PROTEINS: Structure, Function, and Bioinformatics 55:885–894 (2004)

Structure-Based Prediction of DNA-Binding Sites on Proteins Using the Empirical Preference of Electrostatic Potential and the Shape of Molecular Surfaces

Yuko Tsuchiya,[1] Kengo Kinoshita,[2,3]* and Haruki Nakamura[1]
[1]*Institute for Protein Research, Osaka University, Osaka, Japan*
[2]*Graduate School of Integrated Science, Yokohama City University, Yokohama, Japan*
[3]*Structure and Function of Biomolecules, PRESTO, JST, Saitama, Japan*

ABSTRACT **Protein–DNA interactions play an essential role in the genetic activities of life. Many structures of protein–DNA complexes are already known, but the common rules on how and where proteins bind to DNA have not emerged. Many attempts have been made to predict protein–DNA interactions using structural information, but the success rate is still about 80%. We analyzed 63 protein–DNA complexes by focusing our attention on the shape of the molecular surface of the protein and DNA, along with the electrostatic potential on the surface, and constructed a new statistical evaluation function to make predictions of DNA interaction sites on protein molecular surfaces. The shape of the molecular surface was described by a combination of local and global average curvature, which are intended to describe the small convex and concave and the large-scale concave curvatures of the protein surface preferentially appearing at DNA-binding sites. Using these structural features, along with the electrostatic potential obtained by solving the Poisson–Boltzmann equation numerically, we have developed prediction schemes with 86% and 96% accuracy for DNA-binding and non-DNA-binding proteins, respectively. Proteins 2004;55:885–894.**

Key words: protein function prediction; three-dimensional structure; Connolly surface; Poisson–Boltzmann equation; statistical potential; computational method; protein informatics

INTRODUCTION

According to the progress of structural genomics projects, many three-dimensional (3D) structures of proteins have been determined before their functions are identified, and the proteins solved in these projects often show little or no sequence similarity to proteins with known structure.[1] Thus, the relation between evolutional linkage and functional similarity will give us only part of the information to infer their function. Therefore, it is essential to use structural information directly to identify their functions.[2–6]

Structure-based function prediction is usually done by searching for proteins with a similar fold. However, it is now well known that proteins with different folds can perform a similar function, and proteins with similar folds can sometimes express a different function.[7] This means that fold-level similarity does not always imply functional similarity. Thus, a similarity search of the local atomic configurations or molecular surface geometries around the functional or conserved site of known proteins has been attempted to infer the molecular function of proteins, for instance, by detecting hydrophobic patches, conserved residues and secondary structures, and cavities.[8–12] These approaches work well for proteins with small active sites but may not be useful for large ligand-binding sites, such as a double-stranded DNA (dsDNA) chain, because the large ligand-binding sites do not necessarily localize on the protein 3D structure.

Protein–DNA interactions play a central role in living things for genetic activities such as replication, transcription, repair, and translation of genomes. For better understanding of biological systems, it is crucially important to understand such processes in terms of molecular interaction, where an essential problem is to determine which proteins interact with DNA, and how they bind to the DNA. To investigate this problem, numerous studies searching for common rules governing the formation of protein–DNA have been performed.[13] Some rules, such as relatively higher conservation of amino acid residues, large positively charged patches,[14,15] specific hydrogen-bonding patterns,[16] and hydrogen-bond networks,[17] were obtained by analyzing the protein–DNA complexes found in the Protein Data Bank (PDB), and the rules were applied for prediction of protein–DNA interaction. However, the common rules have not been found yet,[17,18] and

Grant sponsor: Ministry of Education, Culture, Sports, Science and Technology (MEXT) (to K. Kinoshita). Grant sponsor: Institute for Bioinformatics Research and Development, Japan Science and Technology Corporation (BIRD-JST) (to H.Nakamura).

*Correspondence to: Kengo Kinoshita, Graduate School of Integrated Science, Yokohama City University, 1-7-29 Suehiro-cho, Turumi-ku, Yokohama, 230-0045, Japan. E-mail: kinoshita@tsurumi.yokohama-cu.ac.jp

Received 23 October 2003; Accepted 22 December 2003

Published online 1 April 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20111

the success rate of prediction is still not high. In other words, it is hard to predict where and how proteins interact with DNA, and it is not an easy problem to find which proteins interact with DNA. On the other hand, many structural biologists can occasionally distinguish a DNA-binding protein from a nonbinding protein by inspecting the molecular surface colored by electrostatic potential.[19–21]

These considerations encouraged us to develop another simple method to predict the DNA-binding sites on protein structure, by focusing our attention on the shape of the molecular surfaces and the electrostatic potential on the surface. The former is a main contributor to van der Waals interaction, and the latter is the major factor due to the strongly negative moiety of the phosphate backbone of DNA. We then provided a new evaluation function to predict DNA-binding sites on protein 3D structures.

MATERIALS AND METHODS

Molecular Surface and Electrostatic Potential

Molecular surfaces are generated by using the program MSRoll developed by Connolly,[22] where molecular surfaces are represented by a set of triangular meshes. The electrostatic potential at each vertex on the mesh was calculated by solving the Poisson–Boltzmann equations numerically[23,24] for a precise continuum model by the program SCB,[25] which used a self-consistent boundary algorithm to eliminate the effect of the boundary in the finite-difference method with a 1.0 Å grid size. A partial charge taken from the AMBER parameter[26] was assigned to each atom, and all histidine residues were assumed to be neutral. The dielectric constants of the protein and solvent regions were set at 2.0 and 80.0, respectively, and the ionic strength of 0.1 M was used in every case. The value of the electrostatic potential apart from each vertex by 1.4 Å along with the normal vector was projected onto the molecular surface.

Average Curvatures

In order to represent the shape of molecular surfaces, we used the *average curvature* defined as follows: Two average curvatures for each vertex point on the molecular surface, one to describe the local shape and the other for the global shape, were considered.

At first, such vertices were gathered for each vertex that was located in the vicinity of the considered vertex, and that had normal vectors with a direction similar to the normal vector of the considered vertex. The proximity of the vertices was judged based on whether the distance between the vertices and the considered vertex was less than 4.0 Å or 15.0 Å for the local and global curvature, respectively. Two normal vectors whose angle was less than 90° were considered to have a similar direction.

