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, the mechanisms of life regulation and control, and it plays a primordial role in some viruses. Our knowledge of the importance of RNA in cellular regulation is relatively new, (a bit more than a decade, Craig and Mello, Nature 1998) and along with the detailed structural elucidation of the transcription machine, the ribosome, has propelled interest in RNA understanding to a level which starts to closely resemble that given to proteins and DNA.

In these scheme 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 it's not foreign to use statistical mechanical methods combined with detailed knowledge of atomic level structure. Such methodology is mainly used in this work in a long term effort on understanding the intrinsic structural features of RNA, and how they might relate to it's 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 1

Introduction

1.1 RNA folding

The first high resolution X-ray structure of RNA larger than a dinucleotide was that of yeast $tRNA^{Phe}$ at 3Å in 1974 [1, 2]. Thirtysix years later there are two orders of magnitude more RNA structural information [3], and new information from non-coding RNA's is expected [4]. This fact and the discovery of ribozymes [5, 6] has renewed interest in solving the RNA folding problem, that is, from primary sequence, finding in an automated[i] way the native three-dimensional structure of RNA and its folding pathway. 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 [5, 6] and the complicated folding of large RNA molecules [10]. 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 [11]. 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 [12, 13, 14] of either protein or RNA.

1.2 Is RNA folding a hard or easy problem?

There are two trains of thought regarding RNA folding. One states that RNA folding is less complex than protein folding [15] 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+} [16] (see Figure 1.1), whereas secondary and tertiary elements are not as easily separable in proteins. The other point of view says that RNA folding

[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 [7, 8] uses the Tinoco-Uhlenbeck dinucleotide postulate [9] to find total free energies.

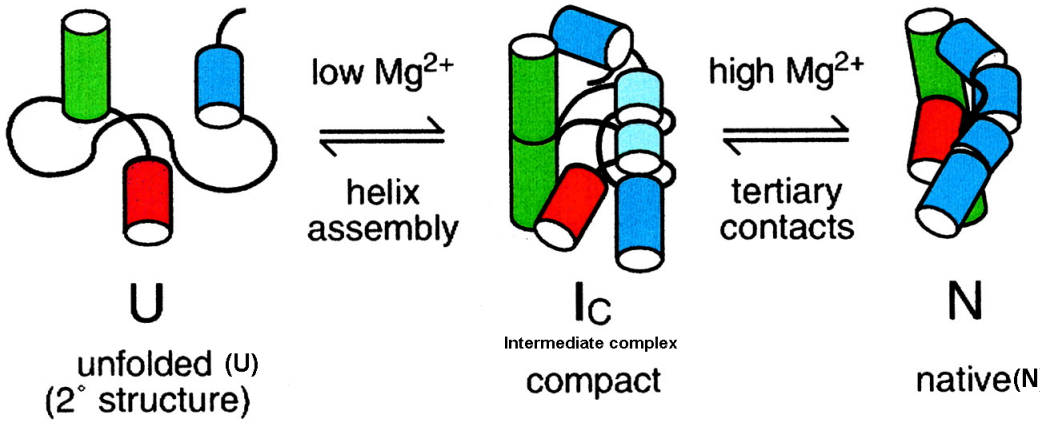

Figure 1.1: Separation of secondary and tertiary interaction in RNA [16]. Double helical secondary structure represented by individual cylinders and tertiary interactions by association of cylinders.

can be at least as complex as protein folding [17, 18] since there is no such thing as hydrophobic burial of regions of RNA as in the case of proteins. Instead, the electrostatic problem of having a complex charged backbone must be dealt with in the case of RNA. For instance, the interactions of the RNA polyanionic backbone with water and cations [19] are not easily simulated with explicit solvent models as can be done for proteins. The aforementioned interactions of RNA need to be modeled implicitly, and must aim to describe long dynamic processes of the order of seconds to minutes, in contrast to the typical time scales of tens of microseconds associated with protein folding. Although secondary and tertiary structure can be separated experimentally, there have been few theoretical efforts to account for the folding of RNA from a random sequence of nucleotides into secondary structures and tertiary structures. What little is know has been investigated at low resolution. Professor Stephen Harvey and associates have simulated yeast $\mathrm{tRNA}^{\mathrm{Phe}}$, [20] and the assembly of the 30S subunit of the ribosome [21] at various levels of detail, initially using only one pseudoatom per helical region, and later one pseudoatom per nucleotide. Recently Major's group at Montreal has proposed a pipeline of two computer algorithms [22], one 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 [23, 24, 25]. This type of work is often referred to as protein structural topology and there is no counterpart for RNA.

1.3 Experimental folding techniques

Traditionally RNA folding and unfolding have been followed calorimetrically and spectroscopically as a function of temperature and cation concentration [26]. 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 [27]. 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 [28, 29, 30, 31]. In the case of single molecule force-induced unfolding, state transitions often occur under non-equilibrium conditions, thereby making it difficult to extract equilibrium information from the data. Recently Bustamante, Tinoco, and associates have shown that using the Crooks fluctuation theorem [32], one can deal with such cases and extract RNA folding free energies from single molecule experiments [33].

1.4 RNA simulations

Network and molecular mechanics-molecular dynamics (MM-MD) methods provide useful information relevant to the RNA folding-unfolding problem, especially for describing fluctuations away from the native conformation. Gaussian network models [34, 35, 36] 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 [37]. Tung and associates used this structure for an all-atom MD simulation of the movement of tRNA into a fluctuating ribosome [38]. 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.

1.4.1 Local nucleotide interactions

The molecular interactions which rule RNA structures at the nucleic acid base level, *i.e.*, local level, are hydrogen bonding and stacking interactions. The former are related to base pairing and the latter, in most cases, to nucleotide steps. These interactions can be explored theoretically at various levels. At

the highest level are ab-initio quantum mechanical calculations which are still too expensive for systems as large as hundreds of atoms. Such calculations, nevertheless, can tell a great deal about local electronic behavior. For example, Hobza and collaborators have found that the stacking interaction of free nucleotide bases is determined by dispersion attraction, short-range exchange repulsion, and electrostatic interaction. No specific $\pi - \pi$ interactions are found from electron correlated ab-initio calculations [39, 40]. 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 [37, 41]. 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 [42] seem to contradict their older publications [41, 43], in which the stacking energies were reported to be relatively insensitive to dimer conformation. The new results use the so-called "coupled cluster singles doubles with triple electron excitations" CCSD(T) method, to account for electron correlation. Using this electron correlation energy correction, the stacking energy differences between dimer conformations turn out to be considerably higher than previously reported.

Single and double strand stacking free energies can be obtained calorimetrically. The most popular method used for obtaining such quantities is differential scanning calorimetry (DSC) [44]. 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 [26]. 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 [45, 46] provides a simple and quantitative picture of the interactions of the double helical nucleic acid polyanion with its counterions, although it does not take into account the discrete nature of charge [26] or the folding of RNA. Poisson-Boltzmann theory offers a more detailed picture of the behavior of charged macroions in solution [47].

The local conformational space of RNA has been studied using a large set of available RNA structures from the Nucleic Acid Database (NDB) [48]. The torsion angles of the nucleotide steps have been clustered in the parameter space using different techniques [49, 50]. The root-mean-square deviations (RMSD) of the distances between closely spaced atoms in the phosphates, sugars, and bases, have

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

also been clustered [51]. The latter studies are aimed at finding the common nucleotide base steps and base-pair building blocks which are given the name of RNA doublets. Recently, the RNA Ontology Consortium (ROC) has proposed a consensus set of RNA dinucleotide conformers integrating the work of various groups [52].

1.4.2 RNA secondary structure algorithms and the lack of tertiary ones

From secondary structure prediction algorithms like Zuker's *mfold* program [53], Hofacker's Vienna RNA package [8], or Mathews Dynaling [54], one obtains a large ensemble of secondary structure graphs. These graphs can be analyzed with graph theory to produce a partition function describing a full arrangement of contacts for the total number of possible secondary structures making possible a "relation of microscopic conformations to macroscopic properties" [55]. 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, followed by a minimization procedure, and then a molecular mechanics (MM) all atom refinement [56, 57, 58]. Another starts from predicted secondary structures and assumes their helical regions adopt the A-form conformation, then mechanically thrusts residues as rigid bodies in the remaining non-helical regions, and finally carry out an MM optimization [59]. Finally, 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 is renamed by Major's group as Nucleic Cyclic Motifs (NCM) [22].

1.4.3 RNA overall fold

Whereas in the case of proteins one can describe the overall fold from the arrangement of secondary structure motifs, *i.e.*, using the helix-ribbon-coil images developed by Jane Richardson [60] (see Figure 1.2), there is still no comparable description of the overall fold of RNA. A ribbon representation of the sugar phosphate backbone helps to understand the folding of small RNA's, but in the case of the ribosome this type of representation is not sufficient, see Figure 1.3.

One can envision that a thorough investigation of the parameter space of translational and rotational degrees of freedom of the helical regions of RNA could give clues as to how we might see an overall

Figure 1.2: Ribbon-coil schematic illustraring the fold and intermolecular units of a dimer of prealbumin, or transthyretin, taken from Richardson *et al.* [61]

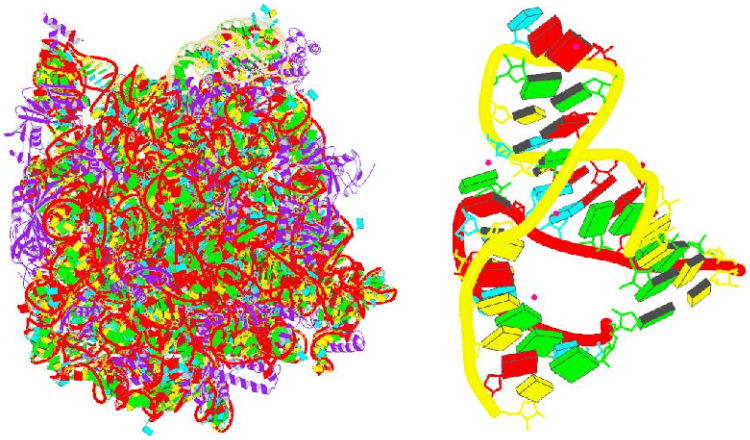

Figure 1.3: *Haloharcula marismortui's* large ribosomal subunit (left) and hammerhead ribozyme (right).The figures were taken directly from the NDB web pages, and show a ribbon representation of the phosphate backbone, and a block representation for the nucleotide bases. From the figures it's clear that,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

fold in RNA structures.

In the case of proteins the SCOP (Structural Classification of Proteins) database [62], classifies proteins, among other classifications, according to recurrent arrangements of secondary structure, that is, folds. The SCOR (Structural Classification of RNA) database [63, 64], 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.

1.4.4 RNA motifs

First, a word of caution must be given to the reader. The term "*RNA motif*" alone is used in the literature to describe three different levels of RNA organization, that is, RNA **sequence** motifs, RNA **secondary structure** motifs, or RNA **3D structure** motifs. We start by making such distinction as it is not always

clearly mentioned in RNA literature, generating a great deal of confusion and bibliographical search frustration for the beginner. The kind of RNA motifs we are dealing with in this thesis are those of the third kind, that is, RNA **3D structure** motifs which we'll address from now on simply as RNA motifs. Yet another source of confusion in understading RNA motifs is the lack of a unique definition. Three popular and somewhat recent definitions are:

- RNA motifs are "*Conserved structural subunits that make up the secondary structures of RNAs.*"[65]
- RNA motifs are "*Ordered stacked arrays of non-Watson-Crick base pairs that form distinct folds on the phosphodiester backbones of RNA strands.*"[66]
- "*An RNA Motif is a discrete sequence or combination of base juxtapositions found in naturally occurring RNA's in unexpectedly high abundance.*"[67]

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 [68, 69, 70], describe [71, 72, 73, 74, 75], and find new [76, 77, 70, 78, 69] motifs.

References

[1] Robertus, J. D., Ladner, J. E., Finch, J. T., Rhodes, D., Brown, R. S., Clark, B. F. C., and Klug, A. (1974) Structure of Yeast Phenylalanine tRNA at 3 Å Resolution. *Nature,* **250**, 546.

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

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

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

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

[10] Batey, R. T., Rambo, R. P., and Doudna, J. A. (1999) Tertiary Motifs in RNA Structure and Folding. *Angewandte Chemie International Edition,* **38**(16), 2326–2343.

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

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

[13] 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.

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

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

[16] 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.

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

[18] 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.

[19] Klein, D. J., Moore, P. B., and Steitz, T. A. (2004) The Contribution of Metal Ions to the Structural Stability of the Large Ribosomal Subunit. *RNA,* **10**(9), 1366–1379.

[20] 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.

[21] 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.

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

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

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

[25] 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.

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

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

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

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

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

[31] 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**(19), 6789–6794.

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

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

[34] 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.

[35] 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.

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

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

[38] 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.

[39] 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.

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

[41] 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.

[42] 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**, 2854 Ű2865.

[43] 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.

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

[45] 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.

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

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

[48] 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.

[49] 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.

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

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

[52] 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.

[53] Zuker, M. (2003) Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction. *Nucleic Acids Research,* **31**(13), 3406–3415.

[54] Mathews, D. H. and Turner, D. H. (Mar, 2002) Dynalign: An Algorithm for Finding the Secondary Structure Common to Two RNA Sequences. *Journal of Molecular Biology,* **317**(2), 191–203.

[55] 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.

[56] Das, R. and Baker, D. (Sep, 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**(37), 14664–14669.

[57] Ding, F., Sharma, S., Chalasani, P., Demidov, V. V., Broude, N. E., and Dokholyan, N. V. (Jun, 2008) Ab Initio RNA Folding by Discrete Molecular Dynamics: From Structure Prediction to Folding Mechanisms. *RNA,* **14**(6), 1164–1173.

