2015

A Volumetric Survey of Focusing and Electrostatic Contributions to Protein RNA Interactions

Devan L. Bicher
Lehigh University

A Volumetric Survey of Focusing and Electrostatic Contributions to Protein RNA Interactions

by

Devan Lester Bicher

A Thesis
Presented to the Graduate Committee
of Lehigh University
in Candidacy for the Degree of
Master of Science
in
Computer Science

Lehigh University
December 2015

Approved and recommended for acceptance as a dissertation in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Devan Lester Bicher
A Volumetric Survey of Focusing and Electrostatic Contributions to Protein RNA Interactions

---

Date

(Committee Chair), Dissertation Director, Chair
(Must Sign with Blue Ink)

Accepted Date

Committee Members

(Committee Member)

(Committee Member)

(Committee Member)

(Committee Member)

Contents

List of Figures

Abstract

Electrostatic focusing has been shown to be a major component in DNA to protein interfaces. Since ribonucleic acid molecules are notoriously less consistent and uniform in structure than their DNA cousins, electrostatic focusing has not yet been evaluated between RNA and protein. This is the first survey of its kind showing the occurrence of focusing with specific amino acids and RNA secondary structures. With unique volumetric tools we quantify the electrostatic focusing phenomena in RNA and look at its impact on the electrostatic component of protein affinity. With those same tools we uncover which amino acids interact with focusing regions and how much they interface with those regions. We also demonstrate what types of secondary structures in RNA are commonly associated with electrostatic focusing.

Chapter 1

Introduction

Ribonucleic Acids (RNA) play a crucial role in molecular biology. It acts as a messenger, delivering important information from DNA in the genetic warehouse of the nucleus. It guides translation in close concert with numerous protein subunits. It contributes to gene expression, and in certain microorganisms it can even be the storage structure of the genetic material itself. RNA is becoming an increasingly popular topic of study. As technology progresses and the mechanisms with which we study biology grow in size and intricacy the mysteries of how RNA guides biology are being revealed. RNA has become a target for therapeutics, and overtime many synthesized drugs have been RNA based. However RNA is still an elusive molecule, much more so than its equally important DNA and protein counterparts.

Unlike its typically double-helical relative, DNA, RNA rarely forms such uniform macromolecules. It is found throughout both the cellular cytoplasm and the nucleus, it can also be either single or double stranded. It can form the familiar double helix, or into irregular, RNA specific secondary structures like stems, loops, and hairpins. This variability makes it much harder to understand and predict modes of interactions with proteins, especially in a structural context [9]. RNA and proteins are in constant motion together in various roles, performing a complicated dance throughout the cell that facilitates life at the cellular level. How these two diverse molecules find, attract, and interact with one another in the chaotic molecular soup of the cell is of great importance for furthering our understanding of biology and its

microscopic mechanisms.

One important factor in DNA protein interaction has yet to be fully illustrated in the more diverse set of RNA protein interfaces. This phenomenon is called **electrostatic focusing** and it has been well documented in DNA protein interactions [10]. Electrostatic focusing works by excluding water from small regions on the surface of molecules. Water, the medium in which all biological life exists and functions, is dielectric, and is constantly shielding charges on molecular surfaces. More specifically the positive side of water is attracted to negative surface regions on DNA or RNA thereby diminishing the magnitude of potential charge than can attract other positively charged molecules that would bind it, like proteins. Electrostatic focusing occurs in small clefts or narrow pockets on the surface of RNA, that are too tight for water to consistently squeeze into. Without water present to dull the magnitude of charge in that region the charge is enhanced, or focused into the surrounding area and it can draw in other positively charged molecules that may bind with it[2].

Since electrostatic focusing has been shown to play an important role in DNA protein contact[10] it is natural to wonder if RNA exhibits similar characteristics; it is in the shape of the cavity region that precludes the presence of focusing, and many RNA structures contain focusing regions. Because the structure of RNA is more irregular than DNA the methods that were applied to DNA depend on it's helical structure, comparable studies in RNA have been more difficult.

In our recent work we have developed unique tools to study the contribution of electrostatic focusing. We try to do this volumetrically, using computational tools to quantify focusing and electrostatics by measuring the volume of space they would take up if present in all directions they exhibits influence. The advantage of volumetric methods is that they can be applied to any structure, sidestepping earlier limitations. Such flexibility allows us to analyze where focusing might exhibit influence in RNA and the adjacent protein based purely on its shape and charge, rather than on sequence or prior knowledge, which do not necessarily apply across the diverse array of RNA structures.

This thesis describes a novel volumetric study of RNA and electrostatic focusing. This work is significant and contributes the first general volumetric survey, to our

knowledge, of the role of RNA shape, charge and secondary structure on protein RNA recognition. We demonstrate a wide breadth of ways in which focusing influences both the RNA that generates it and the protein that interfaces with it. We start by looking at how focusing relates to the electrostatic determinant of molecular affinity. We also study how focusing interacts with amino acids to gain a better understanding of how it affects the building blocks that define proteins and how they function. Lastly, by looking at which secondary structures typically surround focusing regions we can see fundamental patterns that are capable of generating this phenomenon.

Having a better grasp on how RNA and proteins comingle is vitally important to furthering the understanding of microscopic biological mechanisms. With a stronger sense of how RNA not only binds, but also attracts other molecules will aid in drug and therapeutic design strategies as well as disease and general healthcare diagnostics. RNA is a staggeringly complex molecule and it is only by studying its intricate mechanisms that we can hope to gain a better insight into how this structure contributes to the essential functions that make up all life.

Chapter 2

Methods

2.1 Dataset Composition

We began this survey of RNA protein complexes by running some trial experiments on a small set of RNA recognition motif structures with several other representative protein RNA pairs. The initial trials helped guide and refine the experimental methodology in our approach to better understand the complexity of volumetrically based protein RNA electrostatic interactions.

We constructed our dataset to be a nonredundant representation of the protein-RNA complexes available for use in the Protein Data Bank[1]. To construct our dataset we started with any PDB file that explicitly contains both RNA and protein. Structures from our first dataset along with any structure that contained the RNA recognition motif were removed. We parsed the files and removed any that also contained DNA. Our selection of structures with RNA and protein did not eliminate files with RNA and DNA, since many proteins are capable of binding both. This selection criteria eliminates any biased caused by DNA.

We removed structures that contained any RNA or protein component of the ribosome. Many ribosome constituents were too large for our electrostatic analysis software, and we did not want to misrepresent the RNA protein interfaces from the ribosome in our survey.

We also removed all mutants by performing a text search in the PDB files metadata for the term: mutant. In doing this we removed structures that were highly similar to one another, but varied in that they were simple mutations of other structures. Mutant here means that only a few either nucleotides or amino acids in the RNA or protein, respectively, are changed from their wild-type conformation. This selection reduces bias from natural systems.

Finally we removed members with protein sequence similarity above 95%. We found this similarity through use of the sequence alignment and clustering software package clustalw [7, 6]. ClustalW took the proteins sequences of our dataset as input and clustered everything together that had 95% sequence similarity or greater. With all of the structures clustered together we chose one representative structure from each cluster arbitrarily.

Here we reintroduced the original dataset from the first round of experiments, which was about 150 in total. The final dataset contained 355 structures. Some additional structures were removed out of necessity for different reasons: RNA shorter than 25 nucleotides were discarded. Very large RNA chains were also occasionally reduced to chains long enough to completely describe their interface with a given protein, with generous margins. Structures were removed when our analysis software could not process their expansive size, or because the RNA and protein did not make any contact.

2.2 File Preparation

Several cleaning and refining steps were performed on the original files before we could start to collect the data. There are often inconsistencies within the files that must be resolved in order for the operations to run properly. The files were stripped of all header information, comments, remarks, and any extraneous molecules like surface bound water or ions. These molecules are often captured during the crystallization process and remain in the files, but are irrelevant to our experiment. Duplicate overlapping models in the downloaded structures were also removed in

favor of the primary atomic positions. Next, we removed multiple subunits, figure 2.1 A, from each file to isolate only one RNA-protein pair, figure 2.1 B.

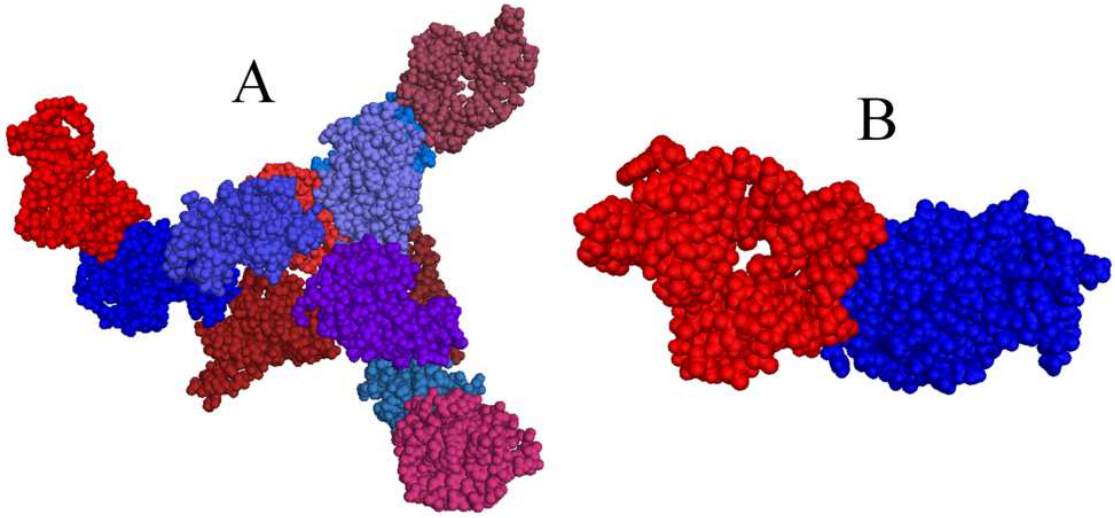

Figure 2.1: Repeating Subunits in One PDB File

Atomic structure is determined through the smallest repeating subunit(A). This file has 6 repeating pairs of RNA(red) & protein(blue). We are concerned only with the single pair (B), located in far left of A).

In cases with multiple subunits where one pair was slightly larger then the rest we always picked the set that had the largest RNA-protein pair. During this subunit elimination process we used visual inspection to remove disjointed RNA chains. Occasionally there were very small sections of discontinuous RNA that were removed; we are only concerned with the interface for whole RNA pieces bound to protein. In few cases the RNA bound to the protein was in two or more large portions, there are many cases of proteins binding more than one chunk of RNA simultaneously. In these scenarios again we choose to keep only the largest chunk of RNA in favor of removing the structure from the dataset altogether.

To perform our survey the RNA and protein had to be separated from one another. We separated out these two molecules from the files using their amino acid designation, whereby the nucleotides in the RNA portion of the files were indicated by single letter nucleotide designation rather than the three letter protein

nomenclature. We also had to ensure that all molecules were properly hydrogenated. Not all PDB files have the hydrogens appropriately placed; occasionally the positions are implied, e.g. off carbon on the protein backbone. Sometimes hydrogens were missing as a result of our cleaning methods, e.g. separating subunits or RNA. Some software like Delphi cannot produce reliable results in cases where hydrogens were missing. The hydrogens for all of the files were systematically removed and then added back on for consistency and completeness[5].