Normal planes including the normal vector of the considered vertex were generated by rotating a normal plane with an interval of 5° around the normal vector. And for each normal plane, the normal cross-section with the molecular surface was approximated using a quadratic curve by least-squares fitting, and then the curvature for the curve was calculated. The average curvature was calculated by taking an average of the set of curvatures obtained for a set of normal cross-sections.

Evaluation Function

In order to describe the structural features of an interaction site between a protein and dsDNA, three parameters—electrostatic potential, local average curvature, and global average curvature—were considered. The 3D space was divided into grids with a size of 0.15 (V) × 0.20 (Å^{-2}) × 0.05 (Å^{-2}). At each grid, the relative frequency of the number of vertices appearing at the dsDNA-binding and nonbinding sites was then counted and designated here as F_{bind} (= $N_{bind}/N_{bind\text{-}total}$) and $F_{nonbind}$ (= $N_{nonbind}/N_{nonbind\text{-}total}$), where N_{bind} and $N_{nonbind}$ are the number of vertices at the dsDNA-binding site and that in the nonbinding site, respectively, and $N_{bind\text{-}total}$ and $N_{nonbind\text{-}total}$ are the total number of N_{bind} and $N_{nonbind}$, respectively.

Using the ratio of these relative frequencies, $Est = F_{bind}/F_{nonbind}$, we estimated the preference of the dsDNA-binding for each vertex on the molecular surface.

Prediction Score

To calculate the prediction score (P_{score}), each protein was viewed from all possible angles, and a particular direction was determined, so that the maximum area of the predicted region that projected onto the direction of the view was obtained. In fact, 36 × 18 possibilities for the azimuth and the zenith angle (i.e., 10° interval) around the center of gravity of the protein were considered. The P_{score} was then obtained as the ratio of the maximum A_{bind} to the whole area (A_{whole}) viewed from the same direction as that for the maximum A_{bind} (i.e., $P_{score} = A_{bind}/A_{whole}$). Here, each area, A_{bind} and A_{whole}, was the sum of the corresponding triangular meshes. The value of P_{score} ranged from 0.0 to 1.0, and the larger P_{score} indicated a better prediction.

Data Sets

Data set 1

Protein–dsDNA complexes were gathered using a protein–nucleic acid complex database[27] and a nucleic acid database,[28] and 63 representative entries were selected according to the Structural Classification of Proteins (SCOP),[29] where one representative having the largest contact area with dsDNA was selected for each SCOP family. This data set was used to obtain the empirical evaluation function.

Data set 2

Complexes between proteins and adenosine triphosphate (ATP) analogs were picked up using Relibase.[30] Here, we considered phosphomethylphosphonic acid adenylate ester (ACP) and phosphoaminophosphonic acid adenylate ester (ANP) as ATP analogs, and 21 representatives were selected based on SCOP family classification.

Data set 3

In order to obtain a list of the proteins that did not form complexes with dsDNA but formed complexes with differ-

ent protein chains, we first picked the entries of hetero-oligomer proteins with the Xquery from the PDB data described by Extensible Markup Language (XML). Among them, 406 proteins were further selected by taking one entry with the best resolution for each SCOP family, which was not annotated as a DNA- or RNA- binding protein.

RESULTS AND DISCUSSION

Data Sets

In this study, we used three sets of entries: (1) a set of complexes of proteins and dsDNA to construct an evaluation function for the prediction of dsDNA-binding sites, (2) a set of complexes between proteins and ATP analogs to compare the dsDNA-(polynucleotide) and ATP analog (mononucleotide)-binding protein surfaces, and (3) the nonredundant protein structures that are not considered dsDNA-binding proteins. We used the structures determined by X-ray crystallography whose resolutions are better than 2.5 Å. All coordinate files were obtained from the PDB,[31,32] and we selected protein subunits and dsDNA chains relevant to the biological activity in each complex. The entries and the chain identifier used in our studies for data set 1, data set 2, and part of data set 3 are listed in Tables I, II, and III, respectively.

Evaluation Function to Predict dsDNA-Binding Surface

Structural features of proteins are described by three parameters: the electrostatic potential, the local average curvature, and the global average curvature (see Materials and Methods section for their definitions). A combination of the minimum and maximum curvatures or that of Gauss and mean curvatures has often been used to describe the shape of the molecular surface.[33,34] However, we used here the average curvatures due to the best prediction performance with this descriptor according to the Jackknife test described later.

A distribution of the three parameters for data set 1 is shown in Figure 1. The vertices in dsDNA-binding and nonbinding sites are colored red and blue, respectively. The protein vertices, which are located within 3.0 Å from any vertices belonging to dsDNA surface, are defined as vertices in dsDNA-binding sites, and other vertices are considered to belong to nonbinding sites.

As seen in Figure 1, the distributions of the electrostatic potential and the global curvature in dsDNA-binding sites are more positive than the distribution in nonbinding sites. In contrast, the peak positions of the local curvatures in dsDNA-binding and nonbinding sites are not so different, but as a whole, the frequency distribution of the dsDNA-binding sites is shifted to some extent. This means that the main contributor that distinguishes a dsDNA-binding site and a nonbinding site is the electrostatic potential, and that the shape descriptor gives rise to minor but meaningful differences.

On the other hand, an ATP-binding site shows a different tendency [Fig. 1(B)] where the main differences between the binding and the nonbinding sites are found in the curvatures; especially the difference in the global average curvature is remarkable. This suggests that a polynucleotide (dsDNA) is mainly recognized through electrostatic interaction, but a mononucleotide (ATP analog) is mainly recognized by van der Waals interaction or shape complementarity.

Based on the distribution shown in Figure 1(A), we created an evaluation function to predict the dsDNA-binding surfaces on the protein surface. The distribution was divided into a grid space with a size of 0.15 (V) $\times$ 0.20 (Å^{-2}) $\times$ 0.05 (Å^{-2}) for each direction, and the relative frequency for each grid was counted. The evaluation function was calculated as the ratio of the relative frequency at each grid, which we call the *Est* value. When the *Est* value was larger than an appropriate threshold value, Est^{th}, the vertex was considered to be dsDNA-binding surfaces and vice versa. After examining several values for Est^{th}, we found that $Est^{\mathrm{th}} = 4.0$ gave the best performance, and it was used in all the following studies.