[58] Jonikas, M. A., Radmer, R. J., and Altman, R. B. (Dec, 2009) Knowledge-Based Instantiation of Full Atomic Detail Into Coarse-Grain RNA 3D Structural Models. *Bioinformatics,* **25**(24), 3259–3266.

[59] Martinez, H. M., Jr, J. V. M., 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.

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

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

[62] 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.

[63] 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.

[64] Klosterman, P. S., Hendrix, D. K., Tamura, M., Holbrook, S. R., and Brenner., S. E. (2004) Three-Dimensional Motifs from the SCOR, Structural Classification of RNA Database: Extruded Strands, Base Triples, Tetraloops and U-turns. *Nucleic Acids Research,* **32**(8), 2342–2352.

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

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

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

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

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

[70] 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**(16), 4755–4761.

[71] Laing, C., Jung, S., Iqbal, A., and Schlick, T. (Oct, 2009) Tertiary Motifs Revealed in Analyses of Higher-Order RNA Junctions. *Journal of Molecular Biology,* **393**(1), 67–82.

[72] Laing, C. and Schlick, T. (Jul, 2009) Analysis of Four-way Junctions in RNA Structures. *Journal of Molecular Biology,* **390**(3), 547–559.

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

[74] Spacková, N. and Sponer, J. (2006) Molecular Dynamics Simulations of Sarcin-Ricin rRNA Motif. *Nucleic Acids Research,* **34**(2), 697–708.

[75] Réblová, K., Spacková, N., Stefl, R., Csaszar, K., Koca, J., Leontis, N. B., and Sponer, J. (Jun, 2003) Non-Watson-Crick Basepairing and Hydration in RNA Motifs: Molecular Dynamics of 5S rRNA Loop E. *Biophysical Journal,* **84**(6), 3564–3582.

[76] Sarver, M., Zirbel, C. L., Stombaugh, J., Mokdad, A., and Leontis, N. B. (Jan, 2008) FR3D: Finding Local and Composite Recurrent Structural Motifs in RNA 3D Structures. *Journal of Mathematical Biology,* **56**(1-2), 215–252.

[77] Mokdad, A. and Frankel, A. D. (April, 2008) ISFOLD: Structure Prediction of Base-pairs in Non-helical RNA Motifs From Isostericity Signatures in Their Sequence Alignments. *Journal of Biomolecular Structure and Dynamics,* **25**(5), 467–472.

[78] Stonge, K., Thibault, P., Hamel, S., and Major, F. (2007) Modeling RNA Tertiary Structure Motifs by Graph-Grammars. *Nucleic Acids Research, 2007, 1ï£¡11,* pp. 1–11.

Chapter 2

RNA Base Steps

The problem of classification of the space of conformations configurations? 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 [i] ribosome [6].

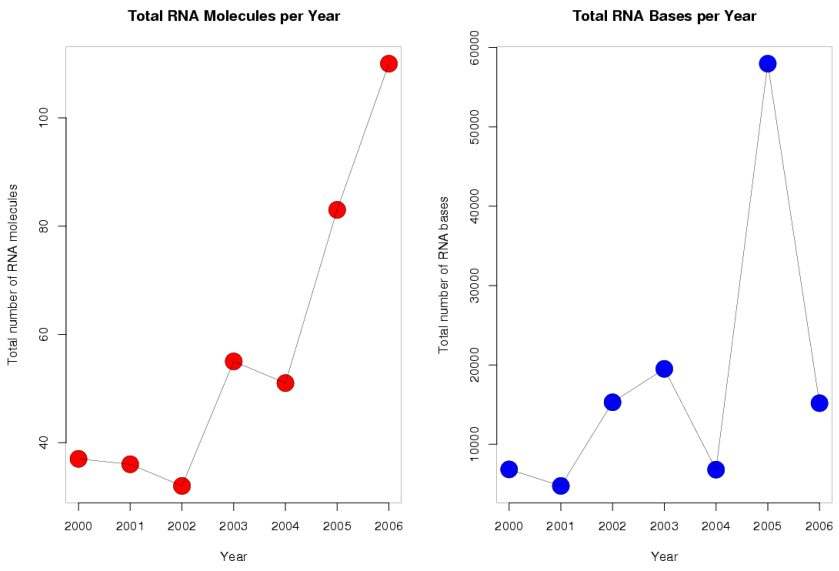

Figure 2.1: **Left:** Total number of RNA structures solved yearly by X-Ray crystallography between 2000 and 2006. **Right:** Total number of RNA bases added to the PDB database between 2000 and 2006.

Between 1972 and 2000 a total of 132 RNA structures with resolution greater than 3 Å, and comprising around 5500 nucleotide bases were found in the Protein Data Bank (PDB), and between 2000 and today a total of 460 RNA structures comprising around 140000 nucleotide bases have been found. That is, the increase in information due to the solution of large RNA structures is two orders of magnitude as

[i] I emphasize the phylogeny of rRNA's here since there is an ongoing discussion among biologists on whether archaea are closer to prokaryotes, or to eukaryotes.

pointed out by Noller [7] in 2005. Looking at the growth of RNA structural information from 2000 until today it's important to point out that, although the total number of RNA structures deposited in the PDB shows exponential growth (see left panel in Figure 2.1), the total number of RNA bases does not show a well defined trend (see right panel in Figure 2.1). This is due to the size preponderance of ribosomal structures. That is, in 2005 nineteen ribosomal structures were deposited in the PDB, whereas in 2006 only four were deposited. So, even though interest in RNA seems to be growing since ribosomal structures have become available in 2000, and two Nobel prizes were awarded for work in RNA in 2006, along with the exciting possibilities of deciphering even larger RNA virus structures, still the growth of the RNA structural field is far from that of proteins if weighed by the growth of RNA structural information in the past seven years. This fact might just be due to the smaller sizes and structural diversity of RNA molecules, which, as can be seen in Figure 2.2, is restricted to "compact" nucleotide ranges. A representative example of these characteristic ranges can be seen in Table 1.1 for structures larger than 300 bases.

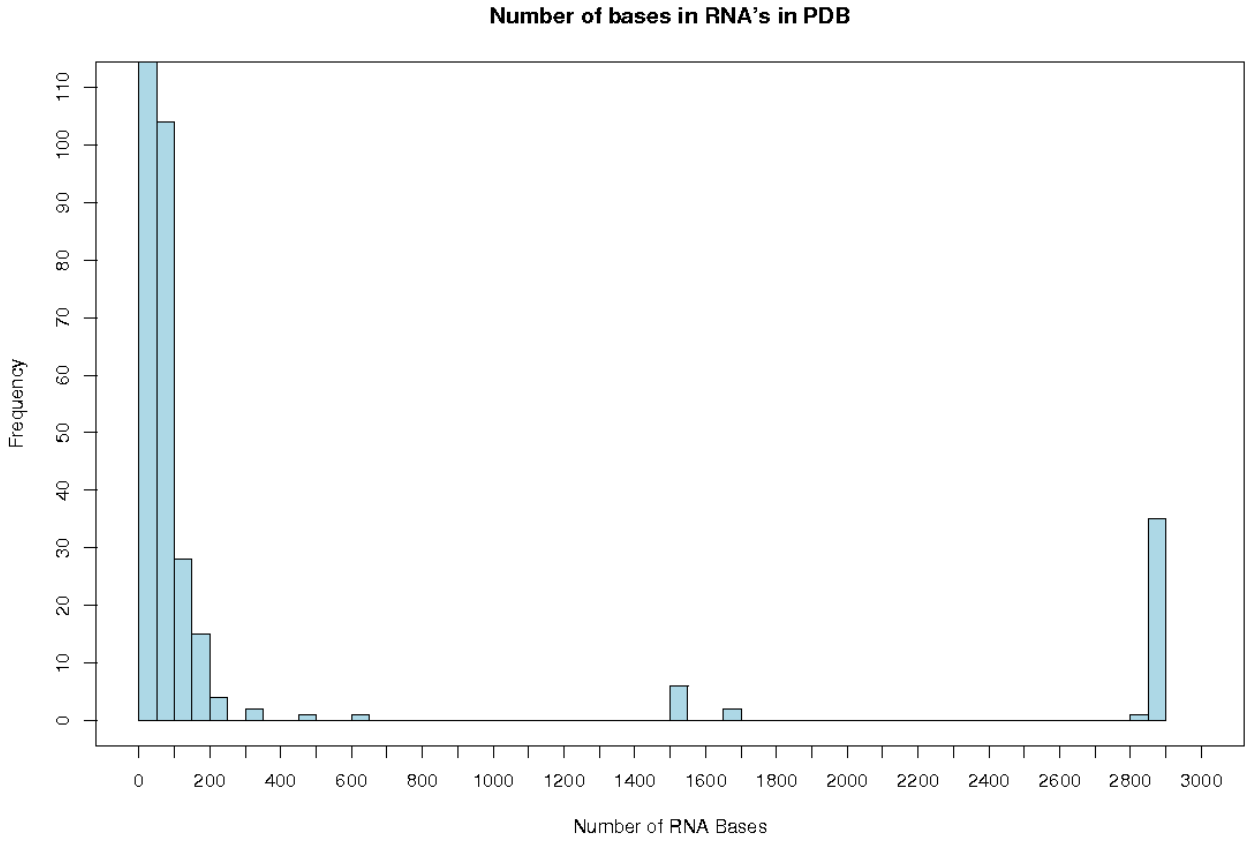

Figure 2.2: Frequency of nucleotide bases in RNA molecules found in the PDB classified by the size of RNA molecules. We define the size as the total number of nucleotide bases present per molecule.

Analysis of RNA conformational information contained in this 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 [8], or a comparison of distances between a reduced set of atoms taken from the nucleotide backbone, sugar, and base [9]. 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 [8, 10, 11, 12, 13]; or alternatively the pseudo-bonds between consecutive P and C4′ atoms and the resulting pseudo-torsion angles η and θ [1, 14, 15, 16]. The third category considers the networks of horizontal hydrogen bonding patterns coming from a definition of interacting edge boundaries in the nucleotide bases [17, 18, 19]. In this report we review one category of the bond based perspective. Namely we review the case where the covalent bonds between backbone atoms give rise to torsion angle space. We also make a first study of 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
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 7 years.

2.1 Dinucleotide Torsion Angles

The covalent bond based perspective as mentioned in the previous section gives rise to six backbone torsion angles and one glycosidic torsion angle. This heptaparametric space has been the subject of several recent studies of RNA dinucleotide steps. Richardson and collaborators [10] have applied van der Waals radius filtering techniques on a database of 8636 nucleic acid residues from RNA X-Ray structures with resolution of 3.0 or better grouping all structures in 42 conformers which they refer to as rotamers. Berman et al. [12] reduced the data space of the large subunit of the *Haloarcula Marismortui* ribosome using Fourier transform filtering. Hershkovitz et al. [11] defined lower and upper bounds

of torsion angle values by "binning" in one dimension. Pyle et al. [16] reduced the heptaparametric space to a biparametric one, defining a virtual bond between consecutive O and C4′ atoms in the RNA backbone.

Hershkovitz and collaborators [13] took a first step towards integrating clustering analysis formally in the study of RNA backbone torsion angles. In particular, they used the k-means partitional clustering algorithm. [ii] It's important to note that Hershkovitz et al. reduce the data set of all torsion angles in rRNA's large subunit, using their binning approach prior to k-means clustering.

Clustering analysis can be divided into two main methodologies, namely, hierarchical clustering and partitional clustering [20]. We have used particular cases of both methodologies to investigate thoroughly if "biased" [iii] data reduction is needed, as has been suggested by various authors [12, 13], or if the use of clustering analysis alone can be used to find in an efficient and clear manner subsets of RNA conformational space which possess a clear structural meaning.

2.1.1 Partitional Clustering for Torsion Angles

For partitional clustering the k-means algorithm as implemented in the software package **R** [21] was employed. [iv]

We consider the 2753 base-steps of the 23S subunit of the ribosome as vectors of seven dimensions composed of the previously mentioned backbone torsion angles α, β, γ, δ, ϵ, ζ, and the glycosidic torsion angle χ. The **R** software package has implemented four different k-means algorithms; *Hartigan-Wong*; *Lloyd*; *Forgy*; and *MacQueen*. For the data set used, that is, the large subunit of the ribosome (PDB code 1jj2), no noticeable differences were found with the four k-means algorithms when we group the data into two partitions, as can be seen in Figures 2.16 through 2.19.

One of the problems with k-means is that the number of clusters is not an emergent property of the data set but a parameter that has to be given to the algorithm. This problem has been given a good amount of attention in the statistical analysis community, in the area of clustering analysis. Hershkovitz

[ii] For a reason that still eludes the author, instead of using the more general and familiar terminology of clustering analysis, they refer to this method as if it was not a clustering analysis method. In one case they call their method scalar quantization, when one torsion angle at a time is clustered. They call it vector quantization when they want to cluster groups of more than one torsion angle at a time.

[iii] By biased data reduction we mean that a reduction of the whole data set is done by taking into account a particular bias imposed by us. This bias can be structural or sequence based.

[iv] Another partitional algorithm that could readily be used in the future is pam (partitioning around medoids). For now we've determined the average silhouette width, which is an analogous quantity to the average distortion **D**, (which will be defined later in the text) and plotted this value against the number of clusters as can be seen in Figure 2.4

et al. [13] use one method which is common in clustering analysis and find a so-called distortion measure **D**, also called the "within clusters sum of squares" in the more common clustering analysis area. That is, the squares of the elements of each cluster is found and then added. This quantity can be plotted against the number of clusters k that were selected as can be seen in Figure 2.3. The "optimal" number of clusters corresponds to the value of k where **D** becomes constant, which in Figure 2.3 is around 60. Where interestingly Figure2.3 is very similar to Figure 8 in the paper of Hershkovitz et al. [13], where the **D** value becomes constant also around 60. The main difference between our plot and theirs is that they exclude dinucleotide steps which are close to the A-type conformation. They further state that the A-type steps amount to over 60$\%$ of the data, meaning that the remaining 40$\%$ accounts for the majority of the conformational diversity of the space of torsion angles, since there is not a significant change in the value where **D** becomes constant.

