AN ABSTRACT OF THE THESIS OF

Liang Zhang for the degree of Master of Science in Computer Science presented on November 26, 2019.

Title: ThreshKnot: Thresholded ProbKnot for Improved RNA Secondary Structure Prediction

Abstract approved: ______________________________

Liang Huang

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 based on base-pairing probabilities. Two typical examples of the latter group are the popular maximum expected accuracy (MEA) method and the ProbKnot method. ProbKnot is 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 probability falls 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 structure prediction accuracy, in its capability to predict pseudoknots, and in its faster running time. This suggests that ThreshKnot should replace MEA as the default partition function-based structure prediction algorithm.

ThreshKnot: Thresholded ProbKnot for Improved RNA Secondary Structure Prediction

by

Liang Zhang

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented November 26, 2019
Commencement June 2020

Master of Science thesis of Liang Zhang presented on November 26, 2019

APPROVED:

Major Professor, representing Computer Science

Head of the School of Electrical Engineering and Computer Science

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Liang Zhang, Author

ACKNOWLEDGEMENTS

I am deeply thankful to my advisor, Professor Liang Huang, for his advice and encouragement. The completion of my thesis would not have been possible without the guidance of Professor Huang.

I would like to express my deepest appreciation to my committee and my Graduate Council Representative Professor Weng-Keen Wong, Professor Xiaoli Z. Fern, Professor Prasad Tadepalli, and Professor Leonard Coop. Thanks for their valuable advice and practical suggestions.

I would also like to thank He Zhang from our research group, who helped me a lot on this research.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THRESHKNOT: THRESHOLDED PROBKNOT FOR IMPROVED RNA SECONDARY STRUCTURE PREDICTION

1. INTRODUCTION

RNAs are involved in multiple processes, such as guiding RNA modifications [1] and regulating a particular disease [2], and their functionalities are highly related to structures. However, physical structure determine techniques, such as X-ray crystallography [3] or Nuclear Magnetic Resonance (NMR) [4], and chemical probing methods [5], though reliable and accurate, are slow and costly. Therefore, fast and accurate computational prediction of RNA structure is useful and desired. Since predicting tertiary structure is challenging [6], 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) [7]. 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 [8, 9, 10, 11].

Most algorithms for RNA secondary structure prediction can be divided into two camps, the classical ones computing a single structure with the minimum free energy (MFE) [12, 13], 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 [14]. 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 [15, 16, 17], (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 [18, 19, 20, 21], (3) we can 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] in $O(n^3)$ time [22, 23].

Two typical (and widely used) examples of partition function-based prediction algorithms are maximum expected accuracy (MEA) [15] and ProbKnot [22]. Both of them 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 much simpler heuristic method that only needs $O(n^2)$ time, and more importantly, it can predict pseudoknots which the former can not.

However, the full potential of ProbKnot has not been fully exploited. In particular, unlike MEA, ProbKnot lacks a hyperparameter to balance the positive predictive value (PPV; a.k.a. precision) and sensitivity (a.k.a. recall) of the output structure. 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 [24], Vienna RNAfold [25], and CONTRAfold [15]), ThreshKnot always outperforms the much more involved MEA algorithm in all three aspects: (1) it can achieve better overall predication 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 $i < k < j < l$.

2. ALGORITHM

2.1. ThreshKnot Algorithm and Pseudocode

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[2]

$$\{(i, j) \mid p(i, j) = \max_k p(i, k) = \max_k p(k, j) \text{ and } p(i, j) \geq \theta\}$$

We split ThreshKnot algorithm into two parts: pruning and selection. The pseudocode of ThreshKnot is as follows:

Algorithm 1 ThreshKnot

```
$P$: base pairing probabilities of an RNA sequence
$P_{max}(i)$: base pairing probability of the most probable pair for nucleotide $i$
$\theta$: probability threshold

procedure PRUNING($P$, $\theta$)
    for each $(i, j)$ in $P$ do
        if $P(i, j) < \theta$ then remove $(i, j)$ from $P$

procedure SELECTION($P$)
    for each $i$ do
        $P_{max}(i) = \max(\max_{i<j\leq n} p(i, j), \max_{1\leq j<i} p(j, i))$
    for each $(i, j)$ in $P$ do
        if $P(i, j) = P_{max}(i) = P_{max}(j)$ then yield $(i, j)$
```

We show a predicted structure sample in Fig. 2.1. These output base pairs are the "most probable" ones whose probabilities are greater than the given threshold θ.

[2] To keep it simple, unlike ProbKnot, ThreshKnot does not remove "helices composed of two stacked pairs".

