Mol. Cells *30*, 99-105, August 31, 2010
DOI/10.1007/s10059-010-0093-0

Molecules
and
Cells

A Novel Sequence-Based Method of Predicting Protein DNA-Binding Residues, Using a Machine Learning Approach

Yudong Cai[1,2,*], ZhiSong He[3], Xiaohe Shi[4], Xiangying Kong[4,5], Lei Gu[6], and Lu Xie[7]

Protein-DNA interactions play an essential role in transcriptional regulation, DNA repair, and many vital biological processes. The mechanism of protein-DNA binding, however, remains unclear. For the study of many diseases, researchers must improve their understanding of the amino acid motifs that recognize DNA. Because identifying these motifs experimentally is expensive and time-consuming, it is necessary to devise an approach for computational prediction. Some in silico methods have been developed, but there are still considerable limitations. In this study, we used a machine learning approach to develop a new sequence-based method of predicting protein-DNA binding residues. To make these predictions, we used the properties of the micro-environment of each amino acid from the AAIndex as well as conservation scores. Testing by the cross-validation method, we obtained an overall accuracy of 94.89%. Our method shows that the amino acid micro-environment is important for DNA binding, and that it is possible to identify the protein-DNA binding sites with it.

INTRODUCTION

Protein-DNA interactions control a variety of vital biological processes, such as gene regulation, DNA replication and repair, recombination, and other critical steps in cellular development (Luscombe et al., 2000). Mutations of DNA-binding regions, such as those on the tumor repressor protein P53, may be directly associated with severe human diseases (Bullock and Fersht, 2001). Thus, the ability to identify the amino acid motifs that recognize DNA can significantly improve our understanding of these biological processes, potentially guiding the functional characterization of DNA-binding proteins in site-directed mutagenesis studies. In addition, such knowledge can contribute to further advances in drug discovery, such as aiding the design of artificial transcription factors (Ahmad et al., 2004; Ho et al., 2007; Hwang et al., 2007; Ofran et al., 2007; Wang and Brown, 2006).

The protein-DNA recognition mechanism is complicated, with interactions consisting of a variety of parameters involving hydrophobicity, alpha and turn propensities, beta propensity, composition, physicochemical properties, and other properties. Researchers have gained insights into the mechanisms of protein-DNA specific binding by using three-dimensional (3D) structures of individual protein-DNA complexes coupled with directed mutagenesis and biochemical analysis. Unfortunately, 3D structures of such complexes are available for fewer than 5% of all known DNA-binding proteins (Ofran et al., 2007). Moreover, it is implausible that researchers will “solve” the structure of every protein-DNA complex through 3D crystal structure resolution. Compared to conventional experimental studies, in silico methods offer the advantages of high efficiency and low cost. Most are based on the idea of geometric or functional inference through homology.

A variety of computational methods have been developed to predict DNA-binding sites. Within these structures, recognition involves partially direct contacts between amino acids and base pairs.

The homeodomain is a DNA binding motif that is found in numerous transcription factors throughout a large variety of species from yeast to humans. Analysis of all 84 independent homeodomains from D. melanogaster reveals the breadth of DNA sequences that can be specified by this recognition motif (Noyes et al., 2008). The homeodomain consists of approximately 60 amino acids that fold into a stable three-helix bundle preceded by a flexible N-terminal arm. Usually, interactions with the 5 to 7 base pair DNA binding sites are formed by a single “recognition” helix motif in the major groove and the N-terminal arm in the minor groove (Noyes et al., 2008). Researchers have used a computational approach for predicting human DNA-binding sites in proteins from amino acid sequences, using a random forest model with a hybrid feature (Wu et al.,

[1]Institute of System Biology, Shanghai University, Shanghai 200244, People’s Republic of China, [2]Centre for Computational Systems Biology, Fudan University, Shanghai 200433, People’s Republic of China, [3]Department of Bioinformatics, College of Life Sciences, Zhejiang University, ZheJiang 310058, People’s Republic of China, [4]Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences (CAS) and Shanghai Jiao Tong University School of Medicine, People’s Republic of China, [5]State Key Laboratory of Medical Genomics, Ruijin Hospital, Shanghai Jiaotong University, Shanghai 200025, People’s Republic of China, [6]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, Germany, [7]Shanghai Center for Bioinformation Technology, Shanghai 200235, People’s Republic of China
*Correspondence: cai_yud@yahoo.com.cn

