

REPORT

Open Access

# MetaDBSite: a meta approach to improve protein DNA-binding sites prediction

Jingna Si<sup>1,3†</sup>, Zengming Zhang<sup>1†</sup>, Biaoyang Lin<sup>1</sup>, Michael Schroeder<sup>2</sup>, Bingding Huang<sup>1,2\*</sup>

From The 4th International Conference on Computational Systems Biology (ISB 2010)  
Suzhou, P. R. China. 9-11 September 2010

## Abstract

**Background:** 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.

**Results:** 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 solely uses sequence information of proteins. A large dataset of DNA-binding proteins is constructed from the Protein Data Bank and it serves as a gold-standard benchmark to evaluate the metaDBSite approach and the other six predictors.

**Conclusions:** The comparison results show that metaDBSite outperforms single individual approach. We believe that metaDBSite will become a useful and integrative tool for protein DNA-binding residues prediction. The MetaDBSite web-server is freely available at <http://projects.biotech.tu-dresden.de/metadbsite/> and <http://sysbio.zju.edu.cn/metadbsite>.

## Background

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 in nature. With recent advances in DNA sequencing such as next-generation sequencing technology, genome sequences 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

\* Correspondence: [bhuang@biotech.tu-dresden.de](mailto:bhuang@biotech.tu-dresden.de)

† Contributed equally

<sup>1</sup>Systems Biology Division, Zhejiang-California International NanoSystems Institute, Zhejiang University, Kaixuan Road 268, 310029, Hangzhou, China  
Full list of author information is available at the end of the article

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

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].

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 protein-DNA binding residues for relevant researchers.

## Results and discussion

### Performance on PDNA-316 benchmark dataset

The detailed procedure of metaDBSite is illustrated in Figure 1. 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 but 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, the DISIS method, with high accuracy and specificity but low sensitivity, is incomparable with the other methods. By integrating the prediction results from the six methods, MetaDBSite has achieved 77% sensitivity value and it is much higher than each single method. This sensitivity is 8 percentages higher than that of DPBind method, 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 to be the best 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.



had similar tendency. The specificity in 3.5 Å was not the highest. Therefore, after considering the overall performance of these three measurements together, we chose 3.5 Å as the distance threshold to define the real DNA-binding residues in this study.

### 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 by metaDBSite. 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, those residues in red are the real DNA-binding residues defined with 3.5 Å distance threshold. The difference between residues in red and in blue can be seen directly from Figure 3, which is the false positive in the prediction. Here in this protein, the prediction accuracy is 89% and specificity is 88% while sensitivity is 1.0.

### Using structural information to eliminate false positives

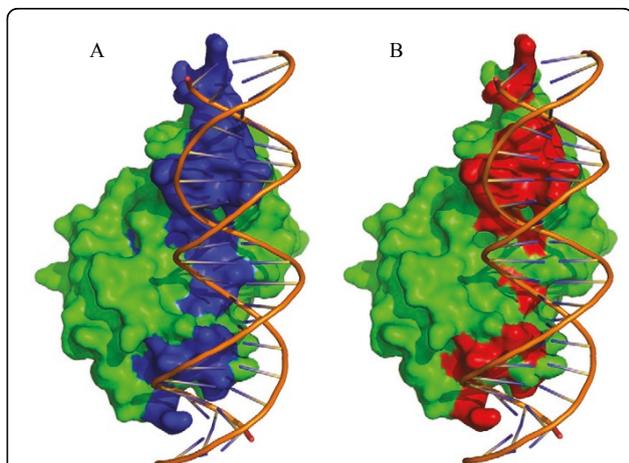
It is noted that those six methods and metaDBSite are all using the information of protein sequence only. Since the dataset of PDNA-316 is derived from PDB and the structures for all proteins are known. This structural information of proteins could be used to improve the prediction of metaDBSite. To do this, we used spatial clustering based on the coordinates of the CA atoms of those predicted residues in metaDBSite in the next step, trying to eliminate those false positive predictions (FP). After clustering, those clusters with small number of

residues are treated as false positive and thus are removed. We have tried several different parameters in this clustering procedure. All the results have shown that the accuracy and specificity are both increased but the sensitivity decreased (data not shown). In the protein-DNA complex structure, because of the spiral of the DNA molecule, the real DNA-binding residues defined with a distance cut-off do not tend to gather together spatially. Some of the real DNA-binding sites can be 3 or less residues and may locate at an isolated position on protein surface. Therefore, when we eliminate those small clusters, some of TPs may also be removed with FPs at the same time. And this is the reason why we can increase specificity and decrease sensitivity after this clustering post-process.

