KnotSeeker: Heuristic pseudoknot detection in long RNA sequences

JANA SPERSCHNEIDER[1,2] and AMITAVA DATTA[1]
[1]School of Computer Science and Software Engineering, University of Western Australia, Perth, WA 6009, Australia
[2]Institut für Informatik, Albert-Ludwigs-Universität Freiburg, 79085 Freiburg, Germany

ABSTRACT

Pseudoknots are folded structures in RNA molecules that perform essential functions as part of cellular transcription machinery and regulatory processes. The prediction of these structures in RNA molecules has important implications in antiviral drug design. It has been shown that the prediction of pseudoknots is an NP-complete problem. Practical structure prediction algorithms based on free energy minimization employ a restricted problem class and dynamic programming. However, these algorithms are computationally very expensive, and their accuracy deteriorates if the input sequence containing the pseudoknot is too long. Heuristic methods can be more efficient, but do not guarantee an optimal solution in regards to the minimum free energy model. We present KnotSeeker, a new heuristic algorithm for the detection of pseudoknots in RNA sequences as a preliminary step for structure prediction. Our method uses a hybrid sequence matching and free energy minimization approach to perform a screening of the primary sequence. We select short sequence fragments as possible candidates that may contain pseudoknots and verify them by using an existing dynamic programming algorithm and a minimum weight independent set calculation. KnotSeeker is significantly more accurate in detecting pseudoknots compared to other common methods as reported in the literature. It is very efficient and therefore a practical tool, especially for long sequences. The algorithm has been implemented in Python and it also uses C/C++ code from several other known techniques. The code is available from http://www.csse.uwa.edu.au/~datta/pseudoknot.

Keywords: RNA pseudoknots; minimum free energy; dynamic programming; heuristic algorithms; minimum weight independent set; RNA structure prediction

INTRODUCTION

A central dogma in biology states that sequence determines structure determines function. This has been successfully applied to protein tertiary structure prediction. Over the past decades, the protein folding problem has attracted worldwide attention from many research groups and is seen as the holy grail of biochemistry. However, proteins are not the only important catalytically active macromolecules. It is clear that RNA can no longer be seen solely as a carrier of genetic information from DNA to proteins. RNA easily keeps up with the countless functions and structures proteins exhibit, adopts diverse three-dimensional folds, and can act like a catalyst. It is an extremely versatile molecule and facilitates various functions, including translational regulation, intron splicing, gene expression, and cell regulation. Novel noncoding RNAs are discovered continuously and the exciting RNA world is far from being fully explored.

Recent studies on RNA emphasize the fact that pseudoknots are a prevalent structural part, occurring in most classes of RNA (e.g., mRNA, tmRNA, rRNA, ribozymes, aptamers) (Staple and Butcher 2005). Pseudoknots are functionally diverse and can induce viral ribosomal frameshift or readthrough, be part of the catalytic core of ribozymes, or promote telomerase activity (Brierley et al. 2007). Especially retroviruses (e.g., HIV), coronaviruses (e.g., SARS), and most plant viruses use pseudoknots for proliferation and replication (Baril et al. 2003; Thiel et al. 2003). This draws attention to the high relevance of pseudoknots in antiviral drug design.

RNA secondary structure prediction methods by free energy minimization require $O(n^3)$ time and $O(n^2)$ space using dynamic programming (Zuker and Stiegler 1981).

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; or Amitava Datta, School of Computer Science and Software Engineering, University of Western Australia, Perth, WA 6009, Australia; e-mail: datta@csse.uwa.edu.au.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.968808.

However, including a tertiary structure element like the pseudoknot dampens the optimism of solving the RNA structure prediction problem. The general pseudoknot structure prediction problem is NP-complete (Lyngso and Pedersen 2000). Practical dynamic programming algorithms run in $O(n^6)$, $O(n^5)$, or $O(n^4)$ time for a restricted class of pseudoknots (Rivas and Eddy 1999; Akutsu 2000; Dirks and Pierce 2003; Reeder and Giegerich 2004). Dynamic programming methods for pseudoknot prediction guarantee an optimal solution in regards to the minimum free energy (MFE) model, yet suffer from two major drawbacks: a high running time even for a restricted class of pseudoknots and decreasing accuracy for long sequences due to sparse knowledge about pseudoknot thermodynamics. Nevertheless, if presented with a sequence fragment exactly harboring a pseudoknot, dynamic programming methods are able to fold it into the correct structure with high base-pair accuracy (Huang et al. 2005). Detection of true positive pseudoknots as a first step in RNA structure prediction can greatly improve the overall performance. The route followed in this article is to perform efficient pseudoknot detection preliminary to structure prediction. The advantage is clear: if we can find pseudoknots with high accuracy, the remaining sequence can be folded in $O(n^3)$ time according to the MFE model.

Apart from dynamic programming, several other techniques exist for RNA structure prediction including pseudoknots. Early methods comprise Monte Carlo simulations (Abrahams et al. 1990), genetic algorithms (Gultyaev et al. 1995; van Batenburg et al. 1995), stochastic context-free grammars (Brown and Wilson 1996; Cai et al. 2003), and maximum weighted matching (MWM) based on graph theory (Tabaska et al. 1999). Elaborated ab initio folding simulations are performed in KineFold (Xayaphoummine et al. 2003). Recent heuristic procedures include iterated loop matching (ILM) (Ruan et al. 2004) and HotKnots (Ren et al. 2005). ILM is a hybrid method employing dynamic programming and comparative information, which iteratively chooses the highest scoring helical region and adds it to the predicted structure. HotKnots expands this idea by considering several alternative secondary structures and returning a fixed number of suboptimal folding scenarios. HPknotter is a detection tool for pseudoknots based on structural matching and dynamic programming kernels (Huang et al. 2005). PLMM_DPSS was recently designed for predicting a limited class of pseudoknots with very high sensitivity (Huang and Ali 2007). All of these heuristic approaches do not guarantee returning an optimal solution with regard to the MFE model, however run in reasonable CPU time.

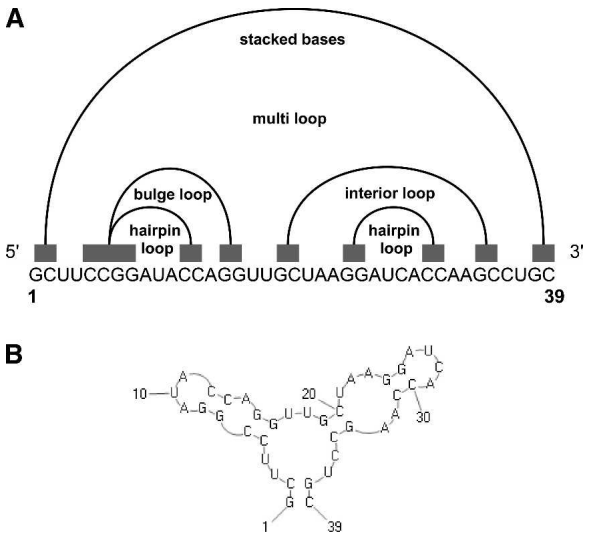


FIGURE 1. (*A*) Representation of RNA secondary structure elements as intervals on the line. Note that all secondary structure elements can be nested, but have to be noncrossing. (*B*) Corresponding secondary structure.

