E11-2013-137

T. P. Akishina, V. V. Ivanov, V. A. Stepanenko

MASSIVE CALCULATIONS OF ELECTROSTATIC POTENTIALS AND STRUCTURE MAPS OF BIOPOLYMERS IN A DISTRIBUTED COMPUTING ENVIRONMENT

Акишина Т. П., Иванов В. В., Степаненко В. А. E11-2013-137
Массовые вычисления электростатических потенциалов
и карт структур биополимеров в распределенной компьютерной среде

К ключевым факторам, определяющим процессы транскрипции и трансляции, относятся распределения электростатических потенциалов ДНК, РНК и белков. Расчеты электростатических распределений и карт структур биополимеров на ЭВМ являются времяемкими и требуют больших вычислительных ресурсов. Нами разработаны процедуры для организации массовых расчетов электростатических потенциалов и карт структур биополимеров в среде распределенных вычислений, содержащей несколько тысяч процессорных ядер.

Работа выполнена в Лаборатории информационных технологий ОИЯИ.

Сообщение Объединенного института ядерных исследований. Дубна, 2013

Akishina T. P., Ivanov V. V., Stepanenko V. A. E11-2013-137
Massive Calculations of Electrostatic Potentials
and Structure Maps of Biopolymers in a Distributed Computing Environment

Among the key factors determining the processes of transcription and translation are the distributions of the electrostatic potentials of DNA, RNA and proteins. Calculations of electrostatic distributions and structure maps of biopolymers on computers are time consuming and require large computational resources. We developed the procedures for organization of massive calculations of electrostatic potentials and structure maps for biopolymers in a distributed computing environment (several thousands of cores).

The investigation has been performed at the Laboratory of Information Technologies, JINR.

Communication of the Joint Institute for Nuclear Research. Dubna, 2013

1. INTRODUCTION

Origin, evolution, functions and regulation of promoter DNA are presently analyzed basing on their sequence alone. This analysis is unsufficient since it is the physicochemical properties of DNA that control the process of gene trascription and its regulation. Electrostatic interactions comprise an essential component of these processes. In this work, a computational approach is developed allowing one to calculate electrostatic potentials of long DNA sequences and transcription factors. This approach allows one to calculate electrostatic potentials of promoter sequences and transcription factors for both prokaryotic and eukaryotic species.

The purpose of the procedure is to provide a capability for automatic calculation of electrostatic potentials around biopolymers using multigrid method by solving Poisson–Boltzmann equation. The idea is to leverage automated approach, in order to be able to calculate a big number of promoters at a time in seven steps rather than performing 28 steps manually for each promoter. Calculation utilities for each promoter were already available, but the time consumption required to calculate a single promoter was unacceptable, moreover, due to the human factor involvement type mistakes were unavoidable and troubleshooting of such issues added up to the overall time consumption. This procedure permits one to leverage automated approach via consuming a great number of promoters and transcription factors (several thousands biopolymer structures at a time).

2. PROCEDURE OF ELECTROSTATIC CALCULATIONS

Thus, electrostatic interactions are of primary importantce in the multistep process of protein-DNA recognition. In the first step of this process, which occurs approximately at the electrostatic sliding surface of DNA, which is about 15 Å away from the DNA longitudional axis [1], the electrostatic interaction is the only physical factor since Coulomb electrostatic forces decay with distance much lower than other forces like hydrogen binding, London dispersion, etc.

Even more important, calculation of electrostatic potential distribution 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 sequencies was performed using the well-known cluster analysis technique. Such a classification allows one to elucidate structure-function relationships [2]. The drawback of such a classification is that it has no explicit physical basis. Besides, DNA electrostatic properties are already known to correlate with its sequences, but that was earlier established for short DNA chains only.