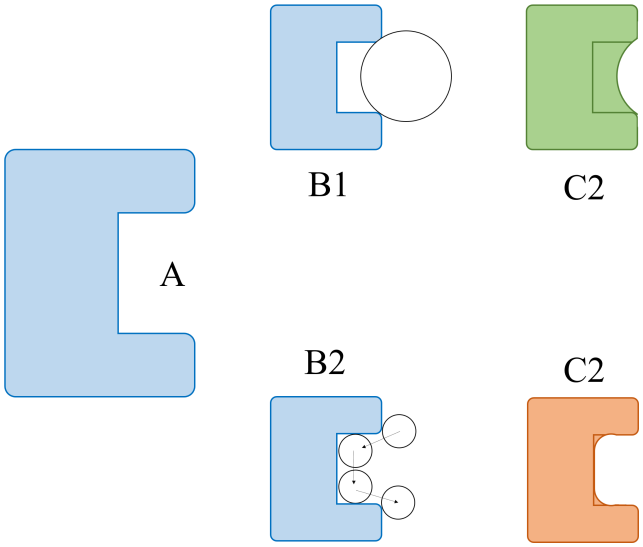

Figure 2.2: Creation of Molecular and Envelope Surfaces

Input molecule (A), the envelope surface (C1) is generated by rolling a 4 Å radius probe (B1 circle) around the atomic structure. The molecular surface (C2), generated from A by probes (B2 circles).

With RNA and proteins separated, we created molecular surfaces, figure 2.2(C2), and envelope surfaces, figure 2.2(C1) for each component. A command in the surfaceExtractor [4] volumetric molecular analysis software package computes the surface based on the size of a probe radius: a hypothetical ball that roles around the

3-D atomic arrangement, figure 2.2 B1 and B2. The radius is a parameter at runtime: 1.5 Å for the molecular surface to mimic a water molecule, balls in 2.2 B2, and 4 Å for the envelope surface, the ball in figure 2.2 B1. The molecular surface is the physical boundary around the outer edge of the molecule formed by its constituent atoms, and it is generally considered the region accessible by the water that surrounds all biological molecules. The envelope surface is larger than the molecular and presents fewer details of the nuances of its atomic surface; it is the surface immediately accessible to other macromolecules.

2.3 Electrostatic Isopotential Contours

To create the isopotentials we start with the program Delphi[8]. Delphi creates a large 3-D grid of electrostatic potential quantified points around the RNA or protein. Delphi is run twice; one run creates a uniform grid representing the raw potential values of the electrostatic charges on the atoms of the structure. The other run creates a non-uniform grid which represents a closer simulation to what might be present in biological systems, in the presence of non-uniformly charged water. With the Delphi grids created, surfaceExtractor[4] is run to create a surface, an isopotential contour, around the structure based on the values in the non-uniform Delphi grid. Figure 2.4, shows hypothetical red (B1) negative levels around RNA, and blue (B2) positive levels around protein. This contour represents the surface within which contains a charge greater than the input charge value, or less than depending on whether a positive or negative charge is chosen, respectively.

The charge value is one of the input parameters to surfaceExtractor, as well as the electronegativity: positive for protein and negative for RNA. The different choices of electrostatic levels were (+/-)[1.25, 2.5, 5.0, and 7.5] (kT/e). Where k is boltzmans constant, T is temperature, and e is the charge on an electron. These values were selected to survey a range of potentials influential in biological systems. Potentials are present around charged atoms of molecules and exhibit the strongest effect nearest the atoms that produce the charge. The charge diminishes the greater

the distance from the source charge and approaches zero as the distance approaches infinity.

For the choice of electrostatic levels, much smaller than 1.25 (kT/e) and the isopotential surface is unrealistically far away from the surface, it would not influence binding significantly in a real setting. If the values were much larger than 7.5 (kT/e) the opposite is true: the potential contour is too close to the molecular surface, sometimes even being inside of it, and is unrealistic for approximating any sort of actual binding influence. A smaller potential level, i.e. 1.25, results in a larger contour or surface around the molecule while a larger potential, i.e. 7.50, results in a smaller surface; smaller surfaces with a smaller volume, see figure 2.4.

2.4 CSG Operations

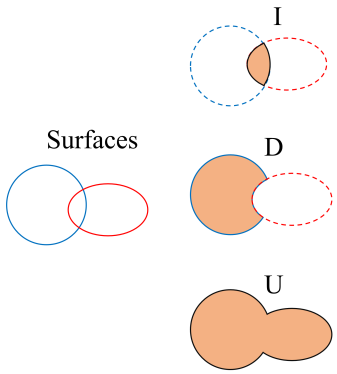

Figure 2.3: VASP CSG Operations

From 3-D Input Surfaces, blue & red on the left, VASP[4] can calculate the following CSG operations: (I) the Intersection between the two surfaces,(D) the Difference of one surface subtracted from the other, & (U) the Union of the two surfaces.

VASP, Volumetric Analysis of Surface Properties, is designed to perform volumetric set operations by comparing regions in space using Constructive Solid Geometry (CSG)[4]. These operations are used throughout the rest of our methods.

The functions operate on the volumes that we have created for experimentation, like the isopotentials or molecular surface, and they include intersections, unions, and differences, figure 2.3. Intersections compute the volume of the overlapping region shared between the input surfaces in 3-D space. Difference subtracts the overlapping shared volume of the second surface from the first. Union integrates the two surfaces together and produces a new surface that is a combination of the two.

2.5 Isopotential Intersections

We used the volume of intersection between isopotential contours as a statistic in our survey. We call this quantity electrostatic **complementarity**, figure 2.4(D), and it is a proxy for binding affinity. Complementarity is the volumetric CSG intersection, figure 2.3 (I), between positive isopotential on the protein, figure 2.4(B2), and negative on the RNA, figure 2.4(B1). More specifically it is the overlapping region of the volume of electrostatic potential at one level, either 1.25, 2.5, 5.0, or 7.5 (kT/e), with the oppositely charged contour at the same level, e.g. negative potential level of 2.5 around the RNA with positive potential also at 2.5 around the protein, figure 2.4(C). This represents how much the surface charges of the molecules interact at the four different potential levels we chose, acting as an adequate proxy for the electrostatic component of binding affinity[3]. This complementarity metric was part of the motivation behind choosing different potential levels because it shows the interplay between different electrostatic quantities and in a real setting isopotentials have influence across a wide array of potential levels.

2.6 Focusing Contours and Degree of Focusing

We created the focusing contours using the Delphi grids[8] by running surfaceExtractor again[2]. This time it computes the difference, see figure 2.3, between the uniform and non-uniform electrostatic fields around the given RNA surface to create the focusing contours for a specific value. This mimics the real life focusing effect by

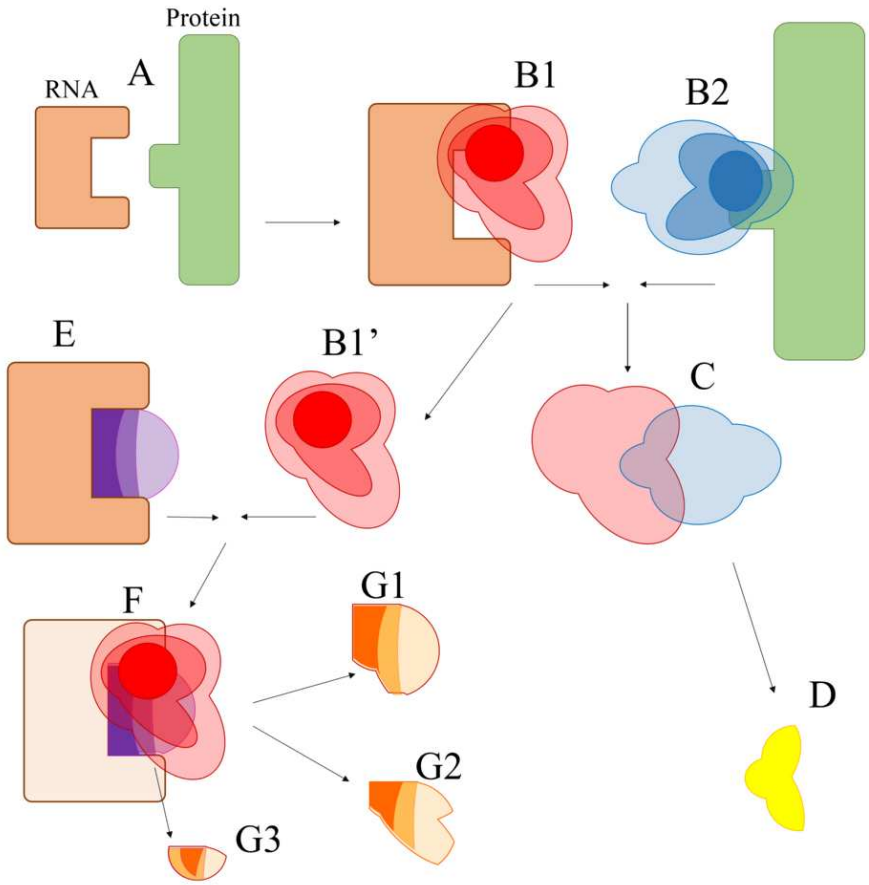

Figure 2.4: Generating Complementarity and Limited Focusing

Given (A) potential contours are generated around RNA(B1) & protein(B2). They are intersected(C) to generate the volume for quantifying complementarity(D). The RNA potential(B1') & the focusing region, purple portion of (E), are intersected (F) to form the volumes(G1, G2, G3) that quantify limited focusing.

reporting the difference between the surface charges with and without the presence of water; focusing exists because of the limited presence of water in small or narrow regions. The values represent the various levels of focusing intensity, figure 2.4 the purple portion of(E); the values we chose were 0.65, 1.25, and 2.50(kT/e) and they simply represent the difference between the two potential grids. The selected focusing values were part of a larger set of interval data ran during our trial experiments,

but the larger values were either empty or low enough to be meaningless[2]. Similar to the isopotential values, a lower focusing level results in a larger volume. However, unlike potential the effect is because smaller magnitude differences can occur farther away from the surface where water is starting to dull the focusing effect, but the greatest magnitude of differences occur closer to the surface where the charge, and therefore focusing, is more concentrated.

2.7 Limited Focusing

After obtaining the focusing contours we used CSG to intersect those contours at each of the three levels(0.65, 1.25, and 2.50 kT/e) with each of the four different isopotential volumes for the same RNA structure, creating 12 different sets of values, which we compare against complementarity at the respective potential level 1.25, 2.5, 5.0, or 7.5 (kT/e). These results are in section 3.2. We called this potential-focusing intersection **limited focusing**, figure 2.4 (G1, G2, G3). Since the focusing contour was limited to within the isopotential contour. It was calculated in favor of using the raw focusing contour to get a better sense of the degree of focusing within the specific area where electrostatic potential has it most pronounced effect since focusing typically only occurs at places of strong potential.

