doi:10.1016/j.jmb.2006.02.053 J. Mol. Biol. (2006) 358, 922–933

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Efficient Prediction of Nucleic Acid Binding Function from Low-resolution Protein Structures

András Szilágyi and Jeffrey Skolnick*

Center of Excellence in Bioinformatics, University at Buffalo, State University of New York, 901 Washington St Buffalo, NY 14203, USA

*Corresponding author

Structural genomics projects as well as *ab initio* protein structure prediction methods provide structures of proteins with no sequence or fold similarity to proteins with known functions. These are often low-resolution structures that may only include the positions of C^{α} atoms. We present a fast and efficient method to predict DNA-binding proteins from just the amino acid sequences and low-resolution, C^{α}-only protein models. The method uses the relative proportions of certain amino acids in the protein sequence, the asymmetry of the spatial distribution of certain other amino acids as well as the dipole moment of the molecule. These quantities are used in a linear formula, with coefficients derived from logistic regression performed on a training set, and DNA-binding is predicted based on whether the result is above a certain threshold. We show that the method is insensitive to errors in the atomic coordinates and provides correct predictions even on inaccurate protein models. We demonstrate that the method is capable of predicting proteins with novel binding site motifs and structures solved in an unbound state. The accuracy of our method is close to another, published method that uses all-atom structures, time-consuming calculations and information on conserved residues.

Keywords: DNA-binding; protein–DNA interactions; function prediction; structural genomics; dipole moment

Introduction

Structural genomics projects aim to solve the experimental structures of all existing protein folds. The rationale behind these projects is that knowing a protein's structure will help with identifying its function. Because the targets of structural genomics projects are proteins for which there is currently little information, it is expected that the structures determined in the framework of these projects will include many proteins with novel functions. Because of the multifunctional nature of proteins and the existence of multifunctional folds, the analysis of sequences, motifs, and the identification of the fold are insufficient to reliably predict the function of these novel proteins; identifying the function of these proteins will be a great challenge. Reliable function prediction will involve prediction and/or analysis of active sites, binding sites or other structural properties indicative of protein function.[1,2]

In recent years, with the increasing computing power available to researchers and the development of new, efficient techniques, large-scale protein structure prediction has become feasible. Comparative modeling methods, threading-based approaches and *ab initio* structure prediction techniques have been applied to the entire protein sequence database[3] and complete genomes.[4] The structures provided by these projects, as well as those from some experimental techniques, are often low-resolution structures where atomic coordinates are not accurate, and some atoms may be missing. In many cases, all we have is a rough C^{α}-backbone that shows the global fold of the protein and the approximate location of each residue without the detailed conformation of each side-chain. It is desirable that we have efficient methods for the prediction of protein function from such limited information and approximate structures.

Present addresses: A. Szilágyi, Institute of Enzymology, Hungarian Academy of Sciences, Karolina ut 29, H-1113 Budapest, Hungary; J. Skolnick, Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology, 250 14th St NW, Atlanta, GA 30318, USA.

Abbreviations used: ROC, receiver operator characteristic; AUC, area under the ROC curve; RMSD, root-mean-square deviation; PDB, Protein Data Bank.

E-mail address of the corresponding author: skolnick@gatech.edu

DNA-binding proteins have a vital role in the maintenance and the biological processing of genetic information, including the (regulated) transcription, replication, repair, packaging and rearrangements of DNA. It has been estimated that 2–3% of proteins encoded by prokaryotic genomes and 6–7% of those encoded by eukaryotic genomes are DNA-binding proteins.[5] The structures and binding modes of DNA-binding proteins are diverse: Luscombe *et al.*[5] classified the known DNA-binding structures into eight major classes and a total of 54 subgroups.

There have been several investigations into the patterns of protein–DNA interactions.[6–10] In an early study, Seeman *et al.*[11] proposed some rules that describe how certain amino acids can discriminate between different bases. Several studies focused on the hydrogen bonding patterns between amino acids and bases.[6–8,10,12,13] Although some typical patterns have been revealed, e.g. the interaction between Asn and A and between Lys and G, the rules of binding and recognition vary substantially among different protein families. In addition, it has been demonstrated that residues and bases not directly in contact with each other also can play important roles in protein–DNA recognition *via* various mechanisms, e.g. by water-mediated contacts[14] or sequence-dependent, binding-induced conformational changes and distortions.[15,16]

Because of the great diversity of protein–DNA binding patterns and modes, identifying DNA-binding proteins based on structural features is a challenging task, especially if the goal is to devise a method that is not limited to a particular family of DNA-binding proteins, e.g. the helix-turn-helix group. There are two published studies that aim to solve this problem.[17,18] Stawiski *et al.*[18] utilized 12 parameters extracted from the detailed atomic structure of the protein. The computation of some of these parameters requires the analysis of "electrostatic patches" (finding positively charged surface patches from a Poisson–Boltzmann continuum electrostatic potential) and an analysis of surface clefts. These parameters require a detailed, accurate all-atom structural model. Three of the 12 parameters rely on a conservation analysis of the sequence, which involves a PSI-BLAST search for related sequences. The 12 parameters are then fed into a three-layer artificial neural network with one hidden layer containing three units, and the prediction is obtained from the output. With a threshold value of 0.5 to decide between negative and positive predictions, 44 out of 54 DNA-binding proteins and 236 out of 250 non-DNA-binding proteins were correctly classified in a leave-one-out cross-validation, which is equivalent to a Φ correlation coefficient (categorical correlation coefficient between prediction and truth; also known as the Matthews correlation coefficient) of 0.738. Ahmad & Sarai[17] presented a different approach to the problem based on simple principles: they created a linear predictor (a two-layer neural network with no hidden layer) that only uses bulk electrostatic properties (the total charge, dipole moment and quadrupole moment of the molecule) to predict DNA binding. As tested by a cross-validation scheme, their method predicted ~63 of 78 DNA-binding and ~96 of 110 non-DNA-binding proteins, equivalent to a Φ correlation coefficient of 0.68. This method is fast and efficient and could work with C^{α}-coordinates only. We believe, however, that there is a flaw in the way Ahmad & Sarai constructed their data sets (see details later), and therefore, the performance of the method is overestimated. It should be noted that Ahmad *et al.*[19] also devised a method to predict DNA-binding proteins from the amino acid composition alone. The accuracy of this method was found to be moderate (68.6% sensitivity at a 63.4% specificity on a large set of sequences); however, because of the way the data set was constructed, these results are not comparable with those obtained from structure-based methods.

Shanahan *et al.*[20] also developed a method to identify DNA-binding proteins from structural information. However, this method relies on detecting particular structural motifs (helix-turn-helix, helix-hairpin-helix and helix-loop-helix), and although a large percentage of known DNA-binding proteins contain this motif, the reliance on the presence of a particular motif clearly limits the applicability of the method when DNA-binding proteins with other or possibly yet unknown motifs are to be detected.

