**SNB-PSSM: a spatial neighbor based PSSM used for protein-RNA binding site prediction**

Yang Liu, Weikang Gong, Zhen Yang, Chunhua Li \*

Faculty of Environmental and Life Sciences, Beijing University of Technology, Beijing 100124, China

\*All correspondence should be addressed to Chunhua Li (E-mail: chunhuali@bjut.edu.cn)

Table S1. Numbers of binding and non-binding residues in RB198, RB44 and RB111 datasets.

|  |  |  |
| --- | --- | --- |
| Dataset | Num. of binding residues | Num. of non-binding residues |
| RB198 | 7878 | 43150 |
| RB44 | 1956 | 4521 |
| RB111 | 3305 | 34255 |

Table S2. Performance comparison of standard PSSM, smoothed PSSM and SNB-PSSM on RB198 dataset (5-fold cross-validation).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *SN* | *SP* | *ACC* | *MCC* |
| Standard PSSM | 0.48±0.01 | 0.87±0.02 | 0.73±0.01 | 0.38±0.01 |
| Smoothed PSSM | 0.51±0.03 | 0.86±0.02 | 0.74±0.02 | 0.40±0.02 |
| SNB-PSSM | 0.50±0.03 | 0.88±0.02 | 0.75±0.02 | 0.42±0.03 |

Table S3. Test comparison of the trained standard PSSM, smoothed PSSM and SNB-PSSM methods on RB44 dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *SN* | *SP* | *ACC* | *MCC* |
| Standard PSSM | 0.48 | 0.75 | 0.67 | 0.23 |
| Smoothed PSSM | **0.77** | 0.52 | 0.62 | 0.27 |
| SNB-PSSM | 0.56 | **0.76** | **0.70** | **0.32** |

Table S4. Test comparison of the trained standard PSSM and smoothed PSSM based sequence window (size of 7) methods respectively and the trained SNB-PSSM based structure window method with the same size on RB44 dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Windowed method | *SN* | *SP* | *ACC* | *MCC* |
| Standard PSSM based | 0.62 | 0.68 | 0.66 | 0.30 |
| Smoothed PSSM based | **0.81** | 0.51 | 0.61 | 0.32 |
| SNB-PSSM based | 0.59 | **0.75** | **0.69** | **0.34** |

**Window size selection for SNB-PSSM based structure window method**

In order to optimize the performance of the SNB-PSSM based structure window method, the size of window needs to be set properly. Table S5 gives the test performances of different window size sets on RB44 (see Table S6 for details). From Table S5, roughly, with the increase of the structure window size, the better prediction can be achieved. Thus, the half of the shortest length of protein sequences 25 was adopted as the window size. When the window size is set to 25, the method attains the better prediction with *SN*, *SP*, *ACC* and *MCC* of 0.64, 0.78, 0.74 and 0.40 respectively. Therefore, our method applies the SNB-PSSM based structure window method with window size 25.

Table S5. Test of the SNB-PSSM based structure window method with different window sizes on RB44 dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Window size | *SN* | *SP* | *ACC* | *MCC* |
| 7 | 0.59 | 0.75 | 0.69 | 0.34 |
| 9 | 0.65 | 0.70 | 0.68 | 0.35 |
| 11 | **0.67** | 0.69 | 0.68 | 0.34 |
| 13 | 0.67 | 0.70 | 0.69 | 0.35 |
| 15 | 0.61 | 0.76 | 0.71 | 0.37 |
| 17 | 0.62 | 0.76 | 0.72 | 0.37 |
| 19 | 0.63 | 0.75 | 0.72 | 0.38 |
| 21 | 0.65 | 0.77 | 0.73 | 0.39 |
| 23 | 0.64 | 0.77 | 0.73 | 0.39 |
| 25 | 0.64 | **0.78** | **0.74** | **0.40** |

Table S6. Test of our trained models (5-cross validation) with different window sizes on RB44 dataset.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Window size | cost parameter C | kernel function parameter γ | *MCC*1 | *MCC*2 | *MCC*3 | *MCC*4 | *MCC*5 |
| 7 | 10 | 0.0005 | 0.37 | 0.37 | 0.35 | 0.37 | 0.26 |
| 9 | 10 | 0.0002 | 0.38 | 0.37 | 0.34 | 0.37 | 0.27 |
| 11 | 10 | 0.00015 | 0.38 | 0.36 | 0.33 | 0.38 | 0.28 |
| 13 | 10 | 0.000096 | 0.38 | 0.37 | 0.35 | 0.38 | 0.28 |
| 15 | 10 | 0.000089 | 0.40 | 0.39 | 0.36 | 0.39 | 0.30 |
| 17 | 10 | 0.000089 | 0.40 | 0.39 | 0.37 | 0.39 | 0.31 |
| 19 | 8 | 0.000094 | 0.41 | 0.40 | 0.37 | 0.40 | 0.32 |
| 21 | 8 | 0.000096 | 0.41 | 0.41 | 0.38 | 0.41 | 0.33 |
| 23 | 8 | 0.000094 | 0.43 | 0.41 | 0.39 | 0.41 | 0.33 |
| 25 | 8 | 0.000096 | 0.45 | 0.43 | 0.39 | 0.41 | 0.34 |

