

On the Accuracy of Sequence-Based Computational Inference of Protein Residues Involved in Interactions with DNA

¹Zhenkun Gou and ^{1,2,3}Igor B. Kuznetsov

¹Gen*NY*sis Center for Excellence in Cancer Genomics,
Department of Epidemiology and Biostatistics,

²Department of Biological Sciences,

³Department of Computer Science, University at Albany,
One Discovery Drive Rensselaer, 12144 New York, USA

Abstract: Methods for computational inference of DNA-binding residues in DNA-binding proteins are usually developed using classification techniques trained to distinguish between binding and non-binding residues on the basis of known examples observed in experimentally determined high-resolution structures of protein-DNA complexes. What degree of accuracy can be expected when a computational methods is applied to a particular novel protein remains largely unknown. We test the utility of classification methods on the example of Kernel Logistic Regression (KLR) predictors of DNA-binding residues. We show that predictors that utilize sequence properties of proteins can successfully predict DNA-binding residues in proteins from a novel structural class. We use Multiple Linear Regression (MLR) to establish a quantitative relationship between protein properties and the expected accuracy of KLR predictors. Present results indicate that in the case of novel proteins the expected accuracy provided by an MLR model is close to the actual accuracy and can be used to assess the overall confidence of the prediction.

Key words: Multiple linear regression, DNA-binding protein, KLR

INTRODUCTION

Protein-DNA interactions form the basis of such fundamental biological activities as transcription and replication. The exact location of DNA-binding interface on a DNA-binding protein can be readily identified if a high-resolution structure of its complex with DNA was determined experimentally. However, the number of experimentally solved structures of protein-DNA complexes that can be used to locate the interaction interface is very limited. This limitation leads to a significant gap in the knowledge base: A protein may be known to interact with DNA, its unbound structure may be known, but the actual location of the interaction interface is unknown. In such cases, when a high-resolution structure of a protein-DNA complex is not available, the approximate location of the interaction interface can be determined by bioinformatics methods. Bioinformatics methods are a group of computational approaches that utilize the fact that the location of interaction sites on a particular protein can be identified even without the knowledge about its interaction partner. These methods are based on the observation that interface residues share certain common properties that distinguish them from the rest of surface residues. A number of bioinformatics methods for predicting the location of protein sites involved in interactions with DNA have been developed (Ahmad and Sarai, 2005; Bharwdaj and Lu, 2007; Jones *et al.*, 2003; Kuznetsov *et al.*, 2006; Saito *et al.*, 2006; Tjong and Zhou, 2007; Tsuchiya *et al.*, 2004; Wang and Brown, 2006; Yan *et al.*, 2006).

Corresponding Author: Igor B. Kuznetsov, Gen*NY*sis Center for Excellence in Cancer Genomics,
University at Albany, One Discovery Drive Rensselaer, 12144 New York, USA
Tel: (518) 591-7156 Fax: (518) 591-7151

Bioinformatics predictors are usually developed using machine learning methods trained to distinguish between DNA-binding residues and non-binding ones on the basis of known examples observed in a set of experimentally determined protein-DNA complexes. The main assumption of such approach is that a predictor of DNA-binding residues obtained in such a way will be able to generalize and successfully predict DNA-binding residues in proteins that contain novel DNA-binding motifs not present in the training set. In order to assess the ability to generalize, predictors are usually tested using n-fold cross-validation. In this procedure, a set of protein-DNA complexes is randomly partitioned into n groups. The predictor is trained on n-1 groups and tested on the remaining group. The process is repeated n times, so that each group is used for testing once. Traditionally, papers that describe a new predictor of DNA-binding sites only report accuracy averaged over n cross-validation runs. However, the average accuracy provides only a rough estimate of the ability to generalize averaged over all classes of proteins used for testing. It cannot be used to estimate what degree of accuracy can be expected in the case of a particular novel protein. Recently, we have shown that the performance of machine learning methods for predicting DNA-binding residues can vary dramatically on the level of individual proteins, from a nearly random assignment to a perfect prediction (Kuznetsov *et al.*, 2006). This implies that when a prediction is done for a novel protein, the actual accuracy can be far off the expected average value determined by cross-validation. In this study, we ask the following two questions:

- How successfully a machine learning method for predicting DNA-binding residues in DNA-binding proteins can perform on proteins from a structural class that was not used to train the method?
- Is it possible to provide a reasonable blind estimate of the expected prediction accuracy for a novel DNA-binding protein based on its properties?

