Published online 16 May 2012 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* ***W29–W34***
doi:10.1093/nar/gks412

Rtips: fast and accurate tools for RNA 2D structure prediction using integer programming

Yuki Kato[1,*], Kengo Sato[2], Kiyoshi Asai[3,4] and Tatsuya Akutsu[5]

[1]Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), 8916-5 Takayama, Ikoma, Nara 630-0192, [2]Department of Biosciences and Informatics, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223-8522, [3]Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561, [4]Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-41-6, Aomi, Koto-ku, Tokyo 135-0064 and [5]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan

Received January 31, 2012; Revised April 6, 2012; Accepted April 20, 2012

ABSTRACT

We present a web-based tool set `Rtips` for fast and accurate prediction of RNA 2D complex structures. `Rtips` comprises two computational tools based on integer programming, `IPknot` for predicting RNA secondary structures with pseudoknots and `RactIP` for predicting RNA–RNA interactions with kissing hairpins. Both servers can run much faster than existing services with the same purpose on large data sets as well as being at least comparable in prediction accuracy. The `Rtips` web server along with the stand-alone programs is freely accessible at http://rna.naist.jp/.

INTRODUCTION

RNAs are versatile molecules for biological processes, working as messengers, regulators or catalysts in living cells. In particular, considerable attention has been focused on functions of regulatory non-coding RNAs. It is widely believed that there is a strong correlation between the 3D structure of an RNA molecule and its function. Since experimental determination of RNA 3D structure is difficult and their structures are hierarchical, computational prediction of secondary structures from a given single sequence or multiple sequences provides a major key to elucidating the potential functions of RNAs. Furthermore, interaction with another RNA or protein is often necessary for functional RNAs to perform their programmed tasks, and prediction of interacting structures is also an important problem in bioinformatics.

Taking as input either a single RNA sequence or a pair of RNA sequences, major software seeks to find an optimal secondary structure under a certain scoring function, given that the predicted structure has no complex motifs such as pseudoknots in intramolecular base pairings and kissing hairpins in intermolecular bindings. More specifically, a pseudoknot is typically formed from the base pairings between the unpaired bases of a loop and those outside the loop, whereas a kissing hairpin is caused by loop–loop interaction between two hairpin-type RNAs. Example predictive web tools are `mfold` (1), `RNAfold` (2) and `CentroidFold` (3) for RNA secondary structure prediction, and `PairFold` (4), `RNAhybrid` (5) and `IntaRNA` (6) for RNA–RNA interaction prediction. One reason why the complex motifs are disregarded is that the capability of handling such structural motifs results in high computational cost. However, it is observed that not a few number of these motifs occur in living cells, and thus these motifs should be considered in prediction algorithms to achieve more accurate prediction and avoid missing potential RNA genes in genome-wide sequence analysis. To this end, researchers have developed several tools together with web servers that can explicitly deal with such complicated motifs at the cost of computational efficiency such as `NUPACK` (7) and `pknotsRG` (8) for predicting secondary structures with pseudoknots, and `inteRNA` (9) for predicting RNA–RNA interactions with kissing hairpins. To summarize, it is desirable to clear the trade-off between the efficiency of a prediction algorithm and the class of predictable structures in order to broaden its applications.

To address this challenging problem, we have recently proposed two novel prediction methods, `IPknot` (10) for RNA secondary structure prediction including pseudoknots and `RactIP` (11) for RNA–RNA, interaction prediction including kissing hairpins, both of which employ integer programming (IP). Experimental validations of `IPknot` and `RactIP` indicate that our prediction methods are sufficiently accurate and quite fast even on large data sets as compared with several state-of-the-art methods [see (10,11)]. For easy access

*To whom correspondence should be addressed. Tel: +81 743 72 5232; Fax: +81 743 72 5239; Email: ykato@is.naist.jp

and use of those tools, we develop Rtips, a web server for Rna sTructure prediction using IP Scheme that comprises IPknot and RactIP. The website is free and open to all users, and there is no login requirement.

METHOD OVERVIEW

The methodology common to IPknot and RactIP is to combine the following two procedures when an RNA sequence or a pair of RNA sequences is given:

(1) approximate a posterior probability distribution over a space of complex structures by its factorization;
(2) maximize expected accuracy of a predicted structure by solving the corresponding IP problem.

