BIOINFORMATICS DISCOVERY NOTE

Vol. 29 no. 20 2013, pages 2524–2528
doi:10.1093/bioinformatics/btt453

Structural bioinformatics

Advance Access publication August 9, 2013

Analysis of base-pairing probabilities of RNA molecules involved in protein–RNA interactions

Junichi Iwakiri[1,2,*], Tomoshi Kameda[2], Kiyoshi Asai[1,2] and Michiaki Hamada[1,2,*]

[1]Graduate School of Frontier Sciences, The University of Tokyo, 5–1–5 Kashiwanoha, Kashiwa, Chiba 277–8562, Japan and [2]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2–4–7 Aomi, Koto-ku, Tokyo 135–0064, Japan

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Understanding the details of protein–RNA interactions is important to reveal the functions of both the RNAs and the proteins. In these interactions, the secondary structures of the RNAs play an important role. Because RNA secondary structures in protein–RNA complexes are variable, considering the ensemble of RNA secondary structures is a useful approach. In particular, recent studies have supported the idea that, in the analysis of RNA secondary structures, the base-pairing probabilities (BPPs) of RNAs (i.e. the probabilities of forming a base pair in the ensemble of RNA secondary structures) provide richer and more robust information about the structures than a single RNA secondary structure, for example, the minimum free energy structure or a snapshot of structures in the Protein Data Bank. However, there has been no investigation of the BPPs in protein–RNA interactions.

Results: In this study, we analyzed BPPs of RNA molecules involved in known protein–RNA complexes in the Protein Data Bank. Our analysis suggests that, in the tertiary structures, the BPPs (which are computed using only sequence information) for *unpaired* nucleotides with intermolecular hydrogen bonds (hbonds) to amino acids were significantly lower than those for unpaired nucleotides *without* hbonds. On the other hand, no difference was found between the BPPs for *paired* nucleotides with and without intermolecular hbonds. Those findings were commonly supported by three probabilistic models, which provide the ensemble of RNA secondary structures, including the McCaskill model based on Turner's free energy of secondary structures.

Contact: iwakiri@cb.k.u-tokyo.ac.jp or mhamada@cb.k.u-tokyo.ac.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 12, 2013; revised on July 30, 2013; accepted on August 3, 2013

1 INTRODUCTION

Protein–RNA interactions play important roles in various aspects of biological processes, such as splicing, translation and RNA silencing. Recently, several experimental techniques have been developed to investigate protein–RNA interactions (Licatalosi *et al*., 2008; Ray *et al*., 2009). In addition, the increasing number of determined 3D structures of protein–RNA complexes deposited in the Protein Data Bank (PDB) (Berman *et al*., 2000) has promoted many computational studies to understand the detailed mechanisms underlying the protein–RNA interactions (Allers and Shamoo, 2001; Bahadur *et al*., 2008; Ellis *et al*., 2007; Gupta and Gribskov, 2011; Iwakiri *et al*., 2012; Kim *et al*., 2006; Kondo and Westhof, 2011; Sonavane and Chakrabarti, 2009). Most of these studies consistently showed the importance of electrostatic interactions between the positively charged amino acids of the proteins and the negatively charged phosphate groups of the RNA. However, only a few studies have considered the role of RNA secondary structures in the protein–RNA interactions (Gupta and Gribskov, 2011; Iwakiri *et al*., 2012), although many RNA binding proteins are known to recognize the specific secondary structures of their partner RNA (Ray *et al*., 2009). In addition, the secondary structures are phylogenetically conserved and are closely related to the functions of RNAs (Burge *et al*., 2012).

One difficulty for analyzing RNA secondary structures in protein–RNA complexes in the PDB is that only snapshots of the secondary structures are available but they are variable in reality (Gopal *et al*., 2012). Moreover, the probability of forming a specific RNA secondary structure, such as the minimum free energy structure, often becomes extremely small (typically $<10^{-10}$) (Hamada, 2013). These facts are known as the *ambiguity* or *uncertainty* of RNA secondary structures. To tackle this uncertainty, a probability distribution of secondary structures has been considered in many analyses of RNAs (Adachi *et al*., 2011; Halvorsen *et al*., 2010). Specifically, base-pairing probabilities (BPPs) that provide probabilities for each base pair with respect to an ensemble of RNA secondary structures are often used for RNA secondary structure analysis (see Section 2.4 for the detailed definition of BPPs). For example, the BPPs played essential roles in the analysis of an RNA aptamer (Adachi *et al*., 2011), in which an unstable stem in the RNA aptamer was confirmed by both computational analysis of BPPs and biochemical experiments. This indicates the importance of considering the ensemble in the analysis of RNAs. However, there is no study that considers the ensemble of RNAs in the analysis of protein–RNA interactions.

