

Kernel-based machine learning protocol for predicting DNA-binding proteins

Nitin Bhardwaj, Robert E. Langlois, Guijun Zhao and Hui Lu*

Bioinformatics Program, Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607, USA

Received June 20, 2005; Revised August 2, 2005; Accepted October 18, 2005

ABSTRACT

DNA-binding proteins (DNA-BPs) play a pivotal role in various intra- and extra-cellular activities ranging from DNA replication to gene expression control. Attempts have been made to identify DNA-BPs based on their sequence and structural information with moderate accuracy. Here we develop a machine learning protocol for the prediction of DNA-BPs where the classifier is Support Vector Machines (SVMs). Information used for classification is derived from characteristics that include surface and overall composition, overall charge and positive potential patches on the protein surface. In total 121 DNA-BPs and 238 non-binding proteins are used to build and evaluate the protocol. In self-consistency, accuracy value of 100% has been achieved. For cross-validation (CV) optimization over entire dataset, we report an accuracy of 90%. Using leave 1-pair holdout evaluation, the accuracy of 86.3% has been achieved. When we restrict the dataset to less than 20% sequence identity amongst the proteins, the holdout accuracy is achieved at 85.8%. Furthermore, seven DNA-BPs with unbounded structures are all correctly predicted. The current performances are better than results published previously. The higher accuracy value achieved here originates from two factors: the ability of the SVM to handle features that demonstrate a wide range of discriminatory power and, a different definition of the positive patch. Since our protocol does not lean on sequence or structural homology, it can be used to identify or predict proteins with DNA-binding function(s) regardless of their homology to the known ones.

INTRODUCTION

The number of genes encoding for DNA-binding proteins (DNA-BPs) in the human genome has been pegged at 6–7% by comparative sequence analysis (1). These proteins play key roles in molecular biology, such as recognizing specific nucleotide sequences, regulation of transcription, maintenance of cellular DNA, DNA repair, DNA packaging and recombination and control of replication (1). Protein–DNA interactions also play other crucial roles in the cell. In eukaryotic cells chromosomal DNA is packaged into a compact structure with the help of histones. Restriction enzymes are DNA-cutting enzymes found in bacteria that recognize and cut DNA only at a particular sequence of nucleotides to serve a host-defense role. Being at the core of such momentous processes, protein–DNA interactions have received a commensurate interest (2–5). There have been studies to detect (6,7), design (8) and predict them using a probabilistic recognition code (9). There have also been works towards analyzing protein–DNA recognition mechanism (10) and binding site discovery (11).

DNA-BPs represent a broad category of proteins, known to be highly diverse in sequence and structure. Structurally, they have been divided into 54 protein-structural families (1). With such a high degree of variance, using conventional annotation methods rooted in database searching for sequence similarity (12), profile or motif similarity (13) and phylogenetic profiles (14) may not lead to reliable annotations. In this context, a DNA-BP prediction protocol that takes into account the structural information and does not depend on sequential or structural homology to proteins with known functions will be very useful.

Previously, there have been a few bioinformatics methods developed towards automated identification and prediction of DNA-BPs. Cai and Lin (15) used pseudo-amino acid composition to identify proteins that bind to RNA, rRNA and DNA. Ahmad *et al.* (16) integrated structural information with a neural network approach for the prediction of DNA-BPs.

*To whom correspondence should be addressed at Department of Bioengineering, University of Illinois at Chicago, 851 S. Morgan Street (M/C 063), Room 218, Chicago, IL 60607, USA. Tel: +1 312 413 2021; Fax: +1 312 413 2018; Email: huilu@uic.edu

Stawiski *et al.* and Jones *et al.* (17,18) characterized electrostatic features of proteins for an automated approach to DNA-BP and DNA-binding site prediction. Ahamd and Sarai (19) showed that overall charge and electric moment can be used to identify DNA-BPs. Tsuchiya *et al.* (20) combined structural features with electrostatic properties of the proteins. Accuracy rates achieved in these methods varied from 65% to 86% depending on both the features used and the validation method adopted.

In this work, we build a support vector machine (SVM)-based classification model to distinguish DNA-BPs from non-binding ones with high accuracy and investigate different features (21). Application of SVM in bioinformatics to various topics has been explored (22–24). Implementations of SVM to protein fold recognition, including our own, has achieved superior performance over neural-networks (25,26).

The goal of the current work is 2-fold: the implementation of a robust protocol using SVM for DNA-BPs prediction and the development of meaningful descriptors. We characterize the structural and sequential features of DNA-BPs and use them to develop the protocol. The validation showed the current protocol outperformed other published data. The results indicate the plausibility of an application of kernel-based machine learning methods to identify and predict DNA-BPs. The approach will be refined as more knowledge becomes available about the determinants of protein–DNA binding so that more features will be included.

Organization of the paper is as follows. First, in Materials and Methods, we describe the features that are used, the implementation of SVM and the evaluation method adopted. In Results, we present the discriminative power of individual features and their combined performance using SVM. In Discussion, we interpret the origin of our improved performance and suggest possible future research directions.

MATERIALS AND METHODS

Dataset

A positive dataset of 121 DNA-BPs was obtained from a union of datasets used in previous related studies (16,17,27). The complexes in the dataset were better than 3 Å in resolution. A negative dataset of 238 non-DNA-BPs with resolution better than 3 Å was also adopted from an earlier study (17). These proteins have <35% sequence identity between each pairs. The 121 DNA-BPs are further reduced to 83 proteins to ensure that there is no >20% sequence identity between any pairs. A complete list of all the cases can be downloaded from our webpage (<http://proteomics.bioengr.uic.edu/pro-dna>).