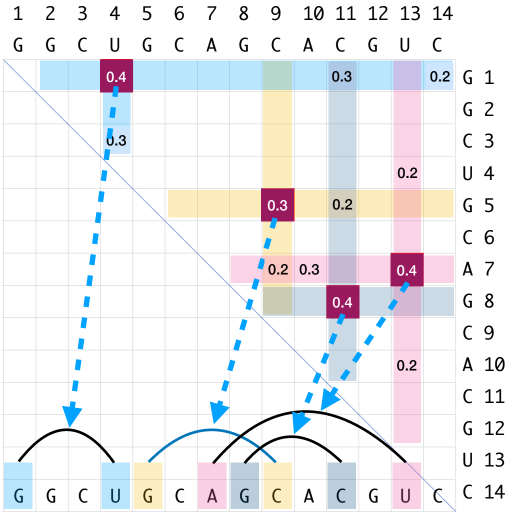

Figure 2.1: An example of ThreshKnot prediction with $\theta = 0.1$. ThreshKnot only outputs the "most probable base pairs".

2.2. Prediction Runtime

After obtaining base-pairing probabilities, ThreshKnot takes $O(n^2)$ time in the worst case, whereas MEA takes $O(n^3)$ time (see Table 3.1 for time complexities); this is indeed confirmed in practice by Figure 2.2A. Furthermore, Fig. 2.3 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 4.1, 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.

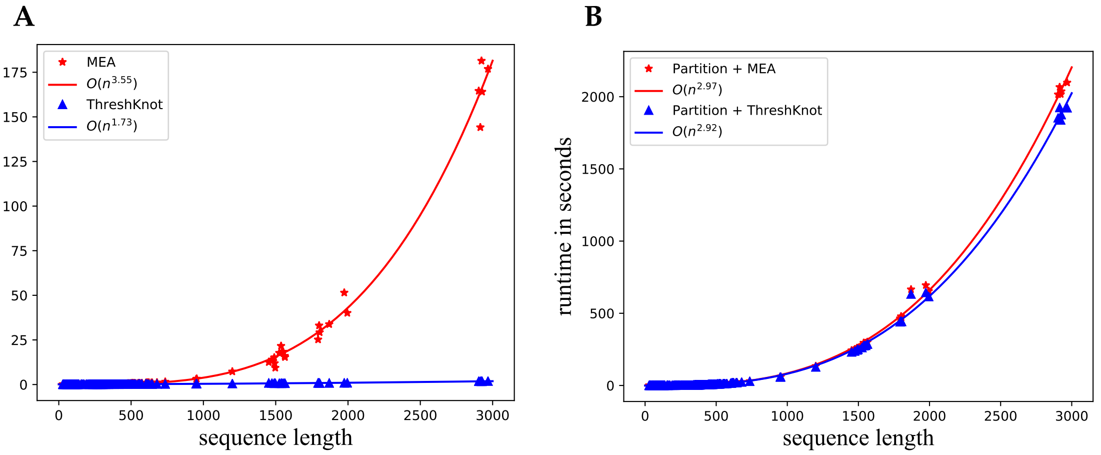

Figure 2.2: Runtime comparison: ThreshKnot (θ = 0.3) vs. MEA (γ = 1.5). **A**: excluding the time for computing base-pairing probabilities (ThreshKnot is substantially faster than MEA). **B**: including the time for computing base-pairing probabilities.

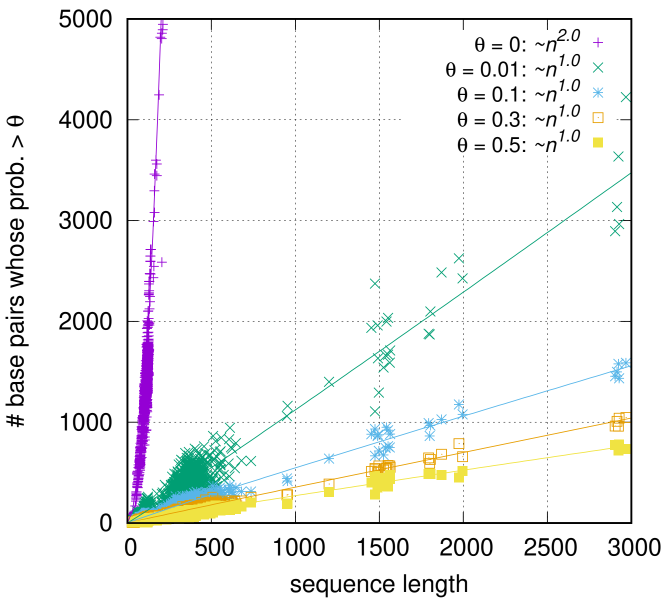

Figure 2.3: The number of base pairs whose probabilities $>$ threshlds θ

That being stated, in both ThreshKnot and MEA, the overall runtime is still dominated by the $O(n^3)$-time first step (see Figure 2.2B).

3. EXPERIMENTS

3.1. Dataset

We use the ArchiveII dataset [27], a diverse set of RNA sequences with accepted structures.[3] Following LinearFold [28], we only consider full sequences (i.e., excluding the individual folding domains 16S/23S rRNAs) and remove those sequences found in the S-Processed set [29] (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*.

3.2. Software and Computing Environment

We use the following software:

1) RNAstructure 6.1: https://rna.urmc.rochester.edu/RNAstructure.html

2) CONTRAfold 2.02: http://contra.stanford.edu/contrafold/download.html

3) Vienna RNAfold 2.4.13: https://www.tbi.univie.ac.at/RNA/

4) IPknot: https://github.com/satoken/ipknot

5) 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.

