A mathematical model for RNA 3D structures

by

Sixiang Zhang

(Under the Direction of Liming Cai)

Abstract

RNA (ribonucleic acids) tertiary (3D) structure prediction is crucial for understanding the relationship between RNA structures and their functions. RNA 3D structure prediction remains challenging in spite of advancements in recent years. In this thesis, we propose a new method for RNA 3D structure prediction with a novel mathematical model. We model an RNA 3D structure as a collection of interacting helixes with a succinct geometric characterization for every pair of consecutive nucleotides on the RNA sequence. Given a small set of parameters, such as various angles between segments, the model geometrically projects any consecutive segment of the RNA sequence into a single helix in the 3D space, enabling effective assembly of RNA 3D structure. Tests on RNA sequences from the Protein Data Bank have shown the success of our method on prediction of 3D structures involving double-helices, hairpin loops, and bulges.

Index words: [RNA, Nucleotide, Watson-Crick pair, Helix, Gaussian-like distribution, Rotation matrix]

A mathematical model for RNA 3D structures

by

Sixiang Zhang

B.S., Dalian Jiaotong University, China, 2018

A [Thesis] Submitted to the Graduate Faculty of the
University of Georgia in Partial Fulfillment of the Requirements for the Degree.

[Master] of [Science]

Athens, Georgia

2021

A mathematical model for RNA 3D structures

by

Sixiang Zhang

Major Professor: Liming Cai

Committee: Ismailcem Budak Arpinar

Giorgis Petridis

Electronic Version Approved:

Ron Walcott

Vice Provost for Graduate

Education and Dean of the

Graduate School

The University of Georgia

August 2021

Acknowledgments

First of all, I am incredibly grateful to my advisor Dr. Liming Cai for his unwavering support at the stage of my graduate studies, for his infinite patience, passionate answers to my questions, and practical suggestions. He has been helping me wholeheartedly during the research process and in writing my thesis.

Moreover, I would also like to extend my deepest gratitude to Dr. Russell Malmberg, who is always patient and extremely helpful in writing my thesis.

Besides my advisor and Professor Malmberg, I would like to extend my sincere thanks to Dr. Ismailcem Budak Arpinar and Dr. Giorgis Petridis. They serve as my committee members for their helpful advice, patience, and constructive comments.

Last but not least, Special thanks to my fellow labmates in the RNA Informatics group: Dr. Robert Robinson, Di Chang, Fereshteh Rabiei Dastjerdi, Zahra Jandaghi, Salvatore LaMarca, and David Robinson.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

RNA (Ribonucleic acid) is a complex macromolecule essential for all living lives (Wan & Chatterjee, 2018). RNA, very similar to DNA (Deoxyribonucleic acid), is made up of nucleotides, including a nucleobase, ribose sugar, and phosphate group. There are several aspects that can distinguish RNA from DNA: First, RNA contains the sugar ribose instead of sugar deoxyribose (Pliatsika, n.d.); Second, RNA has the four nitrogenous bases (A, C, G, U) ("Structure and Function of RNA", 2021). Finally, RNA sequences are usually much shorter, and are single-stranded instead of double strands in DNA. Most single-stranded RNA molecules fold back to themselves to form complicated 3D structures.

RNA structure plays a vital role in RNA functions. Understanding how an RNA constructs its structure would offer an insight into the function of the RNA (Doudna & Cate, 1997). Even with only four nucleotides as simple building blocks, many RNAs have complex tertiary shapes defining their activities and functions ("RNA: The Versatile Molecule", 2016). For example, there are various types of RNAs: mRNA (messenger RNA), rRNA (ribosomal RNA), and tRNA (transfer RNA), which have different functions (Clancy, 2014); Other RNAs are engaged in gene expression and other activities. However, fundamentals and parameters to determine RNA structures remain unclear. In particular, elucidating RNA structure by biological experiments usually is very time-consuming. Nevertheless, revealing the mystery

behind the structure folding of RNAs may be aided by correctly predict the RNA structure through computer programs. Computational prediction of RNA structures is, therefore, an essential topic in bioinformatics, which is also the focus of this thesis.

RNA structure may be categorized into three levels: the primary sequence of nucleotide bases; secondary structure that is formed by Watson-Crick base pairs like A-U, G-C, and wobble pair G-U; and tertiary structure, the 3D shape. An RNA secondary structure consists of different elements (Figure 1.1). A stem consists of two single-strand segments on which nucleotides form stacked base pairs across the strands. Unpaired single-strand segments form hairpin loops, bulges, and junctions. The most common element of RNA secondary structure is the stem-loop (hairpin), which is the focus of this thesis.

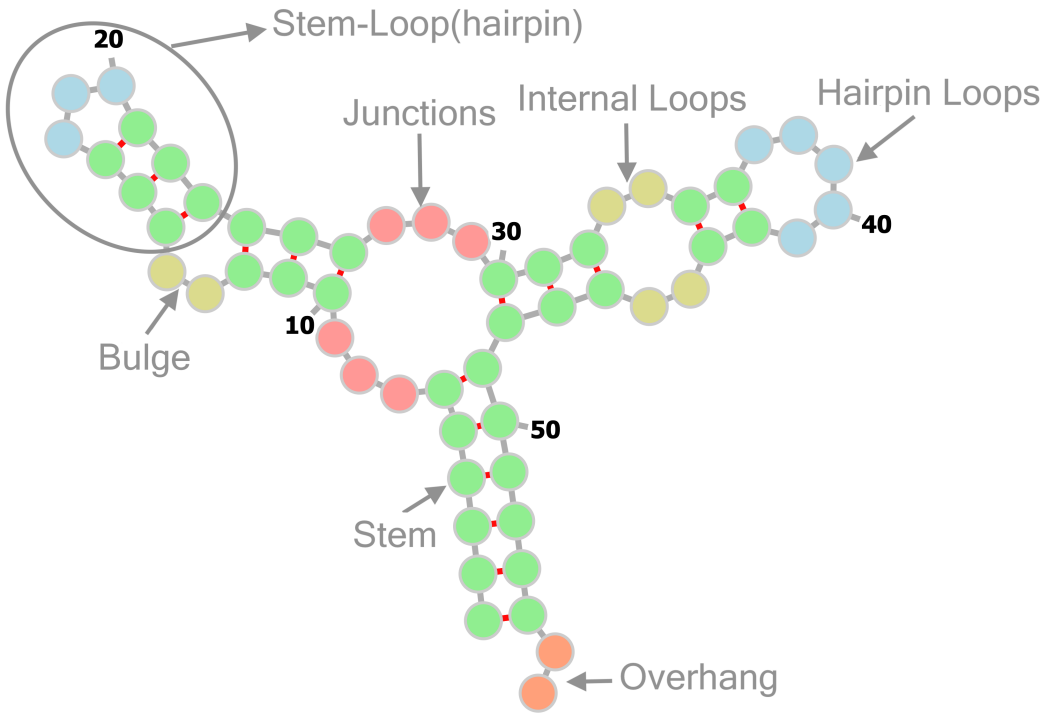


Figure 1.1: Structural elements in RNA secondary structure (created by ViennaRNA Web Services(Kerpedjiev et al., 2015)).

At a higher level, under appropriate conditions, RNA molecules may fold into a 3D structure, with its secondary structure as a scaffold ("RNA Structure, Function, and Recognition iBiology", 2014), in which the different building blocks (helices and the unpaired regions) are precisely arranged in space (Figure 1.2). Three experimental methods have been developed to elucidate for the tertiary structure: X-ray crystallography, nuclear magnetic resonance (NMR), and the phylogenetic method combined with computer modeling and experimental approaches (Westhof & Pascal, 2006).

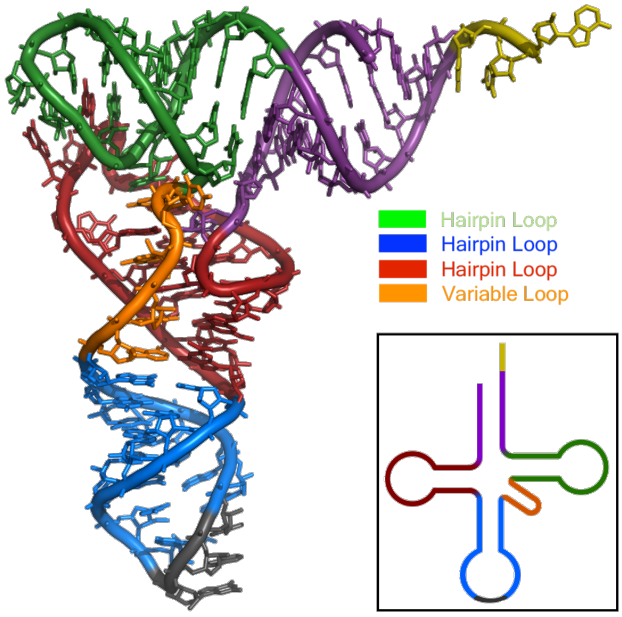


Figure 1.2: The tRNA molecule includes three hairpin loop, formed "L" shape (modified from "TRNA-phe yeast"(Yikrazuul, 2010)).

