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, the mechanisms of life regulation and control, and it plays a primordial role in some viruses. Our knowledge of the importance of RNA in cellular regulation is relatively new, (a bit more than a decade, Craig and Mello, Nature 1998) and along with the detailed structural elucidation of the transcription machine, the ribosome, has propelled interest in RNA understanding to a level which starts to closely resemble that given to proteins.

In these scheme 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 it's not foreign to use statistical mechanical methods, combined with detailed knowledge of atomic level structure. Such methodology is mainly used in this work in a long term effort on understanding the intrinsic structural features of RNA, and how they might relate to it's 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.

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

1.1 RNA folding

The first high resolution X-ray structure of RNA larger than a dinucleotide was that of yeast $tRNA^{Phe}$ at 3Å in 1974 [1, 2]. Thirty years later there are two orders of magnitude more RNA structural information [3], and new information is expected [4]. This fact and the discovery of ribozymes [5, 6] has renewed interest in solving the RNA folding problem, that is, from primary sequence, finding in an automated[i] way the native three-dimensional structure of RNA and its folding pathway. The RNA folding problem is usually seen as analogous to the protein folding problem, due both to the discovery of the enzymatic behavior of RNA [5, 6] and the complicated folding of large RNA molecules [10]. To take advantage of this analogy, a unified conceptual framework for describing RNA and protein folding, called the kinetic partitioning mechanism (KPM), has been developed by Thirumalai and Hyeon [11]. This and other methods are based on defining an adequate partition function for describing the correct conformational ensemble of folded, partially folded, and unfolded structures [12, 13, 14] of either protein or RNA.

1.2 Is RNA folding a hard or easy problem?

There are two trains of thought regarding RNA folding. One states that RNA folding is less complex than protein folding [15] because RNA is made up of a four letter alphabet of similar nucleotide units instead of a 20 letter alphabet of dissimilar amino acids. Therefore the number of possible sequential combinations is smaller. It is also well known that secondary and tertiary interactions can be separated in the case of RNA by the absence or presence of Mg^{2+} [16] (see Figure 1.1), whereas secondary and tertiary elements are not as easily separable in proteins. The other point of view says that RNA folding can be at least as complex as protein folding [17, 18] since there is no such thing as hydrophobic burial of regions of RNA as in the case of proteins. Instead, the electrostatic problem of having a complex charged backbone must be dealt with in the case of RNA. For instance, the interactions of the RNA polyanionic backbone with water and cations [19] are not easily simulated with explicit solvent models as can be done for proteins. The aforementioned interactions of RNA need to be modeled implicitly, and must aim to describe long dynamic processes of the order of seconds to minutes, in contrast to the typical time scales of tens of microseconds associated with protein folding. Although secondary and tertiary structure can be separated experimentally, there have been few theoretical efforts to account for the folding of RNA from a random sequence of nucleotides into secondary structures and tertiary structures. What little is know has been investigated at low resolution. Professor Stephen Harvey and associates have simulated yeast $tRNA^{Phe}$, [20] and the assembly of the 30S subunit of the ribosome [21] at various levels of detail, initially using only one pseudoatom per helical region, and later one pseudoatom per nucleotide. Recently Major's group [22] at Montreal has proposed a pipeline of two computer algorithms, one makes secondary structure predictions, and the other assembles 3D structures based on the best scoring secondary structures. The key to the success of the 3D prediction

[i] The term automated is used here to mean a theoretical model of tertiary folding, which could use experimental measures of secondary structure association in the same way that the traditional secondary structure folding model [7, 8] uses the Tinoco-Uhlenbeck dinucleotide postulate [9] to find total free energies.

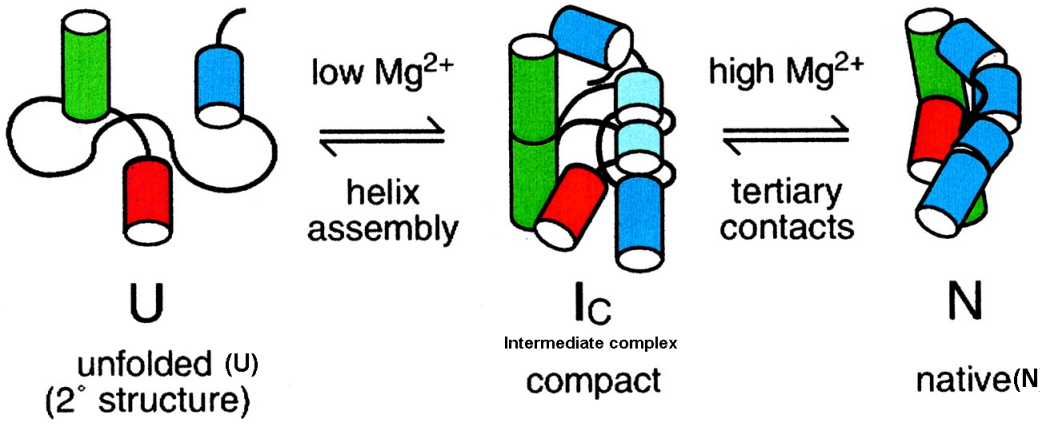

Figure 1.1: Separation of secondary and tertiary interaction in RNA [16]. Double helical secondary structure represented by individual cylinders and tertiary interactions by association of cylinders.

is attributed to what they call Nucleotide Cyclic Motifs (NCM), which are based on a graph theoretical description common to secondary and 3D structure. By contrast, in the case of proteins many groups have simulated the transition from secondary to tertiary structure, including some calculations which account for the strong coupling of secondary and tertiary structure [23, 24, 25]. This type of work is often referred to as protein structural topology and there is no counterpart for RNA.

1.3 Experimental folding techniques

Traditionally RNA folding and unfolding have been followed calorimetrically and spectroscopically as a function of temperature and cation concentration [26]. While this approach works well for studying two-state folders, *i.e.*, structures which populate only two states (native and melted), in general RNA's are not two-state folders. RNA seems to go through a rugged free energy landscape of conformations in the process of folding [27]. The experimental solution to this problem is offered by single molecule techniques like fluorescence resonance energy transfer (FRET) and mechanical micromanipulation, in which the ends of RNA are attached to micron sized beads which are then pulled apart and monitored with a laser light trap [28, 29, 30, 31]. In the case of single molecule force-induced unfolding, state transitions often occur under non-equilibrium conditions, thereby making it difficult to extract equilibrium information from the data. Recently Bustamante, Tinoco, and associates have shown that using the Crooks fluctuation theorem [32], one can deal with such cases and extract RNA folding free energies from single molecule experiments [33].

1.4 RNA simulations

Network and molecular mechanics-molecular dynamics (MM-MD) methods provide useful information relevant to the RNA folding-unfolding problem, especially for describing fluctuations away from the native conformation. Gaussian network models [34, 35, 36] which treat RNA at less than atomic detail have been used to describe the motions of large RNA structures like the ribosome. Examples of the predicted normal modes of motion of the ribosome can be seen at: http://ribosome.bb.iastate.edu/70SnK mode. Using MM, Sanbonmatsu and coworkers obtained a static atomic model of the 70S ribosome structure through homology modeling [37]. Tung and associates used this structure for an all-atom MD simulation of the movement of tRNA into a fluctuating ribosome [38]. This type of simulation might be useful in a reverse-folding approach to the RNA folding problem. To the best of our knowledge, such calculations haven't as yet been done for RNA.

