BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 6 2012, pages 792–798
doi:10.1093/bioinformatics/bts044

Structural bioinformatics

Advance Access publication January 27, 2012

TurboKnot: rapid prediction of conserved RNA secondary structures including pseudoknots

Matthew G. Seetin[1] and David H. Mathews[1,2,*]

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

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Many RNA molecules function without being translated into proteins, and function depends on structure. Pseudoknots are motifs in RNA secondary structures that are difficult to predict but are also often functionally important.

Results: TurboKnot is a new algorithm for predicting the secondary structure, including pseudoknotted pairs, conserved across multiple sequences. TurboKnot finds 81.6% of all known base pairs in the systems tested, and 75.6% of predicted pairs were found in the known structures. Pseudoknots are found with half or better of the false-positive rate of previous methods.

Availability: The program is available for download under an open-source license as part of the RNAstructure package at: http://rna.urmc.rochester.edu.

Contact: david_mathews@urmc.rochester.edu

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

Received on July 14, 2011; revised on November 25, 2011; accepted on January 23, 2012

1 INTRODUCTION

In many cases, RNA functions in cells without being translated into proteins. These non-coding RNAs (ncRNAs) carry out a wide range of functions, including catalysis, intracellular protein trafficking, immunity and gene regulation (Doudna and Cech, 2002; Gesteland *et al.*, 2005; Marraffini and Sontheimer, 2010; Nissen *et al.*, 2000; Tucker and Breaker, 2005; Walter and Blobel, 1982; Wu and Belasco, 2008). The functions of these ncRNAs depend on their structures.

RNA structure is hierarchical, beginning with the sequence of nucleotides (the primary structure), then the set of all canonical, i.e. GC, AU and GU base pairs (the secondary structure), and ultimately the full 3D positioning of all the atoms in the molecule (the tertiary structure). The secondary structure tends to form on a faster time scale and with stronger interactions than those interactions that fold the tertiary structure, so the secondary structure can usually be predicted independently of knowing the overall tertiary structure (Tinoco and Bustamante, 1999).

The most accurate method of secondary structure determination is comparative sequence analysis. It relies on the hypothesis that sequences that serve the same function share the same structure. But comparative sequence analysis is impractical in many situations because it requires significant human insight and a large number of diverse sequences (Gutell *et al.*, 2002; Woese and Pace, 1993). Commonly, the secondary structure for a single sequence is predicted using free energy minimization. These methods use a set of thermodynamic parameters that predict the folding free energy change of a given structure and a dynamic programming algorithm to find the structure with lowest free energy change (Mathews and Turner, 2006; Mathews *et al.*, 2004).

The thermodynamic parameters can be used with a different type of dynamic programming algorithm to predict a partition function, which can be used to predict the base pairing probability for all possible canonical pairs (McCaskill, 1990). Base pair probabilities can be used to assemble structures called maximum expected accuracy (MEA) structures (Do *et al.*, 2006; Hamada *et al.*, 2009; Knudsen and Hein, 1999; Lu *et al.*, 2009). Briefly, MEA prediction seeks to maximize the sum of probabilities of each base pair for all paired bases and the probabilities of being single-stranded for all unpaired bases. Empirically, MEA structure prediction works because that base pairs predicted to be the most probable are the most likely to be in the known structure (Mathews, 2004).

Although comparative sequence analysis is not yet automated, a number of methods have been developed to improve structure prediction by using the information available in homologous sequences. (Bernhart *et al.*, 2008; Harmanci *et al.*, 2007, 2008; Mathews and Turner, 2002; Steffen *et al.*, 2006; Will *et al.*, 2007; Xu *et al.*, 2007). These methods use a supplied input sequence alignment, predict the structure for each sequence and then find the common structure, or align and fold the sequences simultaneously. These programs have been reviewed previously (Bernhart *et al.*, 2008).

The recently published TurboFold algorithm is one such method that improves secondary structure prediction by predicting a conserved structure using two or more homologous sequences (Harmanci *et al.*, 2011). It refines base pairing probabilities of an arbitrary number of sequences using their probabilistic alignments in an iterative manner. Structures are then assembled using a MEA algorithm from the final set of base pairing probabilities.

Because of computational complexity, most programs that predict secondary structure do not predict pseudoknots. Formally, a pseudoknot occurs if an RNA has two base pairs, one between bases i and j, and a second between i' and j', such that $i < i' < j < j'$. Pseudoknots make up only a small fraction of base pairs, but they are often found in ncRNAs. Predicting the minimum free energy secondary structure that includes pseudoknots of any topology has been proven to be NP-hard (Lyngsø and Pederson, 2000). In spite of this, a number of single-sequence methods have been developed to

*To whom correspondence should be addressed.

predict structures with pseudoknots (Akutsu, 2000; Condon *et al.*, 2004; Dirks and Pierce, 2003; Reeder and Giegerich, 2004; Rivas and Eddy, 1999; Uemura *et al.*, 1999). These methods limit the set of possible pseudoknot topologies to accelerate the search. A number of available methods has been previously reviewed (Liu *et al.*, 2010).

