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

October, 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 translational 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 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. Also I would like to thank comments and help with the persistence length code to Luke Czapla, and Guohui Zheng who were very kind and prompt in answering in full detail technical questions concerning their code. Constant help on answering 3DNA related questions were always kindly and promptly answered by Xiang-Jun Lu.

I 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 be.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

RNA plays a primordial role in life, and perhaps also in the early history of its origins [1, 2, 3, 4]. In molecular biology RNA is a central player in the transcription and translation steps of what is known as its central dogma, i.e., DNA makes RNA (via transcription) and RNA makes protein (during translation). In the last decade of the twentieth century Fire and Mello [5] found that RNA also plays a role previously thought to be the job of proteins. That is, RNA can regulate translation using non-coding RNA's (ncRNA's). Another fundamental discovery about RNA came in 2000 with the elucidation of the structure at atomic level detail of a large non-coding RNA, the ribosome [6, 7, 8].

Since its very beginnings, structural understanding of RNA has proven to be a very complex problem. It was not until 1956, three years after the famous *Nature* triad of papers by Watson and Crick, Wilkins, Stoke, and Wilson, and Franklin and Gosling [9, 10, 11] on the double-stranded structure of DNA, that Alex Rich and David Davies were able to produce double-stranded RNA from polyriboadenylic acid (poly-rA) and polyribouridylic acid (poly-rU) to produce a neatly difracting X-ray pattern typical of a double-helical structure. It was not until 1965 that Robert Holley was able to obtain the complete sequence of yeast alanine tRNA, and also its secondary structure from cleavage of the whole structure into smaller fragments [12], and it was only in 1973, that the first complex, but small, tRNA structure, was solved at full atomic detail [13, 14, 15]. Fifty seven years have passed since the description of the double-helical structure of DNA, but still RNA faces more challenges with the possibility of finding a whole new zoo of non-coding RNA structures [16], and the possibility of new engineered ones [17].

1.1 RNA Chemistry

RNA is a polynucleotide chain, that is, a polymer whose monomeric unit is the nucleotide. The nucleotide unit is composed of three chemically distinct entities: base, sugar, and phosphate. The bases can be of two types, purines (R), i.e. adenine (A) and guanine (G), and pyrimidines (Y), i.e. cytosine (C) and uracil (U) as shown in Figure 1.1.

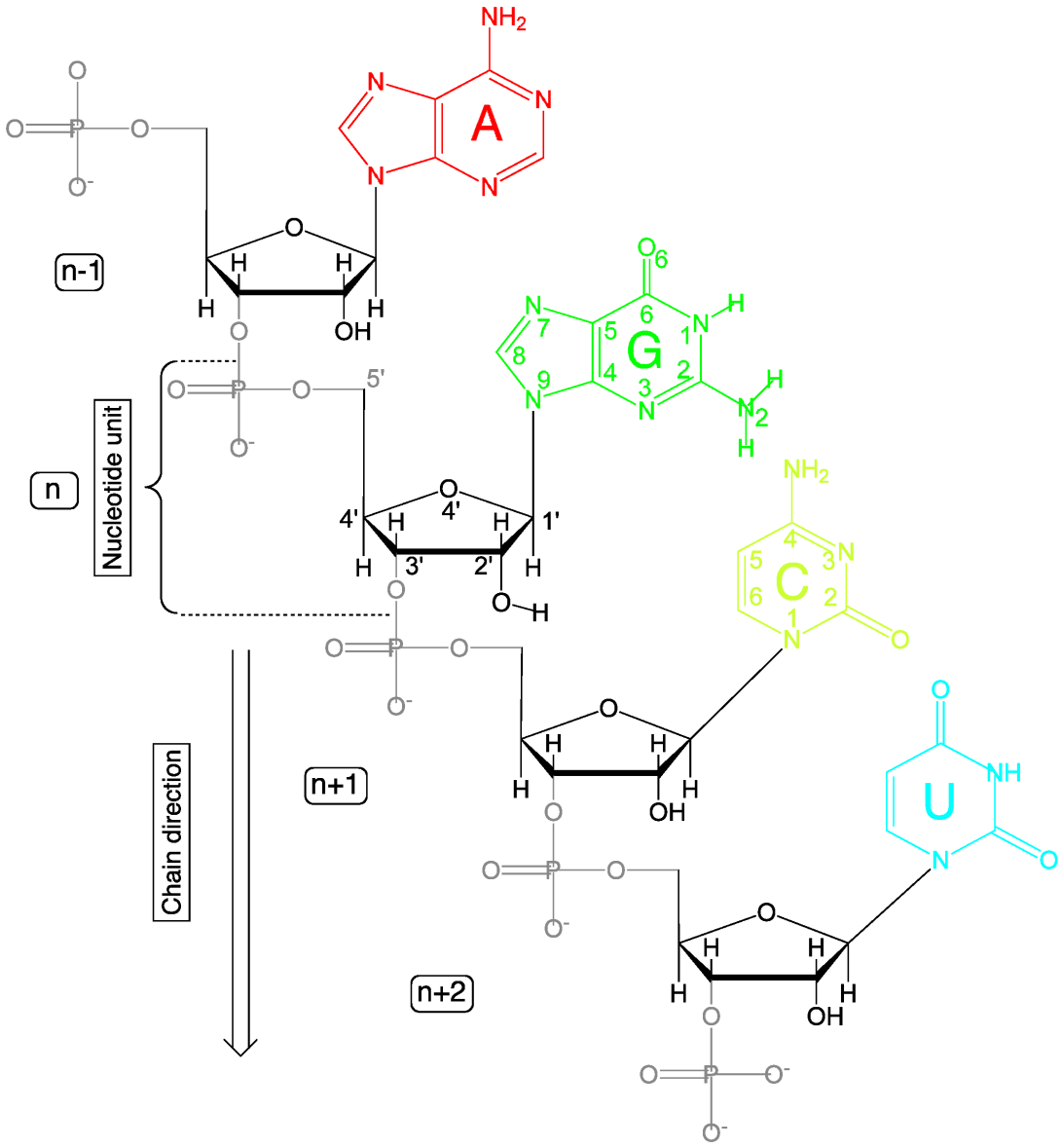


Figure 1.1: A single strand of RNA drawn in the 5′ to 3′ sense showing the three chemical entities which compose it- base, sugar, and phosphate. The four bases (A, G, C, U) are colored according to the NDB (Nucleic Acid Database) convention [18], the phosphate is colored gray, and the sugars black. The 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].

The heterocyclic bases can form a diversity of base pairs through hydrogen bonding and associate in at least 28 distinct classes, first proposed by Saenger [20] and illustrated in Figure 1.2. A system of base-pair nomenclature, which conforms to Saenger's groups, has been developed by Lee-Gutell [21], Leontis-Westhof [22], and Lemieux-Major [23] in order to classify the arrangements of bases seen in high-resolution structures.

The other non-covalent interactions which are common to the nucleotide bases are those of stacking through London dispersion forces and electrostatic interactions. It has been hypothesized that π electron interactions could also account for stacking, but very precise quantum calculations [24, 25] have show otherwise thus far.

The sugar and phosphate groups can adopt a variety of conformations, typically defined by the values of the torsion angles described by the planes formed by four succesive atoms. In the case of the sugar, the torsion angles are constrained by the closure of the five-membered ring to distinct ranges corresponding to two principal puckered arrangements in which one or two of the five atoms lie out of the plane defined by the other four or three atoms. The prefered sugar pucker in RNA is the $\mathrm{C}3' - \mathrm{endo}$ form, but in cases where a base intercalates between two sequential bases, the sugar pucker frequently change to the less-prefered $\mathrm{C}2' - \mathrm{endo}$ conformation. Standards to describe the conformations resulting from the specific sets of torsion angle values which sugars and phosphate can attain have been developed and can be seen in textbooks [20], on the web [26], and in the IUPAC recommendations [19]. We refer the reader to these sources for a more detailed description, and limit ourselves to the brief description of these torsion angles shown in Figure 1.3.

1.2 RNA Folding

The first high-resolution X-ray structure of RNA larger than a dinucleotide was that of yeast phenylalanine transfer RNA ($\mathrm{tRNA}^{\mathrm{Phe}}$), solved at 3Å in 1974 [13, 14, 15]. Thirty six years later there are two orders of magnitude more structural information about RNA [27], and new information from non-coding RNA's is expected [16]. This fact and the discovery of ribozymes [28, 29], which are catalytic RNA molecules, has renewed interest in solving the RNA folding problem, that is, starting from the primary sequence, finding in an automated[i] way the native three-dimensional structure of an RNA molecule and

[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 [30, 31] uses the Tinoco-Uhlenbeck dinucleotide postulate [32] to find total free energies That is, it is assumed that the total free energy of the

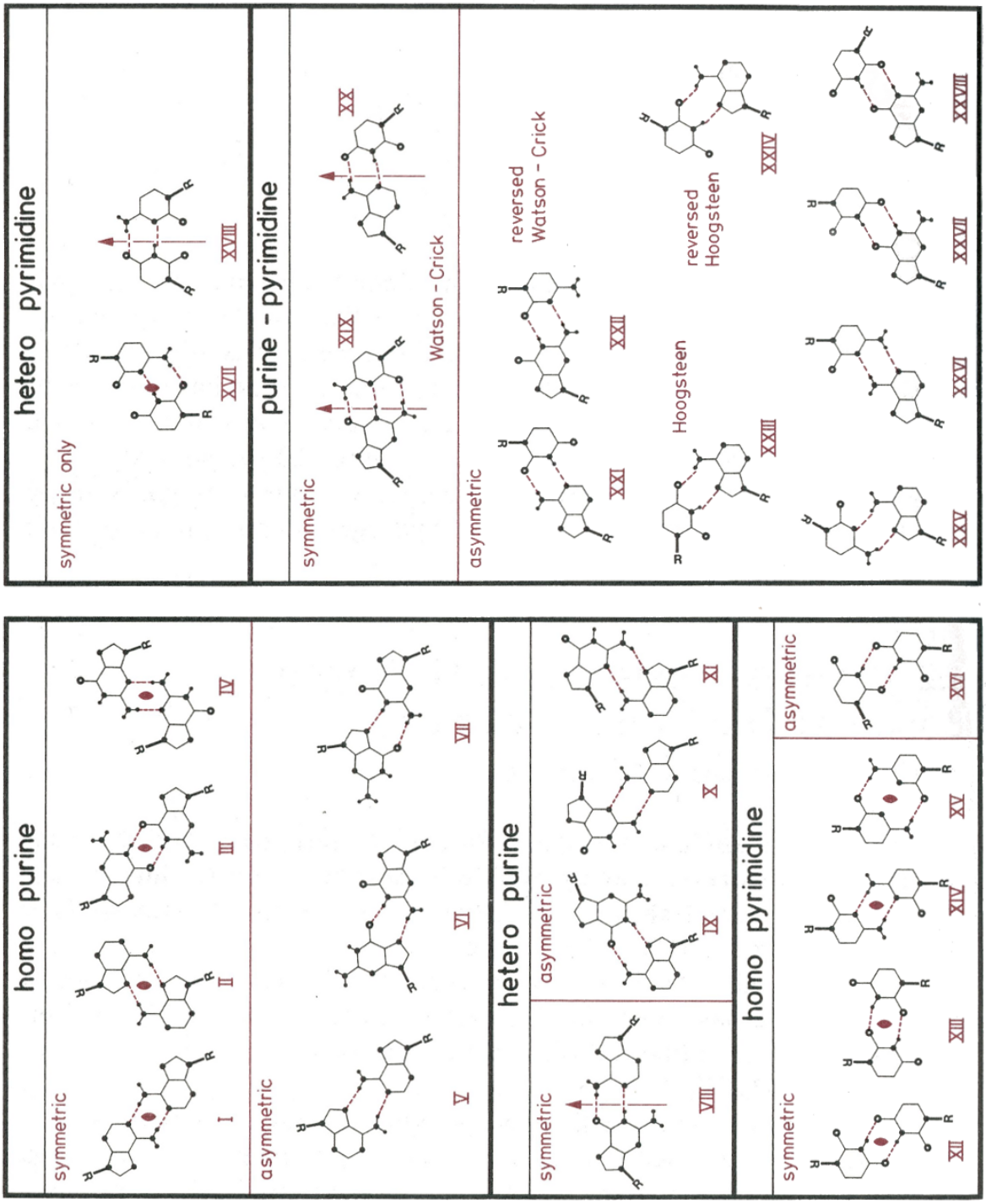


Figure 1.2: Saenger base-pairing classes, reproduced with permission from his book, "Principles of Nucleic Acid Structure". [20].

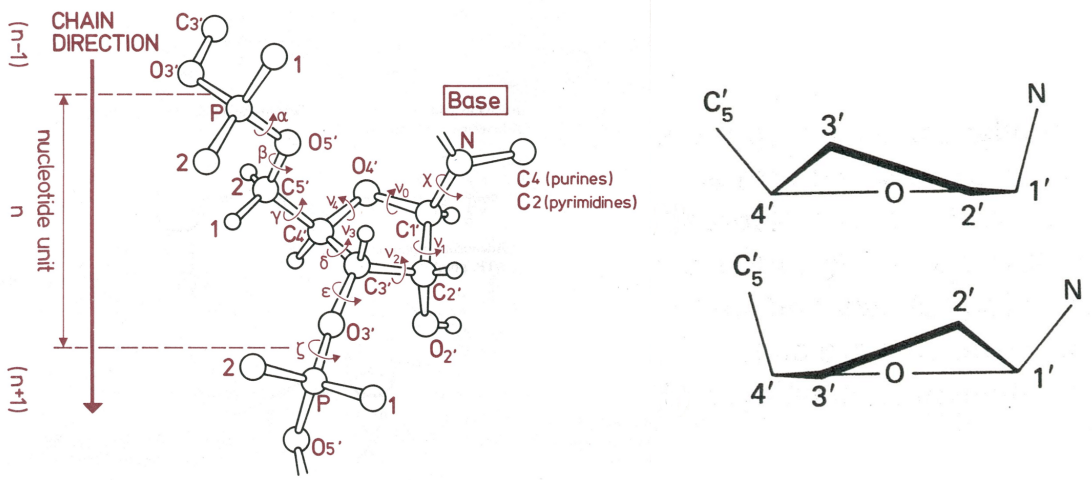


Figure 1.3: **Left:** Sugar, and sugar-phosphate backbone torsion angles. **Right:** The most common sugar pucker conformations in RNA, that is, $C3' - \text{endo}$ and $C2' - \text{endo}$, reproduced with permission from Wolfram Saenger's, "Principles of Nucleic Acid Structure". [20].

the folding pathway that it follows. The RNA folding problem is usually seen as analogous to the protein folding problem, due to both the discovery of the enzymatic behavior of RNA [28, 29] and the complicated folding of large RNA molecules [33]. 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 [34]. This and other methods used to characterize RNA and protein folding depend upon the partition function used to describe the correct conformational ensemble of folded, partially folded, and unfolded structures [35, 36, 37] of either protein or RNA.

1.3 Is RNA Folding a Hard or Easy Problem?

There are two trains of thought regarding the mechanism of RNA folding. One states that RNA folding is less complex than protein folding [38] 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+} [39] (see Figure 1.4), and that the secondary structure motifs of RNA are more limited in number than those of protein. By contrast 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 [40, 41] since there is no such thing as hydrophobic burial of regions of RNA as in the case of proteins. Instead,

RNA polymer is the sum of the free energies of the individual base-pair steps which constitute it.

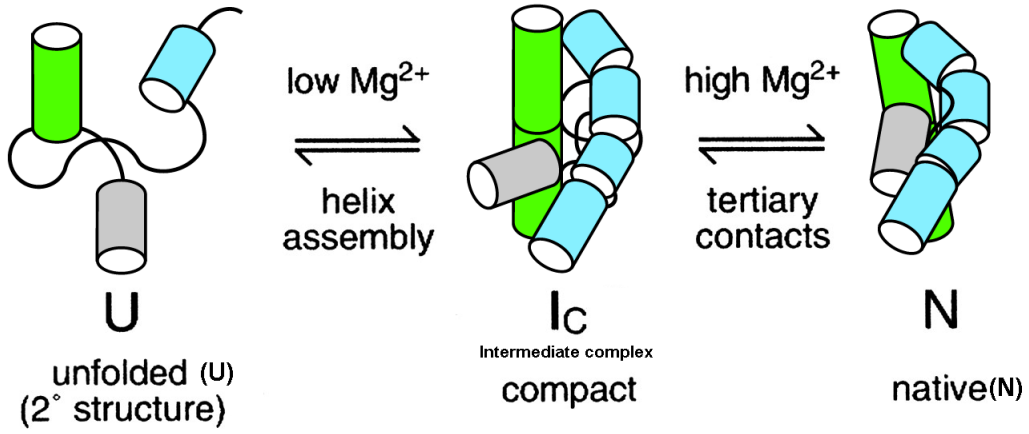


Figure 1.4: Separation of secondary and tertiary interactions in RNA [39]. Double helical secondary structure represented by individual cylinders and tertiary interactions by association of cylinders. Color coding stands for separate helical regions of RNA, and the connecting black strings represent single stranded loop structures. The cartoon corresponds to the Mg^{2+} dependent folding of the *Azoarcus* ribozyme, where at low magnesium ion concentrations an intermediate and somewhat loose ensemble of conformations denoted by I_C is formed by association of helical segments. At higher Mg^{2+} concentrations the helices arrange into a more ordered and compact structure which is catalitically active and is denoted as the native state N.

the electrostatic problem stemming from a complex charged backbone and polar base side groups must be dealt with in the case of RNA. For instance, the interactions of the RNA polyanionic backbone with water and cations [42] are not easily simulated with explicit solvent models like those used to treat 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. Stephen Harvey and associates have simulated the folding of yeast $tRNA^{Phe}$ [43] and the assembly of the 30S subunit of the ribosome [44] at various levels of detail, initially using only one pseudoatom per helical region, and later one pseudoatom per nucleotide. Recently François Major's group at Montreal has proposed a pipeline of two computer algorithms to study RNA structure [45]. One pipeline makes secondary structure predictions, and the other assembles 3D structures based on the best scoring secondary structures. 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 [46, 47, 48]. This type of work is often referred to as protein structural

topology and there is no counterpart for RNA.

1.4 Experimental Folding Techniques

Traditionally RNA folding and unfolding have been followed calorimetrically and spectroscopically as a function of temperature and cation concentration [49, 50]. 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 [51]. 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 that are then pulled apart and monitored with a laser light trap [52, 53, 54, 55]. 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. Bustamante, Tinoco, and associates have shown that by using the Crooks fluctuation theorem [56], one can deal with such cases and extract RNA folding free energies from single-molecule experiments [57].

1.5 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 [58, 59, 60], which treat RNA at less than atomic detail, have been used to describe the global motions of large RNA structures like the ribosome. Examples of the predicted normal modes of motion of the ribosome can be seen at the Robert Jernigan group web site: `http://ribosome.bb.iastate.edu/70SnKmode`. Using MM, Tung and Sanbonmatsu have obtained a static atomic model of the 70S ribosome structure through homology modeling [61]. Tung and associates have used this structure in an all-atom MD simulation of the movement of tRNA into a fluctuating ribosome [62]. 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 have not as yet been done for RNA.

1.5.1 Local Nucleotide Interactions

The molecular interactions that guide RNA structures at the nucleic acid base level, *i.e.*, local level, are, as noted in Section 1.1, 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 interactions of free nucleotide bases are determined by dispersion attraction, short-range exchange repulsion, and electrostatic interaction. No specific $\pi - \pi$ interactions are found from electron correlated ab-initio calculations [24, 25]. This is why force field methods have been so successful in the study of nucleic acids, since the simple empirical potentials used in such studies mimic well the quantum mechanically obtained energy profiles [61, 63]. 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 [64] seem to contradict their earlier work [63, 65], 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-strand and double-strand stacking free energies can be obtained calorimetrically [66]. One of the most popular methods used for obtaining such quantities is differential scanning calorimetry (DSC) [67]. 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 [49]. On the other hand, the stacking free energies for some sequences[ii] are 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 [70, 71] provides a simple and quantitative picture of the interactions of a regular double-helical nucleic acid polyanion with its counterions, but it does not take into account the discrete nature of charge [49] or the folding of RNA. Poisson-Boltzmann theory offers a more detailed picture of

[ii] Free energies for 5' unpaired nucleotides (e.g., UC/A UU/A) are quite small (i.e., < 0.4 kcal/mol) and are termed weakly stacking bases [68, 69].

the behavior of charged macroions in solution [72, 73].

The local conformational space of RNA has been studied using a large set of available RNA structures from the Nucleic Acid Database (NDB) [74]. The torsion angles of the nucleotide steps have been clustered using different techniques [75, 76]. The root-mean-square deviations (RMSD) of the distances between closely spaced atoms in the phosphates, sugars, and bases, have also been clustered [77]. The latter studies are aimed at finding the common nucleotide base steps and the base-pair building blocks, which have been 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 [78].

1.5.2 RNA Secondary Structure Algorithms and the Lack of Tertiary Ones

From secondary structure prediction algorithms like Zuker's *mfold* program [79], Hofacker's Vienna RNA package [31], or Mathew's Dynaling software [80], one obtains a large ensemble of secondary structure graphs, i.e., 2D representations of the double-stranded helical stems, hairpin loops, and bubbles formed by the constituent bases. These graphs can be analyzed with graph theory to produce a partition function that describes the full arrangement of contacts for the total number of possible secondary structures, and allows the construction of a "relation of microscopic conformations to macroscopic properties" [81]. So far this type of model has not been generalized to take into account tertiary structural features, *i.e.*, interhelical interactions of RNA. 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, applying a minimization procedure, and then performing a Molecular Mechanics (MM) all-atom refinement [82, 83, 84] of the structure. Another starts from the predicted secondary structures, assumes that the helical regions adopt the canonical A-form structure, mechanically inserts residues as rigid bodies in the remaining non-helical regions, and finally carries out an MM optimization [85]. The third approach entails a pipeline between secondary structure prediction, and tertiary structure assembly. This pipeline uses as a bridging concept between 2D and 3D structure, the graph theoretical definition of a minimum cycle basis, which for the case of nucleic acids has been renamed as a Nucleic Cyclic Motif (NCM) [45].

1.5.3 RNA Overall Fold

Whereas in the case of proteins one qualitatively describes the overall fold in terms of the arrangement of secondary structure motifs, *i.e.*, using the helix-ribbon-coil images developed by Jane Richardson [86] (see Figure 1.5), there is still no comparable description of the overall fold of RNA. A ribbon representation of the sugar-phosphate backbone (see Figure 1.6) helps to understand the folding of small RNA's, but in the case of the large ribosome structure, a representation at the same level of detail does not make sense (these are close to 3000 nucleotides in the large subunit of the archaeal ribosome). In the past two years Holbrook [87] and Sykes [88] have proposed new representations for RNA based on helical region organization. Holbrook makes an analysis of continuous interhelical strands, so called, COINS, and Sykes makes an optmized projection of the 3D helical axes to 2D images, which can later be annotated with, for example, hydroxyl radical footprinting data.

Figure 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.* [89]

One can envision that a thorough investigation of the space of translational and rotational degrees of freedom of the helical regions of RNA could give clues as to how one might see an overall fold in RNA structures. To the best of our knowledge, there is no comparable quantitative description of the folding of proteins.

In the case of proteins, the SCOP (Structural Classification of Proteins) database [91] classifies proteins, among various qualitative descriptors, according to folds, which are recurrent arrangements of secondary structure, that is, a list of consecutive secondary structures with unique topological connections. The SCOR (Structural Classification of RNA) database [92, 93] aims to provide a similar classification to that obtained for proteins, but using RNA motifs instead. This classification focuses on the local folding of small pieces of RNA and does not describe the overall fold. The local classification of RNA motifs in SCOR is also qualitative rather than quantitative.

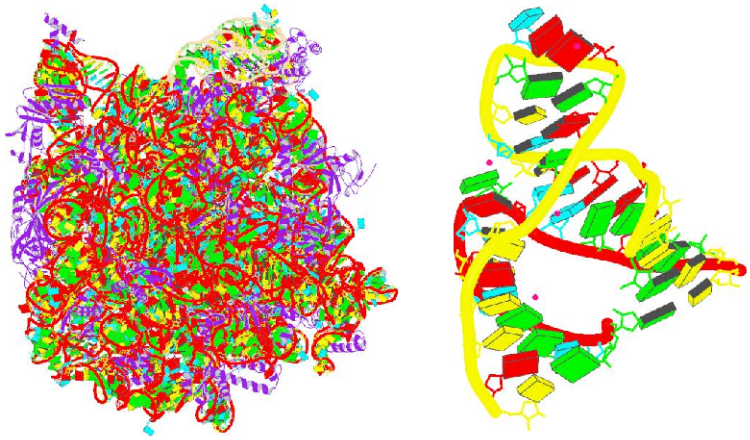


Figure 1.6: Images of the *Haloharcula marismortui's* large ribosomal subunit NDB_ID:RR0033 (left) and the hammerhead ribozyme (right) NDB_ID:UR0029. The figures were taken directly from the NDB web pages, and show a 3DNA-generated [90] ribbon representation of the phosphate backbone, and a block representation of the nucleotide bases. From the figures is clear that, whereas the ribozyme fold can be clearly understood with this representation, the ribosome fold cannot.

1.5.4 RNA Motifs

The term "*RNA motif*" is used in the literature to describe three different levels of RNA organization, namely, RNA **sequence** motifs, RNA **secondary structure** motifs, or RNA **3D structure** motifs. Because these distinctions are not always clearly made, the beginner may become confused and frustrated in bibliographical searches.

The lack of a unique definition of RNA motifs is yet another source of confusion. Three popular and somewhat recent definitions of RNA motifs include:

- "*a discrete sequence or combination of base juxtapositions found in naturally occurring RNA's in unexpectedly high abundance.*"[94]
- "*conserved structural subunits that make up the secondary structures of RNAs.*"[95]
- "*ordered stacked arrays of non-Watson-Crick base pairs that form distinct folds on the phosphodiester backbones of RNA strands.*"[96]

The kind of RNA motifs addressed in this thesis are of the third type, that is, RNA **3D structure** motifs which we henceforth term RNA motifs. From our point of view, RNA motifs are to be understood as unique sets of geometrical arrangements (in the rigid block sense) in three-dimensional space.

Even though there is no unique definition, we can think of three practical tasks regarding RNA motifs. That is, given an RNA 3D structure, automatically identify, describe, and find new motifs. For automatic

identification of RNA motifs, Pyle and collaborators have developed the AMIGOS software, which finds RNA motifs based on specific values of the backbone virtual torsion angles η and θ [97, 98, 99] in a way which resembles a Ramachandran plot analysis. Lemieux and Major [100] employ the MC-Fold software, which implicitly finds RNA motifs based on an algorithm that determines so-called nucleic cyclic motifs, which are just the minimal cycle basis of an RNA secondary structure interpreted as a mathematical graph. Leontis [101] and collaborators have created FR3D (read as FRED), a Matlab Windows executable program which finds RNA motifs based on the isostericity matrices of base-pairs.

Schlick and collaborators have used FR3D to localize RNA helical junctions of order four (i.e. four-way junctions) or higher, and performed a visual analysis to see if the helices in such junctions form coaxial stacks or not, and have classified them accordingly [102, 103]. As mentioned previously in the context of RNA folds, Holbrook, and Sykes, describe helical regions and display them in two-dimensional representations. Sponer's group has studied RNA motifs present in the ribosome using Molecular Dynamics (MD) methods implemented in the AMBER package, including 25ns simulations of the sarcin-ricin domain (SRD) [104], and 80ns simulations of the hydration of loop E in the 5S subunit [105].

The software programs, which perform the task of identifying RNA motifs in RNA structures - namely AMIGOS, MC-Fold, and FR3D - also have the capability to find new RNA motifs.

1.6 Overview

Always keeping in mind the greater scope of the RNA folding problem, this thesis addresses various issues of RNA folding and deformation using information in crystallographic structures from the Protein Data Bank (PDB). These data have been analyzed statistically in terms of a rigorous rigid-body formalism and applied to the study of RNA polymers. In Chapter 2 the consensus clustering technique is used to classify the geometry of dinucleotide of non-A-RNA conformations in terms of the base-step parameters describing the orientation and displacement of succesive bases along the RNA chain. The resulting groups are localized and examined in the context of ribosomal RNA. In addition to consensus clustering, a set of clustering validation techniques is used in order to select, a priori, an optimal clustering methodology. This approach differs from the more common practice in the bioinformatics field of choosing an optimal clustering technique arbitrarily. The validated clustering method is used then to classify the non-A-RNA-like conformations of the 23S subunit of the ribosome into 17 main groups.

In Chapter 3 we explore, using statistical analyses, the identities and spatial arrangements of base pairs in the context of RNA helical regions. We find the most prominent base pairs, where these base pairs are located in the helical environment, and how they deform.

In Chapter 4 we present a web framework with an integrated MySQL database containing information on the RNA base-pair steps within the helical regions of representative high-resolution structures. Properties such as the average step parameters and force constants (derived from inverse covariance analysis of the data) are tabulated and available for user download. The public information in the web framework is used then to determine the persistence length and elastic behaviour of double-stranded RNA's with various repeating sequences and to compare the crystallographically based behavior with other experimental results. Finally in Chapter 5 we provide a new software, "getMotif", which interfaces with the 3DNA suite of programs to perform a rigorous search of existing and potentially new RNA motifs. The computations are based on a simple score that accounts for structural similarity at the base-step-parameter level, providing a new metric [106] for structural understanding of RNA beyond the traditional all-atom RMSD.

References

[1] Woese, C. (1967) The Genetic Code, the Molecular Basis for Genetic Expression, Harper and Row, .

[2] Crick, F. (1968) The Origin of the Genetic Code. *Journal of Molecular Biology,* **38**, 367–379.

[3] Orgel, L. (1968) Evolution of the Genetic Apparatus. *Journal of Molecular Biology,* **38**, 381–393.

[4] Orgel, L. E. (2004) Prebiotic Chemistry and the Origin of the RNA World. *Critical Reviews in Biochemistry and Molecular Biology,* **39**, 99–123.

[5] Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998) Potent and Specific Genetic Interference by Double-Stranded RNA in Caenorhabditis Elegans. *Nature,* **391**, 806–811.

[6] Schluenzen, F., Tocilj, A., Zarivach, R., Harms, J., Gluehmann, M., Janell, D., Bashan, A., Bartels, H., Agmon, I., Franceschi, F., and Yonath, A. (2000) Structure of Functionally Activated Small Ribosomal Subunit at 3.3 Å Resolution. *Cell,* **102**, 615–623.

[7] Ban, N., Nissen, P., Hansen, J., Moore, P. B., and Steitz, T. A. (2000) The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution. *Science,* **289**, 905–920.

[8] Wimberly, B. T., Brodersen, D. E., Clemons, W. M., Morgan-Warren, R. J., Carter, A. P., Vonrhein, C., Hartschk, T., and Ramakrishnan, V. (2000) Structure of the 30S Ribosomal Subunit. *Nature,* **407**, 327–339.

[9] Watson, J. D. and Crick, F. H. (1953) Molecular Structure of Nucleic Acids; A Structure for Deoxyribose Nucleic Acid. *Nature,* **171**, 737–738.

[10] Wilkins, M. H. F., Stokes, A. R., and Wilson, H. R. (1953) Molecular Structure of Deoxypentose Nucleic Acids. *Nature,* **171**, 738–740.

[11] Franklin, R. E. and Gosling, R. G. (1953) Molecular Configuration in Sodium Thymonucleate. *Nature,* **171**, 740–741.

[12] Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, J. R., and Zamir, A. (1965) Structure of a Ribonucleic Acid. *Science,* **147**, 1462–1465.

[13] 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.

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

[15] Stout, C. D., Mizuno, H., Rubin, J., Brennan, T., Rao, S. T., and Sundaralingam, M. (1976) Atomic Coordinates and Molecular Conformation of Yeast Phenylalanyl tRNA. An Independent Investigation. *Nucleic Acids Research,* **3**, 1111–1123.

[16] 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.

[17] Severcan, I., Geary, C., Verzemnieks, E., Chworos, A., and Jaeger, L. (2009) Square-Shaped RNA Particles from Different RNA Folds. *Nanotechnology Letters,* **9**, 1270–1277.

[18] http://ndbserver.rutgers.edu/atlas/atlas_about.html.

[19] IUPAC-IUB (1983) IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). Abbreviations and Symbols for the Description of Conformations of Polynucleotide Chains. Recommendations 1982. *European Journal of Biochemistry,* **131**, 9–15.

[20] Saenger, W. (1984) Principles of Nucleic Acid Structure, Springer-Verlag, London.

[21] Lee, J. C. and Gutell, R. R. (2004) Diversity of Base-pair Conformations and their Occurrence in rRNA Structure and RNA Structural Motifs. *Journal of Molecular Biology,* **344**, 1225–1249.

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

[23] Lemieux, S. and Major, F. (2002) RNA Canonical and Non-Canonical Base Pairing Types: A Recognition Method and Complete Repertoire. *Nucleic Acids Research,* **30**, 4250–4263.

[24] 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.

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

[26] http://www.fli-leibniz.de/ImgLibDoc/nana/IMAGE_NANA.html.

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

[28] 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.

[29] 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.

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

[31] 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.

[32] Borer, P. N., Dengler, B., Tinoco, I., and Uhlenbeck, O. C. (1974) Stability of Ribonucleic Acid Double-Stranded Helices. *Journal of Molecular Biology,* **86**, 843–853.

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

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

[35] Chen, S.-J. and Dill, K. A. (1995) Statistical Thermodynamics of Double-Stranded Polymer Molecules. *Journal of Chemical Physics,* **103**, 5802–5813.

[36] 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.

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

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

[39] 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 States of America,* **100**, 1574–1579.

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

[41] 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.

[42] 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**, 1366–1379.

[43] 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.

[44] 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.

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

[46] Westhead, D., Slidel, T., Flores, T., and Thornton, J. (1999) Protein Structural Topology: Automated Analysis and Diagrammatic Representation. *Protein Science,* **8**, 897–904.

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

[48] 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.

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

[50] Boots, J. L., Canny, M. D., Azimi, E., and Pardi, A. (2008) Metal Ion Specificities for Folding and Cleavage Activity in the Schistosoma Hammerhead Ribozyme. *RNA,* **14**, 2212–2222.

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

[52] Liphardt, J., Onoa, B., Smith, S., Tinoco, I., and Bustamante, C. (2001) Reversible Unfolding of Single RNA Molecules by Mechanical Force. *Science,* **292**, 733–737.

[53] Onoa, B. and Tinoco, I. (2004) RNA Folding and Unfolding. *Current Opinion in Structural Biology,* **14**, 374–379.

[54] Tinoco, I. (2004) Force as a Useful Variable in Reactions: Unfolding RNA. *Annual Review of Biophysics & Biomolecular Structure,* **33**, 363–385.

[55] Hyeon, C. and Thirumalai, D. (2005) Mechanical Unfolding of RNA Hairpins. *Proceedings of the National Academy of Sciences of the United States of America,* **102**, 6789–6794.

[56] Crooks, G. E. (1999) Entropy Production Fluctuation Theorem and the Nonequilirium Work Relation for Free-Energy Differences. *Physical Review E,* **60**, 2721–2726.

[57] Collin, D., F.Ritort, Jarzynski, C., Smith, S. B., Tinoco, I., and Bustamante, C. (2005) Verification of the Crooks Fluctuation Theorem and Recovery of RNA Folding Free Energies. *Nature,* **437**, 231–234.

[58] 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.

[59] 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.

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

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

[62] 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 of the United States of America,* **102**, 15854–15859.

[63] 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.

[64] 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,* **12**, 2854Ű2865.

[65] 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.

[66] Freier, S. M., Sinclair, A., Neilson, T., and Turner, D. H. (1985) Improved Free Energies for G.C Base-Pairs. *Journal of Molecular Biology,* **185**, 645–647.

[67] Marky, L. A. and Breslauer, K. J. (1982) Calorimetric Determination of Base-Stacking Enthalpies in Double-Helical DNA Molecules. *Biopolymers,* **11**, 2185–2194.

[68] Burkard, M. E., Kierzek, R., and Turner, D. H. (1999) Thermodynamics of Unpaired Terminal Nucleotides on Short RNA Helixes Correlates with Stacking at Helix Termini in Larger RNAs. *Journal of Molecular Biology,* **290**, 967–982.

[69] Burkard, M. E., Turner, D. H., and Tinoco, I. The RNA World chapter 10. The Interactions That Shape RNA Structure, pp. 233–264 Cold Spring Harbor Laboratory Press 2nd edition (1999).

[70] 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.

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

[72] Antypov, D., Barbosa, M. C., and Holm, C. (2005) Incorporation of Excluded-Volume Correlations into Poisson-Boltzmann Theory. *Physical Review E,* **71**, 1–6.

[73] Xu, D., Landon, T., Greenbaum, N. L., and Fenley, M. O. (2007) The Electrostatic Characteristics of G.U Wobble Base Pairs. *Nucleic Acids Research,* **35**, 3836–3847.

[74] 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.

[75] 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.

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

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

[78] 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,* **14**, 465–481.

[79] Zuker, M. (2003) Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction. *Nucleic Acids Research,* **31**, 3406–3415.

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

[81] 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.

[82] Das, R. and Baker, D. (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**, 14664–14669.

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

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

[85] Martinez, H. M., Maizel, J. V., 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.

[86] Richardson, J. S. (2000) Early Ribbon Drawings of Proteins. *Nature Structural Biology,* **7**, 624–625.

[87] Holbrook, S. R. (2008) Structural Principles From large RNAs. *Annual Review in Biophysics,* **37**, 445–464.

[88] Sykes, M. T. and Williamson, J. R. (2009) A Complex Assembly Landscape for the 30S Ribosomal Subunit. *Annual Review of Biophysics,* **38**, 197–215.

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

[90] Lu, X.-J. and Olson, W. K. (2008) 3DNA: A Versatile, Integrated Software System for the Analysis, Rebuilding and Visualization of Three-Dimensional Nucleic-Acid Structures. *Nature Protocols,* **3**, 1213–1227.

[91] 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.

[92] 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.

[93] 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**, 2342–2352.

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

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

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

[97] Olson, W. K. (1980) Configurational Statistics of Polynucleotide Chains. An Updated Virtual Bond Model to Treat Effects of Base Stacking. *Macromolecules,* **13**, 721–728.

[98] Malathi, R. and Yathindra, N. (1985) Backbone Conformation in Nucleic Acids: An Analysis of Local Helicity Through Heminucleotide Scheme and a Proposal for a Unified Conformational Plot. *Journal of Biomolecular Structure and Dynamics,* **3**, 127–144.

[99] Duarte, C. M., Wadley, L. M., and Pyle, A. M. (2003) RNA Structure Comparison, Motif Search and Discovery Using a Reduced Representation of RNA Conformational Space. *Nucleic Acids Research,* **31**, 4755–4761.

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

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

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

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

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

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

[106] Parisien, M., Cruz, J. A., Westhof, E., and Major, F. (2009) New Metrics for Comparing and Assessing Discrepancies Between RNA 3D Structures and Models.. *RNA,* **15**, 1875–1885.

Chapter 2

RNA Base Steps

The problem of classification of the space of conformations of RNA is not new, see for example, Olson 1972 [1], Saenger 1984 [2], and Gautheret 1993 [3]. Although only a few researchers addressed this problem before the turn of the twenty first century, now this situation is changing rapidly. The reason for this fast change came in the year 2000, when a vast amount of RNA structural information became available upon the elucidation of the structure of the 30S small ribosomal subunit of *Thermus thermophilus*, a bacterial ribosome [4, 5], and the 50S large ribosomal subunit of *Haloarcula marismortui*, an archaeal ribosome [6].

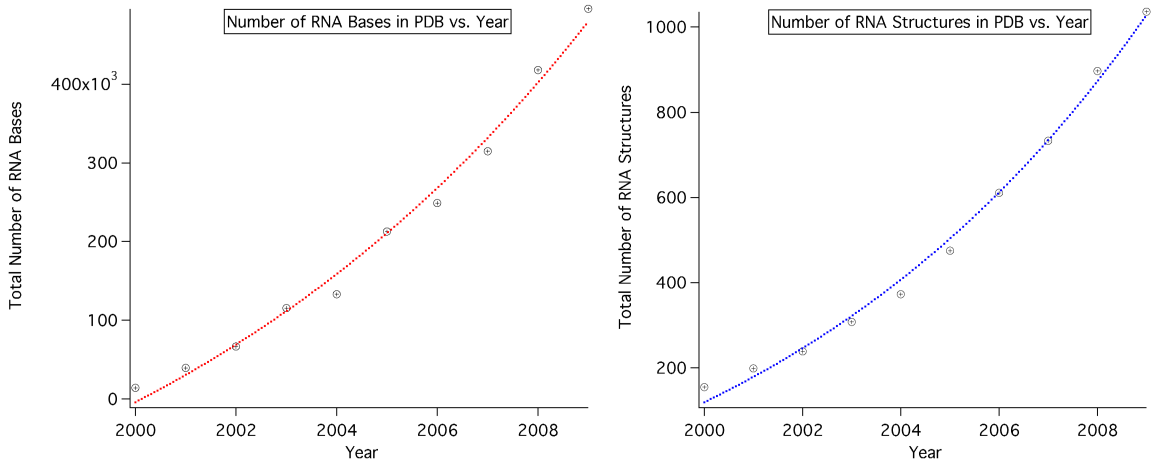


Figure 2.1: **Left:** Total number of RNA bases added to the Protein Data Bank (PDB) between 2000 and 2010 (exponential fit line in red). **Right:** Total number of RNA structures solved yearly by X-ray crystallography between 2000 and 2010 (exponential fit line in blue).