There are more methods to determine the optimal amount of partitions that a data set can be split into. For example, one can also do another partitioning around medoids (PAM), and find an analogous quantity to the within clusters sum of squares, such quantity is called the average silhouette width, and the optimal number of clusters is that which maximizes this quantity. In Figure 2.4 we can see that the maximum corresponds to $k = 3$, nonetheless, we also see various local maxima, for example in $k = 16, 22, 28, 31, 36, 45, 48, 51, 57$, it's interesting to point out that in a very recent preprint by Berman et al. [22] they review the work on RNA backbone conformations and summarize that different research groups find 32, 37, and 42 discrete RNA conformations.

Another way of selecting k is just by visual inspection of the data. In our case if we take any of the scatterplots in Figures 2.16 thru 2.19, we can imagine that in some of them the data seems to be clustered around eight groups[v]. We choose eight groups also because this is the result that Duarte and Pyle obtain for their classification based on RNA pseudotorsions [14]. Following this argument we can color code our scatterplots for the cases where we select k to be eight and sixty as can be seen in Figures 2.5 and 2.6. In Figure 2.5 we see that for k=8, the k-means method clearly does not differentiate the clusters as one would expect, that is, one might expect to find eight clearly separate color regions for the ζ vs α plot, but we see that they overlap too much, this should not be surprising since we are just looking at the projections of a heptadimensional space in a bidimensional one. We have also plotted the cluster centers as black dots of greater diameter and we see that the cluster centers overlap in three separate regions. For Figure 2.6 the case is even more confusing since one would have to distinguish

[v]Checking the data in two dimensions more clearly it seem to me that at most one could say that there can be six groups

60 different colors. This situation doesn't get better when instead of colors we use numbers from one thru sixty to show which points belong to which cluster. The previous results are a good argument to use data reduction approaches, we propose to introduce biased classifications on the data, whether the bias has its origin in taking sequence into account, or from chemical considerations like taking A-type conformations into account, for example, more interesting results can be obtained as has been shown by others [13].

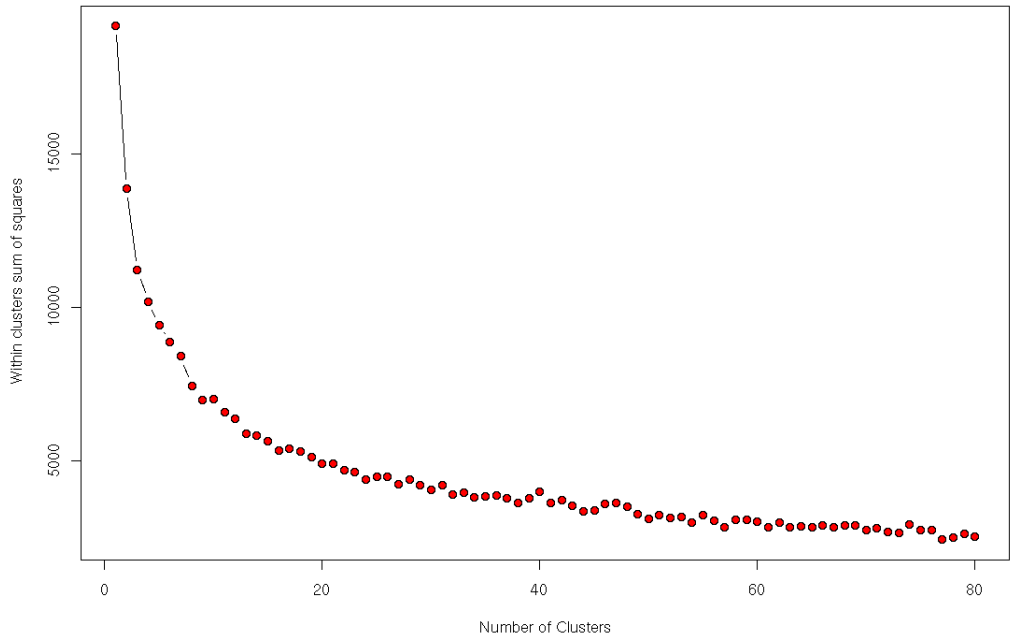

Figure 2.3: Sum of all within clusters sum of squares against number of clusters for data of all torsion angles in 23S rRNA.

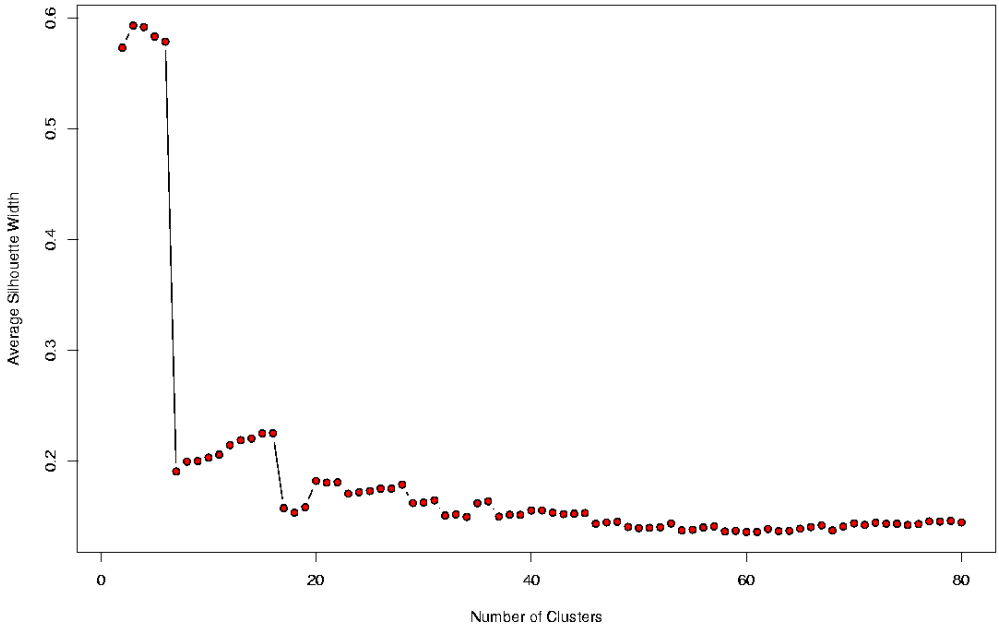

Figure 2.4: Average silhouette width against number of clusters for data of all torsion angles in 23S rRNA. The best clustering method and value of k is then defined as the model that maximizes a.s.w.

2.1.2 Hierarchical Clustering for Torsion Angles

Other authors have used hierarchical clustering to analyze the torsion angles in nucleic acid structure, taking the Fröbenius norm and Ward's method as the distance definition for four different RNA representations (see Reijmers et al. [8]) on a "small" database similar to that of Duarte and Pyle [14]. This databases do not include ribosomal RNA's. The other case where hierarchical clustering has been used did not use torsion angles, but rather a set of 15 atoms belonging to the nucleotides sugar and backbone. The latter study used used the unweighted pair group method (UPGMA) to classify a database of RNA loop structures (see Huang et al. [23]).

In our case we used three distance definitions (Euclidean, Manhattan, and maximum), and four different clustering methodologies, that is, single, complete, average and centroid (see A. We tried to make a consensus analysis of the twelve trees obtained, but these trees are too large. The number of trees, that is, twelve, is also too small compared with the data vectors (2753 step vectors), for the algorithms to find a reasonable consensus. We were not even able to find consensus for two "near" clusters, where the "near" criteria was taken from a tree dissimilarity algorithm implemented in **R** cluster and whose result can be seen in Figure 2.9 where the previously mentioned "near" clusters refer to

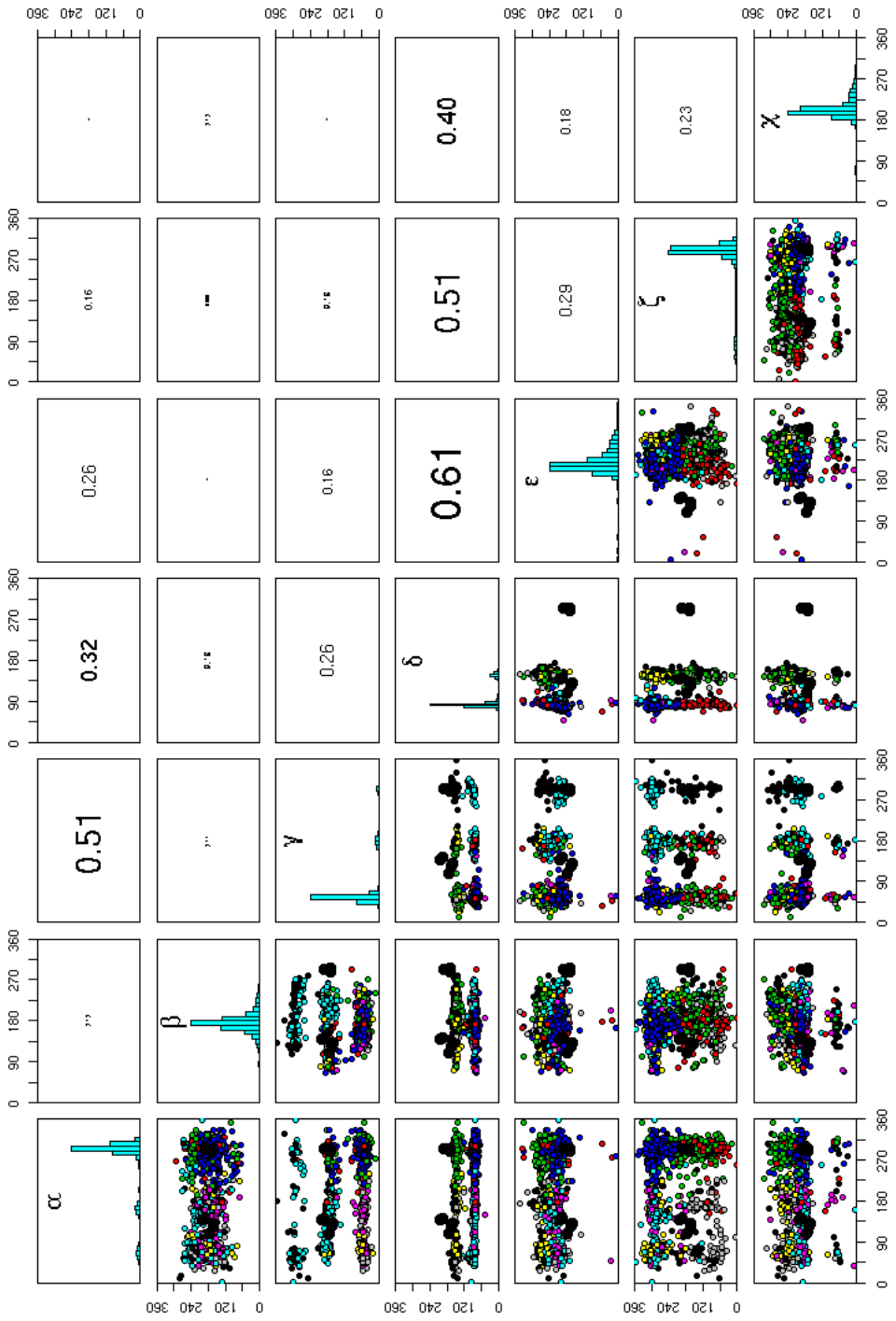

Figure 2.5: K-means clustering of heptadimensional torsion angle vectors of 2753 dinucleotide steps present in 23S rRNA. The number of partitions is **8**. The large black dots represent cluster centers. The upper diagonal matrix displays the values of the linear correlation coefficient r, and a histogram showing the torsion angle distribution is rendered in the diagonal.

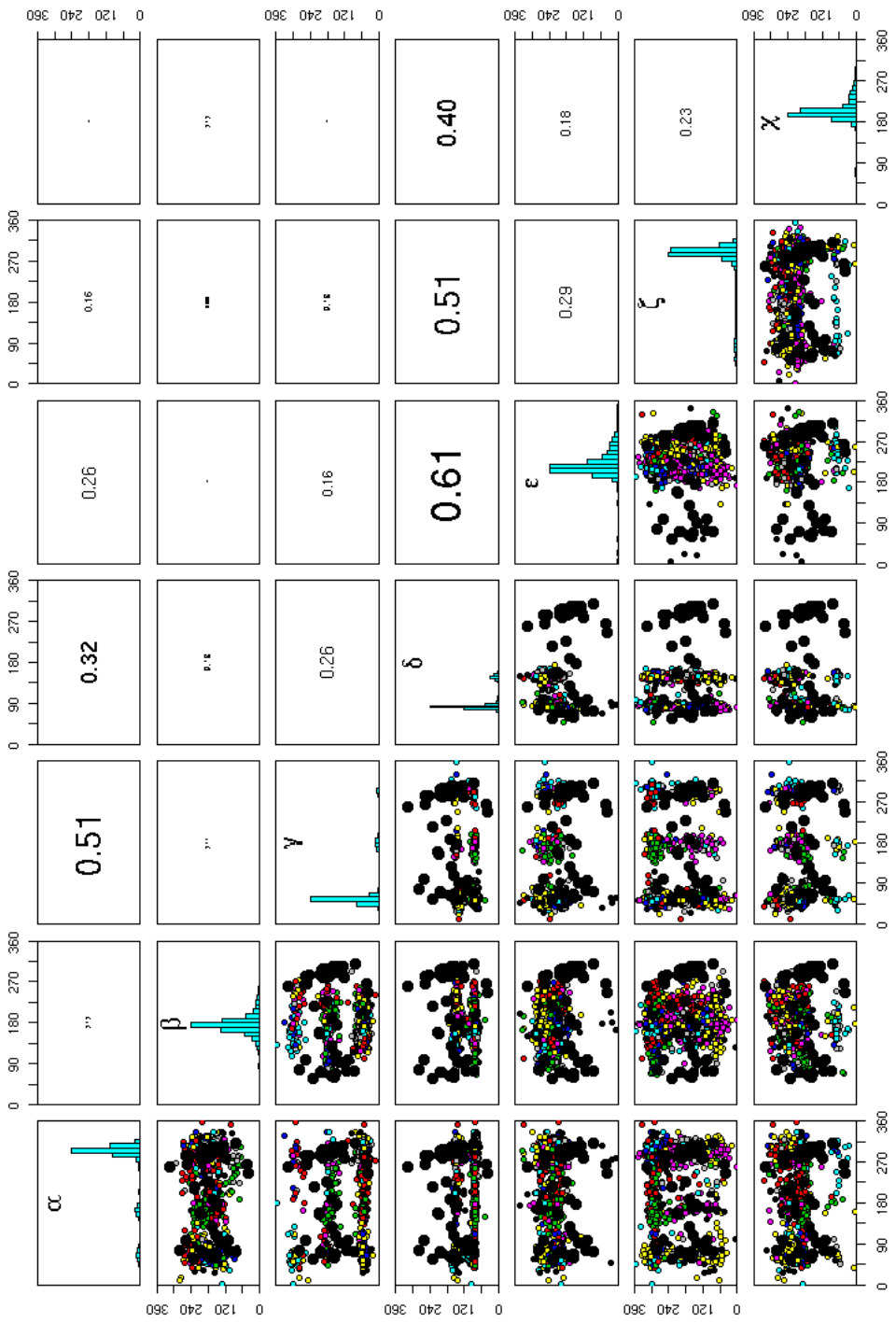

Figure 2.6: K-means clustering of heptadimensional torsion angle vectors of 2753 dinucleotide steps present in 23S rRNA. The number of partitions is **60**. The large black dots represent cluster centers. The upper diagonal matrix displays the values of the linear correlation coefficient r, and a histogram showing the torsion angle distribution is rendered in the diagonal.

