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.

Table of Contents

Abstract ii

Acknowledgements iv

List of Tables vii

List of Figures viii

1. Introduction 1

1.1. RNA chemistry 1

1.2. Standard reference frame and local parameters 3

1.2.1. Base-pair and base-step parameters 5

1.2.2. Local helical parameters 8

1.3. RNA folding 10

1.4. Is RNA folding a hard or easy problem? 10

1.5. Experimental folding techniques 11

1.6. RNA simulations 12

1.6.1. Local nucleotide interactions 12

1.6.2. RNA secondary structure algorithms and the lack of tertiary ones 14

1.6.3. RNA overall fold 14

1.6.4. RNA motifs 15

1.7. Overview 17

References 18

2. RNA Base Steps 25

2.1. Consensus Clustering of Single Stranded Base Step Parameters 28

2.1.1. Combining Fourier Averaging Results and Clustering Analysis 29

2.1.2. Selection of a Clustering Methodology 32

References . . . 43

3. RNA Base-Pairing . . . 37

3.1. Canonical and Noncanonical Base-pairs, Methods Paper . . . 37

3.2. Clustering of Yurong's Classification . . . 37

4. RNA Base Pair Steps . . . 38

4.1. Analysis (Albany Poster) and Django Webserver . . . 38

4.2. Persistence Length vs. Hagerman . . . 38

4.3. AMBER: Persistence Length of Base-Pair Step Patterns . . . 38

5. RNA Motifs . . . 39

5.1. GNRA tetraloop . . . 39

5.1.1. 3DNA-Parser . . . 39

5.1.2. Overlap Scores . . . 39

5.2. Triplets on RNA (comparison to Laing et al.) . . . 40

References . . . 43

6. RNA Helical Regions and Graph Theory . . . 44

Appendix A. Clustering Analysis (CA) . . . 45

A.1. General Methodology . . . 45

A.2. Hierarchical methods . . . 46

References . . . 50

Appendix B. Dimension Reduction . . . 48

B.1. Principal Component Analysis . . . 48

References . . . 49

Appendix A. Figure Supplements . . . 50

S1. Supplement Figures for Chapter 2 . . . 50

Curriculum Vitae . . . 52

List of Tables

2.1. Some large RNA structures (>300 bases) elucidated in the last decade. 28

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. 32

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. 39

A.1. Example of structures, considered as bidimensional vectors, to be clustered using the average linkage method and the Manhattan distance. 48

List of Figures

1.1. A single strand of RNA drawn in the 5′ to 3′ sense showing the main chemical entities which compose it; base, sugar, and backbone. The four bases (A, G, C, U) are colored according to the NDB (Nucleic Acid Database) convention [14], the backbone is colored gray, and the sugars black. The bases G, and C, and the furanose sugar are numbered according to the IUPAC rules [15]. This figure is a reproduction of Figure 2.1, in Wolfram Saenger's book [16]. 2

1.2. Saenger base-pairing classes, reproduced from his book, "Principles of Nucleic Acid Structure". [16]. 4

1.3. **Left:** Backbone and Sugar torsion angles. **Right:** The most common sugar pucker conformations in RNA, that is, C3'endo and C2'endo, reproduced from Wolfram Saenger's, "Principles of Nucleic Acid Structure". [16]. 5

1.4. Standard reference frame of an A-T base-pair. 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. 6

1.5. Illustration of base pair and base step parameters [23] 9

1.6. Separation of secondary and tertiary interaction in RNA [44]. 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. 11

1.7. Ribbon-coil schematic illustraring the fold and intermolecular units of a dimer of prealbumin (PDB_ID:2pab), or transthyretin, taken from Richardson *et al.* [95] 15

1.8. 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 [96] 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. 16

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). 25

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. 27

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. 28

2.4. Dendrogram showing the results of consensus clustering of 20 non-Atype rRNA dinucleotides according to their hexadimensional base-step parameter vectors. 30

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. 31

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 . . . 33

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. 34

2.8. **Left**: With an RMSD cutoff of less than or equal to 10 Å, we identify three steps from the 23S subunit. **Right**: With an RMSD less than or equal to 15 Å we get a total of seven structures, where we clearly see that three of them are farther from the original Group I main structure of Figure 2.5 35

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. 36

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. 37

2.11. Cluster validity scores for stability measures. 38

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. . . . 39

2.13. Cluster validity scores for the non-ARNA dataset. It can be seen clearly that the optimal method for clustering is the hierarchical one, as measured by lower values in the connectivity scores, and higher values in the Dunn score. The optimal number of clusters given by the dunn score is 67, we also see shoulders at $k = 67$, for the connectivity and silhouette scores. 41

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. 42

5.1. GNRA Tetraloop from *Thermus Thermophilus* 23S Ribosomal RNA PDB-ID:1ffk. 40

5.2. Normalized histograms showing the distribution of overlap values in the 23S subunit or *Thermus Thermophilus* rRNA, PDB-ID:1jjk. In histogram (a) all values are included, but in histogram (b) only values greater than zero are included. Notice the high preponderance of zero values, exactly 897 out of a total of 2705. 41

5.3. Dendrogram for consensus clustering of overlap scores in the ribosome. Zero values filtered out and remaining data normalized. 42

A.1. Clustering tree for 5 bidimensional vectors using the Manhattan distance definition and the average linkage clustering method. 49

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. 51

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. 52

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].

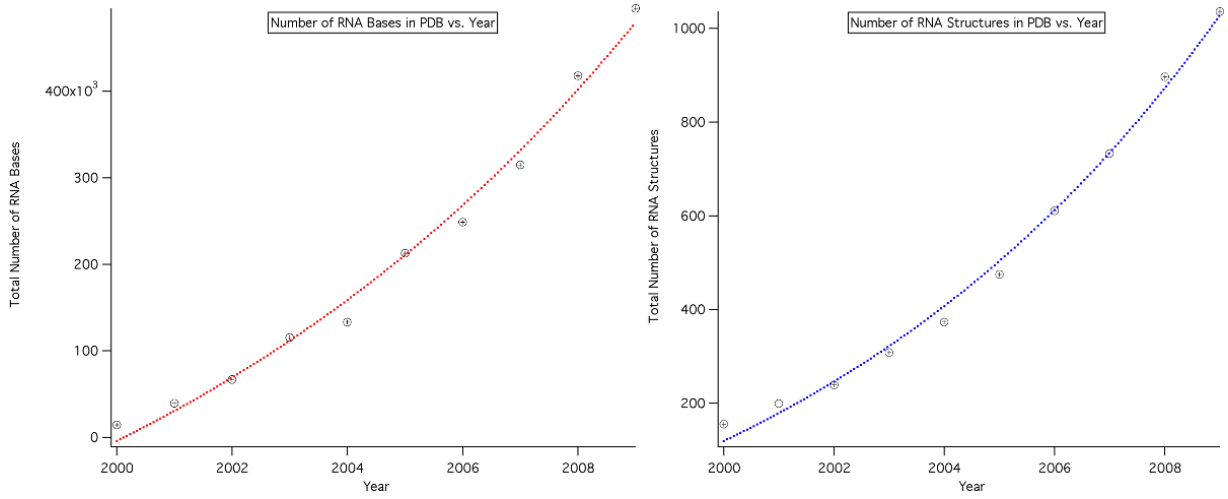


Figure 2.1: **Left:** Total number of RNA bases added to the PDB database between 2000 and 2010 (Exponential fit line in blue). **Right:** Total number of RNA structures solved yearly by X-Ray crystallography between 2000 and 2010 (Exponential fit line in red).

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.

