ThreshKnot: Thresholded ProbKnot for Improved RNA Secondary Structure Prediction

Liang Zhang[a], **He Zhang**[b], **David H. Mathews**[c], **and Liang Huang**[a,b,♣]

[a]School of Electrical Engineering & Computer Science, Oregon State University, Corvallis, OR; [b]Baidu Research, Sunnyvale, CA; [c]Department of Biochemistry & Biophysics, Center for RNA Biology, and Department of Biostatistics & Computational Biology, University of Rochester Medical Center, Rochester, NY

RNA structure prediction is a challenging problem, especially with pseudoknots. Recently, there has been a shift from the classical minimum free energy-based methods (MFE) to partition function-based ones that assemble structures using base-pairing probabilities. Two examples of the latter group are the popular maximum expected accuracy (MEA) method and the ProbKnot method. ProbKnot is a fast heuristic that pairs nucleotides that are reciprocally most probable pairing partners, and unlike MEA, can also predict structures with pseudoknots. However, ProbKnot's full potential has been largely overlooked. In particular, when introduced, it did not have an MEA-like hyperparameter that can balance between positive predictive value (PPV) and sensitivity. We show that a simple thresholded version of ProbKnot, which we call *ThreshKnot*, leads to more accurate overall predictions by filtering out unlikely pairs whose probabilities fall under a given threshold. We also show that on three widely-used folding engines (RNAstructure, Vienna RNAfold, and CONTRAfold), ThreshKnot always outperforms the much more involved MEA algorithm in (1) its higher structure prediction accuracy, (2) its capability to predict pseudoknots, and (3) its faster runtime and easier implementation. This suggests that ThreshKnot should replace MEA as the default partition function-based structure prediction algorithm. ThreshKnot is already available in the widely used RNAstructure software package version 6.2 (released November 27, 2019): https://rna.urmc.rochester.edu/RNAstructure.html

1 Introduction

RNAs are involved in multiple processes, including catalysis, guiding RNA modification, and post-transcriptional gene regulation (Bachellerie et al., 2002, Doudna and Cech, 2002, Karijolich et al., 2015, Serganov and Nudler, 2013, Storz and Gottesman, 2006, Wu and Belasco, 2008). Often, RNA function is highly related to structure. However, structure determine techniques, such as Cryo-Electron Microscopy (Cry-EM) (Ognjenović et al., 2019), X-ray crystallography (Zhang and Ferré-D'Amaré, 2014) or Nuclear Magnetic Resonance (NMR) (Zhang and Keane, 2019), though reliable and accurate, are slow and costly. Therefore, fast and accurate computational prediction of RNA structure is useful and desired. Because tertiary structure modeling is challenging (Miao et al., 2017), many studies focus on predicting the secondary structure, i.e., the double helices formed by base pairing of self-complementary nucleotides (A-U, G-C, G-U base pairs) (Tinoco and Bustamante, 1999). The secondary structure is well-defined, provides detailed information to help understand the structure-function relationship, and is a basis to predict full tertiary structure (Nawrocki and Eddy, 2013, Parisien and Major, 2008, Seetin and Mathews, 2011).

Most algorithms for RNA secondary structure prediction can be divided into two categories, the classical ones computing a single structure with the minimum free energy (MFE) (Nussinov and Jacobson, 1980, Zuker and Stiegler, 1981), and the more recent ones based on the partition function, which is the sum of all equilibrium constants for all possible structures and is the normalization for estimating marginal probabilities of base pairs and motifs (McCaskill, 1990). Generally speaking, there is a trend to shift from the former (MFE-based) methods to the latter (partition function-based) ones for many reasons, including (1) the overall accuracy of partition function-based methods is generally higher than that of MFE-based (Do et al., 2006, Hajiaghayi et al., 2012, Lu et al., 2009), (2) instead of predicting a single structure as in MFE, the partition function captures the whole ensemble of conformations and an RNA molecule (e.g., mRNAs) can be many different conformations at equilibrium (Cordero and Das, 2015, Lai et al., 2018, Long et al., 2007, Lu and Mathews, 2008, Tafer et al., 2008), (3) we can

♣Corresponding author: liang.huang.sh@gmail.com.

also induce the base-pairing probabilities from the partition function, and (4) as a by-product, heuristic algorithms can use the partition function to predict pseudoknots[1] (Bellaousov and Mathews, 2010, Sato et al., 2011).

There are two typical (and widely used) examples of partition function-based prediction algorithms. The first is maximum expected accuracy (MEA) (Do et al., 2006, Knudsen and Hein, 2003), which predicts the structure $\mathbf{y}$ that maximizes the sum of the base-paired and single-stranded probabilities ($p_{i,j}$'s and q_j's, respectively):

$$2\gamma \sum_{(i,j)\in \text{pairs}(\mathbf{y})} p_{i,j} + \sum_{j\in \text{unpaired}(\mathbf{y})} q_j$$

where γ is a hyperparameter that balances the positive predictive value (PPV; a.k.a. precision) and sensitivity (a.k.a. recall) of the output structure. The other one is ProbKnot (Bellaousov and Mathews, 2010), which builds structure of mutually maximal probability pairing partners. Both use base-pairing probabilities to assemble the output structure, but the former requires another $O(n^3)$-time dynamic program for the assembly, while the latter is a simpler heuristic method that only needs $O(n^2)$ time. More importantly, ProbKnot can predict pseudoknots while MEA cannot.

However, the full potential of ProbKnot has not been fully exploited. In particular, unlike MEA, ProbKnot lacks a hyperparameter to balance PPV and sensitivity. To address this problem, we present ThreshKnot (short for Thresholded ProbKnot), which adds a probability threshold θ to disallow any pair whose probability falls below θ. Therefore, a smaller value of θ encourages ThreshKnot to predict more base pairs, and a higher one makes it more selective. By tuning θ, we can balance the PPV (the fraction of predicted pairs in the accepted structure) and sensitivity (the fraction of accepted pairs predicted).

Simple as it is, we show that ThreshKnot leads to more accurate overall predictions, and with three widely-used folding engines (RNAstructure (Reuter and Mathews, 2010), Vienna RNAfold (Lorenz et al., 2011), and CONTRAfold (Do et al., 2006)), ThreshKnot always outperforms the much more involved MEA algorithm in all three aspects: (1) it can achieve better overall prediction accuracy than MEA, (2) it can predict pseudoknots that MEA can not, (3) it is much simpler to implement and runs much faster. This suggests that ThreshKnot should replace MEA as the default partition function-based structure prediction algorithm.

[1] A pseudoknot involves at least two pairs (i, j) and (k, l) such that $i < k < j < l$.

