TargetDBP+: Enhancing the Performance of Identifying DNA-Binding Proteins via Weightedly Convolutional Features

Jun Hu [†,‡,*], Liang Rao [†], Yi-Heng Zhu [¶], Gui-Jun Zhang [†,*] and Dong-Jun Yu [¶,*]

† College of Information Engineering, Zhejiang University of Technology, Hangzhou, 310023, P. R. China, ‡ Key Laboratory of Data Science and Intelligence Application, Fujian Province University, Zhangzhou, 363000, China, and ¶ School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing, 210094, P. R. China

*Email: junh_cs@126.com, zgj@zjut.edu.cn, or njyudj@njust.edu.cn

Support Information

Supporting Texts

Text S1: *UniSwiss* dataset

The *UniSwiss* dataset contains two subset, i.e., the training dataset (*UniSwiss-Tr*) and independent validation dataset (*UniSwiss-Tst*). Both *UniSwiss-Tr* and *UniSwiss-Tst* include many DBPs (DNA-binding proteins) and non-DBPs (non-DNA-binding proteins). Here, the set consisting of DBPs in *UniSwiss-Tr* is denoted as *Tr-P* and the set consisting of non-DBPs in *UniSwiss-Tr* is named as *Tr-N*. Similarity, the set consisting of DBPs and the set consisting of non-DBPs in *UniSwiss-Tst* are named as *Tst-P* and *Tst-N*, respectively.

To construct the union set (denoted as *P*) of *Tr-P* and *Tst-P*, the 48,109 DBPs in UniProtKB/Swiss-Prot[1] (up to May 5, 2020), which is a manually annotated database, are first downloaded. Then the CD-hit software[2] is used to remove the redundant protein sequences such that no two sequences have more than 25 percent sequence identity. To ensure no fragment in the final dataset, any sequence with less than 50 residues in length is removed. Finally, a total of 6,651 non-redundant protein sequences are gained to form *P*. We randomly selected 6,000 sequences from *P* to compose *Tr-P*. The remaining 651 sequences in *P* are used to construct *Tst-P*.

Similarity, in order to construct *Tr-N* and *Tst-N*, all 512,312 non-DBPs in UniProtKB/Swiss-Prot (up to May 5, 2020) are first collected. Then the CD-hit software[2] is used to remove the redundant protein sequences such that no two sequences have more than 25 percent sequence identity. To ensure no fragment in the final dataset, any sequence with less than 50 residues in length is removed. We denoted the remaining non-redundant non-DBPs as *TotalNega*. Finally, we randomly selected 6,000 and 651 non-DBPs from *TotalNega* to compose *Tr-N* and *Tst-N*, respectively.

After *Tr-P*, *Tst-P*, *Tr-N*, and *Tst-N* are collected, UniSwiss, including *UniSwiss-Tr* and UniSwiss-Tst, could be easily constructed. The *UniSwiss* dataset is downloadable at https://github.com/jun-csbio/TargetDBPplus/.

Supporting Tables

Table S1. The *p*-values in Wilcoxon signed rank *t*-test for the difference in the outputted probabilities of belonging to the class of DBPs between different convolutional sliding weight windows on *UniSwiss-Tr* over 10-fold cross-validation tests

#	$\mathbf{w}_{0.1.0}$	$\mathbf{w}_{1.1.1}$	$\mathbf{w}_{1.5.1}$	$\mathbf{w}_{1.9.1}$	$\mathbf{w}_{1.13.1}$	$\mathbf{w}_{1.17.1}$	$\mathbf{w}_{1.21.1}$
$\mathbf{w}_{0.1.0}$		0.7515	0.0018	6.3842e-05	8.0701e-04	0.1771	0.2300
$\mathbf{w}_{1.1.1}$			0.1725	0.1079	0.3671	0.9977	0.8786
$\mathbf{w}_{1.5.1}$				0.0596	0.7278	9.1948e-04	7.4216e-04
$\mathbf{w}_{1.9.1}$					4.3551e-07	3.7731e-04	2.0705e-04
$\mathbf{w}_{1.13.1}$						0.0019	0.0015
$\mathbf{w}_{1.17.1}$							0.1531
$\mathbf{w}_{1.21.1}$							