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

3.3. Evaluation Methods

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 [31] 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:

$$\begin{aligned} F(P,R) &= \frac{2PR}{P+R} \\ \Delta F\big((P',R'),(P,R)\big) &= F(P'-P, R'-R) \end{aligned}$$

Where (P', R') are the PPV and sensitivity of ThreshKnot and (P, R) are the PPV and sensitivty of MFE (we assume $P' > P$ and $R' > R$).

For pseudoknot accuracy, we use the PPV and sensitivity of "crossing pairs", i.e., we restrict ourselves to comparing the set of crossing pairs in the predicted structure to the set of crossing pairs in the accepted structure, and a crossing pair in 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.

3.4. Overall Prediction Accuracy

Below we show ThreshKnot results using the base-pairing matrices generated by RNAstructure. Figure 3.1 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.

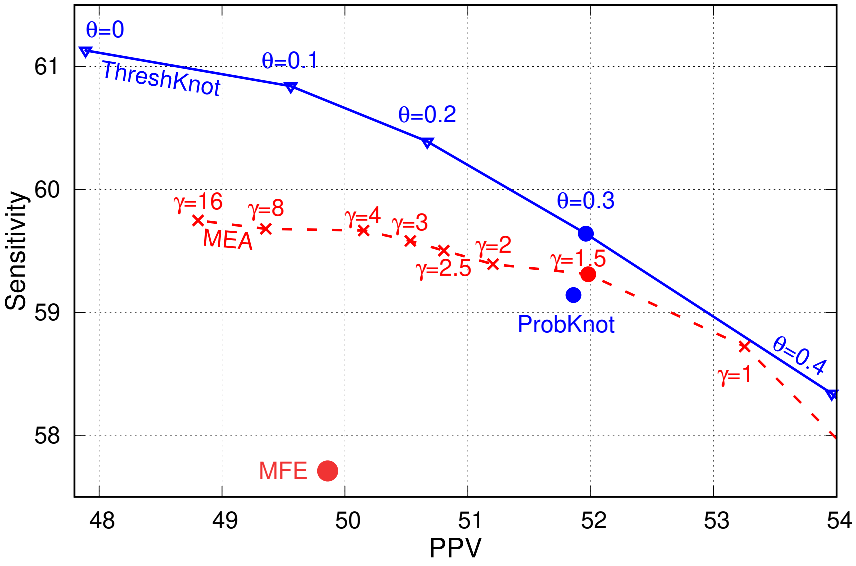

Figure 3.1: Comparison: ThreshKnot, minimum free energy structure prediction (MFE), MEA, and ProbKnot on RNAstructure. ThreshKnot has better PPV and Sensitivity than all other methods.

Figure 3.1 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 to choose the best parameter θ for ThreshKnot and γ for MEA (see Evaluation Methods), i.e. the parameter that maximizes the F score (harmonic mean of sensitivity and PPV). The same $\theta = 0.3$ is chosen con-

			overall		pseudoknot	
		time complexity	PPV	sens.	PPV	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	$O(n^3)+$ILP time	**60.22**	51.46	**16.16**	8.60
	pKiss	$O(n^4)$	44.32	51.03	9.72	**15.29**

Table 3.1: 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 time denotes the time to solve the integer linear program in IPknot.

sistently 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 3.1 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 3.2 details the accuracies on each family and the statistical significance tests.

Table 3.1 also includes two other systems: IPknot [23] and pKiss[4], both of which use 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 Figure 3.3 and Figure 3.4). pKiss, on the other hand, has substantially lower PPV and Sensitivities.

[4] pKiss is the successor of pknotsRG [32]

