W396–W401 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* *Published online 29 May 2009*
doi:10.1093/nar/gkp449

ProteDNA: a sequence-based predictor of sequence-specific DNA-binding residues in transcription factors

Wen-Yi Chu[1], Yu-Feng Huang[1], Chun-Chin Huang[2], Yi-Sheng Cheng[3], Chien-Kang Huang[2,*] and Yen-Jen Oyang[1,4,5,*]

[1]Department of Computer Science and Information Engineering, [2]Department of Engineering Science and Ocean Engineering, [3]Department of Life Science, [4]Graduate Institute of Biomedical Electronics and Bioinformatics, and [5]Center for Systems Biology and Bioinformatics, National Taiwan University, Taipei, Taiwan, ROC

Received March 4, 2009; Revised May 11, 2009; Accepted May 12, 2009

ABSTRACT

This article presents the design of a sequence-based predictor named ProteDNA for identifying the sequence-specific binding residues in a transcription factor (TF). Concerning protein–DNA interactions, there are two types of binding mechanisms involved, namely sequence-specific binding and nonspecific binding. Sequence-specific bindings occur between protein sidechains and nucleotide bases and correspond to sequence-specific recognition of genes. Therefore, sequence-specific bindings are essential for correct gene regulation. In this respect, ProteDNA is distinctive since it has been designed to identify sequence-specific binding residues. In order to accommodate users with different application needs, ProteDNA has been designed to operate under two modes, namely, the *high-precision* mode and the *balanced* mode. According to the experiments reported in this article, under the *high-precision* mode, ProteDNA has been able to deliver precision of 82.3%, specificity of 99.3%, sensitivity of 49.8% and accuracy of 96.5%. Meanwhile, under the *balanced* mode, ProteDNA has been able to deliver precision of 60.8%, specificity of 97.6%, sensitivity of 60.7% and accuracy of 95.4%. ProteDNA is available at the following websites: http://protedna.csbb.ntu.edu.tw/ http://protedna.csie.ntu.edu.tw/ http://bio222.esoe.ntu.edu.tw/ProteDNA/.

INTRODUCTION

In recent years, prediction of residues in a protein chain that may be involved in interaction with the DNA has been a research topic that attracts a high level of interest. Some of the studies were purely based on analysis of the polypeptide sequence (1–5), while the others took the structural information into account (3,6). In this respect, as it has been reported in a recent article that the tertiary structures of a large number of transcription factors (TFs) are mostly disordered (7), sequence-based analysis aimed at identifying the residues in a highly disordered TF that play key roles in interaction with the DNA is essential for obtaining a comprehensive picture of how the TF functions.

Concerning protein–DNA interactions, there are two types of binding mechanisms involved, namely sequence-specific binding and nonspecific binding (8). Sequence-specific bindings occur between protein sidechains and nucleotide bases, while nonspecific bindings occur between protein sidechains and the DNA sugar/phosphate backbone. In molecular biology, sequence-specific bindings correspond to sequence-specific recognition of genes and therefore are essential for correct gene regulation.

This article presents the design of a sequence based predictor named ProteDNA for identifying the residues in a TF that are involved in sequence-specific binding with the DNA. In this article, a residue is regarded as involved in sequence-specific binding with the DNA, if one or more heavy atoms in its sidechain fall within 4.5 Å from the nucleobases of the DNA. Figure 1 illustrates the function carried out by ProteDNA. Figure 1(a) shows the prediction output of ProteDNA for the polypeptide sequence of Yeast TF GCN4 in the complex with Protein Data Bank (PDB) (9) ID 1YSA. Figure 1(b) depicts the output of ProteDNA in the tertiary structure of PDB complex 1YSA. In Figure 1(b), the residues colored by red are those sequence-specific binding residues correctly identified by ProteDNA, while the residue colored by blue is a false negative. In this case, there is no

