RNA FOLDING WITH ENERGY MINIMIZATION

PROJECT REPORT

CS 9832a: Topics in Bioinformatics

Dr. Kaizhong Zhang
Professor, Western University

Aasminpreet Singh Kainth
akainth4@uwo.ca

Aasminpreet Singh Kainth
akainth4@uwo.ca

Abstract

The primary molecules of biological interest i.e. DNA, RNA, and proteins, all have primarily linear structures. Their biological function is dependent upon how they interact with other molecules. These interactions are further dependent upon the shape of the molecules involved. Determining structures using experimental methods such as x-ray, crystallography, and Nuclear Magnetic Resonance, are slow, expensive, and work only in specific environmental conditions. All this combines to make computational prediction of structures an important and messy problem. So, the use of bioinformatics methods for the RNA structure prediction is highly required and several approaches have been proposed. In my report, I am discussing about different algorithms which can be used to determine the RNA secondary structure.

Introduction

RNA, abbreviated for Ribonucleic Acid, molecules are an important component of biological substance. It has become one of the central subject of research in recent year. They are the carriers of genetic information between DNA molecules and proteins. They also play an important role in many biological processes, such as catalysis, protein synthesis, immunity, development and many other important biological processes. It is widely believed that RNA molecules are the closest thing to the molecules from which life originally evolved. RNA molecules can perform the function of coding for proteins (i.e. information storage) usually associated with DNA. RNA molecules can have enzymatic functions usually associated with proteins. Protein synthesis occurs with the help of three RNA types i.e. (A) Messenger RNA (mRNA), (B) Transfer RNA (tRNA), (C) Ribosomal RNAs (rRNA). As Messenger RNA (mRNA) transcripts the genetic information, that is a portion of the DNA, and transfers it to the ribosome, which is outside the nucleus. Then begins the translation by Transfer RNA (tRNA) and ribosome (in which Ribosomal RNAs (rRNA) exist as a construction element). After these procedures, an amino acid chain is constructed and is then referred to as protein. [1] [7]

Understanding structure is a first step toward function understanding. Function is what we're ultimately interested in, Sequences is what we have in plenty. As Basic paradigm of structural biology: Sequence → Structure → Function. There exist many factors which determine the Molecular shape/structure which are (1) Electrostatic forces (i.e. positive/negative charges), (2) The size/shape of molecular subunits such as amino acids and nucleotide bases, (3)

Aasminpreet Singh Kainth
akainth4@uwo.ca

Hydrophobicity of the associated bases, (4) The current environmental conditions which are temperature, salinity, acidity, and (5) Molecular bonds. There are several levels of structures primary structure which usually refers to the raw sequence itself, secondary structure which refers to identifying certain self-interacting features of the structure, ternary structure is the complete 'geometric' description of molecule and quandary structure which identifies how certain parts of structures interact with other structures.