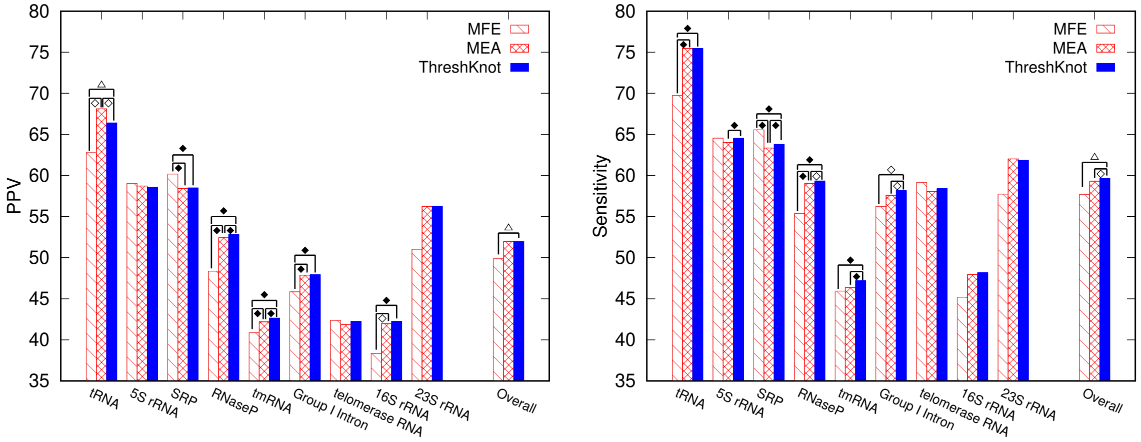

Figure 3.2: Accuracy results of MFE, MEA (γ=1.5), and ThreshKnot (θ=0.3) on RNAstructure. The first nine bars from the leftmost represent PPV (left plot) and sensitivity (right plot) averaged over all sequences in one family. The rightmost bars represent the overall accuracies, averaging over all families. Statistical significance (two-sided) is marked as $^{\blacklozenge}(p<0.01)$, $^{\diamond}(0.01 \leq p < 0.05)$, or $^{\triangle}(0.05 \leq p < 0.06)$.

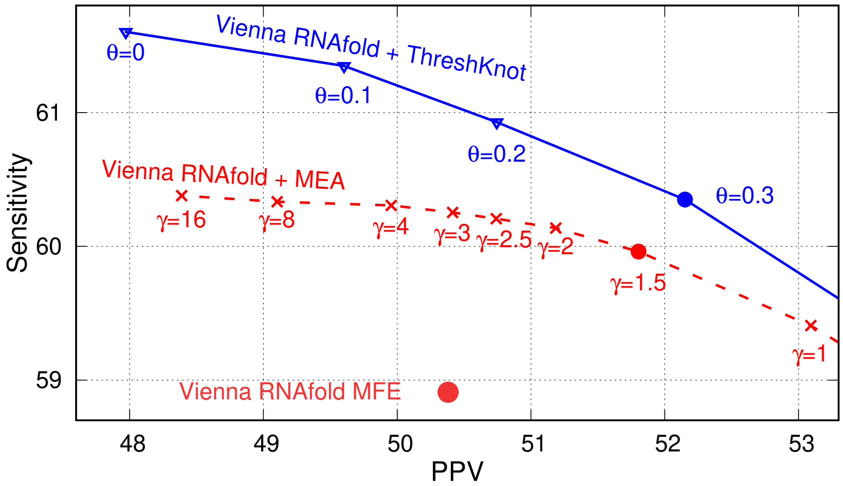

Figure 3.3: Comparison: ThreshKnot, MFE, and MEA on Vienna RNAfold. ThreshKnot has better PPV and Sensitivity than all other methods.

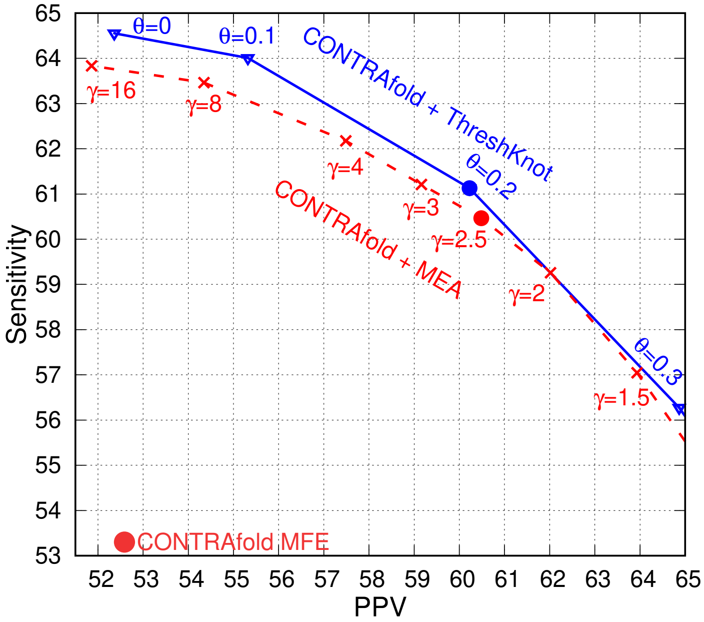

Figure 3.4: Comparison: ThreshKnot, MFE, and MEA on CONTRAfold. ThreshKnot has better PPV and Sensitivity than all other methods.

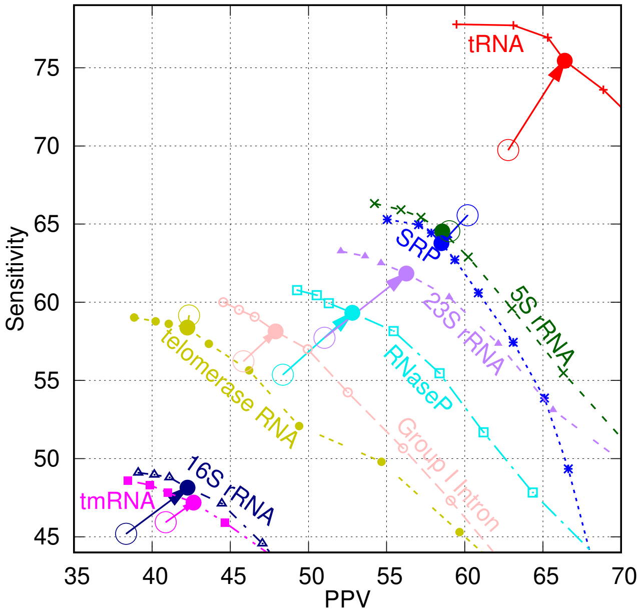

Figure 3.5: ThreshKnot improves 6 out of 9 families over MFE (in both PPV and sensitivity) on RNAstructure. The curves show the ThreshKnot accuracies with varying θ. The arrows point from MFE (hollow circles) to ThreshKnot at $\theta = 0.3$.