Received August 14, 2009; revised April 6, 2010; accepted April 22, 2010; published online July 23, 2010

Keywords: bioinformatics, data mining, machine learning, mRMR, protein-DNA interaction

2009). Researchers have also incorporated high-throughput chromatin modification to improve the accuracy of prediction of transcription factor binding sites (Whitington et al., 2009). Several programs, such as Stubb (Sinha et al., 2003), EvoPromoter (Wong and Nielsen, 2007), and PhylCRM (Warner et al., 2008), have been developed to predict protein-DNA binding sites as significant clusters of transcription factor binding sites, which are detected by comparing orthologous sequences using an evolutionary model of binding sites. For the protein view, a knowledge-based method DNA-binding Domain Hunter (DBD-Hunter) was developed for identifying DNA-binding proteins and associated binding sites (Gao and Skolnick, 2008). It could be calculated that combination of the protein-DNA complex energies leads to enhanced specificity, and the combined energy could explain experimental data on binding affinity changes caused by base mutations (Gromiha et al., 2005). More recent research suggests that mimicked DNA-recognition preferences are compatible with experimental results (Jamal Rahi et al., 2008; Kaplan et al., 2005; Tan et al., 2005; Vavouri and Elgar, 2005). But these methods have at least two major limitations: First, most of the algorithms attribute to the DNA-binding motif preference not the binding proteins analysis; secondly, even dealing with binding protein level, they cannot compare the contribution of each amino acid in the motif to the protein-DNA binding activity.

Several machine learning methods have been applied to predict the protein-DNA binding sites (Ofran et al., 2007; Wu et al., 2009), but the research presented here is the first to compare the contribution of each amino acid in the motif to the protein-DNA binding activity based on the feature selection method, which is a part of the machine learning approaches. .

The main goal of the current study is to develop a protein sequence based method for predicting DNA-binding proteins and associated each amino acid residue preferences from structural genomics targets. The current method employs AAIndex features and a quantized conservation score for each position to represent a fixed-length window running through the protein sequence, for the purpose of prediction DNA-binding sites. The results shows this model achieves 94.89% overall accuracy, and amino acids directly binding to DNA motif or near the direct binding site attributes more sensitivity and specificity.

MATERIALS AND METHODS

Data set

For this study, we used the data set described in the work of Ofran et al. (2007). Specifically, we downloaded all protein-DNA complexes in the Protein Data Bank (PDB, http://www.rcsb.org/pdb/home/home.do) (Berman et al., 2000) were. To reduce the bias of similar sequences, HSSP-value was used as the measure of sequence similarity, and a non-redundant subset, in which no two proteins had HSSP-value > 0, was obtained. All protein-DNA complexes in the PDB with proteins in the non-redundant subset were obtained (Ofran et al., 2007). In these complexes, an amino acid is considered to be in contact with a nucleotide if the distance between any atoms of the two molecules is no more than 6 Å.

To obtain information regarding the microenvironment of each amino acid in the protein sequences, we used a sliding window with length of 9 to scan the sequence (i.e., for one amino acid, the 4 amino acids neighbor on each side were used as the environment of the one in the center). If the amino acid positioned in the center was in contact with a nucleotide, we would classify that 9-amino-acid sample as a positive one. Otherwise, it would be considered negative. The data used in this study can be found in Supplementary Data 1.

Features construction

To represent the properties of each amino acid in each instant, we used features of AAIndex and another feature of conservation.

The features of AAIndex

AAIndex (http://www.genome.ad.jp/aaindex/) is a database of numerical indices representing various physicochemical and biochemical properties of single amino acids and pairs of amino acids. It consists of three sections: AAIndex1, AAIndex2, and AAIndex3. In our study, we used AAindex1, the database for the amino acid index of 20 numerical values. It contains 544 indices for every amino acid, which represent different physicochemical and biological properties. We excluded AAindex1 indices with null values, and, therefore, used 506 indices for encoding samples. Because each amino acid and its nearest 4 amino acids to each side are considered, each sample can be encoded to 506*9 = 4554 features using AAIndex.