Due to the intimate relationship between RNA structures and functions, RNA 3D structure prediction has recently become a widely popular topic. Prediction of RNA three-dimensional structure can be based on either of the following approaches: the dynamics approach and the Monte Carlo approach. The former approach has two main methods: all-atom and coarse-grained. Since all-atom sampling methods are too computationally intensive and not particularly effective, coarse-graining is the primary method in the dynamic approach. The best-known method in this category is iFoldRNA, which uses a 3-bead RNA

model (Krokhotin et al., 2015). Monte Carlo is better known, as there is relatively less demand of arithmetic power, yet still a lot. The two notable methods are Fragment Assembly of RNA (FARNA) (Das & Baker, 2007) and Fragment Assembly of RNA with Full-Atom-Refinement (FARFAR) (Das et al., 2010). Both approaches have to address sampling and scoring issues.

Rarely do people cast this predicting problem as a pure mathematical problem. Thus, in this thesis, we propose a novel approach, a mathematical model, to predict the RNA tertiary structure based on a given RNA secondary structure. In our model, RNA tertiary structure comprises two types of regions: paired and unpaired regions. We believe there are two helical models for two regions, respectively arrangements of helices that interacting in 3D space yields RNA 3D structure. This thesis mainly investigate how to model helices from RNA stem loops.

Chapter 2 will introduce the background for our RNA 3D structure project, the dataset, and the visualization software. It also the goal of this thesis. Chapter 3 discusses our model in detail; we elaborate on the three helical models and choices of parameters. Chapter 4 presents how we implement the model, the test results, and performance evaluation. We conclude with a future plan in Chapter5.

Chapter 2

Background and goals

2.1 Dataset and visualization

This research used RNA sequences from Protein Data Bank (PDB) dataset (Burley et al., 2020), including atomic coordinates and other critical RNA nucleotide and structure information. We selected more than 30 RNAs from PDB, which only consist of the hairpin loop and stems segments, and chose some other RNA sequences, which contain bulges in their stem loops.

In this research, we also used PyMOL (Schrödinger,LLC., 2020) to visualize RNA molecular structures, both resolved and predicted, with atomic coordinates files. PyMOL is a stand-alone molecular visualization program widely used in bioinformatics, which permits annotations of tertiary structure with simple Python scripts.

2.1.1 Atom selections

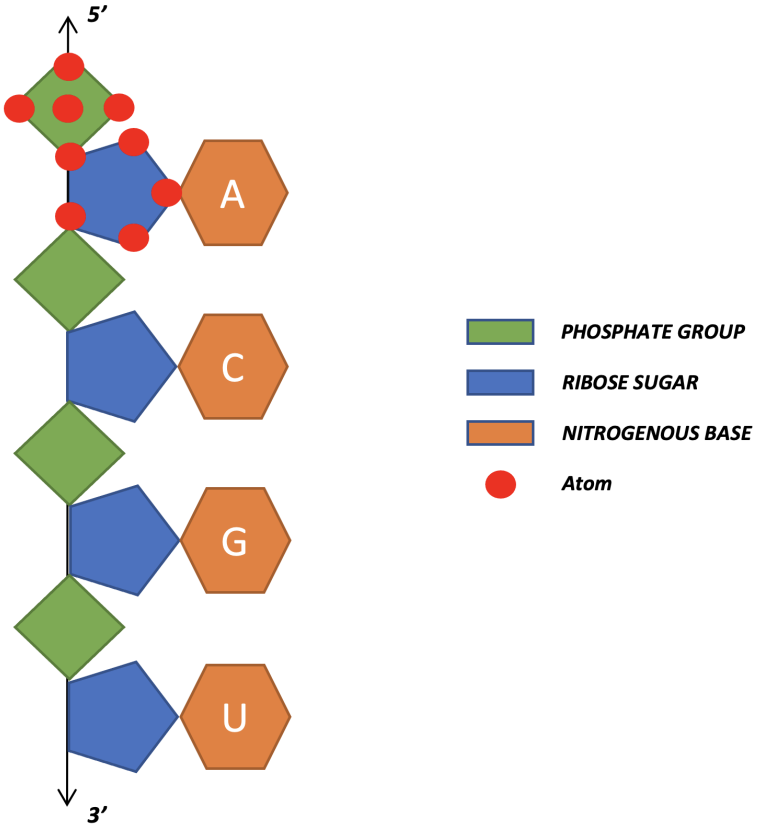
We give some backgrounds for the introduction of our RNA structure model. We begin with the atomic level of RNA; then we provide a mathematical view on the RNA 3D structure. RNAs are polymeric molecules consisting of four nucleotides, each including base (nucleobase),

phosphate groups, and ribose sugar (Figure 2.1(a)). Since our model is coarse-grain, we use a single atom coordinates to represent the nucleotide. C1' atom (Figure 2.1(c)) has been the most common choice of coordinate representation for RNA nucleotides and it work reasonably well for the stem structures (base-pairing parts); However, our tests showed it was not a good fit for the hairpin loop segment since the C1' atoms are closer to the base sidechain and far away from the sugar-phosphate backbone in nucleotides, making it difficult to measure nucleotide-nucleotide distance where sidechains may flip outward (Figure 2.2). Therefore, we have decided to choose atom O5' as the representer, which falls on the sugar-phosphate backbone (Figure 2.1(c)).

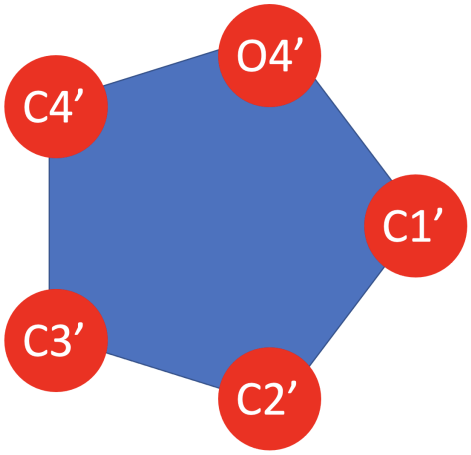
PDB contains tens of thousands of structures, protein and RNA, contributed by research groups in biological sciences; the structures were obtained through experimental methods (X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy) that determine atomic coordinates for all structured molecules. The size of the PDB is still growing since structure determination research is happening in laboratories every day around the world. Not all atomic entries will include coordinates in the molecule structure determined by these methods, however. For instance, hydrogen atoms may not be observed in X-ray crystallography methods("PDB101: Learn: Guide to Understanding PDB Data: Missing Coordinates and Biological Assemblies", n.d.). Therefore, sometimes atoms may be missing from PDB coordinate files. When this happens to O5', we use the P or O3' atom's position as a replacement for O5' because these two atoms also on the sugar-phosphate backbone as O5' atoms (Figure 2.1(c)).

2.2 Data selection

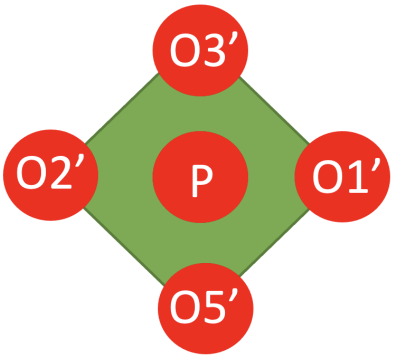
We selected 34 RNA sequences from the Protein Data Bank, 30 of which are hairpin structures and four are with a bulge. The criteria for our choice of data is that they represent cases for our modeling of hairpin loops and bulges in addition to double helices. These hairpin



(a) RNA structure



(b) The atoms arrangement of ribose sugar.



(c) The atoms arrangement of phosphate.

Figure 2.1: (a) RNA structure. Phosphate groups (green, diamonds), ribose sugar (blue, pentagons) and nitrogenous bases (orange, hexagons) , which has four types of base (A, C, G and U) form RNA structure. (b) The chemical structure of ribose sugar, including atoms C1', C2', C3', C4', O4'. (c) The chemical structure of phosphate group, including atoms O1', O2', O3', O5", P.

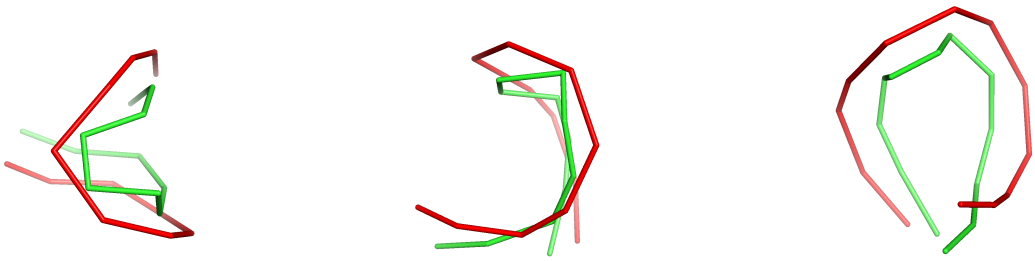


Figure 2.2: These figures are various angles of observation of 1RNG, generated by Pymol(Schrödinger,LLC., 2020). The red curve is formed with nucleotides represented by atom O5', and the green curve is by atom C1'. Moreover, the image on the far left shows that the green curve does not look like a helix.

structure's may contain overhangs. An overhang is one or more unpaired nucleotide at the end of an RNA stem-loops.

