Structural rules for the formation of backbone-backbone interactions between closely packed RNA double helices

By

Fatou TAO

Submitted to:

The Faculty of Health Sciences in
partial fulfillment of the requirements for the
Master degree of Bioinformatics

Department of Biochemistry
University of Montreal

Avril, 2013

To my dear mother and father I love and cherish with all my soul

To my sisters, brothers and nephew

To my professors and colleagues

To all my friends

Acknowledgements

I definitely need to start by thanking God for being by my side throughout this journey. Thanks to Him who kept me on the right path, gave me faith, wisdom and patience, to make the good decisions, to stubbornly rise and pursue my dreams no matter the obstacles I met.

I am deeply indebted to my supervisor, ***Pr. Sergey Steinberg***, for his unfailing support and guidance through the completion of this thesis. My thanks go to him for giving me the chance to experience both part of the Bioinformatics science in wet and dry lab. However, his greatest contribution to my academic education is his teaching of how to look at science beyond the evidence and how to pay attention to apparently meaningless details where the truth might be discovered. This richness did not only serve me as a research student, but also as person.

Thanks to ***Pr. François Major*** and ***Pr. Stephen Michnick*** for accepting to be the judges this thesis. I am grateful to you and appraise the time you will spend on my thesis.

I must also thank the members of my laboratory. Each of them taught me in his own way. Thanks to ***Natalia Kotlova***, ***Tetsu Ishii***, ***Yury Butorin*** and ***Kostantine Bokov*** for sharing with me their immense knowledge, for their assistance and support.

I have a special thought for two ladies who lighted up those years I spent in this university. ***Elaine Meunier***, I don't have enough space to thank you up to everything you did for me. ***Marie Pageau***, you brought me more support and recognition than I could expect. Both of you have been like an ice cream cone under the hot sun of July. Many thanks for your listening to my academic and personal issues.

Finally, I thank my family and my best friend ***Judicaël Zounmevo*** for supporting me through this journey. They have been my cheerleaders I could rely on for better and for worst. I deeply thank ***M. and Mrs TAO***, my beloved parents, for their unerring overseas' support, advice and for all those hours spent on videoconference to lift my spirits when I needed it the most. I dedicate this thesis to my mother and father, for sharing their tremendous experience with me, for caring about my present, my future and for their unfailing belief in me.

May God Bless You.

Abstract

Although backbone-backbone interactions play an important role in stabilization of the tertiary structure of large RNA molecules, the particular rules that govern the formation of these interactions remain basically unknown. One RNA structural element for which the backbone-backbone interactions are essential is the along-groove packing motif. This motif is found in numerous locations in the ribosome structure; it consists of two double helices arranged such that the backbone of one helix is packed in the minor groove of the other helix and *vice versa*. The contact area between the two helices is mostly formed by riboses and totally involves twelve nucleotides. Here we analyze the internal structure of the along-groove packing motif and the dependence of stability of the association of the helices on their nucleotide sequences. We show that the proper positioning of the riboses that allows them to form inter-helix contacts is achieved through the particular choice of the identities of the base pairs involved. For different base pairs participating in the inter-helix contacts the optimal identities can be Watson-Crick, GC/CG, or certain non-Watson-Crick base pairs. The proper choice of the base pairs provides for the stable inter-helix interaction. In some cases of the motif, the identities of certain base pairs do not correspond to the most stable structure, which may reflect the fact that these motifs should break and form during the ribosome function.

Keywords: RNA structure; ribosomal RNA; along-groove packing motif; ribosome structure; recurrent motif; molecular dynamics

Résumé

Les interactions entre les squelettes sucre-phosphate de nucléotides jouent un rôle important dans la stabilisation des structures tertiaires de larges molécules d'ARN. Elles sont régies par des règles particulières qui gouverne leur formation mais qui jusque là demeure quasiment inconnues. Un élément structural d'ARN pour lequel les interactions sucre-phosphate sont importantes est le motif d'empaquetage de deux doubles hélices d'ARN le long du sillon mineur. Ce motif se trouve à divers endroits dans la structure du ribosome. Il consiste en deux doubles hélices interagissant de manière à ce que le squelette sucre-phosphate de l'une se niche dans le sillon mineur de l'autre et *vice versa*. La surface de contact entre les deux hélices est majoritairement formée par les riboses et implique au total douze nucléotides. La présente thèse a pour but d'analyser la structure interne de ce motif et sa dépendance de stabilité résultant de l'association optimale ou non des hélices, selon leurs séquences nucléotidiques. Il est démontré dans cette thèse qu'un positionnement approprié des riboses leur permet de former des contacts inter-hélices, par l'entremise d'un choix particulier de l'identité des pairs de bases impliquées. Pour différentes pairs de bases participant à ce contact inter-hélices, l'identité optimale peut être du type Watson-Crick, GC/CG, or certaines pairs de bases non Watson-Crick. Le choix adéquat de paires de bases fournit une interaction inter-hélice stable. Dans quelques cas du motif, l'identité de certaines paires de bases ne correspond pas à la structure la plus stable, ce qui pourrait refléter le fait que ces motifs devraient avoir une liberté de formation et de déformation lors du fonctionnement du ribosome.

Mots clés: Structure d'ARN; ARN ribosomique; motif d'empaquetage le long du sillon mineur; structure du ribosome; motif récurrent; dynamique moléculaire

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

Å	:	Angstrom
AGPM	:	Along Groove Packing Motif
MD	:	Molecular Dynamic
ns	:	Nanoseconds
PDB	:	Protein Data Bank
ps	:	Picoseconds
R.M.S.D	:	Root Mean Square Deviation
RNA	:	Ribonucleic Acid
rRNA	:	Ribosomal RNA
tRNA	:	Transfer RNA
WC	:	Watson-Crick

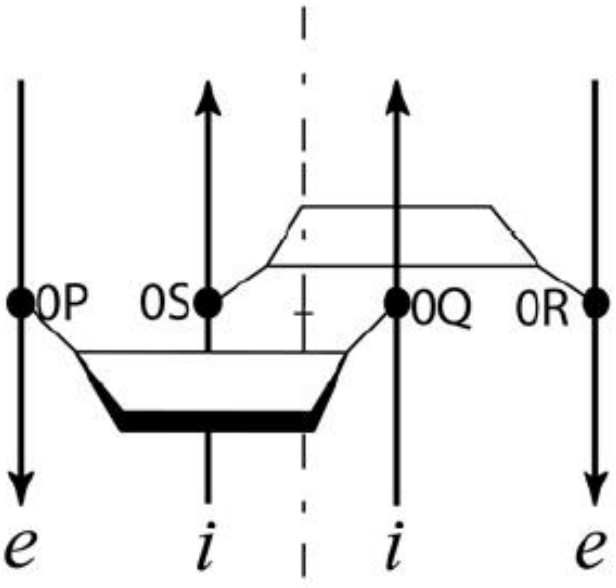
1. Introduction

An essential part of our knowledge on RNA structure is accumulated in the form of recurrent structural motifs, which appear in the same or different molecules and have identical or very similar conformation[1-11, 28, 35, 36]. The fact that such motifs can form in different structural contexts demonstrates a certain level of autonomy of their folding. Therefore, analysis of the aspects that govern formation of RNA recurrent motifs is important for understanding how larger RNA molecules fold and function. In most cases, analysis of the requirements for formation of RNA motifs has been focused on specific interactions that involve nitrogen bases, while the role of the sugar-phosphate backbone has been largely ignored. However, when the backbone participates in the formation of the core of the arrangement, the role of the interactions formed by the backbone can no longer be ignored.

A case of this kind represents the so-called along-groove packing motif (AGPM)[12], which consists of two double helices closely packed *via* minor grooves in the way that a sugar-phosphate backbone of one helix interacts with the minor groove of the other helix and *vice versa* (Figure 1). AGPM has been found in more than a dozen places in ribosomal RNA (rRNA), which makes this motif an important element of the ribosome architecture. Two more cases of AGPM are involved in the association of the P- and E-site tRNAs with the 50S subunit. The recurrence of AGPM and the active involvement of the sugar-phosphate backbone in its formation make this motif an excellent model for studying the general role of the backbone in RNA structure formation. Although since the discovery of AGPM, several studies concerning different aspects of the AGPM formation have been reported[12-15, 37], in none of them has the role of the backbone-backbone interactions been specifically analyzed.

In this paper, we undertake a systematic analysis of the aspects governing the interaction between the two double helices within AGPM. This analysis is based on the available X-ray conformations of the motif, on the collected data for more than sixty-five thousand available nucleotide sequences of AGPM, and on molecular dynamics (MD) simulations of specially modeled AGPM constructs. The analysis demonstrates a very

active role of backbone-backbone interactions in the shaping of the motif. We show that in different parts of the motif, the nucleotide identities are specifically tuned to provide for a stable collision-free interaction between the backbones of both helices. Because backbone-backbone interactions play essential role in the formation of different RNA complexes, including the ribosome, the rules that we discuss here are expected to be of general importance for RNA structure formation.



<u>Figure 1</u>: Description of the along-groove packing motif (AGPM) in a schematic representation.
Trapezoids represent the so-called central base pairs found the 0-level of AGPM and opened toward the minor grooves. The four nucleotides of this layer are named 0P, 0Q, 0R and 0S due to the strand of the helices to which they belong. Thus strands P and Q are respectively the external and internal of one helix while strands R and S are their homologs in the other helix. Arrows represent backbones directed 5'→3'. The internal and external strands of each helix are marked by italic letters i and e, respectively. The internal strand of each helix interacts with the minor groove of the other helix. Rotation of one helix for 180° around the axis of symmetry (dash-dotted line) leads to the superposition of both helices. (b): Juxtaposition of the central base pairs within AGPM. Arrows designate inter-helix hydrogen bonds directed from the donor to acceptor atom. The characteristic geometry of the GU base pair allows one helix to closely pack against the other helix in a collision free manner.

2. Methods

2.1. Insight II Software

2.1.1. Overview

Insight II is a graphic molecular modeling program used in conjunction with the molecular mechanics/dynamics program such as Discover or CHARMm to build and manipulate tridimensional virtual biomolecules, but also to study their molecular properties [42]. For the purpose of our study, we worked with Silicon Graphics Fuel computer in a UNIX environment. Insight II provide us with an adequate environment to model in silico AGPMs, undergo minimization in order to define the optimal structure, meaning the one with the lowest energy and ultimately, to perform dynamics simulations and analysis structural rearrangements within each AGPM mutant providing for a stability inherent to a specific base pair replacement. Figure 2 resumes the main steps of the current work with Insight II.

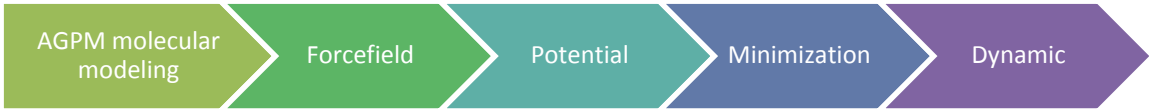


Figure 2: Diagram of the main steps followed in InsightII software.
Each modeled construct has been submitted to specifics forcefield and potential before minimization and dynamics. The details of those steps are presented in the sections below.

2.1.2. Biopolymers module for molecular modeling

The Biopolymer module facilitates the building and modification of peptides, proteins, polynucleic acids and carbohydrates. In the case of the current study, it allows to perform modifications through base pairs replacements, from the initial AGPM molecule construct to its clones. The Biopolymer module has been useful to perform the AGPMs molecular modeling. The mutants AGPM models are based on the conformation of the motif L657 (Figure 3) in the crystal structure of E. coli ribosome (pdb entry code 2aw4) with few base pair modifications. For instance, the [0R; 0S] central base pair of L657 was replaced by GC, resulting in the GU-GC base pair juxtaposition at the 0-level, which is the most favorable and stable arrangement for the central base pairs within AGPM (**Appendix 1**). Moreover, AU base pair [+1P; +1Q] was replaced by GC, which resulted in the GC-GU base pair juxtaposition at the +1-level (Appendix 2). The latter modification increased the stability of the helix formed by strands P and Q, and at the same time, the presence of the GU base pair [+1R; +1S] provided for a relaxed ribose-ribose contact between the internal nucleotides of the Q and S strands (as explained in the sections above). The -1-nucleotides remain unchanged such that [-1P; -1Q] and [-1R; -1S] are represented by a CG-CG base pairs combination; which respects the strong conservation of WC base pairs observed at the -1-level (see Figure 3, Appendix 4) while the -2-nucleotides of the motif L657, -2P and -2Q, as well as, -2R and -2S, do not form base pairs (Figure 3b, g). For the purpose of this analysis, we arranged the latter nucleotides such that they formed two WC base pairs [-2P; -2Q] and [-2R; -2S] (Appendix 3). Finally, we add additional WC-WC base pairs to form a -3-level.

From all those modifications of motif L657, the resulting AGPM construct we used as the initial model and as consensus structure is shown in Figure 4a.

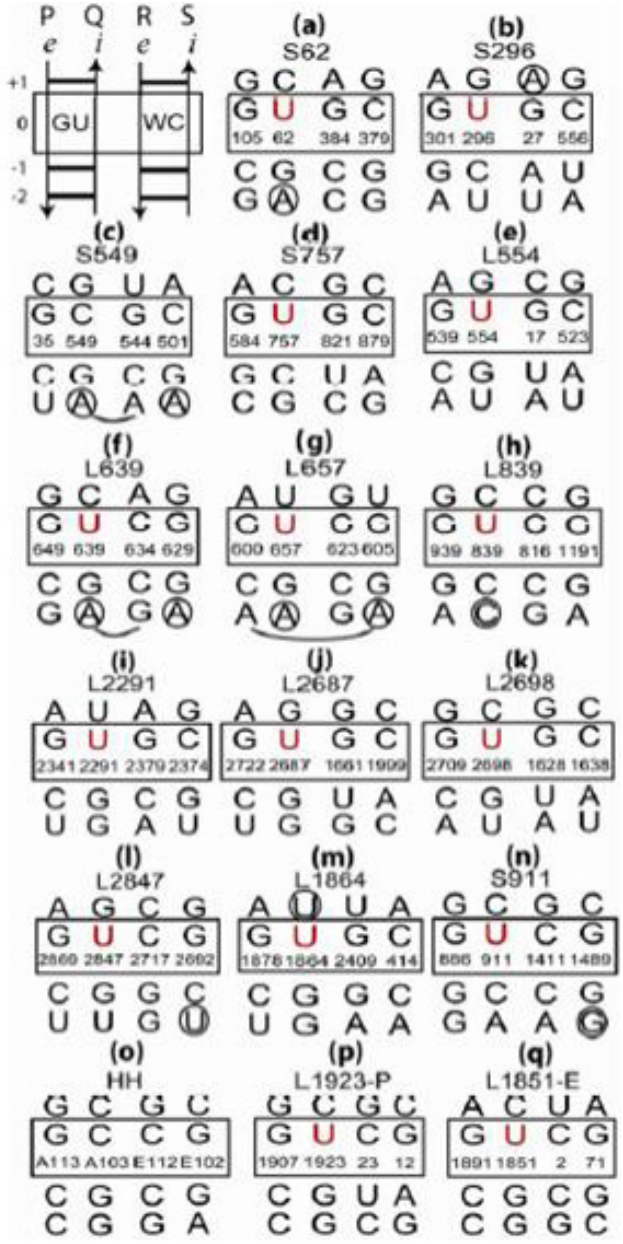


Figure 3: Nucleotide sequences of all known AGPMs identified within ribosomal RNA. Nucleotide sequences of all known AGPMs identified within ribosomal RNA (a-n), between two hammerhead ribozymes (o), and between 23S rRNA and a tRNA bound at the P-site (p) and at the E-site (q). Motifs (a-m) are taken from the structure of the *E. coli* ribosome (pdb entry codes 2avy-2aw4[19]); motif (n) is taken from the complex of the *T. thermophilus* 30S subunit and the initiation factor-1 (IF1) (pdb entry code 1hr0[22]); motif (o) is taken from the structure of the hammerhead ribozyme (pdb entry code 1hmh[27]); motifs (p, q) are taken from the structures of the *T. thermophilus* ribosome (pdb entry codes 2j00-2j01[25] and 1gix-1giy[23]). The positions and orientations of the GU- and WC-containing helices correspond to those shown at the upper left corner. Central base pairs are boxed. U in position 0Q is red. The *E. coli* nucleotide numbering is used for all cases found within the ribosome. The name of each motif starts with letter 'S' or 'L', which reflects the small or large subunit in which it is found, followed by the number in the standard *E. coli* nomenclature of the nucleotide occupying position 0Q in 16S or 23S rRNA.

Each modeled construct has identical nucleotide sequences except for one base pairs located at the level of interest. Moreover, all the constructs are composed of 18 nucleotide-helices forming AGPM and cap on both ends of both helices by GAGA tetraloops, in order to provide for additional stability to the structure during the simulations (Figure 4a, c). In the following discussion, we explain how we limit the effect on this supplemental feature. The simulations were performed on thirty-one different constructs.

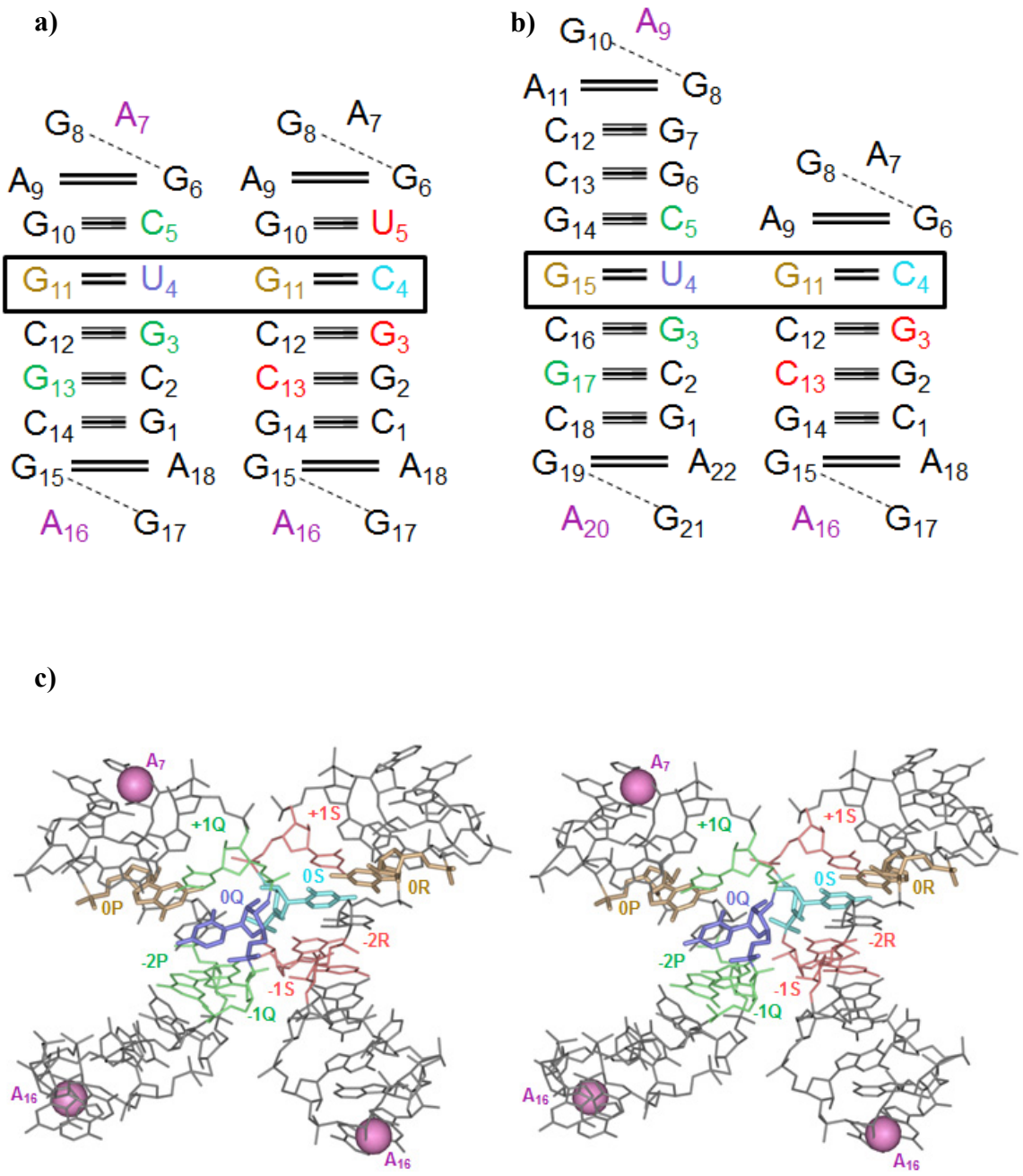


Figure 4: Schematic representation and stereo-view of the AGPM structure.
(a, b) shows the URS and the AGPM construct bearing a AG base pair at +1-level of the Watson-Crick helix respectively. The central base pairs are boxed. Each helix of the URS consists of 18 nucleotides and is capped by a GAGA tetraloop on both ends. For the AGPM in b), the helix containing GU at 0-level (GU-helix) bears two additional layers which increase the number of its nucleotides to twenty-two, while the opposite helix

(WC-helix) has one missing layer. Thus, the latter carries 16 nucleotides. Within base pairs, each solid line represents a distance constraint corresponding to an H-bond between two bases, while a dash line stands for a distance constraint corresponding to an H-bond between a base and a ribose. Details of the constraints used are given in the sections below. The nucleotides involved in the formation of the contact zones are green in the GU-helix and red in the WC-helix. Also, 0Q is dark blue, 0S is light blue, while 0P and 0R are orange. All other nucleotides not involved in inter-helix contacts are black. (c): Stereo-drawing of the tertiary structure of the modeled AGPM construct shown in panel a). The C1' atoms of the pink nucleotides, located within the GAGA tetraloops, were fixed during the molecular dynamic simulations. Their fixation helps to avoid uncontrolled deterioration of the construct at the regions outside the inter-helix contact. Based on such fixation, the WC-helix had enough freedom to dissociate or not from the GU-helix when the requirements for the stability of the overall structure where optimal or absent. Consequently, it becomes possible to estimate the stability of the whole arrangement.