2.8 Normalized Focusing

Additionally we calculated the volume of this limited focusing region divided by the volume of the respective RNA molecule, calculated from the RNA molecular surface. We call the quantity **normalized Focusing** and was calculated because the focusing volume generally scales with the size of the RNA molecule around which it was calculated, i.e. larger RNAs result in larger focusing regions. We want to evaluate how much focusing there is without letting the size of the RNA skew the results. Without presenting this it could be speculated that correlation between complementarity and limited focusing is merely a result of increasing RNA volumes.

We present these normalized intersections versus complementarity in results section 3.3.

2.9 Chunky Focusing Regions

We used one more methodology to evaluate focusing, we called this representation of focusing the **chunky focusing region**. The process for creating chunky focusing regions, figure 2.5, is slightly more detailed than making the other focusing regions, figure 2.4. The process involved creating two protection sets, a set of small radius circles placed around the surface of an input region, green circles in 2.5(C). For the circles to be placed, the closest distance, red lines in 2.5(B), is measured from the focusing surface to the molecular surface, 2.5 down arrows, and the molecular surface to the focusing surface, 2.5 up arrows. If the closest point on the other surface is further than 2 Å the circles are placed at each point, green circles in 2.5(C). SurfaceExtractor executes this process in two steps in which the target surface is switched from the molecular surface, 2.5 organge lines, to the focusing surface, 2.5 purple lines.

The CSG union, figure 2.3(U), of these two set of circles is computed with VASP[4], then turned into a surface with surfaceExtractor, 2.5(D). A CSG difference, figure 2.3(D), of this protection set minus the molecular surface is calculated, 2.5(E), so it does not include anything inside the surface of the molecule. Similarly this volume is intersected with the original focusing contour, 2.5(F), to eliminate anything outside the focusing surface, and beyond the influence of this focusing level.

With that final intersection from 2.5(E) the chunky focusing region has been created, blue area in 2.5(F). For one of our last experiments we separated the different chunks of focusing contained within one file with surfaceExtractor. This step separates one large focusing contour into distinct, non-connected focusing chunks. The blue piece in figure 2.5(F) and the cyan volume in 3.1(B, C, and D) are both examples of a single focusing chunk. This chunky focusing region is a more precise

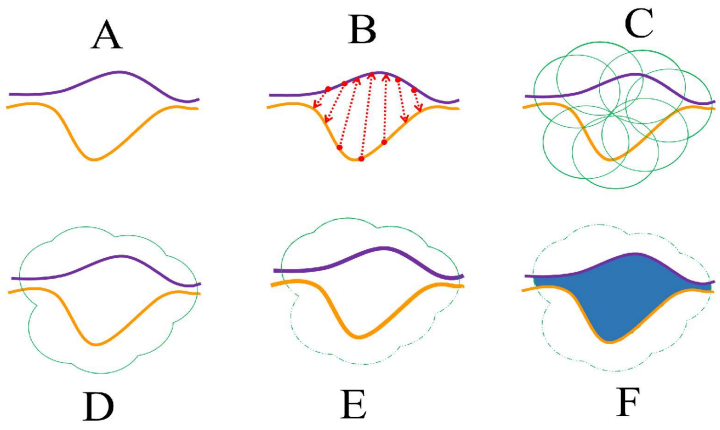

Figure 2.5: Steps for Generating A Chunky Focusing Region

Given focusing surface, purple (A), and molecular surface, orange (A), the closest points from one surface to another, red lines in (B), determine if probe spheres are placed on the respective surfaces, green circles in (C). The union of the spheres (D) minus the molecular surface (E) intersected with the focusing surface creates the chunky region: blue in (F).

approximation of the focusing surface since it eliminates volume contained within the RNA itself. More precision is also achieved with this determination of focusing by eliminating very thin regions of focusing that tend to form around the surface of charged molecules that do not include water due to their small size. These thin regions do not really represent the better defined focusing region we are trying to interpret.

2.10 Amino Acid Volume Intersections

We wanted to study another aspect of focusing using the chunky focusing regions we just described. Specifically we wanted to see which amino acids from the protein counterpart of RNA were intersecting the chunky focusing region. We chose to use the entire chunky focusing region for this experiment, we did not need to limit this consideration to where the focusing and electrostatic potential region intersected

or to any individual focusing chunks. We wanted to find only where amino acids intersected the better defined chunky focusing region. For this experiment we create a volumetric representation of each amino acid from the larger protein structure by parsing out each amino acid and creating a surface using surfaceExtractor in the same fashion as the protein and RNA molecular surfaces, figure 2.2. Each amino acid was CSG intersected with the chunky focusing region, and only non negligible intersection volumes, above 1 $Å^3$, were recorded. This represents which amino acids are frequently contacting the focusing region and to what degree.

2.11 Secondary Structure Distances

The last experiment concerned what types of structures within RNA might produce the largest pieces of focusing, or might regularly make contact with a large volume of focusing. We used a program called DSSR[9] to label the secondary structures within each RNA pdb file in my dataset. The output was an appended RNA pdb file listing the secondary structures and which nucleotides were within each structure in the file. With this information we parsed each RNA file and separated out each secondary structure portion in pdb format.

The different secondary structures, exemplified in figure 2.6, produced by DSSR were stems, iloop, hairpins, junctions, and bulges. Stems are the ladder-like double stranded sections of RNA that resembles the classic DNA double helix, red in figure 2.6. Hairpins are single stranded loops on the end of stems that connect the backbones of the two sides, blue in 2.6. Iloops are internal loops that connect two stems together, yellow in 2.6. Bulges are small chunks of nucleotides that protrude out from the normal ladder-like conformation of stems, purple in 2.6. Junctions are internal loops that connect more than two stems or stem portions together, cyan in 2.6. Not all nucleotides are labelled within each RNA, the green in 2.6 are portions not found to be a part of any secondary structure[9].

With the secondary structures and the distinct focusing chunks, not the whole chunky focusing region, we used surfaceExtractor to report the distance between a

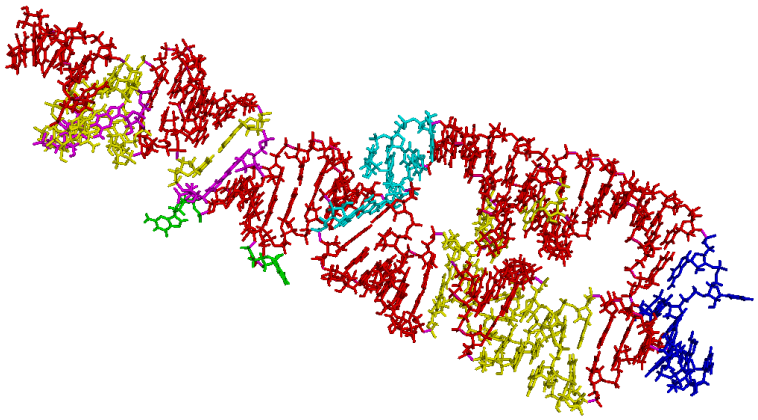

Figure 2.6: Secondary Structure Partitioning of RNA as Labelled by DSSR

We used DSSR[9] to label secondary structures in RNA. Red sections are stems, blue are hairpins, yellow are iloops, purple are bulges, cyan is a junction, & the green was left unlabelled.

surface, the focusing chunk, and a pdb file, the secondary structure. This step was done sequentially: for each secondary structure the distance was calculated to every focusing chunk. The final results were collected using that information, section 3.4, by only reporting the number of times a secondary structure makes contact with a focusing chunk volume, determined to be within a distance of 4 Å.

Chapter 3

Results

3.1 Example Image and Ouput Structures

Figure 3.1 demonstrates several example structures and volumes from our study at a focusing level of 0.65(kt/e). In 3.1(A) the original RNA is in red and protein in blue. In 3.1 (B) and (D) there is a moderately sized focusing chunk, of volume 918.5 $Å^3$, inside a stem secondary structure of the RNA. In 3.1(C) we can see this focusing region passes through a hole in the stem of the RNA. In 3.1 (E) and (F) we show two of several amino acids, from the blue protein in (A), that intersect this particular structure: arginine and alanine with intersection volumes of 125.68 $Å^3$ and 23.83 $Å^3$ respectively . In 3.1(E) the amino acid intersections and distinctions are clearer through the now translucent focusing region; arginine is the larger blue piece in front-left of the smaller alaninine in the back-right. In figure 3.1 the RNA conforms to the familiar double helix shape and is presented as a more classical RNA example from our data for visual clarity. In reality, the structures in our dataset have a diverse array of shape, focusing region definition, secondary structure conformations and amino acid intersections.

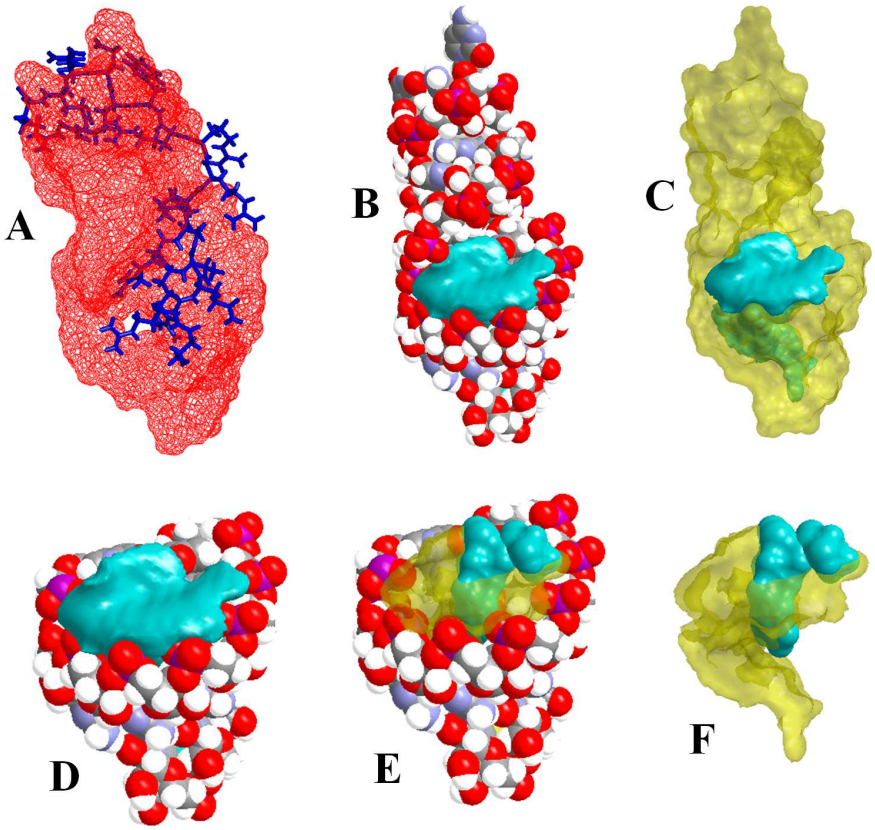

Figure 3.1: Volumetric Examples of Focusing Chunk, Secondary Structure, and Amino Acid

(A) RNA in red, protein in blue. (B) RNA with a small focusing chunk, cyan. (C) RNA, yellow transparent, & the same focusing region, cyan. (D) A single stem secondary structure surrounding the focusing region, cyan. (E) Arginine & alanine in cyan intersecting the same focusing region in yellow surrounded by a stem. (F) Only the focusing region, yellow, & two amino acids that intersect it. D, E, & F are magnified slightly from A, B, & C for clarity.

