LinearPartition: Linear Time RNA Partition Function and Base Pair Probability Calculation

He Zhang[1,2], Liang Zhang[2], David H. Mathews[3,4,5] and Liang Huang[2,1,†]
[1]Baidu Research USA, Sunnyvale, CA 94089, USA, [2]School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97330, USA, [3]Department of Biochemistry & Biophysics, [4]Center for RNA Biology, and [5]Department of Biostatistics & Computational Biology, University of Rochester Medical Center, Rochester, NY 48306, USA
[†]E-mail: liang.huang.sh@gmail.com

RNA secondary structure prediction is a well-known problem, and it has been used for medical design. Compared with MFE-based methods, partition function-based methods have gained more and more attention due to their higher accuracy and ability to predict pseudoknots. However, partition function calculation, as well as the downstream base pair probability prediction, uses cubic algorithm and is slow. This slowness is even more severe than cubic MFE-based methods because of the larger cost in the inner loop. To address this, we present LinearPartition, a novel algorithm that can calculate partition function and base pair probability in both linear runtime and linear memory space with RNA sequence length. LinearPartition, as an extention of LinearFold, inherits LinearFold's efficiency and accuracy. LinearPartition is 10× faster than Vienna RNAfold for the longest sequence (about 3000 nucleotides) in the dataset. Not only fast, LinearPartition is as accurate as Vienna RNAfold when comparing MEA and ProbKnot output structure. Surprisingly, even though LinearPartition uses an inexact search, it achieves better accuracy on longer families (16S and 23S rRNA).

Keywords: RNA secondary structure, partition function, base pair probability, linear time.

1. Introduction

For past decades, our understanding of ribonucleic acid (RNA) is changing. New proofs reveal that noncoding RNAs (ncRNAs) are involved in multiple processes, such as controlling gene expression, guiding RNA modifications,[1] or regulating a particular disease.[2] These functionalities are highly related to RNA's structure, so being able to rapidly determine the structure is extremely useful given the overwhelming pace of increase in genomic data (about 1021 basepairs per year[3]) and given the small percentage of sequences that have experimentally determined structure. While experimental assays still constitute the most reliable way to determine structures, they are prohibitively costly, slow, and difficult.

Due to such limitations computational prediction is required and desired, however, predict full RNA structure is very challenging, even more difficult than protein folding.[4] Alternatively, RNA secondary structure, the double helices folding structure formed by self-complementary nucleotides (A-U, G-C, G-U base pairs),[5] provides detailed information to help understand RNA's functionality,[6] and is a useful as a starting point to further predict full tertiary structure.[7] Furthermore, nesting secondary structure prediction problem, though still challenging, is well-defined in mathematics formation, and can be suitable modeled with the summa-

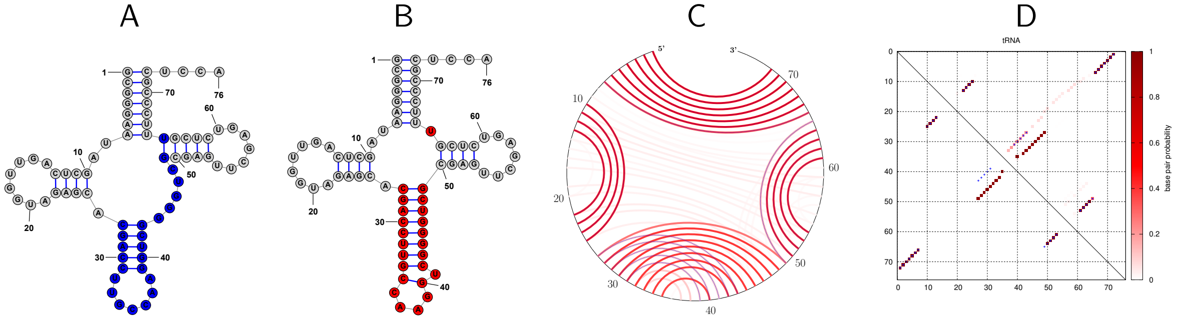


Fig. 1. Comparison of MFE-based method and partition function-based method. **A**: ground truth secondary structure of *E. coli* tRNA^{Gly}; **B**: the corresponding MFE structure. Structural difference are denoted with blue in ground truth structure and red in MFE structure; **C**: the corresponding circular representation. Ground truth base pairs are denoted with dash blue lines. Base pair probabilities are denoted with red solid lines and line shade is proportional to probability value. **D**: the corresponding heatmap representation. MFE structure (lower triangle) misses some ground truth base pairs (blue cross), while base pair probability matrix (upper triangle) covers these correct base pairs.