1.4.1 Local nucleotide interactions

The molecular interactions which rule RNA structures at the nucleic acid base level, *i.e.*, local level, are hydrogen bonding and stacking interactions. The former are related to base pairing and the latter, in most cases, to nucleotide steps. These interactions can be explored theoretically at various levels. At the highest level are ab-initio quantum mechanical calculations which are still too expensive for systems as large as hundreds of atoms. Such calculations, nevertheless, can tell a great deal about local electronic behavior. For example, Hobza and collaborators have found that the stacking interaction of free nucleotide bases is determined by dispersion attraction, short-range exchange repulsion, and electrostatic interaction. No specific $\pi - \pi$ interactions are found from electron correlated ab-initio calculations [39, 40]. This is why force field methods have been so successful in the study of nucleic acids, since the empirical potentials used in such studies mimic well the quantum mechanically obtained energy profiles [37, 41]. A currently debated ab-initio finding is whether small fluctuations in the configurations of neighboring base pairs (dimers) are iso-energetic or not. Recent calculations of Sponer and Hobza [42] seem to contradict their older publications [41, 43], in which the stacking energies were reported to be relatively insensitive to dimer conformation. The new results use the so-called "coupled cluster singles doubles with triple electron excitations" CCSD(T) method, to account for electron correlation. Using this electron correlation energy correction, the stacking energy differences between dimer conformations turn out to be considerably higher than previously reported.

Single and double strand stacking free energies can be obtained calorimetrically. The most popular method used for obtaining such quantities is differential scanning calorimetry (DSC) [44]. These measurements show favorable dinucleotide stacking free energies as large as -3.6 kcal/mol for double strand stacking. Experimentally, the magnitudes of these interactions are found to be sequence dependent [26]. In fact, the stacking free energies for some sequences[ii] are found to be negligible. Thus there may be no accountable stacking interaction at all for some sequences.

Besides taking into account the effects of stacking and hydrogen bonding, it is important to think at the same time about the polyelectrolyte nature of the RNA backbone. Manning's counterion condensation theory [45, 46] provides a simple and quantitative picture of the interactions of the double helical nucleic acid polyanion with its counterions, although it does not take into account the discrete nature of charge [26] or the folding of RNA. Poisson-Boltzmann theory offers a more detailed picture of the behavior of charged macroions in solution [47].

The local conformational space of RNA has been studied using a large set of available RNA structures from the Nucleic Acid Database (NDB) [48]. The torsion angles of the nucleotide steps have been clustered in the parameter space using different techniques [49, 50]. The root-mean-square deviations (RMSD) of the distances between closely spaced atoms in the phosphates, sugars, and bases, have also been clustered [51]. The latter studies are aimed at finding the common nucleotide base steps and base-pair building blocks which are given the name of RNA doublets. Recently, the RNA Ontology Consortium (ROC) has proposed a consensus set of RNA dinucleotide conformers integrating the work of various groups [52].

1.4.2 RNA secondary structure algorithms and the lack of tertiary ones

From secondary structure prediction algorithms like Zuker's *mfold* program [53], Hofacker's Vienna RNA package [8], or Mathews Dynaling [54], one obtains a large ensemble of secondary structure graphs. These graphs can be analyzed with graph theory to produce a partition function describing a full arrangement of contacts for the total number of possible secondary structures making possible a "relation of microscopic conformations to macroscopic properties" [55]. So far this type of model has not been generalized to take into account tertiary structural features, *i.e.*, interhelical interactions of RNA.

[ii] Unpaired terminal nucleotides UC/A UU/A at 1M NaCl.

In the last two to three years a boom in prediction of small ($approx$ 200 nucleotides) RNA 3D structures has started. Basically three types of approaches are being followed. One is that of using a coarse grained model assigning a potential function to it, followed by a minimization procedure, and then a molecular mechanics (MM) all atom refinement [56, 57, 58]. Another starts from predicted secondary structures and assumes their helical regions adopt the A-form conformation, then mechanically thrusts residues as rigid bodies in the remaining non-helical regions, and finally carry out an MM optimization [59]. Finally, a pipeline between secondary structure prediction, and tertiary structure assembly is proposed. This pipeline uses as bridging concept between 2D and 3D structure the graph theoretical concept of a minimal cycle basis, which for the case of nucleic acids they rename as Nucleic Cyclic Motifs (NCM) [22].

1.4.3 RNA overall fold

Whereas in the case of proteins one can describe the overall fold from the arrangement of secondary structure motifs, *i.e.*, using the helix-ribbon-coil images developed by Jane Richardson [60] (see Figure 1.2), there is still no comparable description of the overall fold of RNA. A ribbon representation of the sugar phosphate backbone helps to understand the folding of small RNA's, but in the case of the ribosome this type of representation is not sufficient, see Figure 1.3.

Figure 1.2: Ribbon-coil schematic illustraring the fold and intermolecular units of a dimer of prealbumin, or transthyretin, taken from Richardson *et al.* [61]

One can envision that a thorough investigation of the parameter space of translational and rotational degrees of freedom of the helical regions of RNA could give clues as to how we might see an overall fold in RNA structures.

In the case of proteins the SCOP (Structural Classification of Proteins) database [62], classifies proteins, among other classifications, according to recurrent arrangements of secondary structure, that is, folds. The SCOR (Structural Classification of RNA) database [63, 64], aims to provide a similar classification to that obtained for proteins, but using RNA motifs[iii] instead. This classification focuses on the local folding of small pieces of RNA and cannot describe the overall fold.

1.4.4 RNA motifs

The current review on RNA motifs spans (mainly) the first decade of the XXI century. It is arranged in chronological order from the date of the most recent publication to the oldest one. The header indicates the group leader and current location.

[iii] Leontis and Westhof [65] define RNA motifs as: "Directed and ordered arrays of non-WC (Watson-Crick) base-pairs forming distinctive foldings of the phosphodiester backbones of the interacting RNA strands"

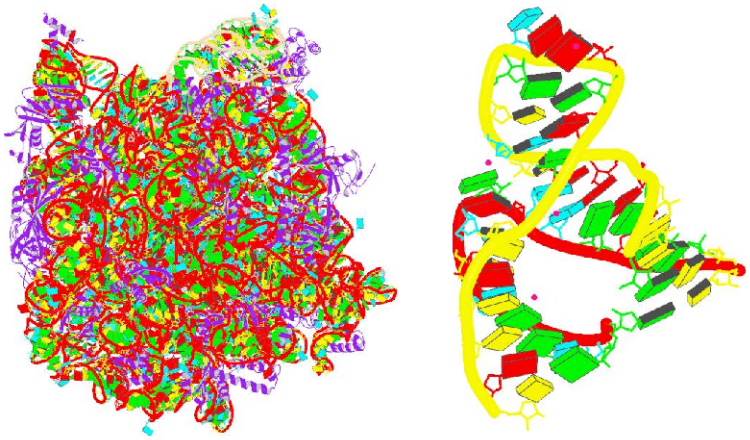

Figure 1.3: *Haloharcula marismortui's* large ribosomal subunit (left) and hammerhead ribozyme (right).The figures were taken directly from the NDB web pages, and show a ribbon representation of the phosphate backbone, and a block representation for the nucleotide bases. From the figures it's clear that, whereas the ribozyme fold can be clearly understood with this representation, the ribosome fold cannot.

