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 efforst 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

List of Tables

List of Figures

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, but starting in the year 2000 a vast amount of RNA structural information has become available with 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].

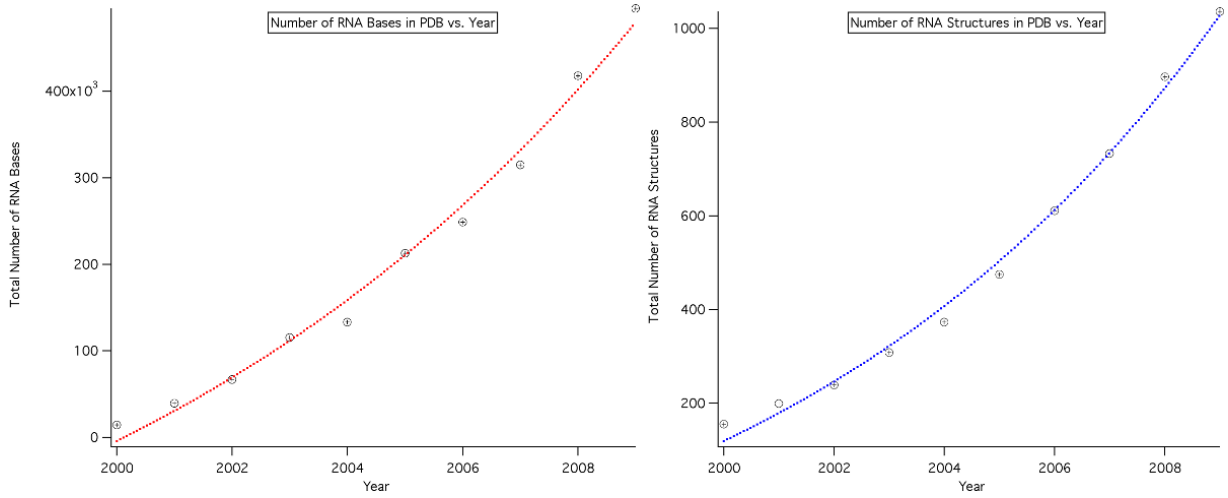

Figure 2.1: **Right:** Total number of RNA bases added to the PDB database between 2000 and 2010 (Exponential fit line in blue). **Left:** Total number of RNA structures solved yearly by X-Ray crystallography between 2000 and 2010 (Exponential fit line in red).

Between 1978 and 2000 a total of 116 RNA structures with resolution greater than 3.5Å, 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. At the present time if we look at the 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.

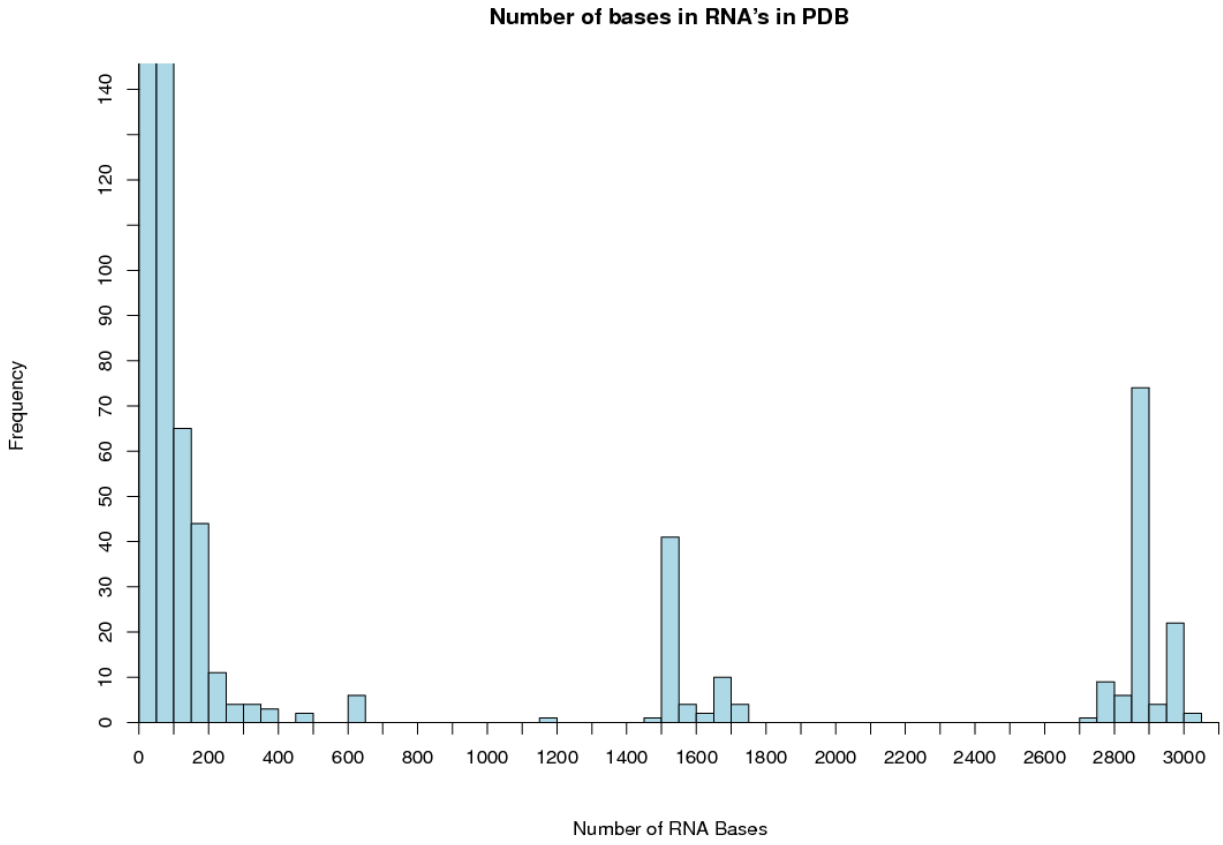

Figure 2.2: Frequency of nucleotide bases in RNA molecules found in the PDB classified by the size of RNA molecules. We define the size as the total number of nucleotide bases present per molecule.