The purpose of this work is to develop a method to identify DNA-binding proteins from low-resolution structures (experimental or predicted) using only bulk, coarse-grained properties that are not specific to a particular binding motif or fold and which are fast to calculate. Our goal is to construct a method that is as efficient and fast as Ahmad & Sarai's[17] but is as accurate as Stawiski *et al.*'s[18] method.

Results

Our goal was to create a classifier that predicts whether a protein is DNA-binding from its sequence and low-resolution structure. To this end, first of all, we need a set of structures for DNA-binding proteins as well as a set of structures for non-DNA-binding proteins. Using the Nucleic Acid Database (NDB),[21] we created a representative set of DNA-binding proteins with a maximum pairwise sequence identity of 35% between any two sequences (see Materials and Methods for details). This set contains 138 chains; hence, its name PD138. As a sample of non-DNA-binding proteins, we used a set created by Ahmad & Sarai,[17] consisting of 110 non-DNA-binding chains (NB110).

To find features that discriminate between DNA-binding and non-DNA-binding proteins, we tested a number of properties computable from the protein's sequence and/or structure. The tested

parameters include the amino acid composition, the total charge and the dipole moment of the molecule, the number of certain amino acids in layers (of various thicknesses) between planes perpendicular to the direction of the dipole moment (which, we assumed, should roughly point towards the DNA-binding site), parameters describing the asymmetry of the spatial distribution of certain amino acids, various parameters related to the secondary structure, including the amino acid composition of helices and strands, various parameters related to the shape of the molecule (e.g. an estimate of the exposed surface per residue). Only features that can be calculated quickly and efficiently were examined. Many features, including those related to secondary structure and the shape of the molecule were found not to discriminate well (or at all) between DNA-binding and non-DNA-binding proteins. Other parameters, such as the total charge, were eliminated because they correlate well with other parameters; therefore, they do not add new information (e.g. the total charge can be calculated from the numbers of positively and negatively charged residues; therefore, it is redundant information when the numbers of charged residues are already provided). A systematic search for the best combination of parameters (see Materials and Methods for details of the procedure) resulted in the following ten features: proportion of Arg, Lys, Asp, Ala and Gly; spatial asymmetry of Arg, Gly, Asn and Ser; dipole moment.

The spatial asymmetry of an amino acid is the asymmetry of the spatial distribution of the residues of that amino acid relative to the center of mass of the entire polypeptide chain. Quantitatively, we measured the asymmetry by how far the center of mass of the set of residues of the given amino acid is from the center of mass of the entire chain. The dipole moment was calculated with the center of mass of the chain as the reference point. Although the magnitude of the dipole moment tends to grow with protein size, we found that it is not necessary to normalize it by the chain length; better discrimination is achieved without normalization. It should be noted that the idea that bulk electrostatic properties such as the total charge and the dipole moment could have predictive power with regard to DNA-binding proteins was introduced by Ahmad & Sarai.[17] We found, however, that these properties alone are not sufficient for a prediction of DNA-binding proteins with an accuracy that is comparable to that of more sophisticated approaches.

Classification method

Our purpose was to use a classification method that is conceptually and algorithmically simple, fast, and provides some insight into which features discriminate best between classes. Logistic regression satisfies our criteria. Binary logistic regression describes the relationship between a dichotomous response variable (in our case: a protein being DNA-binding or not; we assign the numerical values 1 and 0 to the two cases) and a set of explanatory variables. Mathematically, what the regression model describes is not the value of the response variable itself but the probability (p) that it assumes the value one rather than zero. Since p ranges from 0 to 1, linear regression is inappropriate to predict its value directly. Instead, we use the logistic transformation of p, i.e.:

$$\text{logit}(p) = \log(p/(1-p))$$

which is the logarithm of the odds or likelihood ratio that the response variable is 1. Whereas p ranges from 0 to 1, logit(p) ranges from negative infinity to positive infinity, with logit(0.5) being zero. Logistic regression involves fitting to the data an equation of the form:

$$\text{logit}(p) = a + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_nx_n$$

where x_1, x_2,... x_n designate the explanatory variables and b_1, b_2, ... b_n are the coefficients. This simple formulation allows a ranking of the explanatory variables by relative importance.

Logistic regression and cross-validation results

Logistic regression was performed on PD138 as the DNA-binding set and NB110 as the non-DNA-binding set, using the ten parameters specified in the previous section as explanatory variables. In leave-one-out cross-validation, when the threshold for the prediction is set to the value providing the highest Φ (also known as Matthews) correlation coefficient, 124 of the 138 DNA-binding proteins and 92 of the 110 non-DNA-binding proteins are correctly classified, giving rise to a Φ correlation coefficient of 0.74.

Looking at the regression coefficients when the entire set is used for logistic regression, the explanatory variables can be ranked by relative importance (Table 1). Because the variables have different scales (e.g. the maximum amino acid percentages are around 20 while the maximum dipole moment is >600), we used the regression coefficients multiplied by the standard deviations of the corresponding variables for ranking. Arginine content is by far the strongest predictor of DNA-binding, followed by the percentage of glycine and lysine. The dipole moment is the fourth most important variable, followed by the aspartate content. Glycine and aspartate contents have negative coefficients while the dipole moment as well as arginine and lysine contents have positive coefficients. These parameters are followed by the spatial asymmetries of several amino acids. Asparagine and glycine residues have the strongest effect, followed by serine and arginine. Interestingly, serine asymmetry has a negative coefficient, i.e. serine residues are more symmetrically distributed in DNA-binding proteins than in non-DNA-binders. Alanine content has the smallest contribution to the total score, but we found that its

Table 1. Ranking of explanatory variables based on the logistic regression coefficients obtained on PD138/NB110

Property	Coefficient (β_i)	Standard deviation (σ_i)	True weight ($\beta_i\sigma_i$)
Arg content	0.71 (0.48/0.95)	3.54 (2.5/3.8)	2.53 (1.2/3.6)
Gly content	−0.37 (−0.3/−0.4)	3.36 (3.2/3.6)	−1.24 (−0.9/−1.6)
Lys content	0.30 (0.16/0.43)	3.81 (3.4/3.9)	1.16 (0.55/1.7)
Dipole moment	0.012 (0.013/0.012)	87.1 (91/70)	1.06 (1.2/0.85)
Asp content	−0.33 (−0.16/−0.5)	2.56 (2.6/2.7)	−0.84 (−0.4/−1.3)
Asn asymmetry	0.12 (0.12/0.16)	4.12 (3.1/4.3)	0.50 (0.37/0.68)
Gly asymmetry	0.07 (−0.05/0.14)	4.41 (2.2/4.7)	0.32 (−0.11/0.7)
Ser asymmetry	−0.08 (−0.04/−0.14)	3.56 (2.9/3.7)	−0.29 (−0.11/−0.53)
Arg asymmetry	0.02 (0.001/0.07)	3.79 (3.7/3.9)	0.07 (0.005/0.3)
Ala content	0.015 (−0.05/0.08)	3.73 (3.8/3.9)	0.05 (−0.2/0.3)

In parentheses: the numbers for enzymes only and non-enzymes only.

coefficient is larger when the logistic regression is performed on smaller sets such as PD54; therefore we kept it (the coefficients of other properties were much less variable).