- **SCHLICK GROUP at NYU.**

Analysis of four-way junctions and higher order junctions. Using Leontis group software, FR3D, junctions of order four or higher are found and then are classified according to whether they form coaxial stacks or not, and so on. It seems like most of the analysis is based on visual inspection. (2009) [66] [67]

- **LEONTIS GROUP at Bowling Green U.**

Book chapter which brings together many of Leontis papers into the software named FR3D, which unfortunately is coded in matlab and makes a GUI windows executable which is very hard to adapt to efficient analysis of many structures through command line scripts. They also provide new definitions for RNA motifs extending the vocabulary for naming them, that is, they start using terms such as "*3D structural RNA motifs*", and "*modular RNA 3D motifs*", to further distinguish them from sequence alone motifs, or from secondary structure motifs. Another new contribution is to explicitly show examples where sequences of different length form the same motif, as is the case with the GNRA loop. Finally, for differentiating between functional motifs, and structural motifs they show the example of the 23S subunit of rRNA for two different species, where in one case part of a helix is smaller than in the other, but nonetheless the geometry is the same where helices 63 and 101 are contacting each other, therefore presenting a somewhat self-contradictory statement since the final emphasis is put into the three dimensional configuration (structure determining function) which is retained. (2009) [68]

- **SCHROEDER GROUP at Oklahoma U.**

Gives an up to date description of the main programs and algorithms used for RNA secondary structure prediction. The article focuses on the major groups, i.e., Turner-Mathews, Zucker, Hofacker-Stadler, Major, and it also gives the location of their software and a good list of online tools. (2009) [69]

- **FRENKEL GROUP at UCSF.**

ISfold is a matlab program for examination of patterns of nucleotide substitutions from sequence alignments or mutation experiments. It can identify plausible base pair interactions. It Identifies the existence of non-WC base pairs within RNA bulges, internal loops, and hairpin loops, structures that cannot be easily predicted with existing algorithms. The IS in ISfold stands for iso-steric in the same sense as

isostericity is treated by Leontis-Westhof et al. The main author of this paper is now working for Leontis. So it's clear that FR3D is based on ISfold. They are stuck with matlab, which eventually will prove a big hurdle for automatic fast analysis of large databases. (2008) [70]

- **ALLAIN GROUP at ETH Zurich.**

This article deals mainly with RNA Recognition Motifs (RRM). It's important to note here that RRM's are proteins, not RNAs. In this review structural data show that binding affinity and specificity of RRM-RNA and RRM-protein interactions produce structural versatility which explains why proteins that have RRMs have a diverse range of biological functions. (2008) [71]

- **GIEGERICH GROUP at Bielefeld U.**

From the online version of the software:
"*Locomotif is a GUI-based program that allows for the visual design of RNA motifs. The graphical structures are then translated into executable programs to be used for searching a motif in a sequence (plain text or FASTA format)*".
From this description is clear that Locomotif is a secondary structure to sequence motif finder, not a 3D structure motif finder. (2007) [72]

- **MAJOR GROUP at University of Montreal.**

RNA secondary structures are described as graphs with the intention of finding minimun cycle basis in RNA 3D structures using a common graph theoretical algorithm known as Horton's algorithm. Once the minimum cycles are found they can are clustered using single linkage and a so called "*cycle distance metric*" which correlates with the common RMSD metric. The resulting cycles are called cyclic motifs, and later (2008) they have been renamed as nucleotide cyclic motifs (NCM's) and used as the main idea for generating RNA 3D structure predictions from sequence alone. It's interesting to note that the results obtained are not very dependent on backbone conformations but mainly on base-pairing and stacking. [73]

- **SPONER GROUP at Academy of Sciences of the Czech R.**

Molecular dynamics (MD) simulations of the Sarcin-Ricin Domain (SRD) motifs from 23S (E. coli) and 28S (rat) rRNAs using AMBER6 and AMBER7. Unusual stiffness of rRNA building blocks in 25ns simulations, as well as intrinsic structural and dynamical signatures that distinguish them from other rRNA motifs such as Loop E and Kink-turns. (2006) [74]

- **RUZZO GROUP at U. of Washington.**

CMfinder. Software for finding RNA sequence motifs in unaligned sequences. CMfinder uses a Bayesian framework which handles information and folding energy based approaches to predict sequence "structure" in a so-called principled way. The implemented methods work for high and low sequence similarities. (2006) [75]

- **HOLBROOK at LBNL.**

RNA motif definition:
"*Conserved structural subunits that make up the secondary structures of RNAs.*".
Review of identification and classification. They state that structural motifs are held together by tertiary interactions, and are different from sequence or functional motifs. The article discusses the biological roles of functional motifs, binding motifs and their function when complexed with metals and other ligands, and the relationship between sequential and structural motifs in tracing phylogenetic relationships in RNA engineering. (2005) [76] [77]

- **SCHLICK GROUP at NYU.**
A protocol for searching genomes of a set of organisms to find RNA sequences based on pre-defined patterns (In this case aptamer patterns). Once the sequence hits are obtained they are folded into secondary structures using the Vienna package. The resulting sets of secondary structures are validated with statistical significance and thermodynamic stability. (2005) [78]

- **WESTHOF GROUP at Louis Pasteur U.**
Kink-turn and C-loop are two recurrent motifs in ribosomal RNA sequences. These two motifs are analyzed in crystal structures and are compared to sequence alignments of rRNAs from the three kingdoms of life to identify the range of the structural and sequence variations. The sequence variations of the non-Watson-Crick base pairs for each motifs are analyzed using isostericity matrices. These matrices are useful for deriving sequence signatures of recurrent motifs as well as determining the motif conservation through evolution. The observed conservations are helpful in identifying motifs in sequences. (2005) [79]

- **FOX GROUP at U. of Houston.**
Clustering weighted RMSD's for loops (5 to 13 nucleotides) recognized by using a reduced set of atoms per nucleotide, that is, this is not an all atom RMSD. For clustering they use average linkage (UPMGA). (2005) [80]

- **BRENNER GROUP at UC Berkeley.**
A database of secondary structure based motifs which are split into three main classification schemes which are: Structure, Function, and Tertiary Interaction. SCOR stands for Structural Classification of RNA's. It would be better if perhaps it was called secondary structure classification of RNA's. Nonetheless their Tertiary interaction classification matches in some way the RNA motif definitions and one can run a query on a pdbid to see a motif finding result. (2004) [64]

- **LEONTIS at Bowling Green U.**
RNA motif definition:
"*Ordered stacked arrays of non-Watson-Crick base pairs that form distinct folds on the phosphodiester backbones of RNA strands.*"
Motifs are characterized by all sequences that make up identical three-dimensional structures. Review of hairpin loops, asymmetric internal loops (A-minor motifs, K-turn motifs, Sarcin-like motif, C-motif), symmetric internal loops (Chloroplast 5S rRNA loop E), junction loops (Hook-turn motif). (2003) [65]

- **SPONER GROUP at Academy of Sciences of the Czech R.**
MD of non-canonical WC and hydration in RNA motifs. Their experiment involved a total of over 80 ns on bacterial and spinach chloroplast 5S rRNA loop E motifs. (2003) [81]

- **LEONTIS at Bowling Green U.**
Dictates the steps involved in the analysis and annotation of RNA motifs in 3-dimensional structures in detail (Later in 2009 in a book called Non-protein coding RNA's this article is practically reconstructed in the Leontis Group chapter). Annotation involves decomposition of each motif into non-WC base pairs, geometric classification of each base-pair, identification of isosteric substitutions, alignment of homologous sequences, and acceptance or rejection of the null hypothesis that the motif is conserved. (2002) [82]

- **LEONTIS at Bowling Green U.**

They describe in detail the concept of isostericity matrices and how they serve the purpose of describing non-WC base-pairs. They include a very long list of RNA base-pair structures classified according to the isostericity concept. (2002) [83]

- **ZACHARIAS now at Technische Universitat Munchen**
General musings on “non-helical" RNA motifs with no experimental or theoretical work, just musings. (2000) [84]

- **MOORE at Yale U.**
RNA motif definition:
“*An RNA Motif is a discrete sequence or combination of base juxtapositions found in naturally occurring RNA’s in unexpectedly high abundance.*”
Motifs are classified as being inside three possible groups. These are: Terminal loop motifs (U-turns, tetraloops), Internal loop motifs (cross-strand purine stacks, bulged-G, A-platforms, bulge-helix-bulge, metal binding) and tertiary motifs (ribose zippers, tetraloop-helix). (1999) [85]

- **PYLE at Yale U.** Pyle and Duarte re-discover RNA backbone virtual torsion angles ω_v and $\omega_{v'}$, and rename them η and θ, they further produce scatterplots of η vs. θ as Malathi and Yathindra did for yeast tRNAphe. They implement an automated software for generating η, and θ angles called PRIMOS. A detailed account of the re-discovery is made by Leontis and Westhof in 2003. (1998) [86]

References

[1] Robertus, J. D., Ladner, J. E., Finch, J. T., Rhodes, D., Brown, R. S., Clark, B. F. C., and Klug, A. (1974) Structure of Yeast Phenylalanine tRNA at 3 Å Resolution. *Nature,* **250**, 546.