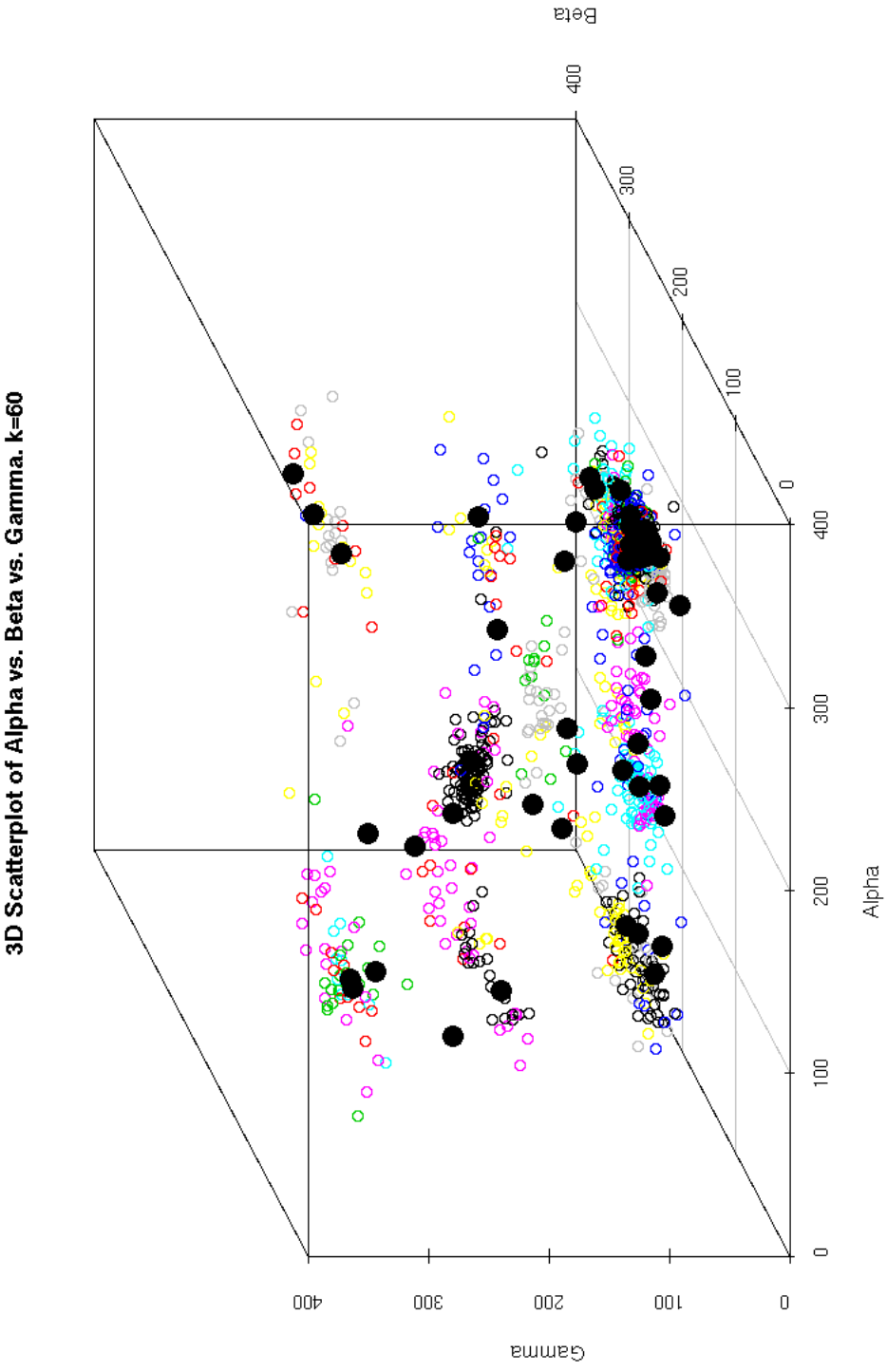

Figure 2.7: K-means clustering of the heptadimensional torsion angle vectors of 2753 dinucleotide steps of 23S rRNA. The axis of the three dimensional scatterplot corresponds to the torsion angles, α, β, and γ. The large black dots correspond with the cluster centers for clustering by using k-means with k=60.

clusters five and nine. One typical suggestion in clustering analysis is to just use the single linkage method since this one uses as its grouping criteria the minimal distance between clusters, therefore giving a direct interpretation of the trees as just a direct minimal proximity relation depending on the metric being used (see Appendix A). [vi] In Figure 2.8 we see twelve clustering trees clustering all base-steps in the large subunit of the ribosome. It's noticeable that trees are more similar when the linkage method is the same than when the metric is the same. The main problem with the dendrograms in Figure 2.8 is very similar to that of determining the number of partitions in partitional clustering. In this case we have to determine which is the optimal tree height which would determine how many meaningful groups we have, for Figure 2.8 we have selected to draw boxes around a group of branches at the height where 36 branches are found. The reason for selecting 36 is because this was one of the maxima in Figure 2.4 and because itÂt's close to 37, the number of discrete nucleotide conformations suggested by Hershkovitz et al. [13]

2.2 Base-step Parameters

To our knowledge there has been no classification of rigid-body base-step parameters for RNA structures deposited at the PDB [vii]. 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 2.10, reproduced from Richardson's work [10], 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.

2.2.1 Combining Fourier Averaging Results and Clustering Analysis

Using the coordinates files of 20 rRNA structures provided by Schneider at al.[12] we have used standard clustering analysis (CA) techniques to classify a set of non-ARNA base-steps using, rather than the torsion angles space, the base-step parameters space, that is, three translational parameters (ShiftD_x, SlideD_y, RiseD_z), and three rotational parameters (Tiltτ, Rollρ, Twistω), which are described by the hexaparametric vector ν:

[vi] The author still has to do consensus of single method trees.

[vii] The effort of putting together a database for such effort could be an interesting project to be considered.

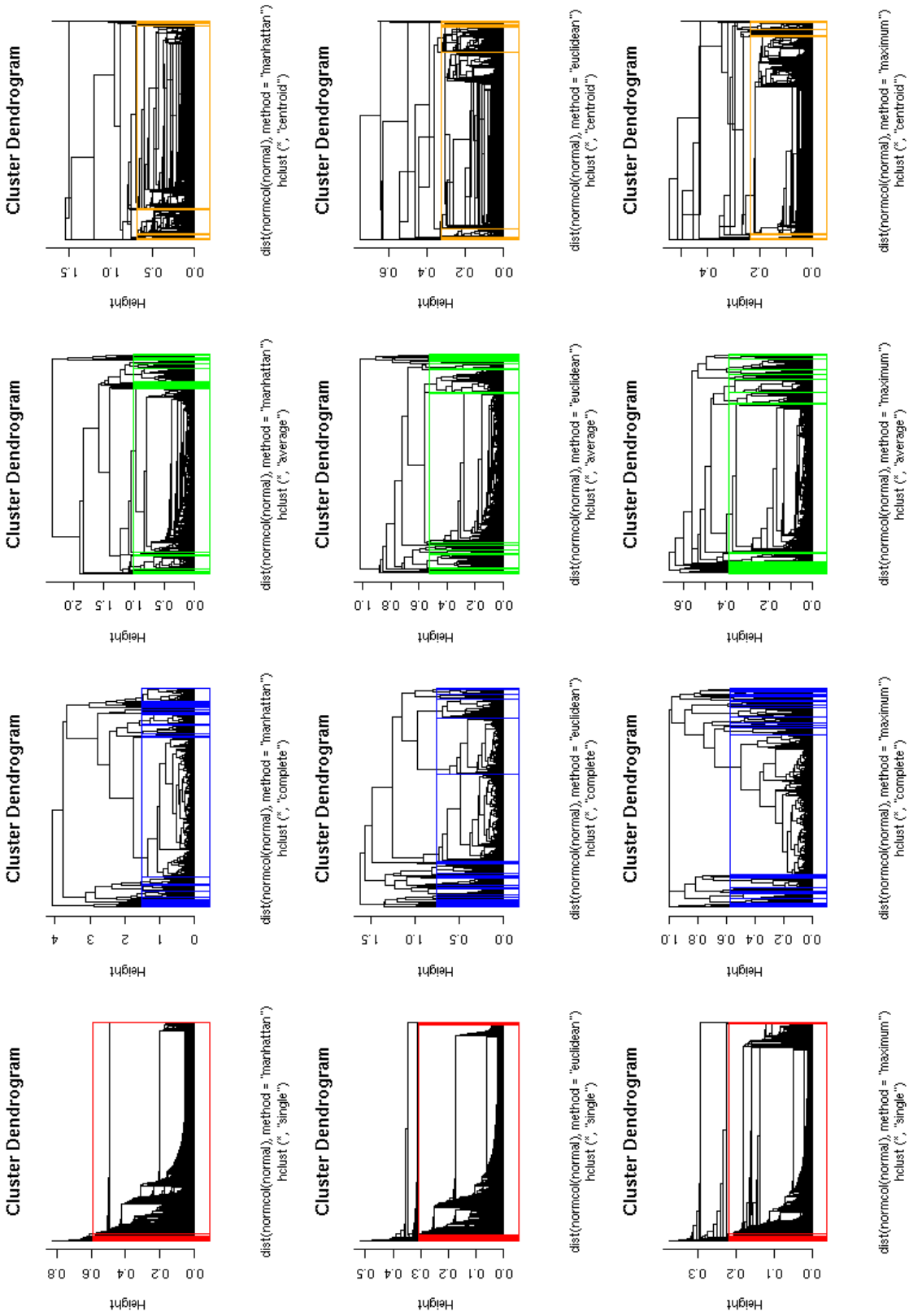

Figure 2.8: Hierarchical clustering for the twelve trees obtained from clustering of torsion angles of the large subunit of the ribosome (PDB-ID:1jj2). We have colored a box around branches for the case where the height of each tree has 36 branches.

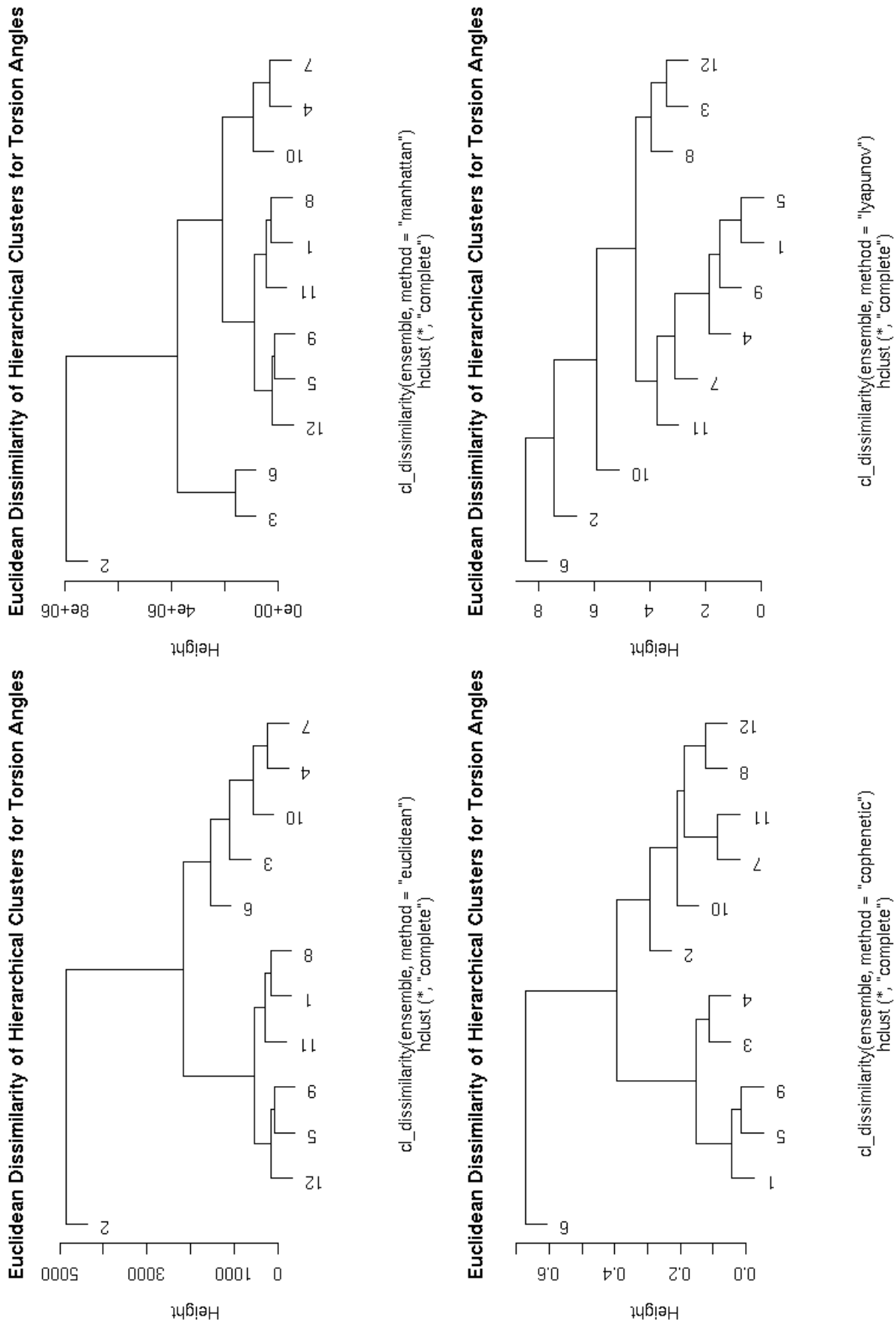

Figure 2.9: Cluster dissimilarities for the 12 combinations of metrics and methods used to obtain hierarchical clusterings of the 2753 heptadimensional torsion angle vectors of 23S rRNA.

