RNA STRUCTURE ANALYSIS VIA THE RIGID BLOCK MODEL

by

MAURICIO ESGUERRA NEIRA

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Chemistry and Chemical Biology

Written under the direction of

Wilma K. Olson

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2010

ABSTRACT OF THE DISSERTATION

RNA Structure Analysis via the Rigid Block Model

by Mauricio Esguerra Neira

Dissertation Director: Wilma K. Olson

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

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

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

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

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

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

Ludwig Wittgenstein

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

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

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

Anonymous Chemist

Acknowledgements

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

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

Table of Contents

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 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 poly-nucleotide 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 uracyl (U) as shown in Figure 1.1.

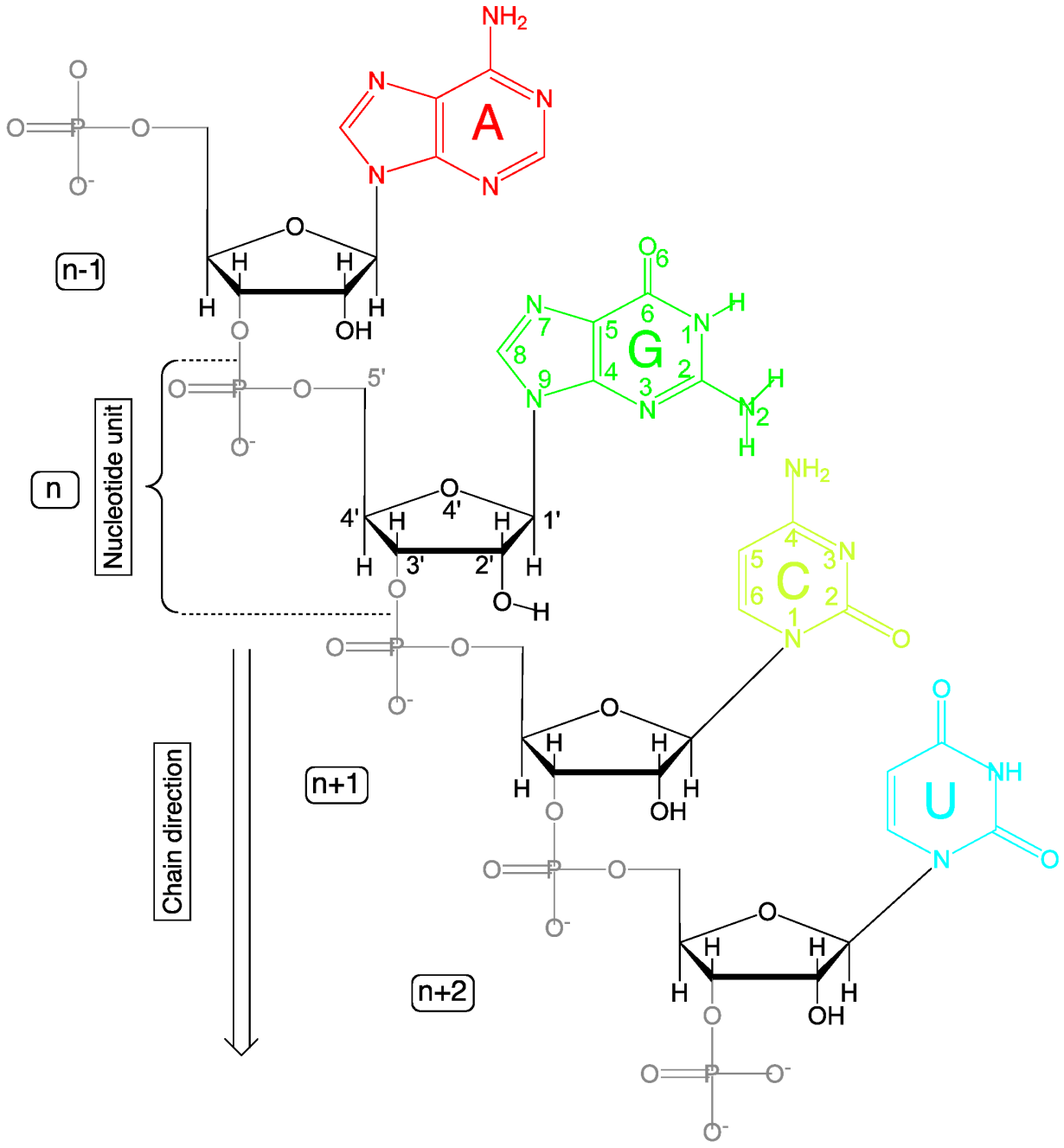


Figure 1.1: A single strand of RNA drawn in the 5′ to 3′ sense showing the three chemical entities which compose it, base, sugar, and phosphate. The four bases (A, G, C, U) are colored according to the NDB (Nucleic Acid Database) convention [18], the phosphate is colored gray, and the sugars black. The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

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 unique puckered arrangements in which 1-2 of the five atoms lie above of the plane defined by the other 3-4-5. The prefered sugar pucker in RNA is the $C_{3'-\mathrm{endo}}$ form, but in cases where a base intercalates between two sequential bases, the sugar pucker frequently change to the less-prefered $C_{2'-\mathrm{endo}}$ 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 $\mathrm{tRNA}^{\mathrm{Phe}}$ 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 the folding pathway that it follows.

[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.

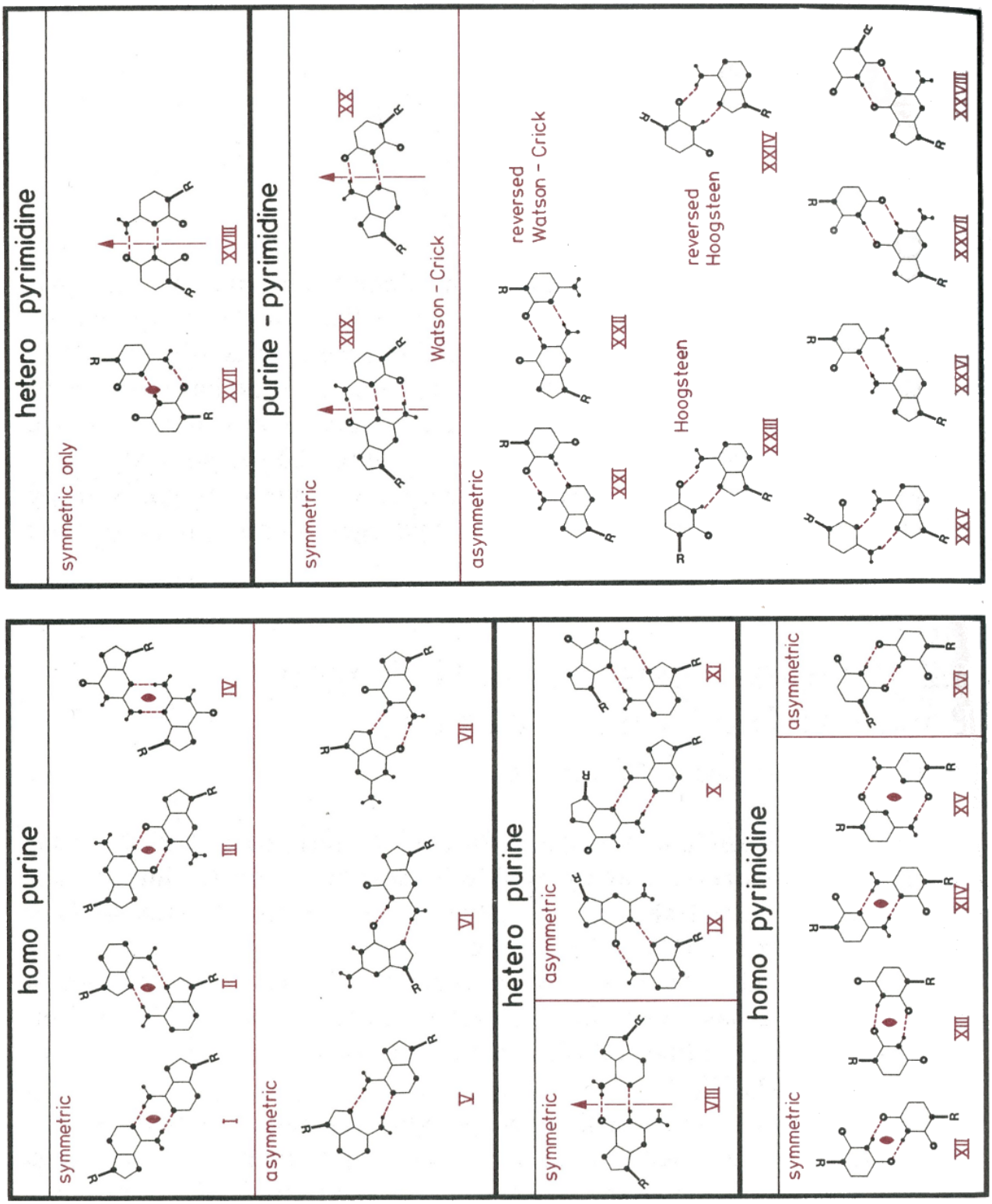


Figure 1.2: Saenger base-pairing classes, reproduced from his book, "Principles of Nucleic Acid Structure". [20].

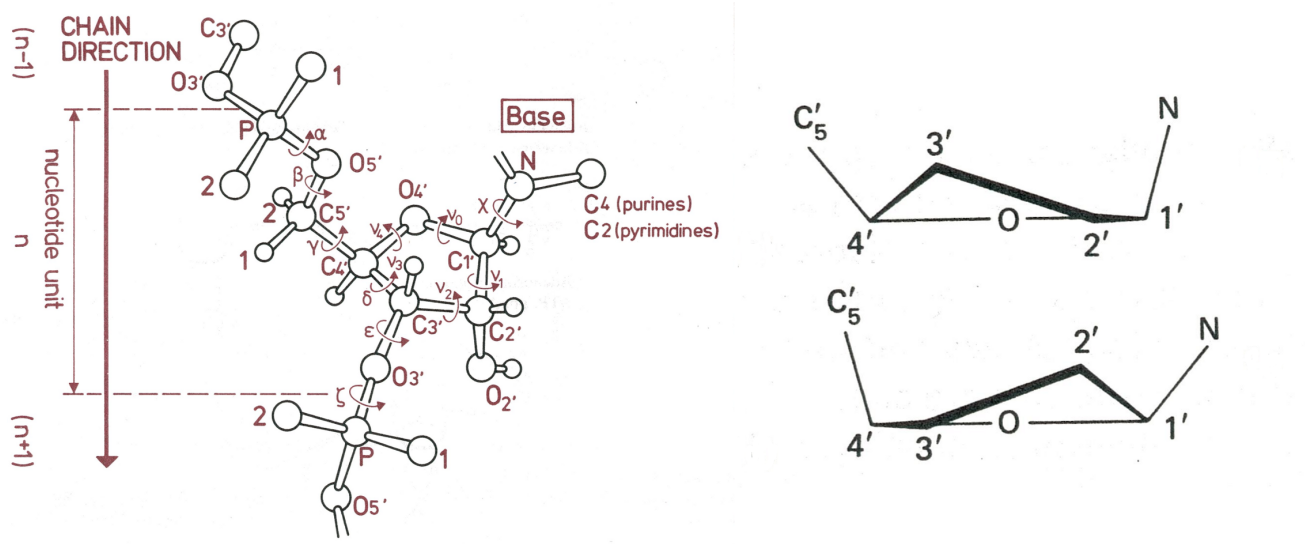


Figure 1.3: **Left:** Sugar, and sugar-phosphate backbone torsion angles. **Right:** The most common sugar pucker conformations in RNA, that is, $C_{3'-\mathrm{endo}}$ and $C_{2'-\mathrm{endo}}$, reproduced from Wolfram Saenger's, "Principles of Nucleic Acid Structure". [20].

The RNA folding problem is usually seen as analogous to the protein folding problem, due both to the discovery of the enzymatic behavior of RNA [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 are based on defining an adequate partition function for describing 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, whereas secondary and tertiary elements are not as easily separable in proteins. The other point of view says that RNA folding can be at least as complex as protein folding [40, 41] since there is no such thing as hydrophobic burial of regions of RNA as in the case of proteins. Instead, the electrostatic problem

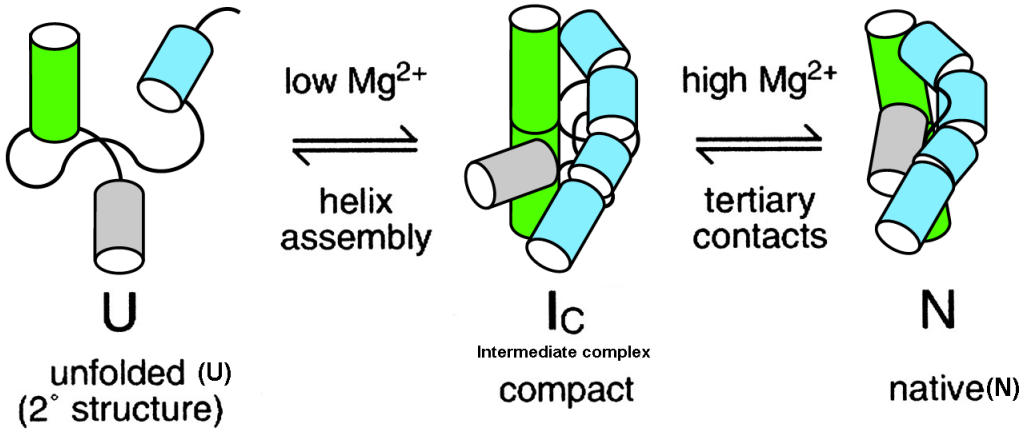


Figure 1.4: Separation of secondary and tertiary interaction in RNA [39]. Double helical secondary structure represented by individual cylinders and tertiary interactions by association of cylinders. Color coding stands for separate helical regions of RNA, and the connecting black strings represent single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

stemming from a complex charged backbone must be dealt with in the case of RNA. For instance, the interactions of the RNA polyanionic backbone with water and cations [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 $\mathrm{tRNA}^{\mathrm{Phe}}$, [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 which 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]. Recently an alternative solution to this problem has been proposed by Thirumalai and associates based on single-molecule force-quenching experiments, by using a so called de Genes "expanding sausage model" [58].