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]. The third category considers the networks of horizontal hydrogen bonding patterns coming from a definition of interacting edge boundaries in the nucleotide bases [18, 19, 20]. In this chapter we study the rigid body based perspective using clustering analysis.

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 deposited at 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.

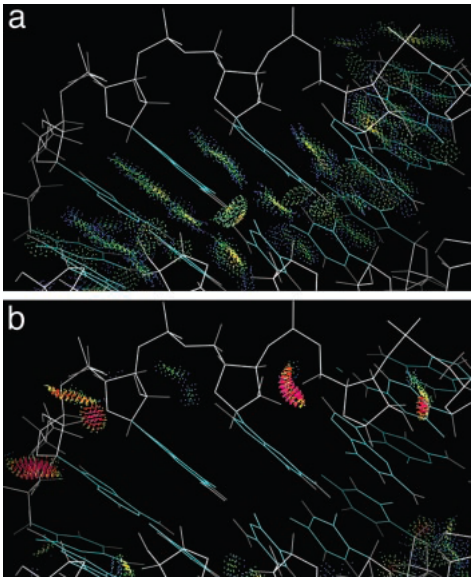

Figure 2.3: Figure taken from Richardson et al. [11] where the blue and green dots in a) mean very accurate van der Waals distances, and in b) the red and orange dots mean steric clashes, that is, distances outside the acceptable van der Waals range.

2.1.1 Combining Fourier Averaging Results and Clustering Analysis

Using the coordinates 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 Figures 2.4 C.1 and 2.5 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 [21, 22].

Group I contains a single 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 are A-RNA related; (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).

When looking at Table 2.2, it's clear that there are 1858 steps (67%) which are not classified into any of the groups. The reason for this is the mixing of Fourier averaging for backbones, and the base step perspective. It might also be that we are not using the other A-RNA like backbone based structures from schneider's paper.

Right now I am doing a validation with clValid, to see if anything pops up regarding the "optimal" number of clusters for the data. Perhaps it would be wise to filter the data by proximity to A-RNA like conformation, say, take all structures which are some RMSD, or manhattan, or euclidean distance appart from the cannonical A-RNA step parameters which are in table such and such.

Leave this argument for second part.

Table 2.2 shows the residue numbers of bases from 23S rRNA which belong to the main categories of Figure 2.5. To match residues of 23S rRNA belonging to the non-Atype clusters, a root mean squared deviation (RMSD) of 15 or less was required between step parameter vectors of 23S rRNA and the mean parameter vectors for the four non-Atype groups identified.

2.1.2 Partitional Clustering for Rigid Body Parameters

The argument I thought could have been made was that with clustering analysis alone on the whole data set, the A-RNA data would split naturally, without recurring to other ideas like Fourier Filtering of Bergman et al.

We also analyze the 2753 base-step parameter vectors in the ribosome. For the partitional clustering case, again, there is no known number of clusters in which the data must group, therefore we've calculated the within clusters sum of squares and also the average silhouette widths, for a particular selection of the number of partitions of the data for $k = [2 - 80]$. From figure 2.6 we can't conclude much. We see that the value of the within clusters sum of squares becomes constant around $k = 47$ and thereÂt's also a change of curvature around $k = 13$. For the case where the average silhoutte width has been computed, that is, figure 2.7, we see that the maximum is for $k = 2$, and there are some interesting maxima at $k = 9, 12$. Now that we have a clue as to which number of partitions the data optimally has we have plotted the k-means results for $k = 13$ and $k = 47$ in Figures number and number, and the PAM results for $k = 2, 9, 12$ in Figure number.

