The Fourth International Conference on Computational Systems Biology (ISB2010)
Suzhou, China, September 9–11, 2010
Copyright © 2010 ORSC & APORC, pp. 266–275

MetaDBSite: a Meta Approach to Improve Protein DNA-Binding Sites Prediction

JingNa Si[1,3] Zengming Zhang[1] Biaoyang Lin[1]
Michael Schroeder[2] Bingding Huang[1, 2,*]

[1] Systems Biology Division, Zhejiang-California International NanoSystems Institute, Zhejiang University, Kaixuan Road 268, 310029, Hangzhou, China.
[2] Bioinformatics Group, Biotechnology Center, Technical University of Dresden, Tatzbergstr 47, 01307, Dresden, Germany.
[3] College of Biological Sciences, China Agricultural University, Beijing 100193, China.
[*] Corresponding Author: Bingding Huang, E-mail: bdhuang@mail.systemsbiozju.org

Abstract Protein-DNA interactions play an important role in many fundamental biological activities such as DNA replication, transcription and repair. Identification of amino acid residues involved in DNA binding site is critical for understanding of the mechanism of gene regulations. In the last decade, there have been a number of computational approaches developed to predict protein-DNA binding sites based on protein sequence and/or structural information. In this article, we present metaDBSite, a meta web server to predict DNA-binding residues for DNA-binding proteins. MetaDBSite integrates the prediction results from six available online web servers: DISIS, DNABindR, BindN, BindN-rf, DP-Bind and DBS-PRED and it only uses sequence information of proteins. A large dataset of DNA-binding proteins are constructed from the Protein Data Bank and serves as a gold-standard benchmark to evaluate metaDBSite approach and the other six predictors. The comparison results show that metaDBSite outperforms the individual approach. We believe that metaDBSite will become a useful tool for protein DNA-binding residues prediction. The MetaDBSite server is freely available at http://projects.biotec.tu-dresden.de/metadbsite/.

Keywords: Protein-DNA interaction; MetaDBSite; Support vector machine

1 Introduction

Protein-DNA complexes perform essential functions in many cellular activities. For example, transcription factors bind to specific DNA sequences in promoters to activate gene expression [1]. Protein-DNA interactions also play important roles in many other biological processes, including DNA replication, DNA repairing, viral infection, DNA packing and DNA modifications [2]. However, the biophysical mechanism of protein-DNA interactions is not clear and the identification of protein-DNA interactions by experimental methods is difficult at present.

Although there are more than 60,000 experimentally determined structures deposited in the current (June 2010) Protein Data Bank database [3] , there are only several hundreds structures on protein-DNA complexes, which is much smaller than the number of protein-DNA complexes existed in nature. With recent advances in DNA sequencing such as the next-generation sequencing,genome sequences for for many organisms were completed in recent years, producing a huge amount of protein sequences, many of which are DNA-binding proteins. Predicting the DNA binding properties of these DNA binding proteins will be very useful in helping understanding their biological functions.

There are several state-of-the-art prediction servers for predicting DNA bindings based on protein sequences, including DISIS [2], DNABindR [4], BindN [5], BindN-rf [6], DP-Bind [7] and DBS-PRED [8]. Table 1 summarizes the detailed characteristics of these six servers. These six web servers are all based on protein sequences and they combined several features derived from sequence information, such as amino acid frequency, evolutionary profile, sequence conservation, predicted secondary structure, predicted solvent accessibility, electrostatic potential, hydrophobicity, BLOSUM62 matrix, position-specific scoring matrix (PSSM) etc [2, 4-5, 7]. Furthermore, various machine learning methods are used in these servers, including support vector machine (SVM) [9], Naïve Bayes classifier, random forest [10] and neural network [11].

Table 1. Summary of detailed characteristics of the six available web servers for DNA-binding sites prediction.