Problem formulation

In simple words, the binary classification problem being studied here can be stated as: can we predict if a given protein belongs to DNA-binding or non-DNA-binding class? When translated into machine-learning lexicon, sequence and structure form the ‘state space’ of the problem. Some characteristics of the state space, called the descriptors, which are believed to be important for classification, are formulated into a fixed length feature vector by ‘feature generation’. Also, every member in the dataset is associated with one of the two ‘class labels’: DNA-binding or non-DNA-binding.

After the formulation of the feature vectors for every member, they are input into the classifier. There are two parts to classification: training and testing. During training, the class of every input vector is known in advance. The classifier then adopts its own method of building a classification model that minimizes the empirical error. During testing, when the class of the input vector is not known beforehand, the classifier then uses the classification model built during training to predict the class of each member and outputs it. To achieve high accuracy in classification, thus, the choice of good descriptors that can distinguish the two classes is very important.

Feature design

The features explored in this study include positive potential surface patches, overall charge of the protein and overall/surface composition. Each class of features is described below.

Overall charge. Overall charge of a protein comprises a single attribute subset in the SVM feature vector. Hydrogen atoms were added to all the proteins using a publicly available tool REDUCE (28). Then charges were assigned to all the atoms employing the CHARMM force field parameters (29). Histidine residues were assigned a neutral charge.

Electrostatic calculations and patch formation. The program Delphi was used for all electrostatic calculations in this study (30–32). It solves the non-linear Poisson–Boltzmann equation using finite-difference methods to calculate the potential at specified points. Potential on the coordinate of every atom of the protein was calculated in the absence of the DNA and reported. The CHARMM force-field was used for assignment of partial charges to all the atoms of the protein and Debye–Huckel boundary conditions were employed. Probe radii of 1.4 Å and a stern (ion-exclusion) layer of 2 Å were specified. Salt concentration and temperature were fixed at 145 mM and 298 K, respectively. The dielectric constants used were 2.0 and 80.0 for protein interior and the solvent. A fine-resolution grid structure with a scale (grids/Å) of 2 was employed. The percentage fill specified was 50%, meaning that protein fills half of the total volume of the grid cubic. The center of the grid architecture was translated to the geometric center of the protein.

The positive surface patches are identified with an iterative growing algorithm. Surface residues are defined as the ones that have more than 40% of their area exposed to water as calculated by DSSP (33). A surface atom with potential higher than 200 kT/e was used as the starting point for the patch and all surface atoms having a positive potential and falling within a distance of 2 Å were added to the patch. Each of these atoms was then used as the center for further expansion of the patch. When the process converges, an atom with positive potential higher than 200 kT/e that doesn’t belong to this patch starts a new patch formation. The size of a patch was defined by the number of atoms it contained. Usually there is more than one patch formed on each protein. These patches are sorted by the size. The size of the largest patch was used as a feature in SVM. We also used the aggregate size of the largest four patches as features but using the size of the largest patch gave the best performance.

Amino acid composition. We computed two compositions (% of the 20 amino acids) for a protein: overall and surface. For overall composition, all residues were used. For surface composition only calculation residues having more than 40% of their surface area accessible to solvent were used. Each kind of composition is a 20-dimensional input feature sub-vector so that amino acid composition becomes a 40-dimensional sub-vector.

Prediction protocol

Classifier. We use SVM for this classification problem. SVM is a binary classification tool that uses a non-linear transformation to map the input data to a high dimensional feature space where linear classification is performed. It is equivalent to solving the quadratic optimization problem:

$$\min_{w, b, \xi_i} \frac{1}{2} w \cdot w + C \sum_i \xi_i \quad 1$$

$$\text{s.t. } y_i(\phi(x_i) \cdot w + b) \geq 1 - \xi_i, \quad i = 1, \dots, m, \quad 2$$

$$\xi_i \geq 0, \quad i = 1, \dots, m,$$

where x_i is a feature vector labeled by $y_i \in \{+1, -1\}$, (x_i, y_i) , $i = 1, \dots, m$, and C , called the cost, is the penalty parameter of the error term. The given model summarizes the so-called soft-margin SVM, which tolerates noise within the data. It does so by generating a separating hyper-plane using the equation $f(x) = \phi(x) \cdot w + b = 0$. Through the representation of $w = \sum_j \alpha_j \phi(x_j)$ we obtain $\phi(x_i) \cdot w = \sum_j \alpha_j \phi(x_j) \cdot \phi(x_i)$. This provides an efficient approach to solve SVM without the explicit use of the non-linear transformation (34). Further $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ is called the kernel function and it is this function that maps the data to a higher dimension. Here we use the polynomial kernel, which is of the form: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d$, $\gamma > 0$ where γ , r and d are kernel parameters. These parameters are searched to give the best model. While a fixed value of $r (=0)$ was used, γ , d and the cost (C) in the soft-margin SVM were optimized based on grid search. Publicly available LIBSVM was used to build a classifier with a given kernel and a set of parameters (35).

Evaluation methods. First, the dataset was divided into training set and testing set. The parameters for SVM were found using cross-validation (CV) over the training set by maximizing the accuracy. Then, the SVM was validated against the untouched test set. This procedure was repeated N times.