2.3 Goals of this thesis

The goal of this research is to establish a mathematical model to predict the 3D structure of RNA. In this work, due to time, we mainly focus on the hairpin structure, which accounts for the most significant proportion of RNA tertiary structure, and we extend the work to other segments. In this these, we also show the accuracy of our model by comparing prediction results with resolved models.

Chapter 3

The Model

In this thesis, we treat the RNA tertiary structure as a combination of various helices. A helix is a three-dimensional curve that rotates or circles around a central line with a constant angle, just like a spring seen in daily life. Here, we envision two different helices for RNA tertiary structures: one is a base-paired double helix, and another is an un-paired single helix. With this model, we predict the 3D structure consisting of helices with a set of parameters, for example, helix diameter, rotational angles, and height or length for each helix segment, determined by the query sequence and its secondary structure.

3.1 Diameter and backbone length

The query RNA sequence contains a stem as its secondary structure (Figure 3.2(b)). The stem is projected into a model of a double helix in the 3D space (Figure 3.2(c)). It has been known that the nucleotide-nucleotide distance of a Watson-Crick pair is nearly 10.5 Å; the backbone distance between two neighboring nucleotides is length b = 3.4 Å, when C1' atoms represent nucleotides (Westhof, 2014). Since we represent nucleotides with their O5' atoms, we need to know the double helix diameter and backbone distance in terms of O5's atoms position. For this, we computed the diameter d (the distance between two O5' atoms

in the base-paired nucleotides) and the length b (the distance between two successive O5' atoms) from more than 30 RNA sequences extracted from PDB. From the two histograms and the P-P plot, it is clear that d and b nearly match the normal distribution (Figure 3.1). Based on the principle that the sample mean $\overline{X}$ from a group of observations is an estimate of the population mean μ (Zhang & Shafer, 2014), we computed the means of d and b to be 16.5 Åand 5.8 Å, respectively (Figure 3.1(a) and Figure 3.1(c)). We have adopted these parameters in the following RNA helix structure models. Furthermore, according to the literature (Warden et al., 2017), a double helix making a complete turn around its axis contains 11 base pairs.

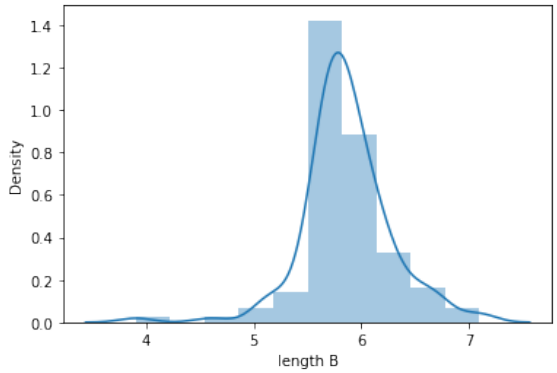
3.2 Base-pairing helix

We now describe the parameters for a double helix in detail (Figure 3.3(a)). Since it takes 11 base pairs to turn around, ∠AOB' = $\frac{360}{11}$= 32.7272° (Figure 3.3(b)). We established that the diameter d is 16.5°A, so the radius r of the cylinder = $\frac{16.5}{2}$ = 8.25 Å, and the length AB = length = 5.8 Å. Then, we calculated the length AB' = $\sqrt{2 * (1 - \cos \angle AOB')} = 4.648$ and the elevating growth of each nucleotide on the helix-axis is equal to length BB' = $\sqrt{b^2 - AB^2}$ = 3.469 Å. With these parameters, we can draw the projected double helix (Figure 3.2(c)).

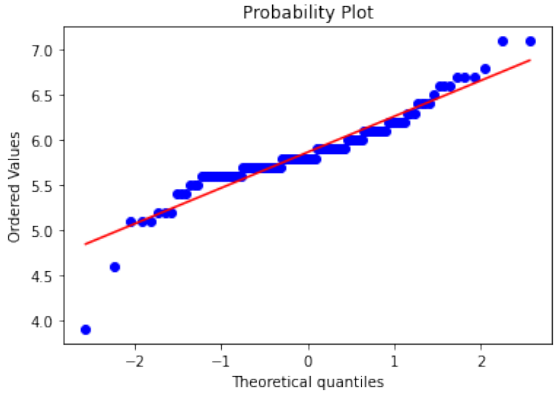
3.3 Un-paired helix

3.3.1 Hair-pin loop structure

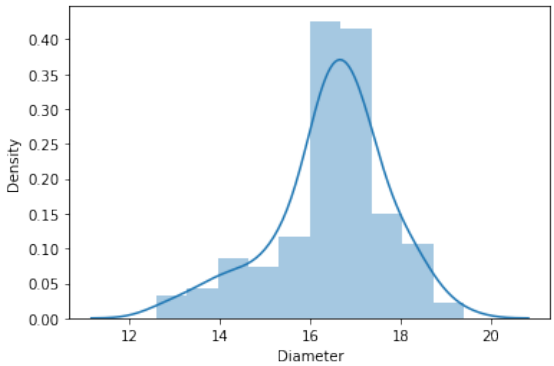
We treat the unpaired strands as a single helix. In particular, for the top loop over the bottom double-helical structure (Figure 3.4), we model this unpaired segment containing n nucleotides, including the enclosing base-pair (X_0 and X_n), as a single un-paired helix.



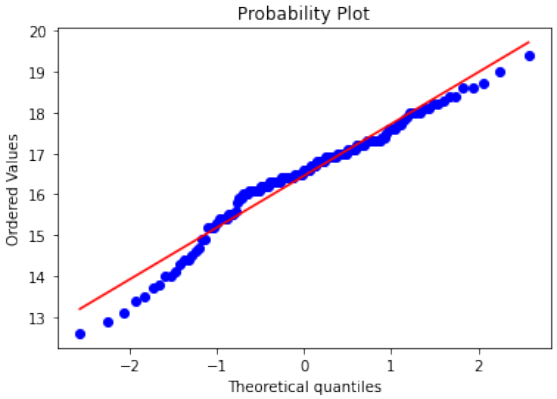
(a) Histogram for length b



(b) P-P plot for length b

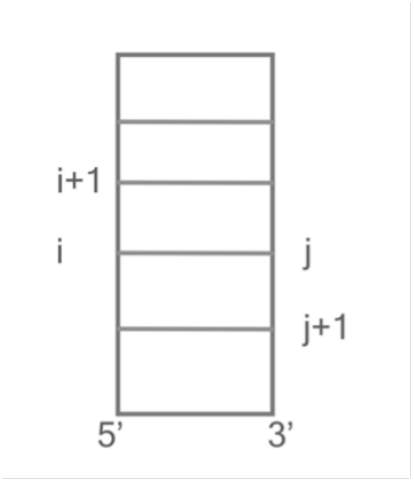


(c) Histogram for diameter d

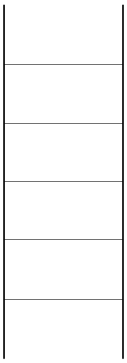


(d) P-P plot for diameter d

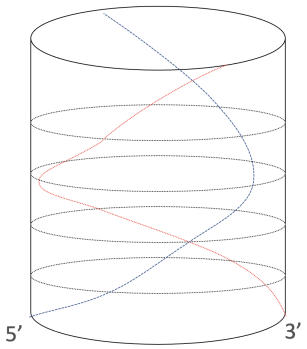
Figure 3.1: (a) The histogram of length b nearly fits the bell shape, a Gaussian distribution. (b) The length b's plot indicates that the data follow approximately a normal distribution, lying close to a diagonal line through the main body of the points. (c) The histogram of diameter d nearly also fits the bell shape, is also nearly a normal distribution. (d) The diameter d's plot indicates that the data also follow approximately a normal distribution, lying close to a diagonal line through the main body of the points.



(a) The arrangement of RNA's stem segment

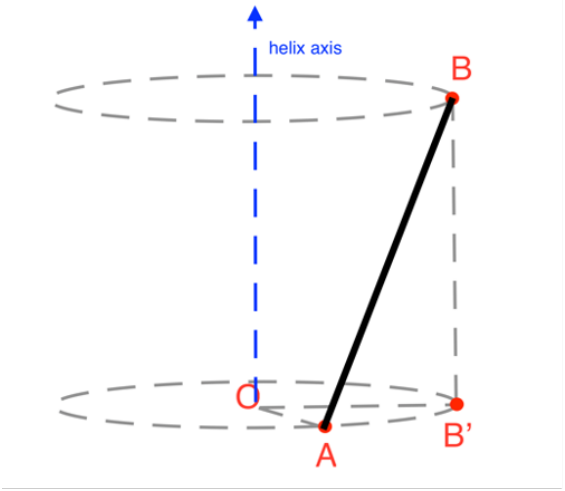


(b) The secondary structure of RNA stem

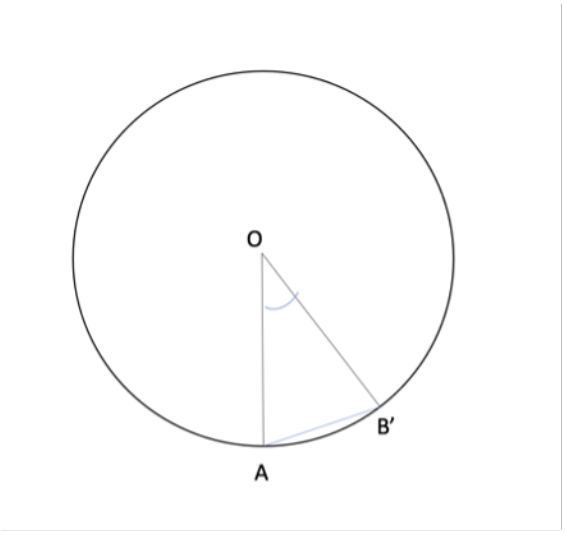


(c) Right(blue) and left(red) helix attached on the cylinder.

