BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 5 2007, pages 634–636
doi:10.1093/bioinformatics/btl672

Sequence analysis

DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins

Seungwoo Hwang†, Zhenkun Gou† and Igor B. Kuznetsov*

Gen*NY*Sis Center for Excellence in Cancer Genomics, Department of Epidemiology and Biostatistics, One Discovery Drive, University at Albany, Rensselaer, NY 12144, USA

Received on August 29, 2006; revised on November 16, 2006; accepted on January 3, 2007

Advance Access publication January 19, 2007

Associate Editor: Thomas Lengauer

ABSTRACT

Summary: This article describes DP-Bind, a web server for predicting DNA-binding sites in a DNA-binding protein from its amino acid sequence. The web server implements three machine learning methods: support vector machine, kernel logistic regression and penalized logistic regression. Prediction can be performed using either the input sequence alone or an automatically generated profile of evolutionary conservation of the input sequence in the form of PSI-BLAST position-specific scoring matrix (PSSM). PSSM-based kernel logistic regression achieves the accuracy of 77.2%, sensitivity of 76.4% and specificity of 76.6%. The outputs of all three individual methods are combined into a consensus prediction to help identify positions predicted with high level of confidence.

Availability: Freely available at http://lcg.rit.albany.edu/dp-bind

Contact: lKuznetsov@albany.edu

Supplementry information: http://lcg.rit.albany.edu/dp-bind/dpbind_supplement.html

1 INTRODUCTION

A reliable computational prediction of DNA-binding sites in DNA-binding proteins is important for studying protein–DNA interactions. There are two types of such prediction methods: structure-based and sequence-based methods. Sequence-based methods have an advantage of not requiring the expensive and time-consuming process of experimental determination of protein structure. Thus, it is important to develop and improve prediction methods based on sequence input alone. Previously, we developed support vector machine predictors of DNA-binding sites (Kuznetsov *et al*., 2006). Here, we extend our previous work to develop a web-server for improved sequence-based prediction of DNA-binding sites by applying three supervised pattern recognition methods: support vector machine (SVM) (Vapnik, 1998), kernel logistic regression (KLR) (Zhu and Hastie, 2005) and penalized logistic regression (PLR) (le Cessie and van Houwelingen, 1992). The outputs from these three individual methods are combined to obtain a consensus prediction to further improve the performance and help identify positions predicted with high confidence. The on-line implementation of the predictors, called DP-Bind, is freely available at http://lcg.rit.albany.edu/dp-bind.

2 TRAINING AND TESTING DP-BIND

A detailed description of the materials and methods used for developing the web server is provided in our previous article (Kuznetsov *et al*., 2006). We therefore only briefly describe the methodology here. We used a non-redundant set of 62 experimentally solved protein–DNA complexes utilized in previous studies (Ahmad *et al*., 2004; Ahmad and Sarai, 2005; Kuznetsov *et al*., 2006; Wang and Brown, 2006). We implemented two types of methods to encode the input protein sequence. One is the single sequence-based encoding that utilizes the input sequence alone. The other is the PSSM-based encoding which accounts for evolutionary conservation of the input sequence and is based on PSI-BLAST (Altschul *et al*., 1997) position-specific scoring matrix (PSSM). Two variants of the single sequence-based encoding are implemented in DP-Bind: binary and BLOSUM62 encoding (Kuznetsov *et al*., 2006). Description of the three machine learning methods implemented in DP-Bind can be found in Supplementary Materials. Each method independently assigns a predicted label (binding or non-binding) to each residue in the input sequence. Then, these three labels can be used to produce a consensus prediction for each residue position. The consensus-based approach has previously been shown to improve performance in a number of applications, such as prediction of protein secondary structure (Cuff *et al*., 1998). We used two types of consensus. One is majority consensus obtained by majority voting. For instance, if two methods predict a given position as 'DNA-binding' and the third predicts it as 'non-binding', the majority consensus label is 'DNA-binding'. The other is strict consensus obtained by unanimous agreement. For instance, if one method disagrees with the other two, no consensus label is assigned to a given sequence position. Thus, the strict consensus retains only high confidence predictions on which all three methods agree.

