Using Evolutionary and Structural Information to Predict DNA-Binding Sites on DNA-Binding Proteins

Igor B. Kuznetsov,* Zhenkun Gou, Run Li, and Seungwoo Hwang
*Gen*NY*sis Center for Excellence in Cancer Genomics, Department of Epidemiology and Biostatistics, University at Albany, Rensselaer, NewYork*

ABSTRACT **Proteins that interact with DNA are involved in a number of fundamental biological activities such as DNA replication, transcription, and repair. A reliable identification of DNA-binding sites in DNA-binding proteins is important for functional annotation, site-directed mutagenesis, and modeling protein–DNA interactions. We apply Support Vector Machine (SVM), a supervised pattern recognition method, to predict DNA-binding sites in DNA-binding proteins using the following features: amino acid sequence, profile of evolutionary conservation of sequence positions, and low-resolution structural information. We use a rigorous statistical approach to study the performance of predictors that utilize different combinations of features and how this performance is affected by structural and sequence properties of proteins. Our results indicate that an SVM predictor based on a properly scaled profile of evolutionary conservation in the form of a position specific scoring matrix (PSSM) significantly outperforms a PSSM-based neural network predictor. The highest accuracy is achieved by SVM predictor that combines the profile of evolutionary conservation with low-resolution structural information. Our results also show that knowledge-based predictors of DNA-binding sites perform significantly better on proteins from mainly-α structural class and that the performance of these predictors is significantly correlated with certain structural and sequence properties of proteins. These observations suggest that it may be possible to assign a reliability index to the overall accuracy of the prediction of DNA-binding sites in any given protein using its sequence and structural properties. A web-server implementation of the predictors is freely available online at http://lcg.rit.albany.edu/dp-bind/. Proteins 2006;64:19–27.**

© 2006 Wiley-Liss, Inc.

Key words: support vector machine; protein–DNA interaction prediction; position-specific scoring matrix; evolutionary conservation; structural information; DNA binding; accuracy; statistical analysis

INTRODUCTION

Proteins that interact with DNA are involved in a number of fundamental biological activities such as DNA replication, transcription, and repair. It is estimated that in the human genome the total number of transcription factors alone can be as high as 3000 or about 10% of all protein-coding genes.[1] Therefore, a reliable identification of DNA-binding sites in DNA-binding proteins is important for functional annotation, in silico modeling of transcription regulation, and site-directed mutagenesis. This process is relatively straightforward if the structure of a protein–DNA complex is known. However, solving the structure of a protein–DNA complex is a much more complicated and time-consuming process than solving the structure of a protein alone. A number of computational methods that use experimentally solved structure of a DNA-binding protein to identify DNA-binding interface based on the electrostatic potential and the shape of molecular surface have been developed.[2,3] However, these methods have at least two major limitations. One problem with using the structure of an unbound protein to predict sites involved in interactions with DNA is that the actual structure of the DNA-bound protein may substantially differ from the unbound form. The other problem is that structure-based computational methods still involve the expensive and time-consuming process of experimental determination of protein structure. Homology modeling is unlikely to solve this problem because it produces low-resolution models that lack atomic level details, especially on side-chain packing.

An alternative to structure-based prediction is to use the properties of amino acid sequence of a DNA-binding protein to predict DNA-binding sites. Amino acid sequence itself is always available for any protein. For most proteins, a profile of evolutionary conservation of sequence positions can also be obtained. This profile provides information on how conserved each sequence position is and can be utilized to predict functional sites. The underlying idea of this approach is that functionally important residues corresponding to functional sites are usually more conserved than the rest of the sequence.[4,5] For a given sequence, a profile of evolutionary conservation is constructed by collecting all homologs of this sequence and

*Correspondence to: Igor Kuznetsov, Gen*NY*sis Center for Excellence in Cancer Genomics, Department of Epidemiology and Biostatistics, University of Albany, One Discovery Drive, Room 206, Rensselaer, NY 12144. E-mail: Ikuznetsov@albany.edu

Received 8 November 2005; Revised 18 January 2006; Accepted 24 January 2006

Published online 27 March 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20977

© 2006 WILEY-LISS, INC.

aligning them together. The resulting multiple alignment is used to assign a conservation score to each position in the sequence. Neural network predictors of DNA-binding sites that utilize amino acid sequence[6] and a profile of evolutionary conservation in the form of a Position Specific Scoring Matrix (PSSM)[7] have recently been developed.

In this work, we use Support Vector Machine (SVM), a supervised pattern recognition method, to predict DNA-binding residues in DNA-binding proteins. SVM is a powerful tool for data classification based on a rigorous statistical theory.[8] It is replacing neural networks in a variety of fields, including computational analysis of molecular data. One of the most attractive features of SVM is that it is capable of dealing with a very large number of variables (usually called features) used to describe the objects being classified. The training of SVM is relatively easy and efficient even for large-scale datasets. Many successful applications of SVM have demonstrated its superiority over neural networks and other machine learning methods.[9,10] We use the following three types of features to construct SVM predictors of DNA-binding residues in a DNA-binding protein: its amino acid sequence, profile of evolutionary conservation of sequence positions, and low-resolution structural information (secondary structure, solvent accessibility, and spatial neighbors of a residue). We use SVM predictors that utilize different combinations of features to study how the performance of a pattern recognition method may be affected by structural and sequence properties of proteins. To the best of our knowledge, this is the first study of its kind that uses a rigorous statistical approach to analyze the performance of a pattern recognition method using groups of structurally similar proteins and to study correlations between the performance and protein properties. Our results indicate that incorporating a properly scaled profile of evolutionary conservation in the form of a PSSM obtained using a nonredundant sequence database significantly improves the accuracy of knowledge-based prediction of DNA-binding residues. We show that the best performance is achieved by an SVM predictor that combines evolutionary information with low-resolution structural information. Our results also show that knowledge-based predictors of DNA-binding residues perform significantly better on proteins from mainly-α structural class and that the performance is correlated with certain structural and sequence properties of proteins.