Recently, a new algorithm was published, ProbKnot, that uses a single sequence partition function to predict pairing probabilities and assigns base pairs if the two bases are mutually maximally probable to pair with one another (Bellaousov and Mathews, 2010). This algorithm can find pseudoknots of any topology. While the partition function does not include terms for pseudoknotted structures, the ProbKnot algorithm was able to find pseudoknotted base pairs with reasonable accuracy compared with other methods, and at minimal computational cost: the algorithm is $O(N^2)$ in addition to the $O(N^3)$ partition function calculation.

Approaches to the pseudoknot prediction problem that employ information from sequence and structural homology are few in number. Like single sequence methods (Abrahams *et al.*, 1990; Isambert and Siggia, 2000), a Monte Carlo alignment and folding method has also been published (Meyer and Miklos, 2007). Also analogous to single sequence methods (Dawson *et al.*, 2007; Jabbari *et al.*, 2008; Ren *et al.*, 2005), another approach identifies common pseudoknot-free structures and then iteratively matches regions that can form pseudoknots (Ruan *et al.*, 2004). A third approach computes a score matrix based on alignment and thermodynamic information and uses a maximum weight matching (MWM) algorithm to assemble an optimal secondary structure (Tabaska *et al.*, 1998; Witwer *et al.*, 2004). The latter two approaches are dependent on a high-quality sequence alignment as input, either one assembled manually or one computed from sequences with a large degree of sequence identity (70% or more). Such alignments are not always available, particularly for newly discovered classes of RNAs.

This contribution reports a new algorithm, TurboKnot, for predicting RNA secondary structures conserved in multiple sequences, including pseudoknots. TurboKnot assembles structures in the same way as ProbKnot, but with the pairing probabilities predicted by a TurboFold calculation using multiple sequences to inform the pairing probabilities. The additional computational cost compared with TurboFold is small compared with the TurboFold calculation itself, just $O(MN^2)$, where M is the number of sequences used and N is the length of the longest sequence. It is benchmarked against two algorithms capable of predicting pseudoknots that also use multiple sequences as input, ILM and Hxmatch (Ruan *et al.*, 2004; Witwer *et al.*, 2004). It is also benchmarked against the single sequence ProbKnot algorithm (Bellaousov and Mathews, 2010). Finally, two algorithms that cannot predict pseudoknots are included in the benchmark, TurboFold (Harmanci *et al.*, 2011) and MEA (MaxExpect) (Lu *et al.*, 2009).

2 METHODS

2.1 Predicting base pair probabilities

For TurboKnot, base pair probabilities were predicted using TurboFold (Harmanci *et al.*, 2011). This algorithm employs the nearest neighbor thermodynamic parameters (Mathews *et al.*, 1999, 2004; Xia *et al.*, 1998). The per-branching-helix bonus parameter, however, was left at -0.6 kcal/mol, consistent with optical melting data (Diamond *et al.*, 2001), rather than using the optimized parameter of Mathews *et al.* (2004). TurboFold was run using four iterations instead of the default of three for an increase in accuracy at the expense of computational cost, but otherwise was run with the default parameters.

2.2 Assembling structures

Structures were constructed using the same method as the ProbKnot algorithm. First, maximal pairing probabilities for each base i were identified from the TurboFold output and stored in an array $P_{max}(i)$. Next, all pairing probabilities for all base pairs, $P(i,j)$ were checked, and if $P(i,j)$ is equal to $P_{max}(i)$ and $P_{max}(j)$, a base pair was assigned between bases i and j. This process can be iterated, where subsequent iterations only examine the bases that remained unpaired from previous iterations, but subsequent iterations were not found to increase the accuracy of the results (data not shown). Finally, all base pairs involved in helices shorter in length than 3 bp were removed (helices were not considered to be interrupted by single bulges or 1×1 internal loops). This structure prediction process was performed independently for each sequence run in a single TurboFold calculation.

2.3 MWM

MWM calculations were performed with Wmatch from MATHPROG (http://elib.zib.de/pub/Packages/mathprog/matching/weighted/) (Micali and Vazirani, 1980). Edges were only supplied to the algorithm for consideration if their computed pairing probability was $\geqslant 10^{-6}$. As with TurboKnot, all helices shorter than 3 bp were removed from the structure.