Between 1978 and 2000 a total of 116 RNA structures with resolution better than 3.5 Å, and comprising around 5500 nucleotide bases were added to the Protein Data Bank (PDB), and between 2000 and today 931 RNA structures comprising 491,158 nucleotide bases. That is, the increase in information due to the solution of large RNA structures, is about two orders of magnitude, as pointed out by Noller [7]. It is clear from the growth of RNA structural information from 2000 until today that both the total number of RNA structures deposited in the PDB, and the total number of nucleotide bases in these

h

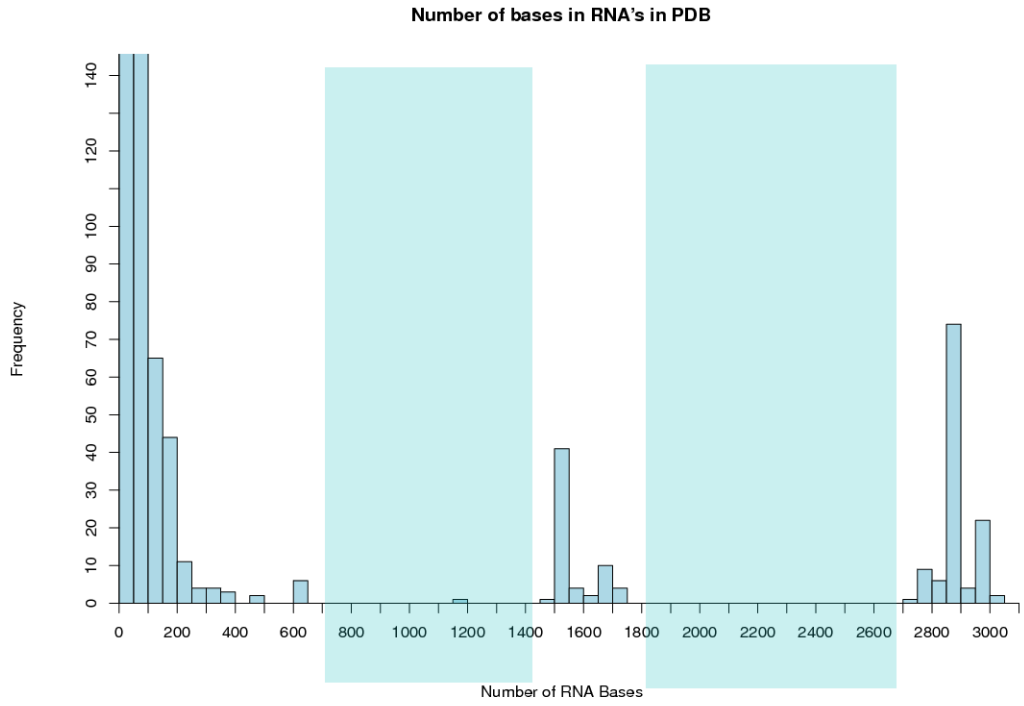


Figure 2.2: Frequency of nucleotide bases in RNA molecules found in the PDB classified by the size of RNA molecules. We define the size as the total number of nucleotide bases present per molecule. Notice the areas in transparent blue boxes, which are devoid of RNA molecules.

structures, is increasing at an exponential rate (as can be seen from the exponential fits of these data in Figure 2.1). It is important to note that such growth comes mainly from ribosomal structures which contain 88 percent of all the RNA bases in the PDB. So, even though structural interest in RNA has been growing since ribosomal structures became available in 2000, and several Nobel prizes have been awarded for work in this field, along with the exciting possibilities of deciphering large RNA [8] structures other than the ribosome, still the growth of the RNA structural field is far from that of proteins, if weighed by the growth in diversity of RNA structural information in the past decade.

If we look at the current distribution of the sizes of RNA structures counted in terms of the number of bases (Figure 2.2), it is clear that there are large patches of chain lengths where there are no RNA structures whatsoever. That is, there are no solved structures of RNA with roughly 600 to 1400 bases or 1800 to 2700 bases. The area of non-coding RNA's holds great promise for finding structured RNA's in such length ranges, as has recently been suggested by the Breaker group [8]. A representative example of the characteristic ranges of RNA structures available to date in the PDB can be seen in Table 2 for structures larger than 300 bases. A comparison between the total number of structures of protein, protein plus nucleic acid, DNA, and RNA, available at the PDB from the seventies until today –

shown in Figure 2.3 – reveals large differences between the number of deposited protein and nucleic acid structures. There is over an order of magnitude more protein than nucleic acid. The rate of growth of DNA vs. RNA structural information is also changing. RNA structures have been growing steadily since the mid-nineties and seem to parallel the growth of DNA structures.

PDBID	Structure Name	Phylogenetic Group	Number of bases	Year
1l8v	Mutant of P4-P6 Domain of Group I Intron	Eukaryota	314	2002
3igi	Group II Intron	Bacteria	395	2009
1fg0	Central Loop in Domain V of 23S rRNA	Archaea	499	2000
2nz4	GlmS Ribozyme	Eukaryota	604	2006
1xmq	30S rRNA	Bacteria	1522	2004
1ffk	50S rRNA Subunit	Archaea	2828	2000

Table 2.1: Types and sizes of some large RNA structures ($>$300 bases) elucidated in the last decade.

The conformational information contained in RNA structures can be examined from three main perspectives: an atom-based perspective; a bond-based perspective; and a third, as yet unexplored to our knowledge, rigid-body-based perspective. In the atom-based perspective, either a direct comparison of backbone atom positions is made [9], or a comparison of the distances between a reduced set of atoms taken from the nucleotide backbone, sugar, and base [10]. The bond-based perspective is divided into three main categories. The first considers the spatial arrangements of the consecutive covalent bonds in the RNA backbone and the glycosidic bond between the sugar and base, that is, the six backbone torsion angles and one glycosidic torsion angle [9, 11, 12, 13, 14]. The second considers the pseudo-bonds between consecutive P and C4′ atoms and the resulting pseudo-torsion angles η and θ [1, 15, 16, 17] [i]. The third category considers the networks of horizontal hydrogen bonding patterns coming from a definition of interacting edge boundaries in the nucleotide bases [19, 20, 21]. In this chapter we study the rigid-body based perspective using clustering analysis and discuss the relationship of these findings to the other previously reported perspectives on RNA conformation.

2.1 Consensus Clustering of Single-stranded Base-step Parameters

To our knowledge there has been no classification of rigid-body base-step parameters for the RNA structures now available in the PDB. It is important to note here that in crystal structures, the RNA bases

[i] The η and θ pseudo-torsion angles are determined by consecutive backbone torsion angles as described by Olson [18] in relating the structural parameters of polynucleotide chains to virtual bond vectors.

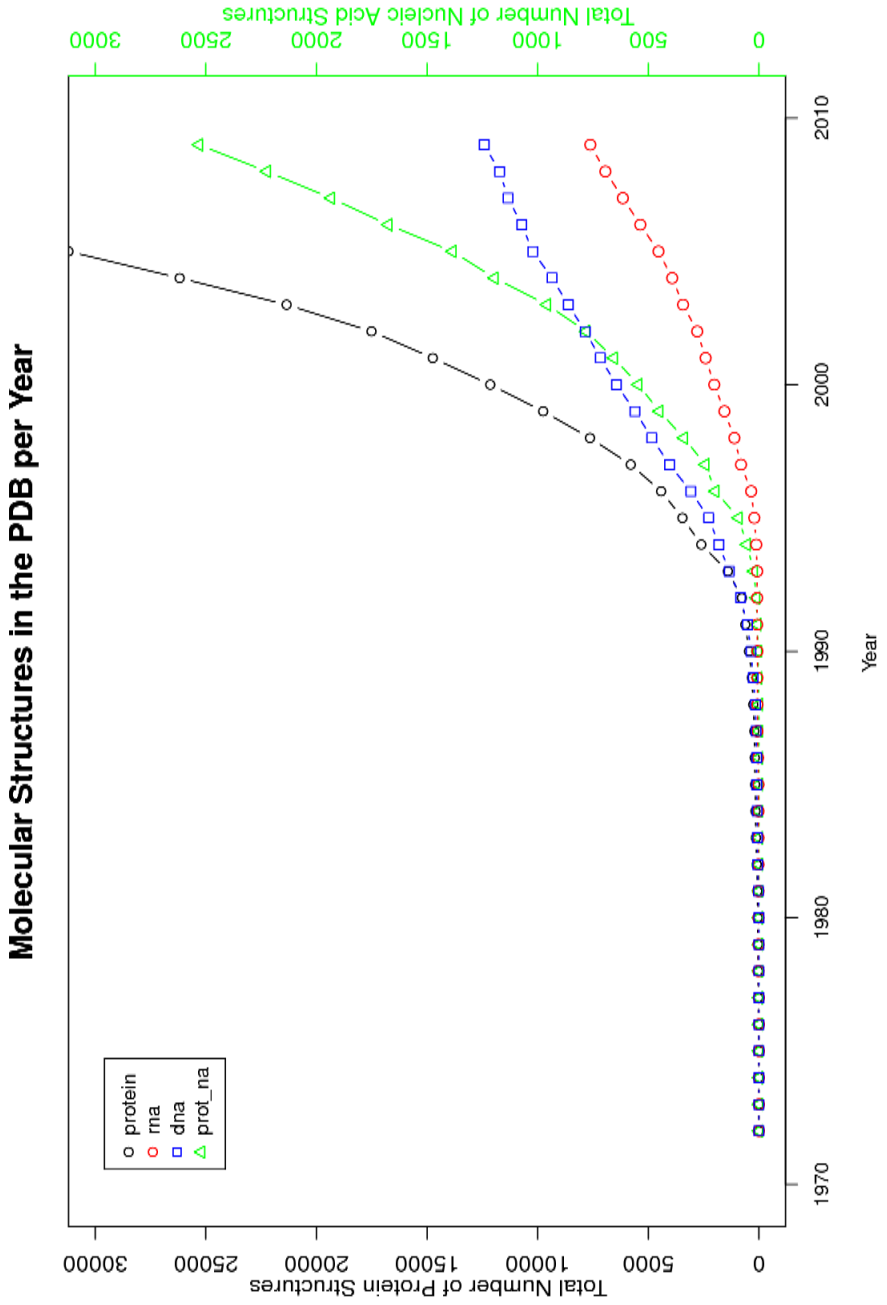


Figure 2.3: The total number of structures available in the PDB up to the end of 2009. The scale of the axis on the left (in black), is ten times that on the right (in green). The black y-axis sets the scale for the number of protein structures. The green y-axis sets the scale for the number of molecular structures containing nucleic acids – 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. The data also tend to show that the number of DNA structures is perhaps growing linearly instead of exponentially. 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 seems to be constant now.

are determined more accurately than the backbone torsion angles, as has been shown by Richardson and collaborators from the analysis of van der Waals steric clashes. This can be seen more clearly in Figure 2.4, reproduced from Richardson's work [11], where the red and orange dots in the backbone atoms region (Figure 2.4 b) denote steric clashes and the green and yellow dots in the base atoms region (Figure 2.4 a) denote very good agreement with expected van der Waals distances.

2.1.1 Combining Fourier Averaging Results and Clustering Analysis

We used standard clustering analysis (CA) techniques (see Appendix B) to classify 18 non-A-RNA, and two A-RNA dinucleotide steps identified by Schneider at al.[13] from grouping the backbone and glycosidic torsion angles found in the 23S ribosomal subunit (PDB_ID:1JJ2). Here we describe these structures using their base-step parameters, that is, three translational parameters (Shift D_x, Slide D_y, Rise D_z), and three rotational parameters (Tilt τ, Roll ρ, Twist ω), in a hexaparametric vector ν:

$$\nu = (D_x, D_y, D_z, \tau, \rho, \omega) \tag{2.1}$$

The results, illustrated in the dendrogram shown in Figure 2.5, were obtained by performing clustering analysis and consensus clustering of the 20 structures kindly provided to us by Schneider et al. [13]. Schneider et al. obtained these structures, which are illustrated in Figures 2.6 and 2.7, by applying a Fourier averaging technique and lexicographical clustering. The methodology we used to produce the dendrogram follows the approach taken by others to recover the Periodic Table classification of the elements from multidimensional property vectors of the elements [22, 23]. Further details of the approach are described in Appendix B

As is clear from Figures 2.5-2.7, which reflect the clustered groups obtained using the clustering techniques described in Appendix B, the 20 structures identified by Schneider et al. fall into seven unique groups based on the relative position of the base side groups, here taken to be an adenine in the 3′ position and a uracil in the 5′ position. Group I contains structure 1 with bases stacked and base-plane normals pointing in opposite directions, Group II includes extended, unstacked conformations with neighboring bases widely displaced and oriented roughly perpendicular to each other in all cases – structures 15, 16, 10, 14. Group III also contains extended conformations but with the uracil on the minor-groove, or sugar [24], edge of adenine and the bases perpendicular to one another. The uracil lies near the "C8-carbon" of adenine in these cases – structures 8, 9, 17. The bases in each of the

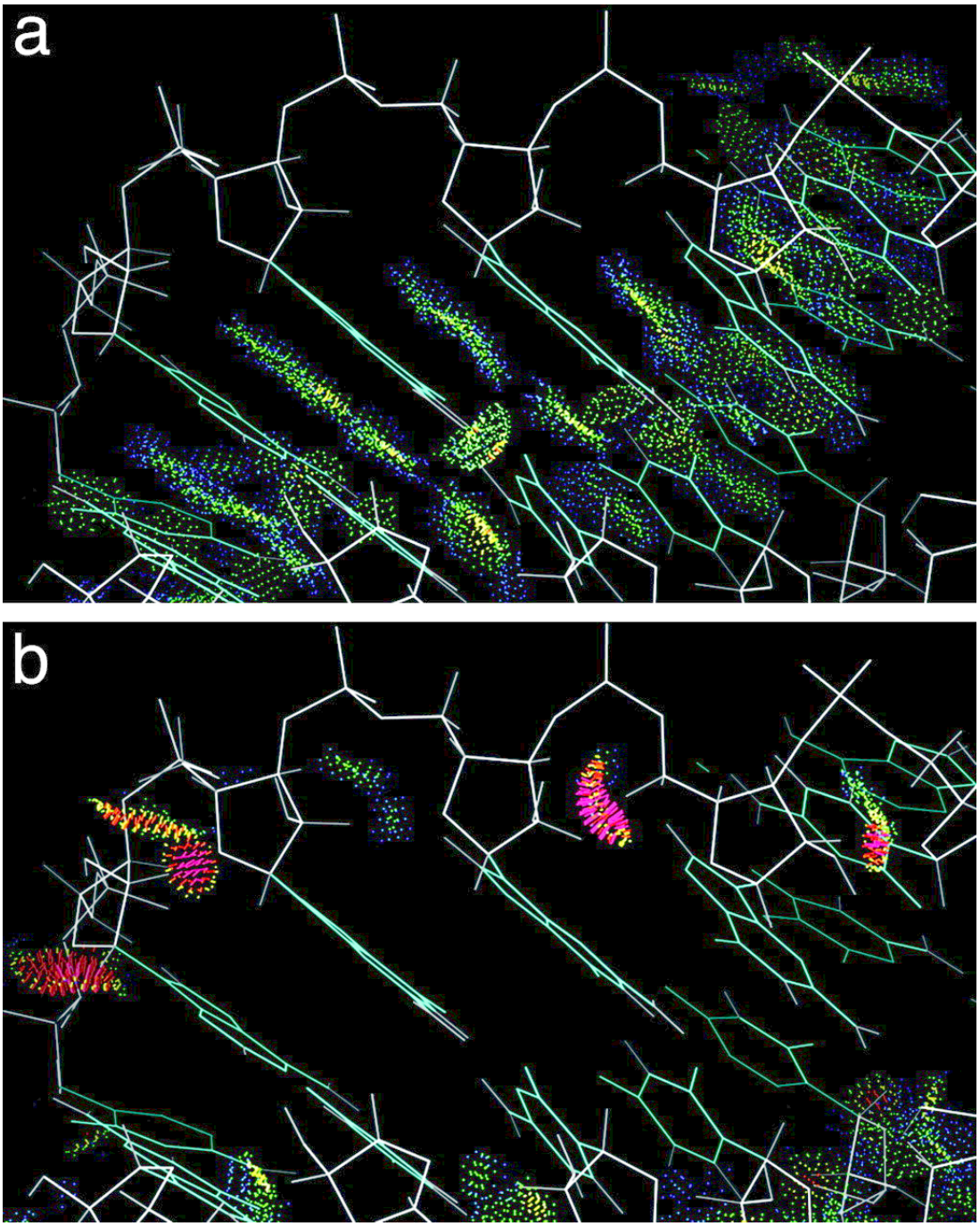


Figure 2.4: Comparison of base vs. backbone structure in RNA, reproduced with permission from Jane Richardson [11] and the publisher. Here the blue and green dots in (a) denote very accurate van der Waals distances, and the red, pink, and orange dots on (b) denote steric clashes, that is, distances outside the acceptable van der Waals range.

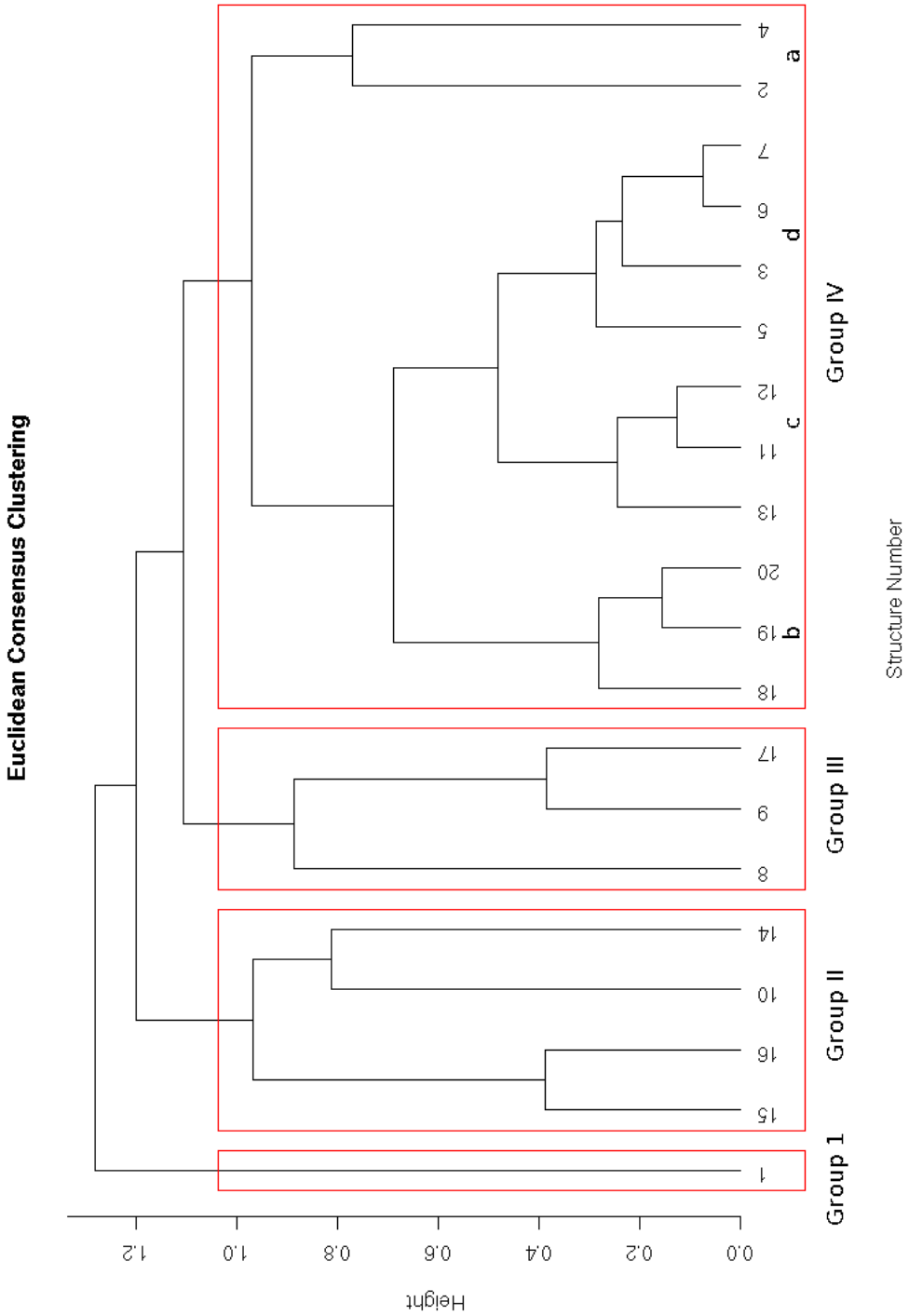


Figure 2.5: Dendrogram showing the results of Euclidean consensus clustering of base-step parameter vectors formed from 18 non-A-type and two A-type rRNA dinucleotides obtained by Schneider et al. [13]. The red rectangles around the branches in the tree have been chosen to highlight a four group clustering solution. The height of the dendrogram represents the similarity between dinucleotide steps across the various clustering methodologies described in detail in Appendix B.

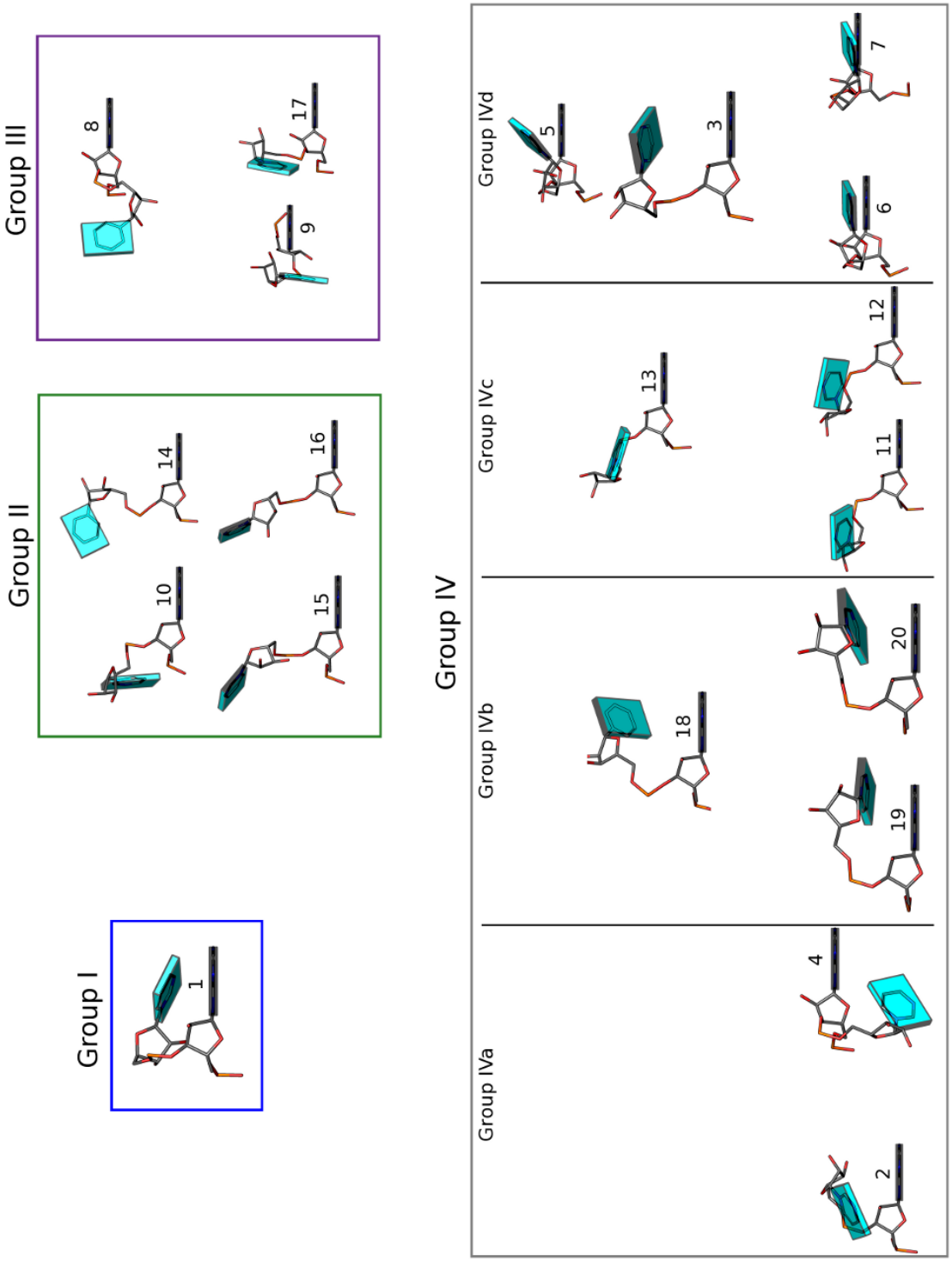


Figure 2.6: Molecular images of non-A-RNA dinucleotide (ApU) structures identified by Schneider at al. [13] and organized by clusters obtained from consensus clustering of their hexadimensional base-step parameter vectors. The structures are centered on the reference frame of the adenine base, with the (shaded) minor-groove, or sugar, edge of the rigid block on adenine facing the viewer.

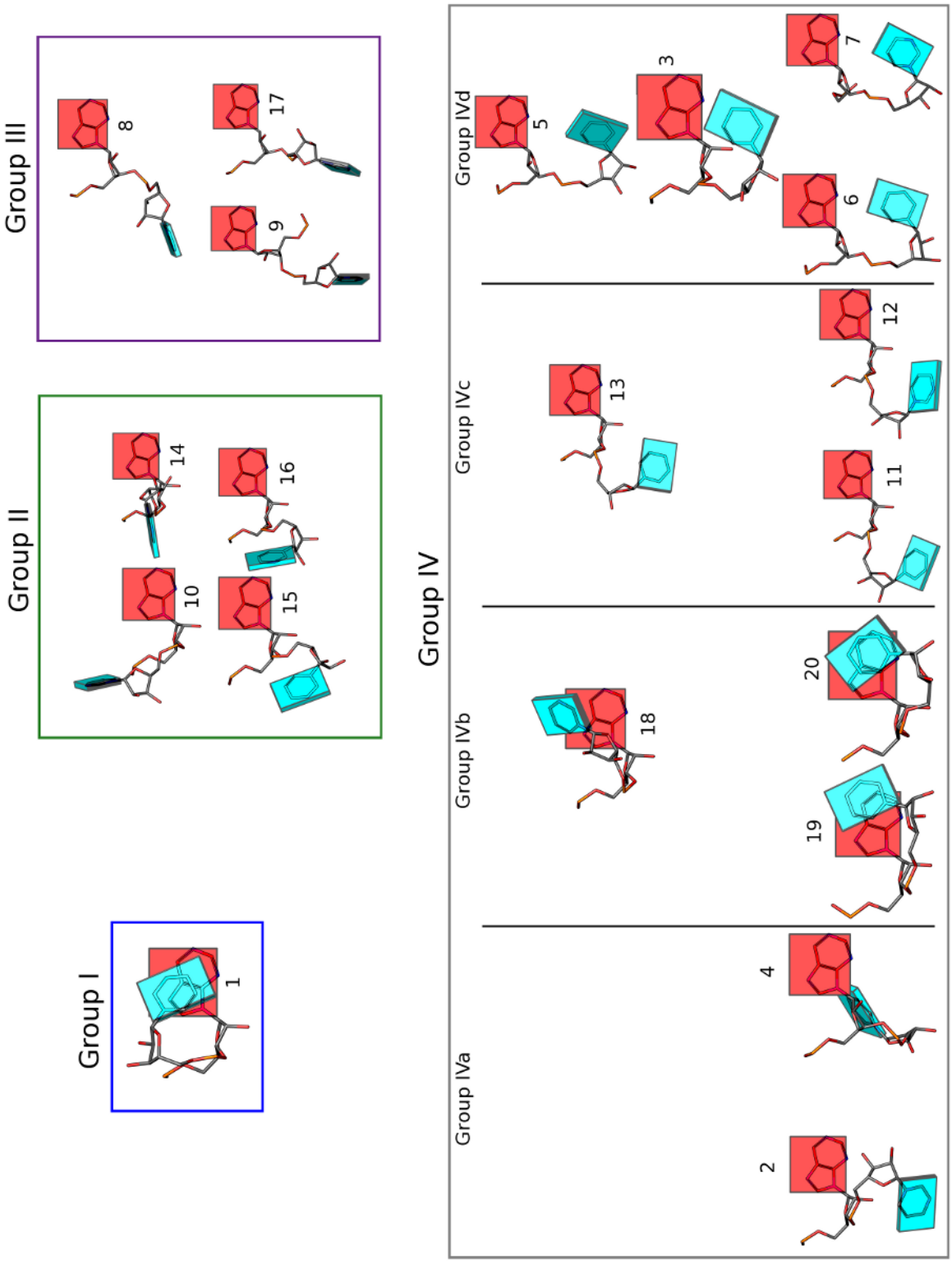


Figure 2.7: Top view of the non-A-RNA dinucleotides (ApU) from Figure 2.6 centered on the reference frame of adenine with its minor-groove, or sugar, [24] edge oriented towards the bottom of the image and the major-groove, or Hoogsteen, [24] edge oriented towards the top.

conformers in Group IV are more parallel to one another but are displaced in four different ways. The bases in Group IVa – structures 2, 4 – are unstacked and neither parallel nor perpendicular. Those in Group IVb – structures 18, 19, 20 – are closely related to A-RNA with parallel and stacked bases. The bases in Group IVc – structures 11, 12, 13 – are parallel but unstacked, with uracil on the major-groove, or Hoogsteen [24], edge of adenine. The bases in Group IVd – structures 3, 5, 6, 7 – are also parallel and unstacked, but with the long axis (y-axis) of uracil perpendicular to that of adenine. As is clear from Figures 2.5 and 2.6, the conformers in subgroups IVc and IVd are closely related, and the dimers in these two subgroups are more closely related to those in subgroup IVb than to those in subgroup IVa.

In order to map the groups, which were obtained by clustering with respect to the 23S subunit of the ribosome, we computed the root-mean-square deviation (RMSD) [ii] between the average of the step parameters of each clustered group and the step parameters of all dinucleotide residues of the 23S subunit (PDB_ID:1JJ2). Before computing the RMSD values we normalized the step parameters using the ratio between the difference of step parameters to their minimum value, and their range. This is expressed in Equation B.1.

The RMSD results for each group are show in histogram plots in Figures 2.8, and 2.9. Not surprisingly we see in the histogram corresponding to Group IV, that most of the dimer steps in the ribosome resemble it more closely (smallest mean RMSD) than the other groups. This is clearly due to the fact that subgroup IVb contains the A-RNA (structure 20) and AII-RNA (structure 19) dimer steps. The mean RMSD values also give a good idea of the structural similarity between each one of the four groups found by consensus clustering and the most populated base-step conformation adopted by RNA steps. That is, we can order the clustered groups by mean RMSD in the order Group I > Group III > Group II > Group IV.

The distribution of conformer groups is presented in an alternative way in Table 2.2, where we count all base steps which fall below an RMSD score of 10. It is important to note here that the computed RMSD is not an all-atom RMSD, as is commonly used after superimposition of molecular structures, but the deviation between an average base-step parameter vector and each step (described as a vector) in the 23S subunit, where there is no need for superimposition since the base-step parameter vectors are computed using a middle step reference frame between the step bases [25]. The reason for choosing 10 as the RMSD cutoff value is based on visual inspection of superimposed reconstructed structures at different RMSD values. For example, for Group I, if we reconstruct the ribosomal steps with an RMSD of

[ii] For the mathematical definition of RMSD see equation B.4 in Appendix B.

10 or less using their base-step parameter values, we obtain the set of superimposed structures shown in the left panel of Figure 2.10. But if we reconstruct the base-steps using the set of parameter values which fall within a larger RMSD cutoff value of 15, as can be seen in the right panel of Figure 2.10, there will be three new structures (whose second bases are shown in the blue "clouds") in the set of superimposed structures. The latter clearly do not overlap well with the structures which fall under the lower cutoff value and therefore are unlikely candidates for being part of such a conformational group.

In Table 2.2 we see that the total number of structures which fall into any of the four conformational groups is only 31 percent of the total number of steps in the 23S subunit of the ribosome, meaning that 69 percent of base steps fall outside the conformational groupings. We believe this result might stem from three factors: 1) the A-RNA structure identified by Schneider et al. (Structure 20 in Figures 2.6 and 2.7) differs from canonical A-RNA in that the rise value is 4.39 Å, rather than the standard value of 3.30 Å obtained for A-RNA fibers by Arnott and collaborators [26] and seen in other RNA structures [27]. This might have an effect on the number of structures which can be grouped under the A-RNA-like Group IVb category; 2) our treatment ignores the richer classification of A-RNA conformations described by Schneider et al. [13] by using only conformers 19 and 20 in Group IVb and omitting the remaining 12 A-RNA-like conformers that they report; and 3) the mix here of results from Fourier averaging of sugar-phosphate backbone torsion angles and base-step parameters consensus clustering, might omit intermediate base-step conformations which could be highly populated.

We therefore chose to investigate the whole space of RNA conformations in the 23S subunit of the ribosome (PDBID_ID:1JJ2) in terms of base-step parameters alone with the clustering analysis validation techniques described in the following section.

2.1.2 Selection of a Clustering Methodology

In order to analyze our dataset of base-step parameters we have used clustering analysis methods. Clustering analysis methods can be broadly classified into two main categories, either partitional or hierarchical. In either case the main problem one faces for classification purposes is that of deciding the optimal number of hierarchies or partitions into which the analyzed data can be split. To obtain a criterion for an optimal number of clusters, and also to decide which method might be better for our dataset, we have used two types of cluster-validation techniques. They are known as internal measures and stability measures. Full details of the definitions of such measures are provided in the literature [28, 29] and in Appendix B. To perform the validation analysis we used a freely available

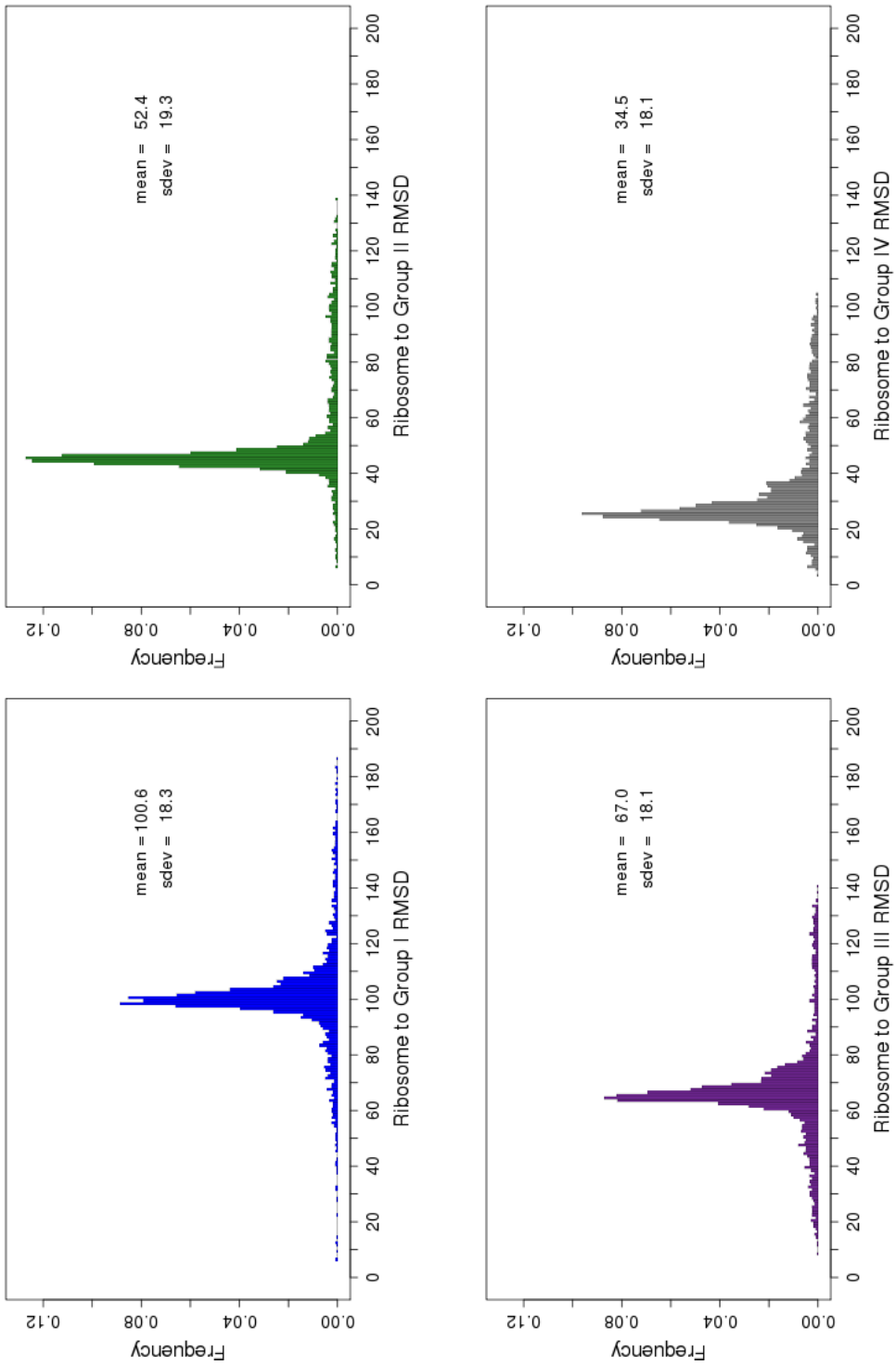


Figure 2.8: Root-mean-square deviation of the dinucleotide steps in the 23S subunit of the ribosome from the main four groups shown in Figures 2.6 and 2.7. The colors of the histograms are the same as those of the outlines surrounding the structures in Figure 2.6

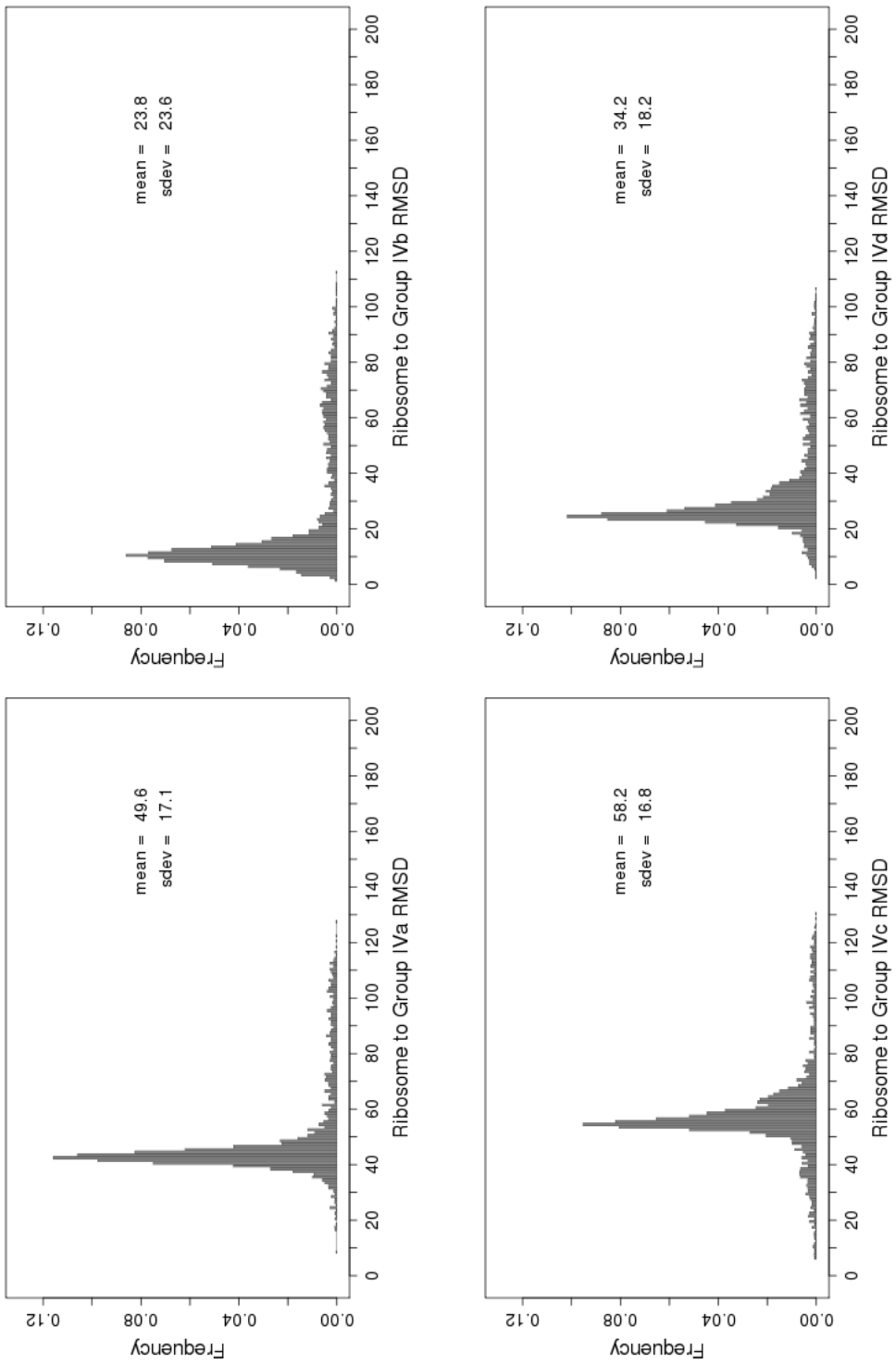


Figure 2.9: Root-mean-square deviation of dinucleotide steps in the 23S subunit of the ribosome from the four types of conformers within the Group IV category shown in Figures 2.6 and 2.7. Note that most of the dimers resemble the A-RNA-like conformers in subgroup IVb in the upper left histogram where there is the highest proportion of small RMSD values.

