Published online 31 July 2009

Nucleic Acids Research, 2009, Vol. 37, No. 12 ***1–S14***
doi:10.1093/nar/gkn000

Supplementary: Computational Learning on Specificity-Determining Residue-Nucleotide Interactions

Ka-Chun Wong [1] **et al.** [*]

[1] Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong

Covariation Analysis Implementation

In the implementation level as shown in Figure S1, we have a matrix for each side. On the DBD sequence side, each row corresponds to a DBD sequence while each column corresponds to the presence of an amino acid residue at an aligned position. On the DNA motif matrix side, each row corresponds to the flattened vector of an aligned DNA motif matrix. We traverse and compute all the possible pair-wise correlations between each column of the DBD numeric matrix and each column of the DNA numeric matrix, resulting in a correlation matrix (heat map). Exact p-value is computed for each Spearman correlation value. The correlation matrix entries with p-value $>= 0.01$ are discarded and assigned zeros. The resultant correlation matrix (heat map) is then clustered for each DBD family. We can observe that there are statistically significant co-variations between residues and nucleotides. To visualize the co-variations, we can obtain a binding pair of protein sequence and DNA sequence from PDB and select the corresponding columns and rows to form its own correlation sub-matrix (heat map) from the correlation matrix (heat map) as shown from Step 7 to Step 8 on Figure S1.

Time Complexity Analysis

The overall approach is summarized in Figure 1, which can be divided into training and testing.

Training Procedure For the model training part of each domain as shown in Figure 1, time complexity is analyzed step by step. (Steps A and B) N' training protein-DNA binding sequence pairs with their structural information are retrieved from PDB. The average lengths of the protein sequences and DNA sequences of the pairs are indicated as L_{aa} and L_{dna} respectively. (Step C) CD-HIT redundancy removal on the protein sequences of the training protein-DNA binding sequence pairs causes $O(N'L_{aa})$, resulting in N non-redundant pairs. (Step D) Let the average number of atoms of amino acids and those of nucleotides be α_{aa} and α_{dna} respectively. All the possible pair-wise residue-nucleotide interactions are examined with their structural atom information in this step, resulting in the time complexity $O(NL_{aa}\alpha_{aa}L_{dna}\alpha_{dna})$. After that, $NL_{aa}L_{dna}$ labeled residue-nucleotide interactions are obtained. (Step E) The isolation of class labels incurs $O(NL_{aa}L_{dna})$. (Steps F to H) We have M protein-DNA binding sequence pairs from the CISBP database. In this study, we have chosen MUSCLE to conduct the multiple sequence alignment with the time complexities $O(M^4+ML_{aa}^2)$ and $O(M^4+ML_{dna}^2)$ for protein and DNA respectively (1). (Step I) MUSCLE is applied again to align the training sequence pairs to the multiple sequence alignment profile pairs constructed in the last step (step H), resulting in $N*O(2^4+2L_{aa}^2)$ and $N*O(2^4+2L_{dna}^2)$ for protein and DNA respectively (1). (Step J) To build $NL_{aa}L_{dna}$ feature vectors for $NL_{aa}L_{dna}$ labeled residue-nucleotide interactions with the help of the alignment profiles built in the previous steps, different feature building methods are involved. The mapping methods require constant time complexity, resulting in $O(NL_{aa}L_{dna})$ in total; The feature building which involves looking up the alignment profiles built causes $O(M^2)$ at most, resulting in $O(NL_{aa}L_{dna}M^2)$ in total. Note that the feature building which involves whole alignment length lookups can be pre-computed in a single pass first. (Steps J and K) Given $NL_{aa}L_{dna}$ feature vectors with F input features, a classification model is trained. In this study, we have chosen Random Forest as the model. For its building, $NL_{aa}L_{dna}$ data vectors with F input features are given for training (building). For each decision tree, a random set of R input features is used for node split. To build a random decision tree, assuming the average depth of those decision trees is D, time complexity $O(DNL_{aa}L_{dna}R)$ complexity is involved. If the Random Forest model has T trees, the total model building time complexity is $O(TDNL_{aa}L_{dna}R)$.

In summary, the overall time complexity of model training is $O(N'L_{aa}) + O(NL_{aa}\alpha_{aa}L_{dna}\alpha_{dna}) + O(NL_{aa}L_{dna}) + O(M^4+ML_{aa}^2) + O(M^4+ML_{dna}^2) + N*O(2^4+2L_{aa}^2) + N*O(2^4+2L_{dna}^2) + O(NL_{aa}L_{dna}) + O(NL_{aa}L_{dna}M^2) + O(TDNL_{aa}L_{dna}R)$. If only dominant complexities are counted, it can be written as $O(N'L_{aa}) + O(NL_{aa}\alpha_{aa}L_{dna}\alpha_{dna}) + O(M^4+ML_{aa}^2) + O(M^4+ML_{dna}^2) + O(NL_{aa}^2) + O(NL_{dna}^2) + O(NL_{aa}L_{dna}M^2) + O(TDNL_{aa}L_{dna}R)$.

[*] To whom correspondence should be addressed. Email: kc.w@cityu.edu.hk

Testing Procedure Given an input protein-DNA binding sequence pair of lengths l_{aa} and l_{dna}, we aim at applying the corresponding trained model to predict its interactions for model testing as shown in Figure 1. Similar to the model training section, time complexity is analyzed step by step. (Steps 1 and 2) The time complexity of the steps 1 and 2 depends on the actual implementation of the PFam database. Nonetheless, most of the queries have already been precomputed by the PFam database. We can safely assume constant complexity here (2). (Steps 3 to 5) The steps have already been computed in the model training part. (Step 6) MUSCLE is applied again to align the input sequence pair to the multiple sequence alignment profile pairs constructed in the step 5, resulting in $O(2^4+2L_{aa}^2)$ and $O(2^4+2L_{dna}^2)$ for protein and DNA respectively (1). (Step 7) As elaborated in the model training section, the time complexity of feature vector building for a protein-DNA binding sequence pair is $O(l_{aa}l_{dna}M^2)$. (Step 8) For the classification (prediction) part, we just need to traverse all the T decision trees of the Random Forest classifier which we have trained. Assuming the average depth of those decision trees is D, the time complexity to obtain the prediction score is $O(TD)$ for each possible residue-nucleotide interaction, resulting in $O(l_{aa}l_{dna}TD)$ in total for all the $l_{aa}l_{dna}$ possible residue-nucleotide interactions on the input protein-DNA binding sequence pair.

