RNA STRUCTURE ANALYSIS VIA THE RIGID BLOCK MODEL

BY MAURICIO ESGUERRA NEIRA

A dissertation submitted to
Dr. Wilma K. Olson Group
Rutgers, The State University of New Jersey

New Brunswick, New Jersey
January, 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 beginning of life, also the mechanisms of life regulation and control. The life regulation part is new, not ten years old. Primodial in understanding the cell. The practical purpuse for the chemist is to understand how RNA folds. It's mainly a mechanical problem, therefore it's not foreing to use statistical mechanics methods, combined with detailed knownledge of atomic level structure.

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

Acknowledgements

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

Table of Contents

Chapter 1
Introduction

1.1 RNA folding

1.1.1 Interest in 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]. This fact and the discovery of ribozymes [4, 5] has renewed interest in solving the problem, that is, from primary sequence, finding in an automated[i] way the native three-dimensional structure of RNA and its folding pathway. The problem is usually seen as analogous to the protein folding problem, due both to the discovery of the enzymatic behavior of RNA [4, 5] and the complicated folding of large RNA molecules [9]. 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 [10]. 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 [11, 12, 13] of either protein or RNA.

1.1.2 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 [14] 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 known that secondary and tertiary interactions can be separated in the case of RNA by the absence or presence of Mg^{2+} [15] (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 [16, 17] 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

[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 [6, 7] uses the Tinoco-Uhlenbeck dinucleotide postulate [8] to find total free energies.

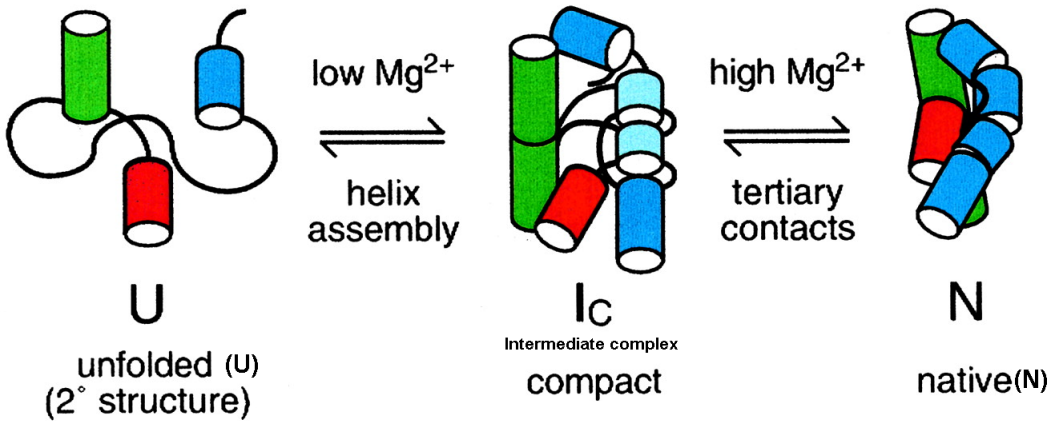

Figure 1.1: Separation of secondary and tertiary interaction in RNA [15]. Double helical secondary structure represented by individual cylinders and tertiary interactions by association of cylinders.

and cations [18] 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 $\mathrm{tRNA}^{\mathrm{Phe}}$ [19, 20] at various levels of detail, initially using only one pseudoatom per helical region, and later one pseudoatom per nucleotide. 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 [21, 22, 23]. This type of work is often referred to as protein structural topology and there is no counterpart for RNA.

1.1.3 Experimental folding techniques

Traditionally RNA folding and unfolding have been followed calorimetrically and spectroscopically as a function of temperature and cation concentration [24]. 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 [25]. 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 [26, 27, 28, 29]. 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 [30], one can deal with such cases and extract RNA folding free energies from single molecule experiments [31].