[2] Kim, S. H. (1974) Three-Dimensional Tertiary Structure of Yeast Phenylalanine Transfer RNA. *Science,* **185**, 435.

[3] Noller, H. F. (2005) RNA Structure: Reading the Ribosome. *Science,* **309**, 1508–1514.

[4] Weinberg, Z., Perreault, J., Meyer, M. M., and Breaker, R. R. (2009) Exceptional Structured Noncoding RNAs Revealed by Bacterial Metagenome Analysis. *Nature,* **462**, 656–659.

[5] Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E., and Cech, T. R. (1982) Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *Cell,* **31**, 147–157.

[6] Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., and Altman, S. (1983) The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme.. *Cell,* **35**, 849–857.

[7] Zuker, M. (1989) On Finding all Suboptimal Foldings of an RNA Molecule. *Science,* **244**, 48–52.

[8] Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P. (1994) Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte fur Chemie,* **125**, 167–188.

[9] Borer, P. N., Dengler, B., Tinoco, J. I., and Uhlenbeck, O. C. (1974) Stability of ribonucleic acid double-stranded helices. *Journal of Molecular Biology,* **86**, 843–853.

[10] Batey, R. T., Rambo, R. P., and Doudna, J. A. (1999) Tertiary Motifs in RNA Structure and Folding. *Angewandte Chemie International Edition,* **38**(16), 2326–2343.

[11] Thirumalai, D. and Hyeon, C. (2005) RNA and Protein Folding: Common Themes and Variations. *Biochemistry,* **44**, 4957–4970.

[12] Chen, S.-J. and Dill, K. A. (1995) Statistical thermodynamics of double-stranded polymer molecules. *Journal of Chemical Physics,* **103**, 5802–5813.

[13] Chen, S.-J. and Dill, K. A. (1998) Theory for the conformational changes of double-stranded chain molecules. *Journal of Chemical Physics,* **109**, 4602–4616.

[14] Thirumalai, D. and Woodson, S. A. (1996) Kinetics of Folding of Proteins and RNA. *Accounts in Chemical Research,* **29**, 433–439.

[15] Tinoco, I. and Bustamante, C. (1999) How RNA folds. *Journal of Molecular Biology,* **293**(2), 271–281.

[16] Rangan, P., Masquida, B., Westhof, E., and Woodson, S. A. (2003) Assembly of core helices and rapid tertiary folding of a small bacterial group I ribozyme. *Proceedings of the National Academy of Sciences of the United Stated of America,* **100**, 1574–1579.

[17] Moore, P. B. The RNA World chapter The RNA Folding Problem, pp. 381–401 Cold Spring Harbor Laboratory Press 2nd edition (1999).

[18] Sorin, E. J., Nakatani, B. J., Rhee, Y. M., Jayachandran, G., Vishal, V., and Pande, V. S. (2004) Does Native State Topology Determine the RNA Folding Mechanism?. *Journal of Molecular Biology,* **337**, 789–797.

[19] Klein, D. J., Moore, P. B., and Steitz, T. A. (2004) The contribution of metal ions to the structural stability of the large ribosomal subunit.. *RNA,* **10**(9), 1366–1379.

[20] Malhotra, A., Tan, R. K., and Harvey, S. C. (1990) Prediction of the Three-Dimensional Structure of Escherichia Coli 30S Ribosomal Subunit: A Molecular Mechanics Approach.. *Proceedings of the National Academy of Sciences of the United States of America,* **87**, 1950–1954.

[21] Stagg, S. M., Mears, J. A., and Harvey, S. C. (2003) A Structural Model for the Assembly of the 30 S Subunit of the Ribosome. *Journal of Molecular Biology,* **328**, 49–61.

[22] Parisien, M. and Major, F. (2008) The MC-Fold and MC-Sym Pipeline Infers RNA Structure from Sequence Data. *Nature,* **452**, 51–55.

[23] Westhead, D., Slidel, T., Flores, T., and Thornton, J. (1999) Protein structural topology: Automated analysis and diagrammatic representation. *Protein Science,* **8**, 897–904.

[24] Gerstein, M. and Thornton, J. M. (2003) Sequences and Topology. *Current Opinion in Structural Biology,* **13**, 341–343.

[25] Meiler, J. and Baker, D. (2003) Coupled Prediction of Protein Secondary and Tertiary Structure. *Proceedings of the National Academy of Sciences of the United States of America,* **100**, 12105–12110.

[26] Bloomfield, V. A., Crothers, D. M., and Jr., I. T. (2000) Nucleic Acids: Structures, Properties and Functions, University Science Books, .

[27] Zhuang, X. and Rief, M. (2003) Single-Molecule Folding. *Current Opinion in Structural Biology,* **13**, 88–97.

[28] Liphardt, J., Onoa, B., Smith, S., Jr., I. T., and Bustamante, C. (2001) Reversible unfolding of single RNA molecules by mechanical force.. *Science,* **292**, 733–737.

[29] Onoa, B. and Jr., I. T. (2004) RNA folding and unfolding. *Current Opinion in Structural Biology,* **14**(3), 374–379.

[30] Tinoco, I. (2004) FORCE AS A USEFUL VARIABLE IN REACTIONS: Unfolding RNA.. *Annual Review of Biophysics & Biomolecular Structure,* **33**, 363–385.

[31] Hyeon, C. and Thirumalai, D. (2005) Mechanical unfolding of RNA hairpins. *Proceedings of the National Academy of Science,* **102**(19), 6789–6794.

[32] Crooks, G. E. (1999) Entropy production fluctuation theorem and the nonequilirium work relation for free-energy differences. *Physical Review E,* **60**, 2721–2726.

[33] Collin, D., F.Ritort, Jarzynski, C., Smith, S. B., Jr., I. T., and Bustamante, C. (2005) Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies. *Nature,* **437**, 231–234.

[34] Wang, Y., Rader, A. J., Bahar, I., and Jernigan, R. L. (2004) Global ribosome motions revealed with elastic network model. *Journal of Structural Biology,* **147**, 302–314.

[35] Bahar, I. and Jernigan, R. L. (1998) Vibrational dynamics of transfer RNAs: comparison of the free and synthetase-bound forms. *Journal of Molecular Biology,* **281**, 871–884.

[36] Wang, Y. and Jernigan, R. L. (2005) Comparison of tRNA Motions in the Free and Ribosomal Bound Structures. *Biophysical Journal,* **89**, 3399–3409.

[37] Tung, C.-S. and Sanbonmatsu, K. Y. (2004) Atomic Model of the Thermus thermophilus 70S Ribosome Developed in Silico. *Biophysical Journal,* **87**, 2714–2722.

[38] Sanbonmatsu, K. Y., Simpson, J., and Tung, C.-S. (2005) Simulating movement of tRNA into the ribosome during decoding. *Proceedings of the National Academy of Sciences,* **102**, 15854–15859.

[39] Sponer, J., Leszczynski, J., and Hobza, P. (1996) Nature of Nucleic Acid-Base Stacking: Nonempirical ab Initio and Empirical Potential Characterization of 10 Stacked Base Dimers. Comparison of Stacked and H-Bonded Base Pairs. *Journal of Physical Chemistry,* **100**, 5590–5596.

[40] Sponer, J., Leszczynski, J., and Hobza, P. (1997) Thioguanine and Thiouracil: Hydrogen-Bonding and Stacking Properties. *Journal of Physical Chemistry A,* **101**, 9489–9495.

[41] Sponer, J., Berger, I., Spackova, N., Leszczynski, J., and Hobza, P. (2000) Aromatic Base Stacking in DNA: From ab initio Calculations to Molecular Dynamics Simulations. *Journal of Biomolecular Structure and Dynamics,* **11**, 1–24.

[42] Sponer, J., Jureka, P., Marchan, I., Luque, F. J., Orozco, M., and Hobza, P. (2006) Nature of Base Stacking: Reference Quantum-Chemical Stacking Energies in Ten Unique B-DNA Base-Pair Steps.. *Chemistry - A European Journal,*.