Table S7. Descriptions on the RNA-binding site prediction servers and the associated references.

|  |  |  |
| --- | --- | --- |
| Method | Description | Reference |
| FastRNABindR | An SVM model using the standard PSSM based sequence window (size of 25). Accessible at: http://ailab.ist.psu.edu/FastRNABindR/ | [1] |
| RNABindR v2 | An SVM model using the standard PSSM based sequence window (size of 21). Accessible at: http://einstein.cs.iastate.edu/RNABindR/ | [2] |
| BindN+ | An updated version of BindN using an SVM model based on PSSM and three biochemical features of amino acids. Accessible at: http://bioinfo.ggc.org/bindn+/ | [3] |
| PPRInt | An SVM model using PSSM. Accessible at: http://www.imtech.res.in/raghava/pprint/ | [4] |
| KYG | Uses a set of scores based on the RNA-binding propensity of individual and pairs of surface residues of the protein, used alone or in combination with PSSM. Accessible at: http://cib.cf.ocha.ac.jp/KYG/.  | [5] |
| PRIP | Uses a set of scores based on the RNA-binding propensity of individual and pairs of surface residues of the protein, used alone or in combination with PSSM. Not accessible via the web server, but results can be obtained via correspondence with the author. | [6] |
| SpaPF | Uses a novel profile of a target residue encoded into a vector of 20×15 elements based on the PSSM with a window of 15 spatially neighboring residues, and SVM classifier algorithm to predict protein-RNA interaction sites  | [7] |

Table S8. Improvement percentages of cases for our method compared with other ones on proteins in RB111 that are classified into four categories according to their chain lengths.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | Short (22) | Medium (56) | Medium long (23) | Long (10) |
| FastRNABindR | 95.4% | 96.4% | 100% | 100% |
| RNABindR v2 | 100% | 96.4% | 100% | 100% |
| SpaPF | 100% | 100% | 100% | 100% |
| BindN+ | - | - | - | - |
| PPRInt | - | - | - | - |
| KYG | - | - | - | - |
| PRIP | - | - | - | - |

“-” means the method is not available and the results can not be obtained.

Table S9. Performance comparison (with *MCC* value) of our method with others on each protein in RB111

