E19-2008-174

Yu. N. Chirgadze[1], V. V. Ivanov[*], R. V. Polozov[2],
V. S. Sivozhelezov[3], E. I. Zheltukhin[2]

STRUCTURAL AND ELECTROSTATIC REGULARITIES IN INTERACTIONS OF HOMEODOMAINS WITH OPERATOR DNA

[1] Institute of Protein Research, Russian Academy of Sciences, Pushchino, Russia

[2] Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences, Pushchino, Russia

[3] Chair of Biophysics, University of Genoa, Genoa, Italy

[*] To whom correspondence should be addressed: tel: (49621) 64887; fax: (49621) 65145; E-mail: ivanov@jinr.ru

Чиргадзе Ю. Н. и др. E19-2008-174

Структурные и электростатические закономерности во взаимодействиях гомеодоменов
с операторной ДНК

Проведено сравнение интерфейсов пяти комплексов белок–ДНК, отобранных исходя из подобия трехмерных структур и свойств контактирующих аминокислотных остатков. Начальная стадия процесса узнавания охарактеризована электростатическими потенциалами на расстоянии около 5 Å от молекулярных поверхностей белков и ДНК. У белков четкий положительный потенциал наблюдается только со стороны, контактирующей с ДНК, а в желобах ДНК — сильный отрицательный потенциал, т. е. одна из функций электростатики состоит в направлении белка в большой желоб ДНК. На близкой стадии узнавания нейтрализация зарядов фосфатов ДНК лизинами и аргининами белка необходима для ослабления электростатического потенциала ДНК, мешающего основаниям ДНК участвовать в образовании атомных контактов белка с ДНК в интерфейсе. Узнающая α-спираль белка образует как инвариантные, так и вариабельные контакты с ДНК посредством определенных специфических боковых групп, причем в некоторых из контактов участвуют молекулы воды. Инвариантные контакты включают высокоспецифичные водородные связи Asn-Ade, неполярные контакты гидрофобных аминокислот, служащие барьерами для фиксации белка на ДНК, и кластер интерфейсных молекул воды, обеспечивающий подвижность, необходимую для диссоциации комплекса белок–ДНК. Одна из молекул воды инвариантна и расположена в центре интерфейса. Инвариантные контакты во всех комплексах образуются мотивом TAAT прямой цепи ДНК. Они выделяют семейство гомеодоменов среди прочих ДНК-связывающих белков. Вариабельные контакты образуются с обратной цепью ДНК и отвечают за специфичность связывания внутри семейства гомеодоменов.

Работа выполнена в Лаборатории информационных технологий ОИЯИ.

Сообщение Объединенного института ядерных исследований. Дубна, 2008

Chirgadze Yu. N. et al. E19-2008-174

Structural and Electrostatic Regularities in Interactions of Homeodomains
with Operator DNA

Interfaces of five DNA-homeodomain complexes, selected by similarity of structures and patterns of contacting residues were compared. The long-range stage of the recognition process was characterized by electrostatic potentials about 5 Å away from molecular surfaces of both protein and DNA. For proteins, clear positive potential is displayed only at the side contacting DNA, while grooves of DNA display a strong negative potential. Thus, one functional role of electrostatics is guiding the protein into the DNA major groove. At the close-range stage, neutralization of the phosphate charges by positively charged residues is necessary for decreasing the strong electrostatic potential of DNA, allowing nucleotide bases to participate in formation of protein-DNA atomic contacts in the interface. The protein's recognizing α-helix was shown to form both invariant and variable contacts with DNA by means of the certain specific side groups, with water molecules participating in some of the contacts. The invariant contacts included the highly specific Asn-Ade hydrogen bonds, nonpolar contacts of hydrophobic amino acids serving as barriers for fixing the protein on DNA, and interface water molecule cluster providing local mobility necessary for the dissociation of the protein-DNA compex. One of the water molecules is invariant and located at the center of the interface. Invariant contacts of the proteins are mostly formed with the TAAT motive of promoter DNA's forward strand. They distinguish the homeodomain family from other DNA-binding proteins. Variable contacts are formed with the reverse strand and are responsible for the binding specificity within the homeodomain family.

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

Communication of the Joint Institute for Nuclear Research. Dubna, 2008

INTRODUCTION

A vast majority of known studies of protein-DNA complexes are related to the binding site of proteins with double-stranded DNA in the B-form in the region of its major groove [1–9]. In paper [4] specific features of protein surface patches of DNA-binding domains, such as accessibility, electrostatic potential, hydrophobicity, as well as residue propensity and conservation have been analyzed. Positive electrostatic potential appears to be the most effective feature for the binding site recognition. In this case [4], about 68% of 56 nonhomologous DNA-binding proteins was correctly described by the prediction. This approach has been improved by adding information about the secondary structure of the protein binding motif [5] and the shape of contacting molecular surface [6]. In this paper we try to deduce the *structural and electrostatic regularities* in the organization of the protein-DNA complex.

Among various complexes with known spatial structures, complexes of transcription factors bound to double-stranded operator B-DNA (Fig. 1) are the most wide-spread ones. In all these cases the main binding site is formed by a single *recognizing α-helix* bound with the region of the DNA major groove. The interaction between polar side groups of protein and DNA presents an essential part of the interactions, as shown earlier [9]. On the other hand, it is well known that the polar groups on the surface of globular proteins are joined into clusters. For example, side groups of charged amino acids form sign-alternating charge clusters which can be considered as protein surface structural invariants [10]. Such clusters are available in 86% of nonhomologous protein structures [11]. We have shown that all polar side groups of protein form polar clusters, and some large clusters play a distinctive functional role in the binding of some protein-DNA complexes [12, 13]. A similar functional role of polar cluster was clearly revealed for different protein-RNA complexes as well [14]. Therefore, at present a key role of polar residue clusters is essentially defined both for protein-DNA and protein-RNA complexes.

Contacts between protein and DNA have been traditionally identified from coordinates of the corresponding DNA/protein pairs of atoms using stereochemical criteria, such as distance thresholds between the atoms for hydrogen bonds or nonpolar contacts. It should be noted that such a definition of contacts is not sufficient for all cases, because it ignores the structural context of occurrence of each given contact, namely «structural relationships» between the path of the protein backbone at each amino acid and the plane of each nucleotide base [15].

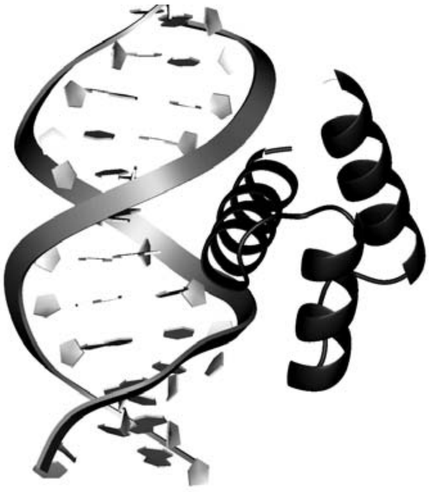

Fig. 1. Spatial backbone model of a complex of transcription factor — homeodomain Msx-1 from *E. coli* with a fragment of operator DNA. Only one protein binding domain A is shown, whose third recognizing α-helix forms contacts with atomic groups of B-DNA in the region of the major groove

In that work, however, one family of proteins, namely the homeodomain family, was identified in which those relationships were found to be very similar within this particular family and very different from all other families. This prompted us to select this family for analysis, considering also that homeodomains are known as the most ancient and conservative.

