Письма в ЭЧАЯ. 2005. Т. 2, № 4(127). С. 82–90

УДК 51-7: 577.21

ON A CLASSIFICATION OF *E. coli* PROMOTERS ACCORDING TO THEIR ELECTROSTATIC POTENTIALS

R. V. Polozov [a], *V. S. Sivozhelezov* [b], *V. V. Ivanov* [c,d], *Yu. B. Melnikov* [e,f]

[a] Institute for Theoretical and Experimental Biophysics, Pushchino, Russia
[b] Institute of Cell Biophysics, Pushchino, Russia
[c] Joint Institute for Nuclear Research, Dubna
[d] Chaos and Innovation Research Unit, Aristoteles University of Thessaloniki, Thessaloniki, Greece
[e] International Solvay Institutes for Physics and Chemistry, Brussels, Belgium
[f] Institute Superiéure de Technologie, Luxembourg

Classification of promoters and other functionally important genome fragments according to their nucleotide sequences and physicochemical properties is a key factor for understanding gene transcription, replication, recombination and their regulation. The classification of genome promoters is usually performed on the basis of analysis of their primary structures. However, such an approach does not allow one to obtain a simple answer because it is the physicochemical properties of DNA that control the process of gene transcription and its regulation. Electrostatic interactions comprise an essential component of those processes. This work presents the approach that allows computation of electrostatic potentials of long nucleotide sequences of DNA for both procaryotic and eucaryotic species. The elecrtostatic potentials of *E. coli* promoters and periodic sequences were calculated. We suppose that the electrostatic characteristics of the genome promoters together with primary structure provide their reliable classification.

Классификация промоторов и других функционально важных элементов генома по их нуклеотидным последовательностям и физико-химическим свойствам является ключевым фактором для понимания процессов транскрипции генов, редупликации, рекомбинации и их регуляции. Обычно классификация промоторов генома проводится на основе анализа их первичных структур. Однако такой подход не позволяет получить однозначного ответа, т. к. за процесс транскрипции и ее регуляцию в основном ответственны физико-химические свойства ДНК. Важную роль в указанных процессах играют электростатические взаимодействия. В настоящей работе развит подход, позволяющий вычислять электростатические потенциалы длинных нуклеотидных последовательностей ДНК как для прокариот, так и для эукариот. Нами вычислены электростатические потенциалы промоторов *E. coli* и периодических последовательностей. Мы полагаем, что электростатические характеристики промоторов генома совместно с первичной структурой обеспечат их надежную классификацию.

INTRODUCTION

One of the most important examples of molecular recognition is interaction of DNA with polymerases and other proteins that play key roles in transcription and its regulation where selective binding of protein to the particular DNA sequence occurs [1]. Specificity of binding can be evaluated in terms of energy, by the difference in free energies for binding the same protein to the specific and average nonspecific DNA site. This value varies from about 40

to over 80 kJ/mole [2], which is quite a large difference considering that only noncovalent forces are involved in protein–DNA binding.

Such a wide range of specificity led to formulating a model of protein–DNA recognition process involving at least three steps [3]. The first is nonspecific binding of a protein to DNA which is energetically driven by the electrostatic complementarity of the DNA and protein contacting surfaces [4]. The second step is one-dimensional diffusion of a protein along DNA chain, which accelerates association rates beyond their three-dimensional diffusion limits [5–7]. During this step, electrostatic interactions of proteins with DNA retain the protein in the immediate vicinity of DNA, thus providing the required reduction in dimension from three to one. The third step is formation of more extensive contacts between DNAs, which occurs when a protein locates its target site. Again, the specific interaction of a protein with its target DNA sequence involves electrostatic interactions [3], but other factors also contribute, e.g., the mutual surface fitting due to the DNA-induced protein refolding [8,9] and protein-induced conformation changes of DNA [10, 11].

