ProbKnot: Fast prediction of RNA secondary structure including pseudoknots

STANISLAV BELLAOUSOV[1,2] and DAVID H. MATHEWS[1,2,3]

[1]Department of Biochemistry and Biophysics, University of Rochester Medical Center, Rochester, New York 14642, USA
[2]Center for RNA Biology, University of Rochester Medical Center, Rochester, New York 14642, USA
[3]Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, New York 14642, USA

ABSTRACT

It is a significant challenge to predict RNA secondary structures including pseudoknots. Here, a new algorithm capable of predicting pseudoknots of any topology, ProbKnot, is reported. ProbKnot assembles maximum expected accuracy structures from computed base-pairing probabilities in $O(N^2)$ time, where N is the length of the sequence. The performance of ProbKnot was measured by comparing predicted structures with known structures for a large database of RNA sequences with fewer than 700 nucleotides. The percentage of known pairs correctly predicted was 69.3%. Additionally, the percentage of predicted pairs in the known structure was 61.3%. This performance is the highest of four tested algorithms that are capable of pseudoknot prediction. The program is available for download at: http://rna.urmc.rochester.edu/RNAstructure.html.

Keywords: RNA partition function; RNA folding; RNA structure prediction

INTRODUCTION

There is a diverse world of functional RNA sequences. Originally in the central dogma of biology, RNA was considered to play a transient role in expressing inherited information as proteins. It was later discovered that, besides this role in generating proteins, RNA has a variety of other functions, such as regulating gene expression (Tucker and Breaker 2005; Storz and Gottesman 2006; Wu and Belasco 2008), catalyzing reactions (Nissen et al. 2000; Doudna and Cech 2002), and trafficking proteins (Walter and Blobel 1982). RNA sequences that do not code for proteins are referred to as noncoding RNA, or ncRNA (Eddy 2001). Many of these ncRNA sequences have well-defined structures, and to understand how these ncRNA sequences perform their functions it is important to know their structure.

Determination of RNA structure is challenging. Primary structure is an ordered sequence of nucleotides. Secondary structure consists of canonical base pairs, i.e., AU, GC, and GU pairs. Secondary structure prediction involves predicting the base pairs that occur in a specified sequence of nucleotides. RNA tertiary structure is the three-dimensional arrangement of atoms. Because RNA structure is generally hierarchical, the secondary structure can be largely determined without knowing the tertiary structure (Tinoco and Bustamante 1999).

Many secondary structure prediction methods are available. The most accurate method is comparative sequence analysis (Pace et al. 1999), which determines base pairs conserved among homologous sequences. The method is highly accurate (Gutell et al. 2002) but requires a large number of homologous sequences and significant human insight, and thus is limited in use. When a single sequence is available, the most popular approach for structure prediction is to predict the lowest free energy structure with a dynamic programming algorithm (Zuker 2003; Mathews et al. 2004; Mathews and Turner 2006; Gruber et al. 2008).

A more recent approach to predict RNA secondary structures is called maximum expected accuracy structure prediction (Knudsen and Hein 2003; Do et al. 2006; Hamada et al. 2009; Lu et al. 2009). Roughly, maximum expected accuracy structures are structures composed of pairs that provide the maximal sum of pairing probabilities. The pairing probabilities can be derived by machine learning methods or by thermodynamic methods using partition functions. Maximum expected accuracy structures have improved accuracy compared with free energy minimization because it has been observed that highly probable base pairs are more likely to be correctly predicted pairs (Mathews 2004).

Reprint requests to: David H. Mathews, Department of Biochemistry and Biophysics, University of Rochester Medical Center, 601 Elmwood Avenue, Box 712, Rochester, NY 14642, USA; e-mail: David_Mathews@urmc.rochester.edu; fax: (585) 275-6007.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.2125310.

RNA (2010), 16:1870–1880. Published by Cold Spring Harbor Laboratory Press. Copyright © 2010 RNA Society.

One important topology for RNA secondary structures is a pseudoknot. This is a type of secondary structure that contains nonnested base pairs. Specifically, a pseudoknot is defined by at least two base pairs, i–j and i′–j′, such that nucleotide i appears before i′, i′ before j, and j before j′ in the sequence. Base pairs in pseudoknots represent a small fraction of base pairs in known RNA secondary structures, but pseudoknots occur in a number of functional RNA sequences (van Batenburg et al. 2001; Condon and Jabbari 2009).

The prediction of secondary structures including pseudoknots is a difficult task. For example, the most popular dynamic programming algorithms for finding low free energy structures do not allow pseudoknots. This allows those dynamic programming algorithms to run quickly and scale well, i.e., O(N^3) in time where N is the length of the sequence. Including pseudoknots in the structure prediction requires higher-order scaling, the use of heuristics, and/or a compromise on the energy model.

It has been proven that the prediction of lowest free energy secondary structures with pseudoknots is NP-hard (Lyngsø and Pederson 2000). In spite of this, a number of innovative and practical approaches have been developed to predict structures with pseudoknots. These approaches can roughly be summarized in six categories. One approach is to use a dynamic programming algorithm to predict structures with a limited topology (Rivas and Eddy 1999; Uemura et al. 1999; Akutsu 2000; Dirks and Pierce 2003; Reeder and Giegerich 2004). A classification of topologies and an explanation of topologies handled by several dynamic programming algorithms are available (Condon et al. 2004). A second approach to predicting pseudoknots is to construct structures using multiple iterations of algorithms that would otherwise not be capable of predicting pseudoknots (Ruan et al. 2004; Ren et al. 2005; Jabbari et al. 2008). One of these algorithms is also capable of using an alignment of multiple homologous sequences to improve its accuracy by finding a consensus structure (Ruan et al. 2004). A third approach is to either simulate a folding pathway or sample structures with a stepwise addition of helices (Abrahams et al. 1990; Gultyaev et al. 1995; Isambert and Siggia 2000; Dawson et al. 2007; Meyer and Miklos 2007). A fourth approach uses the maximum weight matching algorithm to construct structures composed of pairs that give a maximum score (Tabaska et al. 1998; Witwer et al. 2004). These algorithms use alignments of multiple homologous sequences and scoring functions that summarize free energies associated with pairs and covariation of pairs. Recently, a sixth approach, using constrained integer programming has also been applied to finding lowest free energy structures (Poolsap et al. 2009).