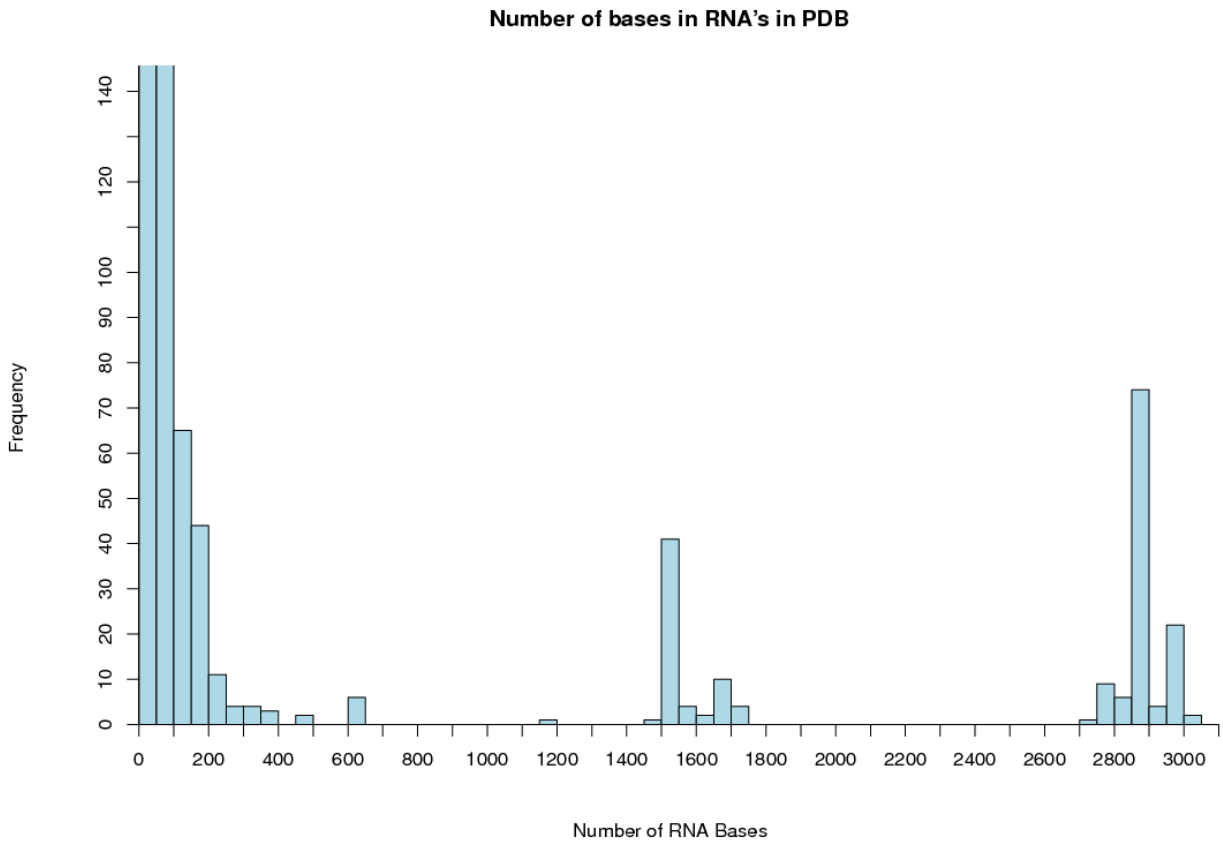


Figure 2.2: Frequency of nucleotide bases in RNA molecules found in the PDB classified by the size of RNA molecules. We define the size as the total number of nucleotide bases present per molecule.

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]

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.

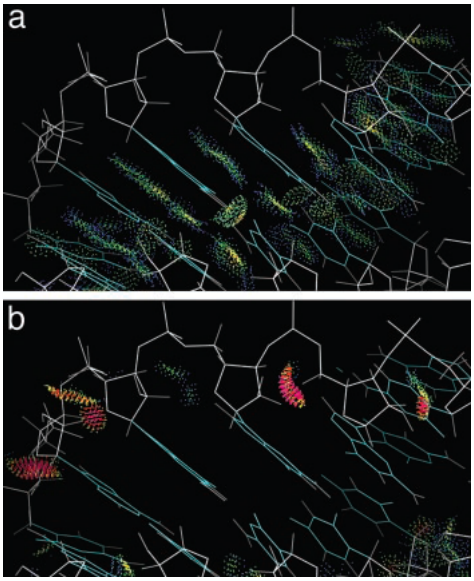


Figure 2.3: Figure taken from Richardson et al. [11] where the blue and green dots in a) mean very accurate van der Waals distances, and in b) the red and orange dots mean steric clashes, that is, distances outside the acceptable van der Waals range.

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 A) 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.

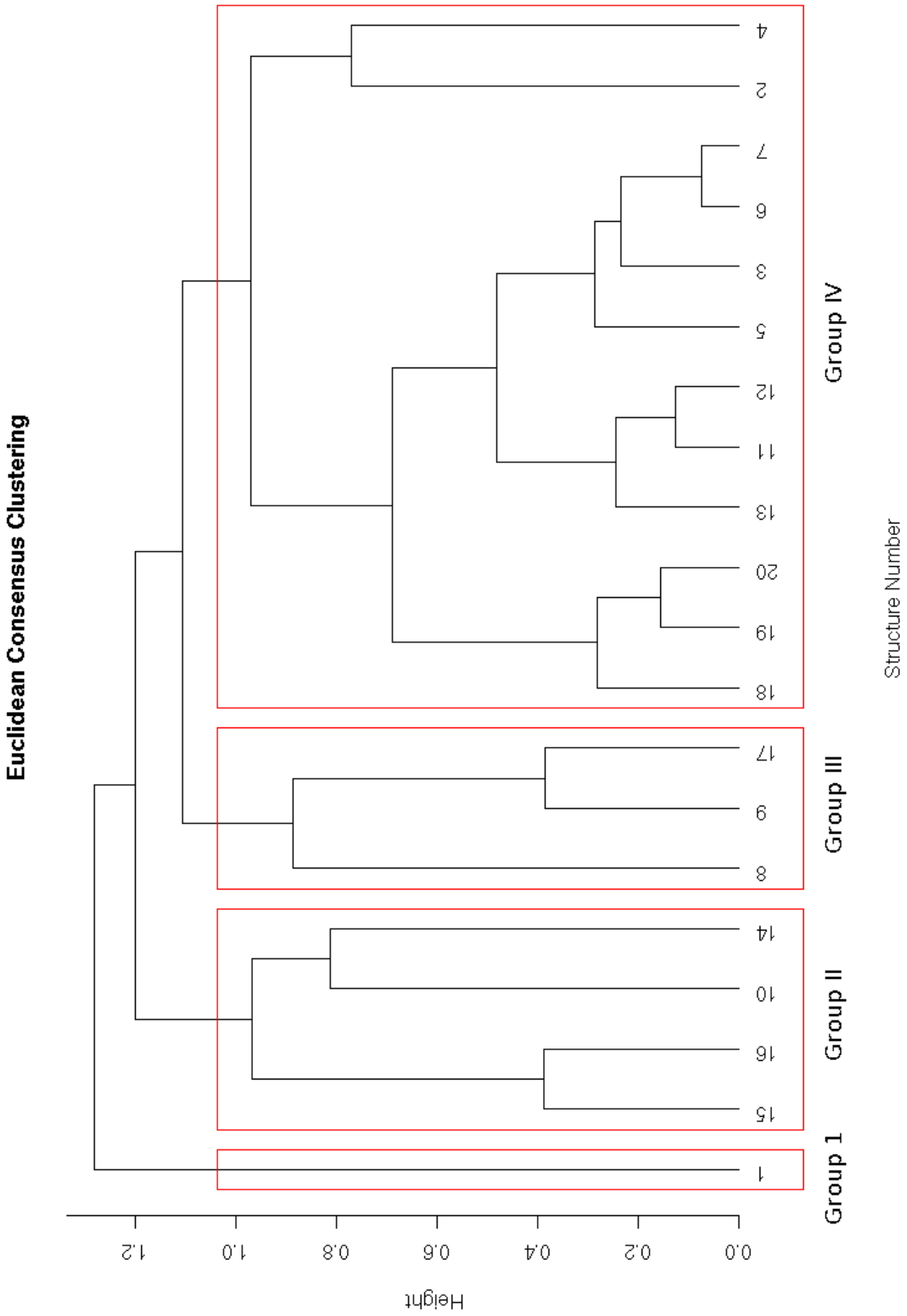


Figure 2.4: Dendrogram showing the results of consensus clustering of 20 non-Atype rRNA dinucleotides according to their hexadimensional base-step parameter vectors.