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 [59, 60, 61], which treat RNA at less than atomic detail, have been used to describe the motions of large RNA structures like the ribosome. Examples of the predicted normal modes of motion of the ribosome can be seen at: http://ribosome.bb.iastate.edu/70SnK mode. Using MM, Sanbonmatsu and coworkers obtained a static atomic model of the 70S ribosome structure through homology modeling [62]. Tung and associates used this structure for an all-atom MD simulation of the movement of tRNA into a fluctuating ribosome [63]. This type of simulation might be useful in a reverse-folding approach to the RNA folding problem. To the best of our knowledge, such calculations haven't as yet been done for RNA.

1.5.1 Local nucleotide interactions

The molecular interactions that rule RNA structures at the nucleic acid base level, *i.e.*, local level, are hydrogen bonding and stacking interactions. The former are related to base pairing and the latter, in most cases, to nucleotide steps. These interactions can be explored theoretically at various levels. At the highest level are ab-initio quantum mechanical calculations which are still too expensive for systems as large as hundreds of atoms. Such calculations, nevertheless, can tell a great deal about local electronic behavior. For example, Hobza and collaborators have found that the stacking interaction of free nucleotide bases is determined by dispersion attraction, short-range exchange repulsion, and electrostatic interaction. No specific $\pi - \pi$ interactions are found from electron correlated ab-initio calculations [24, 25]. This is why force field methods have been so successful in the study of nucleic acids, since the empirical potentials used in such studies mimic well the quantum mechanically obtained energy profiles [62, 64]. 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 [65] seem to contradict their earlier work [64, 66], 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 [67]. One of the most popular methods used for obtaining such quantities is differential scanning calorimetry (DSC) [68]. 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]. In fact, the stacking free energies for some sequences[ii] are found to be negligible. Thus there may be no accountable stacking interaction at all for some sequences.

Besides taking into account the effects of stacking and hydrogen bonding, it is important to think at the same time about the polyelectrolyte nature of the RNA backbone. Manning's counterion condensation theory [71, 72] 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.[69, 70]

the behavior of charged macroions in solution [73, 74].

The local conformational space of RNA has been studied using a large set of available RNA structures from the Nucleic Acid Database (NDB) [75]. The torsion angles of the nucleotide steps have been clustered using different techniques [76, 77]. The root-mean-square deviations (RMSD) of the distances between closely spaced atoms in the phosphates, sugars, and bases, have also been clustered [78]. The latter studies are aimed at finding the common nucleotide base steps and 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 [79].

1.5.2 RNA secondary structure algorithms and the lack of tertiary ones

From secondary structure prediction algorithms like Zuker's *mfold* program [80], Hofacker's Vienna RNA package [31], or Mathews Dynaling software [81], one obtains a large ensemble of secondary structure graphs, i.e. 2D representations of the double-stranded helical stems, hairpin loops, bubbles formed by the constituent bases. These graphs can be analyzed with graph theory to produce a partition function describing a full arrangement of contacts for the total number of possible secondary structures, allowing the construction of a "relation of microscopic conformations to macroscopic properties" [82]. 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 [83, 84, 85]. 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 carry out an MM optimization [86]. The third approach entails a pipeline between secondary structure prediction, and tertiary structure assembly is proposed. This pipeline uses as bridging concept between 2D and 3D structure, the graph theoretical definition of a minimum cycle basis, which for the case of nucleic acids has been renamed as Nucleic Cyclic Motifs (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

[87] (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 ribosome, a representation at such level of detail does not allow to make sense of such a large structure (close to 3000 nucleotides for the large subunit of the archaeal ribosome). In the past two years Holbrook [88] and Sykes [89] 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 3D helical axis to 2D images, which can later be annotated with, for example, hydroxil radical footprinting results.

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.* [90]

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 we 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 [92], classifies proteins, among various qualitative descriptors, according to folds, which are recurrent arrangements of secondary structure, that is, a list of secondary structures with unique topological connections. The SCOR (Structural Classification of RNA) database [93, 94], 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 cannot describe the overall fold. Local classification is also qualitative rather than quantitative.

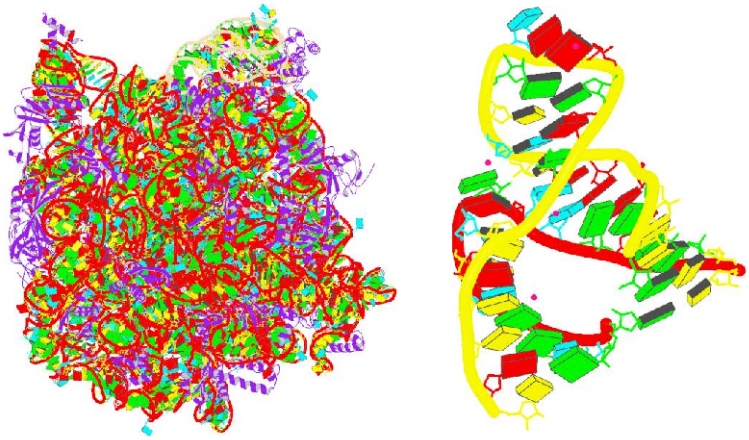


Figure 1.6: Images of the *Haloharcula marismortui's* large ribosomal subunit NDB_ID:RR0033 (left) and the hammerhead ribozyme (right) NDB_ID:UR0029. The figures were taken directly from the NDB web pages, and show a 3DNA generated [91] ribbon representation of the phosphate backbone, and a block representation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

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 result in confused and frustrated bibliographical searches.

The lack of a unique definition of RNA motifs is yet another source of confusion in understading RNA motifs is the lack of a unique definition. 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.*"[95]
- "*conserved structural subunits that make up the secondary structures of RNAs.*"[96]
- "*ordered stacked arrays of non-Watson-Crick base pairs that form distinct folds on the phosphodiester backbones of RNA strands.*"[97]

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 peculiar sets of geometrical (in the rigid block sense) arrangements 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 a software called AMIGOS. This software finds RNA motifs based on specific values of backbone virtual torsion angles η and θ [98, 99, 100] in a way which resembles a Ramachandran plot analysis. Lemieux and Major [101] provide the software MC-Fold, which implicitly finds RNA motifs based on an algorithm to determine so called nucleic cyclic motifs, which are just the minimal cycle basis of an RNA secondary structure interpreted as a mathematical graph. Leontis [102] and collaborators provide FR3D (read as FRED). F3RD is a matlab windows executable program which finds RNA motifs based on the isostericity matrices of base-pairs.

For description of RNA motifs 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 [103, 104]. 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 carried out the description of RNA motifs present in the ribosome using Molecular Dynamics (MD) methods implemented in the AMBER package. They have performed 25ns simulations of the Sarcin-Ricin Domain (SRD) of the ribosome [105], and also 80ns simulations of hydration of loop E in the 5S subunit of the ribosome [106].

The software programs which perform the task of identifying RNA motifs in RNA structures also have the ability to find new RNA motifs, as is the case for AMIGOS, MC-Fold, and FR3D.

1.6 Overview

Keeping always in mind the greater scope of the RNA folding problem, this thesis addresses various issues of RNA structural understanding using RNA crystallographic data from the Protein Data Bank (PDB). Such data has been analyzed statistically in terms of a rigourous rigid-body formalism. In Chapter 2 the consensus clustering technique is used to classify RNA base-step parameters of non-A-RNA conformations, and the resulting groups are localized and understood in the context of rRNA. Chapter 3 reconsiders previous work carried out by Dr. Yurong Xin at the Olson's lab, on classification of RNA base-pairs by resorting again to clustering analysis techniques, and database mining of the WWW available Base Pair Structures (BPS) database. In Chapter 4 we explore, using statistical analysis, the data available on RNA helical regions, and use this information to compute the persistence length of

double-stranded RNA's and compare it to experimental results. In Chapter 5 we provide a new python software, pyRNAmotifs which interfaces with 3DNA to do a rigourous search of existing and perhaps new RNA motifs, and finally in Chapter 6 we propose the measurement and classification of RNA structures using a new graph theoretical index named folding index, based on a helical region "view" of RNA's, which is clearly concordant with the emerging necessity of new metrics beyond RMSD for structural understanding.

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] (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] Hyeon, C., Morrison, G., Pincus, D. L., and Thirumalai, D. (2009) Refolding Dynamics of Stretched Biopolymers Upon Force Quench. *Proceedings of the National Academy of Sciences of the United States of America,* **106**, 20288–20293.

[59] 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.

[60] 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.

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

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

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

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

[69] 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.

[70] 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).

[71] 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.

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

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

[74] 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.

[75] 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.

[76] 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.

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

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

[79] 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.

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

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] 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.

[86] 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.

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

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

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

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

[91] 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.

[92] 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.

[93] 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.

[94] 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.

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

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

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

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

[99] 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.

[100] 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.

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

[102] 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).

[103] 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.

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

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

[106] 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.

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]. This problem had only been addressed by a few researchers 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 due 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].

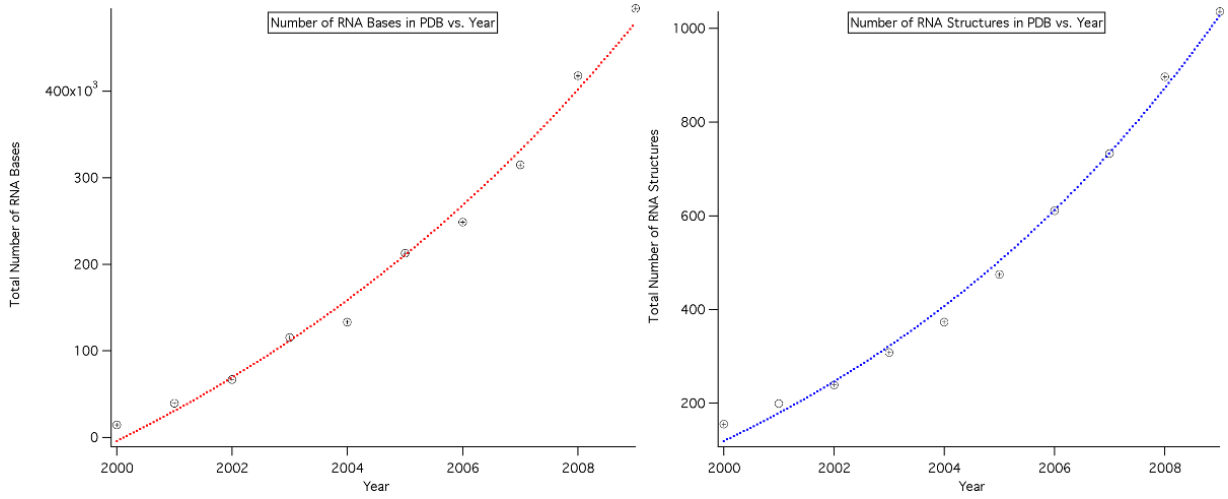


Figure 2.1: **Left:** Total number of RNA bases added to the PDB database between 2000 and 2010 (Exponential fit line in blue). **Right:** Total number of RNA structures solved yearly by X-Ray crystallography between 2000 and 2010 (Exponential fit line in red).

Between 1978 and 2000 a total of 116 RNA structures with resolution greater than 3.5 Å, and comprising around 5500 nucleotide bases are found in the Protein Data Bank (PDB), and between 2000 and today a total of 931 RNA structures comprising 491158 nucleotide bases are found. 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]. Looking at the growth of RNA structural information from 2000 until today, it is clear

that both the total number of RNA structures deposited to the PDB, and the total number of nucleotide bases in these structures, is growing in an exponential way (as can be seen by the exponential fits in Figure 2.1). It's important to note that such growth comes mainly from ribosomal structures which contain 88 percent of all RNA bases in the PDB. So, even though structural interest in RNA is 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 RNA sizes counted by number of bases, as can be seen in Figure 2.2 it's clear that there are great patches where there are no RNA structures whatsoever, roughly between 600 and 1400 bases and between 1800 and 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 Breaker [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. An interesting comparison between the total number of structures of RNA, protein, dna, and nucleic acid plus protein, available at the PDB from the seventies until today can be seen in Supplement Figure S2.

The analysis of RNA conformational information contained in RNA structural data can be divided into 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 direct comparison of backbone atom positions is made [9], or a comparison of 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 consecutive covalent bonds in the RNA backbone and the glycosidic bond between the sugar and base, that is, six backbone torsion angles and one glycosidic torsion angle [9, 11, 12, 13, 14]; or alternatively 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.