In summary, the overall time complexity of model testing is $O(2^4+2L_{aa}^2)$ + $O(2^4+2L_{dna}^2)$ + $O(l_{aa}l_{dna}M^2)$ + $O(l_{aa}l_{dna}TD)$.

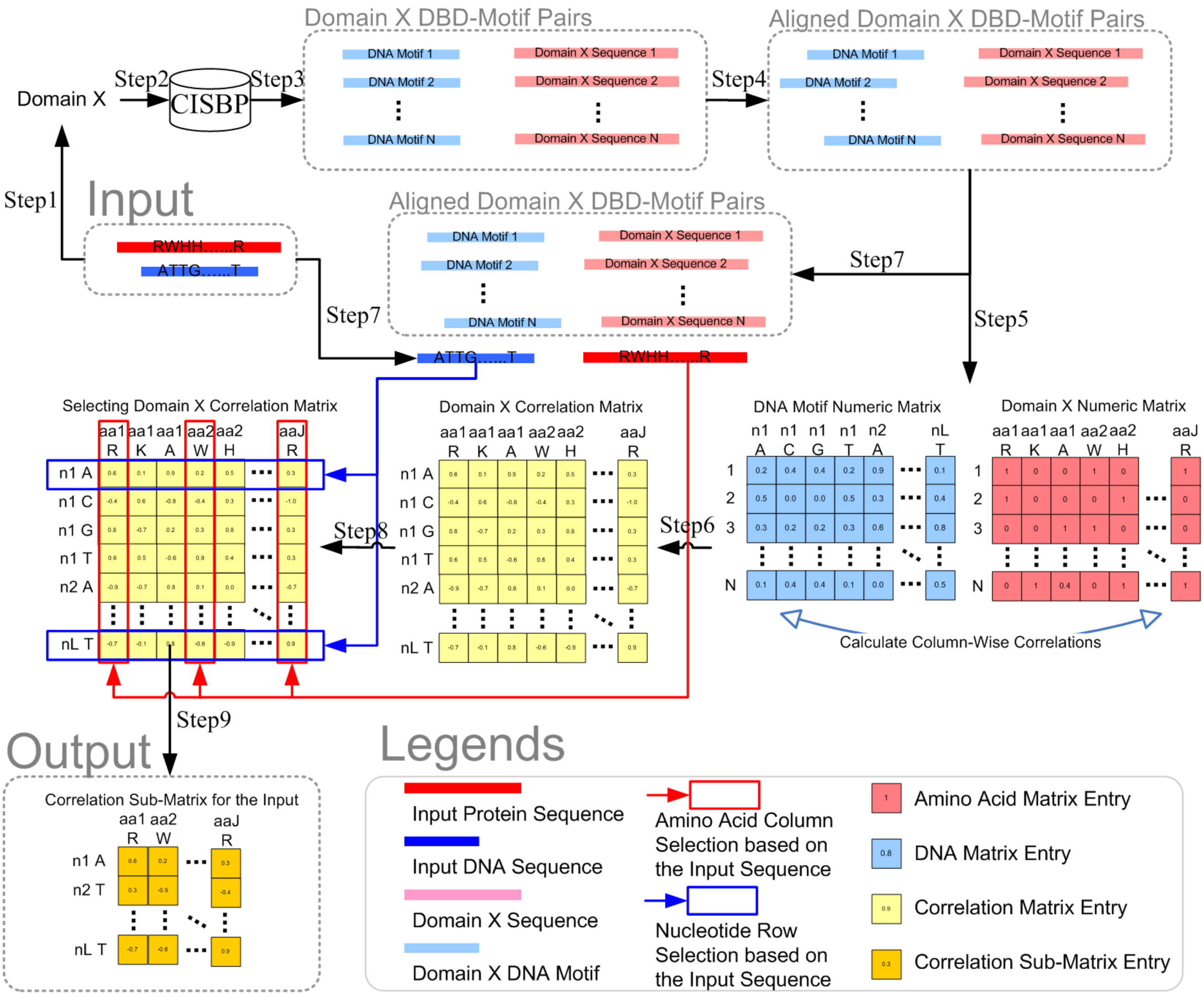
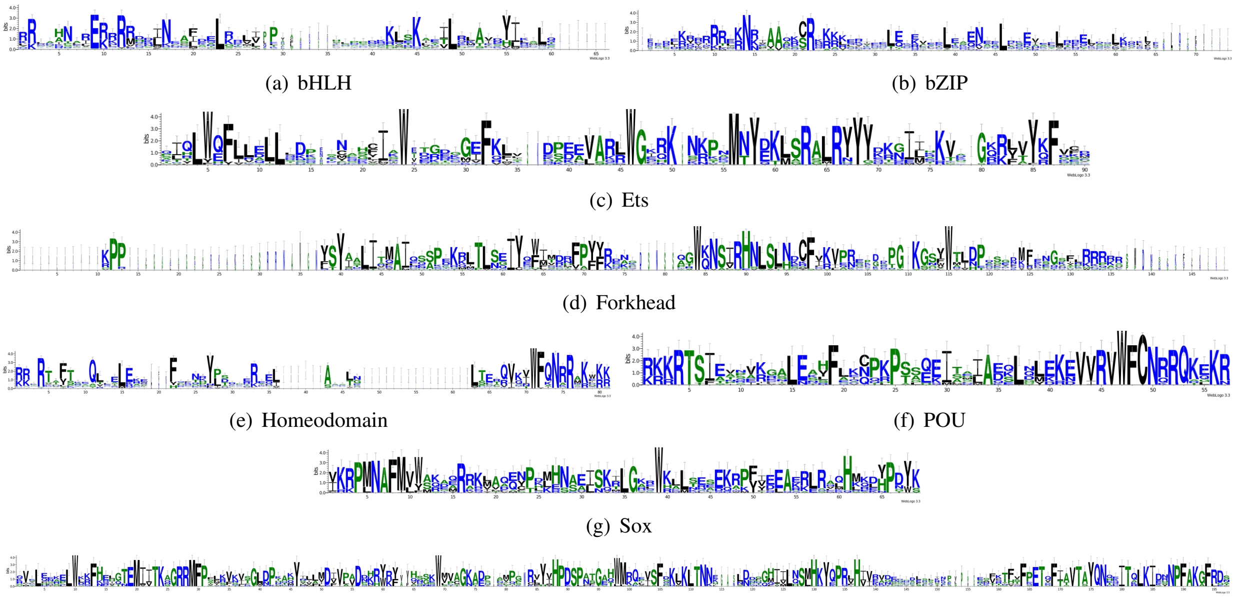