2.4 Assessment

Sensitivity and positive predictive value (PPV) were used to evaluate each algorithm tested. Sensitivity measures the fraction of pairs in each known structure found:

$$\text{Sensitivity} = \frac{\text{True Positive Pairs}}{\text{True Positive Pairs + False Negative Pairs}}$$

PPV measures the fraction of pairs predicted that are correct:

$$\text{PPV} = \frac{\text{True Positive Pairs}}{\text{True Positive Pairs + False Positive Pairs}}$$

Because it is difficult to determine the exact register of pairs by comparative analysis and because pairing can fluctuate because of thermal motions, predicted base pairs between $i$and j were counted as correct if $i$was paired with j, $j-1$, or $j+1$, or if j was paired with $i-1$ or $i+1$. Assessment using only exact matching base pairs is supplied in Supplementary Tables S1 and S2.

All sequences were tested on a randomly chosen subset (with replacement) of seven different subtypes of RNA: 5S RNA, tRNA, RNase P RNA, SRP RNA, tmRNA, telomerase RNA and group I introns (Brown, 1998; Chen *et al.*, 2000; Damberger and Gutell, 1994; Sprinzl *et al.*, 1998; Szymanski *et al.*, 1998; Waring and Davies, 1984; Zwieb and Wower, 2000). For all multisequence methods, five RNAs of each type were selected to be computed together in each calculation. Particular RNA sequences could be chosen more than once as a part of different calculations. These were considered to be separate results, as multiple sequence methods may give different results for the same sequence depending on the other sequences used in the calculation. For consistency, duplicate sequences were also counted as independent calculations when using single-sequence methods to predict structures, even though the resulting structures would be identical. Average values of sensitivity and PPV were computed for each class of RNAs. The overall average is the mean of the individual averages on sequence types.

2.5 Pseudoknot evaluation

Pseudoknots were identified using the algorithm of Smit *et al.* (2008) to identify the fewest pairs that could be removed to remove the pseudoknot from the structure. Predicted pseudoknotted base pairs were only counted

as correct if the $i-j$ pair was in the accepted structure (allowing thermal fluctuation as in Section 2.4), the pair was counted as a pseudoknot by the Smit *et al.* algorithm, and the pair was pseudoknotted. A pair was considered pseudoknotted if it met the $i < i' < j < j'$ criterion, with at least one other pair $i'-j'$ that was also a correct pair compared with the accepted structure (Bellaousov and Mathews, 2010).

3 RESULTS

Over the range of RNA families tested, which have a variety of functions and structures, TurboKnot averaged a sensitivity of 79.8% (Table 1) and a PPV of 72.9% (Table 2). For comparison, TurboKnot was benchmarked against two other multisequence methods that can predict pseudoknots: ILM version 1.0 (Ruan *et al.*, 2004) and Hxmatch version 1.2.1 (Witwer *et al.*, 2004). Each was run with default parameters. Sequences were aligned for input into these algorithms using MUSCLE version 3.6 (Edgar, 2004), which performs well for RNA sequence alignment compared with other available tools (Wilm *et al.*, 2006). In addition to these two methods, benchmarks were run using ProbKnot (Bellaousov and Mathews, 2010), a single sequence method that can predict pseudoknots, TurboFold (Harmanci *et al.*, 2011), the underlying multisequence method that cannot predict pseudoknots and MaxExpect (Lu *et al.*, 2009), a single sequence method that cannot predict pseudoknots. Other single-sequence methods capable of predicting pseudoknots were recently benchmarked (Bellaousov and Mathews, 2010), and ProbKnot compares favorably against these. It is used here as a representative of this class of algorithms. One algorithm, DotKnot, has been published more recently. Its performance on this test set is included in Supplementary Tables S3–S5 for comparison. Similarly, multisequence algorithms that cannot predict pseudoknots were recently benchmarked (Harmanci *et al.*, 2011), and TurboFold compares favorable and is used as a representative of them here.

The average sensitivity and PPV of TurboKnot were better than any other method capable of predicting pseudoknots. Additionally, TurboKnot had a higher sensitivity than TurboFold, albeit at a trade-off in PPV. Since TurboKnot considers a larger structure space than TurboFold, it is natural to find more correct base pairs with a cost to the PPV.

For pseudoknots in particular, TurboKnot found 289 correct pseudoknotted base pairs out of a total of 5897 in the known structures, and it made 672 false positive predictions of pseudoknotted pairs (Table 3). This is more true positives and fewer false positives than any other tested method. When considering

