JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

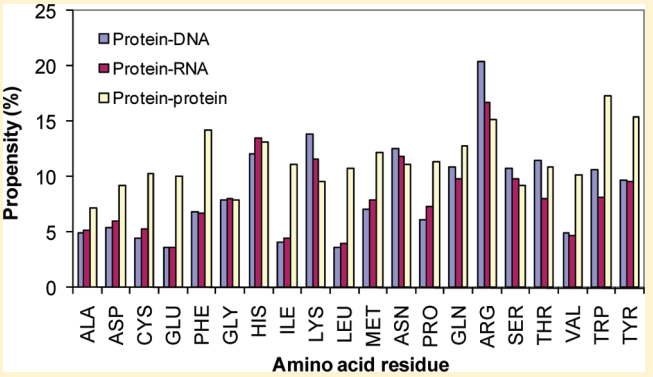
Scoring Function Based Approach for Locating Binding Sites and Understanding Recognition Mechanism of Protein−DNA Complexes

M. Michael Gromiha*,†,‡ and Kazuhiko Fukui‡

†Department of Biotechnology, Indian Institute of Technology Madras, Chennai 600 036, Tamilnadu, India

‡Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan

ABSTRACT: Protein−DNA recognition plays an essential role in the regulation of gene expression. Understanding the recognition mechanism of protein−DNA complexes is a challenging task in molecular and computational biology. In this work, a scoring function based approach has been developed for identifying the binding sites and delineating the important residues for binding in protein−DNA complexes. This approach considers both the repulsive interactions and the effect of distance between atoms in protein and DNA. The results showed that positively charged, polar, and aromatic residues are important for binding. These residues influence the formation of electrostatic, hydrogen bonding, and stacking interactions. Our observation has been verified with experimental binding specificity of protein−DNA complexes and found to be in good agreement with experiments. The comparison of protein−RNA and protein−DNA complexes reveals that the contribution of phosphate atoms in DNA is twice as large as in protein−RNA complexes. Furthermore, we observed that the positively charged, polar, and aromatic residues serve as hotspot residues in protein−RNA complexes, whereas other residues also altered the binding specificity in protein−DNA complexes. Based on the results obtained in the present study and related reports, a plausible mechanism has been proposed for the recognition of protein−DNA complexes.



INTRODUCTION

Protein−DNA interactions play an important role in many vital processes, such as the regulation of gene expression, DNA replication and repair, transcription, and packaging. Understanding the specificity with which proteins recognize target DNA sequences is of considerable theoretical and practical importance, and its basis has been demonstrated through structural analysis of protein−DNA complexes. The availability of three-dimensional structures of protein−DNA complexes in the Protein Data Bank[1] encouraged researchers to analyze the important features for protein−DNA recognition. The investigations have been focused on different perspectives, such as physicochemical properties, conservation of amino acid residues, contribution of noncovalent interactions, and conformational changes of DNA.[2−18]

The structural analysis of protein−DNA complexes based on polarity, size, shape, and packing showed that the binding sites have common features to form direct and water-mediated hydrogen bonds.[10,19] These binding site residues are more conserved than nonbinding residues, and putative hotspots occur as clusters of conserved residues.[14,20] On the other hand, the thermodynamics of binding have been analyzed with the computation of noncovalent interactions, such as electrostatic, hydrogen bonds, hydrophobic, and van der Waals interactions.[11,13,16,21,22] In addition, the influence of cation−π interactions has also been reported.[23] The contributions of energetic terms along with physical and chemical features have been used to understand the recognition mechanism of protein−DNA complexes. Furthermore, knowledge-based statistical potentials have been derived using atomic contacts between protein and DNA, and the potentials have been used to predict the binding specificity of protein−DNA complexes.[24,25]

The conformational change of DNA (also known as intramolecular interactions or indirect readout) is also reported to be important for protein−DNA recognition. It accounts for structural rearrangements of DNA, which have been evaluated mainly with the average and deviations of six base step parameters (shift, slide, rise, tilt, roll, and twist) of DNA upon complex formation.[4,12,26] The relative importance of the two major perspectives on specificity of protein−DNA recognition has been reported with the investigations on: (i) the preference of amino acid residues/nucleotides to form direct contacts between protein and DNA[24,27] as well as (ii) the sequence-dependent

Received: September 21, 2010

dx.doi.org/10.1021/ci1003703 | J. Chem. Inf. Model. XXXX, XXX, 000–000

Table 1. Protein Data Bank codes of 212 Protein−DNA Complexes Used in the Present Study[a]

1a0a_A_CD	1ign_A_CD	1r2z_A_BC	2ac0_A_EFGH	2o8b_AB_EF
1a3q_A_CD	1j1v_A_BC	1r71_A_EIFJGLHK	2aor_A_CD	2oaa_A_CD
1a73_AB_CD	1j3e_A_BC	1r8d_A_D	2aq4_A_T	2odi_A_CD
1am9_AB_FG	1jb7_AB_DGH	1r8e_A_B	2bgw_A_C	2ofi_A_CB
1b01_AB_EF	1je8_A_C	1rep_C_AB	2bnw_A_F	2owo_A_BCD
1b3t_A_CD	1jey_AB_CD	1rh6_A_CD	2bsq_E_I	2p0j_A_CD
1bc8_C_A	1jj4_A_C	1rrq_A_BC	2bzf_A_BC	2p5l_CD_AB
1bdt_AB_EF	1jko_C_AB	1rxw_A_BC	2c5r_A_YZ	2pyj_A_XY
1bf5_A_BC	1jnm_A_CD	1rzr_A_EB	2c7p_A_CD	2qhb_A_EF
1bl0_A_BC	1jt0_C_EF	1sa3_A_CD	2c9l_Z_AB	2qnf_AB_CD
1cez_A_T	1jx4_A_PT	1skn_P_AB	2dnj_A_BC	2qsh_A_WY
1cf7_AB_CD	1k3x_A_BC	1sx5_A_CDEF	2dp6_A_CD	2r1j_L_BA
1ckt_A_B	1k4t_A_CD	1sxq_A_CE	2dpi_A_PT	2r9l_A_CD
1cl8_A_B	1k61_A_EF	1t05_A_TP	2drp_A_BC	2rba_A_CD
1cw0_A_MO	1ku7_A_BC	1t9i_A_CD	2dtu_A_EF	2rbf_AB_CD
1d02_AB_CD	1kx5_D_IJ	1tc3_C_AB	2e1c_A_BD	2rgr_A_CD
1dc1_A_WC	1l3l_A_FH	1tez_A_IJ	2e52_AB_EGIJ	2vjv_A_CE
1dct_A_FMGN	1llm_C_AB	1tro_ACEG_IJKL	2er8_A_EF	2vla_A_LM
1dew_A_UV	1lmb_3_12	1u3e_M_ABC	2ex5_A_XY	2yvh_A_EF
1dfm_A_CD	1lq1_D_EF	1u8b_A_BCDE	2ezv_A_FG	2z3x_A_DE
1diz_A_EF	1m3q_A_BC	1ubd_C_AB	2fcc_A_CD	2zhg_A_B
1dsz_A_CD	1mdy_B_EF	1uut_A_C	2fio_A_CD	3bam_A_CE
1e3o_C_AB	1mjo_A_F	1v15_A_EF	2fkc_A_CD	3bep_A_CD
1efa_AB_DE	1mnn_A_BC	1w0u_A_CD	2fmp_A_TD	3bkz_A_BC
1egw_A_EFGH	1mtl_A_CD	1wb9_A_E	2fr4_AB_MN	3bpy_A_BC
1emh_A_B	1mus_A_BC	1wte_A_XY	2g1p_A_FG	3brg_C_AB
1ewn_A_DE	1nkp_A_FG	1 × 9m_A_CD	2gb7_A_EF	3bs1_A_BC
1f4k_B_DE	1odh_A_CD	1 × 9n_A_CD	2gig_A_EF	3btx_A_BC
1fiu_A_EIJG	1oe4_A_EF	1xo0_B_CD	2h27_A_BC	3c0w_A_BC
1flo_A_EFJ	1omh_A_B	1xpx_A_DC	2h7g_X_YZ	3c25_A_CD
1fok_A_BC	1orn_A_BC	1xsd_A_B	2hdd_A_CD	3c2i_A_BC
1fyl_B_C	1oup_A_GF	1xyi_A_BC	2heo_D_E	3clc_AB_EF
1gd2_EF_AB	1owf_AB_CE	1y8z_A_CD	2hhv_A_BC	3clz_A_EF
1gdt_AB_CDEF	1ozj_A_CD	1yf3_AB_D	2i06_A_BC	3coq_A_DE
1gu4_A_C	1p71_A_CD	1z19_A_DE	2ih2_A_BC	3cro_LR_AB
1gxp_AB_CD	1p7h_L_ACBD	1z63_A_CD	2ihm_A_TP	3dfx_A_XY
1h6f_A_CD	1p8k_Z_ABCD	1z9c_A_GH	2ihn_A_CD	3dvo_AB_EF
1h9d_A_EF	1pp7_U_EF	1zme_C_AB	2irf_G_F	3pvi_A_CD
1hlv_A_BC	1qna_A_CD	1zrf_A_WX	2is6_AB_CD	6cro_A_RU
1hwt_CDGH_EFAB	1qpi_A_MN	1zs4_A_UT	2isz_AB_EF	6pax_A_BC
1i3j_A_BC	1qpz_A_M	1ztw_A_BG	2nq9_A_BCD	
1iaw_A_CDEF	1qrv_A_CD	1zx4_A_TS	2ntc_A_WC	
1ic8_A_EF	1qzh_A_G	2a3v_A_EFGH	2o4a_A_BC	