Figure S1. Constructing the correlation sub-matrix (heat map) for an input pair of protein-DNA binding sequences. Step1: We identify which DBD domain the input pair of protein-DNA binding sequence belongs to (Domain X in this example). Steps 2 and 3: The entire Domain X sequences and the corresponding DNA motifs are retrieved from CISBP. Step 4: The retrieved domain X sequences and the corresponding DNA motifs are aligned. Step5: The DNA motif alignment is transformed into a numeric matrix whereas the domain X sequence alignment is transformed into a binary matrix. Step6: Correlations are calculated between the two matrices. Step7: The input sequence pair is aligned to the existing domain X family alignment to identify their own aligned positions using MUSCLE and STAMP for protein and DNA sides respectively. Step8: The correlation matrix rows and columns corresponding to the input are selected. Step9: The selected rows and columns are concatenated to form the correlation sub-matrix (heat map) for the input.



(a) bHLH

(b) bZIP

(c) Ets

(d) Forkhead

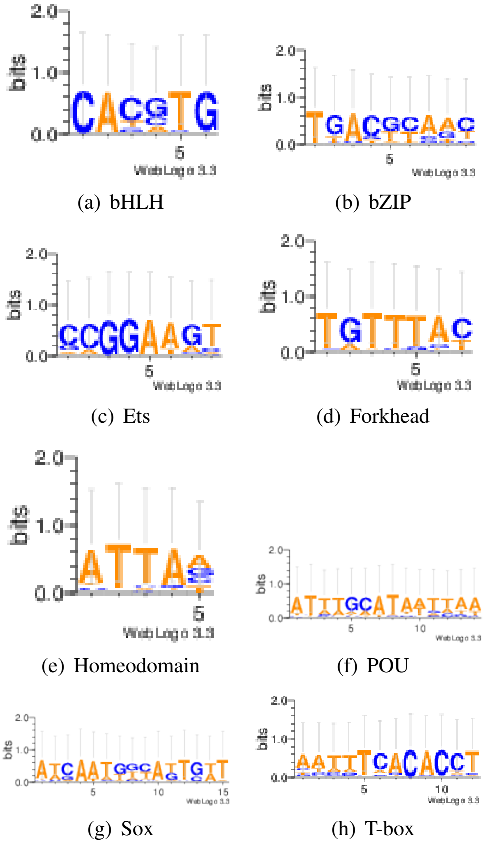
(e) Homeodomain

(f) POU

(g) Sox

(h) T-box

Figure S2. Sequence logos for the DNA-binding domain (DBD) sequences used



(a) bHLH (b) bZIP

(c) Ets (d) Forkhead

(e) Homeodomain (f) POU

(g) Sox (h) T-box

Figure S3. Sequence logos for the DNA binding sites of the DNA-binding domain sequences used

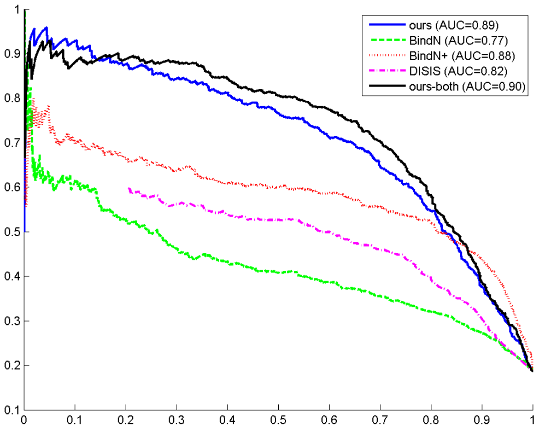


Figure S4. Precision-Recall (PRC) curves for our proposed methods (in Blue and Black), BindN (in Green), BindN+(in Red), and DISIS (in Violet) on the entire DBD families.

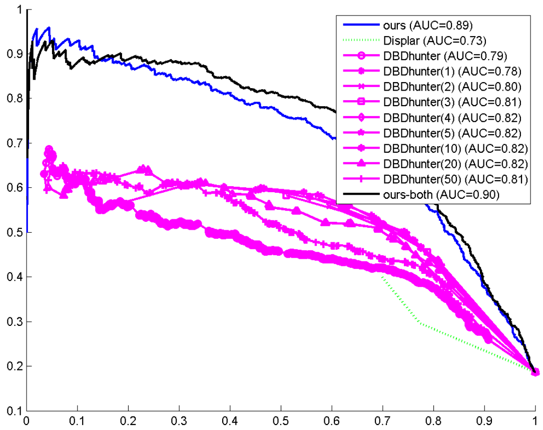


Figure S5. Precision-Recall curves for our proposed methods (in Blue and Black), DBD-Hunter (in Violet), DISPLAR (in Green) on the entire DBD families.

