# Current advances in NA secondary and tertiary structure prediction

and its application for aptamers

by Maxim Shchepetov

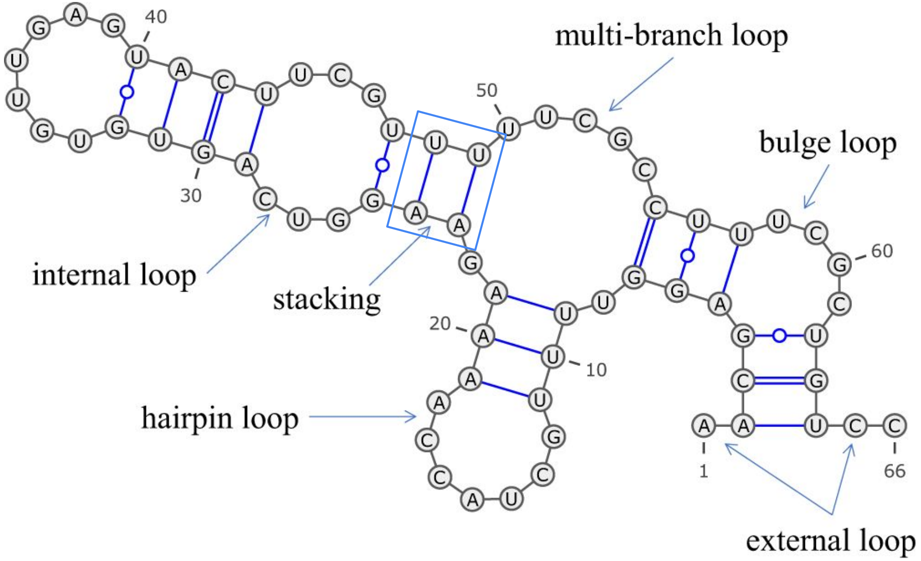
# Secondary structure prediction

# Introduction to 2D

- *Nucleic acid secondary structure* is the base pairing interactions within a single nucleic acid polymer or between two polymers

- Stacking is more important for the structure stabilisation than hydrogen bonds

- RNAs usually include wobble pair (G-U) and have generally more flexible structure than DNA



# Representations

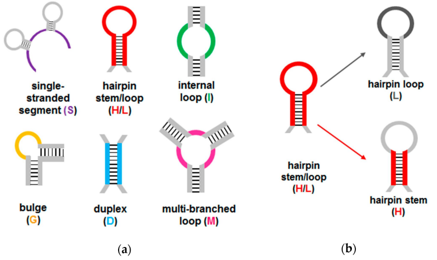
- a) Dot-bracket notation (most common output)
- b) Secondary structure elements (SSE)
- c) Contact table (common internal representation)

a

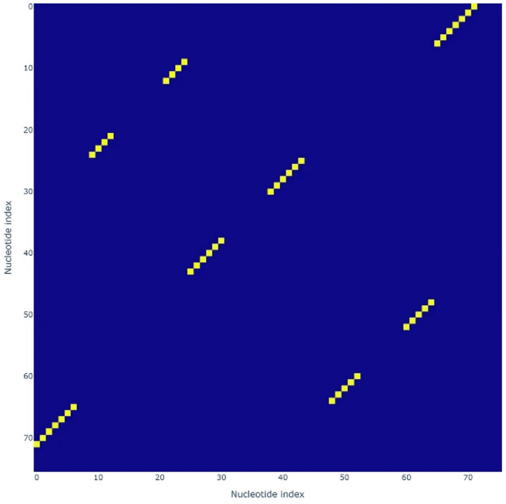
```
GGGGAAUUAGCUCAAAUGGUAGAGCGCUCGCUUUGCUUGCGAGAGGUAGCGGGAUCGAUGCCCGCAUUCUCCACCA
(((((((..((((........))))((((((.......))))))....(((((.......))))))))))))....
```