In approximation of the probability distribution, we aim to decompose it into the product of probabilities defined over smaller base-pairing components, which are computationally easier to handle. The approximate probability distribution, explicitly represented as base-pairing posterior probabilities in the model, is incorporated into the objective function of the IP problem to find a secondary structure with the maximum expected accuracy (MEA). Expected accuracy can be expressed as the expected number of true predictions measured in base pairs. The IP problem is solved by GNU Linear Programming Kit (GLPK) in the web servers, which is freely available software for solving optimization problems. The advantage of using IP formulation is not only its strong descriptive power but also its flexibility and extensibility. In the framework of computing MEA, it is no longer necessary to consider the base pairs that do not contribute to improve expected accuracy, and thus we can take no account of them in advance.

The combination of the above procedures produces drastic speed-up in running time as well as good prediction accuracy. Therefore, the use of this strategy is very powerful to perform prediction even for large RNA sequences with complex motifs. Further details of our methodology can be found in (10,11).

GENERAL REMARKS

The top page of the Rtips web server provides links to respective web-based prediction services together with those to their source codes for stand-alone use and template programs to access the web services.

Each server accepts input by either entering RNA sequences directly or uploading FASTA files. The web interface has several optional parameters that affect prediction results. If the user does not adjust the parameters, the default values will be submitted to the server. Note that the default parameters related to calculating MEA (weight for true base pairs) were determined to obtain good predictions on many data sets and adjustment is hardly needed. Base-pairing posterior probabilities used in both tools are computed by RNAfold with parameters estimated by a Boltzmann likelihood-based method (12), which is based on McCaskill's dynamic programming algorithm (13) and thus we call it the McCaskill model, or by part of CONTRAfold (14), which is a machine learning-based predictor. If an illegal input is submitted to the server, the user will be notified of the inconsistency promptly. Each web interface for input includes automatic loading of several sample data to grasp the behavior of the tool, and provides interpretation of the output in the help page. It should be noted that we limit the size of input data to avoid overloading the servers, and the details of the restriction can be found in the help page of each server. If the size of submitted data is over the designated limit, the user is recommended to run the stand-alone program instead. If the user would like to integrate the functions of our servers with other web services, the template programs will be helpful.

After the job is submitted to the server, a prediction result can be found if the input is compatible. The result can be returned very quickly if the length of the submitted sequence is <400 nt. The user first finds a predicted 2D structure in dot-bracket representation, which can also be downloaded in Vienna format (2). To make the result easier to see, the server provides another graphical representation generated by VARNA (15). These graphics are embedded in the result page as PNG format, and those of original size are also available as PDF files.

IPKNOT SERVER

Input

The input is either a single RNA sequence or a multiple alignment of RNA sequences. If the user would like to know a secondary structure of a single RNA sequence, the sequence can be entered in plain or FASTA format into the field. Instead, the user can submit sequence information by uploading the corresponding FASTA file. Note that the length of the sequence must be at most 1500 nt. IPknot can also accept a multiple alignment of RNA sequences in CLUSTAL W format or multiple FASTA format to predict their consensus secondary structure. In this case, the alignment length is limited to 1500 nt. When pressing the 'Predict' button, the user can get a prediction result in the new page.

There are several parameters that IPknot can adjust. Level is the number of decompositions of a secondary structure where each decomposed substructure must have no pseudoknots. In other words, level can be considered as the number of kinds of brackets for indicating base pairs in dot-bracket representation. For example, level 1 uses just one kind of bracket '( )', level 2 uses two kinds of brackets '( )' and '[ ]', and in level 3, three kinds of brackets '( )', '[ ]' and "{ }" are used. Therefore, IPknot of level 1 is an ordinary secondary structure predictor that does not consider pseudoknots like mfold and RNAfold, and it is almost equivalent to CentroidFold. IPknot of level 2 aims to predict nested pseudoknots, whereas IPknot of level 3 seeks to predict pseudoknotted structures with nested pseudoknots. The server provides three kinds of scoring models that produce base-pairing posterior probabilities. The McCaskill and the CONTRAfold models take no account of pseudoknotted structures in

(a)

```
>MIDV
1     UUCUUUUUUAGUGGCAGUAAGCCUGGGAAUGGGGGCGACCCAGGCGUAUGAACAUAGUGUAACGCUCCCC
1     ....[[.............(((((((]].[[[[[[[.)))))))................]]]]]]].
```

(b)