Many of the above algorithms use rules for predicting the free energy change—i.e., stability—of pseudoknots. Significant progress has been reported in this area as well using several approaches. Two sets of empirical rules were designed for use with dynamic programming algorithms (Rivas and Eddy 1999; Dirks and Pierce 2003). A set of parameters was developed using polymer theory and calibrated to experimentally measured stabilities (Wyatt et al. 1990; Nixon and Giedroc 1998; Theimer et al. 1998; Gultyaev et al. 1999; Theimer and Giedroc 1999, 2000). Another set of parameters was developed using lattice models and self-avoidant walks (Cao and Chen 2006, 2009). Additionally, a set of parameters was developed using polymer theory (Aalberts and Hodas 2005). A recent report provides a technique for refining parameters for predicting pseudoknot stability that utilizes experimental data and the database of sequences with known structure (Andronescu et al. 2010).

This contribution reports ProbKnot, a new RNA secondary structure prediction algorithm that is capable of predicting RNA secondary structures of any topology in O(N^3) time. Base-pair probabilities are first predicted using a partition function (Mathews 2004), which does not include pseudoknotted structures, in O(N^3) time (Xia et al. 1998; Mathews et al. 2004). ProbKnot then assembles a type of maximum expected accuracy structure in O(N^2) time from the base-pairing probabilities, but does so without using a dynamic programming algorithm. By assembling structures from base-pair probabilities determined without pseudoknots, ProbKnot does not require a set of rules for predicting the stability of pseudoknots.

The performance of ProbKnot was benchmarked against other freely available programs that predict pseudoknots: pknotsRG-mfe (Reeder and Giegerich 2004), ILM (Ruan et al. 2004), and HotKnots (Ren et al. 2005); and programs that do not predict pseudoknots: MaxExpect, a maximum expected accuracy approach (Lu et al. 2009) and Free Energy Minimization (Mathews et al. 2004). ProbKnot was able to predict the largest fraction of known base pairs.

RESULTS

ProbKnot

ProbKnot is an algorithm that predicts RNA secondary structure by finding the structure with the most probable base pairs. It assembles structures composed of base pairs, i–j, where the probability of the i–j pair is higher than any i–k or j–k base pair, where k is any other nucleotide in the sequence. This is performed in O(N^2) time by first calculating and storing the pairing probability of the most probable pair for each nucleotide, $P_{max(i)}$. Then each base pair is considered for inclusion in the structure. If the probability of the i–j pair is equal to both $P_{max(i)}$ and $P_{max(j)}$, that pair is included in the structure. The algorithm is additionally capable of supporting multiple iterations in a similar manner by finding the most probable i–j pair for nucleotides that remained unpaired after previous iterations. For benchmarks shown here, only a single iteration was performed.

As a post-processing step, after the structure is obtained, the algorithm removes helices composed of two or fewer stacked base pairs. For these calculations, single nucleotide bulges are considered stacked and therefore do not interrupt helical stacking. So, for example, two pairs separated by a single bulged nucleotide would be considered stacked.

Structure prediction accuracy

The accuracy of ProbKnot was evaluated by predicting structures for sequences with known structure as determined by comparative sequence analysis. Both sensitivity and positive predictive value (PPV) were determined. Sensitivity is the percent of known pairs correctly predicted and PPV is the percent of predicted pairs in the known structure.

For a diverse set of sequences with known secondary structure, ProbKnot had an average of 69.3% sensitivity (Table 1). The performance was compared against three other programs that are capable of pseudoknot prediction and freely available for download. These programs were demonstrated to be among the top performers in structure prediction accuracy in a previous benchmark (Ren et al. 2005). The programs are ILM version 1.0 (Ruan et al. 2004), HotKnots version 1.2 (Ren et al. 2005), and pknotsRG version 1.3 (Reeder and Giegerich 2004). Each was run using default parameters. Additionally, the performance was compared against two other algorithms from RNAstructure, which predicts structures without pseudoknots, free energy minimization (Mathews et al. 2004), and maximum expected accuracy structure prediction (Lu et al. 2009). Overall, ProbKnot had the highest average sensitivity for all methods and the highest PPV among methods that are capable of predicting pseudoknots.

ProbKnot had an average PPV of 61.3% (Table 2), performing best in six out of 10 RNA families including two families with pseudoknots, and performing on the same level with pknotsRG-mfe on the group I intron family that is also known to have pseudoknots. This was the best performance among algorithms that predict pseudoknots, but not as high as MaxExpect, which does not predict pseudoknots. This is consistent with previous observations. Algorithms that predict pseudoknots consider a larger space of possible structures, which leads to a tendency for lower fidelity of structure prediction.

Pseudoknot prediction

The accuracy of pseudoknot prediction was evaluated separately. First, the number of predicted pseudoknotted pairs was tabulated (Table 3). These pairs were found using the method of Smit et al. (2008) to identify the fewest pairs that need to be removed to remove the pseudoknots from a structure. The number of pseudoknotted pairs reported in Table 3 is the sum of the number of pairs that are removed to remove the pseudoknot. Then, the number of these predicted pseudoknotted pairs that are both in the known structure and pseudoknotted in the known structure was determined and reported as the number of correctly predicted pseudoknotted pairs (Table 3).