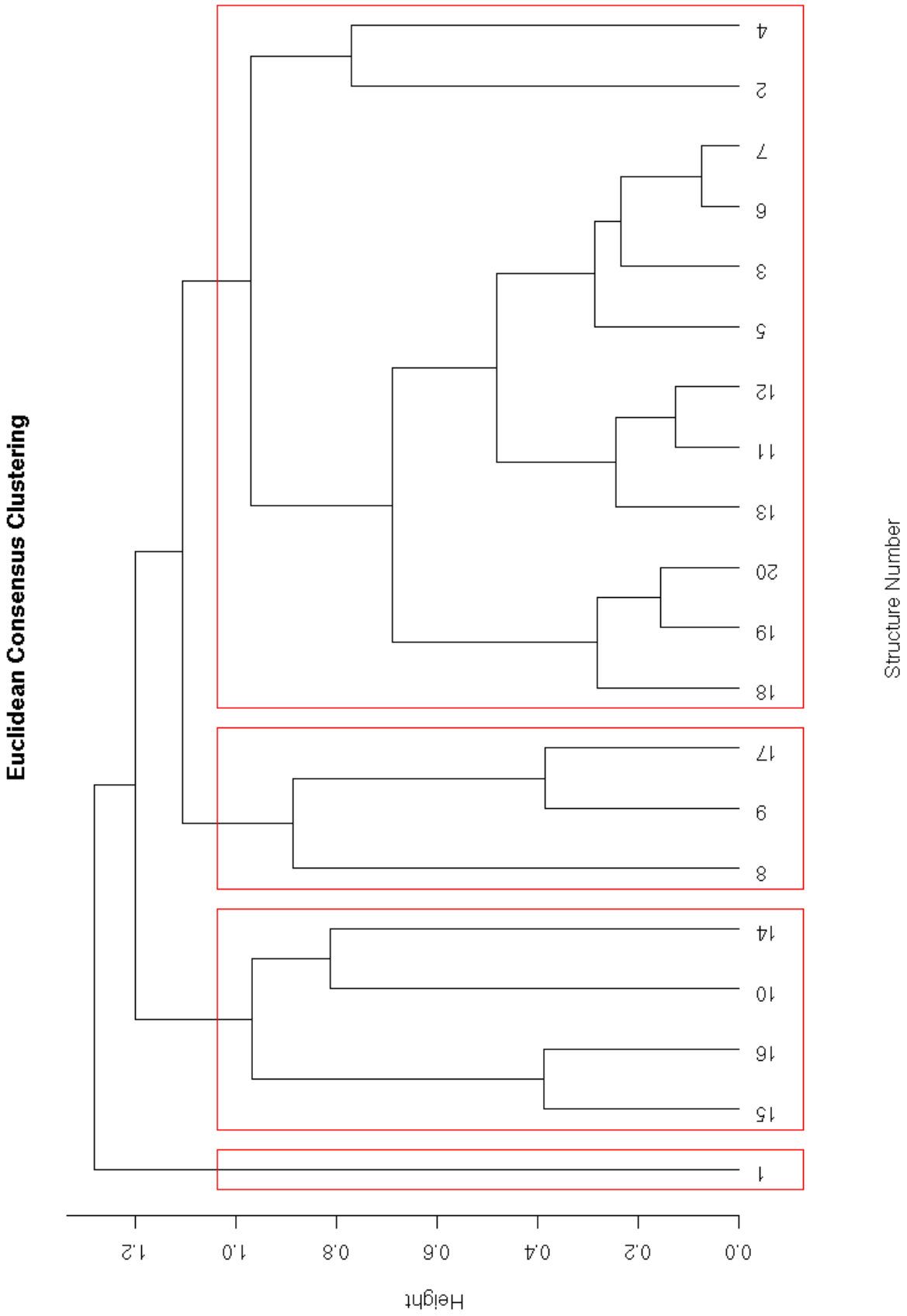

Figure 2.4: Dendrogram showing the results of consensus clustering of 20 non-Atype rRNA dinucleotides according to their hexadimensional base-step parameter vectors.

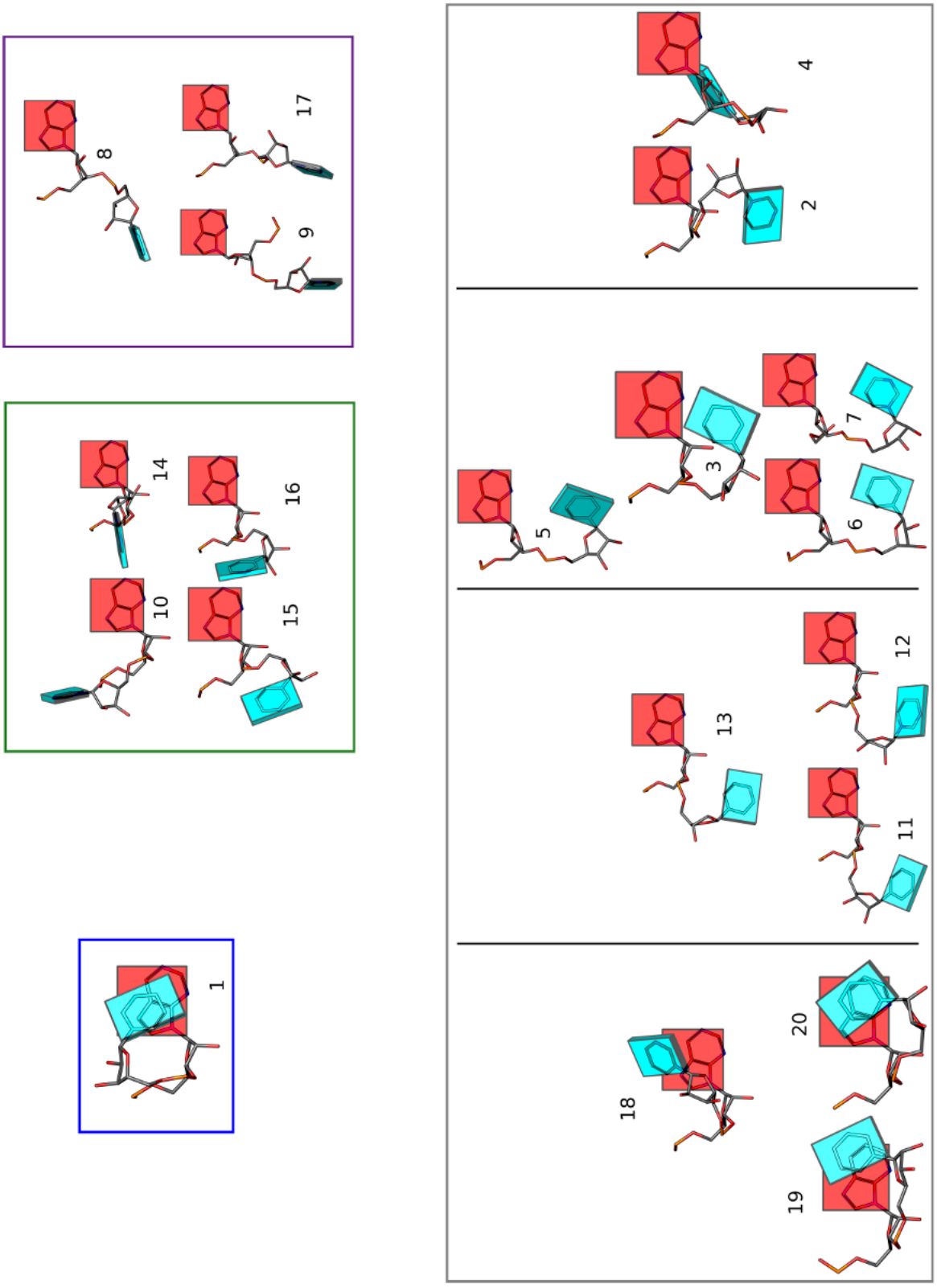

Figure 2.5: rRNA dinucleotide structures organized by clusters obtained from consensus clustering of their hexadimensional base-step parameter vectors.

