Proceedings of the 2005 IEEE
Engineering in Medicine and Biology 27th Annual Conference
Shanghai, China, September 1-4, 2005

Structure Based Prediction of Binding Residues on DNA-binding Proteins

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

Abstract— **Annotation of the functional sites on the surface of a protein has been the subject of many studies. In this regard, the search for attributes and features characterizing these sites is of prime consequence. Here, we present an implementation of a kernel-based machine learning protocol for identifying residues on a DNA-binding protein form the interface with the DNA. Sequence and structural features including solvent accessibility, local composition, net charge and electrostatic potentials are examined. These features are then fed into Support Vector Machines (SVM) to predict the DNA-binding residues on the surface of the protein.**

In order to compare with published work, we predict binding residues by training on other binding and non-binding residues in the same protein for which we achieved an accuracy of 79%. The sensitivity and specificity are 59% and 89%. We also consider a more realistic approach, predicting the binding residues of proteins entirely withheld from the training set achieving values of 66%, 43% and 81%, respectively. Performances reported here are better than other published results. Moreover, since our protocol does not lean on sequence or structural homology, it can be used to annotate unclassified proteins and more generally to identify novel binding sites with no similarity to the known cases.

Keywords—**protein-DNA interaction, function annotation, SVMs, binding site prediction.**

I. Introduction

Structure of a protein governs the biological function it is involved in [1]. Inherent to structure are many functional sites that help proteins to recognize their targets for transport, transcription, catalysis or signal transduction. Identifying these functional sites would help in assigning the protein to its set(s) of functions [2].

DNA-binding sites on the surface of a DNA-binding protein are one type of such functional sites. These sites show many structure and sequence based characteristics that contribute to the DNA binding e.g. electrostatic features, charge complementarity and amino acids preference [3-7]. An automated framework for annotation of such features would aid in identification of binding residues on the surface of the protein.

Manuscript received May 2, 2005.

All authors are affiliated with Bioinformatics Program, Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607, USA.

This work is partially supported by startup funds to H.L. from UIC. R.E.L. acknowledges the support from NIH training grant (PI. John Solaro).

* To whom correspondence should be addressed:
Hui Lu, Department of Bioengineering, University of Illinois at Chicago (UIC), Chicago, IL 60607, USA. Email: huilu@uic.edu.

Previously, there have been studies to identify DNA-binding proteins and the corresponding DNA-binding sites on the basis of their characteristics. Ref. 8 used electrostatic properties, accessibility and hydrophobicity of the residues to predict DNA-binding sites with 68% accuracy. Ref. 9 trained neural networks to identify binding residues using their neighbors and solvent accessibility. Although they reported an accuracy of 79%, their sensitivity was below 40%.

Here, we explore this issue in the framework of Support Vector Machines (SVM). We train SVM on an ensemble of possibly significant properties and subsequently predict binding residues using our "learned" model. SVM has been used with success in various areas of bioinformatics [10-12]. Earlier we successfully applied SVM to protein fold recognition & prediction of DNA-binding proteins [13,14]. Our goal here is to combine these methods to achieve a more poised success in prediction of DNA-binding residues.

II. Materials and Methods

A. Dataset

In this study, we used a dataset of 115 protein-DNA complexes of crystallographic resolution better than 3 Å. The dataset was obtained by combining datasets of previous associated studies [9,15] and included proteins from various structural families. A complete list of the proteins used here is available at http://proteomics.bioengr.uic.edu/pro-dna.

B. Definition of a surface binding residue

First, hydrogen atoms were added to all the structures using a publicly available software package REDUCE [16]. Every residue on each protein was classified as a 'surface' residue if its surface area exposed was more than 40% of its total area. Otherwise, it was classified as 'buried'. Further, we classified a surface residue as 'binding' if any of its atoms fell within a distance of 4.5 Å to any atom of DNA.

C. Feature Construction

Carefully selecting relevant features capable of discriminating the binding residues from the non-binding ones is the key to achieving a high accuracy in identification of binding sites on a protein's surface. Such evaluation of features entails characteristics congenial to DNA-binding. For example, given the negatively charged backbone of DNA, charge and electrostatic complementarity is expected to play an important role in DNA-binding. In the following

points, we identify an array of such meaningful features for the SVM to locate the binding sites.