From Figure 3, one can observe the predominance of AGPMs having +1AG in one the double helices and as previously mentioned, we capped both ends of the double helices forming AGPM with GAGA tetraloops. For instance, having a +1-AG base pair would have forced the superimposition of two AG base pairs; one at +1-level and the other one standing for the base pair of the tetraloop. Such arrangement is not observed in real structure because it is not structurally possible thus, it cannot did not fit our previous model (URS). Therefore, we modeled a construct which were structurally different from the URS. To overcome the problem of AGs superimposition and perform dynamic simulations on an AGPM harboring AG base pair at +1-level, several measures have been taken and are explained below. Ultimately, their lead to the AGPM shown in Figure 4b.

Measure 1: AG base pair of the tetraloop as +1-base pair

In our first attempt, we removed the +1-base pair in the helix where we wanted to insert the AG base pair. This modification brings the AG base pair of the GAGA tetraloop at the +1-level as shown in Figure 5.

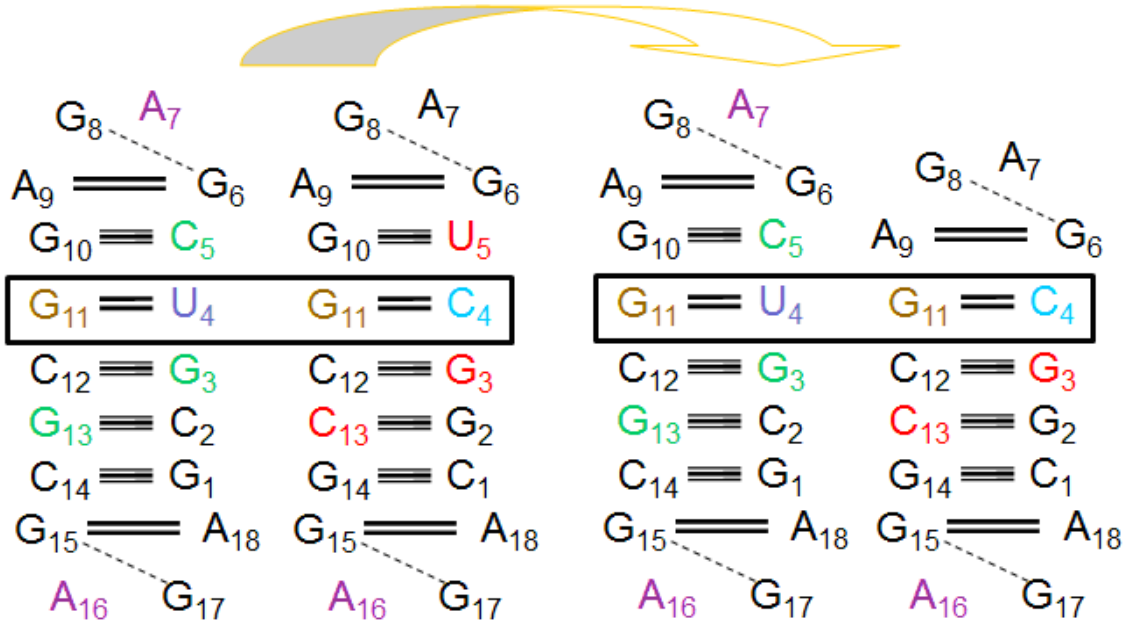


Figure 5: First attempt of modelisation of an AGPM bearing AG base pair at the +1-level of thw Watson-Crick helix.
The URS (left side) is reduced from its +1-level in the WC-helix which is replaced by the closing base pair of the tetraloop (right side).

We thus perform our +1-level analysis using this altered structure. However, we soon notice a parasite phenomenon that affects all performed simulations. Right after the beginning of the simulation the two unpaired nucleotides G and A, from the tetraloop of the shorter double helix (Figure 5), were interacting with the opposite tetraloop. Both A7 nucleotides (Figure 5) of opposite tetraloops come together and stack one on top of the other. This secondary interaction favors the interaction at the 0-level by bringing the two internal nucleotides close enough to interact without collision and prevent the interactions at -1- and -2-levels to occur by pulling upward the GU-helix. This phenomenon resulting of the solution 1 constitutes the problem 2. As our purpose is to study the interactions within the AGPM in response to base pair replacement, we needed to find a way out to prevent any interaction from tetraloop to influence the atom-atom contacts of our mutants. We thus introduce additional constraints between the two tetraloops of interest.

Measure 2: Introduction of distance constraints

We stated that the distance between the two A7 nucleotides should not be less than 15Å, which was the distance observe between both nucleotides before performing the molecular dynamic simulation. In case those nucleotides get close to each other at less than 15Å, we imposed a penalty of 100 to the energy of the overall arrangement. On top of the fact that this restriction didn't solve the initial problem, we decided to lower the distance constraint from 15Å to 5Å with the same penalty on the energy. This new constraint was more realistic and more than enough to enable the molecule to 'breathe' during simulations and still preventing the secondary interaction described above to occur.

For this second attempt, we realise that during the simulation, as the AGPM construct breathes, if for whatever reason the two restrained A7 become close to each other for less than 5Å, the molecule 'explode' because of too high penalty.

A third try was therefore to associate an energetic penalty of 10 to the energy of the overall structure if the distance between the A7 was less than 5Å. Unfortunately, this ultimate attempt concerning the introduction of distance constraints did not solve the

problem of secondary interaction in the tetraloops that favor 0-level atom-atom contact within the AGPM.

Measure 3: Introduction of two additional layers on top of the +1-base pair which is not AG

To prevent the parasite interaction of the A7 nucleotides from the two tetraloop capping the positive side of the mutant AGPM containing a +1AG-base pair, we model a new AGPM having two additional layers on top of the +1-base pair which was not AG (Figure 4b).

At this point, the constraints imposed on this motif were the same than the ones on the initial constructs and they are discussed in the following argumentation. After each AGPM modification, the use of the Discover program helped us to create a proper environment for the AGPM constructs molecular dynamic simulations.

Another issue we had to deal with was the cyclisation of AGPM construct. Building a new constructs necessitates creating and/or breaking of covalent bonds for the introduction of base pair replacement. The biopolymer module enables to do so through its 'Modify' menu. Moreover, one can also create covalent bonds from the 'Ligate' option of the 'Nucleic Acid' menu. The latter was used for each AGPM construct modeling while only the last covalent bonds from each double helices were created, after all modifications and before setting of potentials, through the 'Modify' menu. Besides, the two double helices forming the APGM were unmerged through the 'Unmerge' option of the 'Modify' menu before base pairs manipulations, which allows working on two separated molecules. This strategy enables us to solve the problem of cyclic molecule which led us for a while, to false positive molecular dynamic results.

2.1.3. Discover module

Discover module provides an interface to the Discover program which allows performing minimization, template forcing, derivatives, means square displacements, vibrational frequencies[39]. It also provides tools for performing simulations under various conditions including constant temperature, pressure and stress, periodic boundaries, fixed

and restrained atoms. In the case of the current study where we want to determine the local and global rearrangements on the surface of both interacting helices caused by specific base pairs replacement and how those structural reorganisations influence the ability of one helix to niche optimally in the minor groove the other and vice-versa, particular methods and strategies have been adopted. As we are dealing with dynamic calculations, our results are directly related to the forcefield we use, which is the calculation of the potential energy for a given configuration of atoms[39]. The forcefield represents the single largest approximation in molecular modeling and its quality, applicability to the system at hand and its ability to predict the particular properties measured in the simulation directly determine the validity of the results[39]. Among the available four families of forcefields supported by the Discover program – CVFF, CFF91, ESFF, AMBER, the latter has been used in our study.

2.1.4. Minimization

When an AGPM is built through a process of adding fragments which can generates serious atom clashes, it usually needs to be refined. Thus, the molecule is brought to a stable, sterically acceptable, conformation by the process of minimization. Minimisation is an iterative procedure in which the tridimensional structure is brought to a minimum through adjustments of atoms coordinates and where the obtained molecule with the lowest energy is considered to have the most stable arrangement and the closest resemblance with the real physical AGPM structure. The calculation of the energy follows a classical approach in which the molecule is considered as a set of charged point masses (the atoms), coupled together with springs and the total energy of the system is calculated through the summation of a number of individual energy terms. This calculation encompasses bond stretching, valence angle bending, torsion and nonbond interaction terms which associate an energetic penalty to the structure based upon deviations from an idealised equilibrium geometry. The nonbond interaction energy sums the van der waals attraction and repulsion as well as electrostatic forces for all atom pairs in the structure that are 1-4 nonbonded and above. Moreover, the value of bonds and angles cannot be based upon element type alone. As an example, a carbon-carbon single

bond is longer than a carbon-carbon double bond. For instance, before running energy calculation, the structure needs to have a potential type assigned. In the case of our study, all potentials are fixed.

2.1.5. Steepest descents method

In this low but robust method, each line search produces a new direction that is perpendicular to the previous gradient and the directions oscillate along the way to minimum. Even though the convergence is slow near the minimum because the gradient approaches zero, the steepest descent method is extremely robust. It helps generating a lower-energy structure regardless of the function. For this reason, the steepest descent method is used when the gradient is large and the conformations are far from the minimum, which was our case. We performed 300 iterations per steepest descent minimization. No constraints have been imposed on our mutant AGPMs during the simulations.

2.1.6. Conjugate gradient method

The second move in term of minimization was to use the conjugate gradient method. It allowed us to refine our search of the lowest-energy structure by refining step by step the direction toward the minimum. Moreover, to maintain the integrity of both helices forming AGPM, minor distances constraints were imposed on the lengths of the hydrogen bonds in all base pairs, except the central ones. These constraints were introduced as penalty $K \times (R - 3.3)^2$ added to the energy function when the distance R between the two electro-negative atoms involved in the formation of a corresponding hydrogen bond exceeded 3.3Å. Besides, all nucleotides of all GAGA tetraloops were fixed. Each simulation necessitates 2000 iterations, which enabled the structure to reach convergence.

2.1.7. Molecular Dynamic

Even though the minimization step is of an important use to reach the lowest energy structure, molecules still static and do not represent the reality where molecules are flexible structure subjects to thermal motion. This is the reason why, from the energetically stable structure obtained from previous steps, we used molecular dynamic to simulate the thermal motion of AGPM as a function of time, using the forces acting on the atoms to drive the motion. The forces acting on the atoms can be evaluated with Newton's second law of motion which is F = m * a, where *F* is the force, *m* the masse of an atom and *a* is the acceleration. The acceleration and velocity are then used to calculate new positions for the atoms over a short time step, thus moving each atom to a new position in space. The process iterates, generating series of conformations of the structure which we call trajectory. The velocity of the atoms is directly related to the temperature at which the simulation is run. In our study, the simulations were run at 300K, which provides information on the structural fluctuations that occur around the starting conformation, and the pathways of conformational transitions. The C1' atoms of A7 nucleotides from both GAGA tetraloops in the GU-helix have been fixed during all molecular dynamic simulation, as well as the one from the lower tetraloop of GC-helix. This restriction was necessary to maintain some stability and avoid parasite movements during the simulation. The fourth tetraloop remained unrestrained to give AGPMs some level of freedom and allow them to break or form according to the affinity of interaction of both double helices describing this motif. All mutants AGPM were simulated several times (3 - 10) in a probabilistic process where the number of repeat was determined by the tendency of stability for each simulation. For each repeat, we follow the thermal motion of our mutants throughout one nanosecond (1 000 000 iterations).

2.1.8. Quantitative evaluation of AGPM stability

As mentioned above, AGPM is a motif whose formation involves non-specific interactions of base pairs spreading over four levels. To study the structural rules governing the formation of this motif, we undertake a systematic analysis of inter-helix

interactions at each level, one at a time. This analysis consists in the replacements of base pairs followed by testing of the stability of the modified AGPM. The stability was tested through dynamic simulations of one nanosecond. For each simulation, we monitored the four inter-ribose interactions, which involved the areas of contact described in Figure 7. For backbone-backbone interactions at the +1- and 0-levels where only internal nucleotides are involved (QS zone), the atom-atom contact is made between both C4' of +1Q and +1S and both O2' of 0Q and 0S. For interactions spreading over -1- and -2-levels, the C4'-atom of the -1Q internal nucleotide contacts the -2R external nucleotide via its O2'-atom. The similar situation occurs for the symmetric SP zone where -1S and -2P interact together via their C4' and O2' respectively (Figure 7, Table I).

Each simulation (monitoring 4 contacts at the same time) gave us the overall stability of the studied construct. After several simulations of the same construct, we calculate the average stability, which becomes a characteristic of the given construct. We discuss this stability in the following text. For the current study, we did not use molecular dynamic simulation to generate a space of different conformations AGPM was able to tolerate. We used it as a tool to understand the structural rules governing the AGPM formation through specific base pairs replacements at each level where inter-helix interactions occur. Based on the assumption that each base pair replacement affects the complementarities of shape between the two double helices and thus affects the stability of the modified AGPM, each dynamic provides for a quantitative evaluation of the stability corresponding to a particular base pair replacement. For each construct, 3 to 10 simulations have been made as a probabilistic process which told us how beneficial or detrimental a base pair replacement was for the AGPM structure. As each simulation is defined by four atom-atom contacts, we defined the stability of a specific construct as the sum of all four partial stabilities.

2.1.9. Analysis module

This module helps to trace the history of the simulated AGPMs motion during dynamics through the trajectory function. Data collected from the appropriated file enable to plot profiles of the stability of the AGPMs over one nanosecond. From those profiles

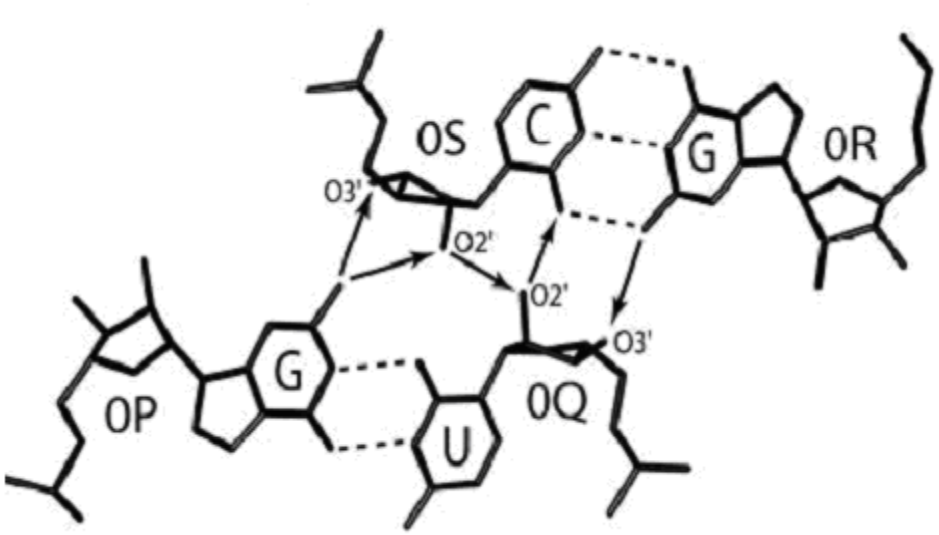
and the different frames witness of the internal rearrangement during dynamic simulations, one can approximate the period of time a modeled AGPM is stable for a quantitative evaluation of the stability, give more insight into the structural constraints guiding the formation of such modified AGPM and how a specific base pair replacement at one level can affect its overall stability.

3. Results

3.1. Background: the general description of AGPM

As mentioned above, AGPM consists of two double helices arranged such that a sugar-phosphate backbone of one helix is packed along the minor groove of the other helix and *vice versa* (Figure 1)[12]. In each helix, the strand that interacts with the minor groove of the opposite helix is positioned closely to the center of the arrangement and is thus called internal. The other strand in each helix stays at the periphery of the arrangement and is called external. The internal strands of both helices go in the same direction, opposite to that of the external strands. The conventional representation of the motif given in Figure 1 demonstrates the existence of symmetry between the two helices[12]. Indeed, within AGPM, the position of each helix can be roughly determined based on the position of the other helix through the rotation for 180º around the common symmetry axis.

At the center of the contact area, there are two base pairs, called central, that pack with each other most closely. Despite the general symmetry of AGPM, the packing of the central base pairs is essentially asymmetric. In most identified cases, one of these base pairs is Watson-Crick (WC), while the other one is GU. In the GU base pair, G and U stay, respectively, in the external and internal strand[12]. The arrangement of these base pairs shown in Figure 6 provides for a close contact between the helices and allows the formation of a network of several inter-helix hydrogen bonds. Such GU-WC arrangement at the center of the contact area is found in most known cases of AGPM. Moreover, as pointed out by Mokdad *et al.*[15], the GU base pairs involved in AGPM are among the most conserved GU base pairs in rRNA. Based on these findings, the coexistence of a WC and GU base pair in the middle of a helix-helix contact has been considered as a signature of AGPM that could facilitate the identification of new cases of the motif. In particular, this pattern enabled us to identify the two AGPMs formed by the P- and E-site tRNAs and 23S rRNA[12].



<u>Figure 6</u>: Hydrogen bonds network at the central part of the AGPM arrangement. This network is possible because of the presence of the GU base pair which allows the displacement of G and U toward the minor and major groove respectively.

Despite the almost universal presence of the GU-WC combination, some cases of AGPM do not follow this pattern, which indicates the existence of other aspects within the structure of the motif that are important for its integrity. The presence of such aspects can also be deduced from the fact that in AGPM, the area in each helix that interacts with the other helix is not limited to one base pair but instead, spreads over several consecutive base pairs. These additional contacts largely represent interactions between the backbones of both helices. Given that the interacting surfaces of the two helices are not flat, each next base pair participates in inter-helix contacts differently from its neighbors within the helix. The latter makes analysis of the aspects responsible for the integrity and stability of AGPM a rather complex exercise and has determined our choice of the approach to be used for this purpose. In this paper, we perform a systematic analysis of the inter-helix interactions in order to understand the role of each pixel of these interactions in the integrity and stability of AGPM. This analysis is based on the MD simulations of thirty-one specially modeled constructs of AGPM harboring a variety of identities for each base pair involved in inter-helix interactions. The results of the simulations are discussed in the context of the statistical data on the occurrence of particular nucleotides in different positions of AGPM. This analysis allowed us to rationalize the identity preferences for all base pairs involved in interaction with the opposite helix. These preferences can now be used for evaluation of stability of particular AGPMs and for further identification of AGPM in yet unresolved RNA structures.

3.2. Nomenclature of different elements of AGPM

To facilitate the discussion of the inter-helix interactions within AGPM, we will use the following nomenclature. For the four strands of these helices, capital letters P, Q, R and S are assigned as shown in Figure 3 (upper left corner). One helix is formed by strands P and Q, while the other one is formed by strands R and S. Strands P and R are external, while strands Q and S are internal. The helix containing GU as the central base pair in the *Escherichia coli* rRNA is composed of strands P and Q and is called the GU-helix (Figure 3). The opposite helix is called the WC-helix. For each base pair of each helix, a number is assigned, so that the central base pairs carry number zero, and the

positive propagation of the numbering corresponds to the 5'→3' direction of the internal chains. The two base pairs of the opposite helices carrying the same number form a layer. In the identity of a base pair, the first and last letter will correspond to the external and internal nucleotide position.

3.3. Collection of the set of AGPMs

The nucleotide sequences of all identified motifs are shown in Figure 3. The original set of 12 motifs[12] presented in Figure 3(a-l) exists in all ribosome structures. In addition, Mokdad *et al.*[15] showed that motif L1864 (Figure 3m) exists in the bacterial 50S subunits of *Deinococcus radiodurans* (pdb entry codes 1kpj-1lnr[18]) and *E. coli* (pdb entry codes 2aw4-2awb[19]) but not in the archaeal 50S subunit of *Haloarcula marismortui* (pdb entry codes 1jj2-1s72[20]), where it was replaced by an A-minor interaction[21]. For this work, we undertook an additional analysis of all ribosome-related crystal structures and found one more motif, S911 (Figure 3n), which is formed between helices h27 and h44 of the 16S rRNA as a result of a conformational rearrangement in the 30S subunit caused by its association with the initiation factor-1 (IF1) (pdb entry code 1hr0[22]).