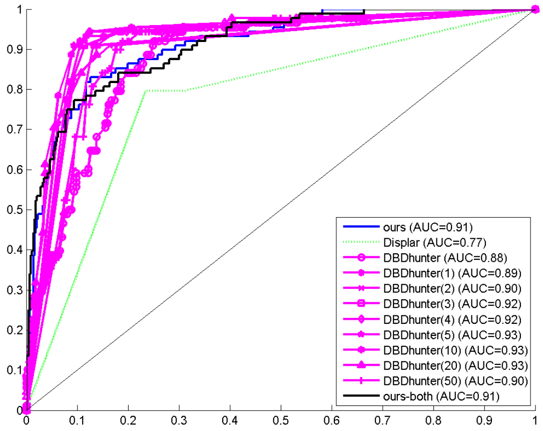


Figure S6. Receiver Operating Characteristic (ROC) curves for our proposed methods (in Blue and Black), DBD-Hunter (in Violet), DISPLAR (in Green) on bHLH family.

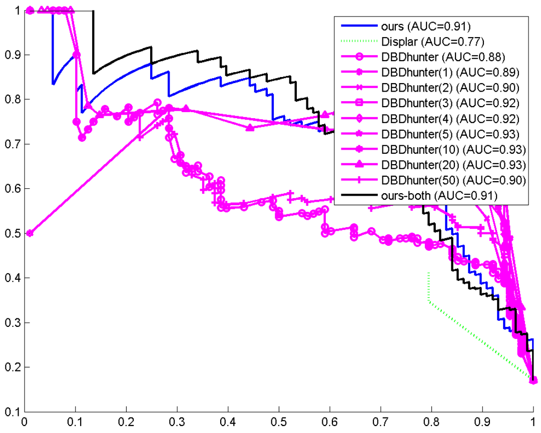


Figure S7. Precision-Recall (PRC) curves for our proposed methods (in Blue and Black), DBD-Hunter (in Violet), DISPLAR (in Green) on bHLH family.

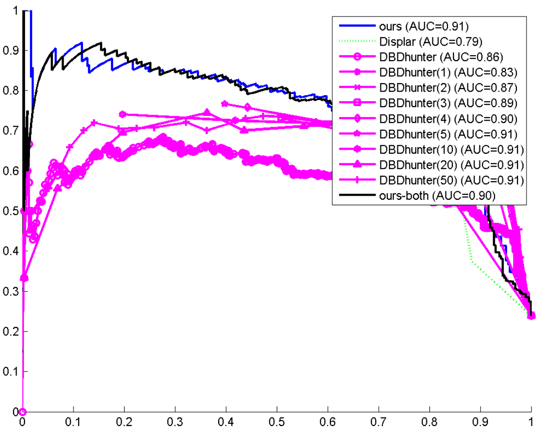


Figure S9. Precision Recall curves for our proposed methods (in Blue and Black), DBD-Hunter (in Violet), DISPLAR (in Green) on Homeodomain family.

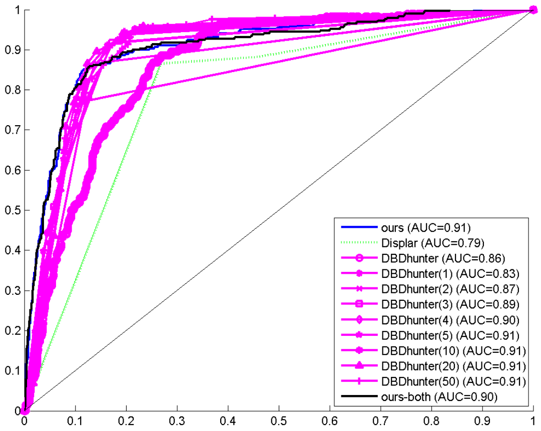


Figure S8. Receiver Operating Characteristic (ROC) curves for our proposed methods (in Blue and Black), DBD-Hunter (in Violet), DISPLAR (in Green) on Homeodomain family.

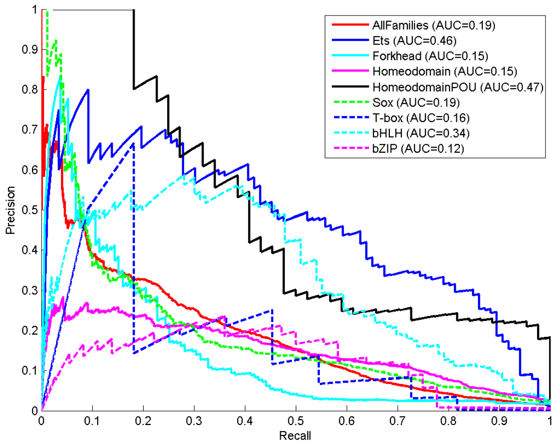
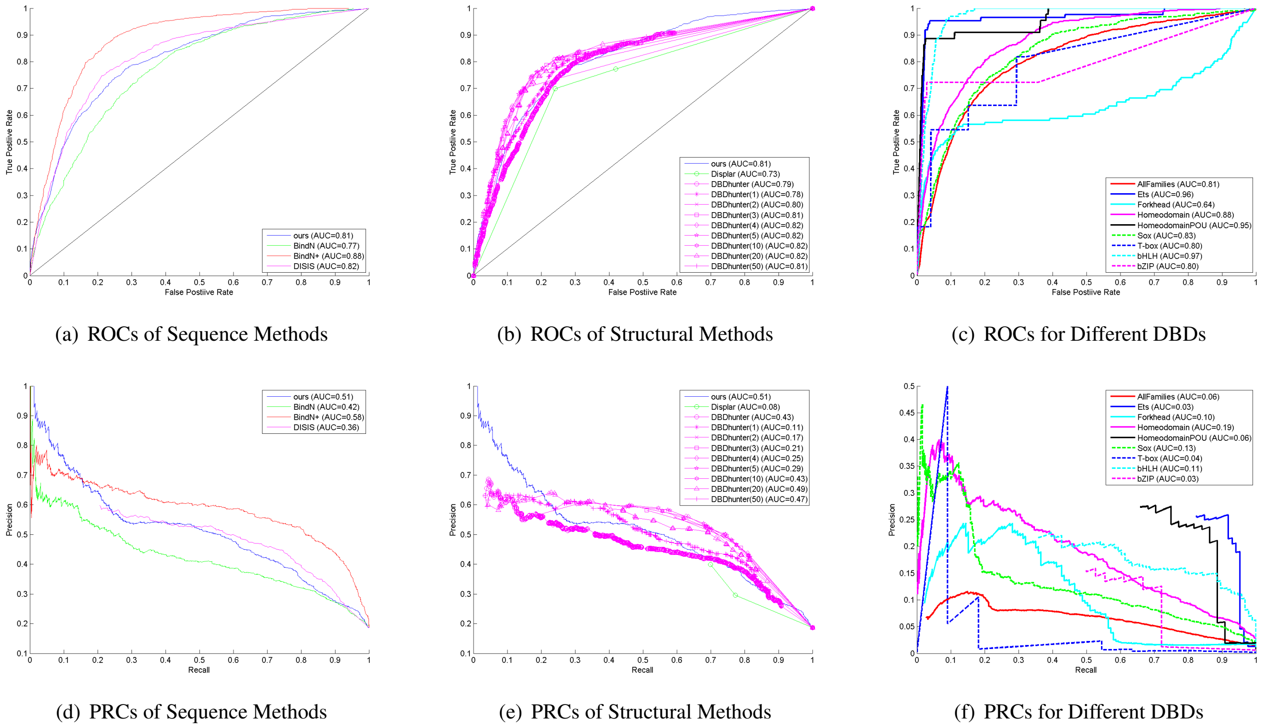