	Machine learning methods	Properties used in training	Online website
DISIS	SVM-light Neural network	Evolutionary profile Conservation Predicted secondary structure Predicted solvent accessibility	http://cubic.bioc.columbia.edu/services/disis
DNABindR	Naïve Bayes classifier	Relative solvent accessibility Sequence entropy Secondary structure Electrostatic potential Hydrophobicity	http://turing.cs.iastate.edu/PredDNA/predict.html
BindN	SVM	The side chain pKa value Hydrophobicity index Molecular mass	http://bioinfo.ggc.org/bindn/
BindN-rf	Random forest	The side chain pKa value Hydrophobicity index Molecular mass Blast-based conservation Biochemical feature Position-specific scoring matrix (PSSM)	http://bioinfo.ggc.org/bindn-rf/
DP-Bind	SVM Kernel logistic regression (KLR) Penalized logistic regression (PLR)	Sequence-based BLOSUM62 PSSM-based	http://lcg.rit.albany.edu/dp-bind/
DBS-PRED	Neural network	Protein sequence information Solvent accessibility Secondary structure	http://www.netasa.org/dbs-pred

However, several limitations impair the application of the above servers: each method constructed their own dataset; had their own definition of binding sites; used different parameters derived from sequences; applied different machine learning methods, produced different accuracy and sensitivity, and calculated at much different speed. Therefore, a better and more consistent prediction server is needed. To meet this goal, we have developed metaDBSite, a meta web server for predicting protein DNA-binding sites based solely on amino acid sequences of proteins. MetaDBSite combined the six available online web servers mentioned in Table 1. MetaDBSite used support vector machine (SVM) learning method to learn and test the data. We constructed a large dataset PDNA-316 from PDB and compared the performance of MetaDBSite and the six servers. We showed that the MetaDBSite has a higher sensitivity in distinguishing DNA binding sites on the benchmark dataset. We believe that metaDBSite will become a useful tool for predicting DNA-protein binding residues for researchers.

2 Methods and Materials

2.1 Benchmark dataset

To evaluate these prediction methods, we derived a large dataset of protein-DNA complexes from current PDB [3]. 865 protein-DNA complexes with resolution better than 3.0 Å were downloaded from PDB and the sequences were submitted to the program H-CD-HIT [12] to get a non-redundant dataset. These 865 proteins are first clustered at a high identity (90%), then the non-redundant sequences are further clustered at a low identity (60%). A third cluster is performed at lower identity (30%). Default clustering parameters were selected in H-CD-HIT. After clustering, we have 316 target proteins in total and it is called PDNA-316 dataset. This dataset is listed in the supplemental data on our metaDBSite web-server.

Several previous studies on DNA-protein binding site prediction [8, 13-15] have used various definitions of DNA-binding sites. Here, we tried different definitions of DNA-binding sites, in order to gain the most appropriate one. In a protein-DNA complex, an amino acid residue in the protein was defined as binding site if the distance between any atoms of this residue and any atoms of the DNA molecule was less than a series of cutoff distance of 3.5 Å, 4.0 Å, 4.5 Å, 5.0Å, 5.5 Å and 6.0 Å. All the other residues were regarded as non DNA-binding sites. On the other hand, we also tried to define binding sites with solvent accessible surface area (ASA). We calculated surface area for each protein residue when DNA chain was absent and present, respectively. The solvent accessible surface area of residues which changed at least 1% (relative surface area) before and after DNA chain appeared were considered to be DNA-binding residues, the other residues were regarded as non DNA-binding residues. Therefore, there are totally seven ways to define binding residues. We do the prediction based on two datasets with each definition. The detailed results will be discussed later in Results and Discussion section. In the final metaDBSite system, distance 3.5 Å was chosen to define the DNA-binding sites.

2.2 Performance measures

Four performance measures were used in MetaDBSite, which are accuracy, sensitivity, specificity, and strength. They are defined below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Strength = \frac{Sensitivity + Specificity}{2}$$

In the formula above, TP is the abbreviation of true positives (residues predicted to be DNA-binding residues that are in fact not interface residues); TN is the abbreviation of true negatives (residues predicted to be non DNA-binding residues that are in fact not DNA-binding residues); FP is the abbreviation of false positives (residues predicted to be DNA-binding residues that are in fact not interface residues)); FN is the abbreviation of false negatives (residues predicted to be non DNA-binding residues that are in fact DNA-binding residues).