In this study, we systematically investigate BPPs of RNA molecules involved in known protein–RNA complexes (taken from the PDB; see Section 2.1 for the detailed dataset). More specifically, BPPs of both paired and unpaired nucleotides (Section 2.2) with and without intermolecular hydrogen bonds (hbonds) to amino acids (see Section 2.3 for the detailed definition) are

*To whom correspondence should be addressed.

systematically analyzed (see Section 3). In the analysis, we tried three probabilistic models of RNA secondary structure (Section 2.4), including the McCaskill model based on Turner's energy model (Mathews *et al*., 1999), when computing BPPs, and confirmed that our results are consistently supported.

2 MATERIALS AND METHODS

2.1 Datasets

For a dataset of protein–RNA complexes, we used 91 complexes created in a previous study (Iwakiri *et al*., 2012). In this dataset, modified nucleotides are deleted and the complexes containing multiple RNA chains are excluded because the calculation of the BPP is limited to standard nucleotides and single-stranded RNAs. The resulting dataset has 68 complexes comprising 2688 nucleotides.

The above dataset includes 27 tRNAs whose secondary structures are structurally similar despite the low similarity between their primary sequences. To avoid this population bias, we created another dataset in which a representative tRNA-related complex (PDBID: 1ASY) was randomly selected for our analysis. The resulting dataset has 42 complexes, comprising 1098 nucleotides.

2.2 Identification of base pairs involved in RNA 3D structures

To be consistent with the definition of a base pair that is used in the calculation of the BPP, canonical Watson–Crick pairs (A–U and G–C) and wobble pairs (G–U) are treated as base pairs. These base pairs are extracted from each RNA tertiary structure using the RNAview program (Yang *et al*., 2003). Nucleotides forming a base pair were categorized as *paired* nucleotides, and the remaining nucleotides were categorized as *unpaired* nucleotides. Note that our definitions of paired and unpaired nucleotides are based on tertiary structures in the PDB.

2.3 Identification of intermolecular hbonds in the protein–RNA interface

In the protein–RNA complexes, amino acids and nucleotides are identified as the interface residues if their closest atom–atom distances are <5.0 Å. Intermolecular hbonds between proteins and RNAs were identified using HBPLUS programs with default settings (the donor–acceptor distance is <3.9 Å, the hydrogen–acceptor distance is <2.5 Å, angles of donor–hydrogen–acceptor (D–H–A) and donor–acceptor–acceptor antecedents (D–A–AA) are >90°) (McDonald and Thornton, 1994). It should be noted that sugar and Hoogsteen edges and phosphate groups are available to form an intermolecular hbond with amino acids even if the nucleotides form base pairs using Watson–Crick edges (i.e. they are paired nucleotides) (Leontis and Westhof, 2001).

2.4 BPPs for RNA sequences

Given an RNA sequence x, a BPP p_{ij} ($i<j$) is equal to the probability that the ith and jth nucleotides of x, x_i and x_j, form a base pair (in an assumed ensemble of RNA secondary structures). This probability is formally defined by the marginal probability.

$$p_{ij} = \sum_{\theta \in \mathcal{S}(x)} I(\theta_{ij} = 1) p(\theta|x) \quad (1)$$

where $p(\theta|x)$ is a probability distribution on the set $\mathcal{S}(x)$ of possible secondary structures of the RNA sequence x, and the indicator $I(\theta_{ij} = 1)$ becomes 1 only when the secondary structure θ includes the base pair (x_i, x_j).

In Equation (1), the probability distribution $p(\theta|x)$ is often given by the Boltzmann distribution [when it is called the McCaskill model (McCaskill, 1990)], which is based on experimentally determined parameters:

$$p(\theta|x) = \frac{1}{Z(x)} \exp\left(\frac{-E(\theta, x)}{RT}\right)$$

where $E(\theta, x)$ is the free energy of the RNA secondary structure θ computed by using Turner's energy parameters (Mathews *et al*., 1999), $Z(x)$ is the normalizing constant (partition function), R is the ideal gas constant [8.314 J/(K mol)] and T is temperature ($T = 37°\mathrm{C} = 310\,\mathrm{K}$ was used in this study).

