An investigation of RNA using the discrete Frenet Frame

Daniel Neiss, Uppsala University
Master thesis
Supervisor: Antti Niemi, Uppsala University

March 2, 2015

Abstract

A brief explanation of RNA and its general structure on different levels is given. The standard continuous Frenet frame is explained. A discrete version of the Frenet frame is explained in detail and constructed for a piecewise linear curve. The results of the application of the discrete Frenet frame to RNA is shown in the form of several distributions. An analysis of these distributions is conducted and gives some results regarding tiny structures in RNA.

Acknowledgments

Acknowledgements go to my supervisor Antti Niemi and his excitement, Xubiao Peng for his advice and help, and Sean Gray for his friendship. I also want to thank my friends and family for their support and love. I also need to thank my frienemies, apparently!

Contents

1 Introduction

In this Master degree thesis the primary focus has been on investigating RNA using the a discrete Frenet frame approach. The discrete Frenet frame is a way to describe piecewise linear chain (polygonal chains) in a manner closely resembling the continuous Frenet frame approach. One may ask what is gained from applying the discrete Frenet frame to this problem and it is a good question that also has a reasonable answer. The discrete Frenet frame is in a sense one of the more "natural" ways to describe a piecewise linear curve rather than describing it using standard cartesian frames. This is similar to using an angular coordinate system to describe the position of stars on the night sky. By using the discrete Frenet frame, since it is a natural setting for investigation of RNA, there is hope to find symmetries, patterns and rules with regards to this setting.

Due to the nature of this project a lot of data from the RCSB protein databank has had to be investigated. This also means that a lot of effort has been put into learning the ropes of several programming aspects that are necessary to carry out such an investigation. Among these were data structure design choices, program flow choices and graphical programming challenges. Some, but not all, of these aspects turned out to be useful for the end results. As an example Java was used to write a PDB file parser, organising the data into a suitable and more easily handled format, and performing the basic construction of the Frenet frames along the RNA chain. However for data visualization purposes Java turned out to be very inadequate even when considering libraries others had written. In an attempt to keep developing an application for Java this led to learning how to use OpenGL. In the end, the time spent on OpenGL was not used, since Matlab turned out to be very capable of presenting data for the scope of this investigation. Thus the Java application was made to output the data into data files fit for Matlab which was used for the rest of the project.

Finally I wish to invoke a quote that very much describes this project:

"A couple of months in the laboratory can frequently save a couple of hours in the library." - Frank Henry Westheimer

On one hand, a lot more progress could have been made if more use of contemporary tools (such as existing programming libraries) and methods had been considered. On the other hand by working and making such tools myself I have become more skillful and achieved a deeper understanding of topics both related to and not immedietly related to the subject at hand. If I had merely relied on existing tools I would probably not have gotten those moments of wondrous realization that came with having to think about these problems myself. The key point is, sometimes it is most important to spend some time reading about what work has already been done; other times it is very worthwhile to approach the work independently. It's a fair compromise and I believe a fair balance between those two choices always has to be considered. This is what I've learned.

2 A brief explanation on the basic structures of RNA

Proteins, RNA (ribose nucleic acid) and DNA (deoxyribose nucliec acid) all have an essential role in cells and their biological processess. These are all composite objects that, together with other components, are essential for life. Most are familiar with DNA being the called the "blueprints" for living organisms, however, the role of RNA isn't as widely known.

In cells RNA has many roles that are vital for the cell to function properly. RNA molecules act as catalysts for many reactions, they play a vital role in controlling gene expression, as well as being able to sense and respond to cellular signals. An example of such a process is the process of protein synthesis, where the RNA in question (mRNA) is used as a transcription to take information from the DNA, and then is sent to the cytoplasm to be translated so that a protein can be formed. In this process transfer RNA (tRNA) delivers aminoacids to the ribosome and ribosomal RNA (rRNA) links them together to form proteins. Hence, while RNA itself may at first glance appear to be quite a simple object structurally, it can fulfill many different roles and has a very complex nature to it.

If one looks at DNA and RNA, one finds that they have many different layers of structure to them such as order, geometry and different chemical properties. RNA's structure at the chemical level is very similar to that of DNA with some key differences. Both DNA and RNA are made up of chains of certain sub-units. These sub-units are called nucleotides and are composed of a sugar, a phosphate and a nitrogen-base (also called nucleobase) that are bound together. The sugar in DNA is the pentose 2'-deoxyribose whereas RNA contains another pentose called ribose. The difference between these two is that the 2'-deoxyribose lacks a hydroxyl group which has been replaced by a single hydrogen at the 2'-position of the pentose. The addition of this hydroxyl group to RNA has a very important consequence in that it repels the phosphate. This makes RNA more prone to hydrolysis and doesn't allow it to twirl into as tight helices as DNA may.

In order for the nucleotides to form a long chain the phoshpates and pentoses in the chain bond together. They do so in an alternating fashion such that a pentose is always followed by a phosphate and vice versa by bonding with a phosphodiester bond. The phosphodiester bond is made up by the phosphate forming a covalent bond to the 5'-carbon of a pentose on one end, and a second covalent bond to the 3'-carbon of a second pentose on the other end. These bonds are further assisted by positive metal ions which prevent the negatively charged phosphordiester bonds from repelling each other [9]. This pentose and phosphate chain is what makes up the "backbone" (or "strand") of the RNA/DNA structure (see figure 1). The remaining nucleobases bonds at the 1'-carbon of the pentose and ends up "sticking out" of the strand. RNA is a single-stranded molecule. However, it may still form a double-strand structure through complementary base pairing of the bases in the strand e.g. in some viruses where it serves as genetic material. DNA however is usually double-stranded which is commonly pictured in the double-helix structure. This is similarly achieved by complementary nucleobases on two strands bonding to each other on the inside of the helix.

There are four nucleobases in DNA named Adenine (A), Thymine (T), Cytosine (C) and Guanine (G). Note that Thymine and Cytosine are pyrimidines while Adenine and Guanine are purines. These purines and pyrimidines constitute a set of complementary bases and through a hydrogen bond may form A-T and C-G bonds with each other. Thus if you have a double-helix of DNA, and know the order of bases on one strand of the DNA, you automatically know the order of the bases on the other strand as well since the bases are in one-to-one correspondence to each other. The nucleobases in RNA are the exact same as in DNA except that

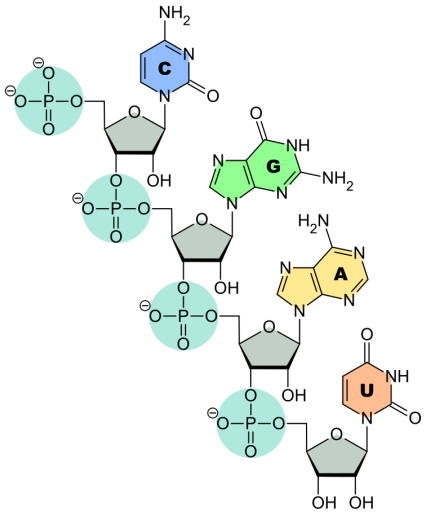

Figure 1: The basic components that form an RNA strand. The alternating phosphate and pentoses that bond together to up the backbone can be seen. One can also the see RNA's four nucleobases, Cyanine, Guanine, Adenine and Uracil. (public domain, taken from *http://commons.wikimedia.org/wiki/File:RNA-Nucleobases.svg*)

Thymine is replaced by another pyrimidine named Uracil (U). The nucleobases in RNA also form complementary bases in the same way using A-U and C-G.

There information and function of DNA/RNA isn't determined only by the nucleobase sequence. The geometry also plays a key-role and is achieved by the molecule folding into different shapes depending on e.g. the nucleobase sequence and surrounding molecules. These folds can quickly become rather complicated to describe, not only on a mathematical and physical level, but also just from a visual perspective. There are however some underlying substructures that can be of help when trying to describe how RNA (as well as DNA/proteins) are folded. These are conventionally named the primary, secondary, tertiary and quaternary structures and they are defined as the following (see [1], [8]):

1. Primary Structure: The primary structure of RNA is defined as the order/sequence of nucleotides in the RNA/DNA. However, RNA and DNA are usually regular in the nucleotide structure (in that they have the same pentose/phosphate molecules all the way through) so it is also commonly defined as the sequence of nucleobases. In most RNA If one looks at a strand of RNA then there are two natural ways to establish this sequence, either from the first end to the other end, or vice versa. Thankfully the ends of the RNA strand aren't exactly the same due to the nature of the phosphatediester-bonds. The phosphatediester-bond is connected to the 3'-carbon on one pentose and the 5'-carbon on the other. Physically any such sequence must eventually terminate on both ends, leaving an "open" phosphate group connected to the 5'-carbon of a pentose on one end, and leaving an "open" hydroxyl group connected to the 3'-carbon of another pentose on the other end. The convention for listing the primary structure is to start the list on the 5'-end of the strand and ending it at the 3'-end (usually just stated "5' to 3'"). An example would be a sequence "AUGTUGA..." for RNA, where we have used the ATGU nucleobases. After the forming of an RNA/DNA, some of the bases may become modified, and these modified bases also have specific names e.g. ψ for pseudouridine.

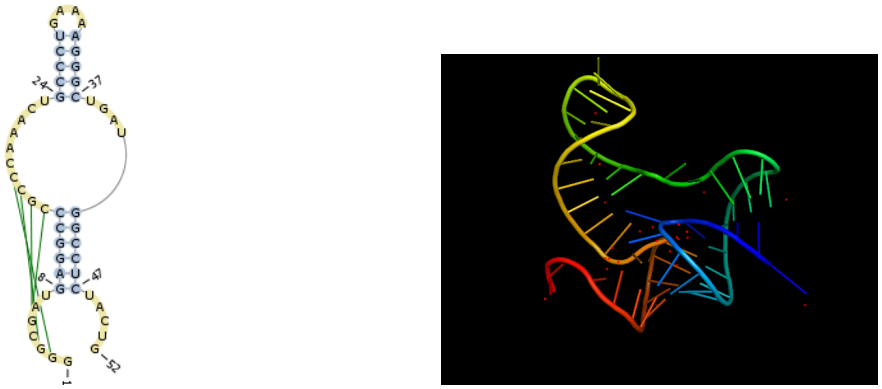

Figure 2: Figures showing the primary structure, secondary structure and tertiary structure of the 4ENC RNA. The primary structure is seen by the AUGC sequence. The secondary structure is seen by the base-pairings and the corresponding shapes they create. The tertiary structure is the collection and arrangement of the secondary structure motifs which creates the whole shape.

A "sequence-motif" may denote a specific sequence of nucleobases that are thought to provide a certain function.

Aside from the "ATGC" and "AUGC" there are also other letters with special meaning that may be used in the primary sequence. These letters are used when there could be "either one or the other" at a specific position in the sequence. Some examples would be W which is used for "A or T" and B which is used for "G or T or U". There exists a symbol for each possible (non-empty) combination, giving 15 symbols in total.

2. Secondary Structure: The secondary structure is the way the backbone of the RNA/DNA folds to create geometric structures. The folding process can be complex and gives a resulting pairing between nucleobases in the strand and can thus also be represented by a list which lists these pairings. This "list" of pairings and non-paired nucleobases is the secondary structure.

