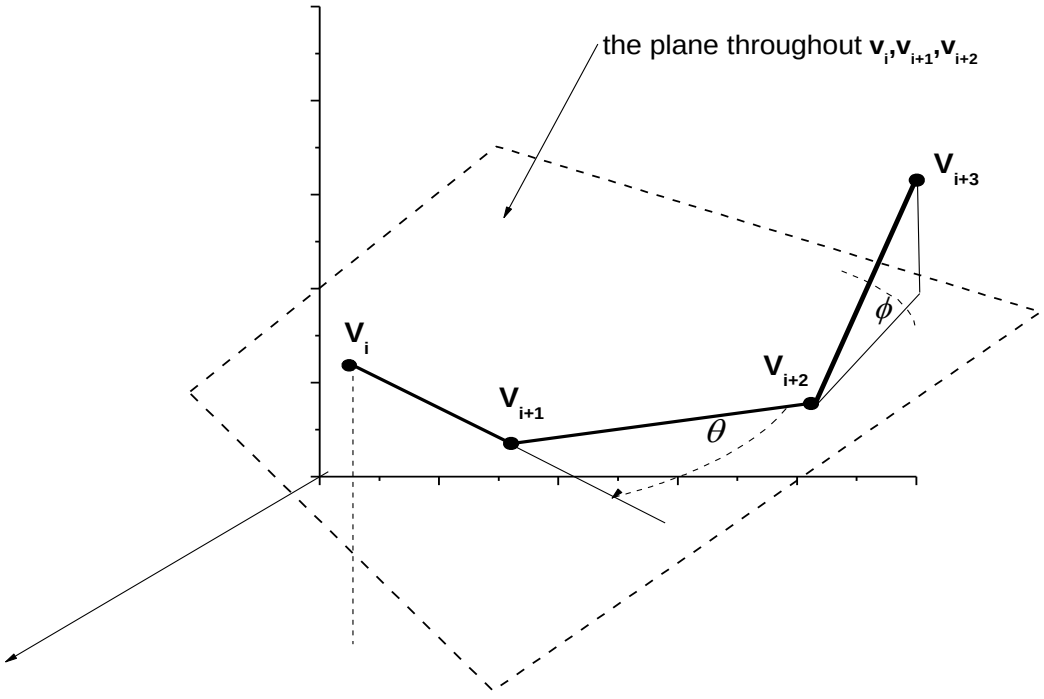

Fig. 4. Four-vertex chain segment explaining the determination of the angles θ and ϕ.

In order to reveal a possible correlation among θ,ϕ and *stretching*, the pairwise scatter plots of these parameters were computed for *Prm* and *Cont* families. Correlations were not found in any case. Figure 5 exemplifies this situation for the *torsion angle* ϕ – *stretching* scatter plots.

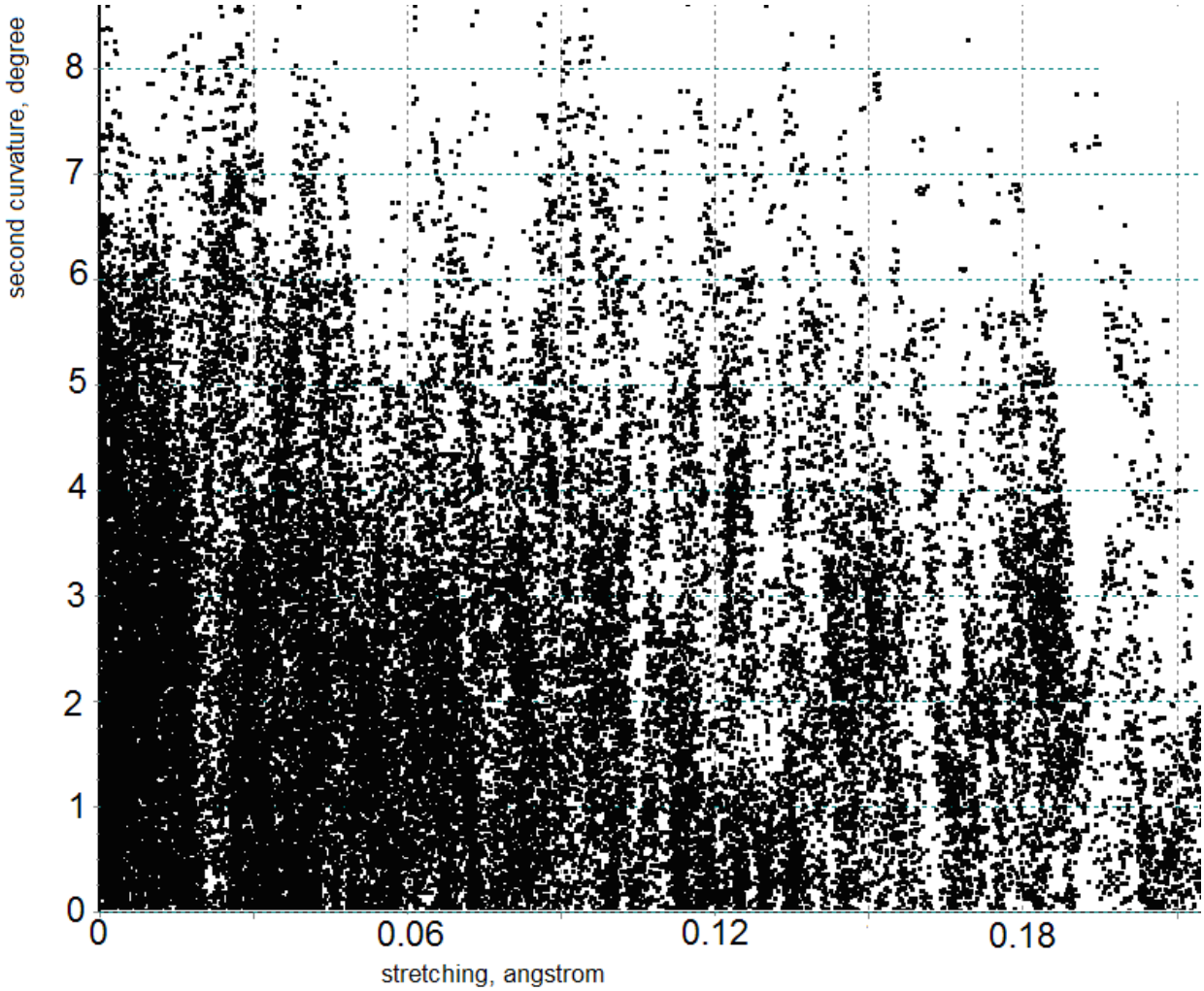

Fig. 5. The scatter plot of the *torsion angle* φ and *stretching* for Prm family.