Electrostatic calculations of biopolymers were performed by solving the nonlinear Poisson–Boltzmann equation that relates the electric potential with the

charge distribution, protein partial charges taken from the AMBER force field [3], mobile solution charged approximated by Boltzmann distribution, with the dielectric constant assumed to be 2 inside the protein and 80 outside the protein. The electrolyte was assumed to be 1:1 ($z_1 = 1$, $z_2 = 1$) at physiological or subphysiological 50150 mM concentrations. Solution is sought with finite difference multigrid method using a sequence of nested finite difference grids, the finest grid having up to $200 \times 200 \times 200$ points so that the interval between grid points is less than 1 Å. We have developed an algorithm of solving the nonlinear Poisson–Boltzmann equation allowing one to efficiently calculate electrostatic potentials for large objects such as proteins, nucleic acids and their complexes [4]. Computation time in our implementation of the multigrid solution of the nonlinear Poisson–Boltzmann equation is proportional to N, where N is the number of nodes in the grid. It allows one to handle large molecular complexes such as ribosomal subunits, and long DNA fragments up to 1000 base pairs, including promoter sequences for both prokaryotic and eukaryotic species.

Visualization was performed using the software package MOLMOL [5], with our modifications according to which the potential was visualized at the surface called the «electrophoretic sliding surface» (15 Å from the cylindrical axis of DNA).

The objects of investigation adopted herein are the promoter DNA responsible for the most important and universal cellular processes of transcription [6, 7].

3. ELECTROSTATICS OF NUCLEIC ACID SURFACES

Interaction of DNA with polymerases and other proteins that play key roles in transcription and its regulation, is one of the most important examples of molecular recognition where selective binding of protein to the particular DNA sequence occurs [8]. 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/mol [9], which is quite a large difference considering that only noncovalent forces are involved in protein-DNA binding.

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 [10]. 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 base pairs long DNA sequences, exemplified in this study by *E. coli* promoter regions, which are 411 base pairs long.

The horizontal axis in all the maps shown (Figs. 1 and 2) coincides with the DNA helix axis. The color scale represents the electrostatic potential in units of $k_{\mathrm{B}}T/q$, which is thermal motion energy $k_{\mathrm{B}}T$ per unit of electric charge q. In those units, red color was chosen to correspond to 1.3, blue to 0.8, and white to intermediate values. In this colour scheme, the visualized electrostatic potential values will span a range of $0.5\ k_{\mathrm{B}}T/q$, so that ten unit charges, which are typically present in protein fragments interacting with DNA, will account for a difference of $5\ k_{\mathrm{B}}T/q$, 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.

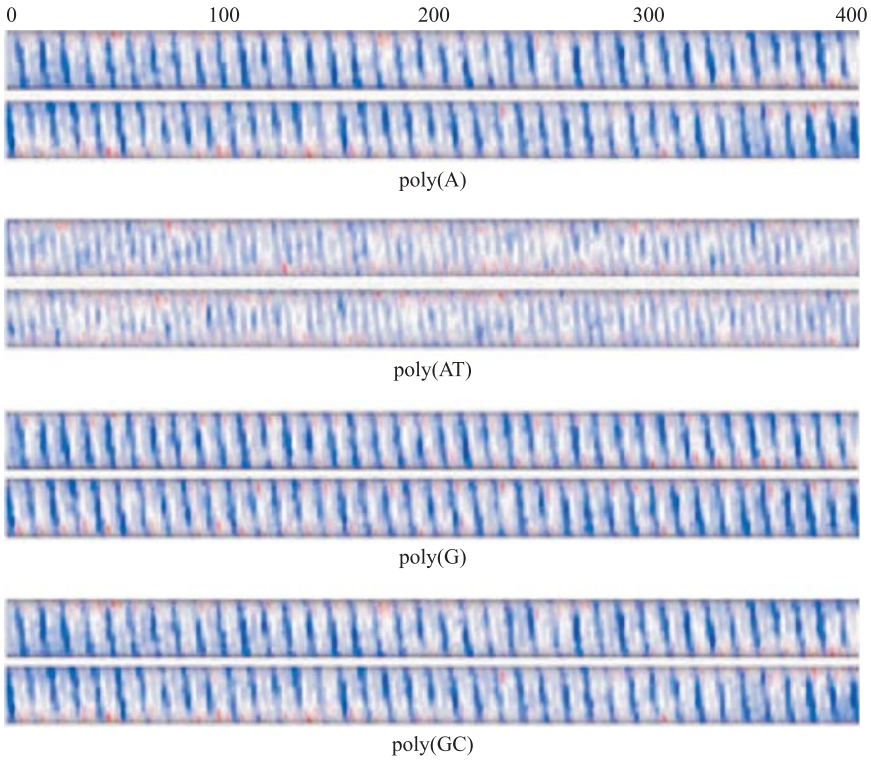

Fig. 1. (Color online). Distribution of electrostatic potential around periodic DNA molecules. Each molecule is shown in two views differing by 180^o rotation around the helix axis