MATERIALS AND METHODS

Datasets

The nonredundant dataset of 62 experimentally solved protein–DNA complexes, PDNA-62, was taken from the work of Ahmad and colleagues (2004).[6] Atomic coordinates for each complex were downloaded from the Protein Data Bank (PDB).[11] Within each complex all identical chains were removed, leaving a total of 66 protein chains interacting with DNA. The NCBI nonredundant protein database (NCBI-NR) was clustered at 90% sequence identity using the CD-HIT program version 2.0.3[12] to generate the NCBI-NR90 database in which all sequences have pairwise sequence similarity below 90%. The NCBI-NR90 database was utilized to run the PSI-BLAST program (see below).

Definition of a DNA-Binding Residue

An amino acid residue in a protein chain is labeled as a DNA-binding residue if the distance from at least one of its atoms to any DNA atom is less than a cutoff distance D. This definition is the same as the one used to train a neural network predictor of DNA-binding residues.[6,7] Our studies have shown that the cutoff distance of $D = 4.5$ Å gives the best separation between binding and nonbinding residues (the highest classification accuracy). We therefore use the cutoff distance of 4.5 Å. Amino acid residues that do not satisfy the above definition of being in contact with DNA are labeled as nonbinding. Some residues in PDB entries, called missing residues, lack information about their atomic coordinates. Such missing residues were excluded from classification into binding/nonbinding.

Construction of a Position Specific Scoring Matrix (PSSM)

We include a profile of evolutionary conservation of residue positions into our prediction of DNA-binding sites by constructing a position specific scoring matrix (PSSM) for each input sequence. For a sequence of length N residues, PSSM is represented by an $N \times 20$ matrix. Each element of this matrix, $m[i,j]$, provides information on evolutionary conservation of residue type j in sequence position i. Information on the overall evolutionary variability of sequence position i is encoded in the entire ith row, $m[i,*]$. We used the PSI-BLAST program version 2.2.10[13] to generate a PSSM for each protein chain. For each protein sequence PSI-BLAST was run for five iterations using the NCBI-NR90 database, the soft masking of low complexity regions, inclusion E value threshold of 10^{-3}, and word size of 2. For all other PSI-BLAST arguments we used the default values.

Support Vector Machine (SVM) classification

Support Vector Machine (SVM) is a very effective supervised machine learning method for pattern recognition.[14,15] An advantage of SVM is that it is based on a rigorous statistical learning theory.[8] SVM has been successfully applied to a variety of high-dimensional nonlinear biological problems, including prediction of biological function and structure from protein sequences.[16–19] When SVM is applied to a two-class pattern recognition, it is presented with a series of labeled objects from two classes (these objects are referred to as positive and negative training examples) and trained to distinguish between these classes. Usually, each object, A, in the sample is represented as an n-dimensional numeric *feature vector,* $V(A) = (v_1, v_2, \ldots, v_n)$, where $v_1, v_2, \ldots, v_n$ are variables (features) describing the properties of object A. During training, SVM maps training examples into a high-dimensional space via a kernel function and constructs a boundary that separates two classes in an optimal way. After successful training, SVM is able to predict with a high degree of confidence to which of the two classes a new, previously unseen, object belongs. In our case, positive examples are represented by DNA-binding residues and negative examples are represented by nonbinding residues. We use SVM implemented in the libSVM package version 2.8.[20] We tried various kernel functions available in libSVM and determined that the Radial Basis Function (RBF) kernel gives the highest accuracy for our dataset.

Feature Selection

Sequence and evolutionary information

In machine learning applications, the 20 amino acid types can be encoded using 20 mutually orthogonal binary vectors of dimension 20.[16] In this case, all pairwise distances between amino acid types are identical and do not take into account the degree of (dis)similarity in their chemical properties. To overcome this limitation, each amino acid type can be encoded as a vector of dimension 20 using a corresponding row from amino acid similarity or dissimilarity matrix.[21] We found that encoding using BLOSUM62 similarity matrix[22] works better on our dataset than binary encoding. Each element of the similarity matrix, $s[i,j]$, was normalized between 0 and 1 using a logistic function:

$$s^{(N)}[i,j] = \frac{1}{1 + e^{-s[i,j]}} \tag{1}$$

This normalized version of BLOSUM62 matrix is utilized to encode a protein sequence with a standard procedure using a sliding window of size w, where w is an odd number.[16] In this procedure, for an amino acid residue a_k in sequence position k, we construct a feature vector by concatenating BLOSUM62 rows describing sequence fragment $a_{k-(w\ -\ 1)/2}, \ldots, a_k, \ldots, a_{k\ +\ (w\ -\ 1)/2}$. This gives us a feature vector S_k of dimension $20 \cdot w$:

$$S_k = (s^{(N)}[a_{k-(w-1)/2},*],\ldots, s^{(N)}[a_k,*],\ldots, s^{(N)}[a_{k+(w-1)/2},*]) \tag{2}$$

where $s[a_i,*]$ is BLOSUM62 row for residue type a_i. If the window extends beyond the sequence termini, empty positions before the first and after the last residue are represented as zero vectors of dimension 20. We will refer to the SVM predictor that uses features given by Equation 2 as seq-SVM.

The evolutionary information from PSSM is encoded using the same approach as described above. First, all elements of a PSSM are scaled between 0 and 1 using Equation 1. Second, for an amino acid residue in position k a feature vector of dimension $20 \cdot w$ is constructed by concatenating corresponding PSSM rows for positions from $k - (w - 1)/2$ to $k + (w - 1)/2$. Our study has shown that setting the window size, w, to 7 gives the best results. Therefore, both sequence and evolutionary information for each residue were encoded by a feature vector of dimension 140. We will refer to the SVM predictor that uses PSSM-derived feature vectors as pssm-SVM.