Success Rates for the Data Sets

With the evaluation function described above, we classified each vertex on the molecular surface into a dsDNA-binding and a nonbinding surface according to the *Est* value. We then judged whether the protein was a dsDNA-binding protein, when a sufficient area on the protein surface viewed from a direction was considered the dsDNA-binding surface. To quantify this judgment, we use a prediction score (P_{score}) as an indicator, which is the ratio of the predicted area to the whole area viewed from the direction where its value becomes maximum (see Materials and Methods section for the precise definition). According to the Jackknife procedure, one entry in data set 1 (dsDNA) was selected as a test entry, and the others in the same data set were used as a training set to calculate the P_{score}. This procedure was repeated until all entries were used as the test entry. With the evaluation function obtained from all the entries in data set 1, we also calculated P_{score} for the entries in data sets 2 and 3. These entries were the negative examples, which were expected to be predicted to be the nonbinding proteins.

Figure 2(A) shows the relative frequency of P_{score} for data sets 1 (dsDNA), 2 (ATP), and 3 (nonbinding). The vertical thick dotted line is the threshold value for the prediction.

Here, we used P_{score} as an indicator, but the absolute value of the predicted area (A_{bind}) is another candidate for a score. To examine which indicator is better for the prediction result, we also calculated A_{bind} in the same manner as described above, as shown in Figure 2(B). As a result, we could not determine a good threshold value for A_{bind} due to a large overlap between the distributions. Furthermore, there is a strong tendency for the larger proteins to have a larger A_{bind}, which is not desirable behavior for our purpose, because a large area is not an important factor for proteins to bind to DNA. Therefore, we adopted P_{score} throughout the following studies.

TABLE I. Entries in Data Set 1

PDB ID	Chain	Protein	P_{score}	CC_v
1a73	AB	Intron-encoded homing endonuclease I-Ppol	0.17	0.50
1am9	AB	Sterol regulatory element binding protein	0.30	0.36
1au7	AB	Pou domain Pit-1	0.33	0.54
1azp	A	Hyperthermophile chromosomal protein Sac7d	0.10	0.32
1b3t	AB	Nuclear protein EBNA1	0.42	0.64
1b94	AB	Restriction endonuclease EcoRV	0.18	0.40
1bc8	C	SAP-1	0.32	0.41
1bdt	ABCD	Wild-type Arc	0.34	0.40
1bl0	A	Multiple antibiotic resistance protein	0.20	0.35
1d02	AB	Restriction endonuclease MunI	0.17	0.55
1d2i	AB	Restriction endonuclease Bgl II	0.17	0.51
1d3u	AB	TATA-binding protein, transcription initiation factor II B	0.12	0.28
1dmu	AM	Restriction endonuclease Bgl I	0.16	0.51
1dnk	A	Deoxyribonuclease I	0.03	0.03
1dp7	PQ	Mhr class II transcription factor hRFXI	0.24	0.51
1ebm	A	Human 8-oxoguanine glycosylase hOGG1	0.05	0.17
1egw	AB	Mads box transcription enhancer factor 2 polypeptide A	0.46	0.51
1eqz	ABCDEFGH	Nucleosome core particle (histone H2A,H2B,H3,H4)	0.48	0.33
1eri	AM	Endonuclease EcoRI	0.12	0.54
1eyu	AB	Restriction endonuclease PvuII	0.09	0.29
1f2i	GH	Zifl2 contains zinc fingers 1 and 2 of Zif268	0.21	0.40
1f4k	AB	Replication terminator protein	0.23	0.35
1fjl	AB	Drosophila paired protein	0.46	0.62
1fjx	A	Methyltransferase HhaI	0.19	0.52
1g2d	C	TATA box zinc finger protein	0.28	0.47
1g9y	AB	Homing endonuclease I-Crel	0.19	0.43
1ga5	AB	Nuclear receptor RevErb α DNA-binding domain	0.43	0.30
1gd2	EF	Transcription factor PAP1	0.28	0.45
1h89	ABC	CAAT/enhancer binding protein beta, Myb proto-oncogene protein	0.30	0.43
1hcr	A	Recombinase Hin	0.23	0.27
1hlv	A	Major centromere autoantigen B CENP-B	0.33	0.46
1hwt	CD	Heme activator protein HAP1	0.39	0.55
1iaw	AB	Restriction endonuclease NawI	0.12	0.50
1ign	A	Repressor activator protein 1 RAP1	0.25	0.47
1ihf	AB	Integration host factor IHF	0.28	0.50
1j59	AB	Catabolite gene activator protein CAP	0.12	0.20
1jey	AB	Ku heterodimer (Ku70, Ku80)	0.23	0.41
1jft	AC	Purine nucleotide synthesis repressor	0.17	0.42
1jxl	A	Y-family DNA polymerase	0.30	0.47
1k78	EF	Paired box protein Pax5/C-Est-1 protein	0.24	0.47
1ku7	A	*Thermus aquaticus* RNA polymerase σ subunit	0.25	0.09
1l31	AC	Bacterial quorum-sensing transcription factor TraR	0.15	0.34
1lq1	AB	Transcription factor Spo0A	0.13	0.41
1mjq	ABCD	Methionine repressor	0.27	0.64
1mnm	ABCD	Transcription regulator MCM1	0.43	0.51
1nfk	AB	Nuclear factor kappa-B p50	0.12	0.48
1qbj	AB	dsRNA specific adenosine deaminase ADAR1	0.21	0.53
1qln	A	Bacteriophage T7 RNA polymerase	0.12	0.17
1qn7	A	Transcription initiation factor TFIID-1	0.22	0.22
1qpi	AB	Tetracycline repressor	0.05	0.34
1qpz	AB	Purine nucleotide synthesis repressor	0.15	0.45
1qrv	A	Chromosomal high mobility group protein D	0.12	0.34
1skn	P	Transcription factor Skn-1	0.34	0.53
1trr	ABDEGHJK	Trp repressor	0.21	0.64
1xbr	AB	T domain from *Xenopus laevis*	0.20	0.47
1ytf	ABCD	Yeast TATA-box binding protein	0.18	0.38
2bam	AB	Restriction endonuclease BamHI	0.19	0.52
2bdp	A	Bacillus DNA polymerase I	0.10	0.35
2irf	GH	Interferon regulatory factor 2 IRF-2	0.25	0.41
3cro	LR	Phage 434 Cro repressor	0.30	0.26
4bdp	A	Bacillus DNA polymerase I	0.12	0.44
4crx	AB	Recombinase Cre	0.33	0.56
6pax	A	Human Pax 6 paired domain	0.25	0.44
Average			0.23	0.42