Among inducible transcription factors, homeodomains are nearly the most important ones in eukaryotes since they control differentiation development of embryonic cells into tissue or organ-specific cells. Initially discovered in fruit flies, these factors have been found in all vertebrates [16]. Structurally, homeodomains are three-helical bundles accompanied, at their N-terminus, by a basic loop. Their *recognizing helix* is the C-terminal helix, and this helix is penetrating the major groove [17]. Interactions of homeodomains with DNAs have been extensively reviewed in [18], but no attempt to reveal the functional meaning of contacts or classify those contacts has been performed. According to the SCOP database [19], there are nearly 30 homeodomains with known structures, and the entire number of PDB entries in the homeodomain-like family is about 150 [20]. Most homeodomains, even those widely differing in primary structures, have very similar tertiary structures, coinciding as a rule with RMS deviation below 1 Å, and within RMS of 0.6 Å in the recognizing helix. Structures of homeodomains very often remain practically unchanged upon binding to DNA [21, 22]. A recent molecular dynamics study has shown that not only the structure but also the mobility of the «engrailed» homeodomain remains the same in the DNA-bound as in the free form [23]. Thus, the homeodomain backbones have rather a rigid conformation, so the complication of major structural changes of protein upon binding with DNA is excluded. This facilitates the functional assignment of each feature of protein-DNA recognition, such as electrostatic fields of DNA and homeodomains, polar contacts including hydrogen bonds, nonpolar contacts, and water molecules mediating interactions in the DNA-homeodomain complex.

The DNA-binding protein domains, in the complexes we have selected, satisfied the following criteria:

— similarity of tertiary structures having three α-helical segments and belonging to the homeodomain family;

— similarity of the binding pattern of polar residues located on the surface of the recognizing α-helix;

— sufficient resolution (better than 2.5 Å) to capture interfacial water molecules.

These selected homeodomains differ from each other within 0.6–0.8 Å in their overall backbone coordinates and within 0.2–0.4 Å in the backbone coordinates of the recognizing helix. Considering that the RMS difference between molecules of the same protein in two complexes within the same asymmetric unit may reach 0.5 Å [24], these differences are practically within the experimental error. All DNAs in the complexes considered herein are close to the canonical B-form. The PDB entries for the selected structures belong to the following homeodomains: Msx-1 (1IG7, [25]), antennapedia (9ANT, [26]), engrailed (3HDD, [21]), paired (1FJL, [27]), and Pit-1 (1AU7, [24]). Thus, these homeodomains are practically identical in terms of their protein backbone, and typical of the homeodomain family. However, the pairwise homology of the amino acid sequences of the recognizing α-helix was below 50%, and pair homology of the nucleotide sequence for the binding DNA-segment was also low. Even with the low homologies, we have clearly observed a common functional meaning of the majority of available contacts between the factor and DNA in these complexes.

Since the protein-DNA recognition is known to have a hierarchical nature [28], we analyzed separately *nonspecific long-range* electrostatic interactions followed by *specific interatomic contact* interactions including close-range van der Waals and hydrogen-bonding contacts.

1. METHODS

Atomic coordinates of complexes were taken from the Nucleic Acid Database [29]. PDB codes are 1IG7-A, 9ANT-A, 3HDD-A, 1FJL-A, and 1AU7-A2, where last digits indicate the protein domains used in the analysis. High-resolution crystallographic data, up to 2.5 Å, were chosen as the only suitable for analysis. Such data give sufficient accuracy of atomic positions and include molecules of structural water, which play a significant role in the binding of protein to DNA. In this work, the residues belonging to the N-terminal «arm», are not analyzed because this region is disordered in the free homeodomains.

In addition to an ordinary used numbering given in the PDB files, we introduced two numbering systems of amino acid residues. One was describing the positions i of amino acids within the recognizing helix beginning with the position $i+ = 0$ assigned to the first amino acid contacting DNA. For example, the position number $i+ = 0$ relates to Lys146 for complex 1IG7. The second numbering system is based on Msx-1 homeodomain (PDB 1IG7) as a count for all other protein's sequences unless otherwise specified.

Since the complexes are embedded in water medium, it is very reasonable to distinguish hydrophilic and hydrophobic interactions as proposed in [14]. The hydrophilic interaction is caused by polar groups of atoms such as =NH, -OH, -NH_2, -NH_3^+, =CO and -COO^-. These groups interact with each other and with water to form hydrogen or ionic bonds. The hydrophobic interactions are determined by nonpolar groups of atoms such as ≡CH, -CH_2-, -CH_3, -SH and -CH=CH-. Evaluation of these interactions for the binding area can be done by calculating the number of contacts of polar and nonpolar atomic groups. The distance limits of atomic contacts were taken from [3]. The direct and water-mediated contacts of polar atoms were determined at distances less than 3.35 Å. These contacts specify hydrophilic interactions and could be related to hydrogen and partially ionic bonds. Contacts between nonpolar atoms were determined at distances less than 3.9 Å. Both types of interatomic interactions were analyzed between binding protein domains of transcription factors and double-stranded DNA fragments in the region of the major groove.

Structural alignments of tertiary structures of five complexes were used to identify the invariant and the variable subsets of the atomic contacts. At the first stage of the procedure we superimposed the axes of the recognizing helixes of all considered transcription factors. At the second stage we superimposed those atoms of amino acids which form common for all complexes contacts with the bases atoms of DNA taken from X-ray data. In this case we have also used three reference points:

— atom C of amide group of Asn 51 ($i+ = 5$);

— atom N7 of adenine (TAAT motif of 5'-chain of DNA, Fig. 3) contacting with Asn 51;

— atom O of central water molecule contacting with Asn 51.

By means of this procedure we determine the common atomic contacts for all five complexes and represent visually their spatial arrangements. All such contacts we considered below as invariant. The procedures were performed by using the MOLMOL software (http://www.mol.biol.ethz.ch/groups/wuthrich_group/software). All structures of the complexes were manually superimposed onto the structure of the Msx-1 homeodomain (any other transcription factor could be also selected), alternately for the recognizing helix and the entire homeodomain. In both cases we have obtained very similar results, further we have used an alignment with the recognizing helixes. The RMS deviation of the Cα atoms for amino acids participating in the contacts according to the above-mentioned distance criteria was found to be within 0.5 Å. While the RMS values for the Cα atoms for the entire homeodomain were found to be below 0.75 Å. These values have been obtained by repeated procedures in order to refine and adjust the structures. A coincidence for positions of contacting atoms of protein and DNA with RMS below 1.0 Å has been defined as a numerical measure of invariant contacts. Other contacts will be considered as variable, and they were not analyzed in detail in this paper.