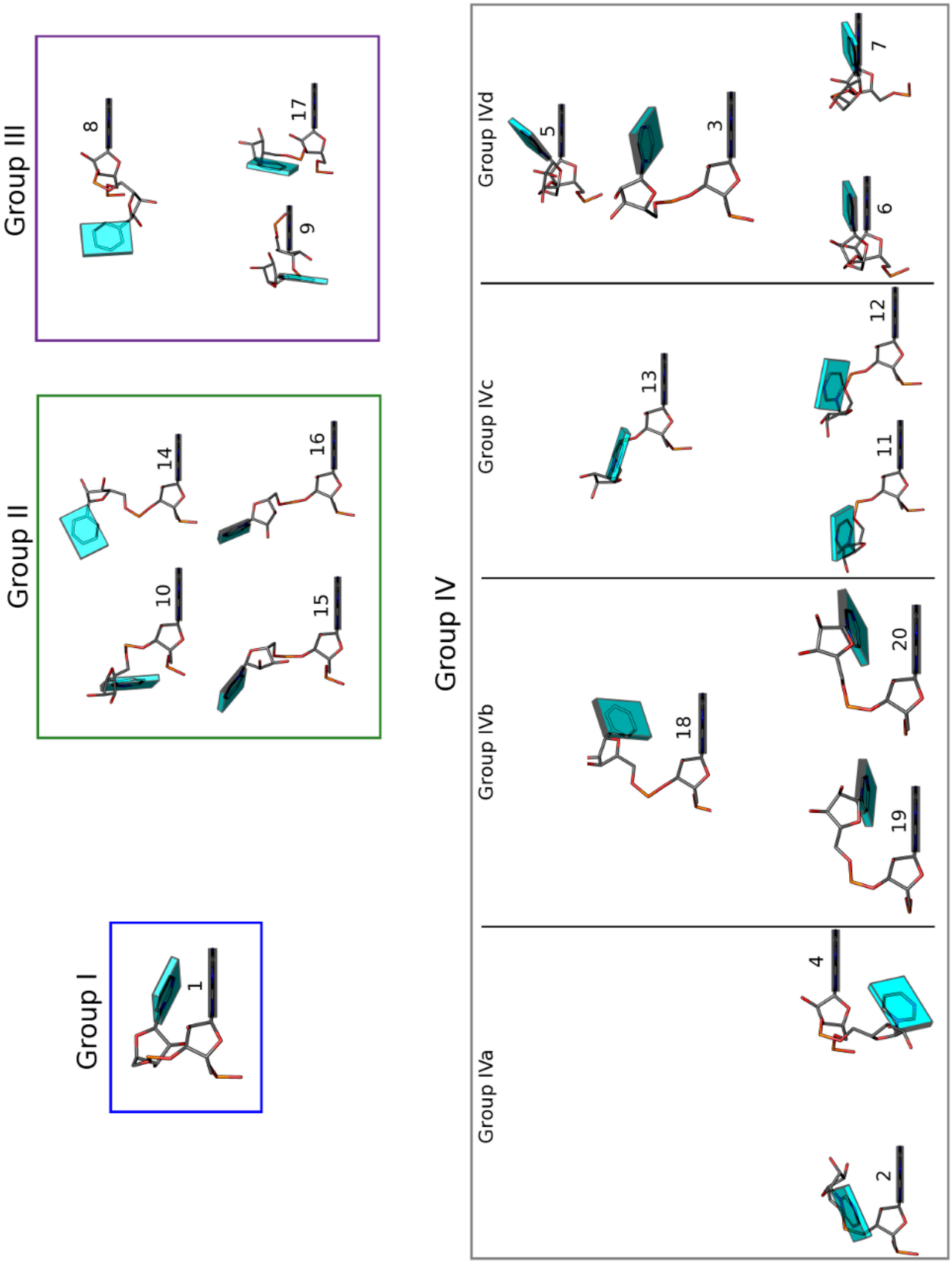


Figure 2.5: RNA dinucleotide structures organized by clusters obtained from consensus clustering of their hexadimensional base-step parameter vectors. The structures have been centered on the reference frame of the first step, that is, the adenine base, and the minor groove face of the rigid block parameter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

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 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

[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.