	RMSD Cutoff Value		
Group	**10**	**15**	**20**
I	3 (0.11)	7 (0.25)	7 (0.25)
II	5 (0.18)	13 (0.47)	25 (0.91)
III	1 (0.04)	5 (0.18)	23 (0.84)
IVa	1 (0.04)	1 (0.04)	7 (0.25)
IVb	807 (29.31)	1696 (61.61)	1965 (71.38)
IVc	9 (0.33)	22 (0.80)	41 (1.49)
IVd	35 (1.27)	99 (3.60)	191 (6.94)
Total	861 (31.28)	1843 (66.95)	2259 (82.06)

Table 2.2: Number of base steps in the 23S subunit of the *Haloarcula marismortui* ribosome with RMSD values less than or equal to 10, 15, and 20 from the average base-step vectors of the four groups of non-A-type RNA dinucleotide conformations. The numbers in parentheses are the corresponding percentages computed with respect to the 2753 base steps present in the 23S subunit.

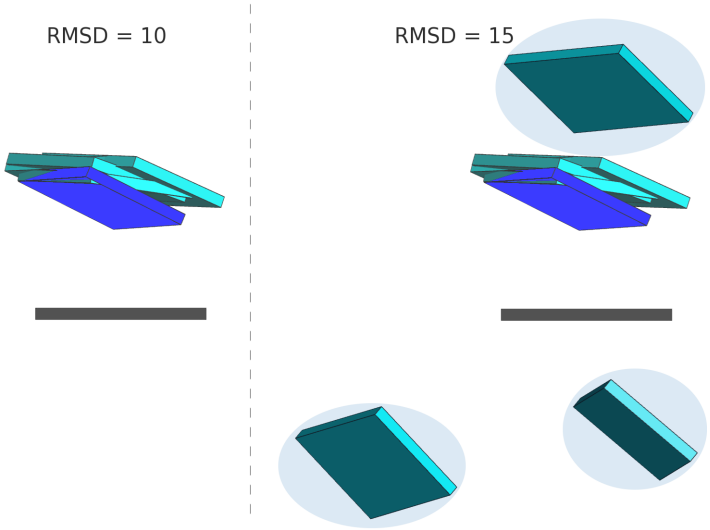


Figure 2.10: Rigid-block representation of dinucleotide steps. The major-groove side of the first nucleotide block is oriented towards the viewer and colored black. **Left:** Drawn in blue, the block representing the Group I cluster from Figure 2.6. Superimposed on the Group I cluster are three structures whose step-parameter RMSD's with respect to the Group I cluster are less than or equal to 10. **Right**: With an RMSD less than or equal to 15 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 2.6, which is again drawn in blue.

cluster validation package called clValid [29], which is implemented in the **R** [30] statistical analysis package.

In Figures 2.11 and 2.12 we present the results for internal and stability validation of clustering techniques for the base-step parameters in the 23S ribosomal subunit. In the clustering analysis literature it is customary to use a variable k to define the number of clusters which we will be using frequently in the text.

In Figure 2.11 we see the validation scores (connectivity, silhouette width, and Dunn index) for cluster groups ranging from $k = 2$, up to $k = 80$, and evaluated using both hierarchical (hierarchical, diana) and partitional (kmeans, pam, som, sota) methods. An optimal connectivity measure must be minimized, and an optimal average silhouette width (silhouette) or Dunn index must be maximized. Keeping in mind the previously described optimal trends, we see that for the three graphs shown in this figure that the hierarchical[iii] method (series of numeral 1's shown in black) performs better in terms of the connectivity and Dunn index for the entire range of eighty clusters $k = (2 - 80)$, and it is also the best performer in silhouette from $k = 12$ onwards.

Stability measures (Figure 2.12) are well suited for highly correlated datasets with linear correlations between many variables, but they are not very useful for our dataset, where the only correlations between variables occur for shift vs. twist and rise vs. tilt, as can be seen from the scatterplots and correlation coefficients reported in Figure 2.14. That is, of the fifteen possible correlated pairs, only two are linearly correlated, and such correlations are not very strong. The values of these correlation coefficients are less than 0.95, meaning that less than 90% of the variation between variables can be accounted for by a linear relationship between them. We, nevertheless, include the cluster stability measures for completeness.

We have computed three measures of stability for each of the hierarchical and partitional methods, namely, the average proportion of non-overlap (*APN*), the average distance (*AD*), and the average distance between means (*ADM*). The details of such measures are given in Brock et al. [29] and in Appendix B. The best clustering methodology will have low values of these quantities. As seen in Figure 2.12, the best performing method according to the *APN* and *ADM* criteria is sota (second-order

[iii] The hierarchical method can be carried out in various ways, for example, using different metrics (i.e. measures the distance between data-points), and different grouping methodologies. In this specific case we are referring to a hierarchical method carried out in an agglomerative, or bottom-up manner, meaning that we start by grouping pairs of steps which are initially ungrouped, and group them progresively with other ungrouped steps using the average method criteria (explained in Appendix B). The criteria in turn, depend on a distance definition, which in this case is that of the Euclidean (square root of the sum of the squared differences between vector elements) metric (defined in Appendix B).

tolerance analysis). One can clearly see from the corresponding series of cyan numerals (5's) that the sota values are smaller consistenly than those for all other methods for almost all cluster sizes up to $k = 70$. The best performers (those with smaller values) over the whole range of k according to the *AD* criteria are pam (series of green 3's) and som (series of blue 4's). Notice that in the *AD* and *ADM* plots, perhaps with the exception of the sota method, all methods seem to follow the same trend, meaning that they would predict the same number of optimal clusters.

2.1.3 Splitting A-RNA-Like and Non-ARNA-Like Steps.

In all cases the validation results tell us that the best overall number of clusters is two, which is not surprising since we do not filter the A-RNA structures from our data set. The two main groups are thus A-RNA type base-steps, and non-A-RNA steps.

We focus our attention on the group of structures which differ from A-RNA. To define an A-RNA-like step we use the standard base-step parameter values of the canonical A-RNA helix determined by Arnott and collaborators [26]. We find that there are 797 non-A-RNA like steps (about 29% of the total number of steps) in the 23S rRNA subunit by counting the steps which have a root-mean square deviation between the ribosomal steps and the standard A-RNA step greater than 18 Å (see Figure 2.13). The standard base-step parameter values for common double-stranded RNA and DNA are listed in Table 2.3.

Structure Name	Shift (D_x)	Slide (D_y)	Rise (D_z)	Tilt (τ)	Roll (ρ)	Twist (Ω)	**Structure Source**	**Method**
A-DNA	0.36	-1.39	3.29	2.5	12.5	30.2	Arnott [31]	fiber-diff.
B-DNA	0.44	0.47	3.33	4.6	1.8	35.7	Arnott [31]	fiber-diff.
A-RNA	0.00	-1.48	3.30	0.0	8.6	31.6	Arnott [31]	fiber-diff.
A'-RNA	0.05	-1.88	3.39	-0.1	5.4	29.5	Arnott [31]	fiber-diff.
AII-RNA	1.01	-2.52	3.33	2.9	9.8	25.1	Schneider [13]	X-ray

Table 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.

Using the filtered dataset, which we refer to as the non-A-RNA dataset, we repeated the clustering validation analysis (Figure 2.16). As was the case for the whole dataset of "A-RNA like" and "non-ARNA-like" base-steps, the best clustering methodology in terms of the validation results is hierarchical clustering. This is clear from each of the validation score plots where the values for the hierarchical method (black number ones) are small for the connectivy and silhouette scores and large for the Dunn index score.

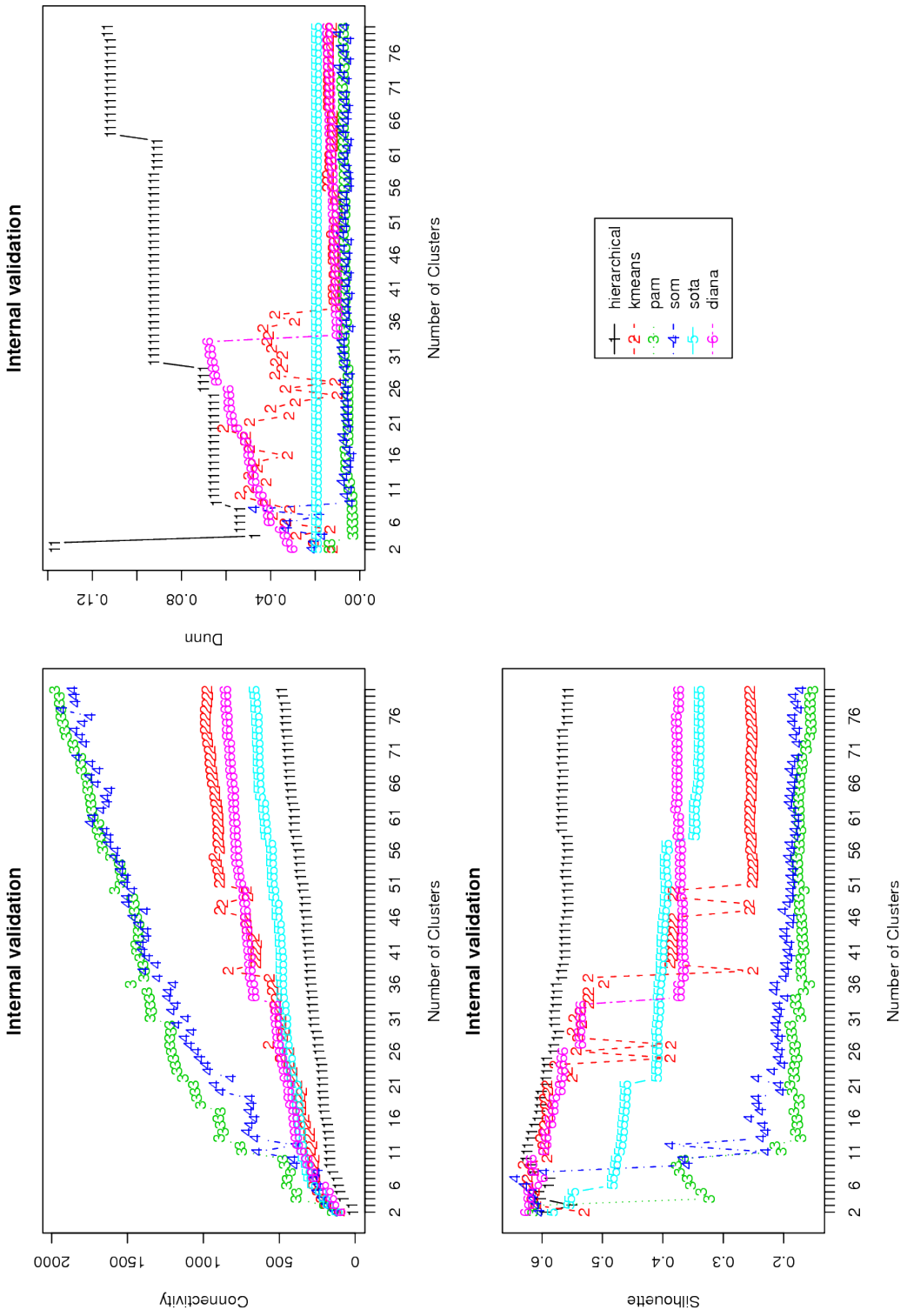


Figure 2.11: Cluster validity scores for internal measures of the distances between all sets of base-step parameters in the 23S subunit of ribosomal RNA (PDB_ID:1JJ2). Notice how the hierarchical method, labeled as one and drawn as series of black numerals, behaves better for the whole range of connectivity (smaller values) and Dunn (higher values) scores, and how it also outperforms all others after $k = 12$ for silhouette (higher values) scores.

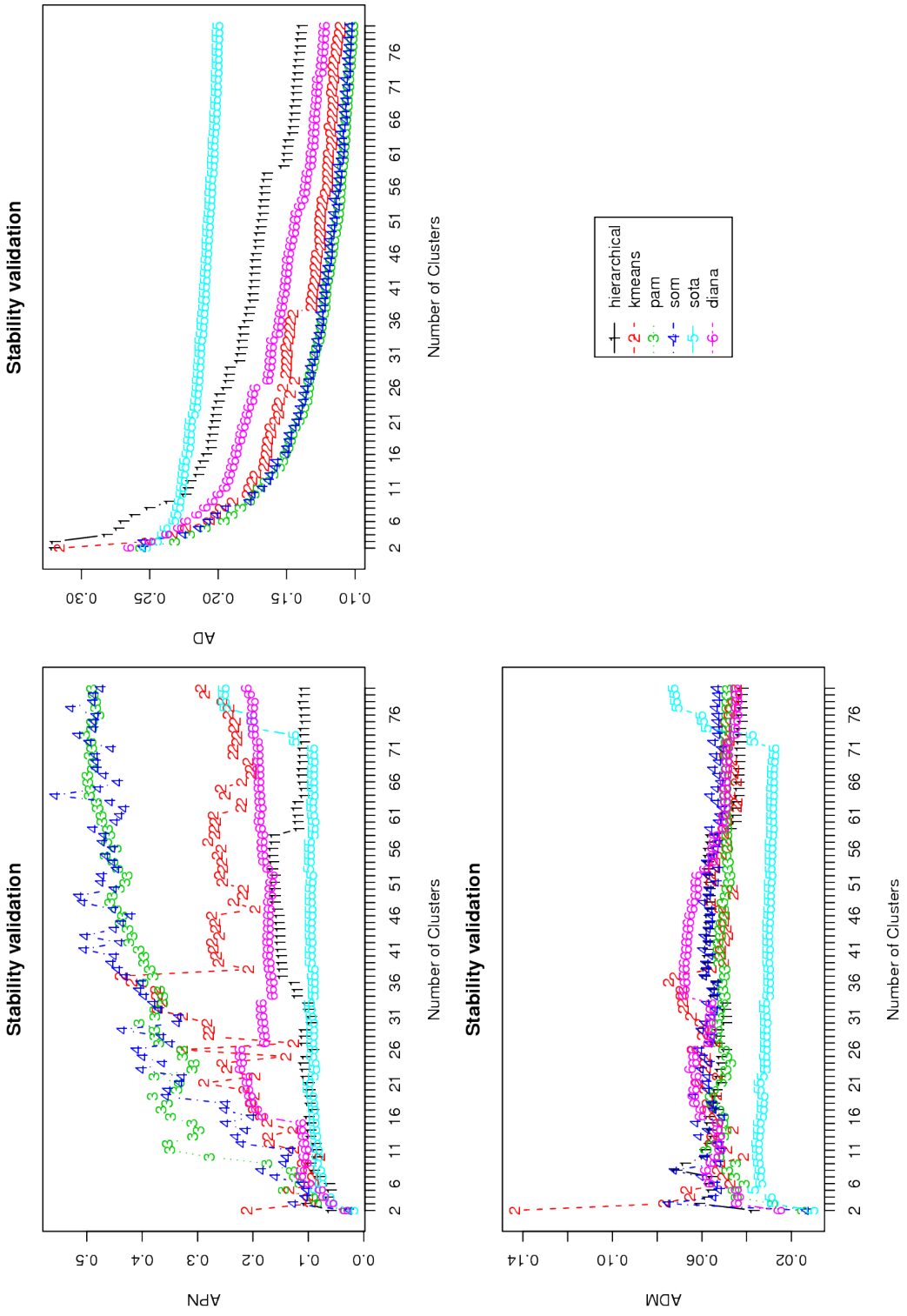


Figure 2.12: Cluster validity scores for stability measures of the distances between all sets of base-step parameters in the 23S subunit of ribosomal RNA (PDB_ID:1JJ2). The average proportion of non-overlap (APN), average distance (AD), and average distance between means (ADM) are plotted against the number of clusters. Note that stability measures work especially well for datasets of highly correlated data, which is not the case for our dataset. These values are shown here for consistency with the validation package clValid.

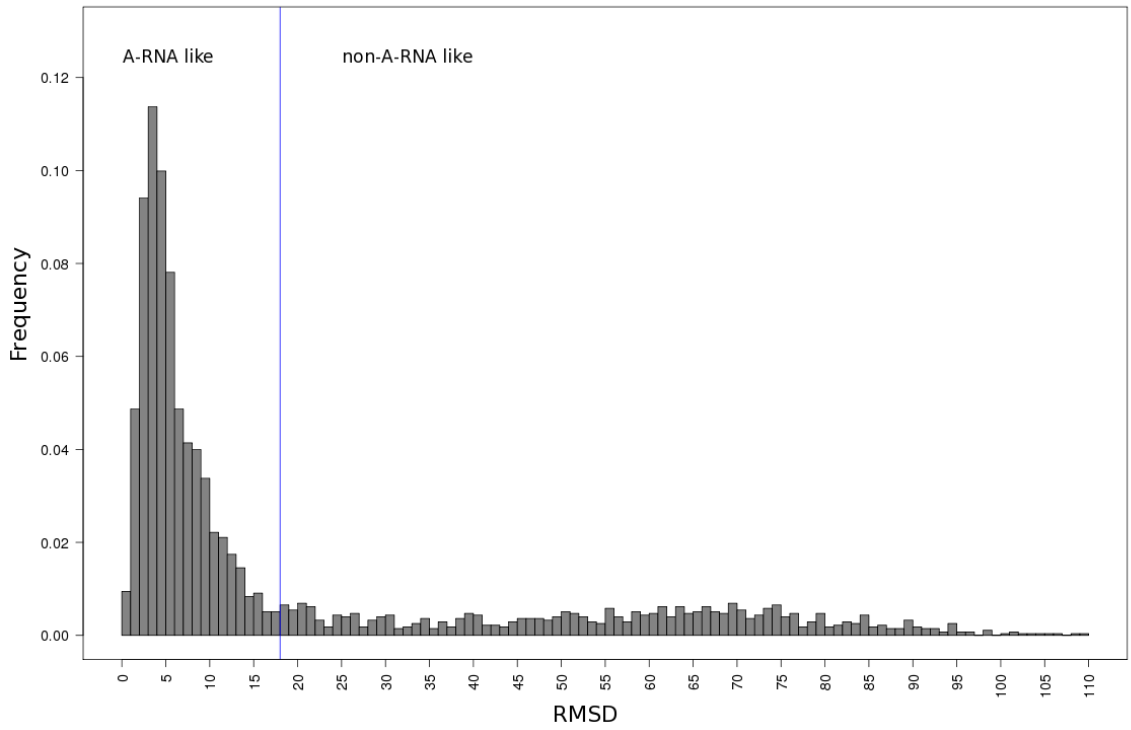


Figure 2.13: Histogram of RMSD values between base-step parameters of the 23S subunit of ribosomal RNA and the standard A-RNA base-step parameters derived from the work of Arnott and collaborators [26]. The black vertical line drawn at an RMSD value of 18 denotes the limit used to distinguish A-RNA-like conformations, from non-A-RNA-like conformations. A histogram which is practically identical to the one shown here is obtained if, prior to normalization of the dataset of step-parameters, their rotational components (i.e. tilt, roll, twist) are expressed in radians instead of degrees.

According to the Dunn index score (upper right plot of Figure 2.16), the optimal number of clusters is $k = 67$. The Dunn index works under the idea of finding the best possible separation and compactness (see Appendix B for definitions of separation and compactness) between clusters. Thus, the grouping of the 797 base-steps into 67 groups should produce well-separated and compact groups. The other two indices (connectivity and silhouette) show, as for the whole dataset, that the optimal number of clusters of the non-A-RNA dataset is two. The presence of shoulders in the connectivity and silhouette plots of the hierarchical clustering data is, nonetheless indicative of optimal clusters. Moreover these shoulders occur at $k = 67$ for the connectivity (upper left) and silhouette (lower left) plots.

We used the 67 clusters, given by the hierarchical method, and their corresponding step-parameter values to reconstruct ApU dinucleotide step structures with the 3DNA software. Figure 2.17 illustrates the first seventeen groups of structures with ten or more members, which account for 80 percent of the non-A-RNA steps. We also plot in the lower right corner of Figure 2.17 the set of 20 structures derived from the work of Schneider et al.[13], and the whole set of non-A-RNA dinucleotide steps in a

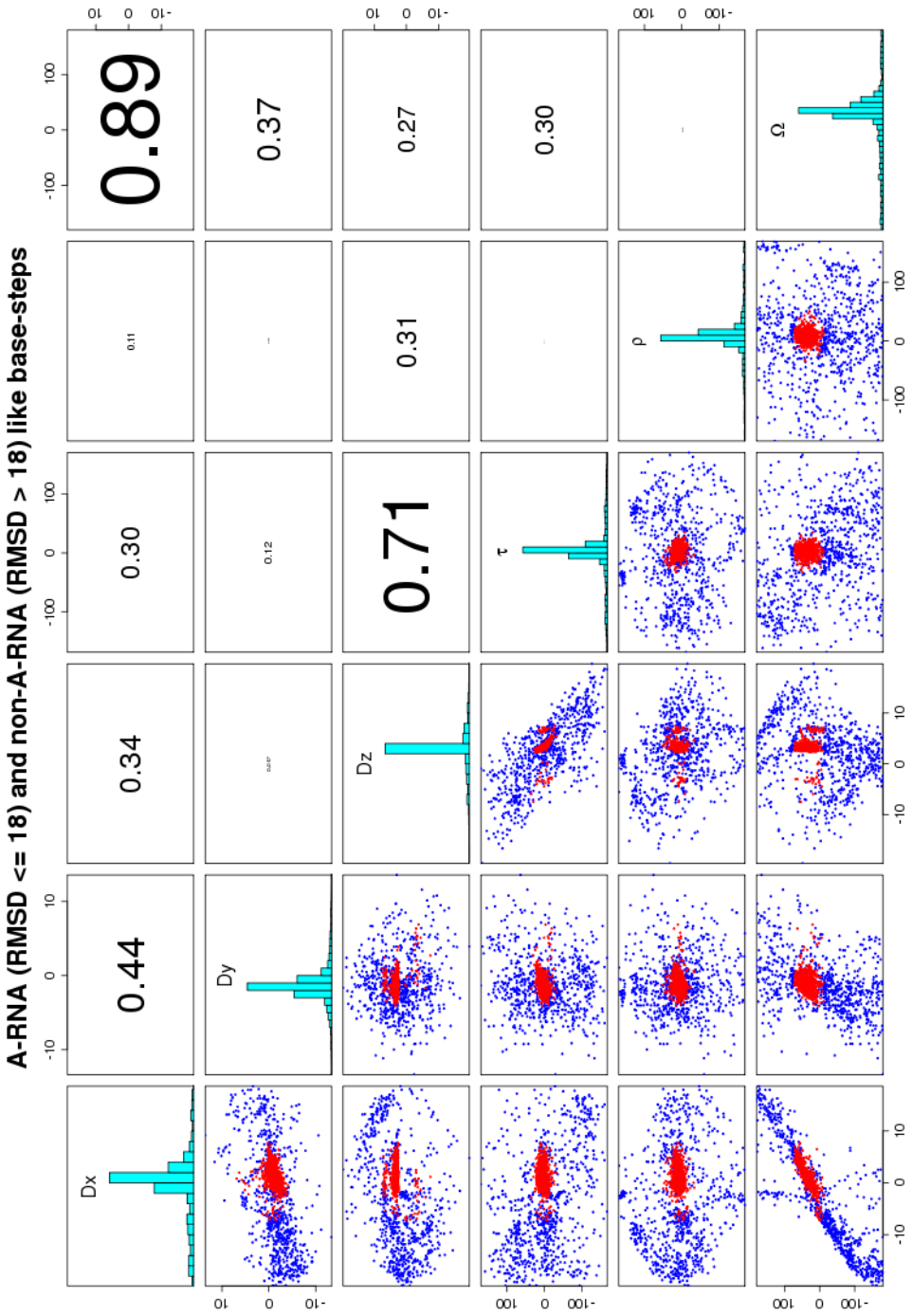


Figure 2.14: Scatterplots for base-step parameters, Shift (D_x), Slide (D_y), Rise (D_z), Tilt (τ), Roll (ρ), and Twist (Ω), for the A-RNA (colored red) and non-A-RNA (colored blue) datasets. The datasets were split apart based on an RMSD cutoff less than or equal to 18, between the base-step parameters of the A-RNA structure determined by Arnott et al. [26] and all the base-step parameters of the 23S subunit of ribosomal RNA (PDB_ID:1JJ2). Here the correlation coefficients of each of the scatterplots (both red and blue points) shown in the lower half of the graph are listed in the mirror position in the upper half of the diagram, i.e., the Shift-Twist (D_x, Ω) correlation coefficient (0.89) of the plotted data shown in the lower left corner, is printed in the upper right corner.

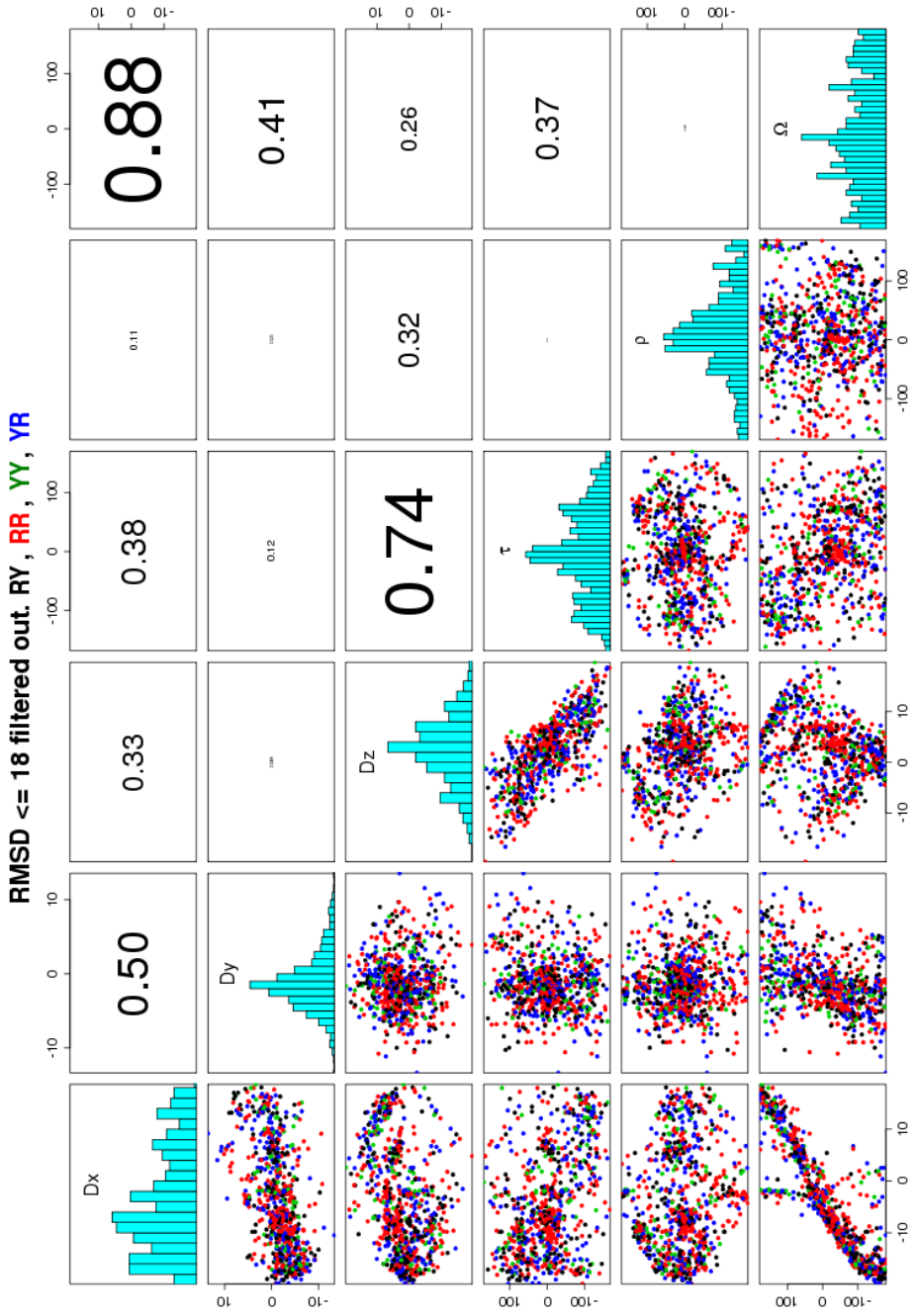


Figure 2.15: Scatterplots for base-step parameters, Shift (D_x), Slide (D_y), Rise (D_z), Tilt (τ), Roll (ρ), and Twist (Ω), for the non-A-RNA dataset (taken as the steps with RMSD's greater than 18). Points are color-coded in terms of base compositions; purine-pyrimidine (black); purine-purine (red); pyrimidine-pyrimidine (green); and pyrimidine-purine (blue). Here the correlation coefficients of each of the scatterplots shown in the lower half of the graph are listed in the mirror position in the upper half of the diagram, i.e., the Shift-Twist (D_x, Ω) correlation coefficient (0.88) of the plotted data shown in the lower left corner, is printed in the upper right corner. As seen from the coloring scheme there is no clear sequence preference for single-stranded purine-pyrimidine (RY), purine-purine (RR), or pyrimidine-pyrimidine (YY) steps.

common reference frame on the adenine of an ApU step. All structures are centered using the standard reference frame embedded in the first base, which in our reconstructions corresponds to a red block representing adenine. The minor-groove, or sugar, edge of the adenine is oriented to the left, the major-groove, or Hoogsteen, edge to the right, and the so-called Watson-Crick base-pairing edge is pointed toward the viewer.

Comparison of the 17 groups of non-A-RNA dinucleotide steps with those found by Schneider and collaborators, shows that there are no steps in the Schneider et al. dataset represented on the major-groove (right) side of the red block representing adenine, that is, the right side of the red adenine block, although there are many such steps in the crystal structure. Many of the "missing" steps lie in the group labeled as g7 in Figure 2.17. The cyan blocks representing uracil in this group orient their planes orthogonally to the major-groove side of the red block representing adenine. On the other hand, although the 17 groups are not as compact as the observed data, they hint of the geometrical preferences of the space of dinucleotide step-parameters.

We include images of the non-A-RNA base-steps in Figure 2.17 in order to give the reader an idea of the complexity of the conformation space described from a base-viewed perspective rather than of the more common backbone perspective. As is well known, small changes in backbone can have drastic effects on the arrangement of bases, for example converting right-handed helices to left-handed helices [32]. On the other hand, large concerted changes in the torsions can preserve the structures of RNA and DNA helices [33].

The composite image of conformations suggests that the task of finding order in this broad range of possible conformations is somewhat analogous to solving a ball puzzle. There are various ball puzzles, but the one which most resembles this case is that of a so-called masterball puzzle, which consists of a sphere with two parallel lines at roughly 60° of latitude north and south from the sphere's equator, and eight meridians, as seen in the left side of Figure 2.18. We believe the puzzle can be effectively solved by using appropriate validated clustering analysis techniques, in the spherical coordinate space (latitude and longitude values) of the dinucleotide steps.

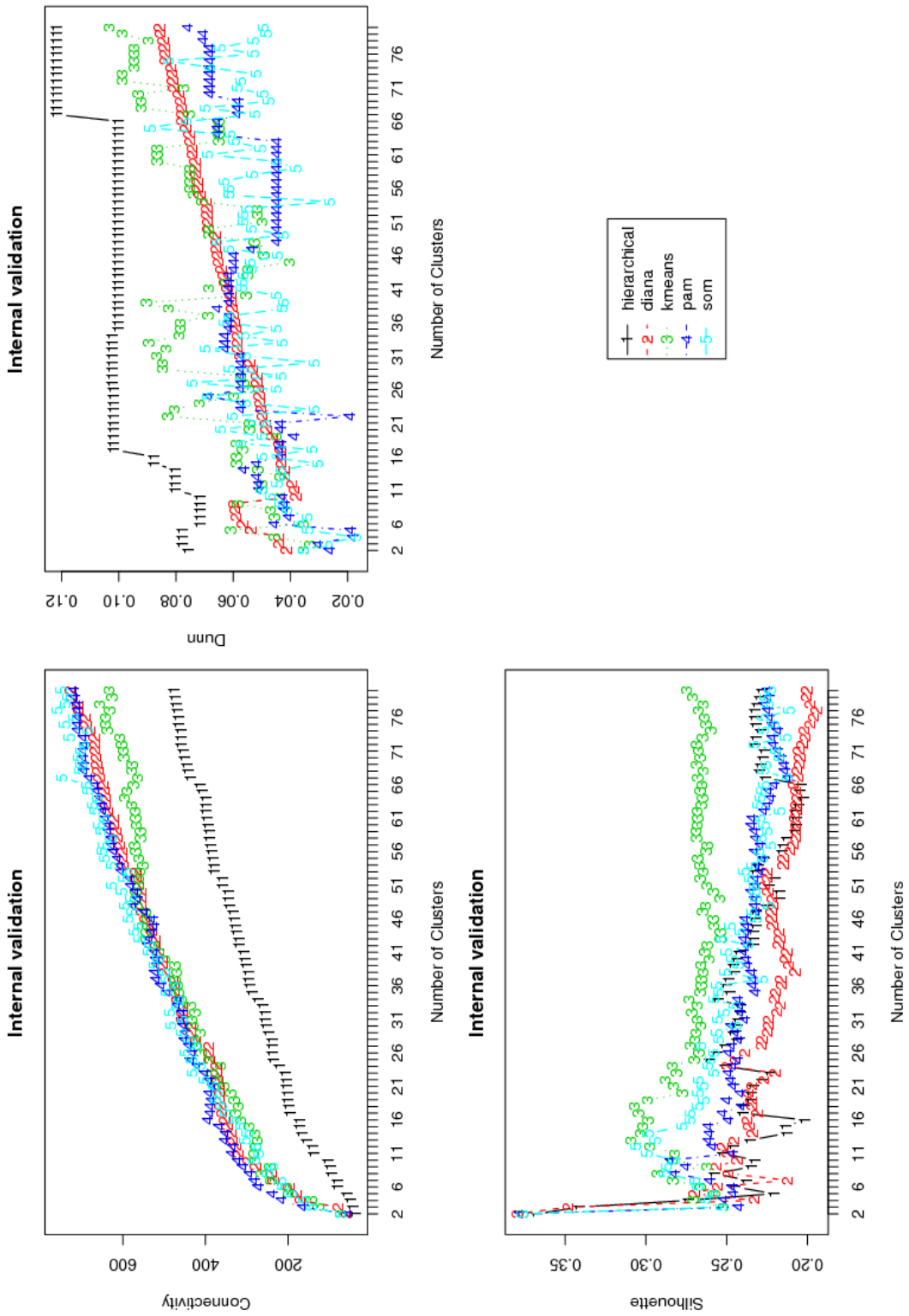


Figure 2.16: Cluster validity scores for the distances between all sets of base-step parameters in the non-A-RNA dataset. It can be seen clearly that the optimal method for clustering is the hierarchical one, as measured by lower values in the connectivity scores, and higher values in the Dunn score. The optimal number of clusters given by the Dunn score is 67. We also see shoulders at $k = 67$, for the silhouette scores, and a "small" shoulder for the connectivity score.

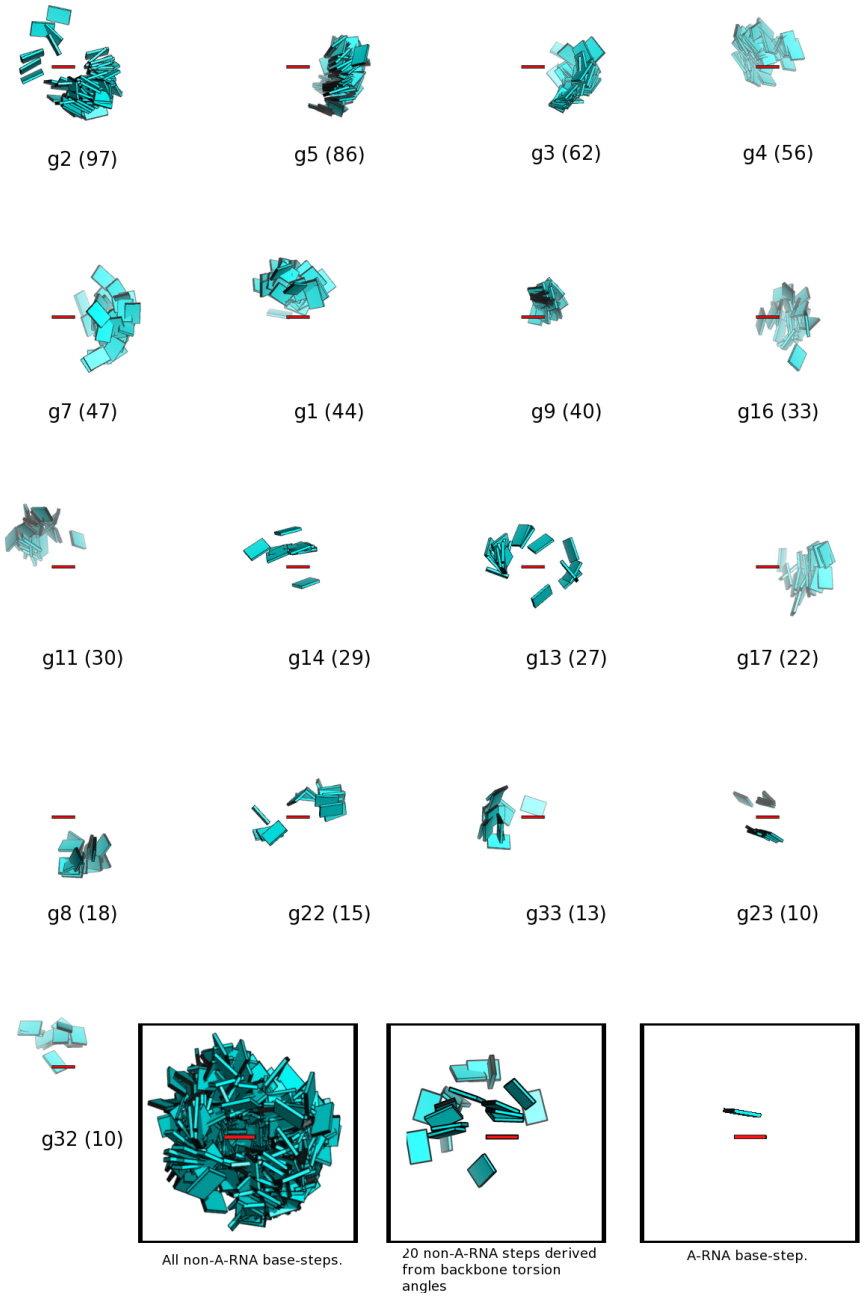


Figure 2.17: Superimposed images of sequential base steps found in the rRNA 23S subunit structure (PDB_ID:1JJ2) and represented in all cases by an ApU step generated from the observed step parameters using 3DNA. Examples are shown for the first 17 of the 67 groups of dinucleotide conformers found using hierarchical clustering. 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 by Schneider et al. [13] from the torsion angles in the same structure. Notice how the only "hollow" section of the "onion" formed by the full space of base-step conformations is that corresponding to the Watson-Crick base-pairing edges, the space that would be occupied by the bases paired to adenine by Watson-Crick or other related non-canonical base-pairing interactions.

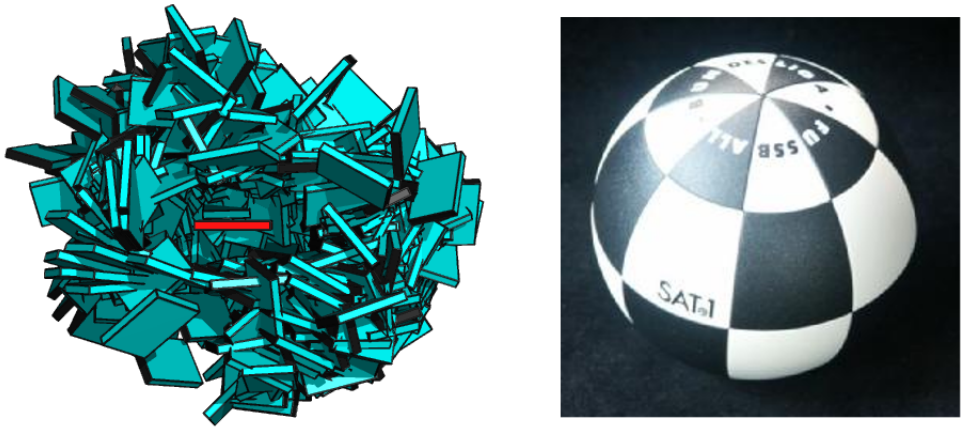


Figure 2.18: Rebuilt-base-step parameters of the 23S subunit of the ribosome using the reference frame of adenine (drawn as a red block) in the left side of the figure, compared to a jumbled masterball puzzle on the right side. The coordinates of the cyan uracils can be mapped in the latitude/longitude (spherical angle) space of the masterball.

References

[1] Olson, W. K. and Flory, P. J. (1972) Spatial Configurations of Polynucleotide Chains. I. Steric Interactions in Polyribonucleotides: A Virtual Bond Model. *Biopolymers,* **11**, 1–23.