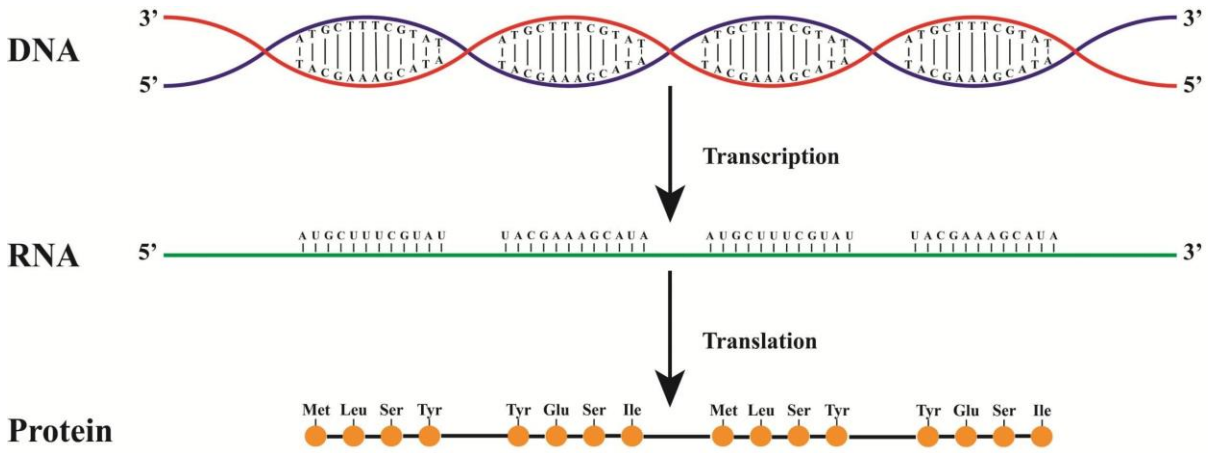


Figure 1 Protein synthesis

The shape of RNA determines its function in a cell. If we could predict exactly what shape the RNA sequence will fold into, we could find the function of that RNA. For illustration, we might be able to design an RNA that could control cancer causing genes. RNA molecules are generally single stranded which fold back onto themselves into predefined 3D or structures shapes. The folding problem seeks the structure or shape of a given sequence. The shape of certain RNAs plays a major role in determining its interaction with other molecules.

RNA typically is a single-stranded biopolymer. RNA consists of Ribose Nucleotides which are nitrogenous bases appended to a ribose sugar. These Ribose Nucleotides are attached by phosphodiester bonds, forming strands of varying lengths. The nitrogenous bases in RNA are adenine, guanine, cytosine and uracil, which replaces thymine in DNA. [1]

Aasminpreet Singh Kainth
akainth4@uwo.ca

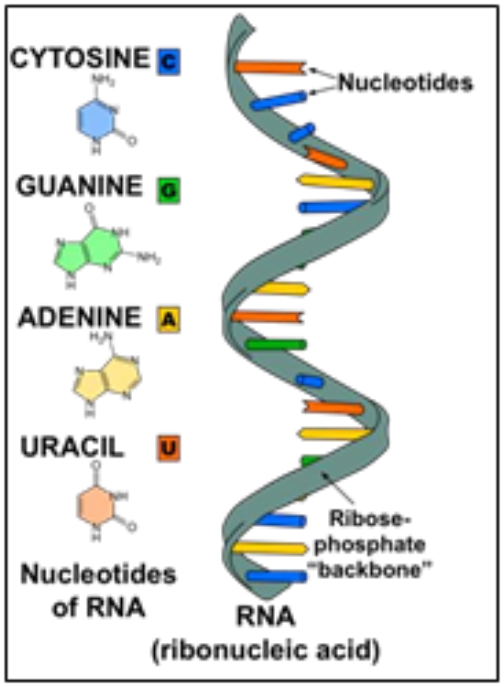


Figure 2 Structure of RNA

As RNA is composed of a combination of four nucleotides: adenine (A); cytosine (C); guanine (G); uracil (U). Since RNA is single-stranded, its component bases tend to bond with other bases. G-C and A-U form complementary hydrogen bonded base pairs also known as canonical Watson-Crick pairs and, G-C base pair is more stable as having 3 hydrogen bonds whereas A-U base pairs with 2 bonds is less stable.[3]

Secondary structure of RNA can be predicted successfully from experimental thermodynamic data on secondary structure. Secondary structure of RNA is much more stable than the tertiary structure. Because the energies involved in the formation of secondary structure are larger than those involved in tertiary interactions. So, secondary structural elements can exist and be stable by themselves. The helices, loops, bulges, and junctions are basic secondary structure elements in RNA.

Aasminpreet Singh Kainth
akainth4@uwo.ca

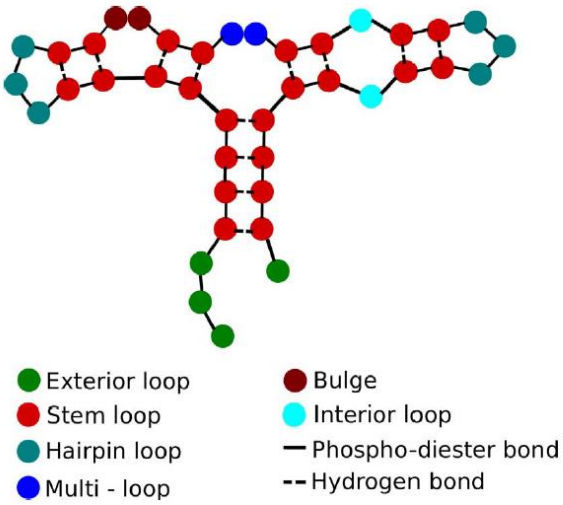


Figure 3 Decomposition of an RNA secondary structure into structural elements.