Two more motifs, named L1923-P and L1851-E, are formed between the D-stem of the P-site tRNA and helix 69 of the 23S rRNA (Figure 3p) as well as between the acceptor stem of the E-site tRNA and helix 68 of the 23S rRNA (Figure 3q)[12]. These motifs exist in two low-resolution structures of the *Thermus thermophilus* 70S ribosome (pdb entry codes 1gix-1giy[23] and 2ow8-1vsa[24], respectively). The high resolution ribosome structure from the same organism (pdb entry codes 2j00-2j01 and 2j02-2j03[25]) confirmed the presence of motif L1923-P, while the E-site tRNA in this structure was positioned differently. It is also observable in the X-ray structure of the 70S ribosome determined at 3.3 Å resolution of the same organism (pdb entry code 3uzn[38]). The structures of the yeast ribosome [39] and of both ribosomal subunits of *Tetrahymena thermophila* [40, 41] confirmed the presence of all AGPMs existing in the bacterial ribosomes except L1864 and S911. Finally, a systematic analysis of all RNA-containing structures in the PDB database[26] revealed a case of AGPM, named HH, which was formed by two hammerhead ribozyme molecules within the same asymmetric unit of the crystal (Figure 3o) (pdb

entry code 1hmh[27]). So far, this case of AGPM has been the only one identified outside the ribosome.

The data on nucleotide sequences of AGPMs were collected from the set of available nucleotide sequences of prokaryotic rRNA, which included 12 107 bacterial and 590 archaeal sequences of 16S rRNA as well as 399 bacterial and 37 archaeal sequences of 23S rRNA[16]. For the tRNA sequence analysis, the database containing 819 bacterial and 220 archaeal tRNA sequences was used[17].

3.4. Principles of helix packing within AGPM: triangle-over-triangle model

Prior to this paper, several studies of AGPM were reported and focused exclusively on the role of the central base pairs and their variations on the stability of the motif[12-15, 37]. Based on the analysis of all cases of AGPM in the available ribosome structures, it was already known that two cases did not follow the standard GU-WC pattern. In particular, motifs S549 in all available 30S subunit structures[19,25,29,30] and L2291 in the 50S subunit of *H. marismortui*[20] have two GC base pairs at the 0-level[12,15]. A similar situation also occurs in motif HH, which forms between two hammerhead ribozyme molecules[27] and where two WC base pairs pack together at the 0-level (Figure 3o). Moreover, further *in vivo* studies of other cases of AGPM showed that the central base pairs can display an array of nucleotide combinations, while still being able to provide for functional ribosomes[14, 37]. These observations downplayed the importance of the central base pairs for the integrity of AGPM and prompted us to consider the role of other inter-helix contacts in the formation of AGPM. Regardless of how effective the interaction between the central base pairs is, it can occur only once per motif. Indeed, because of the spiral character of both helices, the juxtaposition of the base pairs at each level is different. Only at the 0-level, the arrangement of the base pairs is such as it is shown in Figure 6, while even at the neighboring +1 and -1 layers, it is so different that it can no longer be described as a packing of a backbone of one helix in a groove of the other helix. It does not mean, however, that outside the 0-level the two helices do not interact. On the contrary, analysis of the available AGPM conformations shows that the

inter-helix contacts spread over four layers in each helix between -2 and +1. While at the 0-level, these contacts include three out of four bases (bases of nucleotides 0P, 0R and 0S, see Figure 6), at the other levels contacts are mainly formed by elements of the backbones. Most of the backbone contacts are formed by riboses and are thus mainly hydrophobic.

Analysis of the AGPM structure shows that the whole contact area outside the 0-base pairs can be divided in three zones, depending on the particular strands involved in the inter-helix contacts. The first zone, named QS, corresponds to the interaction between chains Q and S. The Q- and S-moieties of this zone are mainly formed by the riboses of nucleotides +1Q and +1S, but also include some atoms of the neighboring nucleotides 0Q and 0S (Figure 7 and Table I. The second zone QR is formed by chains Q and R. It mainly consists of the contact between riboses -1Q and -2R, but also includes some atoms of 0Q and -1R. The third zone PS is symmetrical to zone QR with strands P and S being equivalent to R and Q, respectively. A complete list of the atoms participating in the formation of the three contact zones is given in Table I. On the surface of each helix, these zones form a triangle with the vertices positioned at the riboses of the internal +1- and -1-nucleotides as well as of the external -2-nucleotide (Figure 7b). The interaction of the two helices can thus be seen as superposition of the triangle in one helix on the equivalent triangle in the other helix (shown by orange arrows in Figure 7b). Analysis of the known AGPMs shows that the three contact zones are preserved in all cases regardless of the presence of other features and are thus considered important for the integrity of the motif. The formation of the contacts within the three contact zones depends on the particular positions of the riboses involved and is thus expected to be sensitive to the structures of the corresponding base pairs. In the following sections, we will show how the system of backbone-backbone contacts shapes AGPM and how it restricts the identities of the essential base pairs.

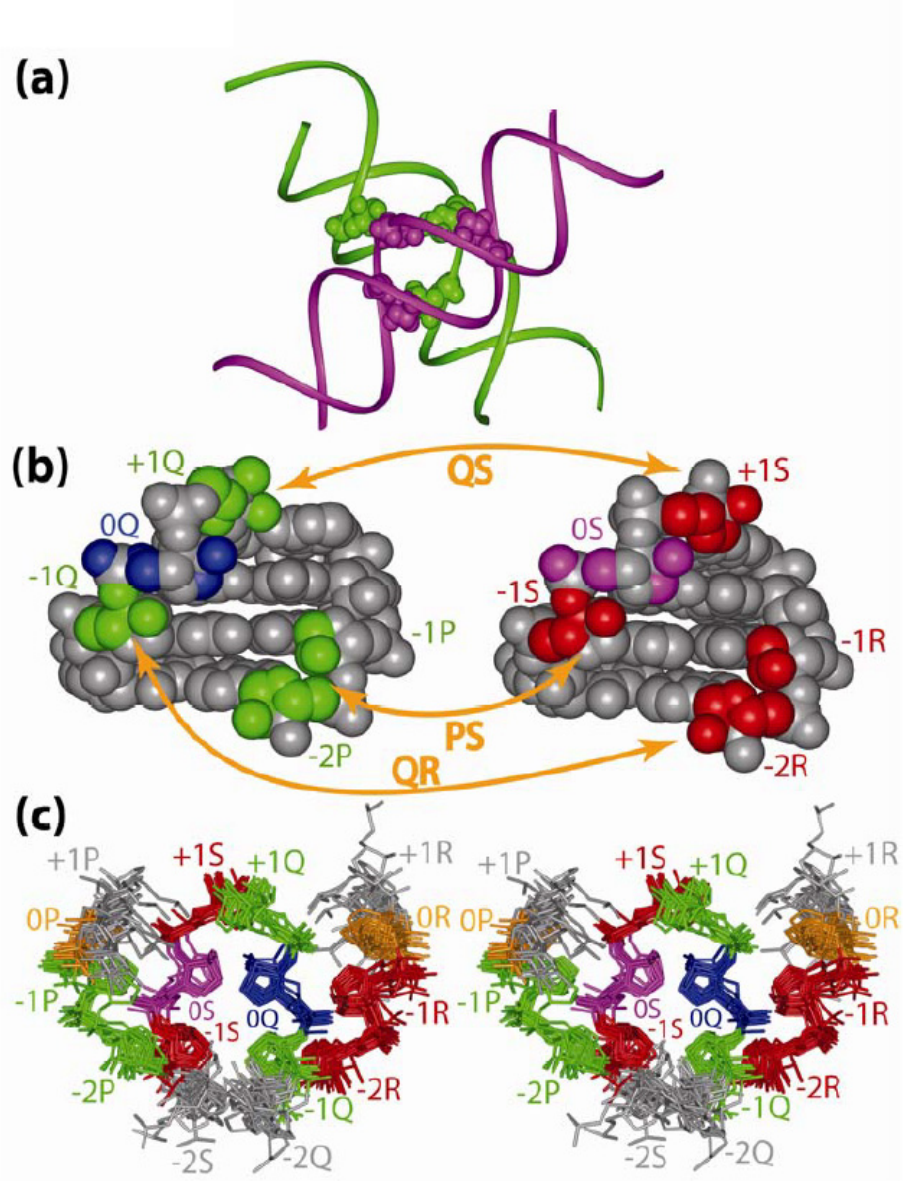


Figure 7: The arrangement of the three contact zones within AGPM.
(a): Schematic representation of the three contact zones when both helices interact together (b): Location of the atoms forming the three contact zones in the GU-helix (left) and in the WC-helix (right). The contact zones in the GU-helix are green; they which encompass nucleotides +1Q, -1Q, -1P and -2P. The symmetrical contact zones in the WC-helix are red; they encompass nucleotides +1S, -1S, -1R and -2R. Also, 0Q is blue; 0S is magenta. The external +1-nucleotides, +1P and +1Q as well as internal -2-nucleotides, -2Q and -2S are grey. Grey nucleotides are not involved in inter-helix contacts, which can explain the freedom allowed to the positions of their nucleotides

(Figure 7c). Orange two-headed arrows indicate the nucleotides forming the three contact zones when both helices interact together. The complete list of the atoms involved in the contact zones is given in Table I.
(c): Stereo view of all known AGPM structures (Figure 3a-p) superposed based on the positions of the C4′ atoms (r.m.s.d. = 0.87 Å). The high resolution structure of the *T. thermophilus* ribosome (pdb entry codes 2j00-2j01 and 2j02-2j03[24]) allowed us to include motif L1923-P (Figure 3p) in this superposition, while the E-site tRNA was positioned differently. For clarity, the bases are not shown. For the atoms of the contact zones, the same colors are used as in panel (b). The 0P and 0R are colored in orange.

Table I : Structure of the contact zones

Zones	Nucleotides	Atoms
QS	+1Q	O4′-C4′-C5′-O2′-O3′
	0Q	C1′-O2′
	+1S	O4′-C4′-C1′-O2′-C5′
	0S	C1′-O2′
QR	-1Q	C4′-C3′-O2′-O3′-C5′
	0Q	O1P-O5′-C5′
	-2R	C1′-O4′-C4′-C2′-O2′-C5′
	-1R	C1′-C2′-O2′
PS	-2P	C1′-O4′-C4′-C2′-O2′-C5′
	-1P	C1′-C2′-O2′
	-1S	C4′-C3′-O2′-O3′-C5′
	0S	O1P-O5′-C5′

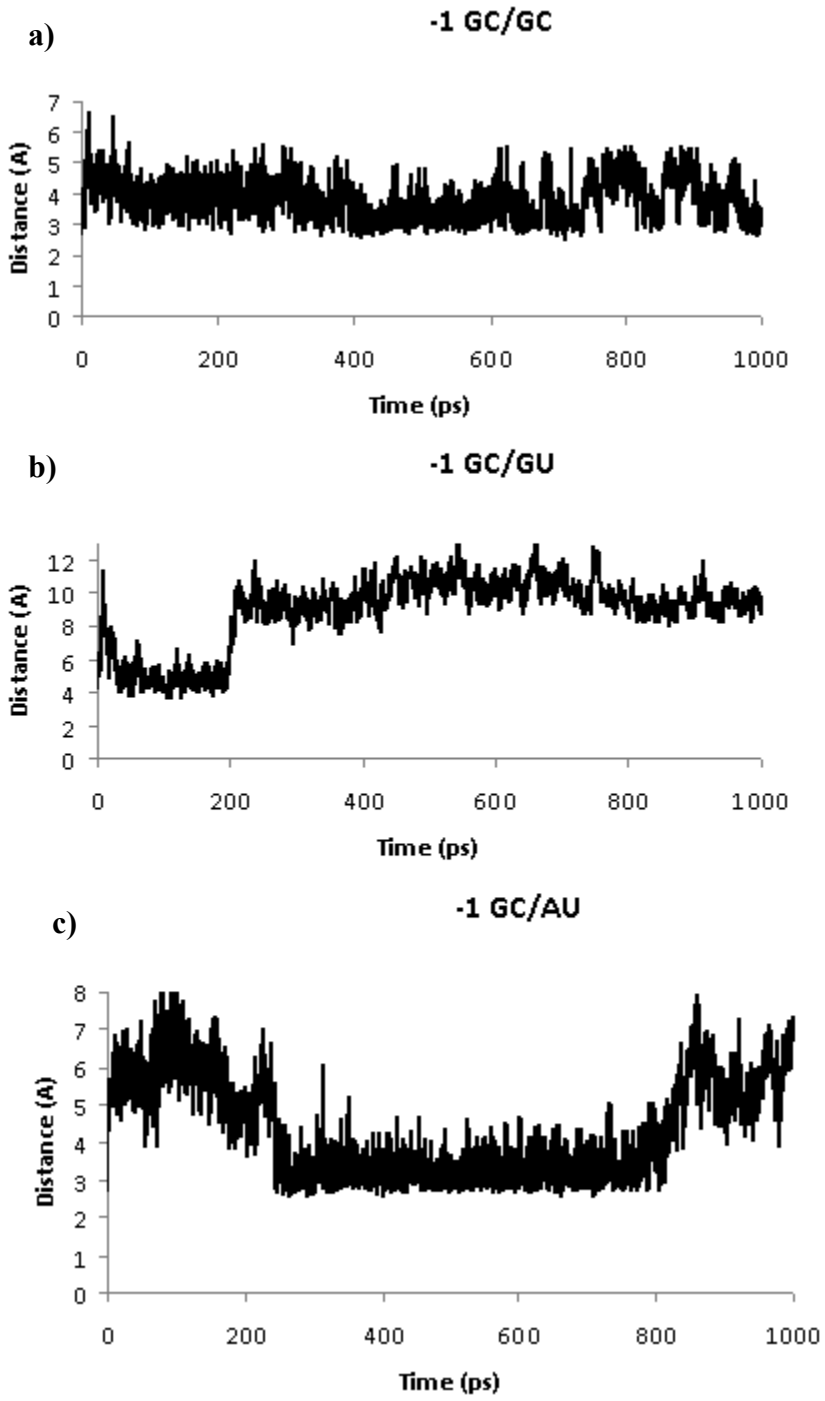
A non-hydrogen atom was considered a part of a contact zone if its distance from the closest non-hydrogen atom of the opposite helix was within 4.2 Å. See also Figure 7.

3.5. Molecular dynamics of specially modeled AGPM constructs

To check how different modifications of each element of the AGPM structure affect the integrity and stability of the whole arrangement, we made different *in silico* constructs and submitted them to MD simulations. Most constructs were made based on the design shown in Figure 4a, which is henceforth named the Universal Reference Structure (URS). It consisted of two closely packed double-helices. Each helix was composed of five base pairs corresponding to layers from -3 to +1 and was capped at both ends by tetraloops GAGA. For URS, the identities of the base pairs were chosen based on the nucleotide sequence of motif L657 in the *E. coli* ribosome (pdb entry code 2aw4[19]). The presence of the two tetraloops made each helix a cyclic structure. The modeling procedure and the particular conditions of the MD simulations are described in the Methods. Figure 4c provides the stereo view of a construct based on the design shown in Figure 4a. For those constructs containing an A-G shared base pair at the +1 level, we used another design shown in Figure 4b.

The effect of base pair replacements on the stability of AGPM was assessed through monitoring the distances within four pairs of atoms. The integrity of the QS contact zone was followed by measuring the distance between the C4′ atoms of nucleotides +1Q and +1S and between the O2′ atoms of nucleotides 0Q and 0S (in the standard AGPM structure these O2′ atoms are connected by a hydrogen bond). In the second zone QR, the contact between riboses of nucleotides -1Q and -2R was monitored following the distance between atoms C4′ of -1Q and O2′ of -2R. Finally, the integrity of the third zone PS was monitored following the distance between the symmetrically positioned atoms C4′ and O2′ of nucleotides -1S and -2P, respectively. At each moment of a simulation, each of the four contacts was considered as existing if the distance between the two assigned atoms did not exceed 4.0 Å (only for the contact between the O2′ atoms of nucleotides 0Q and 0S) or 4.5 Å (for the other three contacts). During a particular MD simulation, each of the four measured contacts could follow one of three patterns of either being stable or undergoing reversible or irreversible breakages. Figure 8 an example of such profiles in

different constructs and when monitoring different distances within the AGPM contact zones.



<u>Figure 8</u>: Example of profiles from different MD simulations.
(a, b) represent the profiles of a base pair replacement at the -1-level leading to -1-GC-GU complex. The profile showed in a) measures the period of time the contact between the O2' atoms of both nucleotides 0Q and 0S last throughout the 1ns the simulation lasts.

This profile is considered as stable while the profile (b), measuring the period of time the contact between the C4’ atoms of both +1Q and +1S breaks irreversibly after 207 ps. The profile in c) shows that the contact between the C4’ atom of -1Q and O2’ atom of -2R breaks instantly after the beginning of the simulation. However, this breakage is reversible and the complex demonstrates stability between 168 ps and 840 ps. In this construct, the base pair GC-AU has been introduced at the -1-level.

The stability of each of the four contacts was evaluated as the fraction of the whole simulation time during which the assigned contact kept its integrity. The stability of the whole construct was calculated as the sum of the stabilities of all four contacts, which thus varied between 0 and 4 for each simulation. For each construct, the final stability was averaged over several simulations. Table II provides the list of all constructs for which MD simulations were performed, the final stability of each construct, as well as the details of all simulations. For URS the stability was 2.4, which was higher than for almost all other constructs. To make the analysis more systematic, the overall energy of most constructs, including URS, was calculated using a specifically modified formula. As discussed below, when base pair [-1P, -1Q] is GC or CG, the amino group of the guanosine forms an H-bond with atom O4' of the 0S nucleotide. In most calculations, the energy of this hydrogen bond was excluded from the formula of the overall energy and was taken into account only when the role of this H-bond was specifically analyzed.

Table II: Summary of the molecular dynamics simulations at -1-, -2-, 0- and +1-level of modeled AGPM used in our analysis

#	Level of interaction	Base pairs combinations	# of simulations[a]	Average stability
1	-1 (-)-H-bond	**CG-CG**	**6**	**2.4**
2		CG-GC	5	1.5
3		GC-CG	5	1.0
4		GC-GC	3	0.3
5		CG-AU	5	1.8
6		CG-UA	5	1.4
7		AU-CG	5	1.1
8		UA-CG	4	1.0
9		CG-GU	4	1.7
10		CG-UG	4	1.3
11		GU-CG	4	0.7
12		UG-CG	5	1.1
13		GC-GU	7	1.4
14		GC-UG	6	1.1
15	-1 (+)-H-bond	CG-CG	5	2.6
16		GC-CG	5	2.4
17	-2	**GC-CG**	**6**	**2.4**
18		GC-UG	6	2.1
19		GC-GU	8	1.5
20		GU-GU	6	1.1
21	0	**GU-GC**	**6**	**2.4**
22		GU-CG	3	1.9
23		GU-AU	3	1.1
24		GU-UA	3	1.8
25		GU-GU	3	1.9
26		GU-UG	3	1.4
27		GC-GC	3	0.9
28		CG-CG	3	1.1
29		UG-UG	3	1.0
30	+1	**GC-GU**	**6**	**2.4**
31		GC-UG	7	0.4
32		GC-GC	9	1.3
33		GU-GC	4	0.6
34		GC-AG	7	2.4

This table shows the results of MD simulations for all AGPM constructs studied in this paper. Each construct is characterized by the particular identity of the base pairs (the third column) located at the particular level (shown in the second column). The identities of all

other base pairs were as in URS, except construct 34, which was made based on the alternative design shown in Figure 4.

3.6. The inter-helix interactions in contact zones QR and PS

3.6.1. The central role of the -1-base pairs

As one can judge from Figure 7 and Table I, among all base pairs participating in the formation of AGPM, only in the -1-base pairs both nucleotides are involved in inter-helix backbone-backbone contacts. In particular, the internal nucleotide -1Q forms a major part of the Q-moiety of zone QR, while its external base pair partner -1P participates in the formation of the P-moiety of zone PS. The same is true for nucleotides -1S and -1R with respect to zones PS and QR. The necessity of the proper fitting for all four -1-nucleotides to the inter-helix contacts within zones QR and PS will impose strong restrictions on the structure of the -1-base pairs. As shown in Figure 3, in all presented examples of AGPM, both -1-base pairs are always WC. Moreover, analysis of the available nucleotide sequences of AGPM shows that in different organisms, the WC identities of both -1-base pairs are maintained in all motifs at the average level of 98% (Table III)[16]. The superposition (Figure 9) of the available AGPM conformations shows that in all of them the two -1-base pairs occupy the same positions, which are symmetrical with respect to the common symmetry axis. Given that a replacement of any of the two -1-base pairs by a non-WC dinucleotide combination will unavoidably affect both contact zones QR and PS, we can suggest that the maintenance of the WC identity of both -1-base pairs is important for the AGPM stability.

Table III: Occurrence of different combinations of the -1-base pairs in AGPMs existing in prokaryotic rRNA.

-1-base pairs	Number of sequences	%
GC/CG-GC/CG	38 250	70.4
GC/CG-AU/UA	14 176	26.1
AU/UA-GC/CG	438	0.8
AU/UA-AU/UA	380	0.7
Total WC-WC	53 244	98.0
Total	54 315	100

For Table III-Table VI, the data were obtained from the available rRNA alignments[16]. In all these tables, "Total" refers to the total number of nucleotide sequences for which the identities of all nucleotides in question are known.

