W562–W567 *Nucleic Acids Research, 2016, Vol. 44, Web Server issue* *Published online 15 April 2016*
doi: 10.1093/nar/gkw251

RBscore&NBench: a high-level web server for nucleic acid binding residues prediction with a large-scale benchmarking database

Zhichao Miao* **and Eric Westhof***

Architecture et Réactivité de l'ARN, Université de Strasbourg, Institut de biologie moléculaire et cellulaire du CNRS, 15 Rue Descartes 67000 Strasbourg, France

Received January 12, 2016; Revised March 13, 2016; Accepted April 2, 2016

ABSTRACT

RBscore&NBench combines a web server, RBscore and a database, NBench. RBscore predicts RNA-/DNA-binding residues in proteins and visualizes the prediction scores and features on protein structures. The scoring scheme of RBscore directly links feature values to nucleic acid binding probabilities and illustrates the nucleic acid binding energy funnel on the protein surface. To avoid dataset, binding site definition and assessment metric biases, we compared RBscore with 18 web servers and 3 stand-alone programs on 41 datasets, which demonstrated the high and stable accuracy of RBscore. A comprehensive comparison led us to develop a benchmark database named NBench. The web server is available on: http://ahsoka.u-strasbg.fr/rbscorenbench/.

INTRODUCTION

RNA– and DNA–protein interactions occur in a large amount of biological processes. The computational prediction of nucleic binding residues on protein is an important step in understanding protein functions. Although the problem of binding site prediction is old (1), the prediction algorithms are not so enlightening and effective as expected.

Recently we developed RBscore (2), which linearly correlates feature values to nucleic acid binding probability in a residue neighboring network approach in order to predict nucleic acid binding residues. RBscore displays merits in several aspects. (i) Physicochemical and evolutionary features (electrostatics, solvation energy and conservation entropy) can be directly related to nucleic acid binding probabilities. (ii) RBscore, standing for RNA Binding score, was trained on RNA binding proteins (RBP) but demonstrates even higher accuracies for DNA binding residue prediction. This underscores, firstly, that RBP and DNA binding proteins (DBP) employ common driving forces in their binding to nucleic acid and, secondly, that RBscore has captured successfully the main factors controlling the binding propensities. There are more than 100 cases in the PDB (3) that report the formation of protein–RNA–DNA complexes. Recently, it has been estimated that about 2% of the human proteome may bind both RNA and DNA, and such proteins are named DRBP (4). Thus, it is now a major challenge to be able to predict nucleic acid specificities on the basis of the protein alone. (iii) One can plot the nucleic acid binding energy funnel on the protein surface using RBscore, with the residues closer to the nucleic acid binding region having higher prediction scores. This shows that a binary classification of binding sites (binding versus non-binding) is far from enough for the binding site prediction problem. Besides, RBscore demonstrates a high-level accuracy with a high stability in the accuracy on all DBP and RBP datasets. It guarantees also a weighted arithmetic mean of Area Under ROC Curve (wAUC) above 0.81, which cannot be achieved by other predictors. Since the running of the RBscore web server starting from June 2014, it has handled more than 5000 jobs submitted by more than 35 users from 11 different countries.

With many strategies to predict RNA- or DNA-binding sites on protein reported every year, fair and rigorous benchmarking is a laborious necessity. However, many programs were only assessed by cross-validation in small-scale datasets and did not fully demonstrate their predictive abilities. Current assessments of binding site prediction programs differ in: (i) the definition of nucleic acid binding sites by distance cutoffs; (ii) the training and test datasets that can induce dataset bias and (iii) the criteria to measure prediction performance. The comparisons in many of the reported works were only based on single cutoff, binary criteria and small-scale datasets validations, which may include bias toward certain methods and lead to dangerous conclusions.