3.2 Electrostatic Complementarity and Focusing

Figures 3.2, 3.3, and 3.4 all show a different perspective on the volume of electrostatic focusing versus complementarity, which roughly serves as a proxy for binding affinity. Complementarity, figure 2.4(D), was measured at four different levels: 1.25, 2.50, 5.00, or 7.50 (kT/e). Here, we show the results at 1.25 kT/e while the findings for

5.00 are provided in the appendix section .1. The data for 2.50 is quite similar to 1.25, as is 7.50 to 5.00. Showing the results for 5.00 shows the general trend of data being clustered further to the left, lower complementarity, as the potentials get much smaller and therefore have less volume to intersect. Isopotential level 1.25 is the one with the greatest volume. Figure 3.2 shows complementarity versus limited focusing, figure 2.4(G1-G3). The values of the focusing levels are listed in the legend to the right, focusing level 0.65 is the one with the greatest volume, followed by 1.25, then 2.50.

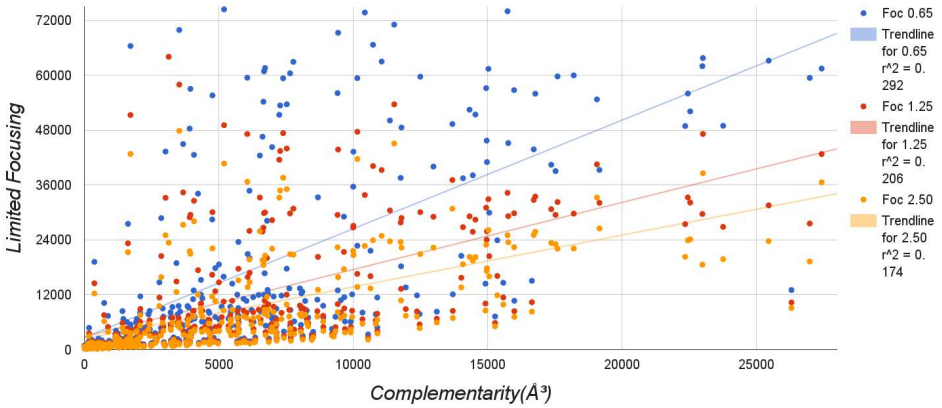

Figure 3.2: Complementarity at 1.25 VS Limited Focusing at 1.25

Complementarity volumes(Å^3) at 1.25(kT/e) versus limited focusing volumes(Å^3) at 1.25(kT/e)

There is a positive linear trend here between limited focusing and complementarity, suggesting that focusing does generally increase with electrostatic affinity. We can see here that when complementarity is low relative to the rest of our dataset, focusing is also usually low. This is not always the case; there are protein-RNA pairs that exhibit high focusing levels at low complementarity. But in general as we move right when complementarity is high, focusing is almost always large. Even at the very lowest volumes of focusing, when moving right, as complementarity increases, the volumes all move off of the condensed lower vertical axis. Then when complementarity is quite high, past the 150,000 Å^3 range focusing never drops nearly as

low as when complementarity is at its lower levels.

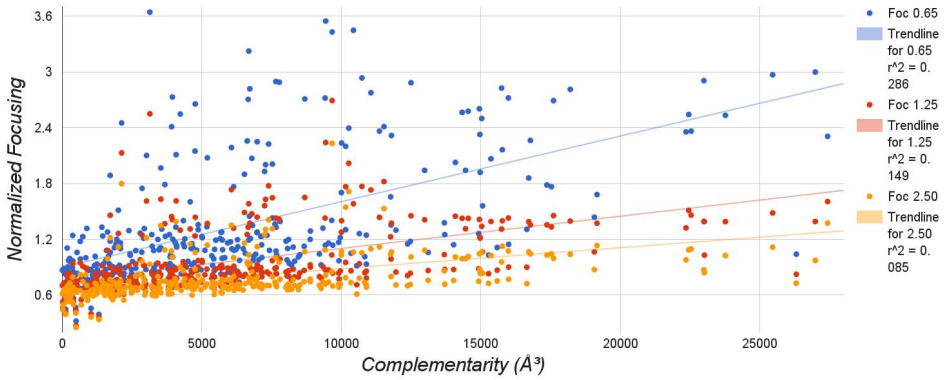

Figure 3.3: Complementarity at 1.25 VS Normalized Focusing at 1.25

Complementarity volumes(Å^3) at 1.25(kT/e) versus normalized focusing volumes(Å^3) at 1.25(kT/e)

Figure 3.3 illustrates the relationship between normalized focusing and complementarity at level 1.25(kT/e), the other graph at 5.00 is in the appendix figure 2. Normalized focusing is limited focusing in figure 3.2 divided by the RNA volume. The point of normalized focusing is to compare complementarity to the quantity of focusing without being biased by how big the RNA molecule is. The desired effect is produced since the data is more spread out across the vertical axis rather than all being clustered along the bottom when compared with some of the very large focusing volumes observed in figure 3.2.

The normalized focusing data in figure 3.3 indicates that RNA structures other than larger RNA molecules exhibit a linearly increasing relationship between focusing and complementarity. Had we not shown this graph one might speculate that the linear trend is only the result of increasingly large RNA molecules creating increasingly large volumes of both potential and limited focusing. Once again here we see the data suggest that when complementarity is low, in general focusing is also low.

Figure 3.4 shows the chunky focusing region relative to complementarity. The other graph at 5.00 is in the appendix figure 3, but they both show the same values

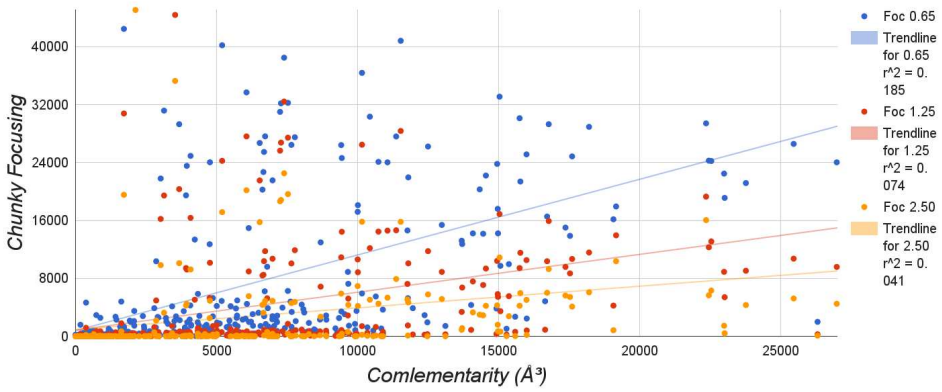

Figure 3.4: Complementarity at 1.25 VS Chunky Focusing

Complementarity volumes($Å^3$) at 1.25(kT/e) versus chunky focusing volumes($Å^3$)

for focusing along the vertical axis, only complementarity along the horizontal axis changes. Recall here that the chunky focusing region does not include the negative isopotential contour of the RNA that was part of the calculation of focusing in the other two graphs, 3.3 and 3.2. This absence of the negative potential region means that the positive linear trend is not dependent on the increasing value of the negative isopotential volume around RNA as it may have been in the first two. The linear trend on this graph also suggests that complementarity does seem to increase as focusing increases. Of course the correlation coefficient, r^2 in the legends, is quite low, so the association between the two is far from ideal for a confident correlation.

The other graphs in the appendix section .1 do all also have a positive linear trend between complementarity and focusing in a very similar way with the above graphs. But due to the nature of how we calculated complementarity the other graphs have the data being increasingly clustered to the lefthand side of the graph. In the case of chunky focusing this makes sense since the values of chunky focusing do not change across the graphs, but the complementarity gets generally smaller, with a few outliers that push everything to the left. But Even despite the clustering of the points on the left with the other graphs, there is still a positive trend.

3.3 Amino Acid and Chunky Focusing Intersections

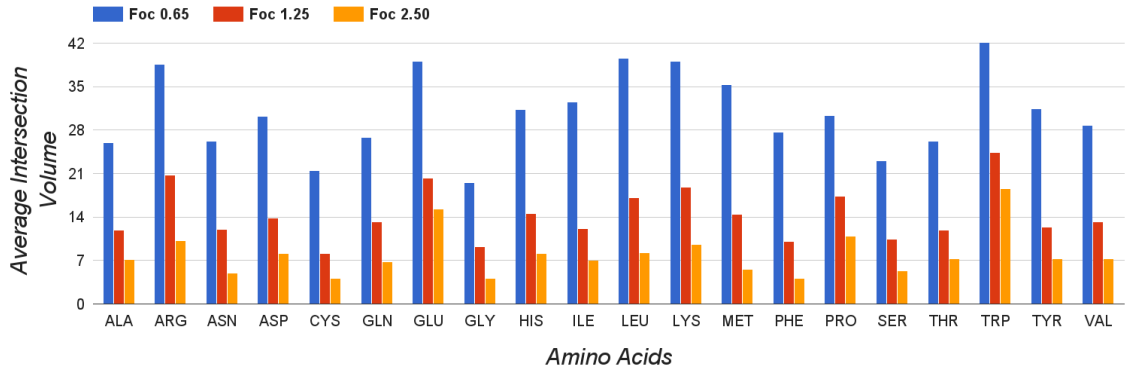

Figure 3.5: Average Intersection Volume Between Amino Acids and the Chunky Focusing Region

Average chunky focusing region intersection volumes(Å^3) for every amino acid at all three focusing levels

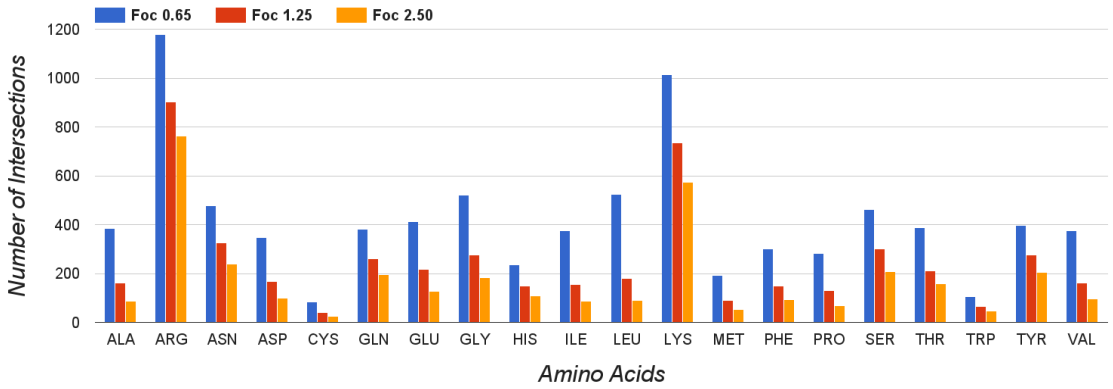

Figure 3.6: The Total Number of Amino Acid Surface Intersections with the Chunky Focusing Region

Total number of times the surface of each amino acid intersects the chunky focusing region at all three focusing levels

Figures 3.5 and 3.6 show the results of intersecting the surface of amino acids

with the chunky focusing region. This data only shows amino acids that have a non-trivial intersection volume, which was limited to 1 $Å^3$. Within my dataset there are thousands of amino acids, most of which do not come close to, let alone intersect this focusing region. Figure 3.5 shows the average volume of intersections for each amino acid at the different focusing levels, the total volume of intersections divided by the number of intersections. Figure 3.6 shows the number of times that amino acids intersect the different focusing regions. Figure 3.6 lends perspective to the averages in 3.5, showing how often the intersections occur.