Electrostatic potentials were determined by solving the Poisson–Boltzmann equation, as specified in [30]:

$$-\nabla(\varepsilon(\mathbf{r})\nabla\varphi(\mathbf{r})) = 4\pi(\rho_0(\mathbf{r}) + \rho_1(\varphi(\mathbf{r})),$$

where $\rho_0(\mathbf{r}) = \Sigma_i z_i q\delta(\mathbf{r}_i)$ is the charge distribution of the protein molecule, $\rho_1(\varphi(\mathbf{r})) = \Sigma_j z_j n_j q\ \exp(-z_j q\varphi(\mathbf{r})/k_\mathrm{B}T)$ is the distribution of the mobile electrolyte charges, $\varphi(\mathbf{r})$ is the electrostatic potential, $\varepsilon(\mathbf{r})$ is the dielectric constant assumed to be 2 inside the protein and 80 outside, $\mathbf{r}(x, y, z)$ is the radius vector of each observation point, z_i and $\mathbf{r}_i$ are the charge and the radius vector of the ith atom of the molecule, z_j and n_j are charges and concentrations of electrolyte ions, respectively, q is the proton charge, k_B is the Boltzmann constant, T is the absolute temperature.

The atomic coordinates of either protein or DNA were taken from the specified PDB files without further modifications. Charges z_i were taken from the AMBER force field [31]. The electrolyte is assumed to be univalent ($z_1 = -1$; $z_2 = 1$) at physiological concentration of 150 mM. Solution is sought with the finite-difference multigrid method using a sequence of nested finite-difference grids, the finest grid having $192 \times 192 \times 192$ points so that the interval between grid points is less than 1 Å. Analytical solution available when $\varepsilon = \mathrm{const}$ is used as the first approximation as well as to establish the boundary conditions. Potentials were mapped onto the surface 5.5 and 3.0 Å away from the van der Waals surface, and the color was coded so that red was the negative ($\varphi < -0.5k_\mathrm{B}T/q$) potential, white — neutral ($-0.5k_\mathrm{B}T/q \leqslant \varphi \leqslant 0.5k_\mathrm{B}T/q$), and blue — positive ($\varphi > 0.5k_\mathrm{B}T/q$). The choice of 5.5 Å distance for potential mapping is explained by the facts that, at this distance away from the van der Waals surface, the surface is approximately 7 Å away from the charges contributing to the potential, which is equal to the Bjorrum length. At this length the interaction energy between two unit charges in water equals $k_{\mathrm{B}}T$. At distances larger than 5.5 Å, the energy of electrostatic interaction becomes comparable to that of the thermal motion. At distances smaller than 3.0 Å, the Poisson–Boltzmann equation may no longer be valid because of electronic fluctuations and correlations [32].

2. RESULTS

2.1. Electrostatic Properties of Interacting Partners for Complexes of Homeodomains with Operator DNA at the Distances 3.0–5.5 Å. For all five protein factors, the map of the electrostatic potential onto 3.0 and 5.5 Å equidistant surfaces is shown in Fig. 2. A region of this surface is facing the major groove of DNA. This region will be further termed the front surface. The opposite surface of the protein will be termed the back surface. Characteristic regions of the positive potential are visible on the front surface and almost completely neutral potential on the back surface.

At the distance of 5.5 Å positive potential occupies almost the entire front surface for the antennapedia (9ANT-A) protein, about 0.75 for the engrailed (3HDD-A) protein, about 0.5 for both the Msx-1 (1IG7-A) and the paired (1FJL-A) proteins, and less than 0.25 for the Pit-1 (1AU7-A2) protein. Overall charges

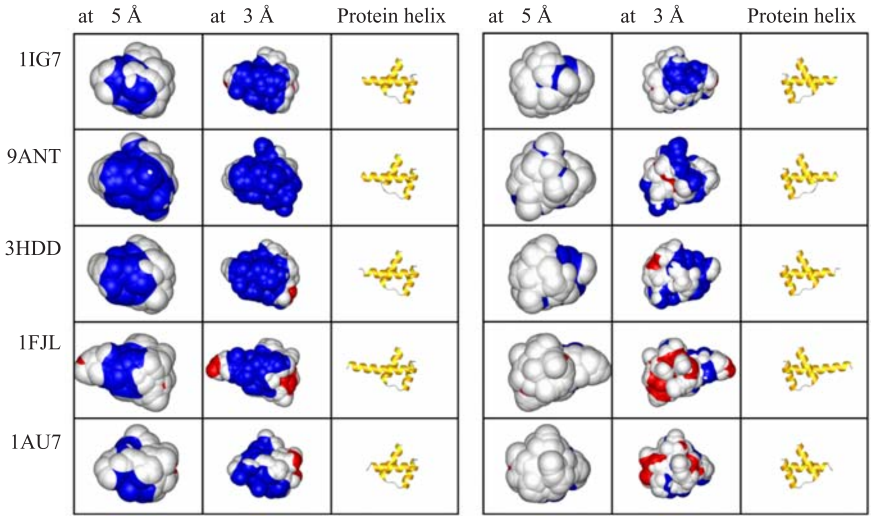

Fig. 2. Electrostatic potentials of protein factor molecules at the distances of 5.5 and 3.0 Å from their van der Waals surfaces. Left sub-table shows the front side of the protein surface contacting DNA, right sub-table shows the back side of the protein surface

for those proteins are +7, +8, +6, +2, +3 q, and dipole moments 339, 517, 439, 468, 557 D, respectively [33]. For the Msx-1 protein, the positive potential region at the front side encompasses the recognizing helix excluding two amino acid residues at its central part. The positive potential region for the engrailed and antennapedia proteins covers not only the entire recognizing helix, but also the laterally adherent regions. The positive potential region of the paired protein is delimited by negatively charged amino acids located at both ends of the recognizing helix. Unlike the remaining four homeodomains, recognizing helix of the Pit-1 protein is entirely encompassed by the neutral potential, while the positive potential is located in the laterally adherent regions. Thus, large areas of positive potential are present only on the front surface of all five homeodomains, while the back surface is almost completely neutral.

For the 3.0 Å equidistant surfaces, the front surface of all proteins is completely covered with positive potential, again with the exception of the Pit-1 protein displaying neutral potential in the region corresponding to the center of the recognizing helix. The back surface is filled with a mosaic distribution of small patches of positive and negative potentials, with the exception of Msx-1, for which a large area of positive potential expands to the back surface, covering about 30% of it.

Electrostatic potentials of the DNA fragments at 3.0 and 5.5 Å distances from their molecular surfaces were shown to be determined by charges on the DNA phosphates in the range of realistic degrees of neutralization by counterions 0.3–0.5 (data not shown).

