

**Project Report**

***Submitted to: Ma’am Huma***

***Course name: Modern Programming Language***

***Semester: BS-Bioinformatics VII***

SUBMITTED BY:  
Adnan younas  
Alina Muhammad Shah  
Akasha amjad  
Oneza Hassan Alvi  
Wajeeha Iqbal

Date: 3rd Jan,2023

***Topic:***

Prediction of RNA secondary structure on the basis

Nussinov algorithm.

Table of content

|  |  |
| --- | --- |
| content | Page no. |
| Introduction | 3 |
| Pseudocode | 4 |
| Algorithm | 5 |
| Flow chart | 6 |
| Code | 8 |
| Output | 13 |
| Scope | 14 |
| Conclusion | 15 |
| Reference | 16 |

Introduction:

*Purpose:* Secondary structure prediction of RNA sequence.

*Objectives:*

* Prediction of RNA sequence.
* A mechanism that once made can help us to do several analyses without the need of applying the methodologies repeatedly
* Bringing accuracy and putting forward a better approach for the analysis.

RNA secondary structure consists of nucleotides that are in one of two states, paired or unpaired, where pairing includes all base-base interactions. In general, most base pairings are adjacent and antiparallel with other base pairings to form secondary structure helices. The most popular approach for predicting RNA secondary structures is based on thermodynamic models, such as Turner's nearest-neighbor model in which secondary structure is decomposed into several characteristic substructures, called nearest-neighbor loops, such as hairpin loops, internal loops, bulge loops etc.

Many machine learning methods have been developed to predict RNA secondary structure, and exhibit good performance by exploiting evolutionary information, as well as statistic information about RNA subsequences. The code is a script for predicting the secondary structure of an RNA sequence, using the Turner model for predicting base pair stability.

Pseudo code:

The script begins by reading in the RNA sequence from the command line and checking for empty input or invalid characters. It then reads in the temperature and pH from the command line, with default values of 25 C and pH 7 if not provided. The script then initializes an array for the secondary structure prediction and defines a threshold for base pair stability.

The script then defines thermodynamic parameters for each base pair, including enthalpy and entropy values. It also defines various parameters for correcting the stability prediction for salt concentration, temperature, and PH. Finally, the script loops through the sequence and predicts base pairs by checking the stability of each pair using the 1s \_stable function, which uses the Turner model to calculate the stability of the base pair based on the thermodynamic parameters and correction factors defined earlier in the script. If the stability of the base pair is above the defined threshold, it is added to the prediction array.

The whole process is called as **Rnai** or **RNA interference** and that’s how these molecules activate and shut down biological pathways.

Algorithm:

Turner model is being used firstly to calculate the stability of the base pair based on the thermodynamic parameters and correction factors. If the stability of the base pair is above the defined threshold, it is added to the prediction array.

***Nussinov algorithm*** is also being used in the script, works by checking the adjacent base pairs If the base pairs are complementary then the RNA structure will bend at that point. Algorithm checks for each and every nucleotide in the sequence present at I and i+ 1 index and will create base pairing and will store the results of prediction.

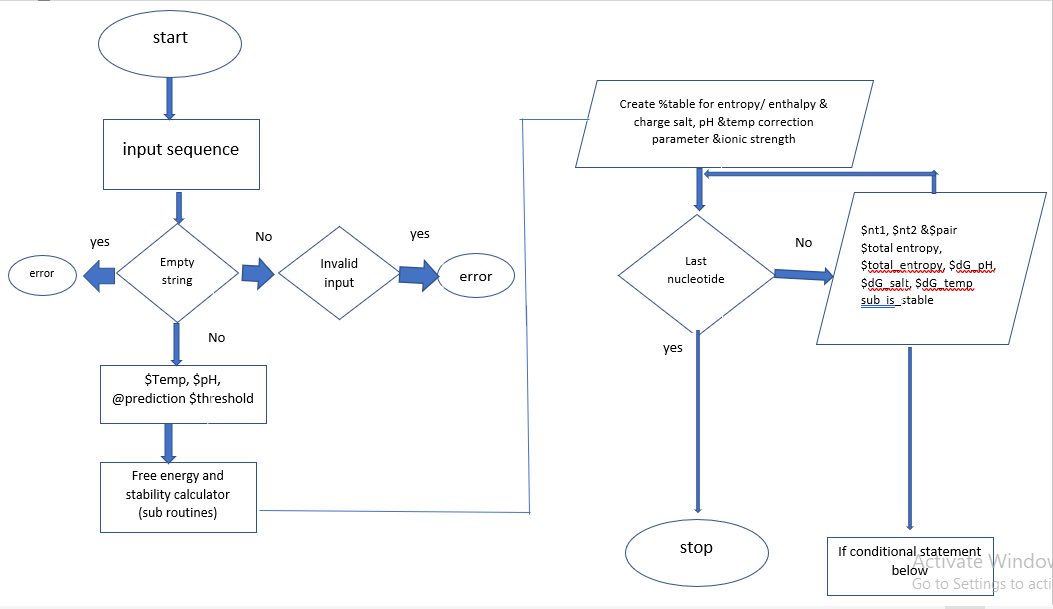
Nussinov algorithm:

