UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTÉ

UPR 9002 – Architecture et réactivité de l'ARN (IBMC)

THÈSE présentée par :

Luigi D'ASCENZO

soutenue le : **29 Septembre 2016**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/Spécialité : Biophysique/Biochimie structurale

Etude des réseaux de reconnaissance biomoléculaire à l'échelle atomique pour les systèmes ARN et ARN/protéines

—

Atomic-scale investigation of recognition networks in RNA and RNA/protein systems

THÈSE dirigée par le:

Dr. AUFFINGER Pascal	IBMC et Université de Strasbourg

RAPPORTEURS :

Prof. LEONTIS Neocles	Bowling Green State University (Ohio, USA)
Prof. BUSSI Giovanni	Scuola Internazionale Superiore di Studi Avanzati (Trieste, Italie)

AUTRES MEMBRES DU JURY :

Dr. WEIXLBAUMER Albert	IGBMC et Université de Strasbourg

INVITÉ :

Prof. WESTHOF Eric	IBMC et Université de Strasbourg

Acknowledgments

First, I would like to acknowledge my thesis supervisor Dr. Pascal Auffinger for his tireless support, inspiring help and his hidden oceans of patience. He has been a marvelous teacher and friend, to the point I consider him a *scientific father*. He has disclosed to me the ways to be a (better) scientist and with his unique sensibility and elegance also a better human being.

A special and heartfelt *grazie* goes to my family, my mother Ida, my father Nino, my sister Carmen and my grandmother Carmela. No matter the circumstances they were the rock solid reference point to always find the way in the middle of the storms I had to face during my life. They have been a Polar Star during nights and my Sun during days.

My deepest gratitude goes to Dr Filip Leonarski and Dr Quentin Vicens. Filip for helping me to deal with MD simulations, where I realized how bad my coding skills are compared to his amazing abilities; further, I thank him for engaging in endless discussions about Science and trains. I wish him all the best for his career and for the forthcoming fifth member of his family. Quentin has been my *scientific bigger brother*, showing me the ways of a decent writer and a thoughtful thinker. I will never forget his daily appearances in our small office, when he came smiling and bringing brand new terrific ideas about our projects.

Special thanks go to Davide “Demon” Porreca, Emanuela “Morwen” Palombo and Annarita “Eiddynil” Vernile. I shared with them the dragons that constantly fly through my head and found in response exquisite friends that will always reside in a precious place of my heart. Without them, I would be a poorer person and probably a worse scientist.

I would like to acknowledge the fundamental role of Prof. Eric Westhof and Prof. Richard Giegé for the success of my work so far. It would take another separate thesis to write down all the advices, suggestions and scientific tricks that I learned from them. For sure, I will jealously keep them among my most valuable secrets.

I would like to acknowledge also Dr Eric Ennifar, the head of our research group, for making me more aware of the pitfalls of X-ray crystallography and for the formative role of our Tuesday after-lunch group meetings. Along with him, I thank Dr Chichau Miao and all the members of the UPR 9002 at the IBMC in Strasbourg that at one point or the other had an impact on my (scientific) life.

A big thanks also to my friends Nicola Ricciuti, Francesca Romana Rizzo, Martina Marcoccio, Davide Marcoccio and Manuele Spanu who filled the holes of my non-scientific life with diverse discussions, boardgames and abundant doses of shared pizza.

I would like to express my sincere gratitude to Prof. Neocles Leontis, Prof. Giovanni Bussi and Dr Albert Weixlbaumer for accepting to be part of my PhD committee. Among them, a special mention to Prof. Leontis (and his wonderful wife Vassiliki) for giving me the opportunity to work for a month in his group in Bowling Green, Ohio.

Additionally, I thank all the people that do not appear in this short list but for many reasons have been responsible for the successful completion of my thesis. I tried to learn, among other skills, to be scientifically concise (see for instance this *concise* 300 pages-long manuscript!) and for this sole reason they are not listed here.

Summary (English)

Recognition networks in RNA and RNA/protein systems are modulated by numerous non-covalent interactions. While hydrogen bonds are well described, less common and often overlooked non-covalent interactions coexist with them. During my thesis, I focused on the structural and functional implications of several of these non-covalent interactions in nucleic acids. To that end, I used computational techniques such as quantum mechanical calculations, molecular dynamics (MD) simulations, as well as screening of chemical structural databases of small molecules (CSD: Cambridge Structural Database) and larger biomolecular structures (PDB).

The main topic of my research is related to the study of a general class of these non-covalent interactions within RNAs, namely the stacking of backbone oxygen atoms with nucleobases to form "oxygen-π" contacts. This stacking occurs in two forms where a nucleobase stacks with: (i) an anionic phosphate oxygen atom; (ii) a ribose O4' atom. Through a survey of PDB crystallographic structures, I verified that the former contacts, belonging to the family of "anion-π" interactions, are recurrently part of the signature of GNRA tetraloops and tRNA anticodon loops. These contacts are of a "capping" type and result from the sharp turn inherent to these loops, incidentally protecting a nucleobase from solvent. I showed by MD simulations that strongly bound (long-lived) water molecules belonging to phosphate first solvation shell are present in these loops. I further established that anion-π interactions do not occur in proteins, while cation-π do occur only in proteins but not in nucleic acids and linked these observations to the electrostatic potentials of their aromatic groups, which differ in proteins and nucleic acids. This finding has fundamental implications on how we perceive stacking interactions in biomolecular systems.

The alternative oxygen-π stacking, involving O4' atoms, occurs in the UNCG tetraloops family and caps a nucleobase in a similar manner. This led us to reconsider the global structural signature of all tetraloops by integrating these two rare non-covalent interaction types. Hence, we suggest that it is possible to categorize all tetraloops in two basic families of turns instead of more than ten as deduced in the literature. Additionally, the widely established correlation between a loop sequence and its structure is not always straightforward, as we highlighted with several ribosomal loops that adopt

unexpected folds, displaying that our capabilities of predicting 3D folds from sequences are still far from perfection. In parallel we discovered that two UNCG-fold tetraloops inside the ribosome are involved in specific long-range RNA interactions, which prompted us to describe UNCG receptors that result from the assembly of several distant structural motifs in opposition to the well-known GNRA receptors that are double strand motifs. As an outcome, UNCG receptors can probably only occur in very large structures of the rRNA size and can give insights on ribosome biogenesis. We also showed that UNCG tetraloops contain a CpG step where a *C2'-endo* pucker (for the cytosine) and a *syn* guanine coexist. This step adopts a "Z-like" conformation that is similar to the CpG dinucleotide steps found in Z-DNA. Although rare, these motifs are present at key locations in RNA riboswitches, ribozymes, aptamers and ribosomes and are relevant for folding, protein binding and immune response. This finding led to the conclusion that Z-conformations are not only specific to DNA but are present at various locations in RNA with functions yet to discover.

Besides non-covalent interactions, biomolecular systems need the assistance of solvent to form and maintain their active folds. Nucleic acids are no exception to that. We wrote two reviews to complement our knowledge of the intricate relationship nucleic acid structures entertain with the solvent. The first addresses the roles monovalent cations play in nucleic acid systems and emphasizes that the hardly biologically relevant Na^+ ion that is predominantly used in *in vitro* experiments often overshadows the more biologically relevant K^+ ion. The second review establishes that anions, frequently used in crystallographic buffers, can interact with nucleotides and that some of them (sulphates, carboxylate containing anions, …) can adopt unexpected protonation states. Therefore, their incidence has to be more carefully considered in solvent attribution procedures. Indeed, we found that Asp and Glu amino acid side chains are sometimes in close contact to phosphate groups and are consequently protonated. In order to better understand these interactions, we surveyed the CSD and classified all interaction types of carboxyl(ate) groups with themselves. We found that "carboxyl" groups can participate in very short hydrogen bonds (≈ 2.5 Å) while "carboxylate" groups show standard hydrogen bond lengths (≈ 2.8 Å). Such short hydrogen bonds involving carboxyl-carboxylate pairs are frequent in proteins. I participated also in the analysis of Mg^{2+} interactions with nucleic acids through a PDB survey. Mg^{2+} ions are essential for nucleic acid structure and function, but their identification in databases is often not reliable. We first analyzed the binding of Mg^{2+} ions to nucleobase imine atoms to infer a set of rules that can promote a more reliable identification of these important ions and other solvent molecules.

The structure and properties of solvent surrounding RNA were also studied by MD simulations at different temperatures. The rationale behind this study are related to gaining a better understanding on how RNA systems are affected by temperature changes at the level of their interactions with the solvent at a pre-melting stage. These solvent interactions and their variation in temperatures are probably highly sequence dependent as shown by preliminary results on RNA duplexes. We developed this methodology to investigate the stability of tetraloops when their sequence is altered and consequently, the stability of the oxygen-π stacking interactions.

To summarize, the results of my PhD work complement our knowledge of several uncommon non covalent interactions in RNA and RNA/protein systems. They are significant for RNA structure and function as well as for improving our understanding of biomolecular recognition networks. As a perspective, my work aims to help structural determination with techniques such as x-ray crystallography and computational methods, as well as elucidating biologically relevant mechanisms related to immune response and RNA folding.

Resumé (Français)

Les réseaux de reconnaissance biomoléculaire dans les systèmes ARN et ARN/protéines sont modulés par de nombreuses interactions non-covalentes. Alors que les liaisons hydrogène sont bien décrites, des interactions non-covalentes rares et souvent négligées coexistent avec elles. Au cours de ma thèse, je me suis concentré sur les implications structurales et fonctionnelles de plusieurs de ces interactions non-covalentes dans les acides nucléiques. À cette fin, j'ai utilisé des techniques informatiques telles que des calculs de mécanique quantique, de simulations de dynamique moléculaire (DM), ainsi que l'analyse de bases de données structurales de petites molécules (CSD: Cambridge Structural Database) et des structures biomoléculaires (PDB).

Le sujet principal de ma recherche est lié à l'étude d'une classe particulière de ces interactions non-covalentes au sein des ARN, le « stacking » ou empilement des atomes d'oxygène du squelette phosphaté avec des nucléobases format des interactions « oxygen-π ». Cet empilement se présente sous deux formes, un empilement de nucléobases avec : (i) un atome anionique d'oxygène d'un groupement phosphate ; (ii) un atome O4' du ribose. Par une analyse des structures cristallographiques de la PDB, j'ai vérifié que les premières interactions, du type « anion-π », font partie de la signature des tétraboucles de type GNRA et de la boucle anticodon dans les ARN de transfert. Ces interactions sont de type « capping » et résultent du tournant du squelette phosphaté inhérent à ces boucles. Par ailleurs, elles protégeant une nucléobase du solvant. J'ai montré par des simulations de DM que des molécules d'eau présentes dans la première couche d'hydratation des phosphates de ces boucles sont fortement liées et montrent temps de résidence élevés. J'ai aussi établi que les interactions anion-π ne se produisent pas dans les protéines. D'un autre côté, les interactions cations-π peuvent apparaitre dans les protéines mais pas dans les acides nucléiques. J'ai lié ces observations aux potentiels électrostatiques des groupes aromatiques qui diffèrent dans les protéines et les acides nucléiques. Cette constatation a des implications fondamentales sur la façon dont nous percevons les interactions d'empilement dans les systèmes biomoléculaires.

L'empilement « oxygen-π » impliquant des atomes O4' apparait dans les tétraboucles UNCG et protège une nucléobase d'une manière similaire aux interactions « anion-π » décrites plus haut. Cela

nous a conduits à reconsidérer la signature structurale globale de toutes les tétraboucles en intégrant ces deux types d'interactions non-covalentes rares. Nous suggérons qu'il est possible de classer tous les tétraboucles en seulement deux familles au lieu de plus de dix comme décrit dans la littérature. En outre, la corrélation établie entre une séquence et sa structure de boucle n'est pas toujours simple, comme nous l'avons souligné pour plusieurs boucles ribosomiques qui adoptent des structures inattendues, montrant que nos capacités de prédiction de structures 3D sont encore perfectibles. En parallèle, nous avons découvert que deux tétraboucles adoptant des repliements de type UNCG au sein du ribosome sont impliquées dans des interactions spécifiques d'ARN à longue distance, ce qui nous a permis de décrire les récepteurs UNCG qui résultent de l'assemblage de plusieurs motifs structuraux éloignés en opposition aux récepteurs GNRA qui impliquent des motifs en double brin. Ainsi, les récepteurs UNCG peuvent apparaître dans de très grandes structures de la taille des ARN ribosomaux et peuvent nous donner des indications sur la biogenèse du ribosome. Nous avons également montré que les tétraboucles UNCG contiennent un pas CpG où un plissement de type *C2'-endo* et une guanine en *syn* coexistent. Ce pas adopte une conformation « Z-like » qui est semblable à ceux formant des doubles hélices d'ADN de forme Z. Bien que rares, ces motifs sont présents à des endroits clés dans les riboswitches, ribozymes, aptamères et ribosomes et sont importants pour le repliement, la reconnaissance des protéines, et la réponse immunitaire. Cette constatation a conduit à la conclusion que les « conformations-Z » ne sont pas seulement spécifique à l'ADN, mais sont présents à divers endroits dans l'ARN avec des fonctions encore à découvrir.

Outre que les interactions non-covalentes, les systèmes biomoléculaires ont besoin du solvant pour former et maintenir leurs plis actifs. Les acides nucléiques ne font pas exception. Nous avons écrit deux articles de revues pour compléter nos connaissances sur la relation complexe des structures d'acide nucléique avec le solvant. La première adresse les rôles que les cations monovalents jouent dans les systèmes d'acides nucléiques et souligne que l'ion Na^+, peu relevant au niveau intracellulaire, est principalement utilisé dans les expériences *in vitro* en remplacement l'ion K^+ qui est dominant dans les milieux intracellulaires. Le deuxième article de revue établit que les anions, fréquemment utilisés dans les solutions cristallographiques, peuvent interagir avec les nucléotides et que certains d'entre eux (sulfates, anions contenant des groupements carboxylate, ...) peuvent adopter des états de protonation inattendus. Par conséquent, leur incidence doit être considérée avec plus d'attention dans les procédures d'attribution du solvant. En effet, nous avons constaté que les chaînes latérales d'acides aminés Asp et Glu sont parfois en contact avec des groupes phosphate et sont par conséquent protonées. Afin de mieux comprendre ces interactions, nous avons exploré la CSD et classé tous les types d'interaction des groupes carboxyl(ates) avec eux-mêmes. Nous avons constaté que les groupes « carboxyle » peuvent participer à des liaisons hydrogène très courtes (≈ 2,5 Å) tandis que les groupes « carboxylate » montrent des longueurs de liaison hydrogène standard (≈ 2,8 Å). Ces liaisons courtes impliquent des paires carboxyle-carboxylate et sont fréquents dans les protéines. J'ai également participé à l'analyse des interactions entre ions Mg^{2+} et acides nucléiques. Les ions Mg^{2+} sont essentiels pour la structure et la fonction des acides nucléiques, mais leur caractérisation dans les bases de données est souvent problématique. Nous avons analysé d'abord les interactions des ions Mg^{2+} avec azotes de type « imine » des nucléobases pour déduire un ensemble de règles qui peuvent favoriser une identification plus fiable de ces ions dans les structures cristallographiques de manière à mieux cerner leur rôle biologique et éviter les erreurs d'interprétations encore trop fréquentes.