[43] Hobza, P. and Sponer, J. (2002) Toward True DNA Base-Stacking Energies: MP2, CCSD(T), and Complete Basis Set Calculations. *Journal of the American Chemical Society,* **124**, 11802–11808.

[44] Marky, L. A. and Breslauer, K. J. (1982) Calorimetric determination of base-stacking enthalpies in double-helical DNA molecules. *Biopolymers,* **11**, 2185–2194.

[45] Manning, G. S. (1977) Limiting laws and counterion condensation in polyelectrolyte solutions. IV. The approach to the limit and the extraordinary stability of the charge fraction.. *Biophysical Chemistry,* **7**, 95–102.

[46] Manning, G. S. (2003) Comments on Selected Aspects of Nucleic Acid Electrostatics. *Biopolymers,* **69**, 137–143.

[47] Antypov, D., Barbosa, M. C., and Holm, C. (2005) Incorporation of excluded-volume correlations into Poisson-Boltzmann theory. *Physical Review E.,* **71**(6), 1–6.

[48] Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S. H., Srinivasan, A. R., and Schneider, B. (1992) The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids.. *Biophysical Journal,* **63**, 751–759.

[49] Murray, L. J. W., III, W. B. A., Richardson, D. C., and Richardson, J. S. (2003) RNA Backbone is Rotameric. *Proceedings of the National Academy of Sciences of the United States of America,* **100**, 13904–13909.

[50] Schneider, B., Moravek, Z., and Berman, H. (2004) RNA Conformational Classes. *Nucleic Acids Research,* **32**, 1666–1677.

[51] Sykes, M. T. and Levitt, M. (2005) Describing RNA Structure by Libraries of Clustered Nucleotide Doublets. *Journal of Molecular Biology,* **351**, 26–38.

[52] Richardson, J. S., Schneider, B., Murray, L. W., Kapral, G. J., Immormino, R. M., Headd, J. J., Richardson, D. C., Ham, D., Hershkovits, E., Williams, L. D., Keating, K. S., Pyle, A. M., Micallef, D., Westbrook, J., , and Berman, H. M. (2008) RNA Backbone: Consensus All-Angle Conformers and Modular String Nomenclature (An RNA Ontology Consortium Contribution). *RNA,*.

[53] Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research,* **31**(13), 3406–3415.

[54] Mathews, D. H. and Turner, D. H. (Mar, 2002) Dynalign: An Algorithm for Finding the Secondary Structure Common to Two RNA Sequences.. *Journal of Molecular Biology,* **317**(2), 191–203.

[55] Chen, S.-J. and Dill, K. A. (2000) RNA Folding Energy Landscapes. *Proceedings of the National Academy of Sciences of the United States of America,* **97**, 646–651.

[56] Das, R. and Baker, D. (Sep, 2007) Automated de Novo Prediction of Native-Like RNA Tertiary Structures. *Proceedings of the National Academy of Sciences of the United States of America,* **104**(37), 14664–14669.

[57] Ding, F., Sharma, S., Chalasani, P., Demidov, V. V., Broude, N. E., and Dokholyan, N. V. (Jun, 2008) Ab Initio RNA Folding by Discrete Molecular Dynamics: From Structure Prediction to Folding Mechanisms. *RNA,* **14**(6), 1164–1173.

[58] Jonikas, M. A., Radmer, R. J., and Altman, R. B. (Dec, 2009) Knowledge-Based Instantiation of Full Atomic Detail Into Coarse-Grain RNA 3D Structural Models. *Bioinformatics,* **25**(24), 3259–3266.

[59] Martinez, H. M., Jr, J. V. M., and Shapiro, B. A. (2008) RNA2D3D: A program for Generating, Viewing, and Comparing 3-Dimensional Models of RNA. *Journal of Biomolecular Structure and Dynamics,* **25**, 573–752.

[60] Richardson, J. S. (2000) Early ribbon drawings of proteins. *Nature Structural Biology,* **7**, 624–625.

[61] Richardson, D. C. and Richardson, J. S. (2002) Teaching Molecular 3-D Literacy. *Biochemistry and Molecular Biology Education,* **30**, 21–26.

[62] Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J. P., Chothia, C., and Murzin, A. G. (2004) SCOP Database in 2004: Refinements Integrate Structure and Sequence Family Data. *Nucleic Acids Research,* **32**, D226–D229.

[63] Klosterman, P. S., Tamura, M., Holbrook, S. R., and Brenner, S. E. (2002) SCOR: a Structural Classification of RNA Database. *Nucleic Acids Research,* **30**, 392–394.

[64] Klosterman, P. S., Hendrix, D. K., Tamura, M., Holbrook, S. R., and Brenner., S. E. (2004) Three-Dimensional Motifs from the SCOR, Structural Classification of RNA Database: Extruded Strands, Base Triples, Tetraloops and U-turns. *Nucleic Acids Research,* **32**(8), 2342–2352.

[65] Leontis, N. B. and Westhof, E. (2003) Analysis of RNA Motifs. *Current Opinion in Structural Biology,* **13**, 300–308.

[66] Laing, C., Jung, S., Iqbal, A., and Schlick, T. (Oct, 2009) Tertiary Motifs Revealed in Analyses of Higher-Order RNA Junctions. *Journal of Molecular Biology,* **393**(1), 67–82.

[67] Laing, C. and Schlick, T. (Jul, 2009) Analysis of Four-way Junctions in RNA Structures. *Journal of Molecular Biology,* **390**(3), 547–559.

[68] Nasalean, L., Stombaugh, J., Zirbel, C. L., and Leontis, N. B. Vol. 13, of Springer Series in Biophysics chapter 1, pp. 1–26 Springer-Verlag Berlin Heidelberg (November, 2009).

[69] Schroeder, S. J. (Jul, 2009) Advances in RNA Structure Prediction from Sequence: New Tools for Generating Hypotheses About Viral RNA Structure-Function Relationships. *Journal of Virology,* **83**(13), 6326–6334.

[70] Mokdad, A. and Frankel, A. D. (April, 2008) ISFOLD: Structure Prediction of Base-pairs in Non-helical RNA Motifs From Isostericity Signatures in Their Sequence Alignments. *Journal of Biomolecular Structure and Dynamics,* **25**(5), 467–472.

[71] Cléry, A., Blatter, M., and Allain, F. H.-T. (June, 2008) RNA Recognition Motifs: Boring? Not Quite. *Current Opinion in Structural Biology,* **18**(3), 290–298.

[72] Reeder, J., Reeder, J., and Giegerich, R. (Jul, 2007) Locomotif: From Graphical Motif Description to RNA Motif Search. *Bioinformatics,* **23**(13), i392–i400.

[73] Lemieux, S. and Major, F. (2006) Automated Extraction and Classification of RNA Tertiary Structure Cyclic Motifs. *Nucleic Acids Research,* **34**, 2340–2346.

[74] Spacková, N. and Sponer, J. (2006) Molecular Dynamics Simulations of Sarcin-Ricin rRNA Motif. *Nucleic Acids Research,* **34**(2), 697–708.

[75] Yao, Z., Weinberg, Z., and Ruzzo, W. L. (Feb, 2006) CMfinder–A Covariance Model Based RNA Motif Finding Algorithm. *Bioinformatics,* **22**(4), 445–452.

[76] Hendrix, D. K., Brenner, S. E., and Holbrook, S. R. (2005) RNA Structural Motifs : Building Blocks of a Modular Biomolecule. *Quarterly Reviews of Biophysics,* **38**, 221.

[77] Holbrook, S. R. (2005) RNA Structure: The Long and the Short of it. *Current Opinion in Structural Biology,* **15**, 302–308.

[78] Laserson, U., Gan, H. H., and Schlick, T. (2005) Predicting Candidate Genomic Sequences that Correspond to Synthetic Functional RNA Motifs. *Nucleic Acids Research,* **33**(18), 6057–6069.