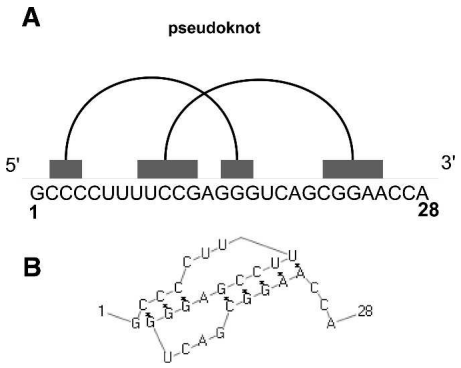


FIGURE 2. (*A*) Representation of a pseudoknot on the line. It consists of two crossing stem intervals. (*B*) Corresponding pseudoknot structure generated by PseudoViewer (Byun and Han 2006).

Overall, there is very high demand for RNA prediction algorithms including pseudoknots. In this article, we present a new approach, called KnotSeeker, for detecting pseudoknots in primary RNA sequences. Our algorithm works as follows: Given an RNA sequence, we find sequence fragments possibly and exactly harboring a pseudoknot using certain criteria. The small number of candidate pseudoknot sequences is folded by a well-established dynamic programming algorithm to see whether they indeed form a stable pseudoknot as the minimum free energy structure. In a second step, these verified pseudoknot candidates are tested if they are likely to exist in the structure with minimal free energy. Only stable, nonoverlapping pseudoknots are returned as the final result. The main advantage of this heuristic pseudoknot detection is that it handles long sequences fast and finds the correct pseudoknots with higher accuracy compared to other methods.

A deeper understanding of pseudoknot thermodynamics will yield better energy parameters and folding results.

However, it is unlikely that this will improve the efficiency of dynamic programming methods. Hence, the KnotSeeker detection tool is a substantial improvement and support for rapid ab initio RNA structure prediction including pseudoknots. Additionally, KnotSeeker will benefit from pseudoknot thermodynamic parameter improvements in the future.

RESULTS

RNA structure and pseudoknots

The foundation of RNA structure formation is continuous base pairing, resulting in so-called helical regions or stems. RNA comprises various secondary structure elements, e.g., single-stranded regions, stacked base pairs, hairpin loops, multiple loops, interior loops, and bulge loops (Fig. 1). Naturally, we can represent a stem s_i with start point a_i and end point b_i as an interval on the line. Formally, we define a stem interval s_i as follows:

- $s_i = [a_i : b_i]$ with an associated stem length $len(s_i)$.
- $[a_i, a_i + 1, \ldots, a_i + len(s_i) - 1]$ is base-paired with $[b_i, b_i - 1, \ldots, b_i - len(s_i) + 1]$.

A pseudoknot basically consists of two crossing stems s_i and s_j (Fig. 2).

In contrast to proteins, RNA can form independently stable secondary structures, which is crucial for RNA folding (Brion and Westhof 1997). The common assumption is that starting with the single-stranded sequence, the majority of secondary structure elements (e.g., hairpin loops in close vicinity) form that determine tertiary structure (Tinoco and Bustamente 1999). There are exceptions to this rule, like secondary structure rearrangements during RNA folding (Tinoco and Wu 1998). However, it is widely accepted in the research community that standard RNA folding is a hierarchical two-step process (Schroeder et al. 2004).

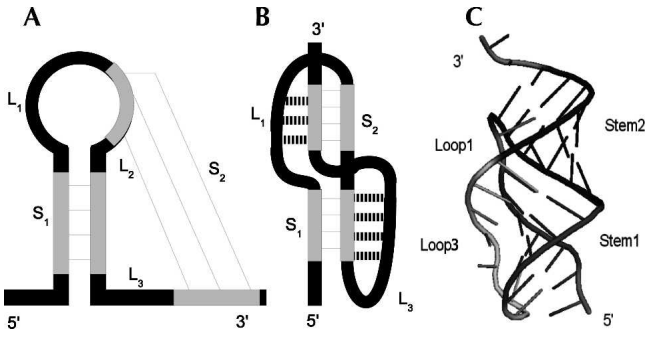


FIGURE 3. Representations of an H-type pseudoknot. (*A*) General formation: base-pairing between a loop and single-stranded region. (*B*) A coaxially stacked pseudoknot: loop–stem interactions are indicated with dotted lines. (*C*) Three-dimensional view of a pseudoknot.

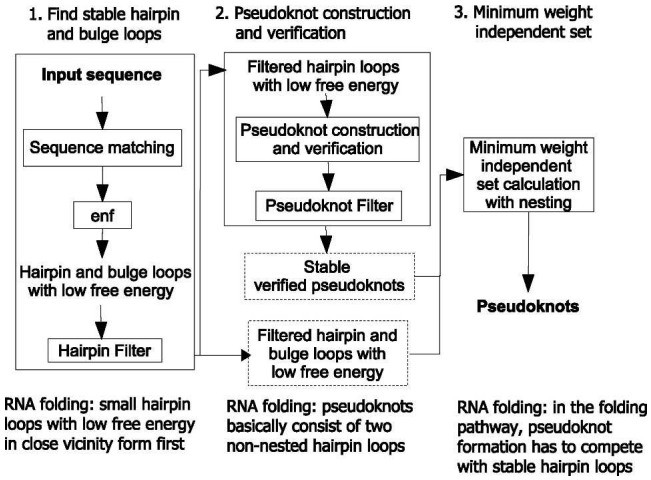


FIGURE 4. New approach for pseudoknot detection and details of its three stages. In the first stage, enf stands for free energy evaluation.

Pseudoknots are tertiary interactions between a loop region and unpaired residues outside the loop. After secondary structure formation, nucleotides in a hairpin loop can base-pair with complementary ones in a single-stranded sequence. These tertiary contacts result in a so-called H-type pseudoknot consisting of two stems (S_1, S_2) and three loops (L_1, L_2, L_3) (Fig. 3). H-type pseudoknots are the simplest and most abundant pseudoknot structures. Over 78% of all 246 in PseudoBase reported pseudoknots are of H-type (van Batenburg et al. 2000).

Pseudoknot folding and formation is a combination of thermodynamics, molecular physics, and sequence composition. It is essential to survey pseudoknots in three-dimensional space for a deeper understanding. The A-form helix forces loops L_1 and L_3 to span the major groove of S_2 and minor groove of S_1, respectively (Pleij et al. 1985). Dependent on the loops, stems, and helical junction, the A-form geometry can deviate from the standard RNA helix. The two stems can coaxially stack with an absent loop L_2. Bent and overtwisted pseudoknot conformations also occur (Giedroc et al. 2000).

Additionally, residues in the loop regions can form tertiary interactions with nucleotides from the minor and major grooves. The shallow and wide minor groove allows tertiary contacts and triple helical regions between S_1 and L_3 (Batey et al. 1999). A-minor interactions resulting from hydrogen bonds between loop adenines and the minor groove are common (Nissen et al. 2001).