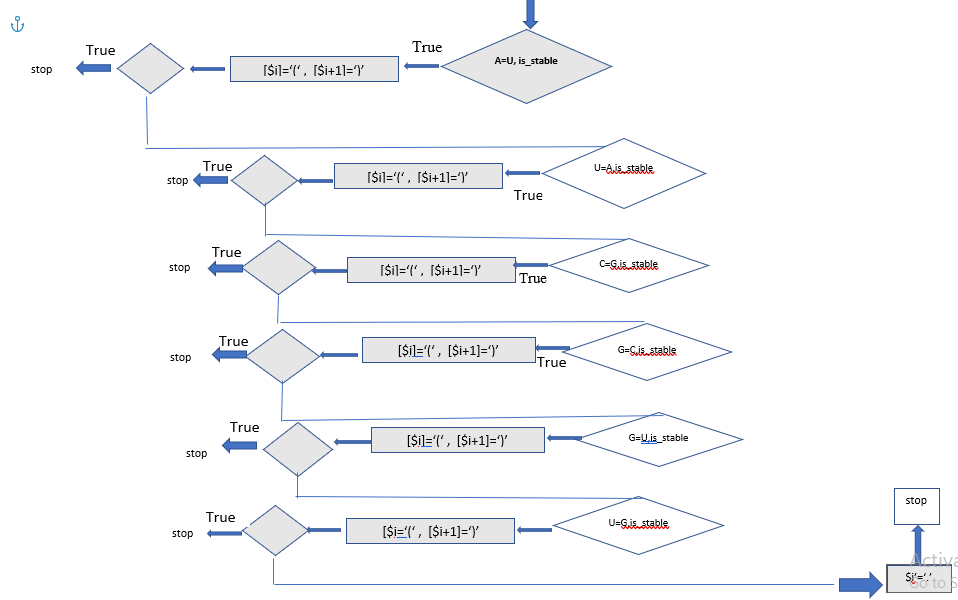
The **Nussinov algorithm** is a [nucleic acid structure prediction](https://en.wikipedia.org/wiki/Nucleic_acid_structure_prediction) algorithm used in [computational biology](https://en.wikipedia.org/wiki/Computational_biology) to predict the folding of an [RNA](https://en.wikipedia.org/wiki/RNA) molecule that makes use of [dynamic programming](https://en.wikipedia.org/wiki/Dynamic_programming) principles.[[1]](https://en.wikipedia.org/wiki/Nussinov_algorithm#cite_note-1) The algorithm was developed by [Ruth Nussinov](https://en.wikipedia.org/wiki/Ruth_Nussinov) in the late 1970s.

***Limitations of Nussinov:***

The Nussinov algorithm does not account for the three-dimensional shape of RNA, nor predict [RNA pseudoknots](https://en.wikipedia.org/wiki/Pseudoknot). Furthermore, in its basic form, it does not account for a minimum [stem loop](https://en.wikipedia.org/wiki/Stem_loop) size. However, it is still useful as a fast algorithm for basic prediction of secondary structure.

FLOW CHART:





Code Interpretation:

EXPLANATION

**CODE**

use strict;

use warnings;

# Read the RNA sequence from the command line

print “enter the rna seq: \n”;

my $Sequence = $ARGV[0];

my $sequence = uc($Sequence);

# Check for empty input

if (!$sequence) {

die "Error: No input RNA sequence provided.\n";

}

# Check for invalid characters in the sequence

if ($sequence =~ /[^ACGU]/) {

die "Error: Invalid character in the RNA sequence.\n";

}

# Read the temperature and pH from the command line (defaults to 25 C and pH 7)

print “enter the temperature: \n”;

my $temperature = $ARGV [1] || 25;

print “enter the pH: \n”;

my $pH = $ARGV[2] || 7;

# Initialize the secondary structure prediction

my @prediction;

#The use strict and use warnings directives are used to enable strict syntax checking and enable warnings, respectively.

#The RNA sequence is read from the command line and stored in the $sequence variable.

