PRIDB: a protein–RNA interface database

Benjamin A. Lewis[1,2,*], Rasna R. Walia[1,3], Michael Terribilini[4], Jeff Ferguson[5], Charles Zheng[5], Vasant Honavar[1,3] and Drena Dobbs[1,2]

[1]Bioinformatics and Computational Biology Program, Iowa State University, [2]Department of Genetics, Development and Cell Biology, Iowa State University, [3]Department of Computer Science, Iowa State University, Ames, IA 50011, [4]Department of Biology, Elon University, Elon, NC 27244 and [5]Computational Systems Biology Summer Institute, Iowa State University, Ames, IA 50011, USA

ABSTRACT

The Protein–RNA Interface Database (PRIDB) is a comprehensive database of protein–RNA interfaces extracted from complexes in the Protein Data Bank (PDB). It is designed to facilitate detailed analyses of individual protein–RNA complexes and their interfaces, in addition to automated generation of user-defined data sets of protein–RNA interfaces for statistical analyses and machine learning applications. For any chosen PDB complex or list of complexes, PRIDB rapidly displays interfacial amino acids and ribonucleotides within the primary sequences of the interacting protein and RNA chains. PRIDB also identifies ProSite motifs in protein chains and FR3D motifs in RNA chains and provides links to these external databases, as well as to structure files in the PDB. An integrated JMol applet is provided for visualization of interacting atoms and residues in the context of the 3D complex structures. The current version of PRIDB contains structural information regarding 926 protein–RNA complexes available in the PDB (as of 10 October 2010). Atomic- and residue-level contact information for the entire data set can be downloaded in a simple machine-readable format. Also, several non-redundant benchmark data sets of protein–RNA complexes are provided. The PRIDB database is freely available online at http://bindr .gdcb.iastate.edu/PRIDB.

INTRODUCTION

Protein–RNA interactions play critical roles in myriad and diverse biological processes, including many recently discovered regulatory functions, in addition to well-studied roles in protein synthesis, DNA replication, regulation of gene expression and defense against pathogens (1–9). Despite their importance, structures of protein–RNA complexes have proven difficult to obtain using experimental structure determination methods; such structures constitute only ~1% of structures in the Protein Data Bank (PDB) (10). For this reason, several computational methods for predicting the interfaces in protein–RNA complexes have been developed (11–21). Virtually all such methods require data in the form of information about structurally characterized protein–RNA complexes and their interfaces.

PRIDB is a repository of protein–RNA interface information derived from structures in the PDB. PRIDB is designed to facilitate detailed analyses of individual protein–RNA complexes of interest and rapid identification of interfacial atoms and residues in both the protein and RNA chains of a chosen complex or user-defined set of complexes. In addition, PRIDB can be used to generate data sets of protein–RNA interfaces for machine learning applications, such as the generation of classifiers for predicting interfaces in protein–RNA complexes for which high-resolution structures are not available.

Related databases/servers

To our knowledge, only one other up-to-date and comprehensive online repository of protein–RNA interfaces is currently available: Biological Interaction Database for Protein-Nucleic Acid (BIPA) (22). BIPA provides a list of protein–RNA (and protein–DNA) complexes from the PDB and displays RNA-binding residues within the linear primary sequence of a chosen protein, or within a multiple sequence alignment of related RNA-binding proteins. PRIDB complements BIPA by providing atomic- and residue-level interfacial information for both the RNA and protein chains of complexes, providing previously published reduced-redundancy data sets and allowing users to make advanced queries and compile custom data sets. Other collections of protein–RNA

complexes and related resources include NDB (http://ndbserver.rutgers.edu/) (23), PRID (http://www-bioc.rice.edu/~shamoo/prid.html) (24), RsiteDB (http://bioinfo3d.cs.tau.ac.il/RsiteDB/) (25), w3DNA (http://w3dna.rutgers.edu/) (26), NPIDB (http://monkey.belozersky.msu.ru/NPIDB) (27), ProNIT (http://gibk26.bse.kyutech.ac.jp/jouhou/pronit/pronit.html) (28) and the RNP Databases http://rnp.uthct.edu/index.html/). Several excellent databases of protein–DNA interfaces are also available, including PDIdb (http://melolab.org/pdidb/) (29) and hPDI (http://bioinfo.wilmer.jhu.edu/PDI/).