1.1.4 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 [32, 33, 34] 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 [35]. Tung and associates used this structure for an all-atom MD simulation of the movement of tRNA into a fluctuating ribosome [36]. This type of simulation might be useful in a reverse-folding approach to the RNA folding problem. To the best of my knowledge, such calculations haven't as yet been done for RNA.

1.1.5 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 [37, 38]. 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 [35, 39]. 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 [40] seem to contradict their older publications [39, 41], 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) [42]. 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 [24]. 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 [43, 44] 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 [24] or the folding of RNA. Poisson-Boltzmann theory offers a more detailed picture of the behavior of charged macroions in solution [45].

The local conformational space of RNA has been studied using a large set of available RNA structures from the Nucleic Acid Database (NDB) [46]. The torsion angles of the nucleotide steps have been clustered in the parameter space using different techniques [47, 48]. The root-mean-square deviations (RMSD) of the distances between closely spaced atoms in the phosphates, sugars, and bases, have also been clustered [49]. 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.

1.1.6 RNA secondary structure algorithms and the lack of tertiary ones

From secondary structure prediction algorithms like Zuker's *mfold* program [50], or Hofacker's Vienna RNA package [7], 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 [51]. So far this type of model has not been generalized to take into account tertiary structural features, *i.e.*, interhelical interactions of RNA.

1.1.7 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 [52] (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.

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.

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

Figure 1.2: Ribbon-coil schematic illustraring the fold and intermolecular units of a dimer of prealbumin, or transthyretin, taken from Richardson *et al.* [53]

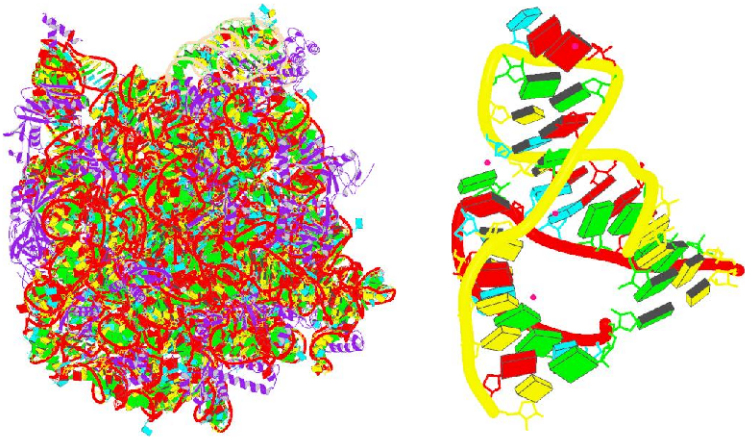

Figure 1.3: *Haloharcula marismortui's* large ribosomal subunit (left) and hammerhead ribozyme (right).The figures were taken directly from the NDB web pages, and show a ribbon representation of the phosphate backbone, and a block representation for the nucleotide bases. From the figures it's clear that,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

In the case of proteins the SCOP (Structural Classification of Proteins) database [54], classifies proteins, among other classifications, according to recurrent arrangements of secondary structure, that is, folds. The SCOR (Structural Classification of RNA) database [55, 56], 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.

Structure, interactions, and reactivity are the conceptual pillars upon which chemistry stands. The aim of this proposal is to try to understand how these concepts relate to the RNA folding problem, by providing a new model for the three dimensional description of RNA.

[iii] Leontis and Westhof [57] 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"

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] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

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

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

[12] 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.

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

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

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

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

[23] 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.

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

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

[26] 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.

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

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

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

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

[31] 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.

[32] 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.

[33] 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.

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

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

[36] 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.

[37] 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.

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

[39] 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.

[40] 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,*.

[41] 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.

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

[43] 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.

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

[45] 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.

[46] 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.

[47] 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.

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

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

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

[51] 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.

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

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

[54] 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.

[55] 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.

[56] 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.

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

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

Chapter on automatic finding of RNA Motifs based on 3DNA Analysis

5.1 GNRA Motif

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

Chapter 6
RNA Helical Regions and Graph Theory

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