There are various RNA secondary structure notations such as parentheses notation, Arc notation, Circle Plot.

Parentheses notation

..(((…..)))……((((…….....)).))…

Arc (‘rainbow’) notation

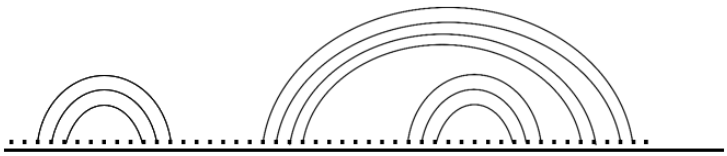


Figure 4 RNA secondary structure notations

There are two main methods to predict RNA secondary structure. First method is Comparative sequence analysis. This method infers base-pairs by determining canonical pairs that are common among multiple homologous sequences. In Comparative sequence analysis, we find structure with the most base pairs. To predict the secondary structure of a single sequence, the most popular methods use Free Energy Minimization with computer algorithms based on dynamic programming. The dynamic programming approach to RNA secondary structure prediction relies on the fact that structures can be recursively decomposed into smaller components with independent energy contributions. In each of the decomposition steps only a single loop (or stacking of two consecutive base pairs) needs to be evaluated. [2]

Aasminpreet Singh Kainth
akainth4@uwo.ca

Nussinov Algorithm

Nussinov Algorithm is a recursive algorithm which calculates the best structure for small subsequence and works its way outward to larger and larger subsequence. The simplest approach to predicting the secondary structure of RNA molecules is to find the configuration with the greatest number of bases paired. The number of possible configurations to be inspected grows exponentially with the length of the sequence. We can employ dynamic programming algorithm to find an efficient solution using Nussinov Folding Algorithm. This algorithm was proposed in 1978. [3]

The key idea of the recursive calculation is that there are only four possible ways of the getting the best structure for *i*, *j* from the best structures of the smaller subsequences. It computes the highest number of nucleotides coupling with 2 structure.

Four ways to get the best structure between position i and j from the structures of the smaller sub sequences.

1. $\boldsymbol{S(i+1, j)}$ – *i* unpaired– Add unpaired position *i* onto best structure for subsequence *i* + 1, *j*; Take lower element.
2. $\boldsymbol{S(i, j-1)}$ – *j* unpaired– Add unpaired position *i* onto best structure for subsequence $\boldsymbol{i}$, $\boldsymbol{j-1}$; Take left element.
3. $\boldsymbol{S(i+1, j-1) + e(i, j)}$– *i*, *j* paired– Add unpaired position *i*, *j* onto best structure for subsequence $\boldsymbol{i}$ + 1, $\boldsymbol{j-1}$; Take the diagonally left Lower Down
4. Bifurcation, or combining two optimal substructures ranging from i <k <j. $\boldsymbol{Max\ i < k < j}$ $\{ \boldsymbol{S(i, k) + S(k+1), j}\}$– bifurcation – combine two optimal substructure *i*, *k* and *k* + 1, *j*. Take left row, bottom column. [3]

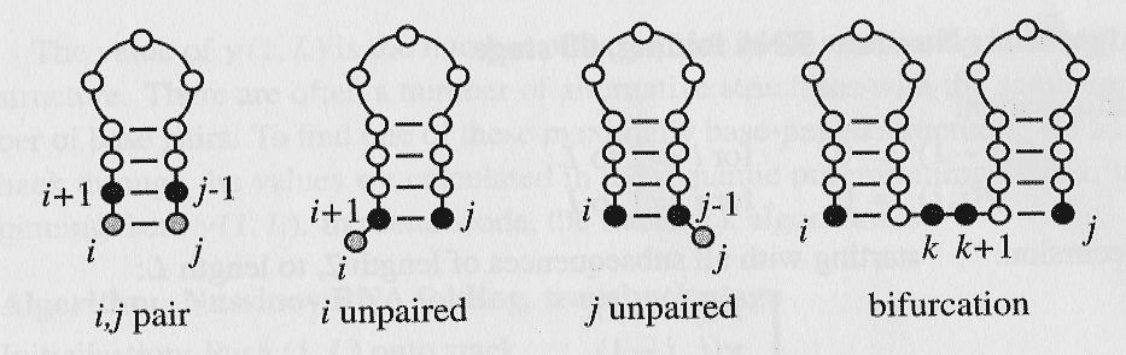


Figure 5 Four ways to get the best structure between position i and j from the structures of the smaller sub sequences.