Thus, electrostatic interactions are of primary importance in the multistep process of the protein–DNA recognition. In the first step of that process which occurs approximately at the electrophoretic sliding surface of DNA, which is about 15 Å away from the DNA longitudinal axis [12], the electrostatic interaction is the only physical factor since Coulomb electrostatic forces decay with distance much lower than other forces like hydrogen bonding, London forces, etc.

Even more importantly, calculation of electrostatic potential distribution along DNA for long chains will open the road to analyzing correlations of DNA functional properties with physical properties of the DNA sequence, particularly, the electrostatic properties. Earlier, correlations were established between the properties and the sequences themselves, and classification of DNA sequences was performed using the well known cluster analysis technique. Such a classification allows one to elucidate structure-function relationships [13]. The drawback of such a classification is that it has no explicit physical basis. In contrast, correlations of electrostatic properties with functions will allow one to establish such a basis. Besides, DNA electrostatic properties are already known to correlate with its sequence, but that was earlier established for short DNA chains only.

On the other hand, the sequences of coding and promoter regions of DNA correlate between DNAs of various biological species, which allows one to identify evolutionary relationships, again via classification by cluster analysis [14]. Once correlations are characterized between electrostatic properties of those regions, the corresponding evolutionary relationships will acquire the physical basis.

Since distributions of electrostatic potentials or fields have distinct geometrical shapes, the classification can be inferred via morphology methods (Procrustean, Minkowski, or other metrics). The most accurate calculation method of electrostatic potentials and energies available for macromolecular systems is numerically solving the Poisson–Boltzmann equation on a rectangular grid [15]. But this method was not used for long DNA sequences recognized by some DNA-binding proteins, because the number of grid points N scales linearly with the DNA length, and computation time typically scales, at best, as N.

In this work we adopt a multigrid method of solving the Poisson–Boltzmann equation in which computation time typically scales as $\ln N$, which allows us to handle several hundreds of base pairs long DNA sequences, exemplified in this study by *E. coli* promoter regions, which are 411 bp long.

1. SOLVING THE POISSON–BOLTZMANN EQUATION

Calculations of electrostatic potential φ of DNA fragments were performed by solving the Poisson–Boltzmann equation, which describes the electrostatic potential in solvent around DNA molecule according to

$$-\bigtriangledown(\epsilon(\mathbf{r})\bigtriangledown\varphi(\mathbf{r})) = 4\pi(\rho_0(\mathbf{r}) + \rho_1(\varphi(\mathbf{r}))), \tag{1}$$

where $\mathbf{r} = (x, y, z) \in R^3$; φ is the sought electrostatic potential; ϵ is the dielectric permeability and ρ_0 is the charge distribution of DNA described by

$$\rho_0(\mathbf{r}) = \sum_i ez_i\delta(|\mathbf{r} - \mathbf{r}_i|), \tag{2}$$

where z_i is the charge of the ith atom of the molecule in units of elementary charge; $\mathbf{r}_i$ is the radius vector of the ith atom; e is the elementary charge (the absolute value of the electron charge); δ is the Dirac delta function, and

$$\rho_1(\mathbf{r}) = \sum_i n_i ez_i \exp(ez_i\varphi/k_BT). \tag{3}$$

When the potential is small enough ($\varphi \ll k_BT/e$), Eq. (1) reduces to its linearized form

$$-\bigtriangledown(\epsilon(\mathbf{r})\bigtriangledown\varphi(\mathbf{r})) + \kappa^2\varphi = 4\pi\rho_0(\mathbf{r}), \tag{4}$$

where

$$\kappa^2 = 4\pi e^2 \sum_i n_i z_i^2/k_BT \tag{5}$$

is the ion density, where n_i is the concentration of ions of the ith kind; z_i is the charge of ion of ith kind in units of the elementary charge; k_B is the Boltzmann constant; T is the absolute temperature assumed to be 300 K.