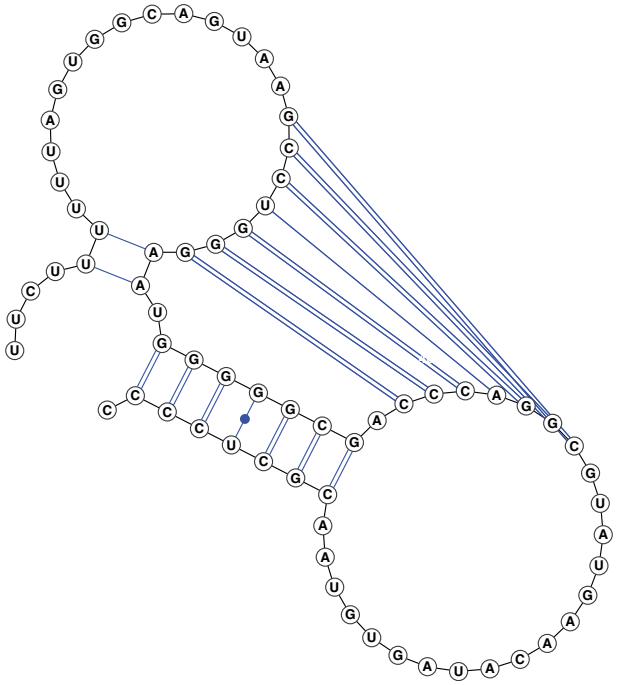

Figure 1. Screenshot of the result page produced by the IPknot server when a sample sequence is submitted. The 'MIDV' sequence shown above is the 6K/TF ribosomal frameshift site of Middelburg virus, which was taken from PseudoBase (16). (**a**) Dot-bracket representation. (**b**) 2D diagram.

each decomposed substructure of IPknot, whereas the NUPACK model considers a certain class of pseudoknots in each substructure. Accordingly, the NUPACK model can be more accurate than the other two models to predict pseudoknotted secondary structures. However, a sequence of length >80 nt is too long for the elaborate NUPACK model to predict fast, and the server rejects the input. Besides, the NUPACK model is not supported for alignment input due to the computational cost. The user can choose whether the base-pairing probabilities of the McCaskill and CONTRAfold models defined over pseudoknot-free structures are refined or not. In the refinement procedure, the base-pairing probabilities are recalculated using the prediction result of the first run of IPknot. It should be noted that the choice of the NUPACK model disallows the refinement due to the computational cost of its iterative use. The weights of arbitrary positive numbers for respective levels can be specified in the web interface. Specifically, they represent the rate of true base pairs in the predicted secondary structure, which determine prediction accuracy. In general, if the weight increases, the algorithm aims to predict more base pairs and sensitivity of a prediction will get better. On the other hand, if the weight decreases, the algorithm tries to predict less base pairs and positive predictive value (PPV) will be enhanced. In this sense, the weights are balanced parameters between sensitivity and PPV.

Output

The user can find a predicted secondary structure with MEA. Figure 1a shows an example of a predicted MEA structure in dot-bracket representation where matching brackets indicate a base pair. Note that different forms of brackets, say '( )' and '[ ]' cross each other, meaning that the predicted structure includes pseudoknots. In addition to a downloadable Vienna file, the server can generate a BPSEQ formatted file for base-pairing information. A 2D diagram of the predicted structure along with its arc representation is displayed by running the VARNA program in the background [see Figure 1b]. Note that in the 2D diagram, an A–U pair is indicated by a single line with a bullet, a G–C pair is shown by a double line and a G–U pair is represented by a single line. In the result page for consensus structure prediction, the user can get the input alignment followed by the MEA common secondary structure in dot-bracket representation (Figure 2). Furthermore, a file that contains the predicted consensus structure as well as all input sequences in FASTA format is also downloadable. Interpretation of the other figures of a predicted structure is the same as that of a single sequence.