TABLE II. Entries in Data Set 2

PDB ID	Chain	Protein	P_{score}	CC_v
1ank	A	Adenylate kinase	0.09	0.22
1cdk	A	cAMP-dependent protein kinase	0.10	−0.05
1dah	AB	Dethiobiotin synthetase	0.00	0.03
1e22	AB	Lysyl-tRNA synthetase hexagonal form	0.02	−0.02
1e9b	AB	Thymidylate kinase	0.10	0.29
1ei1	AB	DNA-gyrase B	0.05	0.38
1gsj	AB	Acetylglutamate kinase	0.05	0.36
1h73	A	Homoserine kinase	0.05	0.39
1i44	A	Insulin receptor tyrosine kinase	0.03	−0.03
1i5a	A	Chemotaxis sensor protein histidine kinase CheA	0.07	0.19
1ia9	AB	Atypical protein kinase domain of a TRP channel	0.05	−0.02
1j7u	B	3′,5′-aminoglycoside phosphotransferase type IIIa	0.02	−0.03
1jbv	A	Folylpolyglutamate synthase	0.03	0.08
1kji	A	PurT-encoded glycinamide ribonucleotide transformylase	0.01	0.02
1lij	A	Adenosine kinase	0.04	−0.04
1mmn	—	*Dictyostelium discoideum* Myosin Motor Domain	0.07	0.41
1ngj	—	ATPase fragment of heat shock cognate protein	0.05	−0.02
1qha	A	Human hexokinase type I	0.06	0.11
1vpe	—	Phosphoglycerate kinase	0.07	0.23
2bif	AB	Fructose-6-phosphate,2-kinase/fructose-2,6-bisphosphatase	0.08	0.22
2mjp	AB	Nucleotide triphosphatase	0.03	0.19
Average			0.05	0.14

TABLE III. Part of the Entries in Data Set 3, with $P_{score} \geq 0.12$

PDB ID	Chain	Protein	P_{score}
1kqf	C	Formate dehydrogenase, nitrate-inducible, cytochome B556	0.26
1ev2	A	Fibroblast growth factor 2	0.25
1h8e	G	Bovine mitochondrial F1-ATPase γ subunit	0.22
1ef1	A	Moesin N-terminal FERM domain	0.19
1fns	A	von Willebrand factor	0.19
1wq1	G	GTPase-activating protein p120GAP	0.18
1dxr	C	Photosynthetic reaction center cytochrome C subunit	0.18
1g4y	B	Calcium-activated potassium channel rSK2	0.17
1e6e	A	NADPH-dependent adrenodoxin oxidoreductase	0.17
1a2x	B	Troponin I	0.16
1gls	A	P-selectin	0.16
1c1y	B	Proto-oncogene serine/threonine protein kinase c-Raf1	0.16
1f7z	I	Bovine pancreatic trypsin inhibitor	0.14
1hzz	C	Proton-translocating nicotinamde nucleiotide transhydrogenase subunit	0.13
1h32	A	Heterodimeric cytochrome C	0.13
1eex	B	Propanediol dehydratase	0.13
1jb0	K	Photosystem I reaction centre subunit X	0.13
1eer	A	Erythropoietin	0.13
1ajs	A	Aspartate aminotransferase	0.12

To determine the threshold of P_{score} for a positive prediction, a correlation coefficient (CC) value[35] was calculated based on the number of correctly predicted entries versus the ATP-binding proteins, or the ATP-binding and non-DNA-binding proteins, by changing the threshold value of P_{score} (Fig. 3). The maximum value (about 0.84, see solid line in Fig. 3) of the CC was obtained for a set of the dsDNA-binding and the ATP-binding proteins, when $P_{score} \geq 0.11$ was used as the criterion for correct prediction. With this threshold value, we could achieve the correct prediction rate of 90.5% for the dsDNA-binding proteins, and 100% for the ATP-binding proteins. In a similar way, we could obtain the maximum value of the CC for a combined data set of all entries (dotted line in Fig. 3) at the threshold of $P_{score} \geq 0.12$. Even with this severe threshold, we obtained CC = 0.76, and success rates of 85.7% and 95.6% for the dsDNA-binding protein and all the other entries, respectively.

Prediction Accuracy for the Location of a Binding Site

Up to now, we have shown the results of predictions regarding whether the given proteins are dsDNA- or non-dsDNA-binding proteins, and we did not mention the prediction accuracy of the location of a binding site. To