Total Number of Nucleotides	**RMSD Limit**	**Group**	**Base-steps**	**Base-step Residue Number**	**Overlaps**
2754	< 15	I	3	892, 2006, 2390	
		II	5	459, 1279, 1653, 1919, 2302	
		III	1	2109	
		IV	35	79, 112, 128, 190, 213, 269, 358, 434, 488, 564, 706, 720, 775, 867, 966, 1292, 1503, 1543, 1614, 1766, 1874, 1908, 1971, 2017, 2257, 2427, 2516, 2540, 2755, 2782, 2810, 2826, 2874, 2882, 2913	
		IVa	1	882	
		IVb	807		
		IVc	9	306, 789, 854, 880, 1107, 1192, 1493, 1818, 2005	
		IVd	35	175, 213, 246, 264, 304, 358, 464, 518, 531, 534, 588, 795, 938, 1214, 1231, 1316, 1340, 1370, 1605, 1745, 1766, 1971, 1976, 2010, 2017, 2291, 2320, 2428, 2469, 2481, 2516, 2532, 2755, 2826, 2882	Only IVd with IV (213, 358, 1766, 1971, 2017, 2516, 2755, 2826, 2882)

Table 2.2: Residue numbers for base-steps with RMSD values less than 15 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* large ribosomal subunit.

We have also filtered the data according to the 16 possible RNA base steps, that is, AA, AG, GA, GG, UU, UC, CU, CC, UA, UG, CA, CG, AU, AC, GU, and GC. Tables showing how many representatives steps there are belonging to non-helical, helical, and watson-crick sets, will be later included and discussed here.

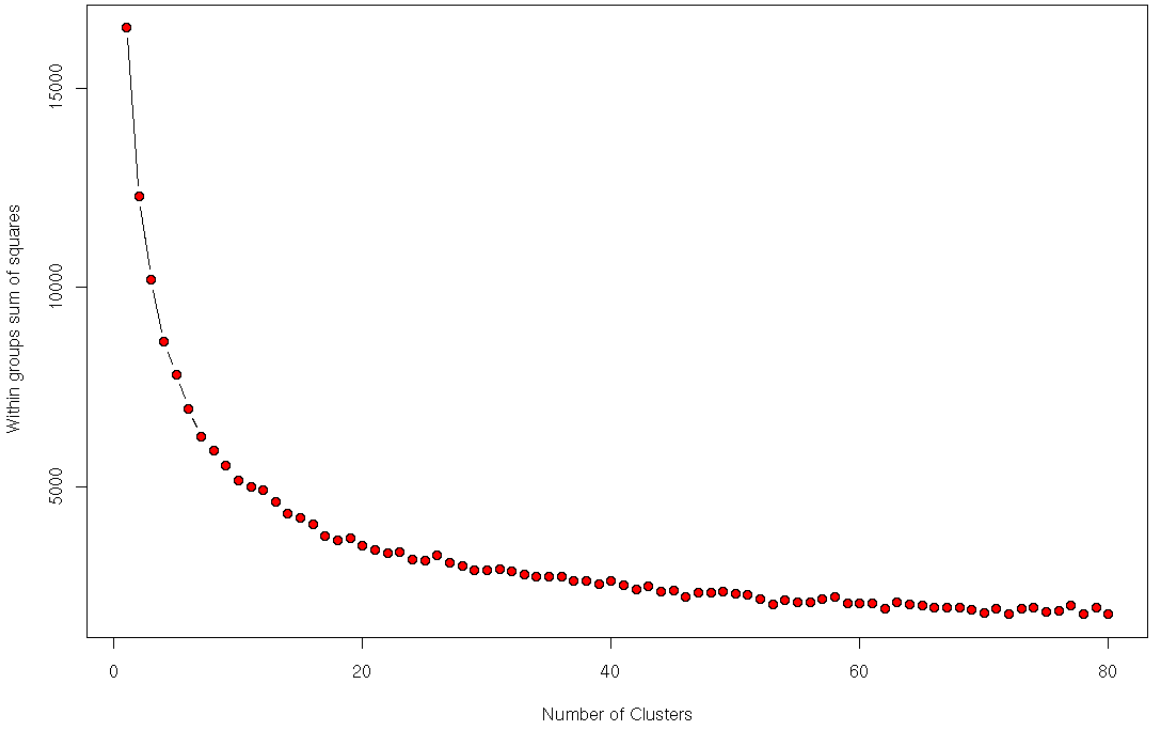

Figure 2.6: Sum of all within clusters sum of squares against number of clusters.

2.1.3 Hierarchical Clustering for Rigid Body Parameters

Also as has been carried out for torsion angles, hierarchical clustering has also been performed on rigid body parameters, the results are yet to be included here. A cluster dissimilarity tree can be seen in Figure 2.8 for the 12 trees resulting from the four clustering methods and three distance definitions used to cluster the base step data.