Since a chain acquires its shape under the cooperative effect of the local deformations, several metric parameters accumulating the local deformations are introduced in order to characterize the global conformation of the chain.

Homogeneity

The magnitude of a given metric varies over a chain and *homogeneity* assesses the range of these variations. Let $\overline{\theta}$ and $\overline{\varphi}$ denote the average angles. The following metric parameters measure *homogeneity* of local deformations.

$$BendHomo\theta(S)=\sum_i\left|\overline{\theta}-\theta_i\right|,\ BendHomo\phi(S)=\sum_i|\overline{\varphi}-\varphi_i|,\ LinkHomo(S)=\sum_i\left|l_i\right|.$$

It was found that *BendHomo*θ and *BendHomo*φ discriminate families *Prm* and *Cont* but *LinkHomo* fails. Indeed, Fig.6 indicates that *h*(*Cont*) < *h*(*Prm*) for *BendHomo*θ, which appears to be true for phosphorous doublet centers and for phosphorous helix centers and remains true despite of *FragSz* variation.

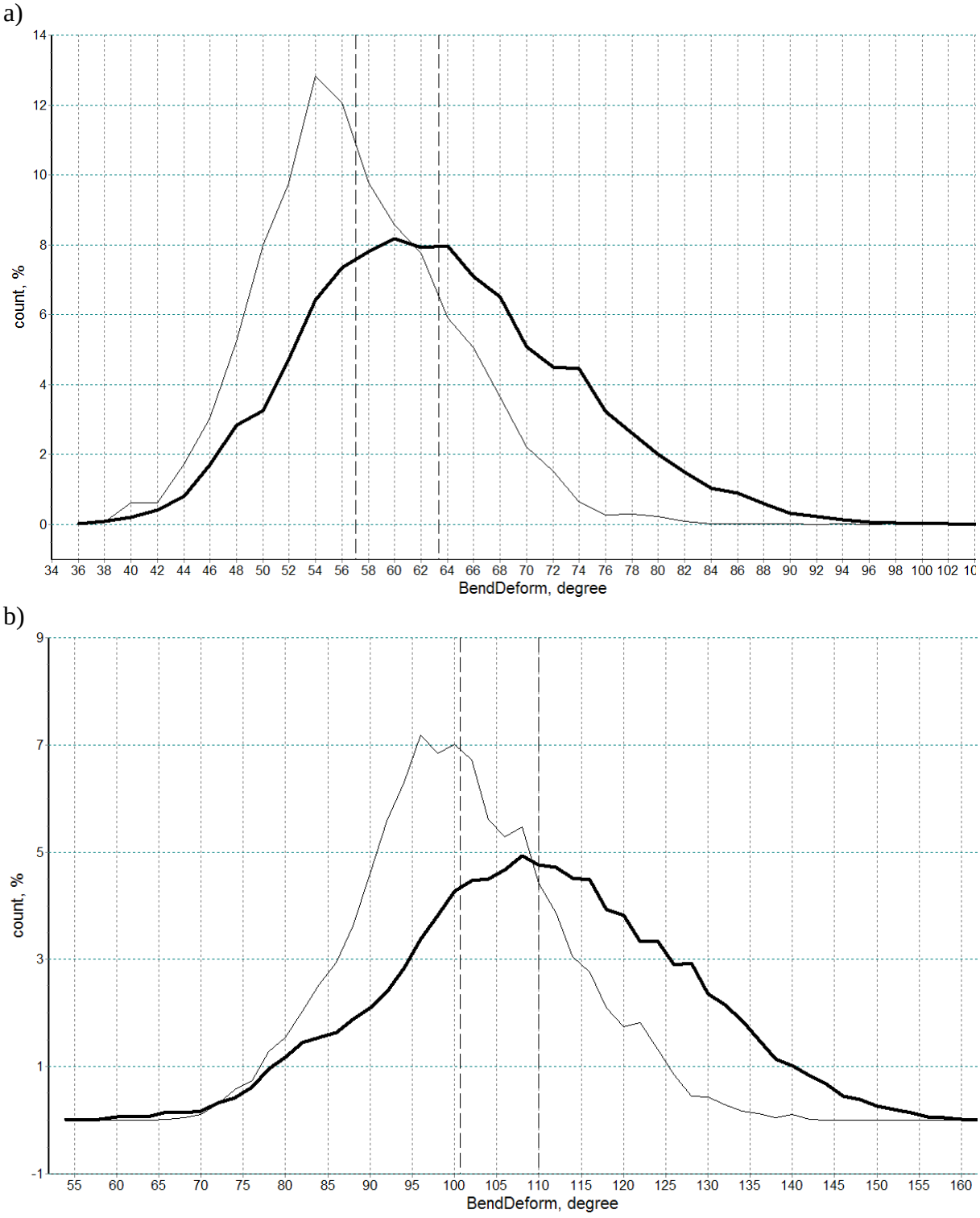

Fig. 6. Histograms *h*(*Prm*) (thick line) and *h*(*Cont*) (thin line) showing the homogeneity of θ measured with *BendHomo*θ for the chain of phosphorous doublet centers (a) and for the chain of phosphorous helix centers (b). *Sq*(*Prm* ∩ *Cont*) are equal to 0.69 (a) and 0.72 (b), respectively. *FragSz* = 60bp.

A similar result holds for the homogeneity measured with *BendHomo*φ. Hence *BendHomo*θ and *BendHomo*φ are discriminative. But the parameter *LinkHomo* is not discriminative because it produces histograms with *Sq(Prm* ∩ *Cont)* ≈ 1.

We tested a lot of metrical parameters to find discriminative ones. Many of them, for example $\bar{\theta}$, appeared to be non-discriminative because inequality *h*(*Cont*) < *h*(*Prm*) $\bar{\theta}$ in Fig. 7a is not stable under the changing chain type, Fig. 7b.