We also tried two machine-learning-based models as the distribution $p(\theta|x)$: the Boltzmann likelihood (BL) model (Andronescu *et al*., 2010) and the CONTRAfold model (Do *et al*., 2006). In these models, internal parameters are *automatically* learned from known RNA secondary structures; in other words, those models do not use experimentally determined parameters. It should be emphasized that recent studies have suggested that machine learning-based models achieve better performance than the energy-based model with respect to the prediction of base pairs in RNA secondary structures (Andronescu *et al*., 2010; Do *et al*., 2006; Hamada *et al*., 2009).

The complete set of BPPs $\{p_{ij}\}_{1 \le i < j < |x|}$, called a BPP matrix (BPPM), can be efficiently computed by the inside–outside algorithm. See, for example, McCaskill (1990) for the details. We used the CentroidFold software (Hamada *et al*., 2009) to compute the BPPM for each probabilistic model. (See Supplementary Section S1 for the detailed command line options.)

Finally, a BPP (denoted by q_i) for a specific single position i in the RNA sequence x is defined by

$$q_i = \sum_{j:j<i} p_{ji} + \sum_{j:i<j} p_{ij} \quad (2)$$

which gives the probability that the nucleotide at position i in x is part of a base pair. In our study, we use the BPPs of Equation (2) for each nucleotide in an RNA sequence.

3 RESULTS AND DISCUSSION

In this study, each nucleotide was categorized into four types based on the combination of intermolecular hbonds and base pairs in the 3D structures: (i) *paired* nucleotides *with* intermolecular hbonds; (ii) *paired* nucleotides *without* intermolecular hbonds; (iii) *unpaired* nucleotides *with* intermolecular hbonds; and (iv) *unpaired* nucleotides *without* intermolecular hbonds. In Supplementary Table S1, we summarize the detailed numbers of nucleotides in each category in the dataset.

The average BPPs for the unpaired nucleotides with forming intermolecular hbonds were significantly lower than those for the nucleotides without hbonds (Fig. 1a). On the other hand, no difference was found between the BPPs of paired nucleotides with and without intermolecular hbonds (Fig. 1b). In the detailed distributions of the BPPs for the four types of nucleotides (Fig. 2), only the unpaired nucleotides with low BPPs (<0.1) were frequently found when the nucleotides were involved in intermolecular hbonds (Fig. 2a). In contrast, unpaired nucleotides with relatively high BPPs (≥0.1) were rarely found when the nucleotides were involved in the hbonds (Fig. 2a), whereas such nucleotides were more frequently found when the nucleotides were *not* involved in the hbonds (Fig. 2b). In addition, similar distributions were observed under the other two probabilistic models (BL and CONTRAfold) for calculating the BPPs (Supplementary Figs S1 and S3). These differences in the