2.2. Patterns of Binding of the DNA Fragments and the Homeodomains in the Complexes. Nucleotides involved in contacts of operator DNA with the protein differ strongly between the two DNA chains, both in terms of the identity of the nucleotides and their positions in the nucleotide sequences (Fig. 3). Here the nucleotides, which contact residues of the recognizing α-helix, are underlined. Practically, all five sequences which are marked as 3'-end display variable contact regions. In fact, the homology is observed only for the limited sequence part 5‘-TAAT which is designated in the figure as the 5'-end. It is the part of the sequence that defines the conservative motive. The subsequent nucleotide region shows only a very weak homology, so it can be regarded variable. The forward (Watson) strand of the conservative motive and the reverse (Crick) strand of the variable region form the specific structural pattern recognized by the homeodomain. Indeed, nucleotides of the forward strand of the 5‘-TAAT motif form contacts with conservative amino acids in positions $i+ = 1$ and 5, while amino acids in positions $i+ = 4$ and 8 form contacts with the reverse DNA strand (Figs. 3, 4).

Let us consider aligned sequences of the recognizing helix of transcription factors more closely. Here the polar residues, which form contacts with B-DNA in the region of the major groove, are hatched. The whole protein domain's

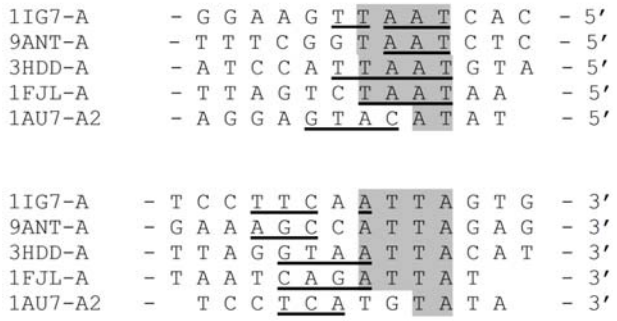
1IG7-A - G G A A G T T A A T C A C - 5′
9ANT-A - T T T C G G T A A T C T C - 5′
3HDD-A - A T C C A T T A A T G T A - 5′
1FJL-A - T T A G T C T A A T A A - 5′
1AU7-A2 - A G G A G T A C A T A T - 5′
1IG7-A - T C C T T C A A T T A G T G - 3′
9ANT-A - G A A A G C C A T T A G A G - 3′
3HDD-A - T T A G G T A A T T A C A T - 3′
1FJL-A - T A A T C A G A T T A T - 3′
1AU7-A2 - T C C T C A T G T A T A - 3′

Fig. 3. Nucleotide sequences of two DNA chains of five considered complexes. Alignment was performed on the promoter part TAAT which is shown in grey color. All binding nucleotides are underlined. Left first column lists PDB codes of complexes

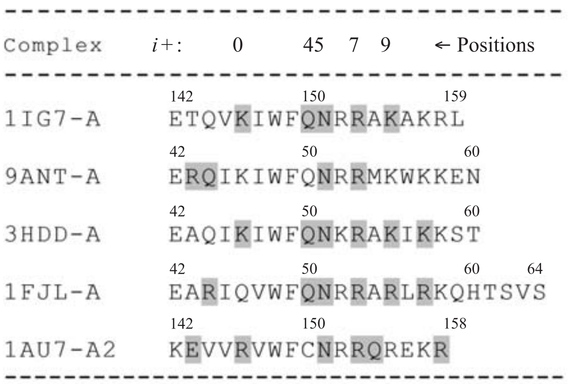
Complex
i+: 0 45 7 9 ← Positions
142 150 159
1IG7-A ETQVKIWFQNRRAKAKRL
42 50 60
9ANT-A ERQIKIWFQNRRMKWKKEN
42 50 60
3HDD-A EAQIKIWFQNKRAKIKKST
42 50 60 64
1FJL-A EARIQVWFQNRRARLRKQHTSVS
142 150 158
1AU7-A2 KEVVRVWFCNRRQREKR

Fig. 4. Amino acid sequences of the recognizing α-helix for five considered transcription factors. Polar residues, which form contacts with B-DNA in the region of the major groove, are shown in grey boxes. Positions $i+$ of these residues, which form contacts with probability more than 60%, are listed in the upper line

chains have lengths of 58, 57, 55, 65 and 68 residues for 1IG7-A, 9ANT-A, 3HDD-A, 1FJL-A, and 1AU7-A2, respectively, and they rather differ in the sequences. Sequences of the third recognizing α-helix are presented in Fig. 4. The average pair homology of the common part of a recognizing helix of 17 residues makes 49%, and homology values vary from 29 to 70% (Table 1).

We marked the positions of binding polar residues with identity in three or more complexes. These distinctive positions $i+$ are equal to 0, 4, 5, 7 and 9,

Table 1. Pair homology by residue identity of recognizing α-helix of factor transcriptions, %

	1IG7	9ANT	3HDD	1FJL	1AU7
1IG7-A	100	70.6	70.6	47.1	47.1
9ANT-A		100	76.5	52.9	35.3
3HDD-A			100	58.8	29.4
1FJL-A				100	41.2
1AU7-A2					100
Note. Averaged homology of all pairs is equal to 49.1%					

and they are occupied by residues Lys (Arg), Gln, Asn, Arg and Lys (Arg). The positions 0, 7, 9 close to the edge of helix are occupied by the positively charged residues. Central positions 4, 5 are occupied by the residues with neutral side groups containing the partially positive amino group, and the partially negative carbonyl group. All considered polar residues are located on the external surface of recognizing α-helix, and only 1–2 nonpolar residues. The lengths of recognizing helices are of 18, 19, 19, 23 and 17 residues. This corresponds to about six turns of the α-helix for factor 1FJL-A and four turns for all the other factors, although the part of the recognizing helix that contacts DNA has the same length, 17 residues.

Let us consider the structurally aligned contacts of transcription factors with DNA in the complex of homeodomain Msx-1 with DNA (Figs. 4, 5). Below, we are comparing such contacts across several complexes, and thus identify the invariant and variable features of the system of protein-DNA contacts. Within such a system, the contacts are spatially arranged in a manner specific to the homeodomain family.

Each of the conservative lysine and arginine residues in positions 0, 7, 9, 11, 12 (Lys146, Arg153, Lys155, Arg157, Arg158) forms polar contacts with phosphate groups directly or via a water molecule. These amino acids are encompassing the recognizing helix, thus forming the periphery of the recognizing helix interaction with DNA bases of the 5‘-TAAT motive.

The position $i+ = 5$ is occupied by a highly conservative amidic side chain Asn151. The asparagine residue forms a highly specific bidentate polar contact with the bold-marked adenine of the 5‘-TA**A**T motive. Position $i+ = 2$ is occupied by conservative hydrophobic Trp148, which belongs to the hydrophobic core of the homeodomain and at the same time has contacts with the sugar-phosphate backbone forward DNA strand of the 5‘-TAAT motive.