Pseudoknot thermodynamic parameters are not very well understood. It is assumed that the free energy of a pseudoknot consists of destabilizing (positive) energy values for loop regions and stabilizing (negative) energy values for stem regions (Gultyaev et al. 1999). The stacked

TABLE 1. Summary of pseudoknot detection results on RNA sequences with less than 300 nt

Sequence			pknotsRG			KnotSeeker			HPknotter			ILM		
ID	nt	PK	*S*	*P*	*r*	*S*	*P*	*r*	*S*	*P*	*r*	*S*	*P*	*r*
5SEColi	120	0	—	—	**0/0**	—	—	**0/0**	—	—	0/2	—	—	0/1
5SDmobilis	133	0	—	—	**0/0**	—	—	**0/0**	—	—	0/1	—	—	0/1
Bacillus-sub	271	1	0	0	0/0	**33.3**	**41.6**	**1/1**	20	37.5	1/3	0	0	—
BMV	282	4	100	100	1/1	**100**	**100**	**2/3**	100	100	2/6	0	0	—
			0	0		**100**	**100**		100	100		62.5	62.5	
			0	0		0	0		0	0		0	0	
			0	0		0	0		0	0		0	0	
BSMV	241	4	**100**	**100**	**3/3**	**100**	**100**	**3/3**	0	0	2/4	0	0	—
			100	**100**		**100**	**100**		100	90.9		90	81.8	
			90	**100**		0	0		90	100		0	0	
			0	0		**74.2**	**88.5**		0	0		0	0	
CAYVV	86	1	0	0	0/1	**100**	**100**	**1/1**	100	100	1/2	0	0	0/2
Cyanophora	290	1	0	0	0/0	**66.7**	**54.5**	**1/1**	66.7	54.5	1/3	0	0	—
DA0260	75	0	—	—	**0/0**	—	—	0/1	—	—	0/1	—	—	0/1
DA1280	73	0	—	—	**0/0**	—	—	**0/0**	—	—	0/1	—	—	0/1
DY4441	73	0	—	—	0/1	—	—	**0/0**	—	—	0/3	—	—	0/1
Dros-mel	81	0	—	—	**0/0**	—	—	**0/0**	—	—	0/1	—	—	0/1
GLV	266	1	**69.6**	**69.6**	**1/2**	69.6	69.6	1/3	26.1	66.7	1/4	0	0	—
HDVanti	91	1	0	0	0/0	**86.9**	**71.4**	**1/1**	0	0	0/2	**100**	**66.7**	**1/1**
HDV	216	1	0	0	0/0	**90.6**	**93.5**	**1/3**	31.2	50	1/4	0	0	0/0
Human-mi	110	0	—	—	**0/0**	—	—	0/1	—	—	0/1	—	—	0/1
Human-telo	210	1	0	0	0/1	**35.5**	**50**	**1/2**	0	0	0/2	0	0	—
IBV	220	1	0	0	0/0	**72.2**	**92.9**	**1/1**	0	0	0/3	72.2	68.4	—
NeRNV	287	5	0	0	1/2	**100**	**100**	**4/4**	100	88.9	4/6	0	0	—
			0	0		**77.8**	**87.5**		44.4	57.1		44.4	57.1	
			0	0		**90**	**100**		40	44.4		0	0	
			0	0		0	0		0	0		0	0	
			100	100		**100**	**100**		100	100		0	0	
satRPV	73	1	**81.8**	**85.7**	**1/1**	**81.8**	**85.7**	**1/1**	0	0	0/1	77.3	68	1/1
STNV1	252	4	0	0	2/2	**100**	**100**	**3/3**	0	0	2/5	0	0	—
			50	42.9		**100**	**85.7**		100	85.7		0	0	
			100	100		**58.3**	**42.1**		100	100		0	0	
			0	0		0	0		0	0		90	90	
TMVdown	105	2	0	0	0/0	**100**	**95.8**	**2/2**	**100**	**95.8**	**2/2**	0	0	0/1
			100	100		**100**	**100**		**100**	**100**		0	0	
TMVup	84	3	**71.4**	**62.5**	**3/3**	**71.4**	**62.5**	**3/3**	**71.4**	**62.5**	**3/3**	85.7	66.7	1/2
			77.8	**87.5**		**77.8**	**87.5**		**77.8**	**87.5**		0	0	
			88.9	**100**		**88.9**	**100**		**88.9**	**100**		0	0	
TMV	214	5	0	0	0/0	**77.8**	**87.5**	**5/5**	77.8	87.5	5/6	0	0	—
			0	0		**81.8**	**90**		81.8	90		0	0	
			0	0		**88.9**	**100**		88.9	100		0	0	
			0	0		**95.8**	**100**		95.8	100		0	0	
			0	0		**100**	**100**		100	100		0	0	
TYMV	86	1	100	80	1/2	**100**	**80**	**1/1**	100	80	1/2	62.5	55.5	1/2

The best results in terms of sensitivity, specificity, and *r* are marked in bold. PK corresponds to the number of pseudoknots as reported in the literature.

stem energy for S_1 and S_2 can be calculated using the additive sum of the nearest-neighbor model. However, the entropic loop energies for L_1 and L_3 still have to be estimated. The loops are not equivalent stereochemically, as they cross different grooves. The simple nearest-neighbor model also neglects the important stem–loop correlations (Cao and Chen 2006). Additionally, stabilizing coaxial stacking and base triples at the helical junction need to be taken into account.

Detection of pseudoknots

The detection of pseudoknots must be clearly distinguished from RNA structure prediction including pseudoknots. Pseudoknot detection is a self-contained step without simultaneous secondary structure prediction aimed to return only pseudoknots. If pseudoknots can be detected with high accuracy, the remaining sequence can efficiently be folded using state-of-the-art secondary structure prediction programs

TABLE 2. Summary of pseudoknot detection results on RNA sequences longer than 300 nt

Sequence			pknotsRG			KnotSeeker			HPknotter			ILM		
ID	nt	PK	*S*	*P*	*r*	*S*	*P*	*r*	*S*	*P*	*R*	*S*	*P*	*r*
BCV	345	1	**100**	**100**	**1/1**	**100**	**100**	**1/1**	100	100	1/3	88.3	100	1/3
EColi	363	4	0	0	0/0	**100**	**100**	**2/2**	100	100	2/6	0	0	—
			0	0		0	0		62.5	62.5		0	0	
			0	0		**66.7**	**57.1**		0	0		0	0	
			0	0		0	0		0	0		42.1	66.7	
HPeV1	709	1	54.5	54.5	1/4	**100**	**100**	**1/8**	100	100	1/14	0	0	—
MHV	315	1	85.7	90	1/3	**85.7**	**90**	**1/2**	85.7	90	1/6	38.1	27.6	—
ORSV	419	11	0	0	2/5	**90**	**90**	**8/8**	**90**	**90**	8/11	0	0	—
			0	0		**77.8**	**87.5**		**77.8**	**87.5**		0	0	
			0	0		**88.9**	**88.9**		**88.9**	**88.9**		0	0	
			0	0		**35.7**	**41.7**		0	0		0	0	
			0	0		66.7	42.9		**100**	**100**		0	0	
			0	0		**88.9**	**100**		**88.9**	**100**		0	0	
			0	0		0	0		**68.7**	**57.9**		75	66.7	
			77.8	87.5		**77.8**	**87.5**		**77.8**	**87.5**		0	0	
			0	0		0	0		0	0		0	0	
			0	0		0	0		0	0		0	0	
			57.1	57.1		**57.1**	**57.1**		**57.1**	**57.1**		0	0	
SARS-TW1	341	1	0	0	0/0	**100**	**100**	**1/3**	100	100	1/5	84.2	64	—
SNV	537	1	0	0	0/1	**92.9**	**100**	**1/5**	92.9	100	1/8	0	0	—
STMV	421	8	0	0	1/1	**91.7**	**91.7**	**5/6**	**91.7**	**91.7**	7/10	0	0	—
			0	0		**100**	**100**		80	100		0	0	
			0	0		0	0		**100**	**80**		0	0	
			0	0		0	0		0	0		0	0	
			0	0		0	0		**100**	**90**		0	0	
			0	0		**88.9**	**100**		**88.9**	**100**		0	0	
			0	0		**63.2**	**50**		**63.2**	**50**		0	0	
			100	100		**100**	**100**		**100**	**100**		0	0	
T2	946	1	**100**	**100**	**1/1**	100	100	1/5	100	100	1/16	0	0	—
T4	1340	1	0	0	0/1	**100**	**100**	**1/8**	100	100	1/17	0	0	—