Three methods are used to evaluate the performance. First is the self-consistency test, also called re-substitution where training (model building) and testing are done on the same dataset. Self-consistency demonstrates how well SVM has turned into internal knowledge. Second evaluation technique adopted is the N -fold CV test. In this method, the whole dataset is randomly separated into N parts. Each time, one part is retained for testing and all others form training dataset. This is repeated until each part forms the testing dataset exactly once. The parameters giving the best average accuracy are kept to form the classification model. Note that, although the parameters are optimized based on the testing set, the decision line and selection of support vectors are based on the training set. Thus CV is different from self-consistency. Two implementations of N -fold CV used here are: 5-fold CV, where the entire

dataset is divided into 5 parts, and leave-one-out (or jackknife test) where N equals the total number of proteins in the dataset, meaning each protein is left out for testing exactly once. Third method of evaluation is the holdout test. The total dataset is randomly divided into two halves with approximately equal number of positive and negative cases. SVM is then trained on one of the two sub-sets with CV to find the best parameters with no regards to the other one. These parameters are then used on the other subset and the performance is reported. It should be noted here that the holdout method is different from 2-fold CV of the whole set, where both the datasets are used for training and testing and the parameters giving the best 'average' accuracy are kept. In holdout test, which is mimicking a true prediction, only one of the two subsets is used for searching the optimum parameters, which are then used to predict the class of every member of the other subset. However this evaluation can have a very high variance depending on the division of the data into training and testing subset. An alternative way to circumvent this is to run the classifier for a number of times and then analyze the performance on the basis of these runs. We iterated this process for 125 times, each time randomly dividing the data into training and testing set. Each time the performance was reported. Apart from accuracy (% of total correct predictions), we also report sensitivity and specificity, which are fractions of positive and negative cases correctly classified, respectively.

RESULTS

We first present various features used in this study and their propensities in binding and non-binding proteins. Following that, we present the model building using SVM with all the features included and the performances using various evaluation methods.

Feature class 1: overall charge

Distribution of overall charge showed significant differences for binding and non-binding proteins (Figure 1). One can see from the figure that 87% of non-binding proteins were negatively charged whereas only 35% of the binding proteins were

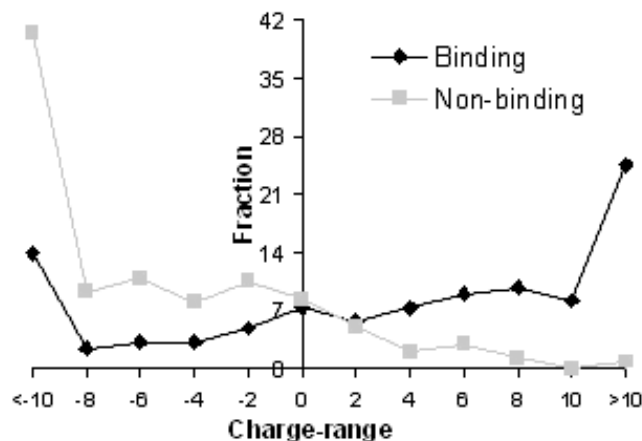


Figure 1. Distribution of overall charge for binding and non-binding cases in Electronic charge units (e). Labels on the x-axis indicate the upper value of the bin. e.g. 2 indicates the bin 0 to 2.

Table 1. Performance of the SVM for different combinations of descriptors, classifiers and evaluation techniques

Descriptor(s)	Classifier	Validation	Accuracy (%)	Sensitivity (%)	Specificity (%)	Parameters		
						<i>D</i>	<i>C</i>	γ
Overall charge	Linear	Jackknife	82.4	58.6	95.7	1	11	0.01
Overall charge + patch size	Linear	Jackknife	83.8	60.3	96.6	1	13	0.02
All	Polynomial	Self-consistency	100	100	100	2	10	0.309
		Jackknife	90.5	81.8	94.9	2	19	0.054
		5-fold CV	89.1	82.1	93.9	2	19	0.051
		5-fold CV (20%)	90.3	67.4	94.9	2	23	0.034
		Leave-half holdout	83.3	82.5	83.5	–	–	–
All	Polynomial	Leave 1-pair holdout	86.3	80.6	87.5	–	–	–
		Leave 1-pair holdout (20%)	85.8	81.6	87.8	–	–	–
		Leave 1-pair holdout (20%)	85.8	81.6	87.8	–	–	–

Reported in the last column are the parameters giving the corresponding accuracies. Sensitivity and specificity were defined as $TP/(TP + FN)$ and $TN/(TN + FP)$, respectively, where T = True, F = False, P = Positive and N = Negative.

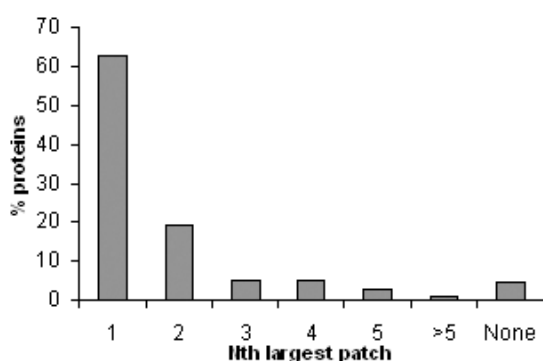


Figure 2. Overlap of the surface positive potential patches with the DNA-binding interface. *x*-axis represents the *N*th biggest patch and *y*-axis exhibits the % of total binding proteins having the largest overlap with that patch. For example, 63% of the proteins had the largest overlap with the 1st largest patch.

negatively charged. Around one-fourth of the binding proteins had an overall charge greater than 10e as compared to only 1% of the non-binding ones. In the intermediate positive range, there were consistently a larger number of binding cases than non-binding ones.

Thus, overall charge can be expected to distinguish the two cases with certain accuracy. Indeed, with this as the only feature combined with a linear classifier adopted from SVM, an accuracy value of 82.4% could be achieved with the jackknife evaluation method (Table 1).