La structure et les propriétés du solvant entourant l'ARN ont également été étudiées par des simulations de DM à différentes températures. Cette étude a visé à une meilleure compréhension de la façon dont les systèmes d'ARN sont affectés par les changements de température au niveau de leurs interactions avec le solvant à un stade de « pré-melting ». Ces interactions du solvant et leur variation avec la température sont probablement très dépendantes de la séquence comme le montrent les résultats préliminaires sur des duplexes d'ARN. Nous avons développé cette méthode afin d'étudier la stabilité des tétraboucles lorsque leur séquence est modifiée et donc la stabilité des interactions « oxygen-π ».

Pour résumer, les résultats de mon travail de thèse complètent notre connaissance de plusieurs interactions non-covalentes rares dans les systèmes ARN et ARN/protéines. Ils sont importants pour mieux comprendre la structure et la fonction des différents ARN présents dans les systèmes cellulaires du vivant, ainsi que pour améliorer notre compréhension des réseaux de reconnaissance biomoléculaire. Comme perspective, mon travail a produit des données qui aideront à la détermination structurale de systèmes ARN avec des techniques telles que la cristallographie aux rayons X et les méthodes informatiques, ainsi que l'étude des mécanismes liés à la réponse immunitaire et le folding d'ARN.

Table of contents

List of Figures

List of tables

Abbreviations/Glossary

A	Adenine
ADAR1	Double strand ribonucleic acid adenosine deaminase
C	Cytosine
CRISPR	Clustered regularly interspaced short palindromic repeats
Cryo-EM	Cryo-electron microscopy
CSD	Cambridge Structural Database
DNA	Deoxyribonucleic acid
DSSR	Dissecting the spatial structure of RNA
EDS	Electron Density Sever
EG	Ethylene glycol
ESP	Electrostatic potential
FRET	Fluorescence resonance energy transfer
G	Guanine
lncRNA	Long non-coding ribonucleic acid
LSU	Large ribosomal subunit
MD	Molecular dynamics
miRNA	Micro ribonucleic acid
mmCIF	Macromolecular crystallographic information file
mRNA	Messenger ribonucleic acid
ncRNA	Non-coding ribonucleic acid
NMR	Nuclear magnetic resonance
PEG	Polyethylene glycol
PDB	Protein Data Bank
RNA	Ribonucleic acid
RNAi	Ribonucleic acid interference
rRNA	Ribosomal ribonucleic acid
SPC/E	Extended simple point charge water model

SSU	Small ribosomal subunit
T	Thymine
tRNA	Transfer ribonucleic acid
$tRNA^{Phe}$	Phenylalanyl transfer ribonucleic acid
U	Uridine
WC	Watson-Crick
Ψ/PSU	Pseudouridine

Thesis overview

This manuscript contains a detailed overview of the work conducted during my thesis between September 2013 and September 2016. I worked under the supervision of Dr Pascal Auffinger at the IBMC – Strasbourg (France), within the UPR 9002 “Architecture et réactivité de l’ARN” directed by Prof. Eric Westhof. The first part of the manuscript contains introductory notions on RNA function and structure, as well as non-covalent interactions and environmental effects that participate in maintaining biomolecular recognition networks. In the following part, I detail some of the procedures and methods utilized for the purpose of studying RNA and RNA-protein systems at the atomic scale. The body of the thesis is split into several sections, with published and unpublished results presented in a logical rather than chronological order.

The first topic is the study of “uncommon” non-covalent interactions in RNA, namely oxygen-π stacking between backbone oxygen atoms and nucleobases. Anion-π and cation-π interactions in nucleic acids and proteins are analyzed in the **Paper 1**:

❖ D’Ascenzo L., Leonarski F. and Auffinger P. Cation-π versus anion-π interactions — biomolecules can’t have both. In preparation.

The stacking between O4’ ribose atoms and nucleobases follows, with an emphasis on its occurrence in “Z-DNA like” fragments within RNA. This topic is discussed in the **Paper 2**:

❖ D’Ascenzo L., Leonarski F., Vicens Q. and Auffinger P. “Z-DNA like” fragments in RNA: a recurring structural motif with implications for folding, RNA/protein recognition and immune response. *Nucleic Acid Res.* **2016**, *44*, 5944-56 (PubMed).

In the following section tetraloop motifs are analyzed in respect of their structural signature and the tertiary interactions in which they take part. A novel perspective in defining tetraloop folds is presented in **Paper 3**:

❖ D’Ascenzo L., Leonarski F., Vicens Q. and Auffinger P. Revisiting GNRA and UNCG folds: U-turns versus Z-turns in RNA hairpin loops. In press in *RNA* (pdf).

The section is closed by a discussion on the novel UNCG tetraloop receptors found inside the ribosome.

Environmental effects are the argument of the next section, mostly studied via the binding of ionic species to nucleic acids and MD simulation with variable temperatures. The interactions of monovalent sodium and potassium ions with nucleic acids is the argument of **Review 1**:

❖ Auffinger P., D'Ascenzo L. and Ennifar E. Sodium and potassium interactions with nucleic acids. *Met. Ions Life Sci.* **2016**, *16*, 167-201 (PubMed).

The binding of magnesium to nucleic acids and the issues encountered in its structural analysis are discussed in **Paper 4** and **Paper 5**:

❖ Leonarski F., D’Ascenzo L. and Auffinger P. Binding of metals to purine N7 nitrogen atoms and implications for nucleic acids: a CSD survey. *Inorg. Chim. Acta* **2016***, 452,* 82-9 (Link).

❖ Leonarski F., D’Ascenzo L. and Auffinger, P. Mg^{2+} ions: do they bind to nucleobase nitrogens? In press in *Nucleic Acid Res* (Link).

Conversely, the binding of anions to nucleic acids and their usage in crystallography are discussed in **Review 2**:

❖ D’Ascenzo L. and Auffinger P. Anions in nucleic acid crystallography. *Methods Mol. Biol.* **2016**, *1320*, 337-51 (PubMed).

The following chapters are devoted to aspartate and glutamate interactions with nucleic acids. A broader perspective on carboxyl-carboxylate interactions is presented in **Paper 6**:

❖ D’Ascenzo L. and Auffinger P. A comprehensive classification and nomenclature of carboxyl-carboxyl(ate) supramolecular motifs and related catemers: implications for biomolecular systems. *Acta Cryst.* **2015**, *B71*, 164-75. (PubMed)

The section concludes with preliminary data of MD simulations on RNA to assess temperature effects on its structure. The manuscript is closed by conclusive remarks, with a perspective on forthcoming MD studies on environmental effects in crowding conditions.

1. Introduction

1.1 Structure-function relationship in biology

Our understanding of living entities is associated with the intimate correlation between structure and function, common to all levels of biological organization. The thin and hollow bones of the birds skeleton lighten their weight, making flying easier, as the unique and still puzzling structure of biological membranes allow cells to exist as we know them. Since the beginning of natural philosophy, early investigators and scientists were fascinated by the concept that the behavioral repertoire of biological beings was *orchestrated from within* (Fernald 2011). This inner force is directly rooted in their structure, a notion expressed in a sublime way by D. A. Spalding in 1873 (Spalding 1873).

> *When, as by a miracle, the lovely butterfly bursts from the chrysalis full-winged and perfect, and flutters off a thing of soft and gorgeous beauty, it but wakes to a higher life, to a new mode of existence, in which, strange though it may sound, it has, for the most part, nothing to learn; because its little life flows from its organization like melody from a music box*.
>
> - *Douglas A. Spalding, 1873* -

We can expand this concept stating that at every level of organization of life functions and properties of biological entities *flow from their organization like melody from a music box*. The same notion applies to biological macromolecules, and among them ribonucleic acids (RNA), which structures embed from their synthesis the potential to achieve very specific and unique functions. This emphasizes the importance of structural studies on biomolecules. In the past decades the structural investigations have been focused on the heterogeneity of biomolecules (especially RNA), showing that they can be considered “modular” systems (Wagner et al. 2007). This concept is associated with the partitioning of macromolecular structures into distinct subunits (or motifs), which contribute to maintain equilibrium states in thermodynamic environments, genetic context and folding kinetics. Even the simplest of those motifs has the potential to combine with others to form large units and to play a specific role within the final biomolecular architecture. However, a less obvious consideration is that many of these structural motifs can behave differently when embedded in large and complex systems compared to when they are isolated. This characteristic can be exploited to generate a great diversity of dynamic building blocks with novel ensembles of biochemical functions. Under this perspective, RNA is structurally and functionally more similar to proteins than to the chemically related DNA. Although RNA structural motifs have been extensively studied and described (Hendrix et al. 2005; Holbrook 2005; Leontis et al. 2006; Butcher and Pyle 2011), there are still overlooked but remarkable characteristics embedded inside their structural signatures (**Sections I** and **II**).

Besides, biomacromolecules operate and interact inside complex cellular environments, in particular physico-chemical conditions that make cells more than a collection of unconnected molecules. Thus, all the structural information are limited without explicit considerations of the context where they are observed, especially for biomolecular recognition. Environmental effects are the *master* of biomolecular structures and interaction networks, generating complex but fascinating biological entities obedient to thermodynamic laws. Overall, the study of these dynamic systems and their fundamental principles is a highly intriguing and rewarding endeavor, as will be shown in **Section III**.

1.2 RNA

1.2.1 An essential biological actor

RNA is, with DNA and proteins, one of the essential component of all known forms of life. It is a common principle of molecular biology that genetic information coded in DNA is transferred through RNA to the final steps of protein synthesis. This *central dogma* was first proposed by Francis Crick in 1956 ("*Ideas on protein synthesis*") and was subsequently summarized as "DNA makes RNA and RNA makes proteins". It was believed that RNA could accomplish only three major roles in the cell: (i) embody a copy of part of the DNA, able to leave the nucleus effectively acting as a messenger, from which the name of messenger RNA (mRNA) was derived; (ii) act as a decoder of the genetic code transcribed in mRNA and amino acids, in the form of transfer RNA (tRNA); (iii) be a structural component of ribosomes (rRNA), playing a fundamental role in the machinery responsible for protein synthesis. However, since 1960 this *monochromatic* picture has been complemented with more colors. In fact, RNA can achieve in the cell a much more interesting ensemble of functions than initially thought (Cech and Steitz 2014). The discoveries made in this fields have been so relevant for biology that numerous authors were awarded with Nobel prizes.

Early RNA molecular and structural biologists focused their investigations on tRNA, because of its importance in decoding, small size (73-93 nucleotides) and availability (Clark 2006). Therefore, it is not surprising that the first breakthrough in the RNA field was the primary structural determination of a tRNA, by R. Holley's group in 1965 (Holley et al. 1965). Later on, at the beginning of 1970s, the tRNA structure was solved by X-ray crystallography by both A. Rich and A. Klug groups (Kim et al. 1973a; Ladner et al. 1975). Concomitantly, in 1970 H. Temin and D. Baltimore independently showed that several RNA tumour viruses contain an enzyme, named reverse transcriptase, that reverse-transcribes the viral RNA genome into DNA, which is then integrated into the host genome and replicated along with it (Baltimore 1970; Temin and Mizutani 1970). The discovery of reverse transcription was the first step of a revision of the classical concept of genetic information flow that happened in the decades following 1960.

Several years later, R. J. Roberts and P. A. Sharp groups showed that genes of higher organisms are mosaics of coding and non-coding sequences, all of which are transcribed to mRNA (Chow et al. 1977; Berget et al. 2000); the non-coding sections of freshly transcribed RNA, called introns, are removed from the transcript and the remaining sequences, called exons, are joined together in the final mature RNA, during a process called RNA splicing. Another major breakthrough in RNA biology happened during 1980s, when T. Cech and S. Altman groups showed that RNAs can also act as enzymes (called ribozymes) for specific biochemical reactions (Kruger et al. 1982; Guerrier-Takada et al. 1983), such as reactions at the core of RNA splicing (Fica et al. 2013) and in the ribosome (that is itself a ribozyme (Steitz and Moore 2003)). It became clear that the so-called non-coding RNAs are more than genetic debris as firstly thought for introns, and can have multiple regulatory roles on many levels from genetic expression to cellular functions (Morris and Mattick 2014). The first regulatory RNA to be identified and sequences has been the *E. coli* 6S RNA in 1971, by G. Brownlee (Brownlee 1971). Following this seminal work, many other regulatory RNAs have been identified, in particular a family of short (~22 nucleotides) fragments, called microRNA (miRNA; (Lee et al. 1993; Reinhart et al. 2000). miRNAs control the expression of numerous genes by base pairing with their mRNA target sequences and are involved in the mechanism of gene silencing through sense-antisense RNA pairing known as RNA interference (RNAi; Fire et al. 1998; Waterhouse et al. 1998). More recently, a new family of regulatory non-coding RNA was found inside the large eukaryotic introns, characterized by elements of more than 200 nucleotide in length called long non-coding RNA (lncRNA) (Mercer et al. 2009). lncRNAs have been involved in the regulation of differentiation, development, epigenetic processes, suspected to participate in aging and play roles in various diseases, thus becoming a very hot research topic (Morris and Mattick 2014). **Fig. 1.1** shows a schematic view of the "central dogma" complemented with discoveries on non-coding RNAs.

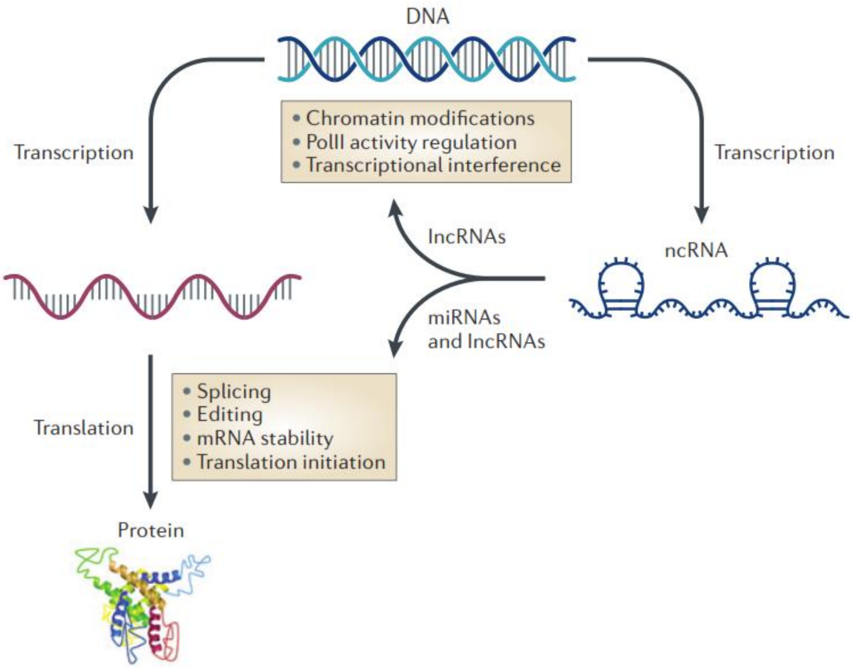


Figure 1.1. 'Central dogma' in the context of regulatory non-coding RNAs. The concept of the "central dogma" has been complemented with aspects of ncRNAs functions. *From:* (Wahlestedt 2013).