   Looking at a protein, RNA or DNA, this folding seems to create certain regular shapes for sub-segments of the strands. These regular substructures are commonly called "secondary structure motifs" (or just secondary structure). The most common example of secondary structure one usually sees is how two strands of DNA folds into a double-helix. For this to work the two strands have to be anti-parallell (in the "5' to 3' sense)", since the nucleobases on each strand has to fit together both as complements and in geometrical manner. There are actually three different double-helix motifs that are relevant called the A-DNA, B-DNA and Z-DNA. Seeing how RNA has different properties than DNA, among them that it is single-stranded, one should expect the secondary structures to be different. RNA may still form a double-helix similar in structure to the A-DNA double-helix. In a strand of RNA the nucleobases may still interact and pair up with each other, making the RNA strand fold "into itself". RNA also has a higher complexity due to the extra hydroxyl group in ribose making it easier to form hydrogen bonds. The results of such a folding could for example be the motifs called stem-loops, where the RNA nucleobases of a segment pairs up to form a "stem" with a loop on one end consisting of the nucleobases that aren't complements to

each other. Another example would be the pseudoknot which is composed of two stem-loops that are "interlaced" e.g. by having a stem-loop and then having a following segment of RNA form another stem-loop by basepairing to the loop-segment of the first. These motifs may serve as sub-motifs for even larger motifs e.g. certain tRNA which often has a cloverleaf motif made up of stem-loops. Note that these DO NOT cover the actual 3-dimensional shape of the molecule.

3. Tertiary structure: The tertiary structure of the molecule is its full 3-dimensional structure with all atoms specified by their atomic coordinates. In essence this describes how they curve around in space to form very complex shapes. Each possible shape may make a difference in the actual function of a protein, DNA or RNA.

   These shapes can be very complex, but one can often identify "tertiary structure motifs", which just as before are regular substructures in the tertiary structure. These motifs can be seen as building blocks for the whole structure, each with their own distinct shape. For DNA, the most easily and most often seen motif is the 3-dimensional double-helix shape. This double-helix makes on average one turn for every 10.5 base-pairs in the secondary structure. RNA may also form double-helix motifs, or even more impressively, DNA and RNA may even form triple- and quadruple helices. A particularly important process for the tertiary structure of RNA is "coaxial stacking". When coaxial stacking occurs two RNA duplexes whose segments lie close to each other link up in a conformation and become connected such that it is stabilized by base-stacking at the interface of the two helices. Coaxial stacking is a component of two common motifs, the kissing loop and the pseudoknot. The pseudoknot has already discussed where as the kissing loop is somewhat similar. The kissing loop is made up of two different stem-loops having a basepairing at each others' loop segments. This gives a coaxially stacked helix structure.

4. Quaternary structure: The quaternary structure is most often used for proteins. However, it may also be defined in a manner for DNA and RNA. An example for RNA would be the structure of two separate RNA interacting in a ribosome.

In the end all this means that what at first seemed like just a simple chain of different nucleotides may actually form a very large and complex objects due to factors such as the basepairs interacting with each other. What one ends up with is a approximately a space curve (the tertiary structure) which in turn is determined in some manner by the primary structure and secondary structure.

3 Summary

This thesis project's topic has been about RNA, visualization and identification of patterns. The RNA has been examined through the perspective of the mathematical Frenet frame of a curve. RNA however has a discrete nature to it rather than a continuous one as it is composed of many repeating units i.e. nucleotides. In response to this a discrete version of the Frenet frame is presented. This version of the Frenet frame gives a chance to set up a local frame at each nucleotide of the RNA in a systematic and natural way.

For the sake of our analysis the position of any local Frenet frame is situated near the center of the associated nucleotide's ribose molecule. All the atom-

coordinates of a given RNA is provided by PDB files downloaded from the RCSB Protein Data Bank [6]. Once all the positions are set up one can then derive the orientation of all the local Frenet frames along the entire RNA backbone.

Using this method 460 different PDB files have been examined. The analysis gives results in the form of distributions. Among these there are results regarding the distance between any two subsequent Frenet frames along the backbone, the distance between any Frenet frame and its associated residue in the nucleotide as well as a spherical distribution of angular position of a residue with regards to its associated Frenet frame. The distance between the residue and its Frenet frame is interesting as it shows that there are two preferred distances residues prefer with regards to the Frenet frame. Even more interesting is the spherical distribution of residues with regards to the Frenet frames. The spherical distribution shows that there is preferred direction for most residues in the data set that seems to correspond to symmetrical structures, however, there are also residues scattered about the rest of the distribution except for two "forbidden" regions. The reason for these forbidden regions as well as how the spherical distribution is related to the the distance distribution is discussed.

Finally an analysis is also carried out that examines how the residues that do not point in the preferred direction behaves. The results seem to indicate that there is some correlation between secondary structure elements such as bulges and hairpin loops and those residues. There also seems to be a connection to inflection/undulation points of the RNA.

4 Some mathematical background: The Frenet Frame and the Discrete Frenet Frame

In investigating RNA it helps to realize that the standard cartesian coordinate system is not always the most appropriate one to use. This should come as no surprise for those versed in natural sciences where choosing an appropriate coordinate system is a common practice. For our purposes we view RNA as a 1-dimensional string that curves around in 3-dimensional space. There are many ways to describe such a space-curve in a natural way (see [7], [11]) but in this report we use the standard Frenet frame as it suits our needs. There is however one issue; a continuous frame is not appropriate for our study of RNA and hence we use a discretized version of the Frenet frame. For a more in-depth discussion about the Discrete Frenet frame read [12] where the following two subsections are explained in great detail. Here we first provide a short introduction to the continuous case and then proceed to introduce the discrete case.

4.1 The continuous Frenet-Serret frame

The continuous Frenet-Serret frame (we will call it Frenet frame) is a natural way to describe a curve in 3-dimensions or higher dimensions (see [14] for an introduction). The continuous Frenet frame describes a curve solely through intrinsic properties of the curve, namely, the curvature and torsion.

Consider a smooth space curve $\gamma : [a, b] \to \mathbb{R}^3$. Assume this curve has a well-defined tangential component everywhere. For the signed arclength parameter s and the curve's position vector $\mathbf{R} \in \mathbb{R}^3$ the unit tangent vector $\mathbf{T}$ at any point is given by:

$$\frac{d\mathbf{R}}{ds} = \mathbf{T}$$

The unit tangent vector doesn't change length but it may still change direction along the curve. Hence we naturally define the principal normal $\mathbf{N}$ and curvature

κ as a measure of how much the curve differs from being a straight line (i.e. how much it curves). Symbolically, this is represented by:

$$\frac{d\mathbf{T}}{ds} = \kappa\mathbf{N}$$

The principal normal vector is defined to be of unit length i.e. $\mathbf{N} \cdot \mathbf{N} = 1$. This principal normal vector is also orthogonal to the unit tangent vector as can be easily seen by taking the derivative of $\mathbf{T} \cdot \mathbf{T} = 1$ which gives:

$$\frac{d}{ds}(\mathbf{T} \cdot \mathbf{T}) = 2\frac{d\mathbf{T}}{ds} \cdot \mathbf{T} = 2\kappa\mathbf{N} \cdot \mathbf{T} = \frac{d}{ds}1 = 0$$

Finally there is also the (unit) binormal vector $\mathbf{B} = \mathbf{T} \times \mathbf{N}$. This vector, in essence, describes how much the curve differs from lying in lying in a single plane. In order to get the corresponding measure of that difference, which we call the torsion τ, consider the following expressions:

$$\mathbf{N} \cdot \mathbf{N} = 1$$
$$\mathbf{N} \cdot \mathbf{T} = 0$$

Applying the derivative $\frac{d}{ds}$ to the former equation gives:

$$\frac{d\mathbf{N}}{ds} \cdot \mathbf{N} = 0$$

Similarly, applying $\frac{d}{ds}$ and then multiplying with $\mathbf{T} \cdot \mathbf{T} = 1$ to the latter equation one yields:

$$(\frac{d\mathbf{N}}{ds} + \kappa\mathbf{T}) \cdot \mathbf{T} = 0$$

Thus the vector $\frac{d\mathbf{N}}{ds} + \kappa\mathbf{T}$ is orthogonal to the tangent vector $\mathbf{T}$. It can also be shown that this vector is orthogonal to the principal normal vector $\mathbf{N}$. This vector is called the binormal vector $\mathbf{B}$ and the torsion τ is introduced by the relation:

$$\frac{d\mathbf{N}}{ds} + \kappa\mathbf{T} = \tau\mathbf{B}$$

Finally, one can arrive at a relation between the derivative of $\mathbf{B}$ and the other basis vectors. If one takes the derivative of the relation $\mathbf{B} = \mathbf{T} \times \mathbf{N}$ and note that some terms vanish, one arrives at:

$$\frac{d\mathbf{B}}{ds} = -\tau\mathbf{N}$$

In summary, one has the following differential equations for the curve which collectively are called the "Frenet formulas":

$$\frac{d\mathbf{N}}{ds} = -\kappa\mathbf{T} + \tau\mathbf{B}$$
$$\frac{d\mathbf{B}}{ds} = -\tau\mathbf{N}$$
$$\frac{d\mathbf{T}}{ds} = \kappa\mathbf{N}$$

Equivalently, in matrix notation the Frenet formulas take the following anti-symmetric form and constitute a linear system of vector ODEs:

$$\frac{d}{ds}\begin{bmatrix}\mathbf{N}\\ \mathbf{B}\\ \mathbf{T}\end{bmatrix} = \begin{bmatrix}0 & \tau & -\kappa\\ -\tau & 0 & 0\\ \kappa & 0 & 0\end{bmatrix}\begin{bmatrix}\mathbf{N}\\ \mathbf{B}\\ \mathbf{T}\end{bmatrix}$$

As such the triplet $(\mathbf{N}, \mathbf{B}, \mathbf{T})$ constitute an orthonormal (right-handed) basis in $\mathbb{R}^3$. Note that the fundamental theorem of space curves states that the curve is unique only up to an euclidean motion in $\mathbb{R}^3$ and only as long as the curve has non-vanishing curvature $\kappa > 0$ everywhere (see [10]). Because of this there are certain trapholes when one has to consider curves with inflection points. These issues are examined more closely in [12].

4.2 The discrete Frenet-Serret frame

While one may view a strand of RNA as a continuous piece of string it can also be viewed as a discrete chain of sub-units (e.g. its nucleotides). As such a continuous frame is not applicable and one has to find a discrete approach to describing the chain. That's where the discrete Frenet frame comes into play. This discretized version of the continuous Frenet frame is thoroughly investigated in [12] and only an overview is given here.

One thing to note is that, in relation to the continuous case, the discrete Frenet frame is not a single frame per-se but actually a sequence of recursively defined frames $\{\mathcal{F}_i\} = \{(\mathbf{n}_i, \mathbf{b}_i, \mathbf{t}_i)\}$ situated at different positions along a curve. If the curve is parametrised by some sequence of points along the curve $\{i\}$ with positions $\{\mathbf{r}_i\}$, we will have to use some specific information of the curve at the points $i-1$ and $i+1$ in order to construct the Frenet frame $\mathcal{F}_i$ at that point. Note that this means that the first and last points of the sequence won't have a Frenet frame associated with them since there is no preceding/succeeding point one can do the recursion on.

Any specific frame in the recursive sequence is defined as the ordered set of orthogonal basis vectors $(\mathbf{n}_i, \mathbf{b}_i, \mathbf{t}_i)$. These vectors correspond to a tangential component, a normal component and a binormal component of the Frenet frame in the sense of a discrete chain rather than a continuous curve. These different components are calculated by first defining the position vectors $\mathbf{r}_i$ of the collection of points (or equivalently the position of the local Frenet frames) with respect to some standard basis.