[a] The first four letters show the PDB code followed by chain information of protein and DNA, respectively.

conformation of DNA structure, which is recognized through protein contacts with the sugar phosphate backbone and/or with nonspecific portions of the bases.[4,12,26,28]

On the other hand, several methods have been proposed for identifying the binding sites in protein−DNA complexes. These methods are mainly based on structure-based statistical potentials as well as the features obtained from amino acid sequences.[24,29−32] The major features are amino acid composition, residue pair preference, secondary structure, solvent accessibility, amino acid properties, and evolutionary information. These features have been used as input parameters in machine learning techniques for predicting the binding sites in protein−DNA complexes.[33−38]

In most of these studies, binding sites have been defined with a criteria based on the contacts between amino acid residues in protein and nucleotides in DNA.[29,34−36,38] The contacts between any atoms in a protein and DNA with a cutoff distance (between 3 and 6 Å) have been used to assign the binding sites. These criteria include the repulsive interactions in which a residue and a nucleotide are close to each other. In addition, the residue−nucleotide pairs with different distances have been treated in the same manner. In this work, a scoring function based approach has been developed for defining the binding sites. The results showed that the binding sites are dominated with positively charged, polar, and aromatic residues indicating the

importance of electrostatic, hydrogen bonding, and aromatic interactions. Furthermore, the preference of interacting partners and of residues in binding segments of different lengths have been analyzed. Based on the results obtained in the present study, a plausible mechanism for understanding the recognition of protein−DNA complexes has been discussed.

MATERIALS AND METHODS

Data Set. Recently, Xu et al.[32] developed a data set of 212 protein−DNA complex structures with two conditions: (i) the structures were solved at less than 3 Å resolution and (ii) no two protein sequences have the sequence identity of more than 35%. However, they have not considered the redundancy occurred among the chains within each protein−DNA complex. The careful survey of the considered protein−DNA complexes showed the presence of duplicate (redundant) chains (protein and DNA sequences) in several of the complexes. These internal redundancies have been removed, and a refined data set of 212 protein−DNA complexes has been constructed. Essentially, the number of protein−DNA complexes is the same, and there are no redundant chains within each complex. The PDB codes along with chain information for protein and DNA are listed in Table 1. Further, nonredundant data sets of 153 protein−protein complexes and 81 protein−RNA complexes have been used for comparative analysis. The development of these data sets has been explained in our earlier articles.[39−41]

Scoring Function for the Interaction between Protein and DNA. The $Score_{energy}$ for the interaction between protein and DNA has been computed using AMBER potential,[42] which is widely used to analyze and understand the recognition mechanism in protein−nucleic acid complexes.[43] It is given by

$$Score_{energy} = \sum[(A_{ij}/r_{ij}^{12} - B_{ij}/r_{ij}^{6}) + q_i q_j/r_{ij}] \quad (1)$$

where $A_{ij} = \varepsilon_{ij}^*(R_{ij}^*)^{12}$ and $B_{ij} = 2\,\varepsilon_{ij}^*(R_{ij}^*)^6$; $R_{ij}^* = (R_i^* + R_j^*)$ and $\varepsilon_{ij}^* = (\varepsilon_i^*\,\varepsilon_j^*)^{1/2}$; R^* and ε^* are, respectively, the van der Waals radius and well depth, and these parameters are obtained from Cornell et al.;[42] q_i and q_j are, respectively, the charges for the atoms i and j; and r_{ij} is the distance between them. The distant-dependent dielectric constant ($\varepsilon = r_{ij}$) has been used to take account of the dielectric damping effect of the Coulomb interactions, as used in other studies on protein−nucleic acid complexes.[43] The $Score_{energy}$ of all the amino acid residues in the considered protein−DNA complexes have been computed, and the residues, which have the $Score_{energy}$ of less than −1 kcal/mol, are identified as binding site residues.

Binding Propensity. The binding propensity for the 20 amino acid residues and 4 nucleotides in protein−DNA complexes has been developed as follows: the frequency of occurrence of amino acid residues (nucleotides) in binding sites (f_b) and in protein (DNA) as a whole (f_t) has been computed. The binding propensity (P_{bind}) is calculated using the equation:[41]

$$P_{bind}(i) = f_b(i)/f_t(i) \quad (2)$$

where, i represents each of the 20 amino acid residues and 4 nucleotides.

Binding Segments. The residues identified as binding sites have been analyzed in terms of binding segments. It is based on the number of consecutive binding residues in amino acid sequences. For example, a four residue binding segment has a stretch of four consecutive binding residues. The behavior of 20 amino residues and 4 nucleotides in binding segments has been analyzed with 1, 2, 3, 4 and more than 4 residues/nucleotides.