Figure 3.2: (a) RNA molecule's growth is always from 5' to 3', which nucleotides always added to the 3' end (Griffiths et al., 2019). (b) Stems' secondary structure only contains the base-paired information. (c) Our model of stem has two helical curves attached to a cylinder.



(a) Stem's helix is formed by a cylinder (Nucleotide A and Nucleotide B are two successive nucleotides.)



(b) Nucleotide B projected B' on the bottom of cylinder.

Figure 3.3: (a) Relative positions of nucleotide B and nucleotide A on the cylinder. Nucleotide B is projected into the same plane as nucleotide A and resulting in B'. (b) The rotation angle by projecting nucleotide B on plane AOB'.

Moreover, this single helix has an axis perpendicular to the bottom double helix axis, and the former has $n-1$ nucleotides, connecting the two helices. As a result, $\angle X_0OX_n = \frac{360}{n-1}$, in the case $n = 6$, $\angle X_0OX_n = 72°$ (Figure 3.5(a)).

As the helix model shown in the (Figure 3.4), the distance between the base-paired nucleotides, i.e., the diameter d, becomes the overall height of the apex loop structure. Moreover, the height for each nucleotide from one to its neighbor is$\frac{d}{n-1}$ Å, in the case $n = 6$, $X_1X_1' = 3.3$ Å(Figure 3.5(b)). Furthermore, the direct distance X_0X_1 between two successive nucleotides still is b. Therefore, the height of the cylinder can be calculated as follows: $X_0X_1' = \sqrt{b^2 - X_1X_1'^2}$, where $b = 5.8$ Å. In the end, with these parameters, we place the projected top loop on the top of the projected double helix top-loop structure.

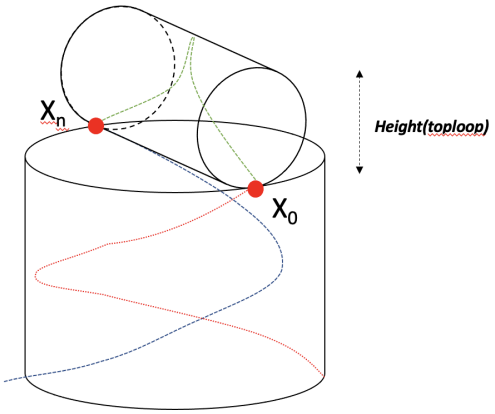
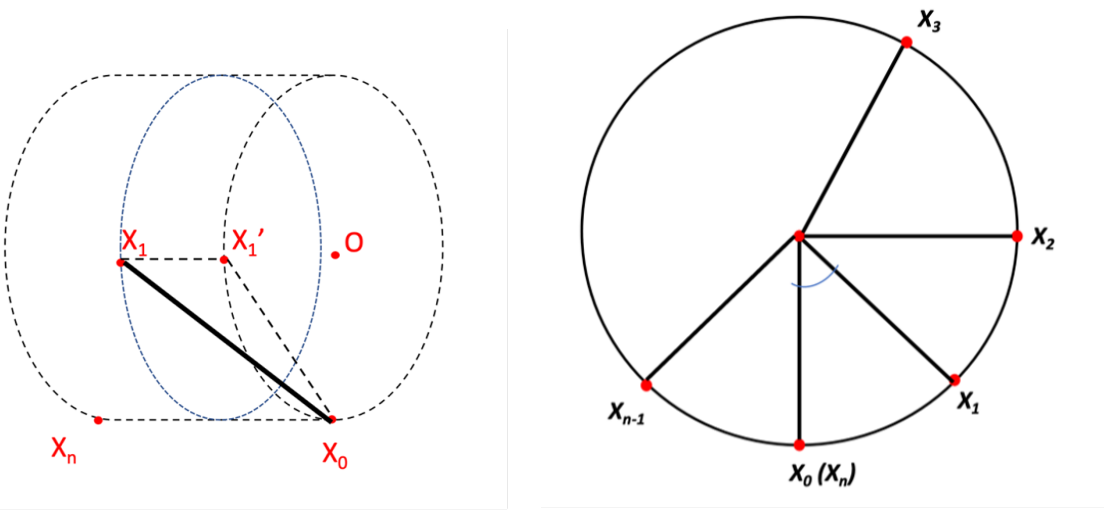


Figure 3.4: This image shows the hairpin structure of RNA, including a stem and a hairpin loop.



(a) A horizontal cylinder forms the structure of the hairpin loop.

(b) All nucleotides are projected on the side of the cylinder.

Figure 3.5: (a)Relative positions of nucleotide X_1 with respect to, the position of nucleotide $X0$ on the cylinder. Nucleotide X_1 is projected into the same plane as nucleotide X0, the projected point being X_1'. (b) The hairpin loops nucleotides are projected on plane X_0OX_1'. This figure also shows the rotation angle for each nucleotide from the top loop segment.

3.3.2 Top-loop position adjustment

However, all stem-loop structures of RNAs viewed with tools (such as Pymol and UCSF Chimera (Pettersen et al., 2004)) have shown that, the top-loop is often not on right top of the bottom double helix. So, the next step is to adjust the hairpin segment to a correct position.

The top-loop structure can be more specific based on from three rotated angles:

1) Angle θ: around the X-axis;

2) Angle γ: around the Y-axis;

3) Angle δ: around the Z-axis;

Furthermore, the same distance d between the two base-paired nucleotides is assumed for the enclosing pairs. Thus, nucleotide B should on the surface of the cylinder, and angles θ and δ should bear some relationship (Figure 3.6), for which we believe that δ varies directly with θ, and vice versa.

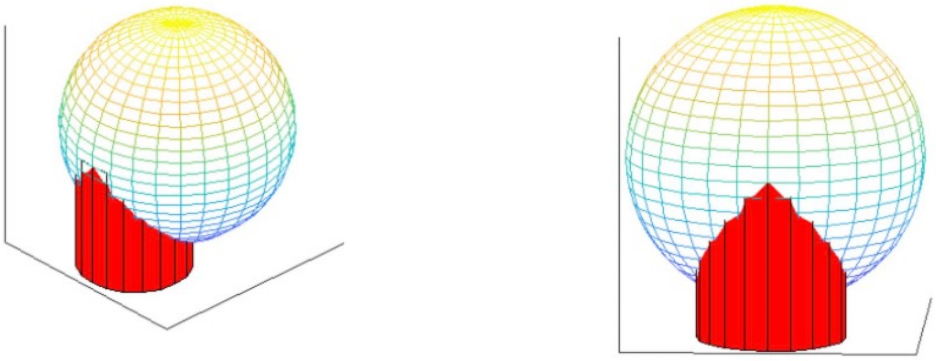


Figure 3.6: These two figures show the relationship between the two angles (created by MATLAB (MATLAB, 2021)). The sphere is centered at X_0, and the radius is $|X_0X_n|$. This sphere simulates all possible positions of X_n. Furthermore, Xn must fall on the cylinder again because Xn is at the same time the nucleotide of the bottom double helix.

We can calculate by following numbers (Figure 3.7) :

$$\begin{cases} \text{Length } AB = d, \text{ and angle } \angle BAB' = \theta; \\ \angle AB'A' = \angle BAB'; \\ \text{Thus, length } AA' = d * \sin(\angle AB'A'); \\ \text{Length } A'B' = d * \cos(\angle AB'A'); \\ \text{So, length } A'B'' = A'B'; \\ \text{Then, we know that A'B" also equals } d * \cos(\angle B'A'B''), \text{ where} \angle B'A'B'' = \delta; \\ \text{As a result, } |\theta| = |\delta| \ . \end{cases}$$

Moreover, due to base-pairing, the angle around the Z-axis should be slight smaller than 0 ($\delta < 0$). Furthermore, there are two directions for rotating around. In this thesis, we assume $\theta > 0$ (i.e., $\theta = -\delta$).

To compute γ and θ, we tested more than 30 RNA sequences to gain some insights into these two parameters.

The two angles θ takes value from 0° to 90°, γ ranges from -180° to 180°. We substituted the angles into our coded program and then compared them with the actual RNA molecules structure. We were left with the result that is closest to the accurate data. We need to mention interesting thing: the results of the two variables are close to a normal distribution (Figure 3.8). Once again, we take the mean of these two variables; setting $\theta = 21.0^\circ$; $\gamma = -12.3^\circ$.