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| PDB ID | Our method | FastRNABindR | RNABindRv2 | SpaPF |
| Short (length ≤100) |
| 2XFZ\_Y | 0.54 | 0.52 | 0.51 | 0.48 |
| 2XS2\_A | 0.57 | 0.55 | 0.52 | 0.49 |
| 3MOJ\_B | 0.33 | 0.34 | 0.32 | 0.32 |
| 3QSY\_B | 0.22 | 0.21 | 0.19 | 0.17 |
| 3V22\_V | 0.18 | 0.17 | 0.15 | 0.14 |
| 3V2C\_Y | 0.32 | 0.31 | 0.29 | 0.27 |
| 3VYX\_A | 0.48 | 0.46 | 0.42 | 0.39 |
| 3WBM\_A | 0.58 | 0.56 | 0.51 | 0.47 |
| 4BTD\_2 | 0.16 | 0.15 | 0.13 | 0.12 |
| 4BTD\_S | 0.40 | 0.39 | 0.36 | 0.34 |
| 4BTD\_X | 0.48 | 0.45 | 0.42 | 0.41 |
| 4BYB\_Y | 0.45 | 0.43 | 0.41 | 0.37 |
| 4I67\_A | 0.52 | 0.48 | 0.49 | 0.46 |
| 4KJ5\_0 | 0.37 | 0.35 | 0.32 | 0.31 |
| 4KJ5\_X | 0.47 | 0.44 | 0.41 | 0.38 |
| 4KJI\_A | 0.27 | 0.25 | 0.23 | 0.22 |
| 4L8R\_C | 0.18 | 0.16 | 0.15 | 0.13 |
| 4LJ0\_A | 0.47 | 0.43 | 0.41 | 0.39 |
| 4M7D\_A | 0.36 | 0.32 | 0.31 | 0.29 |
| 4M7D\_B | 0.56 | 0.54 | 0.53 | 0.48 |
| 4M7D\_C | 0.03 | 0.02 | 0.02 | -0.02 |
| 4M7D\_G | 0.39 | 0.35 | 0.32 | 0.31 |
| Medium (100<length≤400) |
| 2XLK\_A | 0.62 | 0.6 | 0.58 | 0.53 |
| 2Y8Y\_A | 0.22 | 0.21 | 0.19 | 0.15 |
| 3ADD\_A | 0.34 | 0.33 | 0.31 | 0.3 |
| 3AGV\_A | 0.74 | 0.72 | 0.69 | 0.61 |
| 3B0U\_X | 0.37 | 0.35 | 0.34 | 0.33 |
| 3KFU\_A | 0.37 | 0.36 | 0.32 | 0.31 |
| 3LWR\_A | 0.35 | 0.34 | 0.32 | 0.31 |
| 3NMR\_A | 0.48 | 0.46 | 0.43 | 0.41 |
| 3NVI\_A | 0.33 | 0.31 | 0.29 | 0.25 |
| 3OIN\_A | 0.29 | 0.27 | 0.25 | 0.21 |
| 3QJL\_A | 0.18 | 0.17 | 0.14 | 0.13 |
| 3R2C\_A | 0.53 | 0.54 | 0.54 | 0.51 |
| 3R9X\_B | 0.43 | 0.42 | 0.39 | 0.35 |
| 3RTJ\_A | 0.32 | 0.31 | 0.29 | 0.25 |
| 3RW6\_A | 0.41 | 0.4 | 0.37 | 0.35 |
| 3T5N\_A | 0.28 | 0.27 | 0.25 | 0.23 |
| 3U2E\_A | 0.17 | 0.19 | 0.15 | 0.14 |
| 3ULD\_A | 0.36 | 0.35 | 0.32 | 0.31 |
| 3VJR\_A | 0.23 | 0.21 | 0.18 | 0.17 |
| 3ZC0\_A | 0.15 | 0.14 | 0.11 | 0.11 |
| 4AFY\_A | 0.18 | 0.16 | 0.13 | 0.12 |
| 4ATO\_A | 0.31 | 0.3 | 0.28 | 0.27 |
| 4BA2\_A | 0.08 | 0.07 | 0.05 | 0.04 |
| 4BTD\_D | 0.55 | 0.53 | 0.5 | 0.47 |
| 4BTD\_G | 0.36 | 0.34 | 0.33 | 0.31 |
| 4C9D\_A | 0.31 | 0.3 | 0.26 | 0.24 |
| 4DH9\_Y | 0.78 | 0.65 | 0.61 | 0.57 |
| 4DWA\_A | 0.39 | 0.34 | 0.31 | 0.28 |
| 4ERD\_A | 0.59 | 0.54 | 0.51 | 0.47 |
| 4F02\_A | 0.42 | 0.41 | 0.38 | 0.35 |
| 4G0A\_A | 0.26 | 0.24 | 0.23 | 0.2 |
| 4GV3\_A | 0.49 | 0.47 | 0.45 | 0.42 |
| 4H5P\_A | 0.24 | 0.22 | 0.2 | 0.19 |
| 4H8K\_A | 0.22 | 0.2 | 0.18 | 0.17 |
| 4HUB\_G | 0.28 | 0.25 | 0.24 | 0.21 |
| 4IFD\_A | 0.05 | 0.03 | 0.01 | 0.01 |
| 4IFD\_C | 0.21 | 0.18 | 0.14 | 0.13 |
| 4IFD\_G | 0.43 | 0.39 | 0.35 | 0.34 |
| 4IFD\_H | 0.12 | 0.11 | 0.09 | 0.07 |
| 4IG8\_A | 0.30 | 0.28 | 0.27 | 0.25 |
| 4II9\_A | 0.17 | 0.15 | 0.14 | 0.12 |
| 4ILL\_A | 0.24 | 0.21 | 0.19 | 0.14 |
| 4J1G\_A | 0.18 | 0.16 | 0.14 | 0.11 |
| 4J7M\_A | 0.22 | 0.21 | 0.17 | 0.15 |
| 4JVY\_A | 0.15 | 0.14 | 0.11 | 0.08 |
| 4JZU\_A | 0.04 | 0.03 | 0.01 | 0 |
| 4KJ5\_5 | 0.22 | 0.21 | 0.21 | 0.19 |
| 4KJ5\_G | 0.33 | 0.31 | 0.28 | 0.25 |
| 4L8H\_A | 0.47 | 0.44 | 0.42 | 0.41 |
| 4LGT\_A | 0.36 | 0.35 | 0.34 | 0.32 |
| 4M4O\_A | 0.26 | 0.25 | 0.24 | 0.21 |
| 4MDX\_A | 0.60 | 0.58 | 0.52 | 0.49 |
| 4N2Q\_A | 0.08 | 0.07 | 0.05 | 0.04 |
| 4NGD\_A | 0.48 | 0.45 | 0.44 | 0.41 |
| 4NKU\_A | 0.28 | 0.24 | 0.23 | 0.21 |
| 4OL8\_A | 0.48 | 0.43 | 0.41 | 0.38 |
| Medium long (400<length≤800) |
| 3AMT\_A | 0.22 | 0.21 | 0.19 | 0.17 |
| 3KFU\_E | 0.07 | 0.05 | 0.04 | 0.02 |
| 3KFU\_F | 0.29 | 0.28 | 0.26 | 0.23 |
| 3QSY\_A | 0.28 | 0.27 | 0.25 | 0.24 |
| 3RC8\_A | 0.30 | 0.29 | 0.27 | 0.26 |
| 3SFS\_W | 0.38 | 0.36 | 0.34 | 0.32 |
| 3T3O\_A | 0.16 | 0.15 | 0.13 | 0.13 |
| 3TUP\_A | 0.28 | 0.26 | 0.24 | 0.21 |
| 3W3S\_A | 0.34 | 0.32 | 0.29 | 0.27 |
| 3WFR\_E | 0.20 | 0.19 | 0.15 | 0.13 |
| 3ZD6\_A | 0.06 | 0.04 | 0.02 | 0.01 |
| 4AM3\_A | 0.50 | 0.47 | 0.42 | 0.39 |
| 4B3G\_A | 0.18 | 0.16 | 0.15 | 0.13 |
| 4B3O\_A | 0.24 | 0.23 | 0.21 | 0.19 |
| 4BWM\_A | 0.33 | 0.31 | 0.28 | 0.23 |
| 4E78\_A | 0.58 | 0.52 | 0.49 | 0.46 |
| 4GG4\_A | 0.26 | 0.23 | 0.21 | 0.19 |
| 4HKQ\_A | 0.38 | 0.36 | 0.35 | 0.33 |
| 4HOR\_A | 0.23 | 0.21 | 0.19 | 0.17 |
| 4K4Z\_A | 0.22 | 0.21 | 0.19 | 0.17 |
| 4KJ6\_V | 0.48 | 0.46 | 0.45 | 0.43 |
| 4M59\_A | 0.02 | 0.01 | 0.01 | -0.01 |
| 4N48\_A | 0.17 | 0.16 | 0.13 | 0.11 |
| Long (length>800) |
| 2XGJ\_A | 0.18 | 0.17 | 0.16 | 0.13 |
| 3S14\_A | 0.10 | 0.09 | 0.07 | 0.06 |
| 3S14\_B | 0.16 | 0.15 | 0.13 | 0.11 |
| 4ARC\_A | 0.27 | 0.26 | 0.24 | 0.21 |
| 4BOC\_A | 0.12 | 0.1 | 0.09 | 0.08 |
| 4F1N\_A | 0.21 | 0.19 | 0.16 | 0.15 |
| 4FXD\_A | 0.15 | 0.12 | 0.11 | 0.08 |
| 4G7O\_C | 0.34 | 0.31 | 0.29 | 0.27 |
| 4G7O\_D | 0.08 | 0.05 | 0.03 | 0.01 |
| 4OO8\_A | 0.33 | 0.31 | 0.29 | 0.27 |



