利用蛋白質序列資訊以及
二及結構比對資訊來進行蛋白質與
DNA 序列結合專一性殘基
預測發現之研究

Prediction of Specific DNA-Binding Residues in a Transcription Factor based on Analysis of the Polypeptide Sequence

系所單位：　國立台灣大學
工程科學與海洋工程學研究所
資訊與網路實驗室

指導教授：　黃乾綱 老師

參與學生：　R96525072 黃俊欽
R97525023 蔣鈞堯
R97525034 邱莉媛
R97525037 柯佩均
R97525076 黃駿逸

Abstract

Background

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. In this respect, as a recent study has revealed that the tertiary structures of a large number of transcription factors are mostly disordered, sequence based analysis aimed at identifying the residues in a highly-disordered transcription factor that play a key role in interaction with the DNA is essential for obtaining a comprehensive picture of how the TF functions. In this respect, it is further desirable to have a predictor capable of distinguishing those residues involved in specific binding with the DNA, since specific binding corresponds to sequence-specific recognition of a gene and therefore is essential for correct gene regulation.

Results

This article presents the design of a polypeptide sequence based predictor for identifying the specific DNA-binding residues in a transcription factor. The design of the proposed predictor is distinctive by employing a hybrid approach aimed at achieving superior performance. In particular, two prediction mechanisms specialized to make predictions with certain types of protein secondary structure elements have been incorporated. In the experiments reported in this article, the proposed hybrid predictor has been able to deliver overall sensitivity of 59.5%, specificity of 98.8%, precision of 77.4%, and accuracy of 96.3%. Precision of 77.4% implies that about 3 out of 4 predicted binding residues are really involved in specific binding with the DNA. On the other hand, sensitivity of 59.5% implies that the predictor can catch about 6 out of 10 residues involved in specific binding with the DNA.

Conclusions

While the related studies reported in recent years did not distinguish between specific binding and non-specific binding, our study focuses on prediction of residues involved in specific binding with the DNA because specific binding corresponds to sequence-specific recognition of a gene and therefore is essential for correct gene regulation. Though the problem definitions of our study and the related works are not exactly identical, a performance comparison is of interest for obtaining a picture of how well the hybrid predictor proposed in this article works. The experimental results reveal that in comparison with the related works the proposed hybrid predictor is capable of delivering superior performance in terms of the harmonic mean of precision and sensitivity, which is a widely used performance metric in machine learning research. Furthermore, the proposed hybrid predictor is capable of delivering much higher precision than the other predictors. We emphasize precision because it provides the biochemist with a confidence level for designing an experiment to confirm whether a predicted binding residue is really involved in interaction with the DNA.

Background

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 (TF) 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: specific binding and non-specific binding [8]. Specific binding occurs between protein sidechains and nucleotide bases, while non-specific binding occurs between protein sidechains and the DNA sugar/phosphate backbone. In molecular biology, specific binding corresponds to sequence-specific recognition of a gene and therefore is essential for correct gene regulation. Fig. 1 illustrates these two types of binding mechanisms with a complex in the Protein Data Bank (PDB). In Fig. 1, a residue is regarded as involved in 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. On the other hand, a residue is regarded as involved in non-specific binding with the DNA, if the residue is not involved in specific binding with the DNA and one or more heavy atoms in its sidechain are within 4.5 Å from the sugar/phosphate backbone of the DNA.

This article presents the design of a sequence based predictor for identifying the residues in a TF that are involved in specific binding with the DNA. The design of the proposed predictor is distinctive by employing a hybrid approach aimed at achieving superior performance. In particular, two prediction mechanisms specialized to make predictions with different types of protein secondary structure elements have been incorporated. In the experiments reported in this article, the proposed hybrid predictor has been able to deliver overall sensitivity of 59.5%, specificity of 98.8%, precision of 77.4%, and accuracy of 96.3%, based on the following definitions:

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

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.