2.3 SVM learning

In this work, the six predictors were combined into a prediction system called metaDBSite with the assistance of the Support Vector Machine (SVM). As a machine-learning method for two classes of classification, SVM aims to find a rule that put each member in a training set into the corresponding class correctly. Here, the SVM was trained to distinguish DNA-binding residues from non-binding residues.. DNA binding amino acids were regarded to be positive samples, and non-DNA binding amino acids were considered to be negative samples. The residue was defined as binding site if the distance between any atoms of this residue and any atoms of the DNA molecule was less than a cutoff distance of 3.5 Å. Within this context, the PDNA-316 dataset, there are 5342 positive samples and 67396 negative samples.

The detailed procedure of metaDBSite is illustrated in Figure 1. The given sequence is submitted to six web servers and the prediction results are retrieved. Among these six predictors, four of them (i.e., DISIS, DNABindR, BindN, and BindN-rf) return the prediction based on their own scoring functions. The residues with a score above a certain threshold are considered as DNA-binding residues. These scores provide us four input parameters for SVM. For the other two predictors: DP-Bind and DBS-PRED, they only indicate which residues are predicted to bind to DNA or not. Therefore, we simply add a score "+1" to binding sites and "0" to non-binding sites in these two methods. Finally, a total of six parameters are used in the SVM training.

The PDNA-316 datasets were divided into 10 roughly equal subsets. 10-fold cross-validation was performed here. To predict whether a given amino acid in a sequence belongs to the DNA binding site or non-DNA binding site, the subset to which this residue belongs was labeled as the "test" set, whereas the nine remaining subsets were labeled as “training” sets. SVM models were developed for each of the “training” sets. The class label for positive and negative samples was set to +1 and -1, respectively. The ratio of positive to negative samples was about 1:10 in the training set. Using the training set at such a ratio would inevitably cause the SVM model to predict every pair as a negative case. The optimized ratio in the training set was set at 1:1. Each training set was modified by discarding a random selection of the negative samples prior to training. The implemented SVM algorithm was LIB-SVM (http://www.csie.ntu.edu.tw/~cjlin/). The applied kernel function was the radial basis function (RBF). The corresponding parameter settings of SVM learning were automatically optimized by LIB-SVM.

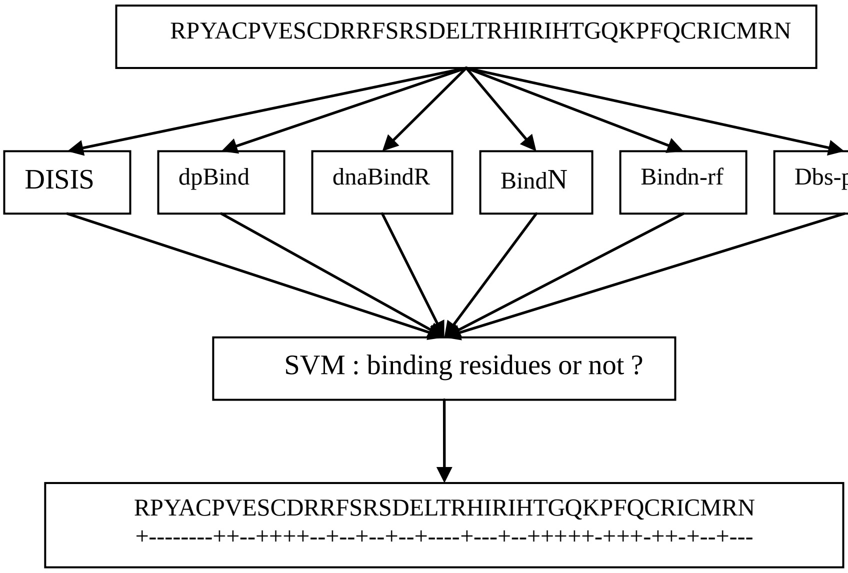

RPYACPVESCDRRFSRSDELTRHIRIHTGQKPFQCRICMRN
DISIS
dpBind
dnaBindR
BindN
Bindn-rf
Dbs-pred
SVM : binding residues or not ?
RPYACPVESCDRRFSRSDELTRHIRIHTGQKPFQCRICMRN
+--------++--++++--+--+--+--+----+---+--+++++-+++-++-+--+---

Figure 1. The prediction workflow of the metaDBSite approach. The protein sequence is submitted to the six predictors and the prediction results are retrieved. Then these predicted results are input into the trained SVM and the final prediction (which residues are DNA-binding sites (marked as “+”) is made.

