doi:10.1016/S0022-2836(03)00031-7 *J. Mol. Biol.* (2003) **326**, 1065–1079

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Annotating Nucleic Acid-Binding Function Based on Protein Structure

Eric W. Stawiski[1], Lydia M. Gregoret[2]* and Yael Mandel-Gutfreund[2]

[1]*Department of Molecular, Cell and Developmental Biology University of California, Santa Cruz, CA 95064, USA*

[2]*Department of Chemistry and Biochemistry, University of California, Santa Cruz, CA 95064, USA*

**Corresponding author*

Many of the targets of structural genomics will be proteins with little or no structural similarity to those currently in the database. Therefore, novel function prediction methods that do not rely on sequence or fold similarity to other known proteins are needed. We present an automated approach to predict nucleic-acid-binding (NA-binding) proteins, specifically DNA-binding proteins. The method is based on characterizing the structural and sequence properties of large, positively charged electrostatic patches on DNA-binding protein surfaces, which typically coincide with the DNA-binding-sites. Using an ensemble of features extracted from these electrostatic patches, we predict DNA-binding proteins with high accuracy. We show that our method does not rely on sequence or structure homology and is capable of predicting proteins of novel-binding motifs and protein structures solved in an unbound state. Our method can also distinguish NA-binding proteins from other proteins that have similar, large positive electrostatic patches on their surfaces, but that do not bind nucleic acids.

Keywords: structural genomics; nucleic acid binding; function prediction; electrostatics; surface patches

Introduction

Structural genomics promises to deliver three-dimensional structures of proteins on a genomic scale.[1] Since the primary goal of this initiative is to obtain structural information about proteins for which there is currently little information, many of the new protein structures solved are expected to lack sequence homologs within the Protein Data Bank (PDB).[2] The ~16,000 protein structures that will be needed to bring the protein structure initiative to completion over the next ten years will likely include many proteins with novel folds and unassigned functions.[3] Identifying the function of proteins with novel folds will be the most challenging task,[4] yet is essential for the success of the structural genomics initiative.

Traditionally, newly discovered proteins are assigned function by searching the databases for homologous proteins with known function.[5] Detection of conserved sequence patterns (motifs) is also used to determine to which known protein family a new sequence belongs and to infer its biological function.[6] However, in the absence of sequence identity to any known protein sequence or motif, predicting function remains a challenge.

It is sometimes possible to predict function based on fold similarity. Many cases have been noted in which proteins with no detectable sequence identity share a common fold as well as a similar function. One example is the structural similarity between the human replication protein A and the human mitochondrial SSB protein,[7] both of which are involved in the binding of single-stranded nucleic acids (NAs). Koppensteiner and co-workers[8] have estimated that roughly two-thirds of proteins with similar topologies carry out similar biological functions. However, the relationship between fold similarity and functional similarity is not always straightforward, and there are many examples where proteins with similar folds carry out very different functions, such as different members of the TIM barrel fold family or SH3 fold proteins.[9,10]

Other techniques have been developed to predict local functional sites, such as enzyme active sites, in proteins. This has been achieved using methods to analyze conserved sequence clusters,[11–15] the shape and size of clefts in proteins,[16,17] hydrophobic patches,[18,19] specific arrangements of

Abbreviations used: NA, nucleic acid; dsDNA, double-stranded DNA; NN, neural network; GLM, generalized linear model; HTH, helix-turn-helix.

E-mail address of the corresponding author: gregoret@chemistry.ucsc.edu

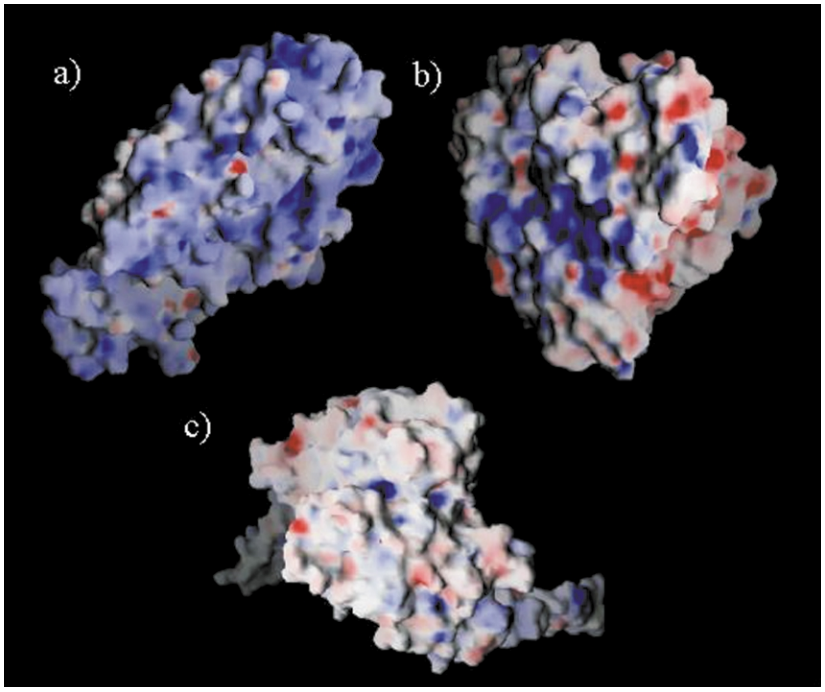

Figure 1. The electrostatic surface potentials of three proteins (color ranges from blue = positive potential, to red = negative potential). (a) The TATA-binding protein (1tbp) shows a large positive patch characteristic of NA-binding proteins. (b) The P450-CAM (1cpt) has a large positive patch that is not involved in NA binding. The patch is known to be involved in binding other proteins involved in electron transfer. (c) Seryl-tRNA synthetase (1ses) is a NA-binding protein that has no obvious positive patch. Images were taken from the Columbia Picture Gallery (http://trantor.bioc.columbia.edu/GRASS/surfserv_enter.cgi) generated by GRASP.[37]

catalytic residues[20–23] or specific residues with perturbed pK_a.[24] It also appears possible to predict function based on more global structural properties. For example, we have studied the protease family and shown that proteases as a whole have unique structural features common across different sequence and mechanistic families.[25] These properties include smaller surface areas, higher packing density, and a secondary structure bias against α-helices. We speculate that these similarities arose in the proteases as a common mechanism to avoid autolysis. The unique properties of proteases were used to train a neural network (NN) to successfully identify proteases in the absence of any homologs in the training set.

Recently we applied a similar approach to the O-glycosidase family. The O-glycosidases hydrolyze the linkage between carbohydrate molecules and are a structurally diverse enzyme family.[9] Members of this family were found frequently as false positives in the protease prediction. By identifying structural properties unique to the O-glycosidases, we have been able to discriminate them from the proteases and to predict new members of the family.[26] We therefore expect that similar methods could be developed for other important protein families. Such methods based on global structural properties could complement existing sequence and fold-based prediction methods.

We now focus on the identification of nucleic-acid-binding (NA-binding) proteins. When the first three-dimensional structures of protein–DNA complexes were solved, it was noted that the surface residues of the proteins were asymmetrically distributed, with an excess of positively charged side-chains at the NA-binding interface.[27] Charge complementarity between protein and DNA is thought to be important for the first step of recognition between protein and DNA *in vivo*.[27,28] Lysine and arginine residues are often involved directly in the recognition, interacting with the negatively charged backbone as well as with the bases themselves.[29–31] Observation of large regions of positive electrostatic potentials on protein surfaces has been suggested to be a good indication of the locations of DNA-binding sites.[32] Recently, the putative biochemical function of the Tubby protein family was assigned by using the crystallographic structure of one member of the family.[33] This protein structure has a large positively charged electrostatic patch, which helped to classify it as a NA-binding protein.

While the detection of a region of positive electrostatic potential can be indicative of NA-binding function, it is certainly not sufficient. Positively charged surfaces can be important for other reasons, such as for binding to negatively charged membranes,[34] receptors,[35] or other proteins.[36] Figure 1 illustrates the electrostatic surfaces of three proteins generated by the GRASP program.[37] The TATA-binding protein shown in Figure 1(a)[38] has a large positive patch at the DNA-binding site. P450-CAM (Figure 1(b)) is an oxygenating enzyme whose large positive surface patch is thought to play a role in binding cytochrome b_5 and putidaredoxin proteins that are involved in electron transfer.[39] It is not known to bind to NAs. Figure 1(c) shows the surface of a single-stranded NA-binding protein, seryl-tRNA synthetase.[40] Here, no obvious positive patches are found on the surface even though this protein binds to tRNA.