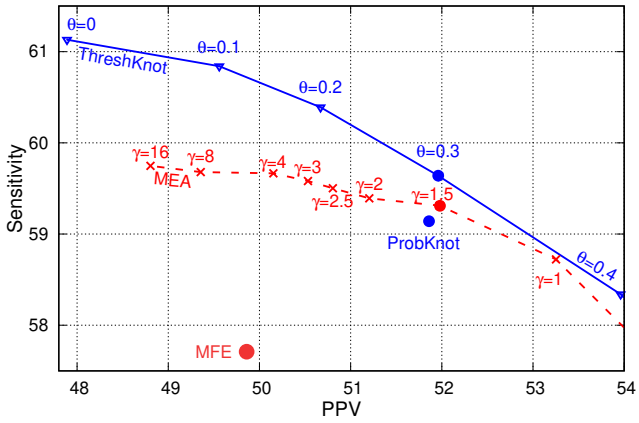

Fig. 1. Comparison between minimum free energy structure prediction (MFE), maximum expected accuracy (MEA), ProbKnot, and ThreshKnot on RNAstructure using the ArchiveII dataset of sequences with well-determined secondary structures. ThreshKnot has better PPV and Sensitivity than all other methods. For MEA and ThreshKnot, calculations were run as a function of the hyperparameter, γ and θ, respectively. Between points, lines are draw. ProbKnot and MFE have no tuning parameter and therefore are single points on these plots. ProbKnot filters out helices of two or less base pairs, while ThreshKnot only filters out single-pair helices.

2 Results

2.1 ThreshKnot Algorithm

ThreshKnot, like ProbKnot, outputs the secondary structure made of "most probable base pairs". i.e., pairs (i, j) whose probability $p_{i,j}$ is the highest among "competing pairs", i.e., $p_{i,j} \geq p_{i,k}$ for all k and $p_{i,j} \geq p_{l,j}$ for all l. But in addition to that, ThreshKnot also rules out any pair whose probability falls below θ, i.e., it returns the set of pairs

$$\{(i, j) \mid p_{i,j} = \max_k p_{i,k} = \max_l p_{l,j} \text{ and } p_{i,j} \geq \theta\}.$$

To keep it simple, unlike ProbKnot which removes helices composed of two or less stacked pairs,

		time complexity	*overall* PPV	*overall* sens.	*pseudoknot* PPV	*pseudoknot* sens.
RNAstructure	MFE	$O(n^3)$	49.86	57.71	-	-
	MEA	$O(n^3)+O(n^3)$	51.98	59.31	-	-
	ProbKnot	$O(n^3)+O(n^2)$	51.86	59.14	7.04	2.59
	ThreshKnot	$O(n^3)+O(n^2)$	51.96	**59.64**	7.62	2.85
	IPknot	$O(n^3)+O(n^2)+$ILP	**60.22**	51.46	**16.16**	8.60
	pKiss	$O(n^4)$	44.32	51.03	9.72	**15.29**

Table 1. Accuracy of ThreshKnot. The gray-shaded $O(n^3)$ denotes the time to compute the partition function and base-pairing probabilities, and light blue shades denote the time for post-processing steps based on those probabilities. ILP denotes the time to solve the integer linear program, which is NP-complete in the worst case but very fast in practice. See the Methods section for the definitions of pseudoknot PPV ($\text{PPV}_{\text{crossing}}$) and Sensitivity ($\text{sens}_{\text{crossing}}$).

ThreshKnot only removes single-pair helices.[2]

2.2 Overall Prediction Accuracy

Below we show ThreshKnot results using the base-pairing matrices generated by RNAstructure; see the Supplementary Information for the results of ThreshKnot on CONTRAfold and Vienna RNAfold. Figure 1 compares ThreshKnot with MEA, MFE, and ProbKnot. We choose θ = 0, 0.1, 0.2, 0.3, 0.4, 0.5, and 0.6 for ThreshKnot, and γ = 0.5, 1, 1.5, 2, 2.5, 3, 4, 8, and 16 for MEA. We evaluate the overall prediction accuracies across all families, reporting both PPV and sensitivity.

Figure 1 shows that the accuracy curve of ThreshKnot with varying θ is always on the upper right side of the accuracy curve of MEA with varying γ. This shows that at a given level of PPV, ThreshKnot always has a higher sensitivity.

We further use Jackknife resampling method (Tukey, 1958) to choose the best parameter θ for ThreshKnot (see Methods) and γ for MEA, i.e. the parameter that maximizes the F-score (harmonic mean of sensitivity and PPV) with respect to MFE F-score. The same $\theta = 0.3$ is chosen consistently across all families for ThreshKnot, and the same $\gamma = 1.5$ is chosen consistently for MEA, suggesting these parameters would be widely applicable to other RNA families. Table 1 summarizes the overall accuracies using these parameters, comparing four methods (MFE, MEA, ProbKnot, and ThreshKnot) with RNAstructure. ThreshKnot's overall sensitivity is significantly higher than MEA (+0.33%, p-value 0.02) and is the best among all methods, while its overall PPV is only marginally and insignificantly lower than MEA (-0.02%, p-value 0.97). Figure 2 details the accuracies on each family and the statistical significance tests.

Table 1 also includes two other systems: IPknot (Sato et al., 2011) and pKiss[3] (Theis et al., 2010), both of which use free energy parameters specialized for pseudoknot prediction in addition to those used by RNAstructure. IPknot has a higher PPV but lower sensitivity than ThreshKnot, and its F-score (55.50) is slightly lower than ThreshKnot's (55.53); however, it is worth noting that the ThreshKnot here is based on RNAstructure, and the ThreshKnot versions based on CONTRAfold and Vienna RNAfold have higher accuracies; see Figs. SI 1 and SI 3. pKiss, on the other hand, has lower PPV and Sensitivities.

Figure 3 shows the ThreshKnot accuracy curve with varying θ for each family, and the corresponding MFE accuracy on that family. Compared with MFE, ThreshKnot improves six (6) out of nine (9) families' accuracies (in both PPV and Sensitivity).

2.3 Pseudoknot Prediction Accuracy

We next evaluate ThreshKnot's abilities to predict pseudoknots, and we use the PPV and sensitivity of "crossing-pairs" to measure the pseudoknot prediction accuracy (see Materials and Methods for details). Table 1 compares ThreshKnot with ProbKnot, IPknot, and pKiss (note that MFE and MEA are unable to predict pseudoknots). ThreshKnot is more accurate in pseudoknot prediction than ProbKnot in both crossing-pair PPV and sensitivity. IPknot and pKiss, on the other hand, are two specialized tools tailored to pseudoknot prediction, and they indeed have higher crossing-pair PPV and sensitivity than ThreshKnot, which is a general-purpose structure prediction tool. Table SI 3 details pseudoknot prediction accuracies for each family.

[2] The ThreshKnot results of not removing any helices are almost identical to those removing single-pair helices.

