RNA STRUCTURE ANALYSIS VIA THE RIGID BLOCK MODEL

by

MAURICIO ESGUERRA NEIRA

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Chemistry and Chemical Biology

Written under the direction of

Wilma K. Olson

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2010

ABSTRACT OF THE DISSERTATION

RNA Structure Analysis via the Rigid Block Model

by Mauricio Esguerra Neira

Dissertation Director: Wilma K. Olson

RNA structure is at the forefront of our understanding of the origin of life, and the mechanisms of life regulation and control. RNA plays a primordial role in some viruses. Our knowledge of the importance of RNA in cellular regulation is relatively new, and this knowledge, along with the detailed structural elucidation of the transcription machine, the ribosome, has propelled interest in understanding RNA to a level which starts to closely resemble that given to proteins and DNA.

In the process of progressively understanding the landscape of functionality of such a complex polymer as RNA, one practical task left to the structural chemist is to understand the details of how structure relates to large-scale polymer processes. With this in mind the fundamental problems which fuel the work described in this thesis are those of the conformations which RNA's assume in nature, and the aim to understand how RNA folds.

The RNA folding problem can be understood as a mechanical problem. Therefore efforts to determine its solution are not foreign to the use of statistical mechanical methods combined with detailed knowledge of atomic level structure. Such methodology is mainly used in this work in a long-term effort to understand the intrinsic structural features of RNA, and how they might relate to its folding.

As a thing among things, each thing is equally insignificant; as a world each one equally significant.

If I have been contemplating the stove, and then am told; but now all you know is the stove, my result does indeed sound trivial. For this represents the matter as if I had studied the stove as one among the many, many things in the world. But if I was contemplating the stove, it was my world, and everything else colorless by contrast with it ...

For it is equally possible to take the bare present image as the worthless momentary picture in the whole temporal world, and as the true world among shadows.

Ludwig Wittgenstein

As a molecule among molecules, each molecule is equally insignificant; as a world each one equally significant.

If I have been contemplating RNA, and then am told; but now all you know is RNA, my result does indeed sound trivial. For this represents the matter as if I had studied RNA as one among the many, many molecules in the world. But if I was contemplating RNA, it was my world, and everything else colorless by contrast with it ...

For it is equally possible to take the bare present image as the worthless momentary picture in the whole temporal world, and as the true world among shadows.

Anonymous Chemist

Acknowledgements

I would first like to give a special thanks to Dr. Yurong Xin, whose patience, help, and collaboration since the very beginning of my joining of the Olson lab have been fundamental for the development of this work. I would like to thank Dr. Olson's extreme patience, and room for freedom on carrying out this research. Finally I thank all colleagues at the Olson lab.

I would like to dedicate this thesis to David and Stella Case, without them these words would not exist.

Table of Contents

Abstract ii

Acknowledgements iv

List of Tables viii

List of Figures ix

1. Introduction 1

1.1. RNA chemistry 1

1.2. RNA folding 3

1.3. Is RNA folding a hard or easy problem? 5

1.4. Experimental folding techniques 7

1.5. RNA simulations 7

1.5.1. Local nucleotide interactions 8

1.5.2. RNA secondary structure algorithms and the lack of tertiary ones 9

1.5.3. RNA overall fold 9

1.5.4. RNA motifs 11

1.6. Overview 12

References 14

2. RNA Base Steps 21

2.1. Consensus Clustering of Single Stranded Base Step Parameters 24

2.1.1. Combining Fourier Averaging Results and Clustering Analysis 24

2.1.2. Selection of a Clustering Methodology 28

References 40

3. RNA Base-Pairing 42

3.1. Canonical and Noncanonical Base-pairs 42

3.2. Clustering of Yurong's Classification . . . 42
References . . . 44

4. RNA Base Pair Steps . . . 45
4.1. Analysis (Albany Poster) and Django Webserver . . . 45
4.2. Persistence Length vs. Hagerman . . . 45
4.3. AMBER: Persistence Length of Base-Pair Step Patterns . . . 45
References . . . 46