[2] Saenger, W. (1984) Principles of Nucleic Acid Structure, Springer-Verlag, London.

[3] Gautheret, D., Major, F., and Cedergren, R. (1993) Modeling the Three-dimensional Structure of RNA Using Discrete Nucleotide Conformational Sets. *Journal of Molecular Biology,* **229**, 1049–1064.

[4] Wimberly, B. T., Brodersen, D. E., Clemons, W. M., Morgan-Warren, R. J., Carter, A. P., Vonrhein, C., Hartschk, T., and Ramakrishnan, V. (2000) Structure of the 30S Ribosomal Subunit. *Nature,* **407**, 327–339.

[5] Schluenzen, F., Tocilj, A., Zarivach, R., Harms, J., Gluehmann, M., Janell, D., Bashan, A., Bartels, H., Agmon, I., Franceschi, F., and Yonath, A. (2000) Structure of Functionally Activated Small Ribosomal Subunit at 3.3 Å Resolution. *Cell,* **102**, 615–623.

[6] Ban, N., Nissen, P., Hansen, J., Moore, P. B., and Steitz, T. A. (2000) The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution. *Science,* **289**, 905–920.

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

[8] 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.

[9] Reijmers, T. H., Wehrens, R., and Buydens, L. M. C. (2001) The Influence of Different Structure Representations on the Clustering of an RNA Nucleotides Data Set. *Journal of Chemical Information and Computer Science,* **41**, 1388–1394.

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

[11] 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.

[12] Hershkovitz, E., Tannenbaum, E., Howerton, S. B., Sheth, A., Tannenbaum, A., and Williams, L. D. (2003) Automated Identification of RNA Conformational Motifs: Theory and Application to the HM LSU 23S rRNA. *Nucleic Acids Research,* **31**, 6249–6257.

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

[14] Hershkovitz, E., Sapiro, G., Tannenbaum, A., and Williams, L. D. (2006) Statistical Analysis of RNA Backbone. *Transactions on Computational Biology and Bioinformatics,* **3**, 33–46.

[15] 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.

[16] Duarte, C. M., Wadley, L. M., and Pyle, A. M. (2003) RNA Structure Comparison, Motif Search and Discovery Using a Reduced Representation of RNA Conformational Space. *Nucleic Acids Research,* **31**, 4755–4761.

[17] Wadley, L. M., Keating, K. S., Duarte, C. M., and Pyle, A. M. (2007) Evaluating and Learning from RNA Pseudotorsional Space: Quantitative Validation of a Reduced Representation for RNA Structure. *Journal of Molecular Biology,* **372**, 942–957.

[18] Olson, W. K. (1980) Configurational Statistics of Polynucleotide Chains. An Updated Virtual Bond Model to Treat Effects of Base Stacking. *Macromolecules,* **13**, 721–728.

[19] Westhof, E. and Fritsch, V. (2000) RNA folding: beyond Watson-Crick pairs. *Structure,* **8**, R55–R65.

[20] 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.

[21] Leontis, N. B., Lescoute, A., and Westhof, E. (2006) The Building Blocks and Motifs of RNA Architecture. *Current Opinion in Structural Biology,* **16**, 279–287.

[22] Restrepo, G., Mesa, H., Llanos, E. J., and Villaveces, J. L. (2004) Topological Study of the Periodic System. *Journal of Chemical Information and Computer Science,* **44**, 68–75.

[23] Restrepo, G., Llanos, E. J., and Meza, H. (2006) Topological Space of the Chemical Elements and its Properties. *Journal of Mathematical Chemistry,* **39**, 401–416.

[24] Leontis, N. B. and Westhof, E. (1998) Conserved Geometrical Base-Pairing Patterns in RNA. *Quarterly Reviews of Biophysics,* **31**, 399–455.

[25] Lu, X.-J. and Olson, W. (2003) 3DNA: A Software Package for the Analysis, Rebuilding and Visualization of the Three-Dimensional Nucleic Acid Structures. *Nucleic Acids Research,* **31**, 5108–5121.

[26] Arnott, S., Hukins, D. W. L., Dover, S. D., Fuller, W., and Hodgson, A. R. (1973) Structures of Synthetic Polynucleotides in the A-RNA and A′-RNA Conformations: X-ray Diffraction Analyses of the Molecular Conformations of Polyadenylic Acid · Polyuridylic Acid and Polyinosinic Acid · Polycytidylic acid. *Journal of Molecular Biology,* **81**, 107–122.

[27] Chandrasekar, R. and Arnott, S. Numerical Data and Functional Relatioships in Science and Technology pp. 31–170 Springer-Verlag (1989).

[28] Handl, J., Knowles, J., and Kell, D. B. (2005) Computational Cluster Validation in Post-Genomic Data Analysis. *Bioinformatics,* **21**, 3201–3212.

[29] Brock, G., Pihur, V., Datta, S., and Datta, S. (2008) clValid: An R Package for Cluster Validation. *Journal of Statistical Software,* **25**, 1–22.

[30] R Development Core Team R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing Vienna, Austria (2009) ISBN 3-900051-07-0.

[31] Arnott, S. Oxford Handbook of Nucleic Acid Structure pp. 1–38 Oxford Science Publications (1999).

[32] Olson, W. K. (1976) The Spatial Configuration of Ordered Polynucleotide Chains. I. Helix Formation and Base Stacking. *Biopolymers,* **15**, 859–878.

[33] Olson, W. K. (1981) Understanding the Motions of DNA. *Biomolecular Stereodynamics,* **1**, 327–343.

Chapter 3

RNA Base-Pairs

3.1 Canonical and Non-canonical Base-Pairs

As shown in Figure 1.2, there is a variety of base-pairing patterns between the heterocyclic bases in nucleic acids due to the many possible hydrogen bonding interactions. The most prevalent hydrogen bonding pattern in nucleic acids is that of the canonical Watson-Crick (WC) base-pair. Patterns other than the WC pairs are known as non-canonical base pairs and are more common in RNA than in DNA. We used the 3DNA [1] software package to find all base pairs in a non-redundant dataset of RNA crystal structures with resolution better than 3.5 Å downloaded from the Protein Data Bank (PDB). We also constrained our search to helical regions, which are defined as stretches of three or more spatially consecutive base-pairs which need not be covalently bonded by the sugar-phosphate backbone between sequentially consecutive base pairs [2].

RNA Type	Number of Structures	G	C	A	U
small helices	78	36	30	16	18
drug-RNA	36	36	33	14	17
protein-RNA	207	37	32	16	16
protein-tRNA	9	34	30	19	17
rRNA	13	37	28	18	17
tRNA	13	34	27	21	19
ribozyme	113	34	29	20	16
Total	469	36	30	18	16

Table 3.1: Description of the types of RNA structures and the base content of each group in the non-redundant dataset of RNA crystal structures with resolution better than 3.5 Å given as a percentage of the total number of structures per group. The listed bases comprise the base pairs in helices of three or more base-pairs. Further details of the structures which compose the dataset, including PDB_ID's and NDB_ID's can be obtained online as supplementary material attached to our recent paper [2].

Our dataset is non-redundant in the sense that from 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*. Table 3.1 shows in detail the number of bases for each RNA type in our dataset of helical structures. It is interesting to see that in general the content of G and C, is higher than that of A and U. The difference might be related to a higher overall stability of G·C base-pairs compared to A·U base-pairs.

In Table 3.2 we show the number of base pairs of different chemical types formed by unmodified nucleotides in our dataset. It is clear from the table that G·C and A·U base pairs dominate the RNA base pairs formed in helical regions, making up 80% of all the base pairs. If we count only those that form canonical WC base pairs (9500 G·C, and 3069 A·U), the number corresponds to 73% of all base pairs in helical regions.

As will be show later (Table 3.3), a considerable number of the A·U pairs associate in a Hoogsteen arrangement. A few of these examples form U·A·U triplets containing a WC and a Hoogsteen [i] base pair in the RNA helical regions.

A	G	C	U	B/B′
384	980	313	3975	A
	128	9913	1282	G
		63	103	C
			187	U

Table 3.2: Composition of base pairs in the non-redundant structural dataset. Note that 9500 out of 9913 G·C and 3069 out of 3975 A·U are canonical WC base pairs. See legend to Table 3.1 for dataset details.

3.1.1 RNA Base-pair Classification

We classified the RNA base pairs in our dataset using three criteria: (1) the Leontis-Westhof edge classification scheme [3], which is based on the identities of the three major interacting edges for hydrogen-bond formation called the WC (W), Hoogsteen (H), and Sugar (S) edges, (2) the rotational and translational rigid-body base-pairing parameters called, shear, stretch, stagger, buckle, propeller and opening, and (3) the location of the base pairs within the helices, that is, their location in either "intact" covalently bonded sugar-phosphate backbones or within "quasi-continuous" helices with breaks in the sugar-phosphate backbone and their positions within these kinds of helices (see below).

We find that $\sim$90% of the base pairs in the RNA helices in our dataset form base pairs in one of seven possible hydrogen-bonding types: canonical WC G·C and A·U pairs, and non-canonical G·U

[i] Hoogsteen base-pairs are illustrated by structure XXIII of the Saenger classification of base-pairs as shown in Figure 1.2

wobble, sheared G·A, Hoogsteen A·U, WC type G·A, and U·U wobble base pairs. Detailed results showing how these seven major RNA base-pairing types are classified according to various schemes, and the details of their hydrogen-bond distances are given in Table 3.3 and in Figure 3.1

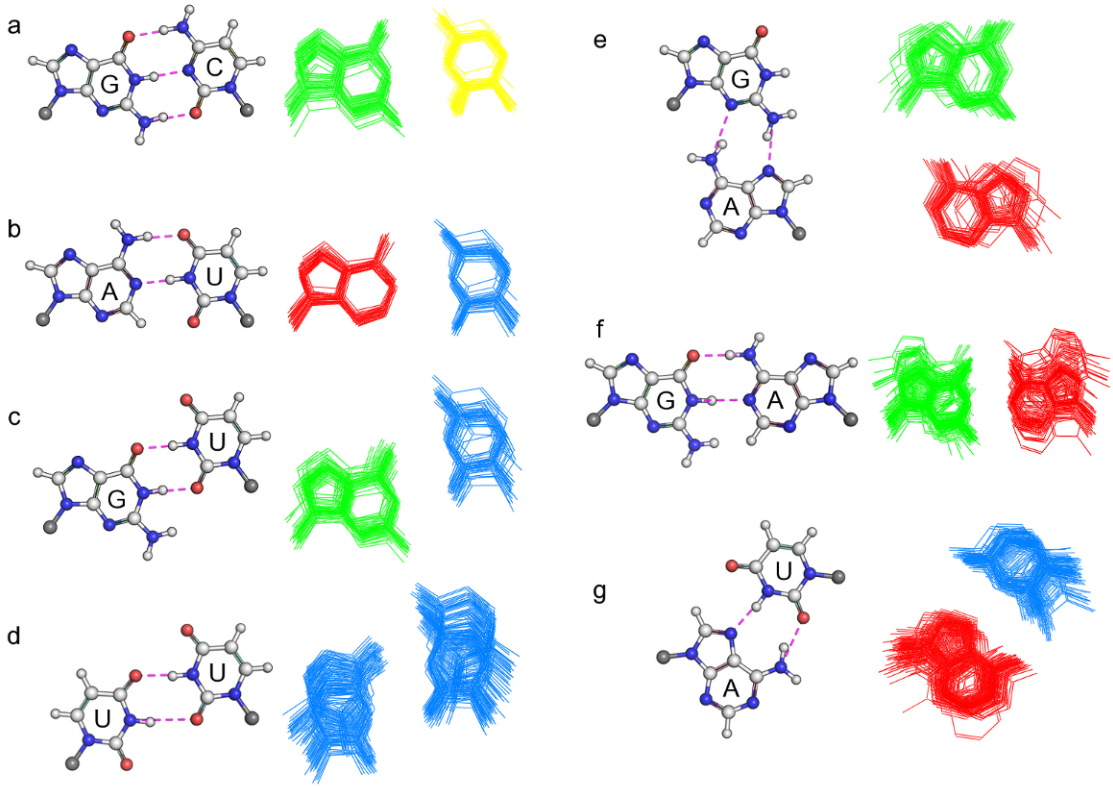


Figure 3.1: Seven most prominent base pairs in RNA helical regions in our structural dataset shown in images (a-g), (a-b) the canonical G·C and A·U Watson-Crick pairs, (c) the wobble G·U pair, (d) the wobble U·U pair, (e) the sheared G·A pair, (f) the Watson-Crick-like G·A pair, and (g) the Hoogsteen A·U pair. The images on the left for each base-pair show the identities of the bases, atom types (oxygen red, nitrogen blue, carbon and hydrogen white, and C1′ atoms gray), and the hydrogen-bond connectivity (magenta colored dashed lines). The rigth side images of each base-pair representation show a superposition of the base pairs in our helical dataset, centered in the middle base triad (MBT) reference frame (The definition of a middle base triad is completely analogous to that of a middle step triad as explained thoroughly in Appendix A).

3.1.2 Base Pairs in Helical Regions

Our classification also includes the locations of the base pairs in helical regions, that is, whether they are in the interior or at the ends of “intact” or “quasi-continuous” helical regions. Figure 3.2 illustrates the two types of helical regions mentioned. The “intact” helical region is depicted in Figure 3.2a. The “quasi-continuous” helical region is shown in Figure 3.2b.

Base-pair		Hydrogen bonds		Sign	Saenger	Leontis-Westhof		Number
Canonical								
G·C	Watson-Crick	N2-H···O2 O6···H-N4 N1-H···N3	2.79(0.17) 2.92(0.18) 2.89(0.13)	-	XIX	cis	W/W	$9500_{x0.90}$
A·U	Watson-Crick	N1···H-N3 N6-H···O4	2.84(0.14) 2.97(0.18)	-	XX	cis	W/W	$3069_{x0.93}$
Non-canonical								
G·U	Wobble	N1-H···O2 O6···H-N3	2.79(0.16) 2.85(0.16)	-	XXVIII	cis	W/W	$1049_{x0.69}$
G·A	Sheared	N2-H···N7 N3···H-N6	2.89(0.17) 3.03(0.18)	+	XI	trans	H/S	$509_{x0.59}$
A·U	Hoogsteen	N6-H···O2 N7···H-N3	2.91(0.21) 2.90(0.17)	+	XXIII	trans	H/W	$354_{x0.71}$
G·A	Watson-Crick	N1-H···N1 O6···H-N6	2.84(0.17) 2.91(0.20)	-	VIII	cis	W/W	$185_{x0.85}$
U·U	Wobble	O2···H-N3 N3···H-O4	2.95(0.24) 2.87(0.15)	-	XVI	cis	W/W	$141_{x0.54}$

Table 3.3: Seven dominant base-pairing types found in RNA helical regions. The first column lists the Gutell and collaborators nomenclature [4] of the base pairs. Column two shows the standard hydrogen bonding pattern and the average hydrogen-bond distances, in Ångstrom units, and standard deviations (in parentheses) associated with each base pair. Column three displays a negative sign if the bases forming the base pair oppose each other and a positive sign if they share the same face [1]. Column four gives the Saenger classification as in Figure 1.2. Column five lists the Leontis-Westhof edge classification obtained through the rnaview program [5], and the last column gives the total number of identified base pairs of each category and the percentage of those which comply exactly with the hydrogen-bonding pattern shown in column two.

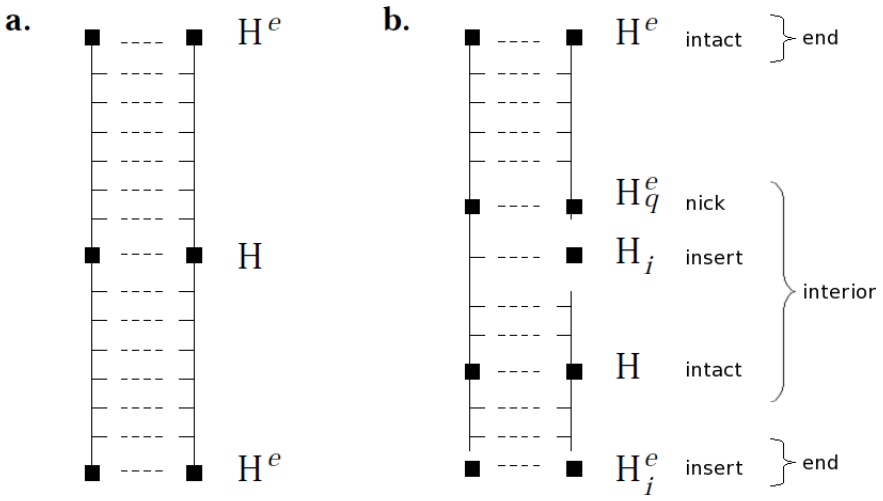


Figure 3.2: Schematic of (a) intact and (b) quasi-continuous helical regions in RNA. Base-pair locations within helical regions are denoted by: intact helix interior (H), intact helix end (H^e), nicked helix end (H^e_q), insert at helix interior (H_i), insert at helix end (H^e_i). Image kindly provided by Dr. Yurong Xin.

The locations of the base pairs in helical regions are summarized in Table 3.4. The A·U Hoogsteen pair ($A{\cdot}U_H$) stands out as being present mainly as an insert [ii] in helical regions, sometimes in nicked regions and rarely in intact ones. The G·A Watson-Crick-like pair ($G{\cdot}A_{WC}$) differs from the A·U Hoogsteen pair, in that, it rarely occurs as an insert, preferring to be in nicks, and sometimes in intact regions. A similar situation happens with the sheared G·A pair ($G{\cdot}A_s$), which also rarely occurs as an insert and is mainly found in intact regions, sometimes in nicks, and often at the end of intact helical stretches. The canonical WC G·C base-pair is two times more likely to occur at the end of helical regions than the canonical A·U base-pair. The helical context of the G·U wobble pair ($G{\cdot}U_w$) is quite similar to that of the canonical base-pairs, with location more closely resembling the context of $G{\cdot}C_{WC}$, than that of $A{\cdot}U_{WC}$. The G·U wobble stands out from the WC base pairs in being slightly more prevalent in nicked regions. The $G{\cdot}U_w$ pair is similar to the canonical A·U pair in that is more commonly seen in interior regions than in ends. Because of the large number of G·C WC base pairs its helical context is practically identical to the one given for all base-pairs.

The mean length of the helical domains in our dataset is 11 base pairs as shown in the histogram of helix length frequencies in Figure 3.3. The value of 11 base pairs coincides with the number of residues per turn in a canonical A-RNA fiber [6] which suggests that the canonical A-RNA conformation is predominant, and is most likely maintained by the canonical WC pairs which constitute 73% of all base pairs. In accordance with Leontis and collaborators [7], we see that helices are usually short and

[ii] Similar to the way intercalating ligands (intercalators) insert themselves in DNA.

Helical context	Base-pair							
	All	G·C_{WC}	A·U_{WC}	G·U_w	G·A_s	A·U_H	G·A_{WC}	U·U_w
Interior								
Intact	0.62	0.62	0.75	0.63	0.34	0.05	0.25	0.74
Nick	0.20	0.20	0.16	0.26	0.25	0.29	0.66	0.13
Insert	0.02	0.01	0.01	0.00	0.00	0.42	0.01	0.03
Ends								
Intact	0.13	0.15	0.07	0.10	0.33	0.05	0.06	0.10
Insert	0.02	0.01	0.01	0.01	0.05	0.19	0.02	0.00

Table 3.4: Distribution of helical location of the seven most abundant base-pairs in our RNA helical regions dataset. The helical context is shown in a secondary structure representation in Figure 3.2.

most are less than 15 residues. On the other hand, we note that the total number of helices longer than 15 residues is not negligible.

3.2 Deformability of Base-Pairs

Figure 3.1 gives a visual representation of the deformability of the seven most predominant base pairs in RNA helical regions. The overlapping structures for each base-pair are centered in the middle base triad (see Appendix for A details). In Table 3.5 we have collected the averages and their corresponding standard deviations for the rigid-body parameters of the base pairs, along with a deformability score, which is extracted from the covariance of the rigid-body parameters. Specifically, V [8] is given by the product of the square roots of the eigenvalues of the covariance matrix of base-step parameters. Thus, volume scores correspond to the accessible conformational volume. Table 3.5 also lists the root-mean-square deviation (RMSD) of the structures which have been superimposed using the middle base triad of the base pairs. The RMSD reported is the mean value of the RMSD's for all base pairs in the predominant groups. The latter RMSD values are computed from the Cartesian coordinates of all atoms in a base pair and the average Cartesian coordinates of the atoms in the specified group [9].

As seen from Table 3.5, the non-canonical base pairs are clearly more deformable than the canonical WC base pairs in terms of their volume scores and RMSD values. This larger deformability comes mainly from the in-plane parameters, that is, Shear, Stretch, and Opening, which determine the hydrogen-bonding patterns. The out-of-plane parameters Stagger, Buckle, and Propeller are roughly comparable to those of the less deformable WC base-pairs. That is, there is more deformability in the non-canonical pairs at the level of the hydrogen-bonding controlling parameters, and less at the level of ones controlling planarity. Another fact that confirms this greater variability is the smaller fraction

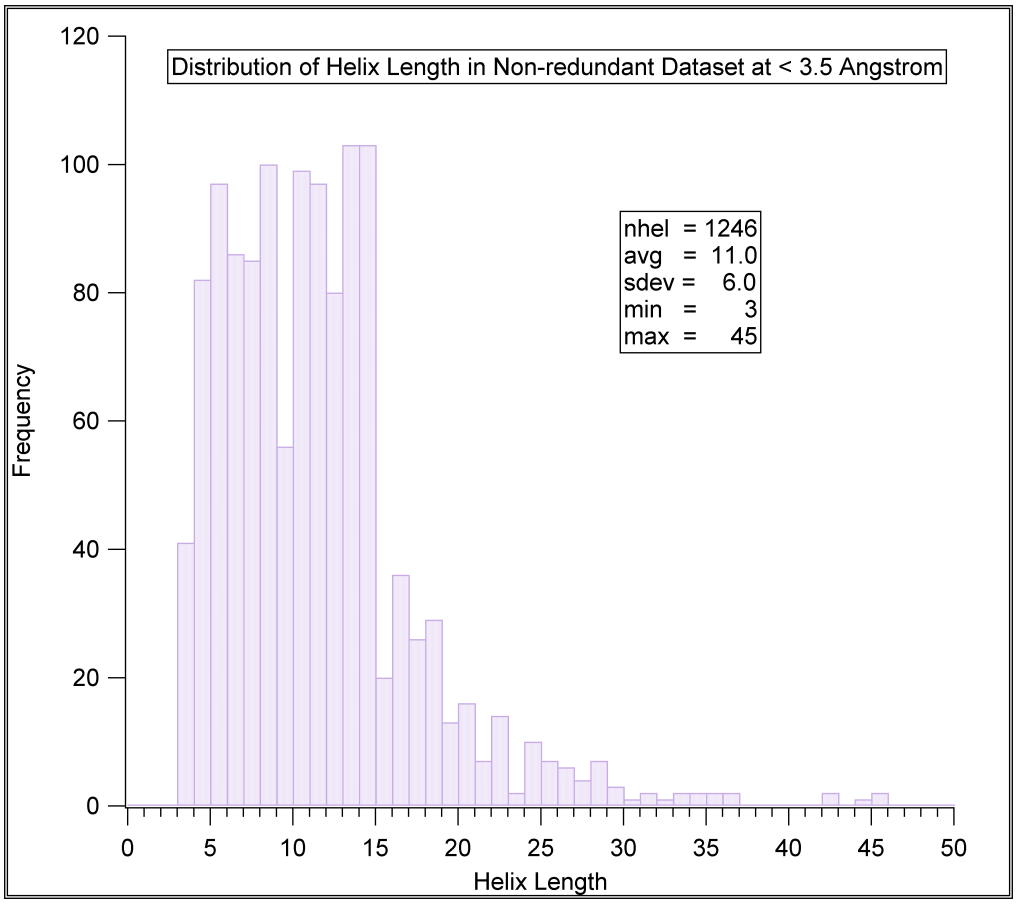


Figure 3.3: Histogram showing the distribution of helical regions in our non-redundant helical dataset composed of structures whose resolution is better than 3.5 Å. The total number of helices (nhel), the average helix length (avg), the standard deviation (sdev), and the minimum (min) and maximum (max) helix lenghts are given in the legend. Note that the helices need not be intact in the sense that the sugar-phosphate backbone might have nicks, but base-pairs are nonetheless stacked sequentially forming quasi-continuous helices.

of base-pairs which comply strictly to the standard hydrogen-bonding pattern in the non-canonical vs. canonical base pairs, as seen in the last column of Table 3.3. For example, whereas 90 percent of the canonical WC G·C pairs comply to the binding pattern shown in column three, only 54 percent of the U·U wobble base pairs associate strictly with the standard O2$\cdots$H-N3 and N3- H$\cdots$O4 interacting pattern.

What occurs with the additional non-canonical base pairs, which are not included in the hydrogen-bond distance averages, is that they are "melting". That is, the WC pairs, as defined here, do not include such melted states, which form fewer hydrogen bonds, or different types of hydrogen-mediated heavy atom bridges. The combination of angular and translational parameters, with respective units in degrees and Angstroms, complicates the analysis of the conformational volume. Nonetheless, the general trends give a clear indication of three main levels of deformation. That is, canonical G·C and A·U base pairs with a mean score of 4.5 are not very deformed compared to the G·U_w, G·A_s, and A·U_H pairs with mean values of around 25, and the remaining U·U_w and G·A_{WC} are more deformable than any other pairs. A more straightforward interpretation of spatial deformability comes from the root-mean-square deviation (RMSD) values of the base-pair groups. These values are computed by reconstructing all-atom base-pair structures using their base-pair parameters. The RMSD values are then obtained from a superposition of base pairs aligned with respect to their standard base-pair reference frames, followed by computation of the deviation for the atoms in each bp with respect to those in the mean structure using the vmd software package [9]. The deformability of base-pairs, as ranked by their RMSD values – U·U_w (0.74 Å) > G·A_{WC} (0.54 Å) $\approx$ G·A_s (0.50 Å) > G·U_w (0.45 Å) $\approx$ A·U_H (0.43 Å) > G·C_{WC} (0.38 Å) $\approx$ A·U_{WC} (0.36 Å) – does not follow exactly the same trend as that of the conformational volume score. The RMSD values, nonetheless, roughly maintain the ordering of the three levels of deformability, although from this perspective the G·A_{WC} base pair may be better grouped with the intermediate deformability non-canonical steps. An interesting observation is the prominence of the base pair with the greatest deformability, the U·U_w base-pair, which is more prominent in the interior of intact helical regions, compared to the location of less deformable non-canonical pairs, which show a preference for nicked or inserted regions in RNA helical regions.

Base-pair	Rigid-body parameters							
	Shear (Å)	Stretch (Å)	Stagger (Å)	Buckle (deg)	Propeller (deg)	Opening (deg)	V (deg^3 $Å^3$)	rmsd
G·C_{WC}	-0.20 (0.41)	-0.15 (0.17)	-0.04 (0.40)	-3.4 (8.4)	-8.7 (8.5)	0.5 (4.5)	6.5	0.38
A·U_{WC}	0.04 (0.34)	-0.14 (0.15)	0.04 (0.39)	-0.3 (8.4)	-9.0 (8.7)	0.9 (5.2)	6.4	0.36
G·U_w	-2.11 (0.84)	-0.52 (0.27)	-0.04 (0.43)	-0.1 (8.2)	-7.4 (7.5)	-0.6 (6.9)	25.1	0.45
G·A_s	6.78 (0.23)	-4.40 (0.55)	0.14 (0.52)	1.5 (11.0)	-3.2 (9.1)	-5.3 (8.2)	27.2	0.50
A·U_H	-4.06 (0.80)	-1.92 (0.84)	0.07 (0.61)	-0.4 (7.3)	1.0 (10.8)	-95.1 (17.4)	25.9	0.43
G·A_{WC}	-2.34 (0.59)	-1.63 (0.31)	-0.09 (0.47)	0.6 (8.8)	-11.1 (7.8)	-0.2 (17.4)	86.2	0.54
U·U_w	0.00 (0.64)	1.52 (0.40)	-0.29 (0.41)	7.8 (10.8)	-10.8 (9.6)	-17.2 (13.5)	49.8	0.74

Table 3.5: Mean values and standard deviations of the six rigid-body base-pair parameters, the conformational accessible volumes scores of each pair [8], and the root-mean-square deviations (RMSD) of base pairs superimposed in the middle base triad (MBT) (see Appendix A) in RNA helical regions.

References

[1] Lu, X.-J. and Olson, W. (2003) 3DNA: A Software Package for the Analysis, Rebuilding and Visualization of the Three-Dimensional Nucleic Acid Structures. *Nucleic Acids Research,* **31**, 5108–5121.

[2] 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.

[3] Leontis, N. B. and Westhof, E. (1998) Conserved Geometrical Base-Pairing Patterns in RNA. *Quarterly Reviews of Biophysics,* **31**, 399–455.

[4] Lee, J. C. and Gutell, R. R. (2004) Diversity of Base-pair Conformations and their Occurrence in rRNA Structure and RNA Structural Motifs. *Journal of Molecular Biology,* **344**, 1225–1249.

[5] Yang, H., Jossinet, F., Leontis, N., Chen, L., Westbrook, J., Berman, H., and Westhof, E. (2003) Tools for the Automatic Identification and Classification of RNA base pairs. *Nucleic Acids Research,* **31**, 3450–3460.

[6] Arnott, S., W.Fuller, Hodgson, A., and Prutton, I. (1968) Molecular Conformations and Structure Transitions of RNA Complementary Helices and their Possible Biological Significance. *Nature,* **220**, 561–564.

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

[8] Olson, W. K., Gorin, A. A., Lu, X.-J., Hock, L. M., and Zhurkin, V. B. (1998) DNA Sequence-Dependent Deformability Deduced from Protein-DNA Crystal Complexes. *Proceedings of the National Academy of Sciences,* **95**, 11163–11168.

[9] Eargle, J., Wrigth, D., and Luthey-Schulten, Z. (2006) Multiple Alignment of Protein Structures and Sequences for VMD. *Bioinformatics,* **22**, 504–506.

Chapter 4

RNA Base-Pair Steps

Before the turn of this century it was still not conceivable that knowledge-based potentials could be obtained for RNA helical regions due to the small amount of crystallographic data available. This was not the case for DNA, where enough data for such potentials have been available since 1998 as shown by Olson and collaborators [1]. As pointed out in Chapter 2, the number of high-resolution X-ray crystal structures of RNA has increased by two orders of magnitude, giving us enough information to develop a dimeric model of double-helical RNA with 10 unique base-pair steps formed by the canonical G·C and A·U Watson-Crick pairs. An extended model with 21 unique dimeric steps can also be constructed by adding the wobble G·U base-pair to canonical G·C, and A·U, but GU·GU (11 cases) and UA·UG (20 cases) dimeric data are still scarce. An illustration of the possible unique dimeric steps which can be formed in RNA is given in Figure 4.1.

For the case of the 91 possible unique base-pair steps, which can be formed from the seven dominant base-pairing types discussed in Chapter 3, we see tendencies of favored sequences inferred from counts of the available data as will be shown in the next section.

The results obtained from the analysis of RNA knowledge-based dimeric information allow us to explore double-helical RNA at the global level, that is, as a polymer chain. Therefore we can compute the persistence length of some RNA sequences as shown in the last section of this chapter.

4.1 Base-Pair-Steps in Intact Helical Regions

From the dataset of base-pairs described in Chapter 2 we also collected base-pair step information and focused our attention on the dimers which are located in intact helical regions, those noted as H in Figure 3.2. From the data shown in Table 4.1 we see that it is more common for a non-canonical base-pair to abut a canonical base-pair than a non-canonical one. The number of steps formed by a non-canonical base pair and a canonical one is six times greater than that of steps formed by two non-canonical pairs. A few non-canonical base pairs occur in the context of a stack of non-canonicals, e.g.,

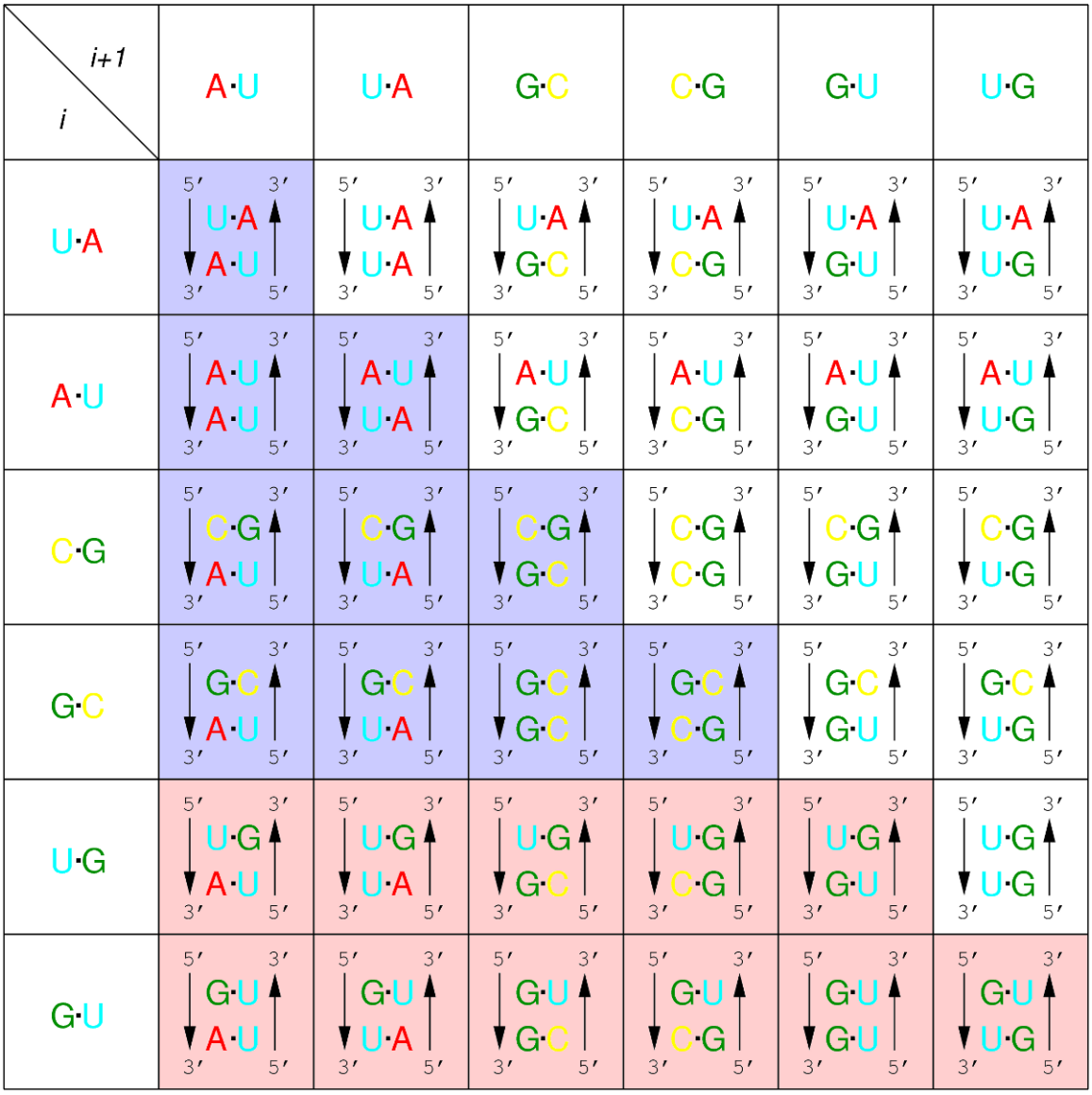


Figure 4.1: Color-coded representation of the 21 unique base-pair dimers of RNA formed by canonical G·C, and A·U Watson-Crick and wobble G·U base pairs. The grey boxes include the 10 unique base-pair steps formed by canonical G·C and A·U, and the pink boxes the additional 11 base-pair steps which result from considering G·U wobble base pairs as dimer building blocks. The coloring scheme used to denote the bases is that used in the NDB, where A is red, U is cyan, G is green, and C is yellow. The first base-pair X in each XpY step in the 5′ to 3′ sense is identified as pair i and the second Y is identified by $i + 1$ in the upper left corner of the figure. The white boxes include the base-pair steps with the equivalent sequence as those in the colored boxes in the mirror location in the grid but with the identities of the strands switched.

the 13 $GA_s{\cdot}GA_s$ and the 20 $GA_s{\cdot}A_sG$ steps, where two sheared base pairs stack together, and the 32 $AG_s{\cdot}GU_w$ where a sheared $G{\cdot}A_s$ base pair stacks next to a Hoogsteen $A{\cdot}U_H$ base-pair, the 42 $GU_w{\cdot}GU_w$ steps, the 31 $UG_w{\cdot}GU_w$ and the 11 $GU_w{\cdot}U_wG$. The majority of stacks between non-canonical base pairs occurs on dimeric steps composed of combinations of G·U wobble and sheared G·A base pairs. Overall the majority of dimeric steps ((604+608+1335)/6755 = 37.7%) are those formed by canonical G·C base pairs making up more than a third of the whole. Furthermore, GG·CC steps (1335/(604+608+1335) = 52.4%) form more than half of these kind of steps.

Overlap values between base pairs are given in parentheses along with the counts of dimeric steps in intact helical regions in Table 4.1. Examination of these values shows that the common trends seen for overlaps of DNA base pairs persist in RNA, i.e., purine-pyrimidine (RY) steps have the greatest overlap values, and pyrimidine-purine (YR) the smallest. When the G·U wobble base pair is taken into consideration the trend remains true, but the values are considerably increased. For example, for the $GU_w{\cdot}UG_w$ dimer the overlap value is 14.4 Å^2 compared to 11.3 Å^2 for $GC_{WC}{\cdot}CG_{WC}$. Other dimers which show a large overlap include those composed of sheared G·A pairs and sheared G·A and U·U wobble pairs. The degree of overlap doesn't necesarily correlate with greater dimer stabilities as judged from the populations of observed dimers. For example, there are more cases of wobble U·U pairs adjacent to canonical G·C pairs than to canonical A·U pairs despite the smaller overlap area.

4.2 RNA Base-pair-steps Database and Web Framework

With the step parameters from the intact helical regions we can construct a harmonic energy function model. That is, we can treat succesive base-pairs as rigid blocks connected by a spring and described by a harmonic potential function $\Psi(x)$:

$$\Psi(x) = \frac{1}{2}\sum_{i,j} F_{i,j}\Delta x_i \Delta x_j \tag{4.1}$$

$$\Delta x_i = x_i - x_i^0 \tag{4.2}$$

Here the x_i represent the spatial arrangements of the base-pair steps in terms of the six rigid-body base-pair-step parameters, that is, the displacements in Shift, Slide, and Rise, and the rotations in Tilt, Roll, and Twist. It has been shown by Go and Go [3] that the correlation of spatial variables can be obtained from the force constants $F_{i,j}$. That is, they show that these correlations are proportional to kT

$C{\cdot}G_{WC}$	$G{\cdot}C_{WC}$	$U{\cdot}A_{WC}$	$A{\cdot}U_{WC}$	$U{\cdot}G_{w}$	$G{\cdot}U_{w}$	$A{\cdot}G_{s}$	$G{\cdot}A_{s}$	$U{\cdot}A_{H}$	$A{\cdot}U_{H}$	$U{\cdot}U_{w}$	$A{\cdot}G_{WC}$	$G{\cdot}A_{WC}$	bp_i/bp_{i+1}
$604_{(4.5)}$	$1335_{(4.1)}$	$747_{(3.0)}$	$574_{(2.9)}$	$77_{(5.0)}$	$192_{(5.3)}$	$66_{(6.9)}$	–	–	–	$18_{(2.9)}$	$4_{(4.8)}$	$5_{(4.2)}$	$G{\cdot}C_{WC}$
	$608_{(11.3)}$	$511_{(4.1)}$	$572_{(9.9)}$	$161_{(2.9)}$	$252_{(13.5)}$	$33_{(10.0)}$	–	–	–	$69_{(6.1)}$	$20_{(8.8)}$	$5_{(7.7)}$	$C{\cdot}G_{WC}$
		$97_{(2.0)}$	$249_{(3.3)}$	$20_{(2.8)}$	$45_{(6.1)}$	$7_{(7.7)}$	–	–	–	$2_{(3.6)}$	–	$3_{(4.3)}$	$A{\cdot}U_{WC}$
			$126_{(8.4)}$	$48_{(2.0)}$	$79_{(11.8)}$	$6_{(9.5)}$	–	–	–	$20_{(8.2)}$	$1_{(6.4)}$	$14_{(5.3)}$	$U{\cdot}A_{WC}$
				$31_{(5.3)}$	$42_{(4.1)}$	$32_{(8.3)}$	$1_{(5.7)}$	–	–	$5_{(2.9)}$	–	–	$G{\cdot}U_{w}$
					$11_{(14.4)}$	–	–	–	–	–	$4_{(8.8)}$	$7_{(9.3)}$	$U{\cdot}G_{w}$
						–	$13_{(10.6)}$	–	–	$6_{(11.1)}$	–	–	$G{\cdot}A_{s}$
							$20_{(9.1)}$	$7_{(3.0)}$	$2_{(3.7)}$	–	–	–	$A{\cdot}G_{s}$
								–	–	–	–	–	$A{\cdot}U_{H}$
									–	–	–	–	$U{\cdot}A_{H}$
										$3_{(1.7)}$	–	–	$U{\cdot}U_{w}$
											–	$1_{(6.3)}$	$G{\cdot}A_{WC}$
												–	$A{\cdot}G_{WC}$

Table 4.1: Counts of unique base-pair steps and overlap areas in Ångstrom^2 (subscripted values in parenthesis) in intact RNA helical regions of RNA structures. The overlap values are those of shared areas between base pairs in a base-pair-step, these areas are defined by the base ring atoms [2]. For details on the dataset composition see Chapter 3 and Table 3.1

with a proportionality coefficient equal to the i, j elements of the inverse matrix of second derivatives with respect to energy, i.e., the F matrix,

$$\langle x_i x_j \rangle = kT(F_n^{-1})_{i,j}, \tag{4.3}$$

where $\langle x_i x_j \rangle$ stands for the correlation between coordinates, k is Boltzman's constant, and T is temperature in Kelvin, an analog to the "classical" treatment for atoms by Wilson et al. [4]. By contrast we take and inverse approach and extract the $F_{i,j}$ from the observed correlation of variables, using the average value of base-pair-step parameters rather than atomic coordinates.