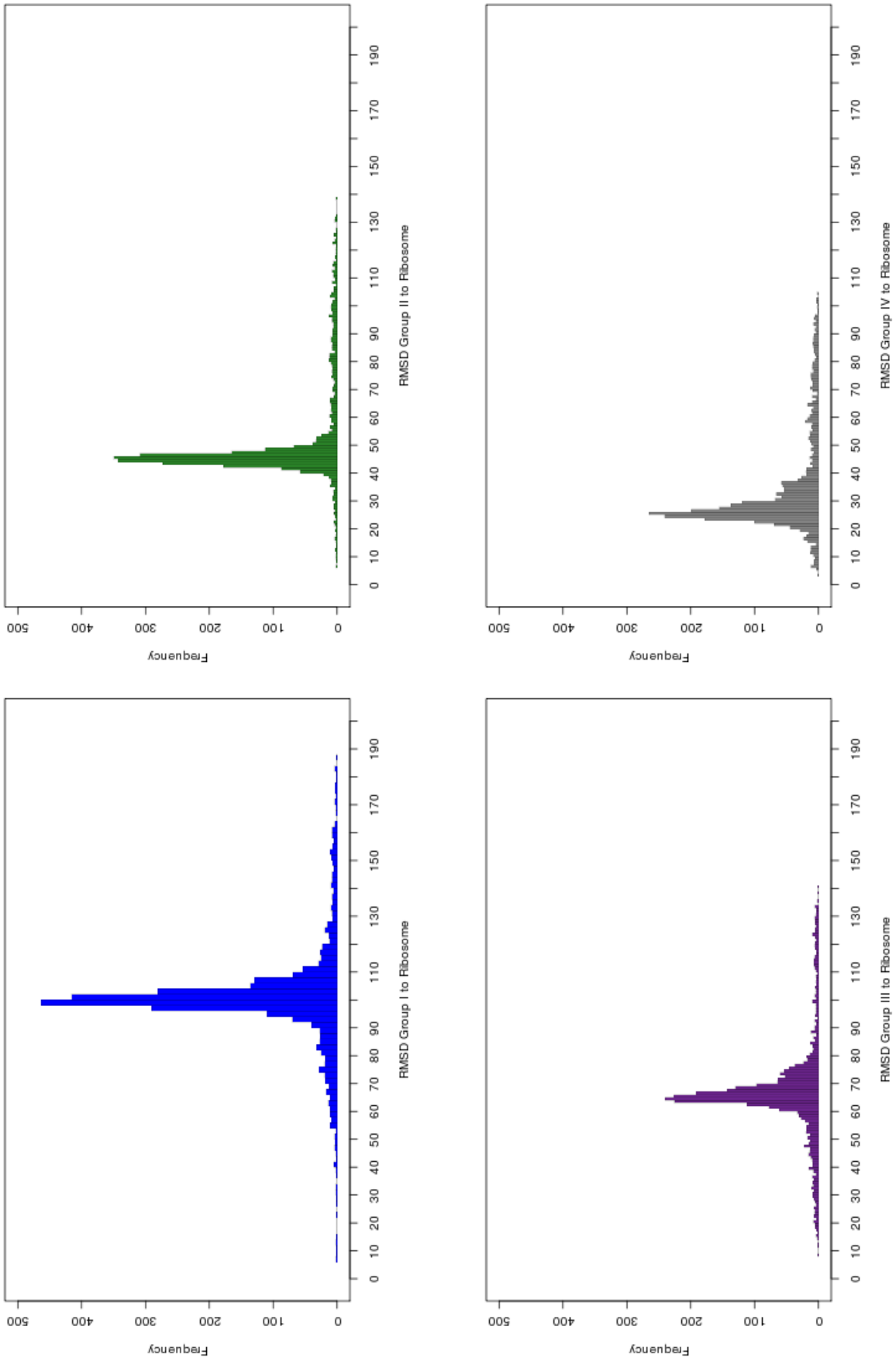


Figure 2.6: Root mean square deviation of the main four groups show in Figure 2.5. The color of the histograms is the same as that of the boxes surrounding the structures of Figure 2.5

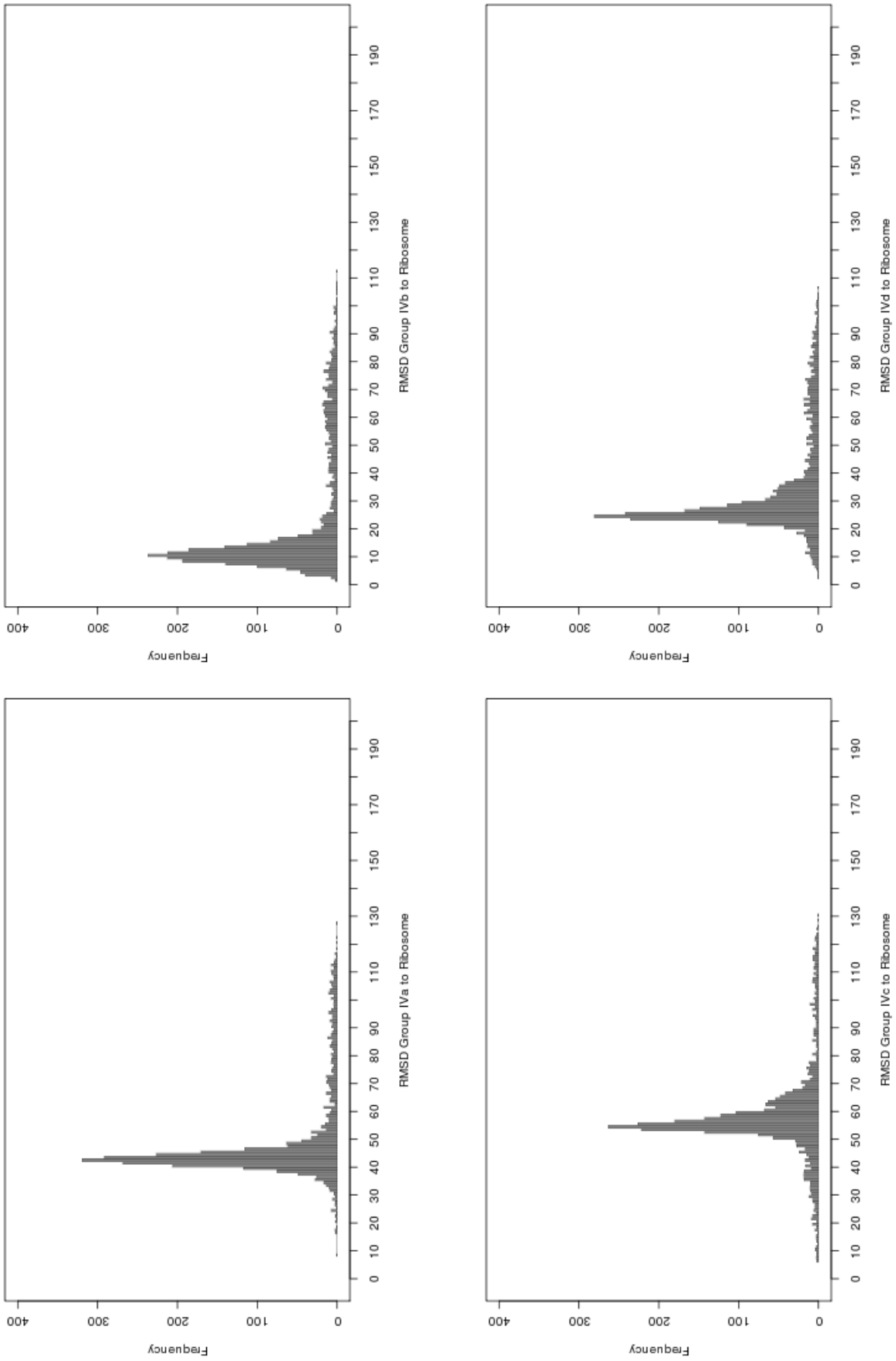


Figure 2.7: Root mean square deviation histograms for the subgroups present in group IV. Since subgroup IVb is composed of A-RNA like conformations we see in the upper left histogram that the highest proportion of small RMSD values belongs to this group.

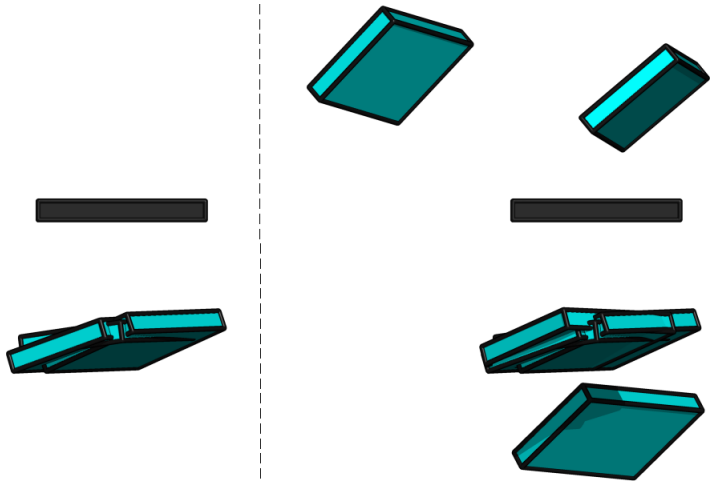


Figure 2.8: **Left**: With an RMSD cutoff of less than or equal to 10 Å, we identify three steps from the 23S subunit. **Right**: With an RMSD less than or equal to 15 Å we get a total of seven structures, where we clearly see that three of them are farther from the original Group I main structure of Figure 2.5

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 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

[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.

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 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.