Arginine and lysine have the largest averages of intersection volume, and they also exhibit many intersections. The other amino acids with large average volumes of intersection, i.e. glutamic acid, tryptophan, and luecine, do not intersect focusing regions frequently, so while they have a few large intersection volumes they do not display this behavior as consistently as arginine and lysine. Based on the data they do not display the behavior often enough to be confident that this is not an anomaly from a few RNA protein pairs.

In figure 3.5 Cysteine, methionine, and tryptophan intersect the least frequently. A few of the amino acids have unusually large average volumes of intersection for focusing level .65, isoleucine, leucine, and methionine, but these averages generally conform with the rest of the amino acids at smaller focusing thresholds. Furthermore, for ILE, LEU, and MET the number of intersections is generally small.

Another anomaly that deserves attention is glutamic acid; it has a rather high intersection average across all three focusing levels. This is particularly unusual considering glutamic acid has a negatively charged side chain. Its number of intersections is not particularly high, so we know the large intersections do not happen often, but the large average is still unusual.

3.4 Secondary Structure Distances from Focusing Chunks

Figures 3.12, 3.7, 3.13, and 3.8 all show the average number of times(vertical axis) RNA secondary structures contact focusing chunks of different volumes(horizontal axis). Since the data is spread across a wide range of volumes, but is the highest in quantity near the lower volumes, it was better for us to break down the number of contacts into averages for different bins defined by the volumes of the chunks, horizontal axes in 3.12, 3.7, 3.13, and 3.8. Figures 3.10 and 3.9 show the total number of secondary structure contacts across the different bin sizes to give a better idea of how the data is distributed amongst the different volumes.

Figure 3.11 shows the distribution of secondary structure types as percentages of the total number of secondary structures in our data. We had a total of 655 individual secondary structure pieces across the 205 RNA molecules in our dataset with secondary structure.

Figures 3.12 and 3.7 how the same information; as do figures 3.13 and 3.8. While figures 3.13 and 3.8 are magnified from 3.12 and 3.7, respectively, to show the large amount of data clustered into volumes of less than 1000 Å^3, the very first bins in 3.12 and 3.7. Figures 3.8 and 3.7 are on the same vertical scale, number of contacts, to give a good perspective on the number of contacts relative to one another across the different chunk volumes. A vertically zoomed in version of 3.8 can be better visualized in the appendix in figure 4. Figures 3.12 and 3.13 show the averages from 3.7 and 3.8, respectively, with the vertical scales set to show 100% of the contacts that are made, regardless of how many. This gives a much better idea visually of the percent breakdown of RNA secondary structure contacts across all of the volume ranges since some ranges have much higher numbers of contacts.

This data represents the secondary structure contacts for the 205 RNA structures that had secondary structure. Not all of the RNA molecules had secondary structures. In order to be classified as having secondary structure RNA molecules

were required to have at least two backbone strands in a double helical conformation. This requirement is a consequence of the way that DSSR identifies secondary structures[9]. This means 150 structures from our set did not contain double stranded RNA. Therefore this analysis excludes RNA molecules lacking secondary structures to evaluate the interaction between focusing regions and secondary structure elements.

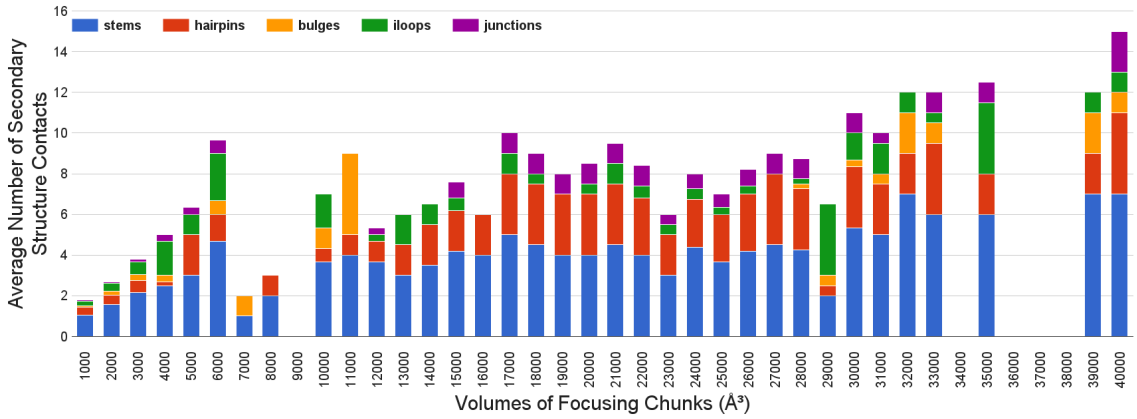

Figure 3.7: Average Number of RNA Secondary Structure Contacts Per Volume of Focusing Chunk for All Sizes

Average number of times each RNA secondary structure contacts focusing chunks, at level 0.65(kT/e), for all volumes

Similar to our other data, much of the volume is clustered around the bottom. This case is slightly different from the graphs in section 3.2 and appendix .1 because here the small pieces of focusing volume are not a result of smaller RNA segments or small focusing regions overall. Instead since this section only contains double stranded RNA molecules there were no small, single chained RNA pieces which generally have very little focusing in the previous two results sections, 3.2 and 3.3. Instead, in this set of data the focusing chunks are individual, discontinuous volumes of focusing that surround the RNA molecule. Due to the large size and twisted shape of some of the RNA pieces there are many small volumes of focusing that represent only distinctly small regions on the surface of the RNA.

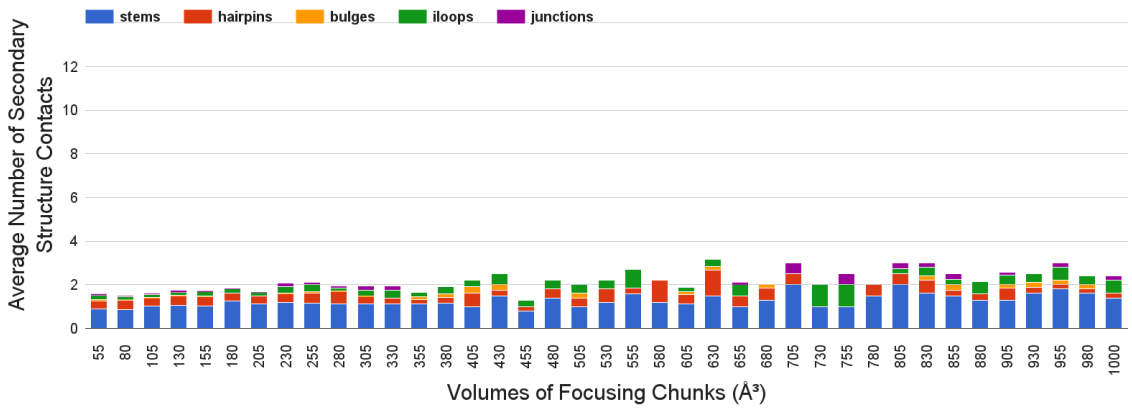

Figure 3.8: Average Number of RNA Secondary Structure Contacts Per Volume of Focusing Chunk Less than 1000Å^3

Average number of times each RNA secondary structure contacts focusing chunks, at level 0.65(kT/e), for volumes no greater than 1000Å^3

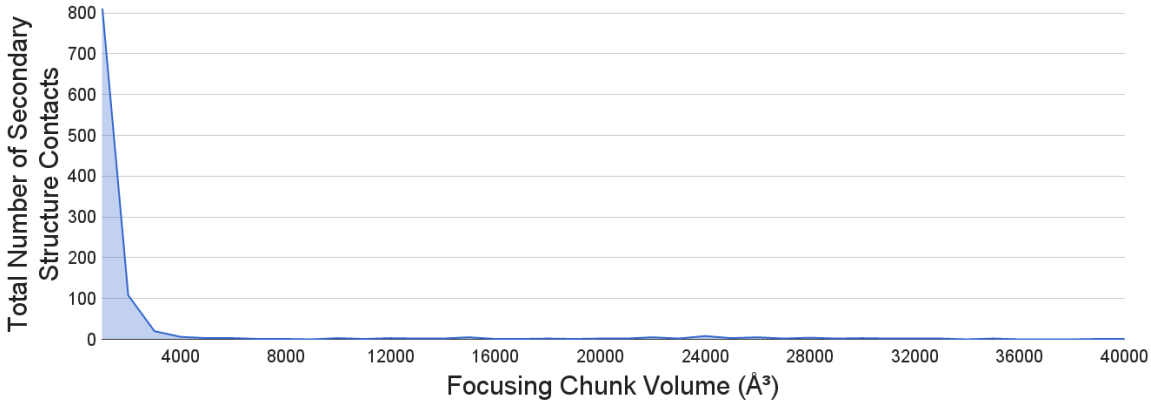

Figure 3.9: Total Number of Secondary Structure Contacts Per Focusing Chunk Volume

Total number of times focusing chunks, at level 0.65(kT/e), of all different volumes sizes make contact with any RNA secondary structure

We can see from the two graphs of the counts of the volume of chunks, 3.9 and 3.10, how many are in the smaller region to the left of the graph, indeed at all three levels of focusing the majority of chunks are below 100 Å^3. Many of the regions of focusing around our structures fall within a small range of chunk sizes. But with