[3] pKiss is the successor of pknotsRG (Reeder and Giegerich, 2004)

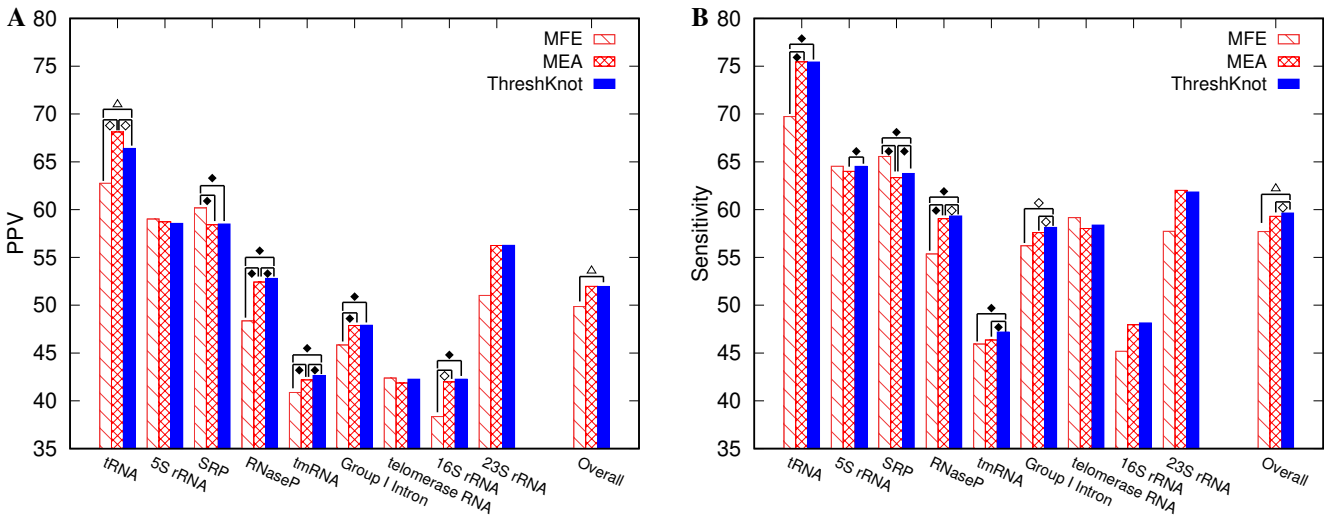

Fig. 2. Accuracies of MFE, MEA (γ=1.5), and ThreshKnot (θ=0.3). In both panels, the first nine bars from the left represent PPV (A) and sensitivity (B) averaged over all sequences in one family, and the rightmost bars represent the overall accuracies, averaging over all families. Statistical significance (two-sided) is marked as $^{\blacklozenge}(p<0.01)$, $^{\diamond}(0.01\leq p<0.05)$, or $^{\triangle}(0.05\leq p<0.06)$.

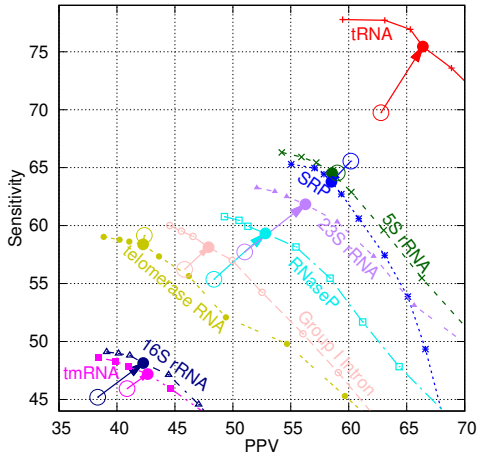

Fig. 3. ThreshKnot improves 6 out of 9 families over MFE (in both PPV and sensitivity). The curves show the ThreshKnot accuracies with varying θ. The arrows point from MFE (hollow circles) to ThreshKnot at $\theta=0.3$.

	base-pair probs	threshold pruning	pair selection
classical (McCaskill)	$O(n^3)$	$O(n^2)$	$O(n)$
LinearPartition	$O(n)$	$O(n)$	$O(n)$

Table 2. The time complexities of ThreshKnot using classical partition function calculation (McCaskill, 1990) and LinearPartition (Zhang et al., 2019).

2.4 Prediction Runtime

We now turn to the comparison of prediction efficiency. After obtaining base-pairing probabilities, ThreshKnot takes $O(n^2)$ time in the worst case, whereas MEA takes $O(n^3)$ time (see Table 1 for time complexities); this is indeed confirmed in practice by Figure 4A. Furthermore, Fig. SI 7 shows that with ThreshKnot, after the $O(n^2)$ threshold pruning step, the number of surviving base pair candidates scales linearly with the length of the RNA sequence (even with a small θ such as 0.01). This is because the vast majority of those $O(n^2)$ pairs have close-to-zero probabilities (also evidenced by Figure 3B in Zuber et al. (2017)). This means the core "selection" step of ThreshKnot only takes $O(n)$ time. Therefore, as summarized in Table 2, there are three steps in the whole ThreshKnot pipeline:

1. $O(n^3)$-time computation of partition function and base-pairing probabilities,
2. $O(n^2)$-time threshold pruning, and
3. $O(n)$-time final pair selection.

That being stated, in both ThreshKnot and MEA, the overall runtime is still dominated by the $O(n^3)$-time first step (see Figure 4B).

3 Discussion

In RNA secondary structure prediction, partition function-based algorithms have become increasingly

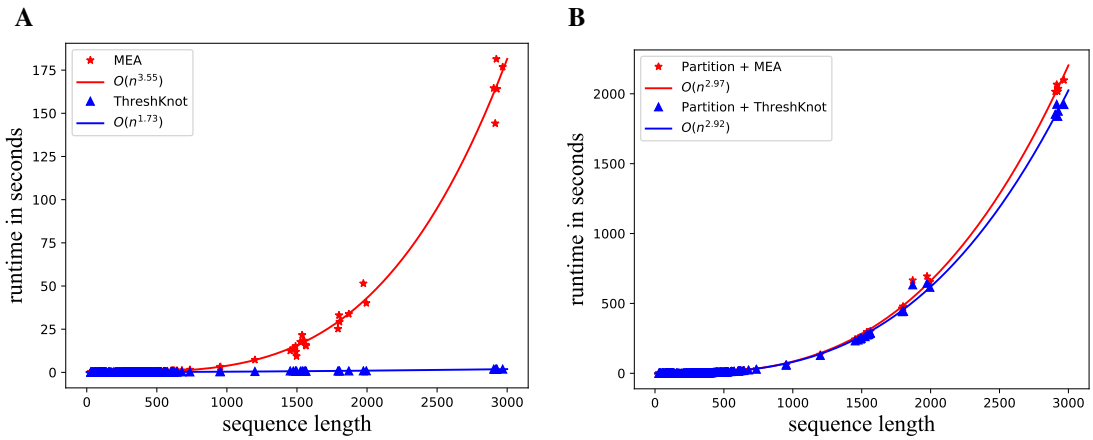

Fig. 4. Runtime comparison for the RNAstructure package on a single thread: ThreshKnot ($\theta = 0.3$) vs. MEA ($\gamma = 1.5$). After obtaining base-pairing probabilities (the Partition part in B), MEA takes $O(n^3)$ whereas ThreshKnot takes $O(n^2)$ in the worst case (also see Table 1 "time complexity"). **A**: excluding the time for computing base-pairing probabilities (ThreshKnot is substantially faster than MEA). **B**: including the time for computing base-pairing probabilities.

popular in recent years. Among these methods, MEA is popular, but our results show that ThreshKnot always outperforms MEA in all three aspects: (1) it can achieve better overall predication accuracy, (2) it can predict pseudoknots that MEA can not, (3) it is much simpler to implement and runs much faster. This suggests that ThreshKnot should replace MEA as the default partition function-based structure prediction algorithm.

