Heuristic RNA pseudoknot prediction including intramolecular kissing hairpins

JANA SPERSCHNEIDER,[1] AMITAVA DATTA,[1] and MICHAEL J. WISE[1,2]
[1]School of Computer Science and Software Engineering, University of Western Australia, Perth WA 6009, Australia
[2]School of Biomolecular, Biomedical, and Chemical Sciences, University of Western Australia, Perth WA 6009, Australia

ABSTRACT

Pseudoknots are an essential feature of RNA tertiary structures. Simple H-type pseudoknots have been studied extensively in terms of biological functions, computational prediction, and energy models. Intramolecular kissing hairpins are a more complex and biologically important type of pseudoknot in which two hairpin loops form base pairs. They are hard to predict using free energy minimization due to high computational requirements. Heuristic methods that allow arbitrary pseudoknots strongly depend on the quality of energy parameters, which are not yet available for complex pseudoknots. We present an extension of the heuristic pseudoknot prediction algorithm DotKnot, which covers H-type pseudoknots and intramolecular kissing hairpins. Our framework allows for easy integration of advanced H-type pseudoknot energy models. For a test set of RNA sequences containing kissing hairpins and other types of pseudoknot structures, DotKnot outperforms competing methods from the literature. DotKnot is available as a web server under http://dotknot.csse.uwa.edu.au.

Keywords: pseudoknots; pseudoknot prediction; intramolecular kissing hairpin; RNA structure prediction

INTRODUCTION

Pseudoknots are versatile structural elements that are abundant in both cellular and viral RNA. The first pseudoknots were experimentally identified in the early 1980s in tRNA-like structures in plant viruses (Rietveld et al. 1982, 1983). Subsequently, the pseudoknot folding principle was established (Pleij et al. 1985) and, over the years, many pseudoknots with an astonishing number of diverse functions have been discovered. Pseudoknots are known to participate in protein synthesis, genome and viral replication, and ribozyme structure and function (Staple and Butcher 2005; Brierley et al. 2007, 2008; Giedroc and Cornish 2009). H-type pseudoknots form when unpaired bases in a hairpin loop bond with unpaired bases outside the loop and have been found essential in the context of programmed -1 ribosomal frameshifting, telomerase RNA, and viral internal ribosome entry sites.

Simple H-type pseudoknots are the best-studied group of RNA pseudoknots and constitute the vast majority of entries in the pseudoknot database Pseudobase (van Batenburg et al. 2001). However, this should not lead to the conclusion that different types of pseudoknots are less frequent or less important in RNA three-dimensional folding and function. A more complex pseudoknot forms when unpaired bases in a hairpin loop bond with unpaired bases in another hairpin loop (Fig. 1). This type of pseudoknot is called an intramolecular kissing hairpin, H-H type pseudoknot, or loop-loop pseudoknot. The hairpin loops can also be located in different RNA molecules, which is referred to as an intermolecular kissing hairpin, RNA-RNA interaction, or a kissing complex (Brunel et al. 2002). Intramolecular kissing hairpins have been reported in different virus families as essential features for viral replication (Melchers et al. 1997; Verheije et al. 2002; Friebe et al. 2005). Kissing hairpins have also been found in some hammerhead ribozymes (Song et al. 1999; Gago et al. 2005), the Varkud satellite ribozyme (Rastogi et al. 1996), or as part of the signal recognition particle (Larsen and Zwieb 1991).

Due to the crossing of three stems, intramolecular kissing hairpin prediction is more complex than prediction of simple H-type pseudoknots. Most discoveries of kissing hairpins have been made in the laboratory with little aid of computational methods due to the lack of practical prediction algorithms. General kissing interactions are hard to predict as it leaves the field of secondary structure prediction,

Reprint requests to: Jana Sperschneider, School of Computer Science and Software Engineering, University of Western Australia, Perth WA 6009, Australia; e-mail: janaspe@csse.uwa.edu.au; fax: 61-8-6488-1089.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.2394511.

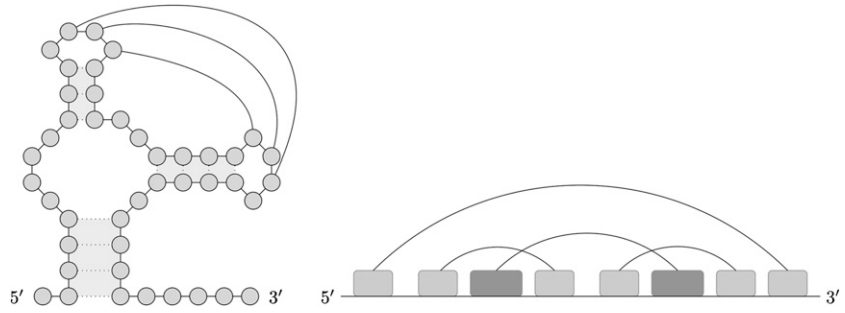

FIGURE 1. Intramolecular kissing hairpin structure and its representation as crossing intervals on the line.

both in terms of the computational complexity and the energy model. Given an RNA sequence, the minimum free-energy (MFE) secondary structure without crossing base pairs can be calculated in $O(n^3)$ time and $O(n^2)$ space under an additive energy model using dynamic programming (Zuker and Stiegler 1981; Lyngsø et al. 1999). Free-energy minimization, including general pseudoknots, has been proven to be an NP-complete problem (Lyngsø and Pedersen 2000a). By restricting the types of pseudoknots that can be predicted, polynomial-time dynamic programming methods can be achieved. Rivas and Eddy (1999) introduced pknots for MFE structure prediction, including a broad class of pseudoknots such as chains of pseudoknots and kissing hairpins which take $O(n^6)$ time and $O(n^4)$ space. More practical algorithms run in $O(n^5)$ or $O(n^4)$ time, depending on the pseudoknot target class; however, kissing hairpins are not included in the recursion schemes (Akutsu 2000; Dirks and Pierce 2003; Reeder and Giegerich 2004). A tree-adjoining grammar algorithm by Uemura et al. (1999) using $O(n^5)$ time and $O(n^4)$ space allows pseudoknot chains of length three under a very simple energy model. Lyngsø and Pedersen (2000b) give a high-level description of a dynamic programming algorithm using $O(n^5)$ time and $O(n^3)$ space, which can predict kissing hairpins. A dynamic programming method requiring $O(n^5)$ time and $O(n^4)$ space for MFE structure prediction was presented by Chen et al. (2009) and includes kissing hairpins and chains of four overlapping stems. Apart from pknots, no implementations are readily available for MFE structure prediction, including intramolecular kissing hairpins.

Due to the computational complexity of dynamic programming for pseudoknot prediction, heuristic algorithms have been developed. A number of heuristic RNA structure prediction methods explicitly include kissing hairpins. FlexStem is a heuristic algorithm with the ability to fold overlapping pseudoknots, i.e., intramolecular kissing hairpins (Chen et al. 2008). HFold is based on the MFE folding for secondary structures and hierarchically calculates a joint structure using the available bases from a given secondary structure. The predicted structure may contain pseudoknots and (nested) kissing hairpins (Jabbari et al. 2008). A range of heuristic RNA structure prediction algorithms cover general types of pseudoknots and may, therefore, implicitly predict kissing hairpins. However, the pseudoknot target class remains elusive and there are no specific energy parameters for kissing hairpins. For example, simple kissing hairpins can be predicted by iterated stem adding procedures such as iterated loop matching (ILM) and HotKnots (Ruan et al. 2004; Ren et al. 2005; Andronescu et al. 2010). It must be noted that the underlying energy parameters may not be tuned for kissing hairpin prediction, and thus only very stable kissing hairpins are likely to be predicted.

Here, we present an extension of the heuristic pseudoknot search method DotKnot for prediction of H-type pseudoknots (Sperschneider and Datta 2010). DotKnot was initially designed as a specialized H-type pseudoknot folding method which returns only the detected pseudoknots for a given sequence. Our main contributions reported here are the following:

- Efficient prediction of a wider class of pseudoknots, namely intramolecular kissing hairpins.
- Prediction of a global structure to allow for performance evaluation with widely used algorithms for secondary structure prediction including pseudoknots.
- Prediction of a number of near-optimal pseudoknot and kissing hairpin candidates for further investigation by the user.

The main idea of the DotKnot method is to assemble pseudoknots in a constructive fashion from the secondary structure probability dot plot calculated by RNAfold (McCaskill 1990; Hofacker et al. 1994). Using a low-probability threshold, pseudoknotted stems can be seen in the dot plot. From the set of stem candidates found in the dot plot, DotKnot derives a candidate set of secondary structure elements, H-type pseudoknots, and kissing hairpins. The presence of the structure elements in the global structure is verified using maximum weight independent set calculations.

