Poster H-65

DP-BIND: a sequence-based application of multiple machine learning methods to predict DNA-binding sites in DNA-binding proteins.

Authors:
Igor Kuznetsov (*University at Albany*)
Seungwoo Hwang (*University at Albany*)
Zhenkun Gou (*University at Albany*)

Short Abstract: An application of three machine-learning methods to sequence-based prediction of DNA-binding sites in a DNA-binding protein. The performance of our predictors is better than that of other sequence-based methods. Outputs of the three individual methods are combined into consensus prediction to further improve performance.

Long Abstract:

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 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 [1-2]. 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. Predictors are available at http://lcg.rit.albany.edu/dp-bind.

Dataset.

We used a non-redundant set of 62 experimentally solved protein-DNA complexes that were utilized previously to develop predictors of DNA-binding sites [3-4]. We label an amino acid residue in a protein chain as DNA-binding if the distance from at least one of its heavy atoms to any heavy atom in DNA is shorter than the cutoff distance of 4.5Å. 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 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 [5] 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 [6]. 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 [7]. 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 [4]. In both the BLOSUM62 and PSSM encoding, we normalize all elements in the matrix between 0 and 1 using the logistic function $f(x)=1/[1+\exp(-x)]$. In all three encoding methods, nearest sequential neighbors of a sequence position are encoded with a standard procedure [6] 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) [8], kernel logistic regression (KLR) [9], and penalized logistic regression (PLR) [10]. For both SVM and KLR we used the Radial Basis Function (RBF) kernel. The SVM algorithm was implemented using the LIBSVM program.

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 test each predictor. We used accuracy (ACC), sensitivity (SN), and specificity (SP) to assess the performance of each predictor:

$ACC=(TP+TN)/(TP+FP+TN+FN)$, $SN=TP/(TP+FN)$, $SP=TN/FP+TN)$

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

Results.

Analysis of the performance of our predictors indicates that:

(1) All three individual sequence-based predictors have similar performance (ACC of 69.7% for SVM, 68.9% for KLR, and 68.6% for PLR).

(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% (78.9% for SVM and 73.7% for PLR).

(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-PSS (ACC of 66.4%, SN of 68.2%, SP of 66.0%) [4].

(4) The strict consensus prediction improves both sequence-based and PSSM-based predictions and achieves ACC of 82.4%, SN of 84.9%, and SP of 83.1%. 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.

References.

1. Jones,S., et al. (2003) Nucleic Acids Res., 31, 7189-7198.
2. Tsuchiya,Y., et al. (2004) Proteins, 55, 885-894.
3. Ahmad,S., et al. (2004) Bioinformatics, 20, 477-486.
4. Ahmad,S. and Sarai,A. (2005) BMC Bioinformatics, 6, 33-38.
5. Altschul,S.F., et al. (1997) Nucleic Acids Res., 25, 3389-3402.
6. Qian,N. and Sejnowski,T.J. (1988) J. Mol. Biol., 202, 865-884.
7. Henikoff,S. and Henikoff,J.G. (1992) Proc. Natl Acad. Sci. USA, 89, 10915-10919.
8. Christianini,N. and Shawe-Taylor,J. (2000) Cambridge, MA.
9. Zhu,J. and Hastie,T. (2005) J.Comp. Graph Stat., 14, 185-205.
10. le Cessie,S. and van Houwelingen,J.C. (1992) Appl. Statist., 41, 191-201.