The Frenet frame $\mathcal{F}_i = (\mathbf{n}_i, \mathbf{b}_i, \mathbf{t}_i)$ at each position $\mathbf{r}_i$ has its components calculated in the following way. The unit tangent vector $\mathbf{t}_i$ is, not surprisingly, defined as the normalized vector with its origin at one Frame in the sequence and pointing to the next:

$$\mathbf{t}_i = \frac{\mathbf{r}_{i+1} - \mathbf{r}_i}{|\mathbf{r}_{i+1} - \mathbf{r}_i|}$$

Since all position vectors are assumed to be known, once all the tangent vectors are calculated, one can calculate the position of any Frenet frame recursively from the ones before it in the sequence. Finding the position of the k:th Frenet frame is easily accomplished (assuming the first Frame is at the origin) by adding together all tangent vectors up to that point, that is:

$$\mathbf{r}_k = \sum_{i=0}^{k-1} |\mathbf{r}_{i+1} - \mathbf{r_i}| \cdot \mathbf{t}_i$$

Next define the unit binormal vector $\mathbf{b}_i$ as:

$$\mathbf{b}_i = \frac{\mathbf{t}_{i-1} \times \mathbf{t}_i}{|\mathbf{t}_{i-1} \times \mathbf{t}_i|}$$

Note that the $\mathbf{b}_i$ is only well-defined when no two consecutive tangent vectors are parallell. This is similar to the continuous case where the normal vector isn't well-defined for points of zero curvature such as inflection points. Note also that $\mathbf{b}_i$ depends on $\mathbf{r}_{i+1}$, $\mathbf{r}_i$ and $\mathbf{r}_{i-1}$. This dependence has the consequence that there will be no Frenet frames defined for the first and last point of the sequence. Finally, define the unit normal vector $\mathbf{n}_i$ as:

$$\mathbf{n}_i = \frac{\mathbf{b}_i \times \mathbf{t}_i}{|\mathbf{b}_i \times \mathbf{t}_i|}$$

The recursively defined vectors $\{\mathbf{n}_i, \mathbf{b}_i, \mathbf{t}_i\}$ are all orthonormal to each other and constitute the discrete Frenet frame $\mathcal{F}_i$ at position $\mathbf{r}_i$ on the curve.

With the sequence of Frenet frames $\{\mathcal{F}_i\}$ in hand one can also introduce the transfer matrix. The transfer matrix $\mathcal{R}_{i+1,i}$ describes how to transform from the Frenet frame $\mathcal{F}_i$ to the Frenet frame at the succeeding point $\mathcal{F}_{i+1}$:

$$\begin{bmatrix} \mathbf{n}_{i+1} \\ \mathbf{b}_{i+1} \\ \mathbf{t}_{i+1} \end{bmatrix} = \mathcal{R}_{i+1,i} \begin{bmatrix} \mathbf{n}_i \\ \mathbf{b}_i \\ \mathbf{t}_i \end{bmatrix}$$

Since the curve lies in $\mathbb{R}^3$ the transfer matrix is intuitively a re-orientation of the first Frenet frame so that it becomes the second Frenet frame i.e. it is a rotation. This rotation can be explicitly determined using Euler angles. Choosing the (z, x, z) Euler angles with inclination angle $\theta \in [0, \pi]$ and azimuthal angles $\psi, \phi \in [-\pi, \pi)$ the transfer matrix takes on the general form:

$$\mathcal{R}_{i+1,i} = \begin{bmatrix} -\sin\psi\sin\phi + \cos\theta\cos\psi\cos\phi & \sin\theta\cos\psi & -\sin\psi\cos\phi - \cos\theta\cos\psi\sin\phi \\ -\sin\theta\cos\phi & \cos\theta & \sin\theta\sin\phi \\ \cos\psi\sin\phi + \cos\theta\sin\psi\cos\phi & \sin\theta\sin\psi & \cos\psi\cos\phi - \cos\theta\sin\psi\sin\phi \end{bmatrix}_{i+1,i}$$

Note once again that here $(\theta, \psi, \phi)_{i+1,i}$ are the angles relating frame $\mathcal{F}_i$ to frame $\mathcal{F}_{i+1}$ and that there is a whole sequence of them along the chain. This general form of the transfer matrix can be significantly simplified due to the construction of the discrete Frenet frame. For instance, from the definition of the binormal vector $\mathbf{b}_i$ it is very easily seen that $\mathbf{b}_{i+1} \cdot \mathbf{t}_i = 0$. If one looks at which entry this corresponds to in the transfer matrix one sees that the following must hold:

$$(\sin\theta\sin\phi)_{i+1,i} = 0$$

Moreover, with some effort, one can show that:

$$\phi_{i+1,i} = 0$$

Rewriting the transfer matrix one then gets the relatively simple form:

$$\mathcal{R}_{i+1,i} = \begin{bmatrix} \cos\theta\cos\psi & \sin\theta\cos\psi & -\sin\psi \\ -\sin\theta & \cos\theta & 0 \\ \cos\theta\sin\psi & \sin\theta\sin\psi & \cos\psi \end{bmatrix}_{i+1,i}$$

Along the diagonal there are two simple terms to pay attention to. Let $\psi_{i+1,i}$ be defined as the *discrete bond angle* and let $\theta_{i+1,i}$ be defined as the *discrete torsion angle*. Then one sees that they are given by:

$$\cos\psi_{i+1,i} = \mathbf{t}_{i+1} \cdot \mathbf{t}_i$$

$$\cos\theta_{i+1,i} = \mathbf{b}_{i+1,i} \cdot \mathbf{b}_i$$

These angles are another way to describe the chain of Frenet frames by. Finally, for the purpose of this investigation of RNA, it is also useful to construct a spherical coordinate system at each Frenet frame $\mathcal{F}_i$ in a systematic way. Consider an object viewed from the frame $\mathcal{F}_i = (\mathbf{n}_i, \mathbf{b}_i, \mathbf{t}_i)$. The position $\mathbf{r}$ of the object is then expressed as:

$$\mathbf{r} = n\mathbf{n} + b\mathbf{b} + t\mathbf{t}$$

The spherical coordinate system at $\mathcal{F}_i$ is then given by the same procedure as for normal cartesian coordinate systems. Namely we introduce a set of coordinates (r, θ, ϕ) (not to be confused with the earlier Euler angles) where r is the radial distance from the origin of the frame, $\theta \in [0, \pi]$ is the inclination angle ranging from north-pole (positive $\mathbf{t}_i$ axis) to the south-pole (negative $\mathbf{t}_i$ axis) and $\phi \in [-\pi, \pi)$ is the azimuthal angle counted as positive being counter-clockwise from the $\mathbf{n}_i$ axis as watched from the north-pole. Then the spherical coordinates are related to the frame's coordinates by:

$$r = |\mathbf{r}| = \sqrt{n^2 + b^2 + t^2}$$
$$\theta = \arccos(t/r)$$
$$\phi = \arctan(b/n)$$

The spherical coordinate system is useful since it enables us to systematically look at spherical distributions along the chain. This is comes in handy when one wishes to relate the Frenet frame (which is intrinsically geometric) to e.g. the orientation of the nucleobase with regards to the frame at each point on the chain. Then, depending on how an RNA curls and what secondary structure it has during a particular segment, one can hopefully see patterns in how the nucleobases during this segment moves around on the spherical distribution.

5 Results

5.1 Results from total data

In this section the application of the discrete Frenet frame approach to RNA is described along with other methods. Among the results there are some answers to questions like how distances between the resulting Frenet frames in a chain and other elements of the chain vary. Moreover the angular distributions of nucleobases with regards to their associated Frenet frames are presented. This has been done in hopes of finding distinct signs in the data which could correspond to different structures in the RNA such as different types of secondary structure or tertiary structures.

The total amount of RNA investigated is about 460 RNA pdb-files. The files were downloaded from the "RCSB Protein Data Bank"[1] in the aforementioned .pdb file format. A Java program was written to parse the contents of these files and to construct a Frenet Frame for each corresponding RNA in the files. In particular, the interesting sections of the pdb files to parse was the contents of the Coordinate section (see [2] for documentation) that contains all the information about the atoms in the RNA.

In order to set up the discrete Frenet frames along the RNA strands one has to choose a good set of position vectors along each strand such that they describe the strand in a proper way. In our case, they have been defined as the average position of the $C4'$, $C3'$ and $C1'$ atoms of the pentose in each nucleotide along the backbone of the RNA-strand using the standard coordinates specified in the RNA's

[1] www.rcsb.org

corresponding pdb-file. To be more clear, for each nucleotide i, the positions vectors are defined in the following way:

$$\mathbf{r}_i = \frac{\mathbf{r}_{i,C4'} + \mathbf{r}_{i,C3'} + \mathbf{r}_{i,C1'}}{3}$$

It's important to note that this isn't the only way one could define the position vectors. Any choice that faithfully represents the geometrical structure of the strand can be used i.e. any choice of vectors that describe the position of each and every nucleotide in the RNA. Good choices of position vectors should preserve the geometrical structure and produce results that are essentially similar to another good choice of position vectors. Here three carbon atoms in the pentose has been used because it was first decided that we look at RNA with a slightly different (non-Frenet) frame. This led us to define a plane (which requires 3 coordinate points) which gave another way to construct a normal vector as well as . This choice of using the 3 carbon atoms remained when applying the Frenet frame but it should still be stressed that any other good choice of position vectors should do just as well.

From this definition of the position vector the Frenet frame is then easily constructed using the tangent vectors between pairwise positions $\mathbf{r}_i$ and $\mathbf{r}_{i+1}$. This choice of position vector is also a good one since the pentose is a unit of the backbone that occurs regularly along the whole strand. One can then get a good sequence of Frenet frames each centered on the corresponding position vector along the whole RNA strand.

In examining the RNA we have only paid attention to internal chains composed of the standard residues Guanine (G), Adenine (A), Uracil (U) and Cytosine (C). Any non-standard residues have thus not been taken into account. This also has the consequence that the sequence of Frenet frames along the RNA strand is partitioned into subsequences with non-standard residues acting as delimiters. An example would be that a chain AGUUC(non-standard)CGU gets split into AGUUC and CGU subchains. This treatment has been done in order to ensure that none of the Frenet frames differ from each other too much as non-standard residues may have a stronger or different influence on the RNA chain than the standard nucleobases do. This also means that only subchains of length 3 or longer will have Frenet frames associated with them.

One more thing to note is that in order to get an angular distribution of the residues a spherical coordinate system has been associated with each Frenet frame in the chain. In this way one can then relate the position of the first atom in the associated residue to the orientation of the Frenet frame. This means that the position of the residue is the position of the $N1$ atom for the pyrimidines Cytosine and Uracil, where as it is the position of the $N9$ atom for the purines Adenine and Guanine.

Now, as for the results, a few different properties have been investigated. The total results with statistics calculated over all of the RNA files will be presented first, with arguments relating to more specific RNAs coming later.

The first result that will be presented is the pairwise distance between local Frenet frames in a given chain. This result does not directly relate to the Frenet frame themselves. One could just as easily have gotten the results by taking the distance between the pentoses in a similar way to how we defined the position of our Frenet frames. Now, since the RNA has a very symmetric molecular structure, as it is a phosphate $\rightarrow$ ribose $\rightarrow$ phosphate sequence, one would suspect that the pairwise distances between frames shouldn't vary too much. There should be a higher and lower bound for this distance, as well as some preferred distance. The distances between Frenet frames i and $i+1$ are given by the length of the non-normalized tangent vector or merely as the length of the difference of their position vectors:

$$d_{i,i+1} = |r_{i+1} - r_i|$$

By calculating these distances for all of the 460 RNA which amount to 40000 Frenet frames, in a chainwise manner, and putting all the results together, one may get the histogram seen in Figure 3.