The accuracy of pseudoknot prediction for structures was also tabulated (Table 4). The number of predicted structures with pseudoknotted pairs was determined. The number of the predicted structures with pseudoknots that were correct was then tabulated as correctly predicted pseudoknotted structures. A predicted pseudoknotted structure was considered correct if it contained at least one correctly predicted pseudoknotted pair. For structures with multiple pseudoknots, such as tmRNA sequences, a structure can be considered correctly predicted if only one pseudoknot is correctly predicted.

ILM has the highest number of correctly predicted pseudoknotted structures and the highest number of correctly predicted pseudoknotted base pairs. Of the predicted pseudoknotted pairs, pknotsRG-mfe has the highest portion of correctly predicted pairs. All algorithms, however, correctly predict only a small fraction of the pseudoknotted base pairs that are in the known structure.

Structure prediction example

Figure 1 shows an example of predicted structure, the *Tetrahymena thermophila* group I intron structure predicted by ProbKnot. Thick lines between the base pairs represent correctly predicted pairs. As can be seen from Figure 1, ProbKnot correctly predicts almost all base pairs with probabilities >70%. Helixes S_1 and S_2 that form pseudoknots in the structure are correctly predicted by ProbKnot.

Time benchmarks

Time trials were performed on sequences ranging in length from 77 to 2904 nucleotides (nt) (Table 5). On the longest sequence, ProbKnot showed the second best time performance, requiring 63 min of runtime to predict both the base-pair probabilities and to assemble the predicted structure. ILM had the best time performance and the dynamic programming algorithm (pknotsRG-mfe) had the slowest time performance.

DISCUSSION

ProbKnot assembles maximum expected accuracy structures using base-pairing probabilities determined from a partition function calculation. Previous approaches for predicting maximum expected accuracy structures used dynamic programming algorithms that do not allow pseudoknots (Do et al. 2006; Hamada et al. 2009; Lu et al. 2009), but ProbKnot is not limited in the topology of structures it can predict. Although the partition function algorithm does not account for pseudoknotted structures, each of the helices in the pseudoknot can occur in different structures (Mathews 2004). ProbKnot takes advantage of this fact to assemble both helices into a single structure.

ProbKnot has some similarities with the maximum weight matching (MWM) methods previously explored to find

TABLE 1. Sensitivities of prediction methods

Type of RNA	Sequences	Base pairs	Pseudoknotted base pairs	ProbKnot (%)	ILM (%)	pknotsRG-mfe (%)	HotKnots (%)	MaxExpect	Free energy minimization
SSU rRNA	88	8749	63	62.2 ± 21.9	61.2 ± 24.4	62.6 ± 25.3	**65.7 ± 25.0**	62.1 ± 23.1	61.4 ± 23.7
	(22)	(8861)	(127)	(47.1 ± 14.3)	(47.2 ± 15.0)	**(47.9 ± 14.5)**	—	(47.2 ± 14.1)	(45.5 ± 14.8)
LSU rRNA	27	3444	6	72.8 ± 12.0	70.6 ± 12.8	68.0 ± 13.5	68.6 ± 11.3	**74.6 ± 11.9**	72.4 ± 17.2
	(5)	(3585)	(6)	(56.2 ± 13.9)	(54.5 ± 12.9)	(51.4 ± 16.0)	—	**(56.6 ± 14.7)**	(55.0 ± 11.2)
5S rRNA	309	10188	0	72.7 ± 25.5	**77.8 ± 21.4**	75.9 ± 24.0	75.2 ± 24.8	72.5 ± 26.4	72.9 ± 26.6
Group I intron	16	1532	91	**72.3 ± 13.6**	70.9 ± 15.7	70.1 ± 18.5	66.8 ± 17.2	71.2 ± 13.9	70.2 ± 13.6
Group II intron	3	503	0	**89.2 ± 2.6**	85.5 ± 0.6	81.7 ± 3.6	81.2 ± 17.5	87.0 ± 5.0	88.1 ± 2.2
RNase P	6	692	68	64.2 ± 16.3	59.6 ± 14.8	51.0 ± 11.5	52.9 ± 13.3	63.5 ± 15.4	**64.6 ± 12.9**
SRP RNA	91	6273	111	66.2 ± 26.1	**72.9 ± 22.5**	70.5 ± 23.5	72.0 ± 22.9	65.9 ± 26.3	68.9 ± 25.4
tRNA	484	10018	0	**88.2 ± 18.0**	77.1 ± 22.3	75.1 ± 24.0	74.9 ± 22.4	85.8 ± 17.9	85.6 ± 19.6
tmRNA	462	45332	10035	**47.2 ± 14.7**	43.7 ± 15.1	42.9 ± 15.7	43.3 ± 15.3	46.0 ± 14.5	45.9 ± 14.3
Telomerase RNA	37	3774	330	57.9 ± 16.0	52.3 ± 17.8	57.5 ± 20.1	54.5 ± 20.8	58.3 ± 15.3	**59.2 ± 16.9**
Total/average	1523	90505	10704	**69.3 ± 16.7**	67.2 ± 16.8	65.5 ± 18.0	65.5 ± 19.0	68.7 ± 17.0	68.9 ± 17.2

Sequences in SSU rRNA and LSU rRNA families were split into domains of no larger than 700 nt (Mathews et al. 1999). The prediction results for full-length sequences shown in parentheses are not considered in the total/average calculation. Underlined results represent the best performance out of all the algorithms that predict pseudoknots. Bold results represent absolute best performance. For a given RNA family, the reported performance is the mean and standard deviation of sensitivity for all sequences of that family. The average sensitivity over all families of RNA is the mean and standard deviation of performance on RNA families.