# Read the RNA sequence from the command line prompt.

# Check for empty input #using an RNA sequence as input (string) containing AGCU and not containing T.

# Check for invalid characters in the sequence

# Check for temperature and ph.

#The @prediction array is initialized to store the predicted secondary structure.

Code Interpretation:

EXPLANATION

**CODE**

This part of the code is defining two hash tables, %enthalpy and %entropy, which store the enthalpy and entropy values for each possible base pair in an RNA sequence.

# Define the nucleotide charge atvarious ph values

The code is then initializing two variables, $total\_enthalpy and $total\_entropy, to 0.

These variables will be used to store the total enthalpy and entropy values for all the base pairs in the RNA sequence.

The code then defines several constants which will be used in the salt correction calculation later in the function.

$R is the gas constant in cal/mol/K, $T is the absolute temperature in K, and $F is Faraday's constant in C/mol.

my $threshold = 0.5;

sub calculate\_stability

{

my %enthalpy = (

'AA' => -7.6, 'AC' => -8.4, 'AG' => -7.8, 'AU' => -7.2,

'CA' => -8.5, 'CC' => -8.0, 'CG' => -10.6, 'CU' => -7.8,

'GA' => -8.2, 'GC' => -9.8, 'GG' => -8.0, 'GU' => -8.4,

'UA' => -6.9, 'UC' => -7.2, 'UG' => -7.8, 'UU' => -7.2

);

my %entropy = (

'AA' => -22.2, 'AC' => -22.4, 'AG' => -21.0, 'AU' => -20.4,

'CA' => -22.7, 'CC' => -19.9, 'CG' => -27.2, 'CU' => -21.0,

'GA' => -22.2, 'GC' => -24.4, 'GG' => -19.9, 'GU' => -22.4,

'UA' => -20.3, 'UC' => -19.0, 'UG' => -21.0, 'UU' => -20.4

);

my %charge = ('A'=> [-0.500, -0.500, 0.500,0.500],‘

C' => [0.500, 0.500, -0.500,-0.500],

'G' => [-0.500, 0.500, 0.500,-0.500],

'U' => [0.500, -0.500, -0.500, 0.5001]);

my $total\_enthalpy = 0;

my $total\_entropy = 0;

# Define the salt correction parameters

my $R = 1.987; # Gas constant in cal/mol/K

my $T = 310.15; # Absolute temperature in K

my $F = 96485.3415; # Faraday's constant in C/mol

Code Interpretation:

EXPLANATION

**Code**

This chunk of code is defining more constants which will be used in the calculation of the stability of the RNA secondary structure

.$I is the ionic strength of the solution in M, and $A and $B are ionic strength correction parameters. $dH and $dS are enthalpy and entropy correction factors in cal/mol and cal/mol/K, respectively.

$pK\_NH, $pK\_CO, and $pK\_OH are the pK values of the NH, CO, and OH groups, respectively.

The code is also initializing three variables, $dG\_salt, $dG\_temp, and $dG\_pH, to 0.

These variables will be used to store the salt correction, temperature correction, and pH correction values, respectively.The code is then looping through the RNA sequence and summing the enthalpy and entropy values for each base pair in the sequence.

It is storing the sum in the $total\_enthalpy and $total\_entropy variables, respectively.

# Define the ionic strength and ionic strength correction parameters

my $I = 0.050; # Ionic strength in M

my $A = 0.50; # Ionic strength correction parameter

my $B = 0.75; # Ionic strength correction parameter

# Define the temperature correction parameters

my $dH = -0.368; # Enthalpy correction factor in cal/mol

my $dS = -1.83; # Entropy correction factor in cal/mol/K

# Define the pH correction parameters

my $pK\_NH = 6.6; # pK of the NH group

my $pK\_CO = 10.7; # pK of the CO group

my $pK\_OH = 13.1; # pK of the OH group

$dG\_salt=0;

$dG\_temp=0;

my $dG\_pH = 0;

for (my $i = 0; $i < length($sequence); $i++) {

my $nt1 = substr($sequence, $i, 1);

my $nt2 = substr($sequence, $i + 1, 1);

my $pair = $nt1 . $nt2;

$total\_enthalpy += $enthalpy{$pair};

$total\_entropy += $entropy{$pair};

Code Interpretation:

EXPLANATION

**CODE**

This chunk of code is performing several calculations related to the stability of the RNA secondary structure .

The code calculates the salt correction,

temperature correction, and pH correction using the constants defined earlier in the function.

It then calculates the Gibbs free energy, $dG, as the sum of the enthalpy correction, temperature correction, salt correction, and pH correction.