The total potential for a system of N base-pair steps would then be given by:

$$U(x_i) = \sum_{n=1}^{N} \Psi_n. \tag{4.4}$$

In order to obtain quasi-Gaussian distributions of the base-pair step parameters and also to exclude extreme conformational deformation in steps, we culled our dataset by restricting it to include the step parameters within 3 standard deviations from the mean values. These data were then used to obtain the force constants by finding the inverse correlation matrix.

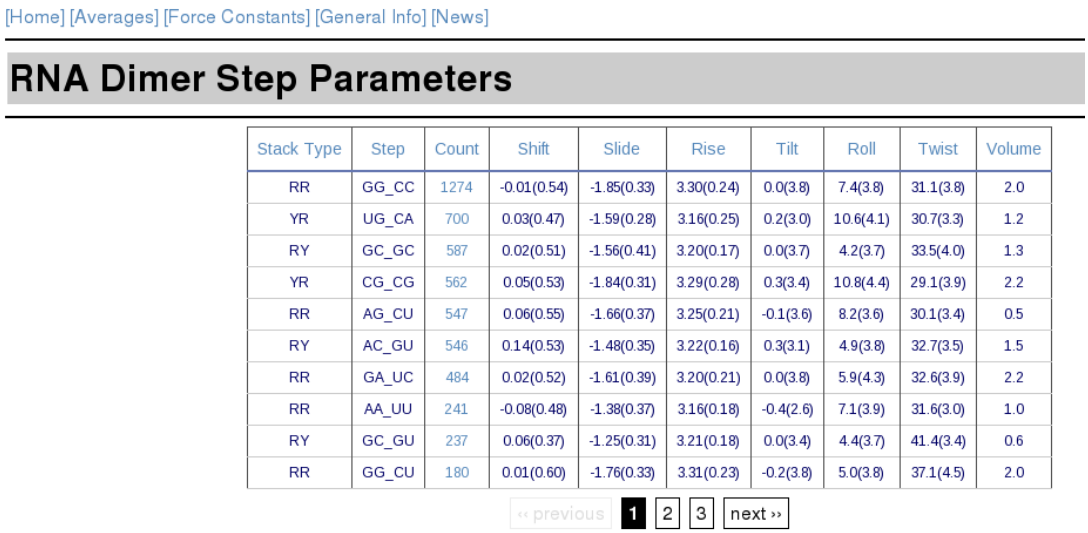
A minimal MySQL database was created to store the mean values and covariance of base-pair-step parameters, their standard deviations, the derived force-constant matrices, and step-deformation measures such as the conformational volume and the RMSD for all atoms other than hydrogen from the average structure. The purpose of creating the database is for ease of sharing information with force-field developers and future automatization of the process of data reduction.

The process of data reduction can be summarized as follows:

- Collection of a non-redundant dataset of RNA X-ray crystallographic structures restricted to resolution better than 3.5 Å , with up-to-date data from the PDB.
- Computation of base-pair and base-pair-step parameters for all structures using 3DNA [2].
- Determination of the Leontis-Westhof classification of base pairs using rnaview [5].
- Construction of a relational database to link various classifications of RNA structure, such as the Leontis-Westhof classification of base pairs, to the helical classification given by 3DNA.
- Selection of steps present in intact helical regions alone.

- Culling unique base-pair step parameters using a three-standard-deviation cutoff.
- Computation of the means, standard deviations, force-constant matrices, and other step-deformation measures.

The database and its web framework can be found at http://rnasteps.rutgers.edu/. The site includes a table of the 21 unique dimers formed by canonical Watson-Crick G·C, A·U, and wobble G·U, shown in Figure 4.2.


[Home] [Averages] [Force Constants] [General Info] [News]

RNA Dimer Step Parameters

Stack Type	Step	Count	Shift	Slide	Rise	Tilt	Roll	Twist	Volume
RR	GG_CC	1274	-0.01(0.54)	-1.85(0.33)	3.30(0.24)	0.0(3.8)	7.4(3.8)	31.1(3.8)	2.0
YR	UG_CA	700	0.03(0.47)	-1.59(0.28)	3.16(0.25)	0.2(3.0)	10.6(4.1)	30.7(3.3)	1.2
RY	GC_GC	587	0.02(0.51)	-1.56(0.41)	3.20(0.17)	0.0(3.7)	4.2(3.7)	33.5(4.0)	1.3
YR	CG_CG	562	0.05(0.53)	-1.84(0.31)	3.29(0.28)	0.3(3.4)	10.8(4.4)	29.1(3.9)	2.2
RR	AG_CU	547	0.06(0.55)	-1.66(0.37)	3.25(0.21)	-0.1(3.6)	8.2(3.6)	30.1(3.4)	0.5
RY	AC_GU	546	0.14(0.53)	-1.48(0.35)	3.22(0.16)	0.3(3.1)	4.9(3.8)	32.7(3.5)	1.5
RR	GA_UC	484	0.02(0.52)	-1.61(0.39)	3.20(0.21)	0.0(3.8)	5.9(4.3)	32.6(3.9)	2.2
RR	AA_UU	241	-0.08(0.48)	-1.38(0.37)	3.16(0.18)	-0.4(2.6)	7.1(3.9)	31.6(3.0)	1.0
RY	GC_GU	237	0.06(0.37)	-1.25(0.31)	3.21(0.18)	0.0(3.4)	4.4(3.7)	41.4(3.4)	0.6
RR	GG_CU	180	0.01(0.60)	-1.76(0.33)	3.31(0.23)	-0.2(3.8)	5.0(3.8)	37.1(4.5)	2.0

‹‹ previous 1 2 3 next ››

Figure 4.2: Snapshot of the unique base-pair-step parameters table for intact helical regions of RNA, showing the fields by which the data can be sorted. In this case the data are sorted by dimer step counts. There are three stack types purine-purine (RR), pyrimidine-purine (YR), and purine-pyrimidine (RY). The steps are denoted in a 5′ to 3′ sense, e.g., GG_CC stands for a G·C base-pair step covalently linked between the G's in the leading strand, i.e., GpG, and the C's in the complementary strand, i.e., CpC.

The values in the table can be sorted by any of the included fields (Stack Type, Step Count, Shift, Slide, Rise, Tilt, Roll, Twist, Volume, and RMSD). Also the scatterplot corresponding to each step is displayed when the users click on the counts column, along with a potential energy contour in the Roll-Twist plane at 4.5 kT (i.e. 3 standard deviations). A snapshot of the roll-twist energy contour for the GG·CC step from the web-framework is shown in Figure 4.3 Note the lack of correlation between the variables compared to the same steps in in DNA [1, 6] and the slightly greater variation in roll compared to twist.

The other values included in the web-framework are the force-constants corresponding to the unique steps. A snapshot of the force-constant matrix for the GG·CC step is shown in Figure 4.4

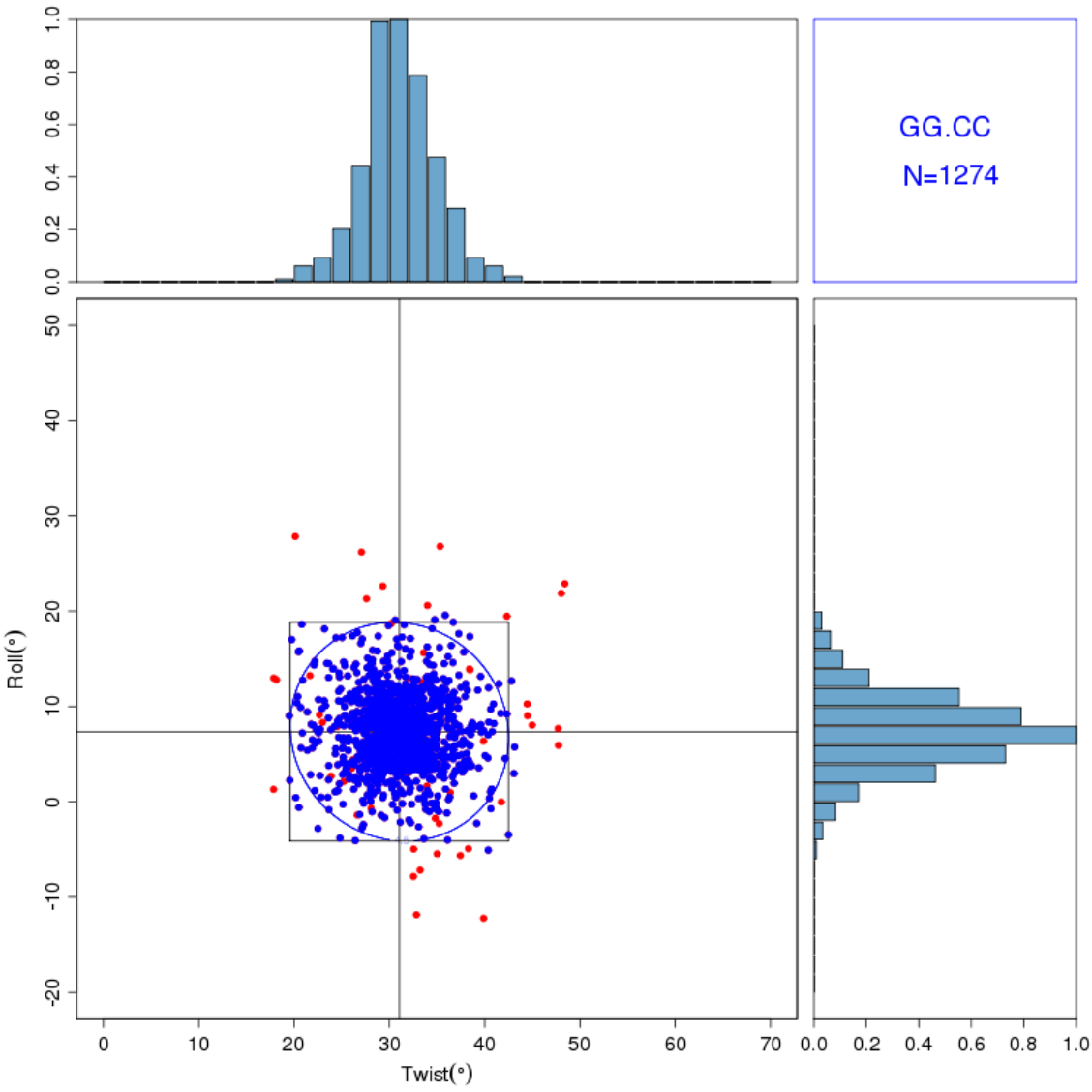


Figure 4.3: Snapshot from the web-framework of a scatterplot in the Roll-Twist plane with an energy contour at 4.5kT. The full data before culling (1335 steps) are depicted by red dots, and the culled data (1274 steps), within 3 standard deviations of the mean, by blue dots.

Parameter	Step	shift	slide	rise	tilt	roll	twist
shift	GG_CC	3.82	-0.02	-0.06	-0.16	-0.01	0.03
slide	GG_CC	-0.02	10.52	1.15	0.02	-0.06	-0.32
rise	GG_CC	-0.06	1.15	18.77	-0.01	-0.16	-0.18
tilt	GG_CC	-0.16	0.02	-0.01	0.08	0.00	0.00
roll	GG_CC	-0.01	-0.06	-0.16	0.00	0.07	0.01
twist	GG_CC	0.03	-0.32	-0.18	0.00	0.01	0.08

‹‹ previous 1 2 3 4 ... 8 9 10 11 **12** 13 14 15 ... 18 19 20 21 next ››

Figure 4.4: Snapshot of the RNA base-pair-steps web-framework where the force constant matrix for the GG·CC dimeric step is shown. The force constant matrix is derived from the covariance of step parameter values following Go and Go [3] and Olson et al. [1].

4.3 Persistence Length of RNA

A quantity commonly used to quantify the stiffness of polymers is the so-called persistence length a. To determine this quantity for DNA or RNA, a variety of theoretical and experimental techniques can be used. Some common experimental techniques to determine a include electron microscopy (EM), gel electrophoresis, sedimentation velocities, electrical birefringence, atomic force microscopy (AFM) , magnetic tweezers, and small angle X-Ray scattering (SAXS). For reviews of such techniques applied to the determination of RNA persistence length, we refer the reader to Hagerman [7], Abels et al. [8], and Caliskan et al. [9]. We compare our simulated results, based on the observed structural properties of RNA and the "realistic" model developed by Olson and collaborators [10, 11] to describe DNA, with the persistence length extracted from various experimental means.

Initial studies started with selected data for the deformabilities of the ten unique Watson-Crick base-pair steps [12]. A more complete picture applied to the study of DNA sequence-dependent deformability became available in 1998 [1]. The base-pair-step deformability data for DNA have been constantly refined as more high-resolution DNA and DNA-protein structures have been added to the Nucleic Acid Database (NDB) [13]. Although such data have been available for DNA since 1998, such was not the case for RNA, until recently [14].

A brief description of the “realistic” model along with a simplified schema of the C++ code developed by Dr. Luke Czapla and modified slightly by the author, is given in Appendix C. This Appendix also includes a brief account of various definitions and models generally used to compute the persistence length in the literature.

Using the mean values and force constants of RNA base-pair steps available at our web framework we can use the “realistic” model, as implemented by Czapla et al [15], to compute the persistence length of RNA helical chains of increasing lengths formed from the ten unique base-pair steps of canonical Watson-Crick G·C and A·U pairs. We used two rest states in our calculations, one which corresponds to a naturally straight chain using the dimeric parameters for an idealized B-like RNA, that is, $x_i^0 = \{0, 0, 3.30, 0, 0, 32.7\}$ with a helical repeat of 11 base-pairs and the standard vertical displacement (rise) but with no intrinsic bending or shearing of succesive base pairs [16, 17]. The other rest states used are the average base-pair-step parameters of the ten unique steps obtained from our web-framework. It is important to note that one cannot have a chain composed of pure GC·GC steps, for example, since the step between two such steps is a CG·CG step. Therefore eight of the ten chains formed by the unique

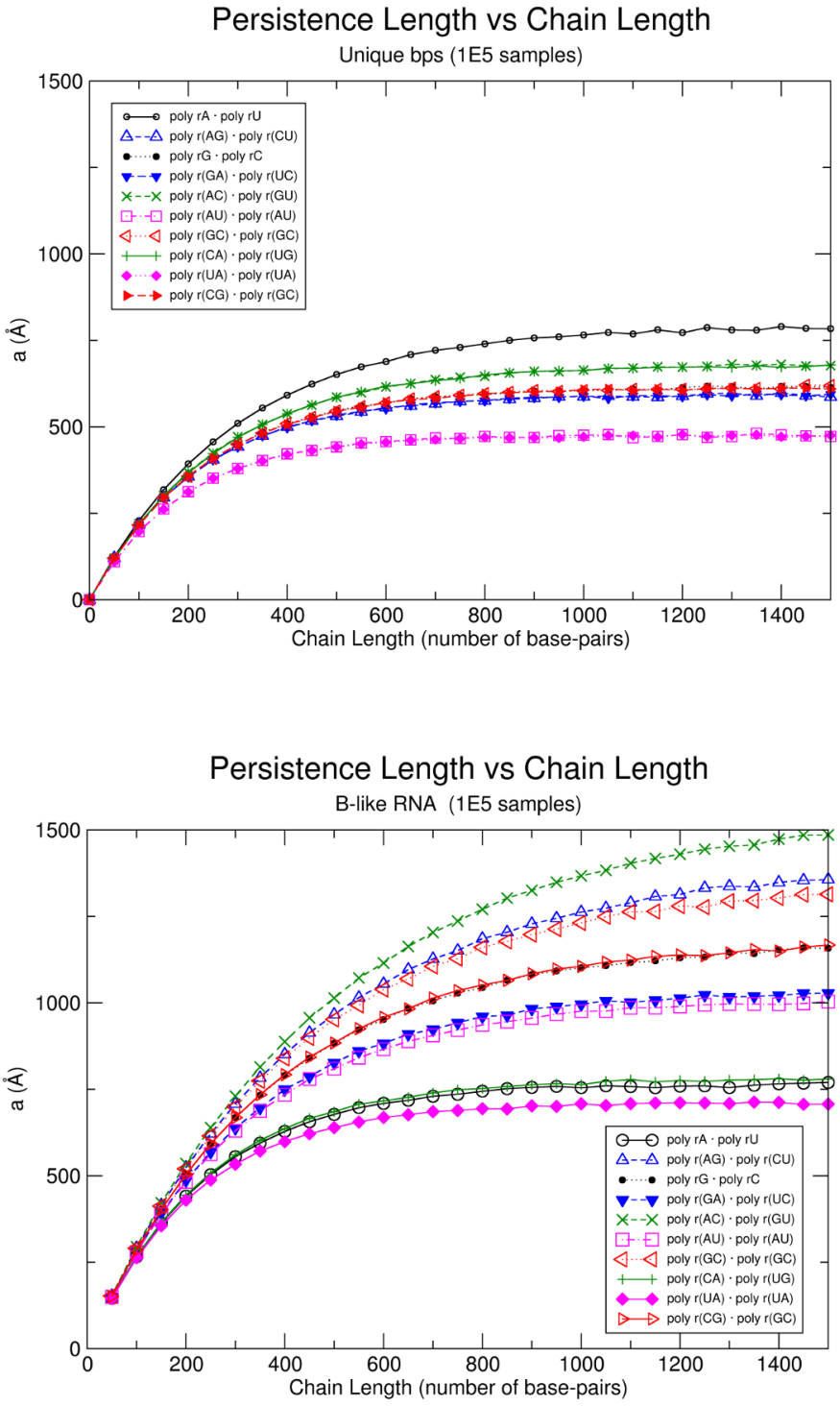


Figure 4.5: Persistence length vs. chain length in base pairs for naturally straigth B-like RNA chains with 11 base pairs per turn (lower frame), "real" chains (upper frame). Note that chains with alternating base pairs sequence are constructed from two types of dimers and that the computed values in a limited configurational sample (10^5 simulated chains) are nearly identical for poly rX· poly rY and poly rY· poly rX chains where X$\neq$Y.

base-pair steps will be mixed (alternating copolymers of XpY and YpX steps), and only two will contain a single kind of step. That is, those chains formed from repetition of the GG·CC or AA·UU dimers.

As expected, the persistence lengths for naturally straight chains are greater than the corresponding persistence lenghts for chains whose rest states come from the averaged crystallographic data (Figure 4.5). Repetition of the averaged crystal rest states yield regular helices with the helical repeats per turn and helical parameters shown in Table 4.2.

"UNREALISTIC" HOMOPOLYMERS									
Type	Step	X-disp (Å)	Y-disp (Å)	H-Rise (Å)	Incl. (°)	Tip (°)	H-Twist (°)	RPT	Pitch
RR	AA·UU	-3.63	0.08	2.79	12.8	0.7	32.4	11.1	31.03
	AG	-4.54	-0.13	2.71	15.4	0.2	31.2	11.6	31.30
	GG·CC	-4.63	0.02	2.80	13.6	0.0	32.0	11.3	31.55
	GA	-3.76	-0.03	2.87	10.4	0.0	33.1	10.9	31.21
RY	AC	-3.39	-0.20	2.97	8.6	-0.5	33.1	10.9	32.34
	AU	-3.58	-0.04	2.85	15.2	-0.7	33.3	10.8	30.81
	GC	-3.33	-0.03	2.99	7.3	0.0	33.8	10.7	31.89
YR	CA	-4.50	-0.02	2.49	19.3	-0.4	32.4	11.1	27.63
	UA	-4.05	0.09	2.52	20.5	0.2	34.2	10.5	26.57
	CG	-5.33	-0.04	2.47	20.6	-0.6	31.0	11.6	28.68
	ARNA	-4.04	0.07	2.81	15.5	0.8	32.7	11.0	30.91
	Blike	-2.72	0.15	3.30	0.0	0.0	31.6	11.4	37.63
HOMOPOLYMERS and COPOLYMERS									
Type	Step	X-disp (Å)	Y-disp (Å)	H-Rise (Å)	Incl. (°)	Tip (°)	H-Twist (°)	RPT	Pitch
RR.RR	AA·UU	-3.63	0.08	2.79	12.8	0.7	32.4	11.1	31.03
	GG·CC	-4.63	0.02	2.80	13.6	0.0	32.0	11.3	31.55
	AG·CU	-4.18	-0.09	2.79	13.2	0.1	32.1	11.2	31.33
RY.YR	AC·GU	-3.90	-0.12	2.75	13.5	-0.5	32.8	11.0	30.20
	AU·AU	-3.79	0.02	2.70	17.6	-0.3	33.7	10.7	28.85
	GC·GC	-4.24	-0.04	2.75	13.3	-0.3	32.5	11.1	30.46
	ARNA	-4.04	0.07	2.81	15.5	0.8	32.7	11.0	30.91
	Blike	-2.72	0.15	3.30	0.0	0.0	31.6	11.4	37.63

Table 4.2: Helical parameter values for RNA homopolymers and alternating copolymers built using 3DNA from the average base-pair-step parameters values of averaged crystallographic data. For corresponding images see Figure 4.6

As is the case with DNA, the poly rA·poly rU chains show a smaller persistence length than the poly rG·poly rC chains, although the difference is not as dramatic as that shown by Maroun and Olson [11] for DNA. That is, the difference for DNA is about $\sim$500 Å, and for RNA the difference is about $\sim$100 Å. The stiffer poly rG·poly rC chains in this case reflect more symmetrical energy surface as is the case for DNA. This can easily be seen from the corresponding energy surfaces available at our web-framework where the energy contour for the GG·CC step is circular and the one for AA·UU base-pair step has the shape of an ellipse.

With 3DNA we can build a rigid-block-model representation of these RNA polymers using the average rest states from the averaged crystallographic data. Examples of these conformations can be seen in Figure 4.6. Note the variation in the central channel as a function of sequence, i.e., narrow for AA·UU

and wide for GG·CC.

For comparison with the sequence-dependent properties of DNA, we computed the persistence lengths of RNA sequences of 1000 base-pair or more steps made up of repeats of the unique base-pair steps, taking a million samples with the Gaussian sampler and assuming a naturally straight (B-like) chain. As can be seen in Table 4.3, the persistence length is generally larger than the obtained for DNA (using corresponding data and with T instead of U). The dependence reflects the greater stiffness of the A-RNA dimers compared to the more flexible B-DNA dimers, which sameple conformations that RNA cannot assume.

We also constructed a mixed sequence RNA homopolymer with a 1:1 ratio of A·U to G·C base-pairs, and 16 equally weighed base-pair steps. As described by Olson et al. [6], the dimers that make up such a chain are guided by a potential obtained by averaging the force constants of the 16 possible base-pair steps. The computed values in Table 4.3 are larger than their experimental counterparts, as is also the case for the DNA values. We therefore introduce a scaling factor $zeta$ (i.e., effective temperature) to scale the persistence lengths to the experimentally observed values (Table 4.3). For example, we can reproduce the values of Abels et al.[8] with a ζ scaling value of 0.75 (638 Å vs. 638 Å deduced from force-extension measurements [i]).

We also computed the persistence length, at increasing chain lengths, of a mixed sequence RNA block copolymer (poly r(AC)$_3$G$_5$· poly rC$_5$(GU)$_3$) as seen in Figure 4.7. We see that, in contrast to the homopolymers and block copolymers in Figure 4.5, the chain-length dependence of the mean extension of the block copolymer seems to be "damped". The same sort of variation in the persistence length occurs in DNA sequences [19]. A set of 100 sampled configurations of a 150 bp long block copolymers chain, obtained using the Gaussian sampling technique and the rebuild feature of 3DNA, is superimposed using the first base-pair-step reference frame of the chains in Figure 4.8 and can be viewed as a sequence of independent images at the web address http://rnasteps.rutgers.edu/rnadimer/media/img/movie.htm. The series of images reveal the tendency for the sequence to curve and therefore reduce the persistence length of RNA. The value of ∼300 Å for the block copolymer is much smaller than that for any of the chains shown in Figure 4.5.

[i] Note that the experimental values are based on an idealized Worm-Like chain model of DNA.

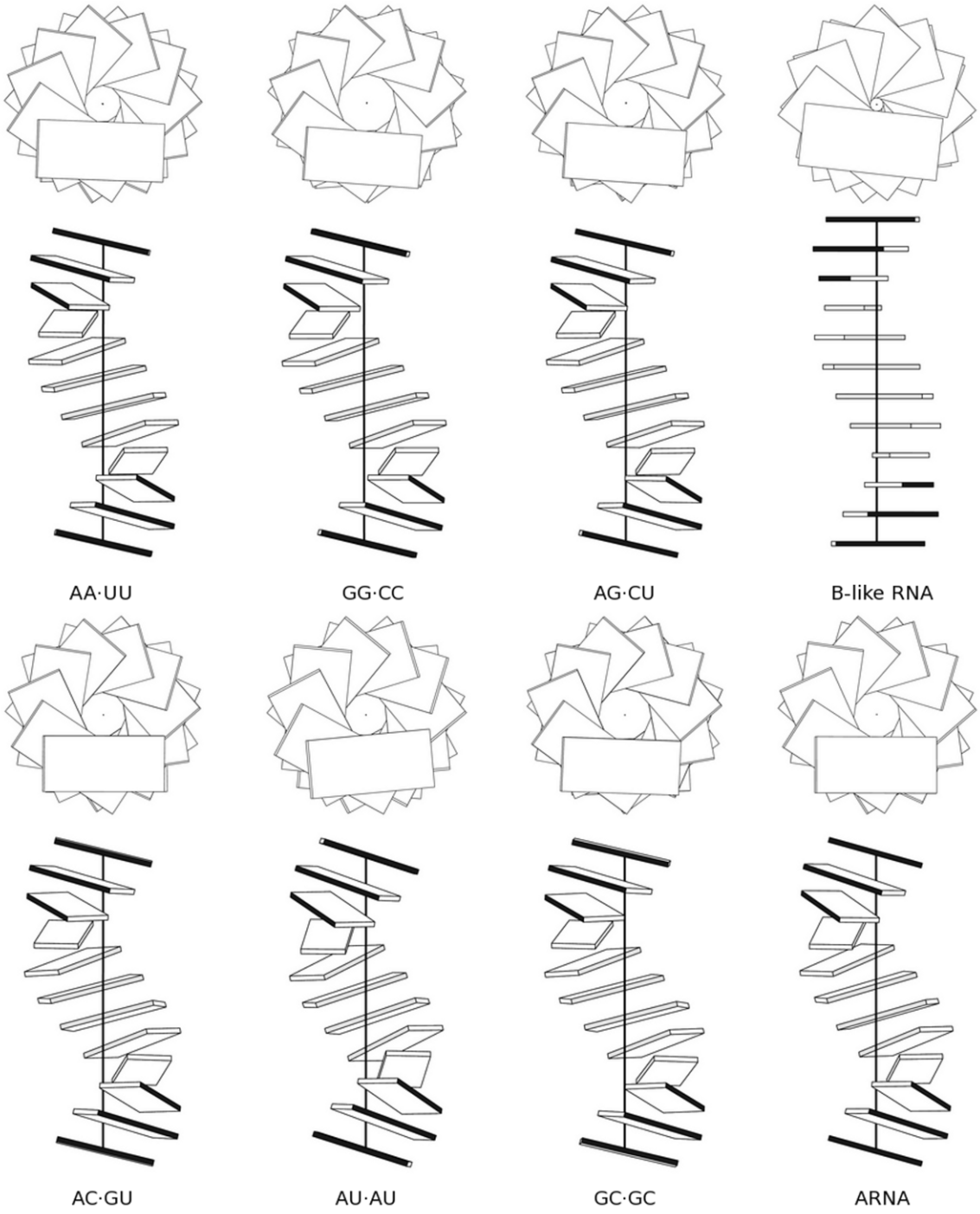


Figure 4.6: Calladine-Drew-like block representations [18] of homopolymers and representative alternating copolymers made from the rest states for the ten unique base-pair steps in RNA from averaged crystallographic data. The structures were built using 3DNA [2].

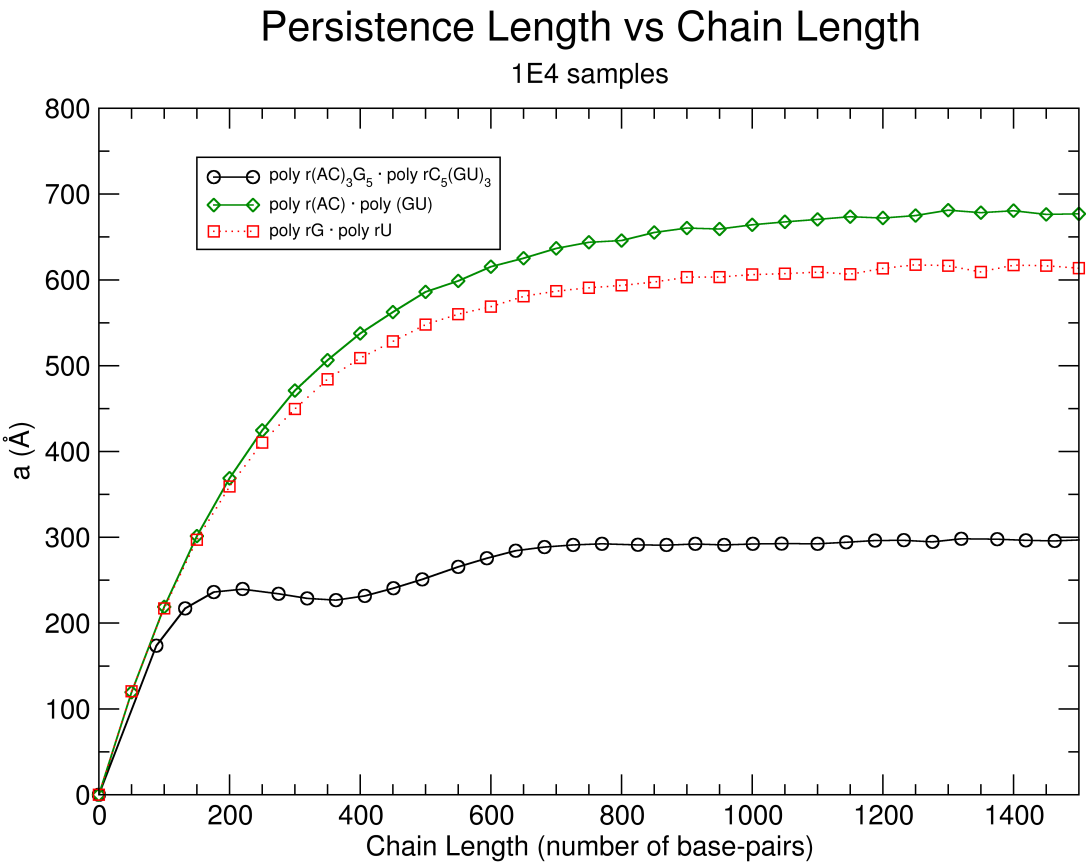


Figure 4.7: Chain-length dependence of the persistence length for represented RNA polymers; the block copolymer poly $r(AC)_3G_5$· poly $rC_5(GU)_3$, the alternating copolymer poly r(AC)·poly r(GU), and the homopolymer poly rG·poly rC. Configurational samples obtained using the Gaussian sampling technique of Czapla et al. [15]. The sinusoidal-like pattern for the block copolymer is indicative of a curved or super helical configuration in RNA induced by sequence.

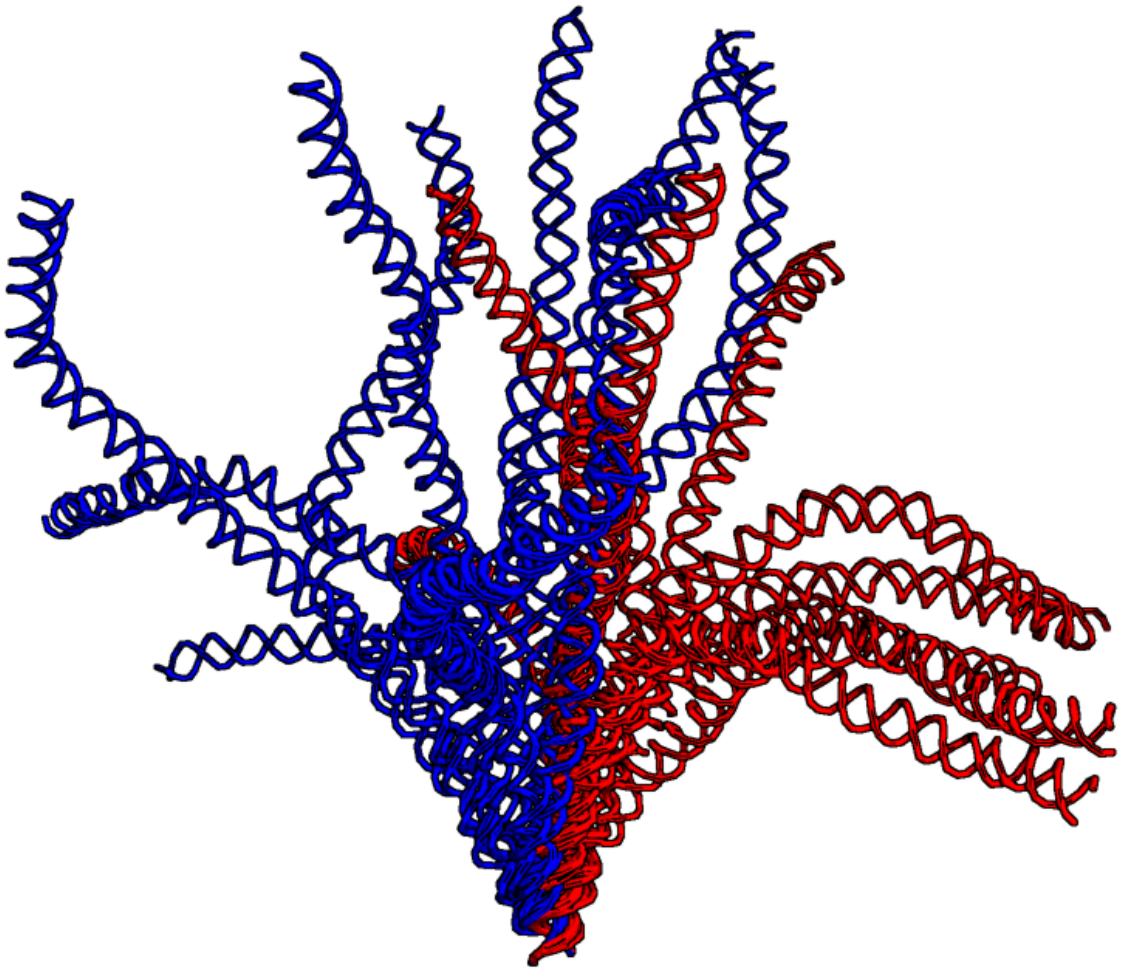


Figure 4.8: RNA chains of the 150-bp block copolymer poly r$(AC)_3G_5$· poly r$C_5(GU)_3$ superimposed in the reference frame of the first base step. Shown are 20 snapshots of the output produced by building the structures from the randomly sampled step-parameters obtained with the Gaussian sampling technique of Czapla et al. [15]. Colored in red are 10 snapshots of simulated "real" chains, and in blue 10 snapshots of intrinsically straigth B-like chains. Each chain is depicted by a ribbon linking sequential phosphorus atoms, reconstructed using 3DNA from the base-step parameters at each dimer step. Note the greater curved arrangements for the red "real" chains which bring their ends in close contact, in contrast to the stiffer B-like chains.

Stack Type	Base-pair Step	a (Å) RNA	a (Å) DNA
RR	AA·UU	765	395
	AG·CU	587	461
	GG·CC	606	405
	GA·UC	588	395
RY	AC·GU	664	625
	AU·AU	476	245
	GC·GC	604	454
YR	CA·UG	662	391
	UA·UA	471	217
	CG·CG	607	269
	Mixed Sequence	818	
	Hagerman	700-800	
	Abels et al. AFM	622	
	Abels et al. FE	638	
	Ideal DNA		500
	ζ (scaling factor)	0.75	0.50

Table 4.3: Persistence lengths for chains of 1000 base-pairs constructed from the ten unique base-pair steps. A scaling factor ζ (i.e., effective temperature) is used to reproduce the experimentally obtained values of persistence length of a random sequence with equal weights of G·C and A·U base-pairs, and with an equally weighed composition of the sixteen unique base-pair steps.

References

[1] Olson, W. K., Gorin, A. A., Lu, X.-J., Hock, L. M., and Zhurkin, V. B. (1998) DNA Sequence-Dependent Deformability Deduced from Protein-DNA Crystal Complexes. *Proceedings of the National Academy of Sciences,* **95**, 11163–11168.

[2] Lu, X.-J. and Olson, W. (2003) 3DNA: A Software Package for the Analysis, Rebuilding and Visualization of the Three-Dimensional Nucleic Acid Structures. *Nucleic Acids Research,* **31**, 5108–5121.

[3] Go, M. and Go, N. (1976) Fluctuations of an Alpha-Helix. *Biopolymers,* **15**, 1119–1127.

[4] Wilson, E. B., Decius, J. C., and Cross, P. C. (1955) Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra, McGraw-Hill, .

[5] Yang, H., Jossinet, F., Leontis, N., Chen, L., Westbrook, J., Berman, H., and Westhof, E. (2003) Tools for the Automatic Identification and Classification of RNA base pairs. *Nucleic Acids Research,* **31**, 3450–3460.

[6] Olson, W. K., Colasanti, A. W., Czapla, L., and Zheng, G. Insights into the Sequence-Dependent Macromolecular Properties of DNA from Base-Pair Level Modeling chapter 14, pp. 205–223 Taylor and Francis (2009).

[7] Hagerman, P. J. (1997) Flexibility of RNA. *Annual Review Biophysics Biomolecular Structure,* **26**, 139–156.

[8] Abels, J. A., Moreno-Herrero, F., van derHeijden, T., Dekker, C., and Dekker, N. H. (2005) Single-Molecule Measurements of the Persistence Length of Double-Stranded RNA. *Biophysical Journal,* **88**, 2737–2744.

[9] Caliskan, G., Hyeon, C., Perez-Salas, U., Briber, R. M., Woodson, S. A., and Thirumalai, D. (2005) Persistence Length Changes Dramatically as RNA Folds. *Phys Rev Lett,* **95**, 268303.

[10] Marky, N. L. and Olson, W. K. (1994) Configurational Statistics of the DNA Duplex: Extended Generator Matrices to Treat the Rotations and Translations of Adjacent Residues. *Biopolymers,* **34**, 109–120.

[11] Maroun, R. C. and Olson, W. K. (1988) Base Sequence Effects in Double-Helical DNA. II. Configurational Statistics of Rodlike Chains. *Biopolymers,* **27**, 561–584.

[12] Olson, W. K., Babcock, M. S., Gorin, A., Liu, G., Marky, N. L., Martino, J. A., Pedersen, S. C., Srinivasan, A. R., Tobias, I., and Westcott, T. P. (1995) Flexing and Folding Double Helical DNA. *Biophysical Chemistry,* **55**, 7–29.

[13] Balasubramanian, S., Xu, F., and Olson, W. K. (2009) DNA Sequence-Directed Organization of Chromatin: Structure-Based Computational Analysis of Nucleosome-Binding Sequences. *Biophysical Journal,* **96**, 2245–2260.

[14] 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.

[15] Czapla, L., Swigon, D., and Olson, W. K. (2006) Sequence-Dependent Effects in the Cyclization of Short DNA. *Journal of Chemical Theory and Computation,* **2**, 685–695.

[16] Chandrasekar, R. and Arnott, S. Numerical Data and Functional Relatioships in Science and Technology pp. 31–170 Springer-Verlag (1989).

[17] Arnott, S. Oxford Handbook of Nucleic Acid Structure pp. 1–38 Oxford Science Publications (1999).

[18] Calladine, C. R. and Drew, H. R. Understanding DNA: The Molecule and How It Works Academic Press (1997).

[19] Maroun, R. C. and Olson, W. K. (1988) Base Sequence Effects in Double-Helical DNA. III. Average Properties of Curved DNA. *Biopolymers,* **27**, 585–603.

Chapter 5

RNA Motifs

As mentioned in Chapter 2, until now the most common perspectives on RNA motif recognition and discovery have been chemical, based on the atoms and bonds. The rigid-body-based perspective involved in the interactions and the internal parameters that describe the 3D folds of RNA motifs, by contrast, has been rather unexplored. In this chapter we address two questions related to this line of investigation:

1. Can the geometric, rigid-block description of base pairing and base stacking help to define RNA structural motifs?

2. Can quantities derived from the 3DNA software package be used to perform an automated search for known RNA motifs, such as the GNRA tetraloop motif, and perhaps to find unknown RNA motifs?

