SEQUENCE-BASED PREDICTION OF DNA-BINDING SITES ON DNA-BINDING PROTEINS

**Gou Z., Hwang S., Kuznetsov B.I.**
Gen*NY*sis Center for Excellence in Cancer Genomics, University at Albany, One Discovery Drive, Rensselaer, NY, USA
*Corresponding author: e-mail: ikuznetsov@albany.edu

**Key words:** protein-DNA interaction, position specific scoring matrix, evolutionary conservation, web-server, DNA binding, prediction, pattern recognition, machine learning

**SUMMARY**

**Motivation:** Identification of DNA-binding sites on DNA-binding proteins is important for functional annotation. Experimental determination of the structure of a protein-DNA complex is an expensive process. Reliable computational methods that utilize the sequence of a DNA-binding protein to predict its DNA-binding interface are needed.

**Results:** We present an application of three machine learning methods: support vector machine, kernel logistic regression, and penalized logistic regression to predict DNA-binding sites on a DNA-binding protein using its amino acid sequence as an input. Prediction is performed using either single sequence or a profile of evolutionary conservation. The performance of our predictors is better than that of other existing sequence-based methods. The outputs of all three individual methods are combined to obtain a consensus prediction. This further improves performance and results in accuracy of 82.4 %, sensitivity of 84.9 % and specificity of 83.1 % for the strict consensus prediction.

**Availability:** [http://lcg.rit.albany.edu/dp-bind](http://lcg.rit.albany.edu/dp-bind).

**INTRODUCTION**

A reliable identification of DNA-binding sites on DNA-binding proteins is important for *in silico* modeling of protein-DNA interactions and functional annotation. Identification of DNA-binding sites is relatively straightforward if the structure of a protein-DNA complex is known. However, solving the structure of a protein-DNA complex is a very complicated and time-consuming process. Several computational methods that use experimentally solved unbound structure of a DNA-binding protein to identify DNA-binding interface based on the electrostatic potential and the shape of molecular surface have been developed (Jones *et al*., 2003; Tsuchiya *et al*., 2004). However, these methods cannot be used if experimentally determined protein structure is not available. An alternative to the structure-based prediction is a sequence-based prediction. In this work, we apply a combination of three supervised pattern recognition methods to improve the prediction of DNA-binding sites in a DNA-binding protein using its amino acid sequence as the only input.
METHODS AND ALGORITHMS

Dataset of protein-DNA complexes. We used a non-redundant set of 62 experimentally solved protein-DNA complexes that were utilized previously to develop DBS-PRED (Ahmad et al., 2004) and DBS-PSSM (Ahmad, Sarai, 2005). We label an amino acid residue in a protein chain as DNA-binding if the distance from at least one of its heavy atoms (atoms other than hydrogen) to any heavy atom in DNA is shorter than the cutoff distance of 4.5 Å. In order to balance the number of examples between binding and non-binding residues, for each protein chain we randomly sampled without replacement the same number of non-binding residues as that of the DNA-binding ones.

SEQUENCE ENCODING

In order to represent the input protein sequence by a numerical feature vector, we used two types of sequence-based encoding and encoding based on PSI-BLAST (Altschul et al., 1997) position specific scoring matrix (PSSM). In the first type of sequence encoding, called binary encoding, the 20 amino acid types are represented by 20 mutually orthogonal binary vectors of dimension 20 (Qian, Sejnowski, 1988). In the second type of sequence encoding, called BLOSUM62 encoding, each amino acid type is represented by a vector of dimension 20 using a corresponding row from the BLOSUM62 amino acid substitution matrix (Henikoff, Henikoff, 1992). In the case of PSSM-based encoding, each sequence position is encoded by a 20-dimensional vector obtained from a corresponding row in the PSSM (Ahmad, Sarai, 2005). In both the BLOSUM62 and PSSM encoding, we normalize all elements in the matrix between 0 and 1 using the logistic function \( f(x) = \frac{1}{1 + \exp(-x)} \). In all three encoding methods, nearest sequential neighbors of a sequence position are encoded with a standard procedure (Qian, Sejnowski, 1988) using a sliding window of size 7.