The code then gets the current and next nucleotides in the RNA sequence and defines a function called is\_stable, which takes as input the current and next nucleotides, temperature, and pH.

The function calculates the stability of the base pair using the Turner model and the calculate\_stability function, and returns 1 if the stability is above the threshold or 0 otherwise.predict\_rna.pl

# Calculate the salt correction

$dG\_salt = ($R \* $T) / $F \* (log(1 - $i / 2) + $A \* $i / (1 - $i / 2));

# Calculate the temperature correction

my $dG\_temp = $dH - $T \* $dS;

# Calculate the pH correction

my $nt = substr($sequence, $i, 1);

my @charges = @{$charge{$nt}};

$dG\_pH += $charges[0] \* ($pK\_NH - $pH) + $charges[1] \* ($pK\_CO - $pH) + $charges[2] \* ($pK\_OH - $pH);

#calculate gibs free energy

my $dG = $dH - &T \* &dS + &dG\_salt + &dG\_temp + &dG\_pH

# Get the current and next nucleotides

my $curr = substr($sequence, $i, 1);

my $next = substr($sequence, $i+1, 1);

sub is\_stable {

my ($curr, $next, %enthalpy, %entropy, $R, $T, $F, $I, $A, $B, $dH, $dS, $pK\_NH, $pK\_CO, $pK\_OH, $temperature, $pH) = @\_ ;

# Calculate the stability of the base pair using the Turner model

my $stability = calculate\_stability($curr, $next, %enthalpy, %entropy, $R, $T, $F, $I, $A, $B, $dH, $dS, $pK\_NH, $pK\_CO, $pK\_OH, $temperature, $pH)

# Return 1 if the stability is above the threshold, 0 otherwise

return $stability > $threshold ? 1 : 0;

}

}

}

Code Interpretation:

EXPLANATION

**Code**

# Predict a base pair if the nucleotides are complementary and the base pair is stable at the given temperature and pH

for (my $i = 0; $i < length($sequence); $i++) {

my $curr = substr($sequence, $i, 1);

my $next = substr($sequence, $i+1, 1);

if ($curr eq 'A' && $next eq 'U' && is\_stable($curr, $next, $temperature, $pH)) {

$prediction[$i] = '(';

} elsif ($curr eq 'U' && $next eq 'A' && is\_stable($curr, $next, $temperature, $pH)) {

$prediction[$i] = '(';

} elsif ($curr eq 'C' && $next eq 'G' && is\_stable($curr, $next, $temperature, $pH)) {

$prediction[$i] = '(';

} elsif ($curr eq 'G' && $next eq 'C' && is\_stable($curr, $next, $temperature, $pH)) {

$prediction[$i] = '(';

} elsif ($curr eq 'G' && $next eq 'U' && is\_stable($curr, $next, $temperature, $pH)) {

$prediction[$i] = '(';

} else {

$prediction[$i] = '.';

}

}

# Print the predicted secondary structure

print “The predicted rna sequence is: \n”;

print join(“” , @prediction), “\n”;

print join("", @prediction), "\n";

# Predict a base pair if the nucleotides are complementary

#In RNA, guanine (G) is able to form a base pair with uracil (U) through a type of base pairing called wobble base pairing. This means that G can form a base pair with either C or U.

# Print the predicted secondary structure.

#The code checks for complementary base pairs and adds parentheses to the @prediction array to denote the base pair. If no complementary pair is found, a period is added to the array.

#The loop ends and the predicted secondary structure is printed using the join function.

#The last dot in the output comes from the fact that the last nucleotide in the sequence does not have a complementary nucleotide to form a base pair with. The loop iterates through the sequence and predicts a base pair for each pair of consecutive nucleotides, so the last nucleotide does not have a nucleotide following it to form a base pair with and is therefore assigned a dot.

#code is attempting to predict the secondary structure of an RNA molecule by looping through the sequence and checking for complementary base pairs. If it finds a complementary pair, it adds parentheses to the prediction array to denote the base pair. If it doesn't find a complementary pair, it adds a period to the prediction array.

OUTPUT:

RESULTS AND DISCUSSIONS:

A picture containing text

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

A picture containing graphical user interface

Description automatically generated

SCOPE:

The program is based on a simple yet a very sophisticated approach towards understanding and bring a code that is very smart to give us the predicted results of RNA which is our final goal. The approach is mainly through determining various aspects like stability is determined by the free energy of the structure, which is calculated using thermodynamic parameters for each base pair and various correction factors.

The script first checks if the input RNA sequence contains any characters other than A, C, G, and U, and exits if it does. It then reads the temperature and pH from the command line arguments or uses default values if they are not provided.