tion of decomposable free energy. Utilizing this decomposable nature, cubic runtime dynamic programming algorithm for nesting secondary structure prediction are proposed,[8] followed by an important paradigm free energy minimization (MFE) method[9] when a single structure is expected. This method gives a practical solution to predict secondary structure, however, it neglects the facts that multiple conformations exit at equilibrium,[10] as well as abandons all pseudoknotted structures. Many RNA sequences, for example mRNAs, exist in a thermodynamic ensemble of structures.[11]

As an alternative, partition function-based method provides a normalization factor from which we can estimate base pairing probabilities[4,10]or statistically sample structures from the ensemble[12,13]The base pairing probabilities also provide confidence estimates for predicted pairs[10,14]As a by-product, the pair probabilities also enable maximum expected accuracy (MEA) structure prediction[15,16]Moreover, although partition function-based method excluded pseudoknotted structures during dynamic programming process, it is able to predict pseudoknotted base pairs and structure by using pair probability matrix, and pseudoknotted prediction systems such as HotKnot,[17] ProbKnot,[18] DotKnot[19] and IPknot[20] all take pair probability matrix as inputs. Therefore, there has been a general shift from the classical MFE-based methods to partition function-based methods. Figure 1 compares MFE-based and partition function-based methods.

However, this partition function-based method, as well as prediction systems based on it such as RNAstructure,[21] CONTRAfold[15] and Vienna RNAfold,[22] suffers the slowness from its $O(n^3)$ runtime and scales poorly with longer sequences. Compared with $O(n^3)$ MFE-based method the slowness is even more severe. Recently, LinearFold,[23] the first linear-time MFE-based (approximate) algorithm for RNA folding, achieves significant efficiency and scalability improvement and higher accuracy than classical $O(n^3)$ MFE-based method, especially on long sequence. Borrowed the efficient linearize idea from LinearFold, we presents LinearPartition, which approximates the partition function and base pair probability matrix in linear time, to address speed bottleneck in existing systems. Similar as LinearFold, LinearPartition incrementally parses a RNA sequence using a left-to-right fashion dynamic programming, and further applys beam prune[24] to narrow search space and only retain states with top b lowest energy, where b is the beam size. Though introducing beam prune results to giving up some possible structures, the well-designed pruning heuristic makes sure that only structures with worse energy are neglected, and partition function is still similar as exact search.

LinearPartition, inherits LinearFold's efficiency and accuracy. LinearPartition is 10× faster than the baseline Vienna RNAfold for the longest sequence (about 3000 nucleotides) in the dataset. Not only fast, LinearPartition even leads to a small improvement in MEA and ProbKnot prediction using the probability matrix computed in linear time. Surprisingly, LinearPartition achieves better accuracy on longer families (16S and 23S rRNA).

2. Results

2.1. *Efficiency and Scalability*

We compare the efficiency of LinearPartition with the baseline Vienna RNAfold (Version 2.4.11) (https://www.tbi.univie.ac.at/RNA/download/sourcecode/2_4_x/ViennaRNA-2.4.11.tar.gz). Vienna RNAfold is a widely-used RNA structure prediction package, and provides partition function and base pair probabilities calculation based on classical cubic runtime algorithm. We use the ArchiveII dataset, a comprehensive set of well-determined structures first curated in the 1990s[25]and updated later with additional structures[26](http://rna.urmc.rochester.edu/pub/archiveII.tar.gz; we removed those sequences found in the S-Processed set). The dataset contains 2889 RNA sequences from 9 families, and the average length is 222 *nt* and max length is 2968 *nt*. We run all programs (compiled by GCC 4.9.0) on Linux, with 2.90GHz Intel Core i9-7920X CPU and 64G memory.