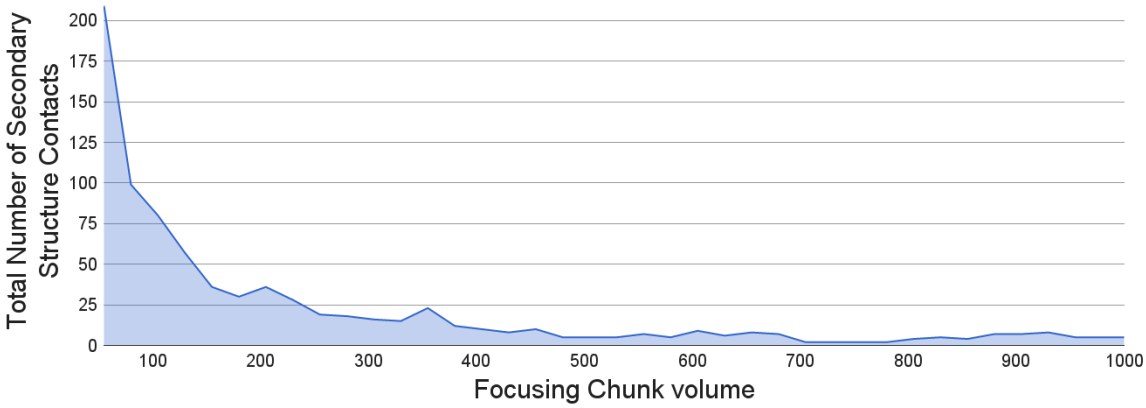

Figure 3.10: Number of Secondary Structure Contacts Per Focusing Chunk Volume Less than 1000Å^3

Total number of times focusing chunks, at level 0.65(kT/e), of different volumes no greater than 1000Å^3 make contact with any RNA secondary structure

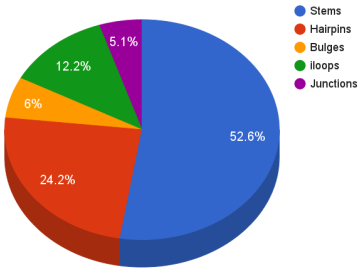

Figure 3.11: Total Number of RNA Secondary Structure Pieces

Percentages of all RNA secondary structures in our dataset

only 205 molecules in this dataset and about a quarter or more of them having very large focusing chunks, of volume greater than around 4000Å^3, it is more likely that many of these molecules all contain several small focusing chunks. In fact at 0.65, the level of the charts above, around 400 chunks have a volume of less than 100 Å^3, 3.10 . This suggests that on average every one of these RNA structures has around two small focusing regions within contact of its surface. Overall the variability of focusing chunk volumes around RNA and its secondary structures is fairly evident considering the size of this set, 205 structures, and the huge number of different

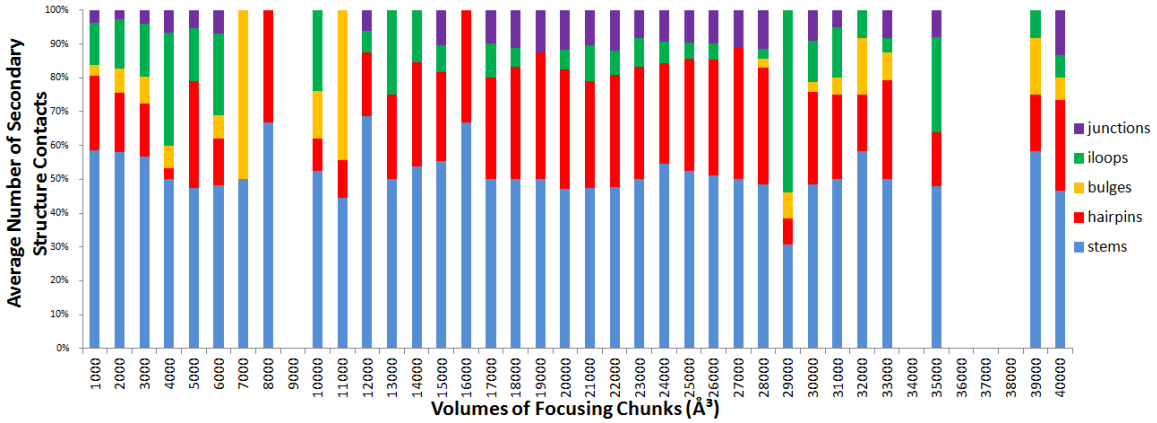

Figure 3.12: Vertically Normalized Average of RNA Secondary Structure Contacts Per Volume of Focusing Chunk

Evenly distributed average of times each RNA secondary structure contacts focusing chunks, at level 0.65(kT/e), for all volumes

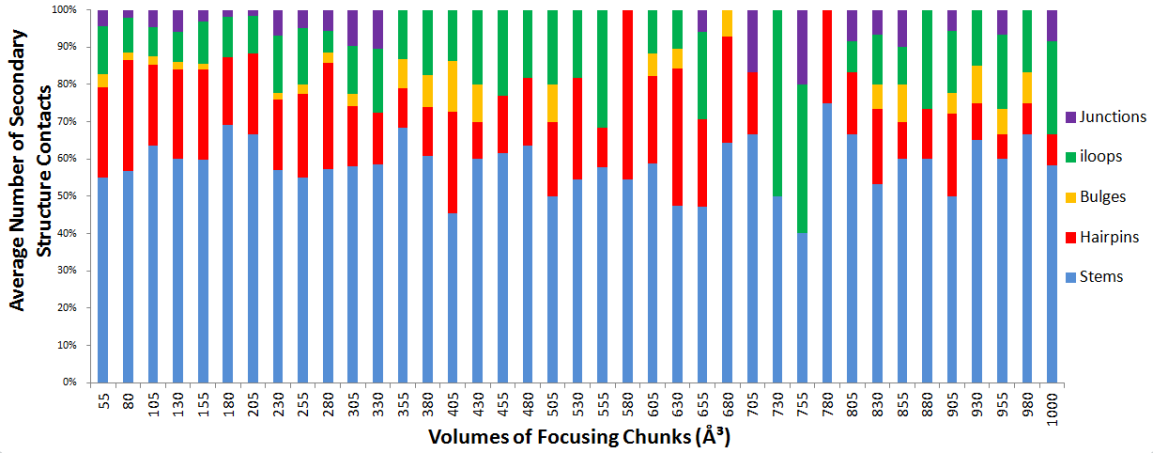

Figure 3.13: Vertically Normalized Average of RNA Secondary Structure Contacts Per Volume of Focusing Chunk Less than 1000Å^3

Evenly distributed average of times each RNA secondary structure contacts focusing chunks, at level 0.65(kT/e), for volumes no greater than 1000Å^3

focusing chunk volumes.

The graphs of the distribution of secondary structure contacts, figures 3.7 and 3.8, make it clear that there is a fairly consistent pattern of secondary structure

distribution around chunks of all sizes. It is not a surprising result that the stems are omnipresent, since the double stranded nature of stems is essential for predicting secondary structure. What is quite interesting is that consistently a quarter of the time when focusing occurs it makes contact with a hairpin. Hairpins are common and important structures in RNA; they are small segments that keep chains of RNA continuous by looping around and connecting the back bone of two adjacent strands[9]. They serve as a sort of cap on the end of the stems and are frequent in double stranded RNA structures, see figure 2.6.

From the pie chart, figure 3.11, we can also see that amongst the RNA molecules in our dataset that have secondary structure, over half of the RNA is comprised of stems. This makes sense, since stems are the ladder shaped backbone pieces that bare resemblance to the DNA double helix and this is often the shape that RNA conforms to in some capacity[9]. Additionally hairpins comprise a quarter of the secondary structures, which is also about the same percentage with which we see them contact the focusing regions.

The last thing to note is the small representation of the remaining three types of the secondary structures. There is not much consistency between bulges, iloops, or junctions, they are infrequently in structures in our dataset. They are similarly absent in contacting the focusing regions, again the two are not the same but they do not seem to play a major role in our data. Beyond their small presence they are inconsistent, if they appeared at some common level, like the regularity of stems and hairpins, we could at least be confident of their infrequent appearance. However they seem to come and go, at both the very high levels, tens of thousands, and the more moderate levels, in the hundreds and low thousands. They are absent occasionally, sometimes together, sometimes separately, meaning that they are not required for focusing. This makes saying much about them difficult, but perhaps it is sufficient to say that they are clearly not necessary for focusing to occur in RNA.

The graphs of these results for focusing level 1.25(kT/e) is shown in the appendix .3 broken down into graphs in the same way, but without the normalized graphs, since the graphs in the appendix have the secondary structure contacts distributed in roughly the same fahsion. The results from level 2.50(kT/e) are not

shown as they have very similar data and do not contribute any new information.

Chapter 4

Discussion

Through our work here we are attempting to show the contribution of focusing towards the electrostatic aspect of affinity between RNA and protein. Knowing the importance of focusing in protein DNA binding[10] we hope to suggest that focusing plays a similar role in the interaction between RNA and proteins. We applied novel software that allows us to volumetrically quantify focusing and how it effects both protein and RNA. This volumetric approach is different from previous methods which are typically based on sequence information that relies on documented focusing phenomena. This sequence based analysis prohibits the ability to define focusing for much of the varied RNA molecules that often do not conform to a general structure in the same way DNA does.

In our survey we quantified focusing in several ways and compared it to electrostatic complementarity to demonstrate its effect on the electrostatic aspect of affinity. We used our method for finding focusing to show how it interacts amongst the amino acids in proteins that bind RNA. We also found where focusing frequently occurs in RNA using relevant new methods for labeling the staggeringly complex set of conformations RNA can adopt[9]. Our survey used a novel volumetric approach, with a wide breadth of coverage across the diverse set of RNA structures to target a comprehensive range of factors that both rely on and are affected by focusing. As far as we are aware no volumetric survey like ours of RNA focusing and electrostatic interactions with respect to protein interfaces has been attempted.

The results from our analysis of intersecting potential regions helped to support our hypothesis that focusing and electrostatic complementarity are connected. Across all of our graphs in section 3.2 and appendix section .1 we saw a linear trend of focusing increasing with complementarity. While we did not see large levels of focusing strictly in places of high complementarity we did not see low levels of focusing in areas with high levels of complementarity. This suggests that regions with a large focusing effect often lead to higher degrees of electrostatic attraction. To avoid speculation that this positive trend was the result of much larger RNA molecules producing uniformly larger focusing and potential regions, we normalized the sizes of focusing calculations proportionally with the size of the RNA. Even with this normalization we still saw a positive trend of focusing rising with complementarity.

A moderate portion of our data in section 3.2 and appendix section .1 was clustered towards the bottom left corner of the graph where both complementarity and focusing were low. We included in our survey RNA pieces that were highly varied in size and shape, some of these included small RNA strands with no secondary structure and little focusing affect. Undoubtedly these tiny RNA components of our data comprised much of the clusters in the bottom corner which reinforces our results and approach. As we had hoped these structures had both very little complementarity and focusing across all three metrics of evaluation, figures 3.4, 3.3, 3.2 and appendix section .1.

In our analysis of amino acid intersections with the chunky focusing region of RNA in section 3.3 we showed that arginine and lysine not only have the most frequent contact with the focusing region around RNA, figure 3.6, but that they have the most consistently large volumes of intersection, figure 3.5. This suggests that across all the RNA protein binding pairs that we studied, arginine and lysine have the strongest connection with the focusing region.

This finding is consistent with what we know about the charges of both RNA and these amino acids. RNA is strongly negatively charged and lysine and arginine have the strongest positive charge of all amino acids. This result gives support to our theory that focusing enhances interactions with proteins since the electrostatically enhanced areas of focusing around the negatively charged RNA are most frequently

interacting with the strongest oppositely charged amino acids of protein.