2.2 RNA Conformations

There are two main RNA conformations, A-RNA ,and A'RNA, and maybe even a third unconfirmed one A"RNA [2]. Their values for their standard torsion angles and step parameters can be seen in Tables 2.3 and 2.4

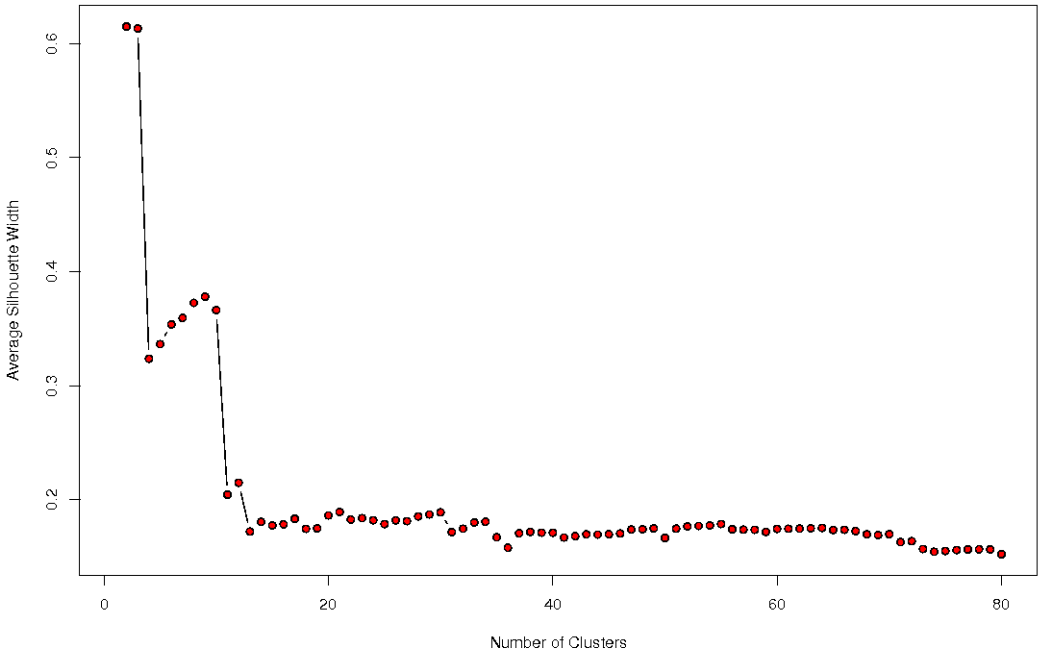

Figure 2.7: Average silhouette width against number of clusters.

Structure Name	α	β	γ	δ	ϵ	ζ	χ	**Reference**
A-RNA	-68.9	179.5	54.5	82.2	-153.9	-70.8	-161.1	Arnott
A'-RNA	-70.0	176.6	60.8	76.7	-153.4	-69.4	-163.4	Arnott
AII-RNA	-65.0	175.1	52.9	81.1	-166.0	-68.0	-157.0	Schneider

Table 2.3: Base step torsion angles for the different known RNA conformations.

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 2.4: Base step parameters for the different known RNA conformations. Notice that the base step parameters are for single bases rather than base-pairs.

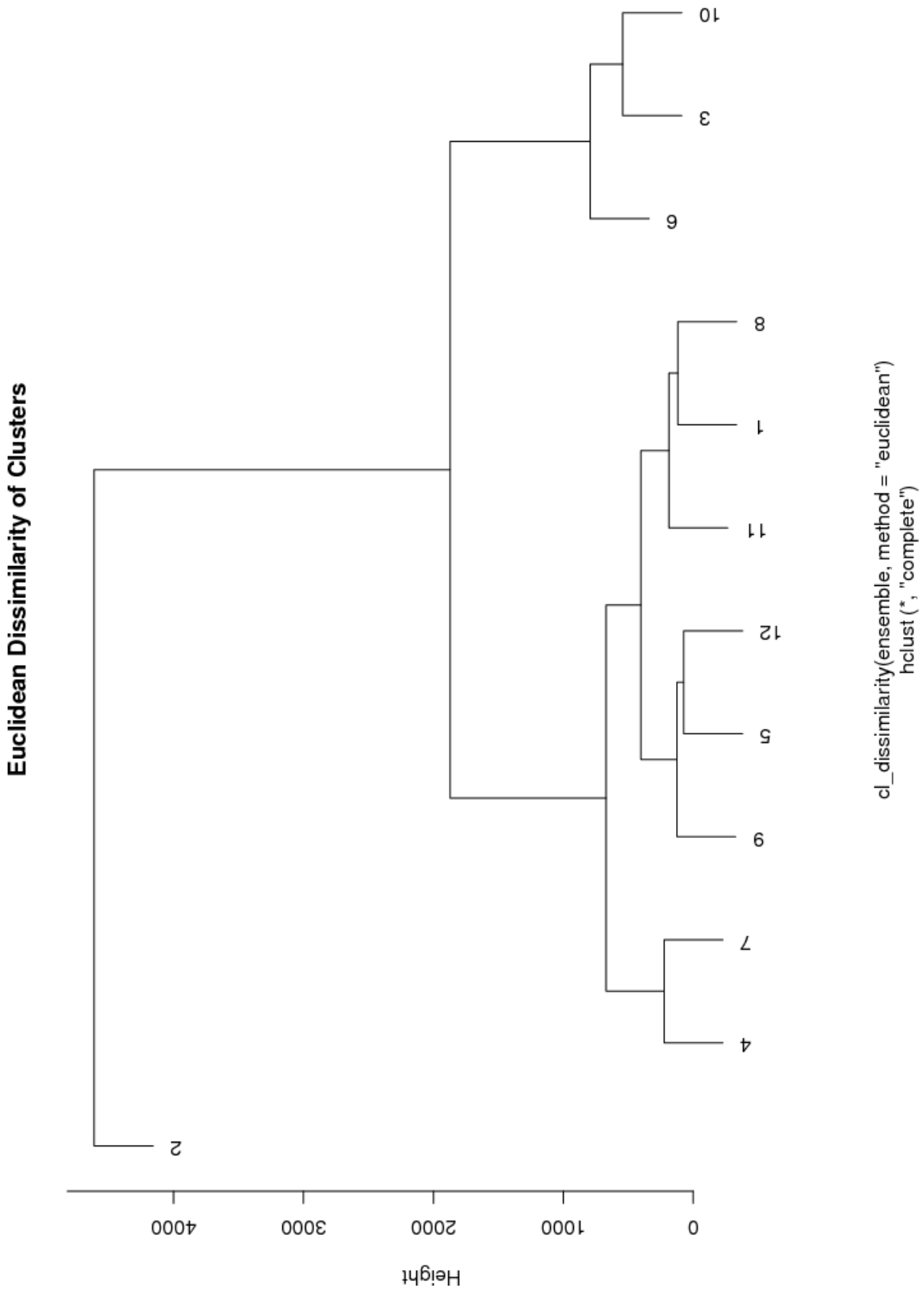

Figure 2.8: Cluster dissimilarities for the twelve hierarchical trees obtained from clustering of the six-dimensional base-step parameters obtained from the large subunit of the ribosome (PDB-ID:1jj2)