Structural information

We used the following features to describe structural environment of a residue: its spatial neighbors, missing residues, solvent accessibility, and the type of secondary structure. Spatial neighbors of a residue in sequence position k are defined as a vector of dimension 20, $(f_1^{(k)}, \ldots, f_{20}^{(k)})$, that contains the normalized frequencies of occurrence of each of the 20 amino acid types within a sphere of radius 12 Å centered at residue k. The position of each residue is described using the coordinates of its C^{α} atom. The normalized frequency of amino acid type i in this sphere is given by:

$$f_i^{(k)} = \frac{n_i^{(k)}}{\sum_{j=1}^{20} n_j^{(k)}} \tag{3}$$

where $n_i^{(k)}$ is the number of occurrences of amino acid type i within a sphere centered at residue k. The nearest N- and C-terminal neighbors are omitted from the count. The radius of 12 Å was chosen because our study has shown that this value gives the best classification accuracy. Solvent accessibility was computed using the DSSP program.[23] We use the normalized solvent accessibility of a residue in position k, $NSA(a_k)$:

$$NSA(a_k) = \frac{SA(a_k)}{\max SA(a_k)} \tag{4}$$

where $SA(a_k)$ is the DSSP solvent-accessible surface area and $\max SA(a_k)$ is the maximum possible solvent-accessible surface area of residue type a_k in Gly-a_k-Gly tripeptide. Secondary structure was also computed using the DSSP program.[23] All types of structure other than helix or β-strand were assigned to coil. Secondary structure was encoded using mutually orthogonal binary vectors: (1,0,0) for helix, (0,1,0) for β-strand, and (0,0,1) for coil.

When we use structural information to make a prediction, we include the data on missing residues by adding one bit to the feature vector describing each sequence position (Eq. 2), so that the final vector is of dimension 21. This 21st bit is set to 0 for a missing residue and 1 otherwise. This 21st bit is always 1 for the central residue in the window because missing residues are excluded from classification into binding/nonbinding.

When we combine sequence with structural features or PSSM with structural features, we concatenate corresponding feature vectors together to form a single feature vector. We will refer to the SVM predictor that uses a combination of sequence- and structure-derived features as seq-str-SVM and to the SVM predictor that uses a combination of PSSM- and structure-derived features as pssm-str-SVM.

Training and Testing

The objective of training is to maximize the ability of an SVM predictor to discriminate between classes while avoiding overfitting. We used leave-one-out cross-validation to train and test our SVM predictor. In this procedure, one of the 62 protein complexes is left out for testing and the remaining 61 complexes are used for training. The process is repeated for each complex, 62 times in total.

These 62 runs are used to compute the average and standard deviation of the measures of classification performance. We use the following measures to assess different aspects of the quality of classification: accuracy (ACC), sensitivity (SN), specificity (SP),[6] and correlation coefficient (CC).[24] These measures are defined as follows:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

$$SN = \frac{TP}{TP + FN} \tag{6}$$

$$SP = \frac{TN}{FP + TN} \tag{7}$$

$$CC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{8}$$

where TP is the number of true positives (correctly predicted DNA-binding residues), FN is the number of false negatives (DNA-binding residue predicted as being nonbinding), TN is the number of true negatives (correctly predicted nonbinding residues), and FP is the number of false positives (nonbinding residues predicted as being DNA-binding).

The ratio of the number of DNA-binding residues to that of nonbinding residues in our dataset is about 1:4. This is one of the most common problems in the application of machine learning to protein data — the number of examples from one class often significantly exceeds the number of examples from the other. If an SVM predictor is trained on such unbalanced dataset, it may become overfitted and will tend to assign observations to the over-represented class. In order to equalize the number of positive and negative examples in both training and testing datasets, for each protein chain we randomly sampled without replacement the same number of nonbinding residues as that of the DNA-binding ones as suggested by Kim and Park (2004).[25] We chose the subsampling approach to balance the data because it significantly reduces learning complexity and allows us to use accuracy as an indicator during parameter optimization. The RBF kernel has two parameters, C and γ, that need to be optimized in order to achieve optimal performance of an SVM classifier. For each combination of features we use an exhaustive grid search to determine the values of C and γ that give the best classification accuracy.

RESULTS

Feature Selection and Prediction Accuracy

The goal of this section is to study how feature selection affects the performance of SVM predictor of DNA-binding sites in DNA-binding proteins. Figure 1 shows the receiver operating characteristic (ROC) curves for all four predictors trained on balanced and tested on unbalanced datasets (all sequence positions in each chain were used for testing). Table I shows the results of predictions made using various combinations of features for both balanced and unbalanced test datasets. One can see that the lowest accuracy, 68.2%, is observed when only the sequence of the protein is used to make a prediction. This is slightly lower than 73.6% accuracy of the sequence-based neural network classifier reported by Ahmad and colleagues (2004).[6] However, our sequence-based SVM classifier has a considerably higher sensitivity (Eq. 6) of 70.2% compared to that of 40.6% in Ahmad and coworkers (2004).[6] Using a profile of evolutionary conservation of sequence positions in the form of a Position Specific Scoring Matrix (PSSM) generated by the PSI-BLAST program[13] significantly increases accuracy of SVM classifier up to 76.0% (Table I). This is 9.3% better than the previously reported accuracy of 66.7% achieved by the PSSM-based neural network classifier of Ahmad and Sarai (2005).[7] Both sensitivity and specificity of pssm-SVM are also considerably better than those of the neural network classifier.