More surprising examples of regulatory RNAs come from bacteria, where they are involved in the adaptive response to environmental factors. Riboswitches are mRNA segments that act in *cis* by binding metabolites in order to regulate gene expression thus functioning as "environmental sensors" (Tucker and Breaker 2005; Garst et al. 2011). Particularly interesting is the T-box riboswitch, that regulates intracellular availability of amino acids through binding aminoacylated or free tRNAs (Henkin 2014; Zhang and Ferre-D'Amare 2015). Other small guide RNA are transcribed and processed from virus-derived DNA sequences incorporated in the bacterial genome, termed Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR); the guide RNAs are complexed with effectors such as Cas proteins to specifically target and cleave viral DNA or RNA (Hale et al. 2009; Marraffini and Sontheimer 2010). This system is the base of the CRISPR/Cas9 genome engineering technique (Kapral et al. 2014).

As showed by this synthetic historical perspective, many new RNA biological functions have emerged since the central dogma. In parallel, structural analysis have made significant steps forward since the tRNA structures to elucidate more complex RNA architectures, such as ribosomes. In particular, biochemical and biophysical investigations allow us to appreciate the landscape of structural features connected to the versatility, dynamicity and plasticity of RNA molecules.

1.2.2 A complex architecture made by simple building blocks

Our current structural knowledge of RNA is due to advances in structure determination using X-ray crystallography, nuclear magnetic resonance (NMR) and more recently cryo-electron microscopy (cryo-EM). However, these techniques have limitations for elucidating the dynamics and convey little information about the folding steps between a linear sequence and a complex tertiary architecture (Cruz and Westhof 2009). These gaps are beginning to be filled with several experimental and theoretical biophysical tools, including single-molecule optical traps, time-resolved fluorescence resonance energy transfer (FRET), hydroxyl radical footprinting and molecular dynamics simulations (MD). Although based on approximate models, these techniques enrich our understanding of the remarkable dynamicity shown by biomolecules.

RNA structure, analogously to proteins, can be described at different level of organization, which combined together yield the complex architectures of molecules such as tRNAs and ribosomes (**Fig. 1.2**). At the primary sequence level, RNA is composed of nucleotides made of a backbone of phosphate groups linking ribose units by phosphodiester bonds. Each ribose has a nucleobase attached at the 1' carbon (glycosidic bond). The four RNA nucleobases are adenine (A), cytosine (C), guanine (G) and uracil (U). From a chemical point of view, adenine and guanine are purines, while cytosine and uracil are pyrimidines. There are local variations to this basic scheme, with more than one hundred characterized modified nucleobases (Sun et al. 2016), ranging from isomers such as pseudouridine (Ψ or PSU) to extensively substituted nucleobases such as wybutosine (Motorin and Grosjean 2005). The RNA backbone is rotameric and can be characterized by six main torsion angles (named from 5' to 3' with lower-case Greek letters from *alpha*, α, to *zeta*, ζ) around the covalent bonds, the χ angle of the

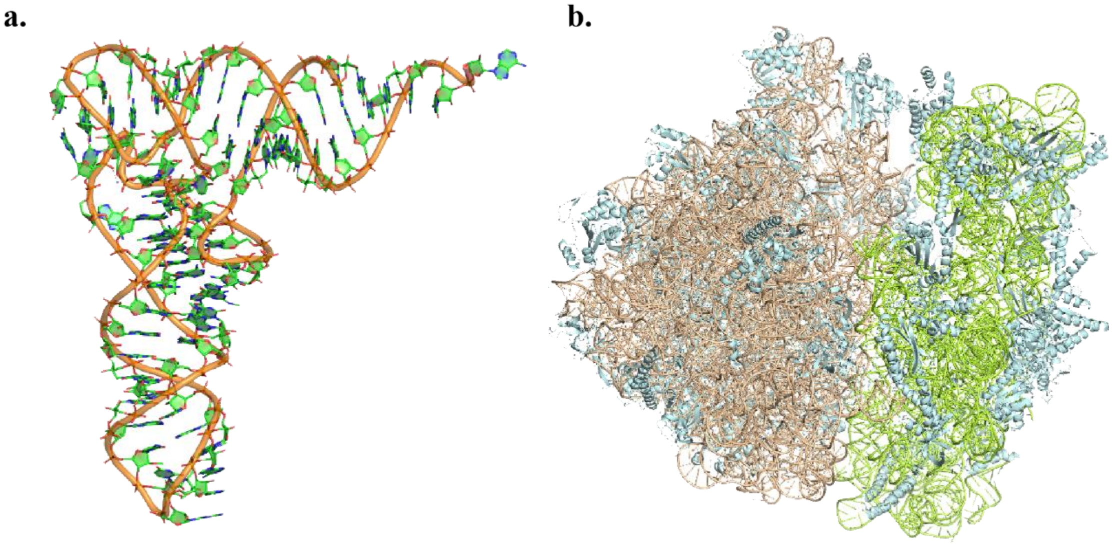


Figure 1.2. Examples of RNA structure. (a) tRNA^{Phe} (PDB: 1EHZ; res.: 1.9 Å; Shi and Moore 2000) **(b)** 70S ribosome from *E. coli* complexed with ribosomal proteins (PDB: 4YBB; res.: 2.1 Å; Noeske et al. 2015).

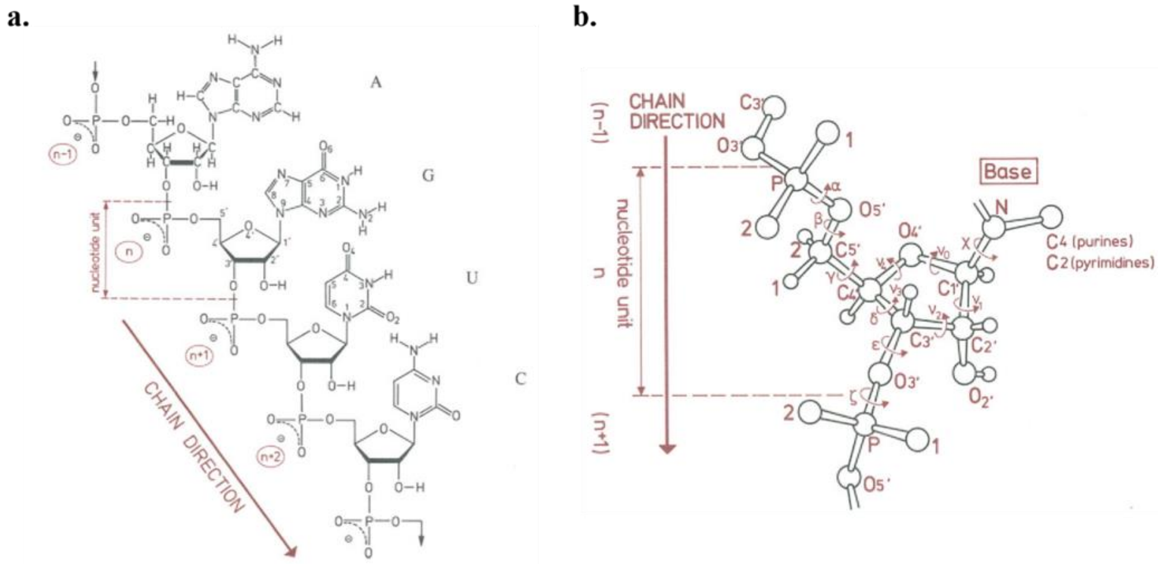


Figure 1.3. RNA nucleobases and backbone conformations. **(a)** RNA strand fragment, with 5’ to 3’ direction shown by an arrow and explicit base atom numbering **(b)** Backbone torsion angles of a nucleotide. *Adapted from:* (Saenger 1984).

glycosidic bond and the sugar pucker (**Fig. 1.3**; Saenger 1984); these parameters generate a great number of accessible stable conformations (Murray et al. 2003). In an attempt to define a consensus classification and nomenclature, 46 discrete clusters of RNA backbone conformers have been identified within RNA structures (Richardson et al. 2008).

The second organization level of RNA is constituted by the Watson-Crick base-paired helical regions, which together define the secondary structure. The scaffold for this structure is provided by non-covalent interactions, mostly base-base stacking and inter-nucleotide hydrogen bonds. Canonical Watson-Crick base pairs, known to promote helical structures, exist together with non-canonical base pairs and hydrogen bonds between nucleobases and the backbone, which promote single strands and other structural motifs (Westhof and Auffinger 2000). N. B. Leontis and E. Westhof proposed a scheme to classify all the possible base pairs, based on the definition of three distinct interacting edges for nucleobases: the Watson-Crick edge (WC), the Hoogsteen edge (H) and the Sugar edge (S, **Fig. 1.4**). This, together with the *cis* or *trans* orientation of glycosidic bonds, originates 12 basic geometric types with at least two hydrogen bonds connecting the bases (Leontis and Westhof 2001; Leontis et al. 2002).

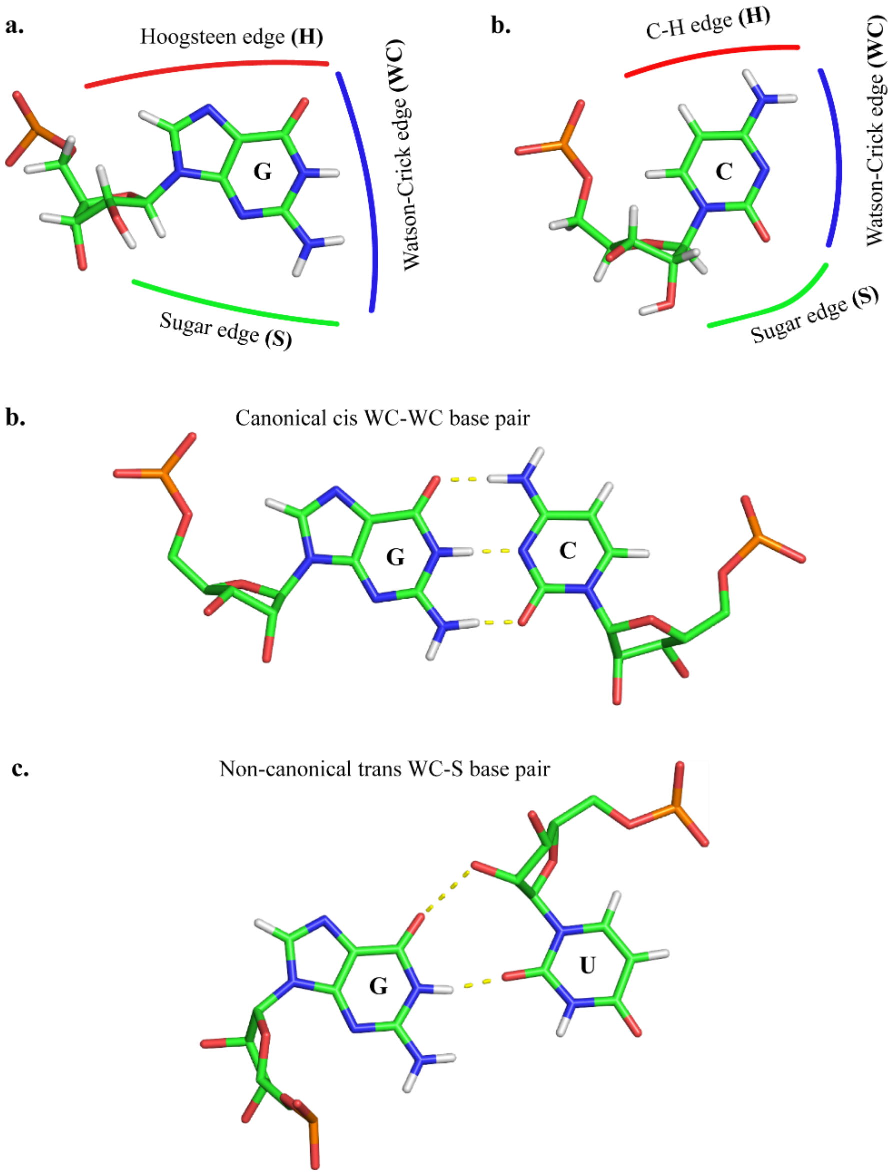


Figure 1.4. Fundamentals of the Leontis-Westhof base pair classification (Leontis and Westhof 2001). **(a)** Edges for purine and pyrimidine bases **(b)** G=C cWW base pair **(c)** G•U tWS base pair.

Appropriate conditions can induce structured RNA molecules to undergo a transition from an unfolded state to a unique 3D fold in which the helices and the unpaired regions are precisely organized in space. This RNA tertiary structure is constituted by secondary elements associated through numerous van der Waals contacts and specific hydrogen bonds (Butcher and Pyle 2011). The formation of additional unusual pairs allows to generate many diverse structural motifs such as hairpin loops. These motifs can be considered as building blocks and at the same time as modules of a higher organization level, which confer to RNA the nature of a modular biomolecule.

1.2.3 A modular biomolecule

During recent decades, the original connotation of "sequence motifs" has extended to structural and tertiary motifs, which have been defined variously as *"directed and ordered stacked arrays of non Watson-Crick base pairs forming distinctive foldings of the phosphodiester backbone of the interacting RNA strands"* (Leontis and Westhof 2003) and *"a discrete sequence or combination of base juxtapositions found in naturally occurring RNAs in unexpectedly high abundance"* (Moore 1999). These descriptions variously capture the nature of RNA structural motifs. They have interested and fascinated structural scientists since the determination of the first RNA structures, namely tRNAs, obtained by X-ray crystallography (Kim et al. 1973b; Ladner et al. 1975). In this structure, the similarity between the anticodon and the TΨC hairpin loop prompted the authors to name them uridine turn or "U-turn", defining a new family of structural motifs (Quigley and Rich 1976). This family has been expanding since, with the inclusion of GNRA tetraloops among others (Jucker and Pardi 1995a). The characteristic geometry of U-turns allows the last three bases of the loop to be stacked and made available for tertiary interactions (**Fig. 1.5a**), as observed in the codon-anticodon recognition and for the long range interactions involved in GNRA receptors (Jaeger et al. 1994; Geary et al. 2008; Fiore and Nesbitt 2013). U-turns constitute stable structural motifs by themselves, but they are simultaneously structural modules or sub-motifs for more complex architectures. Many RNA structural motifs show this multidimensional relevance, providing stability to local structure on one side and maintaining the global fold on the other.