Aasminpreet Singh Kainth
akainth4@uwo.ca

Flow chart for Nussinov algorithm

The flow chart for Nussinov algorithm is described below:

1. Obtain RNA sequence for which we must predict RNA secondary structure.

2. Create a scoring matrix and place sequence – The matrix will be empty initially and then, place the RNA sequence on top as well as left side of the matrix.

3. Set the diagonals of the matrix to zero as well as the lower diagonal.

4. Devise Scoring Scheme – The scoring scheme of Nussinov Algorithm can be done by looking at the bottom, at the left, at the diagonal as well as the two rows beyond the left and bottom.

5. Set each matrix position by calculating 4 conditions.

Aasminpreet Singh Kainth
akainth4@uwo.ca

6. Set the maximum value from the 4 conditions into matrix position.

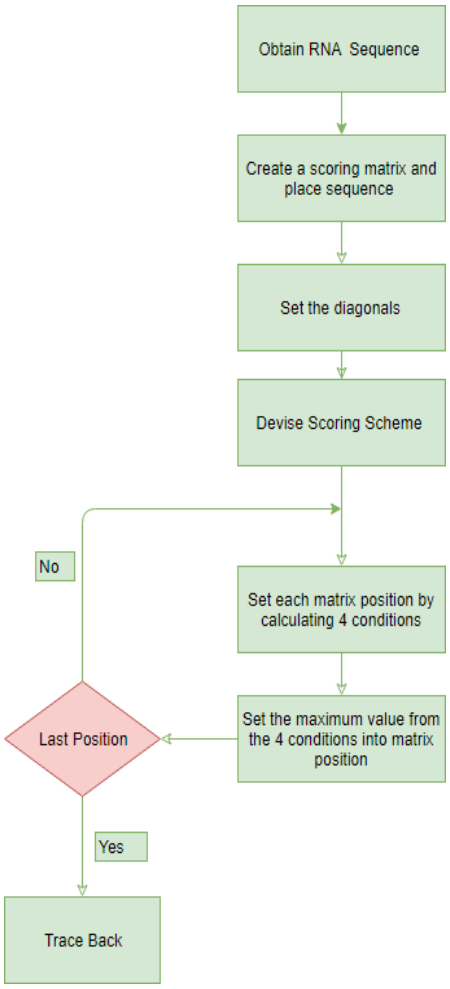


Figure 6 Nussinov algorithm

7. Fill all the matrix position that are above the diagonal in the matrix and check if it's last position, if yes go to 8th step else go to 5th step.

8. Trace back – Once the matrix is filled, a new step has to be done i.e. called the trace back which will help to extract meaning from the score which we have put into the matrix.

Trace back strategy is used to recover the optimal structure.

Aasminpreet Singh Kainth
akainth4@uwo.ca

Algorithm for matrix fill stage:

Algorithm

- Input: Sequence $x = (x1, x2, \ldots, xL)$
- Output: Maximal number $S(i, j)$ of base pairs for $(xi, \ldots, xj)$.
- Initialization:

S(i, i) = 0 *for I = 2 to L.*
S(i, i-1) = 0 *for I = 2 to L;*

for $n = 2$ *to* L *do*
for $j = n$ *to* L *do*
Set $i = j - n + 1$

$$S(i, j) = \max \begin{cases} S(i+1, j) \\ S(i, j-1) \\ S(i+1, j-1) + e(i, j) \\ \boldsymbol{Max}_{i<k<j}\{S(i, k) + s(k+1, j)\} \end{cases}$$