The overall runtime of ThreshKnot is still dominated by the $O(n^3)$-time first step to calculate the partition function (i.e., the McCaskill (1990) algorithm). Fortunately, our forthcoming LinearPartition paper (Zhang et al., 2019) reports an $O(n)$-time algorithm to approximate the partition function inspired by the recently published LinearFold algorithm (Huang et al., 2019), and it outputs just $O(n)$ base pairs with non-zero probabilities instead of all $O(n^2)$ pairs. This implies that we can make the whole ThreshKnot pipeline run in $O(n)$ time with LinearPartition (see Table 2).

Materials and Methods

Dataset

We use the ArchiveII dataset (Sloma and Mathews, 2016), a diverse set of RNA sequences with accepted structures.[4] Following LinearFold (Huang et al., 2019), we only consider full sequences (i.e., excluding the individual folding domains of 16S/23S rRNAs) and remove those sequences found in the S-Processed set (Andronescu et al., 2007) (because CONTRAfold is trained on S-Processed). The resulting dataset contains 2,889 sequences over 9 families, with an average length of 222.2 *nt* and maximum length of 2,968 *nt*.

Software and Computing Environment

We use the following software:

- RNAstructure 6.1:
  https://rna.urmc.rochester.edu/RNAstructure.html
- CONTRAfold 2.02
  http://contra.stanford.edu/contrafold/download.html
- Vienna RNAfold 2.4.13
  https://www.tbi.univie.ac.at/RNA/
- IPknot
  https://github.com/satoken/ipknot
- pKiss
  https://bibiserv.cebitec.uni-bielefeld.de/pkiss

All software were compiled by GCC 5.4.0 on a laptop with Intel Core i7-8550U at 1.8GHz running Ubuntu 16.04.2.

Evaluation Details

We use the standard PPV and sensitivity as follows:

$$\text{PPV}(\hat{y}, y^*) = \frac{|\hat{y} \cap y^*|}{|\hat{y}|}, \quad \text{sens}(\hat{y}, y^*) = \frac{|\hat{y} \cap y^*|}{|y^*|}$$

[4] http://rna.urmc.rochester.edu/pub/archiveII.tar.gz

where $\hat{y}$ is a predicted structure and y^* is the accepted structure (both structures are treated as sets of pairs, i.e., $|\hat{y}|$ is the number of pairs in $\hat{y}$).

Following Mathews et al. (1999), we allow correctly predicted pairs to be offset by one position for one nucleotide as compared to the known structure (see Table SI 1). We also report in Table SI 2 the accuracies using exact matching.

The per-family accuracy is the mean over all sequences in that family, and the overall accuracy is the mean over per-family accuracies from all families.

We use the Jackknife resampling method to choose the best parameter (θ for ThreshKnot and γ for MEA) as follows: each time we held out one family, and evaluate the relative accuracy of ThreshKnot over MFE on the remaining 8 families with θ ranging from 0, 0.1, 0.2, 0.3, 0.4, 0.5, and 0.6. Coincidentally, in each case, the same $\theta = 0.3$ is consistently chosen as the best paramter for ThreshKnot. The same is true for $\gamma = 1.5$ for MEA. The "relative accuracy" is defined as the F-score between the difference in PPV and the difference in sensitivity:

$$\mathrm{F}(\mathrm{PPV}, \mathrm{sens}) = \frac{2 \cdot \mathrm{PPV} \cdot \mathrm{sens}}{\mathrm{PPV} + \mathrm{sens}}$$

$$\Delta\mathrm{F}\Big((\mathrm{PPV}', \mathrm{sens}'), (\mathrm{PPV}, \mathrm{sens})\Big) = \mathrm{F}(\mathrm{PPV}' - \mathrm{PPV}, \mathrm{sens}' - \mathrm{sens})$$

Where $(\mathrm{PPV}', \mathrm{sens}')$ are the PPV and sensitivity of ThreshKnot and $(\mathrm{PPV}, \mathrm{sens})$ are those of MFE (we assume $\mathrm{PPV}' > \mathrm{PPV}$ and $\mathrm{sens}' > \mathrm{sens}$).

For pseudoknot accuracy, we first define the notion of "crossing pairs", notated $\mathrm{crossing}(y)$, in a structure y to be the set of pairs that are crossed by at least one other pair:

$$\mathrm{crossing}(y) = \{(i,j) \in y \mid \exists (k,l) \in y, \; i<k<j<l \text{ or } k<i<l<j\}$$

We then restrict ourselves to comparing the crossing pairs in the predicted structure to the crossing pairs in the accepted structure, and define the pseudoknot PPV and sensitivity to be the PPV and sensitivity on those two subsets:

$$\mathrm{PPV}_{\mathrm{crossing}}(\hat{y}, y^*) = \mathrm{PPV}(\mathrm{crossing}(\hat{y}), \mathrm{crossing}(y^*))$$
$$\mathrm{sens}_{\mathrm{crossing}}(\hat{y}, y^*) = \mathrm{sens}(\mathrm{crossing}(\hat{y}), \mathrm{crossing}(y^*))$$

This means that a crossing pair in the predicted structure $\hat{y}$ is considered correct if it is also a crossing pair in the accepted structure y^*.

All statistical significance tests are done with two-sided permutation test (Aghaeepour and Hoos, 2013).

Code Availability

ThreshKnot is available in the RNAstructure software package v6.2 (released November 27, 2019):

https://rna.urmc.rochester.edu/RNAstructure.html

To run ThreshKnot in the RNAstructure package:

```
./ProbKnot --sequence <infile> <outfile> -t 0.3 -m 2
```

Where `-t` specifies a threshold probability to include a pair; `-m` specifies the minimum length accepted for a helix. We set threshold $\theta = 0.3$ and the minimum helix length as 2 for ThreshKnot using RNAstructure.

Data Availability

The data that support our findings are available from the corresponding author upon request.

ACKNOWLEDGMENTS. This work was partially supported by NSF grant IIS-1817231 (L.H.) and NIH grant R01 GM076485 (D.H.M.).

References

Aghaeepour, N., Hoos, H. H., 2013. Ensemble-based prediction of RNA secondary structures. BMC Bioinformatics 14 (1), 139.

Andronescu, M., Condon, A., Hoos, H. H., Mathews, D. H., Murphy, K. P., 2007. Efficient parameter estimation for RNA secondary structure prediction. Bioinformatics 23 (13), i19–i28.

Bachellerie, J. P., Cavaillé, J., Hüttenhofer, A., 2002. The expanding snoRNA world. Biochimie 84 (8), 775–790.

Bellaousov, S., Mathews, D. H., 2010. ProbKnot: fast prediction of RNA secondary structure including pseudoknots. RNA 16 (10), 1870–1880.

Cordero, P., Das, R., 2015. Rich RNA structure landscapes revealed by mutate-and-map analysis. PLOS Computational Biology 11 (11).

Do, C. B., Woods, D. A., Batzoglou, S., 2006. CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics 22 (14), e90–e98.

Doudna, J. A., Cech, T. R., 2002. The chemical repertoire of natural ribozymes. Nature 418 (6894), 222–228.

Hajiaghayi, M., Condon, A., Hoos, H. H., 2012. Analysis of energy-based algorithms for RNA secondary structure prediction. BMC Bioinformatics 13 (22), 1.