Other examples of RNA structural motifs or sub-motifs (that need to be part of larger assemblies to exist) with multidimensional relevance are: base triples, helices, hairpin loops (other than U-turns), internal loops, bulges, kink-turns, pseudoknots (**Fig. 1.5**). Further instances, such as A-minor and ribose zippers, are only involved in tertiary long range interactions and therefore found especially in larger RNA structures (Cate et al. 1996; Nissen et al. 2001). Altogether, these families constitutes some instances of the still expanding repertoire of motifs found in RNA structures (Hendrix et al. 2005; Holbrook 2005; Leontis et al. 2006; Butcher and Pyle 2011).

The structural modules are the nails and screws (sub-motifs), but also the hinges and locks (motifs) around which RNA fold in spectacular and complex architectures such as ribosomes (Noller 2005). Among the RNA structural motifs, the main focus of my work has been on tetraloops, which are widely spread inside RNA and embed particularly relevant intermolecular stacking interactions.

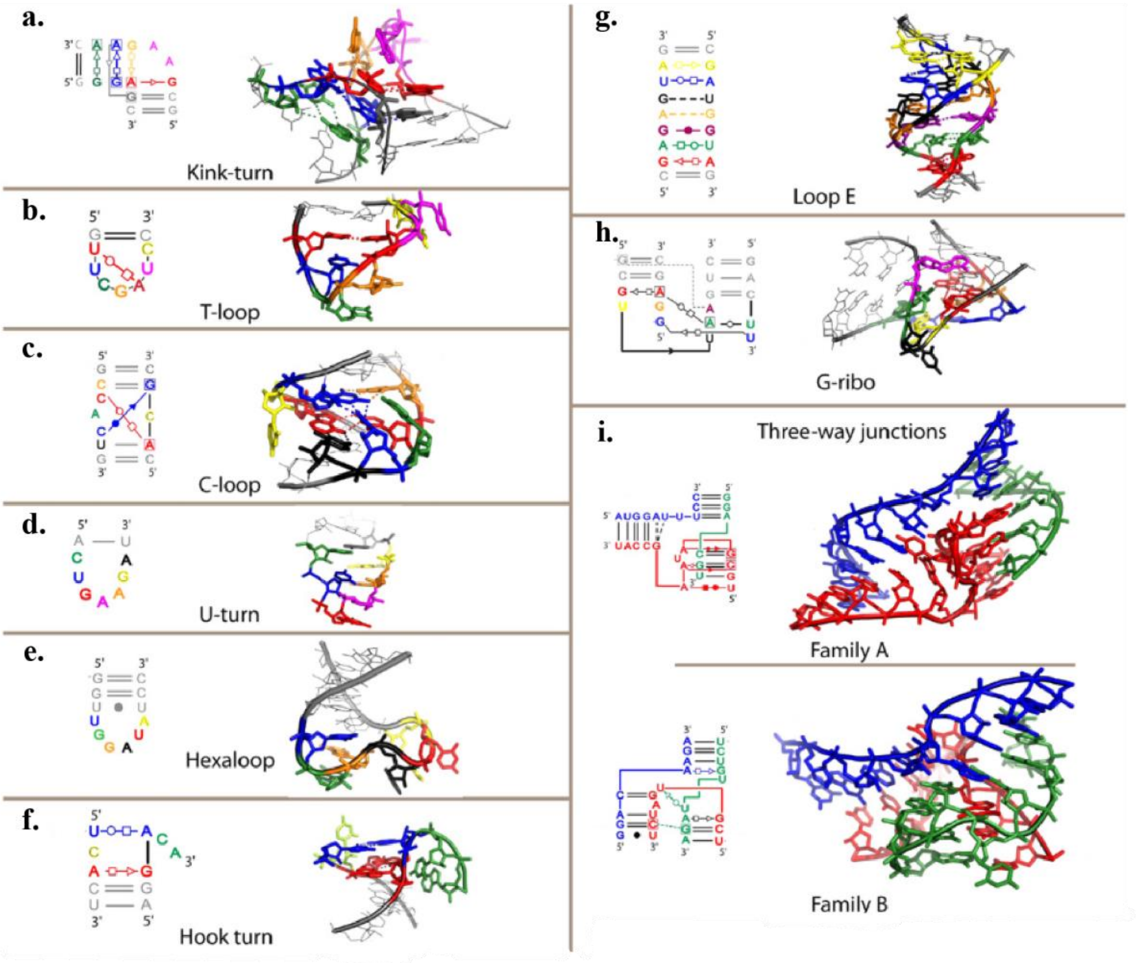


Figure 1.5. Examples of RNA structural motifs. (a) Kink-turn **(b)** T-loop **(c)** C-loop **(d)** U-turn **(e)** Hexaloop **(f)** Hook turn **(g)** Loop E **(h)** G-ribo **(i)** Family A and family B three way junctions. *Adapted from:* (Masquida et al. 2010).

1.2.3.1 Tetraloops

Tetraloops are ubiquitous RNA hairpin loops of four nucleotides that cap helical stems and are closed by at least one canonical Watson-Crick base pair (Cheong et al. 2015). Furthermore, they contribute to long-range 3D interactions (Fiore and Nesbitt 2013) and nucleic acid-protein interactions (Thapar et al. 2014). They are composed of just four nucleobases because a loop of three or less nucleotides is usually not long enough to connect antiparallel base paired strands, whereas increasing the loop size to more than 4 nucleotides may be accompanied by a loss in compactness. Although any combination of four nucleotides could in principle form a tetraloop, some sequences are preferred, mostly due to favorable thermodynamic stability and/or a potential for forming long-range interactions. The most extensively studied tetraloops, identified by their sequences, are GNRA, UNCG (YNMG) and CUYG, where N is any base, R is a purine, Y is a pyrimidine and M is either adenine or cytosine (Moore 1999; Klosterman et al. 2004; Butcher and Pyle 2011; Hall 2015). These loops were also found to be thermodynamically very stable (Antao and Tinoco 1992). This stability stems from specific 3D folds associated with their sequence, which involve extensive networks of hydrogen bonds and stacking interactions (Cheong et al. 1990; Jucker and Pardi 1995b; Jucker et al. 1996; Ennifar et al. 2000; Nozinovic et al. 2010). A notable addition to the classical tetraloop families is the tRNA anticodon loop, which is a tetraloop shaped inside a larger seven membered loop, and as such not closed by a Watson-Crick base pair (Auffinger and Westhof 2001a). Other tetraloop families have been characterized by X-ray crystallography and NMR studies, but in contrast to the previously described motifs, several of them have never been observed as independent motifs. This is for instance the case of GANC tetraloops observed in group IIC introns (Keating et al. 2008). Hitherto, no evidences of the dualistic role for local and global structure exist for these tetraloops. An overview of the most relevant proposed tetraloop families can be found in **Table 1.1**.

It is a common practice to associate a given tetranucleotides sequence (especially belonging to one of the main families) with a known specific tertiary fold. However, this assumption can sometimes be incorrect, affecting the prediction of structures such as long non-coding RNA, which quite the opposite seem to be mostly devoid of defined 3D structures (Rivas et al. 2016). To avoid these issues, it is better to refer to tetraloops associating a structure-based classification such as for U-turns. In fact, U-turn motifs are frequently associated with tetraloop and tetraloop-like structures. Yet, the U-turn motif poses a classical semantic issue. This motif was defined as a U(ridine)-turn, since the best known U-turn is found in tRNA anticodon loops where the universally conserved U in position 33 initiates

Table 1.1. Tetraloop families based on sequence classification. For each family, identified by its sequence, it is reported if the tetraloop has been characterized by X-ray crystallography and/or NMR.

Family	Method[a]	Reference(s)
GNRA	X, N	(Michel and Westhof 1990; Heus and Pardi 1991; Pley et al. 1994)
UNCG (YNMG)	X, N	(Tuerk et al. 1988; Cheong et al. 1990; Ennifar et al. 2000)
CUYG	X[b], N	(Woese et al. 1990; Jucker and Pardi 1995b)
GYYA	N	(Melchers et al. 2006)
GANC	X	(Keating et al. 2008)
UNAC	X, N	(Zhao et al. 2012)
UGAA	N	(Butcher et al. 1997)
UYUM	N	(Zanier et al. 2002; Thapar et al. 2014)
UUUU	N	(Deng and Cieplak 2007)
AGNN	N	(Wu et al. 2004)
AUYG	N	(Duszczyk et al. 2011)
AAGU	N	(Gaudin et al. 2006)
AUYA	X	(Valegård et al. 1997)

[a] Structural method of characterization. X = X-ray crystallography, N = NMR
[b] Only embedded in larger hairpin motif

the turn. Thus, it has the double meaning of a turn starting by a U and also fits the classical U-turn definition that implies for an automobile driver to perform a 180° inversion to face the opposite direction. However, it later became obvious that the U-turn term can also efficiently describe the turns in GNRA loops that are initiated by the 1st residue, a G. Moreover, U-turns starting by an N3-protonated cytosine (C^{+}) were described (Gottstein-Schmidtke et al. 2014). Even the stacking between three last loop residues that form a three-nucleotide long mini-helix is sometimes not conserved. Actually, in related structures this stack is disrupted and the tetraloop becomes a T-loop motif, where the canonical base stacking is replaced by long-range intercalations.

Besides their involvement in RNA interaction networks, tetraloops were generally thought to be initiation sites for RNA structure folding (Uhlenbeck 1990). As such, they are ideal targets for simulation studies on RNA structure and folding, especially based on MD simulations (Deng and Cieplak 2010; Chen and Garcia 2013). These studies gave useful insights into the dynamical steps followed by the RNA structure from an extended conformation to the final tetraloop fold (Haldar et al. 2015; Bottaro et al. 2016). However, they still suffer from issues related to MD parametrization concerning balance of forces and incomplete sampling of the whole conformational space. Although numerous studies are addressing these problems (Bergonzo et al. 2015; Kuhrova et al. 2016), we are still far from being able to systematically simulate the *ab initio* folding of all tetraloop motifs.

All things considered, even relatively simple RNA motifs such as tetraloops still hide surprises in their structure and thus their functions, despite having been studied for decades. The possibility of shedding light on some of these remarkable but often overlooked aspects motivated us to study tetraloops. In particular, we focused on the non-covalent interactions that constitute their structural signature and describe their involvement in more complex interaction networks.

1.3 “Uncommon” non-covalent interactions in biomolecules

The structural and functional variability of RNA molecules, reflected in the diversity of structural motifs, originates from non-covalent interactions outside canonical double helix contexts. Base stacking and hydrogen bonds are the most appreciated of these interactions. Hydrophobicity-driven base stacking is considered a primary promoter of RNA folding, much like the hydrophobic effect that drives protein folding (Butcher and Pyle 2011). On the other hand, besides hydrogen bonds responsible of base pairing, interactions between nucleobases and backbone atoms are diverse, as shown for instance in a classification of nucleobase-phosphate interactions (Zirbel et al. 2009). However, a plethora of less common non-covalent interactions coexist with these two and orchestrate the structures of biomolecules. A systematic characterization of all relevant interactions is currently missing, limiting our complete understanding of important biological phenomena and hampering the parametrization of these interactions in simulation experiments.

Considering the aromatic character of each nucleobase, non-covalent interactions involving the π systems are particularly remarkable for both nucleic acid structures and more generally for (bio)molecular recognition (Krygowski et al. 2014; Persch et al. 2015). Capping interactions, characterized by nucleobase planar faces "capped" by diverse atomic or molecular species, are particularly interesting. This family includes O/NH-π, CH-π, cation-π, anion-π and lone pair-π interactions (**Fig. 1.6**; Salonen et al. 2011).

Cation-π interactions have been found in protein systems and have been extensively described as electrostatic interactions between a cationic species and the negatively polarized aromatic π-system (Dougherty 1996; Ma and Dougherty 1997). Surprisingly, no occurrences of cation-π interactions involving free cations have been reported in nucleic acids so far. Compared to cation-π, anion-π interactions are less intuitive. They are formed by a negative species interacting with an electron deficient (sometimes referred to as "acidic") π-system belonging to a ring substituted with electron-withdrawing groups (de Hoog et al. 2004; Schottel et al. 2008; Wang and Wang 2013). The physico-chemical nature of anion-π interactions in biomolecules has been studied with quantum mechanical calculations by S. E. Wheeler, who showed that exocyclic substituents do not significantly modify π-electron densities (Wheeler and Houk 2010). Rather, the interactions of the ion with the local substituent and heteroatoms induced dipoles ("through-space effects") that more accurately describe the physico-chemical phenomena underlying anion-π interactions (Wheeler and Bloom 2014a; Wheeler and Bloom 2014b). Still, there are ongoing questions in the chemical world about the nature of anion-π interactions (Giese et al. 2015). Computational studies, crystallographic results and gas-phase experiments support the attractive nature of anion-π interactions (Arranz-Mascaros et al. 2013; Estarellas et al. 2013), but other investigations suggest that anion-π may be too weak to compete with other non-covalent interactions (Hay and Custelcean 2009; Giese et al. 2012).

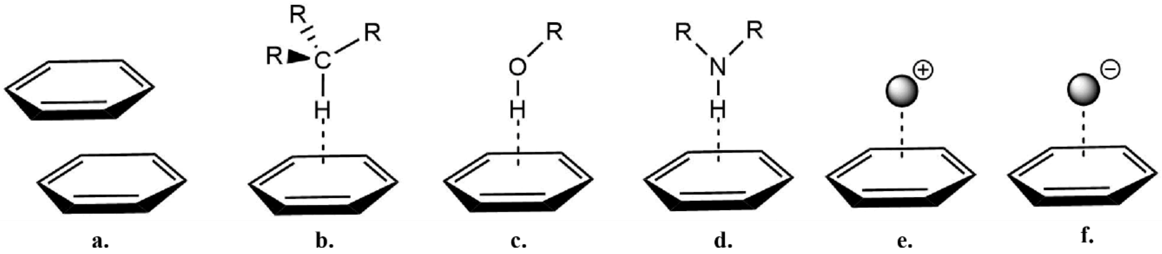


Figure 1.6. Examples of π–interactions. (a) π-π stacking **(b)** CH-π **(c)** OH-π **(d)** NH-π **(e)** Cation- π **(f)** Anion- π. The dashed lines represent an interaction with the whole π system rather than a specific position.

The so-called lone pair-π interactions in nucleic acids have also been investigated to determine their stabilization or destabilization character (Egli and Sarkhel 2007). They are formed by an electron rich atom from a neutral molecule that caps an aromatic nucleobase face. The stacking of O4' ribose oxygen atoms with nucleobases, participating in the architecture of both DNA and RNA, has been included in the family of lone-pair-π interactions (Egli and Gessner 1995). This O4'-π capping contact is characteristic of Z-DNA CpG dinucleotide steps, where specific "slide" values of the purine-pyrimidine arrangement cause the cytosine ribose to be stacked under the guanine plane adopting the less common *syn* glycosidic bond conformation (Wang et al. 1981). During my PhD work, we identified the same interaction pattern in RNA and specifically in UNCG tetraloops (**Paper 2**).