In positions $i+ = 4, 8$ diverse amino acids of variable chemical nature are located. Position $i+ = 4$ is occupied by Gln150 in 1IG7, 9ANT, 3HDD, 1FJL complexes or Cys150 in 1AU7. Position $i+ = 8$ is occupied by Ala154 in

1IG7, 3HDD, 1FJL complexes, Met154 in 9ANT, and residue Gln154 in the structure 1AU7, which forms bidentate contact with adenine of reverse DNA strand (5‘-TAA**T**). The system of contacts of amino acids residues in positions $i+ = 4, 8$ of the considered complexes is variable because these residues are contacting bases of the reverse DNA strand neighboring the TAA**T**, thus forming the 5‘-TAA**TNNN** motive.

In position $i+ = 1$, the conservative Ile/Val147 is located, forming a nonpolar contact with the methyl group of the thymine 7 (numbering as in 1IG7) from the 5‘-TAAT motive. This contact can be strengthened by a contact involving the amino acid in the position $i+ = 4$, such as Cys150 in the 1AU7-A2 structure. Besides, this contact may be accompanied by a nonpolar contact of the thymines complementary to the adenines in the 5‘-TAAT motive with amino acids in other positions such as 8 (Met154 in antennapedia homeodomain).

The protein-DNA interfaces for all five homeodomains contain 10–20 water molecules. Among those water molecules, we can select *the central water molecule*, w_0. This water molecule is specified in corresponding PDB files (1IG7 — residue number of 186, 9ANT — 855, 3HDD — 413, 1FJL — 900, and 1AU7 — 732). It has completely used its hydrogen bonding capacities, and forms hydrogen bonds with Asn151, Gln150, and **T7** (5‘-TAA**T**) but not with other water molecules, which are located more than 3.8 Å away.

3. DISCUSSION

3.1. Electrostatic Potentials of Homeodomains and DNA. The electrostatic potentials of homeodomains indicate that large areas of positive potential are present only on the front surface of all five homeodomains both 5.5 and 3.0 Å away from the protein. The back surface is almost completely neutral at 5.5 Å, and shows mosaic patches at 3.0 Å. The electrostatic fields of DNA and protein at 5.5 Å distance begin to contribute to steering of the protein to DNA through positive potential of the protein interacting with negatively charged phosphates. For this reason, fractions of the front surface areas occupied by the positive potential can be used as rough estimates of the contribution of the protein electrostatic field to long-range protein-DNA recognition. Neither differences in overall charges alone nor differences in dipole moments alone can explain the observed differences in potentials. This is apparently caused by compensation of the contribution of the back surface protein positive charges Lys, Arg to the electrostatic potential by the nearby negatively charged Glu and Asp residues, while contributions of the front surface positive charges remain uncompensated, and thus determine the potential of the front surface. The number of uncompensated charges varies less than two charge units over the five complexes, while the fraction of the front surface positive potential varies from 0.25 to 1, i.e., four times. We conclude that some of the positively charged residues do not participate in *long-range recognition*. The suggested function of such residues in *protein-DNA contact* is neutralization of the charges of DNA phosphates. Only after phosphate charges are neutralized, recognition of the distribution of negative potential caused by the bases in the major groove becomes possible.

The electrostatic potential of DNA is strongly negative for the whole DNA molecule in the physiological solution. At about 5–10 Å distance from the DNA molecular surface, it provides a suitable orientation of the protein molecule [34]. However, tight contact could not be realized due to the lack of a specific interaction at the distances considered. We have found that the electrostatic potential of DNA 3–5 Å away from its molecular surface is dominated by the charges of the phosphate groups rather than those of nucleotide bases. Accordingly, the charges of the phosphate groups of DNA must be neutralized before the nucleotide may take part in close-range protein-DNA recognition. Thus, function of the positively charged amino acids of the homeodomain in its recognition of DNA should be dual. One is to provide a specific orientation of the protein with respect to DNA, and the other is to neutralize the charges of the phosphates, which is necessary for specific, short-range recognition. This is also supported by the observation that the positively charged amino acids, five in the case of 1IG7, are located at the ends of the recognizing helix. They lead to neutralization of the phosphate charges in the vicinity of the recognized 5‘-TAAT motive.

Details of the distribution of electrostatic potentials of proteins at 5.5 Å emerging in the Pit-1 (1AU7-A2) and Msx-1 (1IG7-A) structures could be explained by local conformational features of the recognizing helices. Indeed, at the protein factor Pit-1 neutral potential at the center of recognizing helix is formed by the Ile/Val147 and Cys150, positions 1 and 4, screening the nearby positive charges. In the Msx-1, the sidechain conformation of the Gln150, also position 4, slightly differs from the remaining proteins. This makes the negatively charged oxygen closer to 5.5 Å of the molecular surface, so a spot of the neutral potential appears near the center of the recognizing helix.

3.2. Distribution of Polar Contacts of Protein between DNA Bases and Phosphates. We can find common intermolecular contacts by selecting all conservative or very similar contacts among all complexes. At first, we consider the contacts in complex Msx-1 (code 1IG7, Fig. 5). Here protein atomic groups forming contacts with DNA are shown in black. Direct contacts are presented by filled arrows, and water mediated contacts — by dotted arrows. In this complex, the amount of water mediated polar contacts is less than a half of the total polar contacts. For all five complexes, it also comprises about a half of the total amount of polar contacts. All observed contacts could be divided in two different groups. The first group represents contacts of phosphate groups with positive charged residues of the protein factor, that can be considered as *contacts nonspecific by DNA sequence*. These residues are Lys146, Arg152 and Lys155, their positions

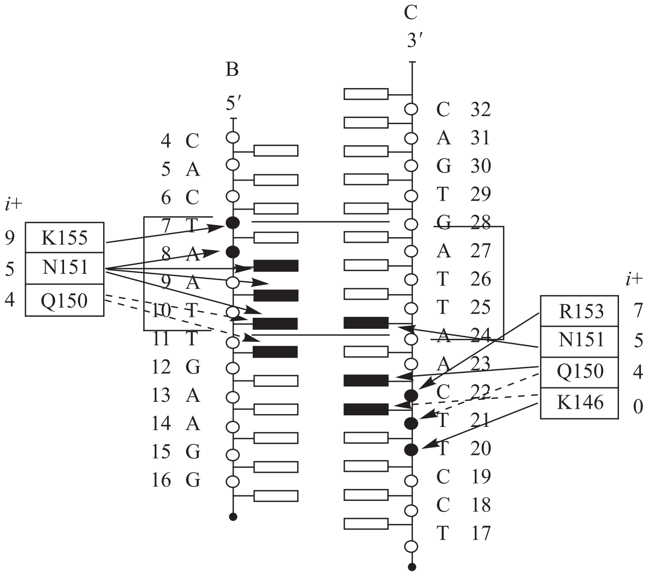

Fig. 5. Contact system between protein and DNA in the region of the major groove of DNA for the complex with homeodomain Msx-1 (PDB code 1IG7). Binding atomic groups of bases and phosphates are shown in black color. Large rectangle shows the invariant promoter sequence part of DNA. Small rectangles contain names of binding protein polar residues. Filled arrow lines designate direct intermolecular contacts. Dotted arrow lines designate structural water mediated contacts. Symbol $i+$ shows binding polar residues positions along the recognizing helix

on the α-helical surface are 0, 7 and 9. The second group represents contacts of the bases with noncharged residues which can be considered as *contacts specific by DNA sequence*. These are Gln150 and Asn151, whose positions are 4 and 5, and in addition also Lys146, position 0. Note that Asn151 is contacting bases of the forward strand of the canonical TAAT motive of promoter region, while Gln150 — the reverse strand outside the motive.