To answer the first question, we apply a strict test of the ability to generalize by using leave-one-class-out cross-validation and show that machine learning methods that utilize evolutionary information can predict DNA-binding residues in proteins from a novel structural class with a reasonable accuracy. To answer the second question, we develop a multiple regression model that uses sequence properties of proteins to estimate the expected accuracy of the prediction of DNA-binding residues in the case of novel proteins. To the best of our knowledge, this is the first study of its kind that addresses these issues.

MATERIALS AND METHODS

We used machine-learning predictors of DNA-binding residues previously developed by our group. Details of the methodology are provided in our publications (Kuznetsov *et al.*, 2006; Hwang *et al.*, 2007). These predictors utilize the profile of evolutionary conservation of residue positions derived from a PSI-BLAST Position Specific Scoring Matrix (PSSM) (Altschul *et al.*, 1997). For a sequence of length N residues, PSSM is represented by an $N \times 20$ matrix. An element (i,j) of this matrix provides information on the evolutionary conservation of residue type j at sequence position i. We use the standard sequence-based descriptor that utilizes a sliding window of size $w*2+1$ to encode the profile of evolutionary conservation of each position and its sequential neighbors (Ahmad and Sarai, 2005; Kuznetsov *et al.*, 2006). In this approach, a feature vector that encodes sequence position k and w neighboring residues on both sides of this position is constructed by concatenating PSSM rows from $k-w$ to $k+w$. We found that the best classification accuracy is obtained by using the window of size seven (Kuznetsov *et al.*, 2006). In this study, we chose to use KLR predictor of DNA-binding residues since its average performance is slightly better than that of

Support Vector Machine (SVM) (Hwang *et al.*, 2007). We used the non-redundant dataset of 62 protein-DNA complexes utilized previously by our group and others to develop machine-learning predictors of DNA-binding residues and described in (Ahmad and Sarai, 2005; Kuznetsov *et al.*, 2006). KLR was used with the Radial Basis Function (RBF) kernel which gives the best performance on our dataset.

We used both leave-one-protein-out and leave-one-class-out cross-validation to train and test our predictors of DNA-binding sites. In the former, one protein-DNA complex is removed from the dataset and the method is trained on the remaining complexes and tested on the removed complex. In the latter, all protein-DNA complexes that belong to a particular structural class are removed from the dataset and the method is trained on the remaining complexes and tested on the complexes from the removed class. The difference between these two approaches is that in the latter all proteins with similar DNA-binding motifs are not used for training. Thus, it provides a very strict test of the ability to generalize on proteins from novel structural classes. In order to assess different aspects of the quality of the prediction, we use the following performance measures: Accuracy (ACC), Sensitivity (SN) and Specificity (SP) defined as follows (Baldi *et al.*, 2000):

$$ACC = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad (1)$$

$$SN = \frac{TP}{(TP + FP)} \quad (2)$$

$$SP = \frac{TN}{(TN + FP)} \quad (3)$$

Where:

TP = No. of true positives (No. of correctly predicted DNA-binding residues)

TN = No. of true negatives (No. of correctly predicted non-binding residues)

FP = No. of false positives (No. of non-binding residues predicted as DNA-binding)

FN = No. of false negatives (No. of DNA-binding residue predicted as non-binding)

Both leave-one-protein-out and leave-one-class-out cross-validation provide ACC, SN and SP for each protein in the dataset (referred to as per-protein ACC, SN and SP).

