# A novel algorithm for ranking RNA structure candidates

Anastacia Wienecke[a,b], Alain Laederach[a,b,1]

[a]Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599

[b]Curriculum in Bioinformatics and Computational Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599

[1]To whom correspondence should be addressed. Phone: (919) 962 4565, e-mail alain@unc.edu

## Abstract

RNA research is advancing at an ever-increasing pace. The newest and most state-of-the-art instruments and techniques have made possible the discoveries of new RNAs, and they have carried the field to new frontiers of disease research, vaccine development, therapeutics, and architectonics. Like proteins, RNAs show a marked relationship between structure and function. A deeper grasp of RNAs requires a finer understanding of their elaborate structures. In pursuit of this, cutting-edge experimental and computational structure-probing techniques output several candidate geometries for a given RNA, each of which are perfectly aligned with experimentally determined parameters. Identifying which structure is the most accurate however, remains a major obstacle. In recent years, several algorithms have been developed for ranking candidate RNA structures in order from most to least probable, though their levels of accuracy and transparency leave room for improvement. Most recently, improvements in both areas are demonstrated by rsRNASP, a novel algorithm proposed by Tan et al. It is a residue-separation-based statistical potential for 3D structure evaluation, and it outperforms the leading algorithms in the field.

## Introduction

RNA is a single-stranded biomolecule with myriad key roles in regulating gene activity (1), catalyzing chemical reactions (2), and encoding plus decoding genetic information (3). Its single-stranded nature enables intramolecular base-pairing, which allows the polymer to fold into intricate 3D structures. A ubiquitous example is the amino-acid-carrier tRNA. Via base-pairing, tRNAs adopt a 2D cloverleaf topology. Interactions between the leaves then create a 3D “L”-shaped architecture, the exact dimensions of which precisely fit through a ribosome’s entry portal and facilitate protein synthesis (4). Alongside tRNAs, a whole host of other structured RNAs play leading and supporting roles in almost every process on the cellular stage; examples include ribozymes (2), mRNAs (5,6), riboswitches (7), spliceosomes (8), ribosomes (9), and miRNAs (10). The rapid development of new structural techniques such as cryo-EM (11,12,13), SHAPE-JuMP (14), smFRET (15), and SAXS (16), as well as the refinement of more traditional techniques like NMR (17) and crystallography (18), has focused the stage lights. Nonetheless each of these methods have experimental limits on their resolution, and as such, computational techniques are required to further refine atomic resolution models from these data.

Like in proteins, RNA structure and RNA function are closely tied (19,20). Even certain single point mutations can augment RNA structures and lead to disease states (21). While the number of genes encoding functional RNAs greatly exceeds the number that encode proteins, to date, only 1% of Protein Data Bank entries include RNAs. A sharper understanding of RNA structure, its various roles, and its patterns of evolution holds much promise in informing the functions of uncharacterized RNAs, clarifying mechanisms of noncoding disease mutations (22), guiding aptamer-based (23) and other therapies (24), and buttressing the science of RNA nanostructures (25,26).

Critical to understanding RNA is the ability to probe their structures reliably and accurately. This is no simple task as RNAs are quite flexible and could even be said to exhibit an RNA version of the Levinthal paradox (27), eg. just a 20-base RNA composed of ten A's and ten U's has almost 10 million distinct folding conformations. Frequently used experimental methods like NMR (17), cryo-EM (11,12,13), and SHAPE-JuMP (14) yield experimental constraints either in the form of distances or density, that when combined with computational structural modeling produce a cloud of probable structures, each of which agrees equally well with the experimental data. Computational methods like fragment assembly (see FARNA) and homology modeling, piece together an RNA's structure based on its chemical similarity to already-known local RNA structures. But whatever the method, a set of candidate structures is output, and ranking these structures to identify the most accurate, or native-like, structure remains computationally challenging (see Figure 1A).