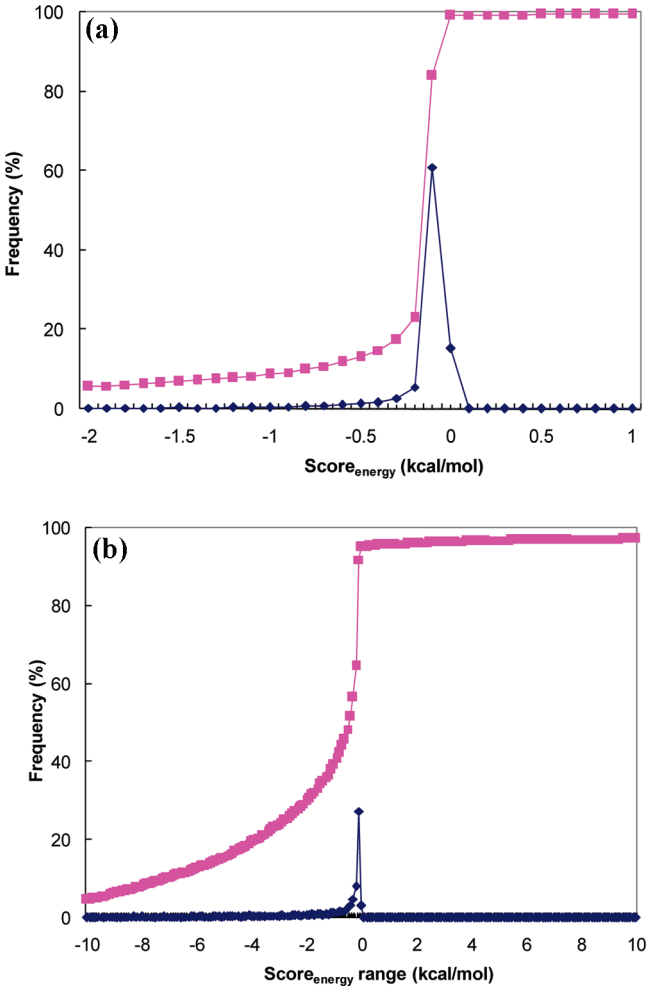


Figure 1. Occurrence of amino acid residues in different ranges of $Score_{energy}$: (a) proteins and (b) DNAs. The diamonds and squares show the fraction and total percentage of residues.

Residue-Pair Preference. The preference of amino acid residues/nucleotides for the interactions between protein and DNA is computed using the equation:[39,44]

$$Pair(i,j) = \sum N_{ij}/(\sum N_i + \sum N_j) \quad (3)$$

where i and j stands for the interacting residues and nucleotides in proteins and DNAs, respectively. $N_{i,j}$ is the number of interacting residues of type i in proteins and j in DNAs. ΣN_i and ΣN_j are the total number of residues of type i and j, respectively, in proteins and DNAs.

RESULTS AND DISCUSSION

Occurrence of Amino Acid Residues at Various Ranges of $Score_{energy}$. In a protein−DNA complex, the $Score_{energy}$ (eq 1) between all pairs of atoms between protein and DNA has been computed. The calculations were repeated for all the complexes, and the $Score_{energy}$ for all the residues at intervals of 0.1 from −15 to 5 kcal/mol have been analyzed. The frequency of occurrence of residues in proteins at different intervals of $Score_{energy}$ (from

dx.doi.org/10.1021/ci1003703 |J. Chem. Inf. Model. XXXX, XXX, 000–000

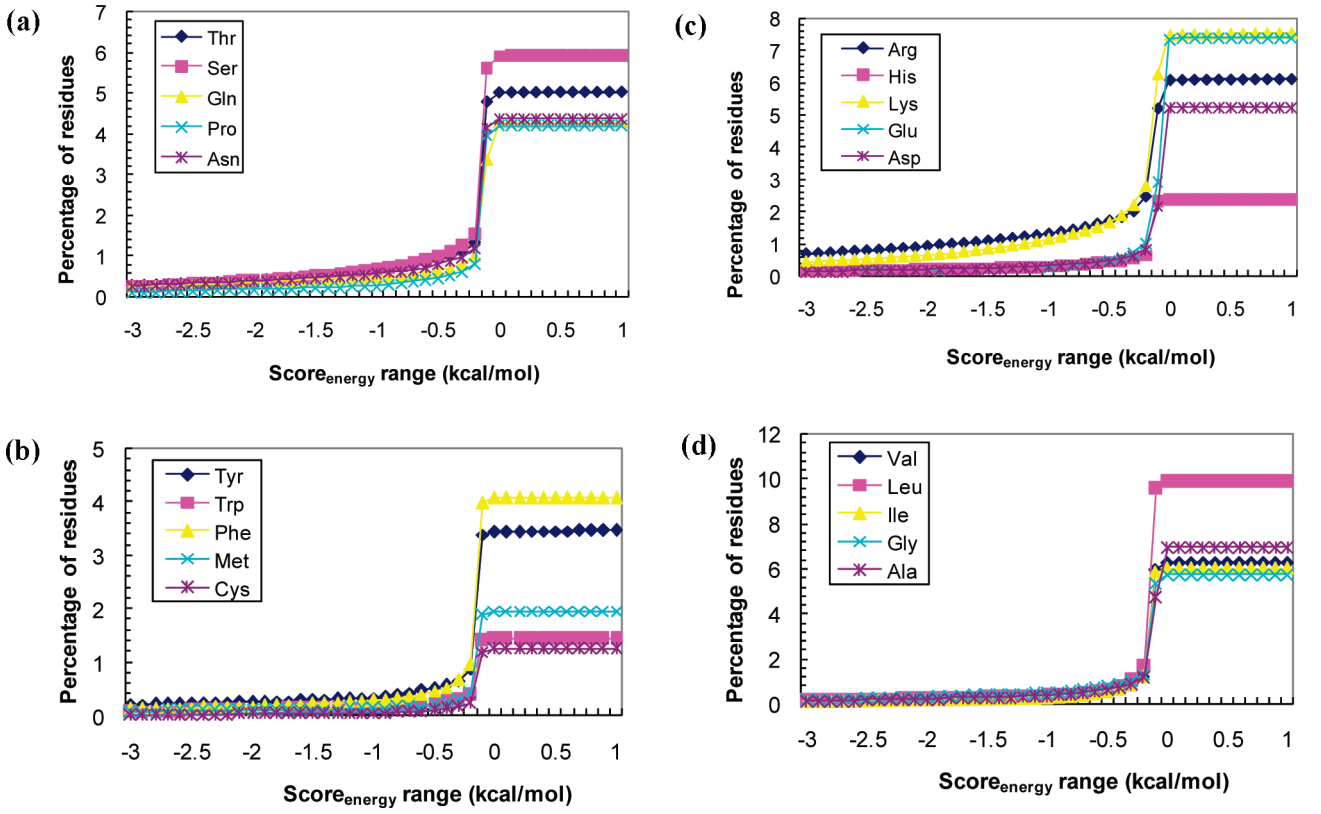


Figure 2. Contribution of the 20 amino acid residues at different $Score_{energy}$ ranges: (a) polar, (b) aromatic and sulfur containing, (c) charged, and (d) hydrophobic.

−2 to 1 kcal/mol) is displayed in Figure 1a. This figure includes the results for both the fraction of residues and the total percentage of residues at each interval. The results showed that 5.7% of the residues have strong interactions with DNA ($Score_{energy}$ < −2 kcal/mol), which is similar to protein−RNA complexes[41] and less than that in protein−protein complexes (7.7%).[39] This observation indicates that the characteristic features of binding are different for protein−protein and protein−nucleic acid complexes. On the other hand, a common behavior observed among all the complexes is that about 70% of the residues have a $Score_{energy}$ in the range of −0.3 to 0 kcal/mol, which might be due to the presence of residues that are far away in three-dimensional (3D) structures in all complexes.[45] Furthermore, 16% of the residues exhibit repulsion energies with DNA, whereas only 6% of the residues showed repulsive energies in protein−protein complexes.[39] Among 52 881 residues 4603 of them have the $Score_{energy}$ less than −1 kcal/mol, and hence 8.7% of residues are identified as binding sites in protein−DNA complexes. Interestingly, protein−DNA complexes have more binding sites than protein−RNA complexes, although the size of DNA (33 nucleotides/complex, on average) is less than RNA (38 nucleotides/complex, on average). It shows that the binding mode of DNA with proteins is widely shared with several amino acid residues.