Return $S(1, L)$

Algorithm for Traceback - Retrieving the Structure:

```
if i < j then
   if S(i, j) = S(i + 1, j) then
      traceback(i + 1, j)
   else if S(i, j) = S(i, j − 1) then
            traceback(i, j − 1)
   else if S(i, j) = S(i + 1, j − 1) + w(i, j)
then
            print base pair (i, j)
            traceback(i + 1, j − 1)
   else for k = i + 1 to j − 1 do
    if S(i, j) = S(i, k) + S(k + 1, j) then
            traceback(i, k)
            traceback(k + 1, j)
            break
end
```

Aasminpreet Singh Kainth
akainth4@uwo.ca

There are few limitations of The Nussinov Algorithm:

1. Does not give accurate structure predictions, as
   I. no stacking of base pairs considered
   II. loop sizes not distinguished
   III. no special scoring of multi-loops
2. Misses:
   I. nearest neighbour interactions:
      Free energy of a structure approximated as the sum of loop energies
      Loop energies depend on loop type and size
      Free energies are dependent on temperature and ionic conditions
   II. stacking interactions [7]
      Major source of stabilizing energy. All 21 combinations measured, accuracy at least 0.1 kcal/mol [8]

		$(i+1, j-1)$					
		CG	GC	GU	UG	AU	UA
(i, j)	CG	-2.4	-3.3	-2.1	-1.4	-2.1	-2.1
	GC	-3.3	-3.4	-2.5	-1.5	-2.2	-2.4
	GU	-2.1	-2.5	1.3	-0.5	-1.4	-1.3
	UG	-1.4	-1.5	-0.5	0.3	-0.6	-1.0
	AU	-2.1	-2.2	-1.4	-0.6	-1.1	-0.9
	UA	-2.1	-2.4	-1.3	-1.0	-0.9	-1.3

Figure 7 stabilizing energy combinations measured

   III. loop length preferences

The computational complexity of algorithm is as follows: (for sequence of length N)

Memory - $O(N^2)$ Time - $O(N^3)$

Aasminpreet Singh Kainth
akainth4@uwo.ca

Energy Minimization methods

Energy minimization algorithm predicts the correct secondary structure by minimizing the free energy (ΔG): Gibbs Free Energy G of a system:

$G = H - TS;$ $\Delta G = 0;$ $\Delta G > 0;$ $\Delta G < 0$

where H is the enthalpy (potential to perform work), T the absolute temperature and S the entropy (measure of disorder). [5] [6]

ΔG = 0 indicates equilibrium; ΔG > 0 indicates an unfavorable process and; ΔG < 0 indicates a favorable process.

Therefore, G is calculated as sum of individual contributions of: loops, base pairs and secondary structure elements. Energies of stems calculated as stacking contributions between neighboring base pairs.

RNA folding algorithm is dictated by biophysics rather than by counting and maximizing the number of base pairs. The most sophisticated secondary structure prediction method for single RNAs is the ZUKER algorithm, an energy minimization algorithm which assumes that the correct structure is the one with the lowest equilibrium free energy (ΔG). [2]

Zuker's Algorithm

The minimum energy structure can be calculated recursively by a dynamic programming algorithm. The principal difference is that the because of stacking parameters, two matrices (called V and W) are kept instead of one. [2]

Aasminpreet Singh Kainth
akainth4@uwo.ca

$W(i, j)$ is the energy of the best structure on i, j.

$V(i, j)$ is the energy of the best structure on i, j given i, j are paired.

Assumptions

In predicting minimum energy of RNA secondary structure, several simplifying assumptions are made:

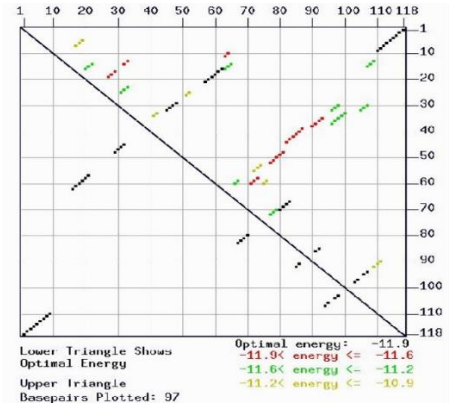
1. The most likely structure is identical to the energetically preferable structure
2. Nearest-neighbor energy calculations give reliable estimates of an experimentally achievable energy measurements
3. Usually we can neglect pseudoknots [8]