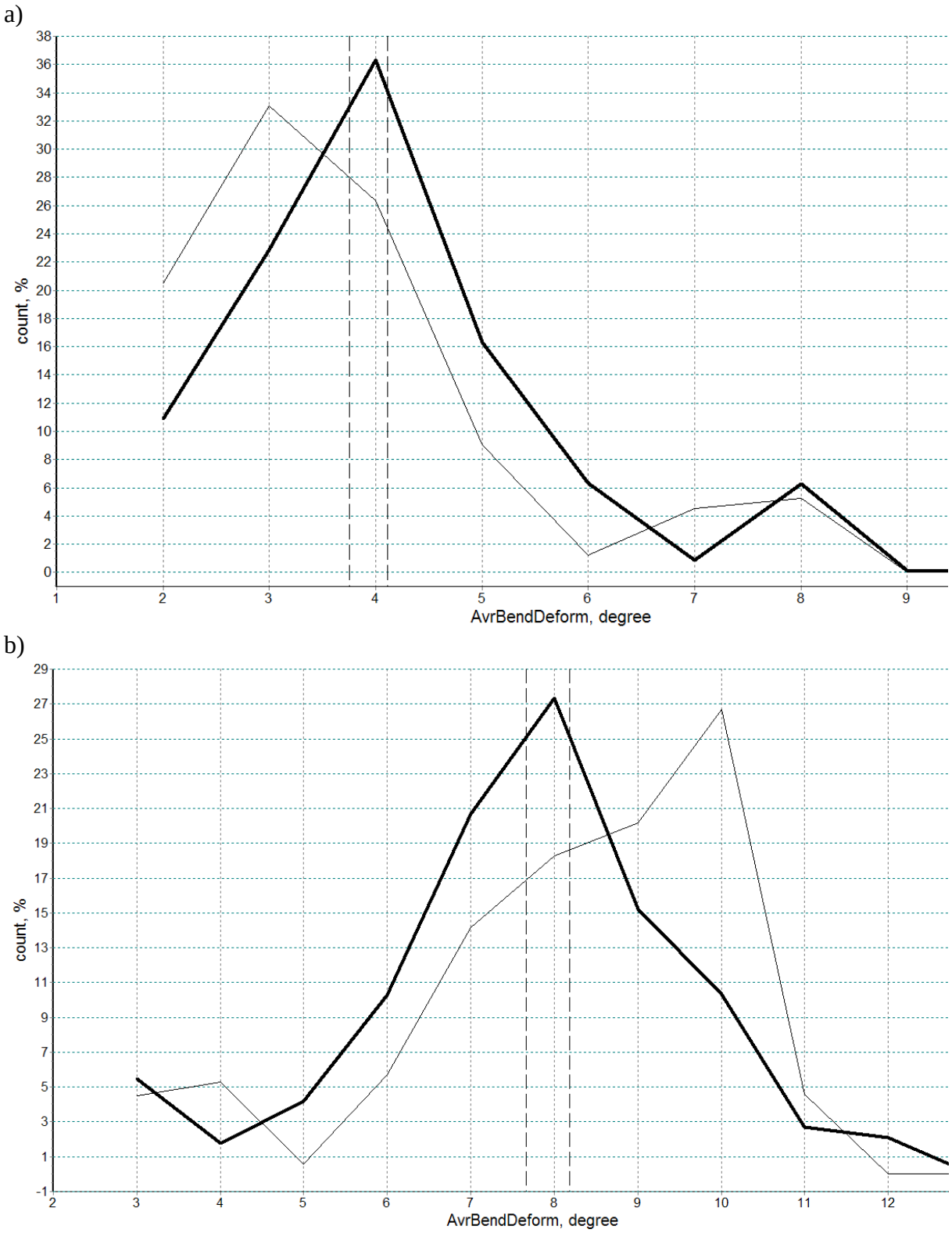

Fig. 7. Histograms $h(Prm)$ (thick line) and $h(Cont)$ (thin line) showing the $\overline{\theta}$ distributions for $FragSz$ = 60bp. (a) The chains of phosphorous doublet centers. $h(Cont) < h(Prm)$, $Sq(Prm \cap Cont) = 0.65$. (b) The chains of phosphorous helix centers. $h(Prm) < h(Cont)$, $Sq(Prm \cap Cont) = 0.70$.

Shape

In the section, we consider several metric parameters provided by package, which measure the global structural deviations of a modeled DNA fragment from the ideal B form.

The parameters measuring *chain* shape. Let $\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_n$ be the vertices of the chain $\boldsymbol{S}$. The *contour length* of the whole chain or fragments equals to the sum of their links lengths, and the *streingthened length* is the distance between their ends. *LengthDeform* measures the difference between these two metrics.

$$LengthDeform(S) = (\Sigma|\boldsymbol{v}_{i+1} - \boldsymbol{v}_i|) - |\boldsymbol{v}_n - \boldsymbol{v}_1| .$$

MaxJut evaluates the maximal deviation of the chain (fragment) vertices from the straight line, connecting their ends (**SL**).

$$MaxJut(S) = max\{dist(\boldsymbol{v}_i) \mid i = 1, \ldots,n\},$$

where $dist(\boldsymbol{v}i)$ denotes distance from vertex $\boldsymbol{v}_i$ to **SL**.

If the carbon doublet centers are the vertices of the chain $\boldsymbol{S}$ then $LengthDeform(S) \approx 0$ and $MaxJut(S) \approx 0$ for ideal B-form.

The parameters measuring the shape of a doublets contour. The ends of the phosphorus and carbon doublets outline the contour of the double helixes in the classical B-form DNA. Several options of the package allow to measure deviations of the doublets helicity from the ideal form. For a given DNA fragment the package processes the ends of all doublets $\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_m$, where $m = 2FragSz$. If we place the origin of coordinates in the midpoint of these doublets, we have $\Sigma\boldsymbol{v}_i = \boldsymbol{0}$. At the first stage aSHAPE finds a straight line L passing through the origin as the solution to $D(L) = \Sigma(\boldsymbol{v}_i, \boldsymbol{l})^2 = minimum$ under the condition $\boldsymbol{l}^2 = 1$, where $\boldsymbol{l}$ is the vector along the axis L. The method of principal components [31] allows to obtain the solution of this, which is interpreted as the axis of ideal double helix approximating all the data. For a modeled fragment of ideal B-form the solution $\boldsymbol{L}$ to the carbon doublets coincides with the axis of double helix and gives $D(L) = 0$ because every carbon doublet is perpendicular to the axis.

At the second stage aSPAPE evaluates the doublets deviation from the helix L axis gained at the first step. Let r_i denotes the distance from the axis L to the vertex $\boldsymbol{v}_i$.
The package provides the following metric parameters:

$$R_{min}(SPP) = min\{r_1,\ldots,r_m\},$$
$$R_{max}(SPP) = max\{r_1,\ldots,r_m\},$$
$$R_{avr}(SPP) = (r_1+\ldots+r_m)/m,$$
$$RadiusRange(SPP) = R_{max}(SPP) - R_{min}(SPP).$$

It is clear that for the doublets of the ideal B-form all these parameters are equal to zero.

Obviously, that the combinatorial use of different metrics allows more deep insight into the fragment shape and permits to make inferences more confident. For instance, Fig. 8 confirms the relationship revealed by Fig. 6 and both figures indicate that the DNA fragments of Prm and Cont are deformed in different ways.