We tested the predictors using leave-one-protein-out cross-validation, and computed the following performance measures (Table 1): accuracy (ACC), sensitivity (SN) and specificity (SP).

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Comparison of performance measures shown in Table 1 leads to the following observations.

(1) All three individual single sequence-based predictors have similar performance.
(2) All three individual PSSM-based predictors have a significantly better performance than corresponding sequence-based predictors. The PSSM-based KLR predictor has the highest prediction accuracy of 77.2%.
(3) The majority consensus only slightly improves the accuracy of single sequence-based prediction. In the case of PSSM-based prediction, its performance is similar to that of individual predictors.

It should be noted that in the case of strict consensus, some sequence positions are not assigned a label when one of the prediction methods disagrees with the other two. As a result, the total number of residues used to compute ACC, SN and SP of the strict consensus is usually smaller than that used to assess performance of the individual methods and the majority consensus. Therefore, the performance measures for the strict consensus shown in Table 1 should not be directly compared to those for individual methods and the majority consensus. The idea of strict consensus is to identify a subset of positions with high confidence labels. This idea is based on the assumption that if all three individual methods assign the same label to a given residue position, it is more likely to be predicted correctly. As shown in Table 1, such unanimously voted positions are predicted with considerably higher ACC, SN and SP (note that the performance of each individual method on such positions is the same as that of the strict consensus shown in Table 1).

Table 1. Performance of the predictors (in percentage)

Predictors	ACC	SN	SP
(a) Sequence-based BLOSUM62 predictor			
SVM	68.2 ± 6.6	70.4 ± 16.5	66.8 ± 9.2
KLR	68.6 ± 5.5	66.8 ± 15.4	68.9 ± 7.8
PLR	67.8 ± 6.9	69.0 ± 13.3	67.0 ± 9.0
Majority consensus	69.1 ± 6.2	69.9 ± 16.1	68.2 ± 8.6
Strict consensus[a]	72.2 ± 7.2	73.1 ± 16.3	71.4 ± 9.8
(b) PSSM-based predictor			
SVM	76.0 ± 9.1	76.9 ± 18.6	74.8 ± 12.5
KLR	77.2 ± 9.3	76.4 ± 18.5	76.6 ± 11.2
PLR	73.0 ± 8.8	73.3 ± 18.4	71.8 ± 12.8
Majority consensus	76.4 ± 9.0	76.9 ± 18.6	75.3 ± 12.0
Strict consensus[a]	80.0 ± 9.4	79.1 ± 19.4	78.6 ± 12.7

Performance measures are averaged over 62 cross-validation experiments. ACC = (TP + TN)/(TP + FP + TN + FN), SN = TP/(TP/FN), SP = TN/(TN + FP). T: true, F: false, P: positives and N: negatives.
[a]The total number of residue positions used to calculate ACC, SN and SP of the strict consensus is less than that used to assess the other methods. Therefore, the strict consensus should not be directly compared to the other methods. See text for details.

The final web-server implementation of each predictor was trained on all 62 protein complexes using optimal parameters determined during the cross-validation experiments.

One should be aware that the results reported in Table 1 show average per protein prediction accuracy. Our previous study has shown that accuracy of the prediction of DNA-binding residues may considerably vary among different proteins (see Figs 2–4 in Kuznetsov *et al.*, 2006, and Supplementary Figs 2–3 for this article). For some proteins the actual accuracy will be higher than the average reported values, for some it will be lower. In the case of each particular protein, accuracy depends on a variety of its structural and sequence properties, as well as the number of homologous sequences used to construct PSSM. For instance, DNA-binding residues in proteins from 'mainly-alpha' and 'few regular structure' classes (annotation from the CATH structural database, Orengo *et al.*, 1997) tend to be predicted with higher accuracy (see Tables II–IV in Kuznetsov *et al.*, 2006, and Supplementary Table 2 for this article). Since results of the prediction appear to be determined by many distinct features, there is no simple way to estimate what level of accuracy will be achieved when predicting DNA-binding residues in a particular protein. A possible practical approach is to use accuracies reported for the proteins from the training set (Supplementary Table 3) that share the same structural family with the test protein to obtain a rough idea what accuracy may be expected for the test protein. However, users should be aware that such empirical approach does not provide rigorous estimates of the expected accuracy and the actual value may still be far off its expectation.