Table 1. Sensitivities of tested methods for all base pairs

Type of RNA	Sequences	Unique Seqs.[a]	Base pairs	Pseudoknotted base pairs[b]	TurboKnot[c,d] (%)	ILM[c,d] (%)	Hxmatch[c,d] (%)	ProbKnot[c] (%)	TurboFold[d] (%)	MaxExpect (%)
tRNA	200	167	4157	0	**97.4 ± 5.5**	70.5 ± 21.1	74.1 ± 25.4	91.4 ± 13.7	93.8 ± 7.0	88.9 ± 14.0
5S rRNA	200	153	6567	0	90.1 ± 5.8	71.5 ± 18.5	69.1 ± 16.4	67.9 ± 27.9	**90.9 ± 5.4**	67.2 ± 28.5
SRP	100	66	6958	86	**78.6 ± 22.5**	17.0 ± 15.4	26.3 ± 22.9	63.2 ± 22.4	77.5 ± 23.9	63.7 ± 27.4
RNase P	80	71	8322	481	**77.3 ± 7.5**	25.6 ± 15.4	30.2 ± 16.1	63.9 ± 11.9	75.0 ± 8.6	63.9 ± 12.6
tmRNA	200	161	23346	4393	**66.5 ± 9.2**	30.1 ± 11.0	30.2 ± 12.6	46.2 ± 12.8	63.3 ± 10.7	45.2 ± 12.8
Telomerase	50	25	5244	477	**83.2 ± 7.2**	26.2 ± 11.4	54.1 ± 17.3	56.1 ± 17.3	82.9 ± 7.4	55.7 ± 16.9
Group I Intron	80	24	7593	460	65.6 ± 15.6	4.2 ± 4.8	4.5 ± 5.7	**67.6 ± 13.9**	59.2 ± 17.2	67.0 ± 14.2
Total/average	910	667	62187	5897	**79.8 ± 11.6**	35.0 ± 26.0	41.2 ± 25.3	65.2 ± 13.0	77.5 ± 13.0	64.5 ± 13.3

Underlined results represent the best performance out of all the algorithms that predict pseudoknots. Bold results represent absolute best performance.
[a]Sequences were selected at random with replacement, so some sequences were chosen more than once.
[b]Pseudoknots were counted using the method of Smit et al. (2008).
[c]TurboKnot, ILM, Hxmatch and ProbKnot can include pseudoknotted pairs in structure prediction. Other algorithms do not include pseudoknots.
[d]TurboKnot, ILM, Hxmatch and TurboFold predict a conserved structure using multiple sequences. Other algorithms use a single sequence.

Table 2. PPVs of tested methods for all base pairs

Type of RNA	Sequences	Unique Seqs.[a]	Base pairs	Pseudoknotted base pairs[b]	TurboKnot[c,d] (%)	ILM[c,d] (%)	Hxmatch[c,d] (%)	ProbKnot[c] (%)	TurboFold[d] (%)	MaxExpect (%)
tRNA	200	167	4157	0	86.9 ± 9.2	81.3 ± 17.3	80.0 ± 22.7	83.2 ± 15.2	**94.8 ± 8.4**	88.0 ± 16.0
5S rRNA	200	153	6567	0	84.3 ± 6.4	81.3 ± 15.9	73.2 ± 15.6	62.3 ± 26.4	**84.8 ± 5.9**	60.6 ± 25.6
SRP	100	66	6958	86	61.9 ± 19.3	21.1 ± 18.4	30.1 ± 22.2	48.7 ± 22.4	**65.2 ± 20.8**	50.0 ± 23.4
RNase P	80	71	8322	481	78.4 ± 7.8	40.5 ± 20.9	45.6 ± 19.6	63.6 ± 11.9	**80.4 ± 7.7**	64.2 ± 12.4
tmRNA	200	161	23346	4393	68.0 ± 12.5	41.6 ± 12.2	44.0 ± 16.4	42.2 ± 12.6	**75.3 ± 10.5**	43.9 ± 13.7
Telomerase	50	25	5244	477	65.3 ± 8.6	27.2 ± 11	49.0 ± 14.5	41.5 ± 13.9	**69.3 ± 9.2**	42.9 ± 14.6
Group I Intron	80	24	7593	460	65.6 ± 14.8	8.1 ± 10.6	11.1 ± 13.6	61.4 ± 13.2	**71.7 ± 15.2**	64.5 ± 13.2
Total/Average	910	667	62187	5897	72.9 ± 10.1	43.0 ± 28.5	47.6 ± 23.7	57.6 ± 14.7	**77.4 ± 10.1**	59.2 ± 15.6

Underlined results represent the best performance out of all the algorithms that predict pseudoknots. Bold results represent absolute best performance.
[a]Sequences were selected at random with replacement, so some sequences were chosen more than once.
[b]Pseudoknots were counted using the method of Smit *et al.* (2008).
[c]TurboKnot, ILM, Hxmatch and ProbKnot can include pseudoknotted pairs in structure prediction. (Other algorithms do not include pseudoknots.)
[d]TurboKnot, ILM, Hxmatch and TurboFold predict a conserved structure using multiple sequences. (Other algorithms use a single sequence.)