A full characterization of the performance of a classifier is provided by receiver operator characteristic (ROC) curves. An ROC curve is a plot of the "hit rate" (i.e. sensitivity, $TP/(TP+FN)$) *versus* the "false alarm rate" (i.e. false positive rate, $FP/(FP+TN)$) as the threshold for the prediction is varied. The ROC curve of our method, as obtained by the leave-one-out cross-validation on the PD138 and NB110 sets, is shown in Figure 1 (continuous line). Table 2 (line 1) shows several performance measures, namely the area under the ROC curve (AUC), the normalized area under the curve up to a false positive rate of 25% (AUC25) and the sensitivity at a false positive rate of 15%.

In order to find out how much the parameters computable from the three-dimensional structure contribute to the accuracy of the prediction, we performed the cross-validation using only the proportions of Arg, Gly, Asp, Lys and Ala; i.e. those parameters that are not dependent on the three-dimensional structure. The dotted line in Figure 1 shows the ROC curve from this test, and line 2 in Table 2 shows the performance measures computable from it. At a false positive rate of 15%, the method using only structure-independent parameters provides a sensitivity of 78.2%, which increases to 86.2% when the structure-derived parameters are added. The AUC25 value is 0.66 and 0.76 without and with the structure-derived parameters, respectively. However, when false positive rates as high as 35% are allowed, the method using only sequence-dependent parameters has a slightly better sensitivity than when the structure-derived parameters are included as well (97% *versus* 94%).

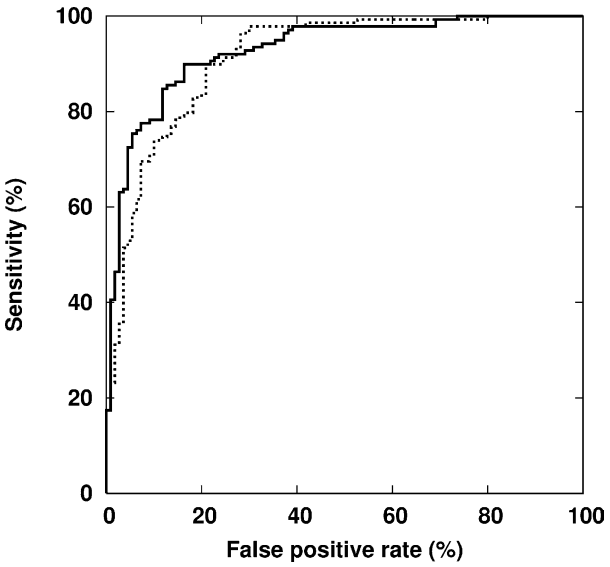

Figure 1. Receiver operator characteristic (ROC) curves for our method from the leave-one-out cross-validation on the PD138/NB110 sets. Continuous line, ROC curve using all ten explanatory variables; dotted line, ROC curve using only the variables computable from the sequence alone.

Prediction on inaccurate structures

To test how well our method performs on low-resolution structures or inaccurate models, we used the TASSER program[22] to create structural models (C^{α}-only) with root mean square deviations (RMSDs) of approximately 1, 2, 3, 4, 5 and 6 Å from native for each chain in PD138 and NB110. We repeated the logistic regression and the leave-one-out cross-validation in a procedure where accurate structures were used as training data but inaccurate structures were used as test proteins. Table 3 shows the maximum Φ correlations as a function of the RMSD from native.

There is only a slight decrease in classification performance as the structures become more and more inaccurate. Line 3 in Table 2 shows various performance measures calculated from a ROC curve (not shown) obtained with structures having an RMSD of ~6 Å from native. For example, at a false positive rate of 15%, our method achieves a sensitivity of 86.2% on accurate structures (see line 1 in Table 2); this drops to 83.3% when structures 6 Å away from native are used.

Prediction on unbound forms

The structures in PD138 are all DNA-bound forms of the proteins. It is known that many

Table 2. AUC and AUC25 values, sensitivities at 15% false positive rate and maximum Φ correlation coefficients from various ROC curves obtained for our method (see the text for details)

	AUC/AUC25	Sensitivity at 15% false positive rate (%)	Maximum Φ
Leave-one-out cross-validation with training on PD138/NB110			
1. All parameters	0.93/0.76	86	0.74
2. Sequence-based parameters only	0.91/0.66	78	0.72
3. Inaccurate structures (6 Å away from native)	0.92/0.72	83	0.69
4. Enzymes	0.85/0.58	74	0.58
5. Non-enzymes	0.96/0.83	92	0.79
Leave-one-out cross-validation with training on subsets of PD138			
6. Enzymes/NB110	0.83/0.56	67	0.57
7. Non-enzymes/NB110	0.96/0.84	94	0.81
Testing on BD54 and UD54, using PD138/NB110 as training			
8. Bound conformations (BD54)	0.93/0.74	85	0.72
9. Unbound conformations (UD54)	0.91/0.70	80	0.68
Leave-one-out cross-validation on other sets			
10. Stawiski *et al.*'s sets (PD54/NB250)	0.93/0.78	89	0.73
11. Ahmad, Sarai's sets (PD78/NB110)	0.95/0.82	92	0.79

Lines are numbered for easier reference.

DNA-binding proteins undergo conformational changes upon binding to DNA; as described by Nadassy *et al.*[23] the extent of these changes varies widely. In practice, predicting DNA-binding function only makes sense if we have an unbound structure. Therefore, it is essential to assess how well our method works on unbound forms of DNA-binding proteins. We constructed a non-redundant set of 54 protein sequences with both DNA-bound and unbound conformations in the PDB (we denote the unbound set as UD54 and the bound set as BD54). The average RMSD between the bound and unbound conformations was 2.39 Å (range: 0 to 10.5 Å). To see how well our method performs on the unbound structures, we applied it to both UD54 and BD54 and compared the results. Since several of the chains in these sets have homologs in PD138, for the testing of each protein, we excluded all its homologs (sequences with >35% identity) from PD138 and recalculated the logistic regression coefficients. Evidently, the number of correct predictions on both sets depends on the threshold value chosen. We computed the ROC curves (not shown) for both sets (using NB110 as the non-binding set) and calculated the usual performance measures from the curves (see lines 8 and 9 in Table 2). There is a small decrease in the performance of the method when the prediction is applied to unbound conformations *versus* bound ones. For example, at a false positive rate of 15%, 85% of bound and 80% of unbound structures of DNA-binding proteins are correctly classified as DNA-binding. It should be noted that the proportion of NMR structures is greater among the unbound conformations than the bound ones (16 *versus* 10 out of 54 structures).