A great need in the field of RNA structure prediction is an accurate, reliable, and efficient way of evaluating and ranking candidate RNA structures. One of the most common methods for achieving a similar end in the protein-world is a type of energy function: the knowledge-based statistical potential. Such a function relies on a reference state, which in the RNA-world, can be determined by information latent in the sequences, chemical bonds, and configurations of a well-characterized training set of RNA structures. Output by this energy function is a potential energy value for each input geometry; the geometry with the lowest energy is taken as the best estimate of the native structure. Several such functions have been proposed and are currently in use: RASP (28), 3dRNAscore (29), DFIRE-RNA (30), and RNA3DCNN (31).

## The New and Noteworthy

In this issue of *Biophysical Journal*, Tan et al propose a new energy function: a residue-separation-based statistical potential (rsRNASP) for 3D RNA structure evaluation. Different from RNA3DCNN, which relies on the “black box” of 3D convolutional neural networks, the inner workings of rsRNASP are transparent. Based on the inverse Boltzman law, rsRNASP relies on three factors: the temperature, the probabilities of nucleotide separation for native state, and the probabilities of nucleotide separation for reference state. These parameters are weighted to factor in distance, an important consideration given the hierarchical nature of RNA folding patterns (32). Output is an energy value in units of $k_B T$; the lower a candidate structure’s energy, the higher its predicted similarity to the true native structure.

Tan et al’s comprehensive testing on three large datasets indicate the high quality of rsRNASP in parsing native from decoy RNA structures. These decoys were computationally generated either by normal mode perturbation, fragment assembly, replica-exchange, shifting atom distances or rearranging dihedral angles. Each RNA in these datasets has one accepted native structure, and several associated decoy structures. While its performance was not perfectly accurate, rsRNASP most successfully and most consistently identified the correct native structure and ranked the decoys. Its performance remained high for small RNAs with very similar decoys, for large RNAs with very variable decoys, and for the most realistic RNA-puzzles dataset. This approach appears to achieve a better balance of short- vs. long-range interactions and it functions at a higher resolution, perhaps explaining its outperformance of 3dRNAscore and RASP, as mentioned by Tan et al.

Tan et al consider their test set III as the most realistic. This set includes a known native structure and multiple computer-generated decoy structures for 1) each of the 22 RNAs in the RNA-puzzles dataset, and 2) each of 20 selected RNAs with known structures in the Protein Data Bank (see Figure 1A). For a given RNA, the energy function ranks the associated array of one native and several decoy structures (see Figure 1B). It receives no input on which structure is native and which is decoy. During testing, this ranking scheme is then compared to a reference list of structures, ordered by a measure of structural deformation. This “deformation index” quantifies the divergence between the shape of each structure and the known native structure; a high deformation index implies a high divergence. Thus, for a given RNA, if the Pearson correlation coefficient between the energy function ranking and the reference list is 1, the energy function accurately identifies structures most similar and most dissimilar to the known native structure (see Figures 1B and 1C for a visualization of rsRNASP's performance on the 71-base-long mc6 RNA riboswitch, 3LA5). For every RNA in their test set III, Figure 1D highlights how rsRNASP, RASP, 3dRNAscore, DFIRE-RNA, and RNA3DCNN compare in ranking the structures of RNAs (see Figure 1D) with a variety of lengths (see Figure 1E).

As more and more RNA structures continue to be probed, it is imperative for there to be an efficient, accurate, and reliable mechanism that ranks their candidate structures. rsRNASP meets this need.