[i] Previously the pseudotorsion angles η and θ were given the names ω_{ν}, and $\omega_{\nu'}$.[18]

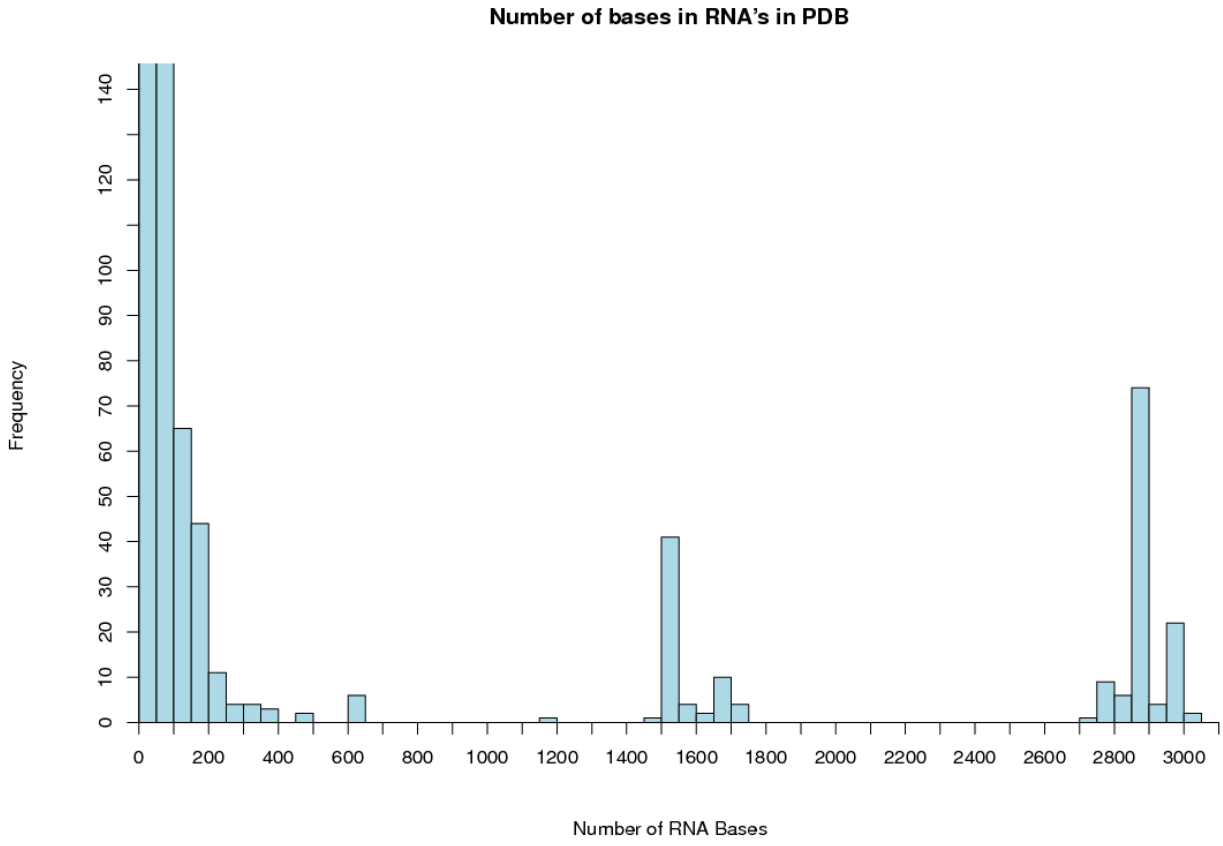


Figure 2.2: Frequency of nucleotide bases in RNA molecules found in the PDB classified by the size of RNA molecules. We define the size as the total number of nucleotide bases present per molecule.

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

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

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 RNA structures available from the PDB. It is important to note here that in crystal structures, RNA bases are determined more accurately than backbone torsion angles, as has been shown by Richardson and collaborators from analysis of van der Waals steric clashes. This can be seen more clearly in Figure 2.3, reproduced from Richardson's work [11], where the red and orange dots in the backbone atoms region denote steric clashes and the green and yellow dots in the base atoms region denote very good agreement with expected van der Waals distances.

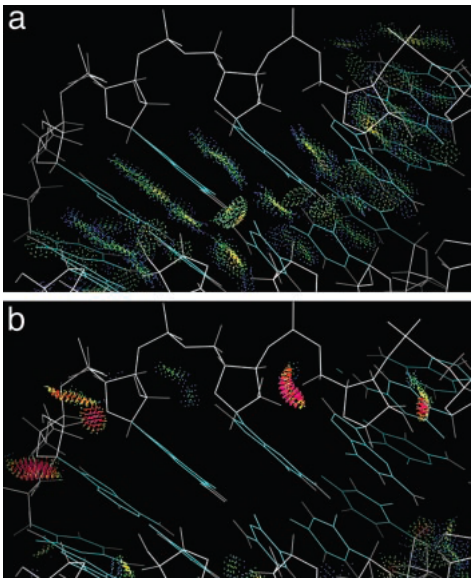


Figure 2.3: Figure taken from Richardson et al. [11] where the blue and green dots in a) mean very accurate van der Waals distances, and in b) the red and orange dots mean steric clashes, that is, distances outside the acceptable van der Waals range.

2.1.1 Combining Fourier Averaging Results and Clustering Analysis

Using the coordinate files of 20 rRNA structures provided by Schneider at al.[13] we used standard clustering analysis (CA) techniques (see Appendix B) to classify a set of non-ARNA base-steps using, rather than the more common torsion angles space, the base-step parameters space, that is, three translational parameters (Shift D_x, Slide D_y, Rise D_z), and three rotational parameters (Tilt τ, Roll ρ, Twist ω), which we describe with the 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 2.4 and whose corresponding structures are shown in Figures 2.5 S1 were obtained by performing clustering analysis and consensus clustering on 20 structures provided by Schneider et al. [13]. These twenty structures were obtained by Schneider applying a Fourier averaging technique, and lexicographical clustering, to torsion angles of 23S rRNA. The methodology we used follows that used by others to recover the periodic table classification from multidimensional property vectors for elements [22, 23].

In Figures 2.5, and S1 we see that Group I contains structure 1 with base-plane normals pointing in opposite directions, Group II includes extended conformations with neighboring bases roughly parallel but not stacked and is formed by structures 15, 16, 10, 14, Group III also contains extended conformations with bases perpendicular to one another and is formed by structures 8, 9, 17, Group IV 18, 19, 20, 13, 11, 12, 5, 3, 6, 7, 2, 4 contains four major subgroups: (a) structures 2, 4 which are unstacked with bases neither parallel nor perpendicular; (b) structures 18, 19, 20 which closely relate to A-RNA; (c) structures 11, 12, 13 which are unstacked and have parallel bases; and (d) structures 3, 5, 6, 7 which are also unstacked and have parallel bases. We also see in Group IV that the conformers in subgroups IV (c) and IV (d) are closely related, and that the dimers in these two subgroups are more closely related to those in subgroup IV (b) than to those in subgroup IV (a).

To account for the representation of the groups obtained by clustering in the 23S subunit of the ribosome we have computed the root-mean-square deviation (RMSD) between the average step parameters of the structures composing each group, and the step-parameters of the 2753 steps present in 23S. That is, for each group we have obtained a set of RMSD values which have been plotted as histograms as shown in Figures 2.6, and 2.7. The results are also summarized in Table 2.2, where we can see that they only constitute 31% of the total amount of steps in the 23S subunit of the ribosome. We used a cutoff of 10 Å[ii] to select the structures which belong to each group, based on visual analysis of superimposed reconstructed structures. For example, for Group I; if we reconstruct the ribosomal steps with an RMSD of 10 Å or less, we get the figure shown in the left panel of Figure 2.8. But if we reconstruct with the set of structures with an RMSD of 15 Å or less we start getting structures, which after being superimposed based on the reference frames of the first base are clearly not related to that group, as can be seen in the right panel of Figure 2.8.

We have also noticed that the starting structures kindly provided to us by Dr. Berman, have a large

[ii] We retain the traditional unit of Angstroms to refers to our RMSD's, but it is important to note that since we are not refering to an all-atom model such unit does not have a direct physical meaning.

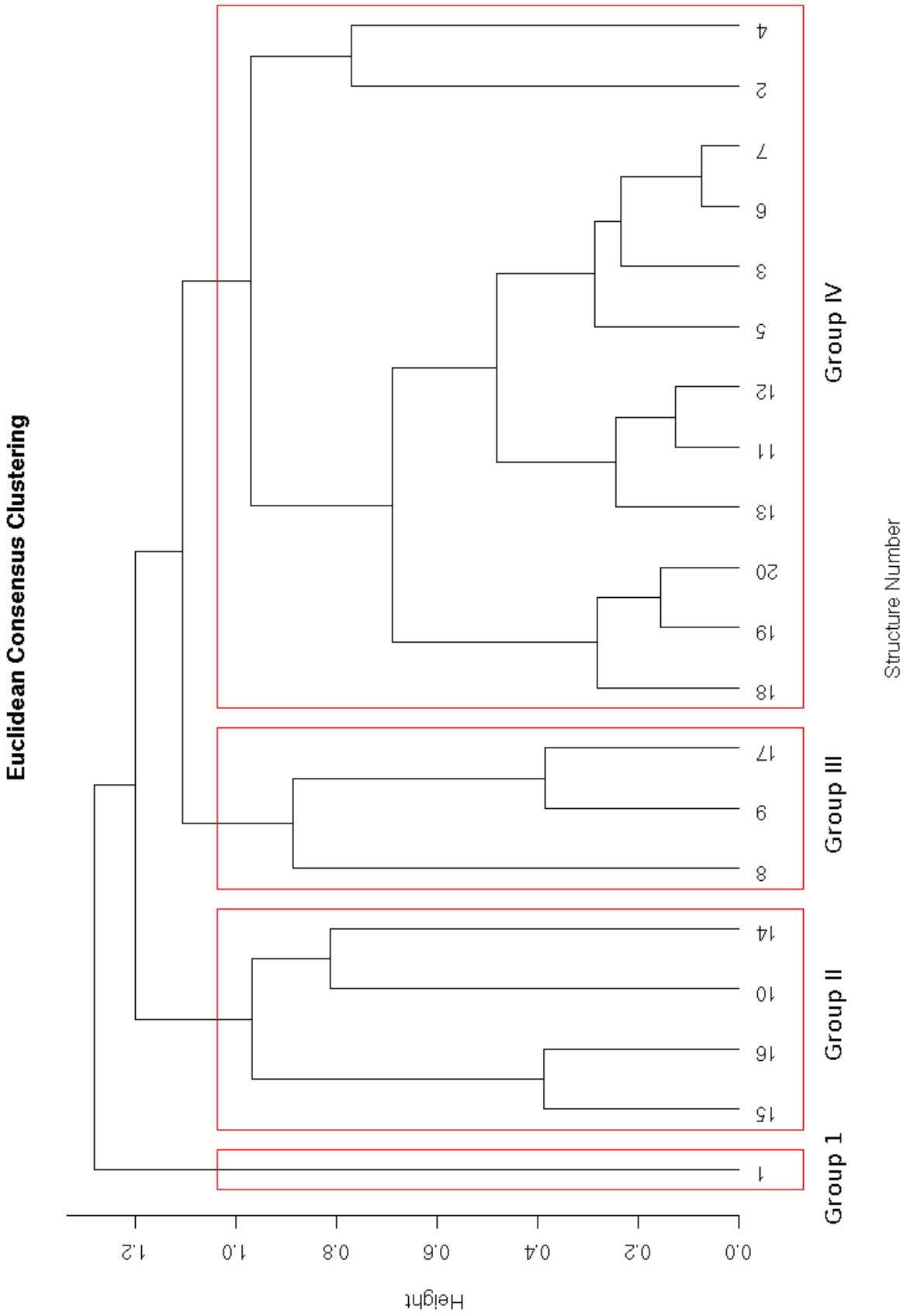


Figure 2.4: Dendrogram showing the results of consensus clustering of 20 non-Atype rRNA dinucleotides according to their hexadimensional base-step parameter vectors.