*To whom correspondence should be addressed. Tel: +886 2 3366 5736; Fax: +886 2 2392 9885; Email: ckhuang@ntu.edu.tw
Correspondence may also be addressed to Dr Yen-Jen Oyang. Tel: +886 2 3366 4888; Email: yjoyang@csie.ntu.edu.tw

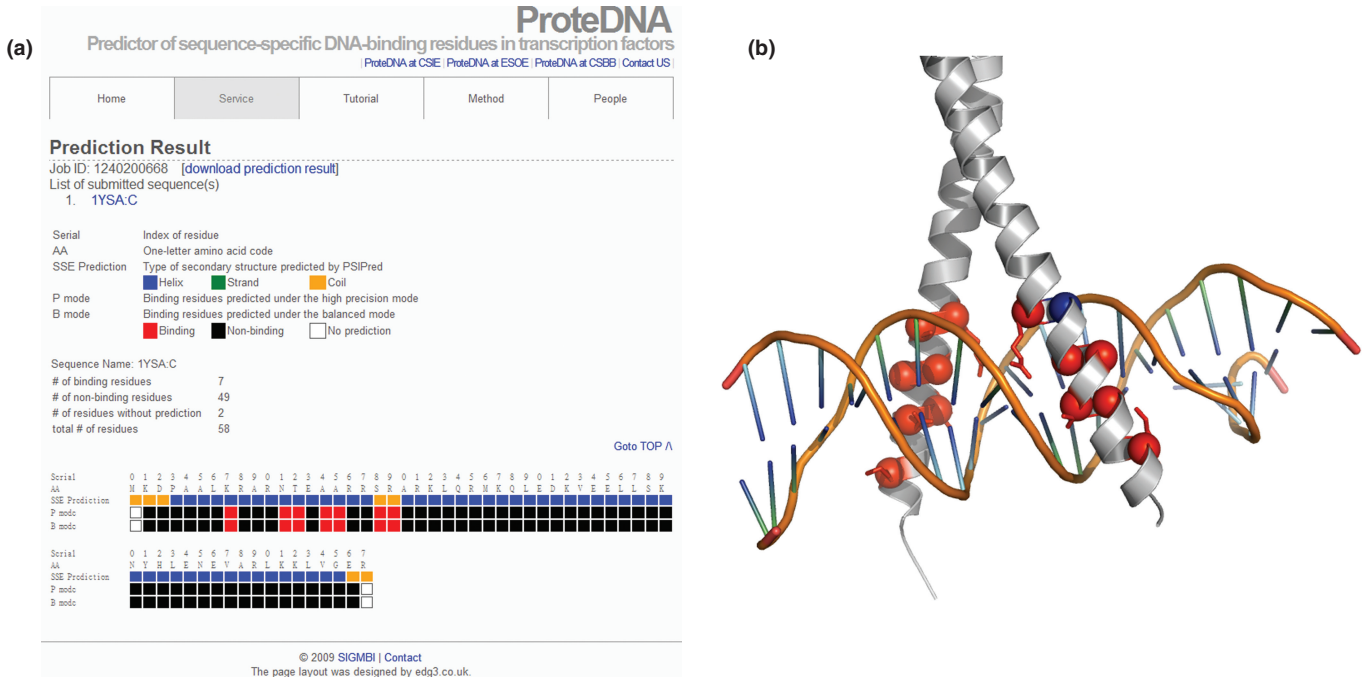

Figure 1. Illustration of the function of ProteDNA. (**a**) The partial prediction output of ProteDNA with the polypeptide sequence of Yeast TF GCN4 in PDB complex 1YSA. (**b**) The tertiary structure of the complex with PDB ID 1YSA. The residues colored by red are those sequence-specific binding residues correctly identified by ProteDNA, while the residues colored by blue are the false negatives. In this case, there is no false positive.

false positive. However, this case contains a residue for which ProteDNA makes no prediction. The reason that causes ProteDNA providing no prediction in some cases is that some TF–DNA complexes deposited in the PDB contain disordered regions and therefore ProteDNA cannot learn any clues in order to make predictions for residues located in a similar polypeptide segment.