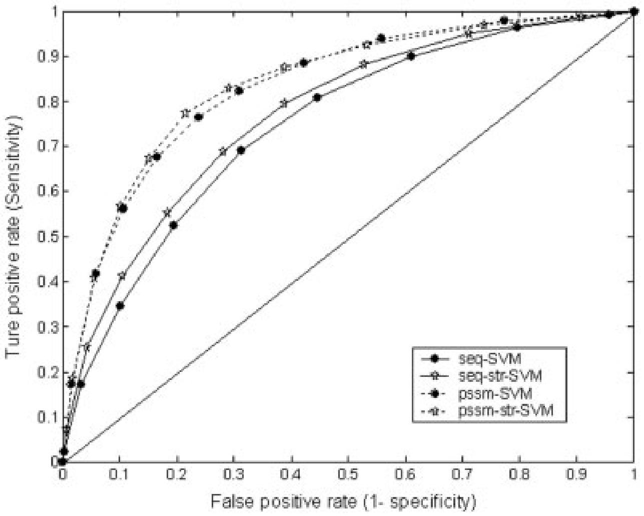

Fig. 1. The ROC curves of the four SVM predictors. The predictors have the following values of the area under the curve (AUC): seq-SVM AUC = 0.748; seq-str-SVM AUC = 0.776; pssm-SVM AUC = 0.836; pssm-str-SVM AUC = 0.840. Results for unbalanced test datasets.

A comparison of the accuracy obtained using the sequence-based SVM (seq-SVM) and the PSSM-based SVM (pssm-SVM) for each protein individually shows that for the majority of the proteins (55 out of 66) using the profile of evolutionary conservation significantly improves accuracy (Fig. 2). However, for some proteins (11 out of 66) the accuracy of pssm-SVM is lower than that of seq-SVM. Partially, this can be explained by the fact that PSI-BLAST finds very small number of sequences homologous to each of these 11 proteins. For example, in the case of 1YUI, seq-SVM has accuracy of 65.4%, whereas pssm-SVM has accuracy of only 50% (prediction at a random level). Only five sequences were used to construct PSSM for 1YUI, which is not sufficient for capturing a reliable profile of evolutionary variability. Another such example of inferior performance of pssm-SVM is 1PAR_B for which only one sequence was used to construct PSSM. Interestingly, a small number of sequences is not always associated with inferior accuracy of the pssm-SVM. In the cases

TABLE I. Measures of the Performance of SVM Predictors of DNA-Binding Sites

SVM Predictor	γ	C	Accuracy	Sensitivity	Specificity	CC
Sequence	0.02	4	69.7 ± 9.3 (68.2 ± 6.6)	70.2 ± 16.8 (70.2 ± 16.8)	69.2 ± 13.7 (66.8 ± 9.2)	0.41 ± 0.191 (0.31 ± 0.144)
Sequence + structure	0.025	4	71.0 ± 9.0 (70.0 ± 7.7)	70.1 ± 17.5 (70.1 ± 17.5)	71.1 ± 15.6 (68.5 ± 12.1)	0.44 ± 0.180 (0.34 ± 0.131)
PSSM	0.025	4	78.9 ± 10.1 (76.0 ± 9.0)	76.9 ± 18.5 (76.9 ± 18.5)	80.9 ± 13.6 (74.7 ± 12.5)	0.60 ± 0.196 (0.45 ± 0.181)
PSSM + structure	0.02	35	82.3 ± 9.4 (78.1 ± 8.0)	79.2 ± 14.9 (79.2 ± 14.9)	85.4 ± 10.3 (77.2 ± 9.9)	0.66 ± 0.184 (0.49 ± 0.174)

Types of SVM predictors: Sequence, only the sequence of the protein was used; Sequence + structure, both the sequence and the structure of the protein were used; PSSM, only the profile of evolutionary conservation of the protein was used; PSSM + structure, both the profile of evolutionary conservation and the structure of the protein were used.

γ, an optimal value of SVM *gamma* parameter; C, an optimal value of SVM *C* parameter; CC, correlation coefficient (Eq. 8).

Accuracy, sensitivity and specificity are scaled between 0% and 100%. First line in each cell gives the values for balanced test datasets, the values for unbalanced test datasets are given in parenthesis.

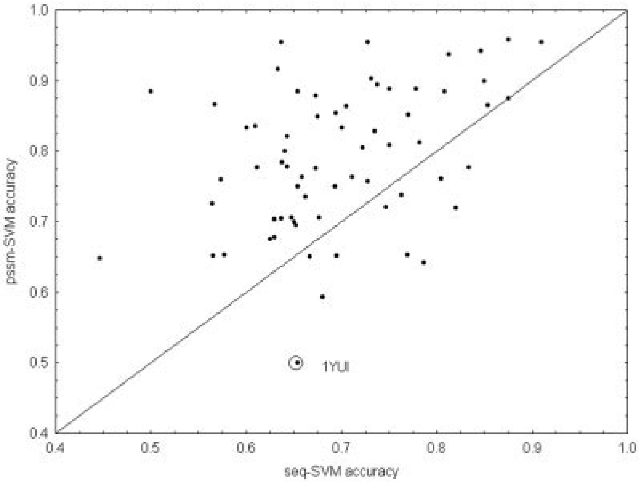

Fig. 2. Scatter plot of the accuracy of SVM predictor that uses single sequence information (seq-SVM) versus the accuracy of SVM predictor that uses the profile of evolutionary conservation (pssm-SVM). Each point corresponds to one protein. Points above the diagonal correspond to proteins on which pssm-SVM performs better. The circled point represents PDB entry 1YUI. Results for balanced test datasets.