There are two main advantages of this heuristic approach. First, it is very efficient and therefore practical for longer RNA sequences. This is important as kissing loop interactions are known to stabilize the overall tertiary folding and are often long-range interactions. In contrast, kissing hairpin prediction using dynamic programming suffers from high computational requirements. For example, pknots is only able to run for sequences shorter than, say, 150 nt. Second, practical dynamic programming methods are fairly restricted with regards to the underlying additive energy model; however, nonadditive H-type pseudoknot energy models have been developed that are based on the important

TABLE 1. RNA types and sequences used for kissing hairpin prediction

RNA Type	Sequence ID	Reference
SRP RNA	ArcFul-SRP, Bsub-SRP, Hs-SRP, Mjann-SRP, Halo-SRP, TheCel-SRP	Zwieb and Müller (1997)
Viral RNA	CoxB3, Echo6, Ent69, HCV, PRRSV, WNV, HCoV229E	Wang et al. (1999), Friebe et al. (2005), Verheije et al. (2002), Shi et al. (1996), Herold and Siddell (1993)
Ribozyme	satRPV, NeuroVS, CChMVd, PLMVd, EColi-P6	Song et al. (1999), Rastogi et al. (1996), Gago et al. (2005), Bussiere et al. (2000), Harris et al. (2001)

interference between opposite stems and loops (Gultyaev et al. 1995; Cao and Chen 2006, 2009). Heuristic methods such as DotKnot, which construct a number of pseudoknot candidates, allow for easy integration of such nonadditive H-type pseudoknot energy models, which can drastically improve prediction accuracy.

In the remainder of this paper, we evaluate the performance of DotKnot and several other methods from the literature and discuss the results for kissing hairpin prediction in detail. Afterward a description of the algorithmic framework of the DotKnot method is given and we show how DotKnot derives a global structure and near-optimal pseudoknots.

RESULTS

Kissing hairpin prediction can be handled by an extended version of DotKnot, pknots, and FlexStem. These methods are chosen for the evaluation because they employ specific energy parameters for kissing hairpins. The secondary structure prediction algorithm RNAfold is also included in the testing to compare results to a hierarchical folding approach where the kissing interaction is added by hand *after* obtaining the MFE structure. Note that only those algorithms that are freely available are compared. No implementation is available for a dynamic programming method for kissing hairpin

TABLE 2. Summary of prediction results using an extended version of DotKnot

Sequence		DotKnot				pknots				FlexStem				RNAfold			
ID	Nt	*S*	PPV	MCC	*r*	*S*	PPV	MCC	*r*	*S*	PPV	MCC	*r*	*S*	PPV	MCC	*r*
ArcFul-SRP	45	100	100	**1**	1/1	31.1	31.1	0.3	0/0	37.5	40	−0.15	0/0	0	0	−0.66	0/0
	310	92.7	92.7	**0.84**	1/1	*	*	*	*	71.8	78.2	0.49	0/0	80.9	82.4	0.6	0/0
Bsub-SRP	43	93.3	100	**0.93**	1/1	60	100	0.64	0/0	33.3	33.3	−0.29	0/0	0	0	−0.7	0/0
	270	82.3	91.9	**0.73**	1/1	*	*	*	*	14.6	15.7	−0.49	0/0	64.6	68.1	0.3	0/0
Hs-SRP	40	93.3	87.5	**0.73**	1/1	73.3	100	0.72	0/0	66.7	66.7	0.17	0/0	73.3	100	0.72	0/0
	299	33.3	36.5	−0.19	1/1	*	*	*	*	22.9	26.4	−0.33	0/0	64.8	70.1	**0.37**	0/0
Mjann-SRP	45	100	100	**1**	1/1	100	100	**1**	1/1	40	40	−0.07	0/0	0	0	−0.6	0/0
	330	95	97.5	**0.91**	1/1	*	*	*	*	79.3	80.7	0.55	0/0	86.8	89.7	0.73	0/0
Halo-SRP	45	66.7	76.9	0.47	0/0	60	75	0.37	0/0	66.7	83.3	**0.52**	0/0	60	75	0.37	0/0
	303	12.4	11.1	−0.47	1/2	*	*	*	*	9	7.7	−0.5	0/0	55.1	49	**0.16**	0/0
TheCel-SRP	45	100	93.8	**0.93**	1/1	33.3	33.3	−0.14	0/0	0	0	−0.57	0/0	33.3	38.5	−0.14	0/0
	318	78.2	78.2	**0.56**	1/1	*	*	*	*	53.6	53.6	0.13	0/0	70	68.8	0.38	0/0
CoxB3	121	60	63.6	0.37	1/1	71.4	86.2	**0.66**	0/0	80	71.8	0.56	0/0	71.4	71.4	0.51	0/0
Echo6	121	97	86.5	**0.86**	1/1	72.7	68.6	0.49	0/0	84.8	71.8	0.61	0/0	81.8	79.4	0.67	0/0
Ent69	121	62.9	56.4	0.3	1/2	71.4	67.6	**0.45**	0/0	71.4	62.5	0.39	0/0	62.9	66.7	0.4	0/0
HCV	75	48.3	60.9	0.07	0/0	69	83.3	0.45	0/0	72.4	80.8	0.42	0/0	69	87	**0.52**	0/0
	343	44.8	52	0.44	0/0	*	*	*	*	48.3	51.9	**0.45**	0/0	44.8	46.4	0.4	0/0
PRRSV	66	40	58.8	0.05	1/1	72	94.7	**0.64**	0/0	44	55	−0.03	0/0	72	94.7	**0.64**	0/0
	459	40	34.5	**0.33**	1/1	*	*	*	*	0	0	−0.09	0/0	28	14.3	0.13	0/0
WNV	96	85.7	90.9	**0.74**	1/1	62.9	73.3	0.38	0/0	80	87.5	0.64	0/0	82.9	90.6	0.71	0/0
satRPV	72	81.8	81.8	**0.68**	1/1	77.3	77.3	0.58	0/0	59.1	76.5	0.47	0/0	59.1	68.4	0.39	0/0
NeuroVS	87	66.7	50	0.21	1/1	87.5	77.8	**0.69**	0/0	50	42.9	0.03	0/0	50	42.9	0.03	0/0
	176	88.1	89.7	**0.78**	0/0	*	*	*	*	72.9	81.1	0.57	0/0	83.1	86	0.7	0/0
CChMVd	74	78.3	62.1	**0.33**	1/1	60.9	51.9	0.11	0/0	30.4	25.9	−0.28	0/0	60.9	53.8	0.15	0/0
PLMVd	75	65.6	95.5	0.51	1/1	71.9	92	0.5	0/0	81.3	100	**0.73**	0/0	81.3	89.7	0.53	0/0
EColi-P6	212	83.6	76.1	**0.64**	1/1	*	*	*	*	49.2	46.9	0.14	0/0	24.6	25.9	−0.18	0/0
HCoV229E	54	100	100	**1**	1/1	79.2	100	0.66	0/0	70.8	89.5	0.31	0/0	79.2	100	0.66	0/0
	224	100	100	**1**	1/1	*	*	*	*	79.2	79.2	0.76	0/0	79.2	90.5	0.83	0/0

In each sequence, one kissing hairpin has been reported in the literature. We use pknots version 1.05, FlexStem version 1.3 and the RNAfold web server (Gruber et al. 2008). The * symbol means that we were not able to run the algorithm to completion due to computational requirements. The ratio r = (number of correctly predicted kissing hairpins) / (number of predicted kissing hairpins) is also reported.

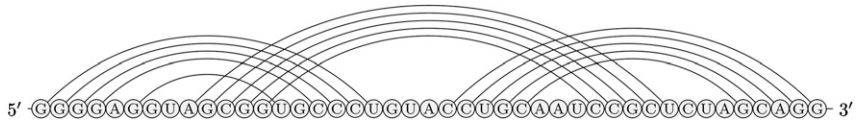

FIGURE 2. *Bacillus subtilis* SRP RNA kissing hairpin structure as found in the Signal Recognition Particle Database (Zwieb and Müller 1997).

prediction described in Lyngsø and Pedersen (2000b), Chen et al. (2009), and HFold (Jabbari et al. 2008).

The test set for the kissing hairpin prediction evaluation is shown in Table 1. The number of kissing hairpins described and verified in the literature is fairly limited. For long RNA sequences including kissing hairpins, such as the signal recognition particle RNA (SRP RNA), structure prediction is also performed for the short sequence exactly harboring the kissing hairpin. This is done in order to compare prediction results to the computationally expensive pknots and to observe whether prediction accuracy improves for all methods if a short kissing hairpin sequence is given.

For each kissing hairpin reference structure in the test set, the predicted base pairs are analyzed and results are shown in Table 2. The number of correctly and incorrectly predicted base pairs in the global structure is counted (TP and FP). The number of base pairs in the reference structure that were not predicted is also reported (FN). Sensitivity S is defined as $S = 100 \times (TP/TP + FN)$, positive predictive value (PPV) as $PPV = 100 \times (TP/TP + FP)$, and Matthews correlation coefficient (MCC) as:

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

The Matthews correlation coefficient is in the range from -1 to 1, where 1 corresponds to a perfect prediction and -1 to a prediction that is in total disagreement with the reference structure.