In this article, the performance of ProteDNA is reported based on the following metrics:

$$\text{precision} = \frac{\text{TP}}{\text{TP}+\text{FP}},\ \text{sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}},$$
$$\text{specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}},\ \text{accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}.$$

where TP, TN, FP and FN stand for the number of true positive samples, the number of true negative samples, the number of false positive samples and the number of false negative samples, respectively. In order to accommodate users with different application needs, ProteDNA has been designed to operate under two modes, namely, the *high-precision* mode and the *balanced* mode. In this respect, the user can select either mode when submitting a query to the web server. The experiments reported in this article show that under the *high-precision* mode, ProteDNA delivers precision of 82.3%, specificity of 99.3%, sensitivity of 49.8% and accuracy of 96.5%. Meanwhile, under the *balanced* mode, ProteDNA delivers precision of 60.8%, specificity of 97.6%, sensitivity of 60.7% and accuracy of 95.4%.

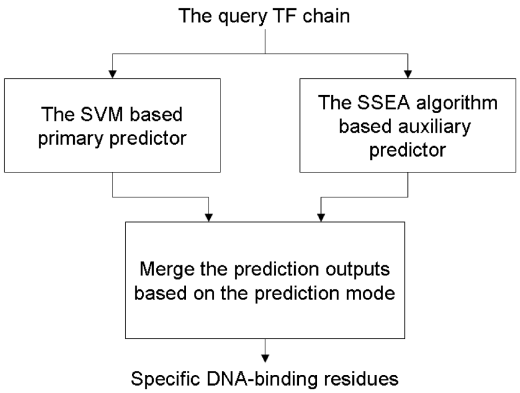

Figure 2. Overview of the architecture of ProteDNA.

METHODS

Overview

Figure 2 presents an overview of the architecture of ProteDNA. The entire hybrid predictor consists of the primary predictor and the auxiliary predictor. The primary predictor is a support vector machine (SVM) with its parameter settings optimized for delivering high precision. As a result, one can expect that sensitivity of the SVM-based primary predictor has been traded, since tuning the parameters of a predictor aimed at raising precision typically means that sensitivity is traded and *vice versa*. Accordingly, as shown in Figure 2, in the design of ProteDNA, we have incorporated a mechanism derived from the secondary structure element alignment (SSEA)

approach first proposed by Gewehr and Zimmer (10) to complement the prediction power of the SVM. With the primary and auxiliary predictors, ProteDNA can operate under the *high-precision* mode as well as the *balanced* mode in order to accommodate users with different application needs. Under the *high-precision* mode, only the SVM-based primary predictor is enabled. On the other hand, under the *balanced* mode, both predictors are enabled and a residue is predicted to be involved in specific binding with the DNA if either the primary or the secondary predictor makes such a prediction.

For evaluating the performance of ProteDNA, we have created a data set containing 253 TF–DNA complexes, among which 227 complexes were extracted from the 691 protein–DNA complexes that Ofran *et al.* (11) collected from the PDB and the remaining 26 TF–DNA complexes are those that were deposited into the PDB during September 2007 and November 2008. During the process to extract the 227 complexes from the Ofran collection, we excluded those complexes that do not contain a TF and then queried the PFAM server (12) to exclude those complexes in which no polypeptide segment is within the DNA-binding domain predicted by the PFAM server. In this respect, we submitted the full sequences of the proteins in the complex to the PFAM server and adopted only those predicted binding domains with the *P*-value computed by the PFAM server <0.01. With this process, we excluded those complexes in which the polypeptide segments just happen to be in the proximity of the DNA but are not really involved in binding with the DNA. It might happen that we accidently excluded some TF–DNA complexes with actual TF–DNA interactions. Nevertheless, it was our intention to be conservative. In the end, 227 out of the 691 complexes initially in Ofran collection were extracted.