MACHINE LEARNING ALGORITHMS

For our two-class (DNA-binding and non-binding residues) classification problem, we applied three machine learning algorithms: support vector machine (SVM) (Christianini, Shawe-Taylor, 2000), kernel logistic regression (KLR) (Zhu, Hastie, 2005), and penalized logistic regression (PLR) (le Cessie, van Houwelingen, 1992). SVM is a margin maximizing classifier that does a linear classification in the feature space, which corresponds to a non-linear classification in the original data space. The feature space is obtained by transforming data from the original data space with a kernel function. Similarly, KLR and PLR are also margin maximizing classifiers. For both SVM and KLR we used the Radial Basis Function (RBF) kernel. The SVM algorithm was implemented using the LIBSVM program (http://www.csie.ntu.edu.tw/~cjlin/libsvm). We implemented the KLR and PLR algorithms in C++.

CONSENSUS PREDICTION

Each of the three machine learning methods independently assigns a label (binding or non-binding) to each position in the input sequence. Then, these three labels can be used to produce a consensus prediction for each sequence position. We used two types of consensus. The first is majority consensus obtained by majority voting (at least two of
three labels are identical). The other is strict consensus which retains only positions with high-confidence predictions on which all three methods agree.

EVALUATION OF THE PREDICTORS

We used leave-one-out cross-validation to train and test each predictor. We used accuracy (ACC), sensitivity (SN), and specificity (SP) to assess the performance of each predictor:

\[
ACC = \frac{TP + TN}{TP + FP + TN + FN}, \quad SN = \frac{TP}{TP + FN}, \quad SP = \frac{TN}{FP + TN}
\]

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

RESULTS AND DISCUSSION

ACC, SN, and SP of the predictors are shown in Table 1. Fig. 1 shows the receiver operating characteristics (ROC) curve for each predictor. ROC curve is more informative than most other measures and allows one to compare the performance of different classifiers by looking at the curve and the area under the curve (AUC). Larger AUC indicates better performance. Analysis of the data presented in Table 1 and Fig. 1 leads to the following observations:

1. All three individual sequence-based predictors have similar performance.
2. All three individual PSSM-based predictors have a significantly better performance than the sequence-based ones, PSSM-based KLR having the highest classification accuracy of 79.2 %.
3. The performance of PSSM-based KLR predictor (ACC of 79.2 %, SN of 76.4 %, SP of 82.0 %) is better than that of the other existing PSSM-based method for predicting DNA-binding sites, DBS-PSSM (ACC of 66.4 %, SN of 68.2 %, SP of 66.0 %).
4. The strict consensus prediction improves both sequence-based and PSSM-based predictions. The majority consensus performs better than individual methods in the case of single sequence-based prediction when evolutionary information is not utilized. It also improves sensitivity of the PSSM-based prediction.

A web server implementation of the predictors, called DP-BIND, is freely available at http://lcg.rit.albany.edu/dp-bind. It can be used for a high-confidence prediction of DNA-binding sites in a DNA-binding protein when its experimentally solved structure is not available.

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Sequence-based BLOSUM62 encoding</th>
<th>PSSM-based encoding</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Sensitivity</td>
</tr>
<tr>
<td>SVM</td>
<td>69.7±9.3</td>
<td>70.2±16.8</td>
</tr>
<tr>
<td>KLR</td>
<td>68.9±7.9</td>
<td>66.7±15.4</td>
</tr>
<tr>
<td>PLR</td>
<td>68.6±8.0</td>
<td>68.9±13.1</td>
</tr>
<tr>
<td>Majority consensus</td>
<td>70.5±8.8</td>
<td>71.3±9.8</td>
</tr>
<tr>
<td>Strict consensus</td>
<td>73.0±9.4</td>
<td>73.4±10.7</td>
</tr>
</tbody>
</table>
Figure 1. Receiver operating characteristics (ROC) curves for predictors that use (a) BLOSUM62 sequence-encoding and (b) PSSM-based encoding.

REFERENCES