RESULTS

Structures with kissing hairpins

Signal recognition particle (SRP) RNA

The signal recognition particle (SRP) is a protein-RNA complex and participates in the translocation of proteins across membranes (Keenan et al. 2001). At the center is the SRP RNA, which typically consists of around 300 nt. A number of SRP RNA sequences with predicted secondary structures are available, for example for *Archaeoglobus fulgidus*, *Bacillus subtilis*, *Homo sapiens*, *Methanococcus jannaschii*, *Halobacterium halobium*, and *Thermococcus celer* (Zwieb and Samuelsson 2000). Tertiary kissing interactions have been established close to the 5′-end using phylogenetic analysis and molecular modeling (Larsen and Zwieb 1991; Zwieb and Müller 1997). The highly conserved kissing hairpin is a compact structure (Fig. 2). We evaluated the predictions for these six SRP RNA sequences and also for the corresponding six short sequences exactly harboring the kissing interaction. DotKnot predicts a kissing hairpin for all sequences except for the short *Halobacterium halobium* SRP RNA and shows the highest MCC for nine of the 12 sequences. Both DotKnot and pknots give perfect predictions (MCC = 1) for the *Methanococcus jannaschii* kissing hairpin and DotKnot also perfectly predicts the kissing hairpin in *Archaeoglobus fulgidus*. The prediction results for all four methods are poor for the long *Halobacterium halobium* SRP RNA. For the short sequence stretch, DotKnot returns a pseudoknot with lower free energy than the kissing hairpin. However, the desired kissing hairpin structure is found as the best near-optimal pseudoknot with lowest free energy (MCC = 1).

Viral replication

A kissing hairpin involving the poly(A)-tail is essential for synthesis of negative-strand RNA in the enteroviral 3′-UTR, e.g., in the Coxsackie B3 virus (Melchers et al. 1997). This tertiary structure element is highly conserved amongst members of the enteroviruses (Mirmomeni et al. 1997). The three sequences Coxsackie B3 virus, human echovirus 6, and human enterovirus 69 are chosen as a test set (Wang et al. 1999; Gardner et al. 2009). DotKnot predicts kissing hairpins in all of the three sequences and shows the best prediction for human echovirus 6 (MCC = 0.86). FlexStem and pknots predict no kissing hairpins and show a significantly lower MCC than DotKnot for the structure; however, pknots returns the best predictions for Coxsackie B3 virus and human enterovirus 69.

Initiation of negative-strand synthesis using a long-range kissing loop structure forming between coding and noncoding

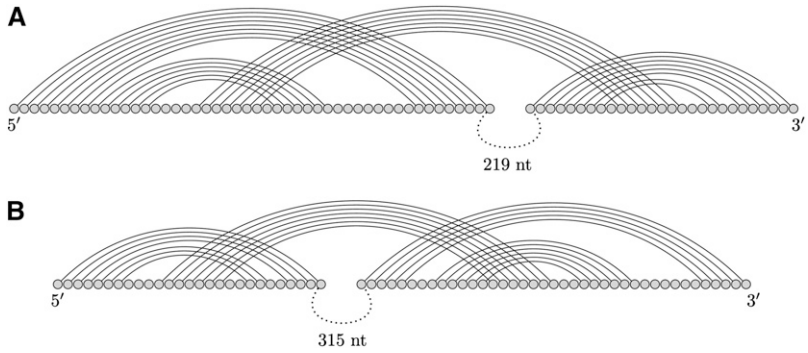

FIGURE 3. Long-range kissing hairpin interactions between coding and noncoding regions in the (*A*) hepatitis C virus (HCV) and (*B*) porcine reproductive and respiratory syndrome virus (PRRSV).

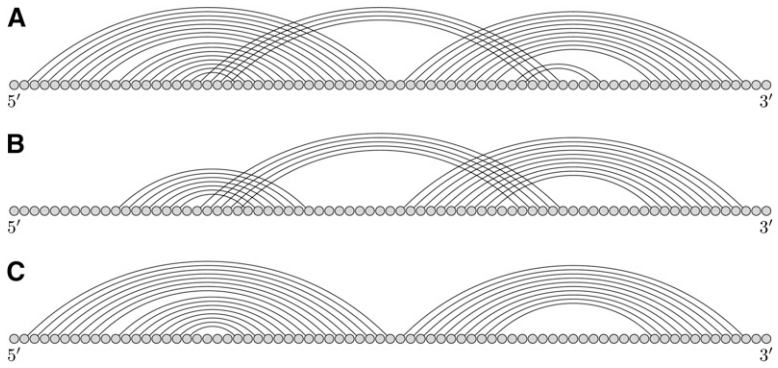

FIGURE 4. (*A*) The peach latent mosaic viroid (PLMVd) P8 pseudoknot is a kissing interaction between the P6 and P7 stems. The proposed structure for the complete 338-nt-long PLMVd viroid is described in Bussiere et al. (2000). (*B*) DotKnot predicts the minimal kissing interaction, however misses parts of stem *S*1 due to a bulge loop (MCC = 0.51). (*C*) FlexStem predicts the noncrossing stems, but no kissing interaction (MCC = 0.73).

regions has been proposed for other virus families such as *Flaviviridae* and *Nidovirales* (Fig. 3). A kissing interaction has been established in the hepatitis C virus (HCV) involving the NS5B coding region and 3′-UTR (Friebe et al. 2005). In the porcine reproductive and respiratory syndrome virus (PRRSV), a hairpin loop in the ORF7 region bonds with another hairpin loop in the 3′-NCR (Verheije et al. 2002). The kissing hairpins in HCV and PRRSV are long-range interactions that span 219 nt and 315 nt, respectively.

For the HCV long-range kissing hairpin, FlexStem and DotKnot predict nested structures with MCCs of 0.45 and 0.44, respectively. However, it is worth noting that the desired kissing hairpin structure is found as the best near-optimal pseudoknot with lowest free energy by DotKnot. For the PRRSV long-range interaction, DotKnot returns a short-range kissing hairpin structure involving one of the two hairpins and has the highest MCC of 0.33. Amongst the top three near-optimal pseudoknots with lowest free energy, DotKnot returns the reference long-range kissing hairpin involving both hairpins. To allow for a performance evaluation of pknots, results were also obtained for short sequences where the long loop between the coding and noncoding region is removed. For the short HCV sequence, RNAfold has the highest MCC as it correctly predicts most base pairs of the kissing hairpin stems S_1 and S_3. No kissing hairpin is predicted by DotKnot, pknots, and FlexStem. DotKnot predicts an H-type pseudoknot for the HCV sequence and has the lowest MCC amongst the kissing hairpin prediction algorithms. For the short PRRSV sequence, pknots and RNAfold correctly identify the noncrossing kissing hairpin stems S_1 and S_3, missing only one base pair. DotKnot is the only method which predicts a kissing hairpin for the PRRSV sequence; however, with lower MCC than the noncrossing predictions returned by pknots and RNAfold.

A compact kissing hairpin in the West Nile virus 3′-NCR is likely to be involved in viral replication (Shi et al. 1996). This structure element was suggested to be conserved in other flaviviruses, such as dengue virus (involving G-A base pairs) and yellow fever virus; however, further phylogenetic and structural investigation is needed. DotKnot returns a kissing hairpin structure and shows the highest MCC of 0.74 for all methods. A pseudoknotted MFE structure is returned by pknots and FlexStem predicts a noncrossing secondary structure with lower MCC than RNAfold.

Ribozymes

During genome replication, the hammerhead ribozyme in the satellite RNA of cereal yellow dwarf virus (satRPV) can alternatively form a compact kissing hairpin structure to inhibit self-cleavage (Song et al. 1999). DotKnot correctly identifies the kissing hairpin and has the highest MCC of 0.68. pknots has the second-highest MCC of 0.58 and predicts a H-type pseudoknot as the MFE structure. RNAfold shows the lowest MCC for all methods due to competing secondary structure elements.

Rastogi et al. (1996) report a kissing interaction in the *Neurospora* VS ribozyme. The pseudoknot is required for self-cleavage activity and forms in the presence of magnesium. In particular, the kissing interaction involves a hairpin loop within a multiloop structure. For the short *Neurospora* VS ribozyme sequence exactly harboring the kissing hairpin, DotKnot is the only method that predicts a kissing interaction (MCC = 0.21). A higher MCC of 0.69 is achieved by pknots for prediction of a noncrossing secondary structure. For the longer sequence, none of the algorithms predict a kissing hairpin. DotKnot returns a structure without any pseudoknots or kissing hairpins, yet it has the highest MCC of 0.78 for all methods. In particular, it perfectly predicts the multiloop structure using the MWIS calculations, whereas FlexStem and RNAfold have a lower MCC for their secondary structure predictions.