DATABASE CONTENTS

Data extraction, interface definition and motif identification

Atomic coordinate information for all 926 protein–RNA complexes in the Protein Data Bank (PDB) on 10 October 2010 was extracted using the REST API advanced search interface. To generate this comprehensive data set (rRB926), no filters based on sequence redundancy, structure resolution or other criteria were applied (see 'Non-redundant Benchmark data sets' below). The complex structures in rRB926 were then scanned to identify interacting amino acids and ribonucleotides using two different definitions: (i) a simple distance-based definition in which a given amino acid residue (AA) in a protein chain is defined as interacting with a ribonucleotide (rNT) in an RNA chain if any atom in AA is within a 5-Å radius of any atom in rNT; and (ii) a rule-based definition based on that of Allers and Shamoo (30), in which interactions are classified as van der Waals, hydrogen-bonding, hydrophobic or electrostatic interactions, involving specific AAs and rNTs. All such interacting AAs and rNTs are defined as 'interface' residues.

ProSite patterns and profiles (31) appearing in any of the protein sequences in the database were retrieved using the ScanProsite REST service (32). RNA structural motifs were identified in RNA sequences using FR3D's (33) pure symbolic search function; specific motif definitions used for these scans are available in the Tutorial and FAQs section of the PRIDB online server.

Non-redundant benchmark data sets

Because PRIDB is intended to be a comprehensive collection of protein–RNA complexes from the PDB, the rRB926 data set was not filtered on the basis of redundancy, structure determination method, resolution or protein/RNA chain length. While it is possible to filter with such criteria using PRIDB's advanced search function, several pre-calculated benchmark data sets, which have been filtered to limit redundancy and to exclude low-resolution structures, are also provided for the user's convenience. These include two previously published data sets, RB109 (17,34) and RB147 (35), as well as a larger, more recently extracted data set (RB199) (B. Lewis, submitted for publication). Complete lists of the PDB IDs for protein–RNA complexes in these data sets, in addition to the pre-calculated interface

residue statistics, can be readily accessed from the 'Datasets' section of the PRIDB homepage.

Implementation and availability

PRIDB runs on the Apache 2.2 web server, using MySQL 14.14 as a database backend with AJAX and PHP 5 for user interface functions. Functions not requiring use of the database (e.g. calculating interface residues for a user-submitted complex) are implemented using standalone Perl 5 scripts and the BioPerl module (36). All PRIDB code is available on request under the Creative Commons Attribution Non-Commercial License. All data currently in PRIDB was obtained from databases or programs which impose no restrictions on academic use.

PRIDB summary statistics

As summarized in Table 1, the current version of PRIDB contains structural information for a total of 926 protein–RNA complexes available in the PDB as of 10 October 2010. These structures contain 9689 total protein chains, among which there are only 1174 unique sequences. While this would seem to indicate that most sequences in the database are repeated several times, this is not the case; 395 of the 1174 (34%) sequences appear only once, and 899 (77%) appear less than eight times (the 'expected' average redundancy). This disparity is due to the large proportion of ribosomal structures in the PDB (and, by extension, in PRIDB); 9 of the top 10 most abundant sequences, each present in more than 70 structures, are ribosomal proteins. The most abundant sequence, repeated more than 100 times, is that of the TRP-responsive attenuation protein, a protein for which numerous multimeric structures have been solved.

As shown in Table 2, PRIDB currently contains 1 475 774 amino acid residues. Based on a 5Å distance cutoff definition for interfacial residues, 397 216 of these residues interact with RNA; of 851 853 ribonucleotide residues in PRIDB, 322 858 interact with protein. On average, 38% of the amino acids in the RNA-binding

Table 1. PRIDB contents: complexes and chains

	Total Number in PRIDB[a]	Unique
Protein–RNA complexes	926	926
Protein chains	9689	1174
RNA chains	2074	746

[a]Total number in PRIDB includes redundant complexes, RNA and protein chains (i.e. chains with identical sequences).