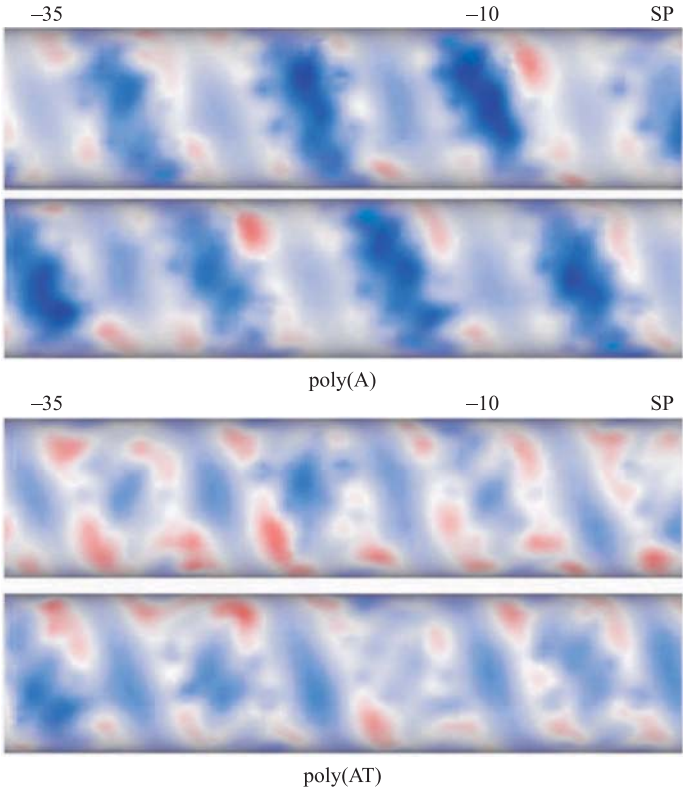

Fig. 2. Distribution of electrostatic potential around accA and uvrA promoter DNAs of *E. coli* from the -35 point to the transcription start point (denoted by SP) shown in two views differing by 180° rotation around the helix axis. The picture is scaled to show finer structure of the specified areas

DNA sequences of *E. coli* promoter regions were taken from [13] and [14]. 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 this 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. 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.

To show the finer structure in functionally important promoter areas (-35, -10, and starting point), the electrostatic potential distribution in these areas are presented for two promoters, accA and uvrA (Fig. 2), scaled to include these areas only.

Long and numerous DNA sequences present a «natural» field of application for distributed computations. Promoter sequences responsible for regulation of transcription of every gene require particular attention. Considering that classification of DNAs based on their electrostatic potential will require pairwise quantitative comparisons of those potentials, distributed computations are possibly the only choice because there are about $35{,}000$ genes in mammalian genomes, and a promoter sequence is at least several hundreds base pairs long for each of them.

The following procedure describes the organization of massive calculations of electrostatic potentials in the Windows environment.

4. CALCULATION PROCEDURE

Below is presented the preparation work that is needed to be done prior to calculation.

4.1. Preparation Instructions

1) Create a folder to store all your files. Sufficient space needs to be allocated especially if you intend to calculate a lot of promoters. We will refer to this folder as Global Destination folder moving forward.

2) Install HyperChem Release 7.

3) Run HyperChem. Go to Setup $\gg$ Molecular Mechanics $\gg$ Amber. Setup $\gg$ Select Parameters Set $\gg$ Amber94.