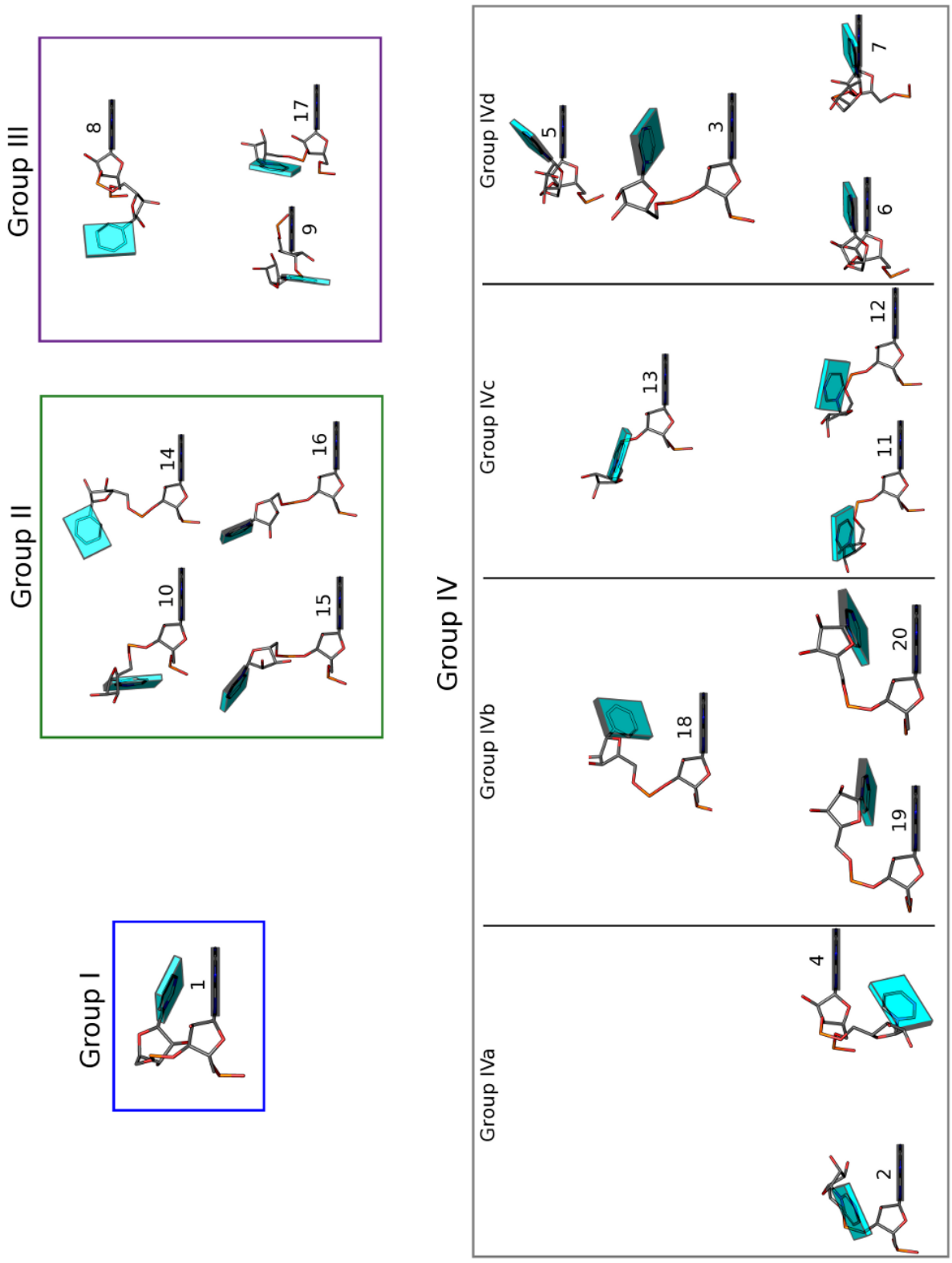


Figure 2.5: RNA dinucleotide structures organized by clusters obtained from consensus clustering of their hexadimensional base-step parameter vectors. The structures have been centered on the reference frame of the first step, that is, the adenine base, and the minor groove face of the rigid block parameter associated to adenine is facing the viewer.

rise in the case of A-RNA, that is, a value of 4.39 Å, which is larger than the 3.30 Å value obtained for the "classical" A-RNA structure from Arnott and collaborators [24]. This might have a significant effect on the amount of structures which can be grouped under the A-RNA like group.

Because of not getting a good representation of the total diversity of base-steps in the 23S subunit of the ribosome, we have opted to perform an analysis based fully on base-step parameters. We believe that the reason for such poor representation is due to the mixing of Fourier averaging for backbones, and the base-step perspective.

Group	Percentage	Number of Base-Steps
I	0.11	3
II	0.18	5
III	0.04	1
IVa	0.36	1
IVb	29.31	807
IVc	0.33	9
IVd	1.27	35
Total	31.28	861

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

2.1.2 Selection of a Clustering Methodology

In order to analyze our dataset of base-step parameters we have decided to use clustering analysis methods. Clustering analysis methods can be broadly classiffied in two main categories, that is, they can be partitional or hierarchical. In either case the main problem one faces for classification purposes is that of deciding which is the optimal number of hierarchies or partitions that the analyzed data is split into. To obtain a criteria 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 detail on the definition of such measures are provided in [25, 26]. To perform the validation analysis mentioned above we used a cluster validation package implemented in the R [27] statistical analysis program called clValid [26].

In Figures 2.10 and 2.11 we present the results for internal and stability validation results exploring the same dataset of base-step paramaters for the 23S subunit of the ribosome that we've used before. In the clustering anlysis literature it's customary to use the variable k to define the number of clusters

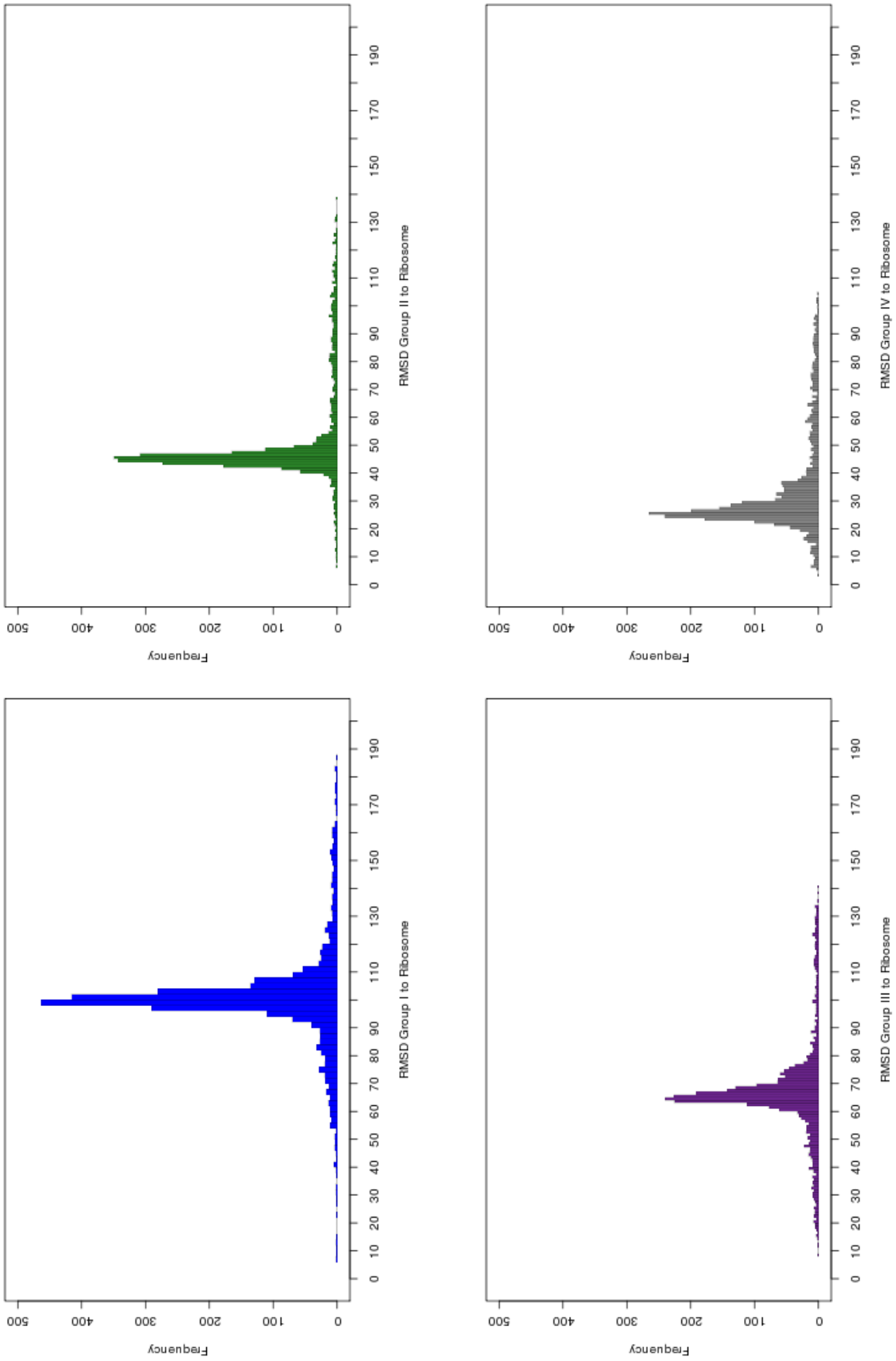


Figure 2.6: Root mean square deviation of the main four groups show in Figure 2.5. The color of the histograms is the same as that of the boxes surrounding the structures of Figure 2.5

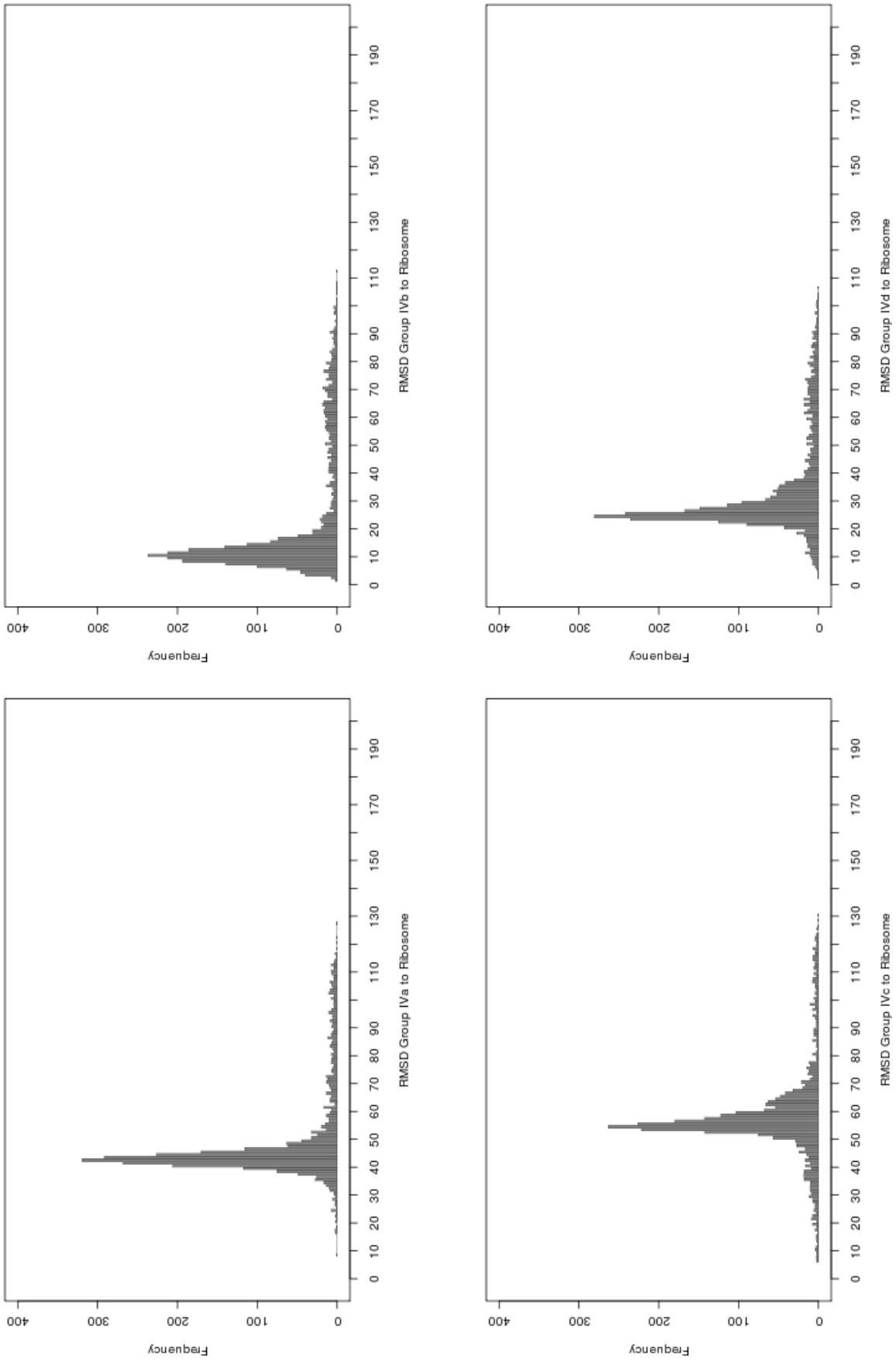


Figure 2.7: Root mean square deviation histograms for the subgroups present in group IV. Since subgroup IVb is composed of A-RNA like conformations we see in the upper left histogram that the highest proportion of small RMSD values belongs to this group.

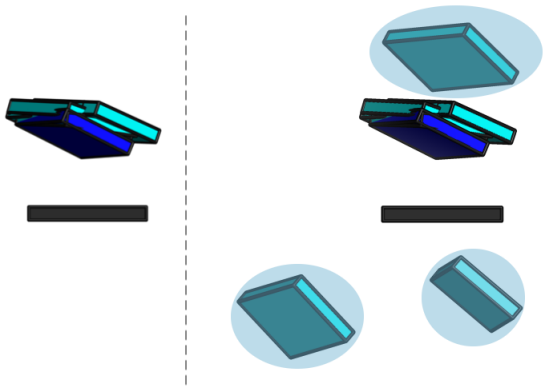


Figure 2.8: Rigid block representation of dinucleotide steps. The major groove side of the first nucleotide block is oriented towards the viewer and shaded gray. **Left:** Drawn in blue, the block representing the Group I cluster from Figure 2.5. Superimposed to the Group I cluster are three structures whose step-parameter RMSD's with respect to the Group I cluster are less than or equal to 10 Å. **Right**: With an RMSD less than or equal to 15 Å we "identify" a total of seven structures from the ribosome. We clearly see that three of them (encircled in cyan blobs) are farther apart from the original Group I main structure of Figure 2.5 which is drawn in blue.

and we will use variable k in that sense in what follows.

Our analysis computed the validation scores for a number of clusters ranging from $k = 2$, up to $k = 80$ clusters, and evaluated hierarchical methods (hierarchical, diana), and partitional methods (kmeans, pam, som, sota). The connectivity measure must be minimized, and the average silhouette width (silhoutte) and dunn index must be maximized. With this in mind, we see that the method labeled as hierarchical[iii] performs better in connectivity and dunn index for the whole range, and it is also the best performer in silhoutte from $k = 12$ onwards.