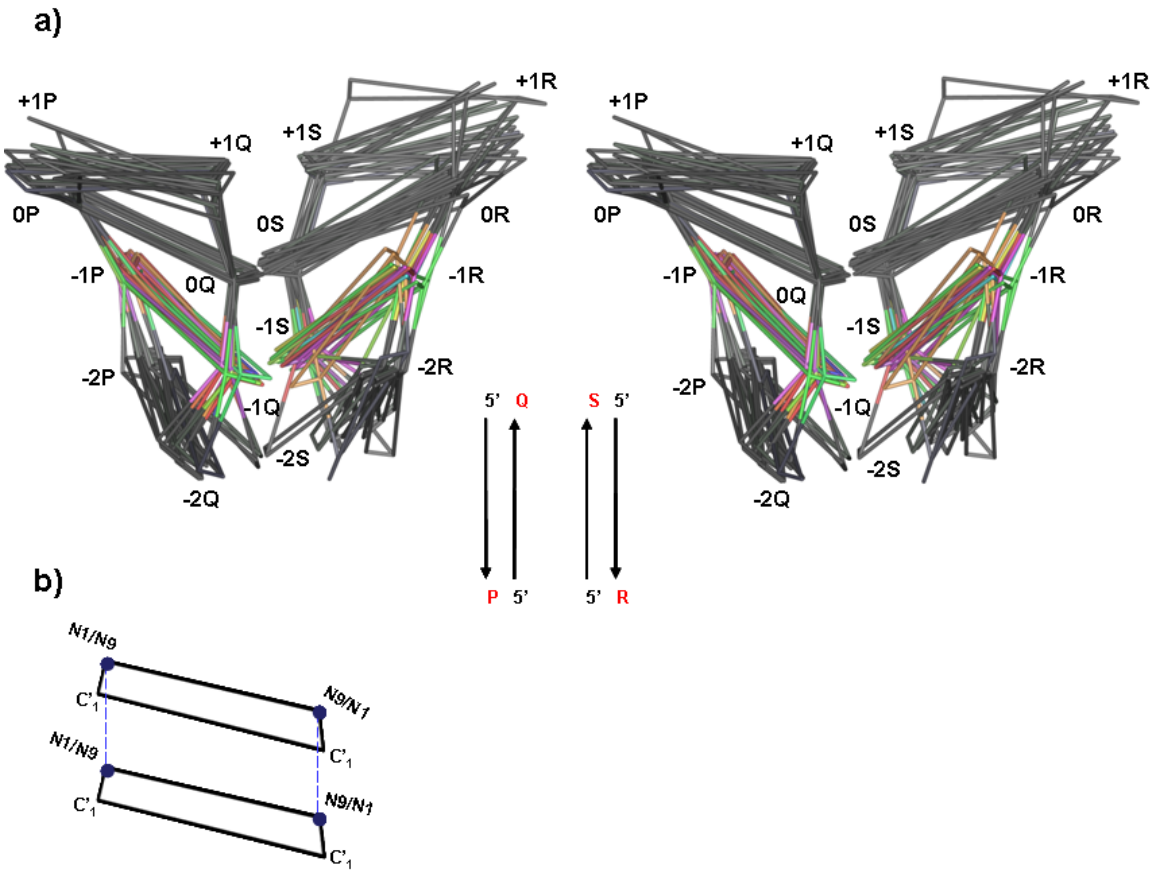


Figure 9: Stereo view of the superposition of the available AGPMs

a) The superposition excludes the motif (2o) from the structure of the hammerhead ribozyme (pdb entry code 1hmh[27]) and motifs (2p, q) from the structures of the *T. thermophilus* ribosome (pdb entry codes 2j00-2j01[25] and 1gix-1giy[23]). The -1-base pairs from each helix are presented in color. The superposition at levels -1 and 0 is optimal while at -2 and +1-levels it is notably poor. The set of four arrows in the middle represents the strands P, Q, R and S (shown in red) and the direction 5' → 3'. The trapeziums in the superposition represent the base pairs as shown in b). The base pairs are linked via their N1/N9 atoms (dotted line).

To test whether indeed both -1-base pairs must be WC, we made six constructs in which, starting from URS, we introduced the GU or UG base pair in either position [-1P;-1Q] or [-1R;-1S]. Compared to other non-WC base pairs, combinations GU and UG provided relatively minor deviations from the WC geometry and thus were expected to fit to these positions better than any other non-WC combination. As one can see in Table II, a replacement of any of the two -1-base pairs by a GU or UG combination resulted in a notable drop of the AGPM stability. Interestingly, the damaging effect of the presence of a GU/UG base pair at the -1-level was notably stronger when it was located in the GU-helix than in the WC-helix. We attribute this effect to the steric intolerance between the two neighboring non-WC base pairs occupying positions [-1P; -1Q] and [0P; 0Q]. Below we will see that on top of being WC, -1-base pairs have additional asymmetric restrictions of their identities caused by the interaction with the 0-base pairs.

3.6.2. The -2-base pairs

Unlike the -1-base pairs, in which all internal and external nucleotides participate in the inter-helix contacts, at the -2-level only the external nucleotides -2P and -2R are involved in the interaction with the opposite helix (Figure 7). It is not surprising, therefore, that contrary to the -1-base pairs, the -2-base pairs harbor a variety of different structural forms. In particular, among the 17 cases of AGPM shown in Figure 3, there are 34 potential -2-dinucleotide combinations, which in the real structures correspond to 16 WC base pairs, 8 non-WC base pairs, while in 10 remaining cases the base pair does not exist at all.

The 16 WC -2-base pairs are distributed between 10 AGPMs: in four motifs S296, S757, L554 and L2698 both -2-base pairs are WC, and in six more motifs only one -2-base pair is WC, while the other base pair is either non-WC or inexistent. When both -2-base pairs are WC, the two -1- and two -2-base pairs form a four-base-pair arrangement seen in Figure 10a, b. All cases of this arrangement are superposable with r.m.s.d. 0.56 Å. Also, each arrangement is symmetrical with respect to the symmetry axis of the whole AGPM. At the center of contact zone QR, the five-member rings of two riboses -1Q and -2R are closely packed with each other. In total, between the Q and R strands there are

about 15 atom-atom van der Waals contacts in which two non-hydrogen atoms are positioned within 4.2 Å of each other. The same interactions occur in the PS contact zone between nucleotides -2P and -1S.

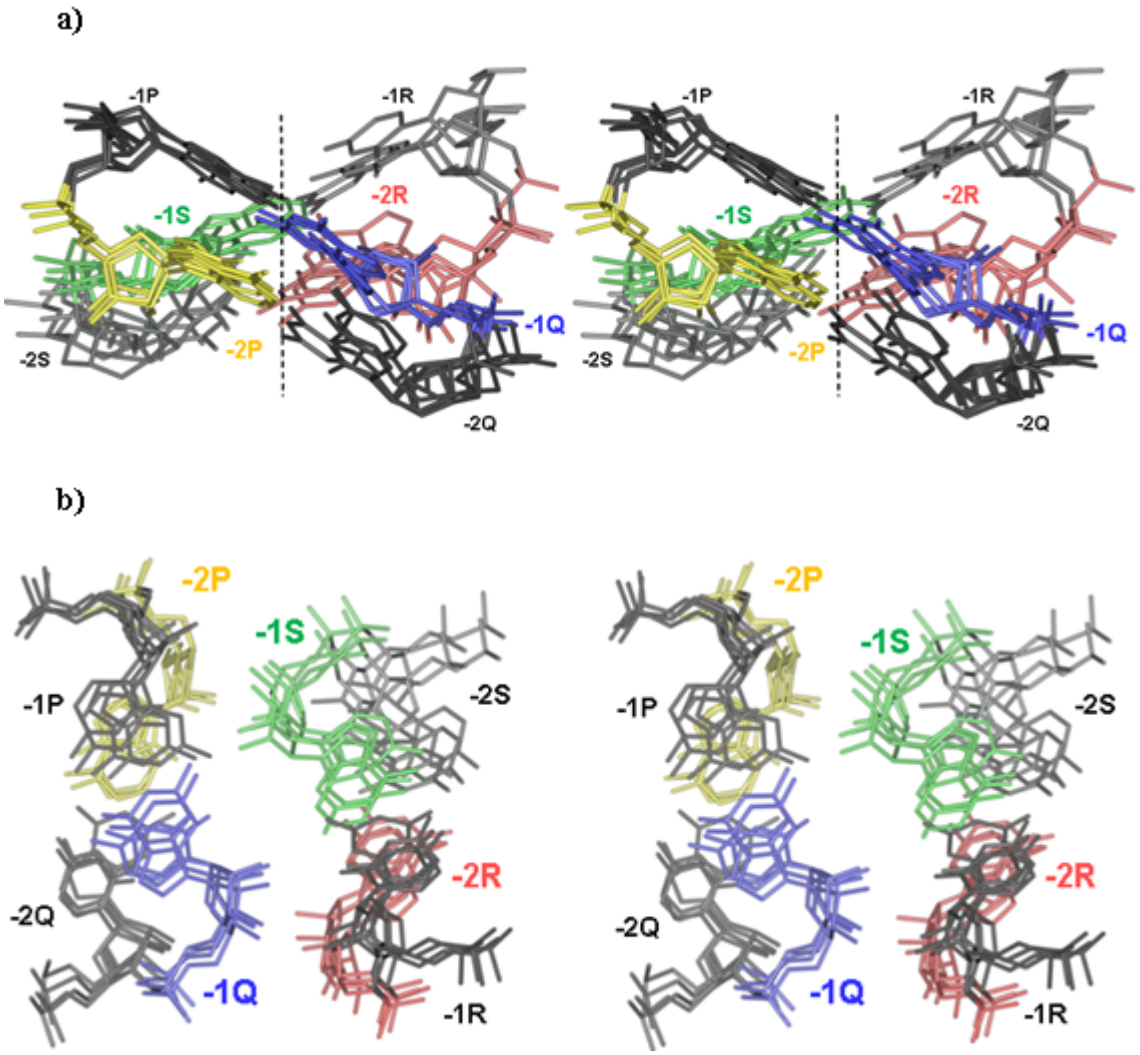
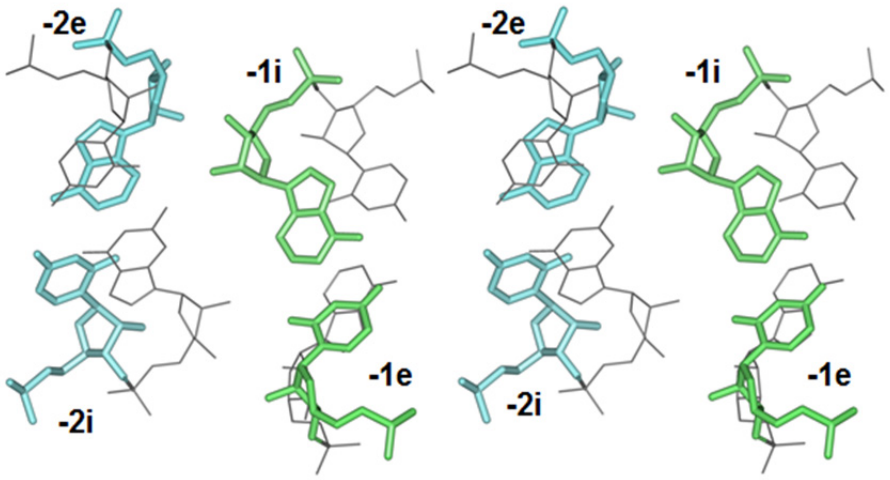


Figure 10: Ribose-ribose interactions within contact zones PS and QR.
(a-b): Two different stereo views of the same superposition of motifs S296, S757, L554 and L2698, in which both -2-base pairs are WC (r.m.s.d. = 0.56 Å). Nucleotides -2P and -1S form a major part of contact zone PS, while nucleotides -1Q and -2R form a major part of zone QR. Nucleotide -1P and -1Q are respectively in yellow and blue. Nucleotides -1R and -1S are respectively in pink and green.

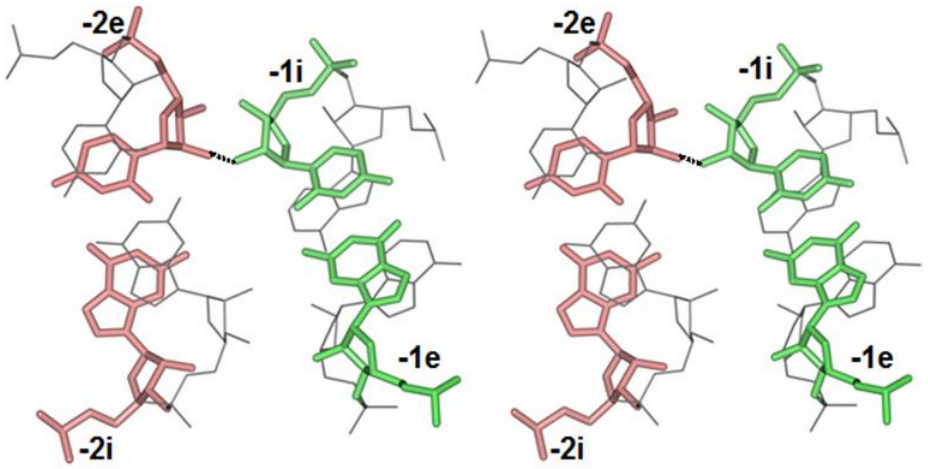
The superposition of the motifs containing only one -2-base pair WC with the motifs having two such base pairs showed that in all cases, a WC -2-base pair stays at the same position with respect to the neighboring -1-base pair regardless of the structure or even of the existence of the -2-base pair in the opposite helix. This fact is taken as an indication that at the -1- and -2-levels the conformations of both helices are optimal and do not change upon the formation of the AGPM.

For the other helix, harboring at the -2-level a non-WC combination, there are two possibilities of having a non-WC base pair or having no base pair. In both cases, the external nucleotide (-2P or -2R) maintains the same position, which, however, is different from that within a WC base pair. Compared to the latter, this external nucleotide is over-twisted by 10˚-15˚, so that its WC edge becomes more open to the major groove (Figure 11). Such over-twist allows the formation of an additional hydrogen bond between the O2´ atoms of the two nucleotides (-2P and -1S) or (-2R and -1Q). Although this hydrogen bond does not exist when the -2-base pair is WC, it is found in most cases when this base pair is either non-WC or non-existent.

a)



b)



c)

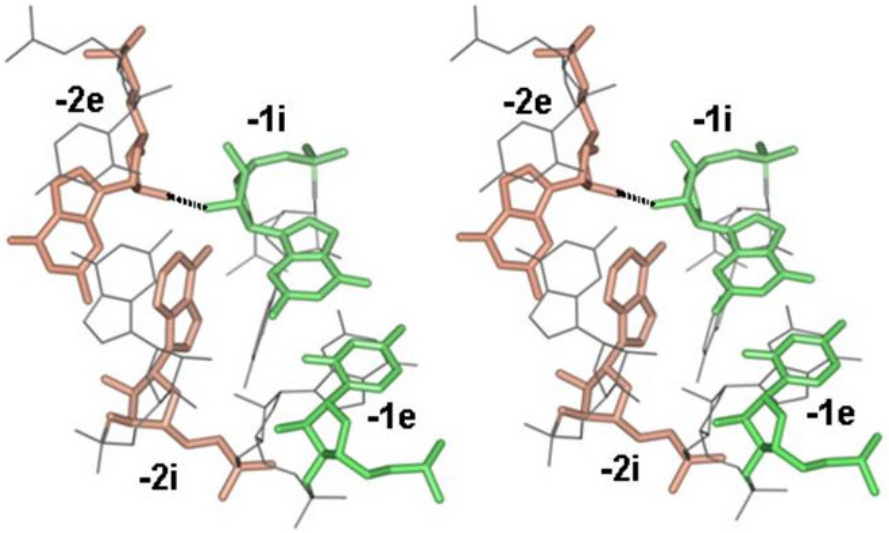


Figure 11: Stereo view of the interaction between the riboses of the external -2-nucleotide (-2e) and of the internal -1-nucleotide (-1i) for different -2-base pairs.
For the -2-base pair, AU is blue (panel a), UG is pink (panel b), and GA is orange (panel c). The -1-internal and external nucleotides are presented in green. In all -2-base pairs, the ribose of the external nucleotide -2*e* is positioned closely enough to the ribose of the internal nucleotide -1*i* of the -1-base pair to form a ribose-ribose contact. Compared to the -2-base pair AU, in both base pairs UG and GA, the external nucleotide -2*e* is slightly over-twisted, which allows the formation of a hydrogen bond between the O2′ atoms of the two riboses (dashed line). The internal nucleotide -2*i* of the -2-base pair shows a strong variability in its position, which is not surprising given that this nucleotide is not involved in inter-helix interactions.

When the external -2-nucleotide (-2P or -2R) is a part of a non-WC base pair, the structure of the pair cooperates with the arrangement described above: in all such cases, the external and internal nucleotides are displaced towards the major and minor grooves, respectively (Figure 11). This displacement allows the over-twist of the external nucleotide and leads to the formation of the above-mentioned inter-helix hydrogen bond. The necessity for such displacement limits the set of the acceptable non-WC dinucleotide combinations to UG (3 times), GA (3), AA (1) and UU (1), all of which are able to support the opening of the base of the external -2-nucleotide towards the major groove. These dinucleotide combinations are also predominant non-WC ones observed at the -2 level in the available sequences of prokaryotic rRNA of AGPM (Table IV)[16]. In total, for those -2-dinucleotide combinations that correspond to a base pair in the ribosome tertiary structure, WC, GA, UG, AA and UU combinations amount to 99% of all cases (Table IV). More specifically, combination GA is almost 600 times more frequent than AG, while combination UG is almost six times more frequent than GU. Such assumption is corroborated with the known AGPMs presented in Figure 3.

Table IV: Occurrence of different identities of the -2-base pairs in AGPMs existing in prokaryotic rRNA.

-2-base pairs	Number of base pairs	% [b]
WC and closely related conformations [a]		
WC	54 815	82.1
UG	1669	2.5
GU	293	0.4
UU	286	0.4
GA	42	0.06
AA	19	0.03
AG	14	0.02
Sheared conformations [a]		
GA	8 507	12.7
AA	856	1.3
UA	320	0.5
UG	27	0.04
AG	1	0.001
Sub-total	66 849	100
Others [c]	42 031	
Total [d]	108 880	

a: The base pair conformations were deduced from the available crystal structures of AGPM.

b: Percentages were calculated with respect to the total number of the -2-dinucleotide combinations that form a base pair.

c: This number relates to those -2-dinucleotide combinations in motifs S62, S549, S911, L639, L657, L839 and L2847 that do not form a base pair.

d: The total number of -2-base pairs is based on 54440 sequences for which the identities of all four nucleotides are known. See also the footnote to Table III.

To check the importance of the correct ribose-ribose interactions between the -1 and -2 base pairs for stability of the whole AGPM, we performed molecular dynamic simulation of several constructs having alternative nucleotide combinations at the -2-level. Starting from URS (Figure 4a, c), we replaced base pair [-2R;-2S] by either the UG or GU base pairs. The resulting constructs were submitted to several MD simulations. For the construct containing combination UG, the stability was 2.1, i.e. slightly lower than of URS. Most of the simulation time, nucleotides -2R and -2S formed the normal UG base pair. In the case of the GU combination, the two nucleotides did not form a base pair, and the stability of the arrangement was notably lower: S=1.5. When both helices contained the GU combinations at the -2-level, the stability of the construct was even lower: S=1.1. These results corroborated our suggestion that base pair UG, which displaces the external uridine towards the major groove, provides for a more stable AGPM structure than GU, which displaces the external guanosine towards the minor groove, thus causing its collision with the opposite helix.

3.7. The inter-helix interactions in contact zone QS

3.7.1. The 0-base pairs

Although the presence at the 0-level of the asymmetric base pair combination GU-WC is considered as a signature of AGPM, the necessity of such asymmetry in view of the rather symmetric arrangement of two base pairs at the -1-level has never been properly understood. To make a step towards this understanding, we performed a simple *in silico* experiment consisted in the extension of both double helices from layer -1 to layers 0 and +1 in the standard A-RNA conformation. As one can see in Figure 12, after such extension, the two helices become colliding with each other at both levels 0 and +1 with the degree of mutual penetration of the two helices of about 1.0 Å (level 0) and 1.5 Å (level +1) (Figure 12). At both levels, the collision mainly affects the internal nucleotides. However, while at the 0-level the inter-helix contacts include not only the riboses but also the bases of the internal nucleotides 0Q and 0S, at the +1-level the inter-

helix contacts are made only by the riboses of +1Q and +1S. This difference makes the strategies for resolving the collisions at levels 0 and +1 different as well.