Validation

We validated prediction performance of IPknot on various data sets. One example of predicting a structure

```
Tomato_mosaic_virus.1   GUGUCUUGGAGCGCGCGGAGUAAACAUAUAUGGUUCAUAUAUGUCCGUAGGCACGUAAAAAAAGCGA
Tobacco_mosaic_virus.1  GUGUCUUGGAUCGCGCGGGUCAAAUGUAUAUGGUUCAUAUACAUCCGCAGGCACGUAAUAAA-GCGA
Rehmannia_mosaic_vir.1  GUGUCUUGGUUCGCGCGGGUCAAGUGUAUAUGGUGCAUAUACAUCCGUAGGCACGUAAUAAA-GCGA
B.pepper.1              GUGUCUUGGAACGCGCGGGUCAAAUAUAAGUGGUUCACUUAUAUCCGUAGGCACGAAAAAUU-GCGU
SS_cons                 (((((((([[[....((((]]]].(((((((((...)))))))))))))))))).............
```

Figure 2. Part of the result page when a multiple sequence alignment is submitted to the IPknot server. The sample alignment of tRNA-like structures was taken from Rfam 10.1 (17).

of a single sequence is a test on 388 non-redundant sequences of length at most 500 nt with at least one pseudoknot, showing 0.567, 0.578 and 0.570 in sensitivity, PPV and Matthews correlation coefficient (MCC), respectively, on average. Although these values may seem small, this is the best prediction performance as compared with other seven competitive methods [see (10)]. Another test on 67 alignments containing five sequences for consensus pseudoknotted structure prediction indicates 0.706, 0.717 and 0.707 in average sensitivity, PPV and MCC, respectively. An example of computation time is 3.95 s on a single sequence of length 989 nt, which was measured on the Linux machine identical to the web server (see the Implementation section for specifications). From the detailed validations in (10), IPknot is quite fast and sufficiently accurate as compared with several state-of-the-art methods.

RACTIP SERVER

Input

The input is a pair of RNA sequences in plain or FASTA format. Notice that each sequence must be put in $5'-3'$ direction. Instead, the user can submit sequence information by uploading two separate FASTA files. Note that the sum of the lengths of two sequences must be at most 1000 nt, otherwise the server rejects the input. The user can get a prediction result in the new page by pressing the 'Predict' button.

The RactIP server offers two options. It provides the two aforementioned scoring models named CONTRAfold and McCaskill that produce internal base-pairing probabilities. In contrast, hybridization probabilities related to external base pairs are calculated by a variant of RNAduplex in the Vienna RNA package with parameters estimated by the Boltzmann likelihood-based method (12). Although the distinct models are used to derive internal base-pairing probabilities and hybridization probabilities, the approximation of the probability distribution enables us to select models separately that yield good predictions. Prediction accuracy depends on the specified weights as in the case of IPknot.

Output

The output is a predicted joint secondary structure with MEA. The MEA structure is first shown in dot-bracket representation, where round brackets '( )' indicate an internal base pair and square brackets '[ ]' denote an external base pair (binding site) [see Figure 3a]. We should draw attention to the fact that there are no internal pseudoknots and external crossing interactions in joint structures predicted by RactIP, which is due to the assumption in the model. The free energy of the predicted joint secondary structure is given by employing RNAeval in the Vienna RNA package. A drawing of the predicted joint structure in arc representation is displayed, where blue arcs represent internal base pairs, red arcs stand for external interactions, and '$5' \rightarrow 3'$' at the bottom shows the orientation of each RNA sequence [see Figure 3b].

Validation

We tested on 23 known RNA–RNA interaction pairs with total length of two sequences at most 300 nt. Five pairs out of 23 that are known to include kissing hairpins were used to evaluate the accuracy of predicted joint structures, indicating 0.963, 0.873 and 0.913 in sensitivity, PPV and MCC, respectively, on average. Looking at binding sites to assess the prediction accuracy on 23 RNA pairs, RactIP yields 0.833, 0.885 and 0.852 in average sensitivity, PPV and MCC, respectively. An example of running time measured on the machine described above is 0.855 s on an RNA–RNA pair of total lengths 306 nt. Detailed validations shown in (11) demonstrate that RactIP is extremely fast and sufficiently accurate as compared with several competitive prediction methods.

IMPLEMENTATION

The web server was implemented on a Linux CentOS 5 machine with Core i7-950 3.06 GHz CPU and 6.00 GB RAM using Apache, XHTML, JavaScript and PHP. The source codes for stand-alone use are written in C++, and the template programs to access the servers and parse the output are written in Perl.