b

LHHHSSSSSSSSSSSSSHHL

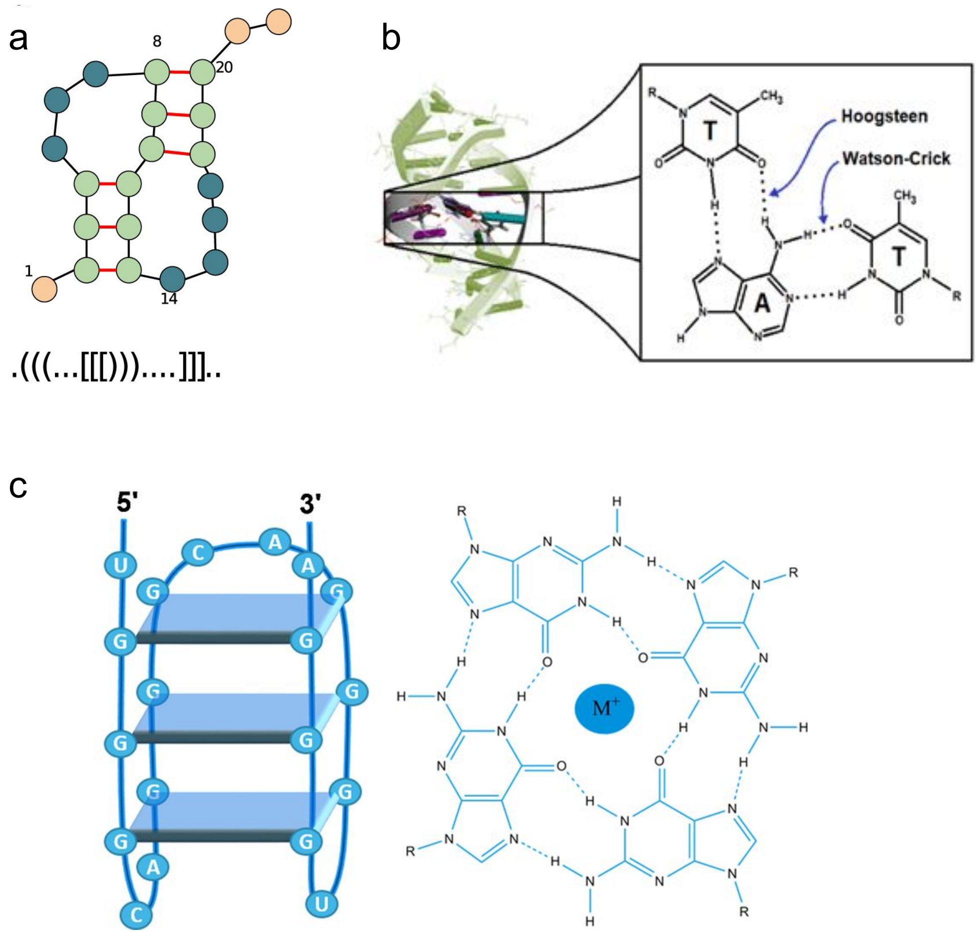


c



# Challenging elements

- a) Pseudoknots
- b) Multipletes
- c) G-quadruplexes



# Metrics

- Precision = TP/(TP+FP)

- Recall = TP/(TP+FN)

- F1-score =

  = 2 x Precision x Recall/(Precision+Recall)

- Accuracy = (TP+TN)/(TP+TN+FP+FN)

- Taminoto accuracy = fraction of characters that is similar in predicted structure and actual structure in dot-bracket notation

Definition of TP, TN, FP and FN.

| | Base pair in predicted structure | Base pair in experimental structure |
|---|---|---|
| TP | Yes | Yes |
| TN | No | No |
| FP | Yes | No |
| FN | No | Yes |