Results

Overview of the design of the proposed hybrid predictor

Fig. 2 presents an overview of the hybrid predictor proposed in this article. 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 in order to raise precision typically means that sensitivity is traded and vice versa. In particular, it has been observed in our experiments that the SVM with the parameter settings adopted in this article is capable of delivering reasonably well precision with respect to identifying the specific DNA-binding residues in α-helix and coil types of secondary structure elements. On the other hand, it has also been observed that the SVM hardly identifies the specific DNA-binding residues in β-sheet elements. Therefore, one straightforward way to improve the overall sensitivity of prediction is to incorporate a mechanism that can accurately identify the specific DNA-binding residues in β-sheet elements. As shown in Fig. 2, in the proposed hybrid predictor, we have incorporated a mechanism based on secondary structure element alignment (SSEA) to complement the prediction power of the SVM. The hybrid predictor then merges the outputs of the primary and auxiliary predictors by referring to the secondary structure elements predicted by the HYPROSP II server [9]. In this respect, the hybrid predictor will rely on the SVM based primary predictor to identify the specific DNA-binding residues in a secondary structure element that HYPROSP II predicts to be either an α-helix or a coil. On the other hand, the hybrid predictor will rely on the SSEA based auxiliary predictor to identify the specific DNA-binding residues in a secondary structure element that HYPROSP II predicts to be a β-sheet. The detailed design of the proposed hybrid predictor will be elaborated in the section **Methods**.

Performance Evaluation

In our study, we have conducted experiments to evaluate the performance of the proposed hybrid predictor. For training the hybrid predictor presented in Fig. 2, we have created a data set containing 228 TF-DNA complexes extracted from the 691 protein-DNA complexes that Yanay Ofran et al. [10] collected from the protein data bank (PDB) [11]. In this process, we first excluded those complexes in the Ofran collection that do not contain a TF. We 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 smaller than 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, 228 out of the 691 complexes initially in Ofran collection remained. This collection of 228 TF-DNA complexes was then employed to generate the training data set and testing data set in the experiments reported in this article.

The performance evaluation was conducted following the leave-one-out approach. Accordingly, the protein chain in each of the 228 TF-DNA complexes was used as the testing case once. In order to avoid bias caused by homologous protein chains, the training data set for the SVM and the template library for the SSEA algorithm were re-generated for each testing protein chain by removing those protein chains in the remaining 227 TF-DNA complexes that have a sequence identity higher than 20% when aligned with the testing protein chain. In our experiment, the BL2SEQ component of BLAST package [13] was invoked to obtain a score of sequence identity between two protein chains.

Table 1 shows how the SVM based predictor in Fig. 2 performed in the leave-one-out process. As mentioned earlier, the parameters of the SVM based predictor has been tuned to deliver high precision. As a result, sensitivity was traded. The results in Table 1 reveal that the SVM based predictor, to a certain extent, is capable of identifying the specific DNA-binding residues in α-helix and coil elements. On the other hand, the SVM based predictor can hardly identify the specific DNA-binding residues in β-sheet elements. Therefore, in order to raise sensitivity of prediction, we have resorted to the SSEA based mechanism to complement the prediction power of the SVM. Table 2 shows how the SSEA based predictor performed in identifying the specific DNA-binding residues in β-sheet elements. Combining the results in Tables 1 and 2, one can easily conclude that the prediction power of the SSEA based mechanism complements that of the SVM. With the SVM based predictor and the SSEA based predictor integrated as shown in Fig. 2, the hybrid predictor has been able to deliver the performance shown in Table 3.

Table 4 shows a breakdown of the experimental results with the hybrid predictor based on the classification of TF-DNA interactions proposed by J.M. Thornton et al. [14]. It should not be a surprise to observe that the hybrid predictor can deliver superior prediction accuracy when dealing with certain types of interactions and delivers inferior prediction accuracy with the other types. In this respect, what a biologist or chemist really cares about is whether the predictor could deliver extremely poor performance with certain types of interactions. The results reported in Table 4 show that the hybrid predictor does not suffer such kind of deficiency.

Discussion

In this section, we will discuss how the proposed hybrid predictor performs in comparison with the related works reported in recent years. In this respect, it must be noted that the problem definition in our study and those of the related studies are not exactly identical. While our study focuses on prediction of residues involved in specific binding with the DNA, all the related studies did not distinguish between specific binding and non-specific binding. Therefore, the results listed in Table 5, which includes the main results extracted from the related studies along with the overall results with the proposed hybrid predictor, 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[15], 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 and et. al.[16], a larger threshold of 4.5 Å, instead of 3.5 Å, has been adopted. In the article by Yan and 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.