1) Charge of each residue: Charge reciprocality of a residue intuitively seems to play a role in binding of a residue to the DNA. So, we used the net charge of a residue as one of the features for classification. We assigned a charge of +1 to *Arg* and *Lys* and -1 to *Asp* and *Glu*. *His* was specified a charge of +0.5 and all others were taken as neutral.

2) Average Potential on a residue: We used Delphi (v4) [17] for all electrostatic calculations in this study. This tool solves the non-linear Poisson-Boltzmann using finite-difference methods to calculate the potential at specified points. Electrostatic potentials at the site of all the atoms in a protein were reported in the absence of the DNA. The CHARMM [18] force-field parameters were employed for assignment of partial charges to the atoms. The Salt concentration and temperature were fixed at 145 mM/liter and 298 K, respectively. The dielectric constants were specified at 2.0 and 80.0 for the protein and the solvent. A fine-resolution grid structure with a scale of 2 (grids/Å) was employed. The center of the grid architecture was translated to overlap with the geometric center of the protein. The protein was made to fill half of the total volume of the grid cubic by specifying a percentage fill of 50. Default values of all other parameters were used. After calculating the potential at the site of every atom of a residue, it was assigned a potential equal to the average of the potentials on all its atoms.

3) Secondary structure. We appraised six secondary structures for every residue in the protein and computed the relative frequency of amino acids in these states to judge if there was an inclination for any particular structural state. DSSP [19] was used to assign every residue to one of the six structural classes.

4) Solvent Accessible Surface Area (ASA). In order to determine the correlation of ASA with a residue's propensity to bind, we calculated the relative ASA of every residue from DSSP.

5) Residue neighbors. Previous studies have shown that DNA-binding in unlikely to involve a single residue-base interaction [9]. Instead, it originates from multiple interactions between the two sides. For a specific locale on a protein to be in the vicinity of the DNA, its neighborhood residues should also be favorable. So, we complied a list of all the neighbors of every residue that were within a distance of 4.5Å from that residue.

D. SVM protocol

SVM solves a binary class problem by finding the maximum margin between two classes of data. . It uses a nonlinear transformation (a kernel function) to map the input data to a higher dimensional feature space where the classes become linearly separable. In other words, this is equivalent to solving the quadratic optimization problem:

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

subject to

$$y_i(\phi(x_i) \cdot w + b) \geq 1 - \xi_i,\ i = 1,...,m,$$

$$\xi_i \geq 0,\ i = 1,...,m,$$

where x_i is a feature vector labeled by $y_i \in \{+1, -1\}$, $(x_i, y_i), i = 1,...,m,$ and C is a parameter. The given model summarizes the so-called soft-margin SVM, which tolerates noise within the data. It does so by generating a separating plane using the equation $f(x) = \phi(x) \cdot w + b = 0$. Through the representation of $w = \sum_j \phi(x_j)$, we obtain $\phi(x_i) \cdot w = \sum_j \alpha_i \phi(x_j) \cdot \phi(x_i)$. This provides an efficient approach to solve SVM without the explicit use of the nonlinear transformation (21).

We use the LIBSVM implementation of SVM and found the polynomial kernel to give the best accuracy (20). The polynomial kernel consists of a family of polynomials represented as follows: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$ (21). Tuning *d, r,* and γ, we can select the best possible model using the weighted-average accuracy of n-fold cross-validation.

III. RESULTS

All the members (residues) of the dataset with their corresponding feature vectors and classes (DNA-binding or non-DNA-binding residue as defined above) associated with each member were given as input to SVM. For each residue, the length of a feature vector was 29 (1 for the net charge, average potential & ASA, 6 for the secondary structure assignment and 20 for residue neighbors). During 'training', SVM projects the data to a higher dimensional feature space and relying on rigorous optimization theory finds the maximal margin hyperplane that best separates the two classes. During 'testing' with a learned model, SVM attempts to correctly predict the class of every member of the test set using their corresponding feature vectors.