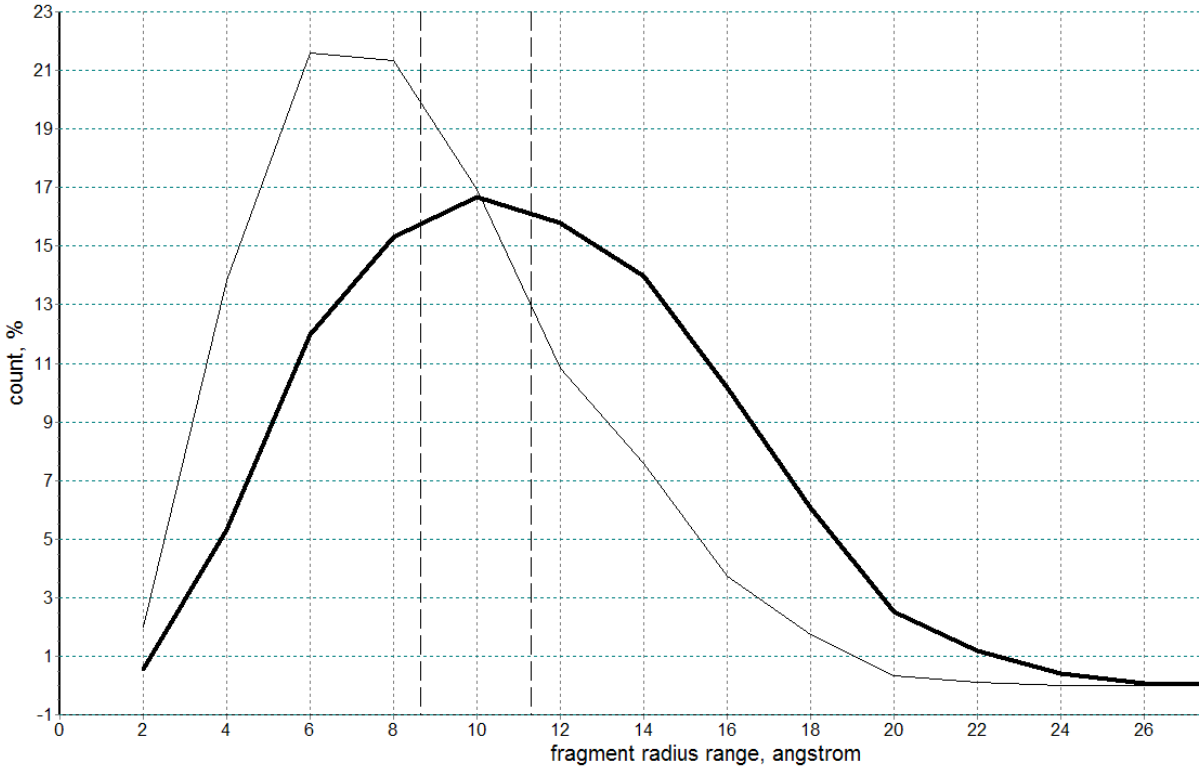

Fig. 8. Histograms of *RadiusRange* for *Prm* (in thick) and *Cont* (in thin) obtained by the chains of phosphorous doublet centers. *Sq*(*Prm* ∩ *Cont*) = 0.73. *FragSz* = 60 bp.

Additional tools

Once the discriminating parameters are detected, the package aSHAPE provides tools for a more detailed study of data ensembles.

Tool-1 allows to select a subensemble *Ens*'(*Fm*) of *Ens*(*Fm*) satisfying either $CutBtm \le f$ or $f \le CutTop$ or $CutBtm \le f \le CutTop$, where the values of *CutTop* and *CutBtm* are assigned by the user.

Tool-2. Suppose that f discriminates families *Fm1* and *Fm2* and $h(Fm1) < h(Fm2)$. If non-overlapped right part of $h(Fm2)$ begins from $\boldsymbol{v}$, *Fm1* does not contain a fragment with $f > v$. Then natural question arises whether each modeled region of *Fm2* possesses a fragment with $f > v$. If the answer is 'yes' then one gains v as divider for *Fm2* and *Fm1*. In order to shed light on the situation, the tool for every modeled region of *Fm2* displays the number of fragments satisfying $f > v$.

Tool-3 executes *FreeWin* procedure. Suppose that for some metric parameter f the fragments of $h(Fm1) \cap h(Fm2)$ called "*hit-fragments*" satisfy $v1 \le f \le v2$. At the first step *FreeWin* searches every i-th modeled region in family *Fm1* for *windows* of size *FragSz*, which are free from *hit-fragments* and saves them in a library Lib_i, $i = 1,\ldots,RegNu$.

At the next step *FreeWin* operates with ensembles of the windows. Each ensemble appears as the combination of *RegNu* free windows taken a one from every Lib_i at a time. In total we have $Lib_1 \times Lib_2 \times \ldots$ window ensembles. The positions of the windows in the ensemble are scattered in a certain range. *FreeWin* finds those which have the minimum range.

Tool-4 executes *Alignment*. Consider a collection of the DNA fragments, which perform similar function in the genome, for instance promoters. If all of them are preliminary aligned in respect to the transcription start points, as in our case, we can expect similar conformation features in these regions, where it is important. The procedure *Alignment* allows one to verify this assumption.

Let us introduce B-window (a batch of windows) as the set of windows of equal size having the same position in every studied DNA region. If *i* enumerates the modeled regions of family *Fm* and *p* enumerates the fragment positions in the region, then we denote by *fip* the value of *f* for the *p-th* fragment in the *i-th* region. The user selects the size of B-window, *WinSz*, and for every B-window position $pWin \in [1, RegSz - WinSz + 1]$ *Alignment* computes the matrix $M(pWin) = \|fip\|$ of size $WinSz \times RegNu$, where *i* counts the regions by rows, and *p*, counting the positions of the current B-window refers to columns. As an output the program generates *WinSz* matrices each of which is subjected to the real *Alignement*, which considers all the combinations associated with matrices $M(pWin)$. Each combination consists of *fip* marked in every row once at a time. It yields the $WinSz^{RegNu}$ combinations. In these combinations the magnitudes of *fip* vary within some range and the task is to estimate the minimum range ρ, which is achieved among the combinations. In time proportional to $|M|log_2|M|$ *Alignment* finds the minimum ρ and saves it as $\rho(pWin)$. At the final stage the software displays the graph of $\rho(pWin)$.