[79] Lescoute, A., Leontis, N. B., Massire, C., and Westhof, E. (2005) Recurrent Structural RNA Motifs, Isostericity Matrices and Sequence Alignments. *Nucleic Acids Research,* **33**(8), 2395–2409.

[80] Huang, H.-C., Nagaswamy, U., and Fox, G. E. (2005) The Application of Cluster Analysis in the Intercomparison of Loop Structures in RNA. *RNA,* **11**, 412–423.

[81] Réblová, K., Spacková, N., Stefl, R., Csaszar, K., Koca, J., Leontis, N. B., and Sponer, J. (Jun, 2003) Non-Watson-Crick Basepairing and Hydration in RNA Motifs: Molecular Dynamics of 5S rRNA Loop E. *Biophysical Journal,* **84**(6), 3564–3582.

[82] Leontis, N. B. and Westhof, E. (2002) The Annotation of RNA Motifs. *Comparative and Functional Genomics,* **3**, 518–524.

[83] Leontis, N. B., Stombaugh, J., and Westhof, E. (2002) The Non-Watson-Crick Base Pairs and their Associated Isostericity Matrices. *Nucleic Acids Research,* **30**, 3497–3531.

[84] Zacharias, M. (Jun, 2000) Simulation of the Structure and Dynamics of Nonhelical RNA Motifs. *Current Opinion in Structural Biology,* **10**(3), 311–317.

[85] Moore, P. B. (1999) Structural Motifs in RNA. *Annual Review of Biochemistry,* **68**, 287–300.

[86] Duarte, C. M. and Pyle, A. M. (1998) Stepping Through an RNA Structure: A Novel Approach to Conformational Analysis. *Journal of Molecular Biology,* **284**, 1465–1478.

Chapter 2
RNA Base Steps

This chapter deals with how starting from a backbone based view of RNA, we can make an interpretation at the step level using the block model.

2.1 Consensus Clustering of Single Stranded Base Step Parameters

2.2 Four Major Non-ARNA Step Groups in the Ribosome

Chapter 3
RNA Base-Pairing

The RNA base-pairs are reviewed again.

3.1 Canonical and Noncanonical Base-pairs, Methods Paper

3.2 Clustering of Yurong's Classification

Chapter 4
RNA Base Pair Steps

4.1 Analysis (Albany Poster) and Django Webserver

4.2 Persistence Length vs. Hagerman

4.3 AMBER: Persistence Length of Base-Pair Step Patterns

Chapter 5
RNA Motifs

5.1 RNA *Structural* Motifs

The following popular definitions of what an "*RNA structural motifs*" is, can be found in recent literature:

- RNA motifs are "*Conserved structural subunits that make up the secondary structures of RNAs.*"[1]
- RNA motifs are "*Ordered stacked arrays of non-Watson-Crick base pairs that form distinct folds on the phosphodiester backbones of RNA strands.*"[2]
- "*An RNA Motif is a discrete sequence or combination of base juxtapositions found in naturally occurring RNA's in unexpectedly high abundance.*"[3]

First, a word of caution must be given to the reader. The term "*RNA motif*" alone, can be used to describe three different levels of RNA organization, that is, RNA sequence motifs, RNA secondary structure motifs, or RNA 3D structure motifs. We start by making such distinction as it is not always clearly mentioned in RNA literature, generating a great deal of confusion and bibliographical search frustration for the beginner. In the remaining of this text it is to be understood that RNA structural motifs refer to specific geometrical arrangements in three dimensional space.

As can be seen from the previous definitions, and from the introduction, there is no unique, or consensus definition of what an RNA structural motif is yet, and it seems like every researcher has it's own, even if they don't declare them. The RNA Ontology Consortium (ROC) has not come to a consensus definition or RNA structural motifs neither. The majority of their work has been centered at understanding RNA backbone conformations, and the influence of isosteric substitutions on RNA structure. The ROC has yet to address the relation of base-staking to RNA structural motifs, which leaves a natural space for the rigid-block interpretation of nucleic acids to fill in.

5.2 GNRA tetraloop

In order to compare our work to that of others on RNA structural motif localization and discovery, we ask the following questions:

1. Can the geometric rigid-block description of base-pairing and base-stacking solve the problem of defining RNA structural motifs?
2. Can we use quantities derived from the 3DNA software package to make and automatic search for a known motif, for example, the GNRA tetraloop motif, and perhaps find unknown motifs?

In the ROC meeting of May, 2009 a reduced dataset of RNA structures found at:
http://docs.google.com/Doc?id=dhrmkfmn_13ftpbjcgq
was made available to participants with the purpose of allowing them to search for RNA motifs, which would later be compared between groups.

We have modestly, and as of yet unsuccessfully, started to aim at solving question number two. Initially we are trying to identify all instances of the well known GNRA tetraloop motif in the 23S subunit

of ribosomal RNA of *Thermus Thermophilus*, PDB-ID:1ffk using results from 3DNA and 3DNA-Parser, and using an automated process which could be later reproduced for any desired dataset. Our hope is that these baby steps will allow us to to tackle the whole ROC dataset.

5.2.1 3DNA-Parser

We started by using Dr. Yurong Xin's 3DNA-Parser hoping that the description of the enclosing base pair in the loop, that is, the sheared G·A, would have a characteristic signature. We found that such is not the case. We know from Major et al. [4] that there should be at least 21 GNRA tetraloops in the 23S subunit of rRNA. We used the G2696 N2697 R2698 A2699 tetraloop as a seed (as can be seen in Figure 1.1) and found out that according to Dr. Xin's helical classification the enclosing G is classified as S_{hq} and A is classified as H_e. We then searched all such instances for G·A base-pairs and we found

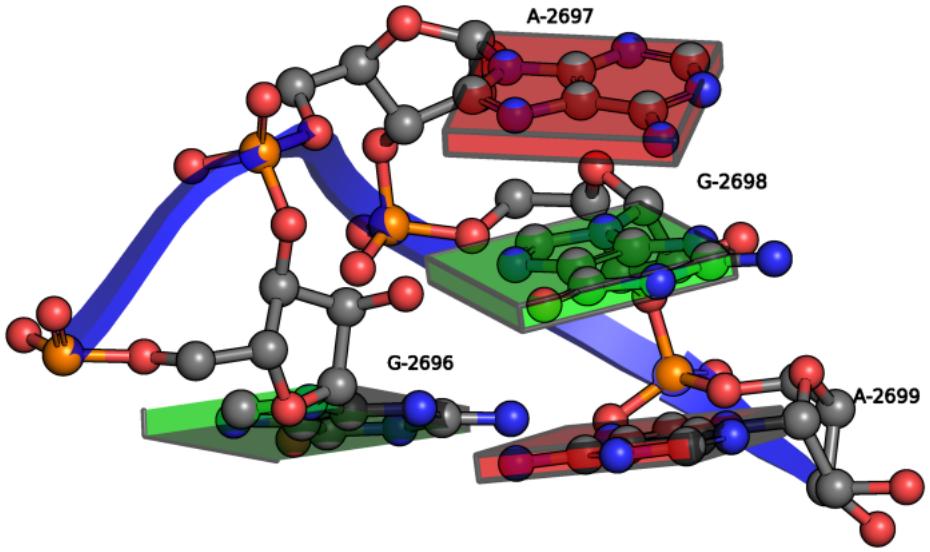

Figure 5.1: GNRA Tetraloop from *Thermus Thermophilus* 23S Ribosomal RNA PDB-ID:1ffk.

seven hits, but none were in fact GNRA tetraloops.

5.2.2 Overlap Scores

We clustered the overlap values impossing a cutoff of values of [1-8]. There are many values which are exactly zero (33%), so, without the cutoff the zero values "overshadow" the data. For this case we obtained a "good" dendrogram as seen in Figure 1.2.

The next step in this analysis will be to find the structures which correspond to this clusters and superimpose and align them using Kabsh's algorithm to be able to determine their RMSD's.