Table 2. PRIDB summary statistics

Type	Total (Interface + Non-Interface)	Number in Interfaces (%)
Amino Acids	1 475 774	414 026 (38)
Ribonucleotides	851 853	326 441 (28)

proteins directly interact with RNA, and 28% of the ribonucleotides in the bound RNAs directly interact with protein. As before, these averages are skewed by the prevalence of ribosome structures; ribosomal proteins account for ~90% of interacting amino acid residues and ~60% of interacting nucleotides.

USER INTERFACE

PRIDB provides a 'Tutorial and FAQs' section with detailed instructions on using PRIDB's web interface; a list and brief descriptions of key capabilities of PRIDB are provided here. Using the 'Basic Search' function, users can retrieve information about protein–RNA complexes using their PDB ID or a keyword. Using the 'Advanced Search' function, users can filter results by specifying:

- the experimental method used to determine the complex structure (e.g. X-ray diffraction, nuclear magnetic resonance);
- a resolution range or threshold (for structures determined using X-ray diffraction, electron microscopy or fiber diffraction);
- the minimum or maximum length of protein or RNA chains within the complex;
- an amino acid or nucleotide subsequence found within the sequence of at least one of the protein or RNA chains in the complex; and

Figure 1. Sample PRIDB output. Amino acid residues and ribonucleotides highlighted in yellow are located in the protein–RNA interface; residues in red font are part of a ProSite or FR3D motif.

- a motif (as defined by ProSite for protein chains or FR3D for RNA chains) found within at least one chain in the complex.

The 'Advanced Search' function also allows users to either specify a different distance cutoff for the distance-based interaction definition or choose the alternative rule-based definition.

As shown in Figure 1, when viewing search results, PRIDB provides:

- a summary of and basic information (name, resolution and structure determination method) about each complex, as well as a link to that complex's PDB entry;
- a linear display of the amino acid and nucleotide residues in each chain of each complex, with residues in the protein–RNA interface highlighted;
- a display of residues (in red font) that are part of a protein or RNA motif, with information about that motif (and a link back to its source) provided on mouse-over;
- a JMol applet for 3D visualization of each complex, with interacting amino acid and nucleotide residues colored (Figure 2A); and
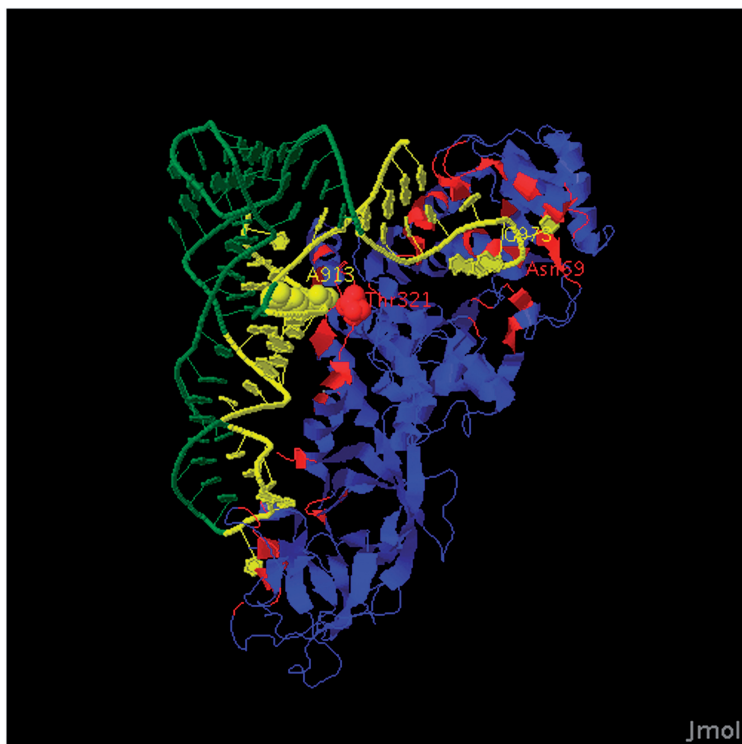
- a link to a dynamically-generated file containing atomic-level interface information for each result in a machine readable format (Figure 2B).