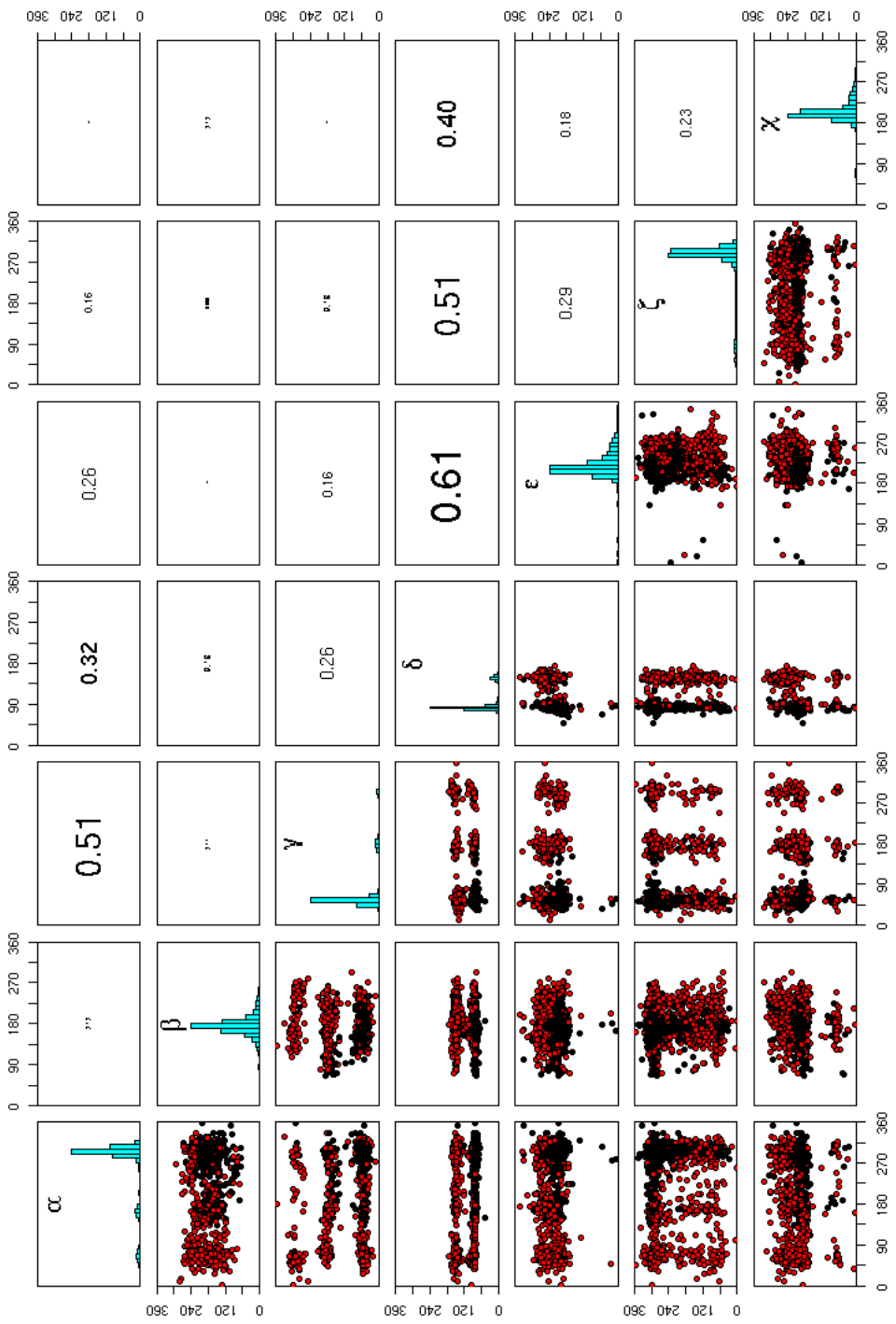

Figure 2.9: K-means of torsion angle vectors of 2753 dinucleotide steps present in 23S rRNA using the *Hartigan-Wong* algorithm. The number of partitions is **2**. The upper diagonal matrix displays the values of the linear correlation coefficient r, and a histogram showing the torsion angle distribution is rendered in the diagonal.