The feature of conservation

Conservation is one of the most important concepts in biology. If an amino acid in a particular position of a particular protein is conserved among different species, it may mean that this amino acid is located in an important region of the protein, and is able to absolutely change the protein's shape and function once it mutated. In our study, we used a conservation score to quantify the conservation status of each amino acid in the protein sequence. First, we used PSI-BLAST (Altschul et al., 1997) to find out all the proteins homologous to the query. Once all the sequences were obtained, we employed ClustalW (Larkin et al., 2007) to do the multi-sequences alignment. The conservation score of each position in the protein was calculated with the CONSCORE approach (Valdar, 2002) based on the output of alignment.

With the AAIndex features and conservation feature for each amino acid, each sample used in our study could be coded into a vector with 4554 + 9 = 4563 dimensions.

Minimum redundancy, maximum relevance (mRMR)

The minimum redundancy maximum relevance (mRMR) method, developed by Peng et al. (2005) is primarily used to deal with microarray data. Here we used it for feature analysis and selection. It ranks each feature according to its relevance to the target and redundancy to other features. In mRMR, a good feature means a maximum relevance to the target and minimum redundancy to other features in mRMR. To calculate relevance and redundancy, mutual information (MI) is used.

In the calculation of MI, the joint probabilistic density and the marginal probabilistic densities of the two vectors should be given. If the variable is continuous, it should be transformed into a new discrete variable first; this can be accomplished by scaling it into several groups according to its value. In mRMR, we used a parameter t to separate each feature in our data into one of three categorical states according to the equation *mean* ± (t · *std*): those with their value smaller than *mean*-(t·*std*), those with values between *mean*-(t·*std*) and *mean*+(t·*std*), and those with values larger than *mean*+(t·*std*), where *mean* is the mean value of the feature in all samples, and *std* is the standard deviation. In our study, t was designated as 1.

Nearest neighbor algorithm (NNA)

Nearest neighbor algorithm (NNA) is a simple but useful algorithm to solve the problem of vector classification (Qian et al.,

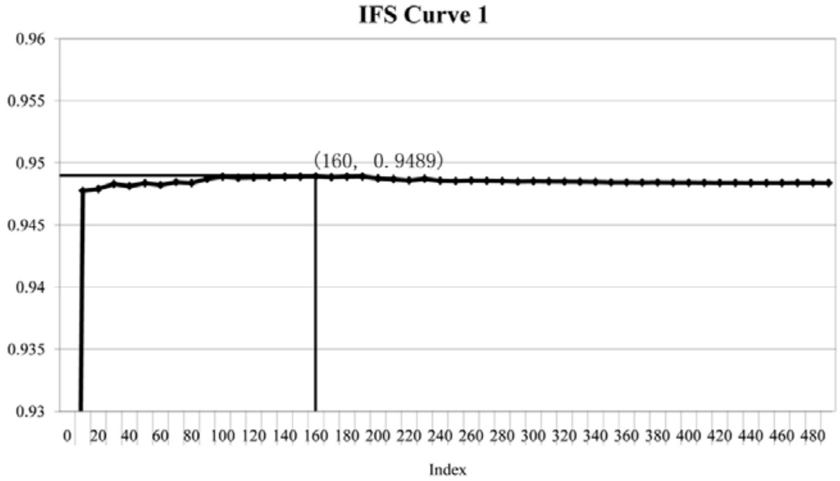

Fig. 1. The first IFS curve. The IFS curve for the 50 results of classifiers based on the 1st, 11th, 21st, …, 491st feature subsets.

2006). It has been widely applied in various other problems in bioinformatics such as protein secondary structure prediction, protein solvent accessibility prediction, and protein cellular localization prediction (Horton et al., 2007; Salamov and Solovyev, 1997; Sim et al., 2005). Its classification is based on the distances between the vector to be tested and all the vectors in the training set. The vector to be predicted would be designated to be in the same class as its nearest neighbor in the training set.