Many people start their RNA Motif identification and classification algorithms by splitting RNA structures into what is helical and what is not, and then finding interactions between these two groups. We believe that we could do a similar exercise with 3DNA by using the scalar product of helical axis vectors and once helical and non-helical regions are found we might be able to use 3DNA Parser to look for characteristic interactions.

5.3 Triplets on RNA (comparison to Laing et al.)

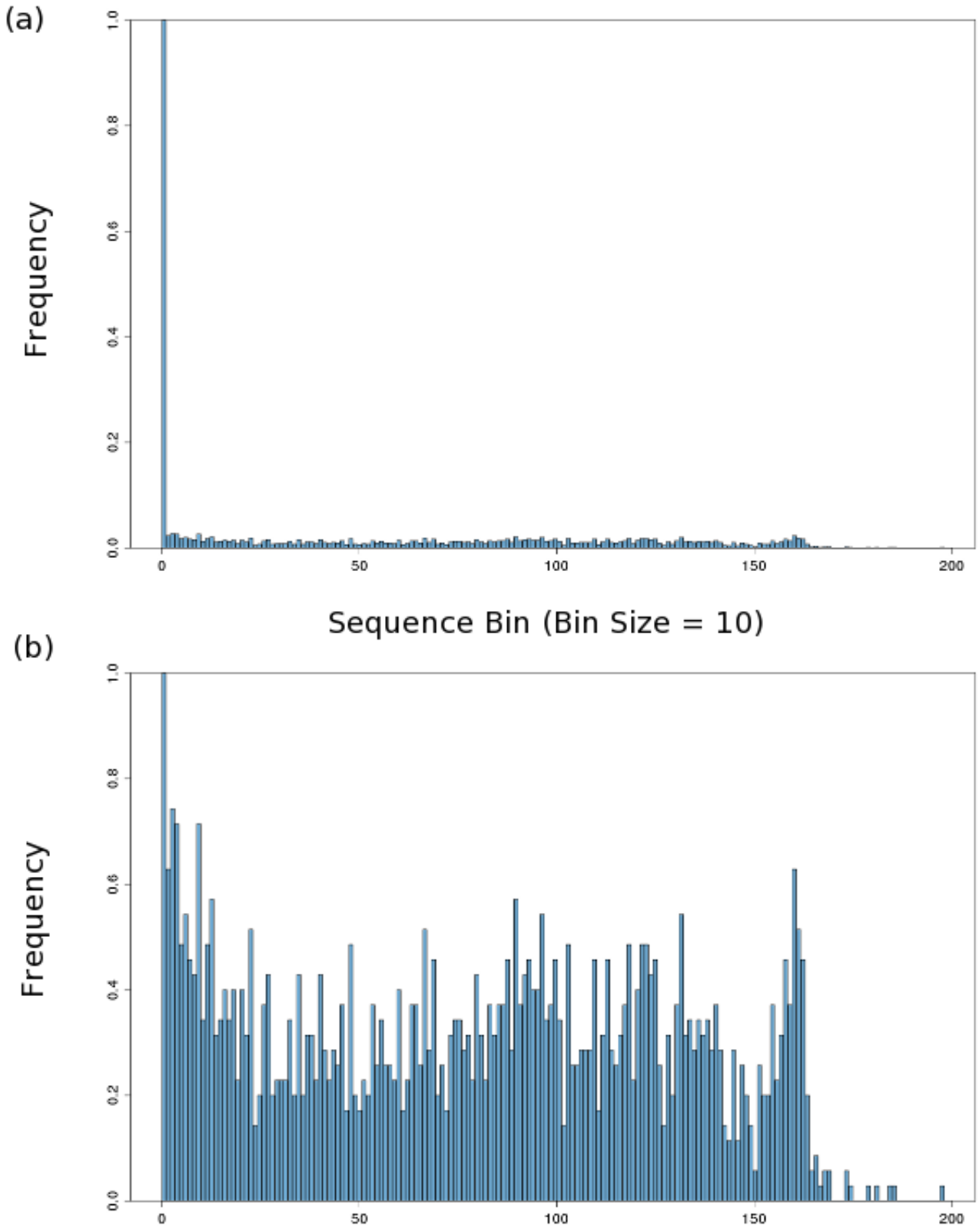

Figure 5.2: Normalized histograms showing the distribution of overlap values in the 23S subunit or *Thermus Thermophilus* rRNA, PDB-ID:1jjk. In histogram (a) all values are included, but in histogram (b) only values greater than zero are included. Notice the high preponderance of zero values, exactly 897 out of a total of 2705.

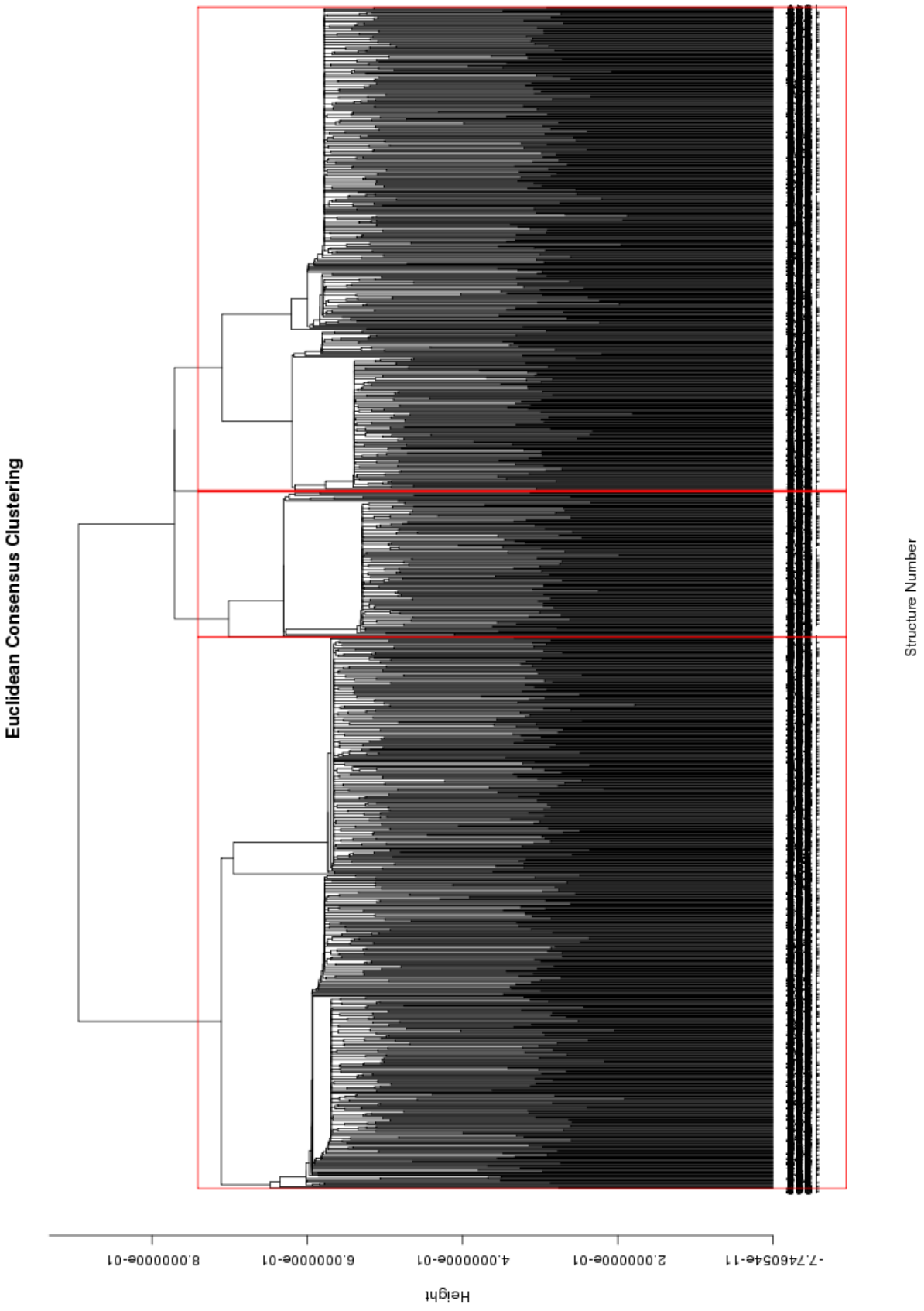

Figure 5.3: Dendrogram for consensus clustering of overlap scores in the ribosome. Zero values filtered out and remaining data normalized.