The binding sites identified with scoring function based methods have been compared with other distance based methods. The distance based methods showed the presence of 6−20% binding site residues with the cutoff of 3.5−6.0 Å.[29,34−36,38] In order to make a fair comparison the distance between amino acid residues and nucleotides has been computed, and the binding site residues are identified with a cutoff of 3.5 Å using the same data set of 212 protein−DNA complexes. This procedure identified 7.9% of residues as binding sites. Further, several binding site residues identified with scoring function based approachs are nonbinding in distance based criteria and vice versa.

Figure 1b shows the $Score_{energy}$ profile for the nucleotides in DNA. It shows that 30% of the nucleotides have the $Score_{energy}$ < −2 kcal/mol, similar to RNA.[41] This result indicates that DNAs prefer to wind the protein and that the nucleotides in DNAs strongly interact with amino acid residues in proteins. Interestingly, only 40% of the nucleotides have the $Score_{energy}$ between −0.3 and 0 kcal/mol, whereas 71 and 45% do in proteins[39] and RNA,[41] respectively. Further, the number of nucleotides that have repulsive energy is less than that of amino acid residues.

Behavior of Different Types of Amino Acid Residues. The behavior of amino acid residues for binding with DNA has been analyzed at different $Score_{energy}$ intervals using their percentage contributions, and the results for four groups of residues [(i) polar, (ii) aromatic and sulfur containing, (iii) charged, and (iv) hydrophobic] are shown in Figure 2. The contribution of polar residues depend on their occurrence in protein−DNA complexes (Figure 2a). This trend differs from protein−RNA complexes in that all the polar residues contribute similar levels to binding with RNA, irrespective of their occurrences.[41] Although the occurrence of Tyr is less frequent than Phe it has higher preference than Phe to bind with DNA (Figure 2b). The positively charged residues are predisposed to interact with DNA as revealed from the contribution of Arg and Lys in Figure 2c. This tendency is similar to protein−RNA complexes.[41] Although the occurrence of His is less frequent than Asp and Glu, all three residues show a similar level of contribution to interact with DNA. This tendency is due to the

dx.doi.org/10.1021/ci1003703 |J. Chem. Inf. Model. XXXX, XXX, 000–000

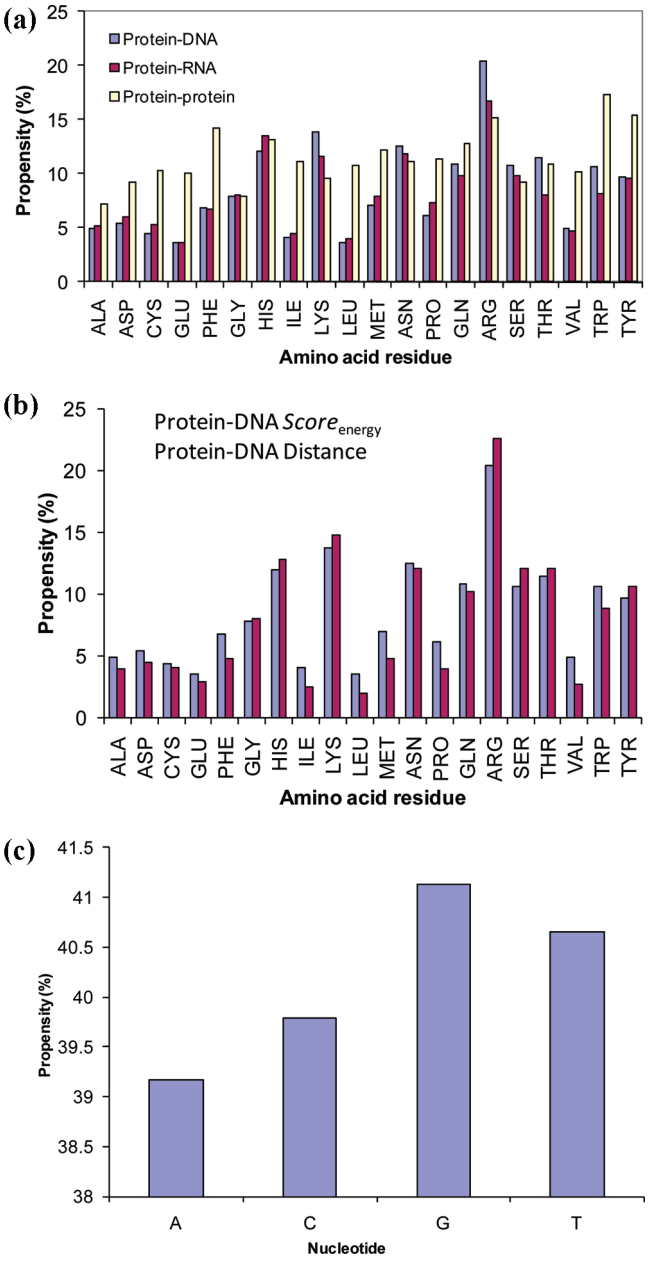


Figure 3. Binding propensity of amino acid residues in protein−DNA complexes: (a) comparison with protein−protein and protein−RNA complexes, (b) comparison between $Score_{energy}$ and distance based methods in protein−DNA complexes and (c) binding propensity of nucleotides.

interaction of positively charged residues with the negative-charged phosphate atom in the main chain of DNA via the electrostatic interactions and the bases of nucleotides through cation-π interactions. The behavior of charged residues in protein−protein and protein−DNA complexes is different in that in protein−DNA interactions the contribution of positively charged residues is dominant, whereas in protein−protein interactions both positively and negatively charged residues contribute similarly to each other.[39] The contribution of Leu is higher than other hydrophobic residues in protein−protein complexes,[39] whereas in protein−DNA complexes all the hydrophobic residues showed similar levels of contribution to interact with DNA (Figure 2d).

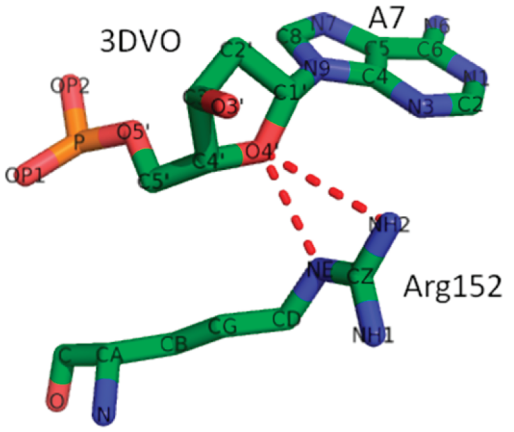


Figure 4. Electrostatic interaction between Arg152 and A7 in SgrAI protein−DNA complex.

Binding Propensity of Residues and Nucleotides in Protein−DNA Complexes. The binding propensity of residues in proteins and nucleotides in DNAs has been computed, and the results obtained for proteins are presented in Figure 3a. This figure also includes the results obtained for protein−protein and protein−RNA complexes. The binding propensities of aromatic, negatively charged, sulfur containing, and hydrophobic residues are remarkably higher in protein−protein complexes than protein−DNA complexes. In protein−DNA complexes the positively charged residues dominate interactions with DNA, and appreciable contribution is observed for polar and aromatic residues. This tendency indicates the dominance of electrostatic interactions, hydrogen bonds, and aromatic interactions for binding. Although several similarities between protein−DNA and protein−RNA complexes were observed, the binding propensity of Lys, Arg, Thr, Ser, Gln, and Asn is high in protein−DNA complexes. Interestingly, these amino acid residues belong to positively charged and polar groups, which form electrostatic interactions/hydrogen bonds with DNA. These results agree well with previous observations reported on structural analysis of protein−DNA complexes.[13,46] Figure 4 shows the electrostatic interaction between the side chain of Arg152 and the main chain of A7 in SgrAI protein with cognate DNA (3DVO) in which the $Score_{energy}$ is −10.4 kcal/mol.