3.4 Bulge

A bulge is a small unpaired segment within one strand of a double helix; We propose to model it with a similar approach modified from our hairpin loop model. The size of a bulge, the

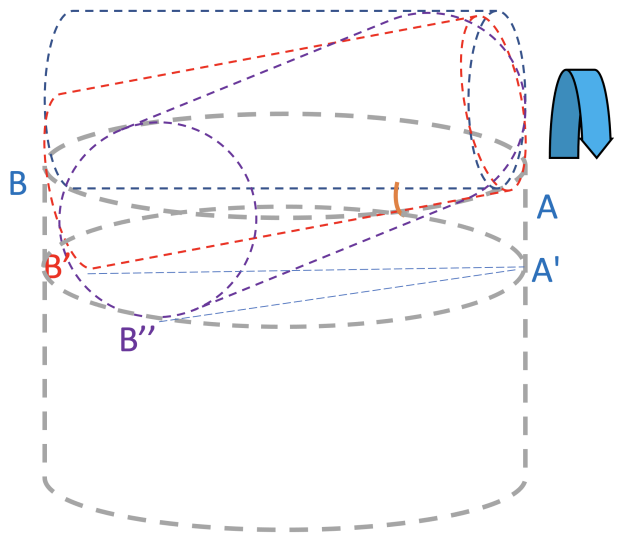


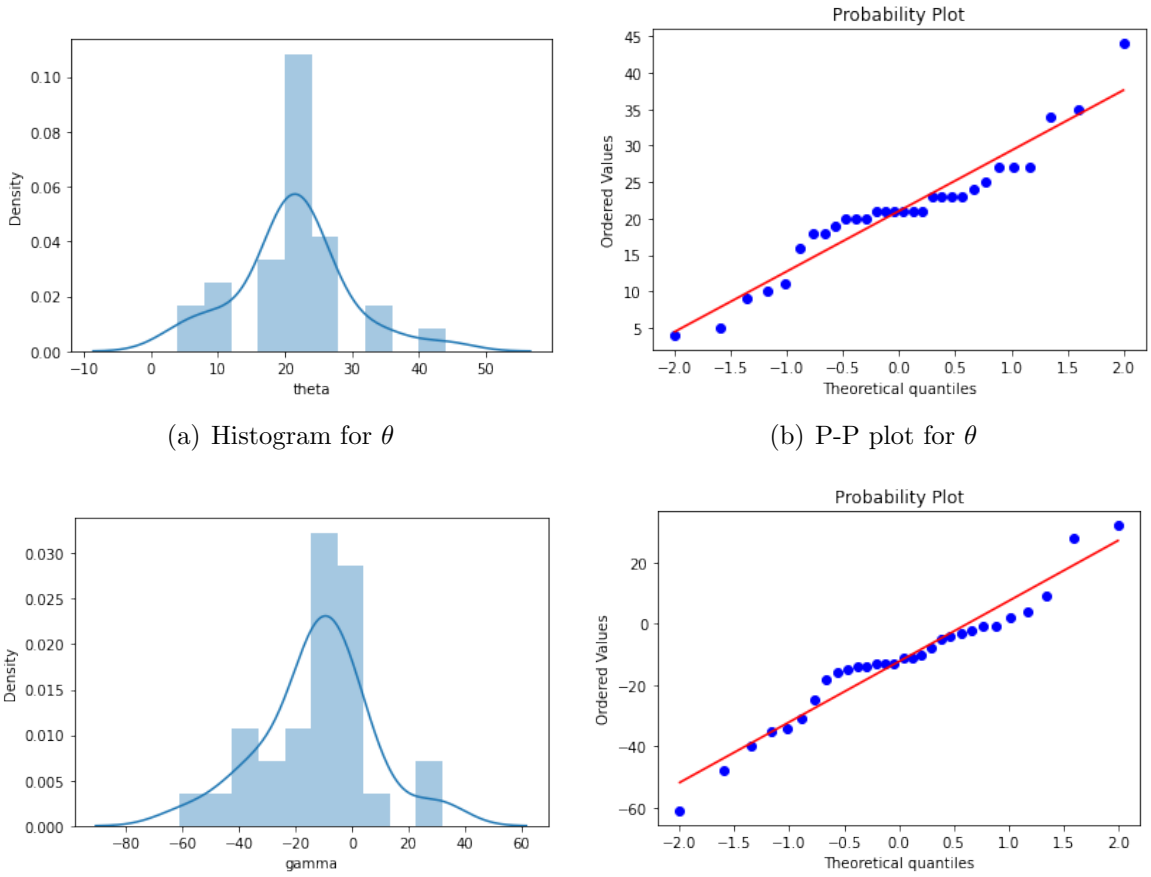
Figure 3.7: The top cylinder rotates at three angles around nucleotide A for its position adjustment. The final position of nucleotide B falls on nucleotide B".

number of unpaired nucleotides, may vary from a single to several nucleotides. Since bulges form intricate structures located within double helices, they may play important structural and functional roles (Hermann & Patel, 2000).

First, we consider a bugle to be modeled as a single helix including k unpaired nucleotides, with two enclosing nucleotides, the same as the top-loop structure (Figure 3.9(a)).

The rotated angle for the bulge segment is $\angle AOX_1 = \frac{360}{k-1}$, in most cases $k = 3$, and $\angle AOX_1 = 180°$ (Figure 3.9(b)).

For the cylinder enclosing the single helix as the bulge model, we need to calculate its height and radius. Our hypothesis is that the bulge is inserted in between the original two neighboring nucleotides. The transformation process is that the bottom nucleotide A remains unchanged, and the top nucleotide B rotates around its base-paired partner nucleotide C. As a result, the distance between the two nucleotides A and B, after the rotation, becomes longer than from their previous distance b (the distance between two successive O5' atoms),



(a) Histogram for θ

(b) P-P plot for θ

(c) Histogram for γ

(d) P-P plot for γ

Figure 3.8: (a) The histogram of θ nearly fits the bell shape, likely a Gaussian distribution. (b) θ's plot indicates that the data follow approximately a normal distribution, lying close to a diagonal line through the main body of the points. (c) The histogram of γ nearly also fits the bell shape, nearly normal distribution. (d) γ's plot indicates that the data also follow approximately a normal distribution, lying close to a diagonal line through the main body of the points. We only have limit time to test limited number of RNAs the relationship between θ and γ will be fully understood in the future work.

which is the height of the bulge column. To be more specific, let us set the three angles for the rotation. Similar to the previous notations (Figure 3.10(a)):

1) Angle θ: around the X-axis;

2) Angle γ: around the Y-axis;

3) Angle δ: around the Z-axis;

The angle γ should equal 0, which implies it does not rotate around the Y-axis. The reason is that if it rotates around the BC-axis (i.e. the Y-axis) , it will not affect the distance between nucleotide B and nucleotide A. Thus, we exclude this variable. Angle δ must be > 0 because all the distance between two enclosing nucleotides in known RNA bulges, are all greater than 5.8 (length b). Therefore, we increase the distance by rotating the Z-axis counterclockwise. Also, for the same reason for angle θ, we conclude $\theta > 0$.

And the height of bulge cylinder can be calculated as follow (Figure 3.10):

$\angle B'CB'' = \delta$, which is the angle around Z-axis;

$\angle ACB = \frac{\angle AOB}{2}$, where $\angle AOB = \frac{360}{11} = 32.7272°$;

$B'C = d$, the diameter we calculate d previously;

Thus, we obtain the length AC and AB':

$AC = d * \cos(\angle ACB)$, and $AB' = d * \sin(\angle ACB)$;

$BB' = h = \sqrt{AB^2 - AB'^2}$, where $AB = b$;

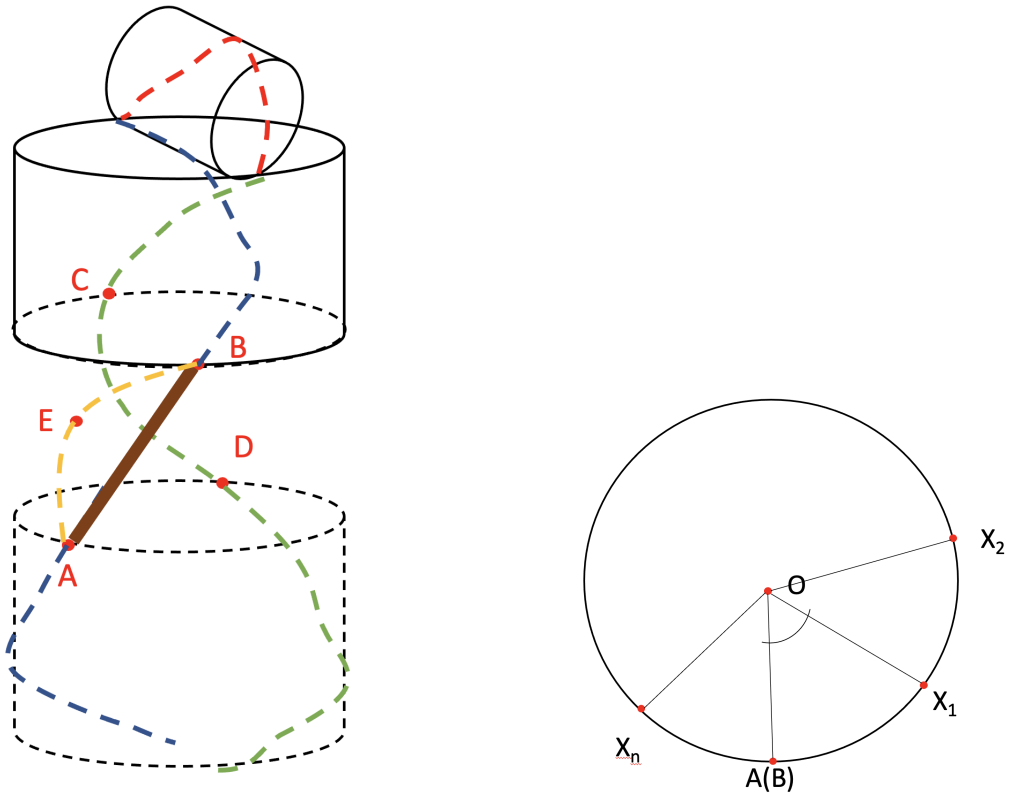
$BB''' = h + d * \sin\theta$, the vertical distance from nucleotide B to plane ADB'.