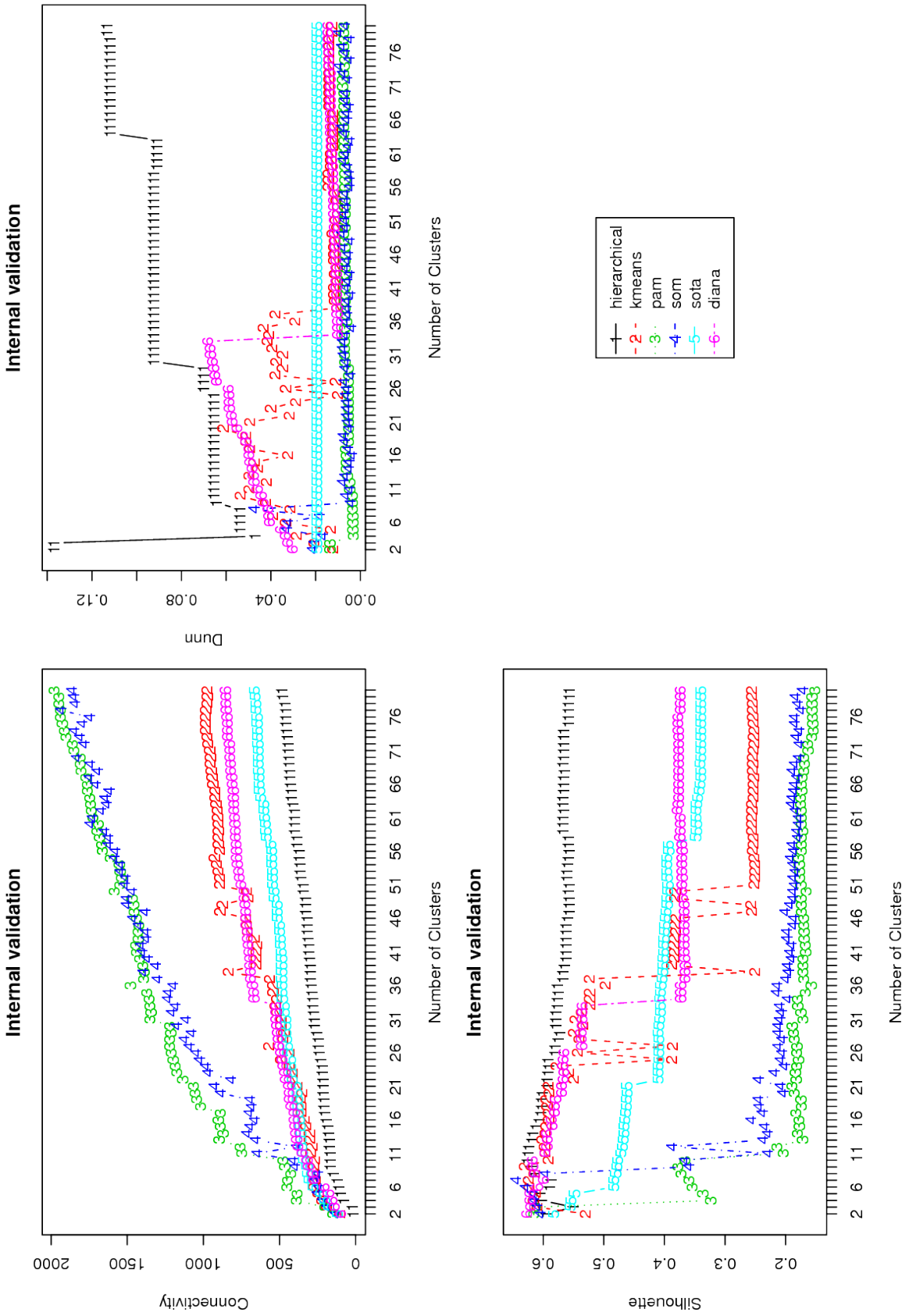


Figure 2.10: Cluster validity scores for internal measures. Notice how the hierarchical method, labeled as 1 in black color, behaves better for the whole range of Connectivity (smaller values) and Dunn (higher values), and it also outperforms all others after $k = 12$ for Silhoutte (higher values) scores.

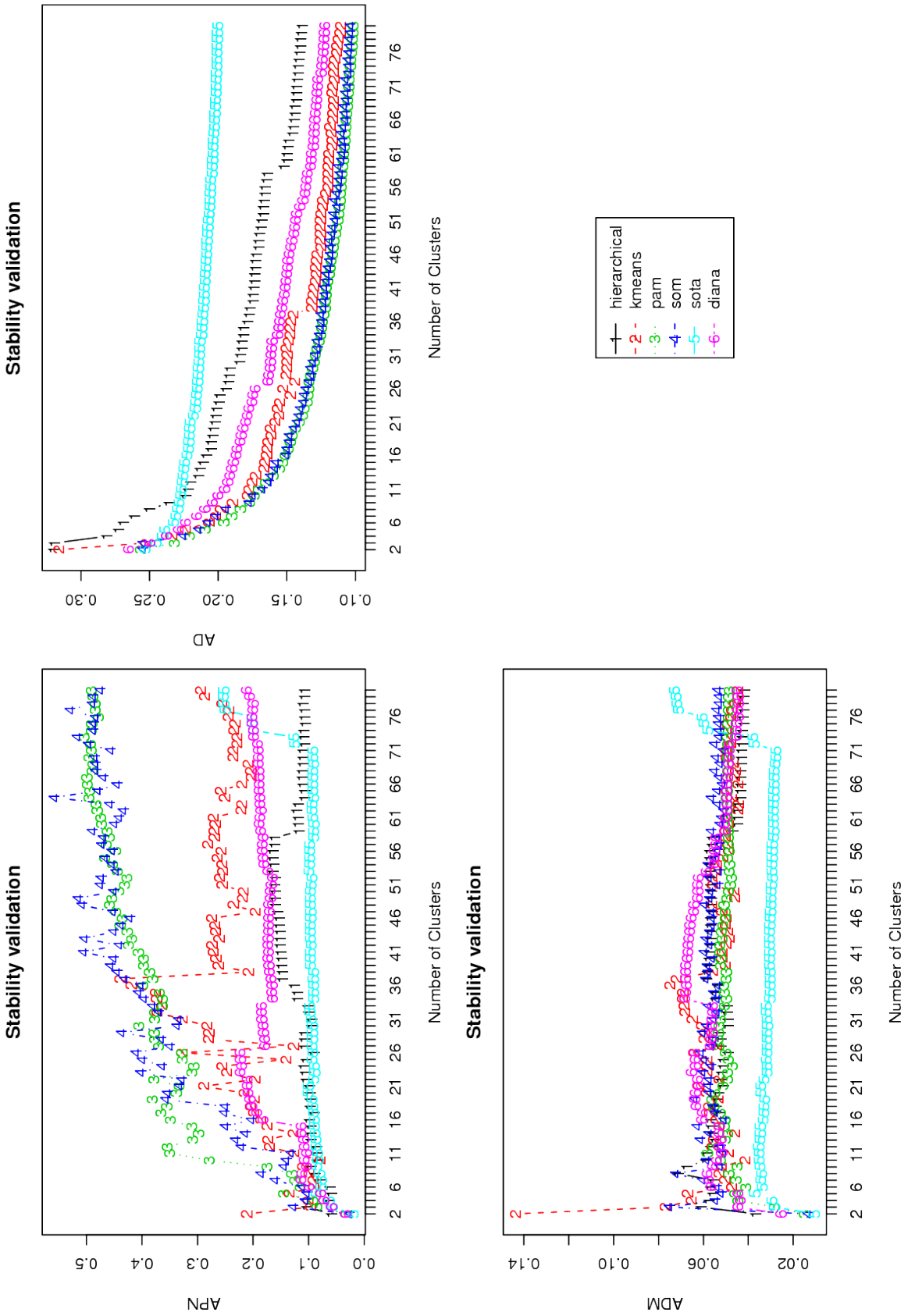


Figure 2.11: Cluster validity scores for stability measures.