New NA-binding proteins continue to be discovered, including those with novel-binding

Table 1. The list of the 54 representative dsDNA-binding proteins grouped according to structural homology[42]

Structural group	Representative proteins
HTH group	3cro, 1fj1, 1ber, 1pnr, 1fok[a], 1gdt*[a], 1hcr[a], 1ign, 1pdn, 1tc3*[a], 1trr*, 1ddn, 1if1, 1vol, 3hts, 1bc8
Zinc-finger group	1aay, 1d66, 2nll, 1tup
Zipper-type group	1a02, 1am9
Other α-helix group	2bop*, 1aoi, 1b3t, 1ckt, 1crz, 1mnm, 1skn
β-Sheet group and β-hairpin/ribbon group	1ytb, 1cma, 1ecr, 1ihf, 1xbr, 1bf4*, 1bdt
Other	1a3q, 1bf5
Enzyme group	3mht*, 3pvi, 1rva, 1qps*, 3bam*, 1vas, 2dnj*, 1cw0, 1bpy, 2bdp, 1t7p, 1hmi, 1ssp*, 1bnk, 1a73, 1a31

A PDB ID with an asterisk indicates that it was incorrectly identified as a non-NA-binding protein.

[a] Denotes proteins with enzymatic activity that do not belong to the enzyme group.

motifs.[41] Based on comparative sequence analysis, it has been estimated that approximately 7–10% of the human genome codes for transcription factors. In addition, genes encoding proteins involved in translation, replication, and splicing (which are thought to be under-represented in the present structural data[41]) are likely to be found as well. In all, perhaps as many as 2500 of the structures solved through structural genomics will be NA-binding proteins. Automated methods for function discrimination will therefore be important for annotating these structures.

Here, we propose a methodology for predicting NA-binding function based on the quantitative analysis of structural, sequence and evolutionary properties of positively charged electrostatic surfaces. Since our method does not require that the protein have a homologous sequence or structure in the databases, it can predict the function of proteins with novel folds and/or unique binding motifs. We show that it is possible to distinguish between actual DNA-binding proteins and proteins that have large positive electrostatic patches but do not bind NAs. Finally, we address the utility of electrostatic patch analysis in assigning NA-binding function to products of structural genomic studies.

Results and Discussion

Our initial goal was to simply distinguish NA-binding proteins from proteins that do not bind NAs. We first characterized the structural features of 54 double-stranded DNA (dsDNA)-binding proteins and compared them to the same features in other proteins that are not involved in NA binding. We concentrated on the positive electrostatic surface patches that are common among NA-binding proteins. In order to be able to validate the patch analysis, we focused our study on DNA-binding proteins that were solved in complex with the DNA and thus have well-defined binding interfaces.

Data set construction

A representative data set of DNA-binding protein structures was constructed based on the protein–DNA complex classification of Luscombe *et al.*[42] The set consisted of the 54 proteins that bind dsDNA with crystallographic resolution better than 3 Å. The proteins in the data set represent 54 different structural families classified based on the structure alignments of the DNA-binding region.[42] From each of the 54 structural families, a representative protein was selected based on lowest resolution and available sequence alignments. The final data set was further examined to make sure that no two proteins shared more than 35% sequence identity. The 54 proteins (Table 1) were further divided into eight groups: seven structural (e.g. helix-turn-helix, HTH) and one enzyme group. A second data set consisting of 250 non-nucleic-acid-binding (non-NA-binding) proteins with resolution better than 2.5 Å was also constructed (see Materials and Methods). For the non-NA-binding proteins, the only other criterion besides resolution to qualify for inclusion in the data set was that the protein did not interact with an NA polymer.

Electrostatic patch concept

In the first step of our analysis we constructed and analyzed positive electrostatic patches for all proteins in the data sets. The UHBD software package[43] was used to compute the Poisson–Boltzmann continuum electrostatic potential at all protein surface points. An in-house software program, PatchFinder, was used to construct the patches (see Materials and Methods). Each protein's largest patch was then analyzed to extract structural and sequence information.

Patch size and the average electrostatic potential were calculated for both the dsDNA and non-NA data sets. Patch size represents both the protein size and its electrostatics. In many protein–DNA complexes, only the DNA-binding domain was crystallized. Therefore, we used the absolute patch size instead of normalizing by the protein's full surface area, which in many cases is not known. Figure 2 shows the number of surface points in the largest positive patch against the normalized surface potential of that patch for each protein. Although dsDNA-binding proteins have, on average, larger positive patches than the non-NA-binding proteins (637(±476) surface points *versus*

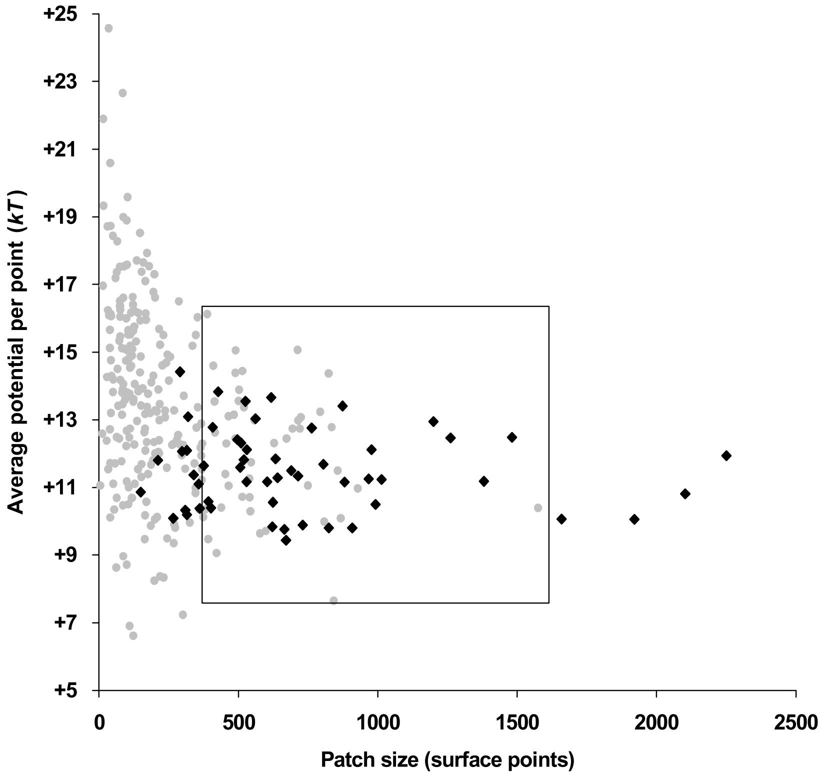

Figure 2. Plot of patch size *versus* average surface potential for NA-binding proteins (◆) and non-NA-binding proteins (●). NA-binding proteins on average have larger positive patches (637(±476) surface points) than non-NA-binding proteins (243(±222) surface points). Several non-NA proteins have patch sizes comparable to that of NA-binding proteins. The largest positive patched non-NA-binding proteins (enclosed in the rectangle), have comparable patch sizes and potentials to that of NA-binding proteins.

243(±222) surface points), they are not always clearly delineated.

To obtain finer separation between DNA and non-NA-binding proteins, a second type of electrostatic patch, the "Lys_{off}" patch, was analyzed. Nadassy and co-workers have shown that of the residues that interact with DNA, roughly one quarter are arginine.[44] To examine whether the high frequency of arginine could be used to identify and characterize the DNA patches, we made computer mutations of all lysine residues in the protein to their hydrophobic isosteres and then re-analyzed the new resulting patches. Hydrophobic isosteres of Lys (Lys_{off}) are structurally identical to Lys but have a net charge of 0.[45] If a patch is dependent upon a large number of lysine residues to give a local net positive charge, such patches should "dry up" when the lysine charges are turned off. Arginine-rich patches, on the other hand, will be retained. The size of (Lys_{off}) patches indicates not only the number of arginine residues in the patch, but also their distribution over the patch. A specific example is shown in Figure 3. Here, the yeast transcription factor, MCM1, retains a large percentage of its patch even after the lysine charges are turned off while the patch of the non-NA-binding protein, cytrochrome *c*2, shrinks dramatically.

Patch and protein-binding site correlation

It has been shown previously that a typical DNA–protein interface spans 1600(±400) Å^2, a considerable portion of the total protein surface.[44] In order for electrostatic patch analysis to be useful for identifying and annotating NA-binding proteins, there should be a strong correlation between the locations of the patches and the locations of the protein–NA-binding interfaces. We indeed found that the largest positive patch of

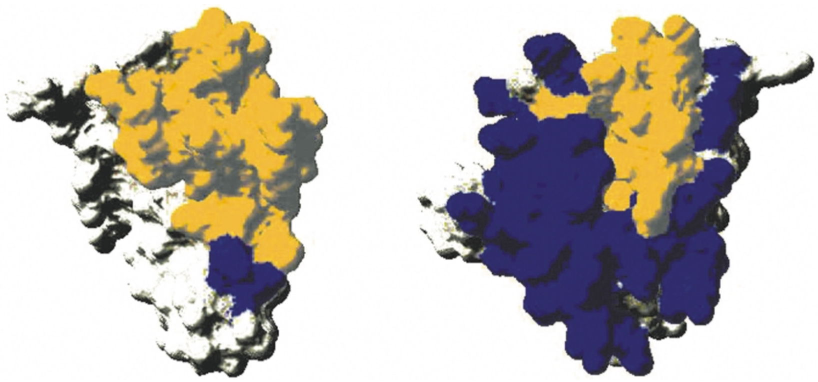

Figure 3. Surface electrostatic patches of two different proteins: on the left, a DNA-binding protein MCM1 (1mnm), and on the right, a non-NA-binding protein, cytochrome *c*2 (1cot). The continuum positive patch is shown in dark blue, and the overlap with its respective Lys_{off} patch (see the text) is colored in yellow. In this example the DNA-binding protein (left) retains most of its patch, while the non-NA-binding protein (right), shows a dramatic decrease in patch retention within its Lys_{off} patch.

a dsDNA-binding protein encompasses, on average, 80% of the protein–DNA interface. There was no overlap between the largest positive patch and the DNA–protein interface in only three of the 54 proteins (1gdt, 1ssp, 2dnj). These outliers typically have a large cluster of positive surface charges not involved in NA binding as defined by their three-dimensional crystal structure complexes. Based on the high degree of overlap between the patch and DNA interface, we conclude that patch analysis could be useful as a quantifiable, fold-independent feature for DNA-binding protein classification as well as binding site identification.