Application of additional tools

Figures 6 and 9 show that the *homogeneity* measured for metric parameter *BendHomo* discriminates *Cont* and *Prm* so that $h(Cont) < h(Prm)$.

We are interested to know where the fragments of lower homogeneity are disposed within the promoter regions. Let us declare a fragment to be inhomogeneous when $BendHomo \geq 25°$(Fig. 9). The cutting procedure under condition $CutBtm < 25°$selects the ensemble *Ens* of the inhomogeneous fragments and computes their number within each modeled promoter region. Fig. 10 demonstrates that five promoters have two inhomogeneous fragments, one of promoters contains 126 such fragments, while fifteen promoters are free of inhomogeneous ones.

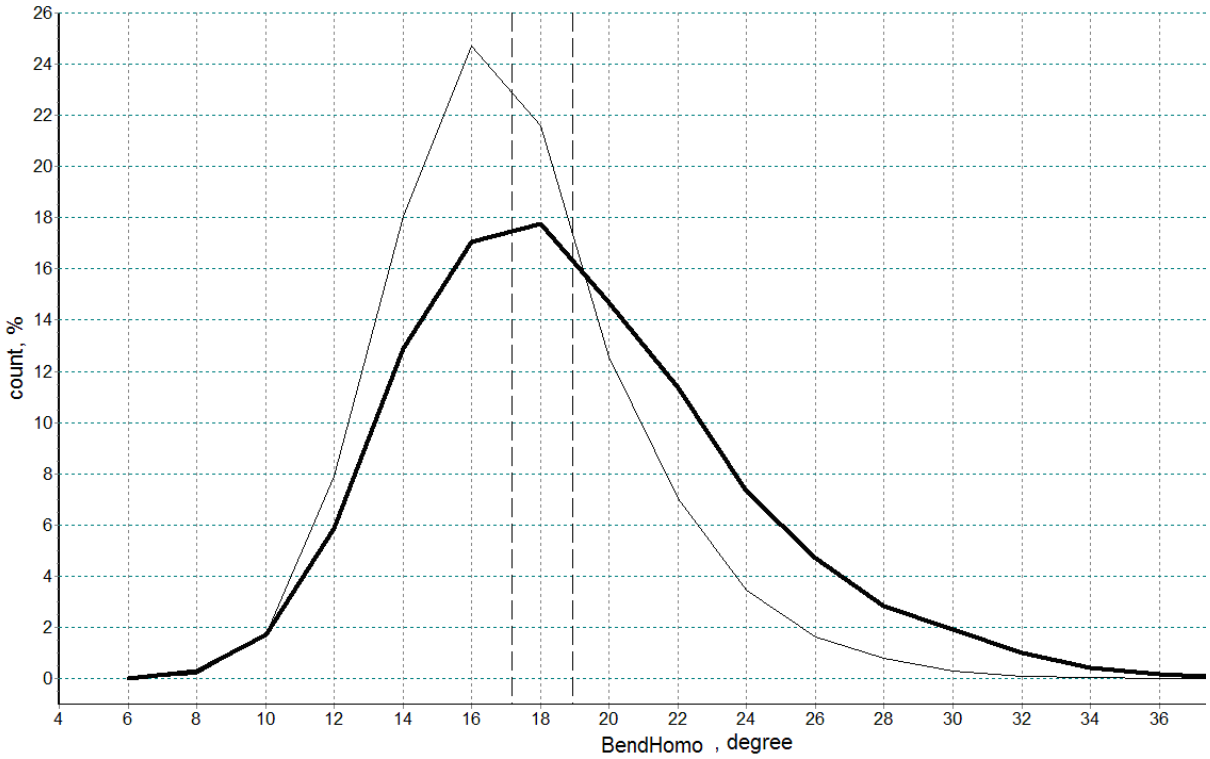

Fig. 9. Histograms $h(Prm)$ (thick line) and $h(Cont)$ (thin line) showing the distribution of $f = BendHomo$, obtained by the chains of phosphorous doublet centers. $FragSz$ = 20bp. Dashed lines indicate an average values. $Sq(Prm \cap Cont) = 0.8$.

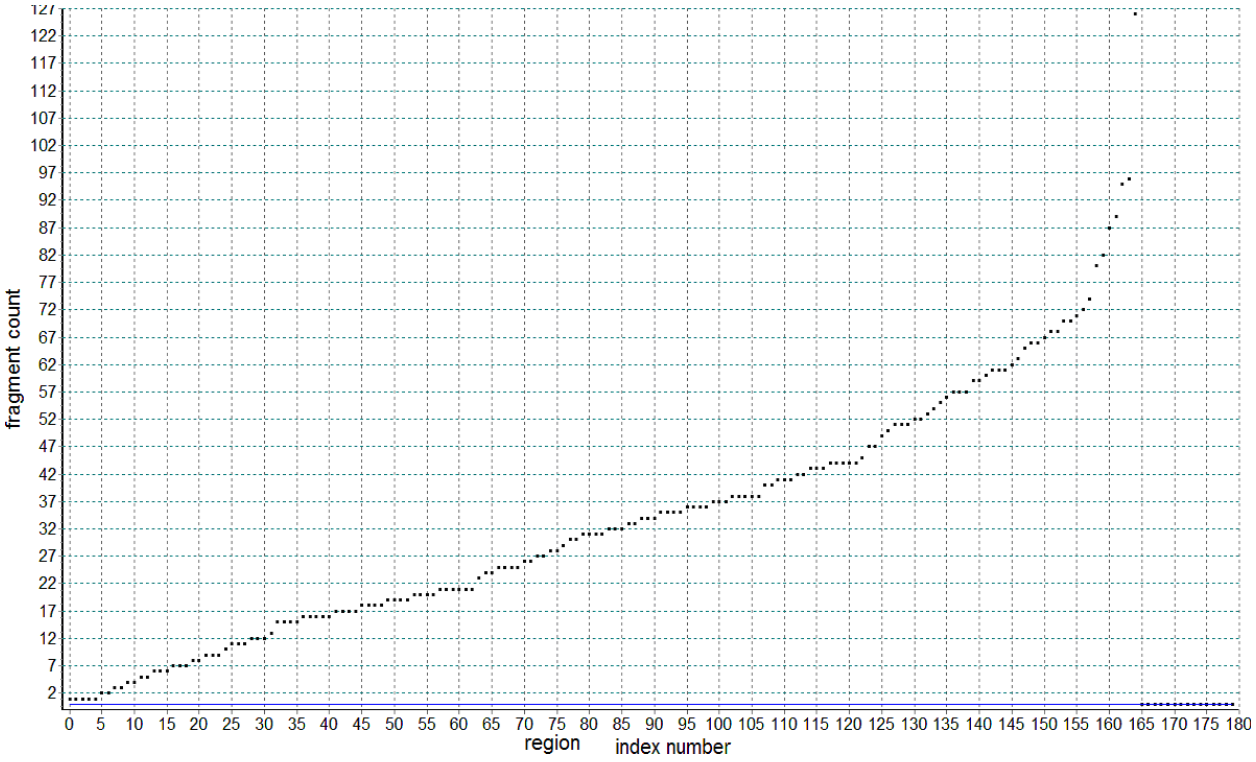

Fig. 10. The ranked plot showing the number of inhomogeneous fragments in every modeled promoter region. *FragSz* = 20bp. The chains of phosphorous doublet centers.