Figure S10. Precision Recall (PRC) curves for our proposed method on the DBD family data. Each line corresponds to a DBD family.



(a) ROCs of Sequence Methods

(b) ROCs of Structural Methods

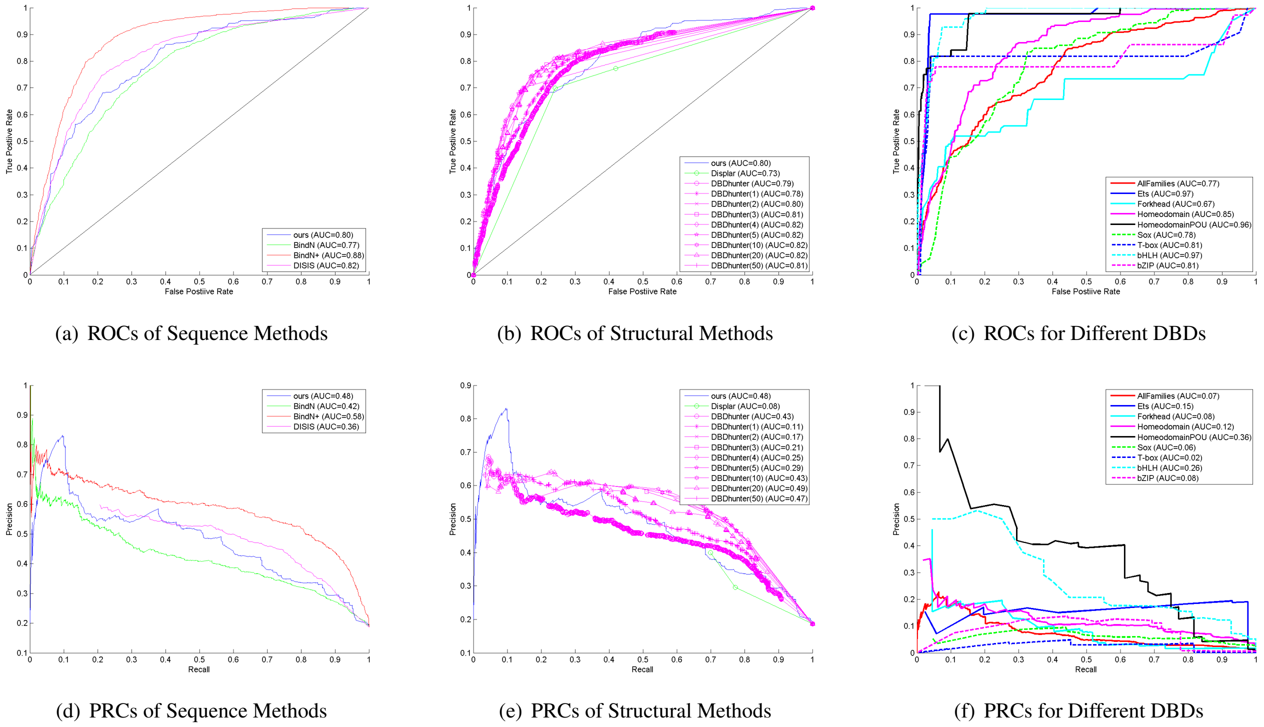
(c) ROCs for Different DBDs

(d) PRCs of Sequence Methods

(e) PRCs of Structural Methods

(f) PRCs for Different DBDs

Figure S11. The overall performance if Naive Bayes is used instead of Random Forest.