For all other complexes the systems of intermolecular polar contacts appear to be quite similar (Table 2). Overall, contacts of the proteins are distributed almost equally between DNA phosphates and DNA bases. However, a significant regularity was disclosed when we considered the contacts of protein polar residues in definite positions of amino acids in the primary structure (Table 3). The positive charged residues in positions 7 and 9 are bound mainly with phosphates. Oppositely, the residues of the N-terminal part of the helix in positions 0, 4, and 5 are bound mainly with the bases. This distinctive regularity allows one to discriminate groups of contacts according to their probable functions. It can

Table 2. Intermolecular contacts with DNA of invariant positions of polar residues for recognizing α-helix of factor transcriptions

Residue position $i+$	Polar residues	Complexes of factor with DNA, PDB code				
		1IG7-A	9ANT-A	3HDD-A	1FJL-A	1AU7-A2
9	Lys (Arg)	1P -	- -	1P -	2P -	- -
7	Arg	2P 3B	2P -	2P -	1P -	1P -
5	Asn	- 1B	1P 2B	1P 2B	1P 4B	1P 2B
4	Gln	1P 3B	- -	- 3B	1P 2B	- -
0	Lys (Arg)	1P 1B	- -	- 1B	- -	1P 2B
Note. Contacts with phosphates and bases of nucleotides designate as P and B. Corresponding number is equal to amount of contacts						

Table 3. Average amount of factor contacts with DNA per each type of polar residues

Residue position $i+$	Polar binding residue	Amount of contacts	
		Phosphates	Bases
9	Lys (Arg)	0.8	0
7	Arg	2.0	0.6
5	Asn	0.8	2.2
4	Gln	0.4	1.6
0	Lys (Arg)	0.4	0.8
	Total:	4.4	5.2

be also noted that the contacts are shared nearly equally between two chains of DNA. As seen in Fig. 3, the contacts in the forward and reverse DNA chains are shifted by two or three nucleotides with respect to each other.

3.3. Set of Invariant Contacts for All Considered Complexes. The contact system of complex 1IG7 is similar in some common sense to the other four complexes (Table 4). Here the conservative polar contacts are marked in bold. The complicated net of contact interactions between protein and DNA in the region of the major groove becomes much simpler when we consider only contacts *common for all complexes*. We imply a contact to be common if it occurs at least in 4 of total 5 complexes. The most significant polar contacts are summarized in Table 5. In this table, we have also included contacts of charged residue Lys146 ($i+ = 0$) though this residue does not always form an exact ionic contact. Note that the total amount of invariant contacts in all considered complexes reaches above 80%. There is also one specific contact of the thymine methyl group with the hydrophobic amino acid in position 1 (data not shown).

A general scheme of conservative contacts in the complex 1IG7 is presented in Fig. 6. It contains interatomic contacts of polar and nonpolar atomic groups, and

Table 4. Contacts between polar groups of protein and DNA in complex 1IG7-A

$i+$	Protein	Contact	DNA	DNA	Contact	Protein	$i+$
	residue		Chain B	Chain C		residue	
9	**Lys 155**	**NZ --------- O2P**	**T7**	**A24**	**N6 --- w --- OD1**	**Asn151**	**5**
2	Trp 148	NE1 --- w --- O2P	A8	A23	-	- -	-
2	**Trp 148**	**NE1 --- w --- O1P**	**A8**	**C22**	**O2P --------- NH1**	**Arg 153**	**7**
5	**Asn 151**	**ND2 --- w --- O2P**	**A8**	C22	N4 ---------- OE1	Gln 150	4
5	**Asn 151**	**ND2 --------- N7**	**A8**	**T21**	**O2P --- w --- NH1**	**Arg 153**	**7**
5	Asn 151	ND2 --- w --- O5*	A8	T21	O5* --- w --- NH1	Arg 153	7
5	**Asn 151**	**ND2 ---------- N7**	**A9**	T21	O2P --- w --- NH2	Arg 153	7
5	**Asn 151**	**OD1 ---------- N6**	**A9**	T21	O1P --- w --- NH2	Arg 153	7
5	**Asn 151**	**OD1 --- w_0 --- O4**	**T10**	**T21**	**O2P --- w --- OE1**	**Gln 150**	**4**
4	**Gln 150**	**NE2 --- w_0 --- O4**	**T10**	T21	O5* --- w --- OE1	Gln 150	4
4	Gln 150	NE2 --- w ---- O4	T11	T21	O4 --- w --- NZ	Lys 146	0
4	Gln 150	OE1 --- w ---- O4	T11	**T20**	**O2P --------- NZ**	**Lys 146**	**0**

Note. Common for all five considered complexes contacts are marked in bold. Letter w designates structural water molecule, w_0 — water molecule of central binding cluster

Table 5. Conservative contacts between polar groups of protein and DNA in complexes of double-stranded operator DNA with transcription factors

$i+$	Protein	Contact	DNA	Complex, PDB code				
	residue		nucleotide	1IG7	9ANT	3HDD	1FJL	1AU7
	Nonsignificant (–) DNA chain							
9	Lys 55	NZ --------- O2P	T	+	–	+	(+)	–
2	Trp 48	NE1 --- w --- O1P	A	+	+	+	+	+
5	Asn 51	ND2 -- w_0 --- O2P	A	+	+	+	+	–
5	Asn 51	ND2 --------- N7	A	+	+	+	+	(+)
5	Asn 51	OD1 --- w_0 --- O4	T	+	+	+	+	+
4	Gln 50	NE2 -- w --- O4	T	+	+	+	+	–
	Significant (+) DNA chain							
7	Arg 53	NH --------- O2P	T(C)	+	+	+	+	+
0	Lys 46	NZ --------- OP	T	+	–	–	–	+

Note. Signs + or – designate presence or absence of corresponding binding contact. Sign in brackets (+) marks insignificant variation in contact: for instance, exchange of Lys by Arg in complex 1FJL or exchange of atom N7 by N6 in complex 1AU7. Total amount of presence in the complexes of all listed conservative contacts is equal to 82.5%

the binding protein residues and DNA bases are shown in grey. Protein molecules have 4 binding polar and 2 nonpolar residues on the surface of the recognizing helix in positions 1, 4, 5, 7, 8 and 9. A DNA molecule has three binding nucleotides: T7 and T10 from one chain and C22 from the other chain. The third essential participant of the complex formation is the set of 10–20 molecules of structural water. It should be noted that this scheme can be considered as common for all five considered complexes with the unique pattern of polar binding residues on the surface of the recognizing α-helix.