From the figure one can see that there is some preferred distance for the Frenet frames with a peak at about 5.89Å. In the regions around the peak there is one fact that is clearly visible. On the left side of the peak, we see that there are very few local Frenet frames that stay closer together than the preferred distance. On the other hand, on the right side of the peak, the Frenet frames are more likely to keep some larger distance between themselves leading to the tail seen in the figure. This might be related to the torsion and curvature of the chain of Frenet frames (or equivalently, the discrete bond- and torsion-angles). Some of the most common secondary structure motifs are motifs such as the hairpin-loop/stem-loop, pseudoknots and similar structures. When two different parts of the RNA strand(s) join together for a longer segment, they typically produce what resembles a helical structure. If these helices are very common, then since the Frenet frame gets a bit closer to each other due to bending of the strand in a helix, the peak on the histogram will be translated to the left. Then, by the same argument, we would expect shorter distances to the left of the peak to correspond to very tight turns in the RNA strands, while longer distances to the right of the peak corresponds to relatively straight segments of the RNA strands. In this project no investigation has been carried out to determine if there is any correlation between curvature/torsion and the distance distribution just discussed though there are certainly arguments such as the one just mentioned that would make one expect a relation.

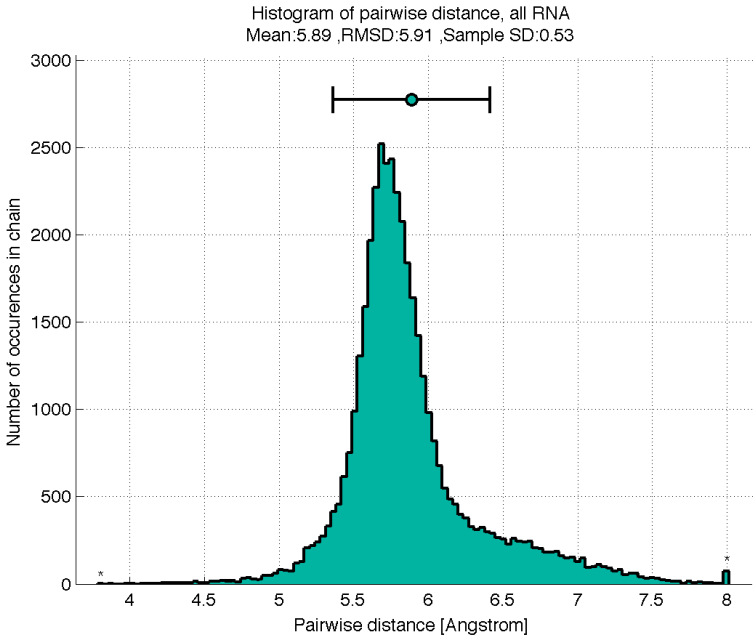

Figure 3: This histogram shows the distribution of the pairwise distances between local Frenet frames. The distribution is based on all 460 pdb files that have been analysed. There is a very notable peak with a trailing edge to the right.

The SampleSD in the figure is the unbiased sample standard deviation (Bessel's correction), which for a sample population of n and measured variable x is given

by:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Following up with a similar result is the distance between the Frenet frame and its associated residue. This distance is yet another measure that does not directly relate to the Frenet frame. The Frenet frames are mainly concerned about orientation while not being concerned with distances as long as one looks at position vectors evenly spaced along the backbone of the RNA. The distance between the Frenet frame and the residue could merely be expressed as the euclidean distance between the origin of the Frenet frame and the position of the residue itself. However, it can also be expressed as the radius component of a spherical coordinate system centered at the Frenet frame as follows:

$$d_{R_i,i} = |\mathbf{r}_{R_i} - \mathbf{r}_i| = r_i$$

Here $\mathbf{r}_{R_i}$ is the position vector of the residue associated with Frenet frame i, where as r_i is the radius-component of the residue's position expressed in the spherical coordinate system derived from the Frenet frame. The easy way to see the second equality is to realize that that the LHS is a vector pointing from the origin of the Frenet frame towards the position of the position of the residue. From that, one can then calculate the components of the position with regards to the local Frenet frame, and from there express it in the derived spherical coordinate systems.

Remember that the orientation of the spherical coordinate system is derived with the tangential component of the Frenet frame being considered the "z-axis" of the spherical coordinate system, that is, it specifies the north- and south-pole. This gives the polar angle θ. Similarly, the binormal and normal components of the Frenet frame span the "xy-plane" and the azimuthal angle ϕ is the angle in the counterclockwise direction from the binormal. This in total gives the residue position in the spherical coordinate system (with its origin at the Frenet frame origin) as (r, θ, ϕ).

Performing this calculation for all Frenet 40000 Frenet frames and residues, and putting the results into a histogram, one gets Figure 4. This time there are two distinctly visible peaks along with a fairly deep valley between them. About 75% of the data lies between 2.4Å and 2.6Å. The exact reason for the two peaks has not been investigated. However, the lower peak seems to hint at something in the upcoming analysis of the spherical distribution as a similar number appears there. Perhaps the lower valley corresponds to very tight turns where the residue ends up flipping over to the outside of the turn. This may also happen in hairpin loops and bulges where uneven base-pairing might leave no room for some of the residues.

There is a more interesting aspect of Frenet frames, namely the fact that their orientations are derived from the geometry of the RNA, and this gives that it is very interesting to investigate the spherical distributions of the residues. Using the spherical coordinate system one can get a proper view of the positions of residues with regards to the orientation of their associated Frenet frame. This is done in hopes of finding some regular structures in the distribution that corresponds to secondary structure or tertiary structure motifs. The total angular distribution for all 460 of the pdb files are given in figure 5. Note that these figures pay no attention to the radial component of residues' position vector and focuses only on the angular components.

At first glance, there seems to be only one small region where residues fall into. Most of them fall into the red region, and the density quickly dissipates along the

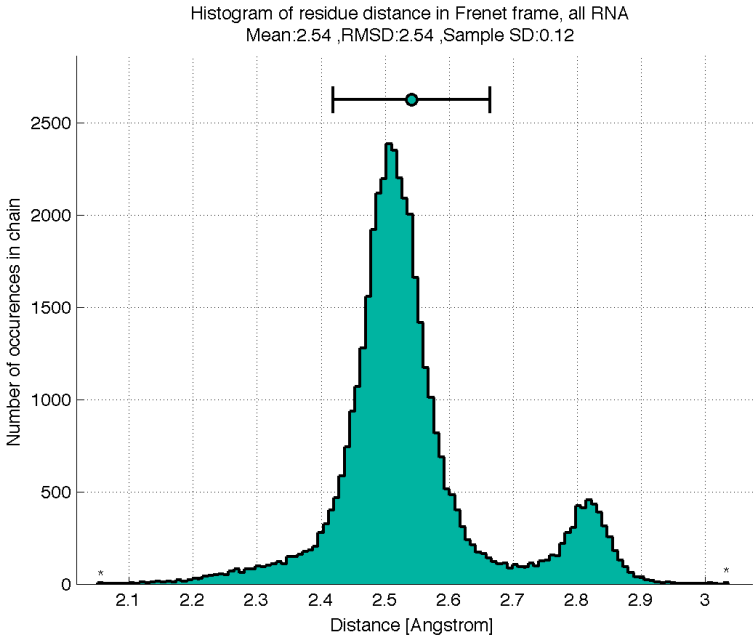

Figure 4: This histogram shows the distribution of distances between local Frenet frames and their associated residues. The distribution is based on all 460 pdb files that have been analysed. There are two peaks, one of which is fairly small but still significant, as well as the valley between the two peaks.

3 "legs" sticking out of the region. However, notice that the red color corresponds to a density of 3000 residues in a bin, and one of the first blue steps correspond to a density of 500 up to 1000 residues in a single bin. This kind of linear plot does not depict the situation truthfully. In order to get a clearer picture of what is going on lets instead look , at the same data, using a logarithmic scale instead. This is given in the figure 6.

From the 10-logarithmic plot, one sees that while the red region still dominates, there is still a considerable amount of residues not pointing towards it. If one defines the red region, which will now be referred to as the "hotspot", to be inside the region $\theta \in [0.959, 1.833], \phi \in [-1.571, 0.3491]$, then the total amount of residues inside amounts to 32.000, while the residues outside of the red region amount to 10.000 scattered over the rest of the sphere. There's also two (or perhaps three, if the southpole is included) "forbidden regions" that contain very few residues. Interestingly enough these numbers match up to the 75% measure that was found for the residue distances. About 76% of the residues in the spherical distribution lies within the hotspot while 24% of them lie around the rest of the distribution. About 87% of the residues in the hotspot also lie in the 2.4Å-2.6Å residue distance region so there is certainly a correlation between them that may be worth investigating further. Many of the RNA with very symmetrical structures such as helices have almost all of their residues pointing towards the hotspot so it may very well be the cause for the peak in distance in the 2.4Å-2.6Å region. Furthermore, if one looks at residues in the hotspot only 1% of them seems to lie in the 2.75Å-2.85Å region of the residue distance plot. Thus the second peak seem to have a larger contribution from residues outside the hotspot.

The forbidden regions seem very interesting. The first forbidden region around the northpole may not be too surprising. The reason that no residues point in that direction is because there is likely to be another residue in that spot, since

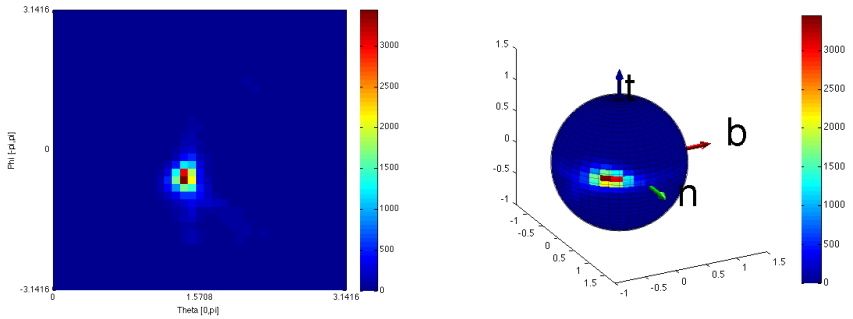

Figure 5: Total spherical distribution of the residues with regards to their associated Frenet frames $\{\mathbf{n}_i,\ \mathbf{b}_i, \mathbf{t}_i\}$. These plots show a spherical distribution and a planar distribution of the θ ϕ coordinates. The planar plot corresponds to a 2D histogram over the angular coordinates, where the color of each bin corresponds to how many elements fall inside this bin. The residues seem to have a preferred orientation with regards to each frame as almost all of them seem to fall into the red region.

the tangent naturally points towards the next Frenet frame in the chain. The size of this region (as in the angle of the conical region) should therefore be linked to the distance between the different Frenet frames in the chain, since if two Frenet frames were very close, the residue would have very little freedom to point in the tangential direction, where as if they were to be very far apart, then the residue could possibly fit into that direction quite easily.

The second forbidden region, which is located just below the equator, could have a number of reasonable causes. The first, and admittedly most boring cause, could be that it corresponds to the location of the previous Frenet frame and its residue, causing hardly any residues to point in that direction. This would not be all that surprising since by far most residues are pointing towards the hotspot. If we have a whole chain of frames and residues that point only towards the hotspot, then we would expect this to create a symmetric structure on the RNA chain (most likely a regular helical shape). Then, if one were to take any specific frame-residue-pair in the chain, and check the position of the next such pair in the chain, it would lie in the tangential direction. Similarly, if one were to check the position of the previous such pair in the chain, could it possibly lie in the second forbidden region? This is certainly a possibility and could explain why that region is forbidden (this would most definitely be possible to check by comparing a spherical distribution plots, one traversing the chain in the 5'-to-3' direction, and one traveling in reverse). However, assume we have one of the chains of residues that do not point into the hotspot. There is no immedietly obvious answer to why such a chain couldn't point into the second forbidden region. If the argument holds this chain would correspond to a more irregular structure and there is no a priori reason why, if the residues are randomly distributed, there shouldn't be any residues in the forbidden region. If there was no physical reason for the forbidden region then the residues should be distributed uniformly (i.e. the distribution of the irregular structures should be similar if one performs any SO(2) rotation around the tangent vector) but this isn't the case. There are certainly some residues in the region, but far less than there are around the rest of the sphere, so there still seems to be signs of this region having a physical significance.