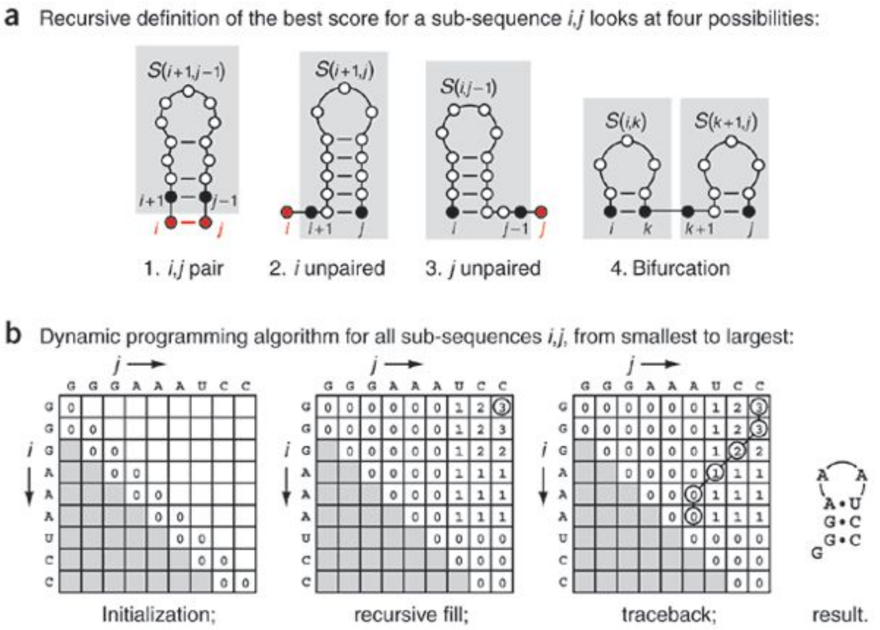
# Basic solutions

- Utilize dynamic programming to obtain structure with minimum free energy or minimum score of other stability function

- Usually rather fast; can generate suboptimal structures and have DNA mode (Mfold)

- Can not predict pseudoknots and multiplets

## Classical approach to 2D structure prediction



a Recursive definition of the best score for a sub-sequence *i,j* looks at four possibilities:

b Dynamic programming algorithm for all sub-sequences *i,j*, from smallest to largest:

| | G | G | G | A | A | A | U | C | C |
|---|---|---|---|---|---|---|---|---|---|
| G | 0 | | | | | | | | |
| G | 0 | 0 | | | | | | | |
| G | | 0 | 0 | | | | | | |
| A | | | 0 | 0 | | | | | |
| A | | | | 0 | 0 | | | | |
| A | | | | | 0 | 0 | | | |
| U | | | | | | 0 | 0 | | |
| C | | | | | | | 0 | 0 | |
| C | | | | | | | | 0 | 0 |

Initialization;

| | G | G | G | A | A | A | U | C | C |
|---|---|---|---|---|---|---|---|---|---|
| G | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 |
| G | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 |
| G | | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 2 |
| A | | | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| A | | | | 0 | 0 | 0 | 1 | 1 | 1 |
| A | | | | | 0 | 0 | 1 | 1 | 1 |
| U | | | | | | 0 | 0 | 0 | 0 |
| C | | | | | | | 0 | 0 | 0 |
| C | | | | | | | | 0 | 0 |

recursive fill;

| | G | G | G | A | A | A | U | C | C |
|---|---|---|---|---|---|---|---|---|---|
| G | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 |
| G | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 |
| G | | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 2 |
| A | | | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| A | | | | 0 | 0 | 0 | 1 | 1 | 1 |
| A | | | | | 0 | 0 | 1 | 1 | 1 |
| U | | | | | | 0 | 0 | 0 | 0 |
| C | | | | | | | 0 | 0 | 0 |
| C | | | | | | | | 0 | 0 |

traceback;

result.

# Basic solutions

| Tool | Best F1-score | Description |
|---|---|---|
| Mfold | 0.42 | free energy minimization, have DNA mode, can generate suboptimal structures, most common implementation |
| RNAfold | 0.52 | free energy minimization, no DNA mode, but allow to turn G-U off, can handle G-quadruplexes via additional algorithm |
| CentroidFold | 0.63 | empiric centroid estimator function, no DNA mode |
| MXfold2 | 0.87 | free energy minimization + deep neural network (SSVM), no DNA mode, python 3.9 implementation |

# Basic solutions tested on aptamers

- 69 DNA structures, 25-57 nt, some has triplexes and G-quadruplexes (G4)

- In case of Mfold authors used best suboptimal solution

Number of accurately (Taminoto accuracy >0.85) predicted structures

| Tool | Accurate structures |
|---|---|
| Mfold | 52% |
| RNAfold | 64% |
| CentroidFold | 36% |

# Basic solutions tested on aptamers

- RNAfold show best results via its ability to predict G4 with satisfactory accuracy
- Structures with pseudoknots are more challenging for Mfold and RNAfold but sample size is small
- All tools show high performance on sequences with triplexes indicating these elements are rare and do not critically affect the rest of 2D structure