Feature class 2: electrostatic patch

Since DNA is negatively charged, a large positive potential patch is deemed to be important in driving the protein to the DNA. The patches are calculated following the procedure described in Materials and Methods. Then the patches are sorted based on their sizes (number of atoms included in the patch) and the size of the largest patch is used as a descriptor.

To ascertain the propriety of this feature, we examined the overlap between our patches and the actual protein–DNA ‘interface’ atoms. A residue was designated as ‘interface’ residue if any heavy atom in that residue is closer than 4.5 Å to the DNA and the all its atoms were classified as ‘interface’ atoms. The overlap between interface and each positive patch was calculated and sorted (Figure 2). We found that in 63% of the DNA-BPs, the largest patch has the biggest

overlap with the DNA-binding interface. For 20% of the proteins, the second largest patch has the largest overlap with the DNA interface. In only 1% (1 out of 120 proteins), the patch that had the largest overlay with the interface was not among the largest 5 in size. In 4% (5 cases), the DNA-binding interface has no overlap with positive patches.

When the above two features are combined for a linear classification of binding and non-binding proteins, we anticipate SVM to achieve a higher accuracy. Figure 3 plots the overall charge against the size of the largest patch for the binding (black) and the non-binding proteins (grey). In the intermediate range of the size of the largest patch, DNA-binding and non-DNA-BPs show some overlap but the separation becomes finer towards the higher extreme of the range. As expected, addition of size of the largest patch to overall charge as a feature vector further increases the accuracy values. With a linear classifier in SVM and these two features, DNA-BPs could be identified with 83.8% accuracy when evaluated on jackknife test.

Feature class 3: amino acid composition

Two kinds of amino acid composition are computed here: overall and surface (Figure 4). In case of overall composition, noticeable differences in binding and non-binding cases were observed with respect to the frequency of Lys and Arg. They both are positively charged amino acids so their over-representation in DNA-BPs is fairly implicit.

As expected, surface composition was more disposed than overall composition for hydrophobic residues such as Trp, Phe, Tyr, Cys, Ile and Met. A higher frequency of Arg and Lys in binding proteins than non-binding ones was observed. A lower level of Asp can also be explained since DNA is negatively charged. Interestingly there is no difference between the frequencies of Glu. Other amino acids have the similar composition in both the binding and unbinding cases. From these observations one can expect that using composition alone won't be as efficient to linearly classify the DNA-BPs. So, we use them in conjunction with previously discussed features with non-linear kernel-based SVM.

Prediction of DNA-BPs

We combined all three classes of feature vectors and used them to train and test the SVM. For self-consistency, we could achieve an accuracy value of 100% (Table 1). This

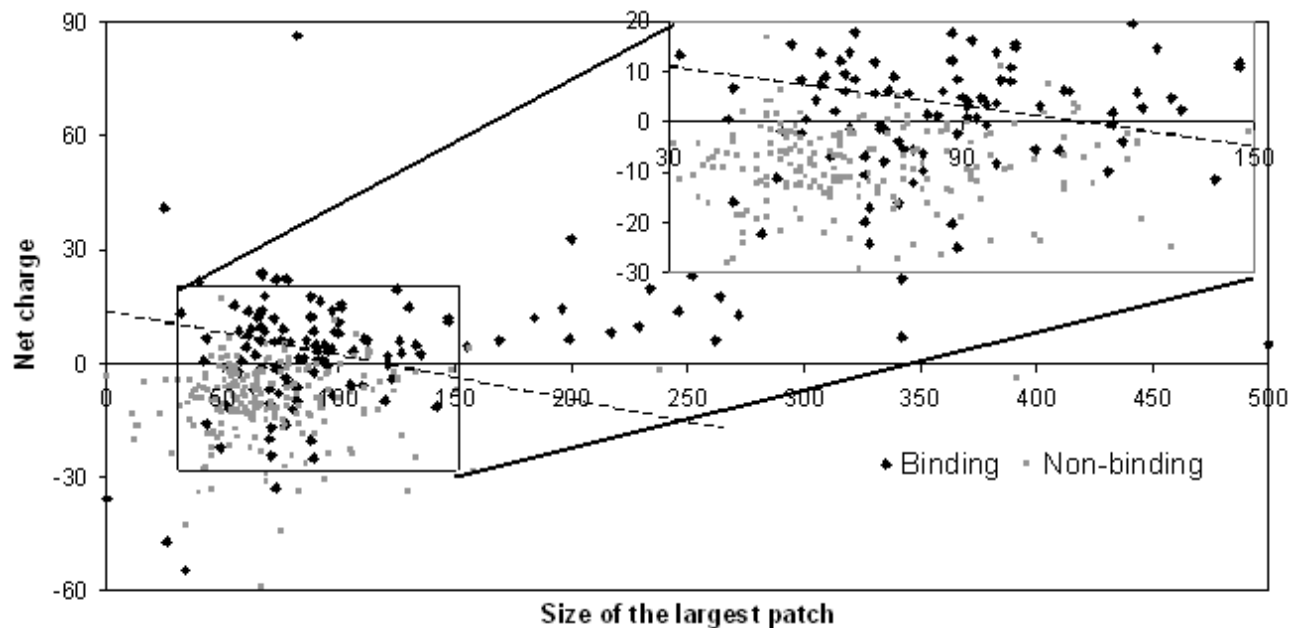


Figure 3. Linear separation of DNA-binding and non-DNA-BPs using only overall charge and size of the largest patch as the features as used by the SVM. Net charge is in electrostatic charge units.