Zuker Algorithm Implementation (MFOLD)

Example Sequence: [9]

GCTTACGACCATATCACGTTGAATGCACGCCATCCCGTCCGATCTGG
CAAGTTAAGCAACGTTGAGTCCAGTTAGTACTTGGATCGGAGACGGC
CTGGGAATCCTGGATGTTGTAAGCT

MFOLD Energy Dot Plot:



Resulted Optimal Structures:

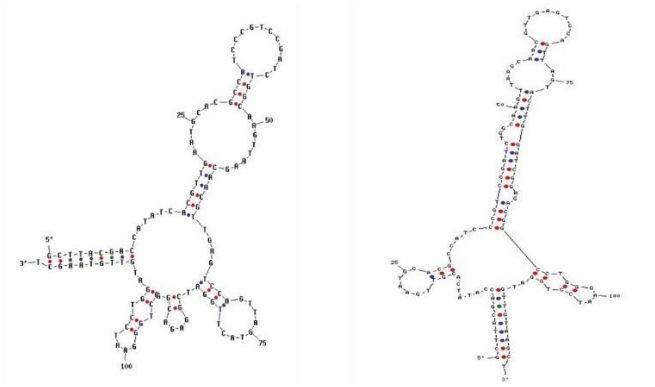


Figure 8 Zuker Algorithm Implementation (MFOLD) [9]

Aasminpreet Singh Kainth
akainth4@uwo.ca

CDPfold – New method

It is the new Method of RNA Secondary Structure Prediction. It combines a convolutional neural network and dynamic programming as well as a sequence alignment method. It is proposed by Zhang et. al in 2019. [4]

Convolutional Neural Network: The word ‘convolutional’ refers to the filtering process. A convolutional neural network (CNN) is a type of neural network Like a normal neural network, a convolutional neural network is made up of multiple layers.

Neural Network: A regular neural network is consisted of an input layer, hidden layers and an output layer.

Where, the input layer accepts inputs in different forms. The hidden layers perform calculations on these inputs. The output layer is then used to deliver the outcome of the calculations and extractions.

Each of these layers has neurons that are connected to neurons in the previous layer, and each neuron has its own weight.

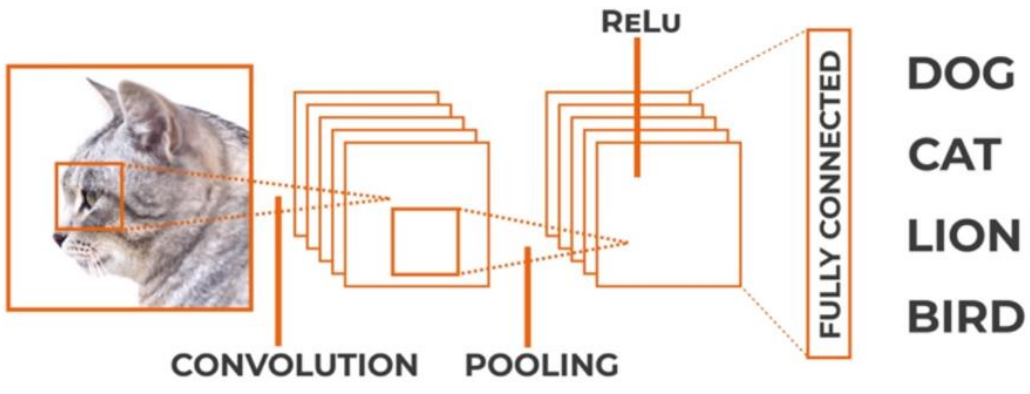


Figure 9 Convolutional Neural Network

The ReLu layer (rectified linear unit layer) acts as an activation function, ensuring non-linearity as the data moves through each layer in the network.

Fully connected layer: allows you to perform classification on your dataset.