Structural features of patches

The next step after defining the electrostatic patches was to find the best features that could be used to discriminate the DNA-binding proteins from other proteins.

Secondary structure content

Since most DNA-binding proteins bind *via* an α-helix interacting with the major groove of DNA,[29,42] we hypothesized that the secondary structure content of the residues within the positive electrostatic patches would be biased towards helical structure. Upon measuring the α-helical and β-sheet composition of the patches, helical composition was indeed found to be favored over β-sheet composition (43% helix *versus* 15% sheet) to a considerably greater degree than in the non-NA-binding data (32% helix *versus* 22% sheet).

Surface area

Lewis and Rees proposed that a greater surface area allows more possibilities for van der Waals contacts and therefore could be associated with regions involved in ligand binding.[46] Since van der Waals contacts are common in protein–NA interactions,[47] we computed the average accessible surface area per residue within the patch. We found that the residues that participate in the NA-binding protein patches possess slightly larger (although not statistically significantly larger) than average surface areas (70(±14) Å^2) compared to patches of non-NA-binding proteins (61(±15) Å^2). This measurement is dependent on both the amino acid composition within the patch and the accessibility of the residues.

Hydrogen-bonding potential

Thornton and colleagues have shown that the DNA–protein interfaces are dense in residues with hydrogen-bonding capacity.[31] Protein–NA interfaces also contain more hydrogen bonds per square angstrom than protein–protein interfaces.[44] We calculated the total number of potential hydrogen bond donors or acceptors within our patches. In agreement with the high concentration of hydrogen bonds found in protein–NA interfaces,[31] we found that the electrostatic patches of DNA-binding proteins have on average more potential hydrogen-bonding groups (20(±12) donors and 25(±14) acceptors) compared to patches of non-NA-binding proteins (8(±7) donors and 8(±7) acceptors).

Surface concavity

In principle, a cleft within a protein could provide a likely location for docking a ligand.[17] Thus, we anticipated that the overlap between the electrostatic patch and a significant protein cleft could potentially help to identify DNA-binding proteins as well as binding sites. The program, SURFNET[48] was used to define the largest clefts in our proteins. The overlap between residues within the cleft (as defined by SURFNET) and the residues within the largest positively charged patch were then examined for DNA-binding and non-NA-binding proteins. This is exemplified in Figure 4, using the *Hha* I methyltransferase protein (3mht). In this example, the residues of the cleft (yellow) and the residues of the patch (blue) overlap exactly (green) where the DNA molecule (pink) docks against the protein. For each protein in our dataset we measured the percentage of the overlap between the two largest clefts and the largest positive patch. The percent overlap was found to be significantly higher among the dsDNA-binding data set (55(±25)%) as compared to the non-NA-binding data set (19(±17)%).

Amino acid frequency and composition

Due to the negatively charged nature of DNA, charged and polar amino acid residues are most

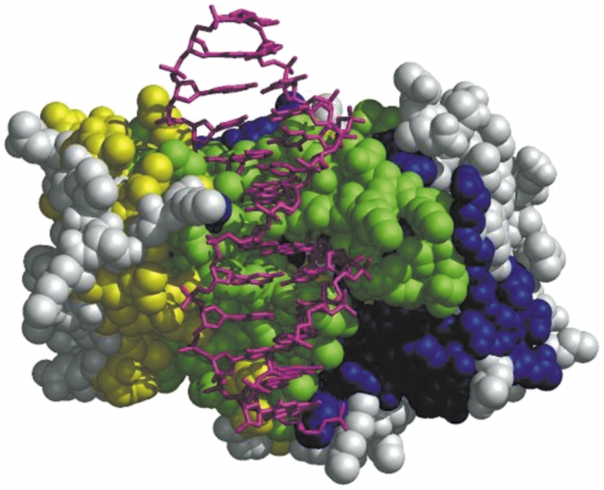

Figure 4. A colored representation of the protein surface of the *Hha* I methyltransferase (3mht). The residues surrounding the largest cleft on the protein's surface (as defined by the program SURFNET[48]) are colored in yellow and the residues involved in the positive electrostatic patch are colored in blue. The overlap region between the patch and the cleft is colored in green. Notice that the overlap region (green) plots exactly to the DNA-binding site.

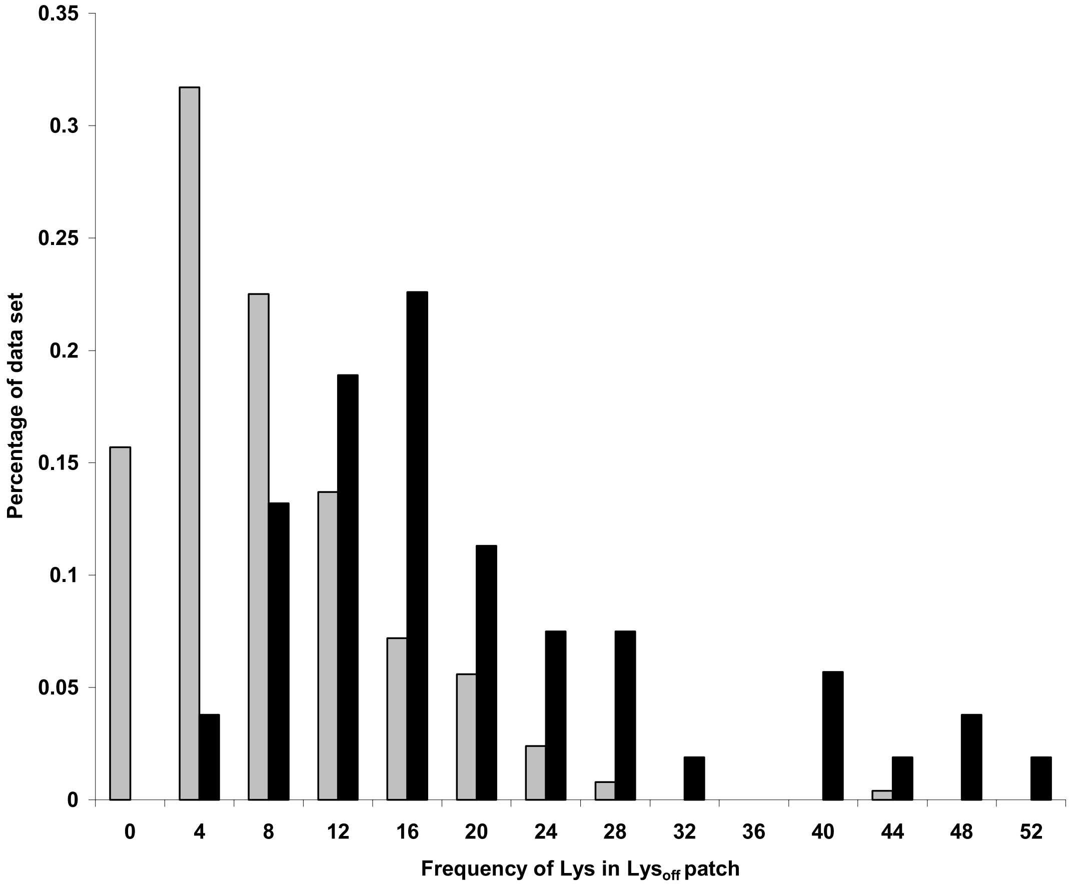

Figure 5. Frequency of Lys in Lys$_{off}$ redefined patches for the dsDNA-binding (black) and non-NA-binding (gray) sets.

common in protein–DNA interactions.[44] To examine whether the positively charged patches of DNA-binding proteins and non-NA-binding protein differ in the frequencies of polar residues, we analyzed the sequence composition of the Lys$_{off}$ redefined patches in both the dsDNA and non-NA data sets. As expected, there was a larger number of polar and arginine residues within the dsDNA-binding Lys$_{off}$ patches (26(±19) residues) compared to non-NA-binding patches (11.5(±9) residues), however, the differences were not significant. As shown in Figure 5, a distinguishing feature between the two data sets was the high number of (uncharged) lysine residues retained in the patches. On average Lys$_{off}$-dsDNA-binding patches had 8.5(±14) lysine residues while non-NA-binding patches had 2.5(±3).