Table 3. Evaluation of tested methods for pseudoknotted base pairs

Type of RNA	Sequences	Unique Seqs.[a]	Base pairs	Pseudoknotted base pairs[b]	Predicted pseudoknotted pairs[b]				Correctly predicted pseudoknotted pairs[c]			
					TurboKnot[d]	ILM[d]	Hxmatch[d]	ProbKnot	TurboKnot[d]	ILM[d]	Hxmatch[d]	ProbKnot
tRNA	200	167	4157	0	41	187	542	109	0	0	0	0
5S rRNA	200	153	6567	0	41	360	1145	98	0	0	0	0
SRP	100	66	6958	86	74	102	1297	118	3	0	0	4
RNase P	80	71	8322	481	54	751	1069	144	9	6	0	32
tmRNA	200	161	23346	4393	545	2659	2799	721	266	73	112	108
Telomerase	50	25	5244	477	52	1534	1303	139	0	28	115	0
Group I Intron	80	24	7593	460	154	585	978	173	11	0	0	12
Total	910	667	62187	5897	961	6178	7830	1502	289	107	227	156

[a]Sequences were selected at random with replacement, so some sequences were chosen more than once.
[b]Pseudoknots were counted using the method of Smit *et al.* (2008).
[c]Correctly predicted pseudoknots are only counted if there is true positive base pair that is pseudoknotted with at least one other true positive base pair.
[d]TurboKnot, ILM and Hxmatch predict a conserved structure using multiple sequences. (ProbKnot uses a single sequence.)

Table 4. Evaluation of time performance

Type of RNA	Avg. Seq. length	TurboKnot[a,b] (min:s)	ILM[a,b,c](min:s)	Hxmatch[a,b,d] (min:s)	ProbKnot[a,e] (min:s)
tRNA	80.0	00:12.5	00:00.4	00:00.2	00:00.7
5S rRNA	119.6	00:34.3	00:00.9	00:00.2	00:01.5
SRP	274.4	06:57.6	00:01.8	00:00.7	00:20.2
RNase P	321.6	07:02.5	00:03.8	00:00.7	00:27.2
tmRNA	367.4	12:25.8	00:04.3	00:00.7	00:42.0
Telomerase	458.8	23:41.2	00:12.9	00:01.2	01:36.4
Group I Intron	326.0	12:04.0	00:04.9	00:00.5	00:35.6

Time necessary to compute the structures of five random RNAs of each time for each algorithm considered that can predict pseudoknots. Timing results for TurboFold and MaxExpect are included in Supplementary Table S2. Calculations were run on a machine with an Intel Core 2 Quad Q6600 processor and 4 GB of RAM running Ubuntu 8.04 64 bit and using version 4.2.4 of the GNU C++ compiler. Times were obtained with the Linux 'time' command.
[a]TurboKnot, ILM, Hxmatch and ProbKnot can include pseudoknotted pairs in structure prediction. Other algorithms do not include pseudoknots.
[b]TurboKnot, ILM, Hxmatch and TurboFold predict a conserved structure using multiple sequences. Other algorithms use a single sequence.
[c]ILM computes one structure in a calculation so this is the time for five successive calculations so as to get structures for each of the five RNAs considered. This also includes the time for Muscle to align the sequences for input, as well as any necessary file format conversions.
[d]Includes the time for Muscle to align the sequences for input and all necessary file format conversions.
[e]Single sequence method times are times necessary for five successive calculations on each RNA individually.

structures with pseudoknots rather than individual pairs, TurboKnot found at least one true positive pseudoknot in 59 structures out of 239 predictions, and out of 426 tested structures that contain at least one pseudoknot (Supplementary Table S6). This is again more true positives and fewer false positives than any other tested method.

The use of an MWM algorithm to assemble structures given the TurboFold pair probabilities was also considered as a control (Supplementary Tables S7 and S8). This approach resulted in the identification of the exact same set of correct base pairs as those based on mutual maximum probability pairs, plus a small number of extra incorrect pairs. The additional, incorrect pairs were so few that they only affected the second decimal places of the PPV percentages. The exception was the tmRNA family, in which a small number of additional correct base pairs were identified. The minor difference between these two approaches was due to how well the TurboFold algorithm refines the pairing probabilities by excluding structures that are not conserved.

As another control, an MWM algorithm was used to predict structures from pair probabilities calculated using a single sequence partition function. This added a number of spurious base pairs (Supplementary Table S9). This emphasized the importance of the ProbKnot approach when single sequences are used.

Time benchmarks were performed on all tested algorithms. Groups of five random sequences were used, with average lengths varying from 80.0 nt to 458.8 nt (Table 4). The Hxmatch algorithm performed the fastest. TurboKnot was significantly slower, but it could still carry out the computations in a reasonable amount of time for a user with a typical desktop computer.

Figure 1 shows a sample structure prediction for *Mycobacterium tuberculosis* RNase P using TurboKnot, ProbKnot or TurboFold (Brown, 1998). The TurboKnot and TurboFold structures are similar, but the TurboKnot algorithm is able to identify most of the pseudoknotted base pairs. ProbKnot has a lower sensitivity and PPV overall for non-pseudoknotted base pairs in this structure. It identifies the four correct pseudoknotted base pairs that TurboKnot missed, but it also finds eight pseudoknotted base pairs that are not in the accepted structure.