In Table 5, the numbers listed with an asterisk have been derived from the numbers reported in the related studies. Since all the four related studies addressed in Table 5 reported three out of the first four performance metrics listed in the table, for each of the related study, we can obtain 3 equations about the following 4 variables:

$$\overset{\wedge}{TP} = \frac{TP}{TP+FP+TN+FN}, \overset{\wedge}{FP} = \frac{FP}{TP+FP+TN+FN},$$
$$\overset{\wedge}{TN} = \frac{TN}{TP+FP+TN+FN}, \overset{\wedge}{FN} = \frac{FN}{TP+FP+TN+FN}.$$

In addition, we have $\overset{\wedge}{TP}+\overset{\wedge}{FP}+\overset{\wedge}{TN}+\overset{\wedge}{FN}=1$. Therefore, for each related study, we can derive the actual values of the fourth performance metrics based on the values of the three performance metrics provided. The only exception is the case for the predictor proposed by Hwang and et. al.[16]. 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 and et. al. Therefore, there is no way to derive the exact values of the other performance metrics for their predictor.

The numbers reported in Table 5 reveal that in comparison with the related works the proposed hybrid predictor is capable of delivering superior performance in terms of the harmonic mean of precision and sensitivity (F-score), which is a widely used performance metric in machine learning research. Furthermore, the proposed hybrid predictor is capable of delivering much higher precision than the other predictors. We emphasize precision because it provides the biochemist with a confidence level for designing an experiment to confirm whether a predicted binding residue is really involved in interaction with the DNA. With the proposed hybrid predictor, the biochemist can expect that on the average three out of the four predicted binding residues are really involved in specific binding with the DNA. On the other hand, the proposed hybrid predictor, on the average, can catch about 6 out of 10 specific binding residues.

Conclusions

This article presents the design of a sequence based predictor aimed at identifying the 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 highly-disordered TF interacts with the DNA. Furthermore, as specific binding corresponds to sequence-specific recognition of a gene and is essential for correct gene regulation, the capability to identify those residues involved in specific binding with the DNA is of particular interest.

In the experiments reported in this article, the proposed hybrid predictor delivered overall precision of 77.4%, sensitivity of 59.5%, specificity of 98.8%, and accuracy of 96.3%. Precision of 77.4% implies that about 3 out of 4 predicted binding residues are really involved in specific binding with the DNA. On the other hand, sensitivity of 59.5% implies that the predictor can catch about 6 out of 10 residues involved in specific binding with the DNA. The experimental results further show that the proposed hybrid predictor is capable of delivering the same level of prediction accuracy when dealing with different types of TF-DNA interactions.

It is anticipated the prediction accuracy delivered by the hybrid predictor will continue to improve as the number of TF-DNA complexes deposited in the PDB continues to grow and leads to continuous increase of the number of training samples that can be exploited. 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.

Methods

As shown in Fig. 2, the hybrid predictor proposed in this article consists of the primary predictor and the auxiliary predictor. This section elaborates the design of the primary and auxiliary predictor.

Design of the Primary predictor

For the design of the primary predictor, we have employed the LIBSVM package [17] with the Gaussian kernel. The model of the SVM has been generated based on a training data set derived from the 228 TF-DNA complexes described above. The training data set was generated by associating each residue in the 228 protein chains with a position specific scoring matrix (PSSM) computed by the PSI-BLAST package with window size set to 11 [5]. In addition, each residue was labeled based on whether it is involved in specific binding with the DNA or not. As mentioned earlier, a residue is regarded as involved in specific binding with the DNA, if one or more heavy atoms in its sidechain are within 4.5Å from the nucleobases of the DNA. The end result was a training data set containing a total of 22097 samples, of which 1416 samples are positive and 20751 samples are negative.

As mentioned earlier, the parameters of the SVM have been set to deliver high precision. In this respect, we have set parameters C and g with the Gaussian kernel to 32 and 0.03125, respectively.