Sequence conservation

Sequence conservation is generally high among protein functional sites.[12,49–51] We therefore examined to what extent the residues that compose the patch are preserved among other proteins in the family. Functional sites involved in DNA binding are expected to show some degree of sequence variability in order to bind specifically to unique DNA sequences. Recently, Luscombe and Thornton showed that, based on a data set of 21 DNA-binding protein families, protein residues that are in contact with the DNA are usually better conserved than the rest of the protein surface.[52]

We extracted for each protein in our dataset a sequence alignment of closely related proteins using PSI-BLAST version 2.1.1.[53] To reduce redundancy, sequences with more than 90% identity were eliminated. In addition, in order to restrict the alignment to proteins with similar function, we only included sequences with at least 35% identity. When we calculated the average conservation of all patch residues based on the sequence variation entropy,

$$-\sum_{a} P_{ai} \log P_{ai}$$

(where P_{ai} is the probability (observed frequency) of amino acid type a (a, represents the 20 different amino acids) in column i in the alignment), we found that in both DNA-binding proteins and non-NA-binding proteins the positive electrostatic patches are generally less variable (more conserved) than the rest of the surface residues. In the next step we concentrated on the positively charged and aromatic residues in the defined electrostatic patch that are likely to be involved in NA binding. Each of the positive and aromatic residues in the patch was characterized by two criteria: (1)

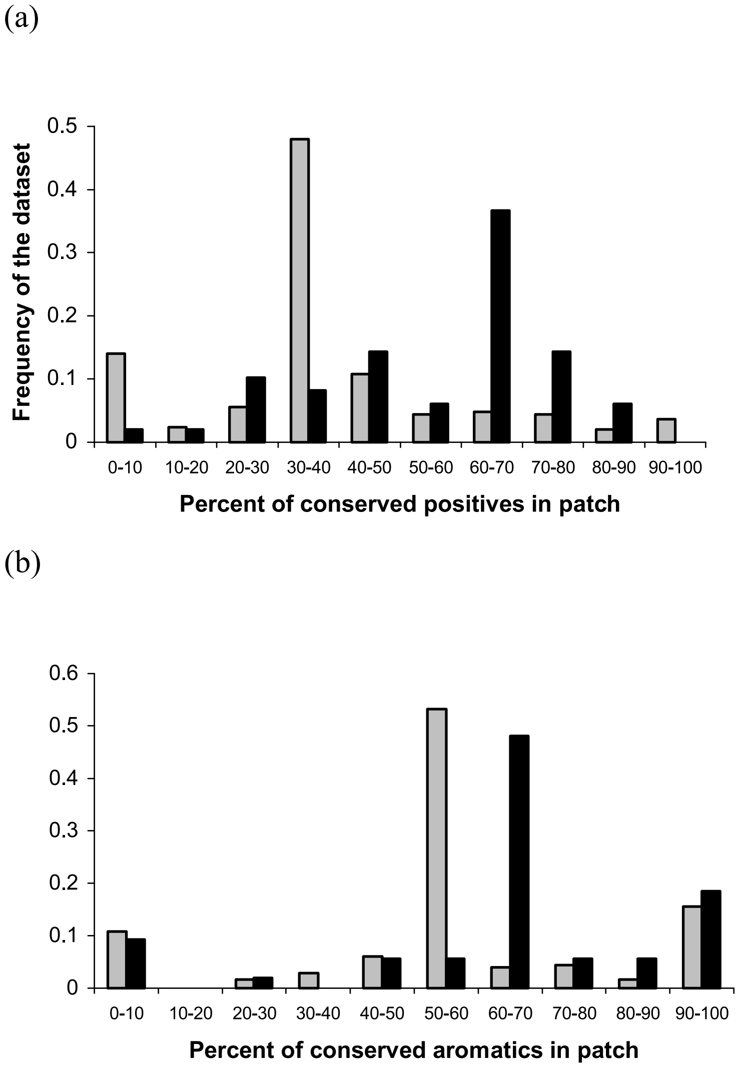

Figure 6. Percent of residues in the electrostatic patches that were conserved at the property level (a) positive charge (b) aromaticity, in dsDNA-binding proteins (black) and non-NA-binding proteins (gray). A residue was considered conserved at the property level if >75% of the sequences in the MSA retained the same property as in the representative structure. The *y*-axis describes the frequency of proteins in the dataset that yielded these fractions of conserved residues in their electrostatic patches.

whether the specific amino acid is retained along the alignment, (2) whether the general property of the residue is retained. Residues were considered to be "conserved" when more than 75% of the residues in an aligned column contained the amino acid type as in the patch (e.g. arginine). "Conservation of property" required only that the general property of the amino acid (e.g. positive charge) is retained in more than 75% of the sequences. The electrostatic patch was further characterized by the percent of the residues in the patch that were conserved at both the amino acid and property level.

Figure 6 shows histograms of the percent of conserved positive and aromatic residues in the electrostatic patches. As can be seen by the shift of the histograms of the dsDNA-binding proteins to the right relative to the non-NA-binding proteins, the electrostatic patches of the dsDNA-binding proteins have a higher percentage of aromatic and positive residues that are conserved at the property level. At the residue level, we only found a distinguishable (but not statistically significant) difference between the percent of conserved arginine in patches of dsDNA-binding proteins (60(±20)% arginine residues conserved in patch) compared to other non-NA-binding protein patches (39(±35)% arginine residues conserved in patch).

Predicting DNA-binding proteins

As summarized in Table 2, most of the sequence/structure patch features discussed above show distinguishable differences between dsDNA-binding proteins and non-NA-binding proteins. When considered alone, these parameters are not sufficient to discriminate between the two groups of proteins. In order to examine the cumulative effect of the observed trends, we used the 12 parameters to train an NN to predict dsDNA *versus* non-NA-binding proteins. The NN uses a standard feedforward, error back-propagation algorithm with two-layer architecture containing three hidden nodes in the first layer and a single output node (see Materials and Methods). The output of the NN gives the probability that a given structure is a dsDNA-binding protein.

Table 2. Summary of 12 features analyzed for dsDNA-binding proteins and non-NA-binding proteins

Feature studied	dsDNA-binding proteins	Non-NA-binding proteins
Molecular weight/residue	115 ± 3	111 ± 4
Patch size (surface points)	637 ± 436	243 ± 222
Percent helix in patch	43 ± 19	32 ± 24
Average surface area/res	70 ± 14	61 ± 15
Hydrogen-bonding potential—average number of donors	20 ± 12	8 ± 7
Hydrogen-bonding potential—average number of acceptors	25 ± 15	8 ± 7
Percent patch/cleft overlap	55 ± 25	19 ± 17
Number of lysine isostere in $\mathrm{Lys_{off}}$ patches	8.5 ± 14	2.5 ± 3
Number of polar residues in $\mathrm{Lys_{off}}$ patches	26 ± 19	11.5 ± 9
Percent of conserved arginine residues	60 ± 20	39 ± 35
Percent of conserved positive residues	54 ± 25	35 ± 36
Percent of conserved aromatic residues	66 ± 32	56 ± 39

One-by-one (jackknife) testing

To cross-validate our results and to prevent over training, a total of 304 NNs were trained and tested, one for each member of the dsDNA and the non-NA data sets. Each member of the data sets was withheld from the training set in turn for testing. The distribution of the NN predicted values for the DNA-binding proteins and the non-NA-binding-proteins are shown in Figure 7(a). At the threshold of 0.5 the NN correctly classified 44 out of 54 dsDNA-binding proteins (81%) and 236 out of 250 non-NA-binding proteins (94%). Of the ten dsDNA-binding proteins predicted erroneously to be non-NA-binding proteins, seven had enzymatic activity (six belonging to the enzyme group and two to the HTH group). The three other dsDNA-binding proteins we missed were the Trp repressor (1trr), the Papillomavirus E2 protein (2bop) and the hyperthermophile chromosomal protein Sso7d (1bf4). Since the majority of the NN errors were enzymes, we decided to train a separate NN on the 34 proteins that do not have enzymatic activity. A cross-validation test was performed on the reduced data set; the results are shown in Figure 7(b) and are summarized in Table 5. Though at 0.5 cutoff, the number of false negatives is still high, from Figure 7(b) we can see that eliminating the enzymes from our data results in a much better separation between the DNA set and the non-NA set. At a cutoff of 0.26 where we obtain the optimal separation between the DNA and non-NA sets, only two out of 34 DNA-binding proteins are predicted to not bind DNA (94% were predicted correctly) and 232 out of the 250 non-NA-binding proteins are predicted correctly as non-NA-binding (93%).