(a) ROCs of Sequence Methods

(b) ROCs of Structural Methods

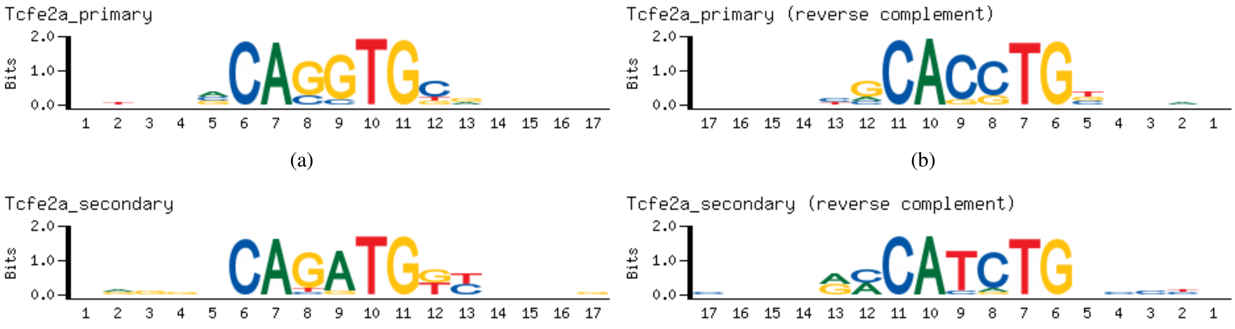
(c) ROCs for Different DBDs

(d) PRCs of Sequence Methods

(e) PRCs of Structural Methods

(f) PRCs for Different DBDs

Figure S12. The overall performance if Adaboost M1 is used instead of Random Forest.


Tcfe2a_primary

Tcfe2a_primary (reverse complement)

Tcfe2a_secondary

Tcfe2a_secondary (reverse complement)

(a)

(b)

(c)

(d)

Figure S13. Sequence logos measured on the bHLH DBD domain of the transcription factor E2-alpha (UniProt code: P21677, UniPROBE code: Tcfe2a) (3)

Table S1. Statistics of human DNA-binding domains collected from CISBP (v0.71)

DNA-Binding Domain Family	DBD Sequence-DNA Motif Matrix Pairs
T-box	13
POU	15
Sox	16
Forkhead	25
Ets	26
bZIP	39
bHLH	48
Homeodomain	142

Table S2. Statistics of extracted DBD sequences from PDB

DNA-Binding Domain Family	DBD Sequence-DNA Sequence Pairs
T-box	2
POU	4
Sox	16
Forkhead	6
Ets	8
bZIP	5
bHLH	10
Homeodomain	22

Table S3. List of input features.

Feature	Description	Data Type
aa-10	The 10th preceding residue	A,R,N,...,V,-
aa-9	The 9th preceding residue	A,R,N,...,V,-
aa-8	The 8th preceding residue	A,R,N,...,V,-
aa-7	The 7th preceding residue	A,R,N,...,V,-
aa-6	The 6th preceding residue	A,R,N,...,V,-
aa-5	The 5th preceding residue	A,R,N,...,V,-
aa-4	The 4th preceding residue	A,R,N,...,V,-
aa-3	The 3rd preceding residue	A,R,N,...,V,-
aa-2	The 2nd preceding residue	A,R,N,...,V,-
aa-1	The 1st preceding residue	A,R,N,...,V,-
aa1	The 1st succeeding residue	A,R,N,...,V,-
aa2	The 2nd succeeding residue	A,R,N,...,V,-
aa3	The 3nd succeeding residue	A,R,N,...,V,-
aa4	The 4th succeeding residue	A,R,N,...,V,-
aa5	The 5th succeeding residue	A,R,N,...,V,-
aa6	The 6th succeeding residue	A,R,N,...,V,-
aa7	The 7th succeeding residue	A,R,N,...,V,-
aa8	The 8th succeeding residue	A,R,N,...,V,-
aa9	The 9th succeeding residue	A,R,N,...,V,-
aa10	The 10th succeeding residue	A,R,N,...,V,-
nt-5	The 5th preceding nucleotide	A,C,G,T,-
nt-4	The 4th preceding nucleotide	A,C,G,T,-
nt-3	The 3rd preceding nucleotide	A,C,G,T,-
nt-2	The 2nd preceding nucleotide	A,C,G,T,-
nt-1	The 1st preceding nucleotide	A,C,G,T,-
nt1	The 1st succeeding nucleotide	A,C,G,T,-
nt2	The 2nd succeeding nucleotide	A,C,G,T,-
nt3	The 3nd succeeding nucleotide	A,C,G,T,-
nt4	The 4th succeeding nucleotide	A,C,G,T,-
nt5	The 5th succeeding nucleotide	A,C,G,T,-
aa	The current residue	A,R,N,...,Y,V
nt	The current nucleotide	A,C,G,T
aa-nt	The current residue and nucleotide pair	AA,RA,NA,...,YT,VT
hydropathyIndex	Hydropathy Index of the current residue	numeric
mass	Mass of the current residue	numeric
npsa	Non-Polar Surface Area of the current residue	numeric
polarity	Polarity of the current residue	n,p,pn,pp
residueBurial	Estimated Hydrophobic Effect For Residue Burial of the current residue	numeric
sea10	Occurring Percentage for Solvent Exposed Area less than 10 square angstrom of the current residue	numeric
sea1030	Occurring Percentage for Solvent Exposed Area between 10 and 30 square angstrom of the current residue	numeric
sea30	Occurring Percentage for Solvent Exposed Area higher than 30 square angstrom of the current residue	numeric
sideChainBurial	Estimated Hydrophobic Effect for side chain burial of the current residue	numeric
surface	Surface Area of the current residue	numeric
volume	Volume of the current residue	numeric
pH	pH at the isoelectric point of the current residue	numeric
corr	Spearman Rank Correlation of the current residue and nucleotide pair in the family alignment	numeric
pvalue	P-value for the Spearman Rank Correlation of the current residue and nucleotide pair	numeric
aa-totalCorr	The Sum of Correlations between the current residue and all the input nucleotides	numeric
aa-totalCorr-count	The Count of Correlations between the current residue and all the input nucleotides	numeric
aa-avgCorr	The Mean of Correlations between the current residue and all the input nucleotides	numeric
stat-aa-totalCorr	The Sum of Positive Correlations between the current residue and all the input nucleotides with P-value <0.01	numeric
stat-aa-totalCorr-count	The Count of Positive Correlations between the current residue and all the input nucleotides with P-value <0.01	numeric
stat-aa-avgCorr	The Mean of Positive Correlations between the current residue and all the input nucleotides with P-value <0.01	numeric
stat-aa-totalCorr-negative	The Sum of Negative Correlations between the current residue and all the input nucleotides with P-value <0.01	numeric
stat-aa-totalCorr-count-negative	The Count of Negative Correlations between the current residue and all the input nucleotides with P-value <0.01	numeric
stat-aa-avgCorr-negative	The Mean of Negative Correlations between the current residue and all the input nucleotides with P-value <0.01	numeric
dna-totalCorr	The Sum of Correlations between the current nucleotide and all the input residues	numeric
dna-totalCorr-count	The Count of Correlations between the current nucleotide and all the input residues	numeric
dna-avgCorr	The Mean of Correlations between the current nucleotide and all the input residues	numeric
stat-dna-totalCorr	The Sum of Positive Correlations between the current nucleotide and all the input residues with P-value <0.01	numeric
stat-dna-totalCorr-count	The Count of Positive Correlations between the current nucleotide and all the input residues with P-value <0.01	numeric
stat-dna-avgCorr	The Mean of Positive Correlations between the current nucleotide and all the input residues with P-value <0.01	numeric
stat-dna-totalCorr-negative	The Sum of Negative Correlations between the current nucleotide and all the input residues with P-value <0.01	numeric
stat-dna-totalCorr-count-negative	The Count of Negative Correlations between the current nucleotide and all the input residues with P-value <0.01	numeric
stat-dna-avgCorr-negative	The Mean of Negative Correlations between the current nucleotide and all the input residues with P-value <0.01	numeric
aa-presence	The Sum of Correlations between the current residue and all the possible nucleotides in the family alignment	numeric
aa-presence-total	The Sum of Correlations between all residues at the current residue’s aligned position and all the possible nucleotides in the family alignment	numeric
dna-presence	The Sum of Correlations between the current nucleotide and all the possible residues in the family alignment	numeric
dna-presence-total	The Sum of Correlations between all nucleotides at the current nucleotide’s aligned position and all the possible residues in the family alignment	numeric
MI	Discrete Mutual Information for the current residue and base pair in the family alignment	numeric
advMI	Continuous Mutual Information for the current residue and base pair in the family alignment	numeric
MIp	Discrete Corrected Mutual Information for the current residue and base pair in the family alignment	numeric
advMIp	Continuous Corrected Mutual Information for the current residue and base pair in the family alignment	numeric
aa-entropy	Entropy of the current residue’s aligned position in the family alignment	numeric
nt-entropy	Discrete Entropy of the current nucleotide’s aligned position in the family alignment	numeric
adv-nt-entropy	Continuous Entropy of the current nucleotide’s aligned position in the family alignment	numeric
polarity-entropy	Entropy of the current residue’s aligned position in the family alignment using polarity symbols	numeric
avg-hydropathyIndex	Average Hydropathy Index of the current residue’s aligned position in the family alignment	numeric
avg-Ph	Average pH of the current residue’s aligned position in the family alignment	numeric
avg-mass	Average Mass of the current residue’s aligned position in the family alignment	numeric
aa-blosum	Average BLOSUM62 Score of the current residue to all the other residues at the same aligned position in the family alignment	numeric
nt-nuc44	Average NUC44 Score of the current nucleotide to all the other nucleotides at the same aligned position in the family alignment	numeric
aa-obsCount	The occurring fraction of non-gap residues at the current residue’s aligned position in the family alignment	numeric
nt-obsCount	The occurring fraction of non-gap nucleotides at the current nucleotide’s aligned position in the family alignment	numeric
aaMSAind	Aligned Position of the current residue	numeric
dnaMSAind	Aligned Position of the current nucleotide	numeric
aaSeqPos	Input Sequence Position of the current residue	numeric
dnaSeqPos	Input Sequence Position of the current nucleotide	numeric
class	Class Label to indicate whether the current residue and nucleotide pair binds or not	’NotBinding’,’Binding’