DISCUSSION

The presented web tool set Rtips can predict sets of canonical base pairs from a set of input RNA sequences quite fast and accurately even if a secondary structure to be predicted is complicated. The proposed methods in Rtips are heuristic in the sense that they superimpose

(a)

```
>DIS
1     CUCGGCUUGCUGAGGUGCACACAGCAAGAGGCGAG
1     ((((.(((((((..[[[[[[.)))))))...))))
>DIS
1     CUCGGCUUGCUGAGGUGCACACAGCAAGAGGCGAG
1     ((((.(((((((..]]]]]].)))))))...))))
```

(b)

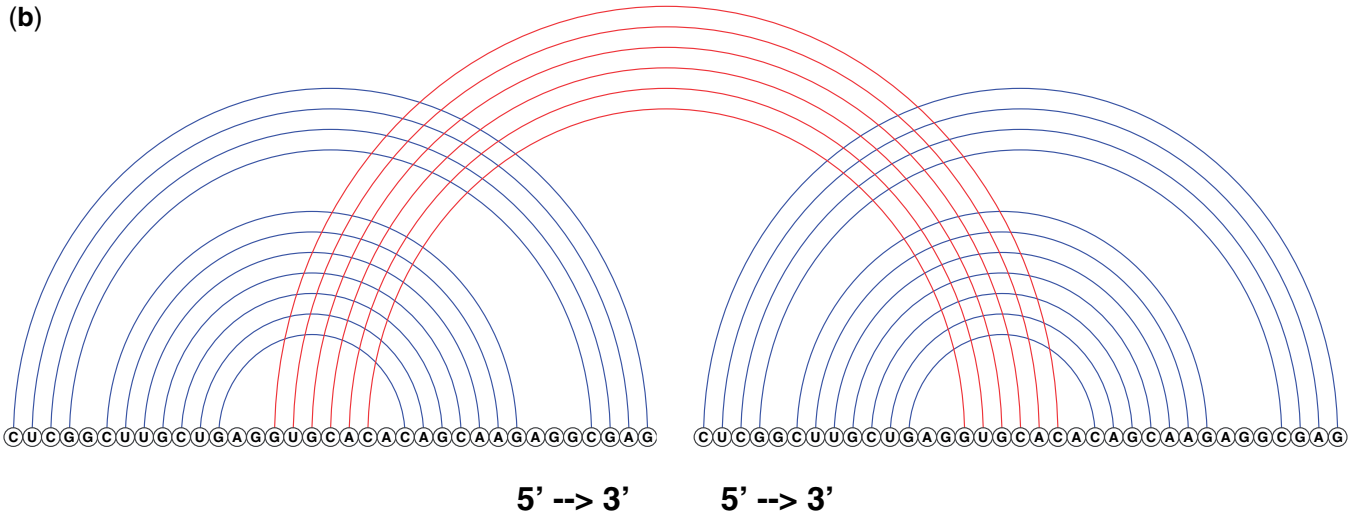

Figure 3. A sample output produced by the RactIP server. The above 'DIS–DIS' pair is caused by interaction between the partially self-complementary loops of the dimerization initiation sites of the HIV-1 genomic RNA (18). (**a**) Dot-bracket representation. (**b**) Arc representation.

prediction results of primitive base-paired substructures to compose more complex secondary structures.

Other heuristic web tools that adopt the superimposition include HotKnots (19,20) for predicting secondary structures with pseudoknots and PETcofold (21,22) for predicting RNA–RNA interactions of multiple RNA sequences. IPknot is at least comparable in accuracy to HotKnots 2.0 (20) and can run an order of magnitude faster on large RNAs as shown in tests on various data sets in (10). The literature (21) reports that accuracy of RactIP is lower than that of PETcofold on condition that a set of homologous sequences is available, but running time of RactIP is much shorter. Equally importantly, RactIP needs no multiple alignment of RNA sequences that are expected to be homologous.

Our methodology will be powerful and useful enough to be applied to other important problems in RNA bioinformatics, including RNA structural alignment, prediction of non-canonical base pairs and genome-scale analysis associated with structure prediction. We have just got off to a good start to address these tasks and provide a potential extension of the server.

ACKNOWLEDGEMENTS

The authors would like to thank all people who are involved in discussion about improvement on the server and testing the robustness.

FUNDING

Grant-in-Aid for Young Scientists (B) from Japan Society for the Promotion of Science [#22700313 to Y.K., #22700305 to K.S.]. Funding for open access charge: Grant-in-Aid for Challenging Exploratory Research from Japan Society for the Promotion of Science [#23650153].

Conflict of interest statement. None declared.