All these interactions coexist in systems that are constantly affected by the context in which biomolecules function. Considerations about environmental effects are therefore fundamental to better comprehend the local interaction properties.

1.4 Relation between biomolecules and the environment

Biomolecules inside living cells fold into stable functional forms in a complex crowded environment composed of many small and large solutes, all solvated in water medium. These unique conditions are different from the environments commonly reproduced *in vitro* or *in cristallo* experiments and have been shown to affect structures and functions of biomolecules in still largely unknown ways (Nakano et al. 2014; Sharp 2016). For instance, studies conducted on ribozymes showed that the presence of crowding agents affects folding, activity and limit the number of cations required for catalysis compared to the higher ion requirement in diluted *in vitro* experiments (Desai et al. 2014; Paudel and Rueda 2014).

Biomolecules are always solvated and physical parameters such as temperature modulate their structure and thus their functions. The polyanionic nature of nucleic acids is responsible for the interaction of a large number of diverse charged species, together with water molecules. The solvation layer of water molecules plays a crucial role for nucleic acid function, both indirectly, by stabilizing native conformations, and directly by actively participating in biological processes (Ball 2008). In fact, well-ordered water molecules in the first hydration shell of nucleic acids showing long residency times (up to nanoseconds) in MD simulations (Auffinger and Westhof 2000; Auffinger and Westhof 2001b; Kürova et al. 2014) can be considered an integral part of nucleic acid structure (Westhof 1988). On the solute side, charged species interacting with nucleic acids are for a large part metal ions, with Mg^{2+} recognized as the most relevant cation for RNA folding, structure and function (Woodson 2005; Auffinger et al. 2011; Erat et al. 2012). The relevance of metal ions and their distribution in cellular compartments was pioneered by R. J. P. Williams, who coined the term "metallome" during his studies on the evolution of life's chemistry (Williams 2001).

The metallome must be recognized as fundamental feature of a cellular compartment which is linked but not quantitatively to the proteome and the genome since it is related also to environmental availabilities and to energy supply.

- Robert J. P. Williams, 2001 -

Several inorganic anionic species have also been found to interact with the first hydration shell of nucleic acids, through interactions with the *electropositive* regions of nucleobases (Auffinger et al. 2004). The structural data available on binding of charged solute molecules to nucleic acids come mostly from crystallographic structures. However, as discussed in **Paper 5** and **Paper 6** for Mg^{2+} ions, there are several issues with identification and attribution procedures for small solute atoms. A critical and educated eye reveals many likely errors hidden inside structural data.

The involvement of RNA in complex biomolecular interaction networks extend the considerations on solute interactions to protein recognition. Theoretically, all 22 amino acids can interact with multiple RNA binding sites to form classical hydrogen bonds or “uncommon” non-covalent interactions of other types. However, statistical approaches have provided some insights on these interactions and revealed that the positively charged Arg and Lys amino acids display the strongest preference for interacting with nucleotides and especially with the negatively charged phosphate groups (Treger and Westhof 2001; Coulocheri et al. 2007). Hydrophobic Leu, Ala, Ile and Val amino acids interact less frequently than Arg or Lys with nucleotides, by forming van der Waals contacts with the bases or sugar moieties rather than phosphates type or engaging the peptide carbonyl and imino groups. Interestingly enough, negatively charged Asp and Glu residues are quite frequent at the interface between nucleic acids and proteins. These amino acids form pseudo-base pairs with H-bonding groups on the edges of nucleobases, with some surprising occurrences where the carboxylate group appears to be protonated (Kondo and Westhof 2011). In fact, the local chemical environment in a biomolecular system can significantly change the microscopic pKa values of ionizable groups (Pace et al. 2009). Interestingly, the occurrence of neutral Asp and Glu side chains has already been shown in proteins (Wohlfahrt 2005; Fisher et al. 2012). Considering their diffusion, studies on carboxyl(ate) groups and their assembly properties are particularly relevant for crystal engineering (Desiraju 2013).

All things considered, studying the complex effects of environment on biomolecular systems is a daunting task. Information on solvent binding and molecular recognition can be gathered by structural studies, but to grasp the dynamicity of biological systems techniques such as MD simulations are needed. MD allows also to implement a certain degree of control on physical parameters such as temperature, giving approximated results on a model of how complex systems are modulated by environmental effects. As a final remark, it is always important to keep in mind the truth that the Gibbs free energy formula tells us. This universally known equation goes as $G = H - TS$, where G is the free energy, H the enthalpy, T the temperature and S the entropy, all parameters usually expressed as differences between a starting and a final state of a process. In other terms, the local energetic aspect of interactions in Nature (H) is meaningless if taken alone as driving force, since it always coexists with factors determined by contextual conditions (TS). Thus, a holistic vision of biomolecular interaction networks is bound to include environmental effects.

2. Methods

Most of my thesis work has been devoted on analyzing biomolecular structural data available in online structural databases. In this section the general steps of database surveys are presented, followed by information on how structural data have been elaborated during my investigations. Further details on the format of structural files, redundancy criteria and the procedures adopted for very large ribosomal structures are also presented. Altogether, the information contained in this section will complement the methodological details presented within published material.

2.1 Database surveys

Structural data on nucleic acids, proteins and smaller molecules were extracted from two online databases: the Protein Data Bank (PDB; Berman et al. 2012) and the Cambridge Structural Database (CSD; Allen 2002; Groom et al. 2016). The PDB is a worldwide known repository of biological macromolecular structures, born from a project of X-ray diffraction data deposition in 1970s. As of August 2016, it contains ~120,000 structures of proteins, nucleic acids and protein/nucleic acid complexes, obtained by X-ray diffraction, NMR, cryo-EM and other techniques. ~110,000 are structures obtained by X-ray diffraction and among them ~7000 structures contain nucleic acids. The CSD is a repository for small-molecule organic and metal-organic crystal structures, containing over 800,000 entries from X-ray and neutron diffraction analyses. Structures deposited in this database are of a significant smaller scale than PDB structures, thus being solved at generally higher resolution. The characteristics of the surveys run on these two databases follow.

2.1.1 CSD survey

Searches in the CSD have been performed with the *ConQuest* software (Bruno et al. 2002), setting filters in order to exclude disordered and error-containing structures. A remarkable feature of the *ConQuest* search engine is the possibility to draw atoms, bonds and distances to find chemical fragments inside molecules. Therefore, it is very handy to isolate precise interaction motifs. When needed, explicit hydrogen atom positions have been used. As a general quality control, the searches were restricted to structures with crystallographic *R*-factor values ≤ 0.05, with the exception of remarkable examples. Unfortunately, even high resolution structures deposited in the CSD are not free of errors (Spek 2009). To assess this issue, the *Mercury* visualization software (Macrae et al. 2008) has been used for structural analyses and some of the structures associated with unreasonable geometrical parameters were eliminated after visual inspection.

2.1.2 PDB survey: retrieving and processing structural data

A general resolution cutoff of 3.0 Å was applied in retrieving PDB X-ray diffraction structures. This value is chosen as a compromise between having the best resolution structures available and including biologically relevant ribosomes that diffract at lower resolution than smaller RNAs. As of

August 2016, in the PDB there are ~5800 structures containing nucleic acids at resolution ≤ 3.0 Å, which contain ~635,000 nucleotides. The structural data are stored on the PDB servers inside textual PDB files of a defined format. A more in-depth analysis of PDB file format and their issues is presented in the following part, while the special case of ribosome structures is analyzed later. This data has been retrieved and parsed using *Perl* scripts, which integrate *DSSR* (Dissecting the Spatial Structure of RNA; Lu et al. 2015) part of the *3DNA* analysis tools package (Lu and Olson 2003). These tools allow to obtain structural information at the nucleotide level, including for instance glycosidic angle values, backbone torsion angles and sugar puckers. Furthermore, *DSSR* dissects the tertiary structure of nucleic acids, identifying canonical and non-canonical base pairs (including those with modified nucleotides), nucleobase-backbone hydrogen bonds, nucleobase stacking and also structural motifs such as hairpin loops, internal loops and pseudoknots. A general distance cutoff of 3.5 Å has been used for hydrogen bonds. All the data obtained for each nucleotide, together with information present in the PDB file such as crystallographic B-factors and occupancy, have been stored in *MySQL* tables. A single row contains all the structural information on each nucleotide and is identified by a univocal *IDnucleo*. This 10 character primary key is composed by the PDB identifier of the structure bearing the nucleotide, together with the residue number and chain identifier of the nucleotide retrieved from the PDB file.

Concomitantly, *PyMOL* (Schrödinger, L.L.C) scripts have been used to obtain: *(i)* symmetry-generated molecules with symmetry information stored inside PDB files; *(ii)* files containing nucleotides together with their solvation sphere; *(iii)* files containing nucleotides and the proximity environment; *(iv)* many other files with isolated nucleobase, backbone phosphates, base pairs, capping molecules and so on. These files are connected with the information inside the database by sharing the same 10 character *IDnucleo*.

2.1.2.1 PDB File Format vs mmCIF File Format

Most of the PDB entries are distributed in the PDB File Format (file extension *.pdb*) following the specification described in the Contents Guide Version 3.30 (2012). The PDB File Format was created in 1976 to be human-readable, allowing international researchers to exchange protein coordinates through a common database system. Its original format was limited to 80 columns, based on the width of the computer punch cards that were used at the time to exchange the coordinates (Berman 2008). Nowadays, it is no longer being modified or extended to support new content. A limitation of the PDB File Format is that large structures containing more than 62 chains and/or 99,999 ATOM records cannot be fully represented, so these structures were split among multiple PDB files identified with the label *SPLIT*. For this and other limitations, in 2014 the standard PDB archive format became PDBx/mmCIF, which can contain even the largest entries in a single file. The mmCIF File Format (file extension *.cif*), developed under the auspices of the International Union of Crystallography (IUCr), was created to extend the Crystallographic Information File (CIF) data representation used for describing small molecule structures and associated diffraction experiments (Brown and McMahon 2002). A dictionary for mmCIF files (identified from now on as "CIF files" for short) is available at http://mmcif.wwpdb.org. CIF files have no limitations for the number of atoms, residues or chains that can be represented in a single PDB entry. They consist in category of information represented as tables and keyboard value pairs with explicit relationship with one another, making the data content fully software accessible. All of the data items in the PDB format have corresponding items data in CIF files.

Analyzing the file format, a PDB file consists of a series of records each identified by a keyword (e.g. *HEADER*) of up to 6 characters. The format and content of fields within a given record is dependent on the record type. A CIF file, on the other hand, is composed by a series of *_name value* pairs. The name is distinguished from the value by a leading underscore. For instance, the *COMPND* record, describing the macromolecular content of an entry, of the PDB entry 1CBN would be represented as follows:

```
HEADER    PLANT SEED PROTEIN                      11-OCT-91   1CBN
```

That in the analogous CIF file becomes:

```
_struct.entry_id                  '1CBN'
_struct.title                     'PLANT SEED PROTEIN'

_struct_keywords.entry_id         '1CBN'
_struct_keywords.text             'plant seed protein'

_database_2.database_id            PDB
_database_2.database_code          1CBN

_database_PDB_rev.num                      1
_database_PDB_rev.date_original 1991-10-11
```

This is not very efficient if for each data name there are multiple values, for example in PDB *ATOM* records, where the coordinates and information on every atom of the structure are stored within a line. This issue is dealt in CIF files with using a *loop_* construct. An example of the PDB *ATOM* records follows:

```
ATOM      1  OP3   G A   1      50.193  51.190  50.534  1.00 99.85           O
ATOM      2  P     G A   1      50.626  49.730  50.573  1.00100.19           P
ATOM      3  OP1   G A   1      49.854  48.893  49.562  1.00100.19           O
```

That in the analogous CIF file becomes:

```
loop_
    _atom_site.group_PDB
    _atom_site.id
    _atom_site.type_symbol
    _atom_site.label_atom_id
    _atom_site.label_comp_id
    _atom_site.label_asym_id
    _atom_site.label_seq_id
    _atom_site.label_alt_id
    _atom_site.cartn_x
    _atom_site.cartn_y
    _atom_site.cartn_z
    _atom_site.occupancy
    _atom_site.B_iso_or_equiv_esd
    _atom_site.auth_asym_id
    _atom_site.auth_atom_id
    _atom_site.pdbx_PDB_model_num
  ATOM  1  O  OP3 G   A  11 . 50.193  51.190  50.534  1.00  99.85 A OP3 1
  ATOM  2  P  P   G   A  11 . 50.626  49.730  50.573  1.00 100.19 A P   1
  ATOM  3  O  OP1 G   A  11 . 49.854  48.893  49.562  1.00 100.19 A OP1 1
#               (data omitted)
```

The structure of the *_name* is on the form *_category.extension*, where a category contains a natural grouping of data items intuitive to a crystallographer. There is no restriction on the length or contents of *_name* (compared to the 6 character limit of a PDB keyword), but special characters are not allowed. While there is no formal specification in *_name* beyond the category and extension, the extension is usually represented as an informal hierarchy of parts, with each part separated by an underscore. Although CIF files are unpractical for visual line-by-line analysis, their structure makes automatic parsing a lot easier compared to the more "reading-friendly" PDB files.

2.1.2.2 Dealing with large ribosomal structures

CIF files are particularly useful to describe ribosomes, because they allow to contain all the information on the subunits and the proteins in just one file, making data editing and visualization easier. As of August 2016, the PDB contains 138 ribosome structures with resolution ≤ 3.0 Å, obtained with X-ray (135) or cryo-EM (3). These structures can contain solely the large or small subunit, or the whole 70S/80S ribosome. The majority of these ribosomes come from prokaryotic organisms such as *E. coli* (15), *T. thermophilus* (51) and *H. marismortui* (57), while the only eukaryotic organism with 7 ribosomal structures at resolution ≤ 3.0 Å is yeast (*S. cerevisiae*). **Table S1** of **Paper 6** contains a list with all ribosomal structures at resolution ≤ 3.0 Å in the PDB, along with further useful crystallographic information.

The large number of residues of ribosome structures requires specific criteria in order to include their nucleotides inside the *MySQL* database. Sixty-four over the mentioned 138 ribosome structures have only CIF files available, and during the data elaboration showed issues related to residue chain identification and numbering. To assess these issues, the chain identifiers of the structure have been renamed and numerated with a conserved arbitrary system based on the nature of the subunit. Moreover, many ribosomal structures are composed by up to four biological assemblies in the asymmetric unit. The asymmetric unit contains the unique part of a crystal structure, thus the smallest portion of a crystal structure from which the complete unit cell can be generated by applying symmetry operations. From this unit cell, by copy and translation the whole crystal is generated. Conversely, a biological assembly is the structure that is believed to be the functional form of the macromolecule and is generally the unit of interest, in the case of ribosomes the 70S/80S complex. This means that many ribosomal 70S/80S assemblies can be present simultaneously in the same asymmetric unit, so in the same PDB structure described by a CIF file. To assess the issue, only one biological assembly was kept for these structures, the choice of which was based on the lowest all-atom average *B-factor*. Moreover, when numbering discrepancies emerged, we chose the structures with numbering consistent with the 2D structures found at http://apollo.chemistry.gatech.edu/RibosomeGallery (Petrov et al. 2013).