Accuracy (nt) of the 2D structure prediction programs on the test set of DNA aptamers

| Tool | Overall Taminoto Accuracy | with Triplexes | with G4 | with Pseudoknots |
|---|---|---|---|---|
| Mfold | 0.76 | 0.95 | 0.45 | 0.68 |
| RNAfold | 0.84 | 0.91 | 0.78 | 0.70 |
| CentroidFold | 0.72 | 0.85 | 0.46 | 0.87 |

# State-of-the-art solutions

| Solution | Best F1-score | Method | Description |
|---|---|---|---|
| Booy et al. | 0,975 | CNN (ResNet) | G-U and pseudoknots allowed, multiplets allowed |
| Qiu | 0,97 | LSTM, CNN | G-U allowed, pseudoknots and multiplets not allowed |
| DMfold | 0,937 | Bi-LSTM | G-U and pseudoknots allowed, multiplets not allowed; best on tRNA, worst on tmRNA (F1=0,619) |
| Zhang et al. | 0,924 | CNN, MLP, DP | G-U allowed, pseudoknots and multiplets not allowed; best on tRNA, worst on 5sRNA (F1=0,823) |
| UFold | 0,91 | CNN (U-Net) | G-U and pseudoknots allowed, multiplets not allowed |
| REDfold | 0,906 | CNN (ResNet) | G-U and pseudoknots allowed, multiplets not allowed; accuracy=0,895 |

# Some other tools

- IPknot - free energy minimization tool which can predict 2D structure with pseudoknots (F1 ~ 0.6)

- SFold - RNA acessibility prediction (Sirna), general statistical folding features (Srna)

- LocARNA - RNA multiple alignment using 2D structure information

- RNAalifold - consensus 2D structure for RNA alignment

- qsfinder, pqsfinder - best G4 prediction

# Tertiary structure prediction

# Introduction to 3D

- Output is PDB (coordinate file)
- Usually 2D structure is used in prediction (higher F1-score in 2D generally causes lower RMSD and improvement in other metrics [7])
- Time-consuming computations
- For long sequences accuracy dramatically decline (lack of data + long RNA usually can form more than one stable conformation)

# Metrics

- RMSD (root mean square deviation of atomic positions) < 3 A - 4 A indicate highly similar structures

- INF (interaction network fidelity), can be calculated for Watson-Crick, stacking and non-canonical interactions measuring accuracy of their prediction

- MCQ (mean of circular quantities) - a measure of torsion angle space simularity (<15 angles means highly similar structures)

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}(\vec{x}_{1,i} - \vec{x}_{2,i})^2}{n}}$$

$$INF = \sqrt{PPV \times STY}$$

$$\text{where } PPV = \frac{|TP|}{|TP| + |FP|}$$

$$\text{and } STY = \frac{|TP|}{|TP| + |FN|}$$

# Approaches to 3D prediction

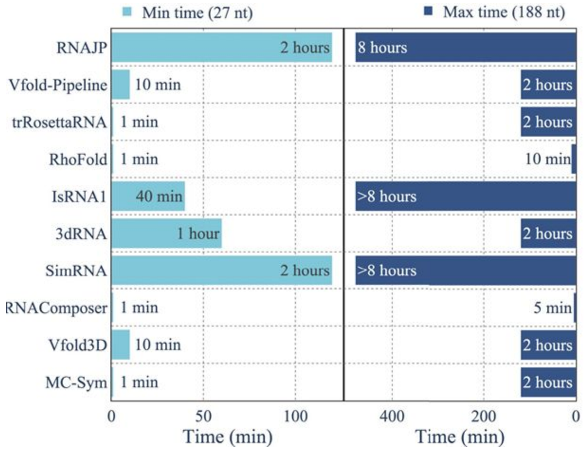
| Approach | Solutions | Method | Limitations |
|---|---|---|---|
| Ab initio | SimRNA, IsRNA1, RNAJP | Use 1-3 atoms per nucleotide for MD/Monte-Carlo sampling evaluated by energy-based scoring function | non-canonical pairs not considered, time-consuming, not accurate scoring functions |
| Template based | MC-Sym, RNABuilder, Vfold3D, RNAComposer, 3dRNA/3dDNA | Utilize nt or SSE motifs to select a template structure from database for each part of the sequence + clashes refinement | non-canonical pairs considered, faster and more well documented than ab initio; can be inaccurate and limited in sequence length due to lack of data |
| DL | RhoFold, trRosettaRNA | Deep learning with attention blocks for ab initio scoring function or direct prediction | Heavy standalone versions, lack of data can cause low accuracy for unseen families |