REFERENCES

1. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
2. Gruber,A.R., Lorenz,R., Bernhart,S.H., Neuböck,R. and Hofacker,I.L. (2008) The Vienna RNA Websuite. *Nucleic Acids Res.*, **36**, W70–W74.
3. Sato,K., Hamada,M., Asai,K. and Mituyama,T. (2009) CentroidFold: a web server for RNA secondary structure prediction. *Nucleic Acids Res.*, **37**, W277–W280.
4. Andronescu,M., Aguirre-Hernández,R., Condon,A. and Hoos,H.H. (2003) RNAsoft: a suite of RNA secondary structure prediction and design software tools. *Nucleic Acids Res.*, **31**, 3416–3422.
5. Krüger,J. and Rehmsmeier,M. (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res.*, **34**, W451–W454.
6. Smith,C., Heyne,S., Richter,A.S., Will,S. and Backofen,R. (2010) Freiburg RNA tools: a web server integrating IntaRNA, ExpaRNA and LocARNA. *Nucleic Acids Res.*, **38**, W373–W377.
7. Zadeh,J.N., Steenberg,C.D., Bois,J.S., Wolfe,B.R., Pierce,M.B., Khan,A.R., Dirks,R.M. and Pierce,N.A. (2011) Software news and updates NUPACK: analysis and design of nucleic acid systems. *J. Comput. Chem.*, **32**, 170–173.

8. Reeder,J., Steffen,P. and Giegerich,R. (2007) pknotsRG: RNA pseudoknot folding including near-optimal structures and sliding windows. *Nucleic Acids Res.*, **35**, W320–W324.
9. Aksay,C., Salari,R., Karakoc,E., Alkan,C. and Sahinalp,S.C. (2007) taveRNA: a web suite for RNA algorithms and applications. *Nucleic Acids Res.*, **35**, W325–W329.
10. Sato,K., Kato,Y., Hamada,M., Akutsu,T. and Asai,K. (2011) IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*, **27**, i85–i93.
11. Kato,Y., Sato,K., Hamada,M., Watanabe,Y., Asai,K. and Akutsu,T. (2010) RactIP: fast and accurate prediction of RNA–RNA interaction using integer programming. *Bioinformatics*, **26**, i460–i466.
12. Andronescu,M., Condon,A., Hoos,H.H., Mathews,D.H. and Murphy,K.P. (2010) Computational approaches for RNA energy parameter estimation. *RNA*, **16**, 2304–2318.
13. McCaskill,J.S. (1990) The equilibrium partition function and base pair probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
14. Do,C.B., Woods,D.A. and Batzoglou,S. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90–e98.
15. Darty,K., Denise,A. and Ponty,Y. (2009) VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.
16. van Batenburg,F.H.D., Gultyaev,A.P., Pleij,C.W.A., Ng,J. and Oliehoek,J. (2000) PseudoBase: a database with RNA pseudoknots. *Nucleic Acids Res.*, **28**, 201–204.
17. Gardner,P.P., Daub,J., Tate,J., Moore,B.L., Osuch,I.H., Griffiths-Jones,S., Finn,R.D., Nawrocki,E.P., Kolbe,D.L. and Eddy,S.R. (2011) Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res.*, **39**, D141–D145.
18. Paillart,J.C., Skripkin,E., Ehresmann,B., Ehresmann,C. and Marquet,R. (1996) A loop–loop "kissing" complex is the essential part of the dimer linkage of genomic HIV-1 RNA. *Proc. Natl. Acad. Sci. USA*, **93**, 5572–5577.
19. Ren,J., Rastegari,B., Condon,A. and Hoos,H.H. (2005) HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. *RNA*, **11**, 1494–1504.
20. Andronescu,M., Pop,C. and Condon,A. (2010) Improved free energy parameters for RNA pseudoknotted secondary structure prediction. *RNA*, **16**, 26–42.
21. Seemann,S.E., Richter,A.S., Gesell,T., Backofen,R. and Gorodkin,J. (2011) PETcofold: predicting conserved interactions and structures of two multiple alignments of RNA sequences. *Bioinformatics*, **27**, 211–219.
22. Seemann,S.E., Menzel,P., Backofen,R. and Gorodkin,J. (2011) The PETfold and PETcofold web servers for intra- and intermolecular structures of multiple RNA sequences. *Nucleic Acids Res.*, **39**, W107–W111.