## Figure Captions

**Figure 1.** rsRNASP's ranking of native and decoy 3LA5 RNA structures, and visualizing the performance of Tan et al's rsRNASP relative to four other energy functions. (A) Commonly-used experimental and computational techniques output several candidate RNA structures. A reliable way of choosing the most accurate, or native-like, structure is crucial. (B) Visualization of 3D 3LA5 RNA structures at ten positions of the rsRNASP ranking. These structures include the native and nine computationally-generated decoys (taken from Tan et al's test set III). rsRNASP determines rank by computing an energy value, in units of $k_BT$, which is reported below each structure; the lower the energy value, the higher the rank and the higher the predicted similarity to the native. Since this is a test case this ranking scheme is compared to an independent reference list, which is determined by the deformation index. This index measures the deviation of a given structure from the accepted native structure in the Protein Data Bank. The gray box highlights the native 3LA5 structure, which rsRNASP correctly identifies and gives a rank of 1. In practice, the true native structure is unknown. (C) Scatterplot of the rsRNASP rank vs. the reference list rank for all 42 3LA5 structures. The Pearson correlation coefficient (PCC) is 0.88, indicating a strong predictive ability of rsRNASP. Datapoints outlined in black are featured in part B. (D) Scatterplot of the ranking accuracy, as measured by the PCC, of rsRNASP, RNA3DCNN, 3dRNAscore, DFIRE-RNA, and RASP for each of the 42 RNAs in Tan et al's most realistic test set (test set III). Each RNA has an associated array of structures. The energy functions have no knowledge of which structure is the "native" and which are the "decoys". A PCC of 1 indicates that the energy function correctly identifies the native structure, and ranks the decoys in order of their structural similarity to the native deposited in the Protein Data Bank. (E) Violin plot displaying the distribution of RNA lengths in test set III, with the middle bar representing the median.

## Acknowledgments

This work was supported but the US National Institutes of Health by grants R35 GM140844 and R01 HL111527 to A.L.

## References

1. O'Brien J. Hayder H. Zayed Y. Peng C. **Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation.** *Front Endocrinol (Lausanne)*. 2018; **9**: 402

2. Lilley D.M. Eckstein F. (Eds.). **Ribozymes and RNA catalysis**. Royal Society of Chemistry. 2007

3. Li J. Liu C. **Coding or Noncoding, the Converging Concepts of RNAs**. *Front Genet*. 2019; **10**: 496

4. Agirrezabala X. Valle M. **Structural Insights into tRNA Dynamics on the Ribosome.** *Int J Mol Sci*. 2015; **16**: 9866-9895

5. Hiller M. Zhang Z. Backofen R. Stamm S. **Pre-mRNA secondary structures influence exon recognition.** *PLoS Genet*. 2007; 3: e204

6. Mustoe A.M. Corley M. Laederach A. Weeks K.M. **Messenger RNA Structure Regulates Translation Initiation: A Mechanism Exploited from Bacteria to Humans.** *Biochemistry*. 2018; **57**: 3537-3539

7. Serganov A. Nudler E. **A decade of riboswitches**. *Cell*. 2013; **152**: 17-24

8. Wilkinson M.E. Charenton C. Nagai K. **RNA Splicing by the Spliceosome.** *Annu Rev Biochem*. 2020; **89**: 359-388

9. Watson Z.L. Ward F.R. Méheust R. et al. **Structure of the bacterial ribosome at 2 Å resolution**. *Elife*. 2020; **9**: e60482

10. Roden C. Gaillard J. Kanoria S. et al. **Novel determinants of mammalian primary microRNA processing revealed by systematic evaluation of hairpin-containing transcripts and human genetic variation.** *Genome Res*. 2017; **27**: 374-384

11. Kappel K. Zhang K. Su Z. *et al.* **Accelerated cryo-EM-guided determination of three-dimensional RNA-only structures.** *Nat Methods.* 2020; **17**: 699–707

12. Koning R. Gomez-Blanco J. Akopjana I. *et al.* **Asymmetric cryo-EM reconstruction of phage MS2 reveals genome structure *in situ*.** *Nat Commun.* 2016; **7**: 12524

13. Zhang K. Li S. Kappel K. *et al.* **Cryo-EM structure of a 40 kDa SAM-IV riboswitch RNA at 3.7 Å resolution.** *Nat Commun.* 2019; **10**: 5511

14. Christy T.W., Giannetti C.A., Houlihan G. et al. **Direct Mapping of Higher-Order RNA Interactions by SHAPE-JuMP.** *Biochemistry*. 2021; **60**: 1971-1982

15. Manz C. Kobitski A. Samanta A. *et al.* **Single-molecule FRET reveals the energy landscape of the full-length SAM-I riboswitch.** *Nat Chem Biol*. 2017; **13**: 1172–1178