The best results in terms of sensitivity, specificity, and *r* are marked in bold. PK corresponds to the number of pseudoknots as reported in the literature.

in $O(n^3)$ time and $O(n^2)$ space. There are several programs for RNA structure prediction including pseudoknots; however HPknotter is the only tool performing sheer pseudoknot detection so far (Huang et al. 2005). HPknotter can improve RNA secondary structure prediction including pseudoknots. However, it suffers from a high number of returned false positive pseudoknots. HPknotter finds pseudoknots using the following steps: First, RNAMotif's structural matcher returns a great number of possible pseudoknot fragments for a given input sequence. The NUPACK energy calculation tool is used for removing hits with lower non-pseudoknotted MFE structure. Second, pseudoknot verification is performed by pknots, NUPACK, or pknotsRG to see if a filtered hit indeed folds into the desired pseudoknot structure. A minimum weight independent set calculation returns a mutually disjoint pseudoknot set as the result.

The new approach for pseudoknot detection followed in this article is presented in Figure 4. Unlike HPknotter, KnotSeeker is based on RNA folding assumptions and free energy minimization considering stable secondary structure elements. Detailed descriptions of the three main steps can be found in the Materials and Methods section.

Experimental results

We tested KnotSeeker on 34 sequences covering various RNA classes. The sequence lengths range from 73 nucleotides (nt) to 1340 nt. As KnotSeeker is designed for detecting pseudoknots in primary RNA sequences, we report sensitivity and specificity only for pseudoknotted base pairs using the following notation as in Baldi et al. (2000):

- Sensitivity $S = \dfrac{100 \times \mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$
- Specificity $P = \dfrac{100 \times \mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$

TP (true positive) corresponds to the number of correctly predicted base pairs in the predicted pseudoknot, FN (false negative) to the number of base pairs in the published pseudoknot that were not predicted, and FP (false positive) to the

TABLE 3. A comparison of running times for all RNA sequences

ID	Length	pknotsRG	KnotSeeker	HPknotter	ILM
5SEColi	120	0.3 sec	2.8 sec	20 sec	0.3 sec
5SDmobilis	133	0.4 sec	5.3 sec	34 sec	0.3 sec
Bacillus-sub	271	3.9 sec	10.2 sec	1 min 46 sec	0.9 sec
BCV	345	8.4 sec	3.5 sec	2 min	0.8 sec
BMV	282	4.5 sec	5.9 sec	2 min 8 sec	0.8 sec
BSMV	241	2.3 sec	1.9 sec	1 min 3 sec	0.8 sec
CAYVV	86	0.2 sec	0.6 sec	18 sec	0.2 sec
Cyanophora	290	5.6 sec	3.5 sec	2 min 12 sec	0.9 sec
DA0260	75	0.2 sec	0.6 sec	17 sec	0.2 sec
DA1280	73	0.2 sec	0.5 sec	23 sec	0.2 sec
DY4441	73	0.1 sec	0.5 sec	24 sec	0.2 sec
Dros-mel	81	0.2 sec	0.5 sec	36 sec	0.2 sec
EColi	363	11.1 sec	17.5 sec	1 min 47 sec	1.5 sec
GLV	266	3.5 sec	5.9 sec	1 min 24 sec	0.7 sec
HDVanti	91	0.2 sec	1.9 sec	15 sec	0.3 sec
HDV	216	1.6 sec	12.7 sec	1 min 8 sec	0.5 sec
HPeV1	709	2 min 57 sec	37.6 sec	4 min 39 sec	11 sec
Human-mi	110	0.3 sec	0.6 sec	47 sec	0.3 sec
Human-telo	210	1.7 sec	17.3 sec	1 min 40 sec	0.6 sec
IBV	220	1.9 sec	6.3 sec	1 min 46 sec	0.5 sec
MHV	315	6 sec	3.3 sec	1 min 9 sec	1.1 sec
NeRNV	287	1.2 sec	5.8 sec	1 min 7 sec	0.6 sec
ORSV	419	21.4 sec	11.3 sec	3 min 32 sec	2 sec
SARS-TW1	341	9 sec	3.9 sec	1 min 40 sec	1.2 sec
satRPV	73	0.6 sec	1.2 sec	11 sec	0.2 sec
STMV	421	21.6 sec	9.3 sec	3 min 23 sec	2.3 sec
SNV	537	53.8 sec	26.5 sec	3 min 15 sec	4.7 sec
STNV1	252	3 sec	8.7 sec	1 min 41 sec	0.7 sec
TMVdown	105	0.2 sec	1 sec	24 sec	0.3 sec
TMVup	84	0.2 sec	0.4 sec	20 sec	0.2 sec
TMV	214	1.6 sec	3 sec	1 min 24 sec	0.6 sec
TYMV	86	0.2 sec	0.6 sec	10 sec	0.2 sec
T2	946	9 min 28 sec	47.7 sec	5 min 24 sec	34 sec
T4	1340	41 min 3 sec	1 min 39 sec	7 min 39 sec	1 min 54 sec

number of incorrectly predicted base pairs in the predicted pseudoknot. We also report the ratio r = (number of correctly predicted pseudoknots)/(number of predicted pseudoknots).

We compared the results to three other methods, namely the dynamic programming algorithm pknotsRG (mfe mode) (Reeder and Giegerich 2004) and the heuristic approaches HPknotter (general descriptor) (Huang et al. 2005) and ILM (Ruan et al. 2004). We also obtained the results achieved by HotKnots (Ren et al. 2005), which heuristically finds 20 structures with lowest free energy. However, we discovered that HotKnots did not return any correct pseudoknots in the best structure with lowest free energy for our test sequences. A comparison of the remaining suboptimal folding scenarios with the other algorithms would be biased and thus we excluded HotKnots in our evaluation. We were unable to obtain results from pknots (Rivas and Eddy 1999) and NUPACK (Dirks and Pierce 2003) due to running out of memory for sequences longer than 150 nt and 200 nt, respectively. We discovered that ILM tends to predict very complex pseudoknots for long sequences. In many cases, we found that long-range pseudoknotted helices with several internal H-type pseudoknots cover the whole sequence. Therefore, it is hard to correctly assign the ratio r, because the number of predicted pseudoknots is ambiguous. Whenever this is the case, we omit the r value for the results obtained by ILM.