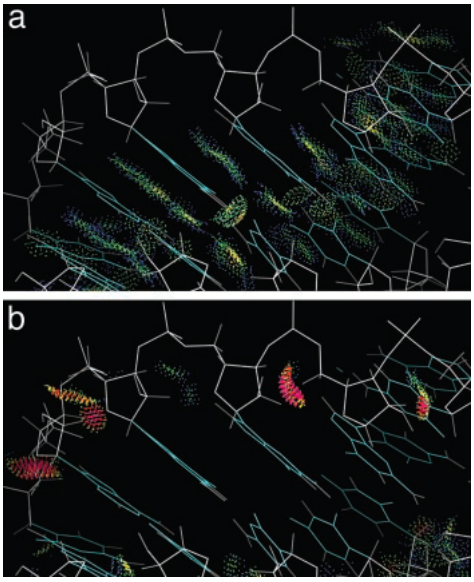

Figure 2.10: Figure taken from Richardson et al. [10] where the blue and green dots in a) mean very accurate van der Waals distances, and in b) the red and orange dots mean steric clashes, that is, distances outside the acceptable van der Waals range.

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

The results illustrated in Figures 2.11 and 2.12 were obtained by performing clustering analysis and consensus clustering on 20 structures provided by Schneider et al. [12]. 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 [24, 25]. Table 2.2 shows the residue numbers of bases from 23S rRNA which belong to the main categories of Figure 2.12. To decide which residues of 23S rRNA belonged 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.

2.2.2 Partitional Clustering for Rigid Body Parameters

The same type of analysis that has been carried out for torsion angles can also be carried out for rigid body parameters, that is, partitional clustering, and hierarchical clustering, are used as standard statistical analysis methods to analyze our set of 2753 base-step parameter vectors. For the partitional

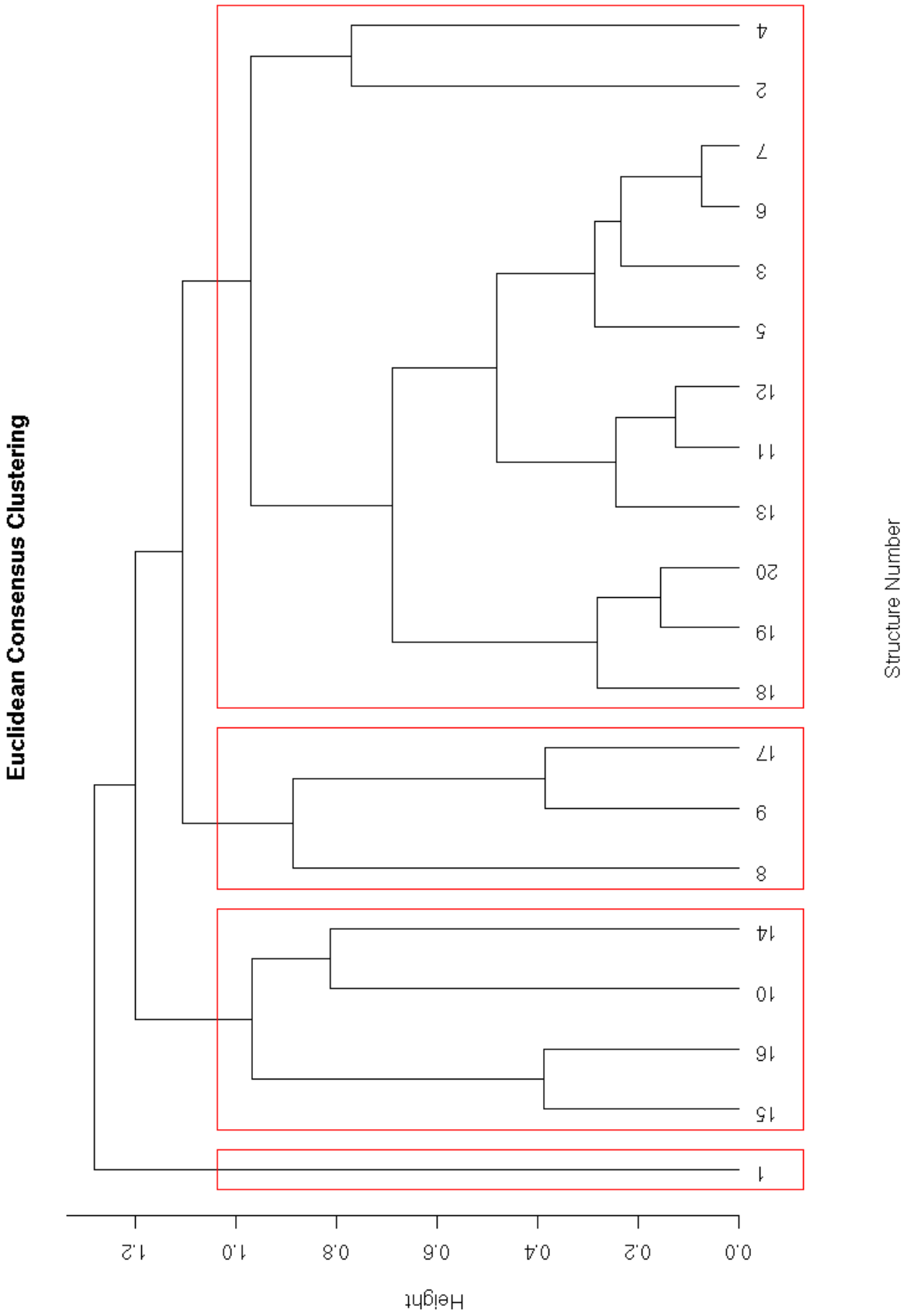

Figure 2.11: Dendrogram showing the results of consensus clustering of 20 non-Atype rRNA dinucleotides according to their hexadimensional base-step parameter vectors.

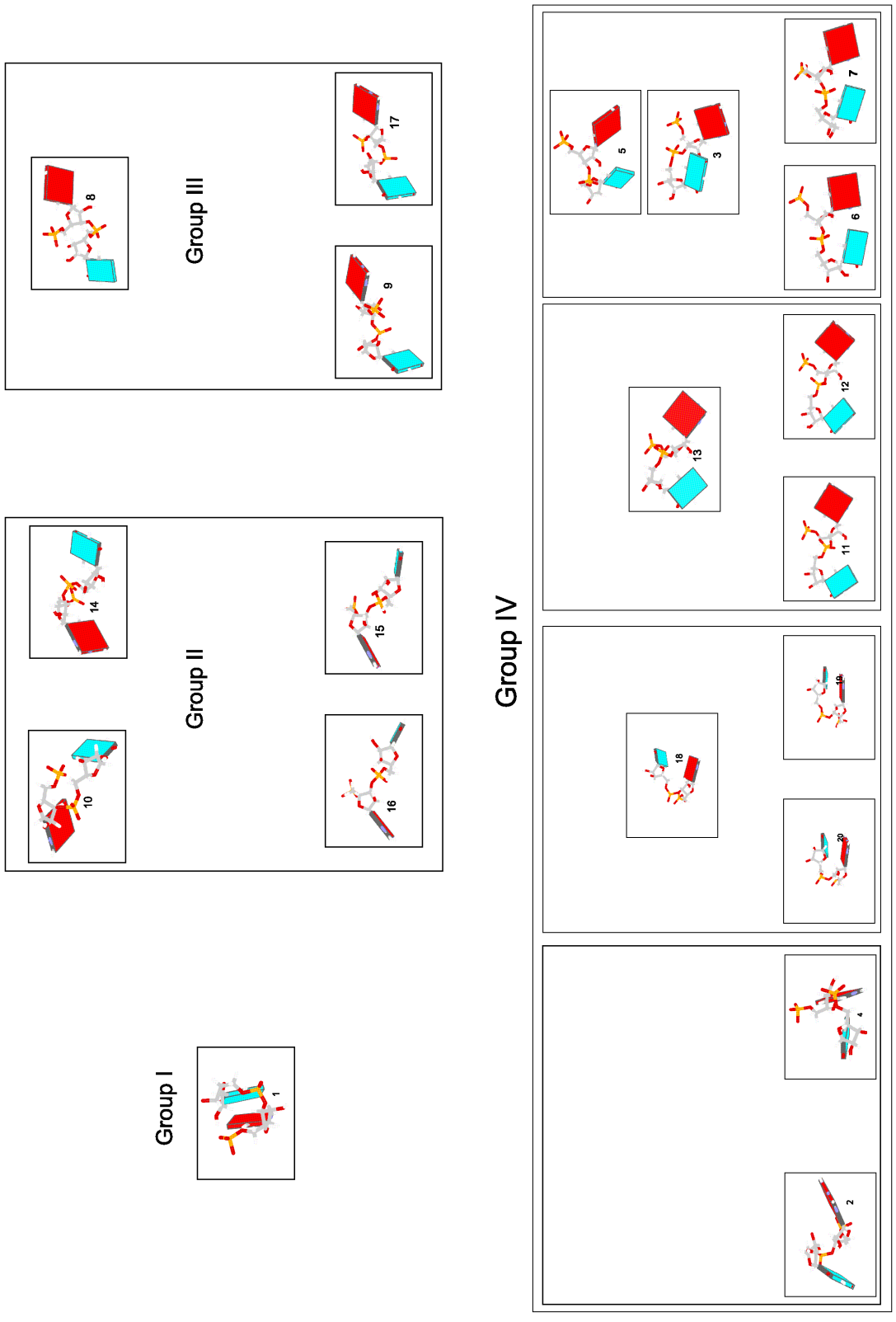

Figure 2.12: rRNA dinucleotide structures organized by clusters obtained from consensus clustering of their hexadimensional base-step parameter vectors.

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.

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 2.13 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 2.14, 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 bla and bla, and the PAM results for $k = 2, 9, 12$ in Figure bla.

We have also prefiltered 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.

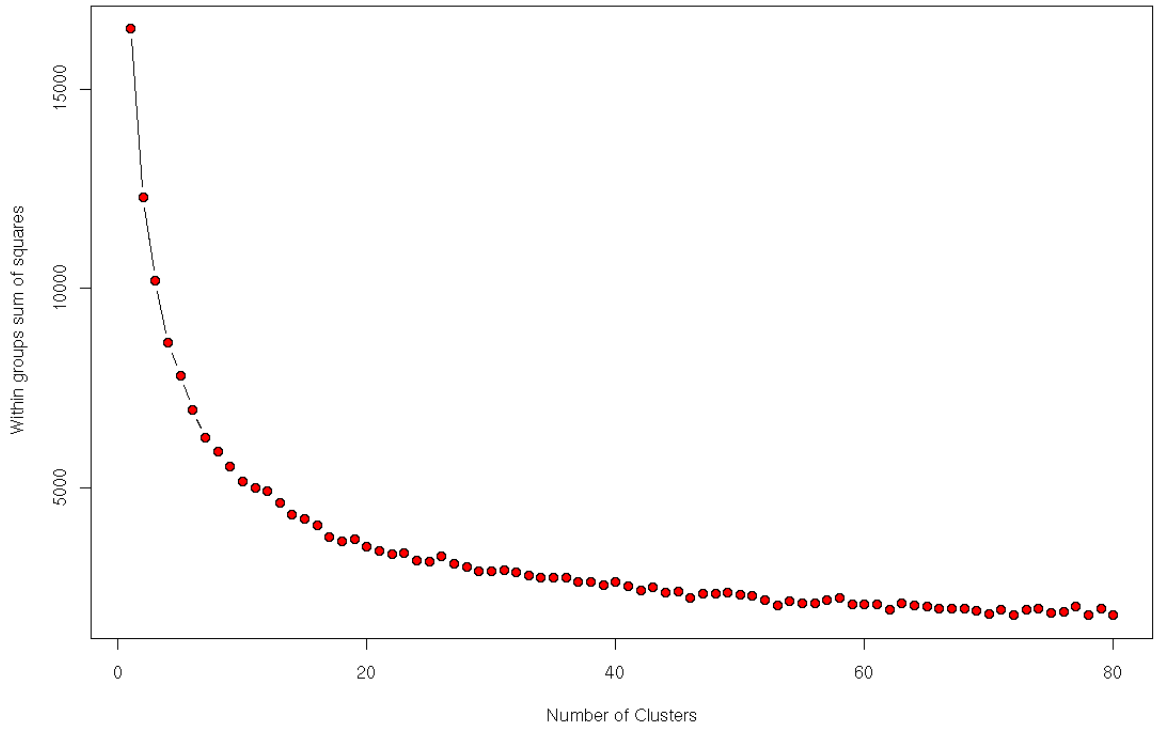

Figure 2.13: Sum of all within clusters sum of squares against number of clusters.