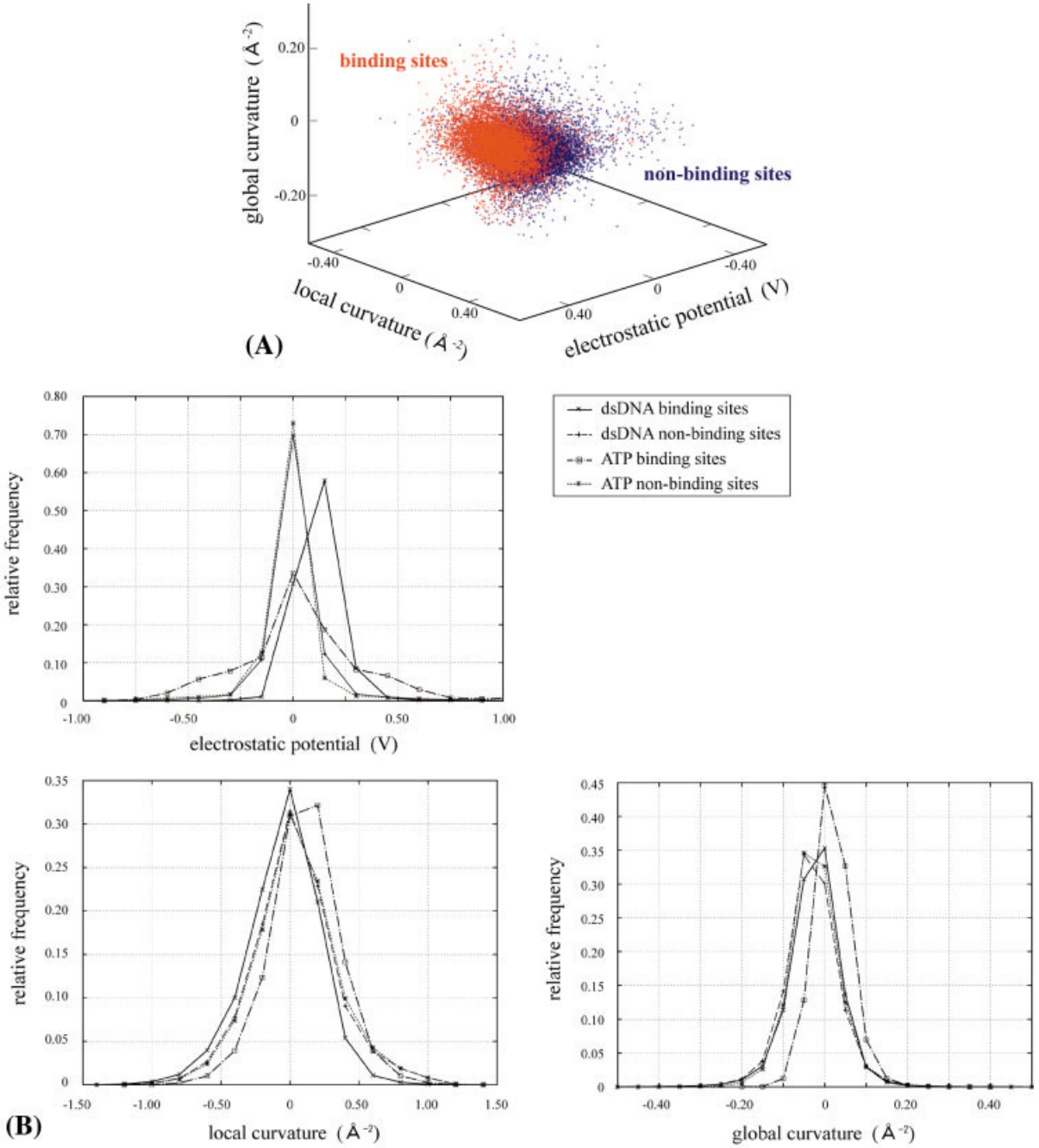

Fig. 1. (**A**) Distribution of the electrostatic potential, the local curvature, and the global curvature of the vertices appearing at dsDNA-binding sites (red dots) and nonbinding sites (blue dots). (**B**) Cross-section of the above distribution along with that of ATP-binding and ATP-nonbinding sites.

evaluate the accuracy of the location prediction, we use the Matthews correlation coefficient again, but now the calculation is based on the number of correctly predicted and incorrectly predicted *vertices,* which we call CC_v. It should be noted that the calculation of CC_v requires information on the actual binding site to determine the number of true positives, true negatives, false positives, and false negatives. Thus, the values for the non-dsDNA-binding protein cannot be calculated, and the results are shown for the dsDNA-binding protein and ATP-binding protein, in Tables I and II, respectively.

In Figure 4, a scatterplot of P_{score} versus CC_v is shown. A higher value of P_{score} indicates that the protein is more likely to have dsDNA-binding regions, and a higher value of CC_v means that the location of the binding site is more correctly predicted. Thus, when the plot is divided into four regions with two thick lines for the convenience of the following discussion, the dsDNA-binding protein in the

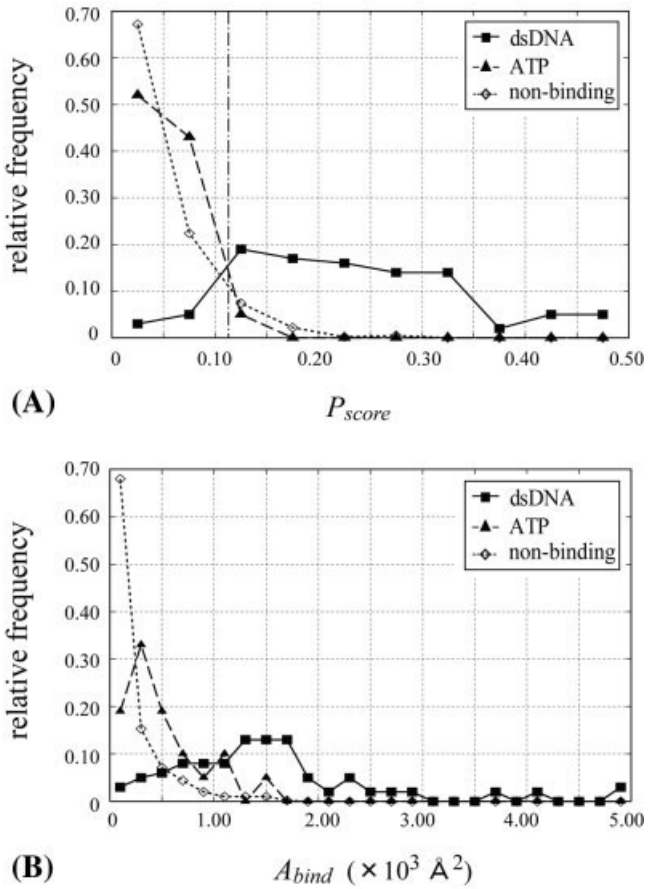

Fig. 2. (**A**) Relative frequency of the number of entries against the prediction score (P_{score}) for data set 1 (dsDNA, solid line), data set 2 (ATP-binding protein, broken line), and data set 3 (non-DNA-binding proteins, dotted line). (**B**) Same as (A) but relative to the absolute value of the predicted area A_{bind}.

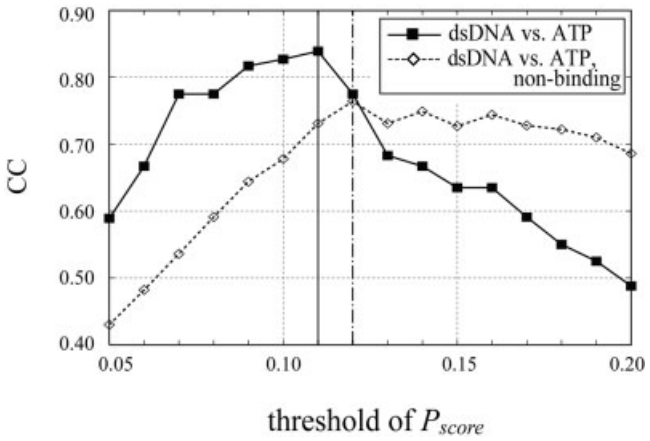

Fig. 3. Correlation coefficient (CC) versus the threshold value of P_{score}. CC was calculated using ATP-binding proteins (data set 2, solid line) as negative entries, or that for the entries in data set 2 and data set 3 (dotted line). Vertical lines indicate the threshold values we used in this study. The vertical solid line is for dsDNA versus ATP-binding proteins, and the vertical broken dotted line is for dsDNA versus ATP-binding or non-DNA-binding proteins.