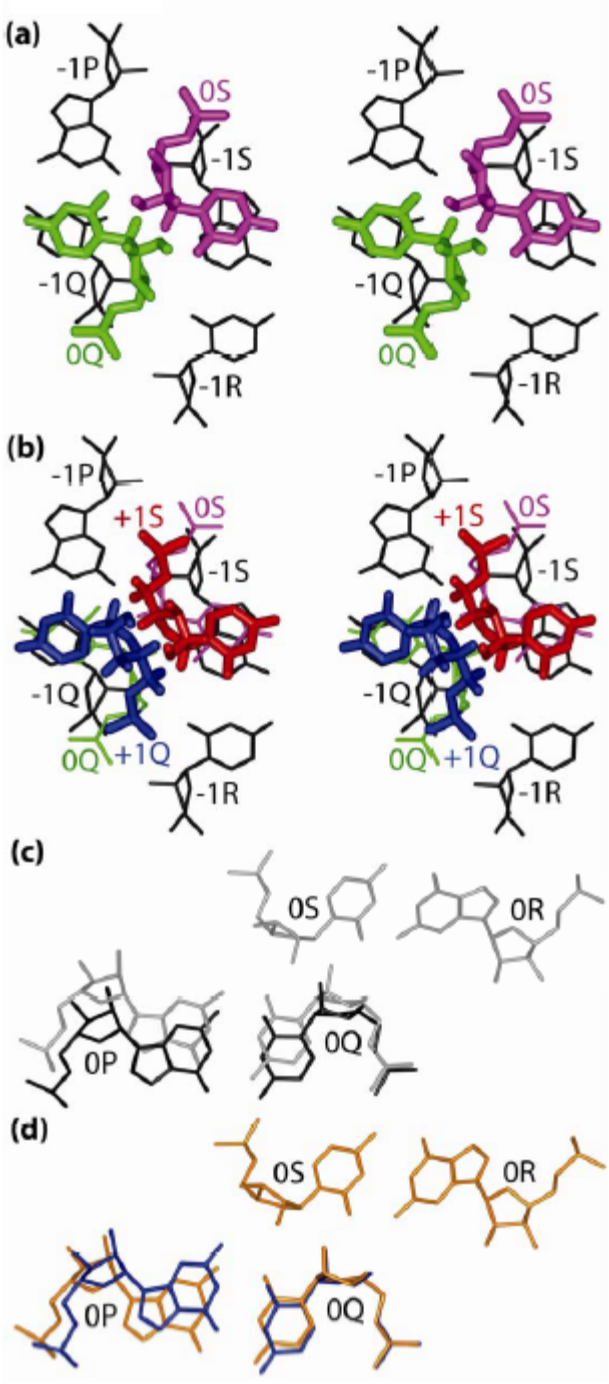


Figure 12: The potential collision of the internal nucleotides at the 0- and +1-levels and its consequences for the structure and position of base pair [0P; 0Q]. (a,b): The extension of both double helices from the -1-level (black) to levels 0 (a) and +1 (b) leads to the collision of the internal nucleotides (green and magenta in (a), blue and red in (b)). (c): the displacement of the WC base pair [0P;0Q] to the major groove allows nucleotide 0Q to avoid the collision with 0S. Grey: the juxtaposition of the central base pairs in the same theoretically obtained structure as in (a) and (b). Black: the adjustment of the position of a WC base pair [0P; 0Q] that allows it to avoid the collision with base pair [0R;0S]. (d): the subsequent adjustment of the position of nucleotide 0P when a WC base pair [0P; 0Q] (brown) is replaced by GU (blue).

At the 0-level, the most convenient way to resolve the collision between the two internal nucleotides consists in introduction of a GU base pair at position [0P;0Q] or [0R;0S]. In this case, the uridine of the GU base pair becomes displaced in the direction of the major groove, i.e. farther from the opposite base pair. Such displacement would allow the uridine to avoid the collision with the opposite internal nucleotide without creation of conformational tensions within the GU-helix. The network of five inter-helix hydrogen bonds (Figure 6) will additionally stabilize this collision-free arrangement. We thus can conclude that the asymmetry between the two base pairs at the 0-level is initially caused by the potential collision between the two WC base pairs and is resolved through replacement of one of them by GU. Correspondingly, the GU base pairs involved in AGPMs are among the most conserved GU pairs existing in rRNA[15] (Table V).

Table V: Occurrence of different combinations of the 0-base pairs in AGPMs existing in prokaryotic rRNA.

0-base pairs	Number of sequences	%
All motifs except S549		
GU-GC	29 524	69.3
GU-CG	10 108	23.7
GU-AU	38	0.1
GU-UA	1 818	4.2
Total GU-WC	41 488	97.3
WC-WC	549	1.3
Others	590	1.4
Total	42 627	100
Motif S549		
Bacteria		
WC-WC	8 811	95.9
GU-WC	250	2.7
Others	129	1.4
Total	9 190	100
Archaea		
GU-WC	381	77.1
WC-WC	12	2.4
Others	101	20.5
Total	494	100

For the combinations shown in this table, the normal order of two base pairs is not respected. GU always takes the first position regardless whether in a real nucleotide sequence it stays at position [0P;0Q] or [0R;0S].

See the footnote to Table III.

Although most known AGPMs have at the 0-level the GU-WC base pair combination, in some motifs both 0-base pairs are WC. In particular, combination GC-GC exists in motif S549 of all known 30S subunit structures[19,25,29,30] and in motif L2291 of the 50S subunit from *H. marismortui*[20], while combination GC-CG is found in motif HH[27]. Also, after introduction of a WC-WC combination at the 0-level of motifs S296[14], L639 and L65737, the ribosome remained functional, albeit at a lower level. Due to the existence of a potential collision between the two internal nucleotides 0Q and 0S, the accommodation of a WC-WC combination at the 0-level of AGPM requires adaptive rearrangements in the conformation of one or both helices. Such rearrangements will make the conformation of at least one of the two helices no longer optimal, thus resulting in a less stable structure of the whole motif.

To test the importance of the GU-WC combination of base pairs at the 0-level, we made two constructs by replacing in URS the GU-GC combination at the 0-level by GC-GC and CG-CG. The two WC-WC combinations demonstrated a substantial loss of stability from 2.4 to 0.9 and 1.1, respectively. We also tested three constructs containing exclusively GU and UG base pairs at the 0-level: UG-UG, GU-UG and GU-GU. Due to the particular structure of the GU base pair with G and U displaced in the minor and major groove, respectively, we expected to have a major collision between the 0-base pairs for combination UG-UG, a mild collision for combination GU-UG and the absence of collision for combination GU-GU. The MD simulations of all three constructs showed that their instabilities correlated with the severity of the collision, so that the most collision-prone construct UG-UG had the minimal stability 1.0, the collision-free construct GU-GU had the highest stability 1.9, while construct GU-UG with a mild collision had stability 1.4, which is between the two extremes. These simulations also showed that although the absence of collision between the two helices is critical for the AGPM stability, the stabilizing interactions between the two helices are also important. The latter aspect explains why the collision-free construct GU-GU (stability 1.9) is still less stable than URS, which has the two 0-base pairs closely packed with formation of a network of several inter-helix hydrogen bonds (Figure 6).

Even if the two base pairs at the 0-level fit to the GU-WC pattern, the stability of AGPM can be notably affected by the identity of the WC base pair. Thus, the UA/AU

base pair at the 0-level deprives the complex of two H-bonds, one within this base pair and the other one between the two 0-base pairs. Correspondingly, the constructs containing the UA and AU base pair at position [0R;0S] had lower stabilities (1.8 and 1.1) compared to the complexes having the GC and CG base pairs (respectively, 2.4 and 1.9). The latter numbers also indicate that GC at position [0R;0S] provides for a somewhat more stable AGPM than CG. We attribute this difference to the more extensive interaction of the external guanosine 0R with other parts of the motif. A more stable AGPM structure represented by situations where at the 0-level, the GU base pair co-exists with either GC or CG explains the predominance of these combinations in the existing AGPMs (shown in Figure 3; the statistics on the identities of the 0-base pairs is provided in **Appendix 1**).

3.7.2. The +1-base pairs

As mentioned above, the two regular WC double helices erected on the two -1-base pairs collide with each other not only at the 0-level, but also at the +1-level. Like at the 0-level, the introduction of GU as one of the two +1-base pairs should relax the tension between the two helices, as it happens, for example, in motif L657 (Figure 3). There is, however, an essential difference between the 0- and +1-levels. While at the 0-level the inter-helix contacts include both backbone-backbone and backbone-base interactions, at the +1-level the inter-helix contact is made only by the riboses of the two internal nucleotides +1Q and +1S (Figure 7). Therefore, the major aspect determining the identities of the nucleotides at the +1 level is whether they can properly position the riboses of the two internal nucleotides with respect to each other. As a result, not only GU, but any non-WC base pair in which the internal nucleotide is displaced in the direction of the major groove can relax the collision between the two helices at the +1-level. In particular, among the AGPMs shown in Figure 3, a WC base pair at the +1-level coexists with a non-WC base pair AG (6 times), AC (2) and GU (1). In all these base pairs, the internal nucleotide is displaced towards the major groove, which allows it to avoid the collision with the internal nucleotide of the opposite base pair. In total, among all nucleotide sequences of AGPMs found in prokaryotic rRNA sequences in which a

WC combination at the +1-level coexists with a non-WC one, the latter in 98.7% of all cases found in motifs S62, S757, L554, L639, L657, L2291, L2687 and L2847 is AG, AC, AA or GU (Table VI).

In two more AGPMs S296 and L1864, one of the two +1- base pairs is WC, while in the opposite helix the +1-dinucleotide combination does not form a base pair. In both motifs, the internal nucleotide of this combination occupies virtually the same position as in the above-mentioned non-WC base pairs, being displaced towards the major groove and keeping the interaction with the internal +1-nucleotide of the opposite helix.

Table VI: Occurrence of different combinations of the +1-base pairs in AGPMs existing in prokaryotic rRNA.

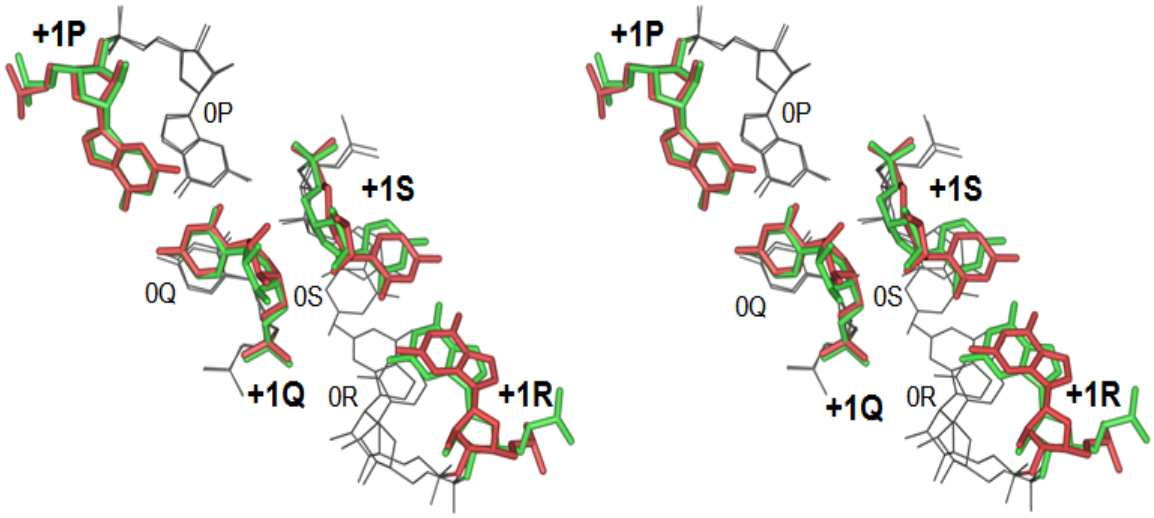
+1-base pairs	Number of sequences	%
Group [WC;non-WC] [a]		
WC-AG	22 543	86.2
WC-AC[b]	2 291	8.7
WC-AA	459	1.8
WC-GU	428	1.6
WC-AU[b]	112	0.4
WC-WC	42	0.2
Others	284	1.1
Total	26 159	100
Group [WC-WC] [a]		
Motifs S911, L839 and L2698		
WC-WC	8 184	89.9
WC-GU	840	9.2
Others	78	0.9
Total	9 102	100
Motif S549		
Bacteria		
WC-WC	8 962	98.3
WC-GU	32	0.4
WC-UG	28	0.3
Others	92	1.0
Total	9 114	100
Archaea		
WC-UG	404	81.6
WC-WC	86	17.4
WC-GU	0	0
Others	5	1.0
Total	495	100

For the combinations shown in this table, the normal order of two base pairs is not respected. WC always takes the first position regardless whether in a real nucleotide sequence it stays at position [+1P;+1Q] or [+1R;+1S].

a: Group [WC;non-WC] includes motifs S62, S757, L554, L639, L657, L2291, L2687 and L2847. Group [WC-WC] includes motifs S549, S911, L839 and L2698 (Figure 3).

b: In the given structural context, all AC and AU base pairs are arranged in the way that the amino group of the adenosine forms a hydrogen bond with atom O2 of the pyrimidine. Such arrangement makes these base pairs similar to the sheared base pairs AG and AA. See also the footnote to Table III.

In spite of the potential collision, six motifs out of the fifteen shown in Figure 3 contain WC base pairs at both +1-positions [+1P;+1Q] and [+1R;+1S]. Although the existence of such structures demonstrates that the simultaneous presence of two WC-base pairs at the +1-level is sterically possible, there are indications that the presence of two WC +1-base pairs can destabilize the arrangement of the two helices. In particular, in all above-mentioned cases, one of the helices is over-twisted compared to the standard A-RNA conformation by the average of 7˚ (Figure 13).



<u>Figure 13</u>: The over-twist observed at the +1-level when both base pairs are either WC or WC-nonWC.
The WC is red and the non-WC is green. When [+1P; +1Q] and [+1R; +1S] are WC, one of the helices is over-twisted compared to the standard A-RNA conformation by the average of 7˚.

To check how the stability of the whole AGPM can be affected by the identities of the +1-base pairs, we performed MD simulations of several constructs. The presence of the GU base pair in position [+1R;+1S] of URS provided for a stable arrangement (S=2.4). The replacement of this GU base pair by GC resulted in a drop of the stability to S=1.3. When this base pair was replaced by UG, in which the internal guanosine was additionally displaced towards the opposite +1-base pair, the stability dropped to S=0.4. Interestingly, when compared to the URS, GC base pair [+1P;+1Q] and GU base pair [+1R;+1S] were exchanged between themselves, the stability of the arrangement dropped to S=0.6. This effect was due to the interference between the two neighboring GU base pairs in positions [+1P;+1Q] and [0P;0Q]. In fact, the ability of the GU base pair in position [+1R;+1S] to eliminate the collision between the two helices without interference with base pair [0P;0Q] was the reason of its introduction into URS.

Given that among all non-WC dinucleotide combinations observed at the +1-level the most frequent is AG (Figure 3), several MD simulations intended to clarify its role. In practically all cases, the AG combination forms a sheared base pair, in which the external A and the internal G are displaced towards the minor and major grooves, respectively. The presence of such base pair at position [+1R;+1S] of URS will cause interference with the neighboring sheared AG base pair at the +2 level (Figure 4b). To avoid such interference, for the corresponding simulations we used an alternative reference structure presented in Figure 4b. The MD simulations showed that the replacement of the GU base pair in position [+1R;+1S] by the sheared AG base pair did not affect the stability of the arrangement (S=2.4).

3.8. The asymmetric requirement for the GC/CG identity of base pairs [-1P; -1Q]: additional H-bond with the ribose of nucleotide 0S

So far, our discussion of the relation between the -1- and 0-base pairs has been limited to the active role of the -1-base pairs in providing the scaffold allowing the 0-base pairs to form optimal interactions between themselves. To play this role, both -1-base pairs must have the WC geometry and be arranged symmetrically with respect to each other. No additional requirements limiting the identities of the -1-base pairs have been

presumed. Further analysis, however, showed that not only the arrangement of the -1-base pairs determines the arrangement of the 0-base pairs, but the influence can also go in the opposite direction from the 0- to -1-level. Due to the fact that the arrangement of the 0-base pairs is essentially asymmetric, these base pairs interact with the -1-base pairs differently. This, in turn, imposes additional asymmetric requirements on the identities of the -1-base pairs.

More specifically, we observed that in most available AGPM conformations, oxygen O4' of ribose 0S and the amino group of the guanosine of base pair [-1P; -1Q] are within the distance of 3.2 Å from each other and form an H-bond between themselves (Figure 14). To make this H-bond possible, base pair [-1P; -1Q] must be either GC or CG. Interestingly, the symmetrical H-bond between the O4' of ribose 0P and the amino group of the guanosine of base pair [-1R; -1S] never exists, because the distance between the corresponding atoms always exceeded 3.7 Å (Figure 14). Therefore, base pair [-1R; -1S] does not have additional restrictions on its identity on top of being WC. Inspection of the available nucleotide sequences of rRNA[16] shows that although both -1-base pairs have a strong preference for GC/CG, their level of conservation is notably different. While [-1P; -1Q] has a GC/CG identity in 98.5% of the sequences, for base pair [-1R; -1S] this number reaches 73%, and the remaining 27% of the sequences have the AU/UA identity (Table III). These data strongly suggest the importance of the H-bond between oxygen O4' of ribose 0S and the amino group of the guanosine of base pair [-1P; -1Q] for the integrity of the AGPM.

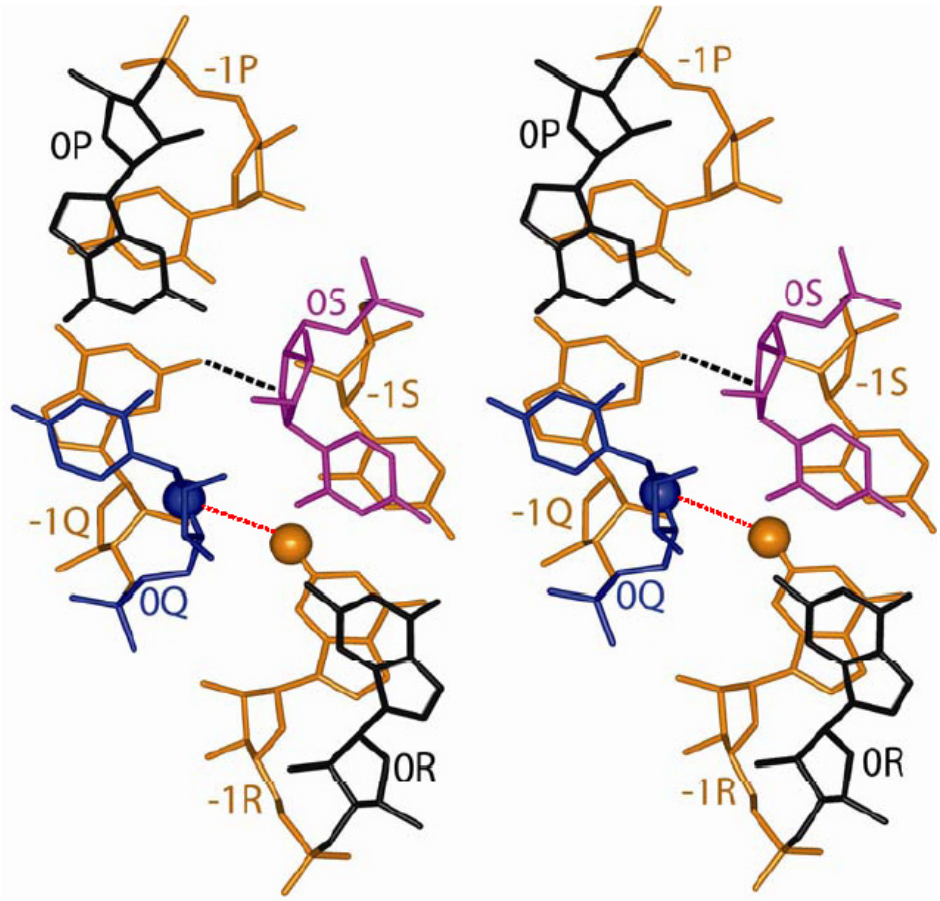


Figure 14: Stereo view of the asymmetry between base pairs [-1P;-1Q] and [-1R;-1S] caused by the displacement of nucleotide 0Q.
In all cases of AGPM, there is a hydrogen bond between atom O4′ of ribose 0S and the amino group of the guanine of base pair [-1P;-1Q] (black dashed line). The analogous bond between atom O4′ of ribose 0Q (blue ball) and the amino group of the guanine of base pair [-1R;-1S] (brown ball) does not exist due to the asymmetric displacement of nucleotide 0Q (red dashed line) which position them far from each other (3.7Å).

To check the importance of the H-bond between the ribose of nucleotide 0S and the amino group of the guanosine in either position -1P or -1Q for the AGPM stability, we ran several MD simulations. As mentioned above, in the simulations described so far this H-bond was excluded from the energy formula, which now allows us to directly address the impact of this H-bond on the stability of the motif. The introduction of the additional hydrogen bond into the formula for the energy calculation modestly increased the stability of URS (combination CG/CG at the -1-level) from 2.4 to 2.6. However, for some other combinations of -1-base pairs, the presence or absence this H-bond in the formula for the energy calculation had a much stronger effect on the AGPM stability. In particular, we noticed that when this H-bond was not taken into account, the presence of the GC combination in either position [-1P; -1Q] or [-1R; -1S] substantially decreased the stability of the motif (the corresponding stabilities were S=1.0 and S=1.5, respectively). In the extreme case when both -1-base pairs were GC, the arrangement behaved as completely unstable (S=0.3). However, when the energy corresponding to this H-bond was added to the energy formula, the stability of this arrangement restored (S=2.4).