Therefore, that brings up another possibility, namely that there is some other physical or chemical cause that prevents the residues from lying in that region.

Regardless of which is the case, the presence of these holes gives an opportunity to investigate another property of the residues in an RNA chain. On a

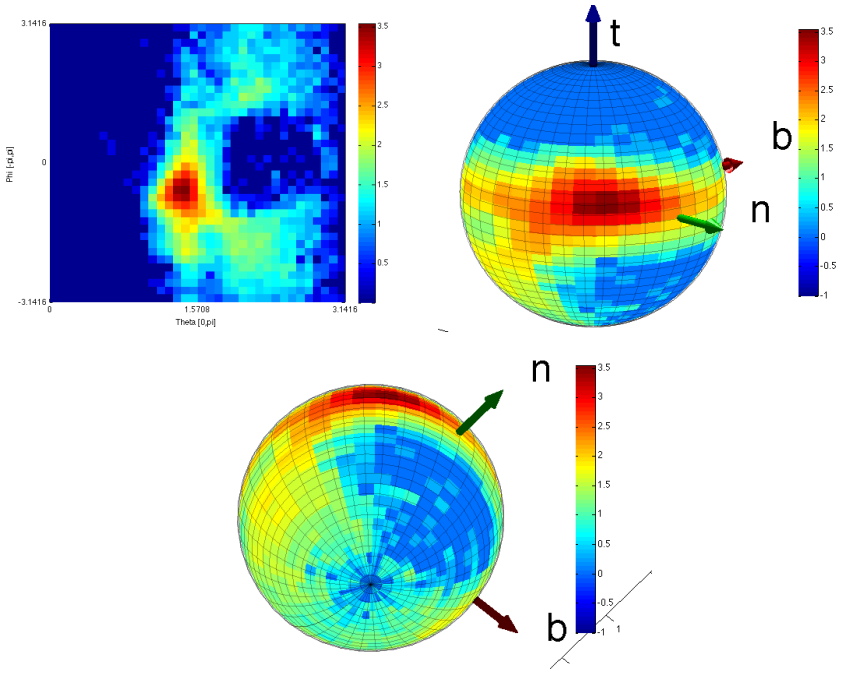

Figure 6: Total spherical distribution of the residues with regards to their associated Frenet frames $\{\mathbf{n}_i, \mathbf{b}_i, \mathbf{t}_i\}$. These plots are in a 10-logarithmic scale over the total angular distribution dataset. Almost all residues lie in the red region, referred to as the "hotspot", but there are also many that are scattered about the rest of the sphere at one or two orders of magnitude less than the red region. In the logarithmic scale, one can identify what appears to be a large forbidden region towards the northpole and another (nearly) forbidden region on the southern hemisphere.

vector field with a zero/singular point one can define an index or winding number around this point. Similarly we wish to examine how the residues of our RNA chains "move around" the forbidden regions in our distribution. Consider the case that we have an RNA, and check where each residue points with regards to their associated Frenet frame. If we move along the chain, checking the residue orientation, eventually we will find a residue that does not point into the hotspot of the distribution. We make a note of this residue, and then continue on to the next one and check if this new residue too does not point into the hotspot, if it doesn't, we draw an arrow connecting the two residues. We continue in this way, until we eventually find a residue that points into the hotspot, and thus we terminate the arrow-chain we have constructed. We do this for all strands in the residue and end up with a collection of such arrow-chains (where each chain may differ in length). The aim of this procedure is then to see if certain RNA motifs, whether primary, secondary or higher-order, correspond to some kind of pattern on the distribution. An example would be if one could identify what corresponds to helix, pseudoknot, coaxial stacking or common motifs like the kink-turn by utilizing an existing database (such as [3]).

In order to visualise this, the spherical distribution has been punctured at the southpole, and then flattened out. This could be done by means of stereographic projection but using such a projection would give a very distorted picture as points near the pole would be projected to points very far away from the origin. Instead, since the spherical distribution only depends on the two coordinates θ_{sphere} and ϕ_{sphere}, one can think of these as a polar plot (r_{pol}, θ_{pol}) in the plane using the

following identifications:

$$r_{pol} = \theta_{sphere}, \quad r_{pol} \in [0, \pi] \tag{5.1}$$
$$\theta_{pol} = \phi_{sphere}, \quad \theta_{pol} \in [-\pi, \pi] \tag{5.2}$$

The polar coordinates are defined on the same intervals as their spherical counterparts are. This choice of polar coordinates corresponds to having punctured the sphere at the southpole. For the polar plot this means that the boundary of the disk defined by its coordinates will correspond to the southpole. This one-to-many map is slightly problematic as will be discussed later.

A normally important and interesting property of a projection of a 2D-surface is its ability to preserve area or angles. This is of no concern here even though a projection such as this is probably bound to introduce quite some distortions. The reason it is of no concern is merely because such properties are not interesting for the subsequent analysis. The important part is that the map is (almost everywhere, not at the southpole) continuous, such that holes and regions still show up as holes and regions in the polar plot, which our map clearly accomplishes.

The resulting figure applied to the total dataset is given in 7 As expected, the details are all still there, with the northpole hole being in the center, the hotspot clearly visible, and with the second hole visible to the right of the hotspot. However, in the subsequent analysis, a grayscale version will be used due to better color properties. Note also the rather inconvenient fact that points near the boundary are in fact very close to each other even if they may appear to be at opposite sides of the disk. This is because the whole boundary corresponds to the southpole.

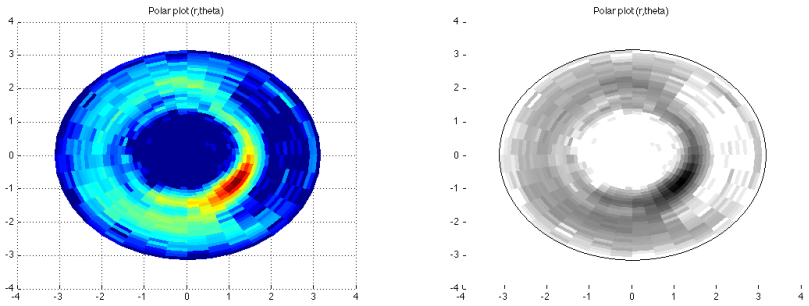

Figure 7: Polar plots corresponding to the spherical angular distribution. This plot has the northpole at the center while the southpole is the boundary. All the important features of the distribution such as holes are still visible. The righthand figure has a different colormap that will be used due to better color properties for the upcoming analysis.

Now one can easily check what happens with chains of residues that are outside the hotspot using the above mentioned arrow-plot. However, the arrow-plots are not taking into account that the boundary is traversible, i.e. all points on the boundary are mapped to the same point which corresponds to the south-pole, hence some residues will appear to move around more than they should. As an example, just consider two points lying near the boundary but at opposite ends of the disk, these will look like they are separated by a large distance, but they are in fact only separated by a small distance since the boundary is traversible. This is an issue that is discussed later on. The following will be a small analysis on some specific RNA. Note, that in all the arrow-plots, the arrows have been overlayed onto the polar plot of the total angular distribution rather than a polar plot of the distribution of the RNA itself. This has been done to get a clearer picture of where the arrows go with regards to the distributions appearance as some RNA have far too little residues to make a clear distribution.

5.2 1CSL

The 1CSL RNA is comprised of two RNA strands forming a straight helical structure. The strands are called Chain A and Chain B respectively. Chain B has some deformations in it compared to Chain A due to uneven base-pairing between them. 1CSL is a relatively small RNA with only 28 residues total but none the less shows an important feature. Chain A has 13 residues, all of which lie in the well-defined peak seen in figure 9, where as chain B has 15 residues and also has most residues in the peak, however it also has two residues which lie far outside. Looking at the base-pairing diagram (obtained by using the tool supplied by [4] , a RNA secondary structure prediction tool), we see that chain B's primary structure does not complement chain A's primary structure perfectly, which introduces some bulges. These bulges are visible on the RNA as sharp turns in the chain B strand. The residues in this sharp turn flip to the outside of the helical structure as can be seen if one loads 1CSL through a viewing-program such as pyMol. This behaviour is present in many other RNA as well.

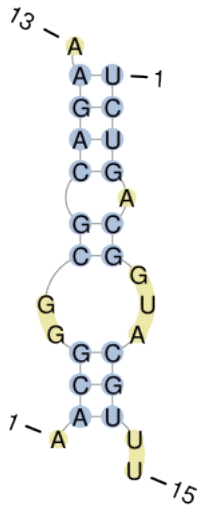

Figure 8: A picture generated by a secondary structure prediction algorithm giving the structure of the 1CSL RNA. From the picture one can identify some basic structures such as the clearly visible bulge that is formed from uneven base-pairing. This picture has been generated using [4].

In the 1CSL RNA, we see the first correlation between distance from a Frenet frame to its associated residue and the residues that are not pointing into the hotspot of the distribution. The A chain which follows an entirely regular structure has all its residues at around the peak distance of 2.4Å. Meanwhile, chain B, which has a bulge, has two residues that lie at 2.85Å away from their associated frames. Correspondingly, since chain A has such a regular structure, which can be seen in figure 10 as it forms a very regular helical structure, it does not have any residues pointing out of the hotspot. If one instead looks at chain B, it has two chains of length 3 that point out of the hotspot. The secondary structure corresponding to these two chains in Chain B can also be seen in the figure. As expected, the sections of the B-chain that are parts of the bulges correspond to residues which do not point into the hotspot of the angular distribution. However, the residues that are outliers in the distance-histogram correspond to the two residues in the pink chain. The residues in the blue chain, despite not belong in the hotspot of the angular distribution, do lie in the peak of the distance-histogram.

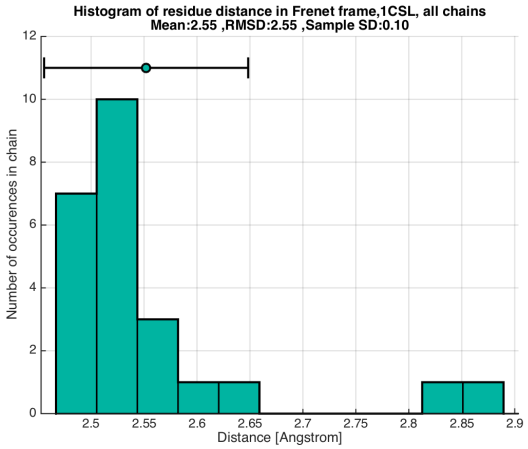

Figure 9: Histogram of the distance between the residues and their associated Frenet frames in the B-chain of the 1CSL RNA. Most of them lie in the expected peak region of 2.5Å, but there are also two outliers around 2.85Å.

The corresponding arrow-plot is also visible in figure 10. The way to understand these plots is to realise a few things. The dots in the plot are there to show where the corresponding residue is located in the angular distribution. The arrows between the residues, as well as the labels S (Start) and E (End), show how the residues change along the chain read in the 5'-to-3' direction of the RNA strand. There are no arrows in or out of the hotspot because including those arrows would just clutter the plot. Note however, that already here, one can see an issue with using a planar and polar plot of the 3-dimensional distribution. It looks like the first step in the blue chain goes all the way from the left side of the plot to the right side of the plot. It is in fact a much shorter distance than what it appears to be because the arrow could instead have traversed the boundary (the south pole) to reach that point. This is an issue and improvements upon it are discussed in a later section.