Together with RBscore, we assessed 18 web servers and 3 stand-alone programs, 25 different approaches in total, on 41 different datasets, including more than 5000 protein chains derived from 3D structures of protein–nucleic acid

*To whom correspondence should be addressed. Tel: +33 0388417047; Email: z.miao@ibmc-cnrs.unistra.fr
Correspondence may also be addressed to Eric Westhof. Tel: +33 0388417046; Fax: +33 0388602218; Email: e.westhof@ibmc-cnrs.unistra.fr

complexes. The results demonstrate: (i) dataset bias and distance cutoff bias in binding site definition exist in some methods; (ii) DBP and RBP appear to follow similar driving forces which have been captured by some of the methods; (iii) the predictors have been greatly improved over the years, but there is still room for sorting out the essential mechanisms of binding for sequence based approaches. According to the results, RBscore also rank as a top-level predictor, which shows high but stable accuracies on all datasets regardless of distance cutoff used to define binding sites.

This assessment work led us to build a database, NBench (5), to benchmark all the prediction programs and to provide all the prediction result data to the scientific community. Hopefully, later development of new nucleic acid binding site prediction programs can take the advantage of NBench database and make direct comparison with the existing 25 approaches to avoid unnecessary bias of dataset, binding site definition or assessment metrics.

RBscore WEB SERVER

RBscore

Several normalization steps were processed before calculating RBscore: (i) alternative locations are cleaned leaving only the first state; (ii) residues with incomplete backbone atoms (N, CA and C) are dropped; (iii) Selenomethionines are taken as methionine, while other HETATM lines in the file are deleted; (iv) incomplete side-chain atoms are predicted by RASP (6,7). The calculation of RBscore is based on electrostatics potential, solvation energy and sequence conservation entropy feature values. Electrostatics potential is measured by programs pdb2pqr (8,9) and APBS (10) in a similar way as PatchFinderPlus (11); implicit model of solvation energy is derived from accessible surface area calculated by NACCESS (12), while sequence conservation entropy is measured by Shannon entropy (13) using Weblogo (14) according to the multiple sequence alignment (MSA) generated by HHblits (15). The program DMS (16) is used to generated the surface grid around the protein surface and to define the residue level neighboring interaction network. A total of 104 weighing factors were trained in the model. Details of the RBscore scoring function were described previously (2).

Besides the structure-based binding site predictor RBscore, we also provide a sequence-based predictor RBscore_SVM, which exploits the facility of machine learning approach, when no structure information is provided. RBscore_SVM first searches a small sequence database of Uniprot (17) for homologous sequences to derive Position Specific Scoring Matrix (PSSM) with PSI-BLAST (18), and then uses a slide-window approach to generate input feature vector for support vector machine (19) (SVM) to build the prediction model. Similar to other sequence-based predictors, RBscore_SVM is less stable in accuracy (5), more dataset dependent, and more distance cutoff dependent and without the ability to depict the binding energy funnel on protein surface. However, it still achieves top-level accuracy among sequence-based predictors, which guarantees a wAUC value >0.71 offering a still good choice when only sequence information is available.

Input

The web server integrates structure-based predictor RBscore and sequence-based predictor RBscore_SVM. When the protein structure is available, the prediction accuracy is generally better than only sequence information as well as more information can be visualized. The input of protein structure can either be a four-letter PDB code or by uploading a PDB formatted file.

When only protein sequence is available, users need to input FASTA formatted sequences or upload a FASTA file of sequences. More than one sequence in the same FASTA file is possible and the prediction results will be in the same order. There are optional parameter settings for the RBscore_SVM model: (i) to set the specificity or sensitivity of the prediction to determine the cutoff value used in binary prediction of binding sites. By default, the specificity is set to 85%; (ii) SVM models used in prediction. The models are derived from RBP or DBP alone or both RBP and DBP. According to the results in NBench, the model trained on both RBP and DBP dataset achieves highest accuracy. Finally, users can input an email address to receive the results after the job finish. Otherwise, the web page is automatically refreshed to show the running log until the prediction results are generated.

Output

The output of the RBscore includes three sections:

(i) Summary of prediction (Figure 1A). It provides basic information of the protein and of the prediction results, including job ID, length of the protein, rough estimation of nucleic acid binding site number based on protein sequence and download link of all the results. For RBscore_SVM prediction, the summary lists the estimated sensitivity, specificity, the resulted threshold for SVM classifier and the predicted number of binding sites.