Design of the auxiliary predictor

As mentioned earlier, the auxiliary predictor incorporates a mechanism based on secondary structure element alignment (SSEA), which was firstly proposed by Gewehr and Zimmer [18]. The SSEA based mechanism refers to a template library containing the sequences of specific DNA binding domains. The template library has been created with the following steps:

1. Each protein chain in the 228 TF-DNA complexes was submitted to the HYPROSP II server as well as to the PFAM server. Then, each residue in the predicted β-sheet elements was examined to determine whether it is involved in specific binding with the DNA. If a β-sheet element contains one or more residues involved in specific binding with the DNA, then the β-sheet element was regarded as involved in specific binding with the DNA.
2. If a DNA binding domain output by the PFAM server contained one or more β-sheet elements involved in specific binding with DNA, then the binding domain was deposited into the template library. In addition, each residue in the domain was labeled by the HYPROSP II with one of the following three types: α-helix, β-sheet, and coil.

With the template library, we then can invoke the following procedure to predict the specific DNA-binding residues in the β-sheet elements of the query TF.

1. Invoke the HYPROSP II server to label each residue in the query TF with one of following three types: α-helix, β-sheet, and coil.
2. Invoke the BLAST package [19] to align the sequence of labels of the query TF with the sequence of labels of each template in the library. The similarity score between the query TF and a template is then computed as follows.

$$\sum_{i}\sum_{j}\sum_{k}\alpha_i(j)\beta_i(k)\exp\{\lambda S[\alpha_i(j),\beta_i(k)]\},$$

where

(i) i is the index of the aligned residue pairs;

(ii) $\alpha_i(j)$ and $\beta_i(k)$ are the PSSM vectors corresponding to the aligned residue pair with index i;

(iii) $S[\alpha_i(j), \beta_i(k)]$ is the score of BLOSUM62 corresponding to residue pair $\alpha_i(j)$ and $\beta_i(k)$;

(iv) is a parameter and has been set to 0.347.

The positions of the specific DNA-binding residues in the 5 templates that give the highest similarity scores are then superimposed to predict the positions of the specific DNA-binding residues in the query TF.

References

1. Ahmad S, Sarai A: **PSSM-based prediction of DNA binding sites in proteins.** *BMC Bioinformatics* 2005, **6:**33.
2. 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.

3. 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.

4. Ferrer-Costa C, Shanahan HP, Jones S, Thornton JM: **HTHquery: a method for detecting DNA-binding proteins with a helix-turn-helix structural motif.** *Bioinformatics* 2005, **21:**3679-3680.

5. Tjong H, Zhou HX: **DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces.** *Nucleic Acids Res* 2007, **35:**1465-1477.

6. Tsuchiya Y, Kinoshita K, Nakamura H: **Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces.** *Proteins* 2004, **55:**885-894.

7. Liu J, Perumal NB, Oldfield CJ, Su EW, Uversky VN, Dunker AK: **Intrinsic disorder in transcription factors.** *Biochemistry* 2006, **45:**6873-6888.

8. Boyer RF: **Concepts in Biochemistry: Structure Tutorials.** In *Concepts in Biochemistry.* 3rd edition edition: Wiley; 2005: 736

9. Lin HN, Chang JM, Wu KP, Sung TY, Hsu WL: **HYPROSP II--a knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence.** *Bioinformatics* 2005, **21:**3227-3233.

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

11. 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.

12. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, et al: **Pfam: clans, web tools and services.** *Nucleic Acids Res* 2006, **34:**D247-251.

13. Tatusova TA, Madden TL: **BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences.** *FEMS Microbiol Lett* 1999, **174:**247-250.

14. Luscombe NM, Austin SE, Berman HM, Thornton JM: **An overview of the structures of protein-DNA complexes.** *Genome Biol* 2000, **1:**REVIEWS001.

15. 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.

16. 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.

17. Chang C-C, Lin C-J: **LIBSVM: a library for support vector machines.** 2001**:**Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

18. Gewehr JE, Zimmer R: **SSEP-Domain: protein domain prediction by alignment of secondary structure elements and profiles.** *Bioinformatics* 2006, **22:**181-187.

19. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.

Figures

Figure 1 - An example of protein-DNA interaction. This complex is with PDB ID 1YSA and contains Yeast TF GCN4.
The atoms colored by red are the heavy atoms in the sidechains of the specific

DNA-binding residues. The atoms colored by light blue are the heavy atoms in the side-chains of the non-specific DNA-binding residues.