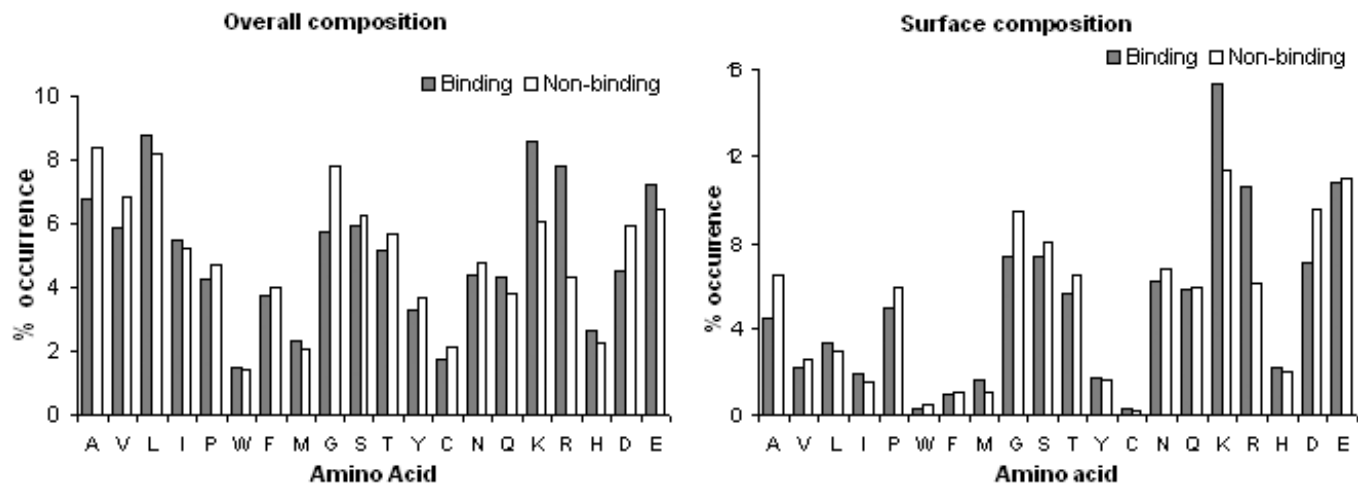


Figure 4. Frequency of different amino acids in overall and surface composition. The difference in the height of a bar for an amino acid for binding and non-binding cases reflects a stronger preference for that amino acid in one case over another.

implies that SVM could cogently capture the intrinsic correlation between feature vectors and the classification being sought.

During CV, SVM is tested on how well it can predict on the basis of optimized parameters chosen during training. For both 5-fold CV and jackknife approaches employed, the performances of the current protocol were almost the same. Using Jackknife test, by optimizing the parameters with a polynomial kernel, we could achieve an accuracy value of 90.5% (325 correct predictions out of 359 cases) (Table 1). Corresponding sensitivity and specificity values were 81.8 and 94.9%. Similarly for a 5-fold CV technique, average correct predictions of 89.1% could be made with the optimum parameters with 82.1% sensitivity and 93.9% specificity (Table 1). Comparable

values of sensitivity and specificity show that the performance of our protocol is poised.

To make sure that the performance of our protocol is not biased towards weak homology, we filtered the proteins using a sequence identity cut-off of 20% i.e. one of the proteins of all the pairs having >20% was removed. This assured that a protein in a test set has no sequence similarity at all to proteins in the train set. For the two sets we obtained accuracy values of 90.3%, which are very close to the accuracy value for the entire dataset (90.5%, Table 1). This shows that the above protocol was not biased due to homology recognition.

Finally, the holdout method is used for validation of the performance of current SVM protocol. The accuracy achieved in this test corresponds to the ones from true blind predictions.

We iterated the method, which displayed a high deviation over 125 runs (Figure 5). Training accuracies varied from 79.8% to 92.7%, and testing accuracies ranged between 76.6% and 88.3%, with an average of 83.4%. Training accuracy was a little higher than testing accuracy because when SVM is being trained, it builds a classification model only on the basis of the training set, which is not optimized for the testing set. Overall the two accuracies showed a good correlation. To directly compare with the previous results (19), we also evaluated our performance for another holdout technique where a pair of DNA-binding and non-binding protein (selected randomly) was removed and the remaining dataset formed the training test. The left pair was then tested upon and this was repeated 500 times. Evaluated over these 1000 (500×2) predictions, we achieved an accuracy of 86.3%, (Table 1) which is higher than previously reported value of 83.9% (19).

Similar to the performance checking in CV, we also performed the leave-pair out hold out for reduced dataset that

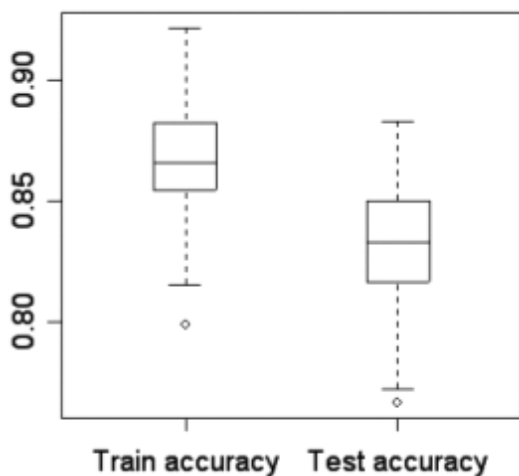


Figure 5. Box and whisker plot for the performance of the SVM for holdout technique. The performance data was collected over 125 runs.

only proteins have <20% sequence identity are included. The accuracy value is 85.8%, which showed our protocol again doesn't depend on sequence homology to perform DNA-BP prediction. Online server has been provided with the current protocol on our webpage (<http://proteomics.bioengr.uic.edu/pro-dna>).