Jackknife cross-validation

The jackknife cross-validation method (Cai et al., 2009) is one of the most effective methods to evaluate the results of prediction, and we used it here to test performances of our classifiers. In the Jackknife cross-validation method, each sample is used as a to-be-test sample and all the other samples are used as the training set for one time. To evaluation the performance, the overall accuracy is used:

$$\text{Overall accuracy} = \frac{\text{correctly predicted samples}}{\text{all samples}} \tag{7}$$

Here, one sample is an oligo-peptide with fixed-length of 9 amino acids as described earlier.

Incremental feature selection (IFS)

Knowing which features are better with the ordered feature set *S* as mentioned above, the next step for feature selection is to obtain which features should be selected. To solve this problem, we used the Incremental Feature Selection (Cai et al., 2009) method. Based on the ordered feature set *S* obtained in mRMR step, we can get *N* feature subset. The i-th subset is defined as:

$$S_i = \{f_1^{'}, f_2^{'}, \ldots, f_i^{'}\}(1 \leq i \leq N) \tag{8}$$

For each feature subset, NNA can be used to construct a classifier for the data set, and Jackknife cross-validation method can be used to evaluate the classifier's performance. The results are plotted to build an IFS curve, with its x-axis to be the index *i*, and its y-axis to be the accuracy.

RESULTS

The results of mRMR

We downloaded the mRMR program used in our study from http://research.janelia.org/peng/proj/mRMR/. The output of mRMR contains two tables: MaxRel list and mRMR list. The latter shows the first 500 indexes of features in the ordered feature set *S,* as mentioned in "Materials and Methods", and it is used in our study for the IFS procedure. The former table shows the relevance of each features to the target variable as defined in Eq(2). In this study, only the mRMR list was used. Please see Supplementary Data 2 for the whole output of mRMR.

The result of IFS

Based on the ordered feature set *S* obtained in the mRMR step, we obtained 500 feature subsets. To improve efficiency, we first built 50 classifiers for the 1st, 11th, 21st, …, 491st feature subsets, and tested them using the Jackknife cross-validation method. The 161st classifier with an index of 161 has the highest overall accuracy of 0.9489. Figure 1 shows the IFS curve drew with these 50 results. To obtain the more optimized feature set, we then built another 20 classifiers for the 20 feature subsets with index from 151 to 170; we also tested them using Jackknife cross-validation. Figure 2 shows the IFS curve for these 20 results. The results of the two IFS analyses can also be seen in Supplementary Data 3. The classifier based on the feature subset with an index of 158 (i.e. the feature subset containing the first 159 features in the ordered feature set *S)*, obtained the highest overall accuracy of 0.9490. Table 1 shows the 159 features selected here with their biological classes. The class of each AAIndex feature can be downloaded from http://www.genome.jp/aaindex/AAindex/Appendix (or see Supplementary Datas 4); 402 features out of the total 506 AAIndex features are clustered into 6 groups, while the remaining 104 features have not been defined.

DISCUSSION

Protein-DNA interactions are central for the regulation of gene expression. Since DNA-binding proteins probably comprise only a small fraction of structural genomics targets, for practical applications it is necessary to develop a method with high precision. To achieve high accuracy, we propose a novel protein sequence-based method for predicting DNA-binding proteins and associated each amino acid residue preferences from structural genomics targets.

We define the protein binding motif of 9 amino acids, with a central direct binding site. The results mirror the fact that amino acid directly binding to DNA motif or near the direct binding site attribute more sensitivity and specificity. Our results are consis-