(ii) Plots of prediction scores and feature scores on protein. The graphical plots of the prediction scores on protein surface are illustrated by JSmol (20) in rainbow color, where blue shows the region with highest prediction scores that most likely to bind nucleic acid, while red shows the least likely parts. Together with RBscore, RBscore predictions based on single features (electrostatics, conservation or solvation energy) as well as feature values of conservation entropy and electrostatics potential are also provided for plotting on protein surface. Feature value plots are similar to CONSURF (21) and APBS tools in pymol (22). Comparing feature values with RBscore plotted on protein, users can intuitively verify the prediction of nucleic acid binding region. Conceptually, nucleic acid are more likely to bind positively charged parts on a protein, since nucleic acids are normally negatively changed resulted from the phosphate group, while functional sites are more conserved than other residues to maintain the function in evolution. Positively charged region and conserved region are also plotted as blue to correspond to the high RBscore region. Additionally, users can load

A

Summary of prediction

Job ID:	1k8wA
Sequence length:	303 residues
Possible binding sites No.:	37 - 59
Download results data:	download
Prediction by RBscore:	
Defined specificity:	85.00%
Estimated sensitivity:	76.31%
Binding sites Number:	43 residues
Threshold in RBscore:	341.5
Prediction by SVM:	
Defined specificity:	85.00%
Estimated sensitivity:	83.46%
Binding sites Number:	39 residues
Threshold in SVM:	-0.437502

C

RBscore: total prediction score
Qscore: prediction based on ELECTROSTATICS
CEscore: Prediction based on CONSERVATION ENTROPY
SOLscore: Prediction based on SOLVATION ENERGY
Entropy: Conservation entropy
RBpred: Binary prediction by RBscore
SVMscore: prediction based on SVM
SVMpred: Binary prediction by SVM
+: RNA binding
-: non-RNA binding

No.	Res	RBScore	Qscore	CEscore	SOLscore	Entropy	RBpred	SVMscore	SVMpred
12	P	251.6	16.0	1265.5	-29.6	2.10	-	-1.000	-
13	Q	309.3	47.7	1430.8	-27.1	0.51	-	-1.000	-
14	G	348.3	63.8	1568.7	-40.1	1.97	+	-1.000	-
15	M	329.2	75.1	1359.7	-18.8	1.44	-	-1.000	-
16	S	475.2	152.2	1777.2	-30.2	1.89	+	0.999	+
17	S	532.4	174.0	1578.3	0.5	2.57	+	-0.999	-
18	N	562.9	231.1	2007.7	14.3	1.89	+	1.001	+
19	D	388.0	109.9	1507.2	-17.1	1.43	+	0.999	+
20	A	-100.0	-100.0	-100.0	-100.0	1.39	-	-1.000	-