Table 3. Maximum Φ correlation coefficients from leave-one-out cross-validation of our method as obtained with inaccurate structures having various RMSDs from the accurate, native structures

RMSD	Φ
0.0	0.74
1.0	0.74
2.0	0.72
3.0	0.73
4.0	0.72
5.0	0.71
6.0	0.69

Performance of the method on various classes of DNA-binding proteins

In their overview of DNA-binding proteins, Luscombe *et al.*[5] classified these proteins by the DNA-binding motif found in each structure. This classification divides DNA-binding proteins into eight major groups comprising a total of 54 structural families. The groups are: (1) helix-turn-helix, (2) zinc-coordinating, (3) zipper-type, (4) other α-helix, (5) β-sheet, (6) β-hairpin/ribbon, (7) other, and (8) enzyme. Stawiski *et al.*[18] unified the β-sheet and the β-hairpin/ribbon groups, thus obtaining a total of seven groups. Since these groups represent classes of DNA-binding proteins that have quite different binding motifs and modes, it is of interest to see how well our prediction method works on each group. We classified the proteins in our PD138 set into these seven groups, using literature data, structural similarities to proteins in known groups, and in some cases, visual inspection. Using this classification, we analyzed the results of the leave-one-out cross-validation procedure on the PD138/NB110 sets, to see which protein groups the false negatives (DNA-binding proteins falsely predicted as non-DNA-binding) come from at various false positive rates. We found that most false negatives usually come from the enzyme group: at a false positive rate of just 11.8%, 91.5% of DNA-binding proteins in the non-enzyme groups are correctly classified but 30.2% of the enzymes are still misclassified.

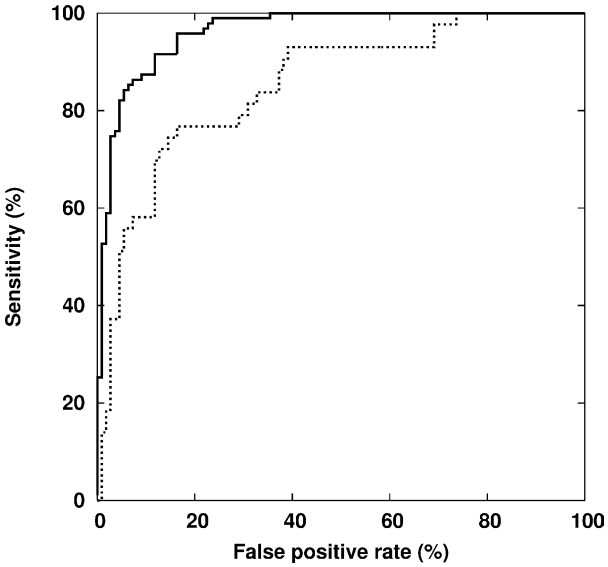

Figure 2. ROC curves for our method from leave-one-out cross-validation for non-enzymes (continuous line) and enzymes (dotted line) in the PD138 set, using the NB110 set as the non-DNA-binding set.

Figure 2 shows separate ROC curves for enzymes and non-enzymes, and lines 4 and 5 in Table 2 show the values of various performance measures computed from the curves. The maximum Φ correlation coefficient is 0.79 for non-enzymes and only 0.58 for enzymes; the sensitivities at 15% false positive rate are 92% and 74%, respectively. This finding is in accordance with Stawiski *et al.*'s own results: their method also produces most false negatives among the enzymes. Comparing the properties of enzymes and non-enzymes, we find that enzymes on average have a smaller total charge than non-enzymes (1.25 *versus* 5.59) and their dipole moment is smaller, too (0.52 *versus* 0.88 per residue).

The marked difference in the performance of our prediction method between enzymes and non-enzymes suggests that separate methods could be used for these two groups, and the regression coefficients should be different when predicting DNA-binding enzymes and non-enzymes. To see whether this approach is viable, we recalculated the logistic regression coefficients for the 43 enzymes and 95 non-enzymes in PD138. The results are shown in Table 1 (numbers in parentheses). We find that the absolute value of most regression coefficients is smaller for enzymes than for non-enzymes; the only exception is the dipole moment. The difference is most remarkable for the percentages of charged residues. This is in accordance with the observation that the total charge of enzymes is, on average, smaller than that of non-enzymes. Obviously, with respect to the explanatory variables we used in our regression model, DNA-binding enzymes are less different from non-DNA-binding proteins than DNA-binding non-enzymes.

In order to test whether different regression coefficients for enzymes and non-enzymes would work better than using the same coefficients to predict both groups, we repeated the leave-one-out cross-validation using only the enzymes and non-enzymes as a training set, respectively (NB110 was still used as the non-binding set). The performance measures calculated from the ROC curves (not shown) are shown in lines 6 and 7 of Table 2. We found that exclusion of the enzymes from the training set did not significantly improve the prediction of non-enzymes: the AUC25 grew from 0.83 to 0.84 and the sensitivity at 15% false positive rate increased to 94% from 92%. On the other hand, the prediction performed worse on enzymes when the non-enzymes were not used for training: the AUC25 dropped from 0.58 to 0.56 and the sensitivity at 15% false positive rate fell from 74% to 67%. Therefore, our method would not benefit from trying to predict DNA-binding enzymes and non-enzymes separately, with two different regression formulae.

Binding motif independence testing

One of our goals when developing our method was to ensure that it could recognize DNA-binding proteins having new folds and new DNA-binding motifs that are not yet present in the structural database. It is possible that the features used in our prediction method somehow contain information about the fold or the particular binding motif of each DNA-binding protein and therefore DNA-binding proteins having different folds or motifs are poorly predicted. To test this, we performed cross-validation on PD138 as the DNA-binding set and NB110 as the non-DNA-binding set using a slightly different scheme than before. With the non-binding set, the usual leave-one-out scheme was used, but with the binding set, we excluded from the "training set" all proteins belonging to the same group as the test protein. (Groups were defined as described in the previous section; proteins in the same group have similar DNA-binding motifs.) The number of proteins correctly predicted as DNA binding in each group, at a false positive rate of 15%, is shown in Table 4. Compared to the plain leave-one-out cross-validation scheme (denoted scheme A in the Table), the number of correct predictions reduced to 26 from 32 in the enzyme group and to 34 from 36 in the helix-turn-helix group when all proteins in the same group as the test protein were left out (scheme B in the Table); in other groups, no reduction was observed. However, it is important to note that some protein groups contain as many as 43 proteins; therefore, excluding all proteins in a group leads to a significant decrease in the size of the training set.

How can we decide if the observed drop in prediction performance results from the fact that all proteins having the same binding motif as the test protein were excluded during cross-validation or from the fact that fewer proteins are used to obtain the logistic regression coefficients? To test this, we performed another cross-validation using a different

Table 4. Data for the binding motif independence testing: number of proteins correctly identified as DNA binding in various cross-validation schemes, at a false positive rate of 15%

Group (total number of proteins in group)	Scheme A: leave out only the test protein	Scheme B: leave out all proteins in the same group as the test protein	Scheme C: leave out the same number of proteins as in Scheme B but from groups other than the test protein's
Helix-turn-helix (42)	36	34	34
Zinc-coordinating (6)	6	6	6
Zipper-type (9)	9	9	9
Other α-helix (18)	18	18	18
Beta sheet (9)	9	9	9
Other (11)	9	10	9
Enzyme (43)	32	26	28

scheme (scheme C in Table 4): for each DNA-binding test protein, we excluded proteins from groups other than the test protein's group. The number of excluded proteins was the same as the number of proteins in the same group as the test protein, and they were picked randomly. In this way, on average 29.8 proteins were excluded for each test protein (and of course, the test protein itself). The results show that this reduction of the size of the data set also caused a slight decrease in performance compared to the plain leave-one-out cross-validation scheme: 28 instead of 32 enzymes and 34 instead of 36 helix-turn-helix proteins are classified correctly. These numbers are only slightly different from those obtained with scheme B; thus, we conclude that most of the reduction in the performance of our method when proteins having the same DNA-binding motif as the test protein are omitted from the training set comes from the reduction of the size of the training set. This suggests that the dependence of our method on the binding motifs present in the training data is very small.