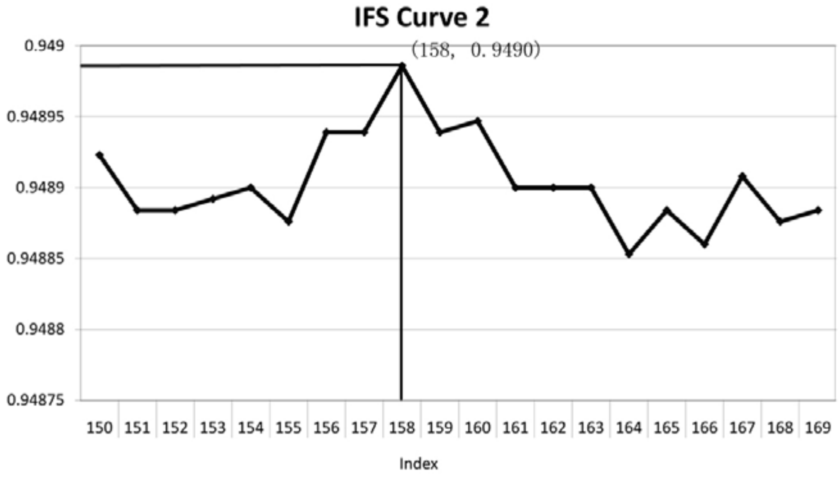

Fig. 2. The second IFS curve. The IFS curve for the 20 results of classifiers based on the feature subsets with indexes from 150 to 170.

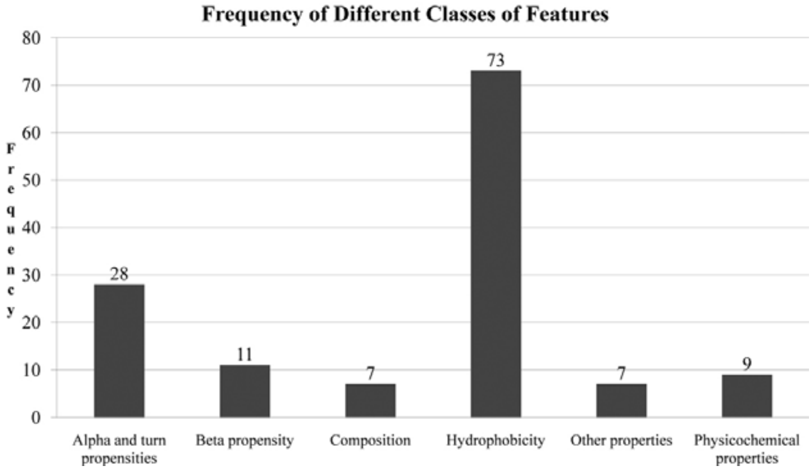

Fig. 3. Classification of features. The frequency of different classes of features selected by IFS procedure.

tent with other sequence-based models (Jones and Thornton, 2004; Luscombe et al., 2000) and achieve 94.90% overall accuracy. The statistics of individual conserved residues and their contributions to the stability of protein-DNA complexes is dependent on the distance to the direct binding amino acid. The complex relationship between protein folding complexes and functions highlights the necessity of looking beyond the global fold of a protein to specific functional sites (Jones and Thornton, 2004). Figure 3 shows the frequencies of different classes of features according to their biological meanings. Compared with all parameters, hydrophobicity is essential to protein-DNA binding activity. The secondary structure-related parameters, such as the alpha and turn propensities and beta propensity, are also irreplaceable in binding activity. Hydrophilic residues such as Asn and Ser, exhibit a preference for binding activity within a helical conformation, which may suggest that they are able to make better contact with the DNA helix when they are in that conformation. This greater tendency to bind when in a helical conformation is shared, to some extent, by Cys, His, and Pro (Ahmad et al., 2004). Two Cysteine residues related to the hydrophobicity and secondary structure in paired domains regulate the DNA binding activity of Pax-8 and provide a new insight into molecular basis for modulation of Pax function (Cao et al., 2005). Conserved residues in the H2 helix and L1 and L3 loops of p53 as novel functional domains contribute to transcription-independent apoptosis by this tumor suppressor protein (Pietsch et al., 2008). For individually mutated, evolutionarily conserved, basic- and hydroxyl-group-containing ily conserved, basic- and hydroxyl-group-containing residues within RAG2, findings support the direct involvement of RAG2 in DNA binding during all steps of V(D)J recombination (Fugmann and Schatz, 2001). In conclusion, binding residues show evidenced overall preference for hydrophobicity and secondary structure due to the protein-DNA interactions.

On the whole, there was a preference for binding residues to occur in hydrophobicity and secondary structure, although there were a number of interesting exceptions to this generalization. However, our method is only a machine learning approach, not one based on first principles. Therefore, in it will be necessary in the future to find more useful bio-chemical and physicochemical features related to DNA-binding to improve the predicting accuracy.