Viroids are 250–400-nt-long single-stranded RNAs that infect plants. Viroids are much smaller than viruses and contain no protective protein coat; therefore, the RNA secondary and tertiary structure of a viroid is critical for its life cycle and infection of the host cell. The group A peach latent mosaic viroid (PLMVd) and chrysanthemum chlorotic mottle viroid (CChMVd) can form hammerhead

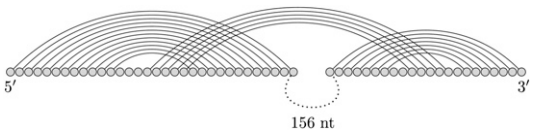

FIGURE 5. Human coronavirus 229E (HCoV-229E) long-range kissing hairpin.

ribozymes and have been proposed to fold into a branched secondary structure containing a kissing loop interaction (Bussiere et al. 2000; Gago et al. 2005). DotKnot is the only method which predicts kissing hairpins for both sequences and has the highest MCC for the CChMVd structure. For the PLMVd kissing hairpin, FlexStem achieves the highest MCC by correctly predicting the noncrossing stems S_1 and S_3. In contrast, DotKnot correctly identifies the kissing interaction S_2, but does not predict parts of stem S_1 because it is interrupted by a bulge loop. A comparison of the prediction results is shown in Figure 4.

Ribonuclease P (RNase P) is a ribozyme which cleaves precursor tRNA molecules. Archaeal and bacterial RNase P RNA structure is highly conserved. The *Escherichia coli* RNase P RNA contains a kissing hairpin structure (P6) nested within a pseudoknot (P4) (Westhof and Altman 1994; Brown 1999; Harris et al. 2001). Only the minimal sequence stretch covering the kissing hairpin P6 is used in the test set, as pseudoknots with internal crossing structures are not considered in the DotKnot algorithm. DotKnot predicts the kissing hairpin interaction with the highest MCC of 0.64. In contrast, RNAfold has a low MCC of -0.18 for the predicted MFE secondary structure. FlexStem predicts two pseudoknots for the sequence (MCC = 0.14).

Programmed -1 ribosomal frameshifting

Programmed -1 ribosomal frameshifting in some group 1 coronaviruses is facilitated by a long-range kissing hairpin (Fig. 5). The kissing hairpin has been confirmed in human coronavirus 229E (HCoV-229E) (Herold and Siddell 1993) and has been suggested for other group 1 members, such as TGEV, HCoV-NL63, and PEDV using phylogenetic analysis (Eleouet et al. 1995; Baranov et al. 2005; Plant et al. 2005). One should note that the TGEV frameshifting site has the potential to fold into a three-stemmed pseudoknot, as observed in the SARS coronavirus (Plant et al. 2005). For both the long and short HCoV-229E sequences, DotKnot returns the kissing hairpin with MCC of 1. All other methods do not find the kissing hairpin and have significantly lower predictive accuracy.

Structures without kissing hairpins

A test set for RNA structures without kissing hairpins is used to assess the prediction of false positive kissing hairpins by DotKnot and to evaluate the prediction results for a number of methods from the literature. The test set contains pseudoknot-free sequences (5S rRNA, tRNA, and miRNA) and sequences which contain one or more pseudoknots and are reported in the pseudoknot database Pseudobase (Table 3; van Batenburg et al. 2001). The sequence lengths range from 52 nt to 419 nt. Predictions are obtained from the practical dynamic programming algorithm pknotsRG (Reeder and Giegerich 2004), the heuristic methods HotKnots (Ren et al. 2005; Andronescu et al. 2010), and FlexStem (Chen et al. 2008), as well as the secondary structure prediction algorithm RNAfold (Hofacker et al. 1994). Results are shown in Table 4.

For our test set of pseudoknot-free structures, DotKnot predicts spurious H-type pseudoknots and kissing hairpins in three of the 12 sequences. For the nested structures, the dynamic programming methods pknotsRG and RNAfold show the highest average MCCs of 0.59 and 0.57, respectively. DotKnot and HotKnots both have an average MCC of 0.55, and FlexStem achieves 0.52.

For our test set of pseudoknotted structures, false positive kissing hairpins are predicted in one of the viral 3′-UTR structures (ORSV), in the *Escherichia coli* tmRNA, and the turnip yellow mosaic virus (TYMV) tRNA-like structure. No spurious hairpins are predicted for any of the remaining sequences in the test set. It should be noted that for the *Escherichia coli* tmRNA and TYMV sequences, prediction of false positive kissing hairpins leads to a higher MCC for the DotKnot predictions.

Our test set includes H-type pseudoknots as well as more complex pseudoknot foldings. For example, the ribozyme structures consist of double pseudoknots and nested pseudoknots feature in the cricket paralysis virus (CrPv) and Plautia stali intestine virus (PSIV) IRES elements. DotKnot is only able to fold H-type pseudoknots and kissing hairpins,

TABLE 3. RNA types and sequences used for pseudoknot prediction without kissing hairpins

RNA Type	Sequence ID	Reference
5S rRNA	5SEColi, 5SDMob, 5SHsap, 5STther	Cannone et al. (2002)
tRNA	DC0010, DC2720, DS0220, DT5090	Sprinzl et al. (1998)
miRNA	ath-mir159c, bta-mir29c, cfa-mir105b, sofmir156	Griffiths-Jones et al. (2006)
Ribozyme	drz-Agam-1-1, drz-Agam-2-1, drz-Tatr-1, HDV, HDVanti	Webb et al. (2009), Ferre-D'Amare et al. (1998), van Batenburg et al. (2001)
IRES	CrPV, PSIV	Hellen (2007), Pfingsten et al. (2006)
3′-UTR	NeRNV, TMV, ORSV	Koenig et al. (2005), van Belkum et al. (1985), Gultyaev et al. (1994)
tmRNA	EColi-tmRNA, LP-tmRNA	Williams (2000)
Viral tRNA-like	LRSVbeta, TYMV	Solovyev et al. (1996), Matsuda and Dreher (2004)
Telomerase	Human-telo, Tetra-telo	Theimer and Feigon (2006)
Riboswitch	SamII	Gilbert et al. (2008)
Retrotransposon	R2retro-Sc, R2retro-Spy	Kierzek et al. (2009)
Frameshifting	BWYV, MMTV, SARS-CoV, VMV	van Batenburg et al. (2001)