One class of non-DNA-BP is RNA-binding protein. Since RNA-binding protein do share similarities as DNA-BP (G. Zhao and H. Lu, unpublished data), it is expected our protocol may not perform as well when the negative cases are actually RNA-binding proteins. For this purpose, a list of 37 RNA-binding proteins was compiled from ref. (36) and the features used above for classification were calculated. Using the model built from our previous dataset (none of the RNA-binding protein was included in optimization), the classification of RNA-binding protein resulted 21 correct prediction, and 16 wrong prediction. Thus we decided to perform a further binary classification model for the DNA- and RNA-binding proteins, this model resulted in 91.3% accuracy for the binary classification. Thus this post-processing can give reasonable accuracy to exclude RNA-binding proteins from the false positives in the DNA-BP predictions. It is an ongoing effort to design features specifically to distinguish the DNA-binding from RNA-binding proteins. The effort will naturally fall into the machine learning prediction categories.

We also attempted to identify DNA-BPs whose structure was solved without the DNA (the unbound cases). We found seven such proteins where native structures are available in both bound and unbound states. We structurally aligned the two states for each of the 7 cases using publicly available tool Superpose (37). RMSD values ranged from 0.53 Å to 5.11 Å between the corresponding structures. In SVM prediction, we used the built model where these 7 proteins were not included in the optimization process. Encouragingly, SVM classified correctly all of proteins using unbound structures. This shows that the above protocol can be used to identify DNA-BPs even before there is any deformation induced due to the binding of the DNA, if any.

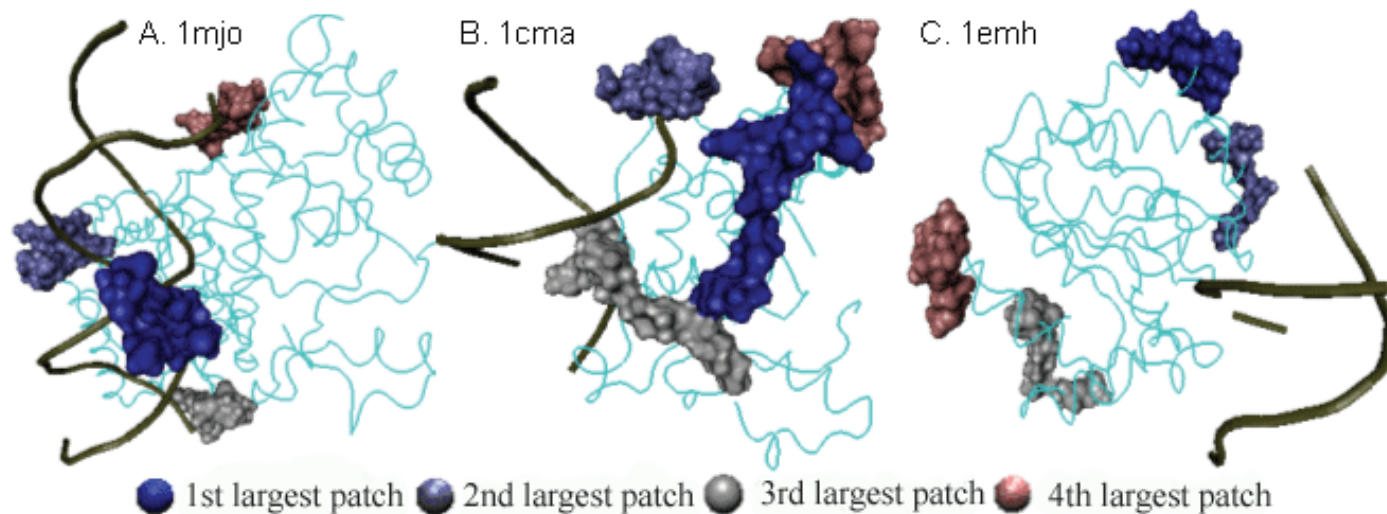


Figure 6. Orientation of the largest four patches with respect to the DNA. Remaining protein is shown in 'tube' representation in cyan. DNA is displayed in golden color with only backbone shown. Patches were mapped on the surface by adapting the PDB files for labeling the atoms in each patch and visualized in VMD (41).

Figure 6 shows three examples of different kinds of prediction results from SVM. In Figure 6A, 1 mjo was correctly classified as DNA-binding. The positive patches are near the DNA-binding surface. The structure of DNA was not used in the prediction and electrostatic calculation, nor was the information of binding sites. In this case, not only the protein is correctly classified, the positive patch can also give indications about the location of the binding site. Figure 6B shows another case where the protein (1 cma) was correctly predicted as DNA-binding one. Although, the largest patch is not so close to DNA, the 2nd and 3rd patches are very close to binding site. In this particular case, if we further want to push the idea of binding site prediction, other information in addition to location of patches would be added, as described in Bhardwaj and Lu (38). Finally there were about 10% of the cases where we incorrectly predicted the DNA binding behavior. Figure 6C is one of these cases (1 emh) that was misclassified as non-binding. In this protein, all the four positive patches are small and are far from the actual binding site (which we did not know during class prediction). Obviously this protein does not have features consistent with most of the other proteins and hence generic rules failed to classify it correctly. Such diverse behaviors show that more work needs to be done to add features to describe different DNA-binding mechanisms.

DISCUSSION

We have implemented a kernel-based method (SVM) towards developing a robust protocol for identification of DNA-BPs. We have appraised the performance of the SVM using self-consistency, CV and holdout evaluation. We achieved an accuracy of 90% for CV and 100% for self-consistency. These values are higher than all previously published studies where the accuracy ranged from 67% to 86% (15–17,19). We also report accuracy value of 86.3% for leave 1-pair holdout technique, which is higher than that reported in a previous study using the same technique (19). Tsuchiya *et al.* (20), who used linear separation, reported 86% and 96% accuracy for DNA-binding and non-DNA-binding cases using self-consistency. Since SVM uses an entirely different strategy to achieve maximal self-consistency accuracy, there is no direct comparison between their values and self-consistency value of 100% reported in this paper.