16. Chen Y. Pollack L. **SAXS studies of RNA: structures, dynamics, and interactions with partners.** *Wiley Interdiscip Rev RNA*. 2016; **7**: 512-526

17. Barnwal R.P. Yang F. Varani G. **Applications of NMR to structure determination of RNAs large and small.** *Arch Biochem Biophys*. 2017; **628**: 42-56

18. Reyes F.E. Garst A.D. Batey R.T. **Strategies in RNA crystallography.** *Methods Enzymol*. 2009; **469**: 119-139

19. Mortimer S. Kidwell M. & Doudna J. **Insights into RNA structure and function from genome-wide studies.** *Nat Rev Genet.* 2014; **15**: 469–479

20. Piao M. Sun L. Zhang Q.C. **RNA Regulations and Functions Decoded by Transcriptome-wide RNA Structure Probing.** *Genomics Proteomics Bioinformatics*. 2017; **15**: 267-278

21. Halvorsen M. Martin J.S. Broadaway S. Laederach A. **Disease-associated mutations that alter the RNA structural ensemble.** *PLoS Genet*. 2010; **6**: e1001074

22. Waldern J.M. Kumar J. Laederach A. **Disease-associated human genetic variation through the lens of precursor and mature RNA structure**. *Hum Genet*. 2021; 10.1007/s00439-021-02395-9

23. Ni S. Zhuo Z. Pan Y. et al. **Recent Progress in Aptamer Discoveries and Modifications for Therapeutic Applications.** *ACS Appl Mater Interfaces*. 2021; **13**: 9500-9519

24. Corley M. Solem A. Phillips G. et al. **An RNA structure-mediated, posttranscriptional model of human α-1-antitrypsin expression.** *Proc Natl Acad Sci U S A*. 2017; **114**: E10244-E10253

25. Grabow W.W. Jaeger L. **RNA self-assembly and RNA nanotechnology.** *Acc Chem Res*. 2014; **47**: 1871-1880

26. Jaeger L. Chworos A. **The architectonics of programmable RNA and DNA nanostructures.** *Curr Opin Struct Biol*. 2006; **16**: 531-543

27. Levinthal C. **How to Fold Graciously**. Mossbauer Spectroscopy in Biological Systems: Proceedings of a Meeting Held at Allerton House, Monticello, Illinois. 1969; 22–24