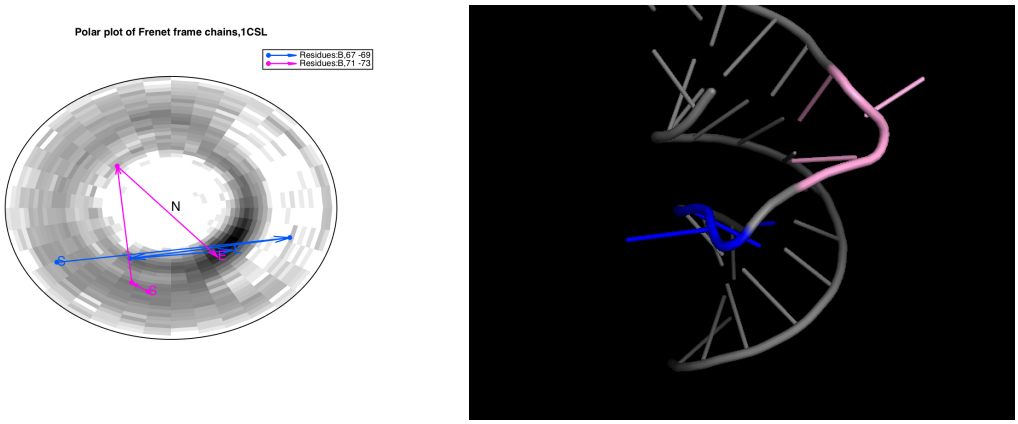

Figure 10: The left plot shows the two chains of 1CSL that have residues which are not pointing into the hotspot. Chain A has no such chains, but chain B has two. Each such chain is 3 residues long and correspond to tight turns in the RNA secondary structure as a direct result of the primary sequences not complenting each other perfectly. The corresponding secondary structure in the RNA itself can be seen in the figure on the right. The colors in the left and right pictures correspond to each other.

5.3 1KXK

The 1KXK RNA is a short RNA and has a highly regular structure. It is single-stranded and forms a helical shape, but has a few bulges and tight turns along the way, and loops in a hairpin-loop at one end. The corresponding arrow-plots in figure 12 and figure 13 reflects this very well. The arrow plots seem to show no particular tendency to flow one way or another such that one can identify a structure. If one looks at the 3-chain in figure 13 one can see that it corresponds to a deformation in the secondary structure which turns out to be a bulge caused by the primary sequence not matching up on both sides of the helix. This causes two of the residues to flip to the outside of the helix. The 2-chains in figure 12 tell a similar story and also show some very clear tight turns in the tertiary structure.

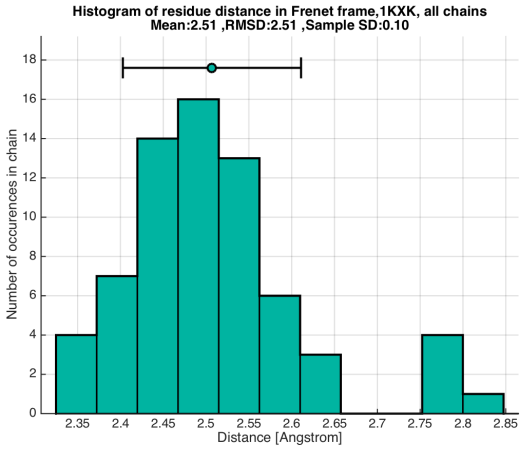

Figure 11: Histogram of the distance between the residues and their associated Frenet frames in the B-chain of the 1KXK RNA. The characteristic peak at 2.5Åis visible and the outliers are as expected around 2.8Å.

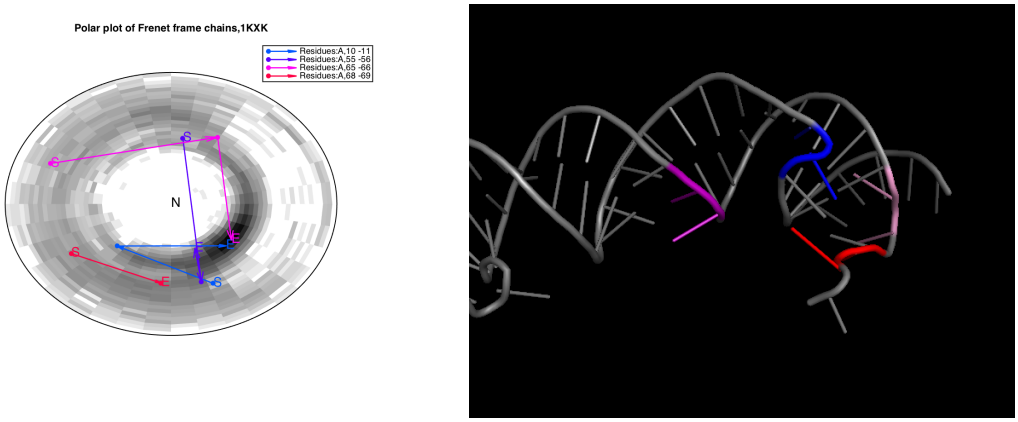

Figure 12: The left plot shows the 2-chains of residues which are not pointing into the hotspot. These chains seem to correspond to tight turns in the tertiary structure as can be seen in the right figure. The colors in the left and right pictures correspond to each other.

Figure 13: The left plot shows a 3-chain of residues which are not pointing into the hotspot. This chain corresponds to a short 2-residue bulge in the the secondary structure due to the bases not complementing each other perfectly. The colors in the left and right pictures correspond to each other.

5.4 1Y26

The 1Y26 RNA is a single-stranded RNA with a pseudo-knot-like structure. It's secondary structures and tertiary structures are very complex and is a prime example of why secondary structure classification of RNA is a hard problem. In the analysis, using the arrow plots, one can see that there are five 1-chains of residues and two 3-chains corresponding to certain parts of the RNA. These can be seen in figure 15 and figure 16. The 1-chains once again seem to contain some important features. The red part seems to be something similar to an undulation point. The blue segment and dark blue segments together seem to correspond to be a proper inflection point. In a similar manner, part of the blue 3-chain seems to have an undulation point as well where as the pink 3-chain has a very tight turn that seems to correspond to a hairpin loop. Thus it seems that undulation and inflection points also has some significance in this analysis.

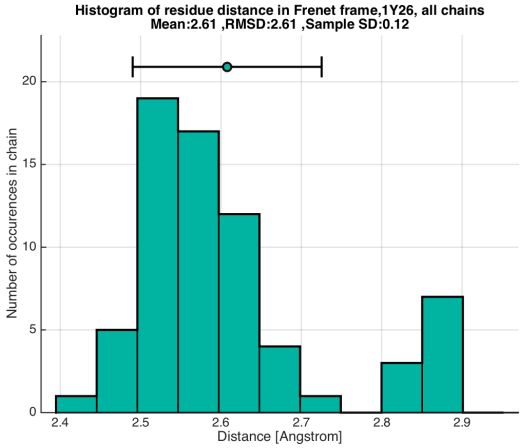

Figure 14: Histogram of the distance between the residues and their associated Frenet frames in the B-chain of the 1Y26 RNA. The characteristic peak at 2.5Åis visible and the outliers are as expected around 2.8Å.

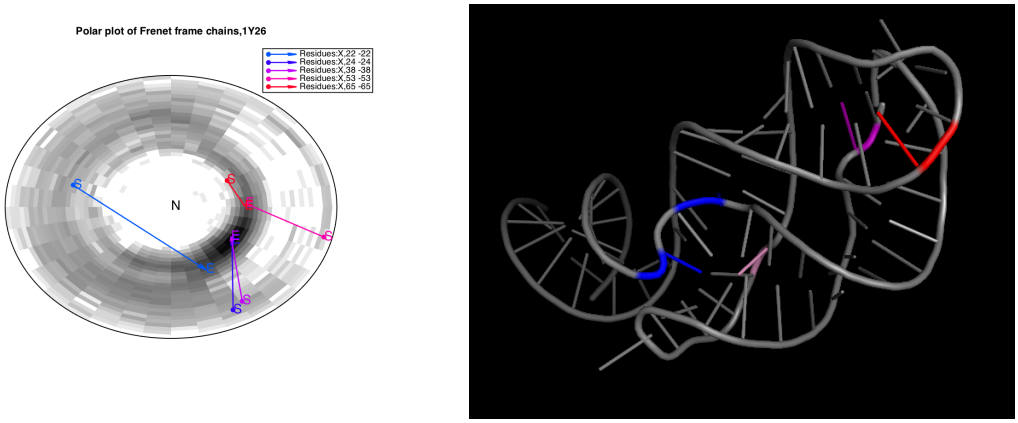

Figure 15: The left plot shows the 1-chains of residues which are not pointing into the hotspot. The red segment seems to be an undulation point. The blue and dark blue segments form an inflection point. The colors in the pictures correspond to each other.

Figure 16: The left plot shows 3-chains of residues which are not pointing into the hotspot. The blue 3-chain seems to correspond to a segment with an undulation point. The pink 3-chain however seems to be a tight turn in the RNA secondary structure. The colors in the left and right pictures correspond to each other.

5.5 1Z43

The 1Z43 RNA is a single-stranded RNA. It's a little bit peculiar in that its distance distribution, seen in figure 17, is a lot more smoothed out compared to the other examples. This might be an issue with the quality of the RNA file or its resolution. The structure of the RNA itself can be seen as two hairpin-loops that join together at their base. It's a very regular structure with a a few inflection points and bulges. The corresponding arrow-plots can be seen in figure 18. The 1-chains all seem to follow similar patterns in the arrow plot. On the right figure, the blue chain has been zoomed onto, and the corresponding Frenet frames have been made visible. The green line is the tangent vector, red is the normal vector and blue is the binormal vector in this picture. As one can see there is a 3-residue bulge (or hairpin loop). Due to how the Frenet frames are constructed, needing the previous ribose position and the next ribose position in the chain to determine the orientation of the current frame, the blue segment's associated frame gets a large influence from the tight turns flipping the direction of its normal vector, however curiously enough the almost symmetrical situation just a few residues earlier does not have this issue. The rest of the arrows in the plot correspond to similar structures that have already been discussed.

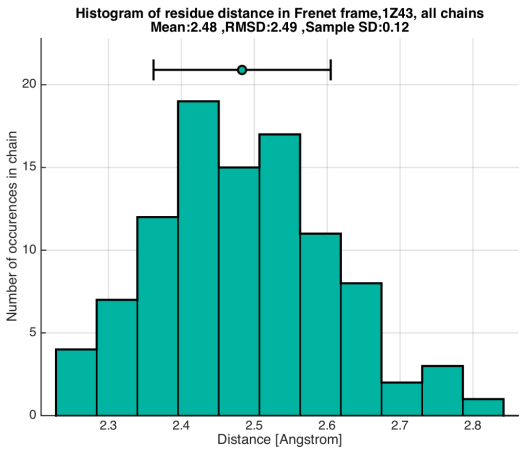

Figure 17: Histogram of the distance between the residues and their associated Frenet frames in the B-chain of the 1Z43 RNA. The characteristic peak at 2.5Åis visible however the distribution is much more smoothed out reaching all the way out to the outliers at 2.8Å.

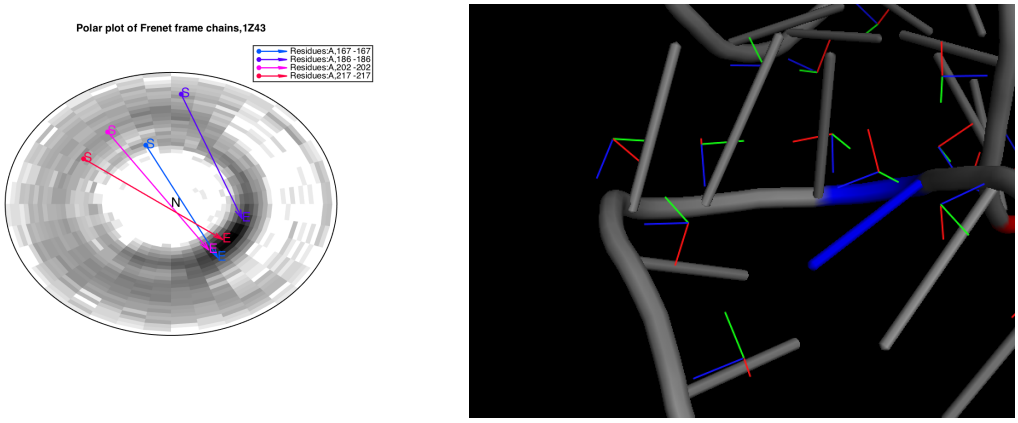

Figure 18: The left plot shows the 1-chains of residues which are not pointing into the hotspot. In the figure to the right the associated Frenet frames have been made visible alongside the RNA chain. The frenet basis vectors $(\mathbf{n}, \mathbf{b}, \mathbf{t})$ correspond to the red, blue and green axes respectively. Due to how the Frenet frame is constructed the blue segment does not exactly correspond to the position of the loop.