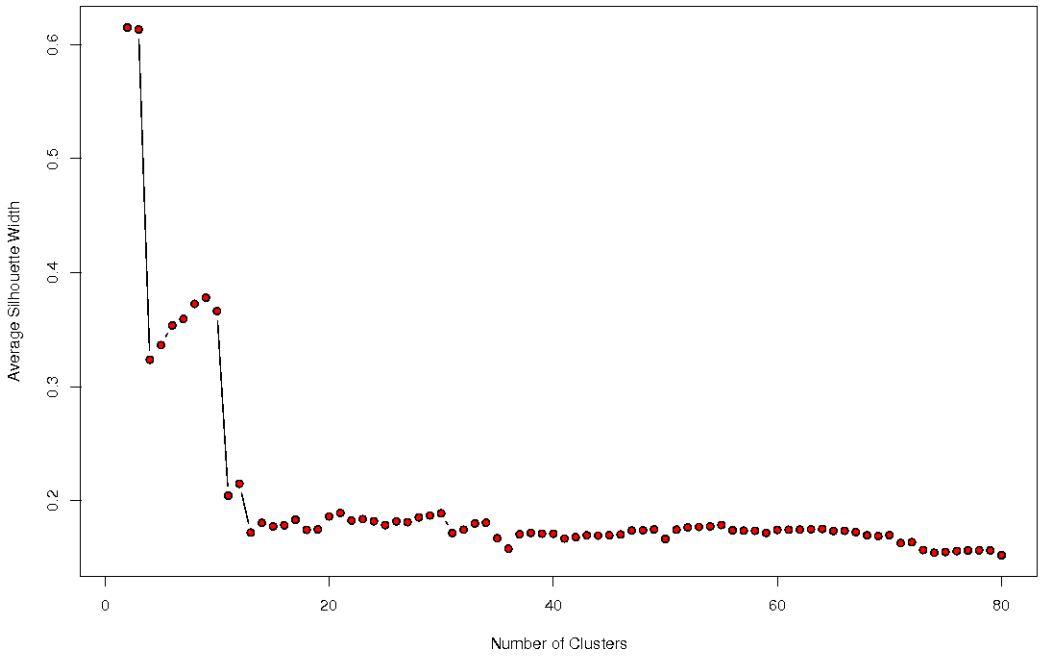

Figure 2.14: Average silhouette width against number of clusters.

2.2.3 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 2.15 for the 12 trees resulting from the four clustering methods and three distance definitions used to cluster the base step data.

2.3 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

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.

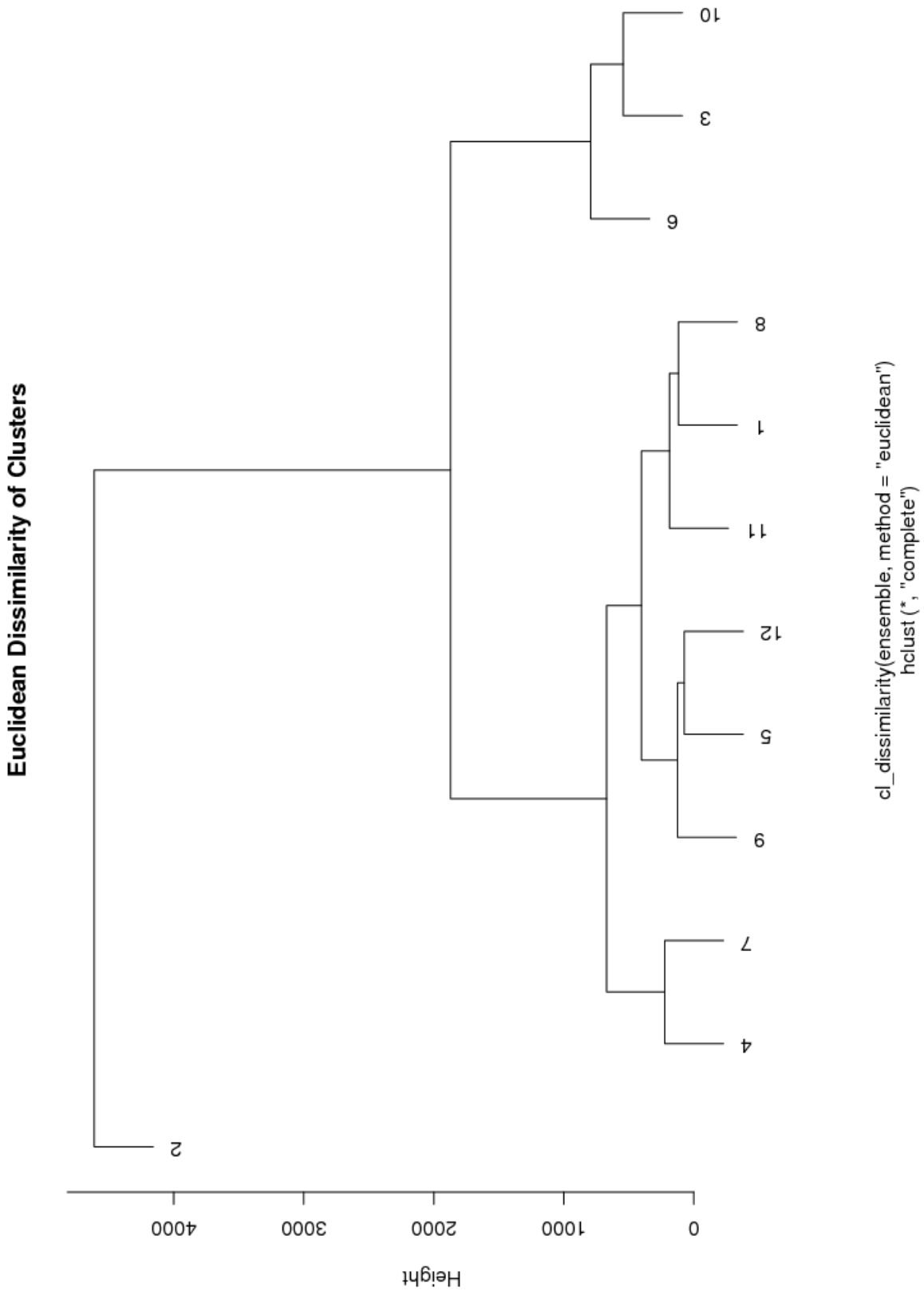

Figure 2.15: Cluster dissimilarities for the twelve hierarchical trees obtained from clustering of the six-dimensional base-step parameters obtained from the large subunit of the ribosome (PDB-ID:1jj2)

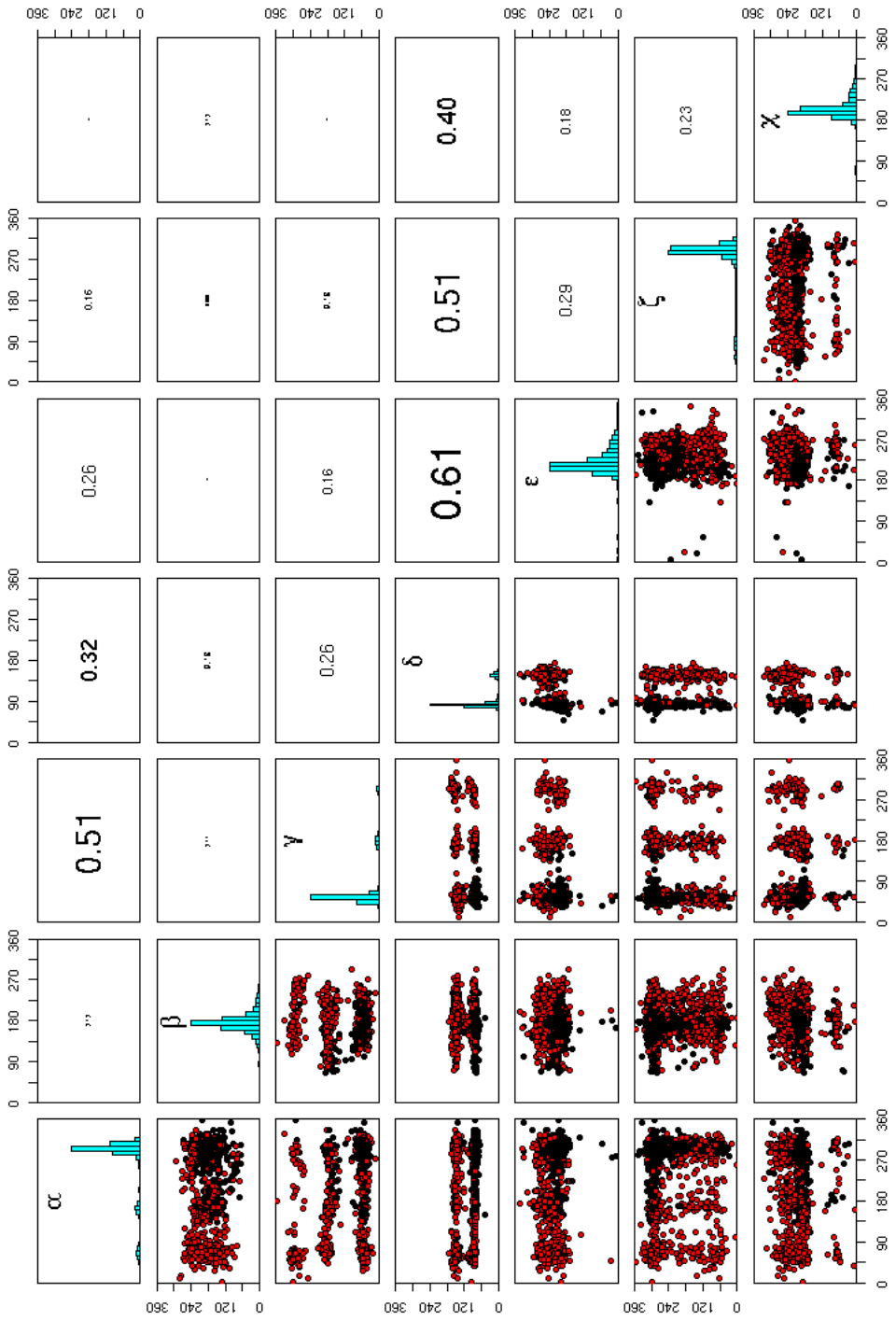

Figure 2.16: K-means of torsion angle vectors of 2753 dinucleotide steps present in 23S rRNA using the *Hartigan-Wong* algorithm. The number of partitions is **2**. The upper diagonal matrix displays the values of the linear correlation coefficient r, and a histogram showing the torsion angle distribution is rendered in the diagonal.

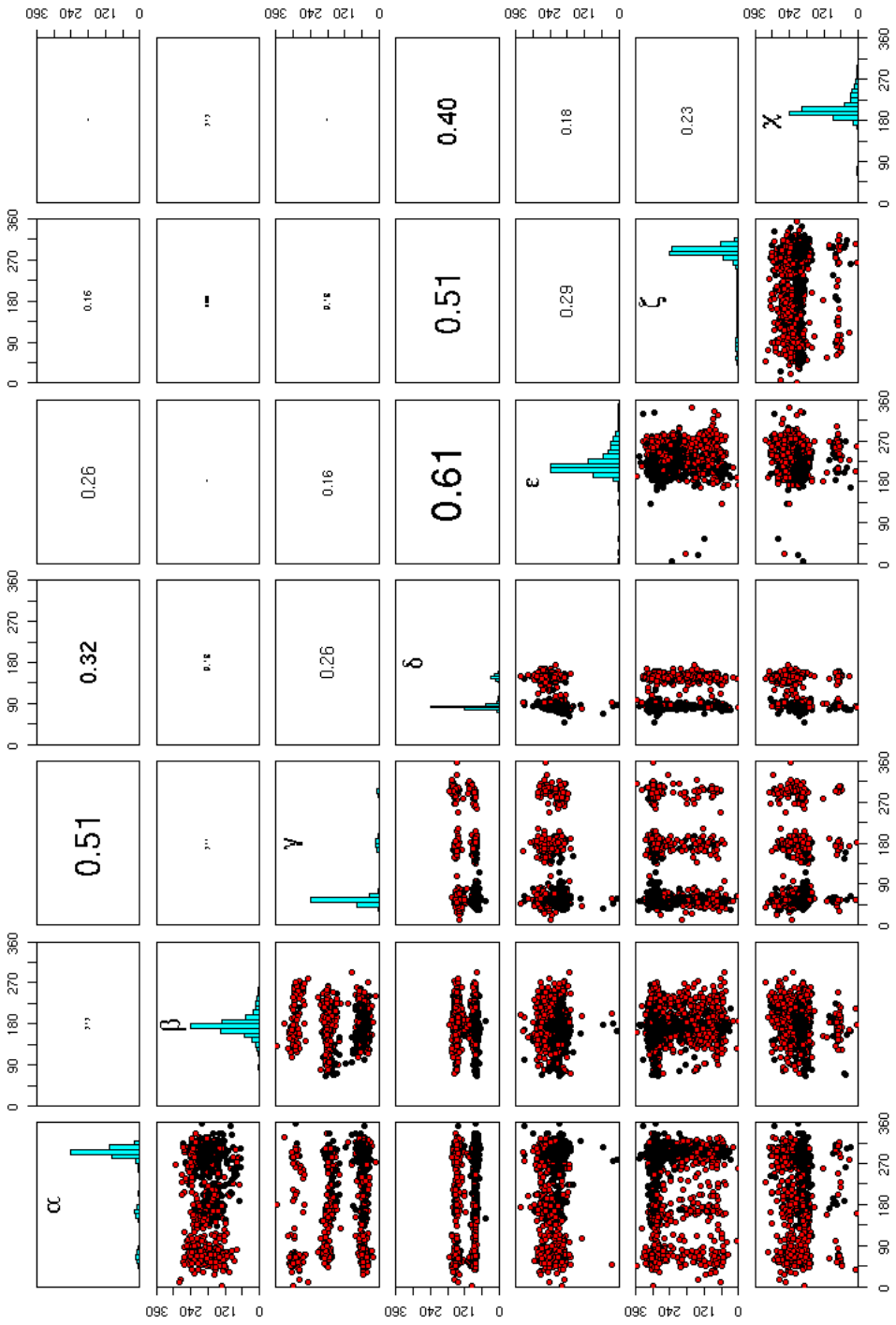

Figure 2.17: K-means of torsion angle vectors of 2753 dinucleotide steps present in 23S rRNA using the *Lloyd* algorithm. The number of partitions is **2**. The upper diagonal matrix displays the values of the linear correlation coefficient r, and a histogram showing the torsion angle distribution is rendered in the diagonal.