To assess how well SVM performs, we employed several validation techniques. In cross-validation, SVM is trained on one subset and then tested on another subset. We divided the dataset into two subsets. Dataset 1, consisting of 40 proteins, was tested with jackknife test (Leave-one-out) at two levels: residue and protein. At the residue level, for every protein, all but one residue was used for training and the class of the "left-out" residue was then predicted. This was repeated until each residue was tested. Accuracy is defined as the fraction of total correct predictions. Sensitivity is TP/(TP+FN), and selectivity is TN/(TN+FP). Unlike previous studies [9] we listed only the 'surface' residues for classification. Taking this reasonable step, we avoided a higher imbalance in the data, otherwise present due to the comparatively lower fraction of binding residues (to non-binding). We reduced the unevenness from 1:10 (1 binding residue for every 10

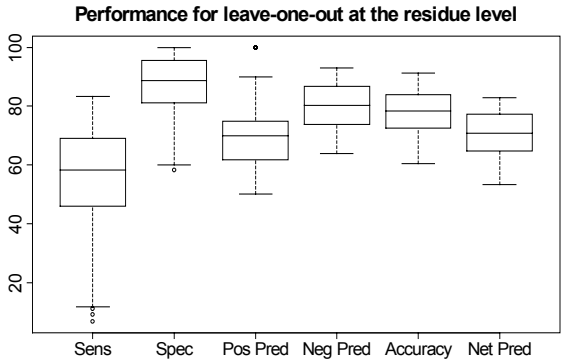

Fig. 1. Box-and-whisker plot of prediction performance for leave-one-out at the residue level. Shown on the horizontal axis is the sensitivity, specificity, positive predication, negative prediction, accuracy and the net prediction. Positive prediction is defined as TP/(TP+FP) and negative prediction as TN/(TN+FN).

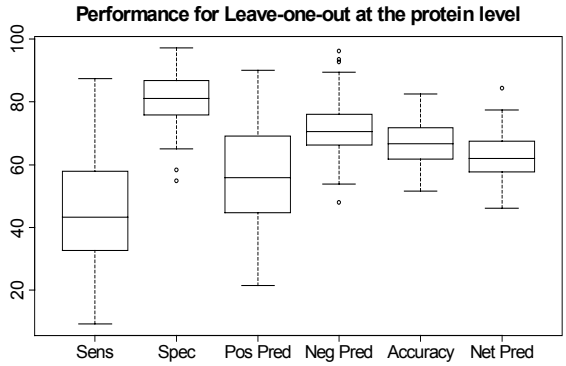

Fig. 2. Box-and-whisker plot of prediction performance for leave-one-out at the protein level.

residues) to 1:5. A high disparity in the ratio between positive and negative classes in the data artificially inflates the accuracy while still giving a poor sensitivity and distorting the picture.

Note, that our dataset is harder than datasets that use internal residues, in that, internal residues are easier to predict as non-binding given that such residues are not exposed to perform the binding function.

Fig. 1 box-plots various performance criteria for leave-one-out at the residue level collected over all the 40 proteins. Our prediction accuracy ranged from 61% to 91% with a mean of 79%; in other words, on an average, we could correctly predict 79% of the residues on a given protein. Although this value is similar to the one reported earlier [9], it is more representative because our data is more balanced in terms of positive and negative cases. Our sensitivity values, however, showed a high fluctuation ranging from 12% to 83%. The mean sensitivity value was around 59%, a 19% improvement over a similar study done earlier that used the same validation technique and reported a sensitivity of 40% [9]. We also report a 11% improvement in net prediction (defined as (sensitivity + specificity)/2), which can be a more meaningful measure of predictability in case of any unevenness in the data.

At the protein level, all residues from one protein were left out. Residues from all other proteins were used for training and the residues of the left-out protein were then used for testing. The percentage of residues correctly predicted is reported as the accuracy for each protein.

At the protein level, the prediction power of SVM was not as high as at the residue level (Fig. 2). The mean accuracy was around 66% whereas the net prediction was comparable to the corresponding value at the residue level. Sensitivity values fluctuated again from 10% to 87% with a mean of around 43%. This means that at the protein level SVM could not identify as many true positives, though it showed almost 81% success (specificity) in recognizing non-binding residues.

Next, we used the 'holdout method' on Dataset 2 consisting of the remaining 75 proteins. One of the classification models (randomly picked) created during the jackknife test at the protein level was used to predict the class of every residue on all the 75 proteins. This resulted in a set of 75 accuracies.