5. RNA Motifs . . . 47
5.1. GNRA tetraloop . . . 47
5.1.1. 3DNA-Parser . . . 47
5.1.2. Overlap Scores . . . 48
5.2. Triplets on RNA (comparison to Laing et al.) . . . 48
References . . . 51

6. RNA Helical Regions and Graph Theory . . . 52

Appendix A. Standard reference frame and local parameters . . . 53
A.1. Base-pair and base-step parameters . . . 53
A.2. Local helical parameters . . . 56
References . . . 59

Appendix B. Clustering Analysis (CA) . . . 60
B.1. General Methodology . . . 60
B.2. Hierarchical methods . . . 61
References . . . 65

Appendix C. Dimension Reduction . . . 66
C.1. Principal Component Analysis . . . 66
References . . . 68

Supplement A. Figure Supplements . . . 69

Curriculum Vitae 72

List of Tables

2.1. Some large RNA structures (>300 bases) elucidated in the last decade. 23

2.2. Number of base-steps with RMSD values less than or equal to 10 Å between the reference base-step vectors from the four groups of non-A-type RNA dinucleotide conformations and all base-step vectors found in the 23S strand of *Haloarcula marismortui*. The percentage is calculated with respect to a total of 2753 base-steps present in the 23S chain of the 50S subunit of the ribosome. 28

2.3. Base step parameters for common DNA and RNA conformations. The base-step parameters are computed for a single-stranded base-step rather than a double-stranded base-pair step. 33

3.1. Classification of RNA Types in Non-Redundant Dataset at less than 3.5 Å (For Base-Pairs in Helices of 3 base-pairs or more). 43

B.1. Example of structures, considered as bidimensional vectors, to be clustered using the average linkage method and the Manhattan distance. 63

List of Figures

1.1. A single strand of RNA drawn in the 5′ to 3′ sense showing the three chemical entities which compose it, base, sugar, and phosphate. The four bases (A, G, C, U) are colored according to the NDB (Nucleic Acid Database) convention [18], the phosphate is colored gray, and the sugars black. The bases G, and C, and the furanose sugar attached to the G are numbered according to the IUPAC rules [19]. This figure is an adaptation of Figure 2.1, in Wolfram Saenger's book, "Principles of Nucleic Acid Structure" [20]. 2

1.2. Saenger base-pairing classes, reproduced from his book, "Principles of Nucleic Acid Structure". [20]. 4

1.3. **Left:** Sugar, and sugar-phosphate backbone torsion angles. **Right:** The most common sugar pucker conformations in RNA, that is, $C_{3'-\mathrm{endo}}$ and $C_{2'-\mathrm{endo}}$, reproduced from Wolfram Saenger's, "Principles of Nucleic Acid Structure". [20]. 5

1.4. Separation of secondary and tertiary interaction in RNA [39]. Double helical secondary structure represented by individual cylinders and tertiary interactions by association of cylinders. Color coding stands for separate helical regions of RNA, and the connecting black strings represent single stranded loop structures. 6

1.5. Ribbon-coil schematic illustraring the fold and intermolecular units of a dimer of prealbumin (PDB_ID:2pab), or transthyretin, taken from Richardson *et al.* [90] 10

1.6. Images of the *Haloharcula marismortui's* large ribosomal subunit NDB_ID:RR0033 (left) and the hammerhead ribozyme (right) NDB_ID:UR0029. The figures were taken directly from the NDB web pages, and show a 3DNA generated [91] ribbon representation of the phosphate backbone, and a block representation for the nucleotide bases. From the figures it's clear that, whereas the ribozyme fold can be clearly understood with this representation, the ribosome fold cannot. 11