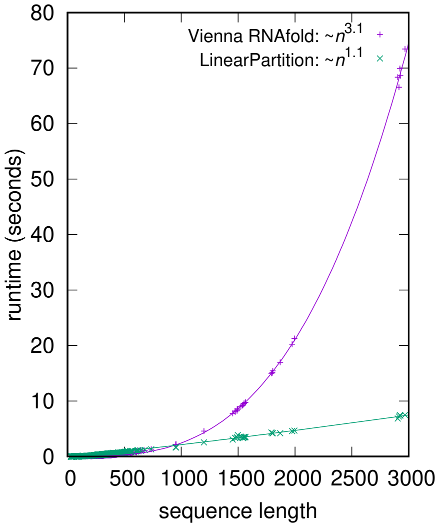


Fig. 2. Runtime comparisons on thze ArchiveII dataset between with the baseline, Vienna RNAfold, and LinearPartition. The curve-fittings were done log-log in gnuplot with $n > 10^3$.

Figure 2 confirms that runtime of LinearPartition scales linearly with sequence length, while the baseline Vienna RNAfold scale cubically. For a sequence of 2,968 *nt* (23S rRNA), LinearPartition takes only 7 seconds while the baselines take 75 seconds. This clearly shows the advantage of LinearPartition on very long ncRNAs.

2.2. *Accuracy*

We next compare accuracy of LinearPartition with the baseline Vienna RNAfold. We take the base pair probability matrices from these two systems, and fed them to standard

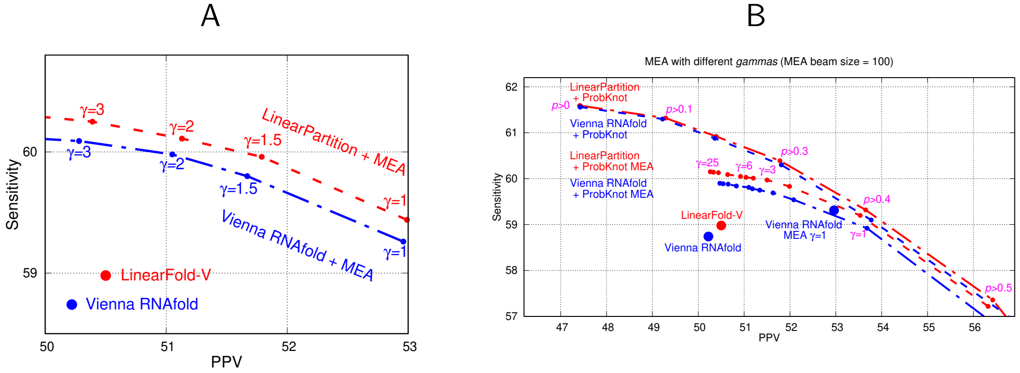


Fig. 3. Accuracy comparison for two systems. **A**: Overall MFE and MEA structure PPV-sensitivity tradeoff of two systems with varying *gamma*. **B**: Overall ThreshKnot structure PPV-sensitivity tradeoff of two systems with varying *gamma* LinearPartition even leads to a small improvement in the downstream MEA predictoin using the probability matrix computed in linear time.

MEA algorithm. We use Positive Predictive Value (PPV, the fraction of predicted pairs in the known structure) and sensitivity (the fraction of known pairs predicted) to measure the accuracy across all families, as well as slipping method to allow base pair to slip by one nucleotide.[26]

Figure 3A shows that (1) MEA-based method is more accurate than MFE-based method for both systems; (2) LinearPartition + MEA is constantly more accurate than Vienna RNAfold + MEA. With the same γ, a hyperparameter balances PPV and sensitivity in MEA algorithm, LinearPartition + MEA enjoys a small improvement in both PPV and sensitivity.

Probknot is another partition-function-based method which is much simpler than MEA, only adds a linear post-processing step after the partition function calculation, and can predict pseudoknots. Recently, ThreshKnot,[?] a simple thresholded version of ProbKnot, leads to more accurate overall predictions by filtering out unlikely pairs whose prob falls under a given threshold, so we also compare ThreshKnot structure accuracy.