We start with the second question, testing the characteristics of the base arrangements in the well known GNRA tetraloop motif. We have also examined other quantities (e.g. endocyclic and exocyclic base-overlaps) obtained with 3DNA [1, 2] to complement the automated search for GNRA tetraloop motifs.

5.1 The GNRA Tetraloop

The GNRA motif was initially found to be an important constituent of the small subunit of the ribosome from comparative sequence analyses [3]. That is, the GNRA sequence, where R refers to purine, was frequently repeated among various organisms in tetraloop regions of RNA, as were the CUUG and UNCG tetraloop sequences. These three sequences account for more than 70% of all tetraloops, i.e. single-stranded loops of four, non-Watson-Crick paired nucleotides, found in the 16S subunit of ribosomal RNA [3, 4].

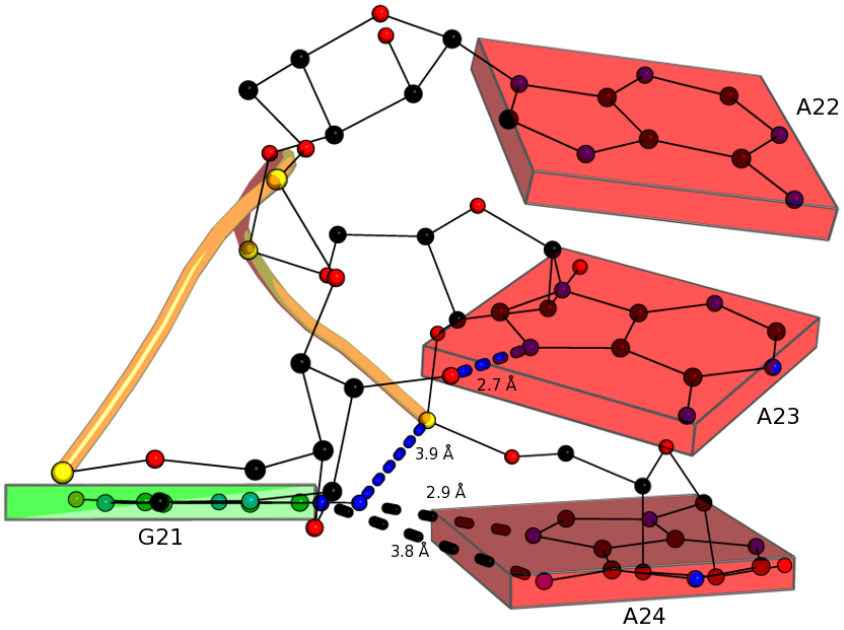


Figure 5.1: The GNRA tetraloop motif in the hammerhead ribozyme PDB_ID:1HMH [7]. Although not a newly recognized GNRA tetraloop, this motif is positively recognized using our rigid-body parameters RNA motif search program "getMotif". The structure was selected from a non-redundant list of RNA structures provided by the RNA ontology consortium (ROC) [8]. The hydrogen bonding interactions detected by Heus and Pardi [5] in NMR experiments are shown by black and blue dashed lines. The yellow tube connects the phosphorus atoms (yellow balls) along the sugar-phosphate backbone. Atoms are represented as balls and CPK colored, i.e., oxygen red, nitrogen blue, carbon black.

The most abundant of all the tetraloops in the ribosome is the GNRA motif, and its structural stability was initially characterized using NMR spectroscopy by Heus and Pardi [5]. They reported that the loop is closed by a non-canonical sheared G·A base-pair and further stabilized by (i) a hydrogen bond between the terminal guanine base and a phosphate, (ii) extensive base stacking, and (iii) a potential hydrogen bond between the hydroxyl group in the ribose at the 3′ end of the terminal guanine and the N7 of the penultimate purine (R) base [5]. These features, which were later confirmed in subsequent X-ray structures [6], can be seen in Figure 5.1.

The description of the GNRA tetraloop motif is a typical case of the problem of RNA motif definition. For example, in the context of sequence alone a GNRA motif would be one which contains, in a consecutive manner, the GNRA pattern of bases. There are GNRA structures, however, which are not formed by a consecutive sequence but rather have the same geometric arrangement of bases in three-dimensional space as the sequentially linked GNRA motif [9, 10]. There are also structures which have the same geometric arrangement and sugar-phosphate backbone connections as the GNRA tetraloop, but do not have the same sequence of bases as in the UCAA, UCAC, CAGA, and CAAC tetraloops

[10]. These other sequences which are geometrically equivalent to the GNRA tetraloop motif form non-canonical base-pairs which are isosteric to the sheared G·A base-pair. That is, the U·A, U·C, C·A, and C·C base-pairs closing the respective tetraloops are isosteric to the G·A base-pair at the end of the GNRA loop [10].

A number of molecular dynamics (MD) studies have explored the conformational space of the GNRA tetraloop [11, 12, 13, 4]. These studies find a set of conformational states for the GNRA tetraloop motif which are closely related to the X-Ray and NMR structures available through the Protein Data Bank [4, 12]. Other MD studies have used the well known GNRA motif conformation as the starting point for simulation of other tetraloops [14]. These other tetraloops do not retain a GNRA-like three-dimensional structure in the simulations [14]. This effect might reflect the fact that the force fields used in such calculations [11] do not correctly predict known tetraloops structures, such as the GAAA tetraloop [13].

5.1.1 GNRA Motif Search Program

We have devised a simple algorithm, “getMotif”, to search for GNRA motifs based on their base-step parameters. The algorithm is summarized in Figure 5.2.

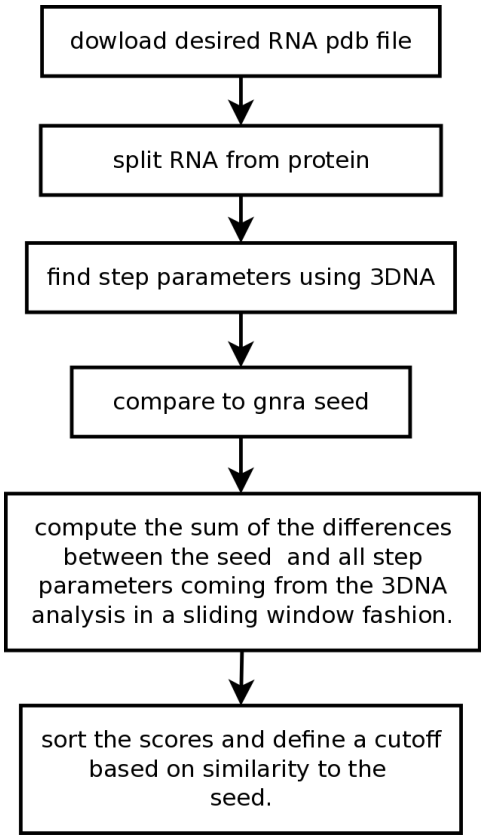


Figure 5.2: Simple algorithm for GNRA motif finding based on base-step parameters.

The algorithm allows for any motif seed to be integrated into it, but for now we limit ourselves to a seed formed by the step parameters of the GNRA motifs found by Lemierux et al. [10] in the 23S subunit of rRNA (PDB_ID:1FFK). We are currently constructing a database of the base-step parameters for known motifs, so that our simple program can be expanded to include various known motifs that a user wishes to localize in an arbitrary RNA structure or set of RNA structures.

The algorithm has been programed as a simple, yet very fast bash script [i] which interfaces with three components – one written in python, the second being the 3DNA package, and the last written in the statistical analysis software **R**. For example, for the large subunit of the ribosome, PDB_ID:1JJ2, the program takes 22.1 seconds to download and analyze the whole structure on an Intel Core Duo of 2.13GHz with 8Gb of RAM.

The program allows the user to query any PDB_ID. That is, the user only needs to input in the command line of a UNIX/LINUX terminal the command, "getMotif" followed by the PDB_ID of the RNA molecule of interest. As a result the user obtains a list composed of residue numbers corresponding to the location of the start of the motif in the structure, and a score which describes how close or far each four-nucleotide sequential structure is from the GNRA motif seed. The advantage of our program over other Windows-based motif recognition software, aside from providing a new analysis based on rigid-body parameters, is that it allows for easy integration of automated scripts for processing large lists of known PDB structures without user intervention for the analysis of every structure. For example, in the RNA ontology consortium (ROC) meeting of May 2009 a reduced dataset of RNA structures found at: https://docs.google.com/Doc?id=dhrmkfmn_13ftpbjcgq was made available to participants for the purpose of allowing them to search for RNA motifs which would later be compared among groups. Using Windows-based software like FR3D [15] it is quite difficult for the user to submit a large job composed of many PDB structures to a queing system, or a cluster computing server. Such a task is made simple using getMotif.

To construct the seed for the GNRA motif we extracted the first 20 GNRA tetraloop motifs found by Lemieux and Major [10] in the large subunit of the ribosome (PDB_ID:1FFK) as summarized in Figure 3 of their results. After computing the base-step parameters with 3DNA for each of these motifs, we arranged their average values in a three by six matrix S, where each row corresponds to one of the three steps in the tetraloop and each column to one of the step parameters. These average values are

[i] For ease of use future coding of the algorithm can be done fully in python without compromising the speed of structure analysis.

show in Table 5.1, along with the corresponsing step parameters of canonical A-RNA. The first step in the GNRA tetraloop, that is, GN, takes up the major distortion in step parameter values, whereas the NR and RA steps are more closely related to a canonical A-RNA like save for lesser sliding and overtwisting.

Step	Shift	Slide	Rise	Tilt	Roll	Twist
GN	-9.77	-1.90	-4.93	71.8	124.0	-57.5
NR	2.74	-0.11	3.04	11.5	6.0	50.6
RA	1.11	-0.20	3.01	9.5	6.1	42.4
A-RNA	0.00	-1.48	3.30	0.0	8.6	31.6

Table 5.1: GNRA motif seed composed of the average base step parameter values for 20 GNRA motifs found in the large subunit of the ribosome PDB_ID:1FFK [10].

Using the information in Table 5.1, we then compute a score to determine the distance between all structures formed by three successive steps in a given RNA structure, and the GNRA motif seed. The score is simply the sum of the elements of the difference matrix X_k, where the difference is between the base-step parameters in the GNRA seed matrix S and the corresponding parameters for three successive steps in each one of the 3 x 6 matrices W_k formed by the set of $k-3$ successive tetranucleotide steps in the RNA structure, where k is the total number of steps.

$$X_k = |S - W_k| \tag{5.1}$$

$$\text{score}_k = \sum_{i,j} \{x_{i,j}\} \tag{5.2}$$

The scores obtained from analysis of the large ribosomal subunit of the ribosome PDB_ID:1FFK can be seen in the histogram displayed in Figure 5.3. We see that the majority of trinucleotides (tristeps) have scores between 350-400. These values most likely correspond to tri-step sequences formed by base-step parameters similar to those of A-RNA-like steps. For our purposes of GNRA motif identification we want to select those steps with scores which reflect a shorter distance to our seed GNRA motif. Therefore we have used a cutoff score of 150, which is represented by the values below the red vertical line in Figure 5.3.

Application of our program to the list of 355 RNA structures provided by the ROC, finds [ii] 75 of these

[ii]The time GNRA tetraloop motifs candidates in the 355 structures were found in 7 minutes and 29 seconds using the program.

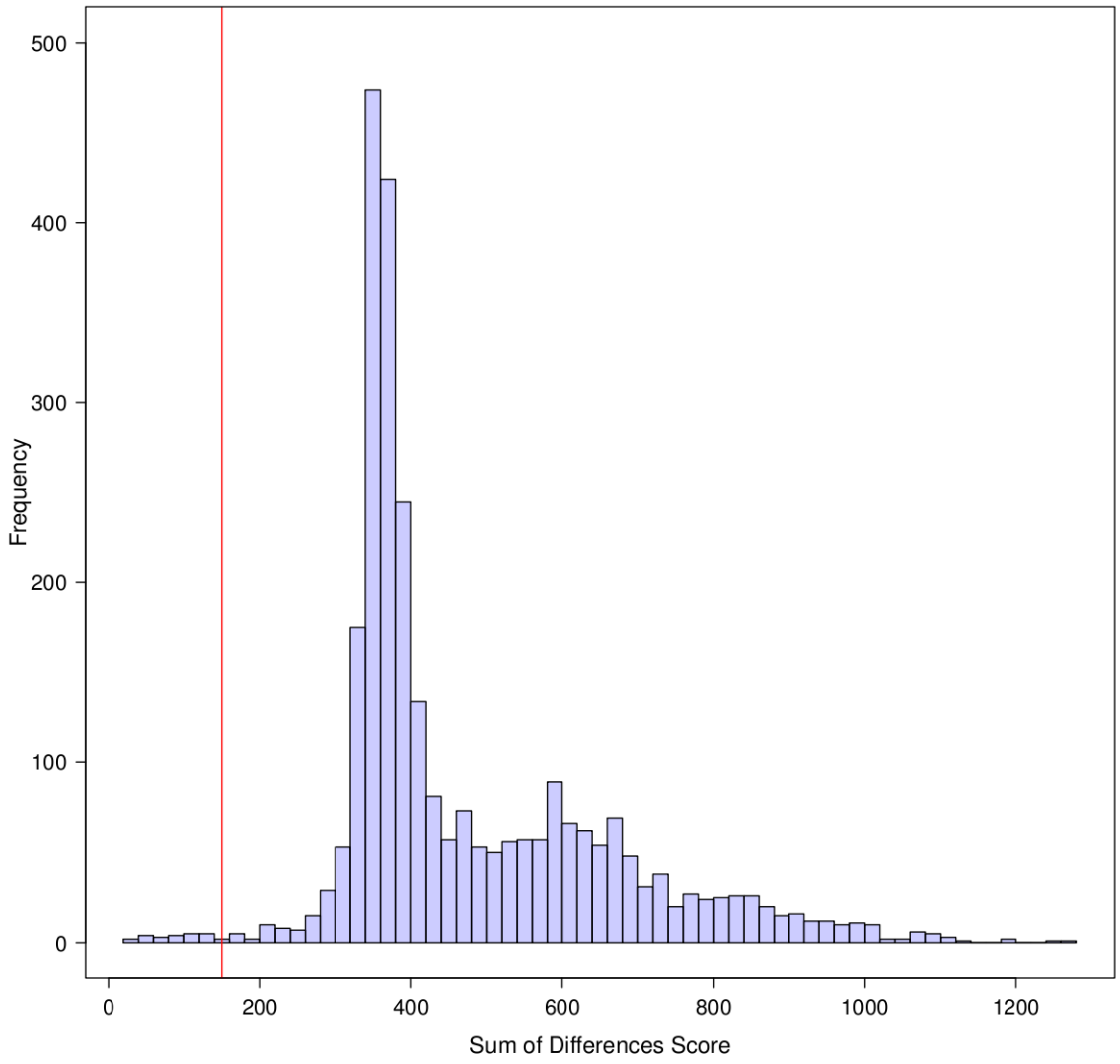


Figure 5.3: Histogram of the sum of differences score between the GNRA seed motif and all sequential tri-nucleotide-steps found in the large subunit of the ribosome PDB_ID:1FFK. A red vertical line is drawn at the cutoff value of 150 used to select the GNRA motif candidates.

structures, to have at least one GNRA 3D motif candidate with a cutoff value of 150, and a total of 211 GNRA motif candidates. The list of all identified motifs in these 75 structures including the parameters at the first step, are presented in Figure 5.4 and Table D.1, which includes an additional check for consistency in the next to last column. The latter entry is the overlap of the ring atoms of the first two bases of the tetraloop, which is zero in the described structures. Additionally the third to last column shows the overlap between bases when additional exocyclic atoms are considered, for example, the N4 amino atoms of cytosines [1]. These numbers are generally small although there are examples where the base overlap is substantial. We see that 67% of the identified GNRA motif candidates start with a GN step, and, interestingly, that almost one third (27%) of the first steps are UN steps, and that a few of the identified motifs start with AN and CN dimers.

To further analyze the results of the motifs identified using “getMotif”, we focused attention on the ribosomal structure (PDB_ID:2J01) that is included in the ROC list. As can be seen in Table D.1 for the 2J01 entries, the program recognizes twenty GNRA tetraloop motif candidates, all of which, in fact are, tetraloops. Of these 13 conform to the GNRA sequence, five start with UN, one starts with A, and one with C. When we align the candidate GNRA motif sequences using the reference frame of the first two bases, we see that they fall into two main groups (Figure 5.5), one with about eleven members and the other with seven members. The two remaining structures fit well with the first two steps, of both groups I and II, but do not fit well on the third step in either group. Group I, which is colored in blue in Figure 5.5 is closer to the common GNRA motif. Group II, which is colored in red folds close to the common GNRA motif but the residues which would form the terminal G·A base-pair are a little too far from each other to form a hydrogen bonding pattern. This variation stems from a large rise (vertical displacements) in the last step. The first two steps are very closely related in every case, suggesting that there is a subtle switch between the tetraloop and a pentaloop conformation, governed by the rise in the last step.

Interestingly, some of the GNRA candidates identified in Group II are part of a highly symmetrical extended "kissing" loop interaction. These can be seen in Figure 5.6 where they are superimposed, and shown in their original orientation in the ribosome. This type of interaction is part of what has been described as the lonepair triloop motif found by sequence covariation techniques [9]. The GNRA motif candidates found by “getMotif” have been named recently by Nasalean et al. [17] as internal and composite motifs. In our case the GNRA motif candidates in Group I correspond to the internal motifs, and the Group II candidates correspond to the composite motifs of Nasalean and collaborators.

As noted above, of the three steps which make up the GNRA tetraloop, the first one is the step

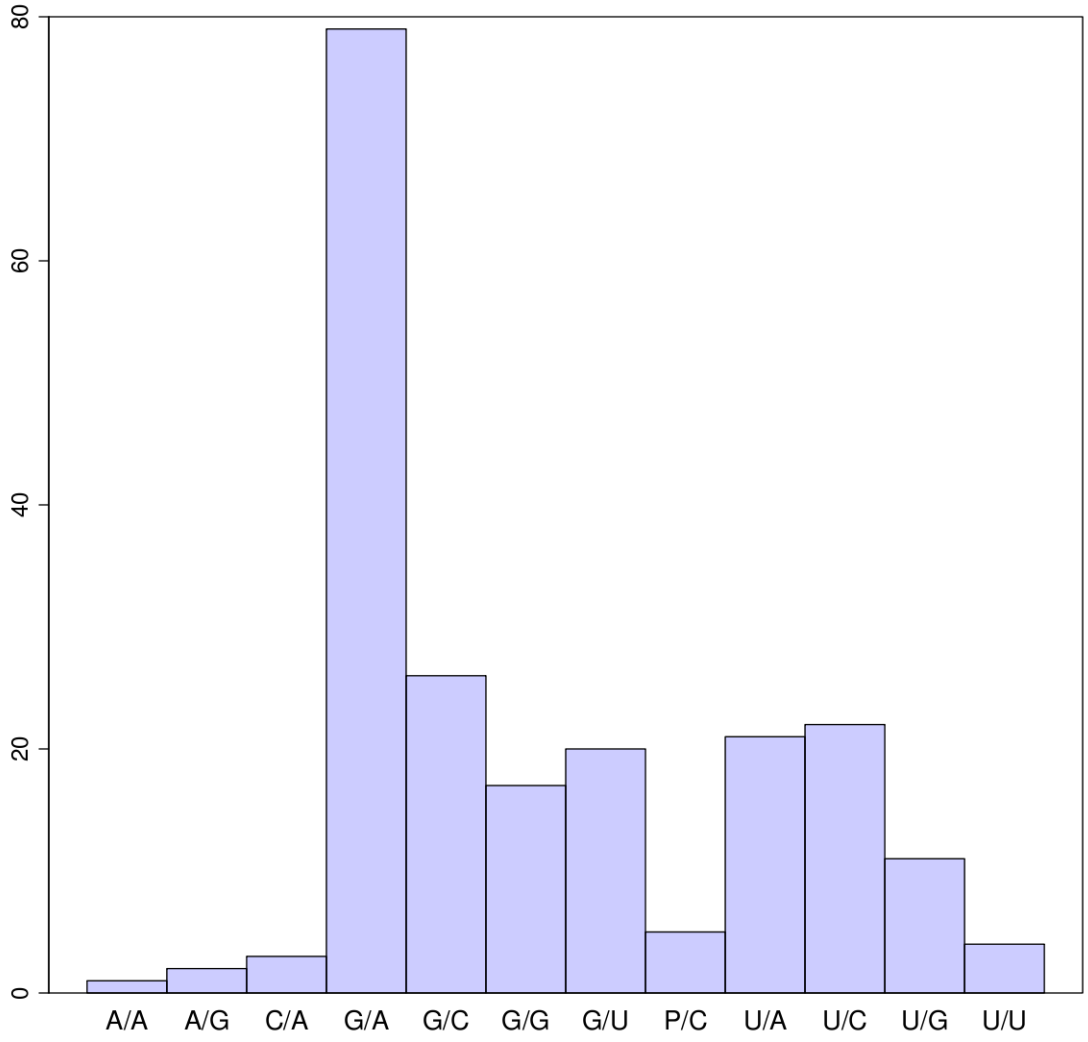
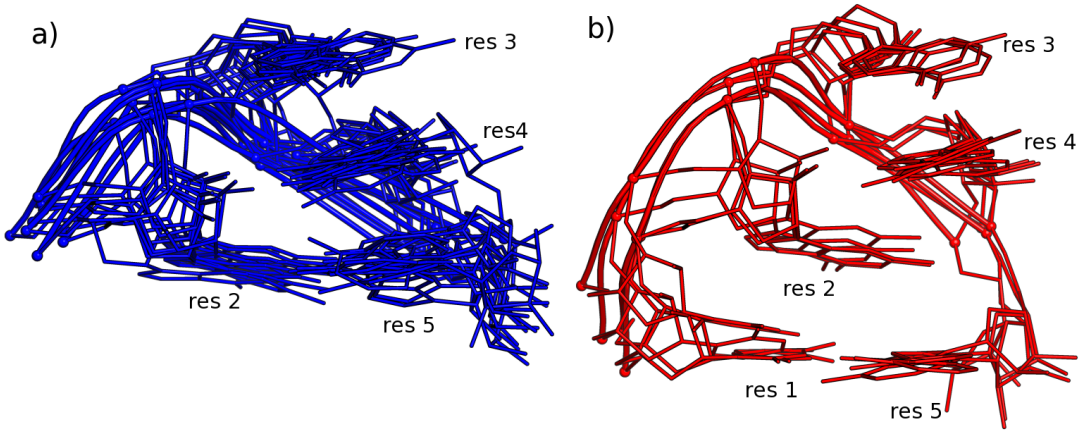


Figure 5.4: Histogram of the frequencies of base-step types at the first steps in the candidate GNRA motifs identified using “getMotif”. The motifs are identified by their similarity to a base-step parameter seed constructed from GNRA motifs found in the 23S subunit of rRNA (PDB_ID:1JJ2).



RMSD_rf (Å)	RMSD_bb (Å)	score	group	resid
0.53	0.32	71	2	1283
0.64	0.43	71	1	463
0.66	0.48	60	2	476
0.70	0.43	40	1	2375
0.76	0.31	99	2	307
0.77	0.46	145	1	2595
0.99	0.49	140	1	1223
1.03	0.41	35	1	630
1.08	0.74	106	1	87
1.19	0.48	71	1	488
1.19	0.31	120	2	500
1.40	0.76	134	1	1364
2.53	0.70	147	1	2357
3.42	3.16	112	1	1865

Figure 5.5: Molecular representation of two main groups of tetranucleotide structures related to the GNRA motif identified using “getMotif” in the 23S subunit of ribosomal RNA (PDB_ID:2J01). All structures are superimposed with respect to the refence frame between residues two and three. In a) Group I structures (RMSD [0.31-0.48]), which are drawn in blue, are close to a typical GNRA motif. The Group II structures (RMSD [0.41-3.16]) in b), which are drawn in red, are closer to a pentaloop. The steps between residues two and three, and three and four are common to the two groups as is clearly seen from the superposition of both groups in c). Residue two from Group II is commonly found forming a sheared G·A base-pair with a sequentially distant base. This type of interaction, which maintains the GNRA motif geometry and is called a lonepair triloop [9], has been previously found from sequence covariation analyisis. In the lower left area of the figure root-mean-square deviations, sum of differences scores, group identities, and residue identities (resid) of the first step in the motif are shown. The simmilariry of structures is quantified in terms of (i) the RMSD values of the sugar-phosphate backbone atoms in the common coordinate frame between residues two and three (RMSD_rf); (ii) the RMSD value of the aligned sugar-phosphate backbone atoms obtained using the VMD [16] software (RMSD_bb); and (iii) the sum of differences score (Equation 5.2).

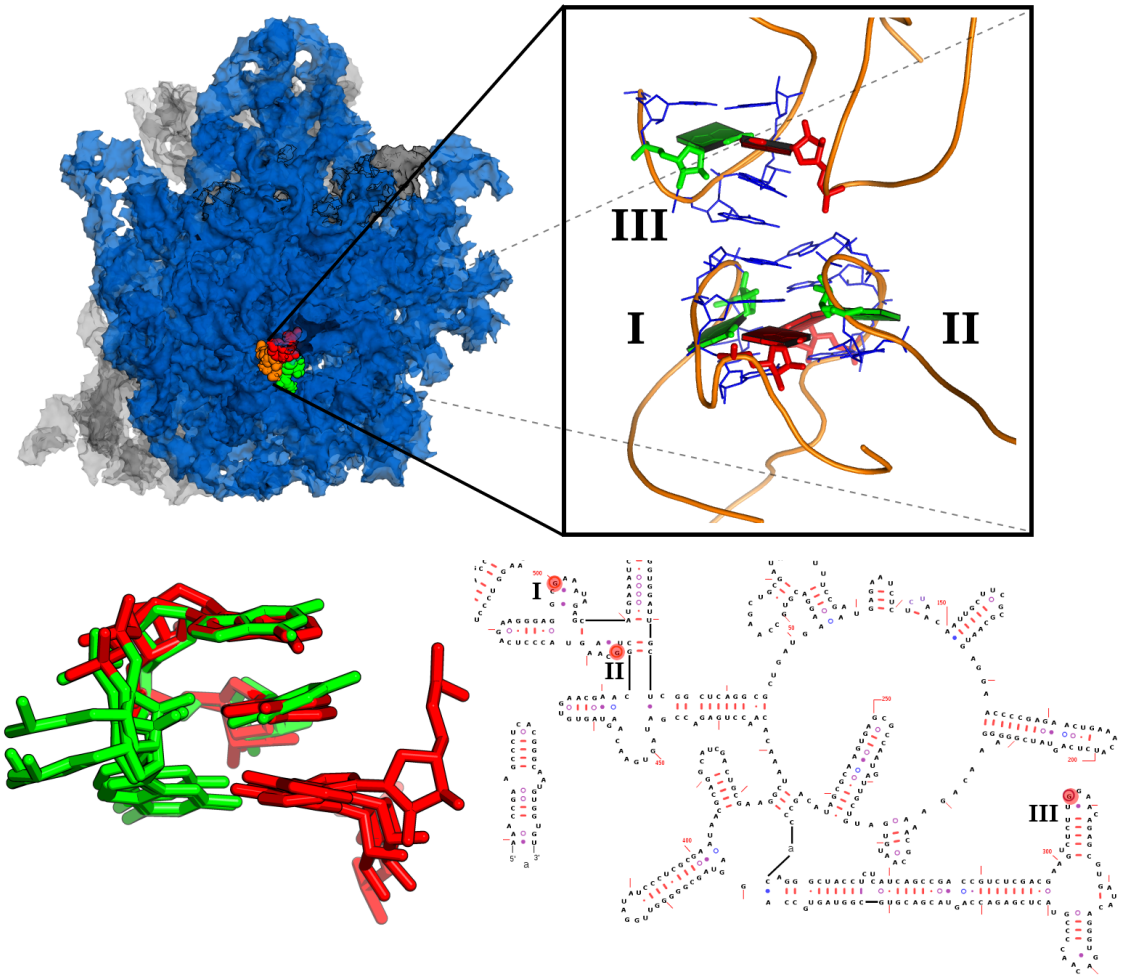


Figure 5.6: Subset of the Group II structures identified with “getMotif”, which correspond to the lonepair triloop motif identified by Gutell and collaborators [9]. In the upper left corner of the figure a space-filling representation of the *Thermus thermophilus* ribosome (PDB_ID:2J01) is shown, where the 50S subunit is drawn in blue, the 30S subunit is drawn in gray respectively, and the three interacting GNRA-like tetraloop motifs, known as lone-pair triloop motifs, are drawn in orange (I), green (II), and red (II) color. The upper right corner of the figure shows a zoomed image of the interacting lone-pair triloop motifs. Each one of the Group II GNRA tetraloop motif candidates are drawn in blue in a stick model representation. Sequentially distant adenines and guanines forming a sheared G·A base-pair are represented as red and green blocks in the identified GroupII structures. In the lower left corner of the figure the three lonepair triloop motifs, or structural GNRA motifs, are superimposed in the reference frame of residues two and three. In the lower right corner of the figure the location of the first residue in the identified GNRA structural motifs is highlighted by a red circle on top of the secondary structure of the large subunit of the ribosome (taken from Gutell Lab’s website [18]) and numbered according to the 3D structural image above.

farthest away from a typical A-RNA-like base-step. We plot the step-parameter values for this first step in the scatterplot in Figure 5.7. It is clear from this scatterplot that these steps, depicted by the blue points, constitute a well defined region in the space of rigid-body step-parameter configurations for the large subunit of the ribosome.

5.2 Results on RNA Motif Recognition via Step-Parameters

We previously asked ourselves if the geometric rigid-block description of base-pairing and base-stacking could help define RNA structural motifs.

We believe that the problem of defining RNA structural motifs is clearly more complicated than what any one structural research methodology, such as the analysis of backbone torsions or all-atom fitting, can alone predict. We have shown that the rigid-body parameters relating sequential bases facilitate motif searches in RNA atomic structures as well as in the rigorous description of structural motifs. It has also been shown in previous work on RNA in the Olson group [19], that the rigid-body parameters describing RNA base-pairs can help to find and describe novel RNA tertiary interactions important for understanding the three-dimensional fold of the ribosome. For example, the noncanonical sugar-edge sugar-edge parallel G·A base-pair acts as a linker of sequentially distant regions in the 50S subunit of the ribosome (PDB_ID:1JJ2), and is related to known RNA motifs such as the kink-turn motif. We think that this type of characterization and its automation with simple programs such as “getMotif” should prove useful to the community interested in ontological characterization of RNA.

Yet another specific question we formulated in this chapter inquires if we can use quantities derived from the 3DNA software package to make and automated search for a known motif, for example, the GNRA tetraloop motif, and perhaps find unknown motifs?

We have found that by defining seeds from known motifs we can find new motifs at the boundaries of the seeds through their base-steps. For example, we started a GNRA motif search using its average base-step parameters from the GNRA’s found in the 23S ribosomal subunit and found a closely related motif which we have called a Group II motif and which in turn is part of a composite motif [17] that is known as the lonepair triloop motif [9].

We can also perform other, simpler, but not as “precise” RNA motif searches, e.g. by following the proposed cluster validation methodology used in Chapter 2, to find an optimal method for clustering other 3DNA-computed properties of smaller dimensionality than the base-step parameters, for example,

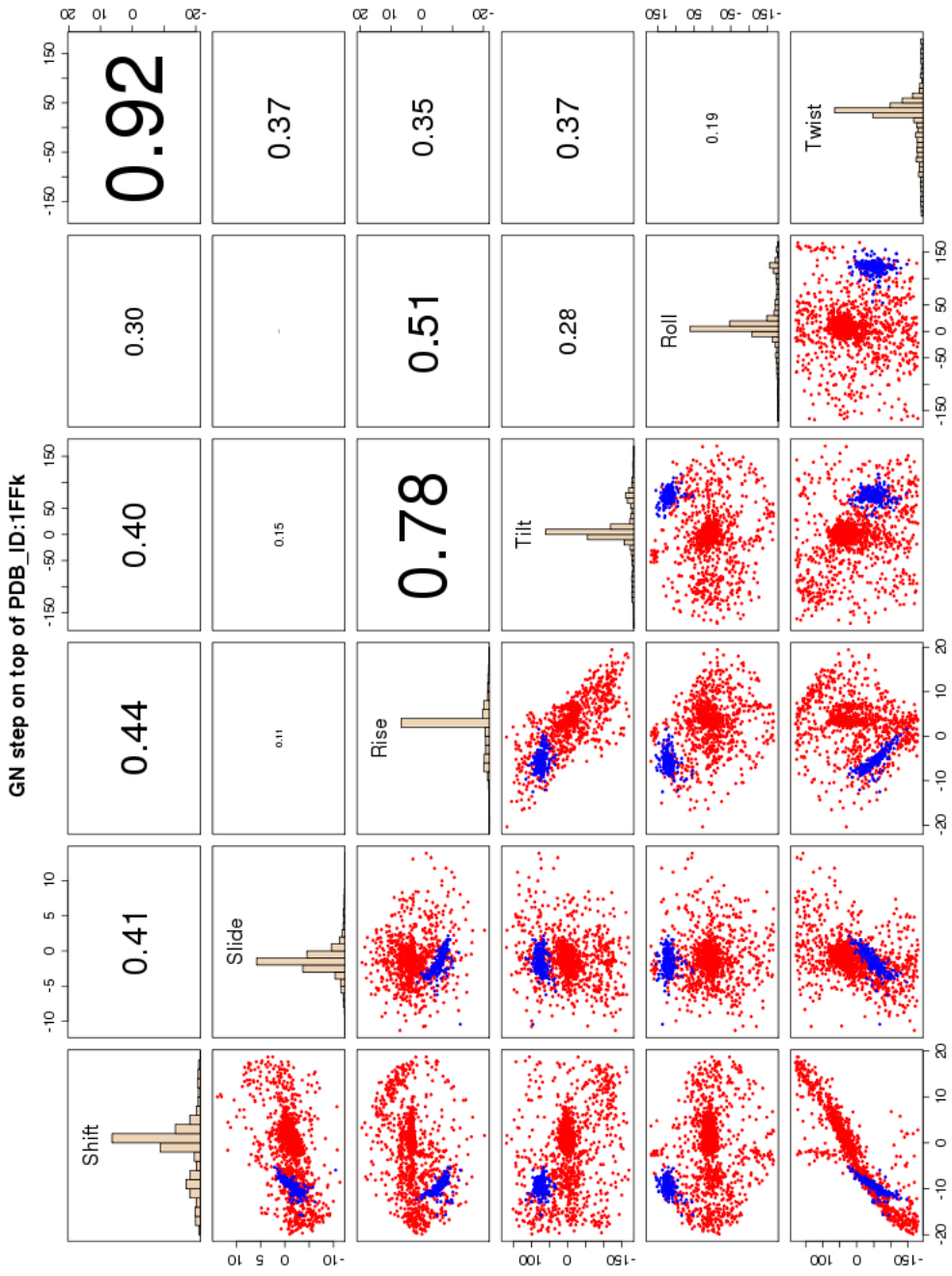


Figure 5.7: Scatterplot of base-step parameter values for the GN step in GNRA motif candidates (drawn as blue points) identified in a list given by the ROC, against a backdrop of the range of values for the step-parameters in all steps of the large subunit of the ribosome PDB_ID:1FFK (drawn as red points). Here the correlation coefficients for each scatterplot are given as explained in Figure 2.14.

sequential base overlaps.

Searches based on other rigid-body parameters, which have not been explored here, could also be useful. These include the helical parameters: x-displacement, y-displacement, helical-rise, inclination, tip, and helical twist, relating sequential bases by a single rotational operation.

References

[1] Lu, X.-J. and Olson, W. (2003) 3DNA: A Software Package for the Analysis, Rebuilding and Visualization of the Three-Dimensional Nucleic Acid Structures. *Nucleic Acids Research,* **31**, 5108–5121.

[2] Lu, X.-J. and Olson, W. K. (2008) 3DNA: A Versatile, Integrated Software System for the Analysis, Rebuilding and Visualization of Three-Dimensional Nucleic-Acid Structures. *Nature Protocols,* **3**, 1213–1227.

[3] Woese, C. R., Winker, S., and Gutell, R. R. (1990) Architecture of Ribosomal RNA: Constraints on the Sequence of "Tetra-Loops". *Proceedings of the National Academy of Sciences of the United States of America,* **87**, 8467–8471.

[4] Depaul, A. J., Thompson, E. J., Patel, S. S., Haldeman, K., and Sorin, E. J. (2010) Equilibrium Conformational Dynamics in an RNA Tetraloop From Massively Parallel Molecular Dynamics. *Nucleic Acids Research,* **38**, 4856–4867.

[5] Heus, H. A. and Pardi, A. (1991) Structural Features That Give Rise to the Unusual Stability of RNA Hairpins Containing GNRA Loops. *Science,* **253**, 191–194.

[6] Pley, H. W., Flaherty, K. M., and McKay, D. B. (1994) Model for an RNA tertiary interaction from the structure of an intermolecular complex between a GAAA tetraloop and an RNA helix. *Nature,* **372**, 111–113.

[7] Pley, H. W., Flaherty, K. M., and McKay, D. B. (1994) Three-Dimensional Structure of a Hammerhead Ribozyme. *Nature,* **372**, 68–74.

[8] Leontis, N. B., Altman, R. B., Berman, H. M., Brenner, S. E., Brown, J. W., Engelke, D. R., Harvey, S. C., Holbrook, S. R., Jossinet, F., Lewis, S. E., Major, F., Mathews, D. H., Richardson, J. S., Williamson, J. R., and Westhof, E. (2006) The RNA Ontology Consortium: An Open Invitation to the RNA Community. *RNA,* **12**, 533–541.

[9] Lee, J. C., Cannone, J. J., and Gutell, R. R. (2003) The Lonepair Triloop: A New Motif in RNA Structure. *Journal of Molecular Biology,* **325**, 65–83.

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

[11] Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1995) A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society,* **117**, 5179–5197.

[12] Sorin, E. J., Engelhardt, M. A., and Herschlag, D. (2002) RNA Simulations: Probing Hairpin Unfolding and the Dynamics of a GNRA Tetraloop. *Journal of Molecular Biology,* **317**, 493–506.

[13] Spackova, N., Reblova, K., and Sponer, J. (2010) Structural Dynamics of the Box C/D RNA Kink-Turn and its Complex with Proteins: The Role of the A-Minor 0 Interaction, Long-Residency Water Bridges, and Structural Ion-Binding Sites Revealed by Molecular Simulations. *Journal of Physical Chemistry B,* **114**, 10581–10593.

[14] Srinivasan, J., Miller, J., Kollman, P. A., and Case, D. A. (1998) Continuum Solvent Studies of the Stability of RNA Hairpin Loops and Helices. *Journal of Biomolecular Structure and Dynamics,* **16**, 671–682.

[15] Sarver, M., Zirbel, C. L., Stombaugh, J., Mokdad, A., and Leontis, N. B. (2008) FR3D: Finding Local and Composite Recurrent Structural Motifs in RNA 3D Structures. *Journal of Mathematical Biology,* **56**, 215–252.

[16] Eargle, J., Wrigth, D., and Luthey-Schulten, Z. (2006) Multiple Alignment of Protein Structures and Sequences for VMD. *Bioinformatics,* **22**, 504–506.

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

[18] Cannone, J. J., Subramanian, S., Schnare, M. N., Collett, J. R., DSouza, L. M., Du, Y., Feng, B., Lin, N., Madabusi, L. V., Muller, K. M., Pande, N., Shang, Z., Yu, N., and Gutell, R. R. (2002) The Comparative RNA Web (CRW) Site: An Online Database of Comparative Sequence and Structure Information for Ribosomal, Intron, and Other RNA's. *BMC Bioinformatics,* **3**, 1–31.

[19] Xin, Y. Non-Canonical Base Pairs in RNA Structures and Folding, Ph.D. thesis PhD thesis Rutgers, The State University of New Jersey (2005).

Appendix A

Standard reference frame and local parameters

In addition to the description of RNA structures at the level of torsion angles, one can also describe structure in terms of the spatial arrangements of adjacent or associated bases. The structural description of RNA used here comes from the program 3DNA [1], which reports three sets of parameters that define the local arrangements of bases.

1. Base-pair parameters,
2. Base (base-pair) step parameters,
3. Base (base-pair) local helical parameters.

The base or base-pair parameters are the quantities that bring into coincidence coordinate frames on two objects using ideas from classical mechanics. The first two sets of parameters are based on Cartesian coordinates of two bases or base pairs, whereas the third set of helical parameters, resembles cylindrical coordinates and are based on the single rotation that brings coordinate frames on two bases or base-pairs into coincidence (Chasles's theorem) [2].