In the case of the stability measures it is important to note here that mainly these measures are well suited for highly correlated data sets, therefore they are not very indicative for our data set, which is correlated in shift and twist, as can be seen from the values on the upper right corner of the pairs scatteplot shown in Figure 2.9. We include the cluster stability measures for completeness.

The stability measurements we have computed are read as being better the smaller their values, of these we have quantified three measures, that is, 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. [26]. As seen in Figure 2.11 the method with the best stability measures is sota for APN, and ADM, almost for the whole range, until it reaches a number of clusters of around 70. For

[iii]The hierarchical label refers precisely to the agglomerative (bottom-up) technique, the euclidean metric, and the average method.

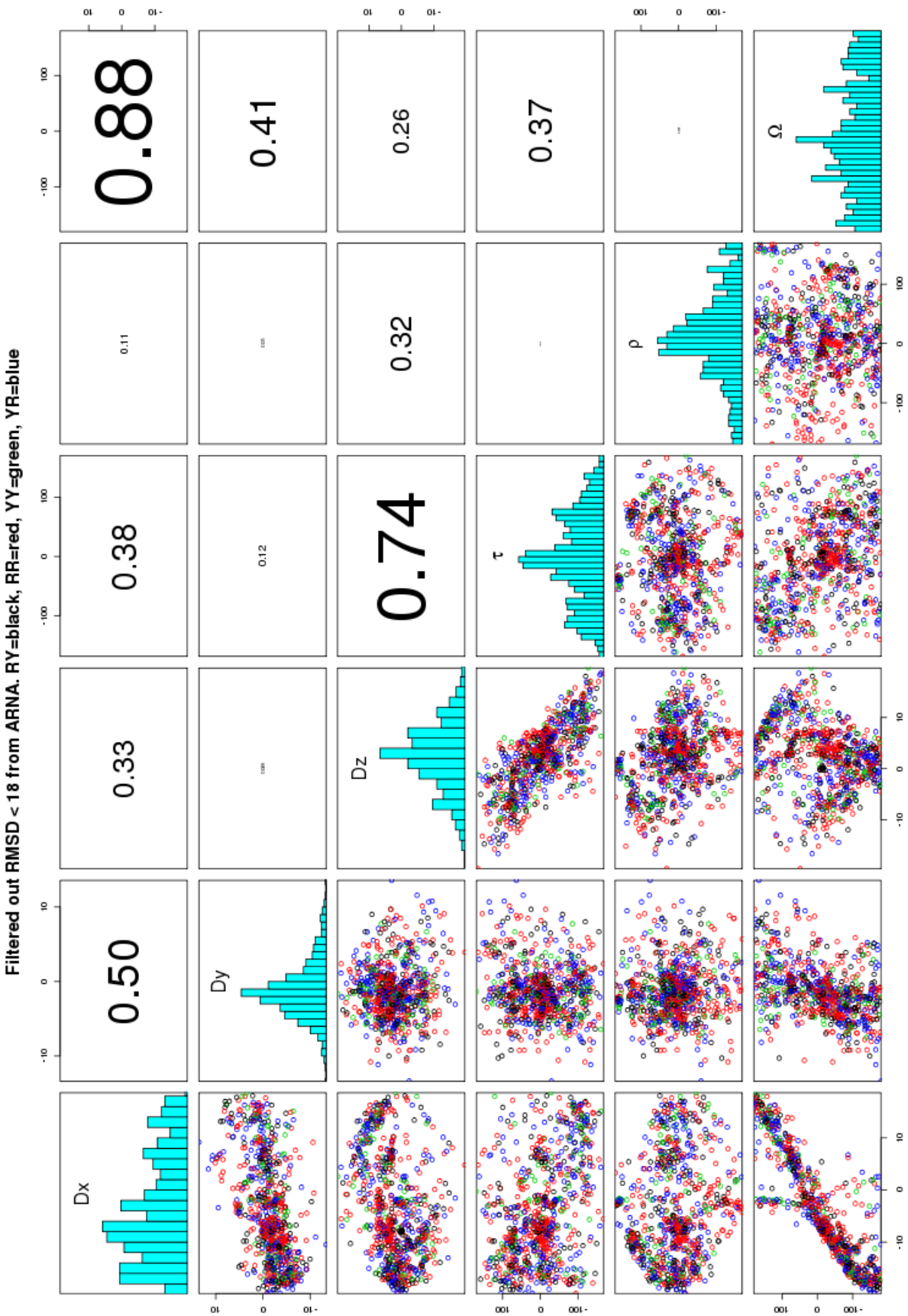


Figure 2.9: Pairs scatterplot for base-step parameters, shift, slide, rise, tilt, roll, and twist, for the non-ARNA dataset colored according to purine-pyrimidine (black), purine-purine (red), pyrimidine-pyrimidine (green), and pyrimidine-purine (blue) steps.

Structure Name	Shift (D_x)	Slide (D_y)	Rise (D_z)	Tilt (τ)	Roll (ρ)	Twist (Ω)	**Reference**
A-DNA	0.36	-1.39	3.29	2.46	12.50	30.19	Arnott [28]
B-DNA	0.44	0.47	3.33	4.63	1.77	35.67	Arnott [28]
A-RNA	-0.08	-1.48	3.30	-0.43	8.64	31.57	Arnott [28]
A'-RNA	0.05	-1.88	3.39	-0.12	5.43	29.52	Arnott [28]
AII-RNA	1.01	-2.52	3.33	2.94	9.75	25.12	Schneider [13]

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.

the AD measure the best performers are pam and sota in the whole range. Notice that the hierarchical method follows the same trend as the other methods, and that in general, apart from the APN measure and the sota method, all methods have a similar behaviour due to the fact that our data set is not highly correlated, that is, it cannot be split into say, two, three, or four, principal components.

In all cases we also see that the best overall number of clusters is two, which is not surprising since we haven't filtered out A-RNA structures from our data set, leaving two main groups; that of A-RNA type base-steps, and those which are not A-RNA like.

We focus our attention in the group of structures which are not so closely related to A-RNA. Therefore we have extracted them from the whole dataset based on Figure 2.12, and end out with a data set of 797 (about 29% of the total number of steps) base-step parameters whose values are greater than an RMSD of 18 Å. These RMSD values have been computed between the base-step parameters of 23S RNA and the standard base-step parameter values derived from Arnott and collaborators [24] work. Standard base-step parameter values for common double-stranded conformations of RNA, and DNA are provided in table 2.1.2.

With the filtered dataset, which we will refer to as non-ARNA dataset, we have again repeated the cluster validation analysis for internal measures as can be seen in Figure 2.13. From this analysis we see again that the best method for clustering our dataset is the hierarchical one. The Dunn index, which works under the idea of finding the best possible separation and compactness between clusters, shows us that the optimal number of clusters is $k = 67$. The other two indices show, as for the whole dataset case, that the optimal number of clusters is two, nonetheless, a common indicative of optimal cluster solutions in the connectivity and silhouette plots is given by the presence of shoulders. We see that there is a shoulder also at $k = 67$ for the connectivity and silhouette plots. We selected the 67 clusters given by the hierarchical method, and took their corresponding step-parameter values to reconstruct the dinucleotide step structures using 3DNA. In Figure 2.14 we draw the first seventeen groups populated

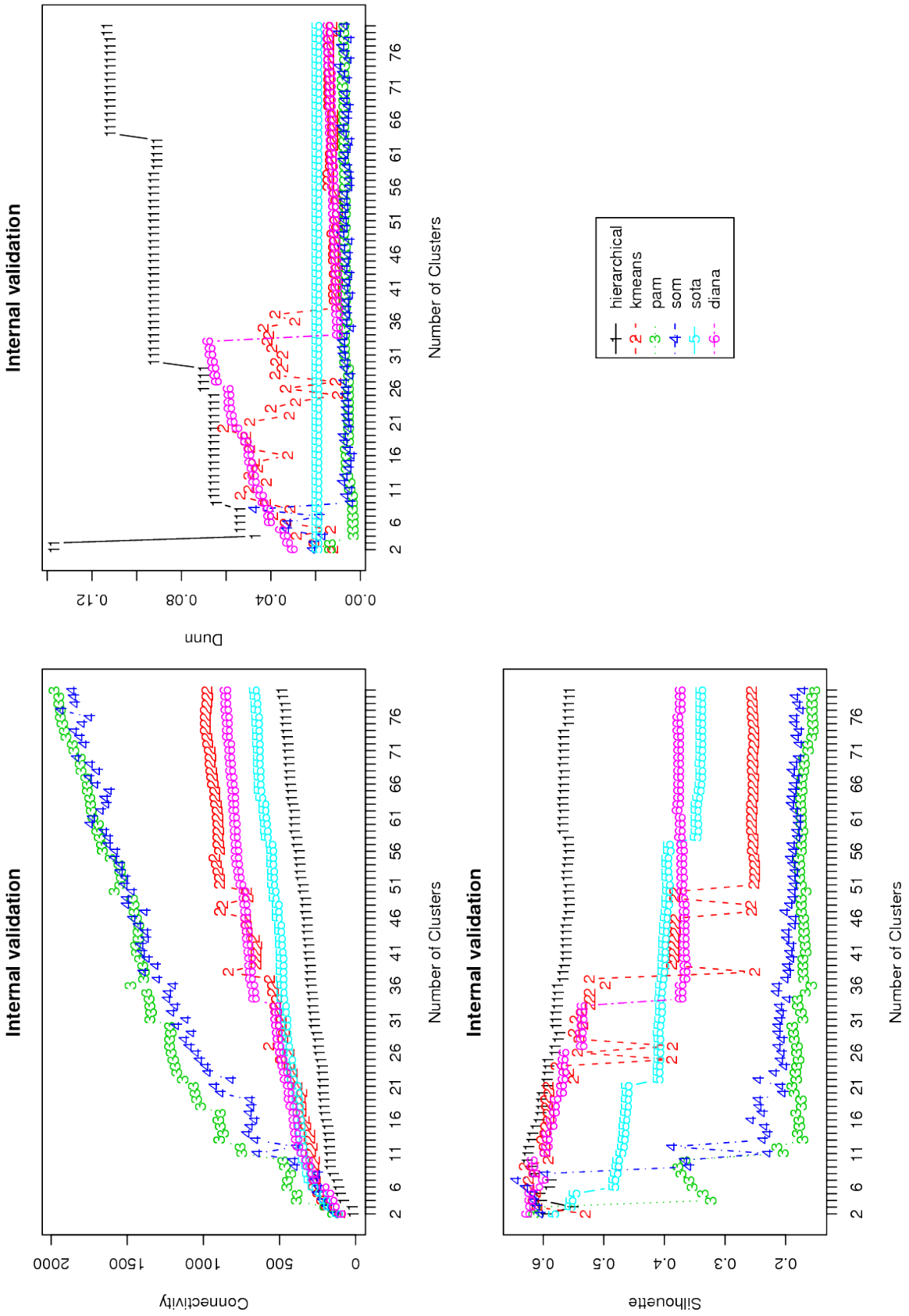


Figure 2.10: Cluster validity scores for internal measures. Notice how the hierarchical method, labeled as 1 in black color, behaves better for the whole range of Connectivity (smaller values) and Dunn (higher values), and it also outperforms all others after $k = 12$ for Silhoutte (higher values) scores.

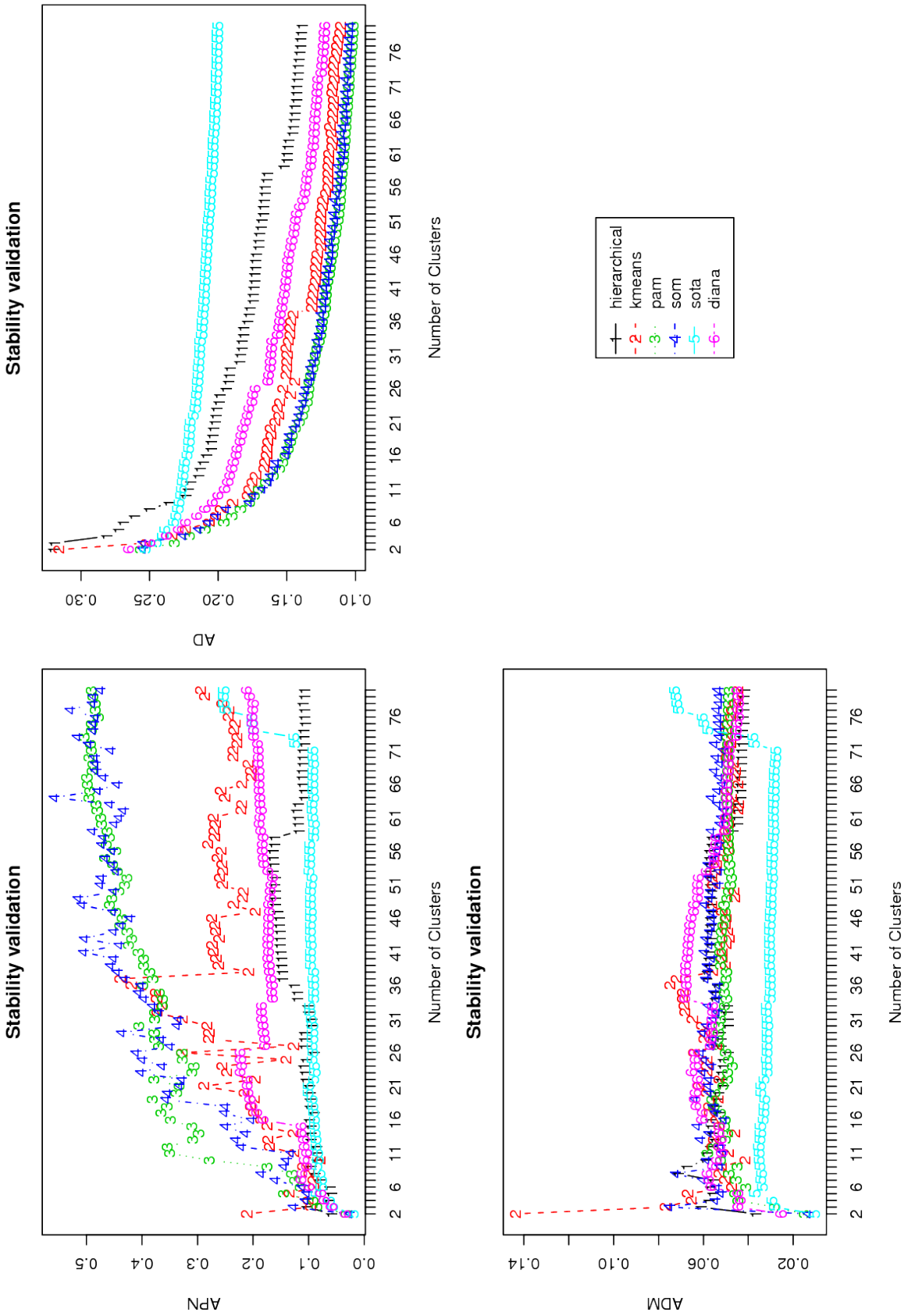


Figure 2.11: Cluster validity scores for stability measures.