In addition to providing machine-readable results files for all searches, pre-computed results files for the non-redundant RB109, RB147 and RB199 data sets described above have been made available. These files, along with the complete PRIDB database (rRB926), can be downloaded from the 'Datasets' section of the website. Users can also generate a machine-readable list of interface residues for any arbitrary collection of complexes by inputting a list of PDB IDs. Results files contain a single line for each pair of interacting atoms listing the specific interacting atoms (by chain name, residue number and atom name) and the distance between them.

Users may also calculate interface residues for protein–RNA complexes that are not in PDB using PRIDB by submitting a structure file in PDB format. A results file containing interface residues (as calculated using PRIDB's 5 Å cutoff) is returned via e-mail.

A Blue: Non-interacting protein residues   Green: Non-interacting RNA residues
Red: Interacting protein residues   Yellow: Interacting RNA residues

PDB ID of structure: 1EUQ

B

PDB ID	PChn	AA#	AA	PAtom	RChn	NT#	NT	RAtom	Distance
1EUQ	A	9	ASN	CA	B	914	A	OP1	4.90277
1EUQ	A	13	GLN	CD	B	915	G	OP1	4.1927
1EUQ	A	13	GLN	OE1	B	915	G	OP1	4.56629
1EUQ	A	13	GLN	NE2	B	915	G	P	4.27964
1EUQ	A	13	GLN	NE2	B	915	G	OP1	3.00861
1EUQ	A	13	GLN	NE2	B	915	G	OP2	4.63184
1EUQ	A	13	GLN	CD	B	916	C	C4	4.88702
1EUQ	A	13	GLN	CD	B	916	C	N4	3.72539
1EUQ	A	13	GLN	OE1	B	916	C	C4	4.27898
1EUQ	A	13	GLN	OE1	B	916	C	N4	3.1138
1EUQ	A	13	GLN	OE1	B	916	C	C5	4.86907
1EUQ	A	13	GLN	NE2	B	916	C	C4	4.76736
1EUQ	A	13	GLN	NE2	B	916	C	N4	3.61701
1EUQ	A	34	GLU	CD	B	975	C	O2'	4.81234
1EUQ	A	34	GLU	OE1	B	975	C	O2'	4.39732
1EUQ	A	34	GLU	OE2	B	975	C	O2'	4.30775
1EUQ	A	34	GLU	CB	B	976	A	C4'	4.62712
1EUQ	A	34	GLU	CB	B	976	A	C3'	4.95418
1EUQ	A	34	GLU	CB	B	976	A	O3'	4.23121
1EUQ	A	34	GLU	CB	B	976	A	O2'	4.30761
1EUQ	A	34	GLU	CG	B	976	A	C5'	4.88732
1EUQ	A	34	GLU	CG	B	976	A	C4'	4.47075
1EUQ	A	34	GLU	CG	B	976	A	O3'	4.63648
1EUQ	A	34	GLU	CD	B	976	A	C5'	3.94281
1EUQ	A	34	GLU	CD	B	976	A	C4'	3.5693
1EUQ	A	34	GLU	CD	B	976	A	O4'	4.27732
1EUQ	A	34	GLU	CD	B	976	A	C3'	4.64741
1EUQ	A	34	GLU	CD	B	976	A	O3'	4.53075
1EUQ	A	34	GLU	CD	B	976	A	O2'	4.67345
1EUQ	A	34	GLU	OE1	B	976	A	O5'	4.85771
1EUQ	A	34	GLU	OE1	B	976	A	C5'	3.53209
1EUQ	A	34	GLU	OE1	B	976	A	C4'	2.7699
1EUQ	A	34	GLU	OE1	B	976	A	O4'	3.24468
1EUQ	A	34	GLU	OE1	B	976	A	C3'	3.84274
1EUQ	A	34	GLU	OE1	B	976	A	O3'	3.98493

Figure 2. (**A**) PRIDB provides a JMol applet for visualizing and manipulating interfaces within 3-D structures. (**B**) PRIDB output can be downloaded as a CSV file.