The pseudoknot detection results are displayed in detail in Tables 1 and 2 and the best results for a sequence are

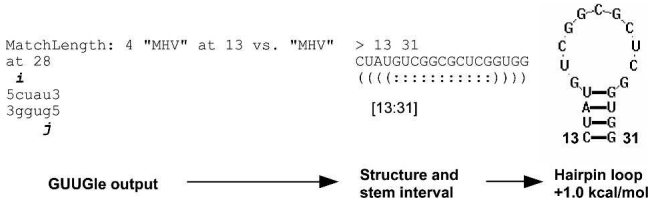


FIGURE 5. GUUGle output, corresponding structural interval representation and hairpin loop with free energy. Note that the stem interval for the GUUGle output correlates to $s_i = [i : j + k - 1]$.

highlighted. One should note that more than one method can give best results for one sequence. For sequences shorter than 300 nt, pknotsRG gives best results for 10 of the 24 sequences. HPknotter and ILM show poor performance, dominating on only 2 and 1 sequences, respectively. KnotSeeker clearly outperforms the other methods with best results on 21 of the 24 sequences. For the set of long sequences (>300 nt), there is a similar scenario. KnotSeeker achieves the best results on 9 of the 10 sequences, whereas pknotsRG and HPknotter dominate on only two of the 10 sequences. Both sensitivity and specificity of the ILM predictions are significantly lower than those for the other methods.

The results emphasize that the strategy followed by KnotSeeker and HPknotter greatly improves the pseudoknot prediction results. The dynamic programming algorithm pknotsRG misses most pseudoknots in the test sequences. This illustrates the limitations of the dynamic programming approach and underlying energy model for long sequences. However, pknotsRG has very high sensitivity and specificity for short sequence fragments exactly harboring a pseudoknot. This becomes clear because KnotSeeker and HPknotter mainly achieve higher sensitivity and specificity because of the correct pseudoknot verification results returned by pknotsRG. Our approach clearly outperforms HPknotter for all test sequences. HPknotter returns many false positive pseudoknots, especially for longer sequences. Even though both procedures follow a similar idea (find sequence fragments possibly harboring a pseudoknot and verify them), KnotSeeker is significantly more accurate. This is due to the fact that our approach is based on RNA folding assumptions and takes into account competing secondary structure elements in the minimum weight independent set (MWIS) calculation.

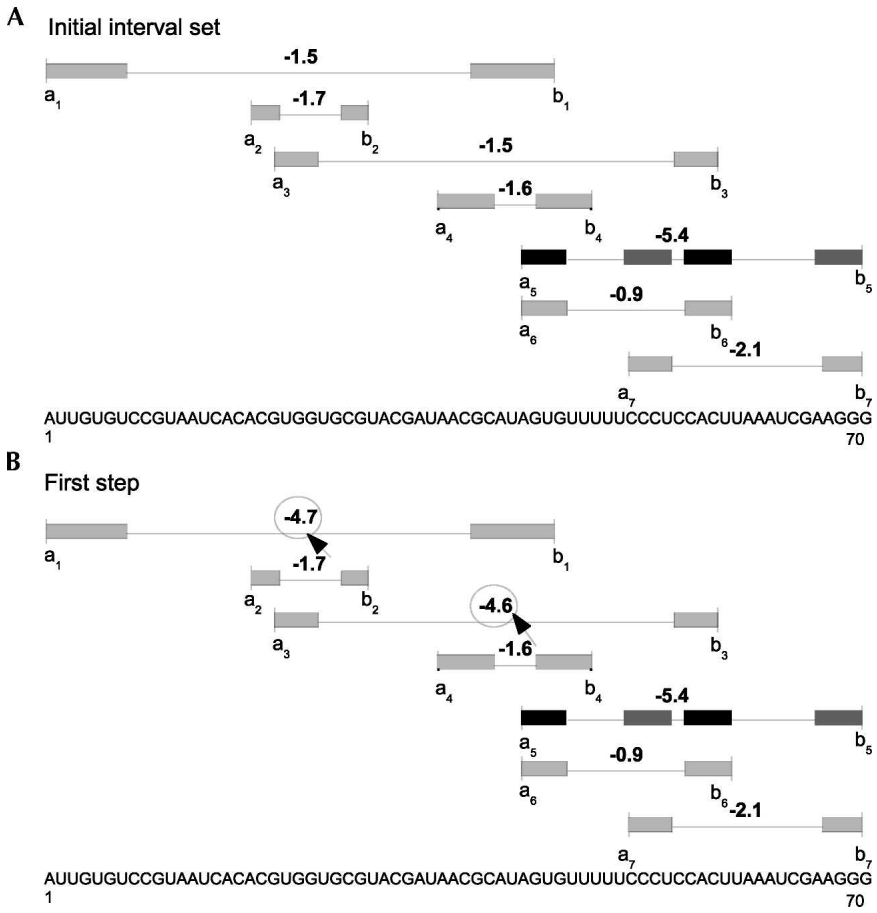


FIGURE 7. (*A*) Initial interval set with six stem intervals and one pseudoknot interval s_5. (*B*) Interval set after first step to include nested structures with new updated weights.

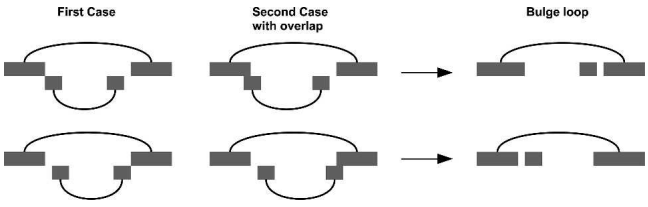
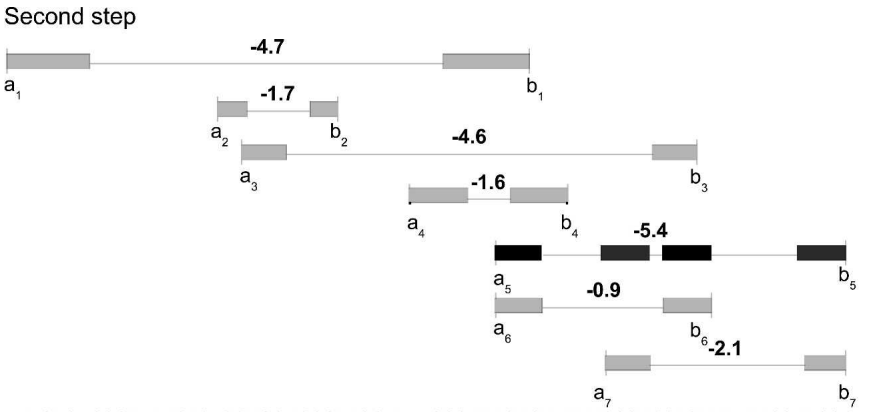


FIGURE 6. Construction of stem intervals with bulge loops of size one from the given sorted list of stem intervals. A partial overlap of size one is allowed in the second case.