Similar evaluation of CV and holdout have been applied to a more difficult (and smaller) test set where all proteins have less than 20% sequence identity. The accuracy achieved there, 90.3% for CV and 85.8% for holdout are comparable with the main results. These numbers demonstrate our protocol does not rely on even remote sequence homology for DNA-BP prediction.

We have used an ensemble of features describing DNA-BPs in order to distinguish them from non-DNA-BPs. Features used here show a varying distinction power. To judge the distinction power of a feature, we calculated Fisher's score (FS) for every feature, j as: $FS_j = (\mu_1 - \mu_2)^2 / (s_1^2 + s_2^2)$ where μ_i and s_i are the mean and standard deviation of the feature in class i , respectively. Higher the FS, the more discriminative power this feature has. The top ten features with highest FS were: overall and surface fraction of Arg, overall charge,

overall fraction of Gly, Asp, Lys, size of the largest patch, and the surface fraction of Asp, Lys and Gly. The greater discriminatory power of these features is also reflected in their contrasting distribution in the corresponding positive and negative cases. While it is expected the positive charged residues Arg, Lys and overall charge contribute to the good performance, there are some very interesting observations such as percentage of Gly plays important role in discriminating DNA-BPs. We have seen that there is less Gly in DNA-BPs than in non-binding ones. This data may suggest that DNA-BPs in general are more rigid than non-binding ones.

Similarly, the top ten features showing highest power to distinguish DNA-BPs from RNA-binding ones were: overall fraction of His, largest patch size, overall charge, overall fraction of Trp, surface composition of Asp, Ile, His and Cys and overall fraction of Asp and Val. In this discrimination, we noticed besides the electrostatic properties, the percentage of His and Trp played important roles. Note, while Fisher's score indicates the predictive power of an individual feature, it is important to remember that the relationships between features that ultimately determine classification. Extracting and interpreting such information from an SVM model is still an open research issue.

As shown in the Results, these features when combined together can predict DNA-BPs from non-DNA-binding ones more accurately than when they are used separately. Apart from establishing the fact there is an inherent correlation between these features, this observation also authenticates the application of kernel-based machine learning methods to take advantage of this correlation.

The current implementation uses three classes of features and the results are highly encouraging. It is natural to assume with the inclusion of more features the performance will increase. A new class of features under development is the geometry of the protein surfaces. Encouragingly, it is easy to include new features in the framework of SVM. Further, when more features available, feature selection can be used to ascertain the features that are contributing more to the overall accuracy and the ones that are just adding noise.

Besides annexation of more features, a potential channel for furthering the prediction power of the SVM would be through development of models specific to a class, a family or even motifs (39). To begin with, trends specific to some structural motifs that play a key role in binding to the DNA, such as helix–turn–helix or helix–loop–helix, could be included (40). Also, deciphering the reasons why some particular sites on the surface bind to the DNA and others do not will go a long way in improving our understanding of the determinants of DNA-binding. For misclassified proteins, it would be tempting to uncover the driving forces and agents behind both mobilization of such erratic proteins to the DNA and, selection of the binding sites.

The development of the above protocol is purported to augment contemporary function annotation tools. Since this protocol does not rely on any sequential or structural analogy, it is capable of identifying DNA-BPs even when they are embedded in novel motifs or folds. In the future course of exploration in related fields we contemplate a multitude of kernel-based machine-learning tools to be employed for the purpose of DNA-BPs identification. The method can apparently be polished by addition of more features and apt

fine-tuning of the SVM once the circumstances of protein–DNA-binding are better comprehended.

ACKNOWLEDGEMENTS

This work is partially supported by NIH P01 AI69015 and UIC startup funds to H.L. N.B. acknowledges the support from FMC Inc. Fellowship. R.E.L. is supported from pre-doctoral fellowship in NIH training grant (T32HL07692): Cellular Signaling in the Cardiovascular System. Funding to pay the Open Access publication charges for this article was provided by UIC startup funds.

Conflict of interest statement. None declared.