5.6 3SUH

The RNA 3SUH is a single-stranded RNA. This is just another example of the small structures that have already been mentioned. Looking at figure 20 and figure 21 one can once again see the trend of the chains moving clockwise around the northpole. One can also identify the tight turns which by now are pretty common as well as an inflection point.

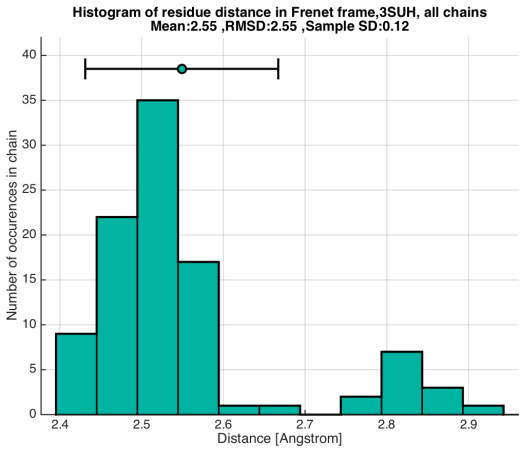

Figure 19: Histogram of the distance between the residues and their associated Frenet frames in the B-chain of the 3SUH RNA. The characteristic peak at 2.5Åis visible and the outliers are as expected around 2.8Å.

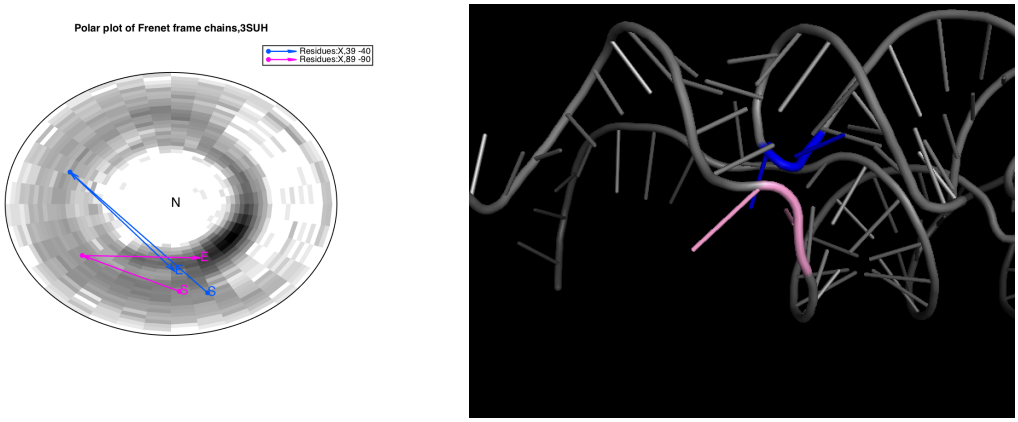

Figure 20: The left plot shows the 2-chains of residues which are not pointing into the hotspot. These chains seem to correspond to tight turns in the secondary structure as can be seen in the right figure. The colors in the left and right pictures correspond to each other.

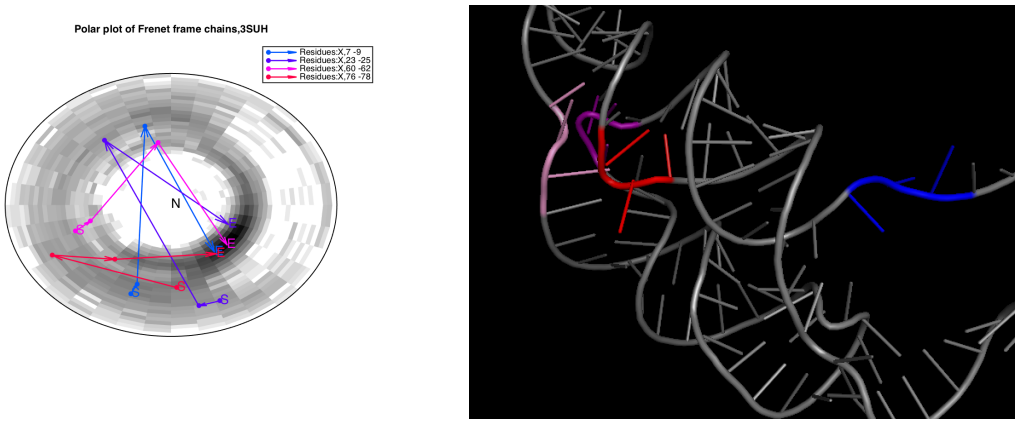

Figure 21: The left plot shows a 3-chain of residues which are not pointing into the hotspot. This chain corresponds to a short 2-residue bulge in the the secondary structure due to the bases not complementing each other perfectly. The colors in the left and right pictures correspond to each other.

There are many other examples one could show, such as the larger RNA 1DDY, 1FG0 and 3ZEX that has really long arrow-chains (up to 25 residues in a single chain) and plenty of smaller arrow-chains. Here, only RNA with less than 100 residues total have been discussed, which compared to 3ZEX which has 4200 residues, makes up quite a small amount of the total investigated data.

6 Discussion about improvements and paths forward

There is potential for further research on the contents of this thesis. First of all, one could check more correlations between data present in the project. This could be things such as the discrete bond angle (which is closely related to the curvature) and the corresponding torsion angle like has been done in [13]. There also has been little explanation for the existence of the second forbidden region so far which is a rather interesting phenomenon.

Second, there is one immediately obvious improvement that can be made. Instead of puncturing the sphere at the southpole and making a polar plot out of it, instead one should puncture it at the northpole and proceed in a similar manner as done in the previous sections. This has the advantage that the northpole is already quite isolated due to the large forbidden region in the tangential direction. All chains of residues will then be confined to the interior of the disk, and the issue of the boundary being traversible will be less significant. Figure 22 should make the idea clear:

Another huge problem is that lots of data has been generated for different RNA but we are also only able to check correlations between parts of that data. It would be ideal to be able to check what correlation the Frenet Frame construction has to specific structures in the primary structure, secondary structure, tertiary structure and so on. To do this one needs these structures explicitly. The primary structure can be easily found since it is specificied for every RNA in every PDB file as a SEQRES keyword. However the more interesting secondary structure is another story entirely. Now, in order to more closely relate specific RNA secondary structure to these findings, one has to have a way to classify those structures first. This is a rather arduous task to do by hand so some way of doing it automatically with a computer is needed. One option to attempt to classify the secondary structure would be to implement secondary structure prediction algorithms. The options range from base-pairing algorithms such as the Nussinov-algorithm (see [5]) to algorithms that use energy minimization methods. However even if one can predict a structure the problem would still remain to classify such structures such that they can be compared to our data. When one has the secondary structure predicted one could construct e.g. the dot-bracket notation RNA and compare patterns in that notation to the data here. An algorithm for the dot-bracket notation of single-stranded RNA would not be a major hurdle to implement with a multi-strand implementation requiring more work. The RCSB database contains 81 single-stranded RNA with good statistical basis (Search criteria: 1 chain in Asymmetric unit, chain length larger or equal to 100). These could provide a decent dataset to run such an analysis on. Matlab has a built-in function to generate dot-bracket sequences for a given primary sequence, however it only works for single-stranded RNA. An example of the dot-bracket notation is given by the 1CSL RNA in figure 8, which is double-stranded, but the pattern is easy to see, just match the parantheses to get base-pairs:

```
>strand_A
AACGGGCGCAGAA
.(((..((((((.
```

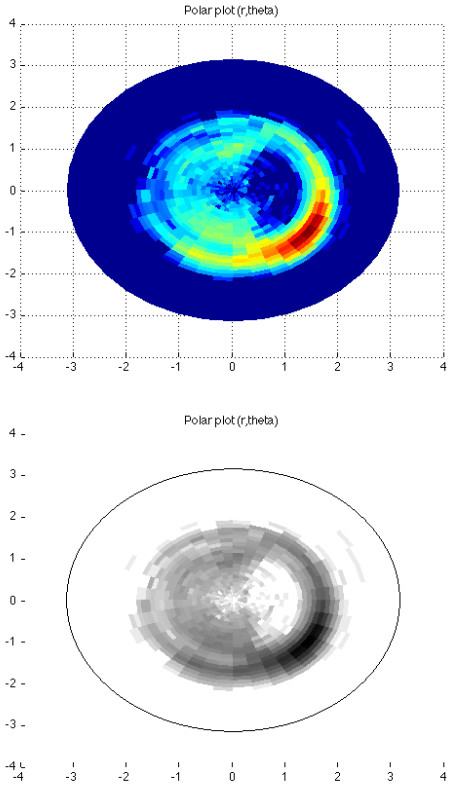

Figure 22: Polar plots corresponding to the spherical angular distribution. The southpole and northpole have switched roles compared to the earlier polar plot. Now the southpole is the center while the northpole is on the boundary. This has certain benefits in visualisation of the data.

```
>strand_B
UCUGACGGUACGUUU
))))..))...))))..
```

Using the dot-bracket notation, even though it is an approximation, one could possibly find basic secondary structures such as hairpin-loops and bulges. With a more advanced algorithm, perhaps even more structures could be found. At this point one could find secondary structures in dot-bracket notation and see if they correspond to specific patterns in the data set presented in this project. Note however, that depending on what algorithm is used to determine the base-pairing, one may get different results for the pairs. This could be an issue with this approach.

Another option would be to be to make use of a database for specific motifs, such as the K-turn database [3] , and perform a similar process. One could also try to find RNA which have an easily identifiable (to the eye) secondary structure, such as the structure of the 1EHZ RNA, and compare such structures to each other.

There's also an idea to combine arrow chains which are close together (such as a 1-chain and a 3-chain that only has single residue in between them) into a single long chain. This could be potentially lead to a better description in identifying structures.

The algorithms that have been made in this work do not take several poten-

tially helpful keywords in the PDB File Format into account, such as the LINK, CONECT and SEQADV keywords. They each have their own use, cited directly from the PDB File Format guide [2]:

"The LINK records specify connectivity between residues that is not implied by the primary structure. Connectivity is expressed in terms of the atom names. They also include the distance associated with the each linkage following the symmetry operations at the end of each record. This record supplements information given in CONECT records and is provided here for convenience in searching."

"The CONECT records specify connectivity between atoms for which coordinates are supplied. The connectivity is described using the atom serial number as shown in the entry. CONECT records are mandatory for HET groups (excluding water) and for other bonds not specified in the standard residue connectivity table. These records are generated automatically."

"The SEQADV record identifies differences between sequence information in the SEQRES records of the PDB entry and the sequence database entry given in DBREF. Please note that these records were designed to identify differences and not errors. No assumption is made as to which database contains the correct data. A comment explaining any engineered differences in the sequence between the PDB and the sequence database may also be included here."