TABLE 4. Summary of prediction results using an extended version of DotKnot

Sequence			DotKnot				pknotsRG				HotKnots				FlexStem				RNAfold			
ID	nt	PK	*S*	PPV	MCC	*r*	*S*	PPV	MCC	*r*	*S*	PPV	MCC	*r*	*S*	PPV	MCC	*r*	*S*	PPV	MCC	*r*
5SEColi	120	0	100	100	**1**	0/0	100	100	**1**	0/0	100	100	**1**	0/0	92.1	79.5	0.7	0/1	100	100	**1**	0/0
5SDMob	133	0	95.6	87.8	0.81	0/0	100	88.2	**0.86**	0/0	100	88.2	**0.86**	0/0	95.6	84.3	0.76	0/0	100	88.2	**0.86**	0/0
5SHsap	119	0	29.7	32.4	−0.15	0/1	29.7	34.4	−0.14	0/0	29.7	34.4	−0.14	0/0	29.7	34.3	−0.11	0/0	29.7	37.9	**−0.05**	0/0
5STther	120	0	25.6	27	**−0.2**	0/0	20.5	23.5	−0.27	0/0	20.5	23.5	−0.27	0/0	25.6	22.7	−0.5	0/0	20.5	21.6	−0.31	0/0
DC0010	73	0	76.2	76.2	0.6	0/0	100	100	**1**	0/0	100	100	**1**	0/0	100	100	**1**	0/0	95.2	95.2	0.92	0/0
DC2720	71	0	35	31.8	**−0.09**	0/0	30	26.1	−0.21	0/0	30	26.1	−0.21	0/0	35	31.8	−0.13	0/0	35	29.2	−0.16	0/0
DS0220	87	0	96.3	86.7	0.83	0/1	48.1	46.4	0	0/0	48.1	44.8	−0.03	0/0	96.3	96.3	**0.93**	0/0	48.1	46.4	−0.02	0/0
DT5090	73	0	63.2	66.7	0.44	0/2	100	100	**1**	0/0	78.9	68.2	0.55	0/0	73.7	63.6	0.48	0/0	78.9	71.4	0.59	0/0
ath-mir159c	225	0	80.3	87.1	0.69	0/0	94.7	96	0.91	0/0	100	100	**1**	0/0	94.7	96	0.91	0/0	98.7	100	0.99	0/0
bta-mir29c	88	0	94.1	100	0.92	0/0	100	100	**1**	0/0	100	100	**1**	0/0	100	100	**1**	0/0	100	100	**1**	0/0
cfa-mir105b	80	0	86.7	96.3	0.8	0/0	100	100	**1**	0/0	93.3	96.6	0.88	0/0	63.3	73.1	0.34	0/0	100	100	**1**	0/0
sof-mir156	137	0	100	100	**1**	0/0	95.9	95.9	0.91	0/0	100	96.1	0.95	0/0	95.9	90.4	0.83	0/1	100	100	**1**	0/0
drz-Agam-1-1	82	1	75	95.5	**0.72**	1/1	82.1	82.1	0.64	1/1	57.1	66.7	0.3	0/0	89.3	86.2	**0.72**	1/1	57.1	66.7	0.3	0/0
drz-Agam-2-1	180	1	86.4	91.9	0.75	1/1	86.4	95	0.79	1/1	90.9	93.8	**0.82**	1/1	81.8	83.1	0.58	1/1	75.8	84.7	0.58	0/0
drz-Tatr-1	88	1	93.1	100	**0.93**	1/1	72.4	72.4	0.47	1/1	82.8	75	0.54	1/1	82.8	72.7	0.49	1/1	69	74.1	0.43	0/0
HDV	87	1	93.8	100	**0.93**	1/1	90.6	93.5	0.81	1/1	37.5	42.9	−0.13	0/0	84.4	79.4	0.57	1/1	37.5	42.9	−0.13	0/0
HDVanti	91	1	100	96.2	**0.97**	1/1	16	14.3	−0.46	0/0	16	14.3	−0.48	0/0	44	34.4	−0.11	1/1	16	14.3	−0.48	0/0
CrPV	190	2	63.6	63.6	**0.37**	2/2	52.7	52.7	0.21	0/0	45.5	51	0.15	0/0	34.5	31.7	−0.19	1/2	52.7	52.7	0.21	0/0
PSIV	194	2	70.7	69.5	0.46	0/0	72.4	68.9	0.46	0/0	72.4	71.2	**0.49**	0/0	39.7	45.1	0.03	0/1	72.4	66.7	0.42	0/0
NeRNV	198	5	70.4	64.4	**0.44**	5/7	48.1	44.8	0.09	1/2	5.6	5.2	−0.49	0/1	38.9	35	−0.04	1/1	31.5	31.5	−0.1	0/0
TMV	214	5	94.3	95.7	**0.9**	5/5	60	66.7	0.34	0/0	60	66.7	0.34	0/0	44.3	44.9	0.01	0/1	51.4	56.3	0.19	0/0
ORSV	419	11	71.3	70.3	**0.44**	9/10	48.5	50.4	0.07	5/5	41.2	45.9	0.01	0/0	39.7	39.7	−0.11	0/2	43.4	46.8	0.02	0/0
EColi-tmRNA	363	4	75	76.5	**0.6**	4/6	50	48.1	0.13	0/0	50	47.7	0.12	0/0	42.3	42.3	0.07	1/2	50	48.1	0.13	0/0
LP-tmRNA	406	4	56.1	54.5	**0.31**	4/8	30.8	28.7	−0.11	0/0	15.9	15.6	−0.27	0/2	34.6	36.6	0.04	0/0	38.3	34.5	−0.02	0/0
LRSVbeta	221	1	90.6	76.2	**0.73**	1/1	84.9	73.8	0.67	0/0	81.1	72.9	0.64	0/0	79.2	65.6	0.55	0/0	86.8	74.2	0.69	0/0
TYMV	110	2	46.7	42.4	**0.1**	1/3	46.7	41.2	0.07	1/1	33.3	30.3	−0.09	0/0	46.7	41.2	0.07	1/1	33.3	30.3	−0.09	0/0
Human-telo	210	1	76	67.9	**0.57**	1/1	54	42.9	0.17	1/1	70	55.6	0.38	0/0	34	22.7	0.23	0/1	64	48.5	0.27	0/0
Tetra-telo	159	1	73.7	70	**0.58**	1/1	65.8	56.8	0.4	0/0	60.5	53.5	0.34	0/0	42.1	35.6	0.06	0/0	71.1	61.4	0.48	0/0
SamII	52	1	78.6	91.7	0.77	1/1	78.6	91.7	0.77	1/1	42.9	54.5	0.22	0/0	92.9	92.9	**0.89**	1/1	42.9	54.5	0.22	0/0
R2retro-Sc	80	1	76.9	83.3	**0.62**	1/1	73.1	82.6	0.59	1/1	57.7	68.2	0.35	0/0	61.5	69.6	0.4	0/0	57.7	60	0.27	0/0
R2retro-Spy	80	1	76.9	80	0.58	1/1	65.4	63	0.31	1/1	80.8	77.8	**0.59**	1/1	80.8	70	0.45	1/1	73.1	82.6	**0.59**	0/0
BWYV	50	1	100	100	**1**	1/1	100	100	**1**	1/1	55.6	62.5	0.47	0/0	55.6	45.5	0.33	1/1	55.6	55.6	0.42	0/0
MMTV	49	1	100	100	**1**	1/1	100	100	**1**	1/1	0	0	−0.3	0/0	100	80	0.83	1/1	0	0	−0.3	0/0
SARS-CoV	82	1	92.3	100	**0.93**	1/1	92.3	92.3	0.85	1/1	73.1	65.5	0.36	0/0	69.2	64.3	0.34	1/1	73.1	65.5	0.36	0/0
VMV	68	1	100	82.4	**0.87**	1/1	50	41.2	0.18	0/0	50	41.2	0.18	0/0	100	60.9	0.66	1/1	50	41.2	0.18	0/0

PK corresponds to the number of pseudoknots in the sequence as reported in the literature. We use pknotsRG version 1.3, HotKnots version 2.0 with default parameters, FlexStem version 1.3, and the RNAfold web server (Gruber et al. 2008). The * symbol means that we were not able to run the algorithm to completion due to computational requirements. The ratio r = (number of correctly predicted pseudoknots) / (number of predicted pseudoknots) is also reported.

whereas HotKnots and FlexStem cover general pseudoknot interactions; yet, DotKnot shows the highest MCC for five of the seven complex pseudoknot structures. Furthermore, DotKnot has the highest MCC for the 3′-UTR, tmRNA, viral tRNA-like, telomerase, and frameshifting structure predictions in our test set. For all pseudoknotted structures, DotKnot shows the highest average MCC of 0.68. Lower average MCCs are achieved by pknotsRG (0.41), FlexStem (0.28), RNAfold (0.2), and HotKnots (0.2).

DISCUSSION

The general pseudoknot prediction problem is intractable due to the vast structure search space. Dynamic programming methods for MFE structure prediction including pseudoknots need to achieve a reasonable balance between the complexity of allowed pseudoknots and computational requirements. A number of heuristic methods include a broad class of pseudoknots which may cover multiple crossing stems or nested pseudoknots. However, it might not be desirable to include arbitrarily complex pseudoknots in an RNA prediction algorithm due to the lack of thermodynamic parameters and knowledge about steric requirements. On the other hand, if a biologically relevant pseudoknot class such as kissing hairpins is known, it can easily be included as a predefined structure type in pseudoknot search programs such as DotKnot.

DotKnot outperforms pknots, FlexStem, and RNAfold for our test set of kissing hairpins. Except for three sequences, DotKnot returns a kissing hairpin structure as the result and has the highest average MCC of 0.56 for the test set. In contrast, the dynamic programming method pknots

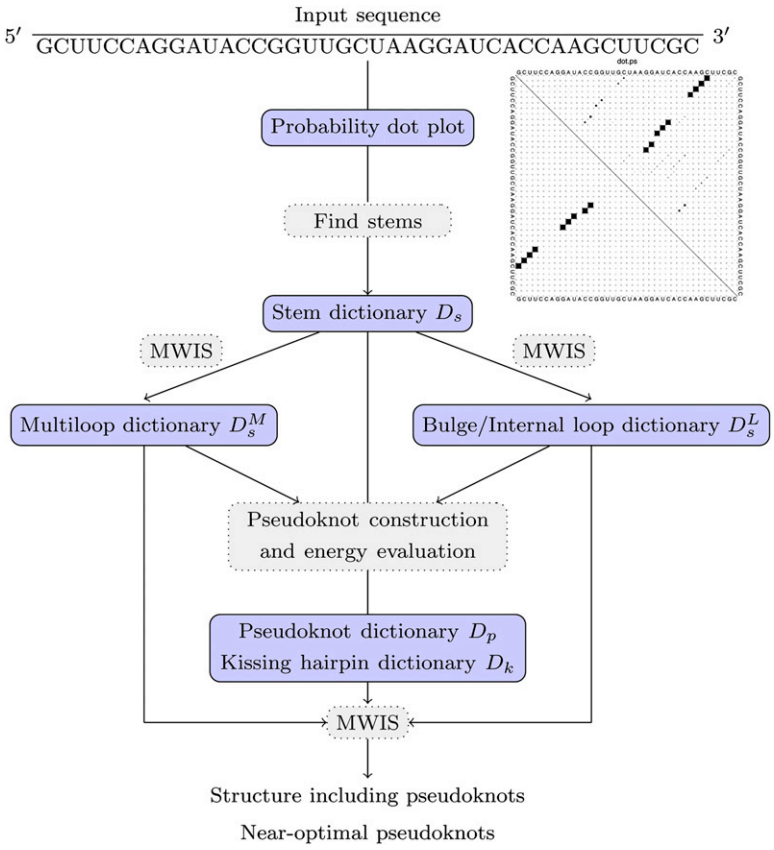

FIGURE 6. Extending DotKnot for the prediction of kissing hairpins. The structure returned as an output may contain both H-type pseudoknots and kissing-hairpin–type pseudoknots. A number of near-optimal H-type pseudoknots and kissing hairpins may also be reported. MWIS stands for maximum weight independent set calculation.

only detects a kissing hairpin for one of the sequences and has an average MCC of 0.46.