Boundary condition for the potential $\varphi(\infty)$ is set using the Debye–Huckel approximation. For the purpose of numerical solution we restrict the infinite region to a big parallelepiped Γ, with the condition imposed on its surface, for $\mathbf{r} = (x, y, z) \in \Gamma$:

$$\varphi(\mathbf{r})|_\Gamma = \sum_i \frac{ez_i \exp(-\kappa|\mathbf{r} - \mathbf{r}_i|)}{|\mathbf{r} - \mathbf{r}_i|}. \tag{6}$$

The problems (4) with boundary condition (2) in the region Γ are solved using the finite element method. We solve the discrete linear system iteratively by the multigrid method (MM). The details of the MM solution algorithm are presented in Appendix.

In order to solve the nonlinear equation (1), we apply the iterations according to

$$-\bigtriangledown(\epsilon(\mathbf{r})\bigtriangledown\varphi^{n+1}) + \alpha\varphi^{n+1} = 4\pi(\rho_0 + \rho_1(\varphi)) + \alpha\varphi^n, \tag{7}$$

where φ^n is the approximation of solution corresponding to the nth iteration. Thus, the linear problem with unknown φ^{n+1} has to be solved on each iteration. The solution to the problem (4) is used as an initial approximation for iterations (7).

2. CALCULATIONS OF ELECTROSTATIC POTENTIALS

All atom models of DNA fragments were constructed using the evaluation version of the HyperChem 7.01 package [16]. DNA was assumed to be in the B form. Charges were assigned to the center of each atom. The values of charges were taken from the AMBER force field [17]. Additional charges of $0.25\,e$ were assigned to O_1 and O_2 atoms of phosphate groups to allow for the well known counter-ion condensation effect, which is retention of part of counter ions near the charged atoms of the phosphate groups. Dielectric constants were taken to be 2 for the DNA interior and 80 elsewhere. Potential was visualized as a topological map on the surface of a cylinder with 15 Å radius centered at the longitudinal axis of DNA, about 5 Å away from DNA sugar-phosphate backbone. Such a surface approximates the electrophoretic sliding surface of the DNA, at which the first stage of DNA–protein recognition is believed to occur.

The horizontal axis in the map shown coincides with the DNA helix axis. The color scale represents the electrostatic potential in units of k_BT/e, which is thermal motion energy k_BT per unit of electric charge e. In those units, red color was chosen to correspond to –1.3, blue to –0.8, and white to intermediate values. In this color scheme, the visualized electrostatic potential values will span a range of 0.5 k_BT/e, so that ten unit charges, which are typically present in protein fragments interacting with DNA, will account for a difference of 5 k_BT/e, which is quite sufficient for the electrostatic steering that happens as the protein approaches the DNA surface. Ion concentrations (1:1 electrolyte was assumed) were $0.15\,M$, which is the physiological value.

DNA sequences of *E. coli* promoter regions were taken from [18] and [19]. The start point of transcription is located at the position 257, so the coding sequence starts further downstream, and the promoter region is upstream from that point.

Figure 1 presents the electrostatic potentials of periodic DNA: poly(A), poly(AT), poly(G), and poly(GC). As one can see from Fig. 1, this electrostatic potential is also periodic in nature. The fact that the periodicity does not appear perfect on the cylindrical surface is explained by the geometry of B form of DNA. One can also see that the potential of poly(AT) sequence is drastically different from the rest of periodic sequences. Particularly, the spots of both the blue (less negative) and red (more negative) potential are smaller and much less intense, indicating that the potential of the poly(AT) DNA sequence deviates from its average value much less than for other periodic sequences. Also, the alternating blue and red bands appear more frequently the electrostatic potential of the poly(AT) sequence indicating that the potential of poly(AT) is of finer structure than for other periodic DNA sequences. Those distinguishing features of the electrostatic potential of poly(AT) show that the electrostatic potential should strongly correlate with the presence of long (AT) runs in the sequence.