References

[1] Holbrook, S. R. (2005) RNA Structure: The Long and the Short of it. *Current Opinion in Structural Biology,* **15**, 302–308.

[2] Leontis, N. B. and Westhof, E. (2003) Analysis of RNA Motifs. *Current Opinion in Structural Biology,* **13**, 300–308.

[3] Moore, P. B. (1999) Structural Motifs in RNA. *Annual Review of Biochemistry,* **68**, 287–300.

[4] Lemieux, S. and Major, F. (2006) Automated Extraction and Classification of RNA Tertiary Structure Cyclic Motifs. *Nucleic Acids Research,* **34**, 2340–2346.

Chapter 6
RNA Helical Regions and Graph Theory

Chapter on RNA Helical Region Recognition and description using graph theoretical descriptors.

Appendix A
Clustering Analysis (CA)

A.1 Hierarchical methods

The hierarchical clustering methods used were:

1. *Single linkage clustering*, where the minimum distance between elements of each cluster is taken as clustering criteria.

$$D(X,Y) = min\{d(x_i, y_j) : x_i \in X, y_j \in Y\} \tag{A.1}$$

   where X and Y are vectors, and $d(x_i, y_j)$ is the distance between cluster elements.

2. *Complete linkage clustering*, where the maximum distance between cluster elements is the clustering criteria.

$$D(X,Y) = max\{d(x_i, y_j) : x_i \in X, y_j \in Y\} \tag{A.2}$$

3. *Average linkage clustering*, the mean distance between elements of each cluster is taken as clustering criteria.

$$D(X,Y) = \frac{1}{N_x * N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} d(x_i, y_j) \tag{A.3}$$

   where N_x and N_y are the number of elements in respective clusters.

4. *Centroid linkage clustering*, uses the distance between cluster centroids, as clustering criteria.

$$D(X,Y) = d(\overline{x}, \overline{y}) \tag{A.4}$$

$$\overline{x} = \frac{1}{N_x} \sum_{i=1}^{N_x} x_i \tag{A.5}$$

$$\overline{y} = \frac{1}{N_y} \sum_{i=1}^{N_y} y_i \tag{A.6}$$

$$\tag{A.7}$$

Structure	Property I	Property II
1	1.00	5.00
2	-2.00	6.00
3	2.00	-2.00
4	-2.00	-3.00
5	3.00	-4.00

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

5. *Ward's Method*, uses the error sum of squares (ESS).

$$D(X,Y) = ESS(XY) - [ESS(X) + ESS(Y)] \tag{A.8}$$

$$ESS(X) = \sum_{i=1}^{N_x} \left| x_i - \frac{1}{N_x} \sum_{j=1}^{N_x} x_j \right|^2 \tag{A.9}$$

As an example lets think of a case where we have five structures. Each one of them is descibed by a bidimensional vector as illustrated in Table A.1.

The first step is to chose a distance definition. We chose Manhattan and the distance values between structures can be displayed in a lower triangular matrix as seen in equation A.10

$$d(X,Y) = \begin{vmatrix} & 1 & 2 & 3 & 4 \\ 1 & & & & \\ 2 & 4 & & & \\ 3 & 8 & 12 & & \\ 4 & 11 & 9 & 5 & \\ 5 & 11 & 15 & 3 & 6 \end{vmatrix} \tag{A.10}$$

Let's calculate explicitily the Manhattan distance between structures 2 and 3,

$$d(2,3) = |-2.00 - 6.00| + |2.00 - -2.00| = 12 \tag{A.11}$$

Now that we have calculated the distances we need a clustering method, in this case, we will use the average linkage clustering method. The first step is to group whatever structures are closer, that is, structures 3 and 5 ($d(3,5) = 3$). Now we find the mean distance between the elements of this cluster and the remaining unclustered structures, that is, structures 1, 2 and 4, we obtain the following mean distances

$$D(\{3,5\},1) = \frac{1}{2*1} * (8 + 11) = 4.5 \tag{A.12}$$

$$D(\{3,5\},2) = \frac{1}{2*1} * (12 + 15) = 13.5 \tag{A.13}$$

$$D(\{3,5\},4) = \frac{1}{2*1} * (5 + 6) = 5.5 \tag{A.14}$$

Since the distances between {3, 5} and all remaining unclustered vectors is higher than the distance between vectors 1 and 2 ($d(1,2) = 4$) then {1, 2} are grouped. The following value, in hierarchical increasing order is 4.5 between {3, 5} and 1 (see equation A.12), but since 1 and 2 are already grouped we can't group {3, 5} with 1. The next value, following the lower to higher hierarchy, is 5 ($d(3,4) = 5$),

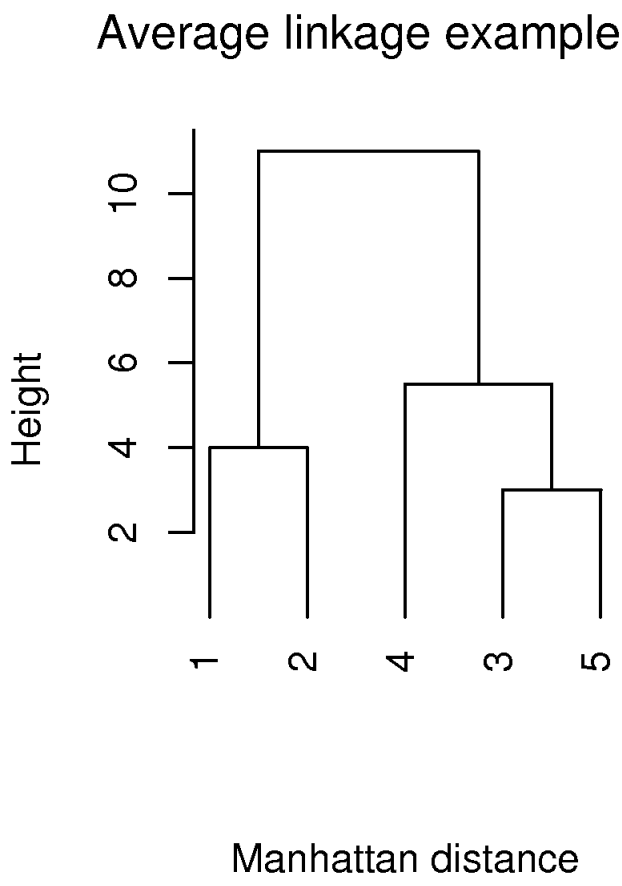

Figure A.1: Clustering tree for 5 bidimensional vectors using the Manhattan distance definition and the average linkage clustering method.

but we have already grouped 3 with 5, so we have to keep advancing in the hierarchy. The next value is 5.5, which corresponds to grouping {3, 5} with 4, so we cluster them. The only remaining possibility for grouping is, group {1, 2} and {4, 3, 5}, so we do it as illustrated in Figure A.1.

Curriculum Vitae

Mauricio Esguerra

La Mala Educacion

1991 High School Diploma from Gimnasio Moderno, Bogota, Colombia.

2000 B. Sc. in Chemistry from Universidad Nacional de Colombia

2010 Ph. D. in Chemistry and Chemical Biology, Rutgers University

Professional Experience

2003-2009 Teaching assistant, Department of Chemistry and Chemical Biology, Rutgers University

Publications

2009 W. K. Olson, M. Esguerra, Y. Xin, X-J. Lu, Methods