We also report the running time for all approaches (Table 3; see Materials and Methods for experimental details). HPknotter runs in the order of minutes for sequences longer than 200 nt. pknotsRG is very efficient on short sequences; however, it becomes computationally expensive for long sequences due to its time requirements of $O(n^4)$. In contrast to that, KnotSeeker is a rapid tool and runs in the order of seconds. It is significantly faster than pknotsRG and HPknotter on longer sequences. KnotSeeker takes less than 2 min to detect the pseudoknot in the very long T4 gene 32 mRNA sequence (1340 nt), whereas pknotsRG requires more than 40 min to fold the sequence. ILM is also a very efficient approach; however with the drawback of low sensitivity and specificity for ab initio structure prediction.

DISCUSSION

Our approach gives the best results for pseudoknot detection when compared to pknotsRG, HPknotter, and ILM. KnotSeeker detects almost all predicted pseudoknots



endpoints scanned	μ(i)	temp_min
a_1	w(1) = 0 - 4.7 = - 4.7	0
a_2	w(2) = 0 - 1.7 = - 1.7	0
a_3	w(3) = 0 - 4.6 = - 4.6	0
b_2	w(2) < temp_min	-1.7
a_4	w(4) = -1.7 – 1.6 = -3.3	-1.7
a_5	w(5) = -1.7 – 5.4 = -7.1	-1.7
a_6	w(6) = -1.7 – 0.9 = 2.6	-1.7
b_1	w(1) < temp_min	-4.7
b_4	w(4) > temp_min	-4.7
a_7	w(7) = - 4.7 – 2.1 = -6.8	-4.7
b_3	w(3) > temp_min	-4.7
b_6	w(6) > temp_min	-4.7
b_5	w(5) < temp_min	-7.1
b_7	w(7) > temp_min	-7.1

s_1 = [1:43], len(s_1) = 7, w(s_1) = -4.7
s_2 = [18:28], len(s_2) = 4, w(s_2) = -1.7
s_3 = [20:57], len(s_3) = 4, w(s_3) = -4.6
s_4 = [34:46], len(s_4) = 5, w(s_4) = -1.6
s_5 = [41:70], len(s_6) = 4, w(s_6) = -5.4
s_6 = [41:58], len(s_5) = 4, w(s_5) = -0.9
s_7 = [50:70], len(s_7) = 4, w(s_7) = -2.1

Result: [18, 28], [41, 70]
Pseudoknots: [41, 70]

FIGURE 8. MWIS calculation using a sweep line strategy. The final result consists of the pseudoknot interval s_5.

and returns significantly less incorrect pseudoknots than HPknotter. Especially for long sequences, our method is a substantial improvement for RNA structure prediction including pseudoknots. Pseudoknot detection prior to structure prediction is a successful and computationally efficient route. This was first demonstrated by HPknotter (Huang et al. 2005) and is now emphasized by the work presented in this article. KnotSeeker returns significantly more accurate results than the heuristic approaches ILM and HotKnots for long RNA sequences. This is mainly due to the fact that we perform a pseudoknot verification step that is consistent with the MFE model, whereas ILM and HotKnots simply combine highest scoring crossing helices. However, one should acknowledge that the heuristic approaches ILM and HotKnots perform well for different frameworks. ILM produces good results for a set of aligned sequences, whereas HotKnots is a reliable heuristic approach for short sequences.

The pseudoknot detection approach KnotSeeker is limited by a few factors. At this stage only certain pseudoknots that can be folded by pknotsRG are detected. These are so-called canonical, recursive pseudoknots (Reeder and Giegerich 2004). Using pknots with a high running time of $O(n^6)$ can improve the results, especially for detecting more complex pseudoknots as in tmRNA or IRES elements (Rivas and Eddy 1999). An experiment using different pseudoknot thermodynamic parameters as in Cao and Chen (2006) or including partition function information (McCaskill 1990) is also an option. Furthermore, one can think of performing an alignment with known pseudoknot classes like retroviral frameshift sites to achieve more accurate results. During the MWIS calculation for nested structures, a free energy evaluation considering the secondary structure can be implemented. This should improve the results considerably and even lead to a fast novel RNA secondary structure prediction method including pseudoknots.

MATERIALS AND METHODS

In this section, we give a detailed description of the algorithmic details and the sequence data used for testing.

The KnotSeeker algorithm

Find stable hairpin and bulge loops

GUUGle is a search tool that finds all exact matches (under RNA base-pairing rules) of a minimum specified length between target and query sequences (Gerlach and Giegerich 2006). It makes use of suffix arrays and runs fast. A target sequence vs. target sequence search can be used to detect helical regions within a sequence.

In the first step, we let GUUGle detect exact matches with length larger or equal to 3 base pairs (bp). These matches correspond to helical regions. GUUGle returns sequence fragments of a certain length k and two indices i,j (Fig. 5). The output usually consists of a large number of matches. The goal is to identify relevant matches. Following the initial assumption, we keep only those intervals with $j - i \geq 6$, according to minimal hairpin loop lengths. Given the sorted stem interval list derived from the GUUGle output, bulge loops of size one are found as well through a simple combination of intervals (Fig. 6).

As the corresponding secondary structure is known for each hairpin or bulge loop, we let RNAeval (Vienna RNA package 1.7) evaluate the free energy using the Turner parameters (Hofacker et al. 1994; Mathews et al. 1999). We only keep secondary structures with free energy < +2.0 kcal/mol with the motivation that stems with low free energy are likely to form in the native structure. Formally, each stem interval $s_i = [a_i : b_i]$ has an associated weight $w(s_i)$ corresponding to its free energy value. To further limit the size of the hairpin and bulge loop set, the following assumption is used: *RNA folding is a two-step process and small structures with low free energy in close vicinity form first.* The set of hairpin loop intervals is parsed as follows:

TABLE 4. Overview of the sequences used in our tests

Type	Sequence ID	Accession no.	Reference
5S rRNA	5SEColi, 5SDmobilis	V00336, X07545	Cannone et al. (2002)
3′UTR	BCV, BMV, BSMV, MHV, NeRNV, ORSV, SARS-TW1, STMV, STNV1, TMV, TMVup	AF220295, V00099, X03854, AF201929, AY751778, U34586, AY291451, M25782, J02399, J02415, AJ011933	van Batenburg et al. (2000)
5′UTR	HPeV1	L02971	Nateri et al. (2002)
TLS	CAYVV, TMVdown, TYMV	U91413, J02415, X16378	van Batenburg et al. (2000)
tRNA	DA0260, DA1280, DY4441	N/A	Sprinzl et al. (1998)
miRNA	Dros-mel, Human-mi	AJ550546, AJ550395	Griffiths-Jones et al. (2006)
tmRNA	EColi, Cyanophora	U68074, U30821	Williams (2000)
mRNA	T2, T4	X12460, J02513	van Batenburg et al. (2000)
frameshift	IBV	M27472	Napthine et al. (1999)
readthrough	SNV	M54993	van Batenburg et al. (2000)
ribozymes	HDV, HDVanti, satRPV	X04451, X04451, M63666	van Batenburg et al. (2000)
IRES	GLV	L13218	Garlapati and Wang (2002)
telomerase	Human-telo	AF221907	Chen et al. (2000)
SRP RNA	Bacillus-sub	X06802	Rosenblad et al. (2003)

Note that 5S rRNA, tRNA, and miRNA are all pseudoknot free.