3.4. Functional Sense of Protein-DNA Contacts. First we select a large group of contacts — *the central binding cluster*. The binding atomic groups of protein and DNA are shown here by grey color (Fig. 7, *a*). This cluster consists of the central molecule w_0 of structural water mediating the binding contacts of two residues Gln150 and Asn151 with the base T10. This nucleotide is a part of the promoter sequence TAAT. For convenience we have used here the designation taken from the complex 1IG7. However, we consider only the common contact system. In fact, the mentioned residues can also form some other contacts. It is worth noting that in these residues both side groups, NH_2 and CO, are forming contacts with DNA. Another distinct feature is that almost all valences, three of total four, of structural water molecule are occupied.

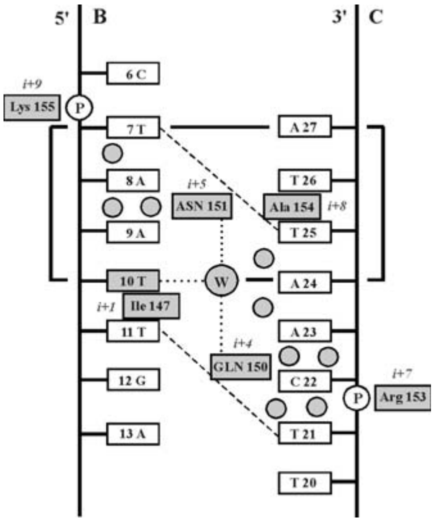

Fig. 6. Contact system between protein and DNA common for all considered factors in the region of the major groove of DNA. Large rectangle shows the invariant promoter sequence part of DNA. The atomic groups forming intermolecular contacts are shown in grey color. Scheme is based on numbering of complex 1IG7

The second group includes contacts of two positively charged residues Arg153 and Lys155 with phosphate groups of nucleotides T7 and C22 (Fig. 7, *b*). As seen above, the negatively charged sugar-phosphate backbone of B-DNA has a strong shielding effect on the field of base groups situated at the bottom of the major groove of DNA. These two residues function as *positively charged compensators of two negatively charged phosphates* of different chains of DNA. And this allows forming the contacts of protein residues with less accessible bases of atomic groups at the bottom of the major groove of B-DNA.

The third group of contacts presents *nonpolar barriers of the protein-DNA binding site* from two sides along the DNA chain (Fig. 7, *c*). These are bridges in DNA molecule composed of methyl groups of closely situated bases T7...T25 and T11...T21 of different sides of B-DNA. In the complex the protein nonpolar residues Ala154 and Ile147 are located here, respectively.

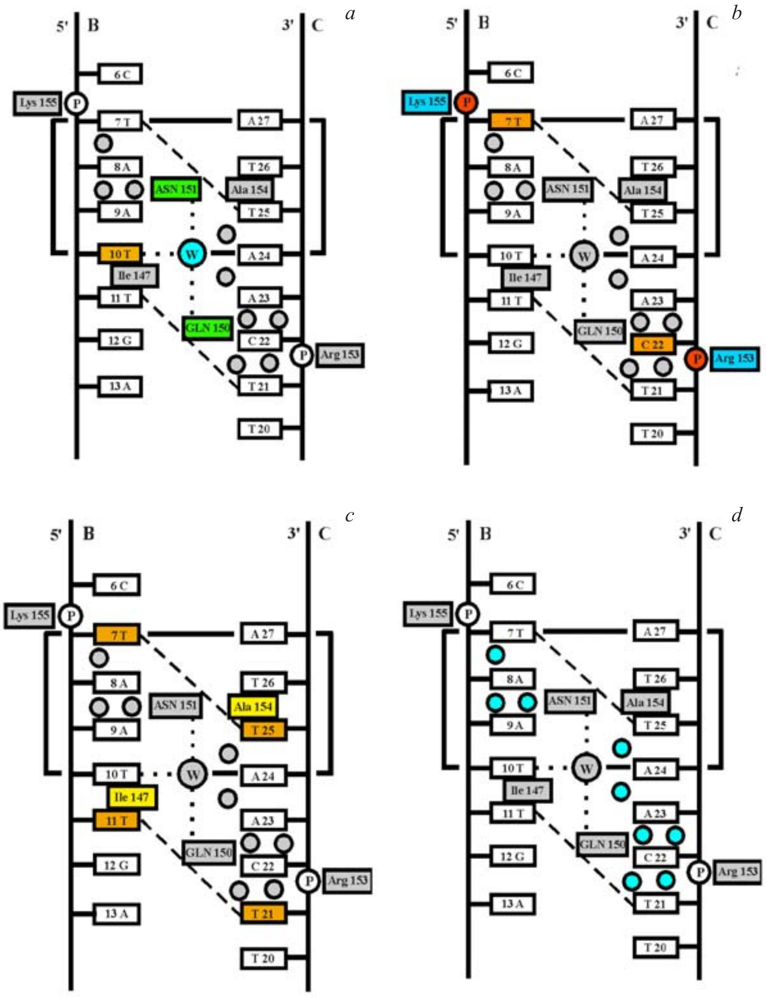

Fig. 7. Groups of homeodomain-DNA contacts with specific functional sense marked in grey: *a*) central binding contact cluster; *b*) positively charged compensators of phosphate negative charges of both DNA chains; *c*) nonpolar fixing barriers of the protein-DNA binding site; *d*) set of 10–20 water molecules providing local conformational mobility

Finally, fourth groups of functionally significant contacts consist of a set of 11–12 closely situated fixed molecules of structural water. These molecules are inside of the whole contact interface region between the protein and DNA (Fig. 7, *d*). This water cluster may provide *local conformational mobility* necessary for the dissociation process. Roles for interface water molecules were first noted in [35] and discussed in the review [36].

We can see that almost all common intermolecular contacts of the recognizing helix have distinctive functions to be performed. However, where is the binding specificity of each individual complex realized? First, even a small variation of similar amino acid residues, such as Lys and Arg, seems to be significant. Second, there is a strong difference in DNA sequences of the noncanonical parts outside the promoter motive TAAT (see Fig. 3). It is this part of DNA that forms contacts with the protein and thus determines some specific details of the interface. Therefore, we can suggest that *DNA sequence immediately following the TAAT motive and located in this motive's complementary strand is mainly responsible for the specificity of binding within the considered group of complexes*. Those nucleotides form contacts with amino acids in positions 4 and 8 which are essentially variable, which is valid also for the entire homeodomain family (data not shown).