REFERENCES

- Luscombe, N.M., Austin, S.E., Berman, H.M. and Thornton, J.M. (2000) An overview of the structures of protein–DNA complexes. *Genome Biol.*, **1**, 1–37.
- Jones, S., Heyningen, P., Berman, H.M. and Thornton, J.M. (1999) Protein–DNA interactions: a structural analysis. *J. Mol. Biol.*, **287**, 877–896.
- Siggers, T.W., Silkov, A. and Honig, B. (2005) Structural alignment of protein–DNA interfaces: insights into the determinants of binding specificity. *J. Mol. Biol.*, **345**, 1027–1045.
- Pabo, C.O. and Necludova, L. (2000) Geometric analysis and comparison of protein–DNA interfaces: why is there no simple code for recognition? *J. Mol. Biol.*, **301**, 597–624.
- Nadassy, K., Wodak, S.J. and Janin, J. (1999) Structural features of protein–nucleic acid recognition sites. *Biochemistry*, **38**, 1999–2017.
- Alexander, M.K., Bourns, B.D. and Zakian, V.A. (2001) One-hybrid systems for detecting protein–DNA interactions. *Methods Mol. Biol.*, **177**, 241–259.
- Sathyapriya, R. and Vishveshwara, S. (2004) Interaction of DNA with clusters of amino acids in proteins. *Nucleic Acids Res.*, **32**, 4109–4118.
- Havraneck, J.J., Duarte, C.M. and Baker, D.J. (2004) A simple physical model for the prediction and design of protein–DNA interactions. *J. Mol. Biol.*, **344**, 59–70.
- Benos, P.V., Lapedes, A.S. and Stormo, G.D. (2002) Is there a code for protein–DNA recognition? *Probab(istical)ly Bioessays*, **24**, 466–475.
- Paillard, G. and Lavery, R. (2004) Analyzing protein–DNA recognition mechanisms. *Structure*, **12**, 113–22.
- Djordjevic, M., Sengupta, A.M. and Shraiman, B.I. (2003) A biophysical approach to transcription factor binding site discovery. *Genome Res.*, **13**, 2381–2390.
- Marcotte, E.M., Pellegrini, M., Ng, H.-L., Rice, D.W., Yeates, T.O. and Eisenberg, D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.
- Lu, X., Zhai, C., Gopalakrishnan, V. and Buchanan, B.G. (2004) Automatic annotation of protein motif function with Gene Ontology terms. *BMC Bioinformatics*, **5**, 122.
- Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
- Cai, Y. and Lin, S.L. (2003) Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequences. *Biochim. Biophys. Acta.*, **1648**, 127–133.
- Ahmad, S., Gromiha, M.M. and Sarai, A. (2004) Analysis of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, **20**, 477–486.
- Stawiski, E.W., Gregoret, L.M. and Mandel-Gutfreund, Y. (2003) Annotating nucleic acid-binding function based on protein structure. *J. Mol. Bio.*, **326**, 1065–1079.
- Jones, S., Shanahan, H.P., Berman, H.M. and Thornton, J.M. (2003) Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Res.*, **31**, 7189–7198.
- Ahmad, S. and Sarai, A. (2004) Moment-based prediction of DNA-binding proteins. *J. Mol. Biol.*, **341**, 65–71.
- Tsuchiya, Y., Kinoshita, K. and Nakamura, H. (2004) Structure-based prediction of DNA-binding sites on proteins using empirical preference of electrostatic potential and the shape of molecular surfaces. *Proteins*, **55**, 885–894.
- Vapnik, V. and Cortes, C. (1995) Support vector networks. *Machine Learning*, **20**, 273–293.
- Dubchak, I., Muchnik, I., Mayor, C., Dralyuk, I. and Kim, S.-H. (1999) Recognition of a protein fold in the context of the Structural Classification of Proteins (SCOP) classification. *Proteins*, **35**, 401–407.
- Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M.Jr and Haussler, D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.
- Jaakkola, T., Diekhans, M. and Haussler, D. (1999) Using the Fisher kernel method to detect remote protein homologies. *Proc. Seventh Int. Conf. Intell. Syst. Mol. Biol.*, AAAI Press, Menlo Park, CA, 149–158.
- Ding, C.H.Q. and Dubchak, I. (2001) Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, **17**, 349–358.
- Langlois, R.E., Diec, A., Perisic, O., Dai, Y. and Lu, H. (2005) Improved Protein Fold Assignment Using Support Vector Machines. *Int J Bioinformatics Research and Applications*, (In Press).
- Jones, S., Shanahan, H.P., Berman, H.M. and Thornton, J.M. (2003) Using electrostatic potentials to predict DNA-binding site on DNA-binding proteins. *Nucleic Acids Res.*, **31**, 7189–7198.
- Word, J.M., Lovell, S.C., Richardson, J.S. and Richardson, D.C. (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.*, **285**, 1735–1747.
- Brooks, B., Brucoleri, R.E., Olafson, B., States, D., Swaminathan, S. and Karplus, M. (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, **4**, 187–217.
- Gilson, M., Sharp, K. and Honig, B. (1988) Calculating electrostatic interactions in bio-molecules: method and error assessment. *J. Comput. Chem.*, **9**, 327–335.
- Jayaram, B., Sharp, K.A. and Honig, B. (1989) The electrostatic potential of B-DNA. *Biopolymers*, **28**, 975–993.
- Sharp, K.A. and Honig, B. (1990) Electrostatic interactions in macromolecules: theory and applications. *Annu. Rev. Biophys. Biophys. Chem.*, **19**, 301–332.
- Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
- Cristianini, N. and Shawe-Taylor, J. (1999) An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, Cambridge, UK.
- Chang, C.-C. and Lin, C.-J. (2005) LIBSVM: A Library for Support Vector Machines. Technical Report. Department of Computer Science, National Taiwan University.
- Chen, Y., Kortemme, T., Robertson, T., Baker, D. and Varani, G. (2004) 'A new hydrogen-bonding potential for the design of protein–RNA interactions predicts specific contacts and discriminates decoys'. *Nucleic Acids Res.*, **32**, 5145–5162.
- Maiti, R., Van Domselaar, G.H., Zhang, H. and Wishart, D.S. (2004) SuperPose: a simple server for sophisticated structural superposition. *Nucleic Acids Res.*, **32**, W590–594.
- Bhardwaj, N., Langlois, R.E., Zhao, G. and Lu, H. (2005) Structure Based Prediction of Binding Residues on DNA-binding Proteins. *27th Annual International Conference of the Engineering in Medicine and Biology Society*, (in press).
- Harrison, S.C. (1991) A structural taxonomy of DNA-binding domains. *Nature*, **353**, 715–719.
- Jones, S., Barker, J.A., Nobeli, I. and Thornton, J.M. (2003) Using structural motif templates to identify proteins with DNA binding function. *Nucleic Acids Res.*, **31**, 2811–2823.
- Humphrey, W., Dalke, A. and Schulten, K. (1996) VMD: visual molecular dynamics. *J. Mol. Graph.*, **14**, 33–38.