TurboKnot overextends some helices as compared with TurboFold when there are valid bases available to pair. The accuracy of the extra base pairs added when TurboKnot extends helices

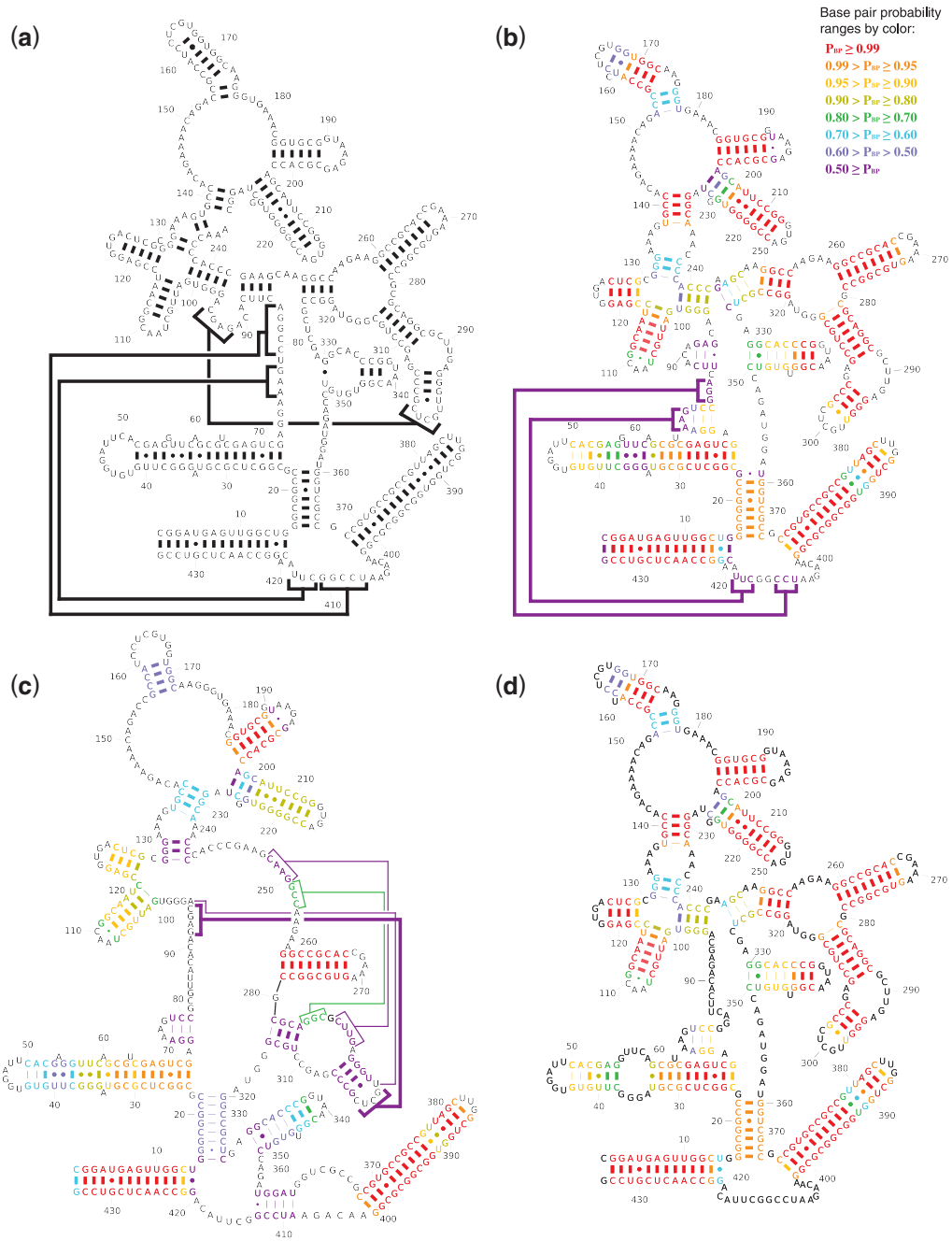

Fig. 1. Accepted and predicted structures for *M.tuberculosis* RNase P. (**a**) The accepted structure of *M.tuberculosis* RNase P. (Brown, 1998). (**b**) The structure predicted by TurboKnot. (**c**) The structure predicted by ProbKnot. (**d**) The structure predicted by TurboFold. Correctly predicted base pairs are shown with thick lines, and incorrect pairs with thin lines. Figures were drawn using the XRNA program (http://rna.ucsc.edu/rnacenter/xrna/).

compared to TurboFold is low; of the base pairs on ends of helices predicted by TurboKnot and not TurboFold, 27.6% were correct in all the structures. This is for cases where a helix is in common in the predictions of TurboKnot and TurboFold. Single nucleotide bulges are not considered to break a helix. These helices that are extended by 1 bp on an end are sometimes correct, so they do increase the sensitivity score of TurboKnot compared with TurboFold but at the expense of PPV.