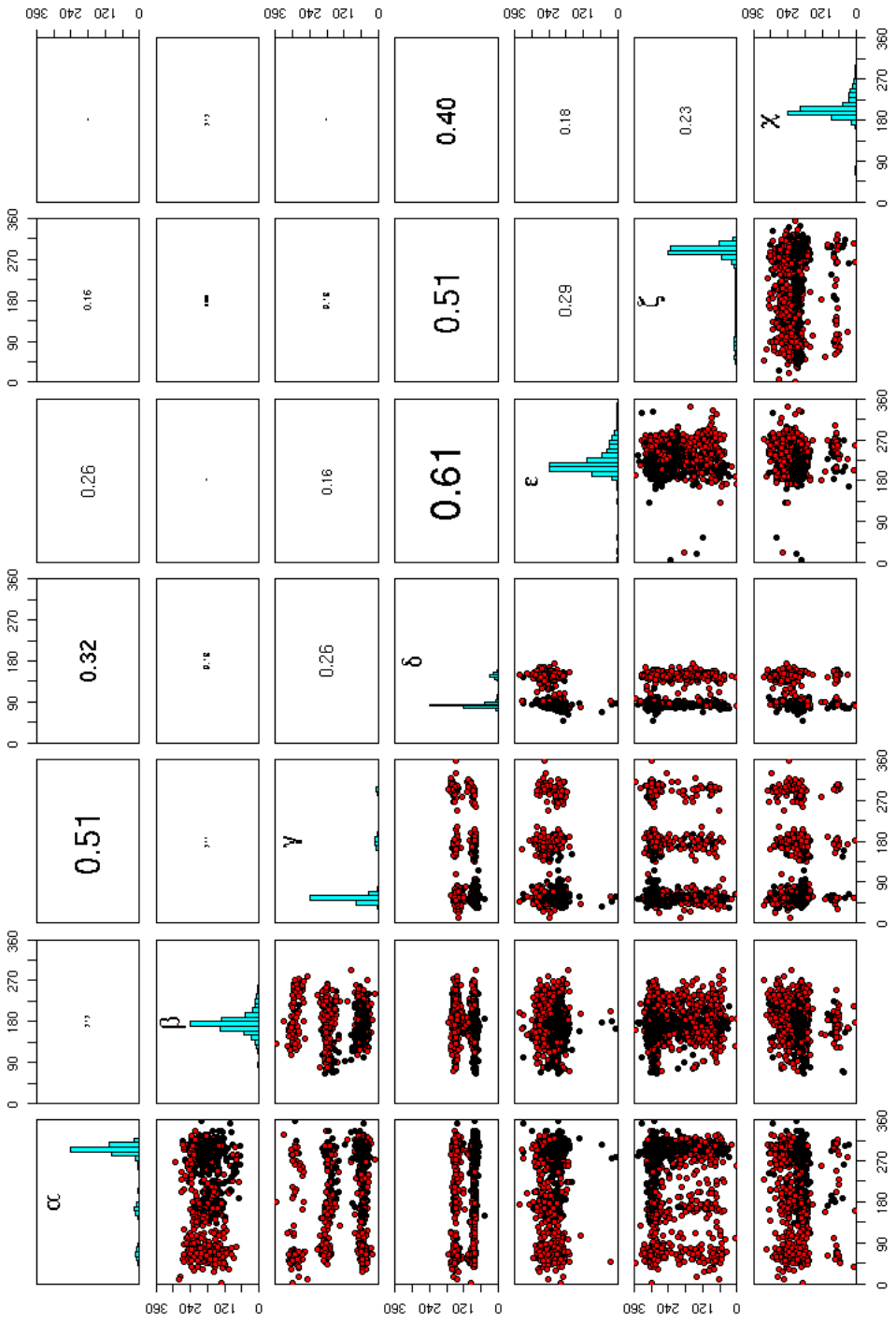

Figure 2.18: K-means of torsion angle vectors of 2753 dinucleotide steps present in 23S rRNA using the *Forgy* algorithm. The number of partitions is **2**. The upper diagonal matrix displays the values of the linear correlation coefficient r, and a histogram showing the torsion angle distribution is rendered in the diagonal.

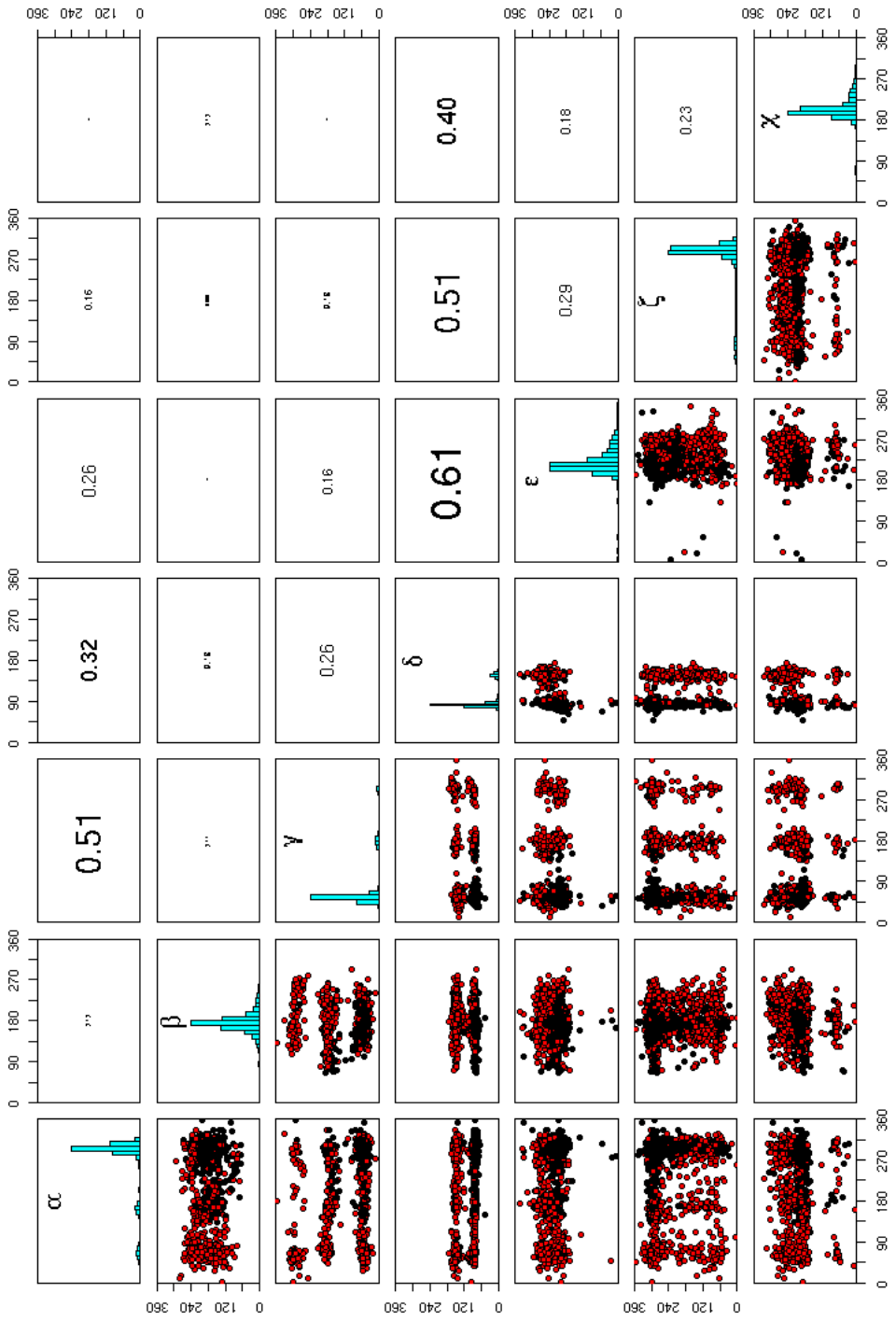

Figure 2.19: K-means of torsion angle vectors of 2753 dinucleotide steps present in 23S rRNA using the *McQueen* algorithm. The number of partitions is **2**. The upper diagonal matrix displays the values of the linear correlation coefficient r, and a histogram showing the torsion angle distribution is rendered in the diagonal.

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.

This chapter deals with how starting from a backbone based view of RNA, we can make an interpretation at the step level using the block model.

2.4 Consensus Clustering of Single Stranded Base Step Parameters

2.5 Four Major Non-ARNA Step Groups in the Ribosome

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**(4), 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. (August, 2000) The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution. *Science,* **289**(5481), 905–920.

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

[8] 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.

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

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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**(16), 4755–4761.

[16] 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.

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

[18] 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.

[19] 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.

[20] Jain, A. K., Murthy, M. N., and Flynn, P. J. (1999) Data Clustering: A Review. *ACM Computing Surveys,* **31**, 265.

[21] R Development Core Team (2007) R: A language and environment for statistical computing. ISBN 3-900051-07-0.

[22] 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.

[23] Huang, H.-C., Nagaswamy, U., and Fox, G. E. (2005) The Application of Cluster Analysis in the Intercomparison of Loop Structures in RNA. *RNA,* **11**, 412–423.

[24] 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.

[25] 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.

Chapter 3

RNA Base-Pairing

The RNA base-pairs are reviewed again.

3.1 Canonical and Noncanonical Base-pairs, Methods Paper

3.2 Clustering of Yurong's Classification

Chapter 4

RNA Base Pair Steps

4.1 Analysis (Albany Poster) and Django Webserver

4.2 Persistence Length vs. Hagerman

4.3 AMBER: Persistence Length of Base-Pair Step Patterns

Chapter 5

RNA Motifs

5.1 GNRA tetraloop

In order to compare our work to that of others on RNA structural motif localization and discovery, we ask the following questions:

1. Can the geometric rigid-block description of base-pairing and base-stacking solve the problem of defining RNA structural motifs?

2. Can we use quantities derived from the 3DNA software package to make and automatic search for a known motif, for example, the GNRA tetraloop motif, and perhaps find unknown motifs?

In the ROC meeting of May, 2009 a reduced dataset of RNA structures found at:
http://docs.google.com/Doc?id=dhrmkfmn_13ftpbjcgq
was made available to participants with the purpose of allowing them to search for RNA motifs, which would later be compared between groups. We have modestly, and as of yet unsuccessfully, started to aim at solving question number two. Initially we are trying to identify all instances of the well known GNRA tetraloop motif in the 23S subunit of ribosomal RNA of *Thermus Thermophilus*, PDB-ID:1ffk using results from 3DNA and 3DNA-Parser, and using an automated process which could be later reproduced for any desired dataset. Our hope is that these baby steps will allow us to to tackle the whole ROC dataset.

5.1.1 3DNA-Parser

We started by using Dr. Yurong Xin's 3DNA-Parser hoping that the description of the enclosing base pair in the loop, that is, the sheared G·A, would have a characteristic signature. We found that such is not the case. We know from Major et al. [4] that there should be at least 21 GNRA tetraloops in the 23S subunit of rRNA. We used the G2696 N2697 R2698 A2699 tetraloop as a seed (as can be seen in Figure 1.1) and found out that according to Dr. Xin's helical classification the enclosing G is classified

as S_{hq} and A is classified as H_e. We then searched all such instances for G·A base-pairs and we found

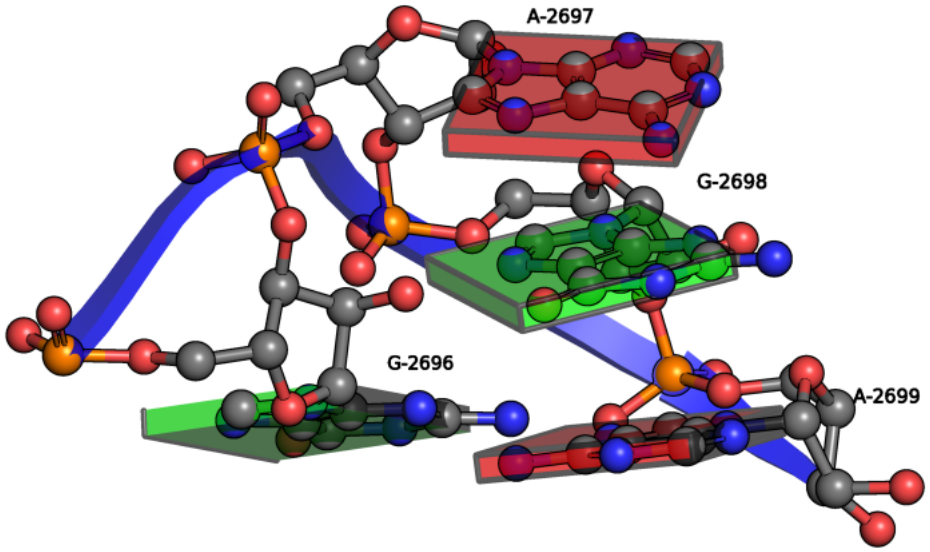

Figure 5.1: GNRA Tetraloop from *Thermus Thermophilus* 23S Ribosomal RNA PDB-ID:1ffk.

seven hits, but none were in fact GNRA tetraloops.