A.1 Base-pair and base-step parameters

In 3DNA one starts with a Protein Data Bank (PDB) formatted [3] file, which is usually based on experimental information[i] and which can be downloaded from the Nucleic Acid Database (NDB) or PDB. This file contains the Cartesian coordinates and other information, such as sequence composition, about the given structure of the atoms. With this experimental data one performs a least-squares fit to a standard reference frame [4]. The octave script at http://www.eden.rutgers.edu/~esguerra/RNA/scripts.html provides a useful tutorial example. The coordinate origin which is embedded in the standard reference frame, is used for the determination of both base and base-pair parameters. In the case of single

[i] A PDB file is sometimes the result of theoretical modeling.

unpaired bases, the program keeps the origin of one base of an ideal Watson-Crick pair. The definition of this frame is illustrated in Figure A.1

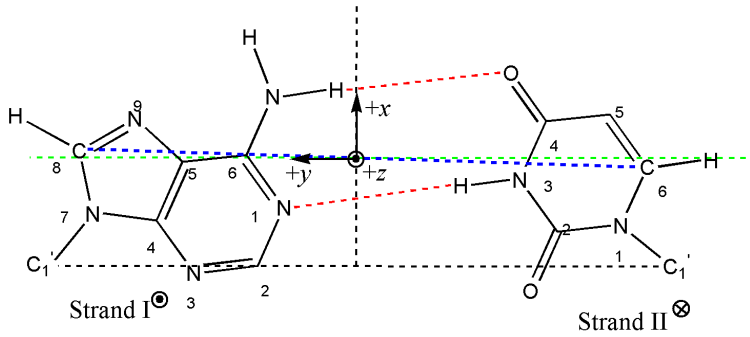


Figure A.1: Standard reference frame of an A-T base-pair [4]. The *y*-axis (dashed green line) is chosen to be parallel to the line connecting the C1′ of adenine and the C1′ of thymine associated in an ideal Watson-Crick base-pair. The *x*-axis is the perpendicular bisector of the C1′ - C1′ line, and 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 normal to the base-pair plane (defined in a positive sense with respect to the leading base, here A) and the direction of the x-axis is defined by the cross product of the $\hat{x}$ and $\hat{y}$ unit vectors.

After the coordinate origins for two consecutive base of base-pairs comprising a step have been computed, then a so-called middle step triad (MST) [5] is defined. Defining the middle step triad is described by the following procedure:

1) Find the angle Γ between consecutive normals, *i.e.*, *z*-axes. Since these are unit vectors, the angle is defined by the scalar product:

$$\Gamma = \cos^{-1}(\hat{z}_i \cdot \hat{z}_{i+1}) \tag{A.1}$$

2) Then find the vector which is perpendicular to the two normals (*z*-axis). This vector is obtained from the cross product of the consecutive *z*-axes (that is, the normal to the plane formed by the two vectors). This axis is called the roll-tilt axis and is normalized to form the unit vector $\hat{r_t}$,

$$\hat{r_t} = \frac{\hat{z}_i \times \hat{z}_{i+1}}{|\hat{z}_i \times \hat{z}_{i+1}|} \tag{A.2}$$

3) To make consecutive *z* vectors coincide, one uses a linear homogeneous transformation $R(\theta)$ about the roll-tilt axis such that the original orientation matrices T_i and T_{i+1}, i.e. the set of unit vectors

along the x, y and z-axes of each base or base-pair specified in the columns of each, are rotated by $\theta = \pm\Gamma/2$ to yield the transformed $T_i^{'}$ and $T_{i+1}^{'}$ orientation matrices.

$$T_i^{'} = R_{rt}(\pm\Gamma/2)T_i \tag{A.3}$$

$$T_{i+1}^{'} = R_{rt}(\mp\Gamma/2)T_{i+1} \tag{A.4}$$

The origin for the middle step triad (MST) is the average of the position vectors r_i and r_{i+1} for the i and $i+1$ reference frames,

$$r_{MST} = \frac{r_i + r_{i+1}}{2} \tag{A.5}$$

4) Again using the dot product one can find the angle between the transformed $\hat{y}^{'}$ vectors. This angle is equal to the magnitude of the Twist (Ω) in the base of consecutive bases or base-pairs. The dot product of the z unit vector for the middle step triad (MST) $\hat{z}_{MST}$ with the vector resulting from the cross product of $\hat{y}_i^{'}$ and $\hat{y}_{i+1}^{'}$ gives the sign of Ω. Since the transformed *x-y* plane is orthogonal to $\hat{z}$ then this applies in the same manner for *x*,

$$\Omega = cos^{-1}(\hat{y}_i^{'} \cdot \hat{y}_{i+1}^{'}) \tag{A.6}$$

$$(\hat{y}_i^{'} \times \hat{y}_{i+1}^{'}) \cdot \hat{z}_{MST} > 0, \quad \text{then } \Omega > 0 \tag{A.7}$$

$$(\hat{y}_i^{'} \times \hat{y}_{i+1}^{'}) \cdot \hat{z}_{MST} < 0, \quad \text{then } \Omega < 0 \tag{A.8}$$

5) With more scalar products one can find other angles, such as the phase angle ϕ,

$$\phi = cos^{-1}(\hat{rt} \cdot \hat{y}_{MST}) \tag{A.9}$$

$$(\hat{rt} \times \hat{y}_{MST}) \cdot \hat{z}_{MST} > 0, \quad \text{then } 180 \geq \phi \geq 0 \tag{A.10}$$

$$(\hat{rt} \times \hat{y}_{MST}) \cdot \hat{z}_{MST} < 0, \quad \text{then } -180 \leq \phi \leq 0 \tag{A.11}$$

6) The roll ρ and tilt τ angles, which are the remaining angular degrees of freedom for step parameters, are defined in terms of the bending angle and the phase angle:

$$\rho = \Gamma cos(\phi) \tag{A.12}$$

$$\tau = \Gamma sin(\phi) \tag{A.13}$$

To get the remaining three translational degrees of freedom for step parameters (Dx, Dy, Dz) one needs to express the displacement vector in the middle step triad frame:

$$[D_x D_y D_z] = T_{MST}(r_{i+1} - r_i) \tag{A.14}$$

The procedure to compute the base-pair parameters is completely analogous. The opening ω, buckle κ, and propeller σ are the analogs of twist Ω, roll ρ, and tilt τ, and the middle step triad is called the middle base triad MBT. The axes which are made to coincide are the *y*-axes and not the *z*-axes as in the base-pair step case [5].

The parameters obtained by this procedure are depicted graphically in Figure A.2.

A.2 Local helical parameters

Local helical parameters are determined using Chasles's theorem, which states [2]:

> "*One can always transport a free rigid body from one position and orientation to another position and orientation by a single continuous motion along a unique axis of rotation.*"

For the case of three-dimensional nucleic acid base steps what this means is that, instead of rotating around a series of reference-frame centered axes and then translating along another reference-frame centered axis, one rotates about and also translates along only one common axis, which is not reference-frame centered. This allows us to define the orientation of a local helical axis (or unique rotational-translational axis) as a unit vector given by equation A.15:

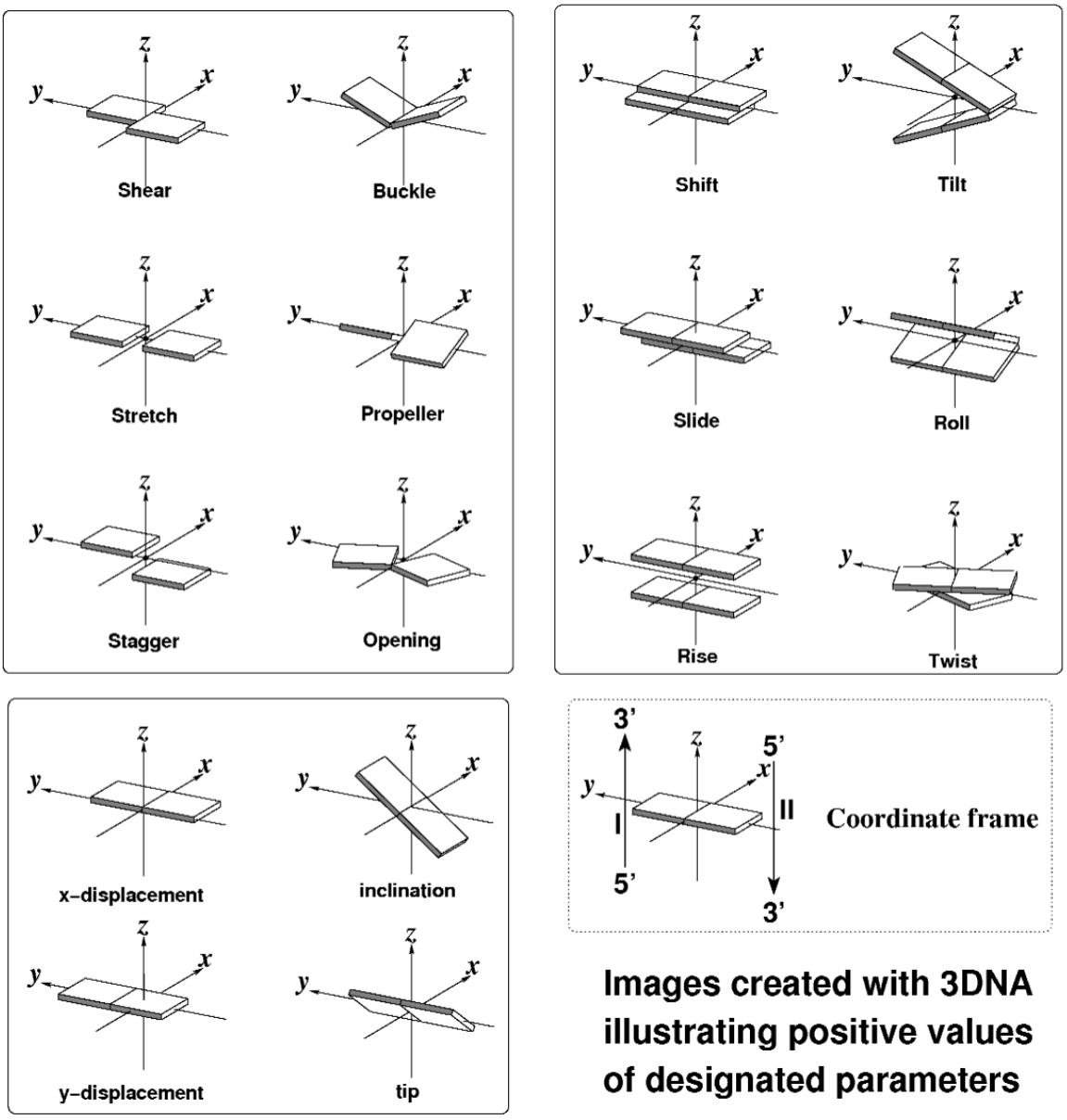


Figure A.2: Illustration of base-pair and base step parameters [1]. As seen in the upper right corner the base and base-pair step parameters correspond to the three translational and three rotational degrees or freedom which describe the geometry of a rigid-body. Thus the three translational degrees of freedom, Shift, Slide, and Rise, are expressed as linear displacements along the x, y, and z axis, and the three rotational degrees of freedom, Tilt, Roll, and Twist, as angular displacements around x, y, and z.

$$h = \begin{bmatrix} h_x \\ h_y \\ h_z \end{bmatrix} \tag{A.15}$$

where:

$$h_x = \frac{\tau}{\Omega_h}, \qquad h_y = \frac{\rho}{\Omega_h}, \qquad h_z = \frac{\Omega}{\Omega_h} \tag{A.16}$$

$$\Omega_h = \sqrt{\tau^2 + \rho^2 + \Omega^2} \tag{A.17}$$

The local helical axis can be defined alternatively [6] as a cross product:

$$h = (x_2 - x_1) \times (y_2 - y_1) \tag{A.18}$$

where the x and y refer to the reference frames on base pairs 1 and 2.

References

[1] Lu, X.-J. and Olson, W. (2003) 3DNA: A Software Package for the Analysis, Rebuilding and Visualization of the Three-Dimensional Nucleic Acid Structures. *Nucleic Acids Research,* **31**, 5108–5121.

[2] Babcock, M. S., Pednault, E. P. D., and Olson, W. K. (1994) Nucleic Acid Structure Analysis; Mathematics for Local Cartesian and Helical Structure Parameters that are Truly Comparable Between Structures. *Journal of Molecular Biology,* **237**, 125–156.

[3] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Research,* **28**, 235–242.

[4] Olson, W. K., Bansal, M., Burley, S. K., Dickerson, R. E., Gerstein, M., Harvey, S. C., Heinemann, U., Lu, X.-J., Neidle, S., Shakked, Z., Sklenar, H., Suzuki, M., Tung, C.-S., Westhof, E., Wolberger, C., and Berman, H. M. (2001) A Standard Reference Frame for the Description of Nucleic Acid Base-pair Geometry. *Journal of Molecular Biology,* **313**, 229–237.

[5] Lu, X.-J., Hassan, M. A. E., and Hunter, C. A. (1997) Structure and Conformation of Helical Nucleic Acids: Analysis Program (SCHNAaP). *Journal of Molecular Biology,* **273**, 668–680.

[6] Bansal, M., Bhattacharyya, D., and Ravi, B. (1995) NUPARM and NUCGEN: Software for Analysis and Generation of Sequence Dependent Nucleic Acid Structures. *Computer Applications in the Biosciences: CABIOS,* **11**, 281–287.

Appendix B

Clustering Analysis (CA)

B.1 General Methodology

We considered each of the 20 non-A-RNA type structures as a vector composed of six base-step parameters. We group these vectors using cluster analysis following an automated process shown to succesfully reproduce well known patterns of the Periodic Table from a selected set of variables, such as, electronegativity, ionization potential, and other elemental properties [1]. The procedure followed here is an adaptation of how clustering is used to re-construct the Periodic Table classification of the elements [2].

We start by normalizing the vectors of step parameters,

$$\overline{x}_{jA} = \frac{x_{jA} - x_{jmin}}{x_{jmax} - x_{jmin}}, \tag{B.1}$$

where x_{jA} is the value of the step parameter j of the structure A and x_{jmin} and x_{jmax} are the minimum and maximum values for a particular step parameter *j* [2]. Then, using the software package **R** [3], we cluster these vectors into groups. These groups can be displayed in a tree representation, also called a dendrogram, or in biology, a phylogenetic tree (see Figure B.1).

To cluster these vectors into groups, it is necessary to define the distance between the vectors. In this work we used three distance definitions. These distances are often referred to as Manhattan, Euclidean and maximum distances. The first two distances are particular cases of what is known as Minkowski's metric:

$$d(X,Y) = \Big(\sum_{i=1}^{N} |x_i - y_i|^k \Big)^{\frac{1}{k}}, \tag{B.2}$$

where $d(X,Y)$ refers to the distance between two vectors X and Y, and N is the dimensionality of the vector. For the case of step parameters, N is six. In the case where *k* is equal to 1, the definition

corresponds to the Manhattan distance (a distance measured by following along the edges of blocks). In the case where *k* is equal to 2, we have the familiar Euclidean distance. The remaining distance, that is, the maximum distance, is defined by:

$$d(X, Y) = max|x_i - y_i|, \tag{B.3}$$

where the distance between vectors X and Y is the maximum difference between vector variables.

Yet another distance definition used in this text, which is more frequently interpreted as a statistical measure of error, is the root-mean-square deviation (RMSD). The root-mean-square deviation can be seen as a mean Euclidean distance, in the sense that it is defined as the square root of the ratio between the sum of the squared differences between vector variables, and the dimensionality of the vectors N,

$$RMSD(X, Y) = \Big(\frac{\sum_{i=1}^{2} |x_i - y_i|^2}{N}\Big)^{\frac{1}{2}}. \tag{B.4}$$

Once the distance is defined, we use a hierarchical method to cluster the set of distances between base-step parameters. The clustering algorithm first finds the two closest vectors (given by one of the distance definitions) and groups them together. Then it compares the distance of the elements in the newly formed group and the elements remaining to be grouped, according to the particular clustering method. For example, the single linkage clustering method takes the minimum distance between elements as the clustering criterion. Such an approach would, as all other agglomerative hierarchical methods do, group together the closest vectors given the distance definition, and then would use the method definition (minimum distance) to compare the distance of the elements of the group to the elements which remain ungrouped, or to the elements of other groups. As new groups are formed the process is repeated following a hierarchical order, that is, whatever distance is smaller gives the grouping criterion. We have used four hierarchical clustering methods, the descriptions of which follow in the next section.

For every possible combination of clustering method and distance definition we obtain a dendrogram. The combination of three distance definitions and four clustering methods leads to 12 clustering trees. These trees are not all exactly the same but show recurring groups of conformers. To find the groups, which are repeated among the trees, a consensus analysis is performed using the clue package

[4], implemented in **R**. The resulting consensus tree is illustrated in Figure 2.5.

B.2 Hierarchical methods

The hierarchical clustering methods include:

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

$$D(X,Y) = min\{d(x_i, y_j) : x_i \in X, y_j \in Y\}. \tag{B.5}$$

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

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

$$D(X,Y) = max\{d(x_i, y_j) : x_i \in X, y_j \in Y\}. \tag{B.6}$$

   Here X, Y, and $d(x_i, y_j)$ have the same meaning as above.

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

$$D(X,Y) = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} d(x_i, y_j). \tag{B.7}$$

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

4. *Centroid linkage clustering*, where the distance between cluster centroids is used as the clustering criterion,

$$D(X,Y) = d(\overline{x}, \overline{y}), \tag{B.8}$$

$$\overline{x} = \frac{1}{N_x} \sum_{i=1}^{N_x} x_i, \tag{B.9}$$

$$\overline{y} = \frac{1}{N_y} \sum_{i=1}^{N_y} y_i. \tag{B.10}$$

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

5. *Ward's Method*, where the error of the sum of squares (*ESS*) is used as the criterion,

$$D(X,Y) = ESS(XY) - [ESS(X) + ESS(Y)], \tag{B.11}$$

$$ESS(X) = \sum_{i=1}^{N_x} \left| x_i - \frac{1}{N_x} \sum_{j=1}^{N_x} x_j \right|^2 . \tag{B.12}$$

As an example let us think of a case where we have five structures. Each one structure is descibed by a bi-dimensional vector as illustrated in Table B.1.

The first step is to chose a distance definition, such as the Manhattan distance. The distance values between structures can then be displayed in a lower triangular matrix (usually referred to simply as the distance matrix),

$$d(X,Y) = \begin{vmatrix} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0 & & & & \\ 2 & 4 & 0 & & & \\ 3 & 8 & 12 & 0 & & \\ 4 & 11 & 9 & 5 & 0 & \\ 5 & 11 & 15 & 3 & 6 & 0 \end{vmatrix} . \tag{B.13}$$

Here we assign labels 1-5 to the rows and columns to make clear the distances between specific pairs of vectors. For example, the Manhattan distance between structures 2 and 3,

$$d(2,3) = |-2.00 - 2.00| + |6.00 - -2.00| = 12, \tag{B.14}$$

is found in the (3,2) element of the lower triangular matrix, i.e., row 3, column 2.

Once we have calculated the distances we pick a clustering method. In this case, we will use the average linkage clustering method. There are two hierarchical techiques, one called agglomerative, or bottom-up, and the other called divisive, or top-down. We will use the agglomerative technique, that is, going from the bottom where no objects are grouped, to the top, where all the objects constitute one final group. The first step then is to group whatever structures are closest, 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

$$D(\{3,5\},1) = \frac{1}{2*1} * (8+11) = 9.5 \tag{B.15}$$

$$D(\{3,5\},2) = \frac{1}{2*1} * (12+15) = 13.5 \tag{B.16}$$

$$D(\{3,5\},4) = \frac{1}{2*1} * (5+6) = 5.5 \tag{B.17}$$

Since the distances between {3, 5} and all remaining unclustered vectors are higher than the distance between vectors 1 and 2 ($d(1,2) = 4$) then {1, 2} is grouped. Because the average distance between {3, 5} and 4 (see equation B.17) in hierarchical increasing order is 5.5 we group these data points. The next value, following the lower to higher hierarchy, is 6 ($d(4,5) = 6$). Since we have already grouped 3 with 5, we have to keep advancing in the hierarchy and find that the only remaining possibility for grouping is, group {1, 2} and {4, 3, 5}. We group them together as illustrated in Figure B.1.

B.3 Clustering Validation Techniques

The main objective of automated clustering methods is to find a reduced number of groups which share common characteristics within a large dataset. The main problem with the practical approaches used to solve this task is that clustering methods do not offer an 'a priori' answer to the optimal number of groups that a large dataset can be split into. In our simple example, shown in Figure B.1, our data split into two main groups. Since the assesment depends upon the distances and clustering methods used in particular cases, then an emergent necessity is to be able to determine the validity of an optimal number of clusters solution. Clustering validation techniques are thus crucial in the analysis of the gigantic amounts of data which are being produced in the the post-genomic boom of biological information [5], such as the structural data dealt with in this thesis.

We have used the package clValid [6], which implements a variety of clustering validation algorithms,

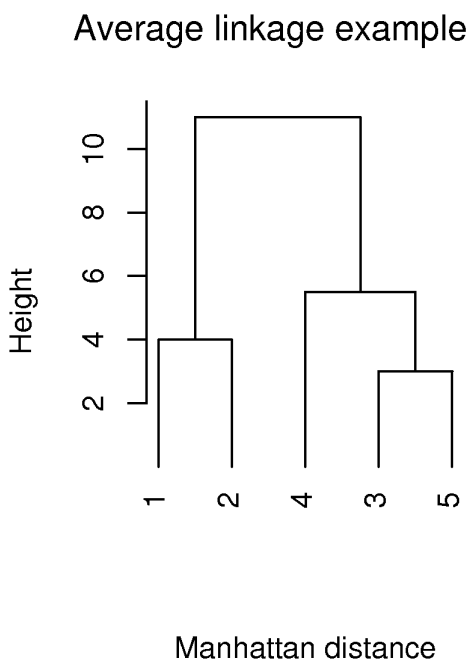


Figure B.1: Clustering tree for 5 bi-dimensional vectors using the Manhattan distance definition and the average linkage clustering method.

using the programming language **R** [7]. The clValid package comprises measures which reflect the compactness, connectedness, and separation of cluster partitions. The concept of compactness of a cluster refers to the extent of intra-cluster variation. The conectedness concept is a more local concept and means that neighboring data elements should belong to the same cluster. The separation concept quantifies the degree of separation between individual clusters. An illustration of these concepts is presented in Figure B.2.

B.3.1 Internal Measures

There are two main types of measures in the clValid package, which are called internal measures and stability measures. In general, internal measures are those that use, exclusively, the dataset, the clustering partitions, and the intrinsic information in the data to quantify the quality of a clustering result. In clValid three indices are used to account for compactness, conectedness and separation. The connectivity index measures cluster connectedness, while the silhouette width and Dunn index provide combined measures of the cluster separation and compactness.

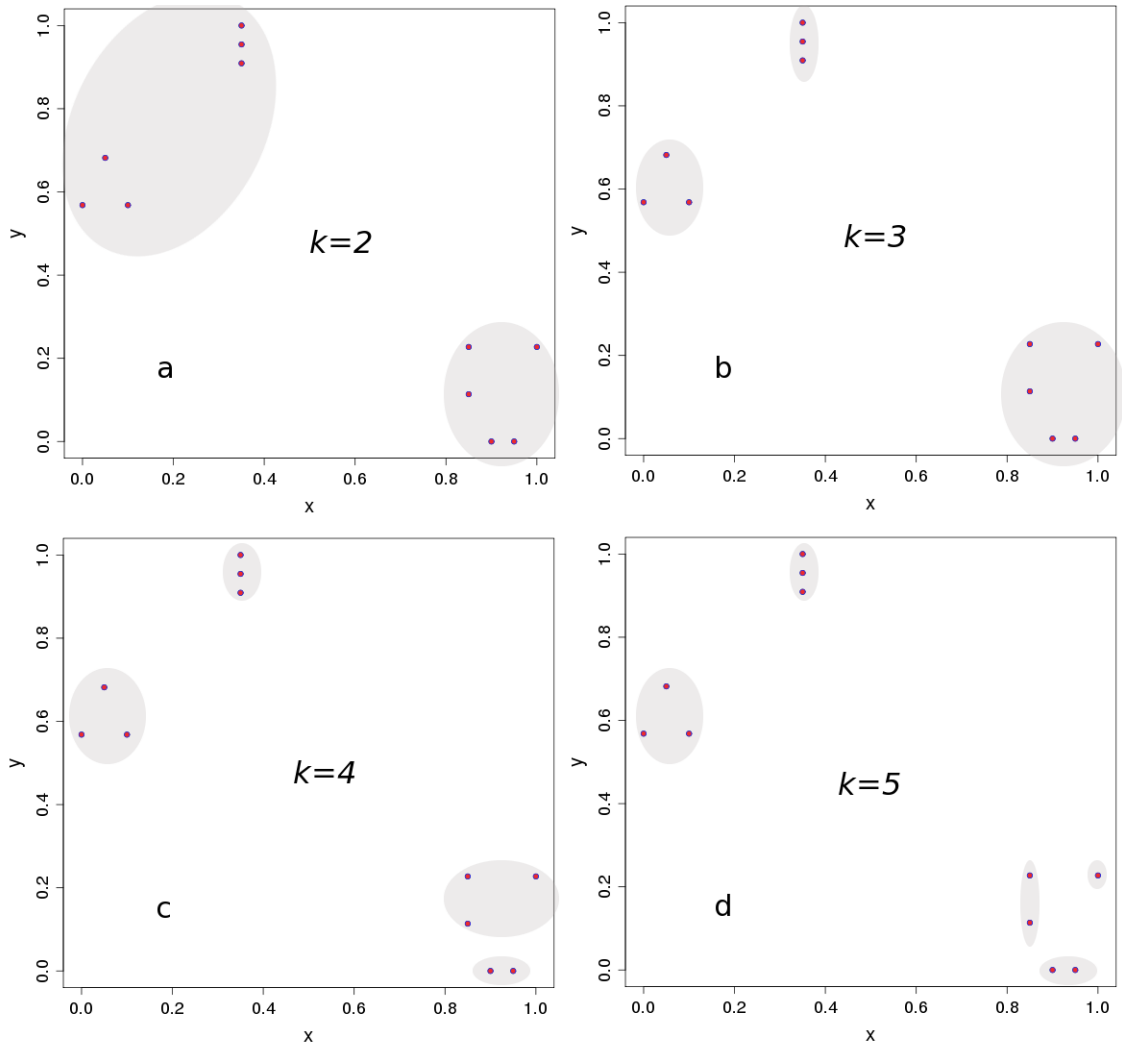


Figure B.2: Illustration of the compactness, connectedness, and separation of a bi-dimensional dataset. Images a-d present the solutions for hierarchical clustering of the Euclidean distances between the datapoints considering $k = [2 - 5]$ clusters. The $k = 3$ case, i.e., three clusters stands out from the other solutions in being composed of more compact and separated groups. This fact is quantified in clValid by the *asw* index and the Dunn index. The $k = 2$ solution in (a), by contrast, is clearly more connected than all other solutions. One can also see that as the data are grouped into more clusters, the connectivity will be progressively lost simply due to the splitting of the data.

Connectivity Index

The connectivity index is defined as:

$$Conn(\mathcal{C}) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_{i(j)}}, \tag{B.18}$$

where $nn_{i(j)}$ stands for the j^{th} nearest neighbor (nn) to the i^{th} observation. If i, and its neareast neighbor $nn_{i(j)}$ are in the same cluster, the term $x_{i,nn_{i(j)}}$ is zero. If they do not belong to the same cluster, then the term is assigned a value of $1/j$. The parameter L resembles a radius, in the sense that it determines how many nearest neighbors are taken into account per connectivity weight.

$$x_{i,nn_{i(j)}} = \begin{cases} 0 & \text{if } i \text{ and } nn_{i(j)} \in C_k \\ \frac{1}{j} & \text{otherwise.} \end{cases} \tag{B.19}$$

The connectivity is defined for a particular clustering partition $\mathcal{C} = C_1, ..., C_K$, that is, for a particular number of clusters k where the total number of observations is N. The connectivity index can give values between zero and infinity, and smaller values mean better connectivity.

Average Silhouette Width

The average silhouette width (*asw*) is the mean of the silhouette values $S(i)$ for n observations,

$$\textit{asw} = \frac{1}{n} \sum_{i=1}^{n} S(i), \tag{B.20}$$

where the silhouette value is defined as:

$$S(i) = \frac{b_i - a_i}{\max(a_i, b_i)}. \tag{B.21}$$

Here a_i is the average dissimilarity of observation i with respect to other observations inside the cluster i to which it belongs, and b_i is the average dissimilarity of observation i to the observations outside its

own cluster, where a_i and b_i are respectively defined as,

$$a_i = \frac{1}{n(C(i))} \sum_{j \in C(i)} d(i,j), \tag{B.22}$$

$$b_i = \min_{C_k \in \mathcal{C}} \sum_{j \in C_k} \frac{d(i,j)}{n(C_k)}. \tag{B.23}$$

The dissimilarity $d(i,j)$ appearing in the expression is usually a distance, such as the Euclidean or Manhattan distance, although formally it does not need to be a metric. Notice that the contributions to the average silhouette width (*asw*), defined in Equation B.21 can only have values in the range $[-1,1]$. According to Kaufman and Rousseeuw [8], acceptable clusterings usually have an *asw* value above 0.5 and those with values below 0.2 should be considered as not well clustered. In this regard, it should be noted that all the values for validation of the hierarchical clustering for single-base step-parameters lie above 0.5 in the lower left plot of Figure 2.11.

Dunn Index

The Dunn index is similar to the *asw* index in the sense that it also quantifies the separation and compactness of clustering solutions. It is an index which measures the ratio between the smallest inter-cluster distance and the largest intra-cluster distance in a partitioning. The index is defined by:

$$D(C) = \min_{C_k \in C} \left(\frac{\min_{C_l \in C} d(C_k, C_l)}{\max_{C_m \in C} diam(C_m)} \right), \tag{B.24}$$

where $\min_{C_{l \in C}} d(C_k, C_l)$ is the minimum inter-cluster distance, and $\max_{C_m \in C} diam(C_m)$ is the maximum intra-cluster distance. The Dunn index can have values between zero and infinity, and a higher value means a better cluster separation and compactness.

B.3.2 Stability Measures

Stability measures are a special type of internal measure, which evaluate the consistency of a clustering result by comparing clustered groups after sequential removal of a column of data [9]. In the following definitions N stands for the total number of data points (rows of data) and M is the total number of columns, dimensions in the data. The dimensions in data points usually are taken as a collection of samples or time points. These measures are specially suited for highly correlated data, as is commonly

the case for high-throughput genomic data, such as micro-arrays. In the cases shown here an average is taken over the combined space of the full data and the reduced data sets.

Average Proportion of Non-overlap (*APN*)

The average proportion of non-overlap (*APN*) is a measure of the proportion of datapoints which are not placed in the same cluster when a comparison is made between full data clustering, and the clustering performed with one column of data, such as one type of base-step parameter, removed. Its value is defined as:

$$APN(\mathcal{C}) - \frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M}\left(1 - \frac{n(C^{i,j} \cap C^{i,0})}{n(C^{i,0})}\right), \tag{B.25}$$

where $C^{i,0}$ is the cluster which contains data point i in the full dataset, and $C^{i,l}$ is the cluster which contains point i in the reduced dimensionality dataset, where the column l has been removed. The value of *APN* ranges between $[0 - 1]$ and is optimal for values closer to zero [9, 6].

Average Distance (*AD*)

The average distance (*AD*) measure computes the distance between datapoints in a particular cluster of the full dimensional dataset and that where one column, or dimension has been removed,

$$AD(\mathcal{C}) = \frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M} - \frac{1}{n(C^{i,0})n(C^{i,l})}\left[\sum_{i\in C^{i,0}, j\in C^{i,l}} d(i,j)\right]. \tag{B.26}$$

Here $d(i,j)$ is the distance between point i in the full dimensional dataset and point j in the reduced dimension dataset. The values of *AD* range between $[0 - \infty]$ and are optimal when they come closer to zero [9, 6].

Average Distance Between Means (*ADM*)

The average distance between means (*ADM*) score measures the distance between the mean of the datapoints within a cluster with the full number of dimensions and the mean of the datapoints of the

same cluster when reduced by one dimension,

$$ADM(\mathcal{C}) = \frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M} d(\bar{x}_{C^{i,l}}, \bar{x}_{C^{i,0}}). \tag{B.27}$$

Here $\bar{x}_{C^{i,0}}$ is the mean value of the data points in the cluster which contains point i in the full dataset, and $\bar{x}_{C^{i,l}}$ is the mean value of the data points in the cluster containing point i with column l removed. The values of *ADM* range between $[0 - \infty]$ and like the *AD* and *APN* values are optimal the closer they are to zero [9, 6].

References

[1] Restrepo, G., Mesa, H., Llanos, E. J., and Villaveces, J. L. (2004) Topological Study of the Periodic System. *Journal of Chemical Information and Computer Science,* **44**, 68–75.

[2] Restrepo, G., Llanos, E. J., and Meza, H. (2006) Topological Space of the Chemical Elements and its Properties. *Journal of Mathematical Chemistry,* **39**, 401–416.

[3] Ihaka, R. and Gentleman, R. (1996) R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics,* **5**, 299–314.

[4] Hornik, K. (2005) A CLUE for CLUster Ensembles. *Journal of Statistical Software,* **14**, 1–25.

[5] Handl, J., Knowles, J., and Kell, D. B. (2005) Computational Cluster Validation in Post-Genomic Data Analysis. *Bioinformatics,* **21**, 3201–3212.

[6] Brock, G., Pihur, V., Datta, S., and Datta, S. (2008) clValid: An R Package for Cluster Validation. *Journal of Statistical Software,* **25**, 1–22.

[7] R Development Core Team R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing Vienna, Austria (2009) ISBN 3-900051-07-0.

[8] Kaufman, L. and Rousseeuw, P. J. (1990) Finding Groups in Data, Wiley - Interscience, .

[9] Datta, S. and Datta, S. (2003) Comparisons and validation of statistical clustering techniques for microarray gene expression data.. *Bioinformatics,* **19**, 459–466.

Appendix C

Persistence Length

Nucleic Acids and other polymers, due to their size, can be understood as mechanical objects [1, 2] and therefore engineering approaches are often used for their understanding. The methodology, which considers the polymer as a long continuous rod, is known as continuum elastic theory. This type of model leaves little space for taking into account the nature of the subunits which make up the polymer. That is, the theory is mainly applicable to homopolymers made up of identical subunits with limited bending, twisting, and stretching motions. In nucleic acids this is not necessarily the case, and a more general approach is needed to take into account the possibility of having different subunits and subunits with different levels of motion in the polymer. Olson and collaborators have developed a sequence-dependent model, refered to here as the "realistic" model [3], to treat long, fluctuating DNA helices. The model uses a harmonic approximation to treat the motions of base-pair steps and uses force constants and rest states derived from X-ray crystallographic data from the Nucleic Acid Database (NDB) [4, 5]. Within the context of the "realistic" model, Czapla et al. [6] have developed a Gaussian sampling methodology to generate random chain configurations and adapted matrix methods to compute global polymer properties, like the persistence length, from expressions developed by Flory [7] and Olson et al. [8, 9]

In what follows we summarize definitions of the persistence length and show how the persistence length is computed using different models.

C.1 Persistence Length Definition

In general there are two parallel perspectives used to define the persistence length of a polymer. One of these has a more intrinsically mathematical, or physical flavor, in which the persistence length is understood as the resistance to deformation of a curve in space (a mathematical object), or a thin rod (a physical object). The other is a stochastic definition, where the persistence length is understood as "a measure of the distance over which the direction of the polymer is maintained" [10]. In both cases

the persistence length can be understood as a measure of polymer stiffness.

Within the context of the "mathematical physics" definition we cite Marko's and Nelson's definitions:

> "Classical elasticity tells us that a thin, straight rod that is bent into an arc has a bending energy $E = Bl/2R^2$, where B is the bending elastic constant of the rod, l is the length of the rod and R is the radius of arc. Setting $R = l$ gives us the energy of a 1 radian bend along the rod, and solving for when $E \sim k_BT$ gives us the length of rod along which a thermally excited bend of 1 radian typically occurs: $l \sim B/k_BT$. This is called the persistence length..." John F. Marko and Simona Cocco, Physics World, March 2003

> "In the elastic rod model of a polymer, the elastic energy of a short segment of rod is $dE = \frac{1}{2}k_BT[A\beta^2 + Bu^2 + C\omega^2 + 2Du\omega]ds$. Here Ak_BT, Ck_BT, Bk_BT, and Dk_BT are the bend stiffness, twist stiffness, stretch stiffness, and twist-stretch coupling, and ds is the length of the segment. (The quantities A and C are also called the bend and twist persistence lengths.)" Philip Nelson, Biological Physics: Energy, Information, Life, 2004.

These definitions, coming from the point of view of physicists, refer to an ideal, continuous thin rod. Marko refers to the bend-persistence length, whereas Nelson's definition uses a more general notion of persistence length, which includes, besides the bend persistence length, twist and stretch persistence lengths.

From the "stochastic" perspective we cite Flory's definition:

> Persistence length is "the average sum of the projections of all bonds $j \geq i$ on bond i in an indefinetely long chain. The bond i is taken to be remote from either end of the chain, i.e., $1 \ll i \ll n$". Paul J. Flory, Statistical Mechanics of Chain Molecules. 1969

This second perspective is more familiar to the chemist, since it assumes some type of bonded connectivity between polymeric units, where the bond can be either "real" or "virtual".

Both perspectives are analogs of one another since the latter definition can be made to appear like the former in the limit where the length of the bonds connecting the monomeric units is zero, and the number of monomers reaches a very large number, formally, infinity.

When talking about persistence length it is usually difficult to have a good idea of the meaning of the quantity by itself. That is, if we are told that the persistence length of DNA is, say, 530 Å under some specific concentration and temperature conditions, we don't know what this is telling us about its stiffness since we don't know any other standard values that can be related to this quantity. To give a better understanding of the meaning of the values of the persistence length, we have collected in Table C.1 values of the persistence length for various filamentous biopolymers. Inspection of the table makes it clear that B-DNA is quite a stiff biopolymer if compared to poly-glycine, or poly-alanine, but flexible if compared to a neurofilament of F-actin.

Polymer	a (nm)	Citation
Polymethylene	0.6	Flory[a]
Polystyrene	0.9	Flory[a]
Polyglycine	0.6	Flory [b]
Poly-L-alanine	2	Flory [b]
Poly-L-proline	22	Cantor and Schimel [11]
B-DNA	53	Rivetti [12]
A-RNA	62-64	Abels [13]
α-helix	80-100	Lakkaraju [14]
Coiled-coil	150-300	Lakkaraju [14]
Neurofilament	500	Nelson [2]
Intermediate filament	1000	Lakkaraju [14]
F-actin	17000	Lakkaraju [14]
Microtubule	5200000	Lakkaraju [14]

[a] Computed using characteristic ratios C_∞ reported in Table 1 of Flory's book [7] using a C-C bond length $\nu = 1.54$ Å. The derivation and formula for C_∞ can be found in section C.2.

[b] Computed using characteristic ratios reported in Table 3 of Flory's book [7] using a virtual bond length $\nu = 3.80$ Å between sequential c-alpha atoms Cα_i-Cα_{i+1}.

Table C.1: Persistence lengths for common polymers, and biopolymers with filament structures.

There are other definitions of persistence length which arise when one wants to take into account the electrostatic nature of polyelectrolytes. For example; Skolnick and Fixman [15] proposed an electrostatic persistence length as a result of an extension of the so-called Porod-Kratky chain to include charges. Another definition comes from Manning, who proposed an ideal case where the charge of a polyelectrolyte (DNA) would be completely neutralized[i] and the persistence length for such neutralized molecule is termed the null persistence length [16]. Yet another definition of DNA persistence length is that of Trifonov et al. [17], who approximate the observed persistence length as a sum of "static" and "dynamic" components. The "static" components come from the static bends like those produced by phased A-tracts in DNA, and the "dynamic" components from the fluctuation of the chain.

C.2 Freely Jointed Chain (FJC)

The end-to-end vector $\mathbf{r}$[ii] is the vector which connects the ends of a polymer chain and is defined as the sum of the vectors connecting the monomeric units in a chain. These connecting vectors can be either

[i] Manning defines DNA* as the null charge isomer of charged DNA.

[ii] Note that we will be using bold-face letters to denote vectors.

"real-bond" vectors or "virtual-bond" vectors and are denoted by $\mathbf{l}$. The magnitude of the end-to-end vector is usually the quantity of interest as given by equation C.2.

$$\mathbf{r} = \sum_{i=1}^{n} \mathbf{l}_i, \tag{C.1}$$

$$r = \sqrt{\mathbf{r} \cdot \mathbf{r}} = \sqrt{\sum_{i,j} \mathbf{l}_i \cdot \mathbf{l}_j}. \tag{C.2}$$

To distinguish between the scalar product of bond vectors with themselves and with all other bond vectors equation C.2 is rewritten as:

$$r^2 = \sum_{i=1}^{n} l_i^2 + 2 \sum_{i \neq j}^{n} \mathbf{l}_i \cdot \mathbf{l}_j \quad . \tag{C.3}$$

This expression takes into account only a single chain conformation. In order to relate the properties of a polymer to its chemical structure, it is necessary to think about the various conformations the chain can adopt due to its flexibility. Therefore, it is important to think of polymer-related quantities in terms of the average properties over all conformations (i.e., an ensemble). The average of the ensemble of end-to-end vectors is denoted by $\langle \mathbf{r} \rangle$, and the average of its squared value, also known as the second moment of the end-to-end distribution or the mean-square end-to-end distance is denoted by $\langle r^2 \rangle$ and is given by the expression:

$$\langle r^2 \rangle = \sum_{i}^{n} \langle l_i^2 \rangle + 2 \sum_{i<j}^{n} \langle \mathbf{l}_i \cdot \mathbf{l}_j \rangle \, . \tag{C.4}$$

When there is no correlation between bonds the average scalar products within the summation vanish.

$$\langle \mathbf{l}_i \cdot \mathbf{l}_j \rangle = 0. \tag{C.5}$$

A chemical interpretation of this is that every bond is allowed to rotate freely around its immediate

neighbors. What is meant precisely by “freely”, is that bond rotation (torsion) angles can randomly assume any value between 0 and 360 degrees, and that there are no bond-angle (valence-angle) constraints whatsoever. When the number of conformations approaches infinity the average cosine between all bond angles will be zero and therefore the scalar product is also zero. Equation C.4 keeps only the bond auto-correlation term under the condition that:

$$\left\langle r^2 \right\rangle = \sum_{i=1}^{n} \left\langle l_i^2 \right\rangle = nl^2. \tag{C.6}$$

This equation is used to describe a so-called freely-jointed chain (FJC), which also corresponds to a 3D random walk.