# Benchmark results

- 65 structures, most are riboswitches, 37-720 nt (most <200 nt)

- While having good RMSD values, the deep learning approaches do not have the best INF and MCQ scores. It means the deep learning approaches can have a general idea of the skeleton structures, but hardly reproduce the specific key RNA interactions [8].

| Tool | RMSD | INF (WC+not-WC+stacking) | MCQ |
|---|---|---|---|
| trRosettaRNA | **6,79** | 0,58 | 33,08 |
| RhoFold | 11,08 | 0,56 | 56,15 |
| MC-Sym | **15,52** | 0,54 | 35,38 |
| 3dRNA | **16,45** | 0,59 | 32,72 |
| Vfold-Pipeline | 18,57 | 0,63 | 25,2 |
| Vfold3D | 18,88 | 0,55 | 30,78 |
| SimRNA | 20,85 | 0,62 | 24,94 |
| IsRNA1 | 21,22 | **0,64** | **24,52** |
| RNAComposer | 21,48 | **0,62** | **25,46** |
| RNAJP | 23,04 | 0,62 | 25,38 |



# Some other tools

- RoseTTAFoldNA - DL solution with specific module for DNA prediction, but need 500 gb of free space for the installation [8]

- RNA Puzzle toolkit, RNAdvisor - tools for RNA structure prediction solutions benchmarking

- Akita - CNN tool for genome DNA 3D conformation prediction

# Conclusion

- Most tools for 2D and 3D structure prediction were developed for RNA, some of them have modifications for DNA, others can be applied for DNA through additional pipeline steps

- Classical 2D prediction tools show satisfactory results on DNA aptamers (RNAfold), their performance can be improved by merging them with G4 and pseudoknots prediction tools and ML/DL parts; other possible approach is to train existing state-of-the-art DL solutions on DNA data

- Existing 3D prediction tools suffer either from time limitations or from accuracy limitations (RMSD for template based and other metrics for DL) or both (ab initio); template based and DL solutions probably can not achieve satisfactory accuracy for DNA without training on DNA data

# References

1. Sullivan R, Adams MC, Naik RR, Milam VT. Analyzing Secondary Structure Patterns in DNA Aptamers Identified via CompELS. Molecules. 2019; 24(8):1572. https://doi.org/10.3390/molecules24081572

2. Lee, S. J., Cho, J., Lee, B. H., Hwang, D., & Park, J. W. (2023). Design and Prediction of Aptamers Assisted by In Silico Methods. Biomedicines, 11(2), 356. https://doi.org/10.3390/biomedicines11020356

3. Afanasyeva, A., Nagao, C., & Mizuguchi, K. (2019). Prediction of the secondary structure of short DNA aptamers. Biophysics and physicobiology, 16, 287–294. https://doi.org/10.2142/biophysico.16.0_287

4. Antunes, D., Jorge, N. A. N., Caffarena, E. R., & Passetti, F. (2018). Using RNA Sequence and Structure for the Prediction of Riboswitch Aptamer: A Comprehensive Review of Available Software and Tools. Frontiers in genetics, 8, 231. https://doi.org/10.3389/fgene.2017.00231

5. Budnik, M., Wawrzyniak, J., Grala, Ł. et al. Deep dive into RNA: a systematic literature review on RNA structure prediction using machine learning methods. Artif Intell Rev 57, 254 (2024). https://doi.org/10.1007/s10462-024-10910-3

6. Saman Booy, M., Ilin, A., & Orponen, P. (2022). RNA secondary structure prediction with convolutional neural networks. BMC bioinformatics, 23(1), 58. https://doi.org/10.1186/s12859-021-04540-7

7. Kulkarni, M., Thangappan, J., Deb, I., & Wu, S. (2023). Comparative analysis of RNA secondary structure accuracy on predicted RNA 3D models. PloS one, 18(9), e0290907. https://doi.org/10.1371/journal.pone.0290907

8. Clément Bernard, Guillaume Postic, Sahar Ghannay, Fariza Tahi, State-of-the-RNArt: benchmarking current methods for RNA 3D structure prediction, NAR Genomics and Bioinformatics, Volume 6, Issue 2, June 2024, lqae048, https://doi.org/10.1093/nargab/lqae048