B

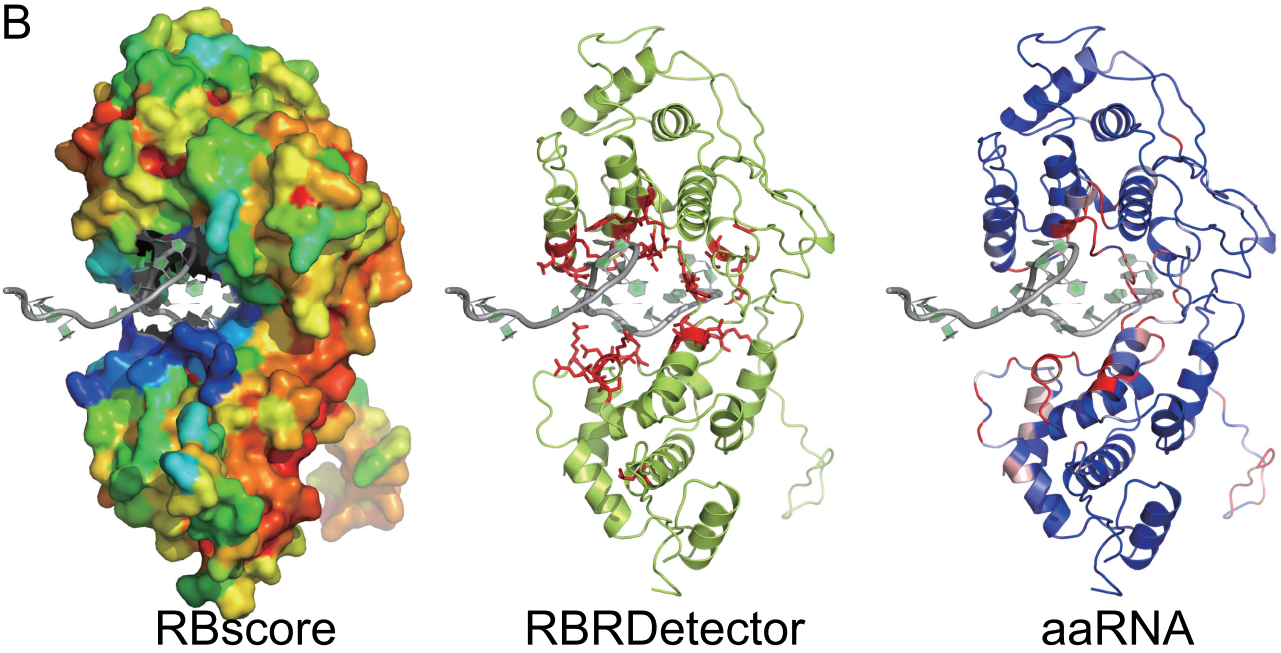

Figure 1. Example of outputs from RBscore. (**A**) Summary of prediction. (**B**) Prediction score mappings on protein structure demonstrated by RBscore (2), RBRDetector (30) and aaRNA (26). (**C**) Detailed residue-wise results.

other molecules or save the figure or molecule coordinates.

Figure 1B compares three existing demonstrations of prediction score mapping schemes. RBRDetector uses a binary color scheme on a cartoon model of the protein while highlighting the predicted binding sites with stick model. Such plots can clearly show the binding sites when the prediction is good but can hardly explain some 'orphan' binding sites that without any other binding site neighbor around. It can neither show the hierarchical binding energy funnel on protein surface. aaRNA uses a hierarchical color scheme of 'blue_white_red', but a cartoon model of protein structure excludes all side-chain atoms which are most important to protein–nucleic acid binding. A rainbow color scheme in RBscore can clearly show the binding energy funnel on protein surface to help users find the binding region intuitively, while a surface model of the protein structure counts for the accessibility of the residues easily excluding unreasonable buried residues as binding sites. Following the hierarchical coloring of binding energy funnel, 'orphan' binding sites can be easily excluded demonstrating a clearer picture of the nucleic acid binding probability.

(iii) Detailed residue-wise result (Figure 1C). It lists the prediction results for each residue, including residue name, conservation entropy, RBscore, RBscore_SVM value and other three prediction values based on single feature applied in RBscore model. Binary binding site prediction is based on the pre-defined specificity or