In order to localize homogeneous fragments we need to employ *FreeWin* procedure. Calculations show that every promoter region has at least two windows of length of 20bp free of inhomogeneous fragments, one in the range [–67, –17], another one - nearby the start point of transcription within [0, +50], the concrete position of free window depends on the region. Thus the start point in every promoter region is surrounded by homogeneous fragments.

The question about the location of fragments having similar shapes is even more important. In order to answer this question one has to choose a chain and a metric parameter reflecting the shape of modeled fragments and to use the procedure *Alignment*. Here, we demonstrate the working capacity of this option using the chain of carbon doublet centers and the metric parameter *MaxJut*. As an example, we chose *FragSz* = 40bp and the size of B-window *WinSz* equal to *FragSz*. Figure 11 shows the plotted *MaxJut* for 180 modeled regions of *Prm* family without alignment. Then we applied *Alignment* to this ensemble (Fig. 12).

We can ask the software to display only the ‘best’ position, the full resulting graph or the desired position (Fig. 12). We think that it is important to explain why we assign *WinSz* = *FragSz* for *Alignment*. Since *Alignment* looks for the combination of fragments, which has the lowest variance ρ of measured parameter, increasing the B-window size gives the better alignment due to the larger number of fragments in the wider B-window. On the other hand, we are interested in narrowing the location of specific fragments. Ordinary, the desired result is achieved by the gradual decrease of the of B-window size.

Fig. 12 testifies that each promoter of *Prm* family has a fragment in [–29, +11] with the shape deformation *MaxJut* in the range $4.72\text{Å} \le MaxJut \le 5.57\text{Å}$. I.e. *MaxJut* variance of aligned fragments is not greater than 0.85Å, whereas the scale of the whole ensemble is 17 Å (Figure 11).

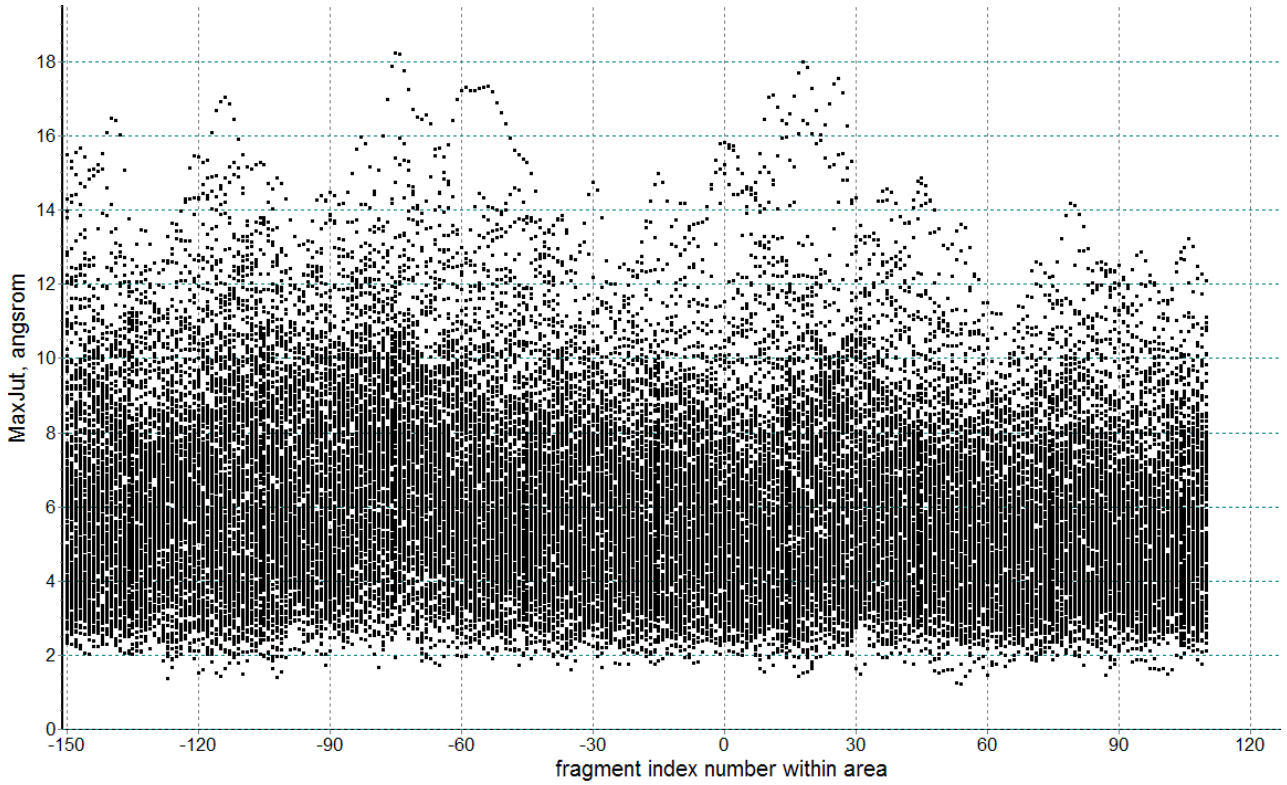

Fig. 11. The ensemble of *MaxJut* values for *Prm* family. The chains of carbon doublet centers has been used. *WinSz* = *FragSz* = 40bp.