The previous criteria allowed to identify the "best" ribosomal structure for every organism. Further, to expand the description to all ribosomal structures, analogous criteria were applied on all ribosomal structures deposited in the PDB, without limitation on resolution or structural method. Also mitochondrial ribosomes have been included in this analysis. The resulting list of the best resolution ribosomal structures (and their subunit with lower average *B-factor*) sorted by organism is presented in **Table 2.1**. When only one ribosomal subunit is available, the structure is considered different from the one with the whole ribosome.

Table 2.1. List of the "best" PDB ribosomal structures sorted by organism. Entries are ordered by increasing biological assembly molecular weight. For each structure is reported: PDB identifier, deposition date, resolution (Å), number of residues, "best" subunit for our criteria, structural method (X for X-ray, C for cryo-EM) and reference. *(Last update: 20/07/2016)*

PDB code	Deposition	Reso.	N° res.	Best sub.	Method	Reference
T. thermophilus (30S)						
2VQE	13/03/2008	2.50	4086	-/A	X	(Kurata et al. 2008)
T. thermophila (40S)						
4BTS	17/07/2013	3.70	30568	-/AA	X	(Weisser et al. 2013)
D. radiodurans (50S)						
5DM6	08/09/2015	2.90	6490	X/-	X	(Kaminishi et al. 2015)
H. marismortui (50S)						
4V9F	02/11/2012	2.40	7583	0/-	X	(Gabdulkhakov et al. 2013)
T. thermophila (60S)						
4V8P	14/09/2011	3.52	43352	A1/-	X	(Klinge et al. 2011)
S. aureus (50S)						
4WFA	14/09/2014	3.39	6252	X/-	X	(Eyal et al. 2015)
S. cerevisiae (60S)						
5APO	17/09/2015	3.41	11657	5/-	C	(Greber et al. 2016)
L. donovani (60S)						
3JCS	21/01/2016	2.80	11330	1-8/-	C	(Shalev-Benami et al. 2016)
T. thermophilus (70S)						
4Y4O	10/02/2015	2.30	21468	2A/2a	X	(Polikanov et al. 2015)
E. coli (70S)						
4YBB	18/02/2015	2.10	20744	DA/BA	X	(Noeske et al. 2015)
T. thermophilus (70S)						

5A9Z	23/07/2015	4.70	11326	AA/BA	C	(Kumar et al. 2015)
B. subtilis (70S)						
3J9W	16/03/2015	3.90	10618	BA/AA	C	(Sohmen et al. 2015)
E. coli (70S)						
5AFI	22/01/2015	2.90	11418	A/a	C	(Fischer et al. 2015)
S. cerevisiae (80S)						
4U4R	24/07/2014	2.80	35344	6/5	X	(Garreau de Loubresse et al. 2014)
S. cerevisiae (80S)						
3J6X	16/04/2014	6.10	17770	2S/1S	C	(Koh et al. 2014)
H. sapiens (80S)						
4UG0	20/03/2015	3.60	20197	L5/S2	C	(Khatter et al. 2015)
O. cuniculus (80S)						
3JAH	10/06/2015	3.45	18529	5/9	C	(Brown et al. 2015)
S. scrofa (80S)						
3J7P	01/08/2014	3.50	19429	5/S2	C	(Voorhees et al. 2014)
P. falciparum (80S)						
3J79 (60S)	02/06/2014	3.20	11190	A/-	C	(Wong et al. 2014)
3J7A (40S)	02/06/2014	3.20	7121	-/A	C	(Wong et al. 2014)
T. aestivum (80S)						
4V7E	22/11/2013	5.50	19110	Aa/Ad	C	(Gogala et al. 2014)
K. lactis (80S)						
4V91 (60S)	21/03/2014	3.70	10675	1/-	C	(Fernandez et al. 2014)
4V92 (40S)	21/03/2014	3.70	6733	-/A2	C	(Fernandez et al. 2014)
D. melanogaster (80S)						
4V6W	27/02/2013	6.00	20752	A5/B2	C	(Anger et al. 2013)
T. brucei (80S)						
4V8M	09/12/2012	5.57	19735	BA-B/AA	C	(Hashem et al. 2013)
Mitochondrial ribosomes						
S. cerevisiae (LSU)						
3J6B	22/01/2014	3.20	11821	A/-	C	(Amunts et al. 2014)
H. sapiens (LSU)						
3J7Y	26/08/2014	3.40	12106	A/-	C	(Brown et al. 2014)
H. sapiens (55S, class I)						
3J9M	08/02/2015	3.50	20405	A/AA	C	(Amunts et al. 2015)
S. scrofa (55S)						
5AJ4	20/02/2015	3.80	19413	BA/AA	C	(Greber et al. 2015)

In the following sections, the examples from ribosome structures will be selected from structures present in this list.

2.2 Data elaboration

The structural data available in *MySQL* tables and textual files generated by *DSSR* and *PyMOL* has been processed via *Python*, *Shell* and *MySQL* scripts, together with statistical analysis tools. In order to isolate the best examples for each structure and to avoid statistical bias we defined a general ensemble of criteria to assess nucleotide redundancy, based on local structural parameters.

2.2.1 Redundancy

A structural analysis of biomolecules with indiscriminate usage of the entire set of structures would be biased towards the most represented structures, which, by size and number, are the ribosomes. The application of redundancy criteria would therefore help to identify the best resolved and modeled representative of each structure class. Sequence-based redundancy criteria for nucleic acids have been used to create widely used non-redundant set such as *RefSeq* (Pruitt et al. 2005). More in-depth redundancy criteria involving also 3D comparison have been proposed (Leontis and Zirbel 2012) and are routinely used in annotation, classification and 3D motif-searching tools such as *FR3D* (Sarver et al. 2008).

We adopted a similar approach, considering the structural parameters of single nucleotides and the solvent molecules in their environment. Non-redundant nucleotides inside PDB structures were tagged as follows. If two nucleotides from different structures share a same residue numbers, chain codes, trinucleotide sequences, ribose puckers, backbone dihedral angle sequences (using the *g*+, *g*-, *t* categorization) and *syn/anti* conformations, they are considered as similar and the one with the best resolution is marked as non-redundant. In case of matching resolutions, the nucleotide with the lowest *B-factor* is selected. Alike, if in a same structure two nucleotides share the same residue numbers and trinucleotide sequences (with different chain codes) as well as ribose puckers, backbone dihedral angle sequences and *syn/anti* conformations, they are considered as similar and the one corresponding to the first biological unit is marked as non-redundant. The former criteria are used to filter similar structures and the latter for filtering structures with multiple related biological assemblies. To note that it is impossible to completely eliminate redundancy from a dataset without eliminating at the same time significant data. These criteria provide an upper limit of a truly "non-redundant" set that marks already close to 3/4 of all ~635,000 nucleotides as redundant.

For specific searches, such as those related to metal ion coordinating to nucleic acids, specific redundancy criteria have been used. These criteria take into account also the redundancy of the ion and are detailed in the **Methods** part of the respective **Paper**.

2.3 Structure visualization and modelling

Structures were visualized and analyzed with *PyMol* and *UCSF Chimera* (Pettersen et al. 2004), the latter also used in association with *Assemble2* for 3D modeling of RNA structures (Jossinet et al. 2010). The general visual inspection of structural features was conducted with a careful evaluation of the most relevant or unusual aspects by eye, integrating structural crystallographic information with electron density maps.

2.3.1 Electron density maps

Electron density maps are a three-dimensional description of the electron density in a crystal structure. They are calculated as function of the *xyz* coordinates by the Fourier transform of the structure factors ($\mathbf{F}_{hkl}$), which are the description of the diffracted waves from the electron "clouds" of atoms in the crystal lattice. Different electron density maps can be calculated and provided for various purposes, such as direct, difference or composite maps. During the structure evaluation I mostly used $2F_o$-F_c (F_o = observed structure factors, F_c = calculated structure factors) composite maps and F_o-F_c difference maps. $2F_o$-F_c composite maps are used to evaluate the fitting of the electron density for the structure, while F_o-F_c difference maps are used to find differences between the true structure and the available model. The F_o-F_c and $2F_o$-F_c electron density maps were retrieved from the Electron Density Server (EDS) at Uppsala University (Kleywegt et al. 2004). When these maps were not available, typically for large ribosomal structures or novel depositions, we calculated them with *phenix.maps* (Adams et al. 2010) using the structure factors deposed in the PDB. During structural evaluation, a general rule of thumb was applied regarding the σ level of electron density maps. A local structure/interaction has been found trustworthy when the densities of residues are defined beyond 2.0 σ on the $2F_o$-F_c map. Density maps for cryo-EM structures were obtained from the EMDataBank (Lawson et al. 2016). These maps are obtained with different procedures compared to X-ray diffraction, but convey analogous structural information.

2.4 Quantum mechanical calculations

Electrostatic potential (ESP) maps, also known as electrostatic potential energy maps, have been widely used to provide a representation of the charge distributions of molecular systems (Naray-szabo and Ferenczy 1995; Ma and Dougherty 1997; Wheeler and Houk 2009). They illustrate with color scales the charge of molecules three-dimensionally, allowing a rapid evaluation of variably charged regions. Knowledge of the charge distributions can be used to determine the most likely/unlikely way for molecules to interact. ESP calculations are accomplished in two steps: (i) definition of isodensity surfaces composed of points with the same electron density contouring a molecule; (ii) calculation of the electrostatic potential energy between an imaginary positive charge (+1) located on every single point of the isodensity surface and the molecule. If the imaginary charge is attracted to the molecule then the calculated potential is negative and if the same charge is repelled, the calculated potential is positive. Electron-rich regions usually display negative potentials and electron-poor or depleted regions display positive ones. To accurately analyze the charge distribution of a molecule, a very large quantity of electrostatic potential energy values must be calculated. A software then imposes the calculated data onto an electron density model of the molecule derived from the Schrödinger equation. Calculations on aromatic groups of proteins and nucleic acids point out some of their physico-chemical properties. For instance, ESP maps were shown to be very useful in the characterization at a quantitative level of interactions involving aromatic π molecules, such as cation-π interactions (Mecozzi et al. 1996).

To assess cation-π and anion-π interactions in biomolecules, we designed molecular models of the three aromatic Phe, Tyr and Trp amino-acid side chains and the A, G, C, T/U/Ψ nucleobases. Their geometries were optimized by solving the Schrödinger equation using the *Hartree-Fock* method with the 6-31G** basis set. Electrostatic potential surfaces were generated by mapping the 6-31G** electrostatic potentials onto surfaces of constant molecular electron density (0.002 e^-/$Å^3$) using the *SPARTAN* software (Wavefunction, Irwine, CA).

2.5 Molecular dynamics simulations

Molecular dynamics (MD) is a computer technique that allows to model the time evolution of the chosen molecular system. First developed in late 70s (McCammon et al. 1977), has advanced to become a method that can be used effectively to understand macromolecular structure-to-function relationship (Levitt 2001; Hospital et al. 2015). The trajectories of atoms and molecules are calculated by integrating equations of motion over a predefined ensemble of time steps. In particular, the force acting on a specific atom is calculated iteratively by taking the derivative of the potential energy function with respect to its position. In **Fig. 2.1** is represented a form of the empirical potential energy function.

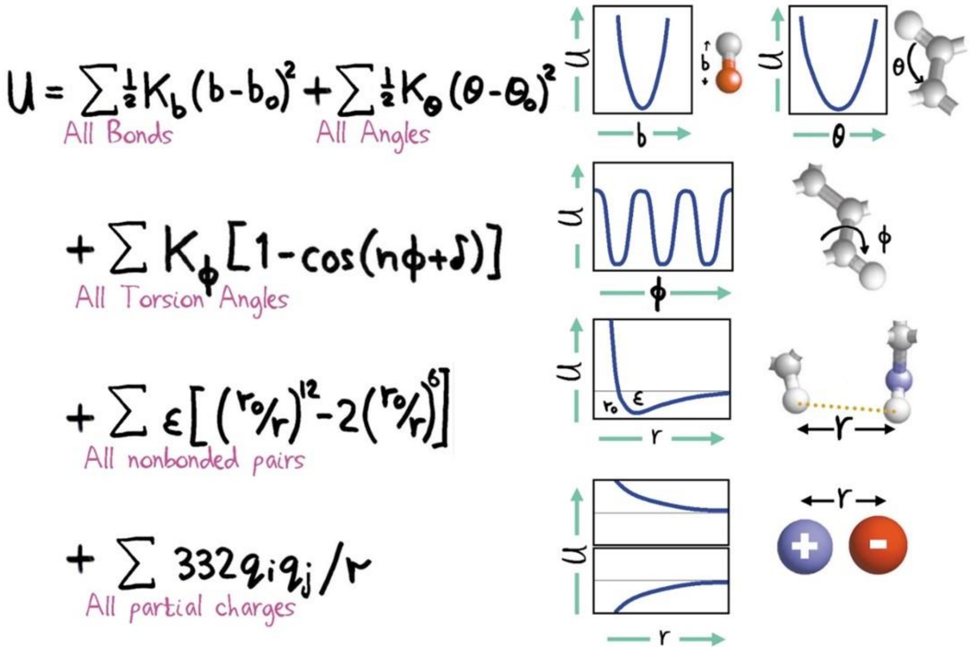

$U = \sum \frac{1}{2} K_b (b-b_0)^2 + \sum \frac{1}{2} K_\theta (\theta-\theta_0)^2$
All Bonds
All Angles
$+ \sum K_\phi [1-\cos(n\phi+\delta)]$
All Torsion Angles
$+ \sum \varepsilon [(r_0/r)^{12} - 2(r_0/r)^6]$
All nonbonded pairs
$+ \sum 332 q_i q_j / r$
All partial charges
U
b
θ
ϕ
r
r_0
ε

Figure 2.1. Total empirical potential energy function for a molecule. Many structural properties of biomolecules can be simulated with such an empirical energy function. *From* (Levitt 2001)

The function *U(r)* describes the potential energy as a function of the position *r* for N particles and is generated by a sum of terms. The first term represents the bond stretching energy between covalently bonded atoms, in a harmonic (ideal spring) force approximation. The second term represents the bond angle bending energy due to the geometry of electron orbitals involved in covalent bonding. The third term represents the energy of dihedral bending of a bond. The fourth and final term represents the non-bonded energy between all atom pairs, decomposed into van der Waals and electrostatic Coulombic energies. The functional form and parameter sets used to calculate the potential energy of a system of atoms constitutes a force field. These parameters of the energy function can be derived from quantum mechanical calculations and improved/tuned by simulations iteratively.

A classical explicit-solvent MD simulation is conducted in a box containing solute and solvent atoms. Periodic boundary conditions are used to mimic an infinite solution and to avoid problems with the molecules reaching the edges of the box. Thus, if a molecule crosses the boundary of the box, it reappears on its opposite side (**Fig. 2.2**).