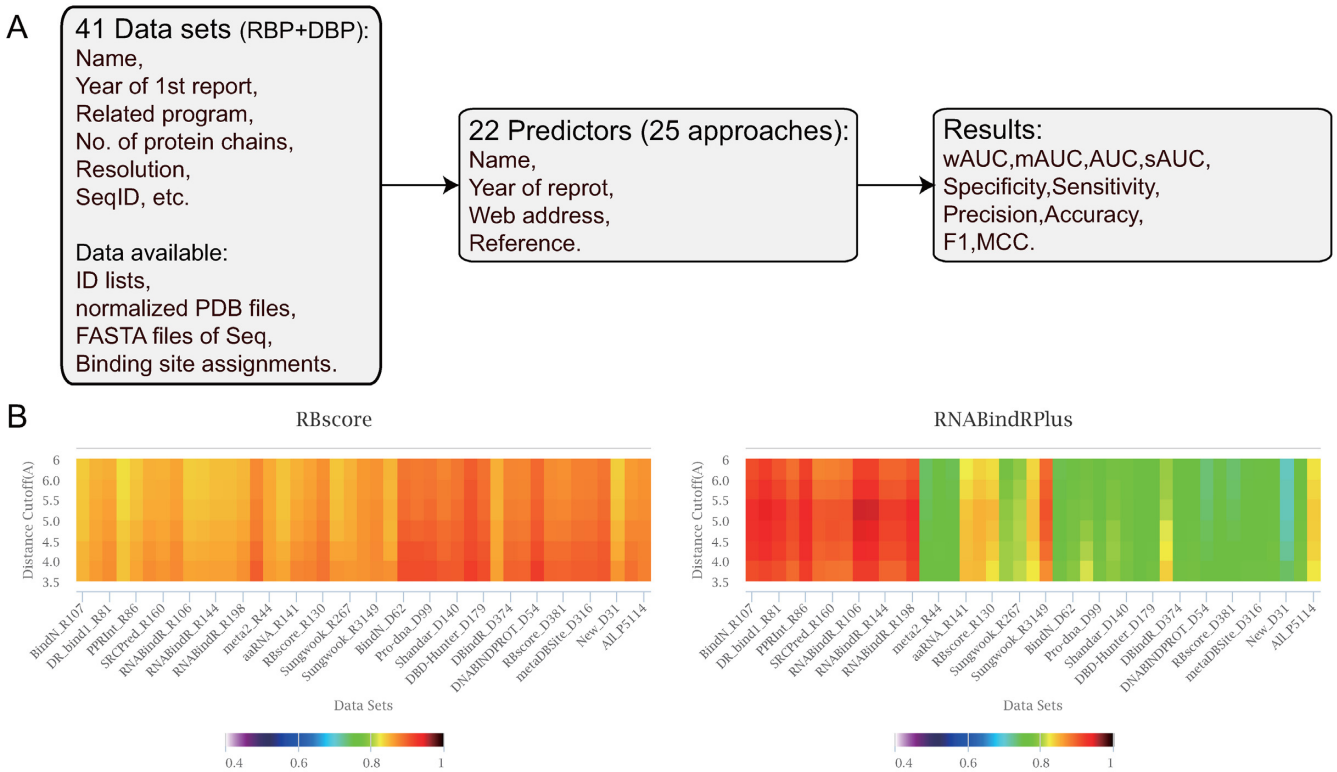

Figure 2. Structure of NBench and examples of heat map. (**A**) Structure of NBench database. The database includes detailed information and raw data of 41 reported datasets of protein–nucleic acid interactions, and it lists all information about the currently available predictors. Besides, it benchmarks all the predictors with various criteria considering datasets and distance cutoffs in defining binding sites. (**B**) Examples of heat maps exported from NBench for comparison.

sensitivity value and the resulted threshold. Generally, binding sites normally have an RBscore > 300 while RBscore_SVM > -0.44. It is more likely to be a nucleic acid binding site, when both RBscore, RBscore_SVM and conservation entropy show high scores.

NBench DATABASE

It is a non-trivial task to compare a new binding site predictor with existing ones to demonstrate its effectiveness. Such a comparison is prone to dataset bias, binding site definition bias and assessment metric bias, as well as the detailed treatments of the datasets. For example, PRBR (23) does not predict binding sites of the N-terminal and C-terminal residues, and may lead to an unfair comparison with other programs who predict on the whole sequence. To minimize the possibility of biased conclusions, NBench contains prediction results of 25 different approaches on 41 datasets to directly benchmark all the programs at the same level.