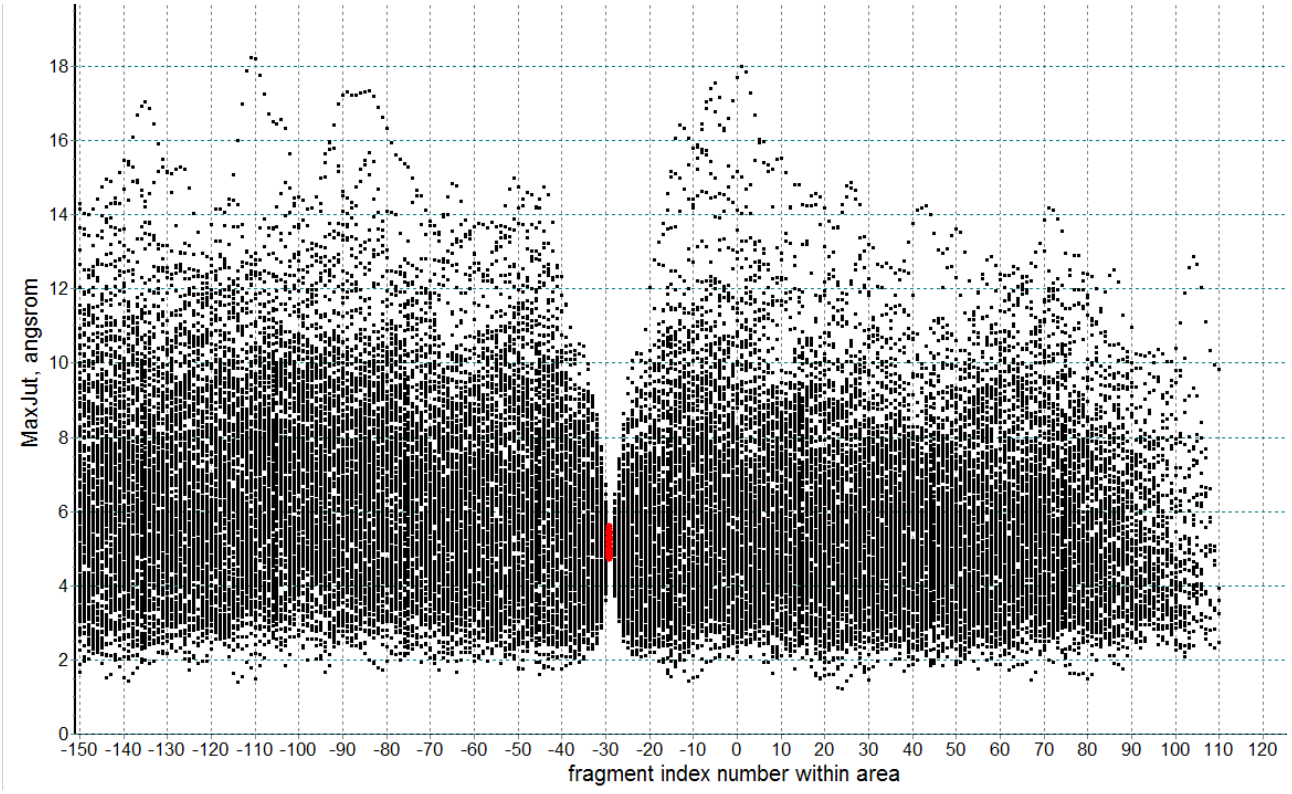

Fig. 12. The output data, exemplifying *Alignment* of *Prm* family in the position –29 for metric parameter *MaxJut*. The range of ρ is shown in red. The chains of carbon doublet centers has been used. *WinSz* = *FragSz* = 40bp.

CONCLUSION

Numerous attempts to construct a computer model of the promoter DNA have been focused so far on the linguistic analysis of nucleotide sequences (see [11] and the references

cited in this publication) or on the analysis of such physico-chemical features as thermodynamic instability [32], conformational flexibility [33] or electrostatic variations [34]. In this paper we propose another approach which exploits “DNA tools” server [14] for computing 3D models of interesting DNA regions, combines them into specific families of the modeled regions and implements comparative 3D analysis among the families with the software aSHAPE. During the aSHAPE debugging we studied the family *Prm* consisting of 3D modeled promoter regions of *E. coli*.

Two promoter regions free of fragments with an extraordinary large variation of the bend angle θ were found in this study. This may reflect structural constraints imposed by the transcription machinery on the conformation of promoter DNA. One of these regions encompasses the transcription start point, which is in direct contact with RNA polymerase and its environment plays an important role in the transition of the open transcription complex into the sate of productive initiation. Curiously, however, that the absence of strongly bent fragments was observed for the 50bp-long transcribed region, which is outside the border of direct contact with the enzyme (usually nearby the position +20). That assumes involvement of more extended transcribed region in the formation of transcription-competent complexes than currently anticipated. The second area without large variations of θ is located nearby the position -67. It is 50bp-long and, therefore, covers at least two important promoter modules: the element -35, recognized by the domain 2.4 of RNA polymerase σ-subunit, and UP-element, interacting with two α-subunits of the enzyme.

The most significant biological result of the study is a revealed ability to align promoters according to their structural peculiarity. The most efficient alignment was observed in the range -45 - -29 (exemplified for the position -29 in Fig. 12), which includes all three functional promoter modules as well as the spacer separating elements -35 and -10. In the contact region with the RNA polymerase all promoters, therefore, have a 40bp fragment with a similar deviation from the standard B-form DNA ($4.72Å \leq MaxJut \leq 5.57$ Å). Perhaps this structural peculiarity provides the primary signal, recognized by the enzyme and positions it near the specific modules.

Methodological feature of the package aSHAPE is its ability to explore the conformational properties of DNA using different conformation *chains*. The scope of the software is not limited by the metric parameters that are already formulated in the package. It is also important that the package can be used for the analysis of various functional modules of the genome, including the transcription terminators, transcription factor binding sites, splice signals, etc.

This work was partially supported by RFBR № 10-04-01218 and № 09-07-00455 grants.

REFERENCES

1. Coulombe B. and Zachary F. B. DNA Bending and Wrapping around RNA Polymerase: a "Revolutionary" Model Describing Transcriptional Mechanisms. *Microbiol. Mol. Biol. Rev*. 1999. V. 63. P. 457–478.
2. Carmona M., Claverie-Martin F., Magasanik B. DNA bending and the initiation of transcription at 54-dependent bacterial promoters. *PNAS*. 1997. V. 94. P. 9568–9572.
3. Bolshoy A., Nevo E. Ecologic Genomics of DNA: Upstream Bending in Prokaryotic Promoters. *Genome Res*., 2000. V. 10. P. 1185–1193.
4. Hirvonen C.A., Ross W., Wozniak C.E., Marasco E., Anthony J.R., Aiyar S.A., Newburn V.H., Richard L. Gourse Contributions of UP Elements and the Transcription Factor FIS to Expression from the Seven rrn P1 Promoters in *Escherichia coli*. *J. Bacteriol*. 2001. V. 183. P. 6305–6314.
5. Kozobay-Avraham L., Hosid S., Bolshoy A. Involvement of DNA curvature in intergenic regions of prokaryotes. *Nucleic Acids Res.* V. 34. P. 2316–2327.