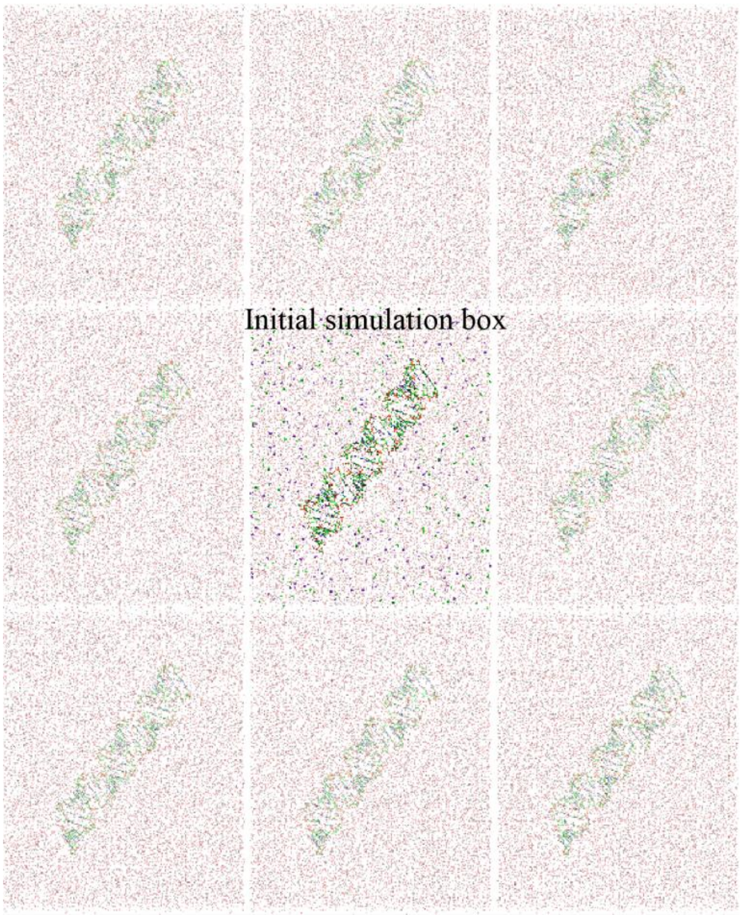


Figure 2.2. Periodic boundary conditions in MD simulations. The box in the middle contains the RNA with solvent in the initial conditions. The box surrounding are exact copies of the system, mimicking an infinite solution.

To treat long-range electrostatic interactions in periodic boundary conditions, the particle mesh Ewald (PME) summation method is routinely used (Cheatham et al. 1995).

For MD simulations performed during my PhD we chose *Amber ff14SB* force field (Cornell et al. 1995). From the 1999 edition the force field was improved numerous times in the description of RNA, with corrections such as *bsc0* to better represent α/γ torsional angles in the nucleotide backbone (Perez et al. 2007) and χ_{OL3} for the glycosidic angle χ (Zgarbova et al. 2011). In the Amber force field to each atom is assigned a particular type, based on the element, but also its hybridization and chemical context. Parameters for each interaction are based on the atom types participating in each interaction. Although this description is functional for the method, it does not include parameters for atom polarization in the potential energy function. In addition, the effect of the local environment on a single atom are included in the atom types, but they are not dynamically changing during a simulation as would appear for water and solute atoms inside a real system when their local context changes. Another limitation of MD simulations is that only non-covalent interactions are allowed to break and form, but covalent bonds cannot be broken. This makes MD a method not suited to analyze reactions that involve breaking-making covalent bonds.

All things considered, MD simulations can give information about the dynamics on an atomic level, not only in terms of RNA dynamics but also in terms of the behavior of hydration shell surrounding RNA. MD is a mathematical attempt to reflect physics and chemistry on the atomic level, thus is approximated and never completely accurate. Also, MD trajectories of the investigated molecule and its solvent environment are highly depending on the starting structure, the parameters of choice and the simulation setup. Nevertheless, a critical and careful eye would be able to use MD simulations in order to gain information on both biomolecular structure and environmental factors.

2.5.1 MD protocol

The structure of RNA duplexes were modeled with *Nucleic Acids Builder*, part of Amber Tools. This molecule was then visually inspected for structural issues, which is relevant especially when using X-ray crystallographic structures as starting point (Hashem and Auffinger 2009). Then, the molecule was placed into a box together with *SPC/E* water molecules (Berendsen et al. 1987) and ions; K^+ was added as counterion and KCl as salt. Periodic boundary conditions and Particle Mesh Ewald have been used. After several equilibration phases, simulations were ran for 50 ns in conditions of constant temperature (with a Berendsen temperature coupling scheme (Berendsen et al. 1984)) and pressure. Trajectories were visualized by *VMD* (Humphrey et al. 1996) and analyzed with *Python* scripts and *CPPTRAJ* software (Roe and Cheatham 2013).

3. Section I.
Stacking contacts between nucleic acid backbone oxygens and nucleobases

3.1 Essential considerations on stacking interactions

The aromatic nature of nucleobases favors hydrophobicity-driven stacking interactions such as the base-base stacking observed in DNA and RNA double helices. Further, other charged or uncharged species can interact with base faces, generating a family of stacking or capping contacts. Among those, I focused on the study of nucleobase capping by backbone oxygen atoms, which generate two types oxygen-π of stacking: *(i)* phosphate-π when it involves a phosphate oxygen; *(ii) O4'*-π when it involves a neutral ribose O4' atom. Both capping contacts are characterized by short (< 3.1 Å) oxygen-to-plane distances and are assisted by other contextual interactions. In fact, one of the central points emerging in this thesis is connected with a concept expressed by J. Dunitz on intermolecular interactions, namely that *short contacts do not necessarily correspond to specific bonding interactions between the atoms involved* (Dunitz and Gavezzotti 2009; Dunitz 2015). **Capping contacts between oxygen backbone atoms and nucleobases can be considered as weak bonding or slightly repulsive interactions and do not constitute the leading force for biomolecular folds, but are nonetheless fundamental to obtain the final biomolecular tertiary structure**. This would explain why both capping contact families exist assisted by hydrogen bonds or other non-covalent interactions and the fact that they appear to be more "tolerated" than actually determinant for the biomolecular structure. Thus, backbone oxygen atoms stacking with nucleobases constitute a *secondary* interaction, with the assistance of stronger constraints, and the term *interactions* to describe them has to be considered in this perspective more as a synonym of *intermolecular proximity*. Anyhow, they are not strongly repulsive interactions, otherwise we would observe a complete different biomolecular structural landscape. In addition, the quasi-absence of reference to energy values for the interactions analyzed during my work is a choice made to avoid the pitfalls of determining local energy values for systems embedded in complex environments, which we cannot precisely describe from the energetic point of view.

Both families of capping interactions have been found ubiquitously in RNA. Phosphate oxygen stacking with nucleobases have been found in the signature of tetraloop motifs. On the other hand, O4' stacking with nucleobases in dinucleotide steps are a "conserved" interaction between RNA and DNA (more specifically Z-DNA), and represent a Z-DNA fragment conserved in RNA with many structural and functional implications.

3.2 Anion-π interactions in nucleic acids

Anion-π interactions in nucleic acids take essentially the form of phosphate-π stacking, between a backbone phosphate and a nucleobase, and is found in biologically relevant tetraloop motifs such as the anticodon loop (Quigley and Rich 1976). Although rarely, this interaction exists also outside

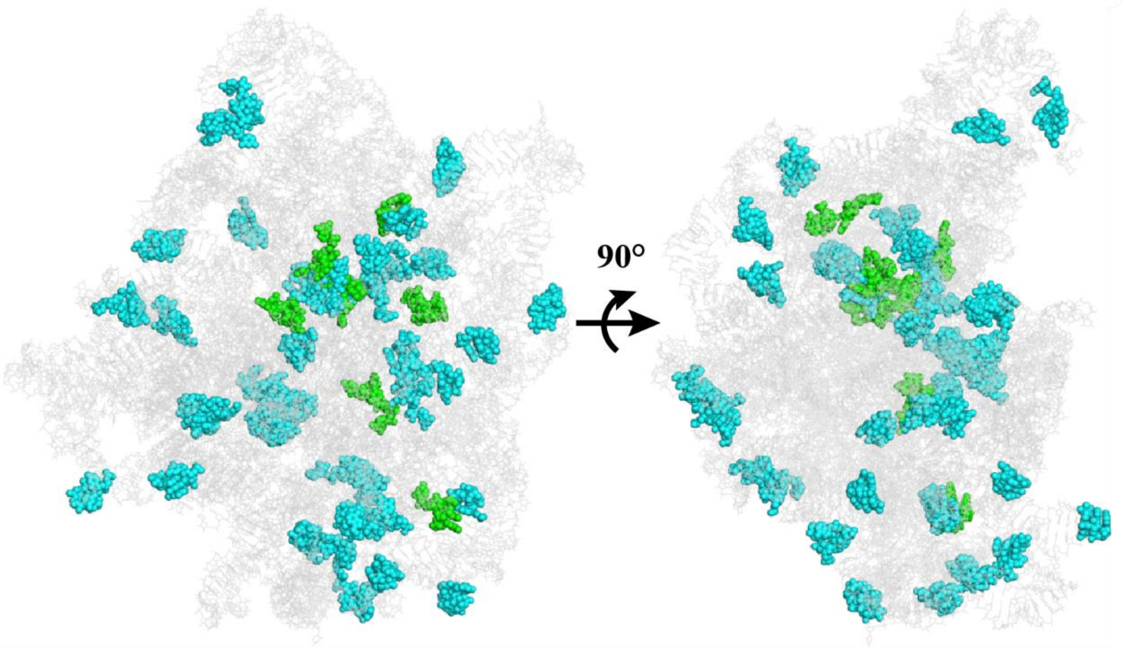


Figure 3.1. Phosphate-π interactions in the 50S subunit of *H. marismortui.* Of 46 total contacts 38 are found within tetraloops (cyan) and 8 in long range contacts (green; PDB: 4V9F res.:2.4 Å; (Gabdulkhakov et al. 2013).

tetraloops, in both short- and long-range contacts. See for instance the 46 phosphate-π stacking examples in the *H. marismortui* 50S ribosomal subunit (**Fig 3.1**).

Except for adenines, all nucleobases have been found to start tetraloops. In this tetraloop sub-family, the 1st nucleobase stacks with the phosphate group of the 3rd nucleotide (generally the *Pro-R* oxygen atom, identified as OP2 in PDB format), as observed for instance in anticodon loops **(Fig 3.2a)**, tRNA T-loops (**Fig. 3.2b**), GNRAs (**Fig. 3.2c**) and also loops starting with a protonated cytosine (**Fig. 3.2d**). In the latter surprising example, a cytosine bearing a proton on its N3 imine nitrogen takes the place of U in a loop with the anticodon topology.

All these tetraloops adopt a fold characteristic of U-turns. The phosphate-π stacking, although sometimes mentioned, was never included up to now in the structural signature of these motifs. A more in-depth analysis of these motifs along with an analysis of ion-π interactions in biomolecules constitutes the core of the **Paper 1**.

Although found ubiquitously in RNAs, phosphate-π stacking has been found to be "assisted" by distinctive hydrogen bonds present in the context, such as the hydrogen bond between the cytosine N3 and the phosphate oxygen in **Fig. 3.2d** (**Paper 1**). More generally, within U-turns the phosphate-π stacking is assisted by a hydrogen bond between a nitrogen on the Watson-Crick edge of the 1st base and the OP2 oxygen of the phosphate group belonging to the 4th nucleotide. The need of a hydrogen-bond donor on the 1st nucleobase explains why guanine and uridine, bearing imino groups

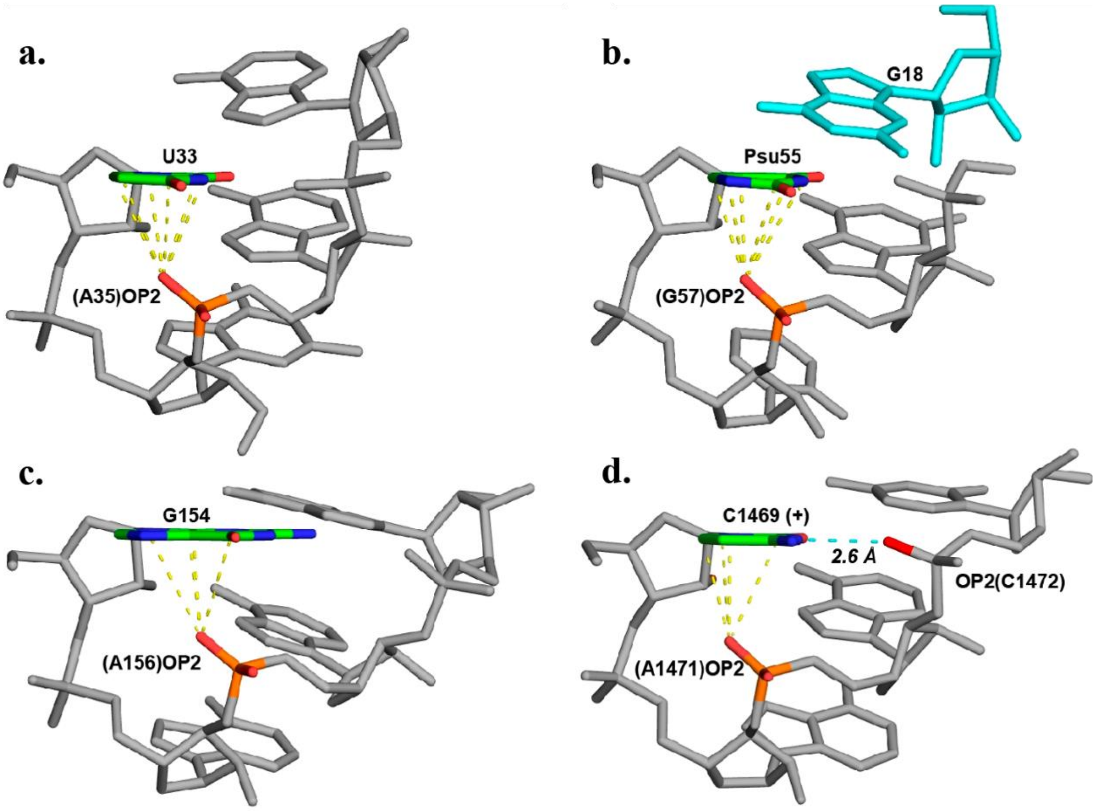
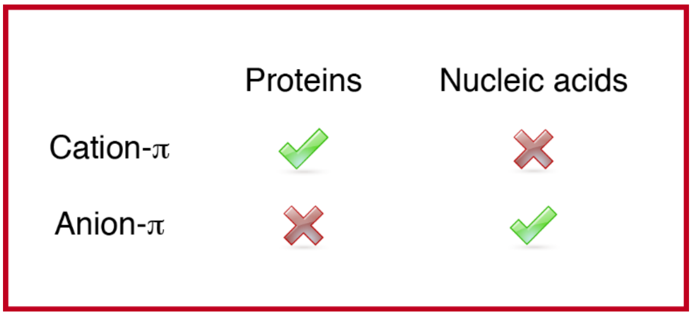


Figure 3.2. Phosphate-π stacking within tetraloops. (a,b) tRNA anticodon and TYC loops (PDB: 1EHZ res.: 1.9 Å; Shi and Moore 2000) **(c)** GAAA tetraloop in the signal recognition particle (PDB: 1HQ1 res.:1.5 Å; Batey et al. 2001) **(d)** Ribosome C$^+$AAC tetraloop (PDB: 1VQ8 res.:2.2 Å; Schmeing et al. 2005). Interatomic distances shorter than 3.5 Å are represented by dashed lines.

on their WC edge (N1 for guanine and N3 for uridine), are almost exclusively favored over cytosine and adenine. In particular, the latter have to be protonated in order to form the base-phosphate hydrogen bond.