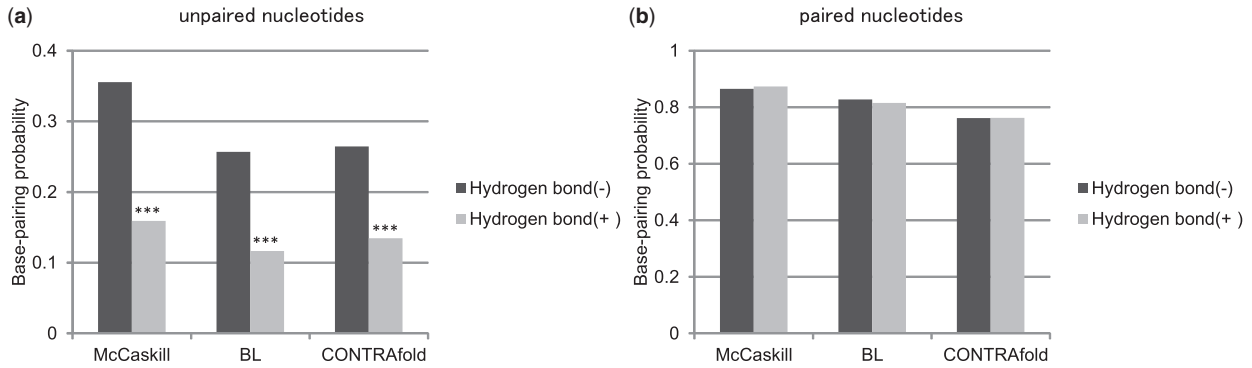

Fig. 1. Statistics of BPPs for RNAs in 68 protein–RNA complexes. Average of BPPs [see Equation (2) for formal definition] for (**a**) unpaired nucleotides and (**b**) paired nucleotides are calculated based on three models, the McCaskill model (McCaskill, 1990), the BL model (Andronescu *et al.*, 2010) and the CONTRAfold model (Do *et al.*, 2006), using CentroidFold software (Hamada *et al.*, 2009). Each nucleotide is classified into one of the two types: forming intermolecular hbonds (+; light gray bars) or not (−; dark gray bars). Asterisks indicate a statistically significant difference, with *t*-test, between the probabilities for the two types: $*P<0.05$, $**P<0.01$, and $***P<0.001$. See Supplementary Table S3 for detailed results of statistical test

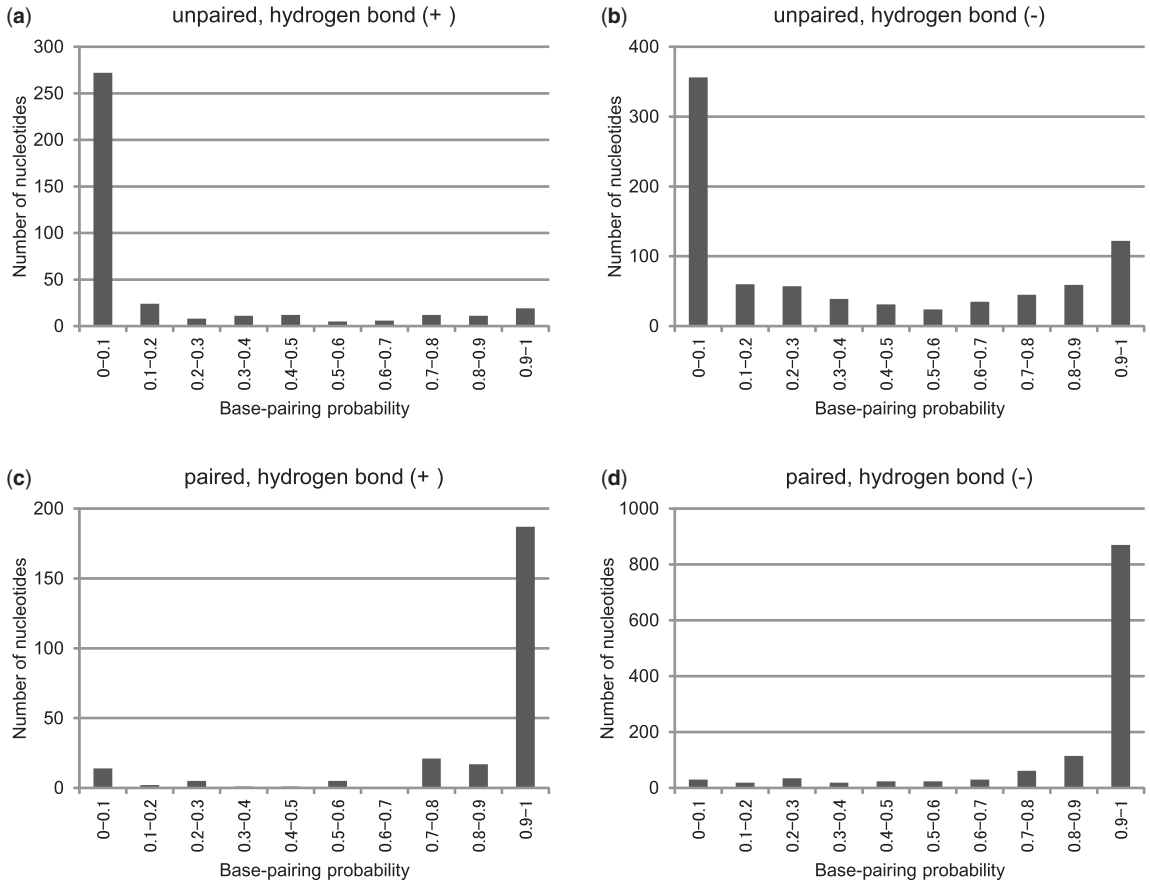

Fig. 2. Frequency distribution of BPPs for (**a**) unpaired nucleotides forming intermolecular hbonds with amino acids, (**b**) unpaired nucleotides not forming hbonds, (**c**) paired nucleotides forming hbonds and (**d**) paired nucleotides not forming hbonds, in 68 protein–RNA complexes. These BPPs are calculated based on the McCaskill model. See Supplementary Figures S1 and S3 for results based on the BL model and the CONTRAfold model, respectively

distributions of the BPPs could cause a significant difference in the average probabilities (Fig. 1a).