3 Results and Discussion

3.1 Performance on PDNA-316 benchmark dataset

Table 2 shows the prediction results for metaDBSite approach (10-fold SVM cross-validation) and the other six methods alone, on PDNA-316 benchmark dataset. It is noted that DISIS gained 19% sensitivity with very high accuracy and specificity. It is also noted that DISIS failed to return any prediction for over 60 proteins in this dataset due to the strict restriction in its web-server parameters. In such a case, small binding sites with very high confidence were found. However, in the same time, many real DNA-binding residues were missing. In a prediction, the balance of exact value and confident level is important. Therefore, high accuracy and specificity with very small sensitivity of DISIS method come to be incomparable with other methods. MetaDBSite's sensitivity of 77% is much higher than each of the single method, and it is 10 percentages higher than BindN-rf, which has the highest sensitivity among the single methods. Moreover, the strength of metaDBSite is 77%, which also holds the line with the best one among the six methods. Although the accuracy of metaDBSite is a little lower than some of the single methods, metaDBSite is still considered meaningful because of the best performance of sensitivity and strength. Sensitivity is the measurement of DNA-binding residues prediction, which is the most interest point for relevant researchers. Strength is considered to be fair evaluation criteria when the datasets are imbalanced in previous studies [8, 16]. In such cases, sensitivity and strength of metaDBSite are also better than each single method; especially sensitivity has gained an obvious improvement. Not only metaDBSite achieves the most prober prediction results, but also it provides the users some analysis and comparison among different methods.

Table 2. The prediction results for metaDBSite (10-fold SVM cross-validation) and the other six methods alone for PDNA-316 benchmark dataset.

ID	accuracy	sensitivity	specificity	strength
metaDBSite	0.77	0.77	0.77	0.77
BindN	0.78	0.54	0.80	0.67
BindN-rf	0.82	0.67	0.83	0.75
DBS-PRED	0.75	0.53	0.76	0.65
DISIS	0.92	0.19	0.98	0.59
DNABindR	0.73	0.66	0.74	0.70
dpBind	0.78	0.54	0.80	0.67

3.2 Comparison of various definitions of DNA-binding sites

In the previous studies, DNA-protein binding sites were defined as the distance between any atoms of one residue and any atoms of the DNA molecule was less than a cutoff value, e.g. 3.5 Å~6.0 Å [8, 13-15]. In order to find out the most proper distance to be the binding distance, we have tried several cutoff distances in this work. On the other hand, we also defined the DNA-binding sites by solvent accessible surface area. Figure 2 shows the overall prediction performance of metaDBSite with different definitions on the PDNA-316 dataset. The sensitivity decreased obviously and successively when the cutoff distance increased. The accuracy at 3.5 Å distance was just lower than that at 5.5 Å distance. But the sensitivity at 5.5 Å was 69%, which was much lower than that of 77% at 3.5 Å. The specificity had similar tendency. The specificity in 3.5 Å was not the highest. However, when considering the overall performance of these three measurements together, 3.5 Å is the best cutoff distance. Therefore, we choose 3.5 Å as the cutoff in our definition of DNA-binding residues.