Figure 3.5 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).

We also test ThreshKnot on Vienna RNAfold and CONTRAfold. ThreshKnot keeps outperforming other systems and engines on CONTRAfold and Vienna RNAfold in terms of overall accuracy. In addition, ThreshKnot improves 8 out of 9 families over MFE on Vienna RNAfold and improves 7 out of 9 families over MFE on CONTRAfold in both PPV and sensitivity, which are better than the results on RNAstructure (See Figure 3.6, Figure 3.7, Figure 3.8, and Figure 3.9).

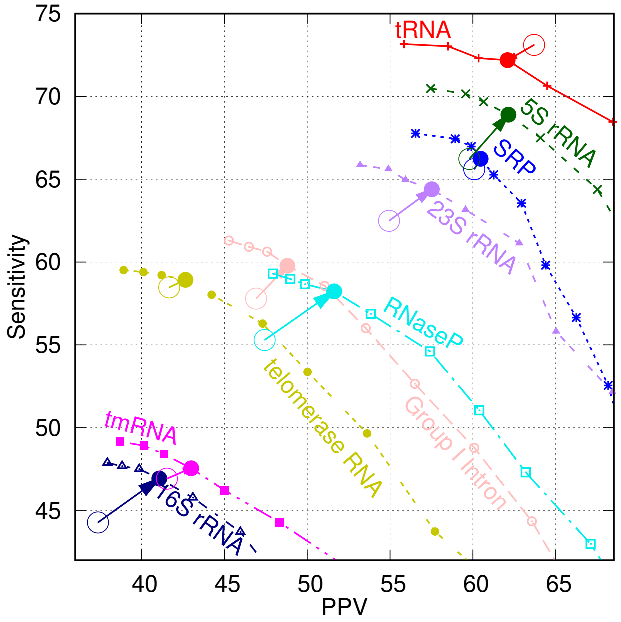

Figure 3.6: ThreshKnot improves 8 out of 9 families over MFE (in both PPV and sensitivity) on Vienna RNAfold. The curves show the ThreshKnot accuracies with varying θ. The arrows point from MFE (hollow circles) to ThreshKnot at $\theta = 0.3$.

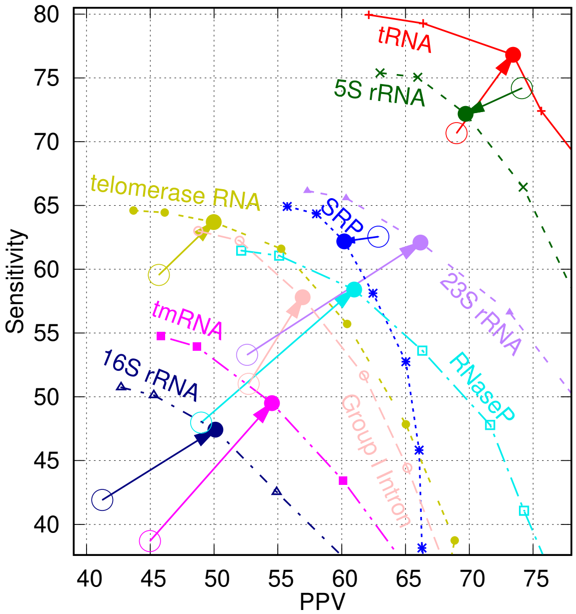

Figure 3.7: ThreshKnot improves 7 out of 9 families over MFE (in both PPV and sensitivity) on CONTRAfold. The curves show the ThreshKnot accuracies with varying θ. The arrows point from MFE (hollow circles) to ThreshKnot at $\theta = 0.2$.