CONCLUSIONS AND FUTURE DIRECTIONS

PRIDB provides researchers with atomic and residue-level information about structures of protein–RNA complexes and their interfaces, facilitating analyses of protein–RNA interactions by pre-computing commonly used information and by providing structural information both interactively onscreen and in a machine-readable format. It allows users to rapidly identify and visualize interfaces in protein–RNA complexes on a residue-by-residue basis and displays identified ProSite or FR3D motifs along with the amino acid or ribonucleotide sequences. PRIDB can be used to generate custom data sets of protein–RNA interfaces for statistical analyses and machine learning applications. The PRIDB server also provides pre-calculated benchmark data sets of protein–RNA complexes for evaluating the performance of interface prediction methods. PRIDB will be updated regularly as new structures are released through PDB, and is intended to be a stable resource for researchers in the field of protein–RNA interactions.

Future versions of PRIDB will include additional protein and RNA motifs from other sources, such as PRINTS (37), PIRSF (38) and other InterPro (39) member databases. In addition, the current JMol 3D visualization capabilities will be extended to user-submitted structures, allowing for more facile manipulation and examination of interfaces in complexes not currently in the PDB.

REFERENCES

1. Fabian,M.R., Sonenberg,N. and Filipowicz,W. (2010) Regulation of mRNA translation and stability by microRNAs. *Annu. Rev. Biochem.*, **79**, 351–379.
2. Hogan,D.J., Riordan,D.P., Gerber,A.P., Herschlag,D. and Brown,P.O. (2008) Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol.*, **6**, e255.
3. Licatalosi,D.D. and Darnell,R.B. (2010) RNA processing and its regulation: global insights into biological networks. *Nat. Rev. Genet.*, **11**, 75–87.
4. Lorkovic,Z.J. (2009) Role of plant RNA-binding proteins in development, stress response and genome organization. *Trends Plant Sci.*, **14**, 229–236.
5. Lukong,K.E., Chang,K.W., Khandjian,E.W. and Richard,S. (2008) RNA-binding proteins in human genetic disease. *Trends Genet.*, **24**, 416–425.
6. Lunde,B.M., Moore,C. and Varani,G. (2007) RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.*, **8**, 479–490.
7. Mansfield,K.D. and Keene,J.D. (2009) The ribonome: a dominant force in co-ordinating gene expression. *Biol. Cell*, **101**, 169–181.
8. Mittal,N., Roy,N., Babu,M.M. and Janga,S.C. (2009) Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks. *Proc. Natl Acad. Sci. USA*, **106**, 20300–20305.
9. Mohammad,M.M., Donti,T.R., Sebastian Yakisich,J., Smith,A.G. and Kapler,G.M. (2007) Tetrahymena ORC contains a ribosomal RNA fragment that participates in rDNA origin recognition. *EMBO J.*, **26**, 5048–5060.
10. 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.
11. Liu,Z.P., Wu,L.Y., Wang,Y., Zhang,X.S. and Chen,L. (2010) Prediction of protein-RNA binding sites by a random forest method with combined features. *Bioinformatics*, **26**, 1616–1622.
12. Murakami,Y., Spriggs,R.V., Nakamura,H. and Jones,S. (2010) PiRaNhA: a server for the computational prediction of RNA-binding residues in protein sequences. *Nucleic Acids Res.*, **38(Suppl.)**, W412–W416.
13. Perez-Cano,L. and Fernandez-Recio,J. (2010) Optimal protein-RNA area, OPRA: a propensity-based method to identify RNA-binding sites on proteins. *Proteins*, **78**, 25–35.
14. Maetschke,S.R. and Yuan,Z. (2009) Exploiting structural and topological information to improve prediction of RNA-protein binding sites. *BMC Bioinformatics*, **10**, 341.
15. Shazman,S. and Mandel-Gutfreund,Y. (2008) Classifying RNA-binding proteins based on electrostatic properties. *PLoS Comput. Biol.*, **4**, e1000146.
16. Wang,L., Huang,C., Yang,M.Q. and Yang,J.Y. (2010) BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst Biol*, **4(Suppl. 1)**, S3.
17. Terribilini,M., Lee,J.H., Yan,C., Jernigan,R.L., Honavar,V. and Dobbs,D. (2006) Prediction of RNA binding sites in proteins from amino acid sequence. *RNA*, **12**, 1450–1462.
18. Wang,L. and Brown,S.J. (2006) Prediction of RNA-binding residues in protein sequences using support vector machines. *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, **1**, 5830–5833.
19. Towfic,F., Caragea,C., Gemperline,D.C., Dobbs,D. and Honavar,V. (2010) Struct-NB: predicting protein-RNA binding sites using structural features. *Int. J. Data Min. Bioinform.*, **4**, 21–43.
20. Kumar,M., Gromiha,M.M. and Raghava,G.P. (2010) SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. *J. Mol. Recognit.*, doi:10.1002/jmr.1061.
21. Wang,C.C., Fang,Y., Xiao,J. and Li,M. (2010) Identification of RNA-binding sites in proteins by integrating various sequence information. *Amino Acids*, doi:10.1007/s00726-010-0639-7.
22. Lee,S. and Blundell,T.L. (2009) BIPA: a database for protein-nucleic acid interaction in 3D structures. *Bioinformatics*, **25**, 1559–1560.
23. Berman,H.M., Olson,W.K., Beveridge,D.L., Westbrook,J., Gelbin,A., Demeny,T., Hsieh,S.H., Srinivasan,A.R. and Schneider,B. (1992) The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.*, **63**, 751–759.
24. Morozova,N., Allers,J., Myers,J. and Shamoo,Y. (2006) Protein-RNA interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. *Bioinformatics*, **22**, 2746–2752.
25. Shulman-Peleg,A., Nussinov,R. and Wolfson,H.J. (2009) RsiteDB: a database of protein binding pockets that interact with RNA nucleotide bases. *Nucleic Acids Res.*, **37**, D369–D373.
26. Zheng,G., Lu,X.J. and Olson,W.K. (2009) Web 3DNA–a web server for the analysis, reconstruction, and visualization of