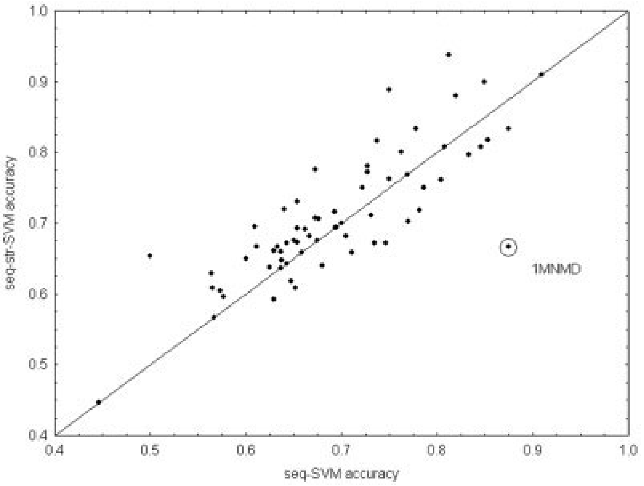

Fig. 3. Scatter plot of the accuracy of SVM predictor that uses sequence information (seq-SVM) versus the accuracy of SVM predictor that uses both sequence and structural information (seq-str-SVM). Each point corresponds to one protein. Points above the diagonal correspond to proteins on which seq-str-SVM performs better. The circled point represents PDB entry 1MNM_D. Results for balanced test datasets.

of 1PVI_B, 1AZQ, 1BHM_A, and 1D02_A, only one to three sequences were used to construct PSSM. However, the accuracy of pssm-SVM for these proteins is higher than that of seq-SVM.

When structural information in the form of spatial neighbors, secondary structure, and solvent accessibility is added to seq-SVM, accuracy increases only by 1.3% to 71%. In the case of pssm-SVM classifier, adding structural information increases accuracy by 3.4% (from 78.9% to 82.3%). If we look at how adding structural features affects prediction of each protein individually (Figs. 3 and 4), we note that all proteins fall into three distinct categories: (1) accuracy is not affected by adding structural descriptors, (2) accuracy is increased by adding structural descriptors, and (3) accuracy is decreased by adding structural descriptors. A closer inspection of Figures 3 and 4 shows that for many proteins in the case of seq-SVM and the majority of proteins in the case of pssm-SVM improvement in prediction accuracy is considerably higher than a few percentage points. For some proteins, the accuracy of the prediction based on the combination of evolutionary and low-resolution structural information can be as high as 100%. One example of how adding more features gradually improves accuracy is represented by homeodomain MAT a1 (PDB id 1YRN). In the case of this protein, sequence-based prediction (seq-SVM) is only 50% accurate, prediction based on the combination of sequence and structural information (seq-str-SVM) is 65.4% accurate, prediction based on evolutionary information (pssm-SVM) is 88.5% accurate, and, finally, prediction based on the combination of evolutionary and structural information (pssm-str-SVM) is 100% accurate. For some proteins, however, adding structural descriptors can considerably lower prediction accuracy: 18 proteins for seq-SVM and 11 proteins for pssm-SVM fall into this category. The most extreme case is 1MNM_D, for which adding structural features to seq-SVM decreases accuracy by 20.8% from 87.5% to 66.7% (see Fig. 3).

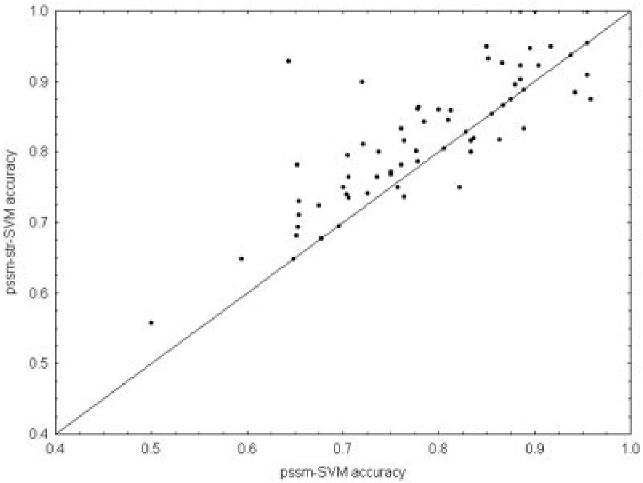

Fig. 4. Scatter plot of the accuracy of SVM predictor that uses the profile of evolutionary conservation (pssm-SVM) versus the accuracy of SVM predictor that uses both the profile of evolutionary conservation and structural information (pssm-str-SVM). Each point corresponds to one protein. Points above the diagonal correspond to proteins on which pssm-str-SVM performs better. Results for balanced test datasets.

Effect of Structural Properties on Prediction Accuracy

In this section we study how the global fold of protein affects prediction accuracy. To perform this study, we compared prediction accuracy between groups of proteins from four major structural classes annotated in the CATH database of structural domains[26]: mainly-α, mainly-β, α/β, and few regular structure (*frs*). We only used proteins that contain domains from the same class. Proteins that have domains from at least two different structural classes were excluded. This gave us a total of 29 mainly-α, 4 mainly-β, 19 α/β, and 4 *frs*-class proteins. We used the *t* test to compare the average prediction accuracy between these classes. The results summarized in Table II show that DNA-binding interface in mainly-α proteins is predicted with significantly higher accuracy than that in α/β proteins. The only exception is seq-SVM that predicts DNA-binding residues with similar accuracy in all four classes. However, when structural information is added to seq-SVM, accuracy also becomes significantly higher in mainly-α class compared to that in α/β class. The size of DNA-binding interface is similar in all four classes, which implies that it does not contribute to the difference in accuracy. Mainly-α proteins have a shorter sequence than α/β. This may partially explain the observed difference in the accuracy of pssm-SVM that utilizes the profile of evolutionary conservation because short sequence implies that most conserved residues are likely to be in contact with DNA. However, the difference in length between mainly-α and α/β proteins is only marginally significant.