### 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 have developed metaDBSite: a meta server for DNA-binding residues prediction based on protein sequence. We evaluated 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> and <http://sysbio.zju.edu.cn/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 results. 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.



**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.

## 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



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

## Materials and methods

### 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 [16] to get a non-redundant dataset. These 865 proteins are first clustered at a high identity (90%), and 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 protein-DNA complexes in total and it is called PDNA-316 dataset. This dataset is listed in the supplemental data on our metaDBSite web-server and can be downloaded freely.

### Defining real DNA-binding sites

Several previous studies on protein-DNA binding site prediction [8,13-15] have used various definitions of

DNA-binding sites. In a protein-DNA complex, an amino acid residue in the protein is defined as DNA-binding site if the distance between any atoms of this residue and any atoms of the DNA molecule is less than a distance threshold. This threshold ranged from 3.5 Å to 6.0 Å in the previous studies. The other residues are 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 molecule was absent and present, respectively. The solvent accessible surface area of residues which change at least 1 Å<sup>2</sup> before and after DNA molecule appeared are considered to be DNA-binding residues, the other residues are regarded as non DNA-binding residues. In the final metaDBSite approach, distance 3.5 Å was chosen to define the real DNA-binding sites.

### Performance measures

Four performance measures were used in MetaDBSite, which are accuracy, sensitivity, specificity, strength,

Mathews correlation coefficient (MCC) and F-measure. They are defined below:

$$\begin{aligned} \text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \\ \text{Sensitivity} &= \frac{TP}{TP + FN} \\ \text{Specificity} &= \frac{TN}{TN + FP} \\ \text{Strength} &= \frac{\text{Sensitivity} + \text{Specificity}}{2} \\ \text{Mcc} &= \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \\ \text{F-measure} &= \frac{2 \times \text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \end{aligned}$$

In the formulas above, TP is the abbreviation of true positives (residues predicted to be DNA-binding residues that are in fact DNA-binding residues); TN is the abbreviation of true negatives (residues predicted to be non DNA-binding residues that are in fact non DNA-binding residues); FP is the abbreviation of false positives (residues predicted to be DNA-binding residues that are in fact non DNA-binding residues); FN is the abbreviation of false negatives (residues predicted to be non DNA-binding residues that are in fact DNA-binding residues). These definitions and measures are comparable to the previous studies.

### 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 as a classifier for two classes, 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 3.5 Å. With this definition, there are 5342 positive samples and 67396 negative samples in the PDNA-316 dataset.

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, DNA-BindR, 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 labelled as the “test” set, whereas the nine remaining subsets were labelled 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.

### Acknowledgements

Funding from the Ministry of Science and Technology of China (grant No. 2008DFA11320) and EU 7<sup>th</sup> framework Marie Curie Action IRESE program (grant NO. 247097) is kindly acknowledged.

This article has been published as part of *BMC Systems Biology* Volume 5 Supplement 1, 2011: Selected articles from the 4th International Conference on Computational Systems Biology (ISB 2010). The full contents of the supplement are available online at <http://www.biomedcentral.com/1752-0509/5?issue=S1>.

### Author details

<sup>1</sup>Systems Biology Division, Zhejiang-California International NanoSystems Institute, Zhejiang University, Kaixuan Road 268, 310029, Hangzhou, China.

<sup>2</sup>Bioinformatics Group, Biotechnology Center, Technical University of Dresden, Tatzbergstr 47, 01307, Dresden, Germany. <sup>3</sup>College of Biological Sciences, China Agricultural University, 100193, Beijing, China.

### Authors' contributions

JNS collected the data, implemented codes, and drafted the manuscript. ZZ helped implementation and developed the web server. BH directed the research and seriously revised the manuscript. BL and MS co-supervised the work and revised the manuscript. All authors read and approved the final manuscript.

### Competing interests

The authors declare that they have no competing interests.

Published: 20 June 2011

### References

1. Ptashne M: Regulation of transcription: from lambda to eukaryotes. *Trends Biochem Sci* 2005, **30**:275-279.
2. Ofraan 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* 2010.
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. Wang L, Brown SJ: **Prediction of DNA-binding residues from sequence features**. *J Bioinform Comput Biol* 2006, **4**:1141-1158.
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. 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.

doi:10.1186/1752-0509-5-S1-S7

**Cite this article as:** Si et al.: MetaDBSite: a meta approach to improve protein DNA-binding sites prediction. *BMC Systems Biology* 2011 **5**(Suppl 1):S7.

**Submit your next manuscript to BioMed Central  
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
www.biomedcentral.com/submit