Figure 3B shows that ???????

per family accuracy

2.3. *Search Quality*

Fig. 4A–B show that our LinearPartition algorithm can indeed approximate the partition function reasonably well. Here we measure root-mean-square deviation (RMSD) between the two probability matrices p and p' (from Vienna RNAfold and LinearPartition, resp.) over the set of all possible pairs pairs(x) on a sequence x (i.e., pairs(x) = $1 \leq i < j \leq |x| \mid x_i x_j \in \mathrm{CG}, \mathrm{GC}, \mathrm{AU}, \mathrm{UA}, \mathrm{GU}, \mathrm{UG}, j - i > 3$):

$$\mathrm{RMSD}(p, p') = \sqrt{\frac{1}{|\mathrm{pairs}(x)|} \sum_{(i,j) \in \mathrm{pairs}(x)} (p_{i,j} - p'_{i,j})^2 (1)}$$

Figure 4A shows the probability matrix for short sequence, e.g. tRNA sequence, from both RNAfold and LinearPartition yield identical matrices (i.e., RMSD=0). Figure 4B shows that RMSD is relatively small across all RNA families in the ArchiveII dataset. The highest deviation is 0.067 for one RNaseP sequence, which means on average, each pair's probability deviation in that worst-case sequence is about 0.067 between the exact algorithm and our linear-time one. With sequence length increasing, RMSD gradually decreases, since the number of possible pairs grows in $O(n^2)$ but the num-

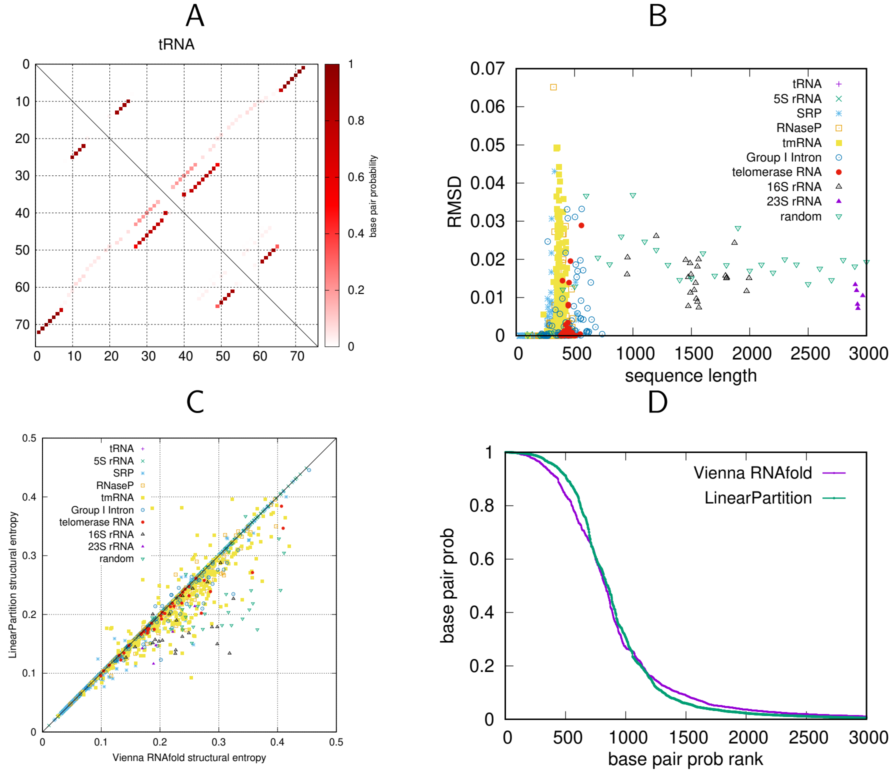


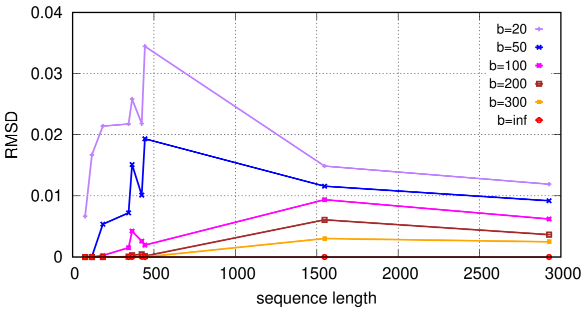
Fig. 4. Comparison of base pair probabilities from Vienna RNAfold and LinearPartition. **A**: LinearPartition (upper triangle) and Vienna RNAfold (lower triangle) result in identical base pair probability matrix for *E. coli* tRNA^{Gly}. **B**: root-mean-square deviation (RMSD) is relatively small between LinearPartition and Vienna RNAfold. **C**: structural entropy comparison. **D**: LinearPartition starts higher and finishes lower than RNAfold in a sorted probability curve for *E. coli* 23S rRNA.