Data availability

NBench lists all the detailed information of the 41 datasets: number of protein, resolution, sequence identity, structural similarity and year of publication and PDB ID list. It provides all the PDB IDs, curated PDB files, sequence files in FASTA format and binding sites definition based on RBscore criterion which considers both distance cutoff and accessible surface area change. Besides, NBench stores all the assessment results of the programs and exhibit them in terms of 2D heat map, as shown in Figure 2. Users of the database can select their interested program, dataset, distance cutoff to define binding sites, assessment criterion and plot the 2D heat map accordingly. In this way, the comparison can be more specific and concrete. Finally, users are allowed to export these heat maps in different formats.

Potential benchmarking

Many predictors targeting the nucleic acid binding site prediction problem are being developed every year, NBench makes the validation of the new predictors easier and straightforward: on one hand, new predictor developers can download the datasets from NBench to run their predictor and compare with the results of other predictors obtained from NBench. Developers can perform their assessments on all the results for comparison, it avoids repeated calling of the other web servers. On the other hand, developers are also encouraged to submit their prediction results to NBench, so new predictions can be benchmarked in a more systematic way by NBench during the maintenance of the database. New data is added to NBench during regular maintenance of the database or upon request. Both approaches suggest better validations of the new upcoming predictors.

SOFTWARE IMPLEMENTATION

The web server was developed on Ubuntu 14.04 linux OS and is running on an Apache2 server and PHP5.3 as server-side scripting language. The server pipeline was written in python2.7 and interacts with RBscore program written in C++. Data of prediction results in NBench is organized and indexed by MySQL. The web pages were written in bootstrap HTML with Javascript, and were tested on the latest versions of Firefox and Chrome.

DISCUSSION

RBscore highlights the point that a nucleic acid binding site prediction is not a binary classifier but is to find the potential binding region to help understanding the underlying essence of protein–nucleic acid interaction, as well as to find the potential binding energy funnel. It is the first automated web server reported to predict DNA- and RNA-binding sites within the same prediction model. Besides, it directly combine feature values into a probability score of nucleic acid binding without complexion and achieves high level accuracy on all datasets regardless of binding site definition bias.

Nucleic acid binding site prediction is an active field of work, no fewer than eight papers (24–31) targeting this problem were published in 2014. Validation of new predictors is a crucial necessity but prone to bias. NBench directly provides normalized datasets and related results from existing approaches, which can be a valuable resource for new predictor validation. We hope RBscore&NBench can help our understanding the essence of protein–nucleic acid binding and support the biological community as a useful tool.

AVAILABILITY

The web server is available on: http://ahsoka.u-strasbg.fr/rbscorenbench/.

FUNDING

French Government [ANR-10-BINF-02-02 'BACNET']. Funding for open access charge: French Government grant [ANR-10-BINF-02-02 'BACNET'].
Conflict of interest statement. E.W. is an Executive Editor of Nucleic Acids Research.