The script then defines a threshold for base pair stability and declares a subroutine for calculating the stability. The subroutine first defines some thermodynamic parameters for each base pair in the form of enthalpy and entropy values stored in hashes. It then initializes some variables to track the total enthalpy and entropy of the structure.

The subroutine then loops through the RNA sequence and sums up the enthalpy and entropy contributions of each base pair. After the loop, it calculates various correction factors for the salt concentration, temperature, and pH of the solution, as well as the stability of the structure based on these factors and the total enthalpy and entropy. Finally, the subroutine returns the stability of the structure.

Conclusion:

The idea of secondary structure prediction of RNA has been introduced previously. But it is one of the most critical aspects of genomics, As the secondary structure of RNA plays one of the major roles in protein structure. But predicting the structure while keeping in mind the necessary conditions temperature, ph. and stability that affect the structure in real time give us a huge benefit in understanding these structures. SO, using bioinformatic tools gives the analyzer a huge edge in computing multiple and long sequences within a short span of time and with ease.

*Terminologies:*

**Gibbs free energy**, also known as the Gibbs function, Gibbs energy, or free enthalpy, is a quantity that is used to measure the maximum amount of work done in a thermodynamic system when the temperature and pressure are kept constant. Gibbs free energy is denoted by the symbol 'G'.

**Enthalpy** a property of a [thermodynamic system](https://en.wikipedia.org/wiki/Thermodynamic_system), is the sum of the system's [internal energy](https://en.wikipedia.org/wiki/Internal_energy) and the product of its pressure and volume.[[1]](https://en.wikipedia.org/wiki/Enthalpy#cite_note-:0-1) It is a [state function](https://en.wikipedia.org/wiki/State_function) used in many measurements in chemical, biological, and physical systems at a constant pressure, which is conveniently provided by the large ambient atmosphere.

**Entropy**, the measure of a system's thermal energy per unit temperature that is unavailable for doing useful work. Because work is obtained from ordered molecular motion, the amount of entropy is also a measure of the molecular disorder, or randomness, of a system.

**A thermodynamic system** is defined as a quantity of matter or a region in space that is of interest. The mass or region outside the system is called the surroundings, and the surface that separates the system and the surroundings is called the boundary.

Reference:

F. Guo et al.Structure of the Tetrahymena ribozyme: base triple sandwich and metal ion at the active site

Mol Cell(2004)

R. Giegerich et al.Abstract shapes of RNA Nucleic Acids Res(2004)

J. Reeder et al.Consensus shapes: an alternative to the Sankoff algorithm for RNA consensus structure prediction Bioinformatics (2005)

J. Chen et al. Prediction of common secondary structures of RNAs: a genetic algorithm approach

Nucleic Acids Res (2000)

D.H. Mathews et al. Predictive methods using RNA sequences D.H. Mathews et al.

Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure Proc Natl Acad Sci USA (2004)

R.D. Dowell et al. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction BMC Bioinformatics (2004)

K.J. Doshi et al. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction BMC Bioinformatics (2004)

H. Isambert et al. Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme

Proc Natl Acad Sci USA (2000)

Jacobson AB, Good L, Simonetti J, Zuker M. Some simple computational methods to improve the folding of large RNAs. Nucleic Acids Res. 1984 Jan 11;12(1 Pt 1):45–52. [PMC free article] [PubMed] [Google Scholar]

van Belkum A, Abrahams JP, Pleij CW, Bosch L. Five pseudoknots are present at the 204 nucleotides long 3' noncoding region of tobacco mosaic virus RNA. Nucleic Acids Res. 1985 Nov 11;13(21):7673–7686. [PMC free article] [PubMed] [Google Scholar]

Salser W. Globin mRNA sequences: analysis of base pairing and evolutionary implications. Cold Spring Harb Symp Quant Biol. 1978;42(Pt 2):985–1002. [PubMed] [Google Scholar]

Groebe DR, Uhlenbeck OC. Characterization of RNA hairpin loop stability. Nucleic Acids Res. 1988 Dec 23;16(24):11725–11735. [PMC free article] [PubMed] [Google Scholar]

Papanicolaou C, Gouy M, Ninio J. An energy model that predicts the correct folding of both the tRNA and the 5S RNA molecules. Nucleic Acids Res. 1984 Jan 11;12(1 Pt 1):31–44. [PMC free article] [PubMed] [Google Scholar]

Sprinzl M, Hartmann T, Weber J, Blank J, Zeidler R. Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res. 1989;17 (Suppl):r1–172. [PMC free article] [PubMed] [Google Scholar]