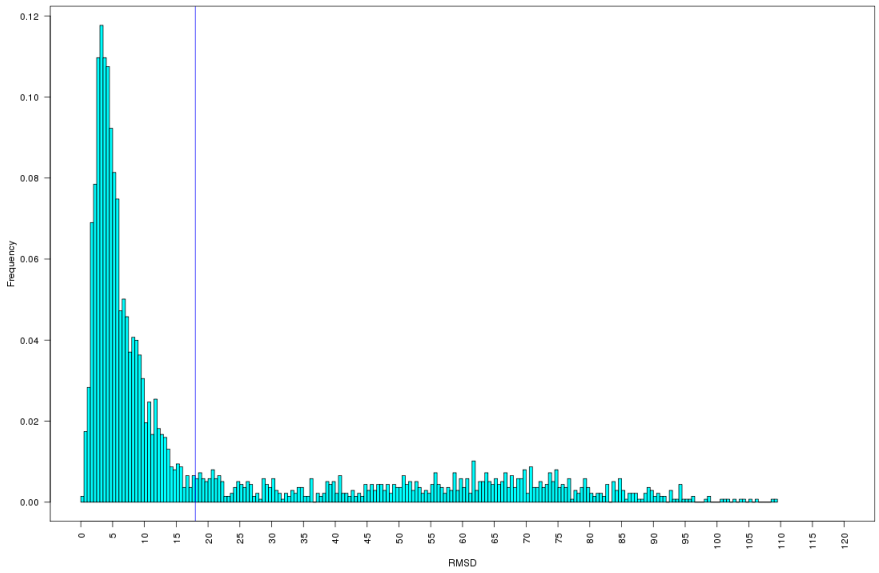


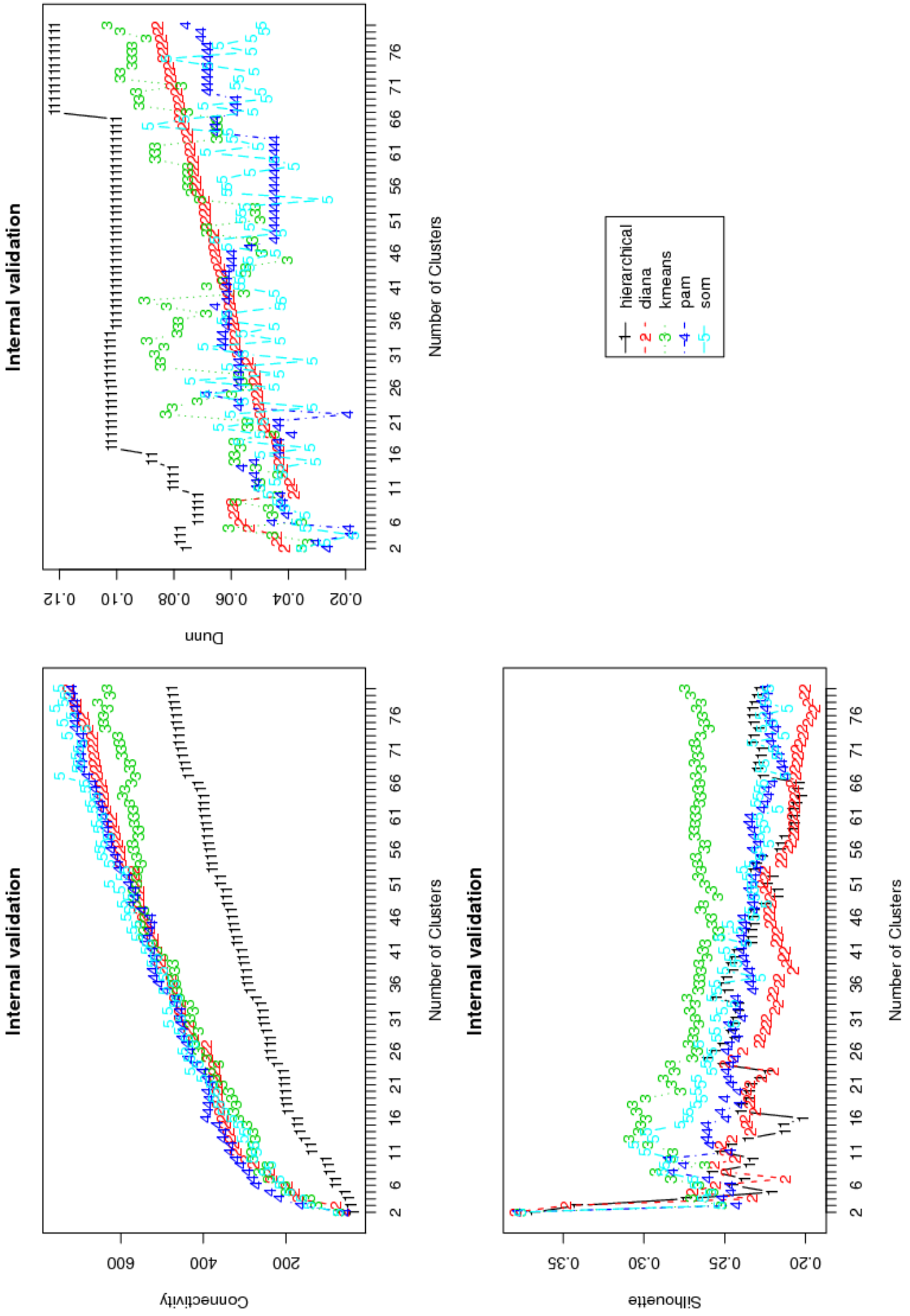
Figure 2.12: RMSD values between base-step parameters of the 23S subunit of ribosomal RNA and the standard base-step parameters derived from Arnott and collaborators [24] work.

by ten or more structures on them, which account for 80 percent of the total amount of steps in the non-ARNA set. We also plot in the lower right corner of Figure 2.14 the set of 20 structures derived from the work of Schneider et al.[13], and the whole set of non-ARNA dinucleotide steps. All structures are centered using the standard reference frame embeded in the first base, which in our reconstructions corresponds to a red block representing adenine, whose minor groove face is oriented left, its major groove is oriented to the right, and its watson-crick base-pairing face is oriented towards the viewer.

When comparing the 17 groups of non-ARNA dinucleotide steps with those coming from the work of Schneider and collaborators we see that in their set of structures there are no steps represented at the major groove side of the red block representing adenine, that is, the right side of the red adenine block. We also see, that even though the 17 groups represented are not as compact as one would desire, they start to give an indication of geometrical preferences on the space of dinucleotide step-parameters. For example, it is remarkable to see in group 7, labeled as g7.png in Figure 2.14 that the blocks representing uracyl in cyan color, orient their planes orthogonally to the major groove side of the red block representing adenine.

The reason we choose to include the figure of all non-ARNA base-steps in Figure 2.14 is that of giving the reader an idea of the complexity of the space of base-step conformations described from a base viewed perspective instead of the more common backbone perspective, this also suggests that the task of finding order in this broad range of possible conformations is analog to the task of peeling

an onion. We believe the onion can be effectively peeled into parts by using appropriate validated clustering analysis techniques.


Internal validation
Connectivity
Number of Clusters
Internal validation
Dunn
Number of Clusters
hierarchical
diana
kmeans
pam
som
Internal validation
Silhouette
Number of Clusters

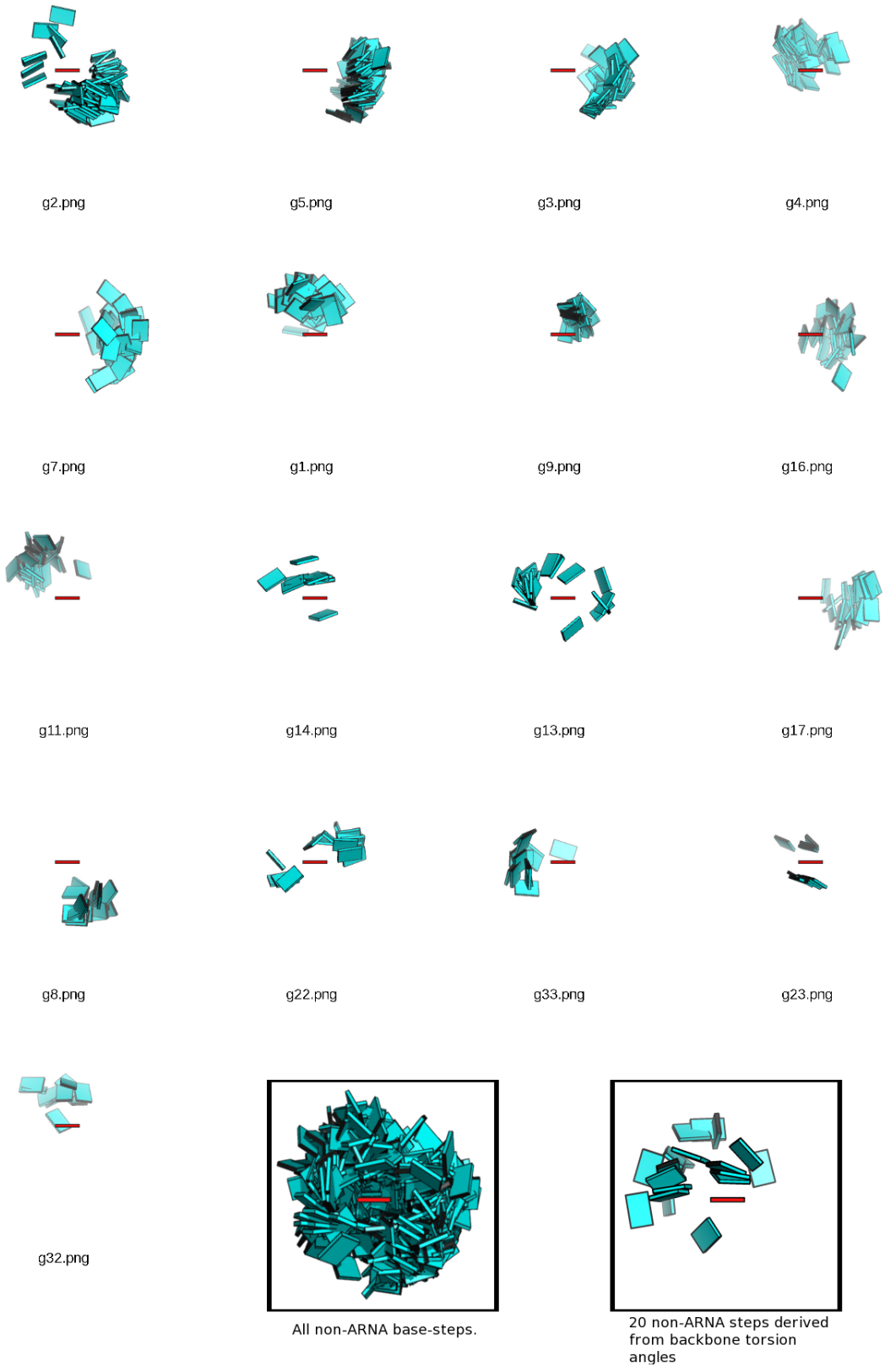


Figure 2.14: 17 out of the 67 groups clustered using the hierarchical clustering algorithm are drawn in a photograph contact sheet fashion. Each group is centered on the base reference frame of the adenine block drawn in red. In the lower right corner of the "contact sheet" the full space of 797 reconstructed steps is shown, along with the 20 steps derived from schneider et al. work. Notice how the only "hollow" side of the "onion" formed by the full space of base-step conformations is that corresponding to the watson-crick base-pairing region.

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] 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.

[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] 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.

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

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

[27] 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.

[28] Arnott, S. (1999) Oxford Handbook of Nucleic Acid Structure, Oxford Science Publications, .