3 DESCRIPTION OF THE WEB SERVER

The web server has a simple user interface that consists of three input fields: sequences to be analyzed, selection of encoding method and e-mail address. Description of each field as well as the output format can be found by clicking the corresponding help hyperlink. The user can input amino acid sequences in FASTA format by either pasting them or uploading as a file. Alternatively, the user may input protein identifiers (GenBank or UniProt) instead of FASTA sequences. Each input sequence can be as long as 1000 residues. The web server accepts as many as 100 sequences for the single sequence-based encoding. Only one sequence is accepted for the PSSM-based encoding due to a heavy computational load required to run PSI-BLAST. The default encoding method is the PSSM-based encoding that yields the most accurate prediction (see Table 1). The user can choose either to receive results by E-mail (default) or manually retrieve them using a temporary URL provided upon submission.

The output from DP-Bind consists of three parts: a header describing its format, input sequence and results of the prediction in tabular format (Supplementary Fig. 1). The results part consists of ten columns. The first column displays residue index (sequence position). The second column displays amino acid residue. Columns 3–8 show the outputs from SVM, KLR and PLR predictors. The output from each method consists of a predicted binding label and the probability of that label, where labels 1 and 0 stand for DNA-binding and non-binding residues, respectively. Columns 9 and 10 show the majority and strict consensus, respectively. If the strict consensus cannot be

obtained (one method disagrees with the other two) the position is marked with 'NA'. We advise users to use the majority consensus for single sequence-based prediction and the strict consensus for identifying sites predicted with high confidence.

4 CONCLUSION

To the best of our knowledge, three web servers are available at the moment for sequence-based prediction of DNA-binding sites: DBS-PRED (Ahmad *et al*., 2004), DBS-PSSM (Ahmad and Sarai, 2005) and BindN (Wang and Brown, 2006). Another study without web server implementation was also reported (Yan *et al*., 2006). Our performance measures are higher than those reported by these four studies (see Supplementary Table 1 for details). Although a direct comparison of all methods requires the use of identical assessment procedure, our improved performance measures suggest that DP-Bind is a competitive tool for sequence-based prediction of DNA-binding residues in DNA-binding proteins.

ACKNOWLEDGEMENTS

We thank Run Li for initial data preparation, and Eric Warnke, Frank Doyle and Chittibabu Guda for their help in web server development. This work was supported by grant 1R03LM009034-01 from the National Library of Medicine of the National Institutes of Health.

Conflict of Interest: none declared.

REFERENCES

Ahmad,S. *et al*. (2005) PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics*, **6**, 33–38.

Ahmad,S. *et al*. (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, **20**, 477–486.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Cuff,J.A. *et al*. (1998) JPred: a consensus secondary structure prediction server. *Bioinformatics*, **14**, 892–893.

Kuznetsov,I.B. *et al*. (2006) Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. *Proteins*, **64**, 19–27.

le Cessie,S. and van Houwelingen,J.C. (1992) Ridge estimators in logistic regression. *Appl. Statist*., **41**, 191–201.

Orengo,C.A. *et al*. (1997) CATH – a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Vapnik,V.N. (1998) *Statistical Learning Theory*. John Wiley and Sons, New York.

Wang,L. and Brown,S.J. (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res*., **34**, W243–W248.

Yan,C. *et al*. (2006) Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinformatics*, **7**, 262.

Zhu,J. and Hastie,T. (2005) Kernel logistic regression and the import vector machine. *J. Comp. Graph. Stat*., **14**, 185–205.