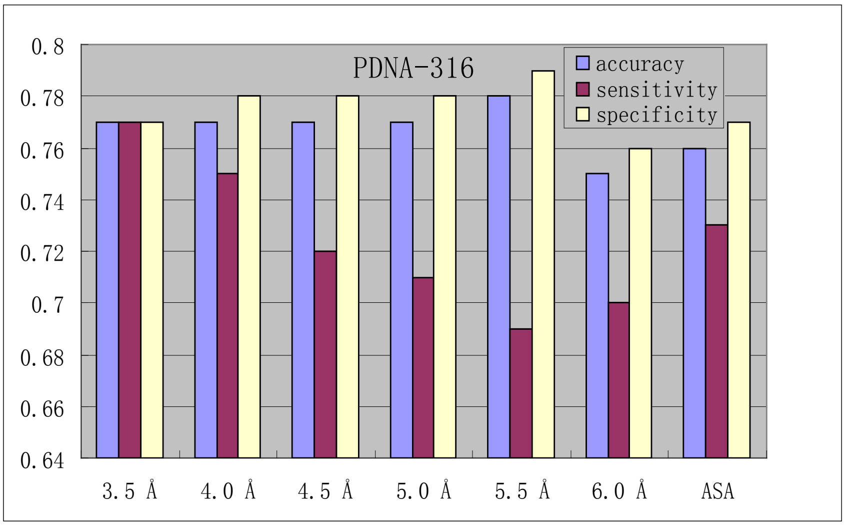


Figure 2 Performance of metaDBSite with DNA-binding site definitions using different distance cut-off and ASA method on the PDNA-316 benchmark dataset.

3.3 Representative example

MetaDBSite reveals its advantage in distinguishing DNA-binding residues sufficiently. In our test dataset, more than 100 proteins were spotted with the sensitivity value of 1.0, which means all the real DNA-binding residues are recognized correctly. Figure 3 shows an example of these proteins (PDB ID: 1REP, Chain: C). In Figure 3A, those residues in blue are the predicted DNA-binding residues by metaDBSite. In Figure 3B, residues in red are the real DNA-binding residues defined with 3.5 Å distance cutoff. The difference between residues in red and in blue can be seen directly from Figure 3, which is the noise of false positive. False positive samples are inevitable in any prediction system. Here in this protein, the prediction accuracy is 89% and specificity is 88% while sensitivity is 1.0.

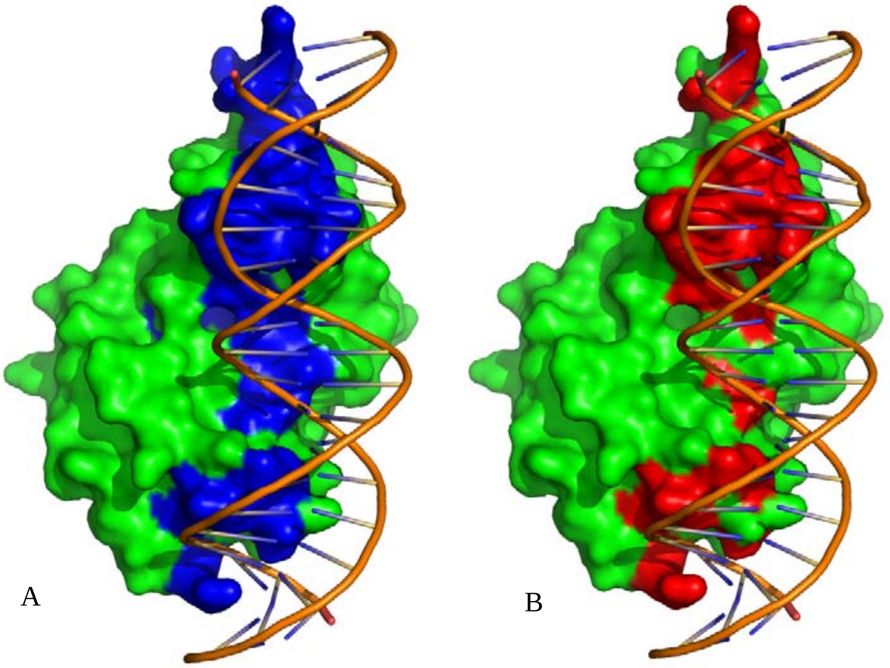
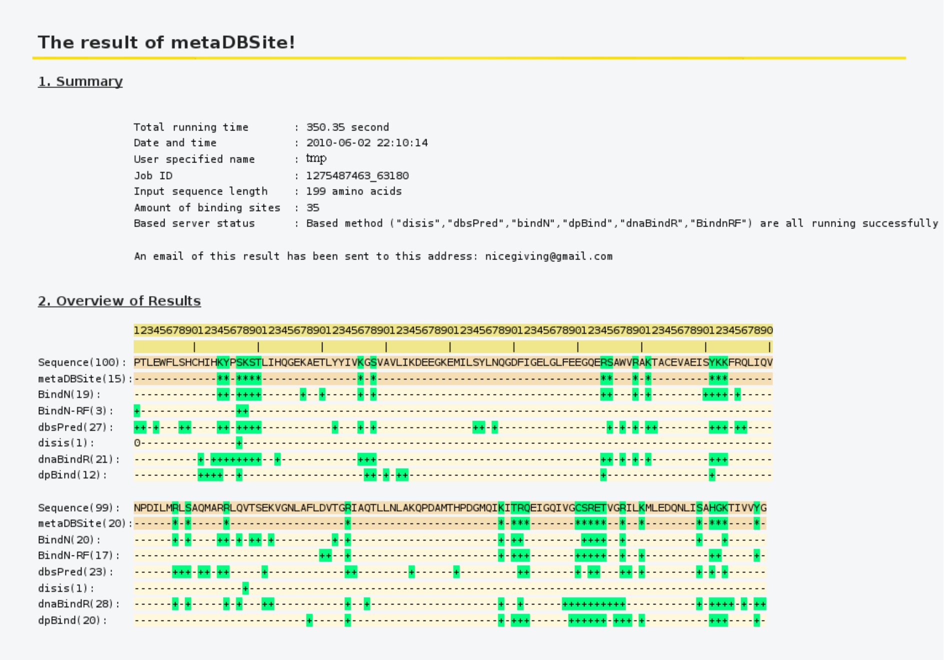