Similar testing of the binding motif independence was also performed by Stawiski *et al.*[18] with their method. Although their results are not directly comparable with ours because of the different (and smaller) data set they used, our results are consistent with theirs regarding the finding that the enzyme group is the most sensitive to the exclusion of the proteins in the same group from the training set.

Comparison with other methods

There are two published methods that aim to solve the same problem as our method: Stawiski *et al.*[18] and Ahmad & Sarai.[17] A comparison of two prediction methods is only adequate if the same data sets and the same evaluation method are used to assess their performance. The authors of both studies created their own data sets. Stawiski *et al.* used 54 DNA-binding (PD54) and 250 non-binding (NB250) proteins; Ahmad & Sarai employed 78 DNA-binding (PD78) and 110 non-binding (NB110) chains. Both studies used leave-one-out cross-validation.

Stawiski *et al.* obtained a Φ correlation coefficient of 0.74; Ahmad and Sarai do not calculate a Φ correlation coefficient, but on looking at their data, this can easily be computed and turns out to be 0.68. Neither study presented an ROC curve for their prediction schemes; only a single value of sensitivity and false positive rate was published in both publications.

When we apply our approach to Stawiski *et al.*'s data set (PD54/NB250), the Φ correlation coefficient from the leave-one-out cross-validation (with the best threshold) is 0.73, i.e. just 0.01 less than that of Stawiski *et al.*'s method. We calculated the ROC curve (not shown) for our method on Stawiski *et al.*'s protein sets; line 10 of Table 2 shows the corresponding performance measures. The performance of our method is quite close to that of Stawiski *et al.*'s: Stawiski *et al.*'s method reaches a sensitivity of 81.5% at a false positive rate of 5.6%, while our method reaches the same sensitivity at a false positive rate of 6.0%, a difference of just 0.4%.

When we apply our approach to Ahmad & Sarai's data set (PD78/NB110), the maximum Φ correlation coefficient obtained from the leave-one-out cross-validation is 0.79. This is significantly higher than the value 0.68 from Ahmad & Sarai's own method. However, our re-analysis of Ahmad & Sarai's data indicates that the way the PD78 set was created by these authors is biased. The authors used 62 protein–DNA complex structures having $<25\%$ pairwise sequence identity. They then split these 62 complexes into chains to obtain the PD78 set, without re-checking the pairwise sequence identity between the chains. In fact, we found that 20 out of these 78 chains have an identical partner within the set (100% sequence identity) and an additional 17 chains have a partner having $>35\%$ sequence identity. Thus, the PD78 set is highly redundant; leave-one-out cross-validation on this set will obviously overestimate the performance of any prediction method. In addition, the PD78 set appears to be somewhat biased with regard to the total charge and dipole moment of the chains: the average charge is 5.7 and the average dipole moment per residue is 0.97, compared to an average charge of 4.0 (4.2) and an average dipole moment per residue of 0.78 (0.77) for the set PD138 (PD54). Since Stawiski *et al.*'s PD54 set and our independently constructed PD138 set have a similar distribution with regard to charge and dipole moment, we believe that these sets represent the actual distribution of DNA-binding proteins in nature better than Ahmad & Sarai's PD78 set.

We calculated the ROC curve (not shown) of our method as obtained by leave-one-out cross-validation on Ahmad & Sarai's protein sets; the corresponding performance measures are shown in line 11 of Table 2. The curve runs higher than the one obtained with either our own protein sets (PD138/NB110) or Stawiski *et al.*'s protein sets (PD54/NB250), again a reflection of the redundancy of Ahmad & Sarai's DNA-binding protein set (PD78). Ahmad & Sarai's method, when tested with their own data sets, has a sensitivity of 80.7% at a false positive rate of 12.7%; our method (as tested with the same protein sets) reaches the same sensitivity at a false positive rate of just 4.5%. This indicates that our method performs significantly better than Ahmad & Sarai's.

Discussion

We have developed a conceptually simple and efficient method that uses logistic regression with ten explanatory variables, each one easy and fast to calculate, to predict whether a protein is DNA-binding from its sequence and a low-resolution, C^{α}-only structure. An analysis of the regression coefficients shows that the proportions of charged residues, especially arginine, are the most important discriminators between the two groups. Just as important is the glycine content of the protein, with DNA-binding proteins tending to contain fewer glycine residues than non-DNA-binding ones. Actually, just these sequence-based parameters are sufficient for predicting DNA-binding with reasonable accuracy; however, including some structure-based parameters further increases the performance of our classifier (e.g. from 78% to 86% sensitivity at a 15% false positive rate). Asymmetries of the spatial distributions of certain residues, especially asparagine and glycine, are good indicators of DNA-binding capacity; serine tends to be more evenly distributed in DNA-binding proteins than non-DNA-binding ones. The dipole moment of the molecule also helps discriminate between the two classes. The ranking reflects a few known facts about DNA–protein interactions. Three of our ten explanatory variables are the percentages of arginine, lysine and aspartic acid in the sequence; although glutamic acid is not included, these parameters are strongly correlated with the net charge of the polypeptide chain. Positive net charge should facilitate the binding of the protein to the negatively charged DNA and, plausibly, its dipole moment could be important for finding the correct orientation for binding.[23] Asymmetric spatial distributions of certain amino acids are indicative of residues typically present at protein–DNA interfaces. Arginine, with its unique capacity to form multiple and bifurcated hydrogen bonds[12] and to simultaneously bind multiple DNA bases, is involved in almost all DNA-binding sites.[23] Asparagine, like glutamine, also occurs commonly at protein–DNA interfaces, with its hydrogen binding properties being essential for molecular recognition. Glycine has not received special attention in most analyses of protein–DNA binding; we have found, however, that it is favored in the minor grooves of the DNA double helix while depleted in the major groove (unpublished data). It is plausible that its small size and flexibility makes glycine a good candidate for motifs recognizing the minor groove; this is consistent with investigations of small, minor-groove recognizing peptides where glycine has been found to allow the chain to bind deeply in the groove.[24,25] Interestingly, although DNA binding surfaces have been found to be enriched in serine by others (e.g. Cheng *et al.*[12]), in our study, the regression coefficient for the serine spatial asymmetry is negative, meaning that greater asymmetry is associated with non-DNA-binding proteins. This is probably due to the various functional roles serine may play in other proteins.