4) The following files need to be copied to Global Destination folder: ARAD.INP, PARAM.INP, showcyl.mac, cyl_-35.mac, cyl_-75.mac, otherwise an error will be returned during calculation.

5) The following files need to be copied to Windows\System32: nucacid.exe, renum.exe, eleful.exe, outsol.exe, grep.exe, cut.exe.

6) Open PR421_v3.xls.

7) On [Program] sheet specify the value for Global Destination Folder (D2), HyperChem installation folder (D4), images folder (E5).

8) PR421_v3.xls ≫ [PR421-1] contains the sequences for calculation. The tool consumes the lines starting from the first line until it runs into a blank line. Hence, e.g., in order to calculate 20 first promoters, you need to insert a blank row before 21st row.

4.2. Calculation Instructions. This subsection contains instructions for calculations. PR421_v3.xls represents a framework for calculation. The following sheets are in the scope:
[Program] main tool interface.
[Manual steps] a copy of manual steps for a single sequence for informational purposes.
[PR421-1] promoter sequences for calculation.

1) Go back to [Program] tab in PR421_v3.xls file.

2) Click on [**Step 1**] button. It will take about 1–10 seconds to complete. Upon completion for each promoter a separate folder will be created in Global destination folder. Moreover, the following files will be copied to each folder: a text file with the sequence, ARAD.INP, PARAM.INP. showcyl.mac, cyl_-35.mac, cyl_-75.mac.

3) Click on [**Step 2**] button. It takes a few seconds to create nucacid_command.bat in global destination folder.

4) Run the bat. ≫ upon completion .scr file will be created in each folder.

5) Click on [**Step 3**] button. A modification to .scr file will be made in each folder to include hydrogens and exclude connectivities.

6) Click on [**Step 4**] button. HyperChem.bat files will be created in Global Destination Folder. Run the file once it is created and wait for this action to complete. On this step the script (.scr file) will be run for each folder (promoter sequence) in HyperChem application.

7) Click on [**Step 5**] button. ModifyFiles.bat is created in Global Destination Folder as a result of this operation. Run the file once it is created. The following modifications will be made in each folder: proper lines numeration in ATOM.ENT will be established and saved as atom.inp with the help of renum utility; all lines starting with anything other than ATOM will be excluded from atom.hin; atom.tmp will be broken down to columns using space separator, the following columns will be kept in atom.xls B-2 (atom number), G-7 (atom charge), C-3 (atom identificator), all other columns will be removed.

8) Click on [**Step 6**] button. Atom.xls in each folder will be opened and modified to break data down to columns and numerate column A; columns order changed to ensure the following order: number, charge, name; in column H absolute charge fir O1P and O2P is reduced by 0.25; .xls is transferred to charge.inp.

9) Click on [**Step 7**] button. elefull_outsol.bat is created on output. Run this file once it is created. The following outputs will be provided upon completion of this step in each folder: U.dat which includes calculated potential will be created;

fort.7 file with cylinder data which will be used to project the potential calculated in U.dat.

10) Click on [**Step 8**] button. On this step images visualization for each promoter is done.

5. ORGANIZATION OF MASSIVE CALCULATIONS AT JINR CICC

Massive calculations of structure maps were organized by setting up multi-task calculations within a network of distributed computations in which several thousands of usual processor cores can be involved. In this case, many copies of the same programs for different fragments of protein are run from a personal computer in batch mode, and further results are transferred to personal computers for detailed analysis or output of maps. Advantage of this approach is absence of need to use parallelized codes in a different programming language.

The computational farms of JINR CICC (Central Information Computing Complex) forming a Linux-cluster [15, 18] are based on a Unix-like operating system with the distributed file system AFS (Andrew File System) [16], which is introduced at JINR for a uniform file space for users.

The basic principle of AFS is splitting disk space of the user into three main directories, home, scratch, and tmp (src). The home directory for long data storage is the safest in terms of safety from unauthorized access and various failures. Scratch are the working directories for storage of large volumes of data. Tmp (src) are temporary directories, data storage, for example, while a program is being run. Batch tasks are implemented in AFS by the PBS tasks (Processing Batch System) [17], allowing one to manage tasks over a wide set of configurations of computational nodes.