The following two observations may explain why DNA-binding interfaces in mainly-α proteins are predicted with higher accuracy. First, the most significant difference between mainly-α and α/β proteins is observed in the case of pssm-SVM. Second, adding structural descriptors to seq-SVM significantly increases the accuracy for mainly-α proteins. Taken together, these observations suggest that, due to identical helical periodicities observed in mainly-α proteins, protein–DNA interaction pattern and corresponding evolutionary conservation pattern among proteins from this class are more similar than among proteins from other classes. Because different proteins from mainly-α class contain similar patterns, pattern recognition methods based on structural information and sequence conservation are more successful in recognizing protein–DNA interactions in this class. Sequence information alone does not seem to capture helical patterns well because the performance of seq-SVM is similar for all classes. DNA-binding interface in proteins from *frs* class also seems to be easier to predict than that in β-sheet containing classes (mainly-β and α/β). However, the total number of proteins in *frs* and mainly-β classes is only four, which makes the results of statistical tests unreliable.

Correlations between Protein Properties and Prediction Accuracy

In this section, we study what structural and sequence properties can affect the prediction accuracy of our SVM classifiers. In order to do this, we compute the Spearman rank correlation between accuracy of a classifier and a particular property of interest. We used secondary structure content and number of DNA-binding residues (interface size) to quantify numerically basic structural properties of each protein. We studied correlations with four types of secondary structure: α-helix, β-strand, coil (everything other than helix or β-sheet in DSSP assignment), and abrupt turn (S in DSSP assignment). We consider the abrupt turn separately because it is a special type of irregular structure that involves 180° change in chain direction and reflects the overall complexity of the packing of regular secondary structure. We also computed correlations between accuracy and sequence length and between accuracy and the number of sequences used to construct PSSM.

The results summarized in Table III show that the accuracy of all classifiers is positively correlated with the percentage of positions in α-helical conformation and negatively correlated with the percentage of positions in coil and abrupt turn. This implies that the more helical positions a protein contains, the higher the prediction accuracy becomes for this protein. On the contrary, an increase in the percentage of positions in coil and abrupt turn conformation considerably reduces accuracy. Abrupt turns often occur between two β-strands. These observations support the conclusion made in the previous section that the prediction of DNA-binding interface is much more challenging in proteins that have a large number of long-range interactions with complex topology and residue periodicities than in proteins with a high content of α-helices. It is interesting to note that there is no correlation between accuracy and sequence length or interface size when prediction is made using sequence only, whereas the accuracy of pssm-SVM shows a significant negative correlation with both sequence length and interface size,

TABLE II. Pairwise Comparison of the Four Structural Classes of Proteins Using the *t* Test

	α vs. β	α vs. α/β	α vs. *frs*	β vs. α/β	β vs. *frs*	α/β vs. *frs*
Accuracy of seq-SVM	n/s	n/s	n/s	n/s	n/s	n/s
Accuracy of seq-str-SVM	0.74 ± 0.084 0.64 ± 0.038 p = 0.036	0.74 ± 0.084 0.69 ± 0.094 P = 0.045	n/s	n/s	n/s	n/s
Accuracy of pssm-SVM	n/s	**0.81 ± 0.098** **0.72 ± 0.098** ***p* = 0.0029**	n/s	n/s	n/s	**0.72 ± 0.098** **0.88 ± 0.065** ***p* = 0.007**
Accuracy of pssm-str-SVM	n/s	**0.86 ± 0.087** **0.76 ± 0.086** ***p* = 0.0005**	n/s	n/s	0.78 ± 0.042 0.87 ± 0.056 *p* = 0.039	0.76 ± 0.086 0.87 ± 0.056 *p* = 0.02
Sequence length	n/s	94 ± 50	n/s	n/s	n/s	n/s
Interface-size	n/s	n/s	n/s	n/s	n/s	n/s
Number of sequences in PSSM	**1080 ± 1177** **120 ± 126** ***p* = 0.0002**	n/s	n/s	120 ± 126 1574 ± 2138 *p* = 0.009	n/s	n/s

Abbreviations: Seq-SVM, a predictor that uses only the amino acid sequence of the protein; Seq-str-SVM, a predictor that uses both the amino acid sequence and the structure of the protein; Pssm-SVM, a predictor that uses only the profile of evolutionary conservation of the protein; Pssm-str-SVM, a predictor that uses both the profile of evolutionary conservation and the structure of the protein; interface size, the number of DNA-binding residues; number of sequences in PSSM, the number of sequences used to construct the profile of evolutionary conservation (PSSM). Each column shows the results of the comparison of the average accuracy between two structural classes. The following four classes were used: mainly-α, mainly-β, α/β, and few regular structure (*frs*). Classes were assigned according to the CATH database of structural domains.[26] The difference between two classes is considered as not significant (n/s) if the *p* value of the *t* test is greater than 0.05. If the *p* value is less than 0.05, the average and standard deviation for both classes are given in the corresponding cell. The difference between two classes is considered to be highly significant if the *p* value is less than 0.05/7 (0.0071), a Bonferroni-adjusted α-level for seven tests. This adjustment takes into account the fact that we perform seven independent statistical tests on the same pair of classes. Highly significant differences are shown in boldface type. Results for balanced test datasets.

TABLE III. Correlations between the Accuracy of SVM Classifiers and Protein Properties

	seq-SVM Accuracy	seq-str-SVM Accuracy	pssm-SVM Accuracy	pssm-str-SVM Accuracy
Sequence length	n/s	n/s	−0.30 $p = 0.015$	**−0.46 $p = 10^{-4}$**
Interface size	n/s	n/s	**−0.49 $p = 3*10^{-5}$**	**−0.57 $p = 10^{-6}$**
Number of sequences in PSSM	n/s	n/s	**0.38 $p = 2*10^{-4}$**	**0.35 $p = 0.004$**
Percentage α-helix	0.26 $p = 0.038$	**0.36 $p = 0.003$**	**0.51 $p = 2*10^{-5}$**	**0.53 $p = 5*10^{-6}$**
Percentage β-strand	n/s	n/s	**−0.37 $p = 0.002$**	**−0.43 $p = 3*10^{-4}$**
Percentage coil	−0.28 $p = 0.024$	−0.32 0.01	**−0.35 $p = 4*10^{-4}$**	**−0.35 $p = 0.004$**
Percentage abrupt turn	**−0.34 $p = 0.005$**	−0.32 $p = 0.008$	**−0.53 $p = 7*10^{-6}$**	**−0.49 $p = 4*10^{-5}$**

Abbreviations: Percentage α-helix, percentage of positions in α-helical conformation (H, according to the DSSP assignment); percentage β-strand, percentage of positions in β-strand conformation (E, according to the DSSP assignment); percentage coil, percentage of positions in coil conformation (everything other than helix and sheet); percentage abrupt turn, percentage of positions in abrupt turns (S, according to the DSSP assignment).