Figure S1. Distribution of TM-score values of the modelled structures from RB111 datasets.

**Reference**

[1] El-Manzalawy Y, Abbas M, Malluhi Q, et al. FastRNABindR: Fast and Accurate Prediction of Protein-RNA Interface Residues[J]. PLoS One. 2016, 11(7): e158445.

[2] Walia R R, Xue L C, Wilkins K, et al. RNABindRPlus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins[J]. PLoS One. 2014, 9(5): e97725.

[3] Wang L, Huang C, Yang M Q, et al. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features[J]. BMC Syst Biol. 2010, 4 Suppl 1: S3.

[4] Kumar M, Gromiha M M, Raghava G P. Prediction of RNA binding sites in a protein using SVM and PSSM profile[J]. Proteins. 2008, 71(1): 189-194.

[5] Kim O T, Yura K, Go N. Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction[J]. Nucleic Acids Res. 2006, 34(22): 6450-6460.

[6] Maetschke S R, Yuan Z. Exploiting structural and topological information to improve prediction of RNA-protein binding sites[J]. BMC Bioinformatics. 2009, 10: 341.

[7] Chen W, Zhang S W, Cheng Y M, et al. Identification of protein-RNA interaction sites using the information of spatial adjacent residues[J]. Proteome Sci. 2011, 9 Suppl 1: S16.