Such an implementation of a distributed computation network necessitated development of a number of the programs operating under Unix OS, and creations of programs for file exchange with Windows OS.

A script program is an executable text file containing commands of PBS for running and monitoring of processes of mapping within a distributed computation network.

The monitoring program with the graphic interface, working under Windows OS which in this case monitors calculation results in the AFS system. At the initial stage of software development, the operator usually handles monitoring.

6. CALCULATION OF BIOPOLYMER STRUCTURE MAPS

For development of console versions of mapping programs described herein and written in the Delphi language, the cross-platform programming environment Lazarus [19] was chosen.

Executables of console programs are run from a command line with four parameters: program name, name of the studied PDB file (whether entire or for selected atomic regions), name of the emulation file of the interface settings, and name of the map of a fragment of a macromolecule which is the program output.

Examples of command lines are below:

./ **SURFACE-2008-compact** protein1_1h8a-A.pdb **loadpar_surf.txt** protein1_1h8a-A

./ **PROT-Zcompact** prot7_helix_1trr-A_80-91.pdb **loadpar_prot.txt** prot7_helix_1trr-A_80-91

./ **helix-DNA-Zcompact** dna6_3cro-L_A14-20_B2-8.pdb **loadpar_dna.txt** dna6_3cro-L_A14-20_B2-8

The third parameter, the text file of emulation of the interface settings, can bear any name, and for each program its content differs by a small number of the interface parameters. The main difference of these parameters will be that the interface file for helical proteins will contain scaling parameter along Z axis, while three more parameters will be added of rotation around coordinate axes before loading of the input PDB file for DNA.

After all parameters are inserted in a command line and «*Enter*» key is pressed, calculations will result in three files with the corresponding names:

1) The CHT file — the saved map of a protein complex or DNA/RNA.

2) The SAV file comprising the exact copy of the PDB file (necessary for map visualization).

3) The new INFO.txt file containing results summary of the calculation (map type, computation duration, computation date, a directory name for results).

6.1. Script Program Controlling Massive Calculations. The task can be started on a farm of a cluster manually for single calculations, via a command line. However, when the number of such calculations reaches about one hundred, an issue emerges of simplification of this process. Such actions can be automated in a script program. According to design of the AFS system, tasks for mass computations must be started from temporary scratch directories available over at all nodes of the Linux cluster, using script program, where the PBS commands are used as the tools.

This allows one to determine a concrete farm of Linux cluster and the node within it, task management, monitoring task the status and sending results to the user in the specified directory, emailing notifications to the user about success or failure. All these options can be used in a script program.

Program script was developed in the distributed AFS file system in shell (bash) language version 3.0. For correct realization of the script program objective, we divided it into two script modules.

6.1.1. Functions of the First Script Module

1) Initialize all variables.

2) Search for input PDB files, the interface emulation file loadpar.txt, and the console program.

3) Create structure of catalogs corresponding to names of PDB files in a temporary directory:

/[temp dir]/[users]/[username]/[protein2008] (for globular proteins);

/[temp dir]/[users]/[username]/[helix_protein] (for the helical of proteins);

/[temp dir]/[users]/[username]/[helix_dnarna] (for DNA/RNA).

4) Copy on one input loadpar.txt, file the executed program and the second script module into these sections, respectively.

5) Run the second script module with parameters transferred to it for each of PDB files from each temporary directory by qsub command of a PBS system.

6) Create a hierarchy of catalogs for results in the home catalog of the user HOME.

6.1.2. Functions of the Second Script Module

1) Copy of the executable program, the file of protein datafile, and the interface emulation module from a temporary directory to the cluster.farm.

2) Start into the console program, estimate duration, and save, by appending to INFO.txt, information on date of creation and time of calculations of the input PDB fragment.

3) Copy results to the specified catalog of the home directory of the user on the remote server.