This feature itself shows that the calculations of the DNA electrostatic potential can contribute to the classification of DNA sequences in a manner similar to the analysis of the sequence itself, leading to expansion of the entire field of bioinformatics. Particularly, instead of building classifications based on the sequence alone, at least one physical property can be allowed for in building classifications, namely, the electrostatic potential.

Figure 2 shows the electrostatic potential of several promoter regions of *E. coli*, together with the adjacent coding regions. Qualitatively, the electrostatic potentials of these regions noticeably differ from the potentials of periodic sequences. The main difference is apparent presence of a strong dipolar component in the electrostatic potential across the DNA double

helix. Indeed, the intense blue spots (less negative potential) are located well away from the intense red spots (more negative potential). In contrast, the periodic DNA sequences (Fig. 1) exhibit a more homogeneous distribution of the electrostatic potential across the double helix, visually more similar to the quadrupolar distribution.

Of the six promoters shown, two top promoters, uvrA and uvrB-P1, show the maximal anisotropy of the electrostatic potential. Indeed, red and blue spots are larger and more intense than for the remaining four promoters. Two bottom promoters, accA and accBC, show the least anisotropy, and the middle two, uvrD-P1 and uvrD-P2, are intermediate in that respect. For all the six promoters, the direction of the dipole moment varies in a sequence-dependent manner.

The data obtained suggest that, first, the promoter and coding regions have electrostatic potential greatly differing from that of periodic sequences. Secondly, the electrostatic potential differs with the type of promoters, mostly in the asymmetry of the distribution of positive and negative patches of the electrostatic potential. Finally, both the amplitude and the direction of the dipole moment across the DNA double helix change along the helix axis.

To show the finer structure in functionally important promoter areas (–35, –10, and starting point), the electrostatic potential distribution in those areas are presented for two promoters, accA and uvrA (Fig. 3), scaled to include those areas only. In the –35 area, a quasi-periodic potential distribution appears, in which red and blue spots are alternating. No explicit anisotropy of potential is observed in that region. In contrast, large areas of red and blue appear in the area from –10 to the starting point on the opposite sides of the cylinder, which suggests the anisotropy of the electrostatic potential in the vicinity of the starting point.

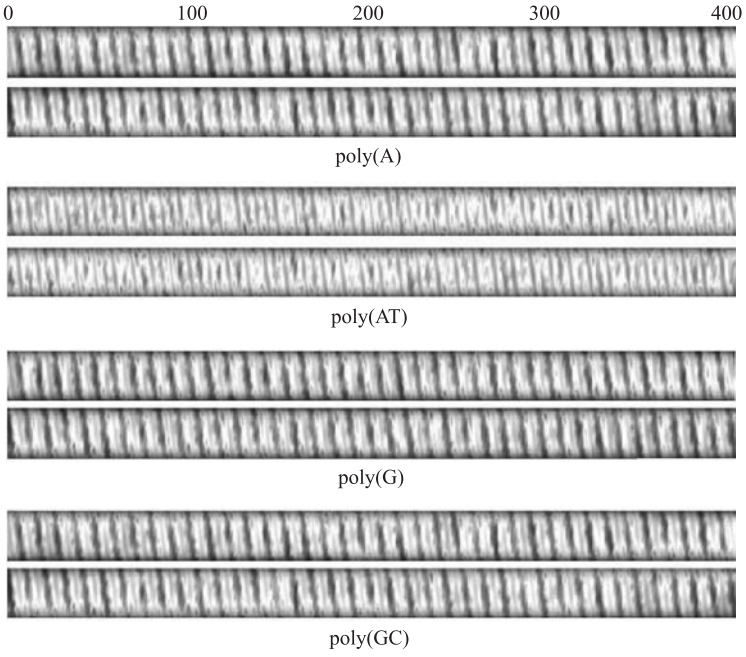

Fig. 1. Distribution of electrostatic potential around periodic DNA molecules. Each molecule is shown in two views differing by 180° rotation around the helix axis