RESULTS AND DISCUSSION

Prediction of DNA-Binding Residues in Proteins from Novel Structural Classes

Previously, we established that the KLR predictor of DNA-binding sites trained on the same dataset as the one utilized in this study and leave-one-protein-out cross-validation achieves ACC of 77.2%, SN of 76.4% and SP of 76.6% (Table 1), (Hwang *et al.*, 2007). In order to assess how the predictor would perform on novel DNA-binding proteins that do not share any structural similarities with known DNA-binding motifs, we split the dataset into four major structural classes annotated in the CATH database of structural domains (Orengo *et al.*, 1997). We only used proteins that contain domains from a single class. Proteins that have domains from two or more different structural classes were excluded.

Table 1: Performance of KLR predictor in leave-one-protein-out and leave-one-class-out cross-validation averaged over the entire data set

	ACC	SN	SP
Leave-one-protein-out	77.2±9.33	76.4±18.48	76.6±11.18
Leave-one-class-out	71.8±11.40	70.9±19.86	71.9±15.38

Values in each cell give the average and standard deviation of the performance measure computed per protein, ACC: Accuracy; SN: Sensitivity; SP: Specificity

Table 2: Class-specific performance of KLR predictor in leave-one-class-out cross-validation

Structures	ACC	SN	SP
Mainly- α	73.3±7.67	67.0±20.10	76.0±10.70
Mainly- β	72.4±5.49	64.5±9.97	73.4±5.58
α/β	63.2±9.35	67.5±17.16	59.7±14.39
Few regular structure (frs)	78.0±4.01	82.1±11.00	77.6±3.17

The row for a particular structural class shows the results when this class was used for testing in leave-one-class cross-validation. Values in each cell give the average and standard deviation of the performance measure computed per protein, ACC: Accuracy; SN: Sensitivity; SP: Specificity

- Mainly- α class: 29 proteins
- Mainly- β class: 4 proteins
- Mixed α/β class: 19 proteins
- Few regular structure (frs) class: 4 proteins

We used these four classes to perform leave-one-class-out training and testing of the KLR predictor as described in Methods. The results of both cross-validations (Table 1, 2) can be summarized as follows:

- The average performance measures computed over all four classes (this corresponds to averaging over the four cross-validation runs) are slightly lower than those obtained from leave-one-protein-out cross-validation. However, the decrease in accuracy is only 5.4% and a comparison of the results of leave-one-protein-out cross-validation to those of leave-one-class-out using the *t*-test shows that this difference is not statistically significant ($p > 0.05$). One may expect that the accuracy of knowledge-based predictors will improve further when more experimentally determined non-redundant structures of protein-DNA complexes become available for training.
- Analysis of class-specific performance in leave-one-class-out cross-validation indicates that the lowest performance is observed for proteins from α/β class. This low performance cannot be explained only by the small size of the training set (37 proteins), since in the case of mainly- α proteins the size of the training sample is considerably smaller (27 proteins), while the performance is considerably better. Another interesting observation is that prediction accuracy is similar for both mainly- α and mainly- β classes, despite very different sample sizes and the fact that proteins from the former are dominated by local helical interactions and proteins from the latter are dominated by long-range interactions in β -sheets.

Overall, the accuracy achieved on three out of the four structural classes (mainly- α , mainly- β and frs) is well above 70% and reasonable sensitivity and specificity. This indicates that machine learning methods for predicting DNA-binding sites are capable of making reasonably successful predictions for novel proteins from at least three out of the four structural classes studied.

Using Protein Properties to Estimate the Expected Accuracy of the Prediction of DNA-Binding Residues in Novel Proteins

Here, we study how sequence properties of proteins can be utilized to estimate the expected accuracy of machine learning predictors of DNA-binding residues in the case of novel proteins. We used the following protein sequence properties that show a statistically significant correlation with

prediction accuracy: The frequencies of the 20 amino acid types, sequence length and the number of homologous sequences used to construct PSSM for a given protein. Since the location of DNA-binding interface is known for each protein in our dataset, we can use a MLR model to establish a quantitative relationship between protein properties and prediction accuracy for a given KLR predictor. In this model, sequence properties are used as predictor variables to obtain an estimate of expected accuracy for a given protein. If the MLR model is good, the expected accuracy it provides will be close to the observed accuracy. This will mean that in the case of novel proteins for which the structure of protein-DNA complex is unknown, the expected accuracy provided by MLR can be used as a measure of the overall confidence of the predicted DNA-binding interface. We compare MLR results obtained from leave-one-protein-out cross-validation to those from leave-one-class-out cross-validation. This comparison allows us to study how MLR models perform on proteins from novel structural classes (Table 3).