- Given two intervals $[i : k]$ and $[i : l]$ with $k < l$. If $w([i : k]) < w([i : l])$, then delete longer interval $[i : l]$.
- Given two intervals $[k : i]$ and $[l : i]$ with $k < l$. If $w([k : i]) > w([l : i])$, then delete longer interval $[k : i]$.

As an output for the first step, we get a list of filtered hairpin and bulge loops with their corresponding free energy values.

Pseudoknot construction and verification

A simple H-type pseudoknot basically consists of two crossing stems with low free energy. Given the list of filtered hairpin loops from the first step, two entries can be combined to potentially form a pseudoknot. An examination of entries in PseudoBase led us to use certain pseudoknot loop length restrictions similar to HPknotter (van Batenburg et al. 2000; Huang et al. 2005). The goal of this heuristic is to keep the set of candidate pseudoknots as small as possible while considering naturally occurring stem and loop lengths.

- 1 nt $\leq$ size (Loop L_1) $\leq$ 20 nt.
- 0 nt $\leq$ size (Loop L_2) $\leq$ 35 nt.
- 1 nt $\leq$ size (Loop L_3) $\leq$ 75 nt.

Overall, we assume that a pseudoknot has to have a length $\leq$90 nt, as this returns the most significant results. These simple pseudoknots are among the best studied, whereas thermodynamics of very long pseudoknots are not well understood. Additionally, the following observation was made by us during preliminary testing: the two hairpin loop intervals potentially forming a pseudoknot need to have a combined free energy sum of less than −2.5 kcal/mol. This improves the runtime drastically, as only a small portion of intervals need to be combined as a pseudoknot candidate. Two important points should be noted. First, certain secondary structure rearrangements during pseudoknot formation are allowed, e.g., stems can partially overlap. Second, three-stemmed pseudoknots with an additional stem in their loops are also naturally included in pseudoknot construction.

Given the list of possible pseudoknots, we test with pknotsRG in $O(n^4)$ time and $O(n^2)$ space if they actually fold into stable pseudoknots (Reeder and Giegerich 2004). This verification is fast, as the list of candidates is small and the test runs on short sequence fragments exactly harboring a potential pseudoknot. Our pseudoknot filter procedure returns the desired and verified pseudoknots. However, pknotsRG returns several false positive verified pseudoknots, which do not occur in the native structure. This issue is tackled in the next step.

Minimum weight independent set

The verified pseudoknots plus filtered hairpin and bulge loops form our candidate structure elements set. To eliminate false positive pseudoknots from the second step, an MWIS calculation is performed. This corresponds to the following RNA folding assumption: *in the folding pathway, pseudoknot formation has to compete with stable secondary structure elements*.

The MWIS problem on a weighted interval set can be solved in linear time and space with an additional $O(n \log n)$ sorting step (Hsiao et al. 1992). It is based on a sweep line strategy and returns the set of nonoverlapping intervals with minimum weight as an output. For the MWIS calculation required here, one additional assumption regarding RNA folding has to be added. There can be nested structures; a hairpin or bulge loop can have several internally nested hairpin and bulge loops or pseudoknots. Like before, no two structure elements are allowed to overlap. For the calculation we assume that nesting results in an additional −1.5 kcal/mol free energy gain for the outer stem interval, as this is an energetically favorable process. The final output consists of the pseudoknots that are likely to occur in the native structure with minimum free energy. The following notations and assumptions are required for the MWIS algorithm:

- Let $s_i = [a_i : b_i]$ be a structure element interval with an associated weight $w(s_i)$ corresponding to its free energy value.
- Let $S = \{s_1, s_2, \ldots, s_n\}$ be the set of candidate structure elements.

- The sorted endpoints list $L = \{e_1, e_2, \ldots, e_{2n}\}$ is given.
- $temp_{\min}$ is a temporary variable that stores the MWIS weight of the set of intervals whose right endpoints have been scanned.
- $\mu(i) = w(s_i) + \min\{\mu(j) \mid b_j < a_i\}$ for any $1 \leq i \leq n$.

The MWIS algorithm scans the endpoints list. If the endpoint scanned is a left endpoint a_i, the weight $w(s_i)$ plus $temp_{\min}$ is stored in $\mu(i)$. If the endpoint scanned is a right endpoint b_i, $\mu(i)$ is checked to see whether it is smaller than $temp_{\min}$ or not. For $\mu(i) < temp_{\min}$, the value of $temp_{\min}$ is replaced by $\mu(i)$. At the end of the calculation, the MWIS weight of S is stored in $temp_{\min}$ and the resulting interval set can be recovered through a traceback step.

First step: Including nested structures

The first step delivers nested intervals and their corresponding updated energy values. The sorted endpoints list is scanned from left to right. If the right endpoint b_i of a hairpin or bulge loop interval s_i is discovered, a search is performed to find all stems and pseudoknots contained in the interval $[a_i + len(s_i) - 1 : b_i - len(s_i) + 1]$. However, the resulting set $S_{nested}(i)$ can have overlapping structure elements. Therefore, a standard MWIS calculation is performed on the set $S_{nested}(i)$ to find only nonoverlapping nested structures of minimum weight. The updated weight $w(s_i)$ of the outer stem s_i is the weight of the MWIS plus an additional -1.5 kcal/mol. This value turned out to give the best results during preliminary testing. The output of this first step is the list of structure elements with new updated weights accounting for nested structures.

Second step: MWIS calculation

In the second step, an overall MWIS calculation is performed on the new structure element candidate set including nesting in linear time and space. The result consists of pseudoknots, hairpin loops, and bulge loops with combined minimum free energy. As this approach is designed for pseudoknot detection, the final output only returns pseudoknots. The different steps of the MWIS calculation are illustrated in Figures 7 and 8.

Sequence test data

An overview of the sequences selected for testing is provided in Table 4. We chose both pseudoknotted and pseudoknot-free sequences from the literature.

Experimental and implementation details

The KnotSeeker pipeline was implemented in Python 2.5 incorporating several existing programs, namely GUUGle (Gerlach and Giegerich 2006), RNAeval (Vienna RNA package 1.7; Hofacker et al. 1994), and pknotsRG (Reeder and Giegerich 2004). The experiments and time measurements for pknotsRG (mfe mode), KnotSeeker, and ILM were carried out with a dual Intel 1.66 GHz processor and 1 GB main memory. The results for HPknotter were obtained from its Web server, which returns also the computation time.

Received December 20, 2007; accepted January 11, 2008.

REFERENCES

Abrahams, J.P., van den Berg, M., van Batenburg, E., and Pleij, C.W.A. 1990. Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucleic Acids Res.* **18:** 3035–3044. doi: 10.1093/nar/18.10.3035.

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

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F., and Nielsen, H. 2000. Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics* **16:** 412–424.

Baril, M., Dulude, D., Steinberg, S.V., and Brakier-Gingras, L. 2003. The frameshift stimulatory signal of human immunodeficiency virus type 1 group O is a pseudoknot. *J. Mol. Biol.* **331:** 571–583.

Batey, R.T., Rambo, R.P., and Doudna, J.A. 1999. Tertiary motifs in RNA structure and folding. *Angew. Chem. Int. Ed. Engl.* **38:** 2326–2343.