2.1. **Left:** Total number of RNA bases added to the PDB database between 2000 and 2010 (Exponential fit line in blue). **Right:** Total number of RNA structures solved yearly by X-Ray crystallography between 2000 and 2010 (Exponential fit line in red). 21

2.2. Frequency of nucleotide bases in RNA molecules found in the PDB classified by the size of RNA molecules. We define the size as the total number of nucleotide bases present per molecule. 23
2.3. Figure taken from Richardson et al. [11] where the blue and green dots in a) mean very accurate van der Waals distances, and in b) the red and orange dots mean steric clashes, that is, distances outside the acceptable van der Waals range. 24
2.4. Dendrogram showing the results of consensus clustering of 20 non-Atype rRNA dinucleotides according to their hexadimensional base-step parameter vectors. 26
2.5. RNA dinucleotide structures organized by clusters obtained from consensus clustering of their hexadimensional base-step parameter vectors. The structures have been centered on the reference frame of the first step, that is, the adenine base, and the minor groove face of the rigid block parameter associated to adenine is facing the viewer. 27
2.6. Root mean square deviation of the main four groups show in Figure 2.5. The color of the histograms is the same as that of the boxes surrounding the structures of Figure 2.5 . . . 29
2.7. Root mean square deviation histograms for the subgroups present in group IV. Since subgroup IVb is composed of A-RNA like conformations we see in the upper left histogram that the highest proportion of small RMSD values belongs to this group. 30
2.8. Rigid block representation of dinucleotide steps. The major groove side of the first nucleotide block is oriented towards the viewer and shaded gray. **Left:** Drawn in blue, the block representing the Group I cluster from Figure 2.5. Superimposed to the Group I cluster are three structures whose step-parameter RMSD's with respect to the Group I cluster are less than or equal to 10 Å. **Right**: With an RMSD less than or equal to 15 Å we "identify" a total of seven structures from the ribosome. We clearly see that three of them (encircled in cyan blobs) are farther apart from the original Group I main structure of Figure 2.5 which is drawn in blue. 31
2.9. Pairs scatterplot for base-step parameters, shift, slide, rise, tilt, roll, and twist, for the non-ARNA dataset colored according to purine-pyrimidine (black), purine-purine (red), pyrimidine-pyrimidine (green), and pyrimidine-purine (blue) steps. 32

2.10. Cluster validity scores for internal measures. Notice how the hierarchical method, labeled as 1 in black color, behaves better for the whole range of Connectivity (smaller values) and Dunn (higher values), and it also outperforms all others after $k = 12$ for Silhoutte (higher values) scores. 34
2.11. Cluster validity scores for stability measures. 35
2.12. RMSD values between base-step parameters of the 23S subunit of ribosomal RNA and the standard base-step parameters derived from Arnott and collaborators [24] work. . . . 36
2.13. Cluster validity scores for the non-ARNA dataset. It can be seen clearly that the optimal method for clustering is the hierarchical one, as measured by lower values in the connectivity scores, and higher values in the Dunn score. The optimal number of clusters given by the dunn score is 67, we also see shoulders at $k = 67$, for the connectivity and silhouette scores. 38
2.14. 17 out of the 67 groups clustered using the hierarchical clustering algorithm are drawn in a photograph contact sheet fashion. Each group is centered on the base reference frame of the adenine block drawn in red. In the lower right corner of the "contact sheet" the full space of 797 reconstructed steps is shown, along with the 20 steps derived from schneider et al. work. Notice how the only "hollow" side of the "onion" formed by the full space of base-step conformations is that corresponding to the watson-crick base-pairing region. 39
5.1. GNRA Tetraloop from *Thermus Thermophilus* 23S Ribosomal RNA PDB-ID:1ffk. 48
5.2. Normalized histograms showing the distribution of overlap values in the 23S subunit or *Thermus Thermophilus* rRNA, PDB-ID:1jjk. In histogram (a) all values are included, but in histogram (b) only values greater than zero are included. Notice the high preponderance of zero values, exactly 897 out of a total of 2705. 49
5.3. Dendrogram for consensus clustering of overlap scores in the ribosome. Zero values filtered out and remaining data normalized. 50