Among the unpaired nucleotides with lower BPPs, 43% nucleotides formed intermolecular hbonds with proteins in contrast to <20% nucleotides in the unpaired nucleotides with relatively high BPPs (Fig. 3). Similar differences were obtained when the other two models were used to calculate the BPPs (Supplementary Figs S2 and S4). These results imply that unpaired nucleotides with lower BPPs in the ensemble of possible secondary structures of the RNA more frequently form

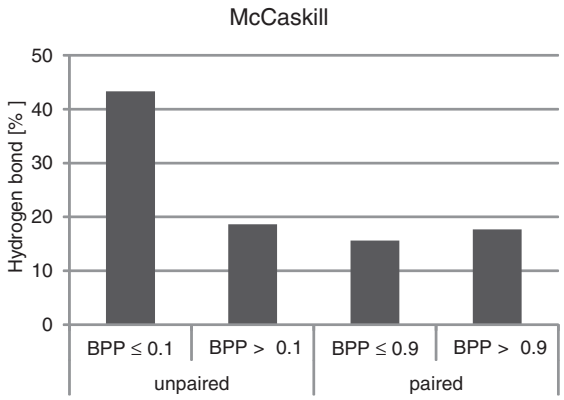

Fig. 3. Percentages of paired and unpaired nucleotides forming intermolecular hbonds in 68 protein–RNA complexes. Unpaired nucleotides are divided into two groups based on their BPPs calculated by using the McCaskill model: strong tendency to become unpaired (BPP ≤ 0.1) or not (BPP >0.1). Paired nucleotides are also divided into two groups: strong tendency to become paired (BPP >0.9) or not (BPP ≤ 0.9). See Supplementary Figures S2 and S4 for results based on the BL model and the CONTRAfold model, respectively

intermolecular hbonds with proteins than unpaired nucleotide with higher BPPs. The importance of the unpaired nucleotides for interacting with proteins is consistently shown in recent studies (Gupta and Gribskov, 2011; Iwakiri *et al*., 2012). In addition, our results suggest that investigating the *ensemble* of the secondary structures of the RNA is also important in the protein–RNA interactions, especially for the unpaired nucleotides.

A previous study suggested that, when using the McCaskill model, BPPs [Equation (1)] of two nucleotides that do *not* form a base pair in the reference structure are often high, falling in the range 0.9–1 [see Supplementary Fig. S13 in Hamada *et al*. (2009)]. Interestingly, a similar tendency was *not* observed for unpaired nucleotides with intermolecular hbonds, whereas it was observed for unpaired nucleotides without hbonds (see Fig. 2a and b), which also supports the idea that unpaired nucleotides with hbonds have a strong tendency to become unpaired in the ensemble of RNA secondary structures.

Our dataset used above included 27 complexes containing various tRNAs. The secondary structures of these tRNAs are known to be a similar cloverleaf structure, even though their primary sequences are not similar. To avoid the bias of the tRNA, a subset containing 42 complexes including a representative tRNA (PDBID: 1ASY) was also created (Supplementary Table S2). This subset was then used to calculate the BPPs based on the three different models (Supplementary Figs S5–S11). These results also showed that the BPPs of the unpaired nucleotides were low when the nucleotides were involved in intermolecular hbonds.

Finally, we investigated a discriminative power of the BPP for predicting the RNA residues either forming intermolecular hbond or not forming the bonds (Supplementary Fig. S12). As anticipated from the previous results, moderate discriminative power was observed for the case of unpaired nucleotides (Supplementary Fig. S13), whereas no discriminative power was seen for the case of paired nucleotides (Supplementary Fig. S14). Note that there was no significant difference of the powers among probabilistic models for RNA secondary structures.

In future work, we plan to incorporate these results into protein–RNA computational docking (Huang and Zou, 2013) or predicting the potential interface of protein–RNA complexes (Lewis *et al*., 2011) because our study provides novel and useful information about the nucleotides (of RNAs) that form intermolecular hbonds with the proteins.