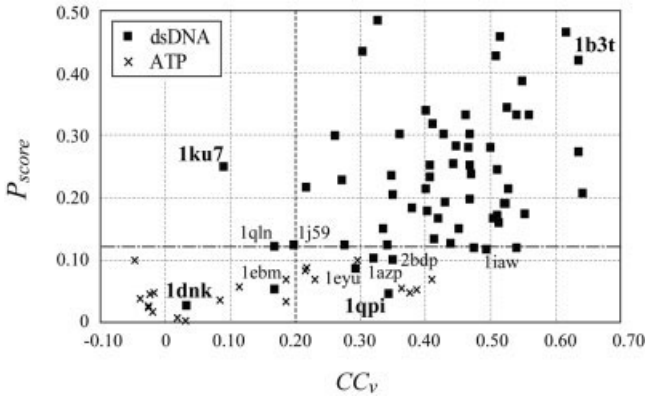

Fig. 4. A scatterplot between P_{score} and CC_v. The horizontal thick dotted line indicates the threshold value of P_{score} used in this study, and the vertical thick dotted line is a tentative threshold line for the discussion (see the text for details). The black squares indicate the entries in data set 1, and the crossmarks represent the entries in data set 2. Parts of the entries of data set 1 placed in the interesting regions are labeled with PDB ID codes, and entries in boldface type are discussed in the text (Fig. 5).

upper right region is correctly predicted both in dsDNA-binding ability and in the location of the binding site, and vice versa for the entries in the lower left region. As a result, 51 out of 63 (81%) entries in data set 1 (dsDNA) belong to the upper right region, thus correctly predicted as the dsDNA-binding protein, and the location of the binding site was also well assigned. As an example of the successful cases, a nuclear protein EBNA1 (PDB code: 1b3t) is shown in Figure 5(A), where $P_{score} = 0.42$ and $CC_v = 0.64$ were obtained.

In contrast, the most unsuccessful prediction was the case for deoxyribonuclease I (PDB code: 1dnk). In this protein, relative to the enzymatic activity, unusually abundant acidic residues are found in the dsDNA-binding sites, where 6 acidic residues (3 Asp and 3 Glu residues) and 1 basic residue (Arg) are involved. These acidic residues seem to construct a negative environment at the binding site, as seen in the red region in Figure 5(B). It may be worth noting that 3 histidine residues exist at this binding site as active-site residues (His-134 and His-252)[36] or as a binding residue (His-44). In our calculation of electrostatic potential, all histidine residues were treated as being in a neutral state, but they can be in a positively charged state according to the environment in which they exist. The change in the charged state of the histidines could affect the electrostatic potential of the binding site, which might be a reason that our prediction failed in this example. To examine this possibility, we recalculated the electrostatic potential by changing the neutral charges of the 134th, 252nd, and 44th histidine residues into positive and obtained sufficiently large P_{score} (0.38) and CC_v (0.28) values [Fig. 5(C)]. A similar rationale may be possible in the case of 8-oxoguanine DNA glycosylase (PDB code: 1ebm), which is another example of an unsuccessful prediction (Fig. 4). In the protein dsDNA-binding site, one histidine residue (His-270) can be observed. Changing the

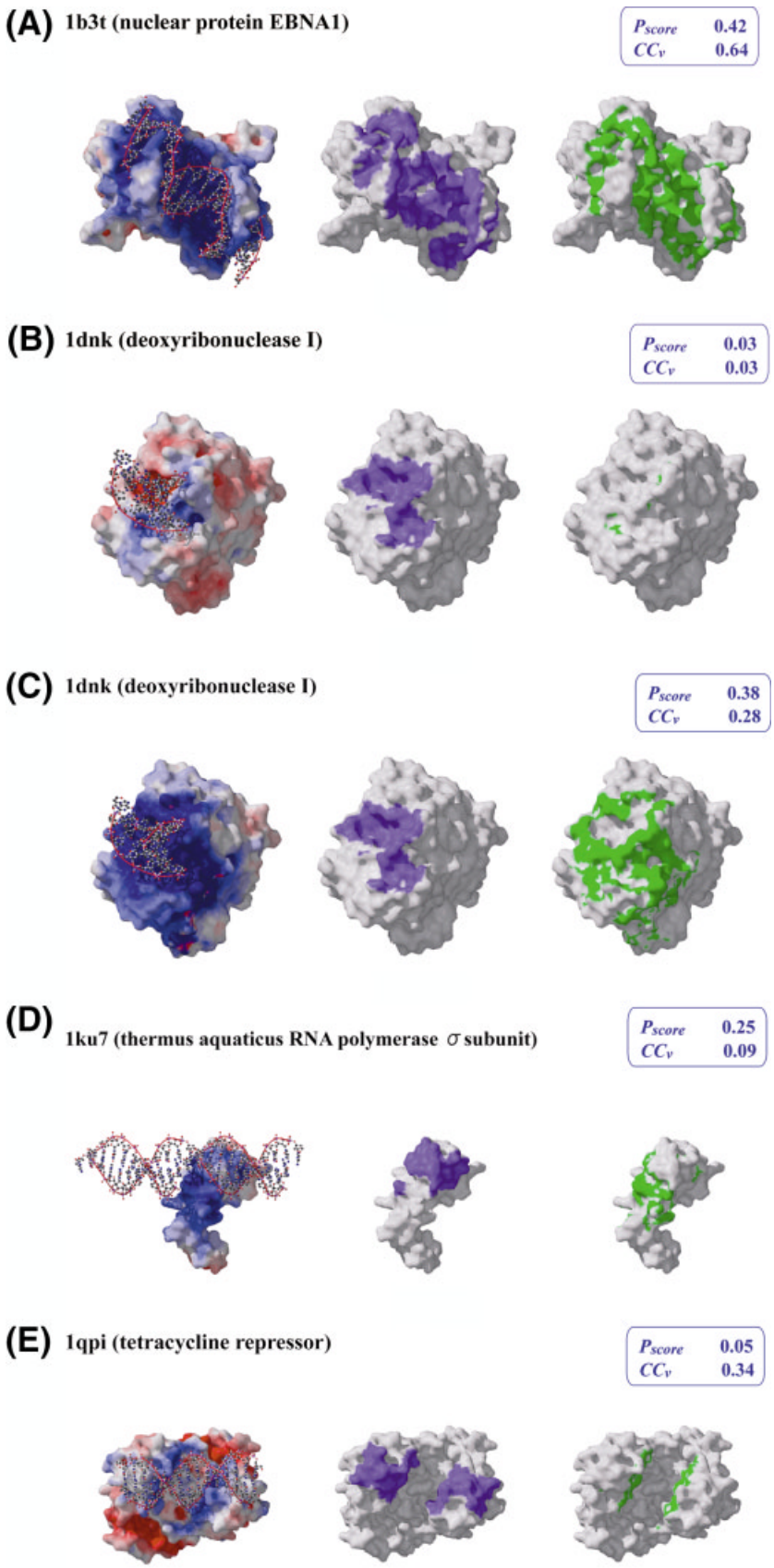
(A) 1b3t (nuclear protein EBNA1)
P_{score} 0.42
CC_v 0.64
(B) 1dnk (deoxyribonuclease I)
P_{score} 0.03
CC_v 0.03
(C) 1dnk (deoxyribonuclease I)
P_{score} 0.38
CC_v 0.28
(D) 1ku7 (thermus aquaticus RNA polymerase σ subunit)
P_{score} 0.25
CC_v 0.09
(E) 1qpi (tetracycline repressor)
P_{score} 0.05
CC_v 0.34