Since the secondary structure section of the PDB File Format contains nothing relevant for RNA, the problem of getting a good dataset for specific secondary structures remains an issue. Moreover, arrow-plots and similar plots have been generated for each and every one of the 460 pdb files, but such information can't all be shown easily in a paper like this. Some of them are rather large as well, such as 3ZEX which is the largest of them all at 4200 residues and chains that go up to 25 residues long. The arrow-plots are a visual aid but the actual geometric content is not easily parsed into a quantitative measure. In other words, the arrow-plots are a qualitative and visual measure but at the level of their current implementation they can't be easily used to make a quantitative measure.

There is also the possibility of extending the Frenet frame approach. Quaternions have long been known to be able to describe 3-dimensional rotations in a beautiful way. They're often used for interpolations between two different orientations of a frame in graphical applications but there may also be symmetries to find in that approach. A quaternion Frenet frame is used in [11] with some success. Will it work with the discrete Frenet frame approach as well?

Last but not least, instead of having a Java program and a Matlab program (as well as some Python, in fact) requiring user interaction in between, all the code should be ported over to a single specific language. The less user interaction needed the less room there is for error in that regard.

7 Conclusions

A few results have been shown by applying the discrete Frenet frame structure to about 460 different RNA. Out of these results the most interesting is perhaps the spherical distribution of the residues along the Frenet frame chain. The most interesting facts about the distribution is the presence of the single high density region as well as the presence of the two holes in the distribution. One of these holes seem to merely be "caused" by the construction of the Frenet frame itself since the hole is located in a region around the tangential axis of the frames. This naturally excludes any residue from pointing in that region since it corresponds to the location of the next Frenet frame which also corresponds to the position of the next nucleotide.

The second hole does not seem to admit a similar explanation. This is due to how the rest of the distribution looks. The rest of the distribution seems to have

residues scattered fairly uniformly around the sphere if one excludes the holes and the high density region. Because of this it would, on an intuitive level, be strange if there was such a symmetry in the distribution such that it produces the second hole. The explicit cause for the hole itself hasn't been examined but it seems to be caused by a physical process rather than being caused by the mathematical construction.

As for the high density region itself there is still the curious fact that it contains about 75% of the residues in the RNA investigated. This figure closely resembles the 75% of residues that lie within the 2.4Å-2.6Å region of the plot regarding the distance between residues and their associated Frenet frames. About 87% of the residues in the high density region lies in the 2.4Å-2.6Å region while only 1% of them lie in the 2.75Å-2.85Å region. The residue distance plot (see figure 4) had two distinct peaks visible and it is clear that the larger peak receives a significant contribution from the residues in the high density region of the spherical distribution. The exact cause for the smaller peak isn't as clear however and it may be signs of some other common structure in the RNA.

Attempting to visualize the flow of residues as they move around outside the high density region provided a few results that seem to point towards smaller irregularities along the RNA. These irregularities comprise things such as undulation/inflection points, small bulges and hairpin loops. The hunt for specific patterns that could represent certain secondary/tertiary structures remains an open question. This is largely because of the fact that this process so far has been done by hand which isn't a feasible process. Moreover it isn't a quantitative way of finding the pattern which is a big disadvantage.

There are a lot of directions one could take from this point, such as investigating other correlations such as between bond/torsion angles and the spherical distribution or trying to find theoretical grounds to why there is a forbidden region in the distribution.

8 Swedish Summary

DNA, ett begrepp som många känner igen, är ofta sedd som ritningarnaför varje organism oavsett om de är stora eller små. Det lärs ofta ut som en av de viktigaste punkterna i normal skolgång. I skuggan har vi däremot också andra stora medspelare inom cellen såsom proteiner och RNA vars funktioner och mekanik är minst lika viktiga och lika, om inte mer, intressanta än de motsvarande frågorna kring DNA. Av speciellt intresse är strukturen hos RNA och protein. Om man jämför med DNA som många känner igen som det populära dubbelhelixmotivet, som uppstår på grund av baspar i DNA-sekvensen, så har RNA/protein liknande strukturer men också väldigt många annorlunda strukturer därutöver. Exempelvis så kan RNA också forma baspar i sin sekvens och forma tre olika typer av helix men utöver dessa kan RNA också forma loops, pseudoknots, bulges och mycket annat. Dessa kan också förekomma flera gånger och i olika kombinationer i en och samma RNA sträng. RNA, till skillnad från DNA, kan bestå av bara en enda sträng där baser paras ihop med andra baser i strängen men det finns också möjlighet att man har flera strängar som kopplar till varandra och sig själva. Dessa kopplingar ger upphov till tredimensionella komplexa former som vid första anblick för oss kanske mest liknar hårbollar. På cellnivå så kan däremot varje liten skillnad i form hos ett RNA eller protein göra stor skillnad för dess funktion i cellen. Skillnad i form kan göra så stor skillnad att t.ex. ett protein kan vara ett enzym för reaktioner i kroppen medan ett annat kan användas för att bygga muskler eller fungera som en budbärare mellan celler.

Det kanske inte är så konstigt att man söker efter en djupare förståelse för hur dessa RNA och proteiner bildas i cellen eftersom de spelar en så stor roll

för våra biologiska processer. Speciellt så vill man förstå hur de viker sig till de former de antar och processerna bakom dessa vikningar. Många sjukdomar såsom Alzheimers och Parkinsons beror på att protein viks fel i kroppen så att kroppens funktioner störs. Om man förstår hur proteiner viks så kan man kanske till och med använda processen för medicinska syften så att man kan tillverka egna protein eller RNA som motverkar eller till och med botar vissa sådana sjukdomar. Att kunna utföra beräkningar och förutse dessa vikningsprocesser är däremot ett mycket svårt problem.

I denna artikel undersöker vi därför RNA från ett nytt perspektiv. Det finns många sätt man kan undersöka RNA till exempel från kemisk eller biologisk synpunkt men man kan också bortse från den fysiska världen och koncentrera sig enbart på geometrin hos RNA. Om man tittar på ett RNA så ser ut som en sträng med hakar som pekar ut längs med hela strängen. Det finns ett mycket naturligt och rent geometriskt sätt att beskriva formen hos denna sträng. Tänk er att ni befinner er i innerstaden och någon ber om en vägbeskrivning till torget. Hur skulle man kunna göra detta? Ett tillvägagångssätt är att slå upp kartan, visa dem var de befinner sig, visa var torget ligger och säga "gå 300 meter norr och 100m väst" vilket är korrekt men kanske inte alltid är det mest intuitiva sättet att beskriva vägen. Vanligare är kanske att säga "följ denna vägen i den riktningen och sväng sen höger vid tredje korsningen" det vill säga att vi förklarar vägen utifrån personens eget perspektiv när hon går längs vägen. På liknande sätt kan vi beskriva en sträng RNA. Istället för att tänka oss att vi befinner oss utanför RNA:t (att vi kollar på kartan) och beskriver dess form så tänker vi oss istället att vi befinner oss på RNA:t och går mellan dess olika punkter, kollar omkring oss, och ser hur det är utformat. På så vis kan vi enkelt ställa oss frågor såsom "vilket håll är basen riktad åt vid varje punkt på RNA:t?", "Hur långt är det mellan denna punkten och den här punkten om jag ska gå dit längs med strängen?" och finna svar.

Vad vi hittar är att RNA verkar ha mer frihet än vad protein har och dess understrukturer är inte så enkla att identifiera. Vi kan därför inte enkelt identifiera olika sekundärstrukturer i hos RNA såsom man har gjort för protein i [13]. Däremot så finns det fortfarande vissa intressanta mönster som är speciellt uppenbara när man kollar på de sfäriska fördelningarna av baserna längs RNA. Det visar sig finnas en koncentration av baser i en viss riktning, två "förbjudna" områden där inga baser pekar och sen en minimal fördelning av baserna på resten av sfären. Det ena förbjudna området kan enkelt förklaras med att eftersom det området motsvarar i vilken riktning nästa del av RNA:t ligger så finns det inte rum för basen att peka åt det hållet. Det andra förbjudna området har däremot inte en lika uppenbar förklaring och kan kräva mer undersökning före man kan dra en slutsats om det.

I denna rapport så ges resultaten av en liten undersökning inom ett annars mycket stort forskningsområde. Det finns många olika sätt att försöka undersöka och lösa vikningsfrågan och vissa metoder kanske är mer framgångsrika än andra. I vilket fall som helst så är det en mycket förundrande fråga, speciellt eftersom de här processerna sker inom allt levande, till och med hela tiden inom vår egen kropp, och svaret på denna fråga kanske dröjer länge ännu.

References

[1] G.P.Moss et al. "Basic Terminology of Stereochemistry". In: *Pure Appl. Chem.* 68.12 (1996), pp. 2193–2222.

[2] J.Callaway et al. *Worldwide PDB Protein Data Bank, PDB File Format documentation.* [Online; accessed 1-March-2015]. 2012. URL: `http://www.wwpdb.org/documentation/file-format.php`.

[3] K.T.Schroeder et al. “A structural database for k-turn motifs in RNA”. In: *Pub. Med.* (2010). DOI: 10.1261/rna.2207910.

[4] M.Antczak et al. “RNApdbee—a webserver to derive secondary structures from pdb files of knotted and unknotted RNAs”. In: *Nucleic Acids Research* 42.W1 (2014), W368–W372. DOI: 10.1093/nar/gku330. URL: http://nar.oxfordjournals.org/content/42/W1/W368.abstract.

[5] R.Nussinov et al. “Algorithms for Loop Matchings”. In: *SIAM Journal on Applied Mathematics* 35.1 (1978), pp. 68–82. DOI: 10.1137/0135006. URL: http://dx.doi.org/10.1137/0135006.

[6] Helen M. Berman et al. “The Protein Data Bank”. In: *Nucleic Acids Research* 28.1 (2000), pp. 235–242. DOI: 10.1093/nar/28.1.235. eprint: http://nar.oxfordjournals.org/content/28/1/235.full.pdf+html.

[7] Richard L. Bishop. “There is More than One Way to Frame a Curve”. English. In: *The American Mathematical Monthly* 82.3 (1975), pp. 246–251. ISSN: 00029890. URL: http://www.jstor.org/stable/2319846.

[8] Samuel E. Butcher and Anna M. Pyle. “The Molecular Interactions That Stabilize RNA Tertiary Structure: RNA Motifs, Patterns, and Networks”. In: *Accounts of Chemical Research* 44.12 (2011). PMID: 21899297, pp. 1302–1311. DOI: 10.1021/ar200098t.

[9] David Elliot and Michael Ladomery. *Molecular Biology of RNA*. 1st ed. Oxford University Press, 2011.

[10] Mary Gray. *Modern Differential Geometry of Curves and Surfaces with Mathematica, Second Edition*. Textbooks in Mathematics. Taylor & Francis, 1997. Chap. 10. ISBN: 9780849371646.

[11] A.J. Hanson and Hui Ma. “Quaternion frame approach to streamline visualization”. In: *Visualization and Computer Graphics, IEEE Transactions on* 1.2 (1995), pp. 164–174. ISSN: 1077-2626. DOI: 10.1109/2945.468403.

[12] Shuangwei Hu, Martin Lundgren, and Antti J. Niemi. “Discrete Frenet frame, inflection point solitons, and curve visualization with applications to folded proteins”. In: *Phys. Rev. E* 83 (6 2011), p. 061908. DOI: 10.1103/PhysRevE.83.061908.

[13] Martin Lundgren, Antti J. Niemi, and Fan Sha. “Protein loops, solitons, and side-chain visualization with applications to the left-handed helix region”. In: *Phys. Rev. E* 85 (6 2012), p. 061909. DOI: 10.1103/PhysRevE.85.061909.

[14] P.Grinfield. *Introduction to Tensor Analysis and the Calculus of Moving Surfaces*. Springer, 2013. Chap. 13, pp. 220–227. DOI: 10.1007/978-1-4614-7867-6.