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

Brion, P. and Westhof, E. 1997. Hierarchy and dynamics of RNA folding. *Annu. Rev. Biophys. Biomol. Struct.* **26:** 113–137.

Brown, M. and Wilson, C. 1996. RNA pseudoknot modeling using intersections of stochastic context free grammars with applications to database search. In *Proceedings of the 1996 Pacific Symposium on Biocomputing* (eds. L. Hunter and T.E. Klein), pp. 109–125. World Scientific Publishing, Singapore.

Byun, Y. and Han, K. 2006. PseudoViewer: Web application and web service for visualizing RNA pseudoknots and secondary structures. *Nucleic Acids Res.* **34:** W416–W422. doi: 10.1093/nar/gkl210.

Cai, L., Malmberg, R.L., and Wu, Y. 2003. Stochastic modeling of RNA pseudoknotted structures: A grammatical approach. *Bioinformatics* **19:** 66–73.

Cannone, J.J., Subramanian, S., Schnare, M.N., Collett, J.R., D'Souza, L.M., Du, Y., Feng, B., Lin, N., Madabusi, L.V., Müller, K.M., 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. and Chen, S.-J. 2006. Predicting RNA pseudoknot folding thermodynamics. *Nucleic Acids Res.* **34:** 2634–2652. doi: 10.1093/nar/gkl346.

Chen, J., Blasco, M., and Greider, C. 2000. Secondary structure of vertebrate telomerase RNA. *Cell* **100:** 503–514.

Dirks, R.M. and Pierce, N.A. 2003. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.* **24:** 1664–1677.

Garlapati, S. and Wang, C. 2002. Identification of an essential pseudoknot in the putative downstream internal ribosome entry site in giardiavirus transcript. *RNA* **8:** 601–611.

Gerlach, W. and Giegerich, R. 2006. GUUGle: A utility for fast exact matching under RNA complementary rules including G-U base pairing. *Bioinformatics* **22:** 762–764.

Giedroc, D.P., Theimer, C.A., and Nixon, P.L. 2000. Structure, stability, and function of RNA pseudoknots involved in stimulating ribosomal frameshifting. *J. Mol. Biol.* **298:** 167–185.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., and Enright, A.J. 2006. miRBase: MicroRNA sequences, targets, and gene nomenclature. *Nucleic Acids Res.* **34:** D140–D144. doi: 10.1093/nar/gkj112.

Gultyaev, A.P., van Batenburg, F.H.D., and Pleij, C.W.A. 1995. The computer simulation of RNA folding pathways using a genetic algorithm. *J. Mol. Biol.* **250:** 37–51.

Gultyaev, A.P., van Batenburg, F.H.D., and Pleij, C.W.A. 1999. An approximation of loop free energy values of RNA H-pseudoknots. *RNA* **5:** 609–617.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., and Schuster, P. 1994. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* **125:** 167–188.

Hsiao, J.Y., Tang, C.Y., and Chang, R.S. 1992. An efficient algorithm for finding a maximum weight 2-independent set on interval graphs. *Inf. Process. Lett.* **43:** 229–235.

Huang, X. and Ali, H. 2007. High sensitivity RNA pseudoknot prediction. *Nucleic Acids Res.* **35:** 656–663. doi: 10.1093/nar/gkl943.

Huang, C.-H., Lu, C.L., and Chiu, H.-T. 2005. A heuristic approach for detecting RNA H-type pseudoknots. *Bioinformatics* **21:** 3501–3508.

Lyngso, R.B. and Pedersen, C.N. 2000. RNA pseudoknot prediction in energy-based models. *J. Comput. Biol.* **7:** 409–427.

Mathews, D., Sabina, J., Zuker, M., and Turner, D.H. 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.

Napthine, S., Liphardt, J., Bloys, A., Routledge, S., and Brierley, I. 1999. The role of RNA pseudoknot stem 1 length in the promotion of efficient −1 ribosomal frameshifting. *J. Mol. Biol.* **288:** 305–320.

Nateri, A.S., Hughes, P.J., and Stanway, G. 2002. Terminal RNA replication elements in human parechovirus 1. *J. Virol.* **76:** 13116–13122.

Nissen, P., Ippolito, J.A., Ban, N., Moore, P.B., and Steitz, T.A. 2001. RNA tertiary interactions in the large ribosomal subunit: The A-minor motif. *Proc. Natl. Acad. Sci.* **98:** 4899–4903.

Pleij, C.W.A., Rietveld, K., and Bosch, L. 1985. A new principle of RNA folding based on pseudoknotting. *Nucleic Acids Res.* **13:** 1717–1731. doi: 10.1093/nar/13.5.1717.

Reeder, J. and 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.

Ren, J., Rastegari, B., Condon, A., and Hoos, H.H. 2005. HotKnots: Heuristic prediction of RNA secondary structures including pseudoknots. *RNA* **11:** 1494–1504.

Rivas, E. and Eddy, S.R. 1999. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.* **285:** 2053–2068.

Rosenblad, M.A., Gorodkin, J., Knudsen, B., Zwieb, C., and Samuelsson, T. 2003. SRPDB: Signal Recognition Particle Database. *Nucleic Acids Res.* **31:** 363–364. doi: 10.1093/nar/gkg107.

Ruan, J., Stormo, G.D., and Zhang, W. 2004. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics* **20:** 58–66.

Schroeder, R., Barta, A., and Semrad, K. 2004. Strategies for RNA folding and assembly. *Nat. Rev. Mol. Cell Biol.* **5:** 908–919.

Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A., and Steinberg, S. 1998. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.* **26:** 148–153. doi: 10.1093/nar/26.1.148.

Staple, D.W. and Butcher, S.E. 2005. Pseudoknots: RNA structures with diverse functions. *PLoS Biol.* **3:** 956–959. doi: 10.1371/journal.pbio.0030213.

Tabaska, J.E., Cary, R.B., Gabow, H.N., and Stormo, G.D. 1999. An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics* **14:** 691–699.

Thiel, V., Ivanov, K.A., Putics, A., Hertzig, T., Schelle, B., Bayer, S., Weissbrich, B., Snijder, E.J., Rabenau, H., Doerr, H.W., et al. 2003. Mechanisms and enzymes involved in SARS coronavirus genome expression. *J. Gen. Virol.* **84:** 2305–2315.

Tinoco, I. and Bustamente, C. 1999. How RNA folds. *J. Mol. Biol.* **293:** 271–281.

Tinoco, I. and Wu, M. 1998. RNA folding causes secondary structure rearrangement. *Proc. Natl. Acad. Sci.* **95:** 11555–11560.

van Batenburg, F.H.D., Gultyaev, A.P., and Pleij, C.W.A. 1995. An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *J. Theor. Biol.* **174:** 269–280.

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.

Williams, K.P. 2000. The tmRNA website. *Nucleic Acids Res.* **28:** 168. doi: 10.1093/nar/28.1.168.

Xayaphoummine, A., Bucher, T., Thalmann, F., and Isambert, H. 2003. Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations. *Proc. Natl. Acad. Sci.* **100:** 15310–15315.

Zuker, M. and Stiegler, P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **9:** 133–148. doi: 10.1093/nar/9.1.133.