Figure 5.

charged state of the histidine residue also brings us larger P_{score} (0.20) and CC_v (0.46) values.

As seen in these examples, the effect of assuming the neutral state of histidine residues seems to contribute to an underestimation of dsDNA-binding ability. Thus, once we obtain information about the charged state of individual histidine residues by NMR experiments or electrostatic calculations,[37–41] the prediction accuracy may be much improved. However, ionization of the interacting histidine residues may be induced upon DNA-binding, as well as the induced protein folding observed in several protein–protein and protein-DNA interactions.[42] The current empirical approach would have some limitations for these systems.

In the usual cases, a large P_{score} indicates a large value of CC_v, but there are some rare exceptions. Here, we discuss two exceptions, one for the σ_4 subunit of RNA polymerase (PDB code: 1ku7, $P_{score} = 0.25$, $CC_v = 0.090$), and the other for the tetracycline repressor (TetR, PDB code: 1qpi, $P_{score} = 0.045$, $CC_v = 0.34$), as representatives from the left upper region and the right lower region in Figure 4, respectively.

The σ_4 subunit is a subunit of RNA polymerase, which is relevant to binding to the promoter region, specifically the -35 element of DNA. This protein plays a central role in the regulation of bacterial transcription, and it is known to interact with various regulator proteins, such as CAP and FNR. The interaction between the σ_4 subunit and regulator proteins is inferred from mutational experiments with the σ_4 subunit of *Escherichia coli*[43] and X-ray crystallography of that of *Thermus aquaticus*.[44] Our prediction result in Figure 5(D) corresponds to the inferred protein interaction sites, where there is a large positive region due to a cluster of lysine or arginine residues (418, 421, 422, and 424). On the other hand, in the dsDNA-binding region elucidated by crystallography, some water molecules exist, and they interact with phosphates in dsDNA and with protein through acidic residues (Glu410 and Glu399). These acidic residues form negative electrostatic moieties at the dsDNA-binding site, which made our prediction unsuccessful.

The TetR protein is known to have a quite high specificity for its operator DNA (association constant $\sim 10^{11}$ M^{-1}).[45] The high specificity comes from a large number of interactions between the bases in DNA and the protein side-chains, which induce unusual electrostatic potential to the dsDNA-binding site [Fig. 5(E)]. Our method is one of the statistical approaches that could elucidate some common aspects of protein–DNA interactions, neglecting specific interactions. Therefore, to predict specific DNA recognition, more physicochemical approaches[46–49] may be required in addition to the statistical approaches.

Fig. 5. Examples of the prediction results. In each example, three figures are prepared. The left one represents a molecular surface colored according to the electrostatic potential from blue (positive) to red (negative); the middle one shows the dsDNA-binding sites in purple; and the right one indicates the predicted region in green. These figures were prepared by Molscript.[50] (**A**) For nuclear protein EBNA1 (PDB code: 1b3t) as one of the successful examples. (**B**) For deoxyribonuclease I (PDB code: 1dnk) as a worst example. The charge state of all histidine residues is assumed to be neutral. (**C**) Same as for (B), but three histidine residues located at the dsDNA-binding site are assumed to be in the positively charged state. (**D**) For the σ_4 subunit of RNA polymerase (PDB code: 1ku7). (**E**) For the tetracycline repressor (PDB code: 1qpi).

ACKNOWLEDGMENTS

We are grateful to the members of the Japanese branch of worldwide Protein Data Bank for their technical support.

REFERENCES

1. Teichmann SA, Murzin AG, Chothia C. Determination of protein function, evolution and interactions by structural genomics. Curr Opin Struct Biol 2001;11:354–363.
2. Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo CA. From structure to function: approaches and limitations. Nat Struct Biol 2000;Suppl 7:991–994.
3. Ota M, Kinoshita K, Nishikawa K. Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. J Mol Biol 2003;327:1053–1064.
4. Ondrechen MJ, Clifton JG, Ringe D. THEMATICS: a simple computational predictor of enzyme function from structure. Proc Natl Acad Sci USA 2001;98:12473–12478.
5. Elcock AH. Prediction of functionally important residues based solely on the computed energetics of protein structure. J Mol Biol 2001;312:885–896.
6. Kinoshita K, Nakamura H. Protein informatics towards function identification. Curr Opin Struct Biol 2003;13:396–400.
7. Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 2001;307:1113–1143.
8. Stahl M, Taroni C, Schneider G. Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network. Protein Eng 2000;13:83–88.
9. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. Analysis of catalytic residues in enzyme active sites. J Mol Biol 2002;324:105–121.
10. Cappello V, Tramontano A, Koch U. Classification of proteins based on the properties of the ligand-binding site: the case of adenine-binding proteins. Proteins 2002;47:106–115.
11. Lichtarge O, Sowa ME. Evolutionary predictions of binding surfaces and interactions. Curr Opin Struct Biol 2002;12:21–27.
12. Binkowski TA, Adamian L, Liang J. Inferring functional relationships of proteins from local sequence and spatial surface patterns. J Mol Biol 2003;332:505–526.
13. Nadassy K, Wodak SJ, Janin J. Structural features of protein–nucleic acid recognition sites. Biochemistry 1999;38:1999–2017.
14. Stawiski EW, Gregoret LM, Mandel-Gutfreund Y. Annotating nucleic acid-binding function based on protein structure. J Mol Biol 2003;326:1065–1079.
15. Jones S, Barker JA, Nobeli I, Thornton JM. Using structural motif templates to identify proteins with DNA binding function. Nucleic Acids Res 2003;31:2811–2823.
16. Mandel-Gutfreund Y, Schueler O, Margalit H. Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. J Mol Biol 1995;253:370–382.
17. Luscombe NM, Laskowski RA, Thornton JM. Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. Nucleic Acids Res 2001;29:2860–2874.
18. Pabo CO, Nekludova L. Geometric analysis and comparison of protein–DNA interfaces: why is there no simple code for recognition? J Mol Biol 2000;301:597–624.
19. Boggon TJ, Shan WS, Santagata S, Myers SC, Shapiro L. Implication of tubby proteins as transcription factors by structure-based functional analysis. Science 1999;286:2119–2125.
20. Hanaoka S, Nagadoi A, Yoshimura S, Aimoto S, Li B, de Lange T, Nishimura Y. NMR structure of the hRap1 Myb motif reveals a canonical three-helix bundle lacking the positive surface charge typical of Myb DNA-binding domains. J Mol Biol 2001;312:167–175.
21. Nagadoi A, Nakazawa K, Uda H, Okuno K, Maekawa T, Ishii S,