Huang, L., Zhang, H., Deng, D., Zhao, K., Liu, K., Hendrix, D. A., Mathews, D. H., 2019. LinearFold: linear-time approximate RNA folding by 5'-to-3' dynamic programming and beam search. Bioinformatics 35 (14), i295–i304.

Karijolich, J., Yi, C., Yu, Y.-T., 2015. Transcriptome-wide dynamics of RNA pseudouridylation. Nature Reviews Molecular Cell Biology 16 (10), 581–585.

Knudsen, B., Hein, J., 2003. Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Research 31 (13), 3423–3428.

Lai, W.-J. C., Kayedkhordeh, M., Cornell, E. V., Farah, E., Bellaousov, S., Rietmeijer, R., Salsi, E., Mathews, D. H., Ermolenko, D. N., 2018. mRNAs and lncRNAs intrinsically form secondary structures with short end-to-end distances. Nature Communications 9 (1), 4328.

Long, D., Lee, R., Williams, P., Chan, C. Y., Ambros, V., Ding, Y., 2007. Potent effect of target structure on microrna function. Nature Structural & Molecular Biology 14 (4), 287.

Lorenz, R., Bernhart, S. H., Zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., Hofacker, I. L., 2011. ViennaRNA package 2.0. Algorithms for Molecular Biology 6 (1), 1.

Lu, Z. J., Gloor, J. W., Mathews, D. H., 2009. Improved RNA secondary structure prediction by maximizing expected pair accuracy. RNA 15 (10), 1805–1813.

Lu, Z. J., Mathews, D. H., 2008. Efficient siRNA selection using hybridization thermodynamics. Nucleic Acids Research 36 (2), 640–647.

Mathews, D. H., Sabina, J., Zuker, M., Turner, D. H., 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. Journal of Molecular Biology 288 (5), 911–940.

McCaskill, J. S., 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 29 (6-7), 1105–1119.

Miao, Z., Adamiak, R. W., Antczak, M., Batey, R. T., Becka, A. J., Biesiada, M., Boniecki, M. J., Bujnicki, J. M., Chen, S.-J., Cheng, C. Y., Chou, F.-C., Ferré-D'Amaré, A. R., Das, R., Dawson, W. K., Ding, F., Dokholyan, N. V., Dunin-Horkawicz, S., Geniesse, C., Kappel, K., Kladwang, W., Krokhotin, A., Łach, G. E., Major, F., Mann, T. H., Magnus, M., Pachulska-Wieczorek, K., Patel, D. J., Piccirilli, J. A., Popenda, M., Purzycka, K. J., Ren, A., Rice, G. M., Jr., J. S., Sarzynska, J., Szachniuk, M., Tandon, A., Trausch, J. J., Tian, S., Wang, J., Weeks, K. M., II, B. W., Xiao, Y., Xu, X., Zhang, D., Zok, T., Westhof, E., 2017. RNA-puzzles round III: 3D RNA structure prediction of five riboswitches and one ribozyme. RNA 23 (5), 655–672.

Nawrocki, E. P., Eddy, S. R., 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29 (22), 2933–2935.

Nussinov, R., Jacobson, A. B., 1980. Fast algorithm for predicting the secondary structure of single-stranded RNA. Proceedings of the National Academy of Sciences 77 (11), 6309–6313.

Ognjenović, J., Grisshammer, R., Subramaniam, S., 2019. Frontiers in cryo electron microscopy of complex macromolecular assemblies. Annual Review of Biomedical Engineering 21, 395–415.

Parisien, M., Major, F., 2008. The mc-fold and mc-sym pipeline infers RNA structure from sequence data. Nature 452 (7183), 51.

Reeder, J., Giegerich, R., 2004. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. BMC Bioinformatics 5 (1), 1.

Reuter, J. S., Mathews, D. H., 2010. RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics 11 (1), 129.

Sato, K., Kato, Y., Hamada, M., Akutsu, T., Asai, K., 2011. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. Bioinformatics 27 (13), i85–i93.

Seetin, M. G., Mathews, D. H., 2011. Automated RNA tertiary structure prediction from secondary structure and low-resolution restraints. Journal of Computational Chemistry 32 (10), 2232–2244.

Serganov, A., Nudler, E., 2013. A decade of riboswitches. Cell 152 (1), 17–24.

Sloma, M., Mathews, D., 2016. Exact calculation of loop formation probability identifies folding motifs in RNA secondary structures. RNA, 22, 1808–1818.

Storz, G., Gottesman, S., 2006. Versatile roles of small RNA regulators in bacteria. In: Gesteland, R., Cech, T., Atkins, J. (Eds.), The RNA World, 3rd Edition. Cold Spring Harbor Laboratory Press, New York, pp. 567–594.

Tafer, H., Ameres, S. L., Obernosterer, G., Gebeshuber, C. A., Schroeder, R., Martinez, J., Hofacker, I. L., 2008. The impact of target site accessibility on the design of effective siRNAs. Nature biotechnology 26 (5), 578–583.

Theis, C., Janssen, S., Giegerich, R., 2010. Prediction of RNA secondary structure including kissing hairpin motifs. In: International Workshop on Algorithms in Bioinformatics. Springer, pp. 52–64.

Tinoco, I., Bustamante, C., 1999. How RNA folds. Journal of Molecular Biology 293 (2), 271–281.

Tukey, J., 1958. Bias and confidence in not quite large samples. Annals of Mathematical Statistics 29, 614.

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

Zhang, H., Keane, S. C., 2019. Advances that facilitate the study of large RNA structure and dynamics by nuclear magnetic resonance spectroscopy. Wiley Interdisciplinary Reviews: RNA, e1541.

Zhang, H., Zhang, L., Mathews, D. H., Huang, L., 2019. LinearPartition: Linear-Time approximation of RNA folding partition function and base pairing probabilities. arXiv preprint 1912.13190 https://arxiv.org/abs/1912.13190.

Zhang, J., Ferré-D'Amaré, A. R., 2014. New molecular engineering approaches for crystallographic studies of large RNAs. Current Opinion in Structural Biology 26, 9–15.
Zuber, J., Sun, H., Zhang, X., McFadyen, I., Mathews, D. H., 2017. A sensitivity analysis of RNA folding nearest neighbor parameters identifies a subset of free energy parameters with the greatest impact on RNA secondary structure prediction. Nucleic Acids Research 45 (10), 6168–6176.
Zuker, M., Stiegler, P., 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Research 9 (1), 133–148.

Supporting Information

ThreshKnot: Thresholded ProbKnot for Improved RNA Secondary Structure Prediction

Liang Zhang, He Zhang, David H. Mathews, and Liang Huang

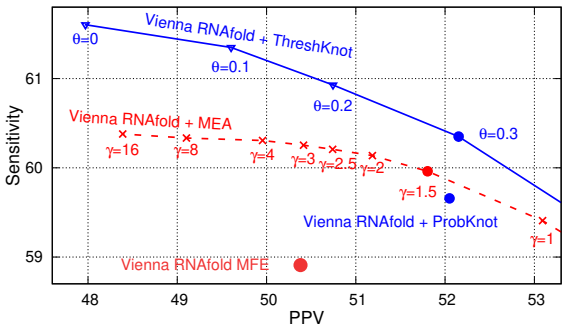

Fig. SI 1. Comparison between ThreshKnot, ProbKnot, MFE, and MEA on ViennaRNA using the ArchiveII dataset. ThreshKnot has better PPV and Sensitivity than all other methods. We also add ProbKnot for comparison.