REFERENCES

1. Kono,H. and Sarai,A. (1999) Structure-based prediction of DNA target sites by regulatory proteins. *Proteins*, **35**, 114–131.
2. Miao,Z. and Westhof,E. (2015) Prediction of nucleic acid binding probability in proteins: a neighboring residue network based score. *Nucleic Acids Res.*, **43**, 5340–5351.
3. 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.
4. Hudson,W.H. and Ortlund,E.A. (2014) The structure, function and evolution of proteins that bind DNA and RNA. *Nat. Rev. Mol. Cell Biol.*, **15**, 749–760.
5. Miao,Z. and Westhof,E. (2015) A large-scale assessment of nucleic acids binding site prediction programs. *PLoS Comput. Biol.*, **11**, e1004639.
6. Miao,Z., Cao,Y. and Jiang,T. (2014) Modeling of protein side-chain conformations with RASP. *Methods Mol. Biol.*, **1137**, 43–53.
7. Miao,Z., Cao,Y. and Jiang,T. (2011) RASP: rapid modeling of protein side chain conformations. *Bioinformatics*, **27**, 3117–3122.
8. Dolinsky,T.J., Nielsen,J.E., McCammon,J.A. and Baker,N.A. (2004) PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.*, **32**, W665–W667.
9. Dolinsky,T.J., Czodrowski,P., Li,H., Nielsen,J.E., Jensen,J.H., Klebe,G. and Baker,N.A. (2007) PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.*, **35**, W522–W525.
10. Baker,N.A., Sept,D., Joseph,S., Holst,M.J. and McCammon,J.A. (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 10037–10041.
11. Shazman,S., Celniker,G., Haber,O., Glaser,F. and Mandel-Gutfreund,Y. (2007) Patch Finder Plus (PFplus): a web server for extracting and displaying positive electrostatic patches on protein surfaces. *Nucleic Acids Res.*, **35**, W526–W530.
12. McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.
13. Shannon,C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423.
14. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
15. Remmert,M., Biegert,A., Hauser,A. and Soding,J. (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.
16. Richards,F.M. (1977) Areas, volumes, packing and protein structure. *Annu. Rev. Biophys. Bioeng.*, **6**, 151–176.
17. Bairoch,A. and Apweiler,R. (1996) The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res.*, **24**, 21–25.
18. 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.
19. Joachims,T. (1999) Making large-scale support vector machine learning practical. In: Bernhard,S, Christopher,JCB and Alexander,JS (eds). *Advances in Kernel Methods*. MIT Press, Cambridge, 169–184.
20. Hanson,R.M., Prilusky,J., Renjian,Z., Nakane,T. and Sussman,J.L. (2013) JSmol and the next-generation web-based representation of 3D molecular structure as applied to proteopedia. *Isr. J. Chem.*, **53**, 207–216.
21. Celniker,G., Nimrod,G., Ashkenazy,H., Glaser,F., Martz,E., Mayrose,I., Pupko,T. and Ben-Tal,N. (2013) ConSurf: using evolutionary data to raise testable hypotheses about protein function. *Isr. J. Chem.*, **53**, 199–206.
22. DeLano,W.L. and Lam,J.W. (2005) PyMOL: a communications tool for computational models. *Abstr. Pap. Am. Chem. Soc.*, **230**, U1371–U1372.
23. Ma,X., Guo,J., Wu,J., Liu,H., Yu,J., Xie,J. and Sun,X. (2011) Prediction of RNA-binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature. *Proteins*, **79**, 1230–1239.
24. Pan,X., Zhu,L., Fan,Y.-X. and Yan,J. (2014) Predicting protein–RNA interaction amino acids using random forest based on submodularity subset selection. *Comput. Biol. Chem.*, **53**, 324–330.
25. Walia,R.R., Xue,L.C., Wilkins,K., El-Manzalawy,Y., Dobbs,D. and Honavar,V. (2014) RNABindRPlus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins. *PLoS One*, **9**, e97725.
26. Li,S., Yamashita,K., Amada,K.M. and Standley,D.M. (2014) Quantifying sequence and structural features of protein-RNA interactions. *Nucleic Acids Res.*, **42**, 10086–10098.
27. Park,B., Im,J., Tuvshinjargal,N., Lee,W. and Han,K. (2014) Sequence-based prediction of protein-binding sites in DNA: comparative study of two SVM models. *Comput. Methods Prog. Biomed.*, **117**, 158–167.
28. Li,B.Q., Feng,K.Y., Ding,J. and Cai,Y.D. (2014) Predicting DNA-binding sites of proteins based on sequential and 3D structural information. *Mol. Genet. Genomics*, **289**, 489–499.
29. Yan,C. and Wang,Y. (2014) A graph kernel method for DNA-binding site prediction. *BMC Syst. Biol.*, **8**(Suppl. 4), S10.

30. Yang,X.X., Deng,Z.L. and Liu,R. (2014) RBRDetector: improved prediction of binding residues on RNA-binding protein structures using complementary feature- and template-based strategies. *Proteins*, **82**, 2455–2471.

31. Chen,Y.C., Sargsyan,K., Wright,J.D., Huang,Y.S. and Lim,C. (2014) Identifying RNA-binding residues based on evolutionary conserved structural and energetic features. *Nucleic Acids Res.*, **42**, e15.