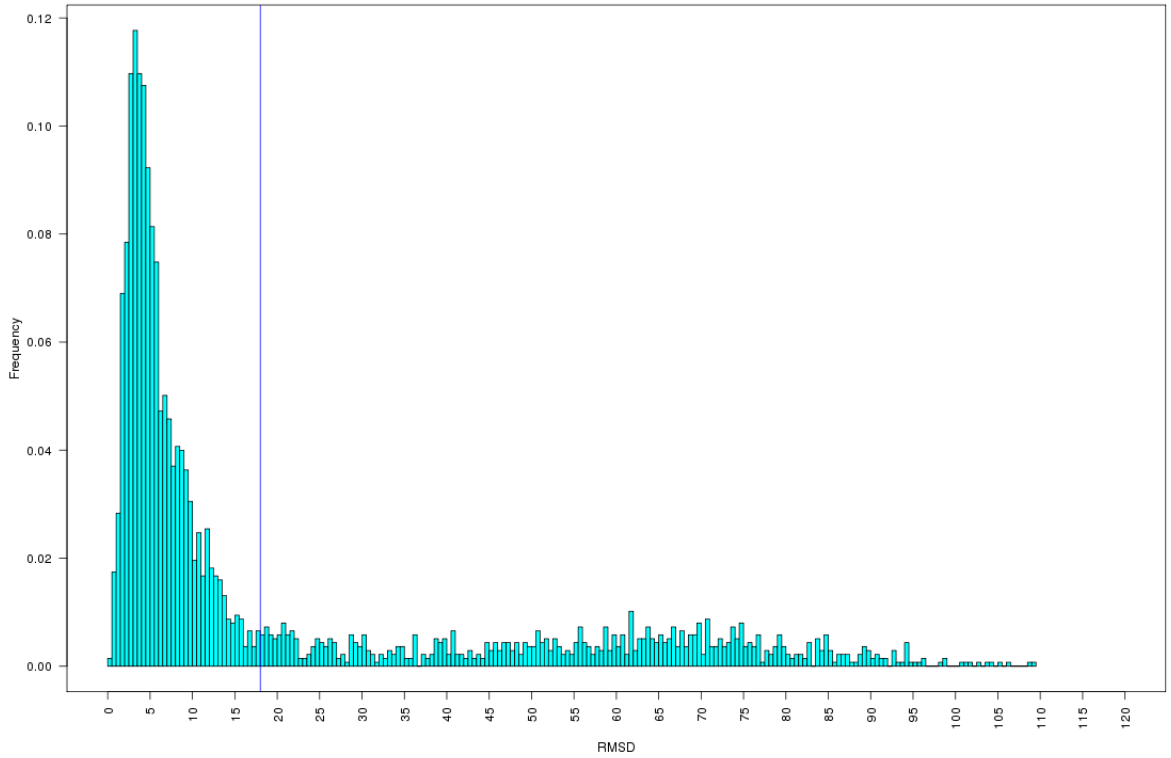


Figure 2.12: RMSD values between base-step parameters of the 23S subunit of ribosomal RNA and the standard base-step parameters derived from Arnott and collaborators [24] work.

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.

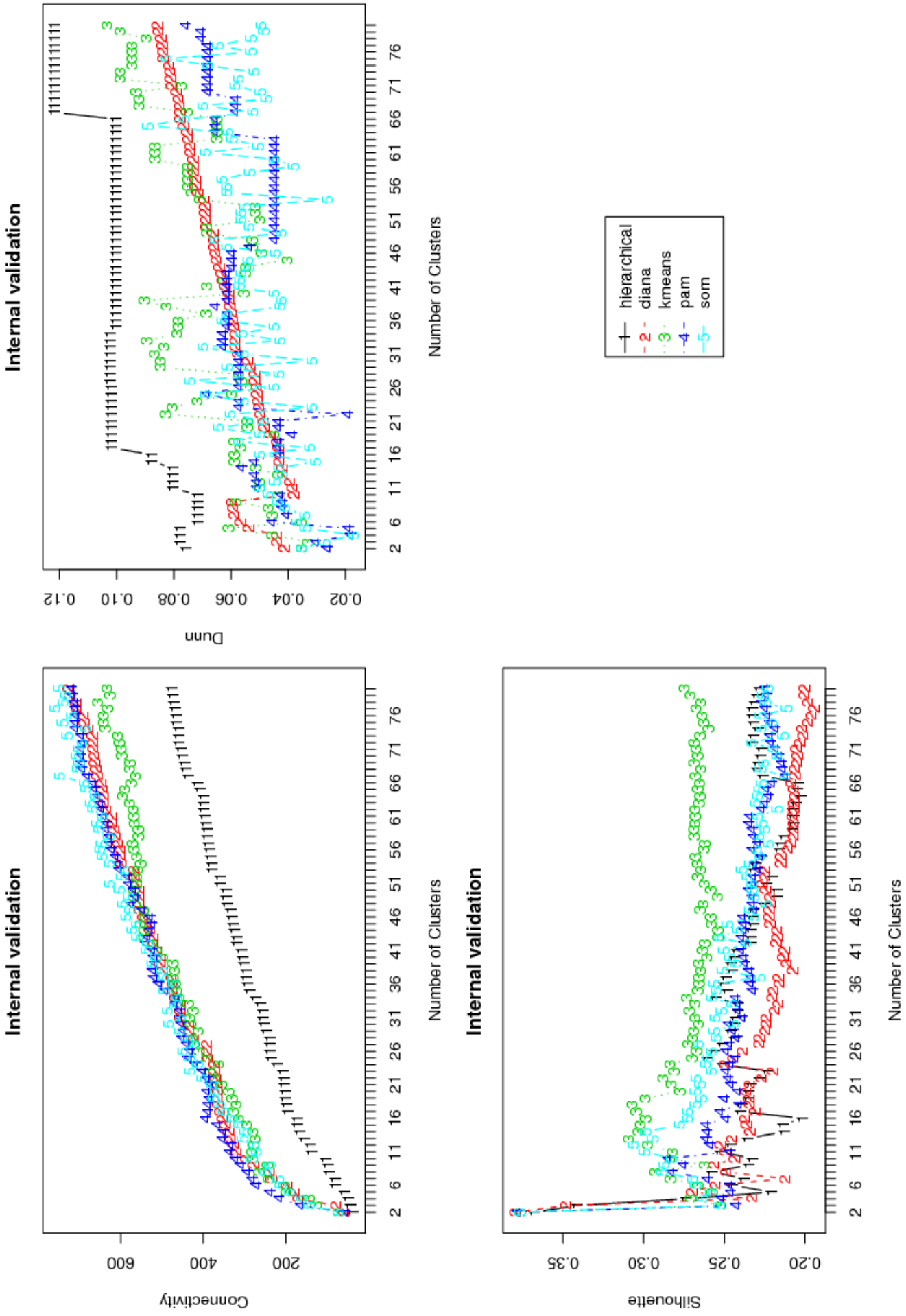
Structure Name	Shift (D_x)	Slide (D_y)	Rise (D_z)	Tilt (τ)	Roll (ρ)	Twist (Ω)	**Reference**
A-DNA	0.36	-1.39	3.29	2.46	12.50	30.19	
B-DNA	0.44	0.47	3.33	4.63	1.77	35.67	
A-RNA	-0.08	-1.48	3.30	-0.43	8.64	31.57	Arnott
A'-RNA	0.05	-1.88	3.39	-0.12	5.43	29.52	Arnott
AII-RNA	1.01	-2.52	3.33	2.94	9.75	25.12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

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.