We found that the performance of our method is lower for enzymes that bind to DNA than for non-enzymes, a finding that is consistent with that of Stawiski *et al.*,[18] obtained with their method. This indicates that DNA-binding enzymes are characteristically different from DNA-binding non-enzymes: their structural properties apparently follow different patterns, and therefore we cannot predict them with equally high sensitivity as the non-enzymes. A plausible explanation for this observation could be that many enzymes only bind to DNA transiently and therefore their properties are not markedly distinct from non-DNA-binding proteins. In particular, we found that enzymes, unlike non-enzymes, do not tend to have a large positive total charge. This is consistent with the finding of Jayaram *et al.*[26] that with proteins that bind to short stretches of DNA in an enveloping mode (which is characteristic of enzymes), electrostatics tends to have an unfavorable contribution to the total free energy of binding.

Despite the fact that enzymes behave differently in our method than non-enzymes, we have demonstrated that the difference is not large enough to justify the development of separate prediction methods for enzymes and non-enzymes based on the ten features that we employ in our approach.

Testing the method on inaccurate structures shows that it is very robust: classification performance only slightly decreases relative to the native structure when structures with an RMSD as much as 6 Å away from native are used. This is understandable considering that none of the properties employed in our method depend strongly on the accurate, detailed atomic structure of the protein.

In actual applications, predicting nucleic acid binding function from a structural model of a protein only makes sense if the available structure does not contain a bound nucleic acid molecule, and therefore, the question of whether the protein is capable of binding a nucleic acid remains open. Many DNA-binding proteins are known to undergo conformational changes upon the binding of

DNA.[23] However, because of the small number of available unbound structures of DNA-binders, the data set PD138 that we used for training (and testing) our method only contains DNA-bound conformations of proteins, with the DNA molecule(s) stripped off. Tests of our prediction method using only bound conformations may overestimate the actual performance of the method in actual applications. Therefore, it is important to test whether the method is capable of providing correct predictions when unbound (free) conformations of DNA-binding proteins are used. We used the bound and unbound conformations of 54 proteins to evaluate the performance of our method on unbound conformations. Although the sensitivity of the method is somewhat less on unbound than on bound conformations (80% *versus* 85% at 15% false positive rate), the difference is small, and some of the difference may be due to the fact that there are more NMR structures (therefore lower accuracy) among the unbound than the bound conformations. This observation, just like the results with inaccurate structures, shows that our method is robust against conformational changes or deviations from the native, bound conformation of a protein. Clearly, the structure-dependent properties that we use as input variables to our method (the dipole moment and the spatial asymmetry of amino acid distributions) vary little upon usual, ligand-induced conformational changes of proteins.

Since we were interested in developing a method that can recognize DNA-binding proteins with new folds or new binding modes or motifs, we tested our method using cross-validation schemes where proteins belonging to the same structural group (as described by Luscombe *et al.*'s classification) as the test protein are removed from the training data. Although we found a slight decrease in the performance of the method when tested this way, we showed that most of this decrease is due to the reduction of the size of the data set and not to the elimination of the proteins in the same group. Therefore, we expect that the method can recognize new kinds of DNA-binding proteins with practically the same sensitivity as those belonging to already known groups.

Comparing our method to two other published methods showed that it has significant advantages over both of them. Stawiski *et al.*[18] presented a method that relies on accurate all-atom structures and requires time-consuming electrostatic and surface calculations and analyses. It also needs information on conserved residues, which requires a PSI-BLAST search and assumes that related sequences are found in the sequence databases. Our method only uses properties that can be computed extremely fast, and it does not need related sequences. Ahmad & Sarai[17] published a method that relies on bulk electrostatic properties of the molecule. Although we utilized this idea when we included the dipole moment of the protein as one of the explanatory variables, we showed that only using these properties is insufficient for reliable prediction. We have pointed out that the way Ahmad and Sarai validated their results was flawed and their method is less accurate than they claimed. Adding a few other properties as inputs to the prediction algorithm, however, greatly improves the classification performance of the method, and that is what we have accomplished. Since our method is both fast and accurate, it can readily be applied in proteome-scale studies, either to experimental structures obtained in structural genomics projects or to structural models obtained from proteome-scale structure prediction.

Materials and Methods

Data sets

A set of 138 DNA-binding protein chain structures (PD138) was created using the following procedure: The Nucleic Acid Database[21] was queried to retrieve all X-ray structures with ≤ 3.0 Å resolution for protein–DNA complexes containing double-stranded DNA. The resulting 576 complexes were split into chains and analyzed. Structures containing less than five DNA base-pairs were discarded, and protein chains shorter than 41 residues or having less than five residues in contact with the DNA were excluded. This resulted in 1130 DNA-binding protein chains. An all-against-all sequence comparison was performed on this set using the ALIGN0 global alignment program[27] from the FASTA2 package, with default gap penalties (no end gap penalties; BLOSUM50 was used as a scoring matrix). From the resulting alignments, pairwise sequence identities were calculated by dividing the number of identities by the length of the shorter sequence. Using the sequence identity data, a culling procedure was used to obtain a subset of the 1130 chains where no two chains have $>35\%$ identity. First, the 1130 chains were sorted by the resolution of their X-ray structure. The top structure (having the best resolution) was kept and all its neighbors (i.e. chains $>35\%$ identical with the top chain) were excluded. This procedure was repeated with the second chain in the list and so on until the list was exhausted. The resulting set contained 138 protein chains. We used literature data, structural similarities to proteins in known groups, and in some cases visual inspection to classify each structure in one of the eight groups of DNA-binding proteins defined by Luscombe *et al.*[5] (the β-sheet and β-hairpin/ribbon classes were unified, resulting in seven actual groups). The structures in each group are as follows (the first four characters are the PDB code and the fifth one is the chain identifier): Helix-turn-helix: 1aisB, 1bc8C, 1c9bA, 1cf7B, 1ddnA, 1dp7P, 1e3oC, 1efaA, 1f4kA, 1fjlA, 1hcrA, 1hlvA, 1ic8A, 1ignA, 1j75A, 1je8A, 1jftA, 1jt0A, 1k78A, 1l3lA, 1lmb3, 1lq1A, 1mnmC, 1ornA, 1perL, 1pp7U, 1pufA, 1pufB, 1qpiA, 1r71A, 1r8dA, 1r8eA, 1repC, 1rh6B, 1saxA, 1sfuA, 1tc3C, 1troA, 2cgpA, 2irfG, 3htsB, 3orcA; zinc-coordinating: 1dszA, 1hwtC, 1meyC, 1ozjA, 1tsrB, 2drpA; zipper-type: 1a0aA, 1am9A, 1dh3A, 1gd2E, 1jnmA, 1llmC, 1mdyA, 1nkpB, 1nlwA; other α-helix: 1b3tA, 1egwA, 1f44A, 1floA, 1fzpB, 1h89C, 1jfiB, 1jj4A, 1ku7A, 1kx5A, 1kx5B, 1kx5C, 1kx5D, 1mnmA, 1p7dA, 1qrvA, 1sknP, 2bopA; β-sheet/hairpin/ribbon: 1bdtA, 1c8cA, 1ecrA, 1h6fA, 1hjcA, 1mjoA, 1owfB, 1p71A, 1qnaA; other: 1a3qA, 1bg1A, 1e3mA, 1j3eA, 1jb7A, 1jb7B, 1jeyA, 1jeyB, 1mnnA, 1p7hL, 1pt3A; enzyme: 1a31A, 1a73A,