4 DISCUSSION

In the same way as the ProbKnot algorithm, TurboKnot is able to assemble structures containing pseudoknots using base pair probabilities. Using probabilities from TurboFold that are functions of structural alignment and homology information in addition to thermodynamic stability significantly improves predictions when considering all base pairs and when considering just pseudoknots. Compared with TurboKnot, some methods were able to find more

true pseudoknotted pairs on some systems. For example, Hxmatch found more true pseudoknotted pairs in telomerase RNA, but in the process it predicted more than twice as many false positive pairs as there were true pseudoknotted pairs in the accepted structures.

Single sequence methods, with or without pseudoknot-prediction capabilities, were reported as having poor performance on the prediction of tmRNA structure (Bellaousov and Mathews, 2010). The results with the sequences used here are consistent with that result. TurboKnot (and TurboFold), however, show significant improvement for this difficult class of RNAs. TurboKnot shows its best pseudoknot-prediction capabilities on this class of RNAs, with 49% of all predicted pseudoknotted base pairs being true positives.

In contrast, Group I introns show slightly worse sensitivity in TurboFold and TurboKnot compared with MaxExpect and ProbKnot, although they did still have higher PPVs. This is likely because the long insertions present in some Group I introns makes accurate sequence alignment difficult. Indeed, ILM and Hxmatch, which are dependent on a sequence alignment algorithm for input, did not perform well with randomly chosen groups of sequences from this class of RNAs. These algorithms may have more success, however, when sequences with higher sequence similarity are used or when using a manually curated alignment, which would allow for more accurate sequence alignment. Such groups of sequences are not always available.

ILM was benchmarked in a previous study using just one sequence rather than an alignment, and, unexpectedly, it appears less accurate when using multiple sequences as input compared with previously reported benchmarks (Bellaousov and Mathews, 2010; Ruan *et al.*, 2004). This is likely because it was hindered by errors made by the sequence alignment algorithm, which aligns sequences but not secondary structure (Edgar, 2004). It fared much better when given manually curated alignments based on structural homology (Ruan *et al.*, 2004). Because Hxmatch, which was given the same input sequence alignments, had similar results on similar classes of RNAs, it is likely tools that consider both secondary structure and alignment at the same time are necessary to accurately align functionally similar RNAs which are not chosen to have a relatively high sequence identity.

While TurboKnot has significantly improved pseudoknot prediction accuracy, more work is needed to raise the accuracy to the current standard for non-pseudoknotted base pairs. Dirks and Pierce (2004) report an algorithm that computes a partition function that includes pseudoknots of some limited topologies in $O(N^4)$ time. This partition function could be used by TurboFold and TurboKnot to consider a larger structure space, ideally increasing the probability of true pseudoknots at the expense of false ones that are predicted currently at an acceptable increase in computational expense.

ACKNOWLEDGEMENTS

The authors thank A.O. Harmanci and G. Sharma for discussions and S. Bellaousov for performing MWM calculations. Computer time was provided by the University of Rochester Center for Research Computing.

Funding: National Institutes of Health (grant number R01HG004002 to D.H.M).

Conflict of Interest: none declared.

REFERENCES

Abrahams,J.P. *et al.* (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. *Discr. Appl. Math.*, **104**, 45–62.

Bellaousov,S. and Mathews,D.H. (2010) ProbKnot: fast prediction of RNA secondary structure including pseudoknots. *RNA*, **16**, 1870–1880.

Bernhart,S.H. *et al.* (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474.

Brown,J.W. (1998) The ribonuclease P database. *Nucleic Acids Res.*, **26**, 351–352.

Chen,J.L. *et al.* (2000) Secondary structure of vertebrate telomerase RNA. *Cell*, **100**, 503–514.

Condon,A. *et al.* (2004) Classifying RNA pseudoknotted structures. *Theor. Comput. Sci.*, **320**, 35–50.

Damberger,S.H. and Gutell,R.R. (1994) A comparative database of group I intron structures. *Nucleic Acids Res.*, **22**, 3508–3510.

Dawson,W.K. *et al.* (2007) Prediction of RNA pseudoknots using heuristic modeling with mapping and sequential folding. *PLoS One*, **2**, e905.

Diamond,J.M. *et al.* (2001) Thermodynamics of three-way multibranch loops in RNA. *Biochemistry*, **40**, 6971–6981.

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.

Dirks,R.M. and Pierce,N.A. (2004) An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. *J. Comput. Chem.*, **25**, 1295–1304.

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

Doudna,J. and Cech,T. (2002) The chemical repertoire of natural ribozymes. *Nature*, **418**, 222–228.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Gesteland,R.F. *et al.* (2005) *The RNA World*, 3rd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Gutell,R.R. *et al.* (2002) The accuracy of ribosomal RNA comparative structure models. *Curr. Opin. Struct. Biol.*, **12**, 301–310.