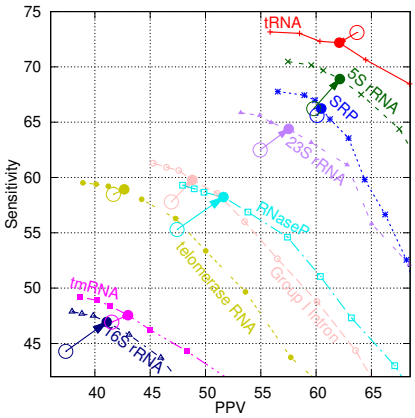

Fig. SI 2. ThreshKnot improves 8 out of 9 families over MFE (in both PPV and sensitivity) on Vienna RNAfold. The curves show the ThreshKnot accuracies with varying θ. The arrows point from MFE (hollow circles) to ThreshKnot at $\theta = 0.3$.

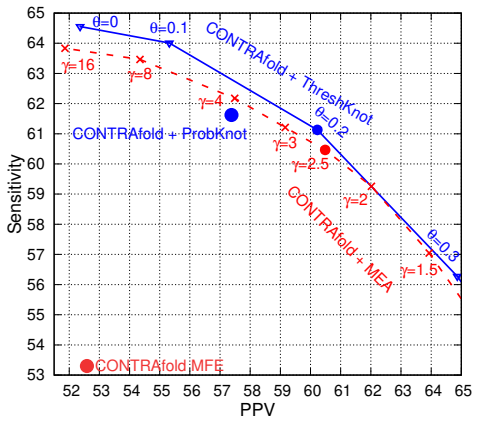

Fig. SI 3. Comparison between ThreshKnot, ProbKnot, MFE, and MEA on CONTRAfold using the ArchiveII dataset. ThreshKnot has better PPV and Sensitivity than all other methods.

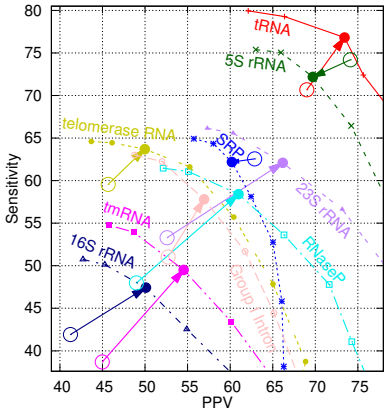

Fig. SI 4. ThreshKnot improves 7 out of 9 families over MFE (in both PPV and sensitivity) on CONTRAfold. The curves show the ThreshKnot accuracies with varying θ. The arrows point from MFE (hollow circles) to ThreshKnot at $\theta = 0.2$.

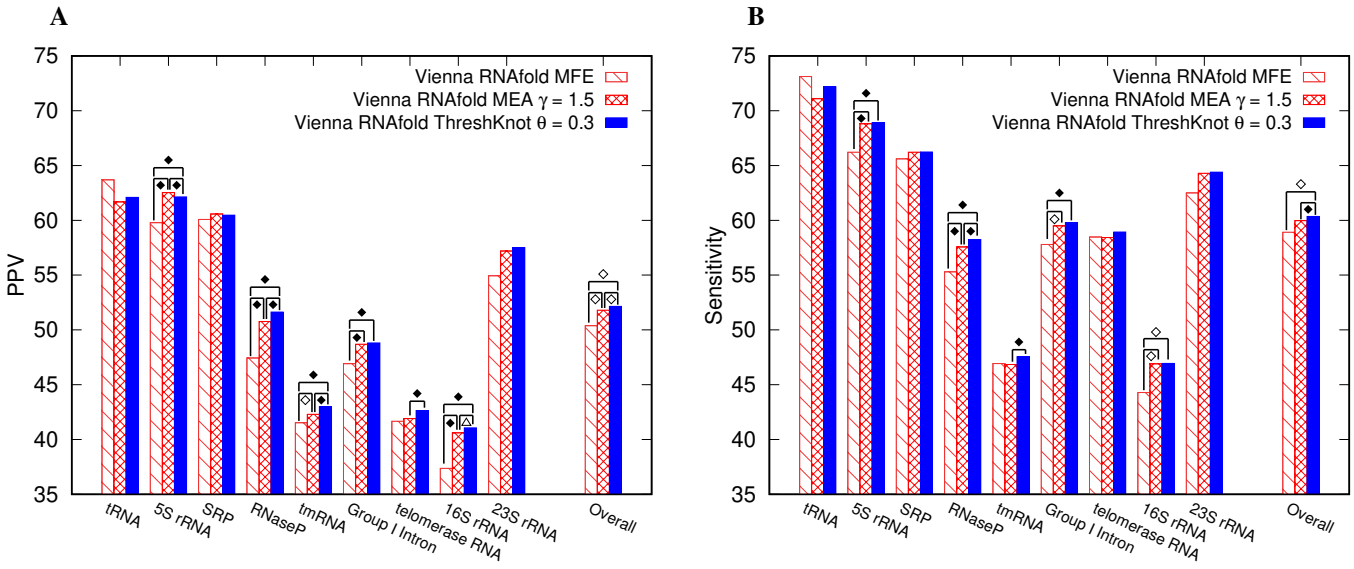

Fig. SI 5. Accuracy results of MFE, MEA, and ThreshKnot using Vienna RNAfold. In both panels, the first nine bars from the left represent PPV (A) and sensitivity (B) averaged over all sequences in one family, and the rightmost bars represent the overall accuracies, averaging over all families. Statistical significance (two-sided) is marked as $^{\blacklozenge}(p<0.01)$, $^{\diamond}(0.01\leq p<0.05)$, or $^{\triangle}(0.05\leq p<0.06)$.