Chapter 3

RNA Base-Pairing

3.1 Canonical and Noncanonical Base-pairs

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

In the helical regions data we quantify:

Abundances (Counts) Deformabilites Helical Context

NON-REDUNDANT DATABASE AND CONSTRAIN TO HELICAL REGIONS.

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

3.2 Clustering of Yurong's Classification

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

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

References

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

Chapter 4

RNA Base Pair Steps

4.1 Analysis (Albany Poster) and Django Webserver

Results shown in Albany and steps part of methods paper.

This gives us the force constant matrices per base-step which are used in the next section.

4.2 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 are used. Some common experimental techniques to determine a are 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 [**?**], Abels et al. [1], and Caliskan et al. [2]. We will use their results for comparison with those coming from the "realistic" model developed by Olson and collaborators [3] to describe DNA. The "realistic" model is dependent on high resolution crystallographic data. Initial studies started with small numbers of data for the deformabilities of the ten unique base-pair steps [3]. A more complete picture applied to the study of DNA sequence dependent deformability became available in 1998 [4]. The base-pair step deformability data for DNA has been constantly refined as more high resolution DNA and DNA-protein structures have been added to the Nucleic Acid Database (NDB) [5]. Although such data has been available for DNA since 1998, it had not been so for RNA, until now [6].

A detailed description of the "realistic model" along with the scheme of the C++ code developed by Czapla and Zheng to implement it, and a brief account of various definitions of persistence length and models from which a can be derived are included in Appendix D

4.3 AMBER: Persistence Length of Base-Pair Step Patterns

I guess it needs some input here in order to work on latex compilation.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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 bases or base pairs and the paramenters are the quantities that bring into coincidence coordinate frames on the two objects using ideas from classical mechanics. The first two sets of parameters are based on Cartesian coordinates, whereas the third set of helical coordinates, resembles cylindrical coordinates and is based on the single rotation that brings coordinate frames on the two bases 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 experimentally derived Cartesian coordinates of the atoms. With this experimental data one performs a least-squares fit to a standard reference frame [4]. This can be done using the octave script at `http://rutchem.rutgers.edu/~esguerra/RNA/scripts.html` as a tutorial example. The coordinate origin which is embedded in the standard reference frame is kept and used for both base and base pairs. In the case of single 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

[i] This is the most common case but the PDB file is sometimes the result of theoretical modeling.

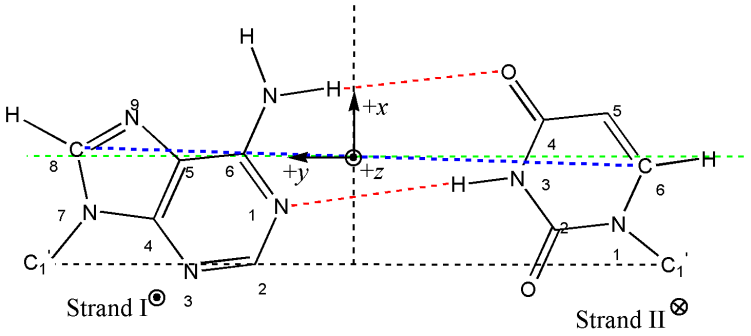


Figure A.1: Standard reference frame of an A-T base-pair [4]. The *y*-axis (dashed green line) is chosen to be parallel to the line connecting the C1′ of adenine and the C1′ of thymine associated in an ideal Watson-Crick base-pair. The *x*-axis is the perpendicular bisector of the C1′ - C1′ line, and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

Once one has determined the coordinate origins for the two consecutive bases or base pairs comprising a step one defines a middle step triad (MST) [5]. This can be described by the following procedure:

1) Find the angle Γ between consecutive normals, *i.e.*, *z*-axis. Since these are unit vectors, the angle is defined by the dot 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*-axis (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} 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 is the average of the position vectors 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). The dot product of the $\hat{z}_{MST}$ unit vector 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 product 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:

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

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

Now to get the remaining three translational degrees of freedom for step parameters (Dx, Dy, Dz) one just 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 is completely analogous to compute the base-pair parameters. 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 middle base triad MBT. The axis which are made to coincide are the *y*-axis and not the *z*-axis 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 three dimensional case of nucleic acid base steps what this means is that, instead of rotating around one reference-frame centered axis 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 one to define the orientation of a helical axis (or unique rotational-translational axis) as a unit vector given by equation 2.15:

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

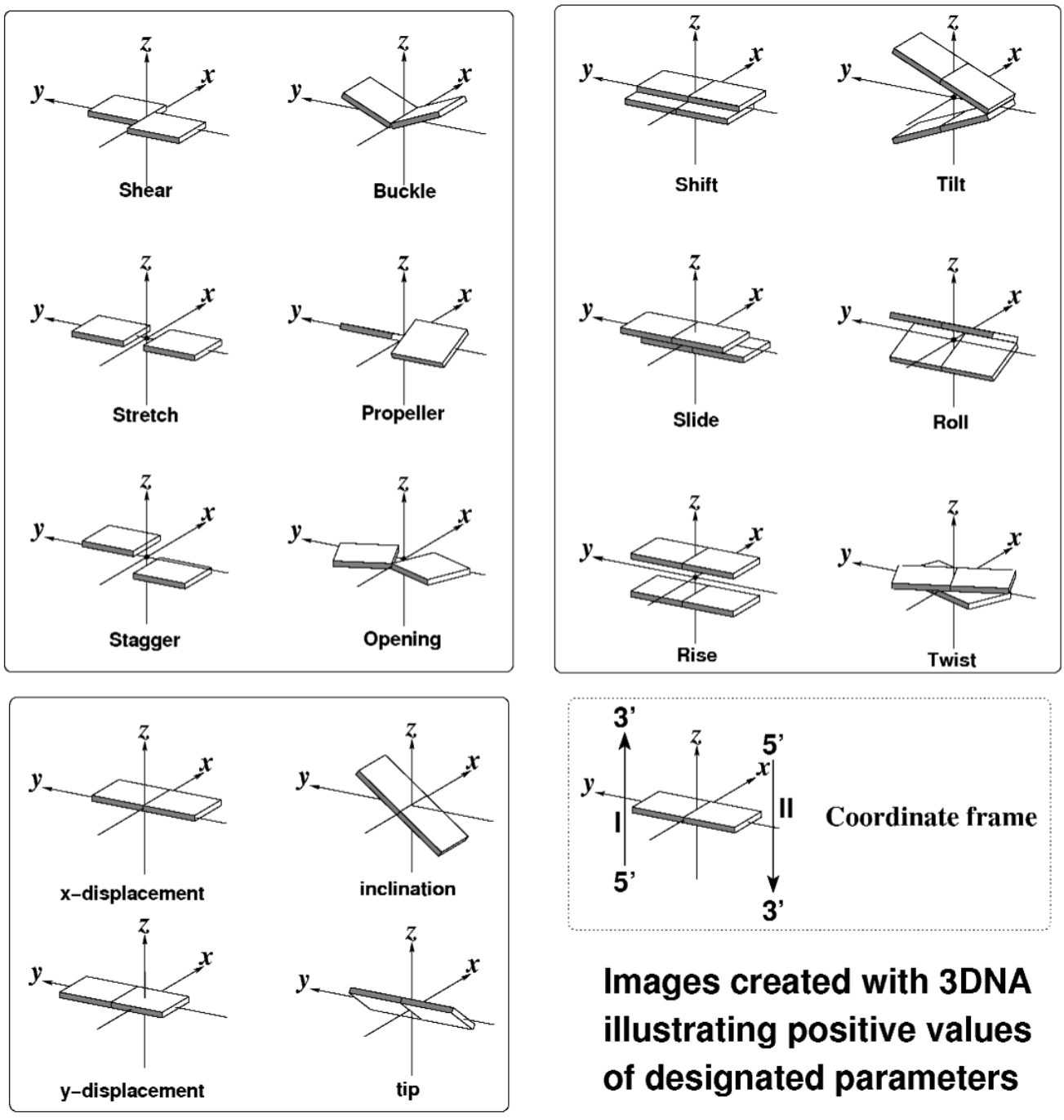


Figure A.2: Illustration of base pair and base step parameters [1]

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 structures as a vector composed of the six base step parameters. We grouped 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 the clustering used to construct the periodic table.

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's 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, 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.

With these distance definitions, we use a hierarchical clustering method.

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 description of these methods follows in the next section, "Hierarchical Methods".

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 2.4.

B.2 Hierarchical methods

The hierarchical clustering methods used were:

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

$$D(X,Y) = min\{d(x_i, y_j) : x_i \in X, y_j \in Y\} \tag{B.4}$$

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

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

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

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

$$D(X,Y) = \frac{1}{N_x * N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} d(x_i, y_j) \tag{B.6}$$

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

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

$$D(X,Y) = d(\overline{x}, \overline{y}) \tag{B.7}$$

$$\overline{x} = \frac{1}{N_x} \sum_{i=1}^{N_x} x_i \tag{B.8}$$

$$\overline{y} = \frac{1}{N_y} \sum_{i=1}^{N_y} y_i \tag{B.9}$$

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

$$D(X,Y) = ESS(XY) - [ESS(X) + ESS(Y)] \tag{B.10}$$

$$ESS(X) = \sum_{i=1}^{N_x} \left| x_i - \frac{1}{N_x} \sum_{j=1}^{N_x} x_j \right|^2 \tag{B.11}$$

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

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.

The first step is to chose a distance definition. We chose the Manhattan distance. The Manhattan distance values between structures can be displayed in a lower triangular matrix as seen in equation B.12

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

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

$$d(2,3) = |-2.00 - 6.00| + |2.00 - -2.00| = 12 \tag{B.13}$$

Now that we have calculated the distances we need a clustering method, in this case, we will use the average linkage clustering method. There are two hierarchical techiques called agglomerative, or bottom-up, and 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 objects constitute one final group. The first step is then to group whatever structures are closer, that is, structures 3 and 5 ($d(3,5) = 3$). Now we find the mean distance between the elements of this cluster and the remaining unclustered structures,

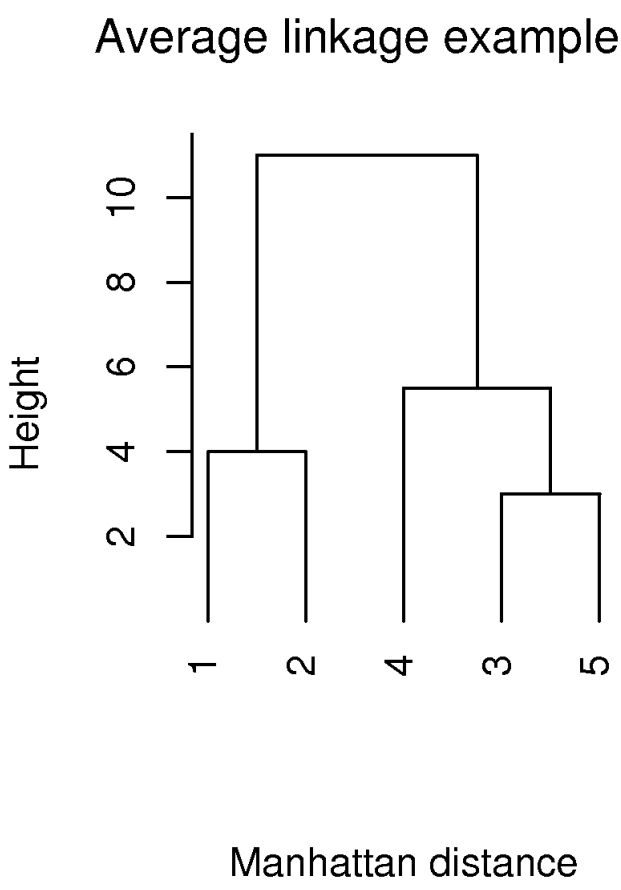


Figure B.1: Clustering tree for 5 bidimensional vectors using the Manhattan distance definition and the average linkage clustering method.

that is, structures 1, 2 and 4, we obtain the following mean distances

$$D(\{3,5\},1) = \frac{1}{2*1} * (8+11) = 4.5 \tag{B.14}$$

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

$$D(\{3,5\},4) = \frac{1}{2*1} * (5+6) = 5.5 \tag{B.16}$$

Since the distances between {3, 5} and all remaining unclustered vectors is higher than the distance between vectors 1 and 2 ($d(1,2) = 4$) then {1, 2} are grouped. The following value, in hierarchical increasing order is 4.5 between {3, 5} and 1 (see equation B.14), but since 1 and 2 are already grouped we can't group {3, 5} with 1. The next value, following the lower to higher hierarchy, is 5 ($d(3,4) = 5$), but we have already grouped 3 with 5, so we have to keep advancing in the hierarchy. The next value is 5.5, which corresponds to grouping {3, 5} with 4, so we cluster them. The only remaining possibility for grouping is, group {1, 2} and {4, 3, 5}, so we do it as illustrated in Figure B.1.

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.

Appendix D

Persistence Length

Nucleic Acids and other polymers can be understood as mechanical objects [1, 2] and therefore engineering approaches can be used for their understanding. Usually the methods followed by the engineering approach consider the polymer as a long continuous rod, and are known as continuum elastic theory. This type of approach leaves little space for taking into account the nature of the subunits which make up the polymer, that is, it is mainly applicable to homopolymers made up of identical subunits. In nucleic acids this is not necesarily the case, a more general approach takes into account the possibility of having different subunits making up the polymer. Olson and collaborators have developed a sequence dependent model, refered to as the "realistic" model [3]. Such model is harmonic and depends on determination of force-constant analogs derived from X-Ray cristallographic data taken from the Nucleic Acid Database (NDB) [4, 5]. Within the context of the "realistic" model Czapla et al. [6] have suggested a gaussian sampling methodology which allows the determination of global polymer properties like the persistence length a following a matrix approach suggested by Flory [**?**].