The comparison of binding site residues in protein−DNA complexes obtained with distance and scoring function based approaches reveals the similarities and differences between them (Figure 3b). Positively charged residues are over represented in the distance based method. The negatively charged residues Glu and Asp as well as other polar residues have similar behavior in scoring function and distance based methods. The propensities of Trp and Phe are higher in the scoring function based method than in the distance based method, whereas an opposite trend was observed for Tyr. Hydrophobic residues are over represented in the scoring function based approach.

The binding propensity of nucleotides in protein−DNA complexes is presented in Figure 3c. We observed that all the nucleotides have a propensity in the range of 39−41% to interact with the amino acid residues in proteins. This result shows that there is no specific preference among the bases of DNA. However Guanine has higher preference than other nucleotides, which is opposite to that observed in protein−RNA complexes.[13,41,46]

Binding Segments in Protein−DNA Complexes. The binding segments are defined as continuous stretches of binding

dx.doi.org/10.1021/ci1003703 |J. Chem. Inf. Model. XXXX, XXX, 000–000

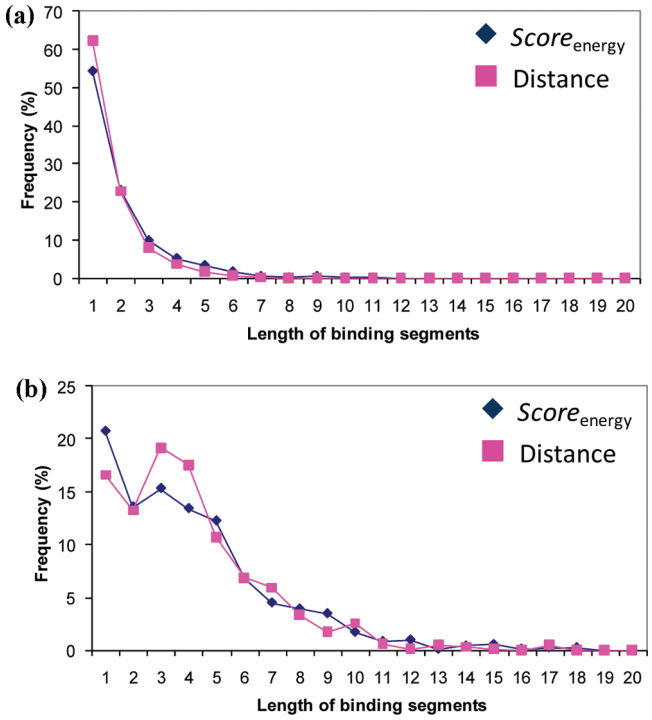


Figure 5. Frequency of occurrence of residues at various lengths of binding segments in (a) proteins and (b) DNAs. The results obtained with distance and $Score_{energy}$ based criteria are shown.

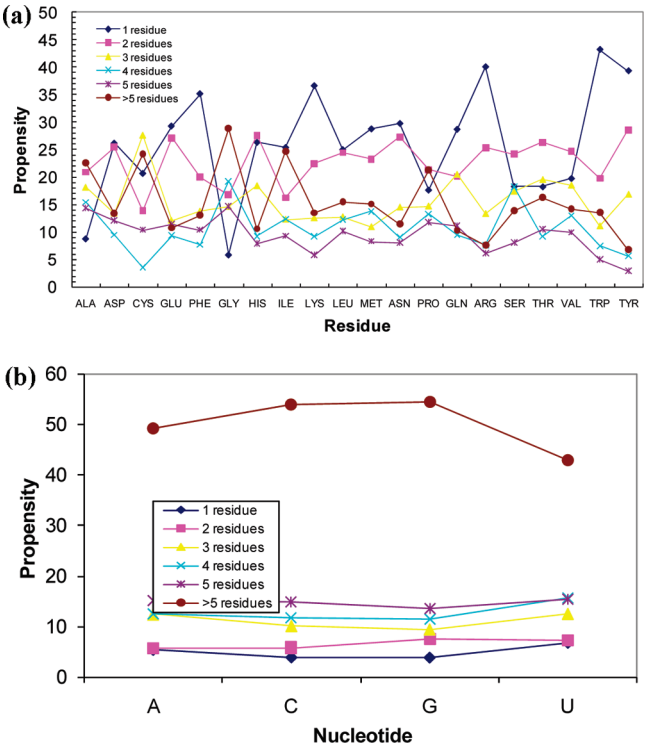


Figure 6. Propensity of binding residues in different binding segments in (a) proteins and (b) DNA.

residues in protein/nucleic acid sequences. The frequency of occurrence of binding segments in proteins and DNAs is presented in Figure 5. Figure 5a shows that 54% of the binding segments is accommodated with single amino acid residues, indicating that the neighboring residues are nonbinding. A similar trend is also observed for the binding site residues identified with the distance based criteria. Two and three-residue segments have the occurrence of 23 and 10%, respectively. In DNA, the interactions of nucleotides with amino acid residues are influenced by a stretch of one to five nucleotides. This trend is similar to the binding site residues obtained with the distance based criteria. Figure 5b shows that the percentage of binding segments accommodated with 1−5 nucleotides are, respectively, 21, 14, 15, 13, and 12%. This result reveals that a stretch of nucleotides in DNAs prefers to interact with amino acid residues in proteins. Similar behavior is also noticed in protein−RNA complexes.[41]

Propensity of Residues/Nucleotides in Different Binding Segments. The influence of amino acid residues in different binding segments has been further analyzed, and the results obtained for proteins are displayed in Figure 6a. We noticed that Lys, Arg, Phe, Tyr, and Trp have a higher tendency to be in single-residue binding segments than others. This shows the tendency of positively charged and polar residues to interact directly with the nucleotides in DNA. The polar residues Ser and Thr prefer to be in two-residue binding segments, indicating their tendency to form dual hydrogen bonds with DNA. Due to the flexibility of Gly, it accommodates all the binding segments and especially long segments.

The influence of residues in single-, two-, and multiple-residue binding segments in protein−DNA complexes is different from that of protein−RNA complexes. In protein−DNA complexes, all the aromatic and positively charged residues dominate in single-residue segments, whereas in protein−RNA complexes, Lys and Tyr showed the preference for other segments also. In addition, specific preferences are noticed for the polar residues Ser and Thr in protein−DNA complexes. On the DNA side, all the bases prefer to be in a continuous stretch of four-, five-, and more than five-nucleotide segments (Figure 6b). This implies that a stretch of nucleotides prefer to bind with protein, showing a different behavior of DNA compared with proteins.

Contribution of Different Types of Atoms in Protein−DNA Interactions. In order to understand the importance of main chain and side chain atoms for binding, the contribution of $Score_{energy}$ due to different atoms has been analyzed in the considered protein−DNA complexes. The atoms in proteins have been classified into seven groups, three for the main chain (C, N, and O) and four for the side chain (C, N, O, and S) atoms. In DNA, the atoms are grouped into P, O, and C in main chain and N, O, and C in side chain. The average $Score_{energy}$ for each atom in protein−protein, protein−RNA, and protein−DNA complexes is presented in Figure 7. The comparative analysis on main chain and side chain atoms in proteins showed that the contributions due to side chain atoms are twice that due to main chain atoms (Figure 7a). In protein−protein complexes, main chain atoms N, C, and O have similar tendency of binding, whereas in protein−DNA complexes, N and O are more preferred than C. A similar tendency is also observed for side chain atoms. This result emphasizes the importance of electrostatic and hydrogen-bonding interactions in protein−DNA complexes.