Based on the number of correctly predicted dsDNA-binding and non-NA-binding proteins the correlation coefficient[54] value was calculated:

$$C = \frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FN} \times \mathrm{FP})}{\sqrt{(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})}}$$

TP and FP are the number of true and false positives, respectively, while TN and FN are the number of true and false negatives, respectively. C ranges from -1 to 1, with $C = 1$ for a perfect prediction, $C = 0$ for a random prediction, and $C = -1$ for an incorrect (opposite) prediction. The correlation coefficient for the full 54 DNA-binding protein set was $C = 0.74$.

Binding motif independence testing

Though the 54 DNA-binding proteins tested above represent different structural families, within the structural groups they still share common binding motifs (e.g. HTH) and thus they could have common structural properties.[55,56] In some cases, the proteins in the same structural group also belong to the same SCOP family (e.g. the engrailed homeodomain (2hdd) and the Hin recombinase (1hcr)). In order to be confident that our NN predictions are not a result of having several proteins sharing the same binding motif within our data set, we tested whether the favorable results are at all influenced by redundancy. Since there is a limited number of known binding motifs used by DNA-binding proteins for generating a non-redundant motif set we withheld, in turn, all members of a particular structural group (motif) and then trained on the remaining proteins from all other groups. We then tested each member of the family on the trained NN in which that group was completely withheld. (An analogous experiment would be used to predict the function of novel proteins for which neither sequence nor fold similarity are available. A robust predictor should be capable of this.)

The motif test showed very similar results to our original jacknife (one-by-one) test, indicating that our method does not rely on having a related protein sharing a similar motif in the training set for its prediction. Table 3 summarizes the NN predictions for the different binding motifs using the motif test and the jackknife test. As shown, even when the most common motif (the HTH) was withheld, the NN still correctly identified 10 of 16 proteins (three of the six proteins missed are again enzymes). This is compared to 13 of 16 proteins correctly predicted in the jackknife test.

The only protein group in which we misclassified 50% of the proteins is the enzyme group. (This group also performed poorly in the jackknife

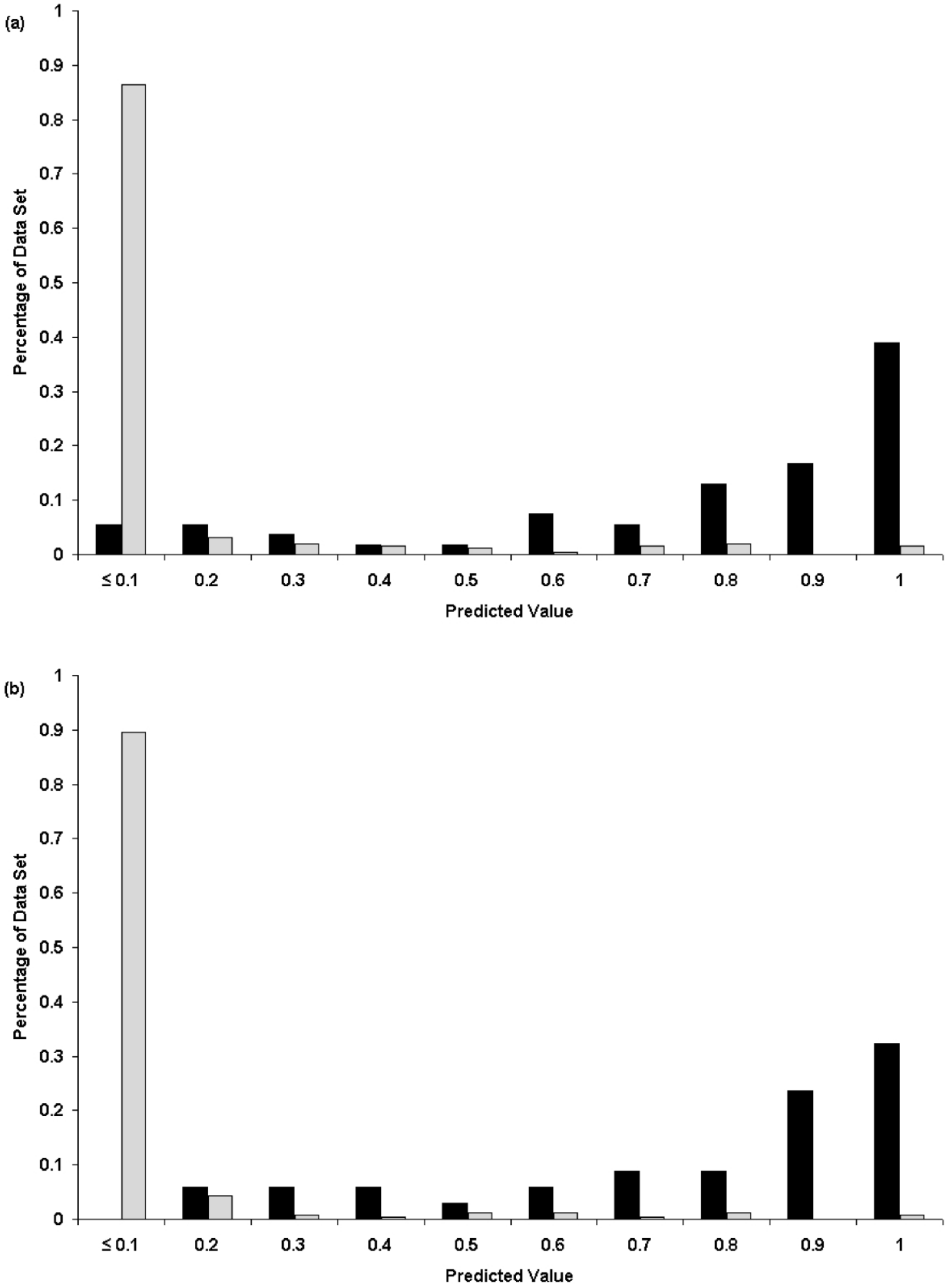

Figure 7. Percentage of the data set for DNA-binding (black bars) and non-NA-binding (grey bars) proteins *versus* their predicted value from the neural net: (a) training and testing on all 54 DNA-binding proteins, (b) training and testing on 34 DNA-binding proteins not including enzymes.

test.) Among the enzymes misclassified are the glycosylases, which typically have negative patches associated with their active sites. It is not difficult to imagine how this would confound our patch analysis. Since other enzymes as well as proteins from the "other" group were also misclassified in both the cross-validation and motif test, we assume that they possess different structural properties from other transcription factors. In order to be able to predict the enzyme group correctly it will be necessary to train a separate NN on enzyme proteins only. This will be achievable once more

Table 3. A summary of the NN predictions for independent binding motifs

Binding motif	Motif test	Generic test
HTH	10/16	13/16
Zinc-finger	4/4	4/4
Zipper-type	2/2	2/2
Other α	6/7	6/7
β-Sheet and β-ribbon	6/7	6/7
Other	2/2	2/2
Enzymes	8/16	11/16

Numbers represent the number of proteins in a specific motif group that were correctly predicted over the total number of proteins within that particular motif group.

DNA-binding enzymes are solved. The overall ability to predict motifs and folds withheld from training proves that the success of our NN predictions does not rely on having several proteins from the same binding motif in our training set. These results also bode well for the ability of our method to detect a novel fold in a real structural genomics setting.

Neural net validation

To better understand which features were most important for distinguishing dsDNA-binding proteins from non-NA-binding proteins, we tested what parameters were most sensitive in terms of NN performance. The NN analysis was repeated 12 times, eliminating one input parameter (e.g. patch size) at a time. Again, the NN was tested using the cross-validation scheme described above. Based on the overall performance of the NN, when eliminating the individual parameters, we found that the most relevant inputs within the NN are (in order of importance): (1) surface area per residue, (2) the percentage overlap between the cleft and the patch, (3) molecular weight per residue within the patch, (4) the number of lysine residues retained in a $\mathrm{Lys_{off}}$ patch, (5) the overall patch size, (6) sequence conservation of positive charges within the patch and (7) the number of hydrogen bond acceptors in the patch. Although other inputs, such as the secondary structure of patch residues, were not as relevant, when we tried to eliminate them, the overall prediction performance decreased and therefore they were retained.

In order to compare our results against another predictive technique, we also tested the ability of a generalized linear model (GLM) to predict the probability that a structure is an NA-binding protein. A GLM, like an NN, is used to describe the relation among variables (parameters) and possibly predict the response (outcome). Unlike NNs, GLM models possess no hidden layers and therefore can be easier to interpret.[57] By applying the GLM, using the log-odds "logit" link function, our prediction results were slightly lower than the NN predictions, correctly classifying 72% of the dsDNA and 96% for the non-NA-binding sets ($C = 0.70$). As for the NN analysis we repeated the GLM analysis 12 times, eliminating each of the 12 parameters in turn. We again found that the overall prediction performance of the GLM was most sensitive to removal of the following features: surface area per residue, the number of lysine residues retained in a $\mathrm{Lys_{off}}$ patch, the overall patch size, and the molecular weight per residue within the patch. In general, the performance of the GLM was much less sensitive than the NN to removing inputs. However when we tried to take out more than one parameter, reducing the input set to ten, the performance of the GLM was severely reduced. Overall, the similarity of the GLM results to the NN results and the high prediction accuracy of the GLM imply that the parameters chosen for this analysis are well suited to the data.