In our final experiment, section 3.4, we identified common regions in RNA based on secondary structure[9] where focusing occurs the most often. The stems that make up a majority of the secondary structure of RNAs are the primary location for focusing. Interestingly we also showed that focusing consistently occurs around hairpins. While focusing is not exhibited near as many hairpins as stems the consistency with which it occurs is similar. Furthermore we found that the other secondary structures in RNA, iloops, junctions, and bulges, are not reliably contributing to the focusing affect around RNA no matter the size of the focusing region. Considering how consistently we saw focusing around both the stems and hairpins of RNA the other secondary structures were too varied in their presence across all sizes of focusing to contribute any sort of reliability towards the affect. Our results certainly do not to say that they cannot contribute to focusing, but rather they are not necessary; this is an interesting finding in itself.

Considering the important role of RNA in molecular biology, studying and understanding this molecule and all of the ways it interacts with proteins is crucial. With a more extensive knowledge of RNA we can continue to open up the doors to drugs and therapies focused around this crucial biochemical machine. The more we know about RNA structure the more we can reveal about diseases and conditions that are the consequence of faulty molecular structures or functions. Diagnostic medicine, predictive drug design, and healthcare in general can all benefit from a wider knowledge base on RNA and its intricacies, including many of the results we have shown here in this work.

The ideal next steps of this experiment would be to test the validity of the computational implications in a rigorous chemical approach. Because our methods are completely computational it is only an approximation of reality. The accuracy of what we have shown is limited by the quality and consistency of the structures we are using from the PDB. Furthermore static molecules and the calculations that we performed on them to simulate reality are only the best computational approximation we have of the attributes of those real-life molecules. In living biological systems there is constant motion and variation; molecules do not always have the

same sequence or tertiary structure we have presented here; and there are numerous ions, molecules and other chemicals competing for interaction.

Without thorough empirical biochemical evaluation we are left only with the accurate speculation that our computational methodology can provide. A physical approach to show how focusing influences real binding would be a fascinating study, and one that would undoubtedly reveal more about what we have shown in this survey. Hopefully we have provided enough evidence to warrant further investigation into the role of electrostatic focusing on RNA and protein interactions. Such an investigation could elucidate crucial mechanisms in the complex process that is RNA-protein binding. Focusing and its role in the interfacing between RNA and protein is clearly worth further analysis, without more attempts like ours to uncover previously unstudied binding mechanisms we have only what we know and might be blind to the possibilities of an even more interesting story about how these crucial biological molecules interact.

Bibliography

[1] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, TN Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[2] Sebastian Blumenthal, Yisheng Tang, Wenjie Yang, and Brian Y Chen. Isolating influential regions of electrostatic focusing in protein and dna structure. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 10(5):1188–1198, 2013.

[3] Brian Y Chen. Vasp-e: Specificity annotation with a volumetric analysis of electrostatic isopotentials. *PLoS Comput Biol*, 2014.

[4] Brian Y Chen and Barry Honig. Vasp: a volumetric analysis of surface properties yields insights into protein-ligand binding specificity. *PLoS Comput Biol*, 6(8):11, 2010.

[5] Ian W Davis, Andrew Leaver-Fay, Vincent B Chen, Jeremy N Block, Gary J Kapral, Xueyi Wang, Laura W Murray, W Bryan Arendall, Jack Snoeyink, Jane S Richardson, et al. Molprobity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic acids research*, 35(suppl 2):W375–W383, 2007.

[6] Mickael Goujon, Hamish McWilliam, Weizhong Li, Franck Valentin, Silvano Squizzato, Juri Paern, and Rodrigo Lopez. A new bioinformatics analysis tools framework at embl–ebi. *Nucleic acids research*, 38(suppl 2):W695–W699, 2010.

[7] Mark A Larkin, Gordon Blackshields, NP Brown, R Chenna, Paul A McGettigan, Hamish McWilliam, Franck Valentin, Iain M Wallace, Andreas Wilm, Rodrigo Lopez, et al. Clustal w and clustal x version 2.0. *Bioinformatics*, 23(21):2947–2948, 2007.

[8] Lin Li, Chuan Li, Subhra Sarkar, Jie Zhang, Shawn Witham, Zhe Zhang, Lin Wang, Nicholas Smith, Marharyta Petukh, and Emil Alexov. Delphi: a comprehensive suite for delphi software and associated resources. *BMC biophysics*, 5(1):9, 2012.

[9] Xiang-Jun Lu, Harmen J Bussemaker, and Wilma K Olson. Dssr: an integrated software tool for dissecting the spatial structure of rna. *Nucleic acids research*, page gkv716, 2015.

[10] Remo Rohs, Sean M West, Alona Sosinsky, Peng Liu, Richard S Mann, and Barry Honig. The role of dna shape in protein–dna recognition. *Nature*, 461(7268):1248–1253, 2009.

Appendix

.1 Additional Results from Section 3.2

Graphs similar to the results in section 3.2, but figures 1, 2, and 3 show the data at potential level 5.00(kT/e).

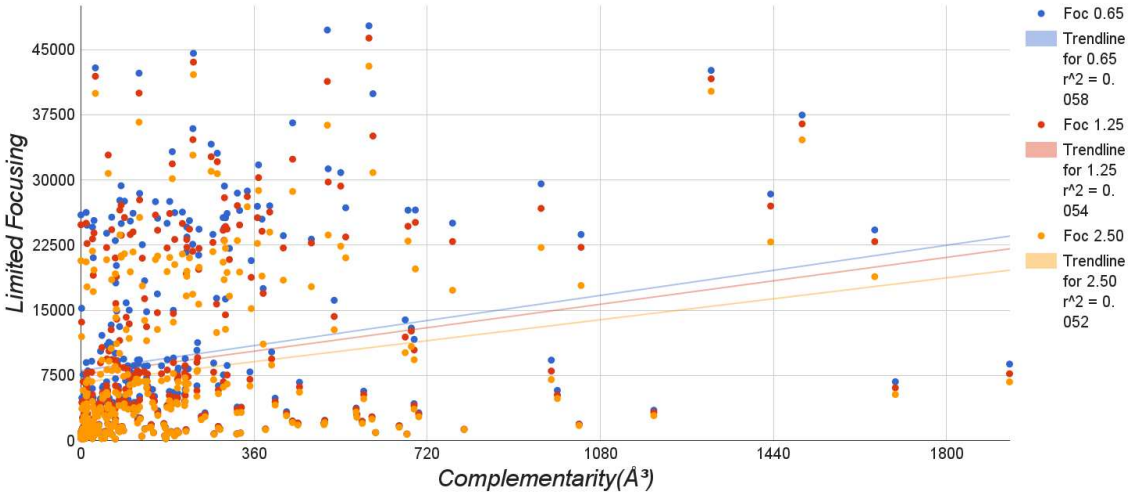

Figure 1: Complementarity at 5.00 VS Limited Focusing at 5.00

Complementarity volumes($Å^3$) at 5.00(kT/e) versus limited focusing volumes($Å^3$) at 5.00(kT/e)

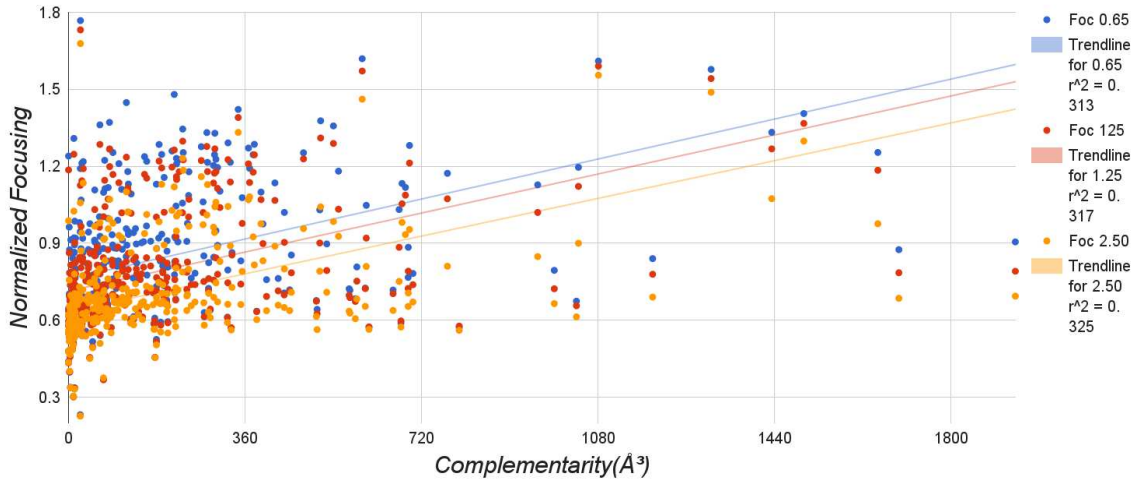

Figure 2: Complementarity at 5.00 VS Normalized Focusing at 5.00

Complementarity volumes(Å^3) at 5.00(kT/e) versus normalized focusing volumes(Å^3) at 5.00(kT/e)

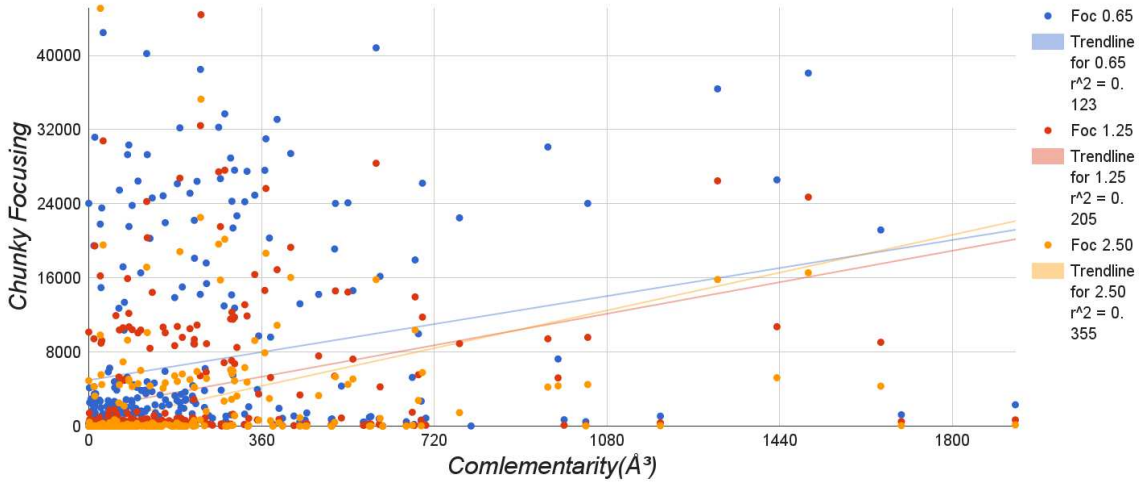

Figure 3: Complementarity at 5.00 VS Chunky Focusing

Complementarity volumes(Å^3) at 5.00(kT/e) versus chunky focusing volumes(Å^3)

.2 Additional Graph from Section 3.4

The same graph as figure 3.8, but with the vertical axis expanded to better show the data. The height on figure 3.8 was bigger to match the same size as the bigger

picture in figure 3.7.