$CB''' = d * \cos\theta$

So, $AB''' = \sqrt{AC^2 + CB'''^2 - 2 * AC * CB''' * \cos(\angle ACB' + \delta)}$

The new height for bulge cylinder is $Hbulge = \sqrt{BB'''^2 + AB'''^2}$

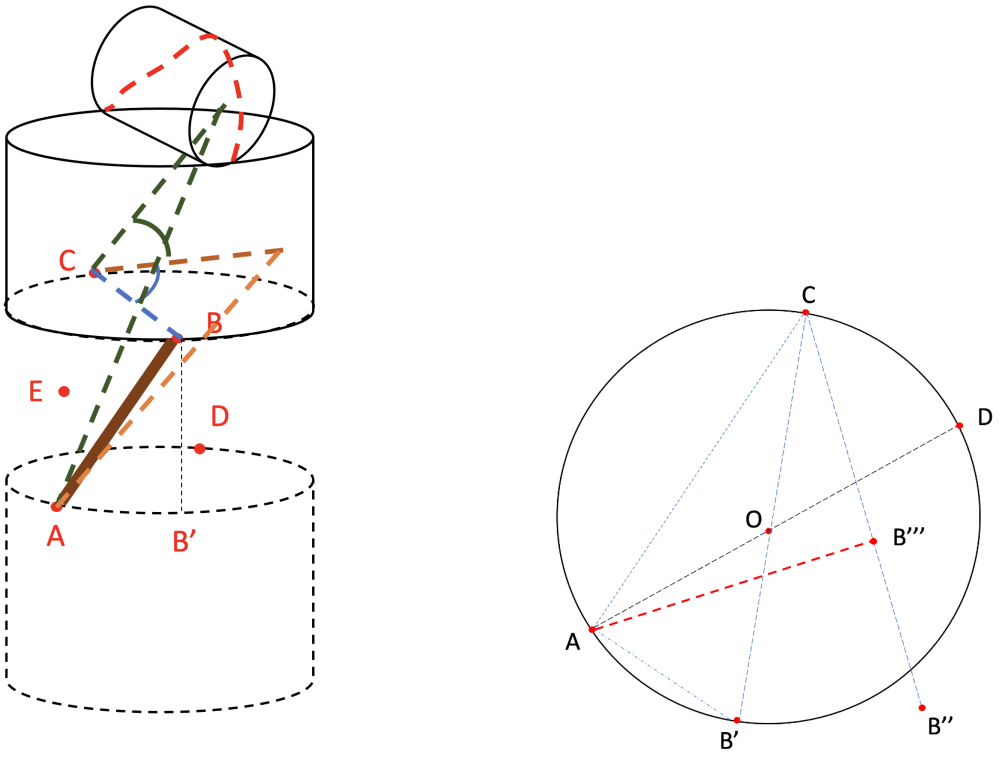
Now we get the height of the bulging cylinder concerning the two rotation angles, which allow us to perform relevant tests on the bulge modeling (see Chapter 4).



(a) This image is the structure of a hairpin with a bulging segment.

(b) This image is all nucleotide of bulge projected on the side of the cylinder.

Figure 3.9: (a) A hairpin structure with a bugle part, nucleotides A and B are the two ends of the bulge, and nucleotide E is the one on the bugle. Nucleotide C and nucleotide D are the base-pairing partners to A and B, respectively. The yellow curve is the bulge, the red is the hairpin loop, and the blue and green correspond to the double helix curves of the stems. (b) The bulge's nucleotides start from the surface of the bottom stem's cylinder and then return to the surface, precisely taking a turn. This projection shows the rotation angle of each nucleotide.



(a) Three rotation angles

(b) The projection of nucleotide b on the plane AB'D

Figure 3.10: (a) Three separate rotations of nucleotide B around nucleotide C. The distance between nucleotide A and nucleotide B changes as B rotates. (b)Nucleotide B with rotation, the projection on the plane changes from B' to B" and then to B"'.

Chapter 4

Implementation and performance

4.1 IMPLEMENTATION

The proposed structure model and prediction algorithm have been implemented in Python (Van Rossum & Drake, 2009). Python has several applicable packages to this project , including numerical computation package (Numpy)(Harris et al., 2020), a visualization library in 2D plotting (Matplotlib) (Hunter, 2007), and the widely used Bio-Python in bioinformatics package (Cock et al., 2009). With these packages, the development time was significantly reduced. This is the main reason we chose programming language Python. Python is also essential to the visualization software, PyMol, which we have used extensively in this research.

This chapter will give detailed accounts on a few important aspects of the implementation. First, we show how to use the package Bio-python to extract the atom O5' locations for each nucleotide from any given PDB file, making it possible to evaluate our model's accuracy. Bio-python is a project that can easily read and write the different sequence file formats and can gain access to online databases and provide structure evaluation methods, such as RMSD (root mean squared deviation).

Then, we will explain how to predict each helical segment with a given set of parameters. We will utilize rotational and translational matrices to arrange different modeled helices to generate a predicted RNA structure in 3D.

Finally, we will introduce the XYZ file format (O'Boyle et al., 2011), which can store a given number of atoms' coordinates of a 3D structure. With such files, we can subsequently display involved atoms with any visualization software (in this work, Pymol). With two such files, we can compute the RMSD value for two given structures.

4.1.1 EXTRACTION FROM PDB

The data structure forms provided by PDB files are SMCRA hierarchic data structures (Hamelryck & Manderick, 2003), which have a unique order (structure/model/chain/residue/atom). Usually, crystal structures from the PDB database only have one model. Models are numbered from 0; Chains are identified with capital letters like A, B, etc.. The following code extracts O5' coordinates from the first residue in chain A of the first model of the structure with PDB id 1Q75.

```
structure = parser.get_structure(1Q75) #choose the PDB 1Q75
model = structure[0]#choose the model0
chain_A = model['A']#choose chain A
res=chain_A[1] #choose residue 1
atom = res["O5'"] #choose atom O5'
```

4.1.2 TRANSLATION AND ROTATION MATRICES

We now describe how to place the model in all desired 3D positions. For this, we needed to rotation and translation operations (Evans, 2001). In particular, we used the following the translation and rotation matrices (Arfken, 1985) in the implementation.

Translation matrices

The following matrix translates coordinates (x, y, z) to (x_1, y_1, z_1):

$$T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ dx & dy & dz & 1 \end{pmatrix} \tag{4.1}$$

And (x, y, z) and (x_1, y_1, z_1) have relationship: $(x_1, y_1, z_1, 1) = (x, y, z, 1) \cdot T$.

The translation matrix is relatively simple. However, rotation around a point in the space, needs to move that point to the origin before applying the following rotation matrices; after the rotation, move the center of rotation back to the original location.

Rotation matrices

Three angles are needed to describe a rotational process. In this research, we use Eulerian angles, three angles around the three different axes, in the following order.

1) Matrix for rotation around Z-axis

This is to rotate coordinates (x, y, z) around the Z-axis counterclockwisely by angle θ to coordinates (x_1, y_1, z_1):

$$R_z = \begin{pmatrix} \cos\theta & -\sin\theta & 0 & 0 \\ \sin\theta & \cos\theta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{4.2}$$

The rotation is expressed as $(x_1, y_1, z_1, 1) = R_z \cdot (x, y, z, 1)^T$

2) Matrix for rotation around X-axis

This is to rotate coordinates (x, y, z) around the X-axis counterclockwisely by angle θ to (x_1, y_1, z_1):

$$R_x = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta & 0 \\ 0 & \sin\theta & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{4.3}$$

The rotation is expressed as $(x_1, y_1, z_1, 1) = R_x \cdot (x, y, z, 1)^T$

3) Matrix for rotation around Y-axis

This is to rotate coordinates (x, y, z) around the Y-axis counterclockwisely by angle θ to (x_1, y_1, z_1):

$$Ry = \begin{pmatrix} \cos\theta & 0 & \sin\theta & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\theta & 0 & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{4.4}$$

The rotation can be expressed as $(x_1, y_1, z_1, 1) = R_y \cdot (x, y, z, 1)^T$

According to Euler's rotation theorem, any rotation can be described by only three angles around three axes (Palais et al., 2009) . As a result, these matrices can apply to any helix structure plotting, just in a different order.