Nishimura Y. Solution structure of the transactivation domain of ATF-2 comprising a zinc finger-like subdomain and a flexible subdomain. J Mol Biol 1999;287:593–607.
22. Connolly ML. Solvent-accessible surfaces of proteins and nucleic acids. Science 1983;221:709–713.
23. Misra VK, Hecht JL, Yang AS, Honig B. Electrostatic contributions to the binding free energy of the lambdacI repressor to DNA. Biophys J 1998;75:2262–2273.
24. Zacharias M, Luty BA, Davis ME, McCammon JA. Poisson–Boltzmann analysis of the lambda repressor–operator interaction. Biophys J 1992;63:1280–1285.
25. Nakamura H, Nishida S. Numerical calculations of electrostatic potentials of protein-solvent systems by the self consistent boundary method. J Phys Soc Japan 1987;56:1609–1622.
26. Cornell WD, Cieplak P, Bayly CI, Gould JR, Merz KM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 1995;117:5179–5197.
27. An J, Nakama T, Kubota Y, Sarai A. 3DinSight: an integrated relational database and search tool for the structure, function and properties of biomolecules. Bioinformatics 1998;14:188–195.
28. Berman HM, Olson WK, Beveridge DL, Westbrook J, Gelbin A, Demeny T, Hsieh SH, Srinivasan AR, Schneider B. The nucleic acid database: a comprehensive relational database of three-dimensional structures of nucleic acids. Biophys J 1992;63:751–759.
29. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res 2002;30:264–267.
30. Hendlich M. Databases for protein–ligand complexes. Acta Crystallogr D Biol Crystallogr 1998;54:1178–1182.
31. Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. Nat Struct Biol 2003;10:980.
32. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
33. Exner TE, Keil M, Brickmann J. Pattern recognition strategies for molecular surfaces: I. Pattern generation using fuzzy set theory. J Comput Chem 2002;23:1176–1187.
34. Kinoshita K, Nakamura H. Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. Protein Sci 2003;12:1589–1595.
35. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405:442–451.
36. Weston SA, Lahm A, Suck D. X-ray structure of the DNase I-d(GGTATACC)2 complex at 2.3 Å resolution. J Mol Biol 1992;226:1237–1256.
37. Oda Y, Yamazaki T, Nagayama K, Kanaya S, Kuroda Y, Nakamura H. Individual ionization constants of all the carboxyl groups in ribonuclease HI from *Escherichia coli* determined by NMR. Biochemistry 1994;33:5275–5284.
38. Takahashi T, Nakamura H, Wada A. Electrostatic forces in two lysozymes: calculations and measurements of histidine pKa values. Biopolymers 1992;32:897–909.
39. Garcia-Moreno B, Dwyer JJ, Gittis AG, Lattman EE, Spencer DS, Stites WE. Experimental measurement of the effective dielectric in the hydrophobic core of a protein. Biophys Chem 1997;64:211–224.
40. Antosiewicz J, McCammon JA, Gilson MK. The determinants of pKas in proteins. Biochemistry 1996;35:7819–7833.
41. Forsyth WR, Robertson AD. Insensitivity of perturbed carboxyl pK(a) values in the ovomucoid third domain to charge replacement at a neighboring residue. Biochemistry 2000;39:8067–8072.
42. Dyson HJ, Wright PE. Coupling of folding and binding for unstructured proteins. Curr Opin Struct Biol 2002;12:54–60.
43. Lonetto MA, Rhodius V, Lamberg K, Kiley P, Busby S, Gross C. Identification of a contact site for different transcription activators in region 4 of the *Escherichia coli* RNA polymerase sigma70 subunit. J Mol Biol 1998;284:1353–1365.
44. Campbell EA, Muzzin O, Chlenov M, Sun JL, Olson CA, Weinman O, Trester-Zedlitz ML, Darst SA. Structure of the bacterial RNA polymerase promoter specificity sigma subunit. Mol Cell 2002;9:527–539.
45. Orth P, Schnappinger D, Hillen W, Saenger W, Hinrichs W. Structural basis of gene regulation by the tetracycline inducible Tet repressor–operator system. Nat Struct Biol 2000;7:215–219.
46. Suzuki M. Common features in DNA recognition helices of eukaryotic transcription factors. EMBO J 1993;12:3221–3226.
47. Suzuki M, Yagi N. DNA recognition code of transcription factors in the helix–turn–helix, probe helix, hormone receptor, and zinc finger families. Proc Natl Acad Sci USA 1994;91:12357–12361.
48. Oda M, Furukawa K, Ogata K, Sarai A, Nakamura H. Thermodynamics of specific and non-specific DNA binding by the c-Myb DNA-binding domain. J Mol Biol 1998;276:571–590.
49. Milev S, Gorfe AA, Karshikoff A, Clubb RT, Bosshard HR, Jelesarov I. Energetics of sequence-specific protein–DNA association: binding of integrase Tn916 to its target DNA. Biochemistry 2003;42:3481–3491.
50. Kraulis PJ. MOLSCRIPT: a program to produce both detailed and schematic plots of proteins structures. J Appl Crystallogr 1991;24:946–950.