ber of highly probable pairs grows in $O(n)$; on the longest 23S rRNA family, RMSD is about 0.015. We also included 30 random RNA sequences with length 100–3,000 and they behave similarly to natural sequences in terms of RMSD.

We assume LinearPartition base pair probability distribution is peakier since it ignores low energy substructure in partition function calculation. We uses structural entropy[27] to measure this, where lower structural entropy indicates that the distribution is dominated by fewer base pairing probabilities. Figure 4C shows LinearPartition distribution is peakier (lower structural entropy) than RNAfold for most sequences.

We also uses *E. coli* 23S rRNA as an example to illustrate the distribution difference. We sort all base pair probabilities from high to low and take the top 3,000 rank. Figure 4D shows LinearPartition probability distribution curve starts higher and finishes lower.

2.4. *Beam Size Impact*



2.5. *Example*

3. Methods

4. Discussion

4.1. *Summary*

4.2. *Analysis*

4.3. *Extensions*

Our algorithm has several potential extensions.

(1) We will linearize the partition function-based heuristic pseudoknot prediction methods such as ProbKnot, IpKnot, and Dotknot by replacing their bottleneck $O(n^3)$-time calculation of the partition function with our LinearPartition. All these heuristic methods uses rather simple heuristic criteria to choose pairs from the base pair probability matrix. For example, the second step of probknot selects base pairs (i, j) where the i–j pairing probability is the largest for both bases i and j. This might appear as $O(n^2)$ in the worst case, but since the linear-time beam search used in LinearPartition only returns $O(nb)$ pairs where b is the constant beam size, this second step is still $O(n)$, giving an overall linear-time method, LinearProbKnot. We can similary get LinearIPknot, LinearProbknot and LinearDotKnot, etc. With these promising substantial results of LinearPartition, we believe LinearProbknot (and LinearIPknot, LinearDotKnot, etc) should be as accurate as, if not more accurate than, their original $O(n^3)$ versions.

(2) Accelerate and Improve bimolecular and multistrand structure prediction. LinearPartition provide important new ways to improve existing bimolecular and multistrand structure prediction algorithm such as AccessFold.[28] LinearPartition will provide a much faster solution to the accessibility calculation. Also, LinearPartition will help include predictions of intramolecular and bimolecular pairs, re have immediate impact on our ability to predict bimolecular structures by improving speed and also providing additional structure information to users.

Aim 3b: Accelerate and Improve bimolecular and multistrand structure prediction Many ncRNAs function by interacting with other RNA sequences by base pairing. We developed several software tools for predicting base pairing structures between two sequences (bimolecular) [81, 68, 23, 24]. There are two important barriers to accurate bimolecular structure prediction. First, RNA sequences form self-structure that prevents bimolecular structure prediction. We account for this in our AccessFold algorithm, which is part of the RNAstructure software package and uses a partition function calculation to approximate the accessibility [23] The second barrier is that many simple bimolecular structures that also include unimolecular pairs, such as the kissing hairpin [13], are tantamount to pseudoknots in our algorithms.

The linear algorithms from aims 1 and 2 provide important new ways to improve our existing AccessFold algo- rithm. First, the linear partition function calculation (Aim 1a) will provide a much faster solution to the accessibility calculation. Second, the linear pseudoknot prediction (aim 2) will provide us with the algorithm needed to elevate AccessFold to include predictions of intramolecular and bimolecular pairs. Currently, AccessFold only provides the pairs between strands, ignoring the pairs within a strand. We will advance AccessFold to incorporate the new algorithms from aims 1 and 2 and test it against a database of bimolecular structures we developed previously [23]. This will have immediate impact on our ability to predict bimolecular structures by improving speed and also providing additional structure information to our users.