Table S4. List of input protein features ranked by information gain on the protein sequences of the PDB data collected.

Rank	Information Gain	H(Class) - H(Class | Attribute)
1	0.07034	37 polarity_entropy
2	0.06724	39 avg_Ph
3	0.05546	42 aa_obsCount
4	0.04358	43 aaMSAind
5	0.04337	35 aa_presence_total
6	0.04051	21 aa
7	0.03931	38 avg_hydropathyIndex
8	0.03823	30 sideChainBurial
9	0.03823	26 residueBurial
10	0.03779	24 npsa
11	0.03686	33 pH
12	0.03673	27 sea10
13	0.03478	31 surface
14	0.03457	23 mass
15	0.03403	29 sea30
16	0.03179	32 volume
17	0.03162	28 sea1030
18	0.03083	40 avg_mass
19	0.03029	36 aa_entropy
20	0.02437	22 hydropathyIndex
21	0.0241	41 aa_blosum
22	0.02152	44 aaSeqPos
23	0.02054	2 aa-9
24	0.01969	8 aa-3
25	0.0193	25 polarity
26	0.01772	3 aa-8
27	0.01732	34 aa_presence
28	0.01643	5 aa-6
29	0.01587	6 aa-5
30	0.01574	7 aa-4
31	0.01485	15 aa5
32	0.01452	9 aa-2
33	0.01423	10 aa-1
34	0.01396	1 aa-10
35	0.01269	20 aa10
36	0.01267	17 aa7
37	0.01195	4 aa-7
38	0.01017	13 aa3
39	0.01006	19 aa9
40	0.0096	11 aa1
41	0.00777	12 aa2
42	0.00747	16 aa6
43	0.00733	14 aa4
44	0.0054	18 aa8

Table S5. List of input protein and DNA features ranked by information gain on the PDB data collected.