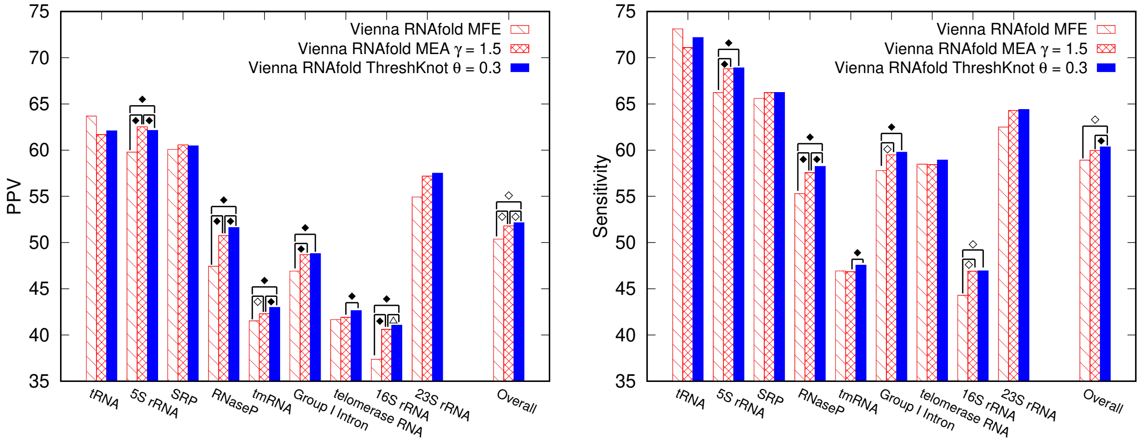

Figure 3.8: Accuracy results of MFE, MEA, and ThreshKnot using Vienna RNAfold. The first nine bars from the leftmost represent PPV (left plot) and sensitivity (right plot) averaged over all sequences in one family. The rightmost bars represent the overall accuracies, averaging over all families. Statistical significance (two-sided) is marked as $^{\blacklozenge}(p < 0.01)$, $^{\Diamond}(0.01 \leq p < 0.05)$, or $^{\triangle}(0.05 \leq p < 0.06)$.

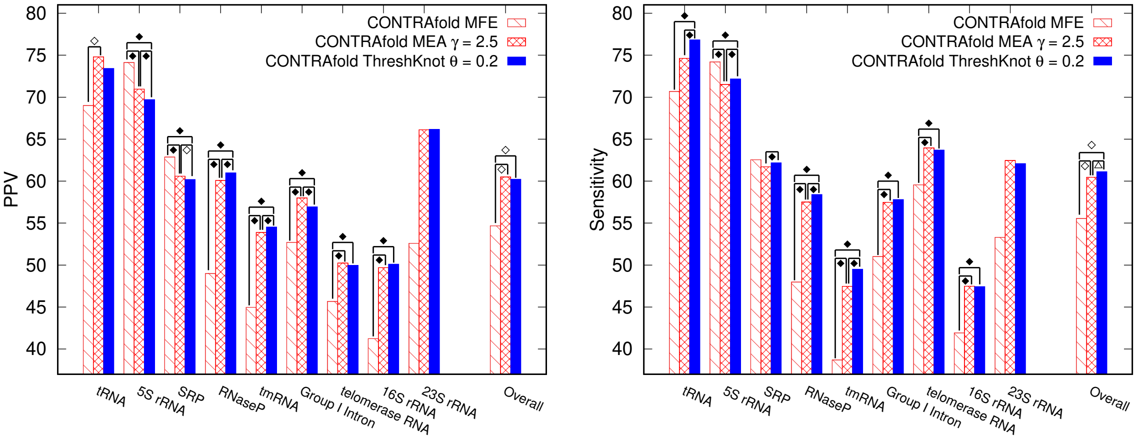

Figure 3.9: Accuracy results of MFE, MEA, and ThreshKnot using CONTRAfold. The first nine bars from the leftmost represent PPV (left plot) and sensitivity (right plot) averaged over all sequences in one family. The rightmost bars represent the overall accuracies, averaging over all families. Statistical significance (two-sided) is marked as $^{\blacklozenge}(p < 0.01)$, $^{\lozenge}(0.01 \leq p < 0.05)$, or $^{\triangle}(0.05 \leq p < 0.06)$.

3.5. 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 3.1 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.

4. DISCUSSION

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 [33] reports an $O(n)$-time algorithm to approximate the partition function inspired by the recently published LinearFold algorithm [28], 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 4.1).

	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 4.1: The time complexities of ThreshKnot using classical partition function calculation [14] and LinearPartition [33].

5. CONCLUSION

In RNA secondary structure prediction, partition function-based algorithms have become increasingly popular in recent years. Among these methods, MEA is popular, but our experiments with the three widely-used folding engines demonstrate 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.

References

1. S. R. Eddy, “Non-coding RNA genes and the modern RNA world,” *Nature Reviews Genetics*, vol. 2, no. 12, pp. 919–929, 2001.

2. J. T. Y. Kung, D. Colognori, and J. T. Lee., “Long noncoding RNAs: Past, present, and future.” *Genetics*, vol. 193, no. 3, pp. 651–669, 2013.

3. S. H. Kim, G. Quigley, F. L. Suddath, and A. Rich, “High-resolution x-ray diffraction patterns of crystalline transfer RNA that show helical regions,” *Proceedings of the National Academy of Sciences*, vol. 68, 1971.

4. L. G. Scott and M. Hennig, “RNA structure determination by NMR,” in *Bioinformatics*. Springer, 2008, pp. 29–61.

5. W. A. Ziehler and D. R. Engelke, “Probing RNA structure with chemical reagents and enzymes,” *Current protocols in nucleic acid chemistry*, 2001.