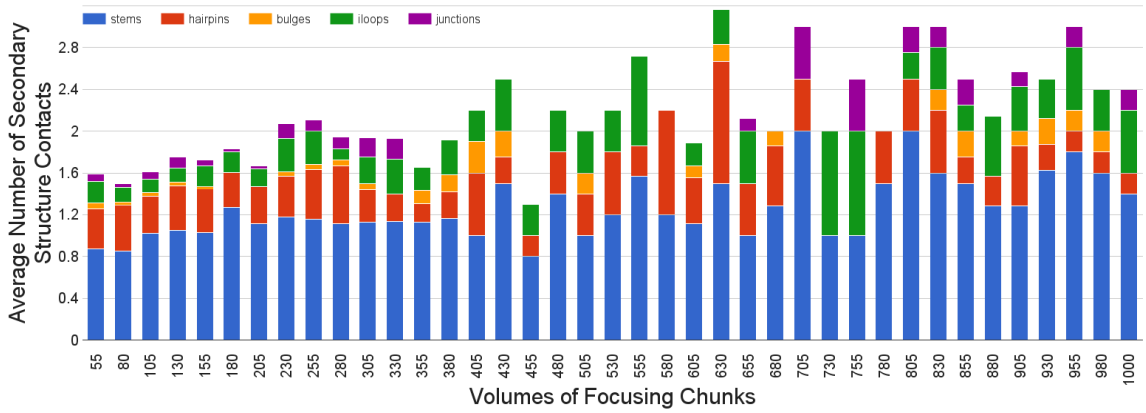

Figure 4: Average Number of RNA Secondary Structure Contacts Per Volume of Focusing Chunk Less than 1000 $Å^3$

Expanded view of 3.8 which shows average number of times each RNA secondary structure contacts focusing chunks, at level 0.65(kT/e), for volumes no greater than 1000 $Å^3$

.3 Additional Results from Section 3.4

Graphs showing secondary structure intersections with chunky focusing regions results similar to section 3.4 from our calculations in section 2.11 The horizontal scale and bin sizes are different than in section 3.4 because this is at a different focusing level, different values were needed to better show the results and to better display secondary structure percentages in appropriate bin widths. These graphs are arranged differently by displaying the counts, 6 and 8, under the averages, 5 and 7 respectively, to show more precisely how the data is spread across the different focusing volume bins.

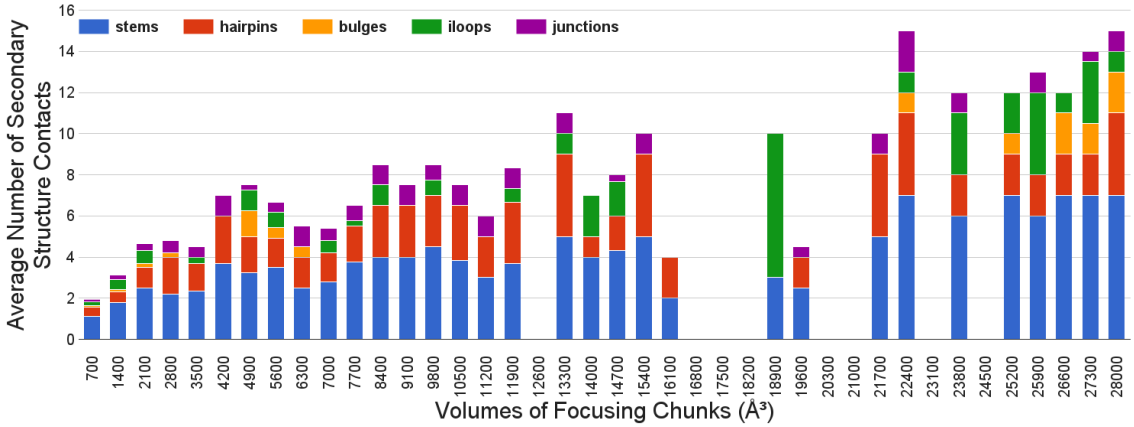

Figure 5: Average Number of RNA Secondary Structure Contacts Per Volume of Focusing Chunk for All Sizes

Average number of times each RNA secondary structure contacts focusing chunks, at level 1.25(kT/e), for all volumes

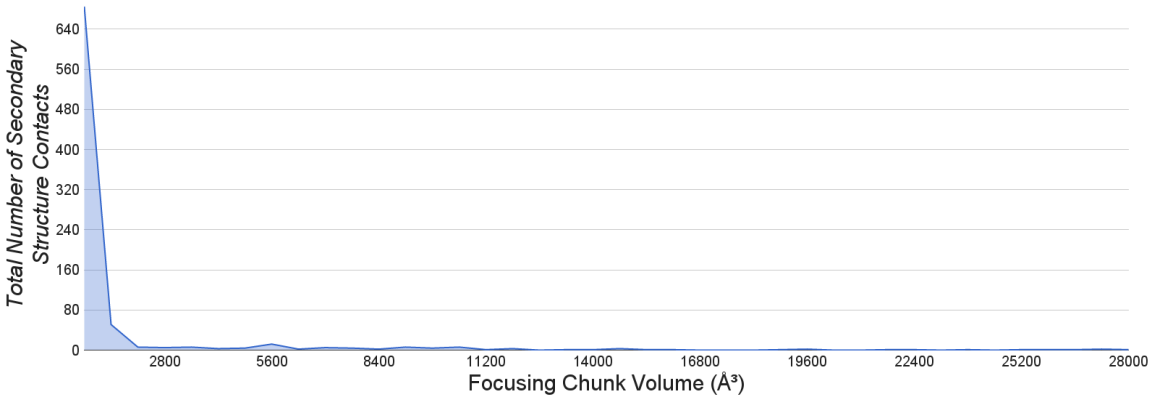

Figure 6: Total Number of Secondary Structure Contacts Per Focusing Chunk Volume

Total number of times focusing chunks, at level 1.25(kT/e), of different volumes of all sizes make contact with any RNA secondary structure

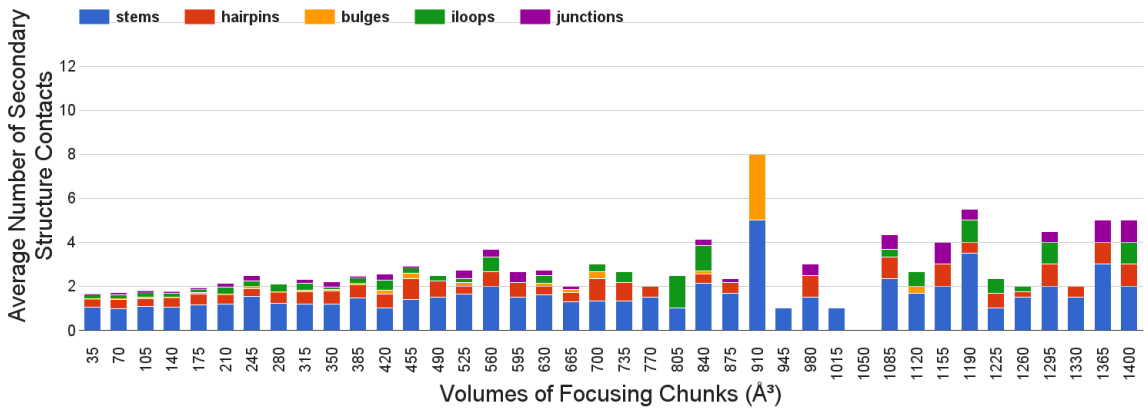

Figure 7: Average Number of RNA Secondary Structure Contacts Per Volume of Focusing Chunk Less than 1400Å^3

Average number of times each RNA secondary structure contacts focusing chunks, at level 1.25(kT/e), for volumes no greater than 1000Å^3

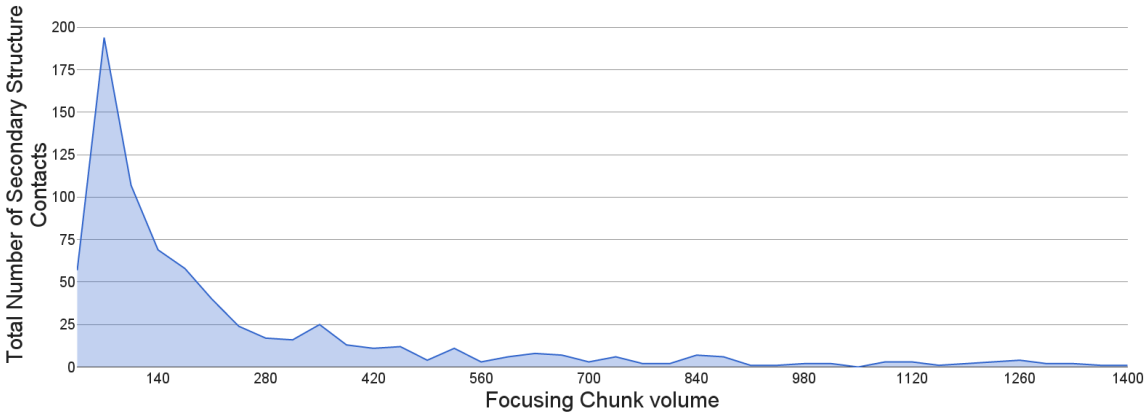

Figure 8: Number of Secondary Structure Contacts Per Focusing Chunk Volume Less than 1400 Å^3

Total number of times focusing chunks, at level 1.25 (kT/e), of different volumes no greater than 1400 Å^3 make contact with any RNA secondary structure

Personal Vita

Devan Lester Bicher

As an undergraduate I attended Lehigh University where, in 2013, I recieved my B.S. in Bioengineering on the Pharmacuetical track, with a minor in computer science. As an undergraduate I was a member of the fraternity Pi Kappa Alpha, I played on the rugby team and worked for College of Arts and Sciences helping to maintain and create websites for the numerous departments within the college.

I have thoroughly enjoyed my time here at Lehigh University both as a graduate and an undergraduate student. I have had many new experiences and have learned a tremendous amount about both my disciplines and, perhaps more importantly, myself. Lehigh has help facilitate this learning experience and it is a wonderful place to grow and develope as a scientist and an individual.

In my free time I enjoy furthering myself by reading, working out and exploring my passions for computer science, biology, and chemistry. I also love spending time with my friends. I am an adventerous spirit and despite my interests in computation I also thoroughly enjoy spending time outdoors where I like camping, hiking, and exploring nature, it is a welcomed reprieve from a life increasingly spent behind a screen.

As may be evident from my combination of degrees I am deeply interested by all aspects of bioinformatics and the interface between biology and computer science. I am becoming increasingly interested in synthetic biology. I truly believe that understanding and harnessing the biological universe is the way of the future and that computer science holds the key to unlocking that potential. Everything about biology is staggeringly complex yet incredibly fascinating. I look forward to my career in a bioinformatics career where I can help the world with my passion for biology and my knowledge of computer science.