4 CONCLUSION

In this study, we analyzed BPPs of RNAs involved in known protein–RNA interactions (complexes). Our findings are summarized as follows: (i) Within unpaired nucleotides average/distribution of BPPs with intermolecular nucleotides, hbond is different from those without hbond (Figs 1 and 2a and 2b). This difference is useful for discrimination between the unpaired nucleotides with hbonds and those without hbonds (Supplementary Fig. S13). (ii) Within paired nucleotides, average/distribution of BPPs with hbond is not different from those without hbond (Figs 1 and 2c and 2d). Hence, BPPs are not useful to discriminate the paired nucleotides with hbonds from those without hbonds (Supplementary Fig. S14). To our knowledge, this is the first study that considers the ensemble of RNA secondary structures in the analysis of protein–RNA interactions. The findings in this study would be useful when developing an algorithm for protein–RNA computational docking and predicting the potential interface of protein–RNA complexes.

ACKNOWLEDGEMENTS

The authors are grateful to the members of Tsukasa Fukunaga and Hisanori Kiryu for valuable comments. They also thank the anonymous reviewers for useful suggestions.

Funding: MEXT KAKENHI (Grant-in-Aid for Young Scientists (A): 24680031 to M.H.) (in part); MEXT KAKENHI (Grant-in-Aid for Young Scientists (B): 23700359 to T.K.) (in part). MEXT KAKENHI (Grant-in-Aid for Scientific Research (A): 30356357 to K.A.) (in part).

Conflict of Interest: none declared.

REFERENCES

Adachi,H. *et al*. (2011) Antagonistic RNA aptamer specific to a heterodimeric form of human interleukin-17A/F. *Biochimie*, **93**, 1081–1088.

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

Andronescu,M. *et al*. (2010) Computational approaches for RNA energy parameter estimation. *RNA*, **16**, 2304–2318.

Bahadur,R.P. *et al*. (2008) Dissecting protein-RNA recognition sites. *Nucleic Acids Res.*, **36**, 2705–2716.

Berman,H.M. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Burge,S.W. *et al*. (2012) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.*, **41**, D226–D232.

Do,C.B. *et al*. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90–98.

Ellis,J.J. *et al*. (2007) Protein-RNA interactions: structural analysis and functional classes. *Proteins*, **66**, 903–911.

Gopal,A. *et al*. (2012) Visualizing large RNA molecules in solution. *RNA*, **18**, 284–299.

Gupta,A. and Gribskov,M. (2011) The role of RNA sequence and structure in RNA–protein interactions. *J. Mol. Biol.*, **409**, 574–587.
Halvorsen,M. *et al.* (2010) Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet.*, **6**, e1001074.
Hamada,M. (2013) Fighting against uncertainty: an essential issue in bioinformatics. *Briefings in Bionformatics*, [Epub ahead of print, doi:10.1093/bib/bbt038, June 26, 2013].
Hamada,M. *et al.* (2009) Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics*, **25**, 465–473.
Huang,S.Y. and Zou,X. (2013) A nonredundant structure dataset for benchmarking protein-RNA computational docking. *J. Comput. Chem.*, **34**, 311–318.
Iwakiri,J. *et al.* (2012) Dissecting the protein-RNA interface: the role of protein surface shapes and RNA secondary structures in protein-RNA recognition. *Nucleic Acids Res.*, **40**, 3299–3306.
Kim,O.T. *et al.* (2006) Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. *Nucleic Acids Res.*, **34**, 6450–6460.
Kondo,J. and Westhof,E. (2011) Classification of pseudo pairs between nucleotide bases and amino acids by analysis of nucleotide-protein complexes. *Nucleic Acids Res.*, **39**, 8628–8637.
Leontis,N.B. and Westhof,E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA*, **7**, 499–512.
Lewis,B.A. *et al.* (2011) PRIDB: a Protein-RNA Interface Database. *Nucleic Acids Res.*, **39**, D277–D282.
Licatalosi,D.D. *et al.* (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, **456**, 464–469.
Mathews,D.H. *et al.* (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.
Ray,D. *et al.* (2009) Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat. Biotechnol.*, **27**, 667–670.
Sonavane,S. and Chakrabarti,P. (2009) Cavities in protein-DNA and protein-RNA interfaces. *Nucleic Acids Res.*, **37**, 4613–4620.
Yang,H. *et al.* (2003) Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res.*, **31**, 3450–3460.