1bl0A, 1cezA, 1cl8A, 1d02A, 1dc1A, 1dctA, 1dewA, 1dfmA, 1dizA, 1dmuA, 1emhA, 1esgA, 1ewnA, 1fiuA, 1g38A, 1g9zA, 1i3jA, 1i6jA, 1iawA, 1jx4A, 1k3xA, 1kc6A, 1m3qA, 1m5rA, 1musA, 1mwiA, 1nk4A, 1oupA, 1p8kZ, 1qumA, 1r2zA, 1rrqA, 1rv2A, 1rztA, 1sl1A, 1sx5A, 1t3nA, 1vasA, 2dnjA, 3pviA, 6mhtA.

For comparison with other, published methods, we also used two other DNA-binding protein sets. Stawiski *et al.*[18] created a representative set of 54 proteins (PD54), based on a classification of DNA-binding proteins by Luscombe *et al.*[5] Ahmad & Sarai[17] constructed another set (PD78) consisting of 78 sequences from 62 complex structures. Because of a mistake in the construction method, however, this set is redundant and therefore inappropriate for unbiased testing. We only used it for comparison between Ahmad & Sarai's and our method.

In order to train and test our prediction method, we also needed sets of proteins that do not bind to DNA. Two published sets were used for this purpose. Ahmad & Sarai[17] created a set of 110 non-DNA-binding proteins (NB110) by excluding the known DNA binders from Rost & Sander's representative 126-protein set RS126.[28] Stawiski *et al.*[18] constructed a set in a similar way using Hobohm & Sander's[29] PDBSELECT database and applying a 25% sequence identity cutoff. This set, NB250, contains 250 chains.

To test how our method performs on unbound conformations of DNA-binding proteins, we created two other sets, one containing bound conformations of DNA-binding proteins and another one containing their corresponding unbound conformations. The procedure to construct these sets was as follows. First, protein chains longer than 40 residues in the PDB were divided into two classes based on whether the PDB entry contained a DNA molecule; this step resulted in 5390 DNA-associated and 64,275 non-DNA-associated chains. A BLAST[30] search was performed for each non-DNA-associated chain to find similar sequences among the DNA-associated chains. For each query sequence, only one hit with a reported 100% sequence identity and an *E*-value $<10^{-10}$ was kept; when no such hit could be found, the chain was discarded. This procedure resulted in a set of DNA-binding chains and another set containing the same sequences in their unbound conformation. The sets were still too large, so further filtering was applied. PDB files not containing the word DNA in their headers were discarded, and so were DNA-associated chains having fewer than six residues in contact with DNA. On the remaining set, an all-against-all BLAST search was performed and the set was culled so that no two sequences align with an *E*-value <0.1. The end result is a non-redundant set of 54 chains in DNA-bound conformation (BD54) and another set containing the same sequences in unbound, free conformation (UD54). We list the PDB codes for UD54 here, with the corresponding chains from BD54 given in parentheses: 1ajyA (1zmeC), 1aqjA (1g38A), 1arqA (1bdtA), 1bfs_ (1le5B), 1bgt_ (1ixyA), 1bno_ (1huoA), 1bw6A (1hlvA), 1ci4A (2bzfA), 1d9nA (1ig4A), 1dbqA (1bdiA), 1eh6A (1t39A), 1enj_ (1vasA), 1ev7A (1iawA), 1exnA (1j5fA), 1f08A (1ksxA), 1f43A (1yrnA), 1f9fA (1jj4A), 1fbuA (3htsB), 1g6nA (2cgpA), 1gdc_ (1r4oA), 1gvjA (1mdmB), 1gxqA (1gxpA), 1hioA (1eqzA), 1hma_ (1e7jA), 1hmy_ (10mhA), 1irf_ (2irfG), 1j0rA (1f4kA), 1j53A (1mgzA), 1jeqB (1jeyB), 1lea_ (1mvdA), 1lfb_ (1ic8A), 1lrp_ (1gfxA), 1lx8A (1rh6A), 1m08A (1pt3A), 1mijA (1xpxA), 1mugA (1mtlA), 1mw9X (1mw8X), 1nikB (1i6hB), 1oy3B (1leiA), 1pqvB (1r9sB), 1q39A (1k3wA), 1qvpA (1c0wA), 1sd4A (1xsdA), 1sso_ (1bbxC), 1tbpA (1ytbA), 1vsrA (1cw0A), 1wtdA (1wteA), 1xv5A (1y6fA), 1xwrA (1zs4A), 1z91A (1z9cA), 2a6mA (2a6oA), 2alcA (1f4sP), 2gcc_ (1gccA), 2hfh_ (2hdcA).

Logistic regression

Logistic regression was performed using Jeffrey Whitaker's public domain Python code†, which uses the iteratively re-weighted least-squares (IRLS) algorithm to find a maximum likelihood estimate of the constant term a and the coefficients $b_1, b_2,\ldots,b_n$ in the regression formula. When prediction is made for new cases, the expression is evaluated with the corresponding values of the $x_1, x_2, \ldots x_n$ explanatory variables, and the response variable is predicted to be 1 if the result is above a certain threshold. Although a threshold of zero (corresponding to a probability value of 0.5) appears to be a natural choice, it is not always appropriate: e.g. it will lead to under-prediction of the positive cases when the data set used for fitting is unbalanced, i.e. it contains significantly more negative cases than positive ones. A better practice is to choose a threshold that maximizes one's preferred measure of classification performance. A good measure of classifier performance is the Φ correlation coefficient (also known as Matthews correlation coefficient), which is the correlation coefficient between two dichotomous variables, in our case the predicted and the actual values of the response variable. Therefore, for the purposes of prediction, where a single threshold was needed, we usually chose a threshold that maximizes the Φ correlation coefficient in a cross-validation test of the method. For the preparation of ROC curves, the threshold was varied.

Cross-validation and measures of classifier performance

Leave-one-out cross-validation (also called jackknifing) was used to assess the performance of our prediction method. Using a given threshold for logit(p), true and false positive and negative predictions were counted and their numbers designated as *TP*, *FP*, *TN*, and *FN*. As one measure of performance, we used the Φ correlation coefficient (also known as Matthews correlation coefficient), defined by:

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

The threshold was varied to prepare ROC curves and to find the highest possible Φ correlation coefficient.

Besides Φ, three additional measures were used to evaluate the performance of our method: (1) the sensitivity at a false positive rate of 15%; (2) the area under the ROC curve (AUC), a well-known measure of classifier performance; and (3) the area under the ROC curve up to a false positive rate of 25%, divided by 0.25 to obtain a measure that varies between 0 and 1 (AUC25). The justification for using AUC25 as a performance measure is that for practical purposes, a false positive rate higher than 25% is rarely acceptable; therefore, the sensitivities at such high false positive rates can be excluded for performance measurement.