6. Z. Miao, R. W. Adamiak, M. Antczak, R. T. Batey, A. J. Becka, M. Biesiada, M. J. Boniecki, J. M. Bujnicki, S.-J. Chen, C. Y. Cheng, F.-C. Chou, A. R. Ferré-D’Amaré, R. Das, W. K. Dawson, F. Ding, N. V. Dokholyan, S. Dunin-Horkawicz, C. Geniesse, K. Kappel, W. Kladwang, A. Krokhotin, G. E. Łach, F. Major, T. H. Mann, M. Magnus, K. Pachulska-Wieczorek, D. J. Patel, J. A. Piccirilli, M. Popenda, K. J. Purzycka, A. Ren, G. M. Rice, J. S. Jr., J. Sarzynska, M. Szachniuk, A. Tandon, J. J. Trausch, S. Tian, J. Wang, K. M. Weeks, B. W. II, Y. Xiao, X. Xu, D. Zhang, T. Zok, and E. Westhof, “RNA-puzzles round III: 3D RNA structure prediction of five riboswitches and one ribozyme,” *RNA*, vol. 23, no. 5, pp. 655–672, 2017.

7. I. Tinoco and C. Bustamante, “How RNA folds,” *Journal of molecular biology*, vol. 293, no. 2, pp. 271–281, 1999.

8. I. Tinoco, O. C. Uhlenbeck, and M. D. Levine, "Estimation of secondary structure in ribonucleic acids," *Nature*, vol. 230, no. 5293, pp. 362–367, 1971.

9. P. E. Auron, W. P. Rindone, C. P. Vary, J. J. Celentano, and J. N. Vournakis, "Computer-aided prediction of RNA secondary structures," *Nucleic acids research*, vol. 10, no. 1, pp. 403–419, 1982.

10. M. Parisien and F. Major, "The mc-fold and mc-sym pipeline infers rna structure from sequence data," *Nature*, vol. 452, no. 7183, p. 51, 2008.

11. M. G. Seetin and D. H. Mathews, "Automated rna tertiary structure prediction from secondary structure and low-resolution restraints," *Journal of computational chemistry*, vol. 32, no. 10, pp. 2232–2244, 2011.

12. R. Nussinov and A. B. Jacobson, "Fast algorithm for predicting the secondary structure of single-stranded RNA," *Proceedings of the National Academy of Sciences*, vol. 77, no. 11, pp. 6309–6313, 1980.

13. M. Zuker and P. Stiegler, "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information," *Nucleic Acids Research*, vol. 9, no. 1, pp. 133–148, 1981.

14. J. S. McCaskill, "The equilibrium partition function and base pair binding probabilities for RNA secondary structure," *Biopolymers*, vol. 29, no. 6-7, pp. 1105–1119, 1990.

15. C. B. Do, D. A. Woods, and S. Batzoglou, "CONTRAfold: RNA secondary structure prediction without physics-based models," *Bioinformatics*, vol. 22, no. 14, pp. e90–e98, 2006.

16. Z. J. Lu, J. W. Gloor, and D. H. Mathews, "Improved RNA secondary structure pre-

diction by maximizing expected pair accuracy," *RNA*, vol. 15, no. 10, pp. 1805–1813, 2009.

17. M. Hajiaghayi, A. Condon, and H. H. Hoos, "Analysis of energy-based algorithms for RNA secondary structure prediction," *BMC Bioinformatics*, vol. 13, no. 22, p. 1, 2012.

18. P. Cordero and R. Das, "Rich RNA structure landscapes revealed by mutate-and-map analysis," *PLOS Computational Biology*, vol. 11, no. 11, 2015.

19. H. Tafer, S. L. Ameres, G. Obernosterer, C. A. Gebeshuber, R. Schroeder, J. Martinez, and I. L. Hofacker, "The impact of target site accessibility on the design of effective siRNAs," *Nature biotechnology*, vol. 26, no. 5, pp. 578–583, 2008.

20. Z. J. Lu and D. H. Mathews, "Efficient sirna selection using hybridization thermodynamics," *Nucleic acids research*, vol. 36, no. 2, pp. 640–647, 2007.

21. D. Long, R. Lee, P. Williams, C. Y. Chan, V. Ambros, and Y. Ding, "Potent effect of target structure on microrna function," *Nature structural & molecular biology*, vol. 14, no. 4, p. 287, 2007.

22. S. Bellaousov and D. H. Mathews, "Probknot: fast prediction of RNA secondary structure including pseudoknots," *RNA*, vol. 16, no. 10, pp. 1870–1880, 2010.

23. K. Sato, Y. Kato, M. Hamada, T. Akutsu, and K. Asai, "Ipknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming," *Bioinformatics*, vol. 27, no. 13, pp. i85–i93, 2011.

24. J. S. Reuter and D. H. Mathews, "Rnastructure: software for rna secondary structure prediction and analysis," *BMC bioinformatics*, vol. 11, no. 1, p. 129, 2010.

25. R. Lorenz, S. H. Bernhart, C. H. Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, "ViennaRNA package 2.0," *Algorithms for Molecular Biology*, vol. 6, no. 1, p. 1, 2011.