The reason for the lower stability of the AGPM constructs containing the GC combinations at the -1-level and the stabilizing role of the above-mentioned H-bond between the ribose of nucleotide 0S and the amino group of the guanosine of base pair [-1P; -1Q] became clear when we analyzed the corresponding MD trajectories that led to the rapid destruction of AGPM. This analysis revealed a particular mechanism of the AGPM destruction in which the GC identities of the -1-base pairs played the key role. Given than in the URS, both 0-base pairs were purine-pyrimidine (GU and GC, Figure 4a), the presence of a GC base pair at the -1-level facilitated a particular rearrangement of base pairing in which a nucleotide assigned for a -1-base pair broke its normal base pairing and formed a new base pair with the neighboring nucleotide assigned for the 0-base pair (for example, base pair [0P; -1Q], [-1P; 0Q] or equivalent base pairs in the WC-helix). However, the formation of the H-bond between the ribose of nucleotide 0S and the amino group of the guanosine of base pair [-1P; -1Q] effectively blocked this kind of slippage, thus restoring the stability of the whole arrangement. This explains the very high level of conservation of this H-bond in the existing nucleotide sequences of AGPM.

In the absence of the H-bond, as in constructs having AU/UA at the -1-level of GU-helix, the stabilities of -1AU-CG, -1UA-CG drop to 1.1 and 1.0 respectively. However, the symmetrical replacement in the WC-helix results in a less drastic decrease of stability. Thus -1AU-CG has a stability of 1.8 while for -1UA-CG S=1.4. This difference allows us to corroborate that base pair [-1P;-1Q] is more sensitive to absence of a GC/CG combination than base pair [-1R;-1S].

3.9. The optimal secondary structure of AGPM

Based on the performed MD simulations, the analysis of the available prokaryotic rRNA sequences of AGPM and different backbone-backbone interactions within the motif, we can suggest a consensus secondary structure to which most known AGPMs fit (Figure 15). In this consensus structure, the identities of base pairs are divided in three categories of those that are satisfied in practically all nucleotide sequences, those for which a strong preference exists in the analyzed AGPM structures, and those that are generally acceptable and appear at a notable level. This consensus structure can be used for the search of new cases of AGPM in RNA molecules for which the tertiary structure is yet to be determined.

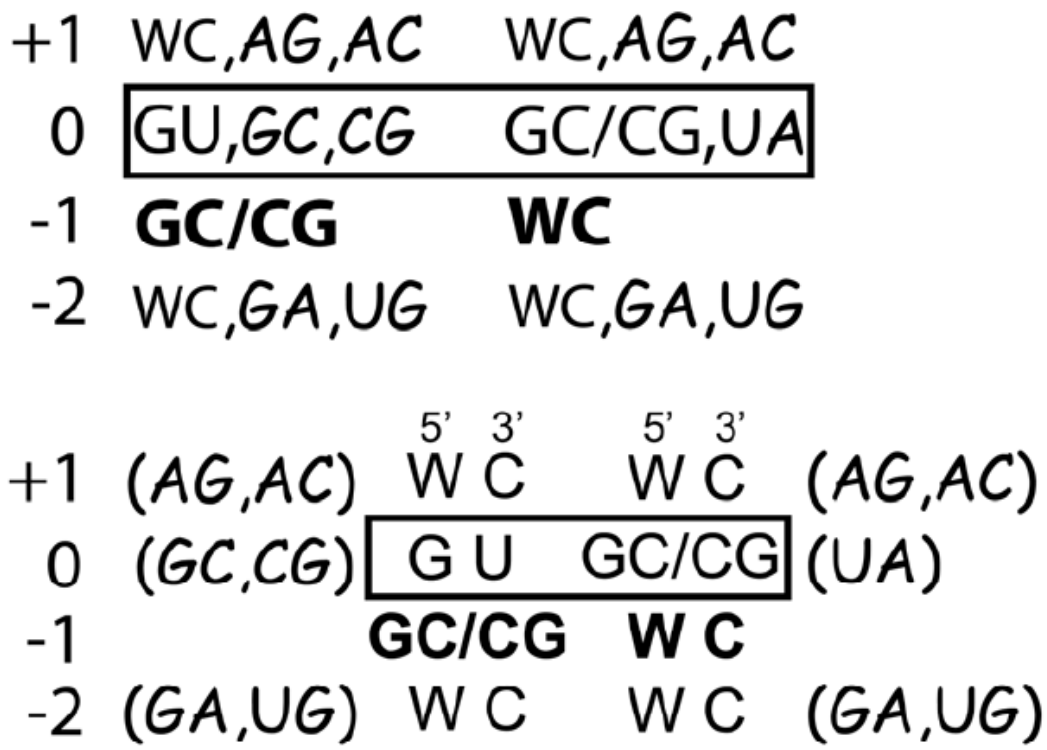


Figure 15: AGPM consensus structure
The consensus secondary structure to which most known AGPMs fit. The base pair identities found in practically all nucleotide sequences, in most sequences, and in a notable number of sequences are shown, respectively, in bold, with the regular and comic fonts. The central base pairs are boxed

4. Discussion and Conclusion

In this paper, we analyze the principles of helix-helix interaction within AGPM, which constitutes an important element of the ribosome architecture. Until now, the nucleotide sequence requirements for the formation of AGPM have been thought to be limited to the coexistence of the GU and WC as central base pairs. The fact that the presence of two central WC base pairs in some cases of AGPM does not interfere with the formation of the motif has been mainly ignored. Here, for the first time, we demonstrate that AGPM cannot be reduced to the interaction of the central base pairs. The contacting area spreads over four base pairs in both helices and includes totally twelve nucleotides. The additional interactions are very important: they are primarily responsible for the maintenance of the juxtaposition of the helices and are able to keep the integrity of the motif even if the central base pairs do not follow the GU-WC pattern.

A specific characteristic of AGPM pertains to the fact that in this motif, most inter-helix interactions occur between the backbones of the two helices. Based on the particular strands involved in the inter-helix contacts, we identified three contact zones QS, QR and PS, and in each zone, the contacts are mainly made by two riboses, respectively, [+1Q;+1S], [-1Q;-2R] and [-2P;-1S]. These interactions are mainly hydrophobic, although on some occasions, hydrogen bonds could also be formed. The ribose-ribose interactions exist in all cases of AGPM and thus constitute an essential aspect of the motif. In each helix, the three riboses form a triangle, which is superposed on the corresponding triangle from the opposite helix. The interaction of the two triangles guides the formation of the whole arrangement and makes it rigid.

The fact that most inter-helix contacts within AGPM are made by the backbones does not mean that the identities of the bases are not important. On the contrary, in order to allow the simultaneous inter-helix interaction within all contact zones, the shapes of both helices should be tuned to each other. This tuning proceeds through the thorough selection of the identities of all base pairs participating in the inter-helix contacts, which would allow the particular positioning of the riboses. At the -1-level, all four nucleotides are involved in the inter-helix contacts; the possibility for formation of these interactions

is guaranteed by the high conservation of the WC identities of both -1-base pairs in all motifs. At levels +1 and -2, only one nucleotide in each base pair is involved in the inter-helix contact. Correspondingly, the identities of the +1- and -2-base pairs are such that they are able to provide the proper position of this nucleotide with respect to the nucleotide from the opposite helix with which it interacts.

An important result of this paper consists in the understanding that the two parts of the motif, encompassing, respectively, layers -1 and -2 and layers 0 and +1 play essentially different roles in the integrity of AGPM. The formation of the inter-helix interactions at layers -1 and -2 provides the scaffold to which the interactions at levels 0 and +1 should fit. Indeed, the geometry of the inter-helix packing of layers -1 and -2 is close to optimal. When all -1- and -2-base pairs are WC, the two helices have the standard A-RNA conformation and form extensive interaction with each other within contact zones PS and QR. When one or both -2-base pairs are non-WC, the external nucleotides of these base pairs are still able to maintain the interactions with the opposite helix. The introduction of non-WC base pairs at the -2-level is thus not dictated by the internal logic of AGPM and seems to be required for a better accommodation of the motif to its immediate surroundings. However, at the levels 0 and +1 the situation is different. The extension of both helices up from layer -1 in the regular A-RNA conformation leads to their collision at levels 0 and +1. A way to avoid this collision would consist in a deformation of the optimal geometry in at least one helix, which can be relaxed through the introduction of non-WC base pairs at both levels. Thus, for levels 0 and +1, unlike for level -2, the presence of non-WC base pairs is mainly determined by the internal logic of AGPM and only to a lesser extent by the interaction with surrounding regions.

The importance of the ribose-ribose interactions for the integrity of AGPM downplays the role of the central base pairs. We show here that the presence of the GU-WC pattern at the zero level is neither necessary nor sufficient for the formation of AGPM. In fact, the role of the GU base pair in position [0P;0Q] to a great extent is limited to providing a way for relaxation of the conformational tensions caused by the displacement of nucleotide 0Q from its regular position due to the potential collision with 0S. Even without this relaxation the motif can still form, although in this case, it is expected to be less stable. The collision between 0Q and 0S originates from the optimal

helix-helix interactions at levels -1 and -2 and is strongly dependent on the WC identity of the -1-base pairs. In other words, the importance of the central base pairs seems to be secondary compared to that of the -1-base pairs.

While at the 0-level, WC-WC combinations occur mostly as exceptions, at the +1-level they are found in about a third of all motifs and are thus more acceptable than at the 0-level. However, even here, the negative effect of such combination is obvious, as one can judge from the fact that in all such cases the optimal geometry of one of the helices is distorted between layers 0 and +1. As determined above from the MD simulations, the presence of such distortion results in a lower stability of the motifs, which, however, should not necessarily be harmful for the ribosome function. One can expect that the ribosome function requires that some motifs break at particular moments of the functional cycle and that a lower stability of the motif would be helpful for its effective breakage. Interestingly, two of the five motifs in Figure 3 that are associated with rRNA and contain a WC-WC combination at the +1-level have already been known to break during the ribosome function. In particular, motif S911 is expected to break and form *de novo* at the initiation of translation[22], while motif L1923-P breaks and forms during the ribosome translocation. A lower stability of these motifs could facilitate the dissociation of the helices when it is functionally required. Another motif of this group, S549, has different kinds of abnormalities. In bacterial rRNA, it has WC-WC combinations at both levels 0 and +1. In archaeal rRNA, it contains a UG base pair at the +1-level. The presence of such abnormalities in S549 indicates that it could play a specific role during the ribosome function. Whether it is true or not would require additional experimental support.

The AGPM structure is characterized by the presence of different types of potentially conflicting interactions. In our analysis, the requirements were always formulated in the form to avoid collision between the two interacting parts from the two helices. The phenomenon of shaping the backbone through formation of particular base pairs has a high chance to be observed not only in AGPM, but in other RNA arrangements as well. Given that in the ribosome and in other RNA-containing complexes there are many occasions when the RNA backbone is involved in essential inter- or intra-molecular interactions, the situations when the shape of the backbone and

its ability to participate in these interactions are regulated by the formation of particular base pairs are expected to be very common[34]. The understanding of the fundamentals of RNA structure and folding will require a systematic consideration of all such cases.

5. References

1. Kim, S. H. & Sussman, J. L. (1976). pi turn is a conformational pattern in RNA loops and bends. *Nature* **260**, 645-646.
2. Quigley, G. J. & Rich, A. (1976). Structural domains of transfer RNA molecules. *Science* **194**, 796-806.
3. Holbrook, S. R., Sussman, J. L., Warrant, R. W. & Kim, S. H. (1978). Crystal structure of yeast phenylalanine transfer RNA. II. Structural features and functional implications. *J. Mol. Biol.* **123**, 631-660.
4. Klein, D. J., Schmeing, T. M., Moore, P. B. & Steitz, T. A. (2001). The kink-turn: a new RNA secondary structure motif. *EMBO J.* **20**, 4214-4221.
5. Szep, S., Wang, J. & Moore, P. B. (2003). The crystal structure of a 26-nucleotide RNA containing a hook-turn. *RNA* **9**, 44-51.
6. Correll, C. C., Freeborn, B., Moore, P. B. & Steitz, T. A. (1997). Metals, motifs, and recognition in the crystal structure of a 5S rRNA domain. *Cell* **91**, 705-712.
7. Correll, C. C., Wool, I. G. & Munishkin, A. (1999). The two faces of the *Escherichia coli* 23 S rRNA sarcin/ricin domain: the structure at 1.11 Å resolution. *J. Mol. Biol.* **292**, 275-287.
8. Lescoute, A., Leontis, N. B., Massire, C. & Westhof, E. (2005). Recurrent structural RNA motifs, Isostericity Matrices and sequence alignments. *Nucleic Acids Res.* **33**, 2395-2409.
9. Hendrix, D. K., Brenner, S. E. & Holbrook, S. R. (2005). RNA structural motifs: building blocks of a modular biomolecule. *Q. Rev. Biophys.* **38**, 221-243.
10. Leontis, N. B., Lescoute, A. & Westhof, E. (2006). The building blocks and motifs of RNA architecture. *Curr. Opin. Struct. Biol.* **16**, 279-287.
11. Noller, H. F. (2005). RNA structure: reading the ribosome. *Science* **309**, 1508-1514.
12. Gagnon, M. G. & Steinberg, S. V. (2002). GU receptors of double helices mediate tRNA movement in the ribosome. *RNA* **8**, 873-877.

13. Sponer, J. E., Reblova, K., Mokdad, A., Sychrovsky, V., Leszczynski, J. & Sponer, J. (2007). Leading RNA Tertiary Interactions: Structures, Energies, and Water Insertion of A-Minor and P-Interactions. A Quantum Chemical View. *J. Phys. Chem. B* **111**, 9153-9164.
14. Gagnon, M. G., Mukhopadhyay, A. & Steinberg, S. V. (2006). Close packing of helices 3 and 12 of 16S rRNA is required for the normal ribosome function. *J. Biol. Chem.* **281**, 39349-39357.
15. Mokdad, A., Krasovska, M. V., Sponer, J. & Leontis, N. B. (2006). Structural and evolutionary classification of G/U wobble basepairs in the ribosome. *Nucleic Acids Res.* **34**, 1326-1341.
16. Wuyts, J., Perriere, G. & Van De Peer, Y. (2004). The European ribosomal RNA database. *Nucleic Acids Res.* **32**, D101-103.
17. Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A. & Steinberg, S. (1998). Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.* **26**, 148-153.
18. Harms, J., Schluenzen, F., Zarivach, R., Bashan, A., Gat, S., Agmon, I., Bartels, H., Franceschi, F. & Yonath, A. (2001). High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. *Cell* **107**, 679-688.
19. Schuwirth, B. S., Borovinskaya, M. A., Hau, C. W., Zhang, W., Vila-Sanjurjo, A., Holton, J. M. & Cate, J. H. (2005). Structures of the bacterial ribosome at 3.5 Å resolution. *Science* **310**, 827-834.
20. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **289**, 905-920.
21. Nissen, P., Ippolito, J. A., Ban, N., Moore, P. B. & Steitz, T. A. (2001). RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc. Natl. Acad. Sci. USA* **98**, 4899-4903.
22. Carter, A. P., Clemons, W. M., Jr., Brodersen, D. E., Morgan-Warren, R. J., Hartsch, T., Wimberly, B. T. & Ramakrishnan, V. (2001). Crystal structure of an initiation factor bound to the 30S ribosomal subunit. *Science* **291**, 498-501.

23. Yusupov, M. M., Yusupova, G. Z., Baucom, A., Lieberman, K., Earnest, T. N., Cate, J. H. & Noller, H. F. (2001). Crystal structure of the ribosome at 5.5 Å resolution. *Science* **292**, 883-896.

24. Korostelev, A., Trakhanov, S., Laurberg, M. & Noller, H. F. (2006). Crystal structure of a 70S ribosome-tRNA complex reveals functional interactions and rearrangements. *Cell* **126**, 1065-1077.

25. Selmer, M., Dunham, C. M., Murphy, F. V., Weixlbaumer, A., Petry, S., Kelley, A. C., Weir, J. R. & Ramakrishnan, V. (2006). Structure of the 70S ribosome complexed with mRNA and tRNA. *Science* **313**, 1935-1942.

26. Murthy, V. L. & Rose, G. D. (2003). RNABase: an annotated database of RNA structures. *Nucleic Acids Res.* **31**, 502-504.

27. Pley, H. W., Flaherty, K. M. & McKay, D. B. (1994). Three-dimensional structure of a hammerhead ribozyme. *Nature* **372**, 68-74.

28. Bhubanananda, S., Prashant K. K. & Simpson J. (2012). Functional replacement of two highly conserved tetraloops in bacterial ribosome. *Biochemistry* **51**, 7618-7626

29. Wimberly, B. T., Brodersen, D. E., Clemons, W. M., Jr., Morgan-Warren, R. J., Carter, A. P., Vonrhein, C., Hartsch, T. & Ramakrishnan, V. (2000). Structure of the 30S ribosomal subunit. *Nature* **407**, 327-339.

30. Schluenzen, F., Tocilj, A., Zarivach, R., Harms, J., Gluehmann, M., Janell, D., Bashan, A., Bartels, H., Agmon, I., Franceschi, F. & Yonath, A. (2000). Structure of functionally activated small ribosomal subunit at 3.3 Å resolution. *Cell* **102**, 615-623.

31. Lu, X. J. & Olson, W. K. (2003). 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* **31**, 5108-5121.

32. Hou, Y. M. & Schimmel, P. (1988). A simple structural feature is a major determinant of the identity of a transfer RNA. *Nature* **333**, 140-145.

33. McClain, W. H. & Foss, K. (1988). Changing the identity of a tRNA by introducing a G-U wobble pair near the 3' acceptor end. *Science* **240**, 793-796.

34. McClain, W. H., Schneider, J., Bhattacharya, S. & Gabriel, K. (1998). The importance of tRNA backbone-mediated interactions with synthetase for aminoacylation. *Proc. Natl. Acad. Sci. USA* **95**, 460-465.

35. Wang, J. & Nikonowicz, E. P. (2011). Solution structure of the K-turn and specifier loop domains from the Bacillus subtilis tyrS T-box leader RNA. *J. Mol. Biol* **408**, 99-117.

36. Butcher, S. E. & Pyle, A. M. (2011). The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks. *Acc. Chem. Res* **44**, 1302-1311.

37. Gagnon, M. G., Boutorine, Y. I. & Steinberg, S. V. (2010). Recurrent RNA motfs as probes for studying RNA-protein interactions in the ribosome. *Nucleic Acids Res* **38**, 3441-3453

38. Demeshkina, N., Jenner, L., Westhof, E., Yusupov, M. & Yusupova, G. (2012). A new understanding of the decoding principle on the ribosome. *Nature* **484**, 256-260

39. Ben-Shem, A., Garreau de Loubresse, N., Melnikov, S., Jenner L., Yusupova G., & Yusupov M. (2011). The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science* **334**, 1524-1529.

40. Rab, J., Leibundgut, M., Ataide, S.F., Haag, A. & Ban, N. (2012). Crystal Structure of the Eukaryotic 40*S* Ribosomal Subunit in Complex with Initiation Factor 1. *Science* **331**, 730-736

41. Kling, S., Voigts-Hoffmann, F., Leibundgut, M., Arpagaus S., Ban, N. (2011). Crystal Structure of the Eukaryotic 60S Ribosomal Subunit in Complex with Initiation Factor 6. *Science* **334**, 941-948

42. Insight II User Guide, October 1995. San Diego: Biosym/MSI

<u>Appendix 1</u> : Table of the statistical spectrum of the identities of the 0-base pairs for all known AGPMs.