A quantity that can be used to check if a chain behaves as a FJC is the characteristic ratio:

$$C_n = \frac{\left\langle r^2 \right\rangle}{nl^2}, \tag{C.7}$$

which will be unity for a FJC. In most cases as $n \to \infty$ the characteristic ratio is greater than one.

C.3 Realistic chains

In general there are correlations between neighboring bonds in polymer chains and so the second term in the right-hand side of equation C.4 has to be taken into account. Using equation C.6 and expanding this term we have:

$$< r^2 >= nl^2 + 2\sum_{j=2}^{n} \langle \mathbf{l}_1 \cdot \mathbf{l}_j \rangle + 2 \sum_{j=i+1}^{n} \sum_{i=2}^{n} \langle \mathbf{l}_i \cdot \mathbf{l}_j \rangle \,. \tag{C.8}$$

Consideration of the correlations between the n bonds of a chain and its intial direction the persistence length was defined by Porod and Kratky and later clarified by Flory. Thus the average sum of projections of n bonds on the direction of the first bond when $n \to \infty$ yields

$$a = lim_{n->\infty} \sum_{i=1}^{n} \langle \mathbf{l}_i \cdot (\mathbf{l}_1/l_1) \rangle \,. \tag{C.9}$$

Then, as $n \to \infty$ the sum in the second term in the right hand side of equation C.8 can be rewritten as:

$$\sum_{j=1}^{n} \langle \mathbf{l}_j \cdot l(\mathbf{l}_1/l_1) \rangle - (\mathbf{l}_1 \cdot \mathbf{l}_1) = al - l^2 \tag{C.10}$$

In a similar fashion one can see that the sum in the third term in C.8 as $n \to \infty$ also becomes $al - l^2$. Then the expression for the second moment of the mean square end-to-end distance in the limit $n \to \infty$ will be given by:

$$\left\langle r^2 \right\rangle = nl^2 + 2n(al - l^2), \tag{C.11}$$

$$\left\langle r^2 \right\rangle = 2nla - nl^2, \tag{C.12}$$

and the limiting characteristic ratio for the chain is then:

$$C_\infty = \frac{\left\langle r^2 \right\rangle}{nl^2}, \tag{C.13}$$

$$= \frac{2nla - nl^2}{nl^2}, \tag{C.14}$$

$$= \frac{2a}{l} - 1. \tag{C.15}$$

C.4 Porod-Kratky or Worm Like Chain (WLC)

If a polymer is modelled as a linear (un-branched) continuous and homogenous chain its curvature at an arclength s is given by the tangential unit vector $\hat{\mathbf{t}}$[iii];

[iii] Note that we will be using from now on the hat ($\hat{\mathbf{t}}$) notation to denote unit vectors.

$$\hat{\mathbf{t}}(s) = \frac{\partial \mathbf{r}(s)}{\partial s}. \tag{C.16}$$

Here $\mathbf{r}(s)$ is the position vector of a point in the chain with respect to the coordinate origin. The end-to-end vector is then defined as the integral sum of the chain curvature over the contour length of the chain:

$$\mathbf{r} = \int_0^L \hat{\mathbf{t}}(s) ds. \tag{C.17}$$

The average mean squared end-to-end distance is then:

$$\begin{aligned}
\left\langle r^2 \right\rangle &= \langle \mathbf{r} \cdot \mathbf{r} \rangle, \\
&= \left\langle \int_0^L \hat{\mathbf{t}}(s) ds \cdot \int_0^L \hat{\mathbf{t}}(s') ds' \right\rangle, \\
&= \int_0^L ds \int_0^L \left\langle \hat{\mathbf{t}}(s) \cdot \hat{\mathbf{t}}(s') \right\rangle ds', \\
&= \int_0^L ds \int_0^L \langle \cos\theta_{ss'} \rangle \; ds', \\
&= \int_0^L ds \int_0^L \exp^{-\frac{|s-s'|}{a}} \, ds', \\
&= 2aL\{1 - \frac{a}{L}(1 - \exp^{-\frac{L}{a}})\},
\end{aligned} \tag{C.18}$$

where for a chain of infinite length $\langle \cos\theta_{ss'} \rangle = \exp^{-\frac{|s-s'|}{a}}$, with a being the persistence length. The last equation in C.18 is used to describe a so-called worm-like-chain (WLC).

When $L \to \infty$ we see that $\left\langle r^2 \right\rangle = 2aL$, and in such case the characteristic ratio for a WLC becomes:

$$C_\infty = \frac{2aL}{nl^2}. \tag{C.19}$$

C.5 Sequence Dependent Model

Equation C.1 can be expressed with respect to the reference frame of the first bond, along the bond-vector, via a series of matrix transformations:

$$\begin{aligned} \mathbf{r} = \mathbf{l}_1 \quad &+ \quad \mathbf{T}_{12}\mathbf{l}_2 + \mathbf{T}_{12}\mathbf{T}_{23}\mathbf{l}_3 \\ &+ \quad ... + \mathbf{T}_{12}\mathbf{T}_{23}...\mathbf{T}_{N-1,N}\mathbf{l}_N. \end{aligned} \tag{C.20}$$

Here the $\mathbf{T}_{N-1,N}$ are matrices that transform the vectors $\mathbf{l}_N$ in coordinate frame N to their representation in coordinate frame N^{-1}. Following Flory [7] equation C.20 can be written as a matrix product in the following way:

$$\mathbf{r} = \begin{bmatrix} \mathbf{E}_3 \ \mathbf{0} \end{bmatrix} \mathbf{A}_{1:N} \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}. \tag{C.21}$$

Here $\mathbf{A}_{1:N}$ is a serial product of generator matrices $\mathbf{A}_n$ associated with consecutive bonds along the chain [7, 8, 9], $\mathbf{E}_3$ is the identity matrix of order 3, and $\mathbf{0}$ is a vector of necessary dimensions for the matrix products to conform.

$$\mathbf{A}_n = \begin{bmatrix} \mathbf{T}_n & \mathbf{l}_n, \\ 0 & 1 \end{bmatrix} \tag{C.22}$$

$$\mathbf{A}_{1:N} = \mathbf{A}_1\mathbf{A}_2...\mathbf{A}_N. \tag{C.23}$$

The generator matrices are set up to perform the coordinate transformations in C.20. The sequence model comes into play in the product of the generator matrices since every generator matrix contains the specific sequence-dependent geometric information associated with the rigid block model for nucleic acids. For complete details of the transformation of step parameters into Euclidean space, and how the polymer configuration space is sampled using the Gaussian sampling technique, we refer the reader to Appendix A of Luke Czapla's doctoral thesis [18].

In addition to being able to express the end-to-end distance with the generator matrices we can

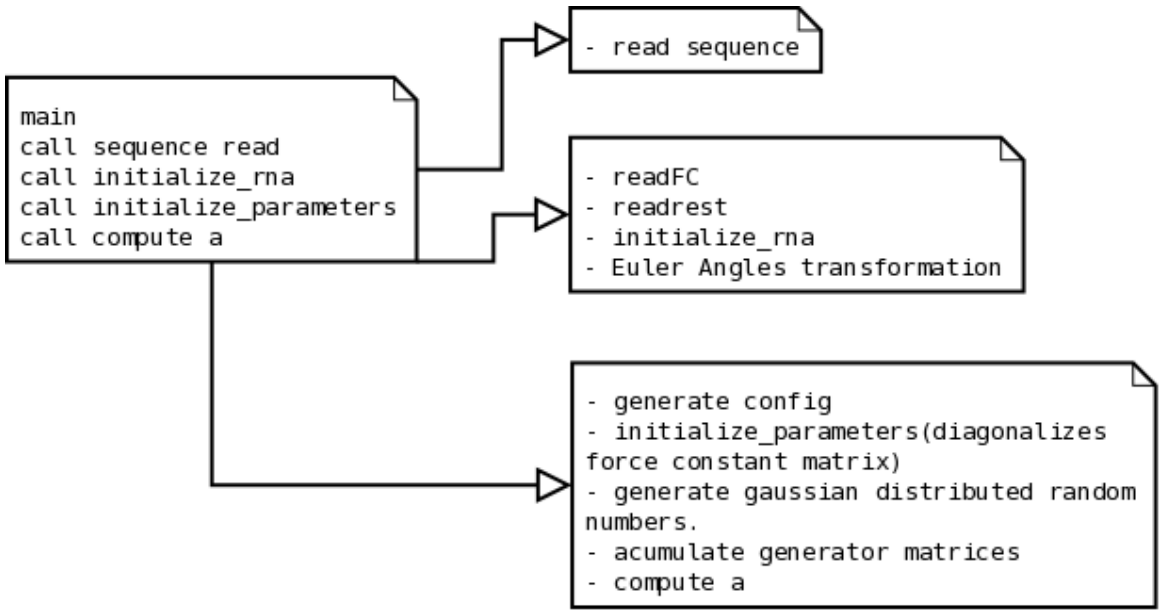


Figure C.1: A condensed diagram of the main functions used to implement a C++ program using the sequence-dependent model based on Gaussian sampling of the known space of base-pair-step parameters. The program is used for calculating persistence lengths only, but it can be easily adapted to compute other global chain properties such as the average end-to-end distance and the global bend and twist angle for the sampled ensemble (e.g. see Maroun and Olson [19]).

also find the persistence length as the component in row three and column four of the average product of generator matrices [iv], that is, the component of the translation vector along the normal of the first base-pair (the assumed initial direction of the chain). Sampling has been performed for a large number of conformations such that the terms in the average rotation matrix $\mathbf{T}$ approach very small numbers.

$$\mathbf{P}_N = \langle \mathbf{A}_1 \rangle \langle \mathbf{A}_2 \rangle \ldots \langle \mathbf{A}_{N-1} \rangle \langle \mathbf{A}_N \rangle . \tag{C.24}$$

$$a = \lim_{N\to\infty} \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} \mathbf{P}_N \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} . \tag{C.25}$$

A sketch of the algorithm used to implement the sequence dependent method in C++ by Luke Czapla is shown in Figure C.1.

[iv] If we assume that the arrangements of succesive base-pair steps are independent of all other residues in the chain, we can replace the average product by the product of average generator matrices.

References

[1] Marko, J. F. and Cocco, S. (2003) The Micromechanics of DNA. *Physics World,* **16**, 37–41.

[2] Nelson, P. (2004) Biological Physics: Energy, Information, Life, W. H. Freeman and Company, .

[3] Olson, W. K., Marky, N. L., Jernigan, R. L., and Zhurkin, V. B. (1993) Influence of Fluctuations on DNA Curvature. A Comparison of Flexible and Static Wedge Models of Intrinsically Bent DNA. *Journal of Molecular Biology,* **232**, 530–554.

[4] Go, M. and Go, N. (1976) Fluctuations of an Alpha-Helix. *Biopolymers,* **15**, 1119–1127.

[5] Olson, W. K., Gorin, A. A., Lu, X.-J., Hock, L. M., and Zhurkin, V. B. (1998) DNA Sequence-Dependent Deformability Deduced from Protein-DNA Crystal Complexes. *Proceedings of the National Academy of Sciences,* **95**, 11163–11168.

[6] Czapla, L., Swigon, D., and Olson, W. K. (2006) Sequence-Dependent Effects in the Cyclization of Short DNA. *Journal of Chemical Theory and Computation,* **2**, 685–695.

[7] Flory, P. J. (1969) Statistical Mechanics of Chain Molecules, Interscience Publishers, .

[8] Maroun, R. C. and Olson, W. K. (1988) Base Sequence Effects in Double-Helical DNA. II. Configurational Statistics of Rodlike Chains. *Biopolymers,* **27**, 561–584.

[9] Marky, N. L. and Olson, W. K. (1994) Configurational Statistics of the DNA Duplex: Extended Generator Matrices to Treat the Rotations and Translations of Adjacent Residues. *Biopolymers,* **34**, 109–120.

[10] Kratky, O. and Porod, G. (1949) X-Ray Investigation of Dissolved Chain Molecules. *Recueil des Travaux Chimiques des Pays-Bas et de la Belgique,* **68**, 1106–1122.

[11] Cantor, C. R. and Schimmel, P. R. (1980) Biophysical Chemistry, W. H. Freeman and Company, .

[12] Rivetti, C., Guthold, M., and Bustamante, C. (1996) Scanning Force Microscopy of DNA Deposited onto Mica: Equilibration Versus Kinetic Trapping Studied by Statistical Polymer Chain Analysis. *Journal of Molecular Biology,* **264**, 919–932.

[13] Abels, J. A., Moreno-Herrero, F., van derHeijden, T., Dekker, C., and Dekker, N. H. (2005) Single-Molecule Measurements of the Persistence Length of Double-Stranded RNA. *Biophysical Journal,* **88**, 2737–2744.

[14] Lakkaraju, S. K. and Hwang, W. (2009) Critical Buckling Length Versus Persistence Length: What Governs Biofilament Conformation?. *Physical Review Letters,* **102**, 118102.

[15] Skolnick, J. and Fixman, M. (1977) Electrostatic Persistence Length of a Wormlike Polyelectrolyte. *Macromolecules,* **10**, 944–948.

[16] Manning, G. S. (2006) The Persistence Length of DNA is Reached from the Persistence Length of its Null Isomer Through an Internal Electrostatic Stretching Force. *Biophysical Journal,* **91**, 3607–3616.

[17] Trifonov, E. N., Tan, R. K., and Harvey, S. C. DNA Bending Curvature p. 243 Academic Press (1987).

[18] Czapla, L. The Statistical Mechanics of Free and Protein-Bound DNA by Monte Carlo Simulation PhD thesis Rutgers, The State University of New Jersey (2009).

[19] Maroun, R. C. and Olson, W. K. (1988) Base Sequence Effects in Double-Helical DNA. III. Average Properties of Curved DNA. *Biopolymers,* **27**, 585–603.

Appendix D

getMotif Results

Table showing the 211 GNRA motif candidates identified by the getMotif program in the list of non-redundant RNA structures provided by the RNA Ontology Consortium (ROC) at https://docs.google.com/Doc?id=dhrmkfmn_13ftpbjcgq:

Table D.1: GNRA Motif Scores.

pdb_id	resnum	3dna_id	step1	shift	slide	rise	tilt	roll	twist	exo	endo	score
1g1x	591	10	G/A	-8.15	0.14	-7.60	83.1	127.0	-23.8	0.00	0.00	86
1g1x	727	52	G/A	-10.61	-2.32	-4.03	76.1	103.1	-67.9	0.00	0.00	119
1gid	150	48	G/A	-8.56	-0.16	-8.01	88.5	126.6	-31.5	0.00	0.00	92
1hmh	21L	19	G/A	-10.22	-1.24	-6.21	94.3	113.6	-59.2	0.00	0.00	68
1j5e	159	149	G/A	-8.59	-0.86	-7.73	82.0	107.0	-24.5	0.00	0.00	102
1j5e	297	288	G/A	-10.98	-2.96	-3.89	84.7	125.0	-81.3	0.00	0.00	61
1j5e	324	315	G/A	-9.76	-2.45	-5.09	67.4	128.5	-56.3	0.00	0.00	62
1j5e	727	706	G/A	-9.44	-1.68	-5.19	80.3	115.1	-48.7	0.00	0.00	103
1j5e	1013	987	G/A	-11.42	-1.84	-6.00	98.9	84.7	-67.4	0.00	0.00	128
1j5e	1166	1144	G/A	-9.47	-0.82	-6.02	85.7	119.5	-55.4	0.00	0.00	50
1j5e	1178	1155	G/A	-9.37	-1.44	-5.85	72.5	104.6	-42.6	0.00	0.00	93
1kxk	34	34	G/A	-8.47	0.23	-7.40	85.8	129.1	-35.9	0.00	0.00	90
1m5k	75	96	G/A	-9.30	-1.93	-4.82	60.0	128.3	-58.4	0.06	0.00	80
1mfq	198	87	G/A	-9.87	-1.06	-4.32	70.0	127.8	-60.2	0.55	0.00	64
1mzp	39	39	G/A	-7.40	0.83	-8.10	80.8	120.2	-13.6	0.00	0.00	115
1q93	14	13	G/A	-10.21	-1.44	-4.95	82.8	119.0	-61.9	0.00	0.00	35
1q96	14	13	G/A	-10.53	-1.84	-4.60	81.2	117.7	-68.2	0.00	0.00	42
1rlg	10	10	G/A	-8.69	-0.16	-7.60	85.9	126.9	-29.1	0.00	0.00	80
1s72	506	495	G/A	-8.33	-0.13	-7.21	76.9	122.2	-27.1	0.00	0.00	98
1s72	691	680	G/A	-10.76	-2.28	-3.24	73.4	109.7	-74.5	0.01	0.00	62
1s72	805	793	G/A	-9.00	-0.37	-7.27	87.4	110.9	-31.9	0.00	0.00	80
1s72	1389	1349	G/A	-9.51	-1.48	-6.63	74.4	120.5	-44.4	0.00	0.00	74
1s72	1629	1588	G/A	-9.09	-0.37	-7.49	93.6	119.6	-34.7	0.00	0.00	81
1s72	2412	2254	G/A	-9.56	-1.82	-5.33	70.2	128.1	-54.9	0.00	0.00	48
1s72	2696	2536	G/A	-9.60	-1.52	-6.88	82.8	129.8	-44.5	0.00	0.00	66
1u6b	24	21	G/A	-7.70	0.88	-8.41	85.9	134.3	-17.4	0.00	0.00	108
1u6b	189	196	G/a	-8.41	-0.26	-6.90	80.1	133.6	-45.1	0.00	0.00	78

Continued on Next Page. . .

Table D.1 – Continued

pdb_id	resnum	3dna_id	step1	shift	slide	rise	tilt	roll	twist	exo	endo	score
1u9s	171	92	G/A	-7.82	0.34	-6.53	75.4	123.1	-12.7	0.24	0.00	93
1un6	83	36	G/A	-9.81	-1.68	-6.22	75.2	115.1	-42.3	0.00	0.00	64
1x8w	150	55	G/A	-9.75	-2.31	-5.83	91.0	98.9	-48.6	0.00	0.00	95
1x8w	323	222	G/A	-10.81	-3.73	-1.40	72.7	111.6	-88.1	0.00	0.00	109
1y0q	131	108	G/A	-8.63	-0.70	-6.37	71.1	124.8	-36.5	0.00	0.00	70
1y0q	205	182	G/A	-7.64	0.18	-8.28	76.6	127.4	-22.2	0.00	0.00	119
2a2e	115	100	G/A	-7.10	0.09	-8.93	72.6	118.8	-9.2	0.00	0.00	120
2avy	159	155	G/A	-9.67	-2.10	-5.98	80.1	104.6	-40.3	0.00	0.00	86
2avy	297	293	G/A	-11.21	-2.35	-4.24	83.3	116.5	-79.4	0.00	0.00	63
2avy	324	320	G/A	-10.00	-2.03	-4.45	69.3	125.1	-63.4	0.05	0.00	64
2avy	727	723	G/A	-10.27	-2.36	-4.59	75.2	112.1	-60.6	0.00	0.00	121
2avy	1013	1009	G/A	-10.50	-4.00	-2.14	71.3	110.1	-72.2	0.00	0.00	106
2avy	1178	1174	G/A	-9.21	-1.06	-6.63	66.1	106.4	-47.9	0.00	0.00	142
2aw4	476	593	G/A	-8.07	-1.16	-7.08	70.6	120.3	-29.3	0.00	0.00	85
2aw4	500	617	G/A	-11.07	-2.79	-1.69	68.9	117.2	-104.4	0.05	0.00	133
2aw4	630	747	G/A	-9.63	-1.81	-5.13	69.2	129.4	-58.6	0.02	0.00	30
2aw4	1283	1381	G/A	-11.00	-2.55	-4.86	77.4	111.0	-74.3	0.00	0.00	134
2aw4	1364	1462	G/A	-9.06	-3.93	-6.32	35.3	146.4	-28.9	0.00	0.00	142
2aw4	1753	1851	G/A	-8.38	-0.53	-6.60	68.5	126.3	-35.9	0.00	0.00	148
2aw4	2357	2412	G/A	-7.83	-0.31	-7.68	55.6	151.4	4.8	0.00	0.00	140
2aw4	2375	2430	G/A	-7.88	-0.05	-6.38	61.7	130.4	-13.6	0.15	0.00	119
2aw4	2659	2714	G/A	-10.68	-3.35	-4.95	75.6	115.0	-69.1	0.00	0.00	99
2gdi	40	31	G/A	-7.15	0.91	-7.33	66.1	130.0	-7.0	0.00	0.00	124
2gis	50	50	G/A	-8.74	0.30	-5.88	74.5	124.5	-43.9	0.10	0.00	77
2gis	74	74	G/A	-7.56	1.61	-8.60	96.8	111.7	-12.4	0.00	0.00	120
2il9	6119A	108	G/A	-7.82	0.41	-8.11	89.0	122.6	-17.1	0.00	0.00	103
2il9	6084B	210	G/A	-8.28	-0.05	-8.08	89.2	119.2	-19.1	0.00	0.00	128
2il9	6119A	240	G/A	-9.32	-0.39	-6.93	83.3	116.2	-40.3	0.00	0.00	62
2j01	476	491	G/A	-9.43	-1.22	-5.68	76.7	122.8	-56.3	0.00	0.00	60
2j01	500	514	G/A	-8.20	-0.48	-7.03	82.5	109.7	-21.9	0.00	0.00	120
2j01	630	642	G/A	-9.93	-1.78	-4.63	70.7	119.8	-54.5	0.06	0.00	35
2j01	1283	1233	G/A	-9.32	-1.92	-6.87	69.5	120.2	-37.2	0.00	0.00	71
2j01	1364	1314	G/A	-9.53	-4.18	-4.84	39.3	151.5	-44.1	0.00	0.00	134
2j01	2375	2253	G/A	-9.44	-1.53	-5.57	72.3	122.4	-49.2	0.00	0.00	40
2oiu	11	11	G/A	-9.94	-1.41	-5.90	82.1	119.2	-53.0	0.00	0.00	54
2oiu	29	29	G/A	-9.94	-2.07	-5.00	82.2	126.0	-56.8	0.00	0.00	41
2oiu	60	60	G/A	-9.84	-1.76	-5.64	76.1	126.4	-60.7	0.00	0.00	68
2pxb	154	25	G/A	-7.76	1.53	-8.06	89.9	126.9	-10.2	0.00	0.00	122
2r8s	150	49	G/A	-8.27	0.17	-8.50	89.1	119.4	-17.9	0.00	0.00	97
2zjr	487	409	G/A	-8.16	-0.05	-6.33	63.5	131.4	-35.4	0.00	0.00	85

Continued on Next Page. . .

Table D.1 – Continued

pdb_id	resnum	3dna_id	step1	shift	slide	rise	tilt	roll	twist	exo	endo	score
2zjr	510	432	G/A	-8.79	-0.81	-5.96	79.8	126.3	-45.7	0.05	0.00	74
2zjr	641	563	G/A	-9.47	-2.60	-6.17	66.6	117.6	-45.0	0.00	0.00	66
2zjr	1296	1199	G/A	-8.34	0.24	-8.38	76.4	126.7	-19.3	0.00	0.00	106
2zjr	2336	2145	G/A	-8.95	-1.04	-7.38	76.3	132.9	-27.0	0.00	0.00	88
2zjr	2354	2163	G/A	-9.05	-1.66	-5.25	67.2	132.8	-50.3	0.08	0.00	80
2zjr	2638	2447	G/A	-9.53	-1.55	-5.94	71.6	126.4	-43.8	0.00	0.00	58
3d0u	143	143	G/A	-11.09	-1.15	-6.56	108.6	106.6	-63.5	0.00	0.00	98
3e5c	19	19	G/A	-9.71	-1.78	-4.78	78.0	130.9	-49.7	0.14	0.00	44
3e5c	42	42	G/A	-9.40	-1.49	-6.38	83.9	126.7	-46.5	0.00	0.00	77
3f4e	19	19	G/A	-9.48	-0.96	-5.83	79.8	111.4	-50.6	0.00	0.00	87
3f4e	70	67	G/A	-6.63	1.40	-8.03	65.7	127.1	3.4	0.00	0.00	129
430d	14	14	G/A	-9.66	-1.25	-6.11	81.4	118.4	-48.2	0.00	0.00	48
1j5e	898	871	G/C	-10.87	-2.26	-4.82	90.1	111.1	-68.7	0.00	0.00	77
1j5e	1030A	1005	G/C	-8.89	-1.00	-8.68	91.7	109.6	-15.0	0.00	0.00	119
1j5e	1316	1293	G/C	-10.63	-4.62	-2.84	51.6	128.2	-76.1	0.00	0.00	112
1mzp	26	26	G/C	-10.02	-1.11	-5.80	86.0	116.9	-53.3	0.00	0.00	45
1s72	482	471	G/C	-8.85	-0.97	-6.61	74.1	128.4	-36.1	0.00	0.00	102
1s72	577	566	G/C	-11.74	-3.96	-0.71	69.1	119.8	-110.1	0.00	0.00	127
1s72	1809	1768	G/C	-8.16	0.17	-6.59	69.4	125.3	-26.1	0.09	0.00	137
1s72	1863	1822	G/C	-10.58	-1.85	-5.04	89.9	112.6	-62.1	0.00	0.00	54
1s72	90	2844	G/C	-10.53	-3.04	-3.92	65.7	124.9	-64.5	0.00	0.00	38
1u9s	205	126	G/C	-11.47	-2.97	-3.58	83.1	121.3	-93.0	0.00	0.00	83
2avy	187	183	G/C	-7.71	1.08	-7.79	79.7	141.2	6.5	0.01	0.00	139
2avy	898	894	G/C	-10.93	-2.70	-4.32	78.9	110.8	-74.4	0.00	0.00	72
2avy	1266	1262	G/C	-11.26	-3.32	-2.09	65.1	119.5	-98.6	0.00	0.00	123
2aw4	2857	2912	G/C	-11.39	-3.28	-0.01	69.0	119.3	-115.8	0.05	0.00	124
2ez6	12	12	G/C	-10.52	-2.60	-4.28	73.6	125.8	-68.3	0.00	0.00	41
2ez6	12	40	G/C	-11.33	-2.55	-3.17	83.2	108.2	-81.8	0.00	0.00	79
2j01	1223	1173	G/C	-11.29	-3.57	-0.91	67.3	132.9	-120.1	0.00	0.00	140
2j01	1753	1679	G/C	-8.63	-1.45	-6.81	68.6	127.2	-31.5	0.00	0.00	141
2j01	1865	1791	G/C	-9.65	-2.64	-3.94	56.3	133.7	-72.9	0.23	0.00	112
2j01	87	2859	G/C	-11.04	-3.99	-1.80	51.4	117.9	-99.3	0.00	0.00	106
2nuf	12	12	G/C	-10.55	-2.37	-4.32	75.6	118.0	-70.1	0.00	0.00	54
2nuf	12	40	G/C	-11.52	-2.98	-3.08	81.8	110.0	-86.7	0.00	0.00	86
2zjr	1664	1567	G/C	-5.94	-10.43	-12.53	96.2	136.7	-67.5	0.00	0.00	117
2zjr	1857	1760	G/C	-8.38	-0.92	-6.53	71.9	123.6	-37.0	0.00	0.00	73
2zjr	2832	2641	G/C	-12.17	-3.29	-0.14	60.8	110.1	-118.0	0.00	0.00	139
3gca	8	8	G/C	-8.34	-1.53	-5.42	50.3	144.1	-31.4	0.48	0.00	99
1j5e	1516	1489	G/G	-9.81	-4.90	-4.98	35.6	130.2	-44.5	0.00	0.00	107
1jid	147	13	G/G	-10.72	-2.38	-4.58	78.3	103.5	-57.9	0.00	0.00	125

Continued on Next Page. . .

Table D.1 – Continued

pdb_id	resnum	3dna_id	step1	shift	slide	rise	tilt	roll	twist	exo	endo	score
1lng	209	70	G/G	-10.53	-2.54	-4.03	74.2	126.1	-76.5	0.00	0.00	56
1nbs	219	99	G/G	-7.98	-0.23	-7.50	82.9	122.6	-31.6	0.00	0.00	144
1s72	314	303	G/G	-8.07	0.31	-7.77	80.0	122.6	-22.9	0.00	0.00	111
1s72	1137	1097	G/G	-12.84	-2.55	-8.75	72.2	56.9	-46.2	0.00	0.00	128
1s72	1794	1753	G/G	-8.02	-0.03	-6.47	66.5	127.0	-22.2	0.07	0.00	96
1s72	2249	2096	G/G	-8.87	-0.35	-7.06	79.5	119.9	-32.0	0.00	0.00	70
2a2e	146	121	G/G	-8.29	-0.13	-7.68	75.8	116.5	-32.8	0.00	0.00	138
2avy	1516	1512	G/G	-9.67	-5.72	-4.20	32.4	130.7	-57.4	0.00	0.00	121
2aw4	307	424	G/G	-9.52	-1.78	-5.20	76.3	132.5	-54.7	0.03	0.00	63
2j01	307	320	G/G	-10.09	-2.03	-5.26	75.5	121.8	-68.7	0.00	0.00	99
2j01	488	503	G/G	-9.36	-0.20	-5.84	84.2	127.7	-48.3	0.16	0.00	71
2v3c	209	68	G/G	-10.54	-1.70	-4.51	95.8	121.2	-80.7	0.00	0.00	78
2zjr	318	265	G/G	-7.76	0.79	-8.13	84.5	119.4	-25.9	0.00	0.00	124
2zjr	499	421	G/G	-8.94	0.03	-6.74	84.0	110.0	-30.9	0.00	0.00	99
2zjr	1236	1139	G/G	-10.93	-2.50	-5.27	83.9	112.9	-63.6	0.00	0.00	127
1j5e	1077	1055	G/U	-11.24	-3.58	-2.08	58.2	121.5	-88.4	0.02	0.00	121
1msy	2659	13	G/U	-12.30	-4.45	1.56	79.5	127.1	-146.9	0.00	0.00	138
1nbs	188	80	G/U	-8.78	0.37	-7.15	75.5	106.9	-33.2	0.00	0.00	102
1s72	469	458	G/U	-10.75	-3.85	-2.10	64.1	135.5	-94.6	0.00	0.00	113
1s72	1055	1015	G/U	-9.69	-1.70	-5.69	74.0	132.8	-47.9	0.00	0.00	50
1s72	2630	2472	G/U	-11.10	-3.58	-0.89	60.5	124.1	-104.1	0.19	0.00	123
1s72	2877	2717	G/U	-10.85	-3.38	-3.21	76.4	125.1	-80.4	0.00	0.00	58
1u9s	100	25	G/U	-10.90	-1.79	-4.92	89.7	114.9	-72.7	0.00	0.00	87
1y0q	102	79	G/U	-9.04	-2.81	-6.94	69.9	142.7	-41.7	0.00	0.00	85
299d	31L	28	G/U	-10.69	-3.34	-3.38	65.6	137.2	-93.8	0.00	0.00	104
2avy	1077	1073	G/U	-10.14	-3.39	-4.62	62.6	118.2	-62.3	0.00	0.00	71
2aw4	463	580	G/U	-10.58	-2.93	-3.95	63.7	126.7	-73.9	0.00	0.00	72
2aw4	1223	1321	G/U	-11.21	-3.77	-2.82	64.6	122.6	-98.4	0.00	0.00	109
2aw4	2595	2650	G/U	-10.16	-2.39	-4.67	65.7	125.3	-58.4	0.00	0.00	63
2j01	463	478	G/U	-10.05	-2.39	-4.71	73.0	137.4	-64.0	0.00	0.00	71
2j01	2595	2473	G/U	-12.04	-3.16	0.21	67.8	119.6	-129.6	0.00	0.00	145
2zjr	474	396	G/U	-9.71	-2.04	-4.86	66.4	137.9	-61.1	0.08	0.00	82
2zjr	2574	2383	G/U	-11.15	-3.26	-2.36	64.1	125.4	-98.3	0.00	0.00	112
488d	31L	28	G/U	-10.60	-3.34	-1.58	60.3	147.0	-116.6	0.25	0.00	145
488d	31L	53	G/U	-10.60	-3.34	-1.58	60.3	147.0	-116.6	0.25	0.00	145
1f7v	955	55	P/C	-7.12	0.12	-7.26	68.8	132.6	0.6	0.00	0.00	139
1h4q	55	53	P/C	-9.81	-2.28	-4.46	63.8	126.0	-65.7	0.00	0.00	71
2czj	55	45	P/C	-8.42	-0.69	-5.17	47.1	146.5	-51.5	2.46	0.00	136
2fmt	55	56	P/C	-10.82	-2.13	-4.36	58.9	113.8	-70.2	0.00	0.00	123
2fmt	55	133	P/C	-9.99	-1.64	-5.69	64.2	114.5	-49.1	0.00	0.00	109

Continued on Next Page. . .

Table D.1 – Continued

pdb_id	resnum	3dna_id	step1	shift	slide	rise	tilt	roll	twist	exo	endo	score
1a9n	13	14	U/A	-10.73	-2.40	-10.13	63.6	104.5	-36.4	0.00	0.00	103
1j5e	863	836	U/A	-9.64	-0.94	-7.05	86.7	96.0	-38.3	0.00	0.00	96
1mms	1066	16	U/A	-7.63	0.16	-7.78	74.4	126.5	-18.4	0.00	0.00	145
1mms	1094	44	U/A	-8.90	-2.08	-6.09	72.2	124.7	-45.3	0.00	0.00	63
1qa6	144	44	U/A	-7.58	-1.46	-6.98	70.2	133.6	-16.7	0.00	0.00	111
1s72	1170	1130	U/A	-10.10	-3.01	-4.35	58.3	129.8	-71.9	0.00	0.00	107
1s72	1198	1158	U/A	-10.10	-2.60	-4.29	65.8	123.5	-69.2	0.00	0.00	86
1s72	1596	1555	U/A	-6.21	1.00	-7.89	66.5	113.0	12.5	0.00	0.00	148
1s72	2598	2440	U/A	-10.21	-2.93	-5.73	88.6	110.8	-60.0	0.00	0.00	102
1y0q	227	204	U/A	-9.36	0.04	-5.02	73.7	117.4	-49.2	0.42	0.00	73
1y39	144	44	U/A	-8.68	-1.43	-6.48	73.5	130.3	-40.3	0.00	0.00	78
2avy	261	257	U/A	-9.17	-2.18	-6.04	50.7	129.4	-55.2	0.00	0.00	135
2avy	863	859	U/A	-10.58	-1.95	-5.56	71.6	107.5	-61.9	0.00	0.00	56
2aw4	1066	1164	U/A	-8.21	-0.52	-7.35	68.5	121.5	-27.0	0.00	0.00	147
2aw4	1094	1192	U/A	-9.34	-2.17	-5.67	68.9	116.8	-49.6	0.00	0.00	85
2aw4	2563	2618	U/A	-9.15	-2.60	-7.80	90.8	120.6	-39.3	0.00	0.00	102
2j01	1391	1341	U/A	-10.19	-3.27	-5.60	85.9	112.4	-60.0	0.00	0.00	118
2j01	2431	2309	U/A	-6.43	1.16	-8.30	70.4	121.6	0.8	0.00	0.00	138
2j01	2563	2441	U/A	-10.19	-3.11	-5.08	74.5	117.9	-65.4	0.00	0.00	86
2zjr	1105	1008	U/A	-7.47	0.26	-7.27	76.7	113.1	-24.6	0.02	0.00	141
2zjr	2542	2351	U/A	-10.62	-3.67	-1.93	73.4	123.2	-94.5	0.00	0.00	126
1eiy	55	55	U/C	-7.47	-1.85	-4.33	84.3	154.7	-84.0	3.49	0.00	144
1ffy	55	56	U/C	-7.09	1.03	-7.45	69.8	129.1	-4.8	0.19	0.00	115
1gax	954	54	U/C	-9.51	-2.30	-4.66	62.1	130.6	-53.1	0.00	0.00	82
1j1u	556	56	U/C	-8.51	-0.60	-6.42	71.0	122.2	-32.2	0.00	0.00	92
1qrs	55	53	U/C	-6.99	1.06	-7.47	68.1	129.6	-5.1	0.21	0.00	119
1s72	253	242	U/C	-8.65	0.19	-6.45	67.1	131.7	-33.4	0.76	0.00	71
1u0b	55	53	U/C	-8.37	-0.18	-6.73	72.5	126.2	-34.0	0.24	0.00	88
1wz2	967	67	U/C	-10.21	-1.97	-3.92	67.2	123.5	-72.9	0.00	0.00	91
2azx	555	54	U/C	-7.43	0.68	-6.77	65.3	121.8	-11.6	0.34	0.00	129
2bte	55	57	U/C	-8.00	0.19	-7.34	67.1	122.4	-22.6	0.01	0.00	112
2cv1	555	54	U/C	-7.61	0.23	-7.27	70.2	131.1	-13.2	0.02	0.00	110
2d6f	955	126	U/C	-7.66	-0.25	-6.64	63.1	123.6	-17.4	0.02	0.00	123
2der	55	55	U/C	-9.31	-2.12	-3.90	68.1	130.0	-65.9	0.00	0.00	97
2du3	954	54	U/C	-10.03	-2.35	-3.93	58.7	122.8	-71.0	0.00	0.00	91
2fk6	55	35	U/C	-8.56	-0.37	-6.11	62.9	127.2	-42.7	0.75	0.00	83
2v0g	55	55	U/C	-9.52	-1.01	-6.09	76.2	107.2	-43.1	0.00	0.00	116
2zni	33	102	U/C	-9.79	-1.66	-5.07	85.9	116.7	-75.0	0.01	0.00	106
3cw5	33	34	U/C	-8.85	-2.53	-5.43	66.2	125.9	-38.2	0.00	0.00	122
3epj	55	53	U/C	-10.07	-2.55	-4.37	72.0	111.9	-63.4	0.00	0.00	131

Continued on Next Page. . .

Table D.1 – Continued

pdb_id	resnum	3dna_id	step1	shift	slide	rise	tilt	roll	twist	exo	endo	score
3epj	55	122	U/C	-9.79	-1.95	-3.90	59.8	129.8	-70.5	0.29	0.00	84
3foz	55	55	U/C	-7.04	0.98	-7.42	67.9	124.4	-5.3	0.18	0.00	119
3foz	55	126	U/C	-8.43	0.82	-7.04	72.9	120.7	-31.1	0.53	0.00	104
1e8o	112	14	U/G	-6.20	1.05	-7.29	58.6	121.9	17.9	0.00	0.00	141
1eiy	33	33	U/G	-11.59	-1.34	-4.91	93.5	124.5	-93.1	0.00	0.00	143
1j5e	692	671	U/G	-5.58	1.58	-9.21	66.9	106.8	13.8	0.00	0.00	138
1u0b	33	32	U/G	-8.07	-0.63	-7.07	51.3	153.1	-17.7	0.00	0.00	120
1yrj	2	2	U/G	-13.07	-1.45	-10.02	77.4	76.1	-51.5	0.00	0.00	149
2j01	714	723	U/G	-5.31	2.14	-8.56	66.1	120.9	22.9	0.00	0.00	147
2j01	2357	2235	U/G	-12.84	-2.86	-6.28	100.7	108.0	-81.1	0.00	0.00	147
2pn4	106	33	U/G	-15.64	-3.03	-7.75	74.6	54.0	-76.5	0.00	0.00	144
2zjr	1927	1810	U/G	-15.83	-3.61	-5.02	73.5	84.2	-94.3	0.00	0.00	134
397d	25	10	U/G	-8.13	-0.26	-12.06	63.2	95.3	-8.4	0.00	0.00	144
3bnq	2	25	U/G	-12.40	-1.86	-9.66	74.4	71.5	-48.3	0.00	0.00	147
1j5e	81	75	U/U	-11.55	-2.76	-2.76	46.2	133.3	-109.0	0.68	0.00	135
1nbs	120	35	U/U	-8.03	-0.26	-7.22	67.9	128.6	-24.9	0.00	0.00	121
1s72	625	614	U/U	-9.09	-2.09	-6.12	66.2	129.8	-41.3	0.00	0.00	114
2aw4	568	685	U/U	-10.47	-3.82	-2.68	58.7	124.7	-84.3	0.00	0.00	130
1u6b	154	161	A/A	-10.21	-2.29	-5.45	74.3	132.3	-67.5	0.00	0.00	90
1mzp	15	15	A/G	-12.81	-3.55	-8.15	54.0	95.3	-66.7	0.00	0.00	127
2j01	2119	2036	A/G	-12.90	-6.24	-5.02	60.0	120.2	-90.1	0.00	0.00	136
1j5e	1359	1336	C/A	-12.43	-1.05	-9.67	115.7	86.3	-62.5	0.00	0.00	144
1s72	1469	1429	C/A	-7.79	-0.50	-6.31	51.6	131.9	-13.4	0.03	0.00	106
2qbz	69	55	C/A	-11.36	-1.19	-6.03	80.4	93.3	-62.4	0.00	0.00	65

Curriculum Vitae

Mauricio Esguerra

Education

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, The State University of New Jersey.

Professional Experience

2000-2002 Teaching assistant, Department of Physics, Universidad Nacional de los Andes, Colombia.

2001-2003 Patent Advisor for Alvaro Castellanos y Cia.

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, **47**, 177-186.