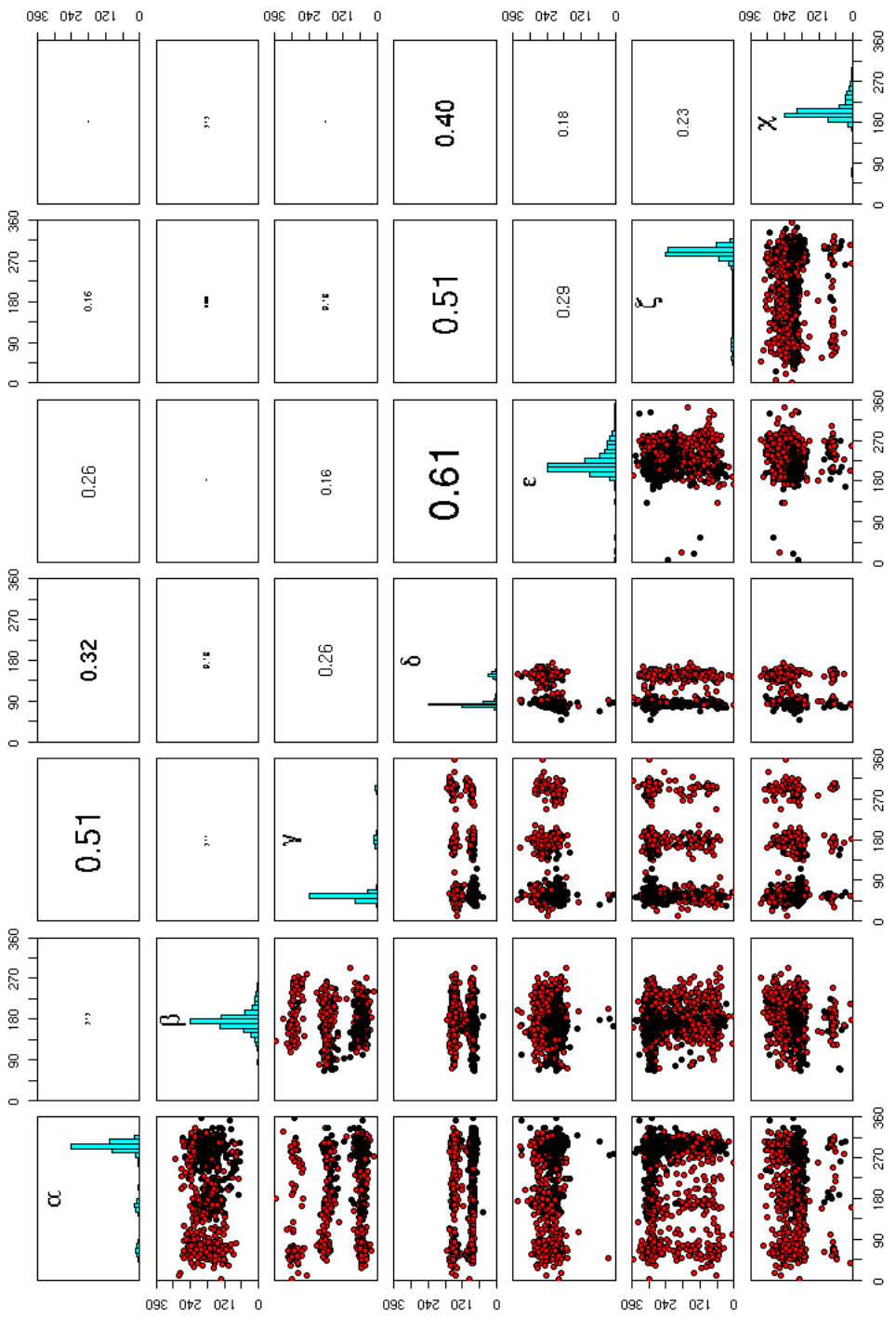

Figure 2.10: K-means of torsion angle vectors of 2753 dinucleotide steps present in 23S rRNA using the *Lloyd* algorithm. The number of partitions is **2**. The upper diagonal matrix displays the values of the linear correlation coefficient r, and a histogram showing the torsion angle distribution is rendered in the diagonal.