Aasminpreet Singh Kainth
akainth4@uwo.ca

The process and Analysis of CDPfold

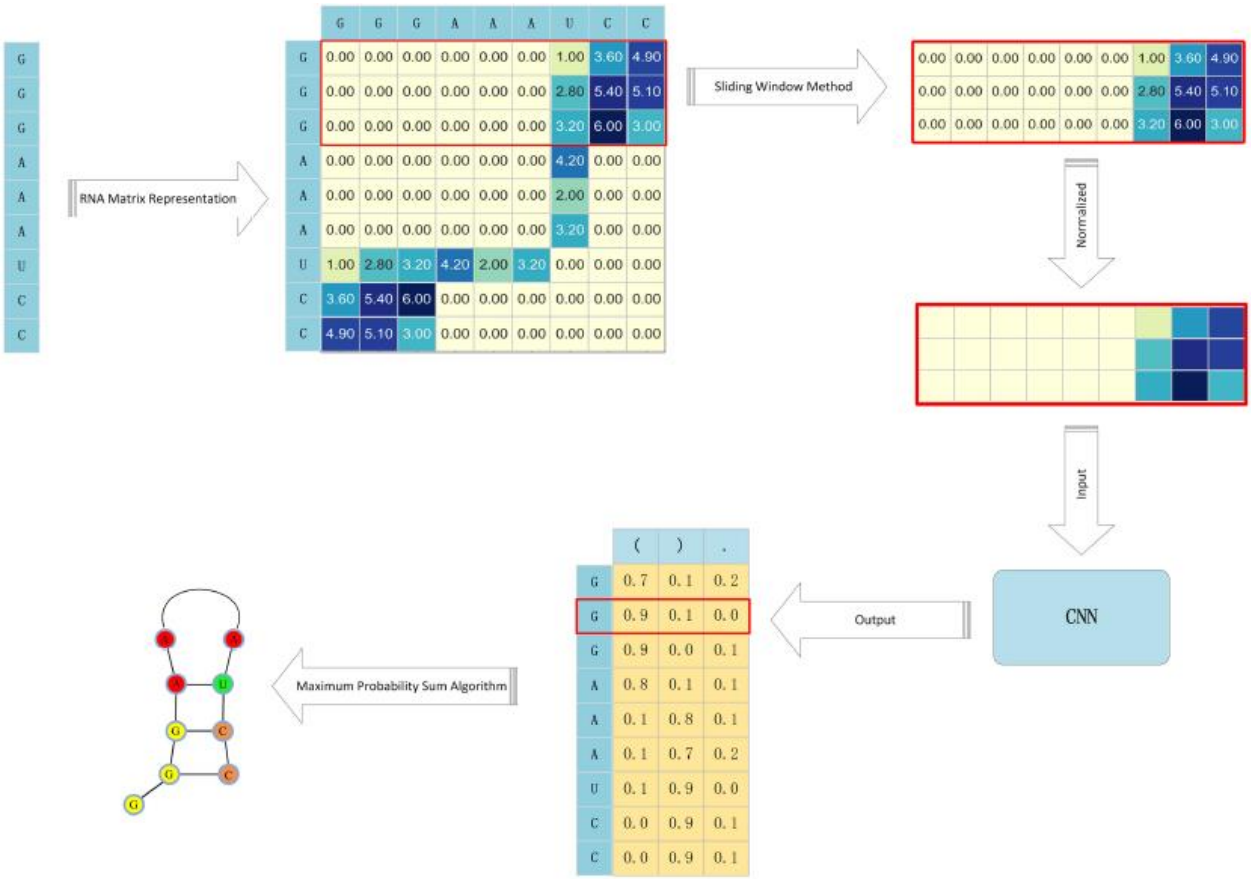


Figure 10 The process and Analysis of CDPfold

Here, a CNN is made to extract the characteristics of actual implicit features from large scale data and predicted the matching probability of each base on the RNA sequence. CNNs can use the currently collected RNA sequences as training samples, which solves the constraints of homologous sequences in comparative sequence analysis. For the probabilistic results obtained by the convolutional neural network, they used the iterative idea of dynamic programming and the definition of the RNA secondary structure to obtain the base matching probability and the maximum RNA secondary structure.