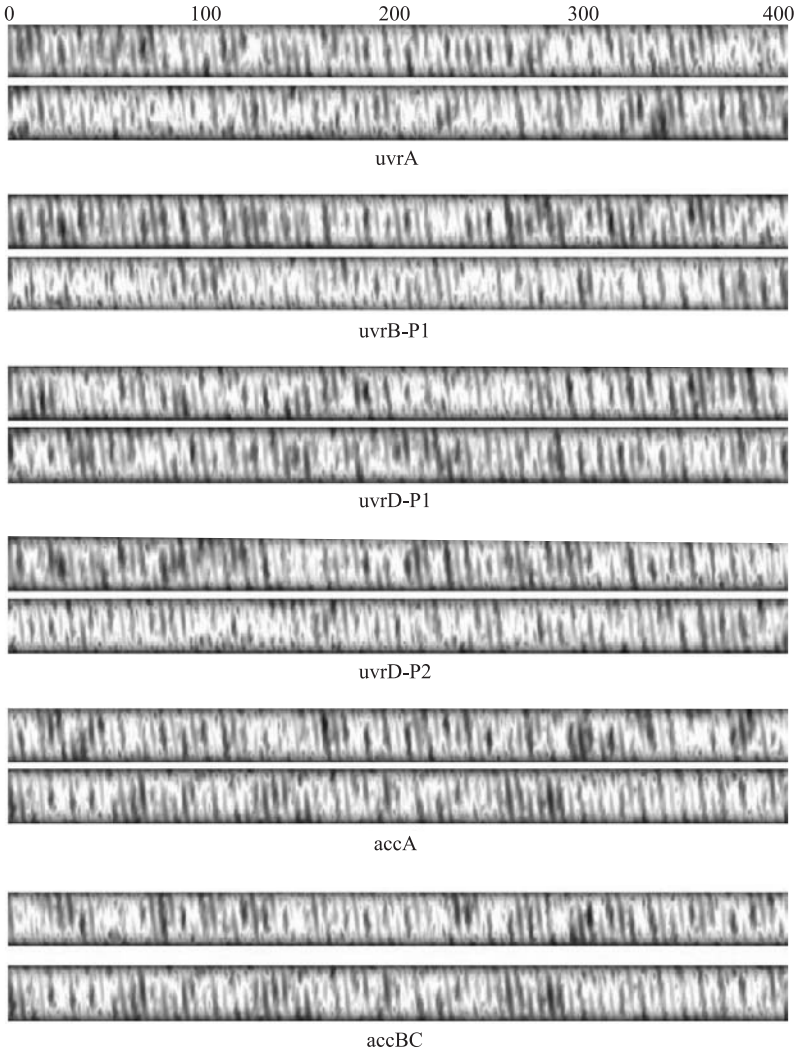

Fig. 2. Distribution of electrostatic potential around promoter DNAs of *E. coli*. Each promoter is shown in two views differing by 180° rotation around the helix axis

The A/T tracks are known to occur more frequently in promoter sequences than in the full genome sequences, and to be distributed nonrandomly in promoter sequences. In Fig. 4 the distribution of the electrostatic potential is presented around the periodic DNA sequences poly(A) and poly(AT) in a fragment equal in length to fragments of promoter DNA sequences of *E. coli* from the –35 area to the transcription start point. One can see that both the promoters accA and uvrA have the electrostatic potential distributions visually more similar to that of poly(AT) sequence than to that of poly(A). Thus, the electrostatic potential distribution of promoters appears to correlate with the contents and positions of A/T tracks along the promoter.