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 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
200
400
600
Number of Clusters
2 6 11 16 21 26 31 36 41 46 51 56 61 66 71 76
Internal validation
Dunn
0.02
0.04
0.06
0.08
0.10
0.12
Number of Clusters
2 6 11 16 21 26 31 36 41 46 51 56 61 66 71 76
hierarchical
diana
kmeans
pam
som
Internal validation
Silhouette
0.20
0.25
0.30
0.35
Number of Clusters
2 6 11 16 21 26 31 36 41 46 51 56 61 66 71 76

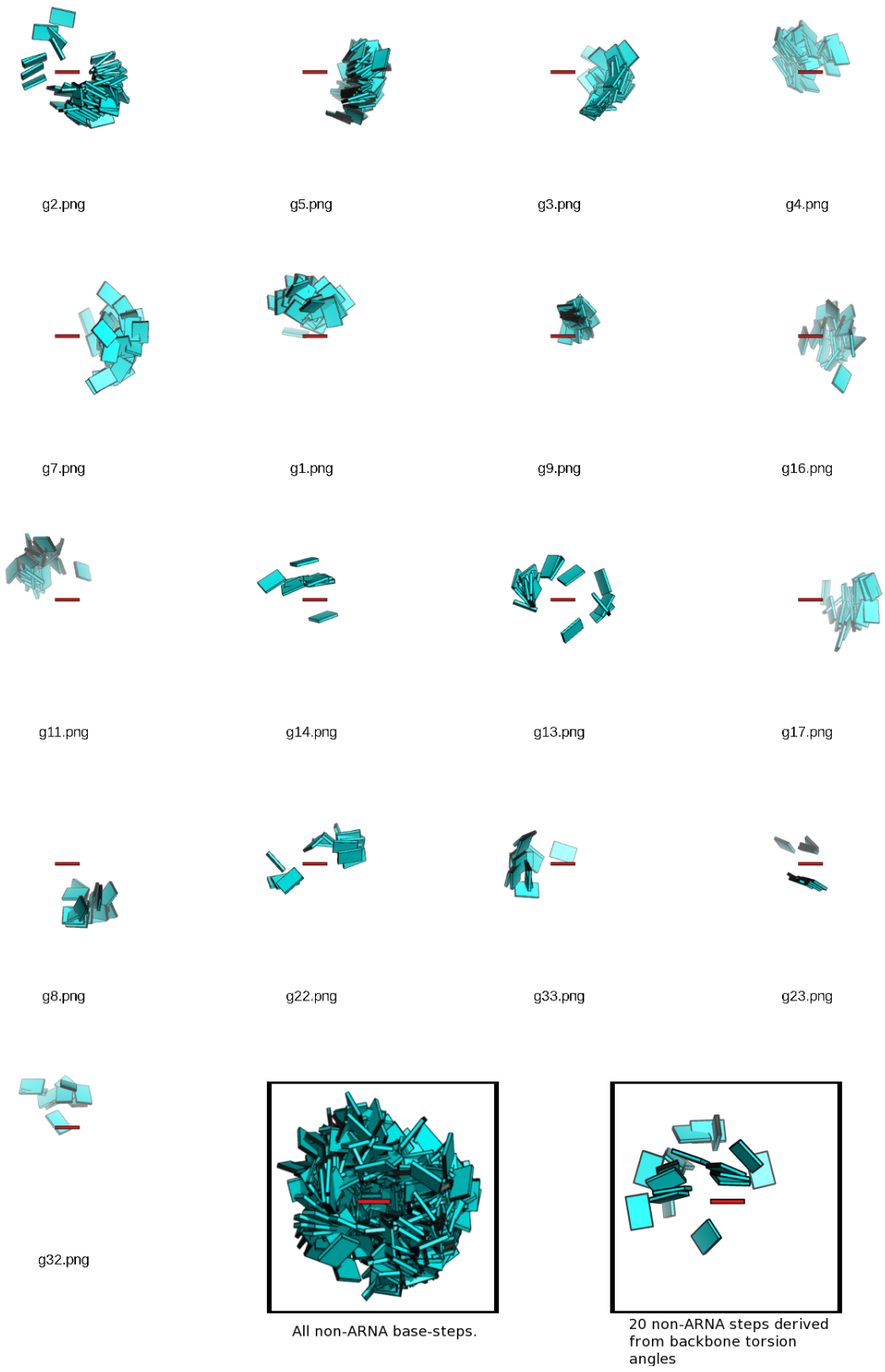


Figure 2.14: 17 out of the 67 groups clustered using the hierarchical clustering algorithm are drawn in a photograph contact sheet fashion. Each group is centered on the base reference frame of the adenine block drawn in red. In the lower right corner of the "contact sheet" the full space of 797 reconstructed 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.

Appendix A

Clustering Analysis (CA)

A.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,

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

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 A.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

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

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:

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

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.

A.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.

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

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.

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

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

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

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.

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

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

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

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

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

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

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 A.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 A.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 A.12

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

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

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

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,

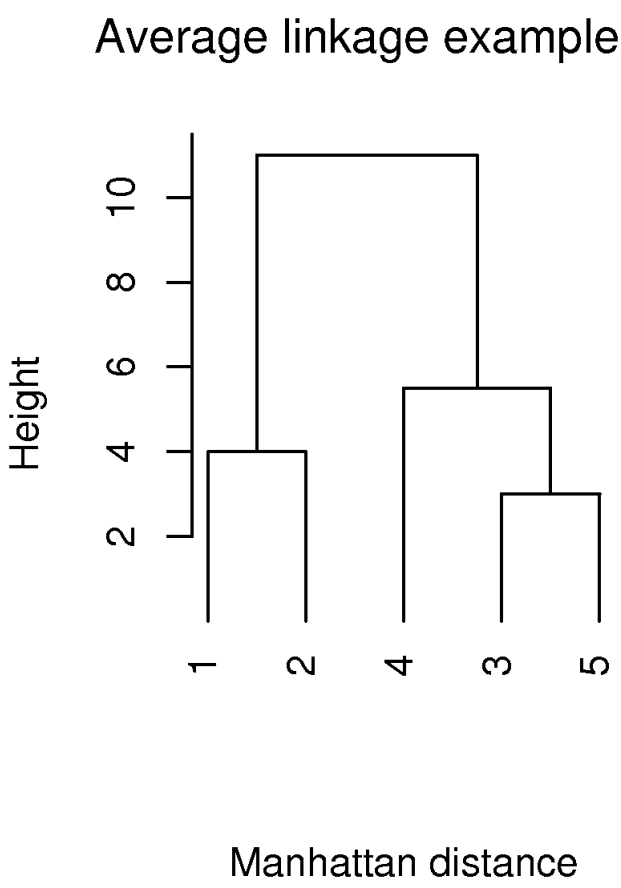


Figure A.1: Clustering tree for 5 bidimensional vectors using the Manhattan distance definition and the average linkage clustering method.

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

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

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

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

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 A.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 A.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.