Large-patched non-NA-binding proteins

While the overall ability of the NN and GLM to predict dsDNA-binding function was encouraging, we would expect the most difficult challenge to come from proteins that have large, positively charged patches but that do not bind NA. We therefore examined the predictive outputs of the NN on a subset of non-NA proteins with large, positively charged patches. The non-NA-binding proteins with large positive patches are shown within the rectangle in Figure 2. Among this subset are oxidoreductases, transferases, electron transport and hydrolases. The average number of positive surface points in the non-NA-binding proteins subset (including 54 proteins) was 591(±215), which is comparable to the average number of surface points in the dsDNA-binding patches (637(±476)). The overall electrostatic potential of the patches in these two data sets was also very similar, with a slightly higher average potential median of 12.0(±1.8) in the non-NA subset *versus* 11.5(±1.2) in the DNA set.

Although both the DNA set and the non-NA subset had similar patch sizes, using the full range of parameters, the NN was able to accurately identify 44 of the 54 large-patched non-NA-binding proteins as non-NA-binding proteins. Similar results were obtained with the GLM (see results summary in Table 5). These results again suggest that we are analyzing important structural and sequence features of the patches that are unique to dsDNA-binding proteins.

Detecting unbound NA-binding proteins

In an actual structural genomics application, if the tested protein has an unknown function, its three-dimensional structure would most likely be solved without its NA ligand. Since many DNA-binding proteins have been shown to undergo conformational changes upon NA binding (summarized by Nadassy *et al.*[44]), our predictions on the protein components of protein/DNA

Table 4. A list of the unbound NA-binding proteins tested

PDB ID	Protein description	Protein classification	Resolution (Å)
1enh	Engrailed homeodomain	Homeodomain	2.1
1etc	ETS domain	Winged helix	NMR
1hcp	Estrogen receptor	Zinc-finger	NMR
1tbp	TATA-binding protein	Beta ribbon	2.6
2cro	434 cro repressor	HTH	2.3
2hts	Heat shock protein	HTH variant	2.2
1bix	DNA repair endonuclease	Enzyme	2.2
1bm8	Mbp1 protein	Winged helix novel	1.7
1c8z	Tubby protein	Novel	1.9
1dm9	Heat shock protein 15	Novel RNA binding	2.0
1h5p	Sp100B SAND domain	Novel	NMR

complexes could be misleading. To test for this, we constructed a representative data set of protein structures that are known to bind double-stranded NAs, but that were solved in a free (unbound) state (ubNA). The data set, shown in Table 4, contained a representative protein for each of the common binding motifs (excluding the zipper type group for which there was no complete structure available), one enzyme, and three proteins with novel-binding motifs. Selected NMR structures were also included.

We then trained the NN with the original dsDNA-binding protein data set (composed of proteins from protein–DNA complexes). Since some of the proteins in our ubNA set had closely related proteins in the training set, for each unbound protein tested, we withheld from the training set all homologs. We also tested the ability to predict the unbound proteins when we withheld from the training set all proteins belonging to the same motif family (similar to the motif test described above). This test was done only for proteins that had a known binding motif.

Overall, when testing on the ubNA protein data set using either the NN or the GLM, we correctly predicted ten out of 11 unbound proteins. The only protein mispredicted is the DNA repair endonuclease (1bix). Among the proteins that we predict correctly are three proteins that have novel-binding motifs. One protein (1dm9) was solved as part of the structural genomics initiative and is classified in the PDB as a hypothetical protein. For comparison, when we ran a BLAST[53] search on each of these three proteins, we did not detect any known DNA-binding proteins among the significant hits. The NN and GLM results suggest that although the DNA-binding proteins can undergo conformational changes upon binding, the overall electrostatic patch properties of the proteins in the free state are similar enough to the properties of the protein in the bound conformation. In addition, this experiment shows that our method is also capable of correctly classifying novel proteins as well as NMR structures. Based on these encouraging results, we believe that this methodology has the potential to predict novel structures resulting from the structural genomic project. In the future it may even be possible to predict the NA-binding function of low-resolution structures generated from predicted models.

Conclusions

The positive electrostatic patches of NA-binding proteins appear to have unique properties that allow one to discriminate them from other proteins that are not involved in DNA binding, specifically those with similar, large, positively charged patches. These properties include molecular weight per residue, patch size, percent α-helix in patch, average surface area per residue, number of residues with hydrogen-bonding capacity, percent of patch and cleft overlap, number of lysine and polar isosteres in Lys_{off} patches, and percent of conserved positive and aromatic residues in the patch. A summary of the properties and their values for both the dsDNA-binding and non-NA-binding proteins is given in Table 2.

Although each one of these properties individually was not sufficient for predicting whether a certain protein is a DNA-binding protein or not,

Table 5. Summary of the NN and GLM predictions for the different testing sets (based on 0.5 threshold)

	Neural network (NN)					Generalized linear model (GLM)				
Test set	TP	TN	FN	FP	C	TP	TN	FN	FP	C
All protein	44	236	10	14	0.74	39	239	15	11	0.70
All proteins–no enzymes	27	232	7	18	0.64	25	243	9	7	0.72
Large patch	44	44	10	10	0.63	39	46	15	8	0.67

TP, true positive; TN, true negative; FN, false negative; FP, false positive; C, correlation coefficient.

in combination they were used successfully to train an NN as well as a GLM to discriminate the dsDNA-binding proteins from other non-NA-binding proteins, including a subset of non-NA-binding proteins that have large positive electrostatic patches usually characteristic of NA-binding proteins. The method is also able to detect proteins with novel-binding motifs, which were never seen by the NN or the GLM. Moreover, we found that it is possible to classify NA-binding proteins that were solved in a free state, implying that these properties are intrinsic properties of the proteins and they are not just a result of the NA binding. These results are summarized in Table 5.

As a new step, we envision refining this method to make more detailed predictions on a protein's function. For example, it would be useful to know whether or not a protein is a transcription factor, does a given protein bind to single or double-stranded NA. Our preliminary (unpublished) results indicate that the features of single-stranded RNA and DNA-binding proteins are subtly different and that it is possible, at some level, to discriminate between single and double-stranded NA-binding proteins.

Electrostatic patch analysis as implemented here, may also be useful for classifying other types of proteins. We have reported that O-glycosidases, which are known to possess negatively charged electrostatic surfaces, may be identified using similar methods.[26] We propose that classification of protein properties based on their common structural and sequence features could eventually lead to an automated function assignment for many other protein families. Eventually, it may be possible to combine multiple classification schemes in order to improve overall function predictions of novel protein structures resulting from the Protein Structure Initiative.

Materials and Methods

Data set construction and calculations

The DNA-binding protein data set used here was constructed based on the protein–DNA complex classification of Luscombe *et al.*[42] The coordinate files for the representative proteins of each structural family (based on resolution and available sequence alignment) were obtained from the nucleic acid database (NDB).[58] All NAs were removed and only single protein chains per complex were considered for calculations. No two proteins in the data set shared more than 35% sequence identity. Overall, 54 protein chains classified to eight structural groups were included (Table 1). All proteins in the dsDNA set were solved by X-ray crystallography with a resolution of better than 3.0 Å. Please see Luscombe *et al.*[42] for more detailed information about the proteins in this data set. The unbound set (ubNA) included 11 DNA-binding proteins solved in free state (Table 4). Among this set we also included three NMR structures.