FlexStem predicts no kissing hairpins for any of the sequences and has the lowest average MCC of 0.18. FlexStem uses the same energy penalties as pknots for overlapping pseudoknots such as kissing hairpins. The energy parameters for pseudoknots estimated by pknots and adopted by FlexStem might be the reason for the poor kissing hairpin predictions. RNAfold has an average MCC of 0.31 for our test sequences. For three of the SRP RNA sequences in our test set, RNAfold does not predict any true positive base pairs for the noncrossing stems. This shows that a hierarchical folding approach where kissing interactions are searched for after obtaining a MFE structure might not always be successful. Thus, specialized pseudoknot folding methods which adopt specific energy parameters are needed for RNA structure prediction including pseudoknots.

DotKnot detects 23 out of 26 kissing hairpins for our test set. This might lead to the conclusion that introducing false positive kissing hairpins is inevitable due to the large number of candidates. However, we find that for our negative control set, DotKnot only predicts false positive kissing hairpins in three of the sequences and outperforms the competing algorithms for the set of pseudoknotted structures. It must be noted that for the pseudoknot-free sequences in our test set, the free energy minimization methods pknotsRG and RNAfold give the best results. DotKnot is a heuristic pseudoknot prediction method which does not aim to compete with free-energy minimization algorithms for secondary structure prediction. DotKnot does not guarantee to find the MFE structure given a sequence; however, we find that it reliably predicts stable secondary structure elements which may compete with pseudoknot formation. DotKnot shows the best average MCC for our test set of pseudoknotted sequences and is a practical pseudoknot prediction tool for finding pseudoknots in longer sequences. A comparison of running times is given in Supplemental Table 1.

We think that the underlying energy parameters for H-type pseudoknots (Cao and Chen 2006, 2009) and kissing hairpin parameters chosen by us are the main reasons for the high predictive accuracy. An RNA folding algorithm can only be as accurate as the quality of underlying energy parameters allows; therefore, laboratory investigations into RNA energy parameters such as long-range kissing hairpin interactions are highly desirable. To overcome the approximate nature of RNA energy parameters, one can gain confidence in predictions by using comparative information. Pseudoknots are known to be highly conserved and the DotKnot method will be extended in the future to take into account multiple alignment information using the probability dot plots.

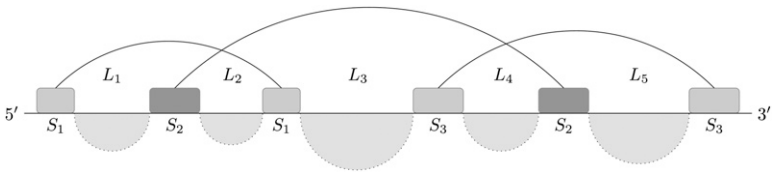

FIGURE 7. Kissing hairpin prediction class where recursive secondary structure elements are allowed in each of the five loops L_1, L_2, L_3, L_4, and L_5.

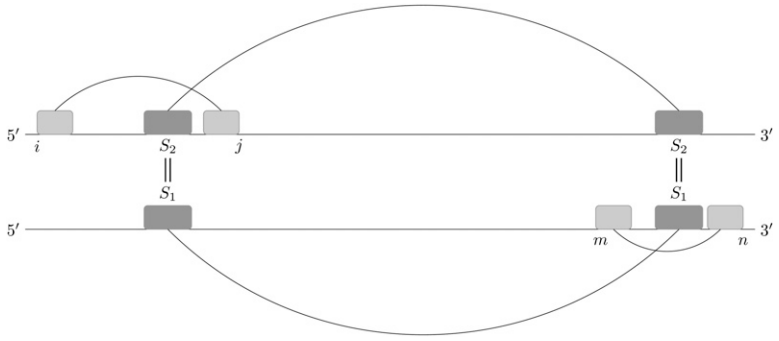

FIGURE 8. A kissing hairpin can be decomposed into two core H-type pseudoknots, where the second stem S_2 of the first pseudoknot equals the first stem $S_1$of the second pseudoknot. Note that for a kissing hairpin, $j < m$ has to hold. Otherwise, a triple helix interaction is formed.

MATERIALS AND METHODS

In this section, we describe the basics of the DotKnot algorithm and its extension to the prediction of a global structure, including H-type pseudoknots and intramolecular kissing hairpins (Fig. 6).

The DotKnot algorithm

The basis of the DotKnot method is the secondary structure probability dot plot calculated by RNAfold (Hofacker et al. 1994). From the dot plot, a set of promising stems is extracted using the base pair probabilities and stored in the dictionary D_s. Note that by setting a low-probability threshold, potential pseudoknot stems can be discovered. Using the stem dictionary D_s, noncrossing secondary structure elements with low free energy are assembled using maximum weight independent set (MWIS) calculations. Stems interrupted by bulges or internal loops are stored in dictionary D_s^L and multiloops are stored in dictionary D_s^M. Stems and secondary structure elements are then used to construct recursive H-type pseudoknots. H-type pseudoknot energies are evaluated with the aid of advanced energy models (Cao and Chen 2006, 2009). The presence of the H-type pseudoknots is verified using a MWIS calculation on the set of all possible structure elements. Outer stems may include nested pseudoknots. In the first version of DotKnot, only pseudoknots were returned as a result. In the current version, the user can choose to additionally see the global structure derived by the final MWIS calculation. For algorithmic details of the first version of DotKnot for detecting H-type pseudoknots, see Sperschneider and Datta (2010).

Our first extension presented here is the ability to predict intramolecular kissing hairpins. The type of recursive kissing hairpin structure allowed in the extended DotKnot method is shown in Figure 7. The crossing of three stems results in five loops which can contain recursive secondary structure elements. Note that we restrict a kissing hairpin structure to be shorter than 400 nt to improve runtime. Furthermore, MFE folding is known to become inaccurate for longer sequences due to the underlying approximate energy parameters (Eddy 2004; Reeder et al. 2006). Long-range interactions are especially hard to predict using MFE folding, as tertiary interactions and forces are likely to further stabilize the structure.

The second extension is that in addition to the best global folding, the best local H-type pseudoknots and kissing hairpins in terms of two criteria are returned. This can help to identify promising pseudoknot foldings and may compensate for the limitations of the energy parameters. DotKnot returns the best pseudoknots in terms of estimated free energy to length ratio (Reeder and Giegerich 2004). This helps to identify local pseudoknots and will favor pseudoknots with compact structure and low free energy. Additionally, DotKnot returns pseudoknots with lowest estimated free energy, regardless of their lengths. For each criterion, a user-set number of pseudoknots are returned.

Kissing hairpin prediction

Assembling kissing hairpin candidates

An intramolecular kissing hairpin is a planar pseudoknot that can be decomposed into two core H-type pseudoknots (Fig. 8). The main idea of the extended DotKnot algorithm is to create a list of H-type pseudoknots, which are subsequently combined into kissing hairpin candidates. Core H-type pseudoknots are assembled from stems which are extracted from the base pair probability dot plot calculated by RNAfold. Each stem $s_i \in D_s$ has two energy weights w_{stack} (s_i) (simple stacking free energy) and $w(s_i)$ (free energy). Only stems s_i are used for kissing hairpin construction where w_{stack} $(s_i) < -5.0$ kcal/mol and $w(s_i) < 2.0$ kcal/mol. During pseudoknot construction, a certain base pair overlap is allowed as crossing stems with a fixed length are combined (see Supplemental Material).

The H-type pseudoknots are stored in a specific manner in order to assemble kissing hairpins efficiently. There are two dictionaries, $D_p^{S_1}$ and $D_p^{S_2}$. Dictionary $D_p^{S_1}$ has a stem as a key and as values the list of corresponding pseudoknots which contain this stem as a first pseudoknot stem S_1. Dictionary $D_p^{S_2}$ has a stem as a key and as values the set of corresponding pseudoknots which contain this stem as a second pseudoknot stem S_2. For each stem in dictionary $D_p^{S_2}$, a key existence test is performed in dictionary $D_p^{S_1}$. If the same stem is found in both dictionaries, the values for the stem entry in $D_p^{S_2}$ are combined with the values for the stem entry in $D_p^{S_1}$ to form a kissing hairpin (Fig. 8). An indication of the stem probability in the secondary structure folding ensemble is given by the confidence indicator which is defined as the average probability of participating base pairs in a stem. We demand that the three kissing hairpin stems have a confidence sum of $>1 \times E^{-3}$. Kissing hairpins consist partly of noncrossing stable secondary structure elements and base pairs below this threshold are unlikely

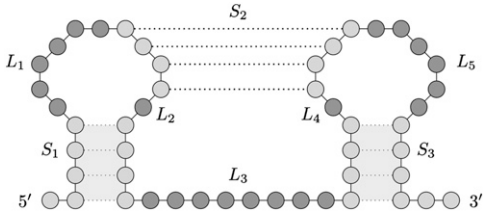

FIGURE 9. Energy estimation for a kissing hairpin: the three stems S_1, S_2, and S_3 contribute stabilizing stacking energies and each unpaired nucleotide in loops L_1, L_2, L_3, L_4, and L_5 is penalized.

to participate in secondary structure formation (Hofacker and Stadler 1999).