5.1.2 Overlap Scores

We clustered the overlap values impossing a cutoff of values of [1-8]. There are many values which are exactly zero (33%), so, without the cutoff the zero values "overshadow" the data. For this case we obtained a “good” dendrogram as seen in Figure 1.2.

The next step in this analysis will be to find the structures which correspond to this clusters and superimpose and align them using Kabsh’s algorithm to be able to determine their RMSD’s.

Many people start their RNA Motif identification and classification algorithms by splitting RNA structures into what is helical and what is not, and then finding interactions between these two groups. We believe that we could do a similar exercise with 3DNA by using the scalar product of helical axis vectors and once helical and non-helical regions are found we might be able to use 3DNA Parser to look for characteristic interactions.

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

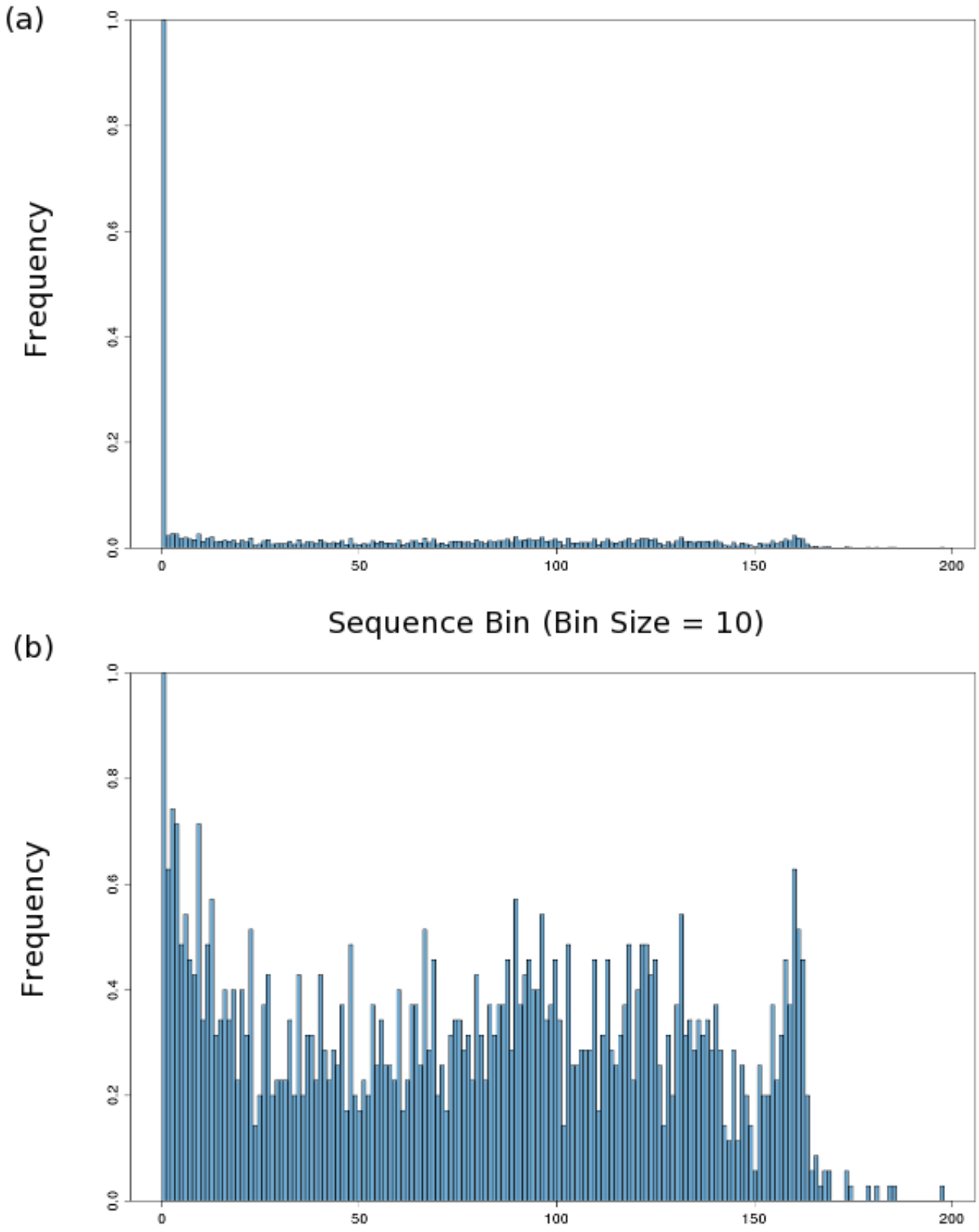

Figure 5.2: Normalized histograms showing the distribution of overlap values in the 23S subunit or *Thermus Thermophilus* rRNA, PDB-ID:1jjk. In histogram (a) all values are included, but in histogram (b) only values greater than zero are included. Notice the high preponderance of zero values, exactly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

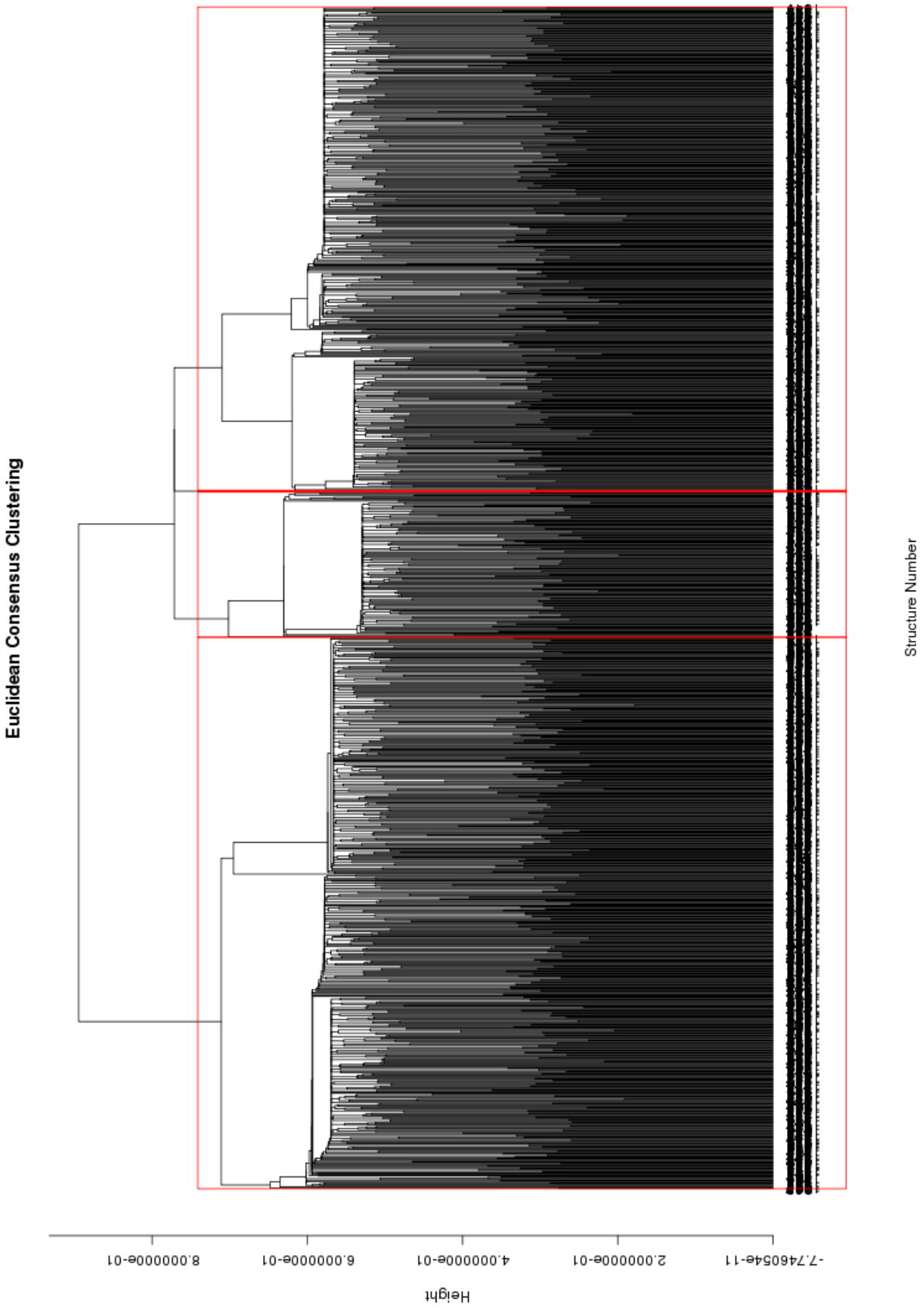

Figure 5.3: Dendrogram for consensus clustering of overlap scores in the ribosome. Zero values filtered out and remaining data normalized.

References

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

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

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

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

Chapter 6

RNA Helical Regions and Graph Theory

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

Appendix A

Clustering Analysis (CA)

A.1 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.

$$D(X,Y) = min\{d(x_i, y_j) : x_i \in X, y_j \in Y\} \tag{A.1}$$

   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.

$$D(X,Y) = max\{d(x_i, y_j) : x_i \in X, y_j \in Y\} \tag{A.2}$$

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

$$D(X,Y) = \frac{1}{N_x * N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} d(x_i, y_j) \tag{A.3}$$

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

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.

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

$$D(X,Y) = d(\overline{x},\overline{y}) \tag{A.4}$$

$$\overline{x} = \frac{1}{N_x}\sum_{i=1}^{N_x} x_i \tag{A.5}$$

$$\overline{y} = \frac{1}{N_y}\sum_{i=1}^{N_y} y_i \tag{A.6}$$

$$\tag{A.7}$$

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

$$D(X,Y) = ESS(XY) - [ESS(X) + ESS(Y)] \tag{A.8}$$

$$ESS(X) = \sum_{i=1}^{N_x} \left| x_i - \frac{1}{N_x}\sum_{j=1}^{N_x} x_j \right|^2 \tag{A.9}$$

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.

The first step is to chose a distance definition. We chose Manhattan and the distance values between structures can be displayed in a lower triangular matrix as seen in equation A.10

$$d(X,Y) = \begin{vmatrix} & 1 & 2 & 3 & 4 \\ 1 & & & & \\ 2 & 4 & & & \\ 3 & 8 & 12 & & \\ 4 & 11 & 9 & 5 & \\ 5 & 11 & 15 & 3 & 6 \end{vmatrix} \tag{A.10}$$

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

$$d(2,3) = |-2.00 - 6.00| + |2.00 - -2.00| = 12 \tag{A.11}$$

Now that we have calculated the distances we need a clustering method, in this case, we will use the average linkage clustering method. The first step is 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, that is, structures 1, 2 and 4, we obtain the following mean distances

$$D(\{3,5\},1) = \frac{1}{2*1} * (8+11) = 4.5 \tag{A.12}$$

$$D(\{3,5\},2) = \frac{1}{2*1} * (12+15) = 13.5 \tag{A.13}$$

$$D(\{3,5\},4) = \frac{1}{2*1} * (5+6) = 5.5 \tag{A.14}$$

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.12), 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.

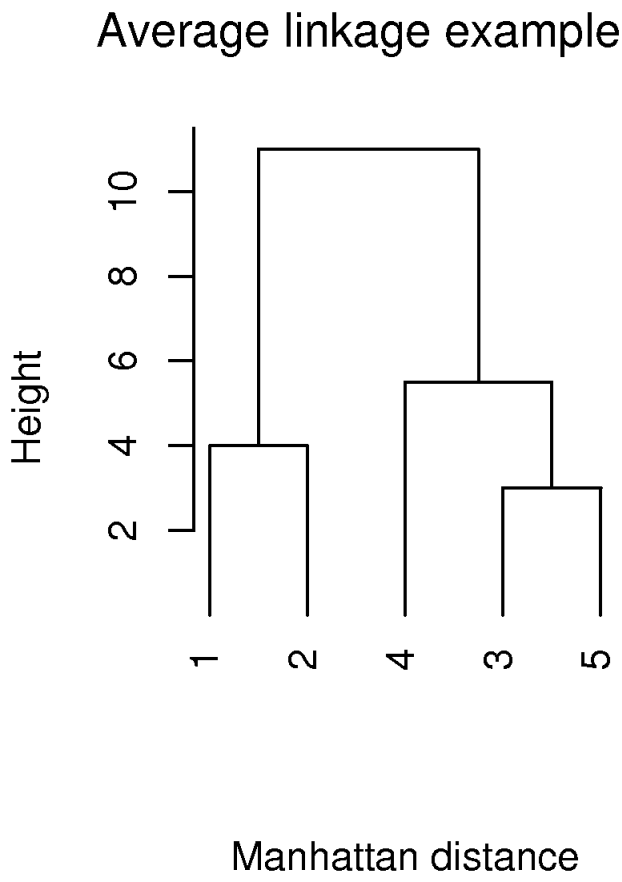

Figure A.1: Clustering tree for 5 bidimensional vectors using the Manhattan distance definition and the average linkage clustering method.

Curriculum Vitae

Mauricio Esguerra

La Mala Educacion

1991 High School Diploma from Gimnasio Moderno, Bogota, Colombia.

2000 B. Sc. in Chemistry from Universidad Nacional de Colombia

2010 Ph. D. in Chemistry and Chemical Biology, Rutgers University

Professional Experience

2003-2009 Teaching assistant, Department of Chemistry and Chemical Biology, Rutgers University

Publications

2009 W. K. Olson, M. Esguerra, Y. Xin, X-J. Lu, Methods