Table S2. The *p*-values in Wilcoxon signed rank *t*-test for the difference in the outputted probabilities of belonging to the class of DBPs between different convolutional sliding weight windows on *UniSwiss-Tr* over jackknife tests

#	$\mathbf{w}_{0.1.0}$	$\mathbf{w}_{1.1.1}$	$\mathbf{w}_{1.5.1}$	$\mathbf{w}_{1.9.1}$	$\mathbf{w}_{1.13.1}$	$\mathbf{w}_{1.17.1}$	$\mathbf{w}_{1.21.1}$
$\mathbf{w}_{0.1.0}$		0.4865	0.0346	8.8253e-04	0.0024	0.7157	0.6756
$\mathbf{w}_{1.1.1}$			0.5485	0.2631	0.5248	0.5687	0.5350
$\mathbf{w}_{1.5.1}$				0.0124	0.2250	0.0014	0.0044
$\mathbf{w}_{1.9.1}$					2.0988e-04	2.7128e-04	1.3400e-04
$\mathbf{w}_{1.13.1}$						6.5060e-04	2.2187e-04
$\mathbf{w}_{1.17.1}$							0.2470
$\mathbf{w}_{1.21.1}$							

Table S3. Performance comparisons of different values of *G* on *UniSwiss-Tr* over ten-fold cross-validation test.

G	*Sen*	*Spe*	*Acc*	*Pre*	*MCC*	F_1	*AUC*
1	74.62	86.63	80.62	84.81	0.617	0.794	0.881
2	77.23	83.57	80.4	82.46	0.609	0.798	0.875
3	74.52	86.00	80.26	84.18	0.609	0.791	0.882
4	75.47	85.40	80.43	83.79	0.612	0.794	0.885
5	75.15	85.43	80.29	83.76	0.609	0.792	0.884
6	80.10	81.53	80.82	81.26	0.616	0.807	0.885
7	79.65	81.82	80.73	81.41	0.615	0.805	0.885
8	78.52	82.25	80.38	81.56	0.608	0.8	0.884
9	75.15	85.43	80.29	83.76	0.609	0.792	0.884
10	77.48	83.23	80.36	82.21	0.608	0.798	0.883
11	78.55	83.32	80.93	82.48	0.619	0.805	0.882
12	78.10	83.72	80.91	82.75	0.619	0.804	0.881
13	78.30	82.52	80.41	81.75	0.609	0.8	0.878
14	72.93	86.8	79.87	84.67	0.603	0.784	0.876
15	74.65	85.42	80.03	83.66	0.604	0.789	0.875
16	73.82	86.05	79.93	84.11	0.603	0.786	0.875
17	74.33	86.48	80.41	84.61	0.613	0.791	0.876
18	76.07	85.58	80.83	84.07	0.619	0.799	0.879
19	74.88	84.87	79.88	83.19	0.600	0.788	0.879
20	75.65	85.55	80.60	83.96	0.615	0.796	0.878
21	78.63	81.55	80.09	81.00	0.602	0.798	0.881

Table S4. Performance comparisons of different values of *G* on *UniSwiss-Tr* over jackknife test.

G	*Sen*	*Spe*	*Acc*	*Pre*	*MCC*	F_1	*AUC*
1	72.77	88.35	80.56	86.2	0.619	0.789	0.882
2	74.7	86.58	80.64	84.77	0.617	0.794	0.876
3	75.77	85.43	80.6	83.87	0.615	0.796	0.883
4	73.22	86.38	79.8	84.32	0.601	0.784	0.875
5	78.22	83.08	80.65	82.22	0.614	0.802	0.885
6	77.85	83.78	80.82	82.76	0.617	0.802	0.886
7	79.55	82.18	80.87	81.7	0.618	0.806	0.885
8	77.08	83.98	80.53	82.8	0.612	0.798	0.884
9	77.93	83.27	80.6	82.32	0.613	0.801	0.884
10	78.22	83.05	80.63	82.19	0.613	0.802	0.883
11	77.23	84.83	81.03	83.59	0.622	0.803	0.882
12	76.95	85.07	81.01	83.75	0.622	0.802	0.882
13	77.2	83.93	80.57	82.77	0.613	0.799	0.879
14	75.78	84.6	80.19	83.11	0.606	0.793	0.877
15	75.72	84.82	80.27	83.3	0.608	0.793	0.876
16	76.45	84.02	80.23	82.71	0.606	0.795	0.876
17	74.3	87.32	80.81	85.42	0.621	0.795	0.877
18	76.45	85.6	81.03	84.15	0.623	0.801	0.880
19	74.12	85.9	80.01	84.02	0.604	0.788	0.880
20	77.45	82.95	80.2	81.96	0.605	0.796	0.882
21	77.97	82.47	80.22	81.64	0.605	0.798	0.881