AGPM	Central base pairs	Eubacteria		Central base pairs	Archaebacteria			AGPM	Central base pairs	Eubacteria		Central base pairs	Archaebacteria	
		N [b]	%		N [b]	%				N [b]	%		N [b]	%
S62	**GU-GC**	8036	73.77	**GU-GC**	280	53.03		L657	**GU-CG**	248	62.469	**GU-GC**	16	43.243
	GU-CG	2676	9	**GU-CG**	195	0			**GU-GC**	132	33.249	**GU-CG**	8	21.622
	CU-GC	25	24.56	**GU-UA**	37	36.93			**GU-UA**	15	3.778	**GC-GU**	4	10.811
	GU-CC	22	8	GU-CC	6	2			GU-CC	1	0.252	**AU-GU**	3	8.108
	GU-UA	21	0.230	**CG-GU**	2	7.008			UU-GA	1	0.252	GC-GC	3	8.108
	GC-GU	13	0.202	GC-CG	2	1.136			TOTAL [a]	397		**GU-UA**	2	5.405
	GC-GC	11	0.193	GA-UA	1	0.379						GU-UG	1	2.703
	UU-GC	11	0.119	AU-GC	1	0.379						TOTAL [a]	37	
	GU-GG	10	0.101	CU-GC	1	0.189								
	AU-GC	10	0.101	GC-GC	1	0.189								
	GU-AU	10	0.092	GU-GU	1	0.189								
	GU-GU	6	0.092	CG-GC	1	0.189								
	GU-AC	6	0.092	TOTAL [a]	528	0.189								
	GA-GC	6	0.055			0.189								
	CU-CG	4	0.055											
	GU-UC	4	0.055											
	GU-CA	3	0.037											
	GG-GC	3	0.037											
	GC-CG	3	0.028											
	GU-AG	3	0.028											
	UG-GU	3	0.028											
	UU-CG	3	0.028											
	GG-CC	1	0.028											
	GU-GA	1	0.028											
	GU-UG	1	0.009											
	TOTAL [a]	10892	0.009											
			0.009											
S296	**GU-GC**	5968	97.93	**GU-GC**	286	71.85		L839	**GU-CG**	265	67.089	**GU-CG**	17	47.222
	GU-AC	27	2	**GU-CG**	61	9			**GU-UA**	63	15.949	GC-CG	10	27.778
	GC-GC	26	0.443	GU-GU	31	15.32			**GU-GC**	56	14.177	**GU-UA**	4	11.111
	AU-GC	25	0.427	**GU-UA**	6	7			AU-UA	4	1.013	AU-CG	2	5.556
	GU-UC	18	0.410	**GU-AU**	6	7.789			CG-CG	3	0.759	CG-CG	2	5.556
	GU-CC	14	0.295	AU-GC	4	1.508			GU-UG	2	0.506	UA-UG	1	2.778
	GU-GU	4	0.230	GU-UG	1	1.508			GU-GG	1	0.253	TOTAL [a]	36	
	GU-AU	3	0.066	AU-CG	1	1.005			UA-CG	1	0.253			
	UU-GC	3	0.049	GU-CU	1	0.251			TOTAL [a]	395				
	CU-GC	2	0.049	GU-AG	1	0.251								
	GU-GA	1	0.033	TOTAL [a]	398	0.251								
	AC-GC	1	0.016			0.251								
	CU-UC	1	0.016											
	AU-CC	1	0.016											
	TOTAL [a]	6094	0.016											
S549	GC-GC	8032	87.39	**GU-CG**	206	41.7		L1864	**GU-GC**	390	98.734	Replaced by GNRA[1]		
	GC-CG	668	9	GU-UG	96	19.43			AU-GC	3	0.759			
	GC-GU	230	7.269	**GU-UA**	85	3			CU-GC	1	0.253			
	GC-AU	95	2.503	**GU-GC**	59	17.20			GU-GA	1	0.253			
	CC-GC	35	1.034	**GU-AU**	31	6			TOTAL [a]	395				
	UC-GC	21	0.381	GC-GC	7	11.94								
	GU-GC	15	0.229	CG-CG	4	3								
	AC-GC	15	0.163	CG-UG	3	6.275								
	GC-UA	14	0.163	CU-CG	1	1.417								
	GC-CC	11	0.152	GC-UA	1	0.81								
	GC-GG	10	0.120	GU-GU	1	0.607								
	GC-UG	9	0.109	TOTAL [a]	494	0.202								
	CC-CG	5	0.098			0.202								
	GC-UC	4	0.054			0.202								
	GU-AU	4	0.044											
	CC-GU	4	0.044											
	GC-AC	2	0.044											
	GG-GC	2	0.022											
	AC-GU	2	0.022											
	GC-AG	1	0.022											
	GA-GC	1	0.011											
	GA-UU	1	0.011											
	GU-CG	1	0.011											
	GG-GU	1	0.011											
	GG-GG	1	0.011											
	GC-GA	1	0.011											
	GG-CG	1	0.011											
	CG-GC	1	0.011											

	CG-CG CC-AU UC-GU TOTAL [a]	1 1 1 9190	0.011 0.011 0.011 0.011											
S757	**GU-GC** **GU-CG** AU-GC CU-GC GU-GU UU-GC **GU-AU** GU-AC GA-GC GC-GC GU-UC GG-GC GU-CC GU-GA UU-CC TOTAL [a]	11737 43 26 18 13 11 10 6 5 5 4 2 2 2 1 11885	98.75 5 0.362 0.219 0.151 0.109 0.093 0.084 0.050 0.042 0.042 0.034 0.017 0.017 0.017 0.008	**GU-GC** GC-AC GU-GU **GU-AU** GG-GC **AU-GU** TOTAL [a]	555 12 9 5 1 1 583	95.19 7 2.058 1.544 0.858 0.172 0.172		L229 1	**GU-GC** **GU-CG** **GU-UA** AU-GC CU-GC TOTAL [a]	382 8 3 2 1 396	96.465 2.020 0.758 0.505 0.253	**GU-GC** **GU-CG** GC-GC AU-GC TOTAL [a]	23 9 3 2 37	62.162 24.324 8.108 5.405
S911	**GU-CG** **GU-UA** AU-CG AU-UA GU-CA GU-UG CU-CG GU-CC UA-CG GU-GG GU-CU UU-CG **GU-GC** GC-CG GU-GA AU-GC GU-AG UU-GC GA-CG AU-GG GU-UC TOTAL [a]	6125 1585 273 104 89 78 14 11 7 7 7 5 4 3 2 2 1 1 1 1 1 8321	73.60 9 19.04 8 3.281 1.250 1.070 0.937 0.168 0.132 0.084 0.084 0.084 0.060 0.048 0.036 0.024 0.024 0.012 0.012 0.012 0.012 0.012	**GU-CG** **GU-GC** GU-UG **GU-UA** AU-CG GU-CA GC-CG TOTAL [a]	231 98 33 11 5 1 1 380	60.78 9 25.78 9 8.684 2.895 1.316 0.263 0.263		L268 7	**GU-GC** **GU-UA** **GU-CG** GU-GU GU-CA UU-GC TOTAL [a]	269 34 23 1 1 1 329	81.763 10.334 6.991 0.304 0.304 0.304	**GU-GC** **GU-CG** **GU-UA** GU-UC TOTAL [a]	23 11 2 1 37	62.162 29.730 5.405 2.703
L554	**GU-GC** **GC-GU** GC-GC AU-GC TOTAL [a]	332 6 2 1 341	97.36 1 1.760 0.587 0.293	**GU-GC** TOTAL [a]	33 33	100		L269 8	**GU-GC** GU-AC TOTAL [a]	313 5 318	98.428 1.572	**GU-GC** TOTAL [a]	37 37	100
L639	**GU-GC** **GU-CG** **GU-UA** **GC-GU** GU-GG GU-CC TOTAL [a]	247 101 34 8 5 1 396	62.37 4 25.50 5 8.586 2.020 1.263 0.253	**GU-CG** **GU-GC** TOTAL [a]	36 1 37	97.29 7 2.703		L284 7	**GU-GC** **GU-CG** **GU-UA** TOTAL [a]	243 49 1 293	82.935 16.724 0.341	**GU-GC** TOTAL [a]	32 32	100

The data were obtained from the available rRNA alignments[2]. For all cases of AGPM, the *E. coli* numbering is used.

a: For the statistics, only those cases where the identities of all four nucleotides are known have been considered.

b: N is the number of sequences which have the corresponding 0-base pairs combination.

Central base pairs in bold are those following the GU-WC pattern. Bold red indicates a GU↔WC; WC↔GU replacement.

References

1. Mokdad, A., Krasovska, M. V., Sponer, J. & Leontis, N. B. (2006). Structural and evolutionary classification of G/U wobble basepairs in the ribosome. *Nucleic Acids Res.* **34**, 1326-1341.

2. Wuyts, J., Perriere, G. & Van De Peer, Y. (2004). The European ribosomal RNA database. *Nucleic Acids Res.* **32**, D101-103.

Appendix 2 : Table of the statistical spectrum of the identities of the +1 base pairs for all known AGPMs.

AGPM	base pairs [+1P;+1Q]-[+1R;+1S]	Eubacteria		base pairs [+1P;+1Q]-[+1R;+1S]	Archaebacteria			AGPM	base pairs [+1P;+1Q]-[+1R;+1S]	Eubacteria		base pairs [+1P;+1Q]-[+1R;+1S]	Archaebacteria	
		N [b]	%		N [b]	%				N [b]	%		N [b]	%
S62	GC-AG	10565	96.299	GC-AG	417	79.127		L657	AU-GU	191	48.111	GC-GC	22	59.459
	UA-AG	143	1.303	AU-AG	68	12.903			UA-GU	99	24.937	GC-GU	15	40.541
	AU-AG	94	0.857	GU-AG	36	6.831			GC-GU	89	22.418	TOTAL [a]	37	
	GC-AC	38	0.346	UC-AG	1	0.190			CG-GU	12	3.023			
	GU-AG	36	0.328	GC-CU	1	0.190			GC-GC	2	0.504			
	CG-AG	20	0.182	UU-AG	1	0.190			AU-GC	1	0.252			
	GG-AG	15	0.137	GC-CG	1	0.190			UU-GA	1	0.252			
	GC-CG	10	0.091	GC-AA	1	0.190			GC-UA	1	0.252			
	GC-AA	8	0.073	GG-AG	1	0.190			UA-AU	1	0.252			
	GC-GC	6	0.055	TOTAL [a]	527				TOTAL [a]	397				
	UC-AG	6	0.055											
	GC-GG	5	0.046											
	CC-AG	4	0.036											
	GC-AU	4	0.036											
	AC-AG	4	0.036											
	GA-AG	3	0.027											
	GC-UG	2	0.018											
	GC-CA	2	0.018											
	AA-AG	1	0.009											
	GC-CU	1	0.009											
	CA-AG	1	0.009											
	UA-GG	1	0.009											
	UA-AC	1	0.009											
	AU-AU	1	0.009											
	TOTAL [a]	10971												
S296	AG-AG	5794	97.690	AG-AG	250	62.814		L839	GC-CG	194	49.114	GC-GC	9	25.000
	AU-AG	40	0.674	AG-AA	124	31.156			GC-UA	90	22.785	GC-CG	8	22.222
	AA-AG	23	0.388	AG-CG	9	2.261			AU-CG	59	14.937	GC-AU	7	19.444
	AG-CG	15	0.253	AG-CA	6	1.508			AU-UA	17	4.304	CG-CG	7	19.444
	AG-AA	15	0.253	AG-UA	4	1.005			GU-CG	13	3.291	UG-CG	3	8.333
	AG-UG	15	0.253	AG-GA	1	0.251			GU-UA	7	1.772	GU-CG	1	2.778
	AG-GG	10	0.169	GG-AA	1	0.251			AU-UG	3	0.759	CG-GC	1	2.778
	AG-AC	6	0.101	CU-CA	1	0.251			UA-GC	3	0.759	TOTAL [a]	36	
	CU-AG	6	0.101	AG-AU	1	0.251			AU-GC	2	0.506			
	GU-UA	3	0.051	AG-UG	1	0.251			UU-CG	2	0.506			
	AG-AU	2	0.034	TOTAL [a]	398				GC-GC	1	0.253			
	GG-AG	1	0.017						GA-CG	1	0.253			
	AC-AG	1	0.017						UG-UA	1	0.253			
	TOTAL [a]	5931							UA-UA	1	0.253			
									GC-AU	1	0.253			
									TOTAL [a]	395				
S549	CG-UA	4505	49.429	UG-CG	372	75.152		L1864	AU-UA	231	75.987	Replaced by GNRA[1]		
	CG-GC	3228	35.418	CG-CG	80	16.162			AC-UA	57	18.750			
	CG-CG	930	10.204	UG-UA	32	6.465			AA-UA	8	2.632			
	CG-AU	276	3.028	GC-UA	5	1.010			CC-UA	3	0.987			
	CG-GU	32	0.351	GG-CG	1	0.202			AG-UA	2	0.658			
	GG-GC	20	0.219	UA-CG	1	0.202			CU-UA	2	0.658			
	UA-UA	13	0.143	UG-GG	1	0.202			GU-UA	1	0.329			
	CG-UC	10	0.110	UG-CC	1	0.202			TOTAL [a]	304				
	UG-GC	10	0.110	UC-UA	1	0.202								
	CG-UG	8	0.088	UG-UG	1	0.202								
	CG-GA	7	0.077	TOTAL [a]	495									
	UA-CG	7	0.077											
	CG-CA	7	0.077											
	UG-CG	6	0.066											
	CC-UA	5	0.055											
	GG-UA	5	0.055											
	CG-AC	4	0.044											
	AG-UA	4	0.044											
	UG-UA	4	0.044											
	CC-GC	4	0.044											
	AG-GC	3	0.033											
	CG-GG	3	0.033											
	CG-AA	3	0.033											
	CU-UA	3	0.033											
	CG-UU	2	0.022											
	CG-CC	2	0.022											

	UA-GC	2	0.022											
	GG-CG	2	0.022											
	GG-GA	1	0.011											
	CU-AU	1	0.011											
	CU-GC	1	0.011											
	CU-CC	1	0.011											
	CC-AU	1	0.011											
	UG-GA	1	0.011											
	UG-CU	1	0.011											
	CA-GC	1	0.011											
	UA-AU	1	0.011											
	TOTAL [a]	9114												
S757	AG-GC	8952	76.324	AG-GC	526	91.638		L2291	AU-AG	251	63.544	GC-AG	31	83.784
	AC-GC	2237	19.072	AG-AU	18	3.136			UA-AG	73	18.481	AU-AG	5	13.514
	AA-GC	378	3.223	AA-GC	16	2.787			CG-AG	46	11.646	GC-GG	1	2.703
	AU-GC	59	0.503	AC-AU	4	0.697			GC-AG	19	4.810	TOTAL [a]	37	
	AU-AU	33	0.281	AA-AU	3	0.523			UA-CG	4	1.013			
	AG-GU	11	0.094	AC-GC	3	0.523			GU-AG	2	0.506			
	GG-GC	8	0.068	AG-AC	2	0.348			TOTAL [a]	395				
	AG-AC	7	0.060	GG-GC	1	0.174								
	AC-GU	5	0.043	AG-CC	1	0.174								
	CG-GC	4	0.034	TOTAL [a]	574									
	AG-GA	4	0.034											
	GU-GC	4	0.034											
	AG-GG	4	0.034											
	UG-GC	3	0.026											
	AA-AU	3	0.026											
	AC-CC	3	0.026											
	CC-GC	2	0.017											
	AC-AC	2	0.017											
	GC-GC	2	0.017											
	AG-AU	1	0.009											
	UC-GC	1	0.009											
	AC-GG	1	0.009											
	AA-GA	1	0.009											
	AC-GA	1	0.009											
	AG-UC	1	0.009											
	AG-CC	1	0.009											
	GG-GU	1	0.009											
	TOTAL [a]	11729												
S911	GC-AU	5554	69.897	GU-AU	197	54.722		L2687	AG-GC	134	40.854	AU-GC	11	29.730
	GC-GC	1958	24.641	GU-UA	98	27.222			AG-CG	118	35.976	AC-GC	8	21.622
	GC-GU	192	2.416	GU-GC	53	14.722			GG-GC	21	6.402	AG-GC	7	18.919
	GU-AU	181	2.278	GU-CG	4	1.111			AG-GU	16	4.878	AG-CG	7	18.919
	GC-AC	11	0.138	GU-GU	2	0.556			AG-AU	13	3.963	GU-GC	3	8.108
	GC-AG	8	0.101	GC-UA	2	0.556			AG-UA	11	3.354	AU-CG	1	2.703
	GC-GG	6	0.076	GU-GG	1	0.278			AA-GC	8	2.439	TOTAL [a]	37	
	GC-UC	5	0.063	GU-UG	1	0.278			UG-GC	3	0.915			
	GC-AA	5	0.063	GU-AA	1	0.278			AU-CG	1	0.305			
	GU-UA	3	0.038	GU-AG	1	0.278			GG-CG	1	0.305			
	GG-AU	3	0.038	TOTAL [a]	360				AG-UG	1	0.305			
	GC-CG	3	0.038						AU-UA	1	0.305			
	GU-GU	2	0.025						TOTAL [a]	328				
	UC-AU	2	0.025											
	GC-CC	2	0.025											
	GC-GA	2	0.025											
	CC-AU	2	0.025											
	AC-AU	1	0.013											
	GC-UA	1	0.013											
	GU-GC	1	0.013											
	AC-GU	1	0.013											
	GC-UU	1	0.013											
	GU-AC	1	0.013											
	GA-GC	1	0.013											
	TOTAL [a]	7946												
L554	AG-GC	129	38.279	GC-GU	10	30.303		L2698	GC-GC	107	32.622	GC-GC	18	48.649
	AG-CG	77	22.849	GC-GC	8	24.242			AU-AU	64	19.512	GC-AU	8	21.622
	AG-AU	47	13.947	GU-GC	6	18.182			GC-AU	33	10.061	GC-GU	6	16.216
	AG-UA	28	8.309	GC-CG	4	12.121			GU-AU	31	9.451	AU-GC	4	10.811
	AG-GU	21	6.231	AU-GC	2	6.061			GC-GU	28	8.537	GC-CG	1	2.703
	AA-UA	14	4.154	GU-CG	1	3.030			CG-GC	20	6.098	TOTAL [a]	37	
	AA-CG	14	4.154	AU-GU	1	3.030			AU-GU	11	3.354			
	AA-GC	2	0.593	GC-UA	1	3.030			UA-AU	9	2.744			
	AA-GU	2	0.593	TOTAL [a]	33				CG-GU	8	2.439			
	AU-UA	1	0.297						UG-GC	4	1.220			
	UA-UA	1	0.297						GU-GC	4	1.220			
	GU-GU	1	0.297						UG-AU	3	0.915			

	TOTAL [a]	337							UA-GU AU-GC AU-UU GU-GU TOTAL [a]	2 2 1 1 328	0.610 0.610 0.305 0.305			
L639	GC-AG AU-AG CG-AG GC-AA GU-AG TOTAL [a]	352 21 11 8 3 395	89.114 5.316 2.785 2.025 0.759	GC-AG CG-AG UA-AG TOTAL [a]	26 10 1 37	70.270 27.027 2.703		L2847	AG-UA AG-CG AG-AU AA-UA TOTAL [a]	158 105 27 3 293	53.925 35.836 9.215 1.024	AG-CG TOTAL [a]	32 32	100

The data were obtained from the available rRNA alignments[2]. For all cases of AGPM, the *E. coli* numbering is used.

a: For the statistics, only those cases where the identities of all four nucleotides are known have been considered.

b: N is the number of sequences which have the corresponding +1 base pairs combination.

References

1. Mokdad, A., Krasovska, M. V., Sponer, J. & Leontis, N. B. (2006). Structural and evolutionary classification of G/U wobble basepairs in the ribosome. *Nucleic Acids Res.* **34**, 1326-1341.

2. Wuyts, J., Perriere, G. & Van De Peer, Y. (2004). The European ribosomal RNA database. *Nucleic Acids Res.* **32**, D101-103.

Appendix 3: Table of the Statistical spectrum of the identities of the -2 base pairs for all known ribosomal AGPMs.