Recursive structure formation in the loops

A kissing hairpin candidate structure has three stems S_1, S_2, and S_3, and five loops L_1, L_2, L_3, L_4, and L_5 (Fig. 7). Given the set of kissing hairpin structures, the five loops are investigated for internal secondary structure elements. Note that internal pseudoknots in the loops are not allowed due to the lack of knowledge about three-dimensional folding. Only secondary structure elements from dictionaries D_s, D_s^L, and D_s^M can form in each of the five loops in a consecutive fashion. Recursive secondary structure elements are found using a MWIS calculation as described in Sperschneider and Datta (2010). Note that for three-dimensional folding reasons in a kissing hairpin structure the following is assumed: There must be at least one nucleotide in loops L_1 or L_2 (L_4 or L_5) that is left unpaired.

Energy evaluation for kissing hairpins

The critical point for kissing hairpin prediction is the underlying energy model. As it is a tertiary structure element, many different types of forces apart from canonical base-pairing are likely to play a role, for example noncanonical base-pairing, base triples, backbone interactions, or ion concentrations (Batey et al. 1999). No experimentally measured energy parameters for intramolecular kissing hairpins have been established to date and thus heuristic energy estimation has to be used. Here, the free energy for each kissing hairpin $k_1,\ldots,k_n$ in dictionary D_k is approximated by adding the stacking energies, including dangling ends for the three stems S_1, S_2, and S_3, plus a length-dependent value for the loop entropies (Fig. 9). An extended version of the parameterized pseudoknot energy model (Rivas and Eddy 1999; Dirks and Pierce 2003; Reeder and Giegerich 2004) is used:

$$\Delta G(k_i) = w_{\text{stack}}(S_1) + w_{\text{stack}}(S_2) + w_{\text{stack}}(S_3) + \alpha + \beta \times (l_1 + l_2 + l_4 + l_5) + \gamma \times l_3,$$

where l_i is the number of unpaired nucleotides in the loop L_i ($i = 1,\ldots,5$). In this model, loop L_3 attracts a penalty γ for each unpaired nucleotide, whereas the four other loops are penalized using the value β. Without the kissing interaction, loop L_3 would not contribute entropic terms according to the Turner model. A kissing interaction between two non-neighboring hairpin loops with a long loop L_3 is thus not unlikely and, therefore, γ is set to 0.0 kcal/mol. Here, the initiation penalty for forming a kissing hairpin is set to α = 9.0 kcal/mol and β to 0.5 kcal/mol. For kissing hairpins with recursive secondary structure elements in the five loops, the stabilizing free-energy weights of the internal elements are added to the overall free energy. The loop entropy for the recursive kissing hairpin is then re-estimated using the remaining number of unpaired nucleotides in the five loops.

Kissing hairpins with negative free energy are stored in the kissing hairpin candidate dictionary D_k. Despite the low number of stem candidates, the number of kissing hairpin candidates is relatively high due to the large structure space. The same length-normalized filtering step as in the first version of DotKnot is used. As an additional measure for kissing hairpin stability, the normalized kissing hairpin free energy must fulfill $\Delta G(k_i)/l_i \leq \varepsilon$ where l_i denotes the length of the kissing hairpin candidate structure k_i. Kissing hairpins are assumed to contribute to the overall stability of an RNA structure. This filtering step helps to eliminate unlikely kissing hairpins with high free energy and improves runtime of the method.

Verification of kissing hairpins in the sequence

The presence of kissing hairpin candidates in an RNA sequence is verified using the MWIS calculation of the DotKnot algorithm (Sperschneider and Datta 2010). The structure elements from the three secondary structure dictionaries D_s, D_s^L, and D_s^M, as well as the recursive H-type pseudoknot candidates stored in D_p and the kissing hairpin candidates stored in D_k participate in the MWIS calculation using free-energy weights. Outer stems are allowed to contain nested structure elements, including pseudoknots and kissing hairpins. The output consists of the (possibly empty) set of detected crossing structures such as H-type pseudoknots and kissing hairpins (Fig. 6). Additionally, the global structure derived by the MWIS calculation, including secondary structure elements and near-optimal pseudoknots, if desired, is displayed.

SUPPLEMENTAL MATERIAL

Supplemental material can be found at http://www.rnajournal.org.

ACKNOWLEDGMENTS

This work is supported by funding from The University of Western Australia.

Received July 29, 2010; accepted October 17, 2010.

REFERENCES

Akutsu T. 2000. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Appl Math* **104:** 45–62.

Andronescu MS, Pop C, Condon AE. 2010. Improved free energy parameters for RNA pseudoknotted secondary structure prediction. *RNA* **16:** 26–42.

Baranov PV, Henderson CM, Anderson CB, Gesteland RF, Atkins JF, Howard MT. 2005. Programmed ribosomal frameshifting in decoding the SARS-CoV genome. *Virology* **332:** 498–510.

Batey RT, Rambo RP, Doudna JA. 1999. Tertiary motifs in RNA structure and folding. *Angew Chem Int Ed* **38:** 2326–2343.

Brierley I, Pennell S, Gilbert RJC. 2007. Viral RNA pseudoknots: versatile motifs in gene expression and replication. *Nat Rev Microbiol* **5:** 598–610.

Brierley I, Gilbert RJC, Pennell S. 2008. RNA pseudoknots and the regulation of protein synthesis. *Biochem Soc Trans* **36:** 684–689.

Brown JW. 1999. The Ribonuclease P Database. *Nucleic Acids Res* **27:** 314. doi: 10.1093/nar/27.1.314.

Brunel C, Marquet R, Romby P, Ehresmann C. 2002. RNA loop-loop interactions as dynamic functional motifs. *Biochimie* **84:** 925–944.

Bussiere F, Ouellet J, Cote F, Levesque D, Perreault JP. 2000. Mapping in solution shows the peach latent mosaic viroid to possess a new pseudoknot in a complex, branched secondary structure. *J Virol* **74:** 2647–2654.

Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, Feng B, Lin N, Madabusi LV, Müller KM, et al. 2002. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal,