Figure 3. Representative protein-DNA complex: replication initiation protein and its binding DNA regions (PDB ID: 1REP). A: Predicted DNA-binding residues are in blue. B: The real DNA-binding residues defined with 3.5 Å are in red. The replication initiation protein is shown in green.

4 Conclusions

DNA-binding residues prediction from protein sequence is of great importance to understand the mechanism of protein-DNA interactions. There have been a lot of research efforts done to discriminate DNA-binding residues from non DNA-binding ones. Various machine learning methods have been applied and different kinds of features based on protein sequence and/or structural information have been used. However, it is hard to directly compare these existing prediction methods because of different data-sets, definitions and evaluation criteria being used. Here, based on the prediction results from six available predictors, we developed metaDBSite: a meta server for DNA-binding residues prediction based on protein sequences. We evaluated the performance of metaDBSite and other 6 predictors on a large data-set using the same definition and criteria. We have shown that MetaDBSite can achieve a better balance of sensitivity and specificity.

MetaDBSite is freely available at http://projects.biotec.tu-dresden.de/metadbsite. The users can simply submit a protein sequence for DNA-binding residues prediction. MetaDBSite will re-direct the submitted sequence to the six predictors automatically and the prediction results are retrieved and analyzed. After the process is finished, the users will be notified by e-mail with a URL to view the prediction result. Figure 4 shows a screenshot of the result page of metaDBSite server. It lists the predicted DNA-binding sites (marked as "*" and "+") for metaDBSite approach and the other 6 predictors. The whole process for each query normally takes no more than 10 minutes with parallel computational processes on a Linux desktop with a CPU of 2.85 HZ and 2 G memory. If any servers fail to return any prediction due to network problem or server shut-down, metaDBSite will automatically ignore them and continue with those successful predictions.



The result of metaDBSite!

1. Summary

```
Total running time       : 350.35 second
Date and time            : 2010-06-02 22:10:14
User specified name      : tmp
Job ID                   : 1275487463_63180
Input sequence length    : 199 amino acids
Amount of binding sites  : 35
Based server status      : Based method ("disis","dbsPred","bindN","dpBind","dnaBindR","BindnRF") are all running successfully

An email of this result has been sent to this address: nicegiving@gmail.com
```