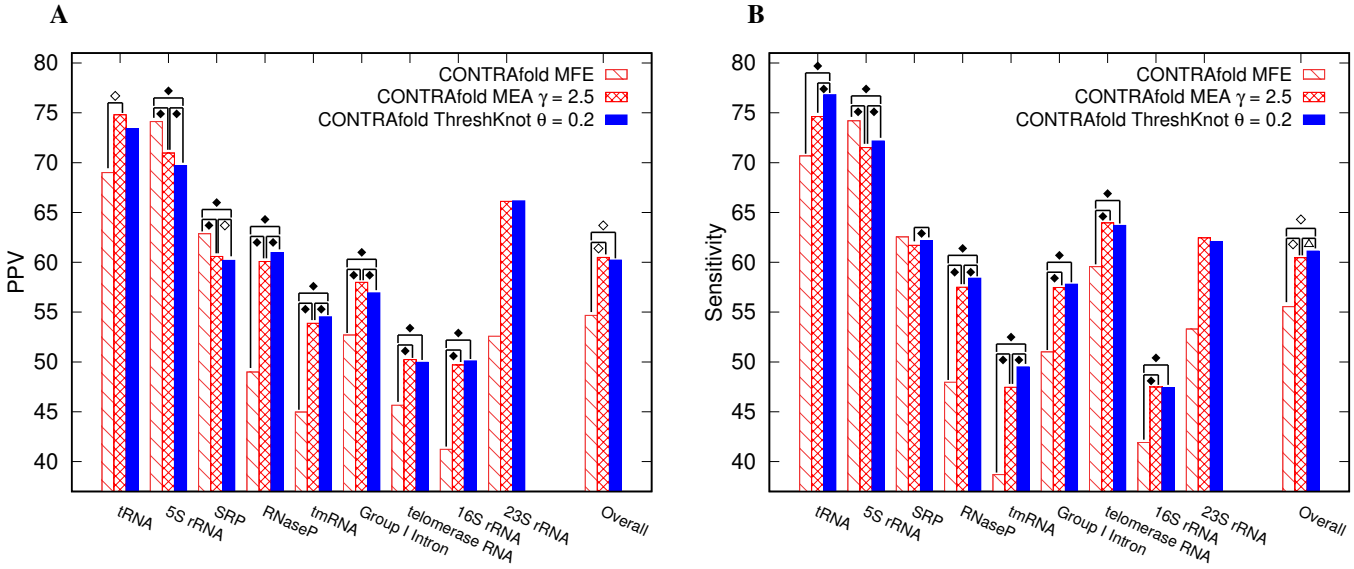

Fig. SI 6. Accuracy results of MFE, MEA, and ThreshKnot using CONTRAfold. In both panels, the first nine bars from the left represent PPV (A) and sensitivity (B) averaged over all sequences in one family, and the rightmost bars represent the overall accuracies, averaging over all families. Statistical significance (two-sided) is marked as $^{\blacklozenge}(p<0.01)$, $^{\diamond}(0.01\leq p<0.05)$, or $^{\triangle}(0.05\leq p<0.06)$.

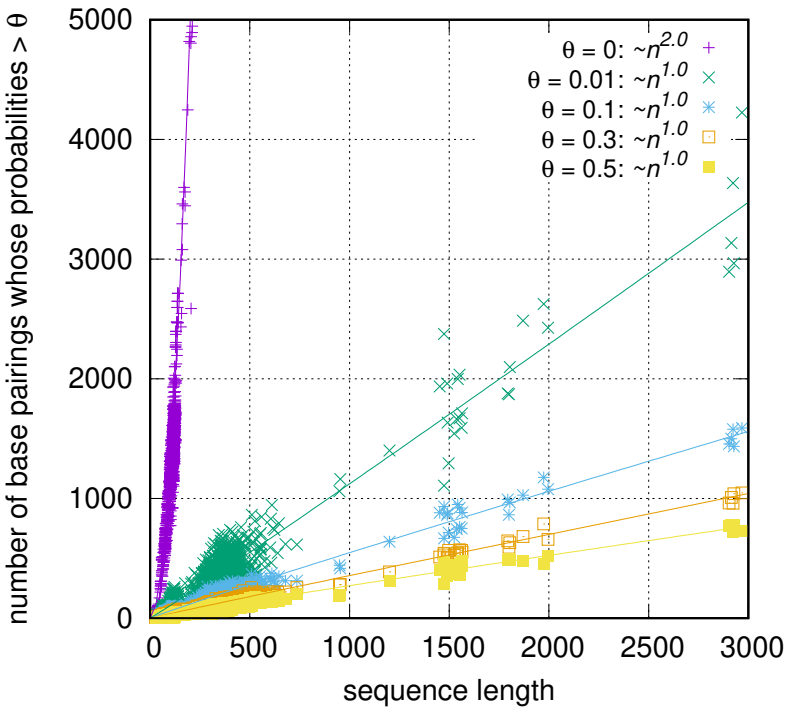

Fig. SI 7. The number of base pairs whose probabilities are greater than the threshlds θ in the base pairing probability matrix from RNAstructure. With any non-zero θ, the number of surviving base pairs is linear in sequence length.

				5S				Group I	telomerase	16S	23S	
		Family	tRNA	rRNA	SRP	RNaseP	tmRNA	Intron	RNA	rRNA	rRNA	Overall
		total seqs	557	1,283	928	454	462	98	37	22	5	3,846
		used seqs	74	1,125	886	182	462	96	37	22	5	2,889
		avg. length of used seqs	77.3	118.8	186.1	344.1	366.0	424.9	444.6	1,547.9	2,927.4	222.2
RNAstructure	MFE	PPV	62.77	59.01	60.18	48.36	40.87	45.84	42.37	38.34	51.02	49.86
		sens	69.73	64.54	65.56	55.36	45.93	56.22	59.15	45.19	57.73	57.71
	MEA	PPV	68.11	58.73	58.41	52.42	42.17	47.87	41.86	41.98	56.25	51.98
	$\gamma = 1.5$	sens	75.45	64.01	63.35	59.05	46.35	57.59	58.03	47.96	62.01	59.31
	ThreshKnot	PPV	66.39	58.56	58.50	52.80	42.65	47.91	42.25	42.27	56.27	51.96
	$\theta = 0.3$	sens	75.44	64.53	63.78	59.33	47.19	58.14	58.38	48.14	61.84	59.64
RNAfold	MFE	PPV	63.69	59.79	60.08	47.43	41.53	46.91	41.67	37.37	54.94	50.38
		sens	73.11	66.22	65.61	55.30	46.93	57.80	58.48	44.29	62.49	58.91
	MEA	PPV	61.68	62.52	60.56	50.76	42.30	48.70	41.91	40.61	57.19	51.80
	$\gamma = 1.5$	sens	71.09	68.80	66.22	57.57	46.85	59.49	58.43	46.90	64.29	59.96
	ThreshKnot	PPV	62.09	62.14	60.46	51.63	43.00	48.81	42.65	41.07	57.51	52.15
	$\theta = 0.3$	sens	72.18	68.90	66.24	58.23	47.56	59.78	58.93	46.94	64.39	60.35
CONTRAfold	MFE	PPV	69.00	74.12	62.87	48.99	44.97	52.71	45.67	41.23	52.59	54.68
		sens	70.67	74.20	62.55	47.98	38.69	51.01	59.56	41.92	53.30	55.54
	MEA	PPV	74.80	70.96	60.58	60.09	53.89	58.00	50.23	49.70	66.13	60.49
	$\gamma = 2.5$	sens	74.63	71.52	61.71	57.51	47.46	57.48	63.95	47.49	62.45	60.47
	ThreshKnot	PPV	73.43	69.71	60.20	60.98	54.53	56.94	49.97	50.11	66.17	60.23
	$\theta = 0.2$	sens	76.82	72.18	62.19	58.41	49.51	57.80	63.70	47.43	62.10	61.13
IPknot		PPV	81.89	62.53	56.63	65.24	55.71	55.98	43.00	52.96	68.07	60.22
		sens	80.25	50.12	49.07	56.89	43.15	48.80	44.44	41.45	48.98	51.46
pKiss		PPV	47.82	47.19	54.45	41.65	36.70	47.09	38.58	38.63	46.74	44.32
		sens	55.16	50.45	59.03	46.80	40.93	57.70	53.38	44.48	51.32	51.03

Table SI 1. Detailed overall prediction accuracies, allowing one nucleotide in a pair to be displaced by one position, on the ArchiveII dataset. This slipping method (Mathews et al., 1999) considers a base pair to be correct if it is slipped by one nucleotide on a strand.