4.1.3 Output format

The input to the prediction task is a query RNA sequence along with its known or predicted secondary structure (by a third-party tool). The output of the prediction is an XYZ file containing a set of atomic 3D coordinates for all the nucleotides in the query sequence. The XYZ file format is a chemical file format supported by many programs. Although there is a formal standard, several variations can be found on the internet. All XYZ files typically have several segments as follows (illustrated with example 1Q75.xyz). The first line contains

the number of atoms in this file; in our cases, this is also the number of nucleotides of a given RNA sequence; the second line usually can be a title, or output filenames, or any preferred meaningful information, here, we fill this line with a PDB ID. The subsequent lines are atoms and corresponding Cartesian coordinates (Figure 4.1).

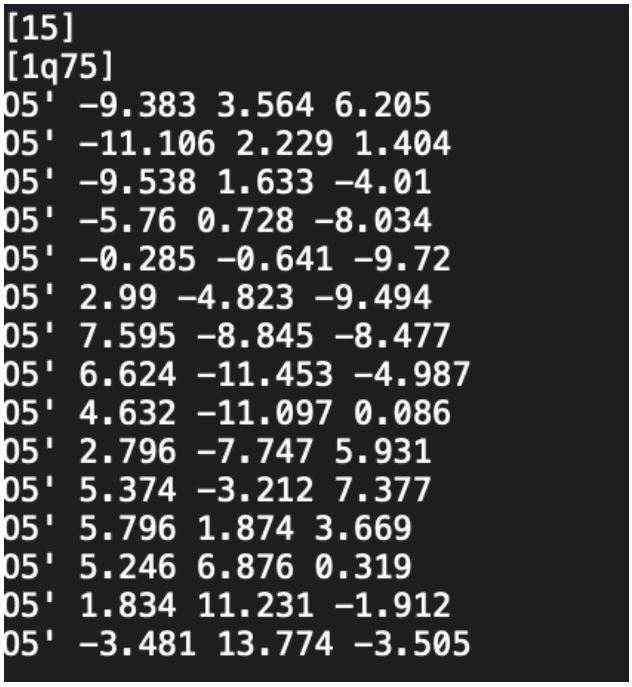


Figure 4.1: Coordinates of atom O5' of RNA with PDB id 1Q75 in XYZ file format.

4.2 PERFORMANCE

In structural bioinformatics, RMSD (Root Mean Square Deviation) is the most common practice to compare two structures (Carugo, 2003), often one is a bioinformatic prediction and the other is an experimentally determined structure. RMSD is often measured in Angstroms and calculated by the following formula (Nguyen et al., 2016)(4.5).

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\|x_i^A - x_i^B\|^2\right)} \tag{4.5}$$

where N is the number of residues, and x_i^A and x_i^B are the ith atoms' coordinates from A and B structures, respectively.

In general, if two structures are identical, the RMSD value should be 0. Since often the two structures to be compared may belong to two different coordinate systems, direct computing their RMSD may result in a very high RMSD value. So RMSD computation is based on transformation between the two coordinate systems to get the actual minimal RMSD value (Kromann, 2021). Biopython's build-in RMSD method and Pymol's alignment function have already applied such an algorithm to get the best RMSD value. We have used RMSD to measure the accuracy of our model predictions. And Pymol also makes it possible to visualize the two structures and their superimposition.

4.2.1 Performance on Stem-loops without bulge

The data in the line graph (Figure 4.2), which corresponds to the table (Table 4.1), shows: the blue line represents the lowest RMSD value obtained by adjusting the top-loop segments to the angle appropriate for each RNA sequence; the orange line represents the RMSD value obtained by averaging over the choices of the two angles since they are nearly Gaussian distribution. Moreover, from this graph, we see the two lines overlap, proving our hypothesis again.

Table 4.1: The 30 RNA Molecules for 3D Structure Prediction in This Thesis

RNA_num	PDB_id	Length(nt)	RMSD($\mathring{A}$)	RMSD_mean($\mathring{A}$)
1	1Q75	15	1.86	1.89
2	2GVO	18	3.09	4.33
3	1ATO	19	2.23	2.54
4	1UUU	19	3.09	3.15
5	2KOC	14	2.07	2.28
6	1RNG	12	1.9	2.07
7	2RLU	19	2.45	2.59
8	6PK9	20	3.05	3.1
9	2Y95	14	2.09	2.43
10	1ZIG	12	1.7	1.72
11	1BZ3	17	2.22	2.36
12	1HS3	13	2.16	2.16
13	1HS1	13	2.21	2.21
14	1HS8	13	2.08	2.08
15	1HS4	13	2.2	2.21
16	1HS2	13	2.18	2.18
17	1LK1	14	2.28	2.52
18	1WKS	17	2.71	2.81
19	1MT4	24	2.68	2.78
20	1E4P	24	2.73	3.12
21	2MXJ	11	2.17	2.23
22	1BN0	20	1.54	2.26
23	1AFX	12	1.46	1.62
24	3PHP	23	2.59	3.41
25	1F9L	22	2.35	2.48
26	2M5U	22	1.64	1.94
27	1JTJ	23	4.44	4.96
28	1ZIF	12	1.83	1.97
29	1ZIH	12	1.12	1.36
30	1K5I	23	2.09	2.34

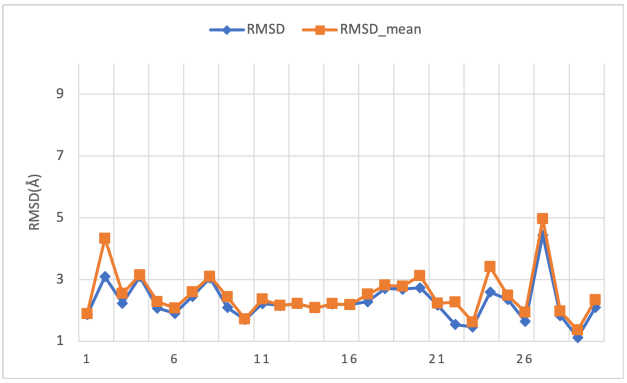
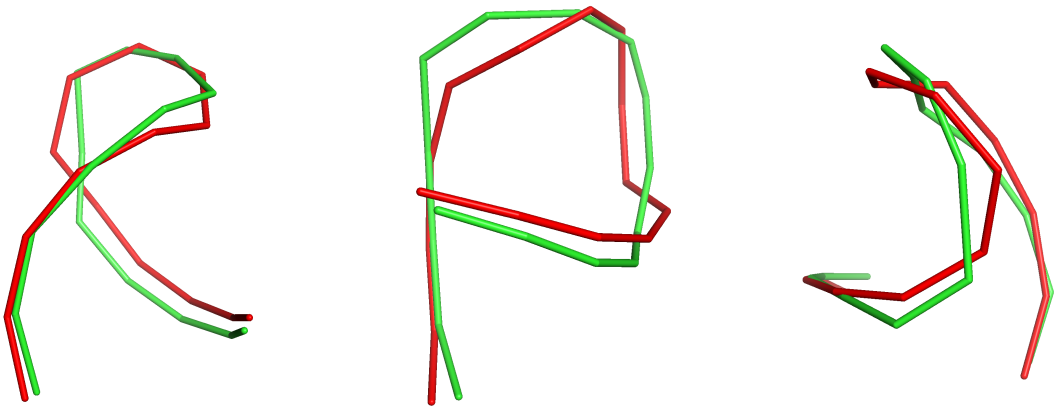


Figure 4.2: This figure is the line graph of RMSD mean (orange) RMSD min (blue).

For example, for RNA of the PDB id(1Q75) Pymol allows us to examine further the closeness of our predicted model to the actual model. We can see from the picture (Figure 4.3) that our predicted model(red) basically overlaps with actual model (green), whether viewed from the side or from the top.

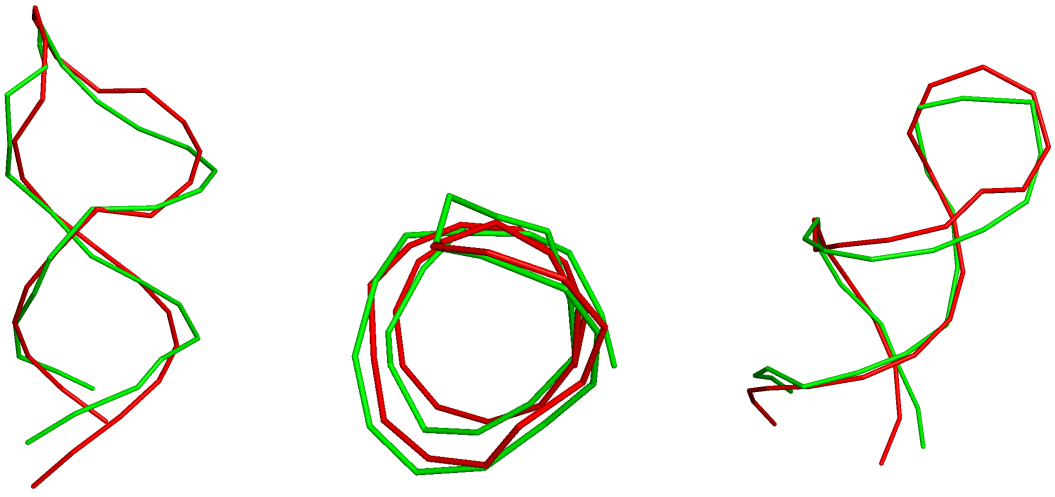


(a) (b) (c)

Figure 4.3: Views from different angles of 1Q75, generated by Pymol(Schrödinger,LLC., 2020): Superimposition between our predicted model(red) and the actual RNA structure(green).