CONCLUSION

Protein-DNA interactions control a variety of vital biological processes, and it is necessary to develop a credible method to predict DNA binding sites in a protein. Here, we have developed a novel prediction approach, using machine learning method, based solely on the sequence of proteins. We obtained an overall accuracy of 94.90% in Jackknife cross-validation test. This result shows the physicochemical and biological properties of environment sequences are important for DNA binding site residues prediction. It may also indicate some mechanisms of protein-DNA interactions.

Table 1. The 159 features and their biological classes selected by IFS procedure

mRMR Order	Feature ID	Position	Feature class
1	2154	5	Hydrophobicity
2	1883	4	Hydrophobicity
3	3130	7	Hydrophobicity
4	2918	6	Hydrophobicity
5	1272	3	Alpha and turn propensities
6	622	2	Composition
7	3868	8	Hydrophobicity
8	494	1	Not defined
9	2351	5	Hydrophobicity
10	1386	3	Other
11	2157	5	Hydrophobicity
12	1843	4	Hydrophobicity
13	2029	5	Alpha and turn propensities
14	2812	6	Physicochemistry properties
15	3296	7	Alpha and turn propensities
16	1987	4	Not defined
17	1377	3	Hydrophobicity
18	2497	5	Not defined
19	3583	8	Hydrophobicity
20	2373	5	Alpha and turn propensities
21	2790	6	Alpha and turn propensities
22	2420	5	Hydrophobicity
23	549	2	Other
24	1561	4	Other
25	3139	7	Beta propensity
26	2027	5	Hydrophobicity
27	2870	6	Hydrophobicity
28	3125	7	Hydrophobicity
29	2446	5	Not defined
30	1141	3	Hydrophobicity
31	1905	4	Hydrophobicity
32	3976	8	Not defined
33	2110	5	Hydrophobicity
34	2298	5	Alpha and turn propensities
35	2939	6	Not defined
36	1613	4	Hydrophobicity
37	3499	7	Not defined
38	707	2	Hydrophobicity
39	2113	5	Hydrophobicity
40	2999	6	Not defined
41	1111	3	Alpha and turn propensities
42	2224	5	Composition
43	1607	4	Hydrophobicity
44	3446	7	Not defined
45	2308	5	Hydrophobicity
46	2419	5	Hydrophobicity
47	1681	4	Alpha and turn propensities
48	3943	8	Hydrophobicity
49	2492	5	Not defined
50	2801	6	Hydrophobicity
51	1492	3	Not defined
52	3445	7	Not defined
53	1858	4	Hydrophobicity
54	2169	5	Hydrophobicity
55	2619	6	Hydrophobicity
56	2451	5	Not defined
57	880	2	Other
58	2168	5	Composition
59	2804	6	Alpha and turn propensities
60	3376	7	Hydrophobicity
61	1251	3	Hydrophobicity
62	2236	5	Hydrophobicity
63	1800	4	Physicochemistry properties
64	2055	5	Composition
65	3587	8	Beta propensity
66	1792	4	Alpha and turn propensities
67	2811	6	Physicochemistry properties
68	2427	5	Not defined
69	1274	3	Alpha and turn propensities
70	2094	5	Hydrophobicity
71	3082	7	Beta propensity
72	1998	4	Not defined
73	2327	5	Alpha and turn propensities
74	2364	5	Hydrophobicity
75	2575	6	Beta propensity
76	2223	5	Composition
77	1921	4	Not defined
78	3362	7	Hydrophobicity
79	3930	8	Hydrophobicity
80	1845	4	Hydrophobicity
81	2760	6	Alpha and turn propensities
82	2359	5	Alpha and turn propensities
83	1352	3	Hydrophobicity
84	2571	6	Hydrophobicity
85	2355	5	Alpha and turn propensities
86	3188	7	Hydrophobicity
87	1875	4	Hydrophobicity
88	2112	5	Hydrophobicity
89	2241	5	Physicochemistry properties
90	2433	5	Not defined
91	3714	8	Beta propensity
92	1617	4	Alpha and turn propensities
93	3215	7	Hydrophobicity
94	2295	5	Hydrophobicity