This study is an attempt to consider contacts of DNA bases with protein side chains within an entire interface. To analyze the interfaces in the set of five proteins together with the 3D structure of entire proteins, we applied a manual structural alignment. The manual alignment used herein combines comparisons of individual contacts in the interfaces with comparisons of the overall fold of the protein domains, thus combining the features of chemical and stereochemical recognition as specified in [37], or the description of contacts with structural relationships as specified in [15]. In this context it may be useful to recall that the structure of the complex only reflects the eventual state of protein-DNA recognition, the recognition being a dynamical process. While strictly defined contacts may determine specificity of binding in the eventual complex, and operating similarly to the «lock-and-key» principle, the contacts that do not satisfy the stereochemical criteria, may be termed «loose contacts». The latter provide the degrees of freedom in the specific docking of protein to DNA, playing the role of dumpers and docking ropes, attenuating the diffusional and collisional mobility of the groups involved in specific contacts, and thus facilitating the specific docking.

CONCLUSION

The comparative analysis of contact interactions of five complexes of the homeodomain transcription factors with fragments of operator DNA has shown that, despite rather low pair homology of the binding helix and the recognized DNA sequence, the regularities relating structural features to function could be very clearly revealed in all considered complexes. Only five examples turned out to be sufficient for identification of both the general and the specific features of the protein-DNA recognition, due to the evolutionary conservation and the rigidity of homeodomain family. We suggest that the found regularities should be valid for the protein-DNA complexes of the entire homeodomain family as we revealed by considering the expanded set of complexes presented in the review [18]. While the invariant contacts likely specify the family of homeodomains, the variable contacts provide specificity of individual complexes within the homeodomain family.

Acknowledgements. The authors thank Prof. A. V. Finkelstein and Prof. A. V. Efimov for useful discussion. This work was supported by the Russian Foundation for Basic Research, Project 07-07-00234.

REFERENCES

1. *Luscombe N. M., Laskowski R. A., Thornton J. M.* // Nucleic Acids Res. 1997. V. 25. P. 4940–4945.

2. *Luscombe N. M., Laskowski R. A., Thornton J. M.* // Nucleic Acids Res. 2001. V. 29. P. 2860–2874.

3. *Jones S. et al.* // J. Mol. Biol. 1999. V. 287. P. 877–896.

4. *Jones S. et al.* // Nucleic Acids Res. 2003. V. 31. P. 7189–7198.

5. *Shanahan H. P. et al.* // Nucleic Acids Res. 2004. V. 32. P. 4732–4741.

6. *Tsuchiya Y., Kinoshita K., Nakamura H.* // Proteins. 2004. V. 55. P. 885–894.

7. *Liu Z. et al.* // Nucleic Acids Res. 2005. V. 33. P. 546–558.

8. *Szilagyi A., Skolnick J.* // J. Mol. Biol. 2006. V. 358. P. 922–933.

9. *Mandel-Gutfreund Y., Schueler O., Margalit H.* // J. Mol. Biol. 1995. V. 253. P. 370–382.

10. *Chirgadze Y. N., Tabolina O. Y.* // Protein Eng. 1996. V. 9. P. 745–754.

11. *Chirgadze Y. N., Larionoiva E. A.* // Protein Eng. 1999. V. 12. P. 101–105.

12. *Chirgadze Y. N., Larionoiva E. A.* // Mol. Biol. (Moscow). 2001. V. 35. P. 709–716.

13. *Chirgadze Y. N., Larionoiva E. A.* // Mol. Biol. (Moscow). 2003. V. 37. P. 232–239.

14. *Chirgadze Y. N., Larionoiva E. A.* // Mol. Biol. (Moscow). 2005. V. 39. P. 892–905.

15. *Pabo C. O., Nekludova L.* // J. Mol. Biol. 2000. V. 301. P. 597–624.

16. *Cappen C.* Homeobox Gene Repertoires: Implications for the Evolution of Diversity / Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. Part 1.1. Genetic Variation and Evolution. (2005) DOI: 10.1002/047001153X.g101209, http://mrw.interscience.wiley.com/ggpb/articles/g101209/frame.html

17. *Gehring W. J. et al.* // Cell. 1994. V. 78. P. 211–223.

18. *Ledneva R. K. et al.* // Mol. Biol. (Moscow). 2001. V. 35. P. 764–777.

19. *Murzin A. G. et al.* // J. Mol. Biol. 1995. V. 247. P. 536–540.

20. *Berman H. M. et al.* // Nucleic Acids Res. 2000. V. 28. P. 235–242.

21. *Fraenkel E. et al.* // J. Mol. Biol. 1998. V. 284. P. 351–361.

22. *Clarke N. D. et al.* // Protein Sci. 1994. V. 3. P. 1779–1787.

23. *Zhao X., Huang X. R., Sun C. C.* // J. Struct. Biol. 2006. V. 155. P. 426–437.

24. *Jacobson E. M. et al.* // Genes Dev. 1997. V. 11. P. 198–212.

25. *Hovde S., Abate-Shen C., Geiger J. H.* // Biochemistry. 2001. V. 40. P. 12013–12021.

26. *Fraenkel E., Pabo C. O.* // Nat. Struct. Biol. 1998. V. 5. P. 692–697.

27. *Wilson D. S. et al.* // Cell. 1995. V. 82. P. 709–719.

28. *Ohlendorf D. H., Matthew J. B.* // Adv. Biophys. 1985. V. 20. P. 137–151.

29. *Berman H. M. et al.* // Biophys. J. 1992. V. 63. P. 751–759.

30. *Polozov R. V. et al.* // Biochemistry. 2006. V. 45. P. 4481–4490.

31. *Wang J. et al.* // J. Comput. Chem. 2004. V. 25. P. 1157–1174.

32. *Netz R. R., Orland H.* // Eur. Phys. J. Soft Matter. E. 2003. V. 11. P. 301–311.

33. *Felder C. E. et al.* // Nucleic Acids Res. 2007. V. 35. P. W512–W521.

34. *Fogolari F. et al.* // J. Mol. Biol. 1997. V. 267. P. 368–381.

35. *Billeter M. et al.* // J. Mol. Biol. 1993. V. 234. P. 1084–1093.

36. *Schwabe J. W.* // Curr. Opin. Struct. Biol. 1997. V. 7. P. 126–134.

37. *Suzuki M., Yagi N.* // Proc. Natl. Acad. Sci. USA. 1994. V. 91. P. 12357–12361.

Received on November 28, 2008.

Редактор *В. В. Булатова*

Подписано в печать 13.04.2009.
Формат 60 × 90/16. Бумага офсетная. Печать офсетная.
Усл. печ. л. 1,37. Уч.-изд. л. 1,98. Тираж 220 экз. Заказ № 56561.

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