4.2.2 Performance on Stem-loops with bulge

On 3D structure prediction of stem-loops with a bulge; due to time constraint, we only tested four RNA sequences(1NBR, 1BVJ, 1TXS,1MKF) (Table 4.2). Nevertheless, for all the four, the RMSDs show decent results of our model. We use RNA with the PDB id 1NBR as an example to illustrate. In 1NBR, the RMSD value between our prediction and the actual model is 2.46 $\mathring{A}$; We attribute this low RMSD value to that the two models of the bulging part is a perfect fit, and that the hairpin models are also very close (Figure 4.4). All these show that choices of parameters are correct in some way. However, due to time constraints, we have yet to test other more RNA sequences. Yet from these four test samples, we have observed that the rotation angle of the X-axis is equal to 0, so we can assume that the bulging part has only two parameters to determine, rotation around the Z-axis and Y-axis. We will conduct further tests on this hypothesis, including more bulge structure and inter-loop structure(symmetrical bulges), in the future.



(a) (b) (c)

Figure 4.4: These figures are various angles of viewing of 1NBR, generated by Pymol(Schrödinger,LLC., 2020). The red curve is our predicted model, and The green curve is the actual RNA structure.

Table 4.2: 3D Structure Prediction of 4 RNA Molecules with bulges

Xangle(°)	Zangle(°)	Rotation_itself(°)	PDB_ID	RMSD($\mathring{A}$)
0	18	19	1NBR	2.46
0	12	10	1BVJ	2.88
0	15	-15	1TXS	3.35
0	21	50	1MKF	3.28

Chapter 5

Conclusion

In this thesis, we present a novel mathematical method to model the 3D structure of RNAs and for structure prediction. First, we established atom O5' as representative for nucleotides rather than other atoms. Then we assumed that the RNA structure is composed of multiple helices, and our method projects helices onto various yet relevant cylinders, assuming that stacking and rotating different cylinders can model the RNA 3D structure. Moreover, we defined two classes of helices, paired-helix, and unpaired-helix. We then conducted tests on the distances between the O5' atoms of base-paired nucleotides and the length between two successive O5' atoms; we have discovered that these quantities are practically in the normally distribution. Accordingly, we picked the mean of each of these quantifies as the desired parameter. We tested more than 30 RNA sequences that have stem-loops structures selected from the PDB, and the results have shown the feasibility of our model. The main reason for selecting a stem-loop structure for testing is that it is the building block of RNA structures.

Due to time constraints, we could not perform more tests for the bulging part. Our future work is to perform more RNA sequences with bulge and then conceptualize the model for other segments of RNA 3D structure.

To the best of our knowledge, this is the first such effort in pure mathematical modeling of RNA 3D structure. While this thesis only covers stem-loops modeling, the ability of the method to use single helices to model well the irregular shapes like top-loop and bulge suggests some mathematics fundamentals underlying bio-molecular structures, a topic remaining to be extensively explored.

Bibliography

Arfken, G. (1985). *Mathematical methods for physicists* (Third). Academic Press, Inc. First edition published in 1967.

Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V., Christie, C. H., Dalenberg, K., Di Costanzo, L., Duarte, J. M., Dutta, S., Feng, Z., Ganesan, S., Goodsell, D. S., Ghosh, S., Green, R. K., Guranović, V., Guzenko, D., Hudson, B. P., … Zhuravleva, M. (2020). RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Research*, *49*(D1), D437–D451. https://doi.org/10.1093/nar/gkaa1038

Carugo, O. (2003). How root-mean-square distance (r.m.s.d.) values depend on the resolution of protein structures that are compared. *Journal of Applied Crystallography*, *36*(1), 125–128. https://doi.org/10.1107/S0021889802020502

Clancy, S. (2014). Chemical rna structure — learn science at scitable. *Nature.com*. https://www.nature.com/scitable/topicpage/chemical-structure-of-rna-348/

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Das, R., & Baker, D. (2007). Automated de novo prediction of native-like rna tertiary structures. *Proceedings of the National Academy of Sciences*, *104*, 14664–14669. Retrieved July 9, 2021, from https://www.pnas.org/content/104/37/14664.short

Das, R., Karanicolas, J., & Baker, D. (2010). Atomic accuracy in predicting and designing noncanonical rna structure. *Nature Methods*, *7*, 291–294. https://doi.org/10.1038/nmeth.1433

Doudna, J. A., & Cate, J. H. (1997). Rna structure: Crystal clear? *Current Opinion in Structural Biology*, *7*(3), 310–316. https://doi.org/https://doi.org/10.1016/S0959-440X(97)80045-0

Evans, P. (2001). Rotations and rotation matrices. *Acta crystallographica. Section D, Biological crystallography*, *57*, 1355–9. https://doi.org/10.1107/S0907444901012410

Griffiths, A. J., Miller, J. H., Suzuki, D. T., Lewontin, R. C., & Gelbart, W. M. (2019). Transcription and rna polymerase. *Nih.gov*. https://www.ncbi.nlm.nih.gov/books/NBK22085/

Hamelryck, T., & Manderick, B. (2003). PDB file parser and structure class implemented in Python. *Bioinformatics*, *19*(17), 2308–2310. https://doi.org/10.1093/bioinformatics/btg299

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hermann, T., & Patel, D. J. (2000). Rna bulges as architectural and recognition motifs. *Structure*, *8*(3), R47–R54. https://doi.org/https://doi.org/10.1016/S0969-2126(00)00110-6

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Kerpedjiev, P., Hammer, S., & Hofacker, I. L. (2015). Forna (force-directed rna): Simple and effective online rna secondary structure diagrams. *Bioinformatics*, *31*, 3377–3379. https://doi.org/10.1093/bioinformatics/btv372

Krokhotin, A., Houlihan, K., & Dokholyan, N. V. (2015). Ifoldrna v2: Folding rna with constraints. *Bioinformatics*, *31*, 2891–2893. https://doi.org/10.1093/bioinformatics/btv221

Kromann, J. C. (2021). Charnley/rmsd. *GitHub*. Retrieved July 7, 2021, from https://github.com/charnley/rmsd

MATLAB. (2021). *9.10.0 (r2021a)*. The MathWorks Inc.

Nguyen, M. N., Sim, A. Y. L., Wan, Y., Madhusudhan, M. S., & Verma, C. (2016). Topology independent comparison of RNA 3D structures using the CLICK algorithm. *Nucleic Acids Research*, *45*(1), e5–e5. https://doi.org/10.1093/nar/gkw819

O'Boyle, N., Banck, M., James, C., Morley, C., Vandermeersch, T., & Hutchison, G. (2011). Open babel: An open chemical toolbox. *Journal of cheminformatics*, *3*, 33. https://doi.org/10.1186/1758-2946-3-33

Palais, B., Palais, R., & Rodi, S. (2009). A disorienting look at euler's theorem on the axis of a rotation. *The American Mathematical Monthly*, *116*(10), 892–909. https://doi.org/10.4169/000298909X477014

Pdb101: Learn: Guide to understanding pdb data: Missing coordinates and biological assemblies. (n.d.). *RCSB: PDB-101*. https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/missing-coordinates-and-biological-assemblies

Pettersen, E., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E., & Ferrin, T. (2004). Ucsf chimeraa visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, *25*.

Pliatsika, V. (n.d.). Dna and rna — computational medicine center at thomas jefferson university. — *Computational Medicine Center at Thomas Jefferson University*. https://cm.jefferson.edu/learn/dna-and-rna/#ref

Rna structure, function, and recognition ibiology. (2014). *iBiology*. https://www.ibiology.org/biochemistry/rna-structure/

Rna: The versatile molecule. (2016). *learn.genetics.utah.edu*. Retrieved June 4, 2021, from https://learn.genetics.utah.edu/content/basics/rna/

Schrödinger,LLC. (2020). *The pymol molecular graphics system* (Version 1.2r3pre). https://pymol.org/2/

Structure and Function of RNA [[Online; accessed 2021-06-27]]. (2021, January 3). https://bio.libretexts.org/@go/page/5177

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Wan, Y., & Chatterjee, K. (2018). *Rna — definition, structure, types, functions*. https://www.britannica.com/science/RNA

Warden, M. S., Tonelli, M., Cornilescu, G., Liu, D., Hopersberger, L. J., Ponniah, K., & Pascal, S. M. (2017). Structure of rna stem loop b from the picornavirus replication platform. *Biochemistry*, *56*, 2549–2557. https://doi.org/10.1021/acs.biochem.7b00141

Westhof, E. (2014). Isostericity and tautomerism of base pairs in nucleic acids [Paris]. *FEBS Letters*, *588*(15), 2464–2469. https://doi.org/https://doi.org/10.1016/j.febslet.2014.06.031

Westhof, E., & Pascal, A. (2006). Rna tertiary structure. https://doi.org/10.1002/9780470027318.a1428

Yikrazuul. (2010). English: X-ray structure of the trnaphe from yeast. data was obtained by pdb: 1ehz and rendered with pymol. *Wikimedia Commons*. Retrieved July 10, 2021, from https://commons.wikimedia.org/wiki/File:TRNA-Phe_yeast_1ehz.png

Zhang, Z., & Shafer, D. (2014). *Introductory statistics*.