Table S5. The *p*-values in Wilcoxon signed rank *t*-test for the difference in the outputted probabilities of belonging to the class of DBPs between different convolutional sliding weight windows on *UniSwiss-Tr* over jackknife tests

Feature	PseConvAAOHM	PseConvPSSM	PseConvPSS	PseConvPSA	PseConvPPDBS	SerialCom	WSerialCom
PseConvAAOHM		0.0938	0.0014	0.0069	0.2652	0.3513	0.0440
PseConvPSSM			2.9684e-04	0.0144	0.6302	9.2878e-12	0.0013
PseConvPSS				0.2384	0.2946	1.6418e-07	2.4341e-04
PseConvPSA					0.6154	1.2229e-05	0.0020
PseConvPPDBS						0.9296	0.7689
SerialCom							1.0913e-04
WSerialCom							

Supporting Figures

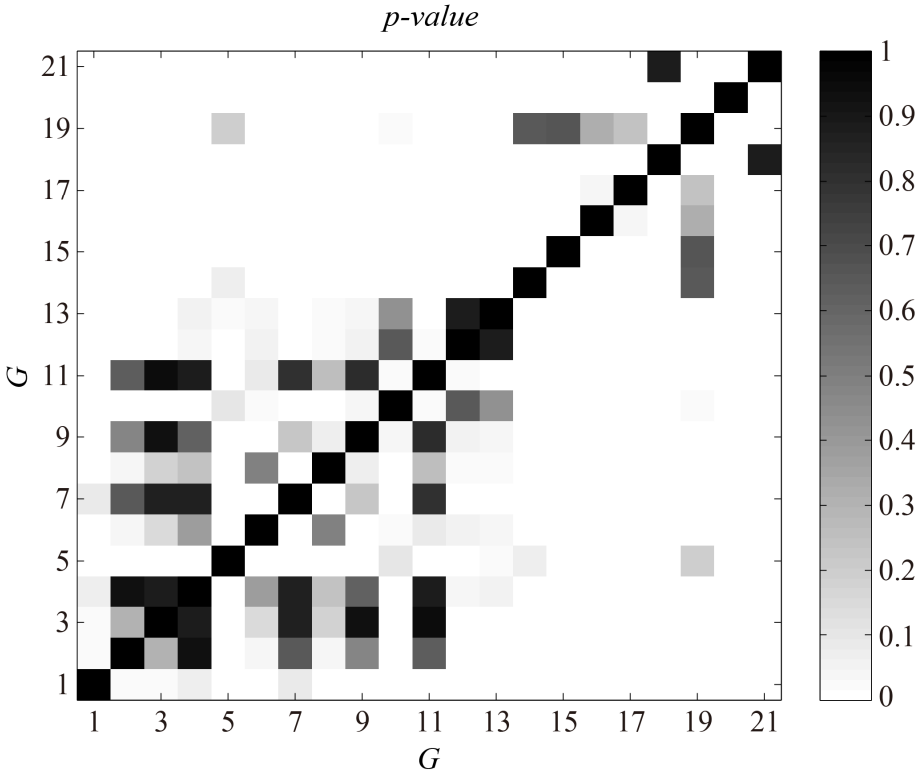

Figure S1. The *p*-values in Wilcoxon signed rank *t*-test for the difference in the outputted probabilities of belonging to the class of DBPs between different values of *G* on *UniSwiss-Tr* over jackknife tests

References

(1) Boutet, E.; Lieberherr, D.; Tognolli, M.; Schneider, M.; Bansal, P.; Bridge, A. J.; Poux, S.; Bougueleret, L.; Xenarios, I., UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. *Plant bioinformatics: methods and protocols* **2016**, 23-54.

(2) Li, W.; Godzik, A., Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, 22, 1658-1659.