AGPM	base pairs [-2P;-2Q]- [-2R;-2S]	Eubacteria		base pairs [-2P;-2Q]- [-2R;-2S]	Archaebacteria			AGPM	base pairs [-2P;-2Q]- [-2R;-2S]	Eubacteria		base pairs [-2P;-2Q]- [-2R;-2S]	Archaebacteria	
		N [b]	%		N [b]	%				N [b]	%		N [b]	%
S62	GA-CG	95	88.15	GA-GC	32	60.95		L657	AA-GA	25	64.35	GA-GA	23	62.162
	AA-CG	55	4	AA-GC	0	2			GA-GA	1	9	GA-AA	9	24.324
	GA-UA	94	8.700	GA-GU	95	18.09			AA-AA	12	31.02	AA-AA	3	8.108
	GU-UA	3	0.673	AA-AU	56	5			GU-GU	1	6	AA-GA	2	5.405
	GA-UG	73	0.424	AA-GU	19	10.66			AG-GA	13	3.333	TOTAL [a]	37	
	GA-GC	46	0.415	GA-CG	8	7			CA-AA	1	0.256			
	GU-CG	45	0.341	GA-AU	8	3.619			GU-GA	1	0.256			
	GA-GG	37	0.323	AA-UG	4	1.524			UA-GA	1	0.256			
	CA-CG	35	0.166	UA-GC	4	1.524			TOTAL [a]	1	0.256			
	GG-CG	18	0.129	AA-CC	2	0.762				1	0.256			
	AA-GC	14	0.092	UU-AU	2	0.762				39				
	AU-UA	10	0.065	GA-CC	1	0.381				0				
	AU-CG	7	0.065	CA-GU	1	0.381								
	GC-CG	7	0.055	AA-UC	1	0.190								
	GA-GU	6	0.046	GG-GC	1	0.190								
	AA-UG	5	0.046	UA-AU	1	0.190								
	AA-GG	5	0.046	GA-CU	1	0.190								
	AA-UA	5	0.037	TOTAL [a]	1	0.190								
	UA-CG	4	0.037		52	0.190								
	GU-GC	4	0.028		5	0.190								
	GA-CU	3	0.028											
	GA-CC	3	0.028											
	GU-UG	3	0.028											
	GA-AG	3	0.018											
	AA-AU	2	0.018											
	GA-UU	2	0.009											
	GA-UC	1	0.009											
	AC-CG	1	0.009											
	TOTAL [a]	1	0.009											
		1												
		10												
		83												
		9												
S296	GC-AU	31	37.95	AU-CG	16	34.76		L839	AC-GA	37	94.88	AC-GA	35	97.222
	AU-UA	02	9	GC-CG	2	4			AU-GA	1	5	UC-GA	1	2.778
	AU-AU	18	22.72	AU-UA	12	27.03			AG-GA	19	4.859	TOTAL [a]	36	
	AU-GC	57	4	AU-UG	6	9			TOTAL [a]	1	0.256			
	UA-AU	17	21.10	CG-CG	98	21.03				39				
	GC-GC	25	9	GC-UA	31	0				1				
	CG-AU	63	7.746	CG-UA	13	6.652								
	UA-UA	3	3.206	AU-GC	9	2.790								
	GU-AU	26	2.080	GC-AU	6	1.931								
	UG-AU	2	1.395	GC-GC	6	1.288								
	AU-GU	17	1.187	AC-CG	4	1.288								
	AU-UU	0	0.318	CG-UG	3	0.858								
	GC-UA	11	0.306	CG-GC	1	0.644								
	GC-UU	4	0.257	GU-UA	1	0.215								
	AU-GA	97	0.208	UA-AU	1	0.215								
	AU-AA	26	0.171	AU-GG	1	0.215								
	GC-CU	25	0.159	UA-CG	1	0.215								
	AU-AC	21	0.147	AU-AU	1	0.215								
	GC-GU	17	0.135	TOTAL [a]	1	0.215								
	AU-UC	14	0.122		1	0.215								
	CG-GC	13	0.110		46	0.215								
	UC-AU	12	0.098		6									
	AU-CA	11	0.061											
	GC-AC	10	0.061											
	UG-UA	9	0.049											
	AC-AU	8	0.049											
	UA-UU	5	0.049											
	UU-UA	5	0.037											
	CA-AU	4	0.037											
	UU-AU	4	0.024											
	AU-AG	4	0.024											
	CG-GU	3	0.012											
	AU-CU	3	0.012											
	AU-UG	2	0.012											
	GC-AG	2	0.012											
	GA-AU	1	0.012											
	AA-AU	1	0.012											

	CG-CU AG-AU AG-UA GG-UU GU-UU TOTAL [a]	1 1 1 1 1 1 1 1 1 1 1 81 72	0.012 0.012 0.012 0.012 0.012 0.012 0.012 0.012											
S549	UA-AA UA-GA UA-UA UG-GA UA-CA CA-AA UG-AA GA-AA GA-GA UA-AG AA-AA UC-AA UC-GA CA-UA AA-UA UG-UA CG-GA UA-UG UG-CA UU-AA UA-GG CA-GA GG-GA UU-CU UA-AC AA-CA UA-AU UU-UA TOTAL [a]	47 07 28 16 13 91 68 33 17 15 10 5 4 4 3 3 3 2 2 2 2 2 2 2 1 1 1 1 1 1 91 00	51.72 5 30.94 5 15.28 6 0.747 0.363 0.187 0.165 0.110 0.055 0.044 0.044 0.033 0.033 0.033 0.022 0.022 0.022 0.022 0.022 0.022 0.022 0.011 0.011 0.011 0.011 0.011 0.011 0.011	UA-GA UA-AA UA-UA UG-AA CA-AA GA-GA UA-UG CA-GA TOTAL [a]	33 5 12 9 38 1 1 1 1 1 50 7	66.07 5 25.44 4 7.495 0.197 0.197 0.197 0.197 0.197		L1864	UG-AA AU-AA UA-AA GC-AA CG-AA AU-AC UA-GA UU-AA GU-AA UG-GA AU-GA UG-AU UA-AC CA-AA AU-AG TOTAL [a]	24 4 75 20 14 10 6 6 6 5 3 3 2 1 1 1 39 7	61.46 1 18.89 2 5.038 3.526 2.519 1.511 1.511 1.511 1.259 0.756 0.756 0.504 0.252 0.252 0.252	Replaced by GNRA [1]		
S757	CG-GC CG-AU CG-CG UA-CG CG-UA UA-GC UG-GC UA-AU UA-UA UG-CG UG-UA CG-GU CG-AC CA-AU CG-CC CG-UU UG-AU GC-AU CC-AU CG-AG CG-AA CA-CG CA-GC CU-GC CC-GC CG-UG GG-GC GG-CG CU-AU CG-UC CG-CA CG-GG GA-UU	38 96 28 20 21 38 68 3 63 9 50 8 47 1 40 6 10 5 37 16 15 13 8 8 8 8 7 7 6 6 5 5 5	32.80 3 23.74 3 18.00 1 5.751 5.380 4.277 3.966 3.418 0.884 0.312 0.135 0.126 0.109 0.067 0.067 0.067 0.067 0.059 0.059 0.051 0.051 0.042 0.042 0.042 0.042 0.042 0.025 0.025 0.025 0.025	CG-GC CG-UA CG-AU CG-CG UA-CG UG-CG UA-UA UA-GU UA-GC CG-AC UG-GC UA-AU CG-GU CA-UA GC-AU CC-GC TOTAL [a]	24 4 79 56 46 34 27 23 19 18 10 6 5 3 1 1 1 57 3	42.58 3 13.78 7 9.773 8.028 5.934 4.712 4.014 3.316 3.141 1.745 1.047 0.873 0.524 0.175 0.175 0.175		L2291	UG-AU UG-CG CG-CG UG-UA CG-GC UG-GC UA-CG UA-AU TOTAL [a]	18 2 12 6 74 7 2 1 1 1 39 4	46.19 3 31.98 0 18.78 2 1.777 0.508 0.254 0.254 0.254	CG-GC CG-CG UG-GC UG-AU CG-CC TOTAL [a]	22 6 4 4 1 37	59.459 16.216 10.811 10.811 2.703

	AG-CG	5	0.025											
	GC-GC	5	0.017											
	AU-CG	3	0.017											
	CU-CG	3	0.017											
	UG-GG	3	0.017											
	CA-UG	3	0.017											
	AA-AU	3	0.008											
	UC-GC	2	0.008											
	UC-CG	2	0.008											
	UA-GU	2	0.008											
	UA-CA	2	0.008											
	UA-UG	2	0.008											
	UA-AC	1	0.008											
	CG-CU	1	0.008											
	GG-AU	1	0.008											
	CC-CG	1	0.008											
	CA-GU	1	0.008											
	GA-AU	1	0.008											
	AU-GC	1	0.008											
	UG-UG	1	0.008											
	CC-UA	1	0.008											
	UA-UU	1	0.008											
	UG-CU	1	0.008											
	GC-UA	1	0.008											
	CG-GA	1	0.008											
	TOTAL [a]	1	0.008											
		1												
		1												
		1												
		1												
		1												
		1												
		1												
		11												
		87												
		7												
S911	GA-AG	78	92.42	AA-AG	22	60.05		L2687	UG-UA	15	46.03	CA-UA	9	24.324
	UA-AG	28	3.731	GA-AG	4	4			UG-GC	1	7	UG-CG	7	18.919
	AA-AG	31	2.869	AA-AA	13	36.72			CG-UA	13	40.24	UG-GC	6	16.216
	GA-CG	6	0.354	UA-AG	7	9			UA-UA	2	4	CG-UA	5	13.514
	GA-AC	24	0.118	GA-UA	5	1.340			UG-CG	9	2.744	UG-AU	4	10.811
	GA-AA	3	0.106	AA-AU	3	0.804			CG-CG	9	2.744	UG-UA	2	5.405
	GA-AU	30	0.106	TOTAL [a]	3	0.804			UG-AU	8	2.439	CA-GC	2	5.405
	GU-AG	10	0.071		1	0.268			CG-UG	5	1.524	CA-CG	1	2.703
	AA-AA	9	0.071		37				UG-UG	3	0.915	CG-CG	1	2.703
	GG-AG	9	0.047		3				AU-UG	3	0.915	TOTAL [a]	37	
	GA-UA	6	0.035						AU-UA	2	0.610			
	GC-AG	6	0.024						CG-GC	2	0.610			
	GA-CU	4	0.012						UG-GA	2	0.610			
	GA-UG	3	0.012						TOTAL [a]	1	0.305			
	UA-AC	2	0.012							1	0.305			
	CA-AG	1	0.012							32				
	TOTAL [a]	1								8				
		1												
		1												
		84												
		70												
L554	CG-GU	76	22.41	AU-AU	7	21.21		L2698	AU-AU	24	75.68	AU-UG	27	72.973
	AU-AU	67	9	CG-GC	7	2			AU-UU	9	4	AU-GA	5	13.514
	CG-AU	54	19.76	AU-GC	4	21.21			AU-CG	49	14.89	AU-CU	2	5.405
	UA-AU	32	4	CG-AU	4	2			AU-GC	10	4	AU-AA	2	5.405
	CG-GC	18	15.92	UG-UU	3	12.12			GC-AU	8	3.040	AU-UU	1	2.703
	GU-AU	18	9	UG-AU	3	1			AU-UG	5	2.432	TOTAL [a]	37	
	CG-UA	17	9.440	UA-GC	2	12.12			AU-UA	4	1.520			
	GC-GU	13	5.310	CG-GU	1	1			AU-GU	2	1.216			
	AU-GU	10	5.310	UA-AU	1	9.091			AU-CU	1	0.608			
	GC-AU	6	5.015	AU-UA	1	9.091			TOTAL [a]	1	0.304			
	CG-CG	6	3.835	TOTAL [a]	33	6.061				32	0.304			
	GC-GC	4	2.950			3.030				9				
	AU-UU	3	1.770			3.030								
	UU-AU	3	1.770			3.030								
	UA-GC	2	1.180											
	GC-CG	2	0.885											
	UA-UA	2	0.885											
	AU-UA	2	0.590											
	UG-GU	2	0.590											
	UU-GC	1	0.590											
	GA-AU	1	0.590											
	TOTAL	33	0.590											

	[a]	9	0.295 0.295											
L639	GA-GA GA-AA GA-UA UA-GA AA-GA GG-AA TOTAL [a]	34 7 25 14 2 2 1 39 1	88.74 7 6.394 3.581 0.512 0.512 0.256	CG-GA GC-GA CG-AA GC-AA TOTAL [a]	24 6 4 3 37	64.86 5 16.21 6 10.81 1 8.108		L2847	UU-GA UU-GU UG-GA GA-GA UU-UG GC-AA UU-GG CG-GA UU-AA UG-GG UG-UG UG-GU UA-GA GA-AA UU-UA UA-UG CG-UG GC-GA CG-GC UC-GA UC-UG UU-AC UA-GG TOTAL [a]	74 69 50 20 12 10 7 7 6 6 5 5 4 3 3 3 3 1 1 1 1 1 1 29 3	25.25 6 23.54 9 17.06 5 6.826 4.096 3.413 2.389 2.389 2.048 2.048 1.706 1.706 1.365 1.024 1.024 1.024 1.024 0.341 0.341 0.341 0.341 0.341 0.341	CG-GA TOTAL [a]	32 32	100

The data were obtained from the available rRNA alignments[2]. For all cases of AGPM, the *E. coli* numbering is used.

a: For the statistics, only those cases where the identities of all four nucleotides are known have been considered.

b: N is the number of sequences which have the corresponding -2 base pairs combination.

References

1. Mokdad, A., Krasovska, M. V., Sponer, J. & Leontis, N. B. (2006). Structural and evolutionary classification of G/U wobble basepairs in the ribosome. *Nucleic Acids Res.* **34**, 1326-1341.

2. Wuyts, J., Perriere, G. & Van De Peer, Y. (2004). The European ribosomal RNA database. *Nucleic Acids Res.* **32**, D101-103.

Appendix 4 : Table of the statistical spectrum of the identities of the -1 base pairs for all known ribosomal AGPMs.

	base pairs [-1P;-1Q]- [-1R;-1S]	Eubacteria		base pairs [-1P;-1Q]- [-1R;-1S]	Archaebacteria			AGPM	base pairs [-1P;-1Q]- [-1R;-1S]	Eubacteria		base pairs [-1P;-1Q]- [-1R;-1S]	Archaebacteria	
		N [b]	%		N [b]	%				N [b]	%		N [b]	%
S62	CG-CG	8900	82.30	CG-CG	281	53.62		L657	CG-CG	393	98.992	CG-CG	37	100
	CG-UA	1676	1	CG-UA	218	6			CG-GG	2	0.504	TOTAL [a]	37	
	CG-GC	55	15.49	CG-AU	10	41.60			CG-UA	1	0.252			
	CG-AU	30	8	CG-GC	6	3			CA-CG	1	0.252			
	UA-CG	28	0.509	CU-CG	4	1.908			TOTAL [a]	397				
	CU-CG	18	0.277	CA-CG	2	1.145								
	CC-CG	16	0.259	CG-UC	1	0.763								
	CG-UG	12	0.166	UG-UA	1	0.382								
	UG-CG	10	0.148	GG-CG	1	0.191								
	GG-CG	8	0.111	TOTAL [a]	524	0.191								
	CG-CA	7	0.092			0.191								
	CG-GG	6	0.074											
	CG-CC	6	0.065											
	CA-CG	6	0.055											
	CG-GU	5	0.055											
	GG-UA	5	0.055											
	CG-CU	3	0.046											
	CG-UC	3	0.046											
	UG-UA	2	0.028											
	AG-CG	2	0.028											
	CG-AG	2	0.018											
	CU-UA	2	0.018											
	AG-UA	2	0.018											
	UA-UA	2	0.018											
	UA-AG	1	0.018											
	CC-GC	1	0.018											
	GG-UG	1	0.009											
	GG-AU	1	0.009											
	UG-CA	1	0.009											
	CA-UA	1	0.009											
	CC-UA	1	0.009											
	UG-CC	1	0.009											
	TOTAL [a]	10814	0.009											
			0.009											
S296	GC-GC	3533	44.79	CG-GC	242	58.59		L839	GC-CG	264	66.667	GC-CG	36	100
	GC-AU	3521	0	CG-CG	47	6			CG-CG	131	33.081	TOTAL [a]	36	
	GU-AU	309	44.63	UA-GC	39	11.38			AG-CG	1	0.253			
	AU-AU	126	7	CG-GU	25	0			TOTAL [a]	396				
	AU-GC	123	3.917	GC-GC	18	9.443								
	CG-AU	75	1.597	GC-CG	15	6.053								
	GC-GU	52	1.559	AU-GU	10	4.358								
	GC-CG	37	0.951	AU-GC	7	3.632								
	CG-GC	28	0.659	GC-GU	3	2.421								
	GC-AC	21	0.469	CC-CG	1	1.695								
	GU-GC	16	0.355	CG-CC	1	0.726								
	GC-UU	12	0.266	CG-AU	1	0.242								
	GC-UC	10	0.203	UG-AU	1	0.242								
	GU-GU	3	0.152	AG-GC	1	0.242								
	UA-GU	3	0.127	UA-CC	1	0.242								
	GG-AU	2	0.038	UA-AU	1	0.242								
	CG-CC	2	0.038	TOTAL [a]	413	0.242								
	GC-CC	2	0.025			0.242								
	GC-CU	2	0.025											
	CG-UC	1	0.025											
	CG-AC	1	0.025											
	AC-AU	1	0.013											
	AU-AC	1	0.013											
	GC-GA	1	0.013											
	CC-GC	1	0.013											
	GG-GC	1	0.013											
	AC-GC	1	0.013											
	GC-AA	1	0.013											
	GC-AG	1	0.013											
	UA-GC	1	0.013											
	TOTAL [a]	7888	0.013											
			0.013											
S549	CG-CG	8888	95.41	CG-CG	501	98.62		L1864	CG-GC	332	84.051	Replaced by GNRA [1]		
	UA-CG	185	6	GC-CG	3	2			GC-GC	61	15.443			
	CG-UA	71	1.986	GG-CG	1	0.591			GG-GC	1	0.253			
	UG-CG	35	0.762	UG-CG	1	0.197			CG-AC	1	0.253			

	CC-CG	24	0.376	CA-CG	1	0.197			TOTAL [a]	395				
	CA-CG	24	0.258	CU-CG	1	0.197								
	GG-CG	23	0.258	TOTAL [a]	508	0.197								
	CG-CC	13	0.247											
	AG-CG	12	0.140											
	CG-CA	9	0.129											
	CG-UG	6	0.097											
	CG-CU	5	0.064											
	CU-CG	4	0.054											
	CG-AU	3	0.043											
	UC-CG	2	0.032											
	CG-GG	2	0.021											
	CC-CA	1	0.021											
	UA-CU	1	0.011											
	CA-CC	1	0.011											
	CA-CA	1	0.011											
	GG-CU	1	0.011											
	CG-AG	1	0.011											
	CC-CC	1	0.011											
	UA-GG	1	0.011											
	AA-CG	1	0.011											
	TOTAL [a]	9315	0.011											
S757	GC-AU	4200	35.24	CG-CG	454	78.14		L229	CG-UA	271	68.608	CG-CG	24	68.571
	GC-GC	3532	7	CG-UA	120	1		1	CG-CG	91	23.038	CG-GC	5	14.286
	GC-UA	3208	29.64	CG-AU	3	20.65			GC-UA	17	4.304	UA-CG	3	8.571
	GC-CG	551	1	UG-CG	1	4			CG-AU	10	2.532	GC-CG	2	5.714
	AU-AU	170	26.92	GC-CG	1	0.516			UA-CG	4	1.013	CG-UA	1	2.857
	AU-UA	66	2	CG-GC	1	0.172			GC-CG	2	0.506	TOTAL [a]	35	
	GC-GU	22	4.624	CG-CA	1	0.172			TOTAL [a]	395				
	AU-GC	17	1.427	TOTAL [a]	581	0.172								
	CG-GC	13	0.554			0.172								
	GC-AC	13	0.185											
	GU-AU	9	0.143											
	CG-UA	7	0.109											
	CC-UA	7	0.109											
	GC-CA	7	0.076											
	GC-UU	7	0.059											
	GC-CC	6	0.059											
	GC-UC	6	0.059											
	CG-AU	5	0.059											
	UC-GC	5	0.050											
	GC-AA	5	0.050											
	AC-GC	5	0.042											
	GC-UG	5	0.042											
	GA-GC	4	0.042											
	GU-GC	4	0.042											
	GA-UA	4	0.042											
	GC-CU	4	0.034											
	GC-GA	3	0.034											
	GG-UA	3	0.034											
	GU-UA	3	0.034											
	GC-GG	3	0.025											
	AC-AU	3	0.025											
	UU-CG	3	0.025											
	GG-AU	2	0.025											
	CC-AU	2	0.025											
	AC-UA	2	0.025											
	GC-AG	2	0.017											
	UC-UA	1	0.017											
	GA-CG	1	0.017											
	CG-CG	1	0.017											
	CC-GC	1	0.008											
	GG-GC	1	0.008											
	GA-AU	1	0.008											
	UC-AU	1	0.008											
	UA-CG	1	0.008											
	TOTAL [a]	11916	0.008											
			0.008											
			0.008											
S911	GC-CG	8259	98.14	GC-CG	385	99.74		L268	CG-UA	287	87.234	CG-CG	26	70.270
	GC-UA	57	6	GC-GC	1	1		7	CG-AU	23	6.991	CG-UA	11	29.730
	GC-CA	18	0.677	TOTAL [a]	386	0.259			CG-CG	18	5.471	TOTAL [a]	37	
	GC-CU	12	0.214						CG-UG	1	0.304			
	AC-CG	11	0.143						TOTAL [a]	329				
	GC-CC	11	0.131											
	GU-CG	9	0.131											
	GC-GC	7	0.107											
	GC-GG	7	0.083											
	GC-UG	6	0.083											

	CC-CG UC-CG GG-CG GA-CG GC-AG GC-AU AC-CA GC-UC TOTAL [a]	5 3 3 2 2 1 1 1 8415	0.071 0.059 0.036 0.036 0.024 0.024 0.012 0.012 0.012											
L554	CG-CG CG-UA GC-CG AU-CG UG-UA AU-UG UA-UA TOTAL [a]	204 110 15 10 1 1 1 342	59.64 9 32.16 4 4.386 2.924 0.292 0.292 0.292	CG-CG CG-UA TOTAL [a]	31 2 33	93.93 9 6.061		L269 8	CG-UA CG-GC CG-AU CG-GU UA-AU UA-UA UA-GC CG-CG CG-GA TOTAL [a]	155 75 62 20 7 6 2 2 1 330	46.970 22.727 18.788 6.061 2.121 1.818 0.606 0.606 0.303	CG-GC CG-CG UA-CG TOTAL [a]	26 8 3 37	70.270 21.622 8.108
L639	CG-CG CG-CA CG-GG CU-GG CG-CC TOTAL [a]	385 5 2 1 1 394	97.71 6 1.269 0.508 0.254 0.254	CG-CG TOTAL [a]	37 37	100		L284 7	CG-GC GC-GC CG-AU CG-CG UA-GC CG-GU UG-CG UA-AU TOTAL [a]	235 20 18 10 6 2 1 1 293	80.205 6.826 6.143 3.413 2.048 0.683 0.341 0.341	CG-GC UA-GC CG-AU TOTAL [a]	21 9 2 32	65.625 28.125 6.250

The data were obtained from the available rRNA alignments[2]. For all cases of AGPM, the *E. coli* numbering is used.

a: For the statistics, only those cases where the identities of all four nucleotides are known have been considered.

b: N is the number of sequences which have the corresponding -1 base pairs combination.

References

1. Mokdad, A., Krasovska, M. V., Sponer, J. & Leontis, N. B. (2006). Structural and evolutionary classification of G/U wobble basepairs in the ribosome. *Nucleic Acids Res.* **34**, 1326-1341.

2. Wuyts, J., Perriere, G. & Van De Peer, Y. (2004). The European ribosomal RNA database. *Nucleic Acids Res.* **32**, D101-103.