Each cell shows the Spearman rank correlation between a particular property given by the row name and the accuracy of a particular SVM predictor given by the column name. A correlation is considered to be highly significant if its *p* value is less than 0.05/7 (0.0071), a Bonferroni-adjusted α-level for seven tests. This adjustment takes into account the fact that for the same SVM predictor we compute correlations with seven different properties. A correlation is considered to be not significant if the *p* value is greater than 0.05. Highly significant differences are shown in boldface type. All other notation is the same as in Table II. Results for balanced test datasets.

meaning that accuracy is lower for long proteins with large DNA-binding interface. The accuracy of pssm-SVM also shows a significant positive correlation with the number of sequences used to construct PSSM, meaning that a large number of homologous sequences tend to improve the prediction. However, this correlation is not perfect, which means that an increase in the number of available homologous sequences does not always correspond to an increase in accuracy. This suggests that other factors may affect the prediction accuracy as well.

Finally, we studied correlation between prediction accuracy and amino acid content. The results summarized in Table IV indicate the following:

1. The accuracy of seq-SVM is correlated only with the content of Tryptophan. All other significant correlations are observed only for pssm-SVM.
2. The most significant correlations are all negative. They include correlations with the percentage of glycine and proline that often occur in coil and abrupt turns. This observation is consistent with the negative correlation observed between accuracy and the percentage of positions in coil and abrupt turn structure. The observed negative correlation between accuracy and the percentage of certain amino acids with strong β-sheet propensity, such as tryptophan, tyrosine, and valine, can be explained by the fact that the percentage of these

TABLE IV. Correlations between the Accuracy of SVM Classifiers and Amino Acid Content

	seq-SVM Accuracy	seq-str-SVM Accuracy	pssm-SVM Accuracy	pssm-str-SVM Accuracy
Trp	−0.25 $p = 0.039$	−0.26 $p = 0.033$	**−0.40 $p = 0.001$**	n/s
Tyr	n/s	n/s	−0.30 $p = 0.015$	−0.36 $p = 0.003$
Val	n/s	n/s	−0.24 $p = 0.048$	n/s
Cys	n/s	n/s	−0.29 $p = 0.017$	−0.29 $p = 0.018$
Gly	n/s	n/s	−0.29 $p = 0.02$	−0.28 $p = 0.02$
Pro	n/s	n/s	**−0.41 $p = 0.0007$**	**−0.46 $p = 0.0001$**
Ser	n/s	n/s	0.31 $p = 0.012$	0.30 $p = 0.015$
Arg	n/s	n/s	0.34 $p = 0.006$	0.37 $p = 0.003$
His	n/s	n/s	n/s	**−0.38 $p = 0.002$**
Glu	n/s	n/s	n/s	0.25 $p = 0.047$

Each cell shows the Spearman rank correlation between the content of a particular amino acid type and the accuracy of a particular SVM predictor given by the column name. Amino acid content is computed as the percentage of positions occupied by a particular amino acid type

A correlation is considered to be highly significant if its p value is less than 0.05/20 (0.0025), a Bonferroni-adjusted α-level for 20 tests. This adjustment takes into account the fact that for the same SVM predictor we compute correlations with 20 different amino acid types.

Highly significant differences are shown in boldface type. Amino acid types that do not have any significant correlations with accuracy are not shown. All other notation is the same as in Tables II and III. Results for balanced test datasets.

residues is positively correlated with the β-sheet content. Tryptophan and tyrosine also do not have a strong propensity for being either interacting or noninteracting residues[6] which makes them hard to classify using a pattern recognition method. A negative correlation with cysteine content is probably explained by the fact that disulfide bridges make protein topology more unusual and affect typical sequence and structural patterns. The observed negative correlations with the percentage of histidine supports previously made assumption that, because of its ability to exist in different charge states, the presence of this amino acid in DNA-binding interface makes its prediction more complicated.[3] The observed negative correlation with the percentage of serine does not have a straightforward interpretation.

3. The observed positive correlation with the content of arginine can be attributed to the fact that this positively charged amino acid has the strongest propensity to be in contact with negatively charged DNA.[6] This strong propensity for interaction can be easily learned by pattern recognition methods. Surprisingly, there is no significant correlation between accuracy and content of lysine, another positively charged amino acid often found in DNA-binding interfaces. The observed positive correlation between accuracy and the content of negatively charged glutamic acid can be attributed to the fact that this amino acid has a strong propensity to be a noninteracting residue.[6]

DISCUSSION

Comparison of the performance of our Support Vector Machine predictor based on the evolutionary information (pssm-SVM) to the previously reported PSSM-based neural network predictor trained and tested on the same dataset[6] indicates that the performance of pssm-SVM is significantly better. An SVM predictor that uses both evolutionary and low-resolution structural information shows the best performance. Previously, it has been reported that using a profile of evolutionary conservation is not helpful for predicting DNA-binding interfaces.[2] Our results do not support this conclusion. A detailed analysis of the performance of pssm-SVM on each protein individually indicates that adding the profile of evolutionary conservation improves the prediction for most proteins in the dataset. However, for some proteins adding evolutionary conservation actually decreases accuracy (Fig. 2). This decrease in accuracy cannot always be explained by a small number of homologs found by PSI-BLAST, because even a PSSM constructed using few available homologs (3 to 5) can significantly increase prediction accuracy. In general, pssm-SVM performs better if a large number of sequences are used to construct PSSM. However, this is not always the case. This suggests that other factors, such as alignment quality and diversity, may affect the performance of pssm-SVM. A potential way of improving its performance is to use manually curated multiple alignments of better quality than those automatically generated by PSI-BLAST and implement a better scheme for weighting homologous sequences in order to remove the effect of over-represented sequence families.