TABLE 2. Positive predictive value of prediction methods

Type of RNA	Sequences	Base pairs	Pseudoknotted base pairs	ProbKnot (%)	ILM (%)	pknotsRG-mfe (%)	HotKnots (%)	MaxExpect	Free energy minimization
SSU rRNA	88	8749	63	56.8 ± 23.7	56.4 ± 26.3	56.8 ± 26.1	**59.8 ± 26.3**	58.0 ± 25.0	54.8 ± 25.3
	(22)	(8861)	(127)	(41.5 ± 15.0)	(40.3 ± 15.3)	(41.5 ± 15.2)	—	**(42.7 ± 14.7)**	(38.3 ± 14.5)
LSU rRNA	27	3444	6	66.5 ± 11.3	64.4 ± 14.1	62.2 ± 13.4	62.2 ± 11.4	**68.4 ± 11.6**	65.0 ± 16.3
	(5)	(3585)	(6)	50.0 ± 14.7	(47.0 ± 13.5)	(44.8 ± 15.3)	—	**(51.6 ± 14.2)**	(47.0 ± 11.6)
5S rRNA	309	10188	0	66.3 ± 24.3	**71.9 ± 21.1**	67.1 ± 21.4	66.2 ± 21.8	65.3 ± 23.6	64.0 ± 23.8
Group I intron	16	1532	91	64.0 ± 14.5	60.8 ± 13.7	64.0 ± 15.8	61.0 ± 16.7	**68.0 ± 15.1**	63.4 ± 13.5
Group II intron	3	503	0	80.8 ± 9.8	78.7 ± 7.8	78.6 ± 8.6	77.9 ± 16.1	**84.9 ± 9.4**	82.7 ± 6.8
RNase P	6	692	68	62.5 ± 17.1	56.8 ± 14.1	48.7 ± 9.1	50.3 ± 11.6	**62.7 ± 15.3**	61.8 ± 12.0
SRP RNA	91	6273	111	50.2 ± 21.5	56.5 ± 20.7	55.5 ± 21.8	**56.7 ± 21.4**	51.3 ± 22.1	52.9 ± 22.2
tRNA	484	10018	0	80.6 ± 18.4	72.3 ± 22.9	74.5 ± 26.0	72.0 ± 24.3	**84.9 ± 19.6**	83.6 ± 22.2
tmRNA	462	45332	10035	42.7 ± 13.8	37.7 ± 13.2	38.6 ± 14.4	39.2 ± 14.3	**44.2 ± 14.5**	41.5 ± 13.9
Telomerase RNA	37	3774	330	43.0 ± 13.4	37.4 ± 13.7	41.8 ± 15.2	39.3 ± 15.3	**43.4 ± 13.4**	42.4 ± 13.7
Total/average	1523	90505	10704	61.3 ± 16.8	59.3 ± 16.8	58.8 ± 17.2	58.5 ± 17.9	**63.1 ± 16.9**	61.2 ± 17.0

Sequences in SSU rRNA and LSU rRNA families were split into domains of no larger than 700 nt. The prediction results for full-length sequences shown in parentheses are not considered in the total/average calculation. Underlined results represent the best performance out of all the algorithms that predict pseudoknots. Bold results represent absolute best performance. For a given RNA family, the reported performance is the mean and standard deviation PPV for all sequences of that family. The average sensitivity over all families of RNA is the mean and standard deviation of performance of RNA families.

TABLE 3. Base-pair statistics: Evaluation of methods in terms of predicted pseudoknotted base pairs

Type of RNA	Nucleotides	Base pairs	Pseudo-knotted pairs	Pseudoknotted pairs predicted				Correctly predicted pseudoknotted pairs			
				ProbKnot	ILM	pknotsRG-mfe	HotKnots v1.2	ProbKnot	ILM	pknotsRG-mfe	HotKnots v1.2
SSU rRNA	33263	8749	63	168	297	141	8	0	5	0	0
	(33263)	(8861)	(127)	(192)	(305)	(45)	—	(0)	(0)	(0)	—
LSU rRNA	12437	3444	6	33	48	4	0	0	0	0	0
	(13341)	(3585)	(6)	(113)	(97)	(4)	—	(0)	(0)	(0)	—
5S rRNA	36925	10188	0	100	103	21	79	0	0	0	0
Group I intron	5518	1532	91	52	64	60	0	4	7	10	0
Group II intron	2006	503	0	8	16	0	0	0	0	0	0
RNase P	2269	692	68	14	14	0	0	3	5	0	0
SRP RNA	24383	6273	111	85	127	150	0	4	5	5	0
tRNA	37502	10018	0	257	349	42	223	0	0	0	0
tmRNA	169099	45332	10035	1683	2429	396	22	280	300	253	14
Telomerase RNA	16452	3774	330	46	162	114	0	0	0	9	0
Total	339854	90505	10704	2446	3609	928	332	291	322	277	14

Pseudoknotted pairs predicted is the sum of pairs removed using the Smit et al. (2008) method. Correctly predicted pseudoknotted pairs is the sum of pairs identified as pseudoknotted that are in the known structure and pseudoknotted. Sequences in SSU rRNA and LSU rRNA subtypes were split into domains of no larger than 700 nt. In parentheses are sums for small and large subunit rRNAs when the whole sequence is folded at once and these sums are not used in the total.