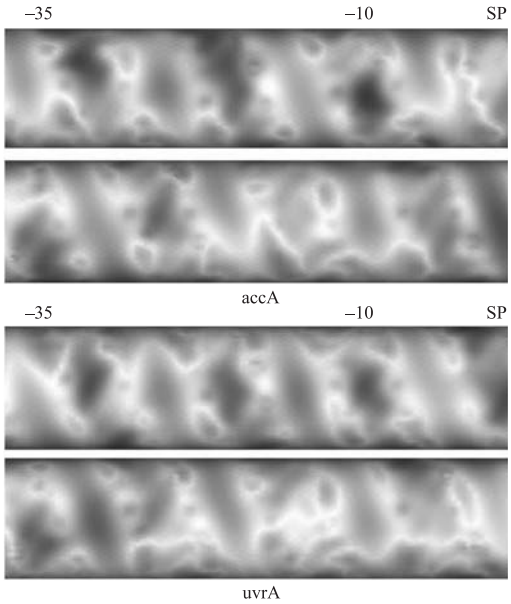

Fig. 3. Distribution of electrostatic potential around accA and uvrA promoter DNAs of *E. coli* from the –35 point to the transcription start point (denoted by SP) shown in two views differing by 180° rotation around the helix axis. The picture is scaled to show finer structure of the specified areas

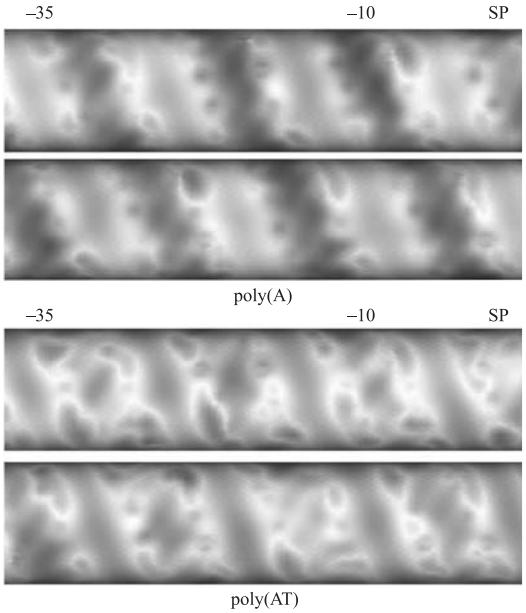

Fig. 4. Distribution of electrostatic potential around periodic DNA molecules poly(A) and poly(AT) in a fragment equal in length to fragments of promoter DNAs of *E. coli* from the –35 area to the transcription start point shown in two views differing by 180° rotation around the helix axis

CONCLUSIONS

Taken together, the data indicate that the electrostatic potential can improve the classification of DNA sequences by providing the physical basis. In such an improved classification, the physical basis will be rendered for drawing structure–function relationships and evolutionary relationships between various DNA sequences, thus contributing to the development of bioinformatics.

Therefore, the entire body of data including the primary structure (sequence), secondary structure (geometry), and physical properties of specific DNA sequences will provide a unified basis for promoter classification, which is a key feature in understanding promoter functioning (transcription), their evolution and regulation.

APPENDIX

To solve the nonlinear equation (1), we apply the iterations by formula (7) where φ^n is the solution approximation corresponding to the nth iteration.

To solve the problems (1) and (2) with the boundary condition (6), we discretize the region Γ with finite elements. The solution approximation is found in the finite dimensional space S with the basis of finite element functions $\Phi^i(x, y, z)$:

$$\varphi = \sum_i u^i \Phi^i. \tag{8}$$

Applying the Galerkin approach [20] to (1), (7) and (8), we obtain a linear algebraic system of equations $Au = f$ from which coefficients u_i can be found. Then we solve this system of linear algebraic equations iteratively by the multigrid method (MM). The MM uses the sequence of nested finite element grids as follows:

$$h_{l-1} = 2h_l; \quad S_{l-1} \subset S_l; \quad l = 1, \ldots, L, \tag{9}$$

where h is the grid step size; S is the corresponding space of basis functions. The final solution should be found on the finest grid number L. It is performed by iterations using a set of auxiliary grids $l = 0, 1, \ldots, L-1$. On each iteration the problem is reduced to a smaller one on the grid $L - l$ for which the same algorithm is applied recursively until the grid number 0 is achieved. The grid 0 is the coarsest (with the biggest step size h_0). It contains a small number of unknowns. Thus, the linear system for that grid can be easily solved by any direct method, for instance, by Gauss elimination.