Rank	Information Gain	H(Class) - H(Class | Attribute)
1	0.028965	74 aa_entropy
2	0.0254295	77 polarity_entropy
3	0.0163194	79 avg_Ph
4	0.0158909	78 avg_hydropathyIndex
5	0.0150191	80 avg_mass
6	0.0143294	85 aaMSAind
7	0.0115361	81 aa_blosum
8	0.0103267	83 aa_obsCount
9	0.0091675	33 aa_nt
10	0.0088123	67 aa_presence_total
11	0.0082388	31 aa
12	0.0080512	45 pH
13	0.007936	38 residueBurial
14	0.007936	42 sideChainBurial
15	0.0079193	41 sea30
16	0.0078905	44 volume
17	0.0078723	35 mass
18	0.0076877	43 surface
19	0.0075964	36 npsa
20	0.0074838	39 sea10
21	0.0072399	40 sea1030
22	0.0063895	34 hydropathyIndex
23	0.0059628	69 dna_presence_total
24	0.0059197	87 aaSeqPos
25	0.0053562	53 stat_aa_avgCorr
26	0.0051385	58 dna_totalCorr_count
27	0.0048563	66 aa_presence
28	0.0041991	37 polarity
29	0.0040737	8 aa-3
30	0.0040557	2 aa-9
31	0.0040007	84 nt_obsCount
32	0.0039398	88 dnaSeqPos
33	0.0037361	48 aa_totalCorr
34	0.0037097	5 aa-6
35	0.0034294	49 aa_totalCorr_count
36	0.0033629	6 aa-5
37	0.0033307	7 aa-4
38	0.0032274	3 aa-8
39	0.0030469	10 aa-1
40	0.0029861	9 aa-2
41	0.002966	15 aa5
42	0.002596	1 aa-10
43	0.0023513	4 aa-7
44	0.0023444	70 MI
45	0.0022853	75 nt_entropy
46	0.0022479	20 aa10
47	0.0021737	62 stat_dna_avgCorr
48	0.0021431	17 aa7
49	0.0021398	59 dna_avgCorr
50	0.0020943	11 aa1
51	0.0020257	65 stat_dna_avgCorr_negative
52	0.001982	46 corr
53	0.0019657	13 aa3
54	0.0019346	76 adv_nt_entropy
55	0.0018553	50 aa_avgCorr
56	0.0017471	12 aa2
57	0.0016383	82 nt_nuc44
58	0.0016214	19 aa9
59	0.0015827	14 aa4
60	0.0015735	86 dnaMSAind
61	0.0013169	60 stat_dna_totalCorr
62	0.0012158	68 dna_presence
63	0.0011903	57 dna_totalCorr
64	0.0010989	16 aa6
65	0.0010665	18 aa8
66	0.0010046	23 nt-3
67	0.0009317	24 nt-2
68	0.0009205	27 nt2
69	0.0009049	28 nt3
70	0.000891	51 stat_aa_totalCorr
71	0.000865	52 stat_aa_totalCorr_count
72	0.0008573	54 stat_aa_totalCorr_negative
73	0.0007884	26 nt1
74	0.0006983	25 nt-1
75	0.0006965	22 nt-4
76	0.0006128	47 pvalue
77	0.0006032	64 stat_dna_totalCorr_count_negative
78	0.0005763	29 nt4
79	0.0005462	63 stat_dna_totalCorr_negative
80	0.0005428	56 stat_aa_avgCorr_negative
81	0.0003977	21 nt-5
82	0.0003533	30 nt5
83	0.0002316	72 MIp
84	0.0002113	61 stat_dna_totalCorr_count
85	0.0001517	32 nt
86	0.0000885	55 stat_aa_totalCorr_count_negative

Table S6. Comparison between the top 25 scoring 8-mers predicted and the top 25 8-mers with the highest median binding intensities measured by Protein Binding Microarray (PBM) (3)

Predicted	PBM Replicate 1 (3)	PBM Replicate 2 (3)
’ACAGGTGC’	’ACACCTGC’	’ACAGGTGC’
’ACAGGTGG’	’GCAGGTGT’	’GCACCTGT’
’CCAGGTGC’	’CACCTGCA’	’CGCACCTG’
’CCAGGTGG’	’TGCAGGTG’	’CAGGTGCG’
’GCAGGTGC’	’CCACCTGC’	’CACCTGTG’
’GCAGGTGG’	’GCAGGTGG’	’CACAGGTG’
’TCAGGTGC’	’GCACCTGT’	’GCACCTGG’
’TCAGGTGG’	’ACAGGTGC’	’CCAGGTGC’
’ACATGTGC’	’AACACCTG’	’GCAGGTGT’
’ACATGTGG’	’CAGGTGTT’	’ACACCTGC’
’CCATGTGC’	’CACCTGCG’	’CGCAGGTG’
’CCATGTGG’	’CGCAGGTG’	’CACCTGCG’
’GCATGTGC’	’CGCACCTG’	’CACACCTG’
’GCATGTGG’	’CAGGTGCG’	’CAGGTGTG’
’TCATGTGC’	’CACCTGTG’	’GCAGGTGG’
’TCATGTGG’	’CACAGGTG’	’CCACCTGC’
’ACAGTTGC’	’CACACCTG’	’CAGGTGCT’
’ACAGTTGG’	’CAGGTGTG’	’AGCACCTG’
’CCAGTTGC’	’ACACCTGG’	’AACACCTG’
’CCAGTTGG’	’CCAGGTGT’	’CAGGTGTT’
’GCAGTTGC’	’GCACCTGC’	’CCAGGTGT’
’GCAGTTGG’	’GCAGGTGC’	’ACACCTGG’
’TCAGTTGC’	’TGCACCTG’	’GCAGGTGC’
’TCAGTTGG’	’CAGGTGCA’	’GCACCTGC’
’ACATTTGC’	’CACCTGCT’	’CACCTGGT’

REFERENCES

1. Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.,* **32**(5), 1792–1797.
2. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., GrifRths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C., and Eddy, S. R. (2004) The Pfam protein families database. *Nucleic Acids Res,* **32**, D138–141.
3. Robasky, K. and Bulyk, M. L. (Jan, 2011) UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.,* **39**, D124–128.