The atomic contribution of DNA is shown in Figure 7b. The contribution of phosphate is remarkably higher in DNA than RNA. Interestingly, the contribution of main chain atoms is stronger than side chain atoms in both protein−DNA and protein−RNA complexes.

dx.doi.org/10.1021/ci1003703 |J. Chem. Inf. Model. XXXX, XXX, 000–000

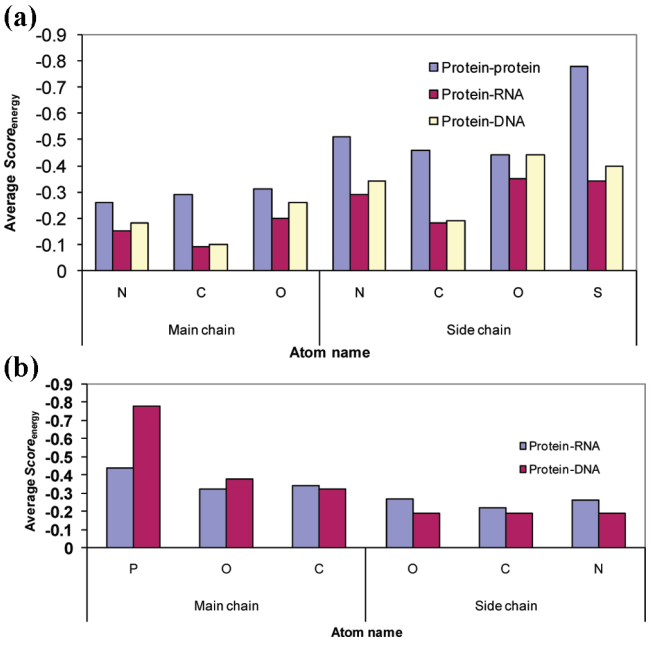


Figure 7. Contribution of main chain and side chain atoms in (a) protein and (b) DNA. The data for protein−protein and protein−RNA complexes are also shown.

Preference of Interacting Residues in Protein−DNA Complexes. The preference for interacting residues/nucleotides between proteins and DNAs has been analyzed by calculating their pair preferences at binding sites. Further, in-depth analysis on the preference of interacting atoms/residues/nucleotides in the main and side chains of proteins and DNAs reveals the importance of specific interactions between them. The main chain atoms of Ala, Gly, Ile, Lys, Leu, Gln, Asn, Arg, Ser, Thr, and Val dominate interaction with the main chain atoms of DNA, irrespective of the bases. This result shows the importance of hydrogen bonds for the interactions between proteins and DNAs. It has been reported that most of the interfacial hydrogen bonds are formed between protein and DNA backbones, which may be important for complex stability rather than specificity.[16,47] The interactions between the main chain atoms in proteins and side chain atoms in RNA are influenced by Ala-C; Cys-A, G; Gly-A, C, G, T; His-C; Ile-C; Lys-C; Met-G; Pro-A; Trp-C; and Tyr-A, C. The side chain atoms of Phe, Trp, Tyr, His, Leu, Lys, Arg, Thr, Asn, Gln, Pro, Met, and Ile interact with main chain atoms of DNA, indicating the influence of hydrogen bonds, aromatic, and electrostatic interactions. The side chain atoms of proteins and RNAs are influenced with His-A, C, G, T; Lys-A, G; Phe-A, C, G, T; Asn-A; Gln-A, G, T; Arg-A, C, G, T; Ile-A, C, G; Leu-A, G; Val-A; Met-A, C, G, T; Trp-A, C, G, T; and Tyr-A, C, G, T, showing the importance of aromatic, hydrophobic, and cation−π interactions. Our overall analysis explored the characteristic features of residues for the recognition between proteins and DNAs through electrostatic, hydrogen bonds, hydrophobic, cation−π, and aromatic interactions.

■ COMPARISON WITH EXPERIMENTS

The results obtained in this work have been compared with experimentally measured changes in binding free energy change upon amino acid substitutions. A search on the protein−nucleic acid interactions thermodynamic database, ProNIT[48] showed the presence of 405 point mutations, which caused a binding free energy of < −1 kcal/mol. A careful survey of these data yielded 244 unique mutants. The binding free energy change of these mutants has been analyzed based on the chemical behavior of wild-type and mutant amino acid residues. We observed that the majority of the mutants are due to the replacement of positively charged (83 mutants), polar (46 mutants), and aromatic (35 mutants) residues. The rest of the data is affected with hydrophobic, negatively charged, and others (pairs of residues with similar chemical behavior). This analysis demonstrates the importance of electrostatic, hydrogen bonds, cation−π, and aromatic interactions for the recognition of protein−DNA complexes with the dominance of electrostatic interactions. Our computational analysis revealed the importance of these interactions and showed good agreement with experiments.

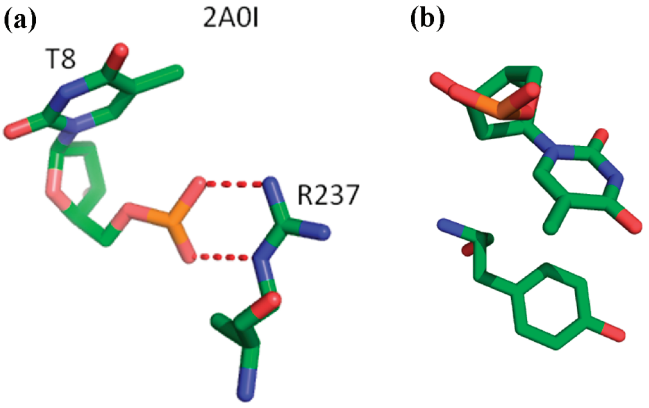


Figure 8. Interactions between amino acid residues and nucleotides in different complexes: (a) F factor relaxase−DNA and (b) telomere binding proteins Cdc13−DNA complexes.

The importance of such interactions has been verified with different specific protein−DNA complexes, and two typical examples, such as F factor relaxase−DNA and telomere binding protein Cdc13−DNA complexes, are discussed below.

I. F Factor Relaxase−DNA Complex. Larkin et al.[49] measured the binding free energies of wild-type and 26 mutants in F factor relaxase−DNA complex (2A0I). They observed that the substitutions of Ser and Arg with Ala drastically altered the binding affinity of the complex. Specifically, R237A changed the binding free energy of 4.39 kcal/mol. The analysis on the energetic contribution of Arg237 in 2A0I showed a strong electrostatic interaction with T8 in DNA in which the $Score_{energy}$ is −2.1 kcal/mol. Interestingly, the major contribution is between the side chain atoms of Arg and T. Figure 8a shows the electrostatic interactions between the residues Arg237 and T8 in F factor relaxase−DNA complex.