The algorithm of one MM iteration for the grid number l is presented in Fig. 5.

```
procedure mgm(l,u_l,f_l)
array u_l,f_l;
if l = 0 then u_0 = A^{-l} f_0 else
begin integer j; array d,v;
u_l := relax(A_l, u_l,f_l);
d := I_l^{l-l}(f_l – A_l u_l); v = 0;
for j :=l until γ do mgm(l– 1; u, d)
u_l := u_l + I_l^{l-l} v;
u_l := relax(A_l, u_l,f_l);
end
```

Fig. 5. Algorithm of one MM iteration for the grid number l

Here the quantities with subscript l are related to grid number l; the notation $u_0 = A^{-1} f_0$ means the direct solution procedure on the grid $l = 0$; I is the interpolation operator that

transfers functions from one grid to another; the *relax* is a simple iteration of Gauss–Seidel type to damp high frequency residual components.

The iteration process finishes when the residual norm satisfies the criteria

$$\frac{\| f_L - A_L u_L \|}{\| f_L \|} \leqslant 10^{-6}. \tag{10}$$

It usually takes 4–5 iterations to converge the iteration process.

Acknowledgements. We are grateful to Prof. I. Antoniou and Prof. V. G. Kadyshevsky for encouragement and support. This work has been partly supported by the Luxembourg Ministry of Culture, Higher Education and Research under Grant BFR01/070.

REFERENCES

1. *Kornberg R. D.* // Trends Cell Biol. 1999. V. 9. P. M46–M49.
2. *Coleman R. A., Pugh B. F.* // J. Biol. Chem. 1995. V. 270. P. 13850–13859.
3. *Ohlendorf D. H., Matthew J. B.* // Adv. Biophys. 1985. V. 20. P. 137–151.
4. *Fogolari F. et al.* // J. Mol. Biol. 1997. V. 267. P. 368–381.
5. *Hsieh M., Brenowitz M.* // J. Biol. Chem. 1997. V. 272. P. 22092–22096.
6. *Jeltsch A. et al.* // EMBO J. 1996. V. 15. P. 5104–5111.
7. *Berkhout B., van Wamel J.* // J. Biol. Chem. 1996. V. 271. P. 1837–1840.
8. *Votavova H. et al.* // J. Biomol. Struct. Dyn. 1997. V. 15. P. 587–596.
9. *Carra J. H., Privalov P. L.* // Biochemistry. 1997. V. 36. P. 526–535.
10. *Guzikevich-Guerstein G., Shakked Z.* // Nature Struct. Biol. 1996. V. 3. P. 32–37.
11. *Strauss-Soukup J. K., Maher L. J., 3rd* // Biochemistry. 1998. V. 37. P. 1060–1066.
12. *Schellman J. A.* // Biopolymers. 1977. V. 16. P. 1415–1434.
13. *Parsons J. D.* // Comp. Appl. Biosci. 1995. V. 11. P. 603–613.
14. *Guan X., Du L.* // Bioinformatics. 1998. V. 14. P. 783–788.
15. *Polozov R. V. et al.* // J. Biomol. Struct. Dyn. 1999. V. 16. P. 1135–1143.
16. http://www.hyper.com/products/description/hyper7.htm
17. http://sigyn.compbio.ucsf.edu/amber/
18. *Ozoline O. N. et al.* // Mol. Biol. 2002. V. 36. P. 682–688.
19. *Chasov V. V. et al.* // Biofizika. 2002. V. 47. P. 809–819.
20. *Fedorenko R.* // USSR Comp. Math. and Math. Phys. 1962. V. 1. P. 1092–1096.

Received on April 19, 2004.