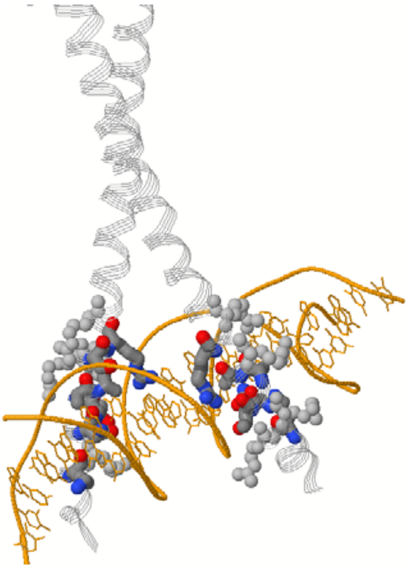
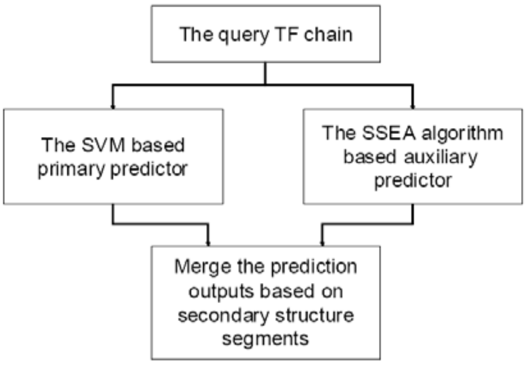


Figure 2 - The overall structure of the proposed hybrid predictor.


The query TF chain
The SVM based primary predictor
The SSEA algorithm based auxiliary predictor
Merge the prediction outputs based on secondary structure segments

Tables

Table 1 - Prediction results with the SVM based primary predictor.

Type of the secondary structure element	# in residues tested	Prediction results							
		TP	TN	FP	FN	Precision	Sensitivity	Specificity	Accuracy
Helix	12781	573	11670	156	382	0.786	0.600	0.987	0.958
Sheet	1465	0	1358	3	104	0.000	0.000	0.998	0.927
Coil	7921	186	7506	58	171	0.762	0.521	0.992	0.971

Table 2 - Prediction results with the SSEA based auxiliary predictor.

Type of the secondary structure element	# in residues tested	Prediction results							
		TP	TN	FP	FN	Precision	Sensitivity	Specificity	Accuracy
Sheet	1465	83	1329	32	21	0.722	0.798	0.976	0.964

Table 3 - Prediction results with the hybrid predictor.

Type of the secondary structure element	# in residues tested	Prediction results							
		TP	TN	FP	FN	Precision	Sensitivity	Specificity	Accuracy
Helix	12781	573	11670	156	382	0.786	0.600	0.987	0.958
Sheet	1465	83	1329	32	21	0.722	0.798	0.976	0.964
Coil	7921	186	7506	58	171	0.762	0.521	0.992	0.971
Overall	22167	842	20505	246	574	0.774	0.595	0.988	0.963

Table 4 - Breakdown of the experimental results with the hybrid predictor in respect of different types of TF-DNA bindings

Type of TF-DNA bindings	# of TFs involved	# in residues tested	Prediction results							
			TP	TN	FP	FN	Precision	Sensitivity	Specificity	Accuracy
Zipper-type	44	3109	213	2821	30	45	0.877	0.826	0.989	0.976
Helix-turn-helix	97	12480	316	11712	123	329	0.720	0.490	0.990	0.964
Zinc-coordinating	57	4792	230	4332	74	156	0.757	0.596	0.983	0.952
β-hairpin/ribbon	30	1786	83	1640	19	44	0.814	0.654	0.989	0.965
Overall	228	22167	842	20505	246	574	0.774	0.595	0.988	0.963

Table 5 - 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**
The proposed hybrid predictor	0.595	0.988	0.963	0.774	0.671
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 (Wang and Brown[15])	0.652	0.728	0.722	0.186*	0.289*
DP-Bind (Hwang and et. al.[16])	0.791	0.786	0.800	–*	–*