(continued)

mRMR Order	Feature ID	Position	Feature class
95	1107	3	Hydrophobicity
96	2857	6	Hydrophobicity
97	2398	5	Other
98	1647	4	Hydrophobicity
99	1338	3	Hydrophobicity
100	2065	5	Hydrophobicity
101	2803	6	Hydrophobicity
102	2052	5	Physicochemistry properties
103	3470	7	Not defined
104	1559	4	Hydrophobicity
105	3266	7	Alpha and turn propensities
106	2049	5	Alpha and turn propensities
107	1778	4	Alpha and turn propensities
108	3869	8	Hydrophobicity
109	2346	5	Other
110	2964	6	Not defined
111	1158	3	Hydrophobicity
112	2099	5	Alpha and turn propensities
113	1981	4	Not defined
114	2230	5	Composition
115	2993	6	Not defined
116	3093	7	Hydrophobicity
117	2246	5	Hydrophobicity
118	1928	4	Not defined
119	3636	8	Hydrophobicity
120	2197	5	Other
121	2410	5	Hydrophobicity
122	3077	7	Hydrophobicity
123	2694	6	Beta propensity
124	1952	4	Not defined
125	2407	5	Physicochemistry properties
126	2202	5	Hydrophobicity
127	2709	6	Hydrophobicity
128	2146	5	Alpha and turn propensities
129	3363	7	Hydrophobicity
130	1576	4	Hydrophobicity
131	2225	5	Hydrophobicity
132	2268	5	Hydrophobicity
133	1864	4	Alpha and turn propensities
134	3142	7	Beta propensity
135	2682	6	Hydrophobicity
136	2191	5	Beta propensity
137	2221	5	Composition
138	2066	5	Alpha and turn propensities
139	1804	4	Alpha and turn propensities
140	3262	7	Beta propensity
141	2587	6	Hydrophobicity
142	2424	5	Hydrophobicity
143	3397	7	Physicochemistry properties
144	1697	4	Hydrophobicity
145	2047	5	Hydrophobicity
146	2856	6	Hydrophobicity
147	2283	5	Alpha and turn propensities
148	2404	5	Hydrophobicity
149	1743	4	Beta propensity
150	3437	7	Hydrophobicity
151	2891	6	Physicochemistry properties
152	2362	5	Alpha and turn propensities
153	1748	4	Alpha and turn propensities
154	2244	5	Hydrophobicity
155	2432	5	Not defined
156	2932	6	Hydrophobicity
157	1879	4	Physicochemistry properties
158	2320	5	Alpha and turn propensities
159	3177	7	Beta propensity

Note: Supplementary information is available on the Molecules and Cells website (www.molcells.org).

ACKNOWLEDGMENTS

The study is supported by National Basic Research Program of China (2004CB518603), Funding of CAS: KSCX2-YW-R-112. We acknowledge Yvonne Poindexter from the Vanderbilt Cancer Biostatistics Center for her editing.

REFERENCES

Ahmad, S., Gromiha, M.M., and Sarai, A. (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., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389-3402.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The protein data bank. Nucleic Acids Res. *28*, 235-242.

Bullock, A.N., and Fersht, A.R. (2001). Rescuing the function of mutant p53. Nat. Rev. Cancer *1*, 68-76.

Cai, Y., He, J., Li, X., Lu, L., Yang, X., Feng, K., Lu, W., and Kong, X. (2009). A novel computational approach to predict transcription factor DNA binding preference. J. Proteome Res. *8*, 999-1003.

Cao, X., Kambe, F., Lu, X., Kobayashi, N., Ohmori, S., and Seo, H. (2005). Glutathionylation of two cysteine residues in paired domain regulates DNA binding activity of Pax-8. J. Biol. Chem. *280*, 25901-25906.

Fugmann, S.D., and Schatz, D.G. (2001). Identification of basic residues in RAG2 critical for DNA binding by the RAG1-RAG2 complex. Mol. Cell *8*, 899-910.

Gao, M., and Skolnick, J. (2008). DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions. Nucleic Acids Res. *36*, 3978-3992.