Supplement A

Figure Supplements

S1 Supplement Figures for Chapter 2

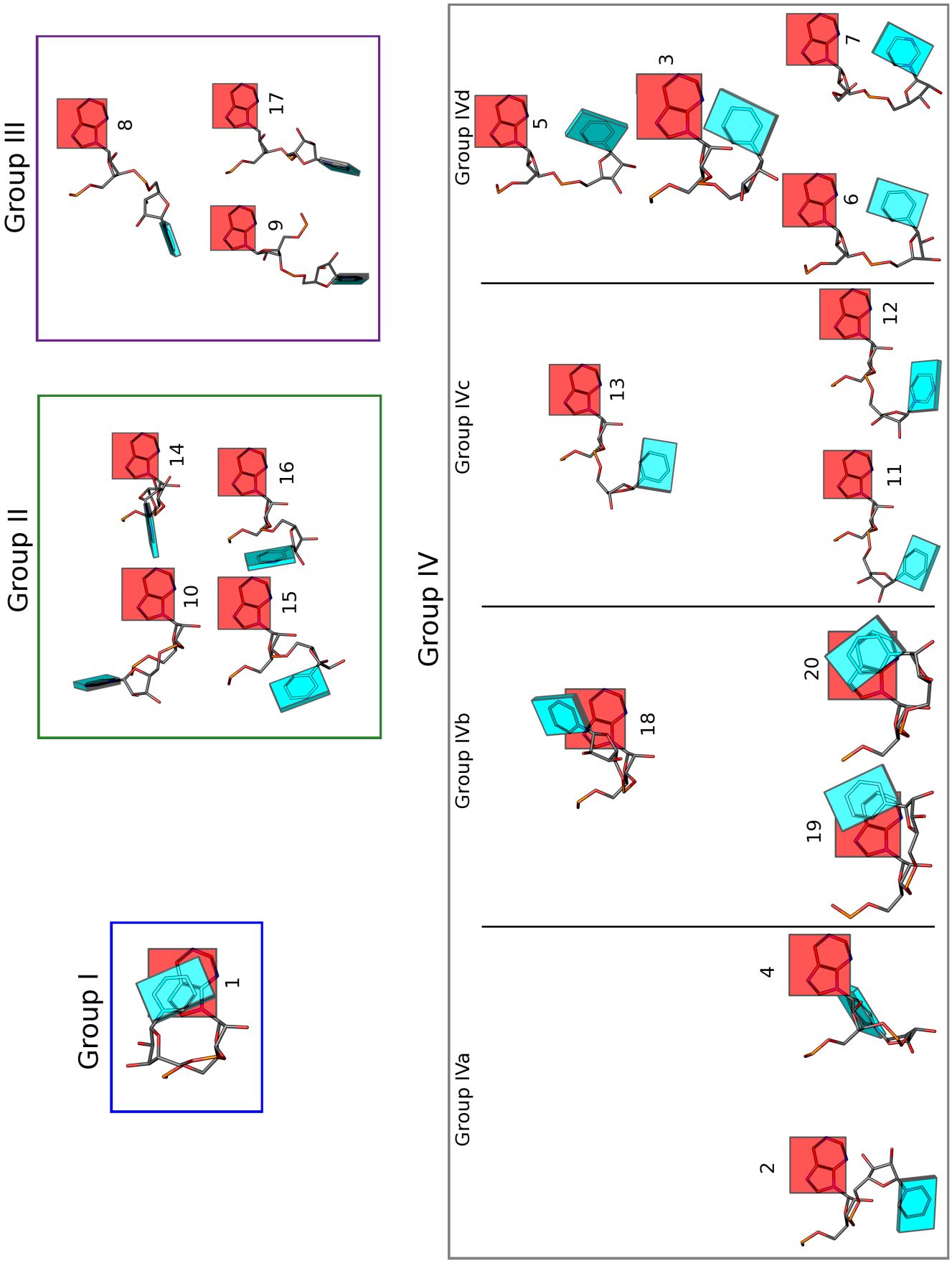


Figure S1: Non A-RNA Type base steps centered on the standard reference frame of Adenine. Top view with the Minor Groove side of Adenine pointing down the page and the Major Groove pointing up.

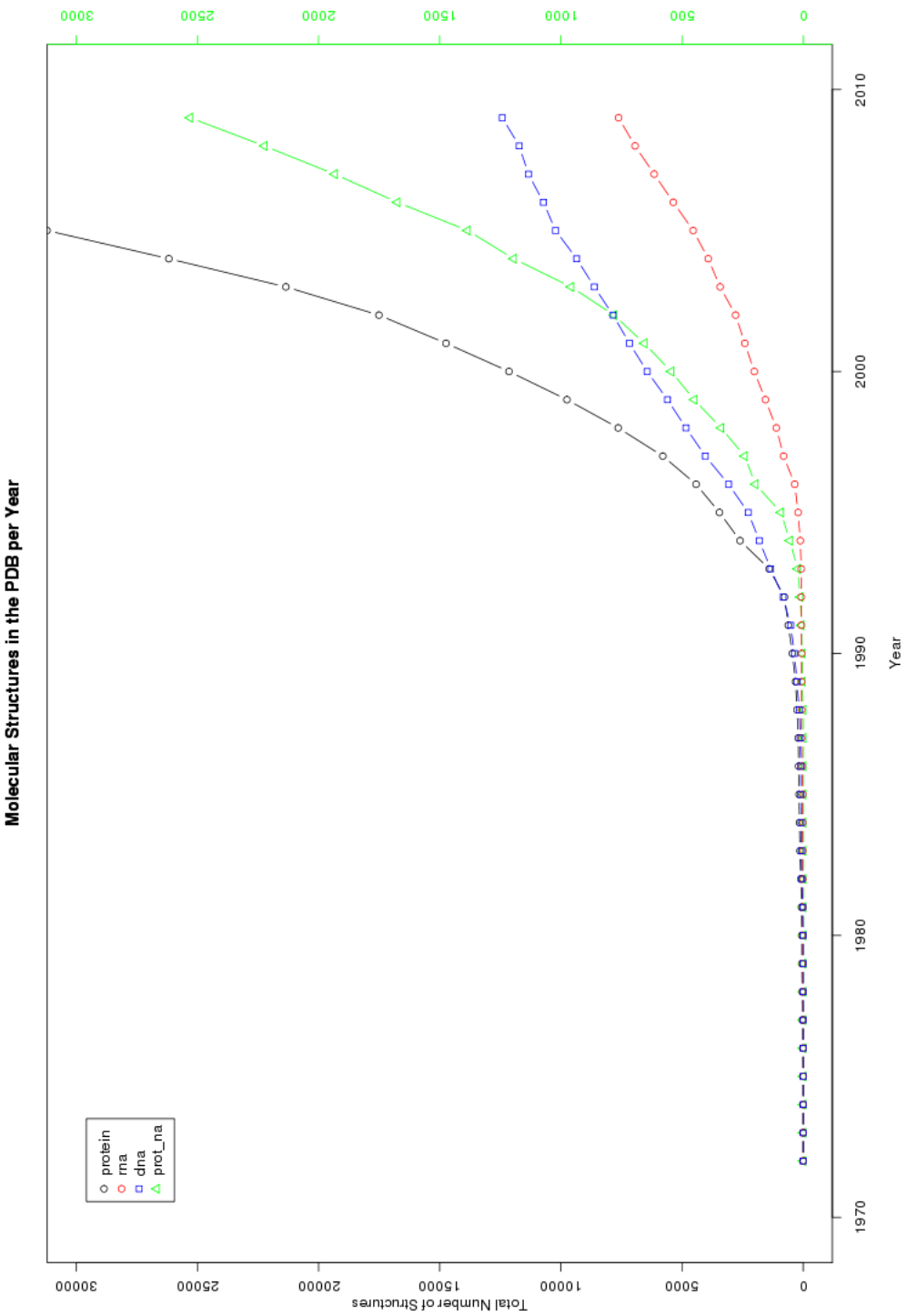


Figure S2: The total number of structures available in the pdb up to the end of year 2009. The scale of the axis in the left (in black), is ten times that in the right (in green). The black y-axis sets the scale for the number of protein structures available in the PDB up to the end of the year 2009. 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.