Design of the primary predictor

For the design of the primary predictor, we have employed the LIBSVM package with the Gaussian kernel (software available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm). The model of the SVM has been generated based on the training data set derived by associating each residue in the 253 protein chains with the evolutionary profiles of the residue and its 10 neighboring residues. The evolutionary profile of a residue is in fact the vector corresponding to the residue in the position specific scoring matrix (PSSM) computed by the PSI-BLAST package with three iterations (5). In addition, each residue was labeled based on whether it is involved in sequence-specific binding with the DNA or not. As mentioned earlier, a residue is regarded as involved in sequence-specific binding with the DNA, if one or more heavy atoms in its sidechain are within 4.5 Å from the nucleobases of the DNA.

As mentioned earlier, the parameters of the SVM have been set to deliver high precision. In this respect, we have set parameter g with the Gaussian kernel to 0.03125 and have set costs for the positive and negative classes to six and four, respectively.

Design of the auxiliary predictor

As mentioned earlier, the auxiliary predictor incorporates a mechanism derived from the secondary structure element alignment (SSEA) approach first proposed by Gewehr and Zimmer (10). The SSEA-based mechanism refers to a template library containing the polypeptide segments of sequence-specific DNA-binding domains. The template library has been created with the following steps:

(i) Each protein chain in the 253 TF–DNA complexes was fed into the PSIPRED predictor of protein secondary structures (13) as well as into the PFAM server (12). Then, each residue in the predicted secondary structure elements was examined to determine whether it is involved in sequence-specific binding with the DNA. If a secondary structure element contains one or more residues involved in sequence-specific binding with the DNA, then the element was regarded as involved in sequence-specific binding with the DNA.
(ii) If a DNA-binding domain output by the PFAM server contained one or more secondary structure elements involved in sequence-specific binding with DNA, then the binding domain was deposited into the template library. In addition, based on the predictions made by PSIPRED, each residue in the domain was labeled as belonging to one of the following three types of elements: α-helix, β-sheet and coil.

With the template library, we then can invoke the following procedure to predict the sequence-specific DNA-binding residues in the query TF, which is a slightly modified version of the original secondary structure element alignment (SSEA) algorithm (10).

(i) Invoke PSIPRED to label each residue in the query TF with one of following three types: α-helix, β-sheet and coil. Then, the BLAST package (14) is invoked to align the sequence of PSIPRED labels of each template in the library with the sequence of PSIPRED labels of the query TF.
(ii) The alignment score between the query protein chain and a template in the library is computed based on the following log average score for profile–profile alignment proposed by von Öhsen and Zimmer (15):

$$\log \sum_{i=1}^{L} \sum_{j=1}^{20} \sum_{k=1}^{20} \alpha_i(j)\beta_i(k) \exp\{\lambda S(j,k)\},$$

where

(a) i is the index of the aligned residue pair and L is the length of the template;
(b) $\alpha_i(\bullet)$ and $\beta_i(\bullet)$ are the two PSSM vectors corresponding to the aligned residue pair with index i;
(c) $S(j,k)$ is the score of BLOSUM62 (16) corresponding to residue types j and k;

(d) λ has been set to 0.347, which is the default value of BLOSUM62.

Under the *balanced* mode, those residues in the query TF chain that are aligned with a sequence-specific DNA-binding residue in the template that yields the highest alignment score are predicted to be involved in the sequence-specific binding with DNA.

WEB SERVICE

Figure 3 shows the webpage for submitting a job to ProteDNA. The user simply needs to provide a polypeptide sequence in the FASTA format and an email address for receiving the output. The output format of ProteDNA has been illustrated in Figure 1(a).