A.1. Standard reference frame of an A-T base-pair [4]. The *y*-axis (dashed green line) is chosen to be parallel to the line connecting the C1′ of adenine and the C1′ of thymine associated in an ideal Watson-Crick base-pair. The *x*-axis is the perpendicular bisector of the C1′ - C1′ line, and the origin is located at the intersection of the *x*-axis and the line connecting the C8 atom of adenine and the C6 atom of thymine. The *z*-axis is the cross product of the $\hat{x}$ and $\hat{y}$ unit vectors. 54

A.2. Illustration of base pair and base step parameters [1] 57

B.1. Clustering tree for 5 bidimensional vectors using the Manhattan distance definition and the average linkage clustering method. 64

S1. Non A-RNA Type base steps centered on the standard reference frame of Adenine. Top view with the Minor Groove side of Adenine pointing down the page and the Major Groove pointing up. 70

S2. The total number of structures available in the pdb up to the end of year 2009. The scale of the axis in the left (in black), is ten times that in the right (in green). The black y-axis sets the scale for the number of protein structures available in the PDB up to the end of the year 2009. The green y-axis sets the scale for the number of molecular structures containing, rna only (in red), dna only (in blue), and protein plus nucleic acid (in green). One can clearly see that the total number of protein, rna, and protein plus nucleic acid structures is growing exponentially. It is also clear that the number of DNA structures is perhaps tending toward a constant number, that is, it might not be growing. It is also interesting to see how the number of RNA structures really lifts off in the middle of the nineties, whereas for DNA the growth started earlier and is settling down. 71

Chapter 3

RNA Base-Pairing

3.1 Canonical and Noncanonical Base-pairs

As seen in Figure 1.2, there can be various base-pairing patterns between heterocyclic bases in nucleic acids due to a variety of possible hydrogen bonding interactions. The most prevalent hydrogen bonding pattern is known as canonical Watson-Crick, all other possible patterns are known as non-canonical base-pairs and are more common in RNA than in DNA. We used 3DNA to find all base-pairs in a non-redundant database of X-ray determined RNA structures from the PDB with resolutions less than or equal to 3.5 Å. We also constrained our search to helical regions in RNA. Such helical regions are composed of 3 consecutive base-pairs or more, and they need not be covalently bonded by the sugar-phosphate backbone between consecutive base-pairs. For more details the reader is refered to Olson et al. [1].

In the helical regions data we quantify:

Abundances (Counts) Deformabilites Helical Context

NON-REDUNDANT DATABASE AND CONSTRAIN TO HELICAL REGIONS.

We use a non-redundant dataset of RNA structures. By non-redundant we mean to say that, for the main source of RNA structural information, which is the ribosome, we used only one of the available structures per organism, that is, one for each of *Deinococcus Radiodurans*, *Haloarcula marismortui*, *Escherichi coli*, and *Thermus thermophilus*.

3.2 Clustering of Yurong's Classification

RNA Type	Counts	G	C	A	U
small helices	78	891	753	404	442
drug-RNA	36	932	862	365	433
protein-RNA	207	4001	3457	1771	1731
protein-tRNA	9	175	155	98	87
rRNA	13	3866	2949	1939	1785
tRNA	13	205	159	124	112
ribozyme	113	2434	2086	1438	1150
Total	469	12504	10421	6139	5740

Table 3.1: Classification of RNA Types in Non-Redundant Dataset at less than 3.5 Å (For Base-Pairs in Helices of 3 base-pairs or more).

References

[1] Olson, W. K., Esguerra, M., Xin, Y., and Lu, X.-J. (2009) New Information Content in RNA Base Pairing Deduced from Quantitative Analysis of High-Resolution Structures. *Methods,* **47**, 177–186.