The non-NA-binding protein data set was constructed from Hobohm and Sander's[59] "pdb select" list of proteins, excluding all proteins that could be involved in binding NAs. The data set was further cleaned excluding sequences with more than 25%. Here we used a more stringent sequence identity cutoff than in the dsDNA-binding set to compensate for the size of the data sets. Overall the non-NA-binding set included 250 proteins. The large, positive-patched non-NA-binding proteins were a subset of the non-NA-binding proteins, and included proteins that possessed an average patch size comparable to that of the DNA-binding data set. The list of 250 non-NA-binding proteins, including the 54 largest positive patches proteins is shown here: 1gnd, 1php, 1lki, 1mrp, 1mai, 1a8e, 1czj, 1a53, 1br9, 1ppn, 1pbv, 1fit, 1bdb, 1lid, 1rcb, 1cot, 1csh, 1a8p, 1atg, 1bg2, 1fmk, 1ndh, 1csn, 1nsj, 1oyc, 1lfo, 1ciy, 1gky, 1opr, 1axn, 1fds, 1drw, 1ido, 1frb, 1lam, 1pda, 1cpt, 1fnc, 1hcz, 1ayl, 1uae, 1a6o, 1gox, 1oaa, 1a7s, 1hcl, 1a17, 1pbe, 1skf, 1a53, 1af7, 1aj2, 153l, 1a1x, 1a44, 1a48, 1a6m, 1a6q, 1a8l, 1a8p, 1a8y, 1aac, 1abe, 1ac5, 1aew, 1ah7, 1aho, 1air, 1ajj, 1al3, 1alu, 1aly, 1amf, 1amk, 1amm, 1amp, 1amx, 1aoa, 1aol, 1aqb, 1arb, 1aru, 1ash, 1ass, 1at0, 1atg, 1auk, 1av4, 1az9, 1b0b, 1b5l, 1b6a, 1ba3, 1bb9, 1bd8, 1bdo, 1bea, 1beo, 1bfd, 1bg6, 1bgc, 1bhe, 1bhp, 1bj7, 1bk0, 1bob, 1bpi, 1bqk, 1bs9, 1btn, 1bv1, 1bx7, 1byb, 1c52, 1ca1, 1cec, 1cem, 1cfb, 1chd, 1clc, 1cnv, 1cpo, 1cpq, 1ctj, 1ctt, 1cv8, 1cvl, 1cyo, 1czj, 1ddt, 1dfx, 1dhn, 1dhr, 1din, 1doi, 1dpe, 1dun, 1dxy, 1eaf, 1ecy, 1edg, 1esc, 1eur, 1ezm, 1fce, 1fkj, 1fua, 1fus, 1g3p, 1gai, 1gca, 1gen, 1gof, 1gpr, 1gsa, 1gym, 1hfc, 1hka, 1hoe, 1hpm, 1htp, 1hxn, 1hyp, 1iae, 1ifc, 1inp, 1iov, 1jdw, 1jer, 1klo, 1koe, 1kpf, 1kte, 1kuh, 1lbu, 1lcl, 1led, 1lit, 1lki, 1lst, 1ltm, 1mai, 1mat, 1maz, 1mba, 1mla, 1moq, 1mpp, 1mrp, 1msk, 1mup, 1nar, 1neu, 1nfp, 1ng1, 1nif, 1nkr, 1nls, 1nnc, 1nox, 1npk, 1obr, 1ops, 1opy, 1osa, 1pbe, 1pbv, 1pdo, 1pea, 1pgs, 1phd, 1phk, 1phm, 1php, 1phr, 1pht, 1plc, 1pmi, 1pne, 1poa, 1poc, 1pot, 1prn, 1ptf, 1puc, 1ra9, 1rb9, 1rcb, 1rcy, 1rec, 1rfs, 1rh4, 1rhs, 1rie, 1rkd, 1rmg, 1rsy, 1rzl, 1sbp, 1sek, 1sfp, 1skf, 1smd, 1sra, 1svb, 1svy, 1tca, 1tde, 1ten, 1tfe, 1thv, 1tml, 1tmy, 1tn3, 1ton, 1try, 1tul, 1uch, 1uok, 1ush, 1utg, 1vls.

Patch calculations

The University of Houston Brownian Dynamics package, UHBD,[43] was used for all electrostatics calculations. Hydrogen atoms were added with the HBPLUS program.[60] The OPLS[61] parameter set was used to assign partial atomic charges and atomic radii. Probe radii of 1.4 Å and 2.0 Å were used to define the molecular surface and the Stern layer, respectively. The temperature was set to 298 K and the ionic strength to 150 mM. Dielectric constants for the protein and solvent were 2.0 and 80.0, respectively. A grid of $65 \times 65 \times 65$ with spacing of 2.0 Å centered at the same point for all proteins was used in all UHBD calculations, in all cases boundary smoothing was applied.

Continuum electrostatic patches were assigned by looking for adjacent surface points that met a given cutoff potential using an in-house program, PatchFinder, based on UHBD output. At first, protein surface points were determined by the UCSF MidasPlus software package.[62] The electrostatic grid points (extracted from the UHBD) were classified as either surface or non-surface points according to their proximity to the molecular surface. The non-surface points were ignored, creating a continuous surface of grid points. Adjoining surface grid points meeting a potential cutoff $\geq +2kT$ were grouped to define the positive patches. Patch sizes were defined according to the number of surface points within the patch. The biggest patch among all positive surface patches in a given protein was chosen to represent the protein.

Structural calculations

Secondary structure assignments were made with the DSSP program.[63] Accessible surface area was calculated by using the method of Lee & Richards[64] as implemented in the program CALC-SURFACE[65] with a default probe radius of 1.4 Å. For molecular surface, the program DMS under the UCSF MidasPlus software package[62] was used.

The program SURFNET[48] was used to identify protein clefts. Residues were defined as interface residues if the accessible surface area of the residue decreased by at least 1 $Å^2$ when the NA coordinates were added back to the structure analysis.

Conservation analysis

For each of the protein structures in our data sets, we created a sequence-based multiple alignment (MSA). PSI-BLAST version 2.1.1[53] was used to search the non-redundant (nr) protein sequence database and generate an MSA from significantly similar sequences. For each protein, we ran ten iterations of PSI-BLAST with an *E*-value threshold of 0.001. To reduce redundancy, sequences with more than 90% identity were eliminated from the MSAs. Furthermore, to ensure that all sequences in the MSA are likely to be structurally related, we included only sequences with more than 35% identity. The overall sequence variation of the patch residues was calculated based on the sequence variation entropy measure. Specific conservation was calculated only for positive and aromatic residues in the electrostatic patch. Columns, in which at least 75% of the residues were identical to the amino acid in the representative sequence, were assigned as a position with conserved identity. Conservation of amino acid property (positive and aromatic) was assigned when at least 75% of the residues in the column shared the same property as the amino acid in the representative structure. All gaps in the alignments were penalized as zero identity. For each amino acid type, the normalized frequency of conserved residues in the patch was analyzed.

Neural network architecture

The NevProp 4 Neural Network package[66] was used for prediction. This package is freely available†. The network used for all predictions had a two-layer architecture with three hidden nodes in the first layer, and a single output. In each hidden unit, the logistic transformation was carried out on the sum of its inputs. Training was performed with a standard feedforward, error back-propagation algorithm. The AutoTrain function was used to prevent over-training. This function attempted to generalize to future data by withholding 50% of the data set at a time and testing on the remaining 50% (also called a "split"). A total of five splits were used. In each split, the mean error criterion is found. NevProp was then retrained with all data without exceeding the mean error criterion reached from any of the splits. The cross-validation scheme (jackknife) used was to train on all but one of the examples. The remaining example was then tested on the trained NN. The cross-validation test was done for each protein in the data sets (54 times for dsDNA proteins and 250 for non-NA-binding proteins).

The inputs used to train the NN are shown in Table 2. The target value for true and false examples was set at 1 and 0, respectively. A threshold of 0.5 was used to evaluate whether the proteins were predicted correctly. In cases where conservation data were not available since the proteins do not share homology with other proteins in the data, the average conservation values for the given data set was used instead. The relevance for the different inputs was evaluated by repeating the full NN analysis 12 times, each time eliminating one of the input parameters. To rank the importance of each input parameter, for each of the 12 analyses, we calculated the correlation coefficient[54] values (*C*) and compared them to the original *C*-value calculated for the generic NN.

Generalized linear model (GLM)

NevProp4 was also used to perform the GLM calculations. A generalized linear model is a regression function that is represented by the following equation:

$$R^{\mathrm{GLM}}(Y|X) = g(X\beta)$$

where β is a vector list of weights, *Y* a vector list of predicted variables, *X* a vector list of descriptor variables and *g* a link function.[66] In this analysis, the log odds "logit" function was used for the link function. The same inputs used for the NN were also used for the GLM. A similar cross-validation scheme to the one described for the NN was used for the GLM.

Acknowledgements

We thank Dr Phil Goodman for the use of and assistance with NevProp4 and Dr Bruno Sunso for statistical advice on the GLM. We also thank Dr Golan Yona for invaluable discussions, Drs Olivier Lichtarge, Hanah Margalit and Ora Furman-Schueler and the anonymous referees for helpful comments on the manuscript. This work was supported by NIH grant GM52885 to L.M.G., an award from the University of California Cancer Research Coordinating Committee to L.M.G., and an American Cancer Society California Division postdoctoral fellowship to Y.M.G.

References

1. Sali, A. (1998). 100,000 protein structures for the biologist. *Nature Struct. Biol.* **5**, 1029–1032.
2. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
3. Vitkup, D., Melamud, E., Moult, J. & Sander, C. (2001). Completeness in structural genomics. *Nature Struct. Biol.* **8**, 559–566.
4. Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N. & Orengo, C. A. (2000). From structure to

† http://brain.unr.edu/FILES_PHP/show_papers.php

function: approaches and limitations. *Nature Struct. Biol.* **7**, 991–994.

5. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
6. Hofmann, K., Bucher, P., Falquet, L. & Bairoch, A. (1999). The PROSITE database, its status in 1999. *Nucl. Acids Res.* **27**, 215–219.
7. Suck, D. (1997). Common fold, common function, common origin? *Nature Struct. Biol.* **4**, 161–165.
8. Koppensteiner, W. A., Lackner, P., Wiederstein, M. & Sippl, M. J. (2000). Characterization of novel proteins based on known protein structures. *J. Mol. Biol.* **296**, 1139–1152.
9. Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147–164.
10. Gerlt, J. A. & Babbitt, P. C. (2001). Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu. Rev. Biochem.* **70**, 209–246.
11. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358.
12. Landgraf, R., Xenarios, I. & Eisenberg, D. (2001). Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.* **307**, 1487–1502.
13. Armon, A., Graur, D. & Ben-Tal, N. (2001). ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.* **307**, 447–463.
14. Sowa, M. E., He, W., Slep, K. C., Kercher, M. A., Lichtarge, O. & Wensel, T. G. (2001). Prediction and confirmation of a site critical for effector regulation of RGS domain activity. *Nature Struct. Biol.* **8**, 234–237.
15. Blom, N., Gammeltoft, S. & Brunak, S. (1999). Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.* **294**, 1351–1362.
16. Stahl, M., Taroni, C. & Schneider, G. (2000). Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network. *Protein Eng.* **13**, 83–88.
17. Laskowski, R. A., Luscombe, N. M., Swindells, M. B. & Thornton, J. M. (1996). Protein clefts in molecular recognition and function. *Protein Sci.* **5**, 2438–2452.
18. Jones, S. & Thornton, J. M. (1997). Analysis of protein–protein interaction sites using surface patches. *J. Mol. Biol.* **272**, 121–132.
19. Gallet, X., Charloteaux, B., Thomas, A. & Brasseur, R. (2000). A fast method to predict protein interaction sites from sequences. *J. Mol. Biol.* **302**, 917–926.
20. Wallace, A. C., Laskowski, R. A. & Thornton, J. M. (1996). Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci.* **5**, 1001–1013.
21. Wallace, A. C., Borkakoti, N. & Thornton, J. M. (1997). TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.* **6**, 2308–2323.
22. Fetrow, J. S. & Skolnick, J. (1998). Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol.* **281**, 949–968.
23. Fetrow, J. S., Siew, N. & Skolnick, J. (1999). Structure-based functional motif identifies a potential disulfide oxidoreductase active site in the serine/threonine protein phosphatase-1 subfamily. *FASEB J.* **13**, 1866–1874.
24. Ondrechen, M. J., Clifton, J. G. & Ringe, D. (2001). THEMATICS: a simple computational predictor of enzyme function from structure. *Proc. Natl Acad. Sci. USA*, **98**, 12473–12478.
25. Stawiski, E. W., Baucom, A. E., Lohr, S. C. & Gregoret, L. M. (2000). Predicting protein function from structure: unique structural features of proteases. *Proc. Natl Acad. Sci.* **97**, 3954–3958.
26. Stawiski, E. W., Mandel-Gutfreund, Y. & Gregoret, L. M. (2002). Progress in predicting protein function from structure: unique features of O-glycosidase. *Pacific Symp. Biocomput.* **7**, 637–648.
27. Ohlendorf, D. H. & Matthew, J. B. (1985). Electrostatics and flexibility in protein–DNA interactions. *Advan. Biophys.* **20**, 137–151.
28. Mandel-Gutfreund, Y., Margalit, H., Jernigan, R. L. & Zhurkin, V. B. (1998). A role for CH…O interactions in protein–DNA recognition. *J. Mol. Biol.* **277**, 1129–1140.
29. Pabo, C. O. & Sauer, R. T. (1992). Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.* **61**, 1053–1095.
30. Mandel-Gutfreund, Y., Schueler, O. & Margalit, H. (1995). Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *J. Mol. Biol.* **253**, 370–382.
31. Jones, S., van Heyningen, P., Berman, H. M. & Thornton, J. M. (1999). Protein–DNA interactions: a structural analysis. *J. Mol. Biol.* **287**, 877–896.
32. Honig, B., Sharp, K. & Gilson, M. (1989). Electrostatic interactions in proteins. *Prog. Clin. Biol. Res.* **289**, 65–74.
33. Boggon, T. J., Shan, W. S., Santagata, S., Myers, S. C. & Shapiro, L. (1999). Implication of tubby proteins as transcription factors by structure-based functional analysis. *Science*, **286**, 2119–2125.
34. Arbuzova, A., Wang, L., Wang, J., Hangyás-Mihályné, G., Murray, D., Honig, B. & McLaughlin, S. (2000). Membrane binding of peptides containing both basic and aromatic residues. Experimental studies with peptides corresponding to the scaffolding region of caveolin and the effector region of MARCKS. *Biochemistry*, **39**, 10330–10339.
35. Büllesbach, E. E. & Schwabe, C. (2000). The relaxin receptor-binding site geometry suggests a novel gripping mode of interaction. *J. Biol. Chem.* **275**, 35276–35280.
36. Sheinerman, F. B., Norel, R. & Honig, B. (2000). Electrostatic aspects of protein–protein interactions. *Curr. Opin. Struct. Biol.* **10**, 153–159.
37. Nicholls, A., Bharadwaj, R. & Honig, B. (1993). GRASP: graphical representation and analysis of surface properties. *Biophys. J.* **64**.
38. Chasman, D. I., Flaherty, K. M., Sharp, P. A. & Kornberg, R. D. (1993). Crystal structure of yeast TATA-binding protein and model for interaction with DNA. *Proc. Natl Acad. Sci. USA*, **90**, 8174–8178.
39. Stayton, P. S. & Sligar, S. G. (1990). The cytochrome P-450cam binding surface as defined by site-directed mutagenesis and electrostatic modeling. *Biochemistry*, **29**, 7381–7386.

40. Belrhali, H., Yaremchuk, A., Tukalo, M., Larsen, K., Berthet-Colominas, C., Leberman, R. *et al.* (1994). Crystal structures at 2.5 angstrom resolution of seryl-tRNA synthetase complexed with two analogs of seryl adenylate. *Science*, **263**, 1432–1436.
41. Danchin, A. (1999). From protein sequence to function. *Curr. Opin. Struct. Biol.* **9**, 363–367.
42. Luscombe, N. M., Austin, S. E., Berman, H. M. & Thornton, J. M. (2000). An overview of the structures of protein–DNA complexes. *Genome Biol.* **1** REVIEWS001..
43. Davis, M. E., Mandura, J. D., Luty, B. A. & McCammon, J. A. (1991). Electrostatics and diffusion of molecules in solution. *Comp. Phys. Commun.* **62**, 187–197.
44. Nadassy, K., Wodak, S. J. & Janin, J. (1999). Structural features of protein–nucleic acid recognition sites. *Biochemistry*, **38**, 1999–2017.
45. Hendsch, Z. S. & Tidor, B. (1994). Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Sci.* **3**, 211–226.
46. Lewis, M. & Rees, D. C. (1985). Fractal surfaces of proteins. *Science*, **230**, 1163–1165.
47. Jones, S., Daley, D. T., Luscombe, N. M., Berman, H. M. & Thornton, J. M. (2001). Protein–RNA interactions: a structural analysis. *Nucl. Acids Res.* **29**, 943–954.
48. Laskowski, R. A. (1995). SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.* **13**, 323–330. see also pp. 307–328..
49. Kisters-Woike, B., Vangierdegom, C. & Müller-Hill, B. (2000). On the conservation of protein sequences in evolution. *Trends Biochem. Sci.* **25**, 419–421.
50. Madabushi, S., Yao, H., Marsh, M., Kristensen, D. M., Philippi, A., Sowa, M. E. & Lichtarge, O. (2002). Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.* **316**, 139–154.
51. Lichtarge, O. & Sowa, M. E. (2002). Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.* **12**, 21–27.
52. Luscombe, N. M. & Thornton, J. M. (2002). Protein–DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J. Mol. Biol.* **320**, 991–1009.
53. Altschul, S. F., Madden, T. L., Schaeffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
54. Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.
55. Pabo, C. O. & Nekludova, L. (2000). Geometric analysis and comparison of protein–DNA interfaces: why is there no simple code for recognition? *J. Mol. Biol.* **301**, 597–624.
56. Suzuki, M. (1994). A framework for the DNA–protein recognition code of the probe helix in transcription factors: the chemical and stereochemical rules. *Structure*, **2**, 317–326.
57. McCullogh, P. & Nelder, J. (1989). *Generalized Linear Models*, 2nd edit., Chapman and Hill, London.
58. Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T. *et al.* (1992). The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.* **63**, 751–759.
59. Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522–524.
60. McDonald, I. K. & Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**, 777–793.
61. Jorgensen, W. L. & Tirado-Rives, J. (1988). The OPLS potential functions for proteins. Energy minimization for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **110**, 1657–1666.
62. Ferrin, T. E., Huang, C. C., Jarvis, L. E. & Langridge, R. (1988). The MIDAS display system. *J. Mol. Graph.* **6**, 13–27. see also pp. 36–37..
63. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
64. Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.
65. Gerstein, M. (1992). A resolution-sensitive procedure for comparing protein surfaces and its application to the comparison of antigen-combining sites. *Acta Crystallog.* **48**, 271–276.
66. Goodman, P. (1996). *NevProp Software Version*, 3rd edit., University of Nevada, Renvo, NV.

Edited by J. Thornton

(Received 30 May 2002; received in revised form 16 December 2002; accepted 18 December 2002)

Supplementary Material comprising three Tables is available on ScienceDirect