RESULTS AND DISCUSSION

In this section, we will report the experiments conducted to evaluate the performance of ProteDNA. In the experiments, we repeated the same testing procedure 20 times with randomly and independently generated testing data sets. The independent testing data set used in each run was derived from 30 TF chains randomly selected from the 253 TF–DNA complexes that we have collected. In order to eliminate possible bias presence in our collection of TF complexes, we took steps to guarantee that no two TF chains used to generate the testing data set in the same run are homologous with a sequence identity higher than 20%. Furthermore, aiming to obtain experimental results that accurately reflect the actual performance observed by the users of ProteDNA, we guaranteed that the training data generated with a TF chain that is homologous to the protein chain under testing by having a sequence identity higher than 20% are removed.

Figure 3. The webpage for submitting a job.

Table 1 shows the overall performance of ProteDNA with 20 independent runs and Table 2 shows a breakdown of the experimental results based on the classification of TF–DNA interactions proposed by Thornton *et al.* (17). In calculating the numbers presented in Tables 1 and 2, the residues for which ProteDNA made no prediction were treated as if ProteDNA had labeled them as non-binding residues. However, the counts of residues listed in Tables 1 and 2 do not include those residues that are located in the disordered regions of the tertiary structures of the testing protein chains. In the Supplementary Data, we have included an additional table in which the residues located in the disordered regions are treated as negative samples. Performance data reported in Tables 1 and 2 and Supplementary Table S1 show how we treat the residues in the disordered regions really does not introduce any material difference.

One interesting observation about the numbers presented in Table 1 is that ProteDNA failed to effectively identify the sequence-specific binding residues in β-sheet secondary structure elements. Our conjecture about this phenomenon is that the number of sequence-specific binding residues in β-sheet secondary structure elements is far fewer than the number of sequence-specific binding residues in either α-helix or coil elements. As a result, ProteDNA cannot learn sufficient clues in order to identify sequence-specific binidng residues in β-sheet elements. Accordingly, as shown in Table 2, ProteDNA failed to effectively identify the sequence-specific binding residues involved in the β-hairpin/ribbon type of binding with

Table 1. Overall performance of ProteDNA

Type of the secondary structure element	No. of residues tested	Prediction results							
		TP	TN	FP	FN	Precision	Sensitivity	Specificity	Accuracy
Performance under the *high-precision* mode									
Helix	33 769	1397	30 916	320	1136	0.814	0.552	0.990	0.957
Sheet	5396	0	5239	0	157	NA	0.000	1.000	0.971
Coil	21 286	355	20 401	57	473	0.862	0.429	0.997	0.975
Overall	60 451	1752	56 556	377	1766	0.823	0.498	0.993	0.965
Performance under the *balanced* mode.									
Helix	33 769	1679	30 299	937	854	0.642	0.663	0.970	0.947
Sheet	5396	39	5208	31	118	0.557	0.248	0.994	0.972
Coil	21 286	417	20 052	406	411	0.507	0.504	0.980	0.962
Overall	60 451	2135	55 559	1374	1383	0.608	0.607	0.976	0.954

Table 2. Breakdown of the experimental results with ProteDNA in respect of different types of TF-DNA bindings

Type of TF-DNA bindings	No. of TFs involved	No. of residues tested	Prediction results							
			TP	TN	FP	FN	Precision	Sensitivity	Specificity	Accuracy
Performance under the *high-precision* mode.										
Zipper-type	146	9587	586	8667	128	206	0.821	0.740	0.985	0.965
Helix-turn-helix	220	27 063	510	25 455	149	949	0.774	0.350	0.994	0.959
Zinc-coordinating	152	12 105	598	11 098	86	323	0.874	0.649	0.992	0.966
β-hairpin/ribbon	38	2618	0	2488	0	130	NA	0.000	1.000	0.950
Others	44	9078	58	8848	14	158	0.806	0.269	0.998	0.981
Overall	600	60 451	1752	56 556	377	1766	0.823	0.498	0.993	0.965
Performance under the *balanced* mode										
Zipper-type	146	9587	643	8496	299	149	0.683	0.812	0.966	0.953
Helix-turn-helix	220	27 063	769	24 994	610	690	0.558	0.527	0.976	0.952
Zinc-coordinating	152	12 105	610	10 925	259	311	0.702	0.662	0.977	0.953
β-hairpin/ribbon	38	2618	39	2365	123	91	0.241	0.300	0.951	0.918
Others	44	9078	74	8778	84	142	0.468	0.343	0.991	0.975
Overall	600	60 451	2135	55 558	1375	1383	0.608	0.607	0.976	0.954