This is a more realistic approach as the model is built on a known case and is tested on the cases that it has not seen so far. The box-and-whisker plot of performance using this technique (fig. 3) shows that the success rates of this validation method is very similar to leave-one-out at the protein level. Specifically, this demonstrates that classification patterns that SVM can recognize for classifying the first 40 proteins also holds good for the remaining 75 proteins.

While the results demonstrate the plausibility of using SVM for identification of the binding residues, they illustrate the need for further improvement. Using the above features, SVM has identified sufficient discriminatory patterns to predict binding residues with a comparatively high accuracy.

IV. CONCLUSION

In the current work we have implemented SVM for the identification of binding residues in DNA-binding proteins with high accuracy. Mean accuracy and net prediction values for the leave-one-out at the residue level are 80% and 70%, and those at the protein level are 65% and 63% respectively. For the holdout evaluation the corresponding values are 75% and 65%. These values depend on the validation technique and are higher than earlier studies by others with various machine learning techniques. We also recorded a much higher sensitivity and positive prediction values than previous similar studies [9]. However, these values need to be pushed further, especially sensitivity and positive prediction. Features that can better characterize the binding residues will be very helpful in correctly identifying binding residues.

Selection of meaningful feature vectors coupled with a powerful kernel function leads to a higher success in prediction of binding residues. In addition to features employed in previous attempts, we used secondary structure,

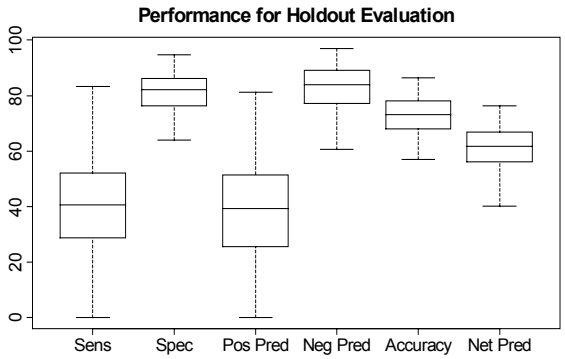

Fig. 3. Box-and-whisker plot of prediction performance for holdout evaluation.

the average potential and the charge on a residue as descriptors. When used in isolation, these features might not effectively differentiate the binding residues from the non-binding ones, but when mapped to a higher dimensional kernel space in the SVM, they provide more information allowing for a higher accuracy in prediction.

The correct prediction values reported above leave room for improvement in the prediction power of SVM. Residue conservation and their occurrence in large positive potential patches could be harnessed to boost the positive prediction of this protocol. For increasing the specificity and negative prediction, clustering of 'predicted' binding residues could be used to eliminate some of the false positives. For example consider the Pu.1 ETS domain (PDB id 1pue). It represents a typical case of binding-residue prediction of the SVM with 70% accuracy (Fig. 4). It can be observed that some of the false positives are isolated on the surface and so could be removed with a distance criterion. This post-prediction scheme could further increase the number of correct predictions. We contemplate that once the determinants of DNA-binding at a particular site are better understood in nature; the above method could be further polished by the addition of more features and fine-tuning of the SVM.

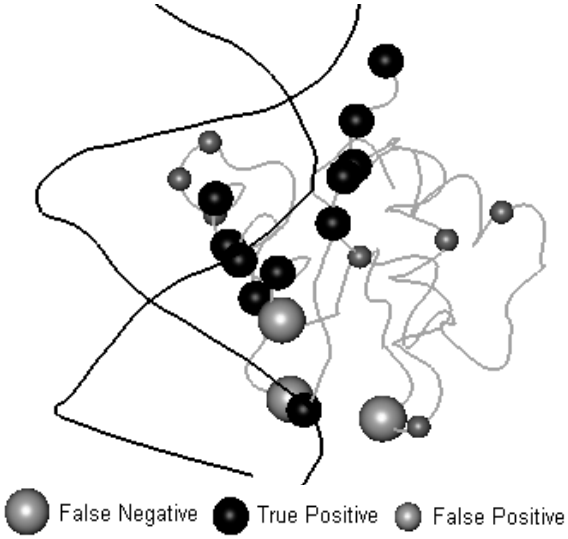

Fig. 4. Prediction of the binding residues on 1pue as an example. DNA and protein are shown in the black and gray tube representation.