In what follows we summarize various definitions of persistence length and how it's computed using different models.

D.1 Persistence Lenght Definitions

"**Bend-persistence length**:
A length scale beyond which the elastic cost of bending is totally negligible" Philip Nelson, Yearbook of Science and Technology, McGraw Hill 1999

"In a randomly shaken rod any particular point in the rod will be pointing in a random direction, but nearby points will be pointing in roughly the same direction, that is, this nearby points are persistent. Points farther away than the bend-persistence length are said to be uncorrelated." Philip Nelson, Yearbook of Science and Technology, McGraw Hill 1999

"**twist-persistence length**:

that is, a rubber rod not only resists bending but also twisting”

"basic mechanical property that quantifies stiffness" Abels et al, Biophysical Journal, 2005, 2737-2744

"length at which the orientation of the sequential bonds which make up a polymer chain, stop being correlated. That is, if you have just two bonds, they will be correlated, which is the case in most molecules, but, in polymers, you have a long chain of sequential bonds. At some length, bonds will become uncorrelated, but up to that length they were correlated, this is what is meant by persistence length, and, in this context it's obvious that is an exclusive property of polymers."

"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

"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 $E\kappa_B T$ gives us the length of rod along which a thermally excited bend of 1 radian typically occurs: $lB/\kappa_B T$. This is called the persistence length..." John F. Marko and Simona Cocco, Physics World, March 2003

"persistence length, a, a measure of the distance over which the direction of the DNA is maintained. Mathematically, a is the mean projection in the limit of infinite chain length of a flexible DNA along its initial direction." [**?**]

"The persistence length a is a measure of the stiffness of a polymer chain and is related to the limiting value of the characteristic ratio at infinite chain length

$$a = \frac{\nu}{2}(C_\infty + 1) \tag{D.1}$$

"

Biopolymers can be either rigid or flexible.

They can be classified according to whether their persistence length (a) is greater, smaller, or similar to the contour length (L) of the polymer.

Notice that for $a \gg L$, there is a definition problem, since L has to be large enough to be a good approximation to the definition of persistence length, which is defined for an infinite chain length.

Model Type	Polymer Characteristic	a to L relation
Rigid Rod	Rigid	$a \gg L$
Gaussian chain	Flexible	$a \ll L$
Worm-like chain	Semi-flexible	$a \approx L$

Worm-like-chain = Porod-Kratky = Freely Rotating Chain in limit l=0 and n=infinity

Rigid biopolymers: actin, microtubules

Flexible biopolymers:

Semi-flexible biopolymers: High force extension DNA.

If the persistence length is of the same order of the length of the polymer, then the polymer is classified as semi-flexible

Polymer	a (nm)
α-helix	80-100
coiled-coil	150-300
Ideal DNA	51
Ideal RNA	70-80

Table D.1: Persistence lengths for some biopolymers with filament structures.

Think about the "energy" based perspective of Nicolas, and the stochastic based perspective of Flory and others.

D.2 end-to-end

The end-to-end vector r is the vector which connects the ends of a polymer chain. It can be defined as the sum of the vectors connecting the monomer units in a chain. These connecting vectors are sometimes called virtual bond vectors l.

From the end-to-end vector the quantity which is usually of interest is it's magnitude.

$$r = \sum_{i=1}^{n} l_i \tag{D.2}$$

$$r^2 = r \cdot r = \sum_{i,j} l_i \cdot l_j \tag{D.3}$$

Equation D.2, can also be written:

$$r^2 = \sum_{i} l_i^2 + 2 \sum_{i \neq j} l_i \cdot l_j \tag{D.4}$$

To describe a polymer it's necessary to think about the various conformations it can adopt due to its flexibility, therefore, it is important to think of polymer related quantities in terms of the average of their possible conformations. For the end-to-end vector the average of its values is denoted as $< r >$, and the average of its norm, also called the second moment of the end-to-end distribution, is denoted by $< r^2 >$:

$$< r^2 >= \sum_{i} < l_i^2 > +2 \sum_{i<j} < l_i \cdot l_j > \tag{D.5}$$

When there is no correlation between succesive bonds we can write:

$$< l_i \cdot l_j >= 0 \tag{D.6}$$

So that equation D.5 keeps only the bond auto-correlation term:

$$< r^2 >= \sum_{i} < l_i^2 >= n < l^2 > \tag{D.7}$$

This equation is used to describe a so-called freely-jointed chain.

D.3 Models

Nelson in book says:

$$dE = 12\kappa_B T[A\beta^2 + Bu^2 + C\omega^2 + 2Du\omega]ds \tag{D.8}$$

A$\kappa\beta$ T = Bend stiffness B$\kappa\beta$ T = Stretch stiffness C$\kappa\beta$ T = Twist stiffness D$\kappa\beta$ T = Twist-stretch coupling

If only the bend stiffness survives then the model is called an inextensible model, also Porod-Kratky, or WLC.

D.3.1 Kuhn - Freely Jointed Chain (FJC)

D.3.2 Porod-Kratky - Worm Like Chain (WLC)

D.3.3 Olson - Realistic

The Hamiltonian for a [7]

D.4 Suggested Reads

From Equilibrium Statistics of Plischke and Bergersen they suggest to read: Des Cloiseaoux and Janik () Rubinstein and Colby (Polymer Physics)

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] 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).

Supplement A

Figure Supplements

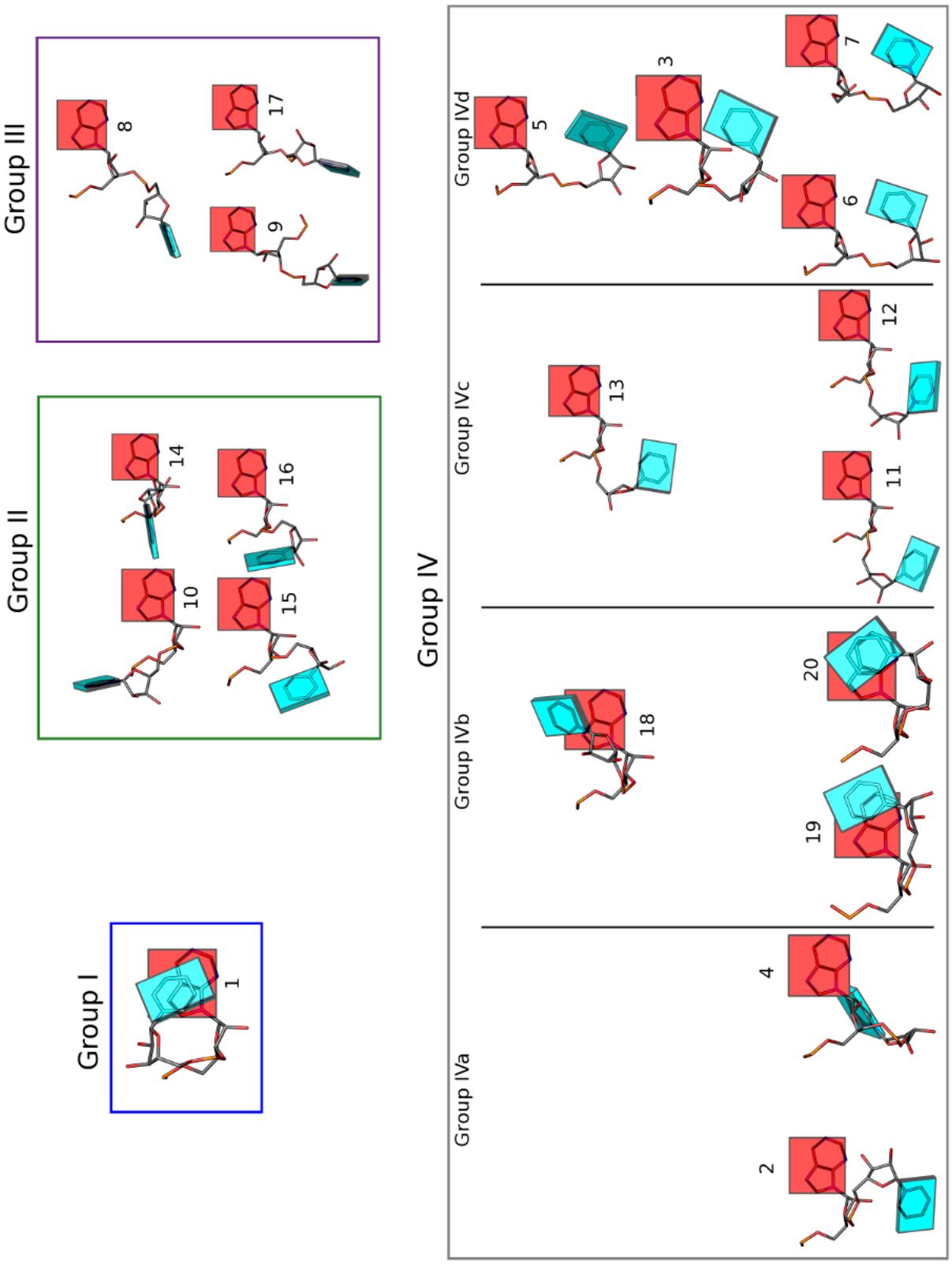


Figure S1: Non A-RNA Type base steps centered on the standard reference frame of Adenine. Top view with the Minor Groove side of Adenine pointing down the page and the Major Groove pointing up.

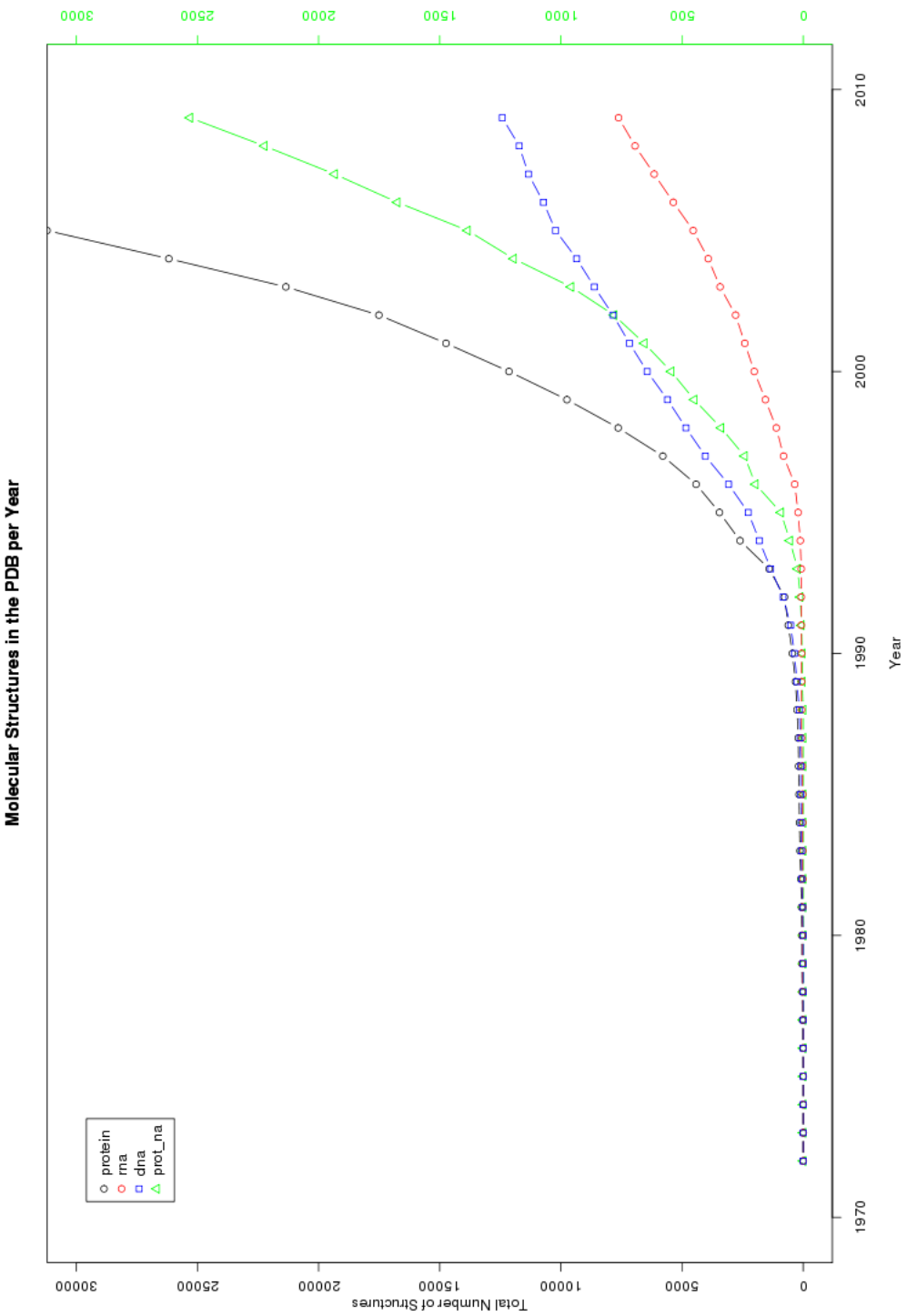


Figure S2: The total number of structures available in the pdb up to the end of year 2009. The scale of the axis in the left (in black), is ten times that in the right (in green). The black y-axis sets the scale for the number of protein structures available in the PDB up to the end of the year 2009.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