Table 3. Performance delivered by alternative predictors of DNA-binding residues, where the *F*-score is the harmonic mean of precision and sensitivity

Predictor	Sensitivity	Specificity	Accuracy	Precision	*F*-score
ProteDNA under the *high-precision* mode	0.498	0.993	0.965	0.823	0.621
ProteDNA under the *balanced* mode	0.607	0.976	0.954	0.608	0.607
Ahmad and Sarai (1)	0.682	0.660	0.664	0.308*	0.425*
Yan and *et al.* (2)	0.410	0.871	0.780	0.439*	0.424*
BindN (18)	0.652	0.728	0.722	0.186*	0.289*
DP-Bind (19)	0.791	0.786	0.800	–*	–*

The numbers with an asterisk are those that have been derived from the numbers reported in the related studies.

DNA, since the interaction region contains a β-sheet element.

In the following, we will discuss how the ProteDNA performs in comparison with the related studies reported in recent years. In this respect, one must note ProteDNA is the only predictor listed in Table 3 that has been designed to identify the residues involved in sequence-specific binding with the DNA, while all the other predictors do not distinguish between sequence-specific binding and nonspecific binding. Therefore, the results listed in Table 3, which includes the main results extracted from the related studies along with the overall results with the ProteDNA, should be regarded as a survey of the latest advances in the field. It must also be noted that most related studies have adopted slightly different definitions of DNA-binding residues. In the article by Ahmad and Sarai (1) and in the article by Wang and Brown (18), a residue is regarded as involved in interaction with the DNA, if one of its heavy atom is within 3.5 Å from a heavy atom of the DNA. In the article by Hwang *et al.* (19), a larger threshold of 4.5 Å, instead of 3.5 Å, has been adopted. In the article by Yan *et al.* (2), a residue is regarded as involved in interaction with the DNA, if its solvent accessible surface area (ASA) in the protein–DNA complex is less than its ASA in the unbound protein by more than 1 $Å^2$.

The numbers listed in Table 3 with an asterisk have been derived from the numbers reported in the related studies. Since all the four related studies addressed in Table 3 reported three out of the four performance metrics listed in the table, for each of the related study, we can obtain three equations about the following four variables:

$$\hat{\mathrm{TP}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}, \hat{\mathrm{FP}} = \frac{\mathrm{FP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}},$$

$$\hat{\mathrm{TN}} = \frac{\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}, \hat{\mathrm{FN}} = \frac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}.$$

In addition, we have $\hat{\mathrm{TP}}+\hat{\mathrm{FP}}+\hat{\mathrm{TN}}+\hat{\mathrm{FN}}=1$. Therefore, for each related study, we can derive the actual value of the fourth performance metric based on the values of the other three performance metrics that were provided. The only exception is precision for the predictor proposed by Hwang *et al.* (19). It can be easily shown in mathematics that accuracy cannot be higher than sensitivity and specificity simultaneously, which is the case with the numbers reported by Hwang *et al.* Therefore, there is no way to derive the exact value of precision for their predictor.

In our view, the general observation concerning the numbers presented in Table 3 is that ProteDNA is capable of delivering superior performance in comparison with the related works. In particular, in terms of the *F*-score, which is the harmonic mean of precision and sensitivity and is widely used for reporting the overall performance of a predictor in machine learning research, ProteDNA can

deliver significantly superior performance, regardless of the mode under which it is operating.