II. Telomere Binding Protein Cdc13−DNA Complex. Anderson et al.[50] carried out binding experiments on telomere binding protein Cdc13−DNA complex (1S40) and reported binding free energies for 19 mutants. The analysis on $\Delta\Delta G$ values shows that the affinity is decreased with the mutations of mainly aromatic residues and that the replacement of Tyr85 caused severe loss of binding. The contribution of aromatic interactions has been analyzed by computing the $Score_{energy}$ between Tyr85 in telomere binding protein Cdc13 and T4 in DNA, and the binding mode is shown in Figure 8b. We observed the presence of aromatic interactions between Tyr85 and T4, and the $Score_{energy}$ is −3.27 kcal/mol.

dx.doi.org/10.1021/ci1003703 |J. Chem. Inf. Model. XXXX, XXX, 000–000

MECHANISM FOR PROTEIN−DNA RECOGNITION

The structural analysis of protein−DNA complexes and the importance of specific amino acid residues for binding prompts us to suggest a mechanism for protein−DNA recognition. Positively charged and aromatic residues have a high binding propensity in proteins, and these residues are dominant in single residue binding segments. In addition, polar residues (Ser and Thr) showed a preference for binding in two-residue segments. On the DNA side, phosphate atoms have high tendency to interact with proteins. Hence, the contribution of electrostatic interaction is appreciable for protein−DNA recognition. This has been supported by experimental data in which the substitution of positively charged residues drastically altered the binding specificity of protein−DNA complexes.[47] Further, electrostatic interactions are reported to play a dominant role in protein−DNA recognition.[16,51,52] Based on these observations, we speculate that the recognition may be initiated with electrostatic interactions.

During the process of complex formation, the polar residues tend to make hydrogen bonds with the atoms in DNA. The present analysis also showed the possibility of forming bifurcated hydrogen bonds between protein and DNA, which was supported by crystallographic data as well as experimental binding specificity.[27,43,47] The contribution of hydrogen bonds to the specificity of protein−DNA complexes has also been stressed by other workers in the field.[10,13,53,54] The importance of aromatic and cation−π interactions is also revealed by the preference of aromatic and positively charged residues. Earlier analysis of protein−DNA complexes showed the preference of forming these interactions for protein−DNA recognition.[23] The hydrophobic interactions contribute to the stability of the complex structures along with hydrogen bonds between main chain atoms of DNA and protein.[16,47]

The above-mentioned mechanism is based on the direct interactions between protein and DNA. On the other hand, the contribution of DNA is reported to be important during the formation of the complex in terms of its flexibility and deformation. Olson et al.[5] reported that the conformational parameters of DNA also play a key role to protein−DNA recognition. This has been supported by the direct relationship between DNA stiffness and binding specificity of protein−DNA complexes.[6] Further, the relative importance of inter- and intramolecular interactions showed that the specificity depends mainly on the complex.[12] In addition, experimental data on binding specificity reveal that the binding is altered with the concentrations of ions, salt conditions, and environmental factors.

Based on the above-mentioned facts, we suggest that the recognition is mainly formed by interactions between protein and DNA, which are complimented by the conformational changes of DNA and local environments. This suggestion agrees well with the recent report on the mechanism of DNA recognition by the restriction enzyme EcoRV.[55]

The work on the thermodynamics of binding using the structures of protein−DNA complexes along with their respective free proteins and DNAs is in progress.

CONCLUSIONS

A scoring function based approach has been proposed for identifying the binding residues in protein−DNA complexes. The results showed the preference of positively charged and aromatic residues to interact with DNA. On the DNA side, the contribution of phosphate atoms is remarkably high for interaction with amino acid residues. Further analysis revealed that most of the binding segments are formed by single residues and that the two-residue segments are enriched in polar residues, indicating a preference for forming bifurcated hydrogen bonds. The results obtained in this work have been compared with the experimental binding energies of protein−DNA complexes, and a good agreement is observed between them. Based on these results, a plausible mechanism is proposed for understanding the recognition of protein−DNA complexes.

AUTHOR INFORMATION

Corresponding Author

*E-mail: gromiha@iitm.ac.in Telephone: +91-44-2257-4138.

ACKNOWLEDGMENT

We thank the reviewers for constructive comments and Dr. Paul Horton for critical reading of the manuscript. We acknowledge Prof. B. Jayaram and Dr. S. Selvaraj for useful discussions. This research was supported by Strategic International Cooperative Program, Japan Science and Technology Agency (JST).

REFERENCES

(1) Berman, H.; Henrick, K.; Nakamura, H.; Markley, J. L. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* **2007**, *35*, D301–303.

(2) Sarai, A.; Kono, H. Protein-DNA recognition patterns and predictions. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 379–398.

(3) Hogan, M. E.; Austin, R. H. Importance of DNA stiffness in protein−DNA binding specificity. *Nature* **1987**, *329*, 263–266.

(4) Gromiha, M. M.; Munteanu, M. G.; Simon, I.; Pongor, S. The role of DNA bending in Cro protein−DNA interactions. *Biophys. Chem.* **1997**, *69*, 153–160.

(5) Olson, W. K.; Gorin, A. A.; Lu, X. J.; Hock, L. M.; Zhurkin, V. B. DNA sequence-dependent deformability deduced from protein−DNA crystal complexes. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 11163–11168.

(6) Gromiha, M. M. Influence of DNA stiffness in protein-DNA recognition. *J. Biotechnol.* **2005**, *117*, 137–145.

(7) Mandel-Gutfreund, Y.; Margalit, H. Quantitative parameters for amino acid−base interaction: implications for prediction of protein−DNA binding sites. *Nucleic Acids Res.* **1998**, *26*, 2306–2312.

(8) Mandel-Gutfreund, Y.; Margalit, H.; Jernigan, R. L.; Zhurkin, V. B. A role for CH···O interactions in protein−DNA recognition. *J. Mol. Biol.* **1998**, *277*, 1129–1140.

(9) Nadassy, K.; Wodak, S. J.; Janin, J. Structural features of protein−nucleic acid recognition sites. *Biochemistry* **1999**, *38*, 1999–2017.

(10) Jones, S.; van Heyningen, P.; Berman, H. M.; Thornton, J. M. Protein−DNA interactions: a structural analysis. *J. Mol. Biol.* **1999**, *287*, 877–896.

(11) Jayaram, B.; McConnell, K.; Dixit, S. B.; Das, A.; Beveridge, D. L. Free-energy component analysis of 40 protein-DNA complexes: a consensus view on the thermodynamics of binding at the molecular level. *J. Comput. Chem.* **2002**, *23*, 1–14.

(12) Gromiha, M. M.; Siebers, J. G.; Selvaraj, S.; Kono, H.; Sarai, A. Intermolecular and intramolecular readout mechanisms in protein-DNA recognition. *J. Mol. Biol.* **2004**, *337*, 285–294.

(13) Lejeune, D.; Delsaux, N.; Charloteaux, B.; Thomas, A.; Brasseur, R. Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure. *Proteins* **2005**, *61*, 258–271.

(14) Ahmad, S.; Keskin, O.; Sarai, A.; Nussinov, R. Protein-DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins. *Nucleic Acids Res.* **2008**, *36*, 5922–5932.