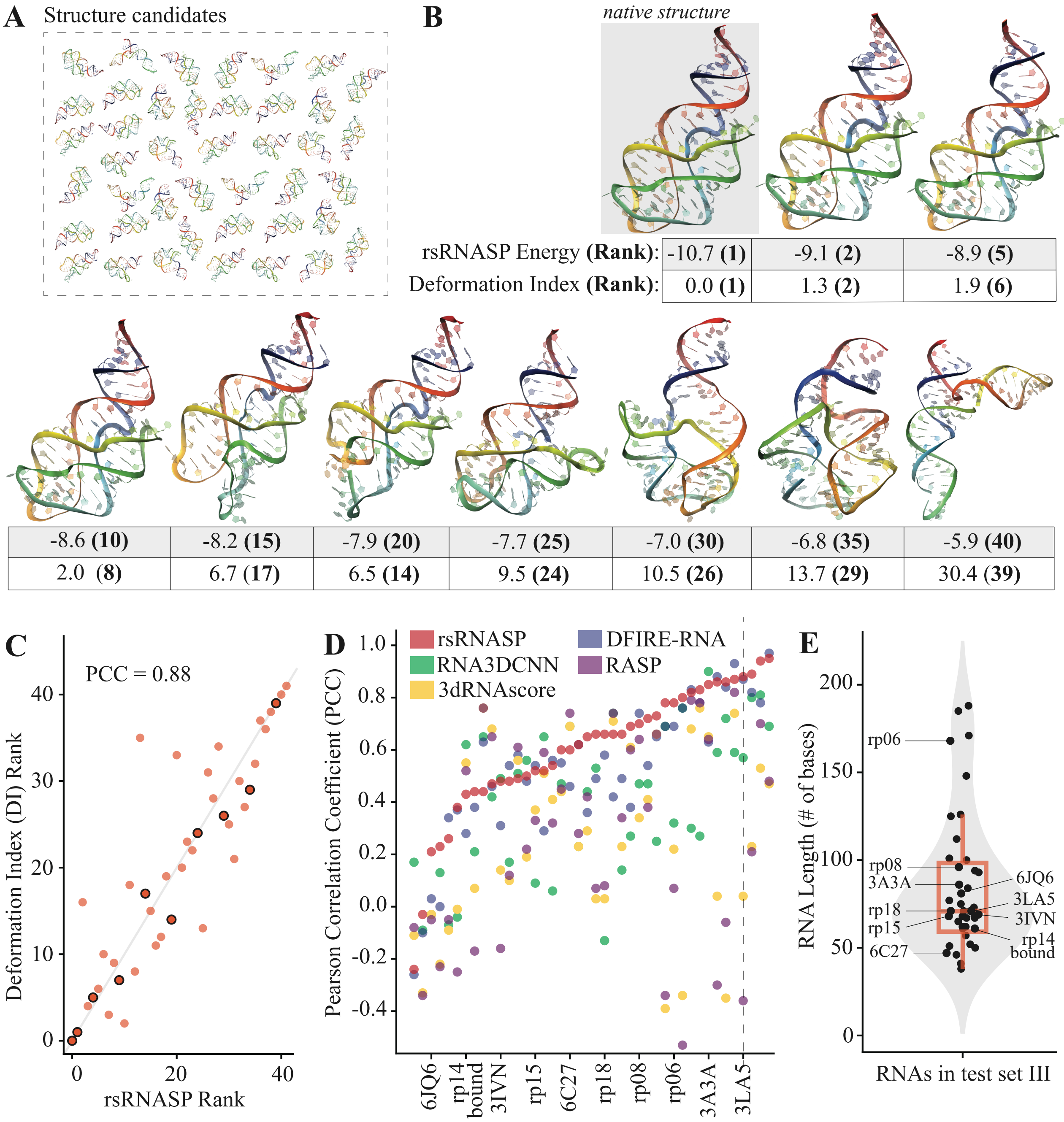
28. Capriotti E. Norambuena T. Marti-Renom M.A. Melo F. **All-atom knowledge-based potential for RNA structure prediction and assessment.** *Bioinformatics*. 2011; **27**: 1086-1093

29. Wang J. Zhao Y. Zhu C. Xiao Y. **3dRNAscore: a distance and torsion angle dependent evaluation function of 3D RNA structures.** *Nucleic Acids Res*. 2015; **43**: e63

30. Zhang T. Hu G. Yang Y. Wang J. Zhou Y. **All-atom knowledge-based potential for RNA structure discrimination based on the distance-scaled finite ideal-gas reference state.** *J. Comput. Bio.* 2019; **27**: 856-867

31. Li J. Zhu W. Wang J. Li W. Gong S. Zhang J. Wang W. **RNA3DCNN: Local and global quality assessments of RNA 3D structures using 3D deep convolutional neural networks.** *Plos Comput. Biol.* 2018; **14**: e1006514

32. Brion P. Westhof E. **Hierarchy and dynamics of RNA folding.** *Annu. Rev. Biophys. Biomol. Struct.* 1997; **26**: 113-137



| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| rsRNASP Energy **(Rank)**: | -10.7 **(1)** | -9.1 **(2)** | -8.9 **(5)** | -8.6 **(10)** | -8.2 **(15)** | -7.9 **(20)** | -7.7 **(25)** | -7.0 **(30)** | -6.8 **(35)** | -5.9 **(40)** |
| Deformation Index **(Rank)**: | 0.0 **(1)** | 1.3 **(2)** | 1.9 **(6)** | 2.0 **(8)** | 6.7 **(17)** | 6.5 **(14)** | 9.5 **(24)** | 10.5 **(26)** | 13.7 **(29)** | 30.4 **(39)** |