TABLE 4. Pseudoknot prediction statistics on structures

			Pseudoknotted structures predicted				Correctly predicted pseudoknotted structures			
Type of RNA	Sequences	Pseudoknotted structures	ProbKnot	ILM	pknotsRG-mfe	HotKnots v1.2	ProbKnot	ILM	pknotsRG-mfe	HotKnots v1.2
SSU rRNA	88	21	34	40	26	2	0	2	0	0
	(22)	(22)	(18)	(21)	(9)	—	(0)	(0)	(0)	—
LSU rRNA	27	2	7	10	1	0	0	0	0	0
	(5)	(2)	(5)	(5)	(1)	—	(0)	(0)	(0)	—
5S rRNA	309	0	26	25	4	17	0	0	0	0
Group I intron	16	16	10	10	10	0	1	2	2	0
Group II intron	3	0	1	2	0	0	0	0	0	0
RNase P	6	6	4	4	0	0	1	1	0	0
SRP RNA	91	23	21	25	34	0	1	2	1	0
tRNA	484	0	75	99	10	54	0	0	0	0
tmRNA	462	459	276	313	54	2	65	64	39	2
Telomerase RNA	37	37	12	23	19	0	0	0	1	0
Total	1523	564	466	551	158	75	68	71	43	2

Pseudoknotted structures predicted is the sum of predicted structures that contain at least one pseudoknotted pair. Correctly predicted pseudoknotted structures is the sum of structures with at least one pseudoknotted pair that is correctly predicted. Sequences in SSU rRNA and LSU rRNA subtypes were split into domains of no larger than 700 nt. In parentheses are sums for small and large subunit rRNAs when the whole sequence is folded at once and these sums are not used in the total.

secondary structures conserved among multiple sequences (Tabaska et al. 1998; Hofacker et al. 2004). The MWM algorithm takes pairing weights as input, where weights are a function of folding free energy change and covariation, and outputs a structure with the greatest sum of these weights. MWM runs in O(N^3) time and is also not limited in topology. It has been noted that MWM methods tend to have poor PPV because the structures are saturated with pairs, but post-processing can remove pairs and improve performance. ProbKnot is distinct because it uses pair probabilities and not folding free energy changes as input. Additionally, the requirement that the pairs included in the structure be the highest pairing probability for pairs possible by either nucleotide provides a stopping rule so that structures are not oversaturated with pairs.

Based on the benchmarks in Tables 1 and 2, ProbKnot has the highest average accuracy for RNA secondary structure prediction among algorithms that predict pseudoknots. It performs on average 2%–4% better in sensitivity and 2%–3% better in PPV. These improvements are considerable, but they leave room for improvement. For example, the average performance for structure prediction on tmRNA, with four pseudoknots, is only 47.2% in sensitivity.

The performance results for ProbKnot were also compared with the performance of two algorithms, MaxExpect (Lu et al. 2009) and free energy minimization (Mathews et al. 2004), which are unable to predict pseudoknots. This comparison was performed to evaluate the benefit for increasing the range of topologies predicted to include pseudoknots. In sensitivity, ProbKnot outperformed both algorithms by ~0.5%–1%. This was expected because ProbKnot has a wider predicting range of possible topologies, and thus it should predict more correct base pairs than other algorithms. Because of the wider range of possible prediction topologies, however, there is wider latitude for incorrectly predicting base pairs and, because of this, PPV decreases compared with MaxExpect.

Given the poor performance of the methods benchmarked here on tmRNA and telomerase RNA, including ProbKnot, there is a need for continued research in predicting pseudoknotted structures. One possible approach for improving ProbKnot is to use a partition function that explicitly includes pseudoknots to predict the base-pairing probabilities. For example, the algorithm reported by Dirks and Pierce is O(N^4) in time and includes a restricted set of pseudoknots (Dirks and Pierce 2003, 2004; Condon et al. 2004). These pair probabilities could be used by ProbKnot to assemble structures of any topology and may yield more accurate structures.

ProbKnot is available in the RNAstructure package (Reuter and Mathews 2010). This includes the source code in C++; text interfaces for Linux, Unix, and Windows; a JAVA graphical interface for Linux and Mac OS-X; and a graphical interface for Microsoft Windows.

MATERIALS AND METHODS

Prediction of base-pairing probabilities

Base-pair probabilities were predicted using a partition function algorithm that includes coaxial stacking (Mathews 2004). This