2. Overview of Results

```
                1234567890123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890
                         |         |         |         |         |         |         |         |         |         |
Sequence(100):  PTLEWFLSHCHIHKYPSKSTLIHQGEKAETLYYIVKGSVAVLIKDEEGKEMILSYLNQGDFIGELGLFEEGQERSAWVRAKTACEVAEISYKKFRQLIQV
metaDBSite(15): -------------**-****---------------*-*-----------------------------------**---*-*---------***-------
BindN(19):      -------------++-++++------+--+-----+-+-----------------------------------++---+-+--------++++-+-----
BindN-RF(3):    +---------------++----------------------------------------------------------------------------------
dbsPred(27):    ++-+---++----++-++++-----------+---+-+---------------++-+-----------------+-+-+-++--------+++-++----
disis(1):       0---------------+-----------------------------------------------------------------------------------
dnaBindR(21):   ----------+-++++++++--+------------+++-----------------------------------++-+-+-+---------+++-------
dpBind(12):     ----------++++--+-------------------++-+-++------------------------------+----------------+---------

Sequence(99):   NPDILMRLSAQMARRLQVTSEKVGNLAFLDVTGRIAQTLLNLAKQPDAMTHPDGMQIKITRQEIGQIVGCSRETVGRILKMLEDQNLISAHGKTIVVYG
metaDBSite(20): ------*-*-----*------------------*-----------------------*-***-------*****--*--*--------*-***----*-
BindN(20):      ------+-+----++-+-++-+---------+-+-----------------------+-++---------++++--+-----------+---+------
BindN-RF(17):   -----------------------------++--+-----------------------+-+++-------+++++--+--+----------++-----+-
dbsPred(23):    ------+++-++-++-----+------------++--------+------+---------++-------+-++---++-+--------+-+-+------
disis(1):       -----------------+---------------------------------------------------------------------------------
dnaBindR(28):   ------+-+-----+-+---++-----------+--+--------------------+--+------++++++++++-----------+-++++-+-++
dpBind(20):     ---------------------------+-----+-----------------------+-+++------++++++-+++-+----------+++----+-
```

Figure 4. Screenshot of result page on the metaDBSite server. The predicted DNA-binding residues are marked "+" for the sixe predictors and "*" for metaDBSite and are all colored green. The non DNA-binding residues are marked "-".

Acknowledges

Funding from the Ministry of Science and Technology of China (grant No. 2008DFA11320) is kindly acknowledged.

References

[1] Ptashne M: Regulation of transcription: from lambda to eukaryotes. Trends Biochem Sci 2005, 30:275-279.

[2] Ofran Y, Mysore V, Rost B: Prediction of DNA-binding residues from sequence. Bioinformatics 2007, 23:i347-353.

[3] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: The Protein Data Bank. Nucleic Acids Res 2000, 28:235-242.

[4] Yan C, Terribilini M, Wu F, Jernigan RL, Dobbs D, Honavar V: Predicting DNA-binding sites of proteins from amino acid sequence. BMC Bioinformatics 2006,

7:262.

[5] Wang L, Brown SJ: BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res 2006, 34:W243-248.

[6] Wang L, Yang MQ, Yang JY: Prediction of DNA-binding residues from protein sequence information using random forests. Bmc Genomics 2009, 10 Suppl 1:S1.

[7] Hwang S, Gou Z, Kuznetsov IB: DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. Bioinformatics 2007, 23:634-636.

[8] Ahmad S, Gromiha MM, Sarai A: Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics 2004, 20:477-486.

[9] Lv S, Wang X, Cui Y, Jin J, Sun Y, Tang Y, Bai Y, Wang Y, Zhou L: Application of attention network test and demographic information to detect mild cognitive impairment via combining feature selection with support vector machine. Comput Methods Programs Biomed, 97:11-18.

[10] Calle ML, Urrea V: Letter to the Editor: Stability of Random Forest importance measures. Brief Bioinform.

[11] De Roach JN: Neural networks--an artificial intelligence approach to the analysis of clinical data. Australas Phys Eng Sci Med 1989, 12:100-106.

[12] Li W, Godzik A: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006, 22:1658-1659.

[13] Tsuchiya Y, Kinoshita K, Nakamura H: PreDs: a server for predicting dsDNA-binding site on protein molecular surfaces. Bioinformatics 2005, 21:1721-1723.

[14] Ahmad S, Sarai A: PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformatics 2005, 6:33.

[15] Jones S, Shanahan HP, Berman HM, Thornton JM: Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. Nucleic Acids Res 2003, 31:7189-7198.

[16] Wang L, Brown SJ: Prediction of DNA-binding residues from sequence features. J Bioinform Comput Biol 2006, 4:1141-1158.