In general, the quality of any multiple linear regression model can be assessed by using the p-value and the squared correlation coefficient, R^2 , which gives the total amount of variability in the predicted variable explained by the MLR model. The higher R^2 , the better the fit provided by the regression. A perfect regression will have $R^2 = 1.0$. A p-value below 0.05 indicates a good MLR model with regression coefficients different from zero. The MLR model that provides an estimate of the expected prediction accuracy for KLR predictor trained using leave-one-protein-out cross-validation explains over 80% of the variability (Fig. 1). The results of MLR model constructed using accuracy obtained from leave-one-class-out cross-validation are shown in Fig. 2. In this case, MLR model explains almost 92% of the variability in prediction accuracy.

Overall, these results suggest that in the case of a novel DNA-binding protein, a simple MLR model based on its properties can be used to provide a reasonably precise estimate of the expected accuracy of the prediction of its DNA-binding interface.

Table 3: The coefficients of multiple linear regression models

Regression parameters	Leave-one-protein-out	Leave-one-class-out
Constant	54.1386	0.0844
b_L	0.0003	0.0003
b_N	0.000019	-0.000028
b_{TRP}	-0.5333	0.0095
b_{ILE}	-0.5281	-0.0063
b_{TYR}	-0.5413	0.0106
b_{PHE}	-0.5367	0.0099
b_{LEU}	-0.5298	0.0014
b_{VAL}	-0.5358	0.0140
b_{MET}	-0.5267	-0.0103
b_{CYS}	-0.5441	0.0008
b_{ALA}	-0.5300	0.0040
b_{GLY}	-0.5382	0.0097
b_{HIS}	-0.5322	-0.0090
b_{PRO}	-0.5454	0.0086
b_{SER}	-0.5326	0.0063
b_{THR}	-0.5299	0.0132
b_{ASN}	-0.5429	0.0080
b_{GLN}	-0.5301	0.0087
b_{ASP}	-0.5348	0.0077
b_{GLU}	-0.5311	0.0056
b_{LYS}	-0.5358	0.0107
b_{ARG}	-0.5359	

The regression model is $Y = \text{Constant} + b_L * L + b_N * N + b_{TRP} * \text{Trp} + \dots + b_{ARG} * \text{Arg}$, where, constant and b_L , b_N , b_{TRP}, \dots, b_{ARG} are regression parameters; L: The sequence length; N: The number of sequences used to construct PSSM; Trp, ..., Arg are the normalized frequencies of the 20 amino acid types in the sequence

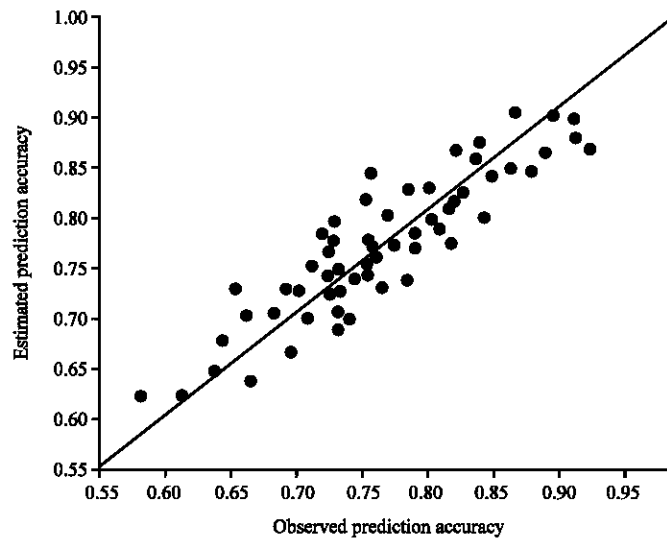


Fig. 1: Scatter plot of the observed per-protein accuracy of KLR predictor obtained from leave-one-protein-out cross-validation vs. accuracy estimated using MLR. $R^2 = 0.805$, F-statistic = 6.77 with p-value = 0.0