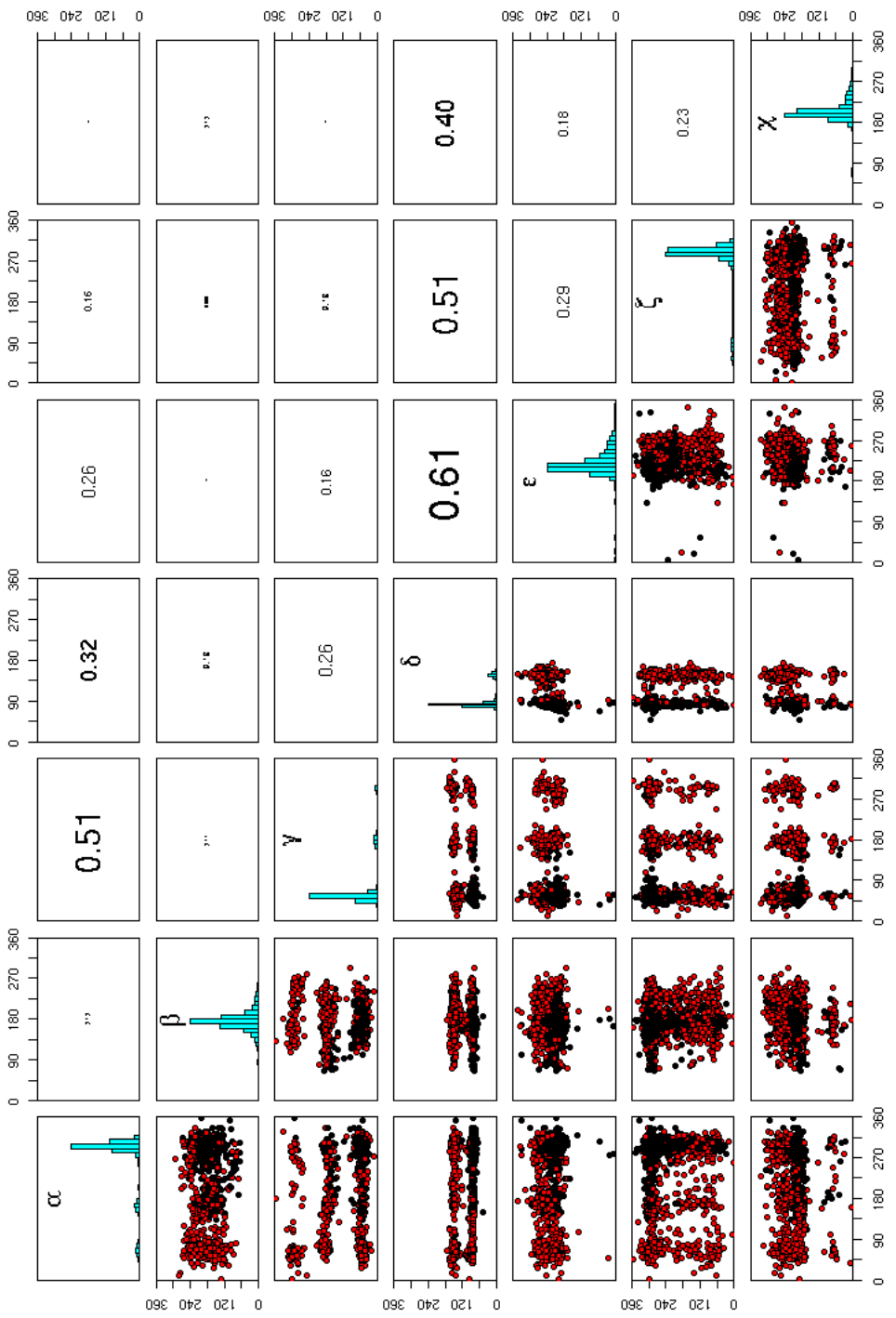

Figure 2.11: K-means of torsion angle vectors of 2753 dinucleotide steps present in 23S rRNA using the *Forgy* algorithm. The number of partitions is **2**. The upper diagonal matrix displays the values of the linear correlation coefficient r, and a histogram showing the torsion angle distribution is rendered in the diagonal.

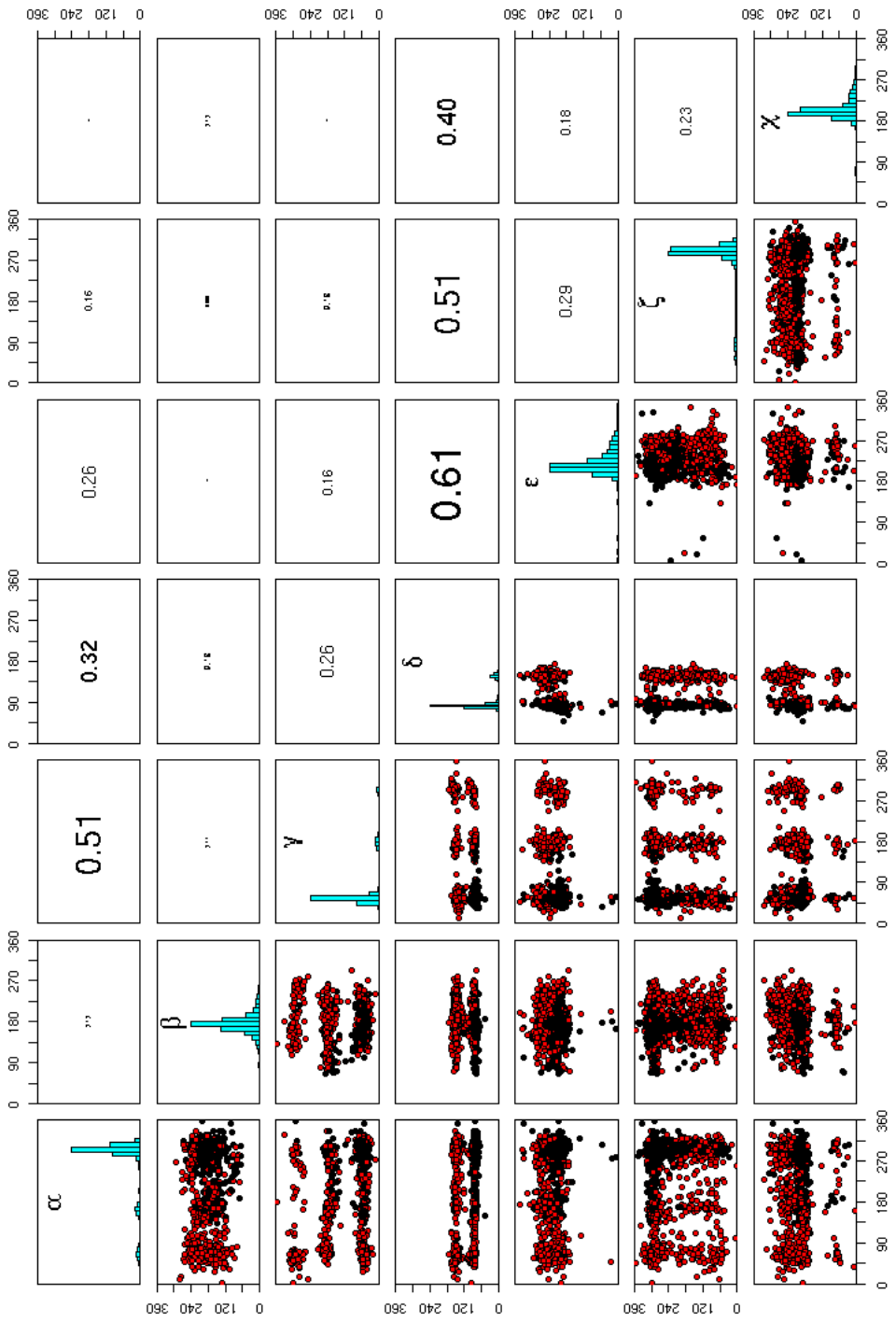

Figure 2.12: K-means of torsion angle vectors of 2753 dinucleotide steps present in 23S rRNA using the *McQueen* algorithm. The number of partitions is **2**. The upper diagonal matrix displays the values of the linear correlation coefficient r, and a histogram showing the torsion angle distribution is rendered in the diagonal.

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.