REFERENCES

[1] H. Hegyi, and M. Gerstein, "The relationship between protein structure and function: a comprehensive survey with application to the yeast genome," *J. Mol. Biol.* 288, pp. 147–164, 1999.

[2] P. Aloy, E. Querol, F. X. Aviles and M. J. E. Sternberg, "Automated Structure-based Prediction of Functional Sites in Proteins: Applications to Assessing the Validity of Inheriting Protein Function from Homology in Genome Annotation and to Protein Docking", *J. Mol. Biol.* 311, pp. 395-408, 2001.

[3] S. Jones, P. Heyningen, H. M. Berman and J. M. Thornton, "Protein-DNA interactions: a structural analysis," *J. Mol. Bio.* 287, pp. 877-896, 1999.

[4] D. H. Ohlendorf and Matthew, J. B., "Electrostatics and flexibility in protein–DNA interactions," *Advan. Biophys.* 20, pp. 137–151, 1985.

[5] Y. Mandel-Gutfreund, H. Margalit, R. L. Jernigan and V. B. Zhurkin, "A role for CH…O interactions in protein–DNA recognition," *J. Mol. Biol.* 277, pp. 1129–1140, 1998.

[6] Y. Mandel-Gutfreund, O. Schueler, and H. Margalit, "Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles," *J. Mol. Biol.* 253, pp. 370–382, 1995.

[7] B. Honig, K. Sharp and M. Gilson, "Electrostatic interactions in proteins," *Prog. Clin. Biol. Res.*, 289, pp. 65–74, 1989.

[8] S. Jones, H. P. Shanahan, H. M. Berman and J. M. Thornton, "Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins." *Nucleic Acid Res*, 31, pp. 7189-7198, 2003.

[9] S. Ahmad, M. M. Gromiha and A. Sarai, "Analysis of DNA-binding proteins and their binding residues based on composition, sequence and structural information," *Bioinformatics*, 20, pp. 477-486, 2004.

[10] M. P. S. Brown *et al.*, "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proc. Natl. Acad. Sci.* USA, 97(1), pp. 262-267, 2000.

[11] C. H. Q. Ding and I. Dubchak, "Multi-class protein fold recognition using support vector machines and neural networks," *Bioinformatics*, 17(4), pp. 349-58, 2001.

[12] T. Jaakkola, M. Diekhans and D. Haussler "Using the fisher kernel method to detect remote protein homologies," in *Proc. of the Seventh International Conference on Intelligent Systems for Molecular Biology* (1999).

[13] R. E. Langlois, A. Diec, Y. Dai and H. Lu., "Kernel based approach for protein fold prediction from sequence" in *Proc. 26th Annual International Conference of the Engineering in Medicine and Biology Society*, pp. 2885-2288, 2004.

[14] N. Bhardwaj, R. E. Langlois, G. Zhao and H. Lu, "Kernel-based machine learning protocol for recognizing DNA-binding proteins" unpublished.

[15] E. W. Stawiski, L. M. Gregoret and Y. Mandel-Gutfreund, "Annotating nucleic acid-binding function based on protein structure" *Journ. of Mol. Bio.*, 326, pp. 1065-1079, 2003.

[16] J. M.Word, S.C. Lovell, J.S. Richardson and D.C. Richardson, "Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation," *J Mol Biol.* 285, pp.1735-47, 1999.

[17] W. Rocchia, E. Alexov and B Honig,. "Extending the applicability of the nonlinear Poisson-Boltzmann equation: Multiple dielectric constants and multivalent ions" *J Phys. Chem. B*, 105, pp. 6507-6514, 2001.

[18] B. Brooks, R.E. Bruccoleri, B. Olafson, D. States, S. Swaminathan, and M. Karplus, "CHARMM: a program for macromolecular energy, minimization, and dynamics calculations," *J. Comput. Chem.* 4, pp. 187-217, 1983.

[19] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, 22, pp. 2577-637, 1983.

[20] C.-C.Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," Technical report, Department of Computer Science, National Taiwan University, 2005.

[21] N. Cristianini and J. Shawe-Taylor, *An Introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press, 1999.