(15) Rohs, R.; West, S. M.; Sosinsky, A.; Liu, P.; Mann, R. S.; Honig, B. The role of DNA shape in protein-DNA recognition. *Nature* **2009**, *461*, 1248–1253.
(16) Zhou, P.; Tian, P.; Ren, Y.; Zou, J.; Shang, Z. Systematization of the themes in protein-DNA recognition. *J. Chem. Inf. Model.* **2010**, *50*, 1476–1488.
(17) Pabo, C. O.; Nekludova, L. Geometric analysis and comparis on of protein−DNA interfaces: why is there no simple code for recognition?. *J. Mol. Biol.* **2000**, *301*, 597–624.
(18) Prabakaran, P.; Siebers, J. G.; Ahmad, S.; Gromiha, M. M.; Singarayan, M. G.; Sarai, A. Classification of protein-DNA complexes based on structural descriptors. *Structure* **2006**, *14*, 1355–1367.
(19) Reddy, C. K.; Das, A.; Jayaram, B. Do water molecules mediate protein-DNA recognition?. *J. Mol. Biol.* **2001**, *314*, 619–632.
(20) Mirny, L. A.; Gelfand, M. S. Structural analysis of conserved base pairs in protein-DNA complexes. *Nucleic Acids Res.* **2002**, *30*, 1704–1711.
(21) Oda, M.; Nakamura, H. Thermodynamic and kinetic analyses for understanding sequence-specific DNA recognition. *Genes Cells* **2000**, *5*, 319–326.
(22) Jen-Jacobson, L.; Engler, L. E.; Jacobson, L. A. Structural and thermodynamic strategies for site-specific DNA binding proteins. *Structure.* **2000**, *8*, 1015–1023.
(23) Gromiha, M. M.; Santhosh, C.; Ahmad, S. Structural analysis of cation-pi interactions in DNA binding proteins. *Int. J. Biol. Macromol.* **2004**, *34*, 203–211.
(24) Kono, H.; Sarai, A. Structure-based prediction of DNA target sites by regulatory proteins. *Proteins: Struct., Funct., Genet.* **1999**, *35*, 114–131.
(25) Donald, J. E.; Chen, W. W.; Shakhnovich, E. I. Energetics of protein−DNA interactions. *Nucleic Acids Res.* **2007**, *35*, 1039–1047.
(26) Gromiha, M. M.; Siebers, J. G.; Selvaraj, S.; Kono, H.; Sarai, A. Role of inter and intramolecular interactions in protein-DNA recognition. *Gene* **2005**, *364*, 108–113.
(27) Seeman, N. C.; Rosenberg, J. M.; Rich, A. Sequencespecific recognition of double helical nucleic acids by proteins. *Proc. Natl Acad. Sci. U.S.A.* **1976**, *73*, 804–808.
(28) Otwinowski, Z.; Schevitz, R. W.; Zhang, R. G.; Lawson, C. L.; Joachimiak, A.; Marmorstein, R. Q.; Luisi, B. F.; Sigler, P. B. Crystal structure of trp repressor/operator complex at atomic resolution. *Nature* **1988**, *335*, 321–329.
(29) Ahmad, S.; Gromiha, M. M.; Sarai, A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics* **2004**, *20*, 477–486.
(30) Bhardwaj, N.; Lu, H. Residue-level prediction of DNA-binding sites and its application on DNA-binding protein predictions. *FEBS Lett.* **2007**, *581*, 1058–1066.
(31) Wu, J.; Liu, H.; Duan, X.; Ding, Y.; Wu, H.; Bai, Y.; Sun, X. Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. *Bioinformatics* **2009**, *25*, 30–35.
(32) Xu, B.; Yang, Y.; Liang, H.; Zhou, Y. An all-atom knowledge-based energy function for protein-DNA threading, docking decoy discrimination, and prediction of transcription-factor binding profiles. *Proteins* **2009**, *76*, 718–730.
(33) Ahmad, S.; Sarai, A. PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinf.* **2005**, *6*, 6.
(34) Wang, L.; Brown, S. J. Prediction of DNA-binding residues from sequence features. *J. Bioinform. Comput. Biol.* **2006**, *4*, 1141–1158.
(35) Kuznetsov, I. B.; Gou, Z.; Li, R.; Hwang, S. Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. *Proteins* **2006**, *64*, 19–27.
(36) Ofran, Y. Mysore, V.; Rost, B. Prediction of DNA-binding residues from sequence. *Bioinformatics* **2007**, *23*, i347–i353.
(37) Ho, S. Y.; Yu, F. C.; Chang, C. Y.; Huang, H. L. Design of accurate predictors for DNA-binding sites in proteins using hybrid SVM-PSSM method. *Biosystems* **2007**, *90*, 234–241.
(38) Wang, L.; Yang, M. Q.; Yang, J. Y. Prediction of DNA-binding residues from protein sequence information using random forests. *BMC Genomics* **2009**, *10*, S1.
(39) Gromiha, M. M.; Yokota, K.; Fukui, K. Energy based approach for understanding the recognition mechanism in protein-protein complexes. *Mol Biosyst.* **2009**, *5*, 1779–1786.
(40) Gromiha, M. M.; Yokota, K.; Fukui, K. Sequence and structural analysis of binding site residues in protein-protein complexes. *Int. J. Biol. Macromol.* **2010**, *46*, 187–192.
(41) Gromiha, M. M.; Yokota, K.; Fukui, K. Understanding the recognition mechanism in protein-RNAcomplexes using energy based approach. *Curr. Protein Pept. Sci.* **2010**, *11*, 629–638.
(42) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P.A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
(43) Pichierri, F; Aida, M.; Gromiha, M. M.; Sarai, A. Free energy maps of interactions for DNA-protein recognition. *J. Am. Chem. Soc.* **1999**, *121*, 6152–6157.
(44) Gromiha, M. M.; Selvaraj, S. Inter-residue interactions in protein folding and stability. *Prog. Biophys. Mol. Biol.* **2004**, *86*, 235–277.
(45) Gromiha, M. M.; Selvaraj, S.; Jayaram, B.; Fukui, K. Identification and analysis of binding site residues in protein complexes: energy based approach. *Lect. Notes Comp. Sci.* **2010**, *6215*, 626–633.
(46) Luscombe, N. M.; Laskowski, R. A.; Thornton, J. M. Amino acid−base interactions: a three-dimensional analysis of protein−DNA interactions at an atomic level. *Nucleic Acids Res.* **2001**, *29*, 2860–2874.
(47) Tomovic, A.; Oakeley, E. J. Computational structural analysis: multiple proteins bound to DNA. *PLoS One* **2008**, *3*, e3243.
(48) Prabakaran, P; An, J.; Gromiha, M. M.; Selvaraj, S.; Uedaira, H.; Kono., H.; Sarai, A. Thermodynamic database for protein-nucleic acid interactions (ProNIT). *Bioinformatics* **2001**, *17*, 1027–1034.
(49) Larkin, C.; Datta, S.; Harley, M. J.; Anderson, B. J.; Ebie, A.; Hargreaves, V.; Schildbach, J. F. Inter- and intramolecular determinants of the specificity of single-stranded DNA binding and cleavage by the F factor relaxase. *Structure* **2005**, *13*, 1533–1544.
(50) Anderson, E. M.; Halsey, W. A.; Wuttke, D. S. Site-directed mutagenesis reveals the thermodynamic requirements for single-stranded DNA recognition by the telomere-binding protein Cdc13. *Biochemistry* **2003**, *42*, 3751–3758.
(51) Qin, S.; Zhou, H. -X. Do electrostatic interactions destabilize protein−nucleic acid binding?. *Biopolymers* **2007**, *86*, 112–118.
(52) Cherstvy, A. G.; Kolomeisky, A. B.; Kornyshev, A. A. Protein−DNA interactions: reaching and recognizing the targets. *J. Phys. Chem. B* **2008**, *112*, 4741–4750.
(53) Cheng, A. C.; Chen, W. W.; Fuhrmann, C. N.; Franke, A. D. Recognition of nucleic acid bases and base-pairs by hydrogen bonding to amino acid side-chains. *J. Mol. Biol.* **2003**, *327*, 781–796.
(54) Mukherjee, S.; Majumdar, S.; Bhattacharyya, D. Role of hydrogen bonds in protein−DNA recognition: effect of nonplanar amino groups. *J. Phys. Chem. B* **2005**, *109*, 10484–10492.
(55) Zahran, M.; Daidone, I.; Smith, J. C.; Imhof, P. Mechanism of DNA Recognition by the Restriction Enzyme EcoRV. *J. Mol. Biol.* **2010**, *401*, 415–432.