Aasminpreet Singh Kainth
akainth4@uwo.ca

RESULTS [4]

For the prediction of an RNA secondary structure gained by the CDPfold, results are attained by two indicators, sensitivity and specificity. [4]

Sensitivity refers to the predicted percentage of all base pairs in the real structure (recall-rate in machine learning)

Specificity refers to the correct percentage of all predicted base pairs (precision-rate in machine learning)

The F-score can be used to measure the precision and recall.

$$\text{F-score} = \frac{2 \times \text{Sensitivity} \times \text{Specificity}}{\text{Sensitivity} + \text{Specificity}}$$

Table shows the accuracy of the designed algorithm compared with other algorithms on the 5sRNA dataset.

TABLE | Comparison of algorithms in 5sRna.

Software	5sRNA		
	Sensitivity	**Specificity**	***F*-score**
mfold	0.693	0.704	0.698
RNAfold	0.694	0.704	0.699
cofold	0.585	0.591	0.588
Sfold	0.703	0.733	0.718
CDPfold	0.932	0.916	0.924

Table 1 Comparison of algorithms in 5sRna

Obviously, the sensitivity and specificity of the designed algorithm are significantly higher than that found in other algorithms.

Aasminpreet Singh Kainth
akainth4@uwo.ca

Using the F-score, they get the predicted effect of the designed generic model on the three types of RNA datasets. They used the same test data to perform experiments under other published algorithms.

TABLE | Comparison of three types of RNA based on their prediction accuracy.

Software	5sRNA	tRNA	srpRNA
Mfold	0.698	0.631	0.566
RNAfold	0.699	0.632	0.577
CDPfold	0.911	0.905	0.823

Table 2 Comparison of three types of RNA based on their prediction accuracy

Discussion

1. Experimentally, the method has had good performance in predicting the accuracy of a RNA secondary structure.
2. Although CDPfold has achieved good results in RNA secondary structure prediction, some problems encountered during the experimental process.
3. The results predicted by the CDPfold method still need to be further corrected in the results predicted by the convolutional neural network.
4. In the current prediction of the RNA secondary structure[4], the prediction of pseudoknots is still a difficult point. In this study, it was found that 5sRNA, srpRNA, and tRNA are free of pseudoknots, while most of RNasePRNA and tmRNA have pseudoknots.
5. The RNA structure representation method used in [4] uses the dot bracket representation. However, the dot bracket representation does not reflect the false knots present in the RNA structure. Therefore, the data containing the pseudoknots are deleted in the experiment.
6. The prediction of the secondary structure of longer RNA sequences is not reflected. This is because the current experimental methods are not perfect enough.

Aasminpreet Singh Kainth
akainth4@uwo.ca

References

[**1**] Recent advances in RNA folding; Jörg Fallmanna, Sebastian Willb, Jan Engelhardta, Björn Grüningc, Rolf Backofenc, Peter F. Stadlera

[**2**] Improved predictions of secondary structures for RNA; JOHN A. JAEGER*, DOUGLAS H. TURNER*t, AND MICHAEL ZUKER*

[3] Fast algorithm for predicting the secondary structure of single-stranded RNA (computer program/polynucleotide/RNA folding); RUTH NUSSINOV* AND ANN B. JACOBSON

[**4**] A New Method of RNA Secondary Structure Prediction Based on Convolutional Neural Network and Dynamic Programming; Hao Zhang1, Chunhe Zhang1, Zhi Li 2, Cong Li 1, Xu Wei 1, Borui Zhang3 and Yuanning Liu1

[**5**] Hogeweg P Hesper, B 1984 Energy directed folding of RNA sequences Nucleic Acids Research, 12 1 Part 1 67 74 doi 10 1093 nar 12 1 part 1 67

[**6**] Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization, D Mathews Rna 2004

[**7**] Haoyue Fu, Lianping Yang, Xiangde Zhang, "An RNA secondary structure prediction method based on minimum and suboptimal free energy structures", Journal of Theoretical Biology, vol. 380, pp. 473, 2015

[**8**] M. E. Nebel and A. Scheid, "Analysis of the Free Energy in a Stochastic RNA Secondary Structure Model," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, no. 6, pp. 1468 1482, Nov. Dec. 2011

[**9**] http://www.bioinfo.rpi.edu/~zukerm/Bio-5495/RNAfold-html/