three-dimensional nucleic-acid structures. *Nucleic Acids Res.*, **37**, W240–W246.

27. Spirin,S., Titov,M., Karyagina,A. and Alexeevski,A. (2007) NPIDB: a database of nucleic acids-protein interactions. *Bioinformatics*, **23**, 3247–3248.

28. Kumar,M.D., Bava,K.A., Gromiha,M.M., Prabakaran,P., Kitajima,K., Uedaira,H. and Sarai,A. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.*, **34**, D204–D206.

29. Norambuena,T. and Melo,F. (2010) The Protein-DNA Interface database. *BMC Bioinformatics*, **11**, 262.

30. Allers,J. and Shamoo,Y. (2001) Structure-based analysis of protein-RNA interactions using the program ENTANGLE. *J. Mol. Biol.*, **311**, 75–86.

31. Sigrist,C.J., Cerutti,L., de Castro,E., Langendijk-Genevaux,P.S., Bulliard,V., Bairoch,A. and Hulo,N. (2010) PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.*, **38**, D161–D166.

32. de Castro,E., Sigrist,C.J., Gattiker,A., Bulliard,V., Langendijk-Genevaux,P.S., Gasteiger,E., Bairoch,A. and Hulo,N. (2006) ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res.*, **34**, W362–W365.

33. Sarver,M., Zirbel,C.L., Stombaugh,J., Mokdad,A. and Leontis,N.B. (2008) FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J. Math. Biol.*, **56**, 215–252.

34. Terribilini,M., Lee,J.H., Yan,C., Jernigan,R.L., Carpenter,S., Honavar,V. and Dobbs,D. (2006) Identifying interaction sites in 'recalcitrant' proteins: predicted protein and RNA binding sites in rev proteins of HIV-1 and EIAV agree with experimental data. *Pac. Symp. Biocomput.*, 415–426.

35. Terribilini,M., Sander,J.D., Lee,J.H., Zaback,P., Jernigan,R.L., Honavar,V. and Dobbs,D. (2007) RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Res.*, **35**, W578–W584.

36. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.

37. Attwood,T.K., Bradley,P., Flower,D.R., Gaulton,A., Maudling,N., Mitchell,A.L., Moulton,G., Nordle,A., Paine,K., Taylor,P. *et al.* (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.*, **31**, 400–402.

38. Wu,C.H., Nikolskaya,A., Huang,H., Yeh,L.S., Natale,D.A., Vinayaka,C.R., Hu,Z.Z., Mazumder,R., Kumar,S., Kourtesis,P. *et al.* (2004) PIRSF: family classification system at the protein information resource. *Nucleic Acids Res.*, **32**, D112–D114.

39. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.