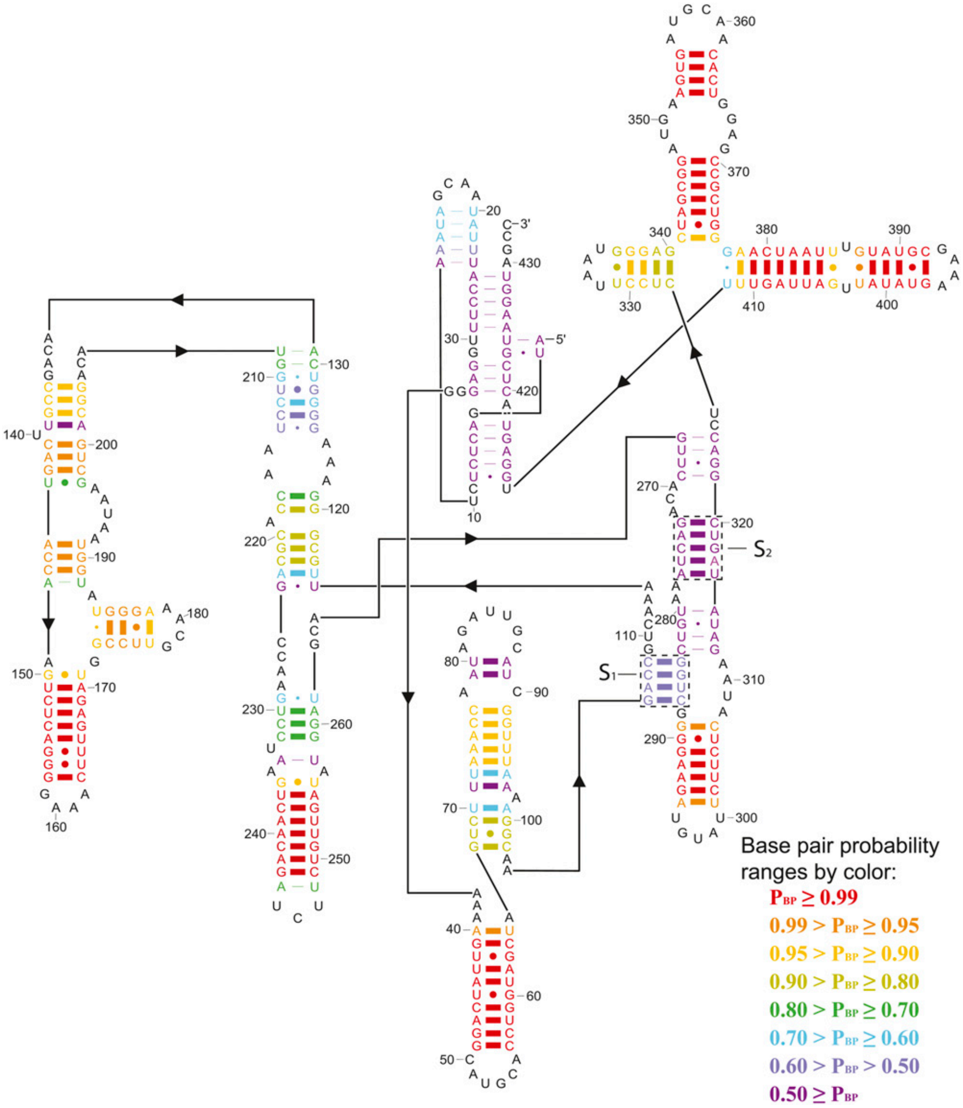

FIGURE 1. Predicted secondary structure of group I intron from *T. thermophila* by ProbKnot. Thick lines represent correctly predicted base pairs; thin lines represent incorrectly predicted base pairs. The boxed helices, labeled S_1 and S_2, are the two helices that define the pseudoknot.

program uses the thermodynamic parameters assembled by Xia et al. (1998) and Mathews et al. (2004) to predict the stabilities of secondary structures. Similar to Lu et al. (2009), the multibranch loop parameter bonus for each branching helix was not optimized as done by Mathews et al. (2004) but was kept at −0.6 kcal/mol, the value suggested by optical melting experiments (Diamond et al. 2001; Mathews and Turner 2002).

Accuracy

All algorithms were tested on 1550 RNA sequences from 10 different families: small subunit rRNA (Gutell 1994), large subunit rRNA (Gutell et al. 1993; Schnare et al. 1996), 5S rRNA (Szymanski et al. 1998), group I intron (Waring and Davies 1984; Damberger and Gutell 1994), group II intron (Michel et al. 1989), RNase P RNA (Brown 1998), SRP RNA (Larsen et al. 1998), tRNA (Sprinzl et al. 1998), tmRNA (Zwieb et al. 1999), and telomerase RNA (Chen et al. 2000). This database is an expansion of a database of structures assembled previously (Mathews et al. 1999) to include the telomerase RNA and the tmRNA, which are pseudoknotted RNA structures. Vertebrate telomerase RNA secondary structure alignments were obtained from the Rfam 9.1 database (Griffiths-Jones et al. 2003, 2005; Daub et al. 2008; Gardner et al. 2009). tmRNA secondary structures were obtained from the tmRDB database (Zwieb et al. 2003). Structures with unknown nucleotides were omitted from the full list of structures in the tmRDB database. Small and large subunit rRNA sequences were divided into domains of ≤700 nt as previously reported (Mathews et al. 1999).

The performance of secondary structure prediction algorithms was evaluated by calculating sensitivity and PPV. Sensitivity measures the percent of known base pairs correctly predicted:

$$\text{Sensitivity} = \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{Number of false negatives}}.$$

TABLE 5. Comparison of time performances of different algorithms

				Saccharomyces		
		Bacillus	*T. thermophila*	*cerevisiae*	*E. coli*	*E. coli*
RNA	*E. coli*	*subtilis*	IVS LSU	A5	small subunit	large subunit
type	arginine tRNA	SRP	group I intron	group II intron	rRNA	rRNA
Length (nt)	77	268	433	631	1542	2904
ProbKnot	0 min 0.1 sec	0 min 2.8 sec	0 min 16.6 sec	0 min 39.3 sec	9 min 13.6 sec	63 min 18.6 sec
HotKnot	0 min 1.3 sec	2 min 3.3 sec	5 min 25.3 sec	8 min 24.8 sec	37 min 11.7 sec	NA[a]
ILM	0 min 0.03 sec	0 min 0.5 sec	0 min 1.8 sec	0 min 8.1 sec	2 min 27.4 sec	35 min 31.6 sec
PknotsRG-mfe	0 min 0.04 sec	0 min 3.1 sec	0 min 18.9 sec	1 min 34.2 sec	59 min 33.6 sec	783 min 9.5 sec
MaxExpect	0 min 0.1 sec	0 min 3.1 sec	0 min 12.0 sec	0 min 40.4 sec	9 min 33.4 sec	66 min 46.1 sec
Free energy minimization	0 min 0.1 sec	0 min 2.4 sec	0 min 9.6 sec	0 min 31.1 sec	7 min 9.9 sec	71 min 49.7 sec

[a]HotKnots did not run using the available resources on the *E. coli* large subunit rRNA. Time calculations were performed on a machine with an Intel Core2 Quad Q6600 processor and 4GB of RAM, running the Ubuntu 8.10 operating system and gcc compiler version 4.3.2. Time results were obtained running the ''time'' command.