Hamada,M. *et al.* (2009) Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics*, **25**, 465–473.

Harmanci,A.O. *et al.* (2007) Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign. *BMC Bioinformatics*, **8**, 130.

Harmanci,A.O. *et al.* (2008) PARTS: Probabilistic Alignment for RNA joinT Secondary structure prediction. *Nucleic Acids Res.*, **36**, 2406–2417.

Harmanci,A.O. *et al.* (2011) TurboFold: Iterative probabilistic estimation of secondary structures for multiple RNA sequences. *BMC Bioinformatics*, **12**, 108.

Isambert,H. and Siggia,E.D. (2000) Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme. *Proc. Natl Acad. Sci. USA*, **97**, 6515–6520.

Jabbari,H. *et al.* (2008) Novel and efficient RNA secondary structure prediction using hierarchical folding. *J. Comput. Biol.*, **15**, 139–163.

Knudsen,B. and Hein,J.J. (1999) RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, **15**, 446–454.

Liu,B. *et al.* (2010) RNA pseudoknots: folding and finding. *F1000 Biol. Rep.*, **2**, 8.

Lu,Z.J. *et al.* (2009) Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA*, **15**, 1805–1813.

Lyngsø,R. and Pederson,C. (2000) RNA pseudoknot prediction in energy-based models. *J. Comput. Biol.*, **7**, 409–427.

Marraffini,L.A. and Sontheimer,E.J. (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat. Rev. Genet.*, **11**, 181–190.

Mathews,D.H. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, **10**, 1178–1190.

Mathews,D.H. and Turner,D.H. (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.*, **317**, 191–203.

Mathews,D.H. and Turner,D.H. (2006) Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol.*, **16**, 270–278.

Mathews,D.H. *et al.* (1999) Expanded sequence dependence of thermodynamic parameters provides improved prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.

Mathews,D.H. *et al.* (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl Acad. Sci. USA*, **101**, 7287–7292.

McCaskill,J.S. (1990) The equilibrium partition function and base pair probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.

Meyer,I.M. and Miklos,I. (2007) SimulFold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. *PLoS Comput. Biol.*, **3**, e149.

Micali,S. and Vazirani,V.V. (1980) An $O(V^{1/2}E)$ algorithm for finding maximum matching in general graphs. In *21st Annual Symposium on Foundations of Computer Science*. IEEE Computer Society Press, New York, pp. 17–27.

Nissen,P. *et al.* (2000) The structural basis of ribosomal activity in peptide bond synthesis. *Science*, **289**, 920–930.

Reeder,J. and Giegerich,R. (2004) Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics*, **5**, 104.

Ren,J. *et al.* (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.

Ruan,J. *et al.* (2004) An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, **20**, 58–66.

Smit,S. *et al.* (2008) From knotted to nested RNA structures: a variety of computational methods for pseudoknot removal. *RNA*, **14**, 410–416.

Sprinzl,M. *et al.* (1998) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, **26**, 148–153.

Steffen,P. *et al.* (2006) RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, **22**, 500–503.

Szymanski,M. *et al.* (1998) 5S rRNA data bank. *Nucleic Acids Res.*, **26**, 156–159.

Tabaska,J.E. *et al.* (1998) An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics*, **14**, 691–699.

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

Tucker,B.J. and Breaker,R.R. (2005) Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.*, **15**, 342–348.

Uemura,Y. *et al.* (1999) Tree adjoining grammars for RNA structure prediction. *Theor. Comput. Sci.*, **210**, 277–303.

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

Waring,R.B. and Davies,R.W. (1984) Assessment of a model for intron RNA secondary structure relevant to RNA self-splicing - a review. *Gene*, **28**, 277–291.

Will,S. *et al.* (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, **3**, e65.

Wilm,A. *et al.* (2006) An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms Mol. Biol.*, **1**, 19.

Witwer,C. *et al.* (2004) Prediction of consensus RNA secondary structures including pseudoknots. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **1**, 66–77.

Woese,C.R. and Pace,N.R. (1993) Probing RNA structure, function, and history by comparative analysis. In Gesteland,R.F. (ed.), *The RNA World*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 91–117.

Wu,L. and Belasco,J.G. (2008) Let me count the ways: mechanisms of gene regulation by miRNAs and siRNAs. *Mol. Cell*, **29**, 1–7.

Xia,T. *et al.* (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick pairs. *Biochemistry*, **37**, 14719–14735.

Xu,X. *et al.* (2007) RNA Sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment. *Bioinformatics*, **23**, 1883–1891.

Zwieb,C. and Wower,J. (2000) tmRDB (tmRNA database). *Nucleic Acids Res.*, **28**, 169–170.