We observe that, on average, adding low-resolution structural information in the form of relative solvent accessibility, secondary structure, and spatial neighbors increases accuracy. However, the low-resolution structural information can also adversely affect prediction results for some proteins. This adverse effect is especially profound in the case of sequence-based prediction when evolutionary conservation is not utilized (Fig. 3). One potential source of this problem may be that the structural descriptors we used do not capture most relevant properties of DNA-binding interfaces such as the electrostatic potential.[2] We therefore need structural descriptors with better discriminative ability with respect to DNA-binding residues. Another potential problem with using structural information is that many proteins undergo considerable conformational changes upon binding to DNA. It is unknown to what extent such conformational changes may affect even low-resolution structural descriptors. If the difference in structural properties between bound and unbound conformations is too profound, a pattern recognition method trained using structures of protein–DNA complexes will not be able to generalize successfully to predict DNA-binding interfaces on unbound proteins.

We also observe that SVM classifiers perform significantly better on proteins from mainly-α structural class and that there is a significant positive correlation between the percentage of residues in α-helical conformation and prediction accuracy. This observation suggests that DNA interaction patterns formed by α-helices are easier to predict using pattern recognition methods. This is somewhat similar to protein structure prediction methods which predict the structure of α-helical proteins much more accurately than the structure of proteins that contain β-sheets. Another possible reason for better performance on mainly-α proteins is that α-helices are over-represented in DNA-binding sites. Because of the lack of nonhelical patterns, a pattern recognition method may become overtrained to predict helical interfaces. In this work we tested only SVM predictors, however, it is most likely that all machine learning methods will perform significantly better on mainly-α proteins because of the aforementioned bias in the limited dataset of experimentally characterized protein–DNA complexes and the nature of interaction patterns observed in this structural class.

CONCLUSIONS

We developed a series of Support Vector Machine classifiers for the prediction of DNA-binding sites in DNA-binding proteins. Our results indicate that including the profile of evolutionary conservation of sequence positions in the form of a properly scaled Position Specific Scoring Matrix obtained using a nonredundant sequence database significantly improves the accuracy of the prediction of DNA-binding sites. The highest prediction accuracy is achieved using a classifier that utilizes a combination of evolutionary conservation and low-resolution structural information.

For some proteins, including evolutionary conservation or low-resolution structural information may actually decrease prediction accuracy compared to the prediction based on protein sequence alone.

Prediction of DNA-binding interfaces in mainly-α proteins that have a relatively simple topology is significantly more accurate than in more topologically complex β-sheet–containing proteins.

The accuracy of SVM classifiers shows significant correlations with sequence length, secondary structure content, and amino acid content. The observed correlations suggest that it may be possible to assign a reliability index to the overall accuracy of the prediction of DNA-binding sites in any given DNA-binding protein using its sequence and structural properties.

A web-server implementation of the predictors is freely available online at http://lcg.rit.albany.edu/dp-bind/.

REFERENCES

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.
2. Jones S, Shanahan HP, Berman HM, Thornton JM. Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. Nucleic Acids Res 2003;31:7189–7198.
3. Tsuchiya Y, Kinoshita K, Nakamura H. Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. Proteins 2004;55:885–894.
4. Landgraf R, Xenarios I, Eisenberg D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol 2001;307:1487–1502.
5. Panchenko AR, Kondrashov F, Bryant S. Prediction of functional sites by analysis of sequence and structure conservation. Protein Sci 2004;13:884–892.
6. Ahmad S, Gromiha MM, Sarai A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics 2004; 20:477–486.
7. Ahmad S, Sarai A. PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformatics 2005;6:33.
8. Vapnik VN. Statistical learning theory. New York: Wiley; 1998.
9. Hearst MA. Support vector machines. IEEE Intelligent Systems 1998;13:18–28.
10. Byvatov E, Fechner U, Sadowski J, Schneider G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. J Chem Inf Comput Sci 2003;43:1882–1889.
11. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, and Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
12. Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 2001;17:282–283.
13. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
14. Boser BE, Guyon IM, Vapnik V. A training algorithm for optimum margin classifiers. In: Haussler D, editor. Proceedings of the fifth annual workshop on computational learning theory. Pittsburgh, PA: ACM Press; 1992. p. 144–152.
15. Cristianini N, Shawe-Taylor J. Support vector machines and other kernel-based learning methods. Cambridge, MA: Cambridge University Press; 2003.
16. Hua S, Sun Z. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J Mol Biol 2001;308:397–407.
17. Karchin R, Karplus K, Haussler D. Classifying G-protein coupled receptors with support vector machines. Bioinformatics 2002;18: 147–159.
18. Kim JH, Lee J, Oh B, Kimm K, Koh I. Prediction of phosphorylation sites using SVMs. Bioinformatics 2004;20:3179–3184.
19. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 2004;337:635–645.
20. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. 2003. Available from: http://www.csie.ntu.edu.tw/~cjlin/libsvm.
21. Atalay V, Cetin-Atalay R. Implicit motif distribution based hybrid computational kernel for sequence classification. Bioinformatics 2005;21:1429–1436.
22. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 1992;89:10915–0919.
23. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
24. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 2000;16:412–424.
25. Kim H, Park H. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. Proteins 2004;54:557–562.
26. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH — a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.