intron, and other RNAs. *BMC Bioinformatics* **3:** 2. doi: 10.1186/1471-2105-3-2.
Cao S, Chen SJ. 2006. Predicting RNA pseudoknot folding thermodynamics. *Nucleic Acids Res* **34:** 2634–2652.
Cao S, Chen SJ. 2009. Predicting structures and stabilities for H-type pseudoknots with interhelix loops. *RNA* **15:** 696–706.
Chen X, He S, Bu D, Zhang F, Wang Z, Chen R, Gao W. 2008. FlexStem: Improving predictions of RNA secondary structures with pseudoknots by reducing the search space. *Bioinformatics* **24:** 1994–2001.
Chen HL, Condon AE, Jabbari H. 2009. An O(n(5)) algorithm for MFE prediction of kissing hairpins and 4-chains in nucleic acids. *J Comput Biol* **16:** 803–815.
Dirks RM, Pierce NA. 2003. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J Comput Chem* **24:** 1664–1677.
Eddy SR. 2004. How do RNA folding algorithms work? *Nat Biotechnol* **22:** 1457–1458.
Eleouet JF, Rasschaert D, Lambert P, Levy L, Vende P, Laude H. 1995. Complete sequence (20 kilobases) of the polyprotein-encoding gene 1 of transmissible gastroenteritis virus. *Virology* **206:** 817–822.
Ferre-D'Amare AR, Zhou KH, Doudna JA. 1998. Crystal structure of a hepatitis delta virus ribozyme. *Nature* **395:** 567–574.
Friebe P, Boudet J, Simorre JP, Bartenschlager R. 2005. Kissing-loop interaction in the 3′ end of the hepatitis C virus genome essential for RNA replication. *J Virol* **79:** 380–392.
Gago S, De la Peña M, Flores R. 2005. A kissing-loop interaction in a hammerhead viroid RNA critical for its in vitro folding and in vivo viability. *RNA* **11:** 1073–1083.
Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, et al. 2009. Rfam: updates to the RNA families database. *Nucleic Acids Res* **37:** D136–D140.doi: 10.1093/nar/gkn766.
Giedroc DP, Cornish PV. 2009. Frameshifting RNA pseudoknots: structure and mechanism. *Virus Res* **139:** 193–208.
Gilbert SD, Rambo RP, Van Tyne D, Batey RT. 2008. Structure of the SAM-II riboswitch bound to S-adenosylmethionine. *Nat Struct Mol Biol* **15:** 177–182.
Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* **34:** D140–D144. doi: 10.1093/nar/gkj112.
Gruber AR, Lorenz R, Bernhart SH, Neubck R, Hofacker IL. 2008. The Vienna RNA websuite. *Nucleic Acids Res* **36:** W70–W74. doi: 10,1093/nar/gkn188.l
Gultyaev AP, van Batenburg E, Pleij CW. 1994. Similarities between the secondary structure of satellite tobacco mosaic virus and tobamovirus RNAs. *J Gen Virol* **75:** 2851–2856.
Gultyaev AP, van Batenburg E, Pleij CW. 1995. The computer simulation of RNA folding pathways using a genetic algorithm. *J Mol Biol* **250:** 37–51.
Harris JK, Haas ES, Williams D, Frank DN, Brown JW. 2001. New insight into RNase P RNA structure from comparative analysis of the archaeal RNA. *RNA* **7:** 220–232.
Hellen CU. 2007. Bypassing translation initiation. *Structure* **15:** 4–6.
Herold J, Siddell SG. 1993. An 'elaborated' pseudoknot is required for high frequency frameshifting during translation of HCV 229E polymerase mRNA. *Nucleic Acids Res* **21:** 5838–5842.
Hofacker IL, Stadler PF. 1999. Automatic detection of conserved base pairing patterns in RNA virus genomes. *Comput Chem* **23:** 401–414.
Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P. 1994. Fast folding and comparison of RNA secondary structures. *Monatsh Chem* **125:** 167–188.
Jabbari H, Condon AE, Zhao S. 2008. Novel and efficient RNA secondary structure prediction using hierarchical folding. *J Comput Biol* **15:** 139–163.
Keenan RJ, Freymann DM, Stroud RM, Walter P. 2001. The signal recognition particle. *Annu Rev Biochem* **70:** 755–775.
Kierzek E, Christensen SM, Eickbush TH, Kierzek R, Turner DH, Moss WN. 2009. Secondary structures for 5′ regions of R2 retrotransposon RNAs reveal a novel conserved pseudoknot and regions that evolve under different constraints. *J Mol Biol* **390:** 428–442.
Koenig R, Barends S, Gultyaev AP, Lesemann DE, Vetten HJ, Loss S, Pleij CWA. 2005. Nemesia ring necrosis virus: a new tymovirus with a genomic RNA having a histidylatable tobamovirus-like 3′ end. *J Gen Virol* **86:** 1827–1833.
Larsen N, Zwieb C. 1991. SRP-RNA sequence alignment and secondary structure. *Nucleic Acids Res* **19:** 209–215.
Lyngsø RB, Pedersen CN. 2000a. RNA pseudoknot prediction in energy-based models. *J Comput Biol* **7:** 409–427.
Lyngsø RB, Pedersen CN. 2000b. Pseudoknots in RNA secondary structures. In *Proceedings of the 4th Annual International Conference on Computational Molecular Biology*, pp. 201–209.
Lyngsø RB, Zuker M, Pedersen CNS. 1999. Fast evaluation of internal loops in RNA secondary structure prediction. *Bioinformatics* **15:** 440–445.
Matsuda D, Dreher TW. 2004. The tRNA-like structure of Turnip yellow mosaic virus RNA is a 3′-translational enhancer. *Virology* **321:** 36–46.
McCaskill JS. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29:** 1105–1119.
Melchers WJ, Hoenderop JG, Bruins Slot HJ, Pleij CW, Pilipenko EV, Agol VI, Galama JM. 1997. Kissing of the two predominant hairpin loops in the coxsackie B virus 3′ untranslated region is the essential structural feature of the origin of replication required for negative-strand RNA synthesis. *J Virol* **71:** 686–696.
Mirmomeni MH, Hughes PJ, Stanway G. 1997. An RNA tertiary structure in the 3′ untranslated region of enteroviruses is necessary for efficient replication. *J Virol* **71:** 2363–2370.
Pfingsten JS, Costantino DA, Kieft JS. 2006. Structural basis for ribosome recruitment and manipulation by a viral IRES RNA. *Science* **314:** 1450–1454.
Plant EP, Prez-Alvarado GC, Jacobs JL, Mukhopadhyay B, Hennig M, Dinman JD. 2005. A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal. *PLoS Biol* **3:** e172. doi: 10.1371/journal.pbio.0030172.
Pleij CWA, Rietveld K, Bosch L. 1985. A new principle of RNA folding based on pseudoknotting. *Nucleic Acids Res* **13:** 1717–1731.
Rastogi T, Beattie TL, Olive JE, Collins RA. 1996. A long-range pseudoknot is required for activity of the Neurospora VS ribozyme. *EMBO J* **15:** 2820–2825.
Reeder J, Giegerich R. 2004. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics* **5:** 104. doi: 10.1186/1471-2105-5-104.
Reeder J, Höchsmann M, Rehmsmeier M, Voss B, Giegerich R. 2006. Beyond Mfold: recent advances in RNA bioinformatics. *J Biotechnol* **124:** 41–55.
Ren J, Rastegari B, Condon AE, Hoos HH. 2005. HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. *RNA* **11:** 1494–1504.
Rietveld K, Van Poelgeest R, Pleij CW, Van Boom JH, Bosch L. 1982. The tRNA-like structure at the 3′ terminus of turnip yellow mosaic virus RNA. Differences and similarities with canonical tRNA. *Nucleic Acids Res* **10:** 1929–1946.
Rietveld K, Pleij CW, Bosch L. 1983. Three-dimensional models of the tRNA-like 3′ termini of some plant viral RNAs. *EMBO J* **2:** 1079–1085.
Rivas E, Eddy SR. 1999. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J Mol Biol* **285:** 2053–2068.
Ruan J, Stormo GD, Zhang W. 2004. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics* **20:** 58–66.
Shi PY, Brinton MA, Veal JM, Zhong YY, Wilson WD. 1996. Evidence for the existence of a pseudoknot structure at the 3 terminus of the flavivirus genomic RNA. *Biochemistry* **35:** 4222–4230.

Solovyev AG, Savenkov EI, Agranovsky AA, Morozov SY. 1996. Comparisons of the genomic *cis*-elements and coding regions in RNA beta components of the hordeiviruses barley stripe mosaic virus, lychnis ringspot virus, and poa semilatent virus. *Virology* **219:** 9–18.

Song SI, Silver SL, Aulik MA, Rasochova L, Mohan BR, Miller WA. 1999. Satellite cereal yellow dwarf virus-RPV (satRPV) RNA requires a douXble hammerhead for self-cleavage and an alternative structure for replication. *J Mol Biol* **293:** 781–793.

Sperschneider J, Datta A. 2010. DotKnot: pseudoknot prediction using the probability dot plot under a refined energy model. *Nucleic Acids Res* **38:** e103. doi: 10.1093/nar/gkg021.

Sprinzl M, Horn C, Brown M, Ioudovitch A, Steinberg S. 1998. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res* **26:** 148–153.

Staple DW, Butcher SE. 2005. Pseudoknots: RNA structures with diverse functions. *PLoS Biol* **3:** 956–959.

Theimer CA, Feigon J. 2006. Structure and function of telomerase RNA. *Curr Opin Struct Biol* **16:** 307–318.

Uemura Y, Hasegawa A, Kobayashi S, Yokomori T. 1999. Tree adjoining grammars for RNA structure prediction. *Theor Comput Sci* **210:** 277–303.

van Batenburg FH, Gultyaev AP, Pleij CW. 2001. PseudoBase: structural information on RNA pseudoknots. *Nucleic Acids Res* **29:** 194–195.

van Belkum A, Abrahams JP, Pleij CW, Bosch L. 1985. Five pseudoknots are present at the 204 nucleotides long 3 noncoding region of tobacco mosaic virus RNA. *Nucleic Acids Res* **13:** 7673–7686.

Verheije MH, Olsthoorn RC, Kroese MV, Rottier PJ, Meulenberg JJ. 2002. Kissing interaction between 3′ noncoding and coding sequences is essential for porcine arterivirus RNA replication. *J Virol* **76:** 1521–1526.

Wang J, Bakkers JM, Galama JM, Bruins Slot HJ, Pilipenko EV, Agol VI, Melchers WJ. 1999. Structural requirements of the higher order RNA kissing element in the enteroviral 3′UTR. *Nucleic Acids Res* **27:** 485–490.

Webb CH, Riccitelli NJ, Ruminski DJ, Luptk A. 2009. Widespread occurrence of self-cleaving ribozymes. *Science* **326:** 953.

Westhof E, Altman S. 1994. Three-dimensional working model of M1 RNA, the catalytic RNA subunit of ribonuclease P from Escherichia coli. *Proc Natl Acad Sci* **91:** 5133–5137.

Williams KP. 2000. The tmRNA website. *Nucleic Acids Res* **28:** 168. doi: 10.1126/science.1178084.

Zuker M, Stiegler P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res* **9:** 133–148.

Zwieb C, Müller F. 1997. Three-dimensional comparative modeling of RNA. *Nucleic Acids Symp Ser* **36:** 69–71.

Zwieb C, Samuelsson T. 2000. SRPDB (signal recognition particle database). *Nucleic Acids Res* **28:** 171–172.