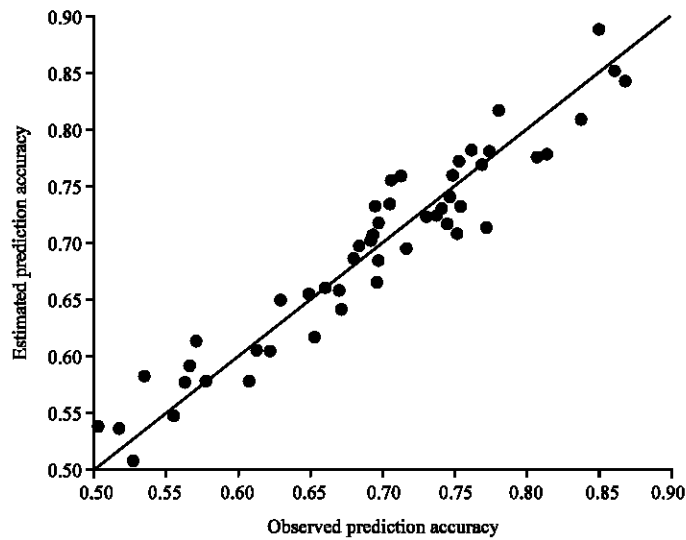


Fig. 2: Scatter plot of the observed per-protein accuracy of KLR predictor obtained from leave-one-class-out cross-validation vs. accuracy estimated using MLR. $R^2 = 0.918$, F-statistic = 14.25 with p-value = 0.0

ACKNOWLEDGMENT

This research was supported in part by grant R03LM009034 from the National Library of Medicine of the National Institutes of Health.

REFERENCES

- Ahmad, S. and A. Sarai, 2005. PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinform.*, 6: 33-38.
- Altschul, S.F., T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman, 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.*, 25: 3389-3402.
- Baldi, P., S. Brunak, Y. Chauvin, C.A. Andersen and H. Nielsen, 2000. Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics*, 16: 412-424.
- Bharwdaj, B. and H. Lu, 2007. Residue-level prediction of DNA-binding sites and its application on DNA-binding protein predictions. *FEBS Lett.*, 581: 1058-1066.
- Hwang, S., Z. Gou and I.B. Kuznetsov, 2007. DP-Bind: A web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics*, 23 (5): 634-636.
- Jones, S., H.P. Shanahan, H.M. Berman and J.M. Thornton, 2003. Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Res.*, 31: 7189-7198.
- Kuznetsov, I.B., Z. Gou, R. Li and S. Hwang, 2006. Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. *Proteins*, 64: 19-27.
- Orengo, C.A., A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells and J.M. Thornton, 1997. CATH-a hierarchic classification of protein domain structures. *Structure*, 5: 1093-1108.
- Saito, M., M. Go and T. Shirai, 2006. An empirical approach for detecting nucleotide-binding sites on proteins. *Protein Eng. Design Selection*, 19: 67-75.
- Tjong, H. and H.X. Zhou, 2007. DISPLAR: An accurate method for predicting DNA-binding sites on protein surfaces. *Nucleic Acids Res.*, 35 (5): 1465-1477.
- Tsuchiya, Y., K. Kinoshita and H. Nakamura, 2004. Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. *Proteins*, 55: 885-894.
- Wang, L. and S.J. Brown, 2006. BindN: A web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res.*, 34: W243-248.
- Yan, C., M. Terribilini, F. Wu, R.L. Jernigan, D. Dobbs and V. Honavar, 2006. Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinform.*, 7: 262.