				5S				Group I	telomerase	16S	23S	
		Family	tRNA	rRNA	SRP	RNaseP	tmRNA	Intron	RNA	rRNA	rRNA	Overall
		total seqs	557	1,283	928	454	462	98	37	22	5	3,846
		used seqs	74	1,125	886	182	462	96	37	22	5	2,889
		avg. length of used seqs	77.3	118.8	186.1	344.1	366.0	424.9	444.6	1,547.9	2,927.4	222.2
RNAstructure	MFE	PPV	61.49	56.55	56.84	46.46	38.65	44.13	39.99	36.52	48.86	47.72
		sens	68.39	61.77	61.67	53.08	43.41	54.07	55.79	43.05	55.28	55.17
	MEA	PPV	65.95	56.36	55.17	50.40	39.58	46.23	39.45	40.43	54.31	49.76
	$\gamma = 1.5$	sens	73.01	61.34	59.67	56.74	43.50	55.64	54.63	46.17	59.88	56.73
	ThreshKnot	PPV	64.15	56.25	55.28	50.83	40.06	46.25	39.90	40.70	54.39	49.76
	$\theta = 0.3$	sens	72.87	61.89	60.10	57.07	44.32	56.12	55.07	46.35	59.78	57.06
RNAfold	MFE	PPV	61.75	57.28	56.58	45.76	39.75	45.49	39.53	35.65	53.20	48.33
		sens	70.98	63.35	61.55	53.28	44.90	56.06	55.40	42.26	60.50	56.48
	MEA	PPV	59.72	60.08	57.13	48.98	40.53	47.13	39.42	38.84	55.54	49.71
	$\gamma = 1.5$	sens	68.89	66.01	62.22	55.48	44.88	57.60	54.89	44.85	62.43	57.47
	ThreshKnot	PPV	60.20	59.73	57.07	49.87	41.20	47.28	40.20	39.38	55.80	50.08
	$\theta = 0.3$	sens	70.04	66.12	62.30	56.19	45.55	57.92	55.50	45.01	62.46	57.90
CONTRAfold	MFE	PPV	67.61	70.68	59.14	47.45	42.96	51.21	43.40	39.84	50.56	52.54
		sens	69.12	70.70	58.61	46.39	36.94	49.56	56.58	40.49	51.24	53.29
	MEA	PPV	73.56	67.94	57.06	58.26	51.68	56.43	47.45	48.09	64.15	58.29
	$\gamma = 2.5$	sens	73.27	68.31	57.94	55.62	45.50	55.94	60.37	45.95	60.56	58.16
	ThreshKnot	PPV	72.19	67.01	56.85	59.17	52.36	55.49	47.28	48.61	64.28	58.14
	$\theta = 0.2$	sens	75.47	69.24	58.56	56.53	47.53	56.36	60.21	46.01	60.30	58.91
IPknot		PPV	80.28	59.65	54.03	63.62	53.93	54.41	40.92	51.78	66.28	58.33
		sens	78.51	47.66	46.66	55.37	41.73	47.48	42.25	40.51	47.67	49.76
pKiss		PPV	45.90	45.14	51.04	40.19	34.56	45.62	37.11	37.21	44.88	42.41
		sens	53.04	48.17	55.13	45.09	38.54	55.90	51.29	42.83	49.27	48.80

Table SI 2. Detailed overall prediction accuracies on the ArchiveII dataset. The accuracies use exact base-pair matching.

	family	tRNA	5S rRNA	SRP	RNaseP	tmRNA	Group I Intron	telomerase RNA	16S rRNA	23S rRNA	Overall
	gold base pairs	1,496	37,727	49,680	17,308	45,332	9,669	3,774	9,135	4,091	178,212
	gold crossing pairs	0	0	0	4,538	26,153	1,164	1,015	568	443	33,881
	predicted base pairs	1,734	41,755	54,455	19,527	50,153	12,433	5,278	10,699	4,498	200,532
RNAstructure	predicted crossing pairs	167	2,064	3,218	1,254	4,510	929	407	880	445	13,874
+ ThreshKnot	correct crossing pairs	0	0	0	139	983	48	13	6	17	1,206
$\theta = 0.3$	PPV	0	0	0	11.08	21.80	5.17	3.19	0.68	3.82	7.62
	sens	NA	NA	NA	3.06	3.76	4.12	1.28	1.06	3.84	2.85
	predicted base pairs	1,776	41,986	54,907	19,575	50,054	12,522	5,272	10,653	4,583	201,328
RNAfold	predicted crossing pairs	183	2,867	2,716	1,263	4,882	945	276	1,013	325	14,470
+ ThreshKnot	correct crossing pairs	0	0	0	185	965	61	5	22	5	1,243
$\theta = 0.3$	PPV	0	0	0	14.65	19.77	6.46	1.81	2.17	1.54	7.73
	sens	NA	NA	NA	4.08	3.69	5.24	0.49	3.87	1.13	3.08
	predicted base pairs	1,610	38,798	50,296	16,756	40,939	10,266	4,808	8,901	3,837	176,211
CONTRAfold	predicted crossing pairs	319	4,998	5,684	1,912	7,893	1,140	650	1,263	561	24,420
+ ThreshKnot	correct crossing pairs	0	0	0	307	2,741	111	48	13	18	3,238
$\theta = 0.2$	PPV	0	0	0	16.06	34.73	9.74	7.38	1.03	3.21	12.02
	sens	NA	NA	NA	6.77	10.48	9.54	4.73	2.29	4.06	6.31
	predicted base pairs	1,494	30,680	41,545	15,165	34,982	8,745	3,874	7,256	2,947	146,688
	predicted crossing pairs	140	3,664	5,499	1,770	6,407	1,096	712	982	358	20,628
IPknot	correct crossing pairs	0	0	0	470	2,155	78	75	37	55	2,870
	PPV	0	0	0	26.55	33.64	7.12	10.53	3.77	15.36	16.16
	sens	NA	NA	NA	10.36	8.24	6.70	7.39	6.51	12.42	8.60
	predicted base pairs	1,755	40,106	54,149	19,596	50,505	12,598	5,301	10,766	4,486	199,262
	predicted crossing pairs	516	8,686	10,284	5,309	21,728	4,199	1,374	3,009	1,416	56,521
pKiss	correct crossing pairs	0	0	0	356	7,273	289	74	76	47	8,115
	PPV	0	0	0	6.71	33.47	6.88	5.39	2.53	3.32	9.72
	sens	NA	NA	NA	7.84	27.81	24.83	7.29	13.38	10.61	15.29

Table SI 3. Detailed pseudoknots prediction accuracies, allowing one nucleotide in a pair to be displaced by one position, on the ArchiveII dataset. This slipping method considers a base pair to be correct if it is slipped by one nucleotide on a strand. For pseudoknot prediction accuracy, we compare all crossing pairs in the predicted structure $\hat{y}$ with all crossing pairs in the accepted structure y^*. A crossing pair in predicted structure $\hat{y}$ is considered correct if it is also a crossing pair in the accepted structure y^*.