Gromiha, M.M., Siebers, J.G., Selvaraj, S., Kono, H., and Sarai, A. (2005). Role of inter and intramolecular interactions in protein-DNA recognition. Gene *364*, 108-113.

Ho, S.Y., Yu, F.C., Chang, C.Y., and Huang, H.L. (2007). Design of accurate predictors for DNA-binding sites in proteins using hybrid SVM-PSSM method. Biosystems *90*, 234-241.

Horton, P., Park, K.J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J., and Nakai, K. (2007). WoLF PSORT: protein localization predictor. Nucleic Acids Res. *35*, W585-587.

Hwang, S., Gou, Z., and Kuznetsov, I.B. (2007). DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. Bioinformatics *23*, 634-636.

Jamal Rahi, S., Virnau, P., Mirny, L.A., and Kardar, M. (2008). Predicting transcription factor specificity with all-atom models. Nucleic Acids Res. *36*, 6209-6217.

Jones, S., and Thornton, J.M. (2004). Searching for functional sites in protein structures. Curr. Opin. Chem. Biol. *8*, 3-7.

Kaplan, T., Friedman, N., and Margalit, H. (2005). Ab initio prediction of transcription factor targets using structural knowledge. PLoS Comput. Biol. *1*, e1.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947-2948.

Luscombe, N.M., Austin, S.E., Berman, H.M., and Thornton, J.M. (2000). An overview of the structures of protein-DNA complexes. Genome Biol. *1*, REVIEWS001.

Noyes, M.B., Christensen, R.G., Wakabayashi, A., Stormo, G.D., Brodsky, M.H., and Wolfe, S.A. (2008). Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell *133*, 1277-1289.

Ofran, Y., Mysore, V., and Rost, B. (2007). Prediction of DNA-binding residues from sequence. Bioinformatics *23*, i347-353.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence *27*, 1226-1238.

Pietsch, E.C., Perchiniak, E., Canutescu, A.A., Wang, G., Dunbrack, R.L., and Murphy, M.E. (2008). Oligomerization of BAK by p53 utilizes conserved residues of the p53 DNA binding domain. J. Biol. Chem. *283*, 21294-21304.

Qian, Z., Cai, Y.D., and Li, Y. (2006). A novel computational method to predict transcription factor DNA binding preference. Biochem. Biophys. Res. Commun. *348*, 1034-1037.

Salamov, A.A., and Solovyev, V.V. (1997). Protein secondary structure prediction using local alignments. J. Mol. Biol. *268*, 31-36.

Sim, J., Kim, S.Y., and Lee, J. (2005). Prediction of protein solvent accessibility using fuzzy k-nearest neighbor method. Bioinformatics *21*, 2844-2849.

Sinha, S., van Nimwegen, E., and Siggia, E.D. (2003). A probabilistic method to detect regulatory modules. Bioinformatics *19*, i292-301.

Tan, K., McCue, L.A., and Stormo, G.D. (2005). Making connections between novel transcription factors and their DNA motifs. Genome Res. *15*, 312-320.

Valdar, W.S. (2002). Scoring residue conservation. Proteins *48*, 227-241.

Vavouri, T., and Elgar, G. (2005). Prediction of cis-regulatory elements using binding site matrices--the successes, the failures and the reasons for both. Curr. Opin. Genet. Dev. *15*, 395-402.

Wang, L., and Brown, S.J. (2006). Prediction of DNA-binding residues from sequence features. J. Bioinform Comput. Biol. *4*, 1141-1158.

Warner, J.B., Philippakis, A.A., Jaeger, S.A., He, F.S., Lin, J., and Bulyk, M.L. (2008). Systematic identification of mammalian regulatory motifs' target genes and functions. Nat. Methods *5*, 347-353.

Whitington, T., Perkins, A.C., and Bailey, T.L. (2009). High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. Nucleic Acids Res. *37*, 14-25.

Wong, W.S., and Nielsen, R. (2007). Finding cis-regulatory modules in Drosophila using phylogenetic hidden Markov models. Bioinformatics *23*, 2031-2037.

Wu, J., Liu, H., Duan, X., Ding, Y., Wu, H., Bai, Y., and Sun, X. (2009). Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. Bioinformatics *25*, 30-35.