CONCLUSIONS

This article presents the design of a sequence-based predictor aiming to identify the sequence-specific DNA-binding residues in a TF. As a recent study has revealed that the tertiary structures of a large number of transcription factors are mostly disordered, a sequence-based predictor is essential for analyzing how a TF interacts with the DNA. Furthermore, it is highly desirable to have a predictor capable of identifying those residues involved in sequence-specific binding with the DNA, since sequence-specific binding corresponds to sequence-specific recognition of a gene and therefore is essential for correct gene regulation.

In the experiments reported in this article, ProteDNA has been able to deliver precision as high as 82.3%, when operating under the *high-precision* mode. Precision of 82.3% implies that about four out of five predicted binding residues are really involved in sequence-specific binding with the DNA. On the other hand, when operating under the *balanced* mode, ProteDNA has been able to deliver sensitivity as 60.7%. Sensitivity of 60.7% implies that ProteDNA can catch about 6 out of 10 residues involved in sequence-specific binding with the DNA.

It is anticipated the prediction accuracy delivered by ProteDNA will continue to improve as the number of TF–DNA complexes deposited in the PDB continues to grow and the number of training samples that can be exploited continues to increase accordingly. Nevertheless, it is computational biologists' primary interest to develop more advanced prediction mechanisms. In this respect, we believe that, as the number of TF–DNA complexes deposited in the PDB increases, we can obtain more insights about the key physiochemical properties that play essential roles in TF–DNA interactions and then we will be able to develop more advanced prediction mechanisms accordingly. In addition, we will exploit the experiences learned in this study in order to design specific predictors for other families of proteins interacting with DNA. We believe that different families of proteins may have very different characteristics. Therefore, concerning a specific type of proteins, a specifically designed predictor should be able to deliver superior performance in comparison with a general-purpose predictor.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Science Council and National Taiwan University. Funding for open access charge: National Science Council, NSC 97-2627-P-001-002.

Conflict of interest statement. None declared.

REFERENCES

1. Ahmad,S. and Sarai,A. (2005) PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics*, **6**, 33.
2. Yan,C., Terribilini,M., Wu,F., Jernigan,R.L., Dobbs,D. and Honavar,V. (2006) Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinformatics*, **7**, 262.
3. Jones,S., Shanahan,H.P., Berman,H.M. and Thornton,J.M. (2003) Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Res.*, **31**, 7189–7198.
4. Ferrer-Costa,C., Shanahan,H.P., Jones,S. and Thornton,J.M. (2005) HTHquery: a method for detecting DNA-binding proteins with a helix-turn-helix structural motif. *Bioinformatics*, **21**, 3679–3680.
5. Tjong,H. and Zhou,H.X. (2007) DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. *Nucleic Acids Res.*, **35**, 1465–1477.
6. Tsuchiya,Y., Kinoshita,K. and Nakamura,H. (2004) Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. *Proteins*, **55**, 885–894.
7. Liu,J., Perumal,N.B., Oldfield,C.J., Su,E.W., Uversky,V.N. and Dunker,A.K. (2006) Intrinsic disorder in transcription factors. *Biochemistry*, **45**, 6873–6888.
8. Boyer,R.F. (2005) *Concepts in Biochemistry*. 3rd ed. Wiley, Hoboken, NJ.
9. 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.
10. Gewehr,J.E. and Zimmer,R. (2006) SSEP-Domain: protein domain prediction by alignment of secondary structure elements and profiles. *Bioinformatics*, **22**, 181–187.
11. Ofran,Y., Mysore,V. and Rost,B. (2007) Prediction of DNA-binding residues from sequence. *Bioinformatics*, **23**, i347–i353.
12. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al*. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
13. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
14. 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.
15. von Ohsen,N. and Zimmer,R. (2001) Improving Profile-Profile Alignments via Log Average Scoring. *Lecture Notes in Computer Science: Algorithms in Bioinformatics*. pp. 11–26.
16. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
17. 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.
18. Wang,L. and Brown,S.J. (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res.*, **34**, W243–W248.
19. 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.