In Fig. 3, the scheme of all computing process realized by both script modules is shown. According to the principle of disk space splitting of AFS, the script program keeps results in the home directory, creating in it a separate catalog with the following structure:

- [catalog_results];
- [catalog_data_type (globular proteins, helical proteins, DNA/RNA)];

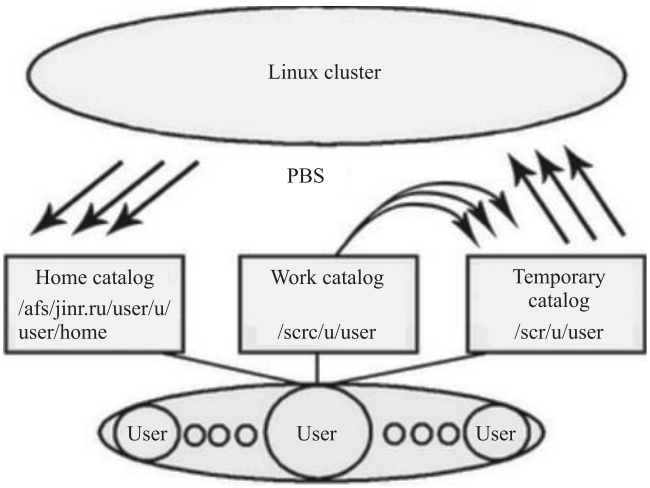

Fig. 3. Flowchart of the script program on the JINR CICC cluster

• [catalog_creation_data];
• [catalog_pdb-filename (contains immediately results)];
• CHT-file, SAV-file, INFO.txt.

To visualize the calculated maps and their further analysis, the corresponding complete versions of programs for DNA, RNA and proteins are used.

7. CONCLUSION

Until very recently, despite the fundamental ideas of Schroedinger (1944) [20], physics had little to contribute to biology simply because of multiplicity and diversity of biology, at the levels of 1) genes and proteins within a single organism; 2) organisms within a single biological species; and 3) different species. Therefore, the ability to process multiple datasets becomes a strict necessity for understanding biological phenomena by means of computational biophysics. This is exactly what was done in this work, namely, very large, genome-wide, biological (DNA) datasets were made accessible to computations of electrostatic potentials that were earlier proven to govern molecular recognition of promoter DNA by proteins, and eventually the function of promoter DNA.

It must be noted that the approach outlined herein and a similar compuuttional tool can be used for calculating and visualizing electrostatic DNA-recognizing proteins eventually highlighting the role of electrostatic in promoter recognition by their natural counterparts, transcription factors. See an example shown below (Fig. 4).

With the help of the developed procedures for organization of massive calculations in a distributed computing environment were calculated electrostatic potentials about 400 proteins («zinc fingers») and more than 100 structure maps of biopolymers.

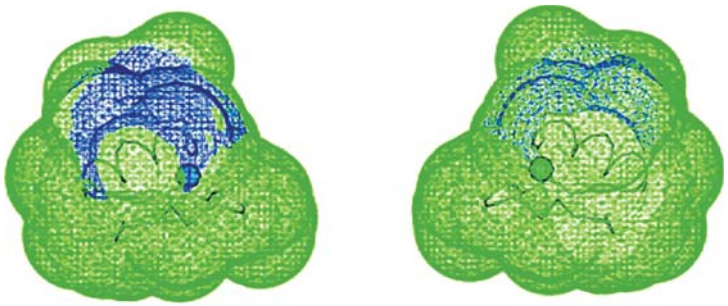

Fig. 4. (Color online). Electrostatic potential of the first zinc motive of the factor of a transcription Zif268 5.5 Å away from its molecular surface. Structure PDB code is 1a1f, subunit is Å, restudies are 100–134. Blue color is positive potential, green neutral (negative is absent). Elements of secondary structure are shown: in the center is the recognizing helix. Cyan color shows the zinc ion

APPENDIX

Model Building, Dielectric Boundary and Charge Assignments. All atom models of DNA fragments were constructed using the evaluation version of the HyperChem 7.01 package [11]. 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 [12]. Additional charges of $0.25\ q$ were assigned to O1 and O2 atoms of phosphate groups to allow for the well-known counterion condensation effect, which is retention of part of counterions 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 sugarphosphate 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.