6. Shultzaberger R.K., Chen Z., Lewis K.A., Schneider T.D. Anatomy of *Escherichia coli* σ_{70} promoters. *Nucleic Acids Res*. 2007. V. 35. P. 771–788.
7. Pul U., Lux B., Wurm R., Wagner R. Effect of upstream curvature and transcription factors H-NS and LRP on the efficiency of *Escherichia coli* rRNA promoters P1 and P2 – a phasing analysis. *Microbiology*. 2008. V. 154. P. 2546–2558.
8. Meysman P., Dang T.H., Laukens K., De Smet R., Wu Y., Marchal K., Engelen K. Use of structural DNA properties for the prediction of transcription-factor binding sites in *Escherichia coli*. *Nucleic Acids Res*. 2011. V. 39. e6.
9. Ozoline O.N., Deev A.A., Trifonov E.N. DNA bendability – a novel feature in *E.coli* promoter recognition. *J. Biomol. Struct. Dynamics*. 1999. V. 16. P. 825–831.
10. Ozoline O.N., Masulis I.S., Buckin V.A. Deformable elements in promoter DNA as a basis for adaptive conformational transitions. *J. Biomol. Struct. Dynamics*. 2001. V. 18. № 6. P. 1002–1003.
11. Shavkunov K.S., Masulis I.S., Tutukina M.N., Deev A.A., Ozoline O.N. Gains and unexpected lessons from genome-scale promoter mapping. *Nucleic Acids Res*. 2009. V. 37. P. 4919–4931.
12. Shatzky-Schwartz M., Shakked Z., Luisi B.F. X-ray and solution studies of DNA oligomers and implications for the structural basis of A-tract-dependent curvature. *J. Mol. Biol*. 1997. V. 267. P. 595–623.
13. McAteer K., Aceves-Gaona A., Michalczyk R., Buchko G.W., Isern N.G., Silks L.A., Miller J.H., Kennedy M.A. Compensating bends in a 16-base-pair DNA oligomer containing a T(3)A(3) segment: A NMR study of global DNA curvature. *Biopolymers*. 2004. V. 75. P. 497–511.
14. Vlahovicek K, Kajan L, Pongor S. DNA analysis servers: plot.it, bend.it, model.it and IS. *Nucleic Acids Res*. 2003. V. 31. № 13. P. 3686–3687. URL: http://hydra.icgeb.trieste.it/dna/index.php (accessed 1 August 2011).
15. *Factor, discriminant and cluster analysis.* Ed. Enyukov I.S. Moscow, 1989. Translation of: Kim J.-O., Mueller C.U., Klecka C. Factor, Discriminant, and Cluster Analysis. (in Russ.).
16. StatSoft. *Electronic Statistics Textbook*. URL: http://www.statsoft.com/textbook/discriminant-function-analysis/ (accessed 1 August 2011).
17. Fedoseeva V.B. DNA: double helix bends: structure and functions. URL: http://medbiol.ru/medbiol/dna_bend/00000a6c.htm#00013419.htm (accessed 1 August 2011).
18. Zheng G, Xiang-Jun Lu, Wilma K. Olson W.K. Web 3DNA – a web server for the analysis, reconstruction, and visualization of threedimensional nucleic-acid structures. *Nucleic Acids Research*. 2009. V. 37. W240–W246.
19. Strahs D., Schlick T. Analysis of A-tract bending: Insights into experimental structues by molecular dynamics simulations. 2000. *J. Mol. Biol.* V. 301. P. 643–663. URL: http://www.biomath.nyu.edu/index/software/Madbend/index.html (accessed 1 August 2011).
20. Xiang-Jun Lu, Shakked Z., Olson W. K. A-form Conformational Motifs in Ligand-bound DNA Structures. *J. Mol. Biol.* 2000. V. 300. P. 819–840.
21. Dickerson R.E. DNA bending: the prevalence of kinkiness and the virtues of normality *Nucleic Acids Res.* 1998. V. 26. P. 1906–1926.
22. Goodsell D.S. and Dickerson R.E. Bending and curvature calculations in B-DNA. *Nucleic Acids Res.* 1994. V. 22. № 24. P. 5497–5503.
23. Shpigelman E. S., Trifonov E. N. and Bolshoy A. CURVATURE: software for the analysis of curved DNA. *Comput. Appl. Biosci*. 1993. V. 9. P. 435–440.
24. Lavery R., Sklenar H. The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dyn*. 1988. V. 6. P. 63–91.

25. Lee S., Park K., Kang C. Z-curve: a computer program calculating DNA helical axis coordinates for three-dimensional graphic presentation of curvature. *Molecules and Cells*. 1999. V. 9. № 4. P. 350–357.
26. Barbic A., Crothers DM. Comparison of analyses of DNA curvature. *J. Biomol. Struct. Dyn*. 2003. V. 21. № 1. P. 89-97.
27. Herisson J., Payen G., Gherbi R. A 3D pattern matching algorithm for DNA sequence. *Bioinformatics.* 2007. V. 23. № 6. P. 680–686.
28. Olson W.K., Bansal M., Burley S.K., Dickerson R.E., Gerstein M., Harvey S.C., Heinemann U., Lu X.-J., Neidle S., Shakked Z., Sklenar H., Suzuki M., Tung C.-S., Westhof E., Wolberger C., Berman H.M. A Standard Reference Frame for the Description of Nucleic Acid Base-pair Geometry. *J. Mol. Biol*. 2001. V. 313. P. 229–237.
29. Suzuki M., Amano N., Kakinuma J., Tateno M. Use of a 3D Structure Data Base for Understanding Sequence-dependent Conformational Aspects of DNA. *J. Mol. Biol*. 1997. V. 274. P. 421–435.
30. Olson W.K. *Nucleic acid structural principles*. URL: http://128.6.69.24/lnotes/BioPhysChem_week5.pdf. (accessed 1 August 2011).
31. *Handbook of applicable mathematics*. Volume VI: Statistics. Part B. John Willey&Sons Ltd, 1984.
32. Kanhere A., Bansal M. A novel method for prokaryotic promoter prediction based on DNA stability. *BMC Bioinformatics*. 2005. V. 6. № 1.
33. Wang H., Benham C.J. Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress. *BMC Bioinformatics*. 2006. V. 7. № 248.
34. Sorokin A.A., Osypov A.A., Dzhelyadin T.R., Beskaravainy P.M., Kamzolova S.G. Electrostatic properties of promoter recognized by *E. coli* RNA polymerase Esigma70. *J Bioinform Comput Biol*. 2006. V. 4. № 2. P. 455–467.

Received August 07, 2012.
Published August 31, 2012.