26. J. Zuber, H. Sun, X. Zhang, I. McFadyen, and D. H. Mathews, "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*, vol. 45, no. 10, pp. 6168–6176, 2017.

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

28. L. Huang, H. Zhang, D. Deng, K. Zhao, K. Liu, D. A. Hendrix, and D. H. Mathews, "LinearFold: linear-time approximate RNA folding by 5′-to-3′ dynamic programming and beam search," *Bioinformatics*, vol. 35, no. 14, pp. i295–i304, 2019.

29. M. Andronescu, A. Condon, H. H. Hoos, D. H. Mathews, and K. P. Murphy, "Efficient parameter estimation for RNA secondary structure prediction," *Bioinformatics*, vol. 23, no. 13, pp. i19–i28, 2007.

30. D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner, "Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure," *Journal of molecular biology*, vol. 288, no. 5, pp. 911–940, 1999.

31. J. Tukey, "Bias and confidence in not quite large samples," *Ann. Math. Statist.*, vol. 29, p. 614, 1958.

32. J. Reeder and R. Giegerich, "Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics," *BMC Bioinformatics*, vol. 5, no. 1, p. 1, 2004.

33. H. Zhang, L. Zhang, D. H. Mathews, and L. Huang, “Linearpartition: Linear-time approximation of RNA folding partition function and base pairing probabilities,” *bioRxiv*, 2019.

Appendix

ThreshKnot: Thresholded ProbKnot for Improved RNA Secondary Structure Prediction

				RNAstructure						Vienna RNAfold						CONTRAfold						IPknot		pKiss	
				MFE		MEA		ThreshKnot		MFE		MEA		ThreshKnot		MFE		MEA		ThreshKnot					
	# of seqs		avg.			γ=1.5		θ=0.3				γ=1.5		θ=0.3				γ=2.5		θ=0.2					
Family	total	used	length	PPV	sens	PPV	sens	PPV	sens	PPV	sens	PPV	sens	PPV	sens	PPV	sens	PPV	sens	PPV	sens	PPV	sens	PPV	sens
tRNA	557	74	77.3	62.77	69.73	68.11	75.45	66.39	75.44	63.69	73.11	61.68	71.09	62.09	72.18	69.00	70.67	74.80	74.63	73.43	76.82	81.89	80.25	47.82	55.16
5S rRNA	1,283	1,125	118.8	59.01	64.54	58.73	64.01	58.56	64.53	59.79	66.22	62.52	68.80	62.14	68.90	74.12	74.20	70.96	71.52	69.71	72.18	62.53	50.12	47.19	50.45
SRP	928	886	186.1	60.18	65.56	58.41	63.35	58.50	63.78	60.08	65.61	60.56	66.22	60.46	66.24	62.87	62.55	60.58	61.71	60.20	62.19	56.63	49.07	54.45	59.03
RNaseP	454	182	344.1	48.36	55.36	52.42	59.05	52.80	59.33	47.43	55.30	50.76	57.57	51.63	58.23	48.99	47.98	60.09	57.51	60.98	58.41	65.24	56.89	41.65	46.80
tmRNA	462	462	366.0	40.87	45.93	42.17	46.35	42.65	47.19	41.53	46.93	42.30	46.85	43.00	47.56	44.97	38.69	53.89	47.46	54.53	49.51	55.71	43.15	36.70	40.93
Group I Intron	98	96	424.9	45.84	56.22	47.87	57.59	47.91	58.14	46.91	57.80	48.70	59.49	48.81	59.78	52.71	51.01	58.00	57.48	56.94	57.80	55.98	48.80	47.09	57.70
telomerase RNA	37	37	444.6	42.37	59.15	41.86	58.03	42.25	58.38	41.67	58.48	41.91	58.43	42.65	58.93	45.67	59.56	50.23	63.95	49.97	63.70	43.00	44.44	38.58	53.38
16S rRNA	22	22	1,547.9	38.34	45.19	41.98	47.96	42.27	48.14	37.37	44.29	40.61	46.90	41.07	46.94	41.23	41.92	49.70	47.49	50.11	47.43	52.96	41.45	38.63	44.48
23S rRNA	5	5	2,927.4	51.02	57.73	56.25	62.01	56.27	61.84	54.94	62.49	57.19	64.29	57.51	64.39	52.59	53.30	66.13	62.45	66.17	62.10	68.07	48.98	46.74	51.32
Overall	3,846	2,889	222.2	49.86	57.71	51.98	59.31	51.96	59.64	50.38	58.91	51.80	59.96	52.15	60.35	54.68	55.54	60.49	60.47	60.23	61.13	60.22	51.46	44.32	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 [27] considers a base pair to be correct if it is slipped by one nucleotide on a strand. Isolated base-pairs are not allowed for ThreshKnot.

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

Table SI 2: Detailed overall prediction accuracies on the ArchiveII dataset. The accuracies use exact base-pair matching. Isolated base-pairs are not allowed for ThreshKnot.

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