File Reference

List of input and output files

File	Description
ARAD.INP	Half-width of Gaussian charge distribution over closest grid points
atom.ent	Cartesian coordinates of promoter atoms
atom.hin	Cartesian coordinates of promoter atoms and their charges
atom.tmp, atom.xls	Temporary files for format interconversion
charge.inp	Promoter atomic charges used as input for electrostatic calculations
elefull.out	Log file of electrostatic calculations
elefull_outsol.bat	.bat file created on Step 7
fort.7	Contains the data for the cylinder to project the electrostatic potential onto
HyperChem.bat	.bat file created on Step 4
ModifyFikes.bat	.bat file created on Step 5
nucacid_command.bat	.bat file created on Step 2
PARAM.INP	Parameters of electrostatic calculations
showcyl.mac, cyl-35.mac, cyl-75.mac	A macro files for visualizing the data in the file U.dat for the entire promoter and functionally important regions
U.dat	A file created for each promoter sequence. Includes the data for the calculated potential

This work was supported in part by the grant of the Russian Foundation for Basic Research 07-07-234.

REFERENCES

1. *Parsons J. D.* // Comput. Appl. Biosci. 1995. V. 11. P. 603–613.
2. *Guan X., Du L.* // Bioinformatics. 1998. V. 14. P. 783–788.
3. *Cornell W. D. et al.* // J. Am. Chem. Soc. 1995. V. 117. P. 5179–5197.
4. *Fedoseyev A. I. et al.* // Abstr. of the Intern. Conf. «Physique en Herbe92», Marseille, France, July 1992;
   Sivozhelezov V. S. // Abstr. Intern. Conf. CSAM93, St. Petersburg, Russia, 1993. P. 121;
   Sivozhelezov V., Nicolini C. // J. Theor Biol. 2005. V. 234, No. 4. P. 479–485.
5. *Koradi R., Billeter M., Wuthrich K.* // J. Mol. Graph. 1996. V. 14, No. 515. P. 29–32.
6. *Polozov R. V. et al.* // Biochemistry. 2006. V. 45. P. 4481–4490.
7. *Polozov R. V. et al.* // Part. Nucl., Lett. 2005. V. 2, No. 4(127). P. 82–90;
   Akishina T. P. et al. Study of Electrostatic Potentials of DNA Promoters // XX Intern. Symp. on Nuclear Electronics and Computing (NEC'2005), Varna, Bulgaria, Sept. 12–18, 2005: Book of Abstr., Dubna, JINR, 2005. P. 11.
8. *Romberg R. D.* // Trends Cell Biol. 1999. V. 9. P. M46–M49.
9. *Coleman R. A., Pugh B. F.* // J. Biol. Chem. 1995. V. 270. P. 13850–13859.
10. *Guan X., Du L.* // Bioinformatics. 1998. V. 14. P. 783–788.
11. *Polozov R. V. et al.* // J. Biomol. Struct. Dyn. 1999. V. 16. P. 1135–1143.
12. http://www.hyper.com/products/description/hyper7.htm.
13. http://sigyn.compbio.ucsf.edu/amber/.
14. *Ozoline O. N. et al.* // Mol. Biol. 2002. V. 36. P. 682–688.
15. http://lit.jinr.ru/ccic/usersguide/
16. http://lit.jinr.ru/ccic/usersguide/index.php?link=3_
17. http://lit.jinr.ru/ccic/usersguide/index.php?link=2_#2.4_
18. http://www.scientificlinux.org/
19. http://www.lazarus.freepascal.org/
20. Schrodinger E. What Is Life? Cambridge: Cambridge University Press, 1944.

Received on December 24, 2013.

Корректор *Т. Е. Попеко*

Подписано в печать 26.02.2014.
Формат 60 × 90/16. Бумага офсетная. Печать офсетная.
Усл. печ. л. 0,75. Уч.-изд. л. 1,26. Тираж 245 экз. Заказ № 58197.

Издательский отдел Объединенного института ядерных исследований
141980, г. Дубна, Московская обл., ул. Жолио-Кюри, 6.
E-mail: publish@jinr.ru
www.jinr.ru/publish/