References

1. S. R. Eddy., Non-coding RNA genes and the modern rna world., *Nature Reviews Genetics* **2**, 919 (2001).
2. J. T. Y. Kung, D. Colognori and J. T. Lee., Long noncoding rnas: Past, present, and future., *Genetics* **193**, 651 (2013).
3. Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha and G. E. Robinson, Big data: astronomical or genomical?, *PLoS Biology* **13**, p. e1002195 (2015).
4. J. S. McCaskill, The equilibrium partition function and base pair probabilities for rna secondary structure, *Biopolymers* **29**, 11105 (1990).
5. I. T. Jr and C. Bustamante., How RNA folds, *Journal of Molecular Biology* **293**, 271 (1999).
6. I. TINOCO, O. C. UHLENBECK and M. D. LEVINE, Estimation of secondary structure in ribonucleic acids, *Nature* **230**, 362 (1971).
7. 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* **10**, 403 (1982).
8. R. Nussinov and A. B. Jacobson, Fast algorithm for predicting the secondary structure of single-stranded RNA, *Proceedings of the National Academy of Sciences* **77**, 6309 (1980).
9. M. Zuker and P. Stiegler, Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information, *Nucleic Acids Research* **9**, 133 (1981).
10. D. H. Mathews., Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization., *RNA* **10**, 1178 (2004).
11. W.-J. C. Lai, M. Kayedkhordeh, E. V. Cornell, E. Farah, S. Bellaousov, R. Rietmeijer, D. H. Mathews and D. N. Ermolenko, mRNAs and lncRNAs intrinsically form secondary structures with short end-to-end distances, *Nature Communications* **9**, p. 4328 (2018).
12. Y. Ding, C. Y. Chan and C. E. Lawrence, RNA secondary structure prediction by centroids in a boltzmann weighted ensemble, *RNA* **11**, 1157 (2005).
13. D. H. Mathews, Revolutions in RNA secondary structure prediction, *Journal of molecular biology* **359**, 526– (2006).
14. J. Zuber, B. J. Cabral, I. McFadyen, D. M. Mauger and D. H. Mathews, Analysis of RNA nearest neighbor parameters reveals interdependencies and quantifies the uncertainty in RNA secondary structure prediction, *RNA* **24**, 1568 (2018).
15. C. Do, D. Woods and S. Batzoglou, CONTRAfold: RNA secondary structure prediction without physics-based models, *Bioinformatics* **22**, e90 (2006).
16. Z. J. Lu, J. W. Gloor and D. H. Mathews, Improved RNA secondary structure prediction by maximizing expected pair accuracy, *RNA* **15**, 1805 (2009).

17. J. Ren, B. Rastegari, A. Condon and H. H. Hoos, Hotknots: heuristic prediction of RNA secondary structures including pseudoknots, *RNA* **11**, 1494– (2005).
18. S. Bellaousov and D. H. Mathews, Probknot: fast prediction of RNA secondary structure including pseudoknots, *RNA* **16**, 1870 (2010).
19. J. Sperschneider and A. Datta, Dotknot: pseudoknot prediction using the probability dot plot under a refined energy model, *Nucleic Acids Research* **38**, e103 (2010).
20. 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* **27**, i85– (2011).
21. D. H. Mathews and D. H. Turner, Prediction of RNA secondary structure by free energy minimization, *Current Opinion in Structural Biology* **16**, 270 (2006).
22. 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* **6**, p. 1 (2011).
23. 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* **35**, i295 (07 2019).
24. L. Huang and K. Sagae, Dynamic programming for linear-time incremental parsing, in *Proceedings of ACL 2010*, (ACL, Uppsala, Sweden, 2010).
25. 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* **288**, 911 (1999).
26. M. Sloma and D. Mathews, Exact calculation of loop formation probability identifies folding motifs in RNA secondary structures, *RNA, 22, 1808–1818* (2016).
27. M. Huynen, R. Gutell and D. Konings, Assessing the reliability of RNA folding usingstatistical mechanics, *Journal of molecular biology* **267**, 1104 (1997).
28. L. DiChiacchio, M. F. Sloma and D. H. Mathews., Accessfold: predicting RNA-RNA interactions with consideration for competing self-structure, *Bioinformatics* **32**, 1033– (2016).