PPV measures percent of predicted base pairs that are correctly predicted:

$$\text{PPV} = \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{Number of false positives}}.$$

Both sensitivity and PPV were evaluated with an allowance for incomplete knowledge of the exact pairing in the known structure. A predicted base pair between nucleotides i and j was considered correctly predicted if i was paired to j, j − 1, or j + 1, or if j was paired to i − 1 or i + 1 (Mathews et al. 1999). Average values were calculated per RNA family and then overall averages were calculated as the mean of the values reported for each family.

Tabulation of pseudoknot content

The number of base pairs in pseudoknots was counted using an implementation of the optimization approach of Smit et al. (2008) as implemented in the RNA class component of RNAstructure (Reuter and Mathews 2010). In this implementation, the scoring function is pairs, so the algorithm counts the fewest number of pairs that would need to be removed to remove the pseudoknot.

ACKNOWLEDGMENTS

We thank A.O. Harmanci, T.D. Romo, G. Sharma, and R. Tyagi for discussions. This work was supported by National Institutes of Health grant no. R01GM076485 to D.H.M. Computer time was provided by the University of Rochester Center for Research Computing.

Received February 11, 2010; accepted June 22, 2010.

REFERENCES

Aalberts DP, Hodas NO. 2005. Asymmetry in RNA pseudoknots: Observation and theory. *Nucleic Acids Res* **33:** 2210–2214.

Abrahams JP, van den Berg M, van Batenburg E, Pleij C. 1990. Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucleic Acids Res* **18:** 3035–3044.

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.

Brown JW. 1998. The ribonuclease P database. *Nucleic Acids Res* **26:** 351–352.

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 JL, Blasco MA, Greider CW. 2000. Secondary structure of vertebrate telomerase RNA. *Cell* **100:** 503–514.

Condon A, Jabbari H. 2009. Computational prediction of nucleic acid secondary structure: Methods, applications, and challenges. *Theor Comput Sci* **410:** 294–301.

Condon A, Davy B, Rastegari B, Tarrant F, Zhao S. 2004. Classifying RNA pseudoknotted structures. *Theor Comput Sci* **320:** 35–50.

Damberger SH, Gutell RR. 1994. A comparative database of group I intron structures. *Nucleic Acids Res* **22:** 3508–3510.

Daub J, Gardner PP, Tate J, Ramsköld D, Manske M, Scott WG, Weinberg Z, Griffiths-Jones S, Bateman A. 2008. The RNA WikiProject: Community annotation of RNA families. *RNA* **14:** 2462–2464.

Dawson WK, Fujiwara K, Kawai G. 2007. Prediction of RNA pseudoknots using heuristic modeling with mapping and sequential folding. *PLoS ONE* **2:** e905. doi: 10.1371/journal.pone.0000905.

Diamond JM, Turner DH, Mathews DH. 2001. Thermodynamics of three-way multibranch loops in RNA. *Biochemistry* **40:** 6971–6981.

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

Dirks RM, Pierce NA. 2004. An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. *J Comput Chem* **25:** 1295–1304.

Do CB, Woods DA, Batzoglou S. 2006. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **22:** e90–e98.

Doudna JA, Cech TR. 2002. The chemical repertoire of natural ribozymes. *Nature* **418:** 222–228.

Eddy SR. 2001. Noncoding RNA genes and the modern RNA world. *Natl Rev* **2:** 919–929.

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.

Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. 2003. Rfam: An RNA family database. *Nucleic Acids Res* **31:** 439–441.

Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. 2005. Rfam: Annotating noncoding RNAs in complete genomes. *Nucleic Acids Res* **33:** D121–D124.

Gruber AR, Lorenz R, Bernhart SH, Neubock R, Hofacker IL. 2008. The Vienna RNA websuite. *Nucleic Acids Res* **36:** W70–W74.

Gultyaev AP, van Batenburg FHD, Pleij CWA. 1995. The computer simulation of RNA folding pathways using a genetic algorithm. *J Mol Biol* **250:** 37–51.

Gultyaev AP, van Batenburg FHD, Pleij CWA. 1999. An approximation of loop free energy values of RNA H-pseudoknots. *RNA* **5:** 609–617.

Gutell RR. 1994. Collection of small subunit (16S- and 16S-like) ribosomal RNA structures. *Nucleic Acids Res* **22:** 3502–3507.

Gutell RR, Gray MW, Schnare MN. 1993. A compilation of large subunit (23S- and 23S-like) ribosomal RNA structures. *Nucleic Acids Res* **21:** 3055–3074.

Gutell RR, Lee JC, Cannone JJ. 2002. The accuracy of ribosomal RNA comparative structure models. *Curr Opin Struct Biol* **12:** 301–310.

Hamada M, Kiryu H, Sato K, Mituyama T, Asai K. 2009. Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics* **25:** 465–473.

Hofacker IL, Priwitzer B, Stadler PF. 2004. Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics* **20:** 186–190.

Isambert H, Siggia ED. 2000. Modeling RNA folding paths with pseudoknots: Application to hepatitis delta virus ribozyme. *Proc Natl Acad Sci* **97:** 6515–6520.

Jabbari H, Condon A, Zhao S. 2008. Novel and efficient RNA secondary structure prediction using hierarchical folding. *J Comput Biol* **15:** 139–163.

Knudsen B, Hein J. 2003. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res* **31:** 3423–3428.

Larsen N, Samuelsson T, Zwieb C. 1998. The signal recognition particle database (SRPDB). *Nucleic Acids Res* **26:** 177–178.