For testing the independence of the method from specific binding motifs, another type of cross-validation was also performed where not only the tested protein

† http://www.cdc.noaa.gov/people/jeffrey.s.whitaker/python/logistic_regression.html

itself but all proteins in the same class (group) were left out from the set used to compute the logistic regression coefficients.

Feature selection

A number of features computable from the protein sequence and structure were considered for inclusion in the set of parameters used for our prediction method. To find the best combination of features, we used an approach combining the ideas of "forward selection" and "backward elimination", two well-known variable selection methods for regression. We started with one feature (the total charge) and added other features one by one, recalculating the maximum Φ correlation coefficient from the cross-validation test as described in the previous section, and keeping those features that increased the correlation coefficient. When no further increase could be reached, we retested the effect of each feature and found that several of them could be eliminated. In this way we arrived at a set of features that resulted in a maximum Φ correlation coefficient; adding new features did not increase Φ and removing any feature decreased it.

Calculation of features

From just the sequence information itself, the proportions of any of the 20 amino acids can be calculated. For features computable from the structure, only the C^{α} positions were used. The spatial asymmetry of an amino acid *AA* was defined as the distance between the geometric center of the molecule and the geometric center of the set of *AA* residues. The dipole moment was calculated from the C^{α} positions (with the center of mass of the molecule as the origin), assigning a charge of $+1$ to Arg and Lys and a charge of -1 to Asp and Glu residues.

Generating inaccurate structures

To test how well our prediction method performs on inaccurate structures (e.g. low-resolution experimental structures or models obtained by structure prediction), we used the TASSER program[22] to generate structures ("decoys") that are protein-like but have various levels of RMSD from native protein structures. TASSER was started with a native structure as input and an ensemble of structures was generated; structures having an RMSD close to 1, 2, 3, 4, 5 and 6 Å from the native structure were selected from this ensemble.

Acknowledgements

This research was supported in part by NIH grant no. GM-48835.

References

1. Skolnick, J. & Fetrow, J. S. (2000). From genes to protein structure and function: novel applications of computational approaches in the genomic era. *Trends Biotechnol.* **18**, 34–39.
2. Skolnick, J., Fetrow, J. S. & Kolinski, A. (2000). Structural genomics and its importance for gene function analysis. *Nature Biotechnol.* **18**, 283–287.
3. Pieper, U., Eswar, N., Braberg, H., Madhusudhan, M. S., Davis, F. P., Stuart, A. C. *et al.* (2004). MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucl. Acids Res.* **32**, D217–D222.
4. Zhang, Y. & Skolnick, J. (2004). Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc. Natl Acad. Sci. USA*, **101**, 7594–7599.
5. 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.
6. Luscombe, N. M., Laskowski, R. A. & Thornton, J. M. (2001). Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucl. Acids Res.* **29**, 2860–2874.
7. Mandel-Gutfreund, Y. & Margalit, H. (1998). Quantitative parameters for amino acid–base interaction: implications for prediction of protein–DNA binding sites. *Nucl. Acids Res.* **26**, 2306–2312.
8. 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.
9. Sarai, A. & Kono, H. (2005). Protein–DNA recognition patterns and predictions. *Annu. Rev. Biophys. Biomol. Struct.* **34**, 379–398.
10. 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.
11. Seeman, N. C., Rosenberg, J. M. & Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl Acad. Sci. USA*, **73**, 804–808.
12. Cheng, A. C., Chen, W. W., Fuhrmann, C. N. & Frankel, A. D. (2003). Recognition of nucleic acid bases and base-pairs by hydrogen bonding to amino acid side-chains. *J. Mol. Biol.* **327**, 781–796.
13. Choo, Y. & Klug, A. (1994). Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc. Natl Acad. Sci. USA*, **91**, 11168–11172.
14. Jayaram, B. & Jain, T. (2004). The role of water in protein–DNA recognition. *Annu. Rev. Biophys. Biomol. Struct.* **33**, 343–361.
15. Crothers, D. M. (1998). DNA curvature and deformation in protein–DNA complexes: a step in the right direction. *Proc. Natl Acad. Sci. USA*, **95**, 15163–15165.
16. Olson, W. K., Gorin, A. A., Lu, X. J., Hock, L. M. & Zhurkin, V. B. (1998). DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proc. Natl Acad. Sci. USA*, **95**, 11163–11168.
17. Ahmad, S. & Sarai, A. (2004). Moment-based prediction of DNA-binding proteins. *J. Mol. Biol.* **341**, 65–71.
18. Stawiski, E. W., Gregoret, L. M. & Mandel-Gutfreund, Y. (2003). Annotating nucleic acid-binding function based on protein structure. *J. Mol. Biol.* **326**, 1065–1079.
19. Ahmad, S., Gromiha, M. M. & Sarai, A. (2004). Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, **20**, 477–486.
20. Shanahan, H. P., Garcia, M. A., Jones, S. & Thornton, J. M. (2004). Identifying DNA-binding proteins using structural motifs and the electrostatic potential. *Nucl. Acids Res.* **32**, 4732–4741.

21. Berman, H. M., Westbrook, J., Feng, Z., Iype, L., Schneider, B. & Zardecki, C. (2003). The nucleic acid database. *Methods Biochem. Anal.* **44**, 199–216.
22. Zhang, Y., Arakaki, A. K. & Skolnick, J. (2005). TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins: Struct. Funct. Genet.* **61**, 91–98.
23. Nadassy, K., Wodak, S. J. & Janin, J. (1999). Structural features of protein–nucleic acid recognition sites. *Biochemistry*, **38**, 1999–2017.
24. Chen, L. & Frankel, A. D. (1995). A peptide interaction in the major groove of RNA resembles protein interactions in the minor groove of DNA. *Proc. Natl Acad. Sci. USA*, **92**, 5077–5081.
25. Geierstanger, B. H., Volkman, B. F., Kremer, W. & Wemmer, D. E. (1994). Short peptide fragments derived from HMG-I/Y proteins bind specifically to the minor groove of DNA. *Biochemistry*, **33**, 5347–5355.
26. Jayaram, B., McConnell, K., Dixit, S. B., Das, A. & Beveridge, D. L. (2002). Free-energy component analysis of 40 protein–DNA complexes: a consensus view on the thermodynamics of binding at the molecular level. *J. Comput. Chem.* **23**, 1–14.
27. Myers, E. W. & Miller, W. (1988). Optimal alignments in linear space. *Comput. Appl. Biosci.* **4**, 11–17.
28. Rost, B. & Sander, C. (1993). Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl Acad. Sci. USA*, **90**, 7558–7562.
29. Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522–524.
30. Altschul, S. F., Madden, T. L., Schaffer, 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.

Edited by Michael J. Sternberg

(Received 27 October 2005; received in revised form 20 February 2006; accepted 21 February 2006)
Available online 10 March 2006