Lu ZJ, Gloor JW, Mathews DH. 2009. Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA* **15:** 1805–1813.

Lyngsø RB, Pederson CN. 2000. RNA pseudoknot prediction in energy-based models. *J Comput Biol* **7:** 409–427.

Mathews DH. 2004. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA* **10:** 1178–1190.

Mathews DH, Turner DH. 2002. Experimentally derived nearest neighbor parameters for the stability of RNA three- and four-way multibranch loops. *Biochemistry* **41:** 869–880.

Mathews DH, Turner DH. 2006. Prediction of RNA secondary structure by free energy minimization. *Curr Opin Struct Biol* **16:** 270–278.

Mathews DH, Sabina J, Zuker M, Turner DH. 1999. Expanded sequence dependence of thermodynamic parameters provides improved prediction of RNA secondary structure. *J Mol Biol* **288:** 911–940.

Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci* **101:** 7287–7292.

Meyer IM, Miklos I. 2007. SimulFold: Simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. *PLoS Comput Biol* **3:** e149. doi: 10.1371/journal.pcbi.0030149.

Michel F, Umesono K, Ozeki H. 1989. Comparative and functional anatomy of group II catalytic introns—a review. *Gene* **82:** 5–30.

Nissen P, Hansen J, Ban N, Moore PB, Steitz TA. 2000. The structural basis of ribosomal activity in peptide bond synthesis. *Science* **289:** 920–930.

Nixon PL, Giedroc DP. 1998. Equilibrium unfolding (folding) pathway of a model H-type pseudoknotted RNA: The role of magnesium ions in stability. *Biochemistry* **37:** 16116–16129.

Pace NR, Thomas BC, Woese CR. 1999. Probing RNA structure, function, and history by comparative analysis. In *The RNA world*, 2nd ed (ed. RF Gesteland et al.), pp.113–141. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Poolsap U, Kato Y, Akutsu T. 2009. Prediction of RNA secondary structure with pseudoknots using integer programming. *BMC Bioinformatics* (Suppl 1) **10:** S38. doi: 10.1186/1471-2105-10-S1-S38.

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.

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

Reuter JS, Mathews DH. 2010. RNAstructure: Software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11:** 129. doi: 10.1186/1471-2105-11-129.

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.

Schnare MN, Damberger SH, Gray MW, Gutell RR. 1996. Comprehensive comparison of structural characteristics in Eukaryotic cytoplasmic large subunit (23S-like) ribosomal RNA. *J Mol Biol* **256:** 701–719.

Smit S, Rother K, Heringa J, Knight R. 2008. From knotted to nested RNA structures: A variety of computational methods for pseudoknot removal. *RNA* **14:** 410–416.

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.

Storz G, Gottesman S. 2006. Versatile roles of small RNA regulators in bacteria. In *The RNA world*, 3rd ed. (ed. RF Gesteland et al.), pp.567–594. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Szymanski M, Specht T, Barciszewska MZ, Barciszewski J, Erdmann VA. 1998. 5S rRNA data bank. *Nucleic Acids Res* **26:** 156–159.

Tabaska JE, Cary RB, Gabow HN, Stormo GD. 1998. An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics* **14:** 691–699.

Theimer CA, Giedroc DP. 1999. Equilibrium unfolding pathway of an H-Type RNA pseudoknot which promotes programmed −1 ribosomal frameshifting. *J Mol Biol* **289:** 1283–1299.

Theimer CA, Giedroc DP. 2000. Contribution of the intercalated adenosine at the helical junction to the stability of the gag-pro frameshifting pseudoknot from mouse mammary tumor virus. *RNA* **6:** 409–421.

Theimer CA, Wang Y, Hoffman DW, Krisch HM, Giedroc DP. 1998. Non-nearest neighbor effects on the thermodynamics of unfolding of a model mRNA pseudoknot. *J Mol Biol* **279:** 545–564.

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

Tucker BJ, Breaker RR. 2005. Riboswitches as versatile gene control elements. *Curr Opin Struct Biol* **15:** 342–348.

Uemura Y, Hasegawa A, Kobayashi S, Yokomori T. 1999. Tree joining grammars for RNA structure prediction. *Theor Comput Sci* **210:** 1330–1348.

van Batenburg FHD, Gultyaev AP, Pleij CWA. 2001. PseudoBase: Structural information on RNA pseudoknots. *Nucleic Acids Res* **29:** 194–195.

Walter P, Blobel G. 1982. Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. *Nature* **299:** 691–698.

Waring RB, Davies RW. 1984. Assessment of a model for intron RNA secondary structure relevant to RNA self-splicing—a review. *Gene* **28:** 277–291.

Witwer C, Hofacker IL, Stadler PF. 2004. Prediction of consensus RNA secondary structures including pseudoknots. *IEEE/ACM Trans Comput Biol Bioinformatics* **1:** 66–77.

Wu L, Belasco JG. 2008. Let me count the ways: Mechanisms of gene regulation by miRNAs and siRNAs. *Mol Cell* **29:** 1–7.

Wyatt JR, Puglisi JD, Tinoco I Jr. 1990. RNA pseudoknots, stability and loop size requirements. *J Mol Biol* **214:** 455–470.

Xia T, SantaLucia J Jr, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH. 1998. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick pairs. *Biochemistry* **37:** 14719–14735.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31:** 3406–3415.

Zwieb C, Wower I, Wower J. 1999. Comparative sequence analysis of tmRNA. *Nucleic Acids Res* **27:** 2063–2071.

Zwieb C, Gorodkin J, Knudsen B, Burks J, Wower J. 2003. tmRDB (tmRNA database). *Nucleic Acids Res* **31:** 446–447.