3.2.1 Paper 1. *Cation-π versus anion-π interactions — biomolecules can't have both* (in preparation)

Graphical abstract



	Proteins	Nucleic acids
Cation-π	✓	✗
Anion-π	✗	✓

Cation-π interactions are recurrent in proteins but have not been observed in the highly aromatic nucleic acid systems. We show that this stems from fundamental electrostatic differences between nucleic acid and protein aromatic groups. In fact, electrostatic potential (ESP) map calculations reveal that nucleic acid aromatic groups bear positive potentials, while protein aromatic groups are characterized by rather negative ones. These features are confirmed by an extensive PDB survey based on the interactions with positive and negative species and the aromatic π systems of nucleic acids and proteins. Anion-π interactions have been found abundantly in RNA (and rarely in DNA), as phosphate-π interactions. These interactions are significant for motifs such as GNRA tetraloops, being part of their structural signature, and are almost exclusively of an "assisted" nature, coexisting with hydrogen bonds involving nucleobases and backbone atoms. We also analyzed by MD the involvement of phosphate-π interactions in stable hydration patterns in tetraloop motifs, showing long residence time (> 1ns) water molecules bound to the stacked phosphate group. In parallel, we strove to find free anions and Asp/Glu side chains stacked over nucleobases, obtaining rare occurrences of "face-to-face" (parallel) stacking contacts and only a few dubious cases of perpendicular ones that do not support the existence of orthogonal anion-π interactions in nucleic acids. Analogous results were obtained searching for cation-π interactions in nucleic acids, with the most abundant occurrences being "face-to-face" arginine-nucleobase stacking interactions and only few ambiguous orthogonal interactions. These parallel interactions with Arg should be considered of hydrophobic nature and do not support the existence of cation-π interactions in nucleic acids. The final step of our survey highlights how the claim that anion-π interactions would exist in protein systems is often related to different types of imprecisions like structural misinterpretations or questionable searching criteria; to confirm this, our search for anion-π interactions in protein did not yield any clear positive result.

Altogether, these data suggest that the expression "π interactions" should be carefully applied, considering how relevant are environmental context factors in their occurrence. These factors are hard to assess. Nonetheless, in hydrated biological systems hydrophobic effects have clearly to be considered as key elements in understanding the right balance of interatomic forces at play in motifs characterized not only by “π interactions”. Progresses in their understanding will surely benefit our comprehension of biomolecular recognition phenomena and will allow their exploitation in drug design and molecular engineering.

CATION-π VERSUS ANION-π INTERACTIONS — BIOMOLECULES CAN'T HAVE BOTH

Luigi D'Ascenzo[1], Filip Leonarski[1,2] & Pascal Auffinger[1*]

[1]Architecture et Réactivité de l'ARN, Université de Strasbourg, Institut de Biologie Moléculaire et Cellulaire du CNRS, 67084 Strasbourg, France
[2]Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

*Corresponding author

For PA, please forward correspondences to:
Telephone: (33) 388 41 70 49
FAX: (33) 388 60 22 18
e-mail: p.auffinger@ibmc-cnrs.unistra.fr

Keywords: DNA; RNA; anion-π; cation-π; non-covalent interaction, tetraloop; anticodon loop; TΨC loop; GNRA; tRNA; pseudouridine; hydration; solvation; aromatic;

Abbreviations: PDB: Protein Data Bank

ABSTRACT

While cation-π interactions are recurrent in proteins, they have not been identified in the highly aromatic nucleic acid systems. Herein, based on electrostatic potential calculations, we suggest that the origin of this effect relates to the hitherto largely unnoticed electrostatic dissimilarity between the aromatic rings of both biomolecular groups that result in inverted cation/anion-π binding profiles. A detailed PDB survey revealed that anion-π interactions take dominantly the form of phosphate-π interactions in RNA, which rarely occur in DNA. They are mostly observed in sharp turns where they are essential for the folding, architecture and function of key structural elements. Further, this survey established that cation-π and anion-π interactions do not occur in nucleic acids and proteins, respectively. The involvement of phosphate-π interactions in capping of otherwise solvent exposed nucleotides is discussed to understand the possible implications of ion-π interactions in biomolecular recognition phenomena and drug development.

	Proteins	Nucleic acids
Cation-π	✓	✗
Anion-π	✗	✓

Graphical Abstract

INTRODUCTION

Aromatic rings are ubiquitous in biomolecular systems where they are involved, besides classical π-π stacking interactions, in unusual non-covalent interaction types[1]. Among those, cation-π interactions **(Fig. 1a)** have well been characterized[1-3]. Intriguingly, these interactions are almost never observed in nucleic acids where they should abound given the aromatic character of nucleobases[4].

Instead, rare stacking interactions between anionic species such as negatively charged phosphate groups and nucleobases were observed. They are commonly called anion-π interactions and were addressed in structures as significant as the first crystallized tRNA molecule[5,6] where they were described as van der Waals contacts stabilized through ion-induced dipole interactions **(Fig. 1b)**. Such anion-π interactions are nowadays perceived as important supramolecular bonds in chemistry[7-10]. Yet, despite their ubiquity in major RNA molecules, a thorough structural survey of these interactions and an understanding of their physico-chemical characteristics in biomolecules are both not currently available.

In order to rationalize the anion-π interaction potential of aromatic groups, it is often stated that electro-attractive substituents induce electron-depleted π-orbitals, called "π-acidic", that create a positive quadrupole moment inducing favorable interactions with anionic species[7,9,10]. Wheeler and Bloom[11] challenged this view by demonstrating that exocyclic substituents do not significantly modify π-electron densities and suggested that these aromatic rings should no longer be considered as electron-depleted. Instead, these authors showed that the interactions of the ion with the local substituent induced dipoles ("through-space effects") describe more accurately the physical phenomena underlying anion-π interactions. In other words, the repulsive anion-to-π-cloud interactions are overcome by favorable charge dipole interactions. This insight was later extended to nitrogen containing heterocycles[12].

In this respect, the π-stacking terminology has been questioned[1,11,13]. Using model systems, Wheeler and Bloom established that aromatic π-electron delocalization hinders π-π stacking[14]. Similarly, although the direct contact of a cation with π-orbitals is stabilizing, the direct contact of an anion with π-orbitals is unfavorable. Consequently, the generic "ion-π interaction" expression might sound misleading as it inclines to consider that the contribution of the delocalized π-electrons drives the association. As a result, the term "π stacking" should merely be regarded as a geometrical descriptor for specific interaction modes in unsaturated molecules[1]. Henceforth, we use the widely accepted "anion-π“ term by stressing that it refers to anionic species stacking with the planar aromatic system of nucleobases in a purely geometrical manner without reference to electronic effects. Conversely, the "cation-π“ term is used here for cationic species stacking with the planar aromatic system of the Phe, Tyr and Thr amino acids.

From a practical point of view, it has repeatedly been emphasized that electrostatic potential (ESP) maps that reflect both electron and nuclear charge distribution provide major insight in the ability of aromatic rings to engage in specific ion-π interactions[2,11]. Here we calculate ESP maps to unambiguously establish that the ion-π binding properties of aromatic amino acid side chains and nucleobases are opposite, therefore suggesting a different ion-π binding potential for proteins and nucleic acid systems. In line with those findings, rigorous PDB surveys assess that cation-π and anion-π interactions cannot occur within nucleic acids and proteins, respectively. We further discuss the implications of the formation of very specific anion-π interactions in nucleic acids in terms of hydrophobic effects and consequently interrogate the physico-chemical origin of these interactions.

RESULTS

Nucleobases and aromatic amino acid rings have opposite polarities

To emphasize their differences, ESP maps were calculated for the Phe/Tyr/Trp amino acid and nucleobase rings. As shown elsewhere[2,15], the negative regions of the ESP maps correlate well with the potential of amino acid aromatic rings to engage in attractive cation-π interactions **(Fig. 2)**. On the contrary, nucleobase ESP maps suggest the presence of neutral to electro-positive regions over the rings. These polarity changes result from the combined effects of the nucleobase heterocyclic nitrogen atoms and exocyclic substituents. The highest positive potentials are calculated for the U/Ψ/T nucleobases where ESP maxima are located over the C2/C4 atoms that are part of the strongly electron-withdrawing carbonyl groups. Comparatively, the adenine ESP profile is rather "neutral" since this nucleobase has only one amino substituent that induces a much smaller dipole than carbonyl groups. Consequently, these maps imply that, in opposition to amino acid rings, nucleobases would favor anion-π over cation-π interactions. They further suggest that U/Ψ/T nucleobases share the strongest tendency to form anion-π interactions, followed by G>C≈A nucleobases.

Phosphate-π contacts are recurrent in RNA systems …

The PDB was searched for stacking of anionic species with nucleobases. We first identified ≈4800 anion-π contacts involving the anionic backbone phosphate groups (hereafter called phosphate-π interactions; **Fig. 1b**). Most of them are found in the structurally highly complex RNA systems (≈97%); the remainders are spotted in double helically dominated DNA structures and are mostly associated with lattice contacts **(Table 1)**. Noteworthy, phosphate-π contacts involving OP2 atoms in RNA are significantly more frequent (≈82%) than those associated with OP1 atoms. Overall, the 1054 non-redundant (see Method section) phosphate-π contacts comprise for the largest part U and G nucleotides (≈86%). For uridines, stacked OP2 atoms cluster over a limited region of the ring in the vicinity of C2 atoms. For guanines, the OP2 distribution spans over the two rings and seems bimodal **(Fig. 3)**. Note that 94 % of the phosphate-π contacts occur on the 3' side of the nucleobases.

Interestingly, the histograms of the stacked OP1/2 to nucleobase plane distances suggest a bonding character for these contacts **(Fig. 3)**. The mean ≈3.1 Å distance **(Table 1)** is below the estimated 3.2 Å sum of the van der Waals radii for carbon/oxygen atoms (note that the average distance between stacked nucleobases is ≈3.4 Å and corresponds roughly to the sum of van der Waals radii of two carbon atoms). These aspects underscore the importance of phosphate-π interactions in RNA systems and hints at their critical role in contributing to supporting the fold of complex structural systems.

… and are of an "assisted" nature …

In order to better apprehend the stacking characteristics of these phosphate-π contacts, we undertook a more precise examination of the associated structural motifs. In most instances (≈73%, **Table 2**), the stacked OP2 (nucleotide *n*) is part of a dinucleotide motif where the (*n+1*)OP2 atom forms also a hydrogen bond with the pyrimidine N3 or guanine N1/N2 Watson-Crick edge atoms of the stacked nucleobase (**Fig. 4a**). In other less frequent motifs (≈8%) the sugar associated with the stacked nucleobase adopts a *C2'-endo* rather than the more common *C3'-endo* pucker allowing the formation of an O2'…OP2 hydrogen bond with the stacked OP2 (**Fig. 4b**). A further and smaller subset (≈6%) is associated with structures where PO_4 and O2' assisted contacts occur simultaneously. A last subset (≈5%) involves contacts between the stacked phosphate and amino acids, nucleobases, or rare instances of intranucleotide stacking. In total, over 93% of phosphate-π contacts occur in combination with other interactions (**Table 2**).

As such, our data suggest that most of these contacts must be considered as "assisted interactions". In other words, phosphate-π contacts would probably not form without the above-mentioned "assisting" interactions.

Hereafter, we call the two most frequent "assisted interactions" involving OP2 atoms "PO_4 assisted" (≈73%) and "O2' assisted" (≈8%) and describe in more details the former. It occurs in tetraloop like motifs such as those found in the anticodon and TΨC loops of tRNA structures (U/Ψ starting loops; **Fig. 4c**) as well as in the highly recurrent GNRA tetraloops (G starting loops; **Fig. 4d**)[5,6,16,17]. These tetraloops have in common an imino group on the Watson-Crick edge of the first nucleotide that is key to the formation of "PO_4 assisted" interactions. In rare instances, when a loop starts with a protonated cytosine (C^+) and consequently bears an imino group on its Watson-Crick edge, its structure is similar to U starting loops[18] stressing even more the importance of this imino group in the formation of these sharp turns (**Supplementary Fig. S1**).

… which might involve stable hydration patterns in tetraloop motifs

In many instances and especially in tetraloop motifs, the stacked OP2 atom is linked to the *(n-1)*OP2 atom through a single water molecule bridge (pyrimidine starting loops; **Fig. 4c**) or a more complex solvation pattern involving sometimes a double water molecule bridge (purine starting loops; **Fig. 4d**). Thus, the stacked OP2 loses only two water molecules out of the three that form its hydration cone when fully exposed to the solvent as in regular RNA helices[19].

To check the stability of the contacts structuring these loops, we performed molecular dynamics (MD) simulations of a $tRNA^{Phe}$ and a sarcin-ricin loop structure, the latter containing a GNRA tetraloop[17]. In these simulations, the stacking of the phosphate group was maintained in the tRNA anticodon and TΨC loops as well as in the GNRA tetraloop with average distances of

≈3.1±0.2 Å over a total of 60 ns of simulations. Surprisingly, the average hydrogen bonded (HB; **see supplementary material**) time of the water molecules attached to the stacked OP2 atom (**Fig. 4c/d**) were in the ≈1-2 ns range for the anticodon and TΨC loops and around ≈1.0 ns for the GNRA tetraloop (**Fig. 4e/f**) with maximum HB times in the ≈3-6 ns range. Such residency times are at least two orders of magnitude larger than those estimated for solvent exposed OP1 atoms derived from the same simulations (≈20 ps). In "O2' assisted" contacts, the hydroxyl group replaces a first shell water molecule of the stacked OP1/2 atom. In most instances a second water molecule forms additional contacts with this oxygen atom. Hence, in the latter configuration, the OP1/2 atom loses only a single water molecule from its hydration shell.

Anion-π contacts involving anionic amino acids or "free" anions are rare in nucleic acids

To better explore the characteristics of anion-π interactions in nucleic acids, we searched for nucleobase contacts involving the negatively charged Asp/Glu side chains. These residues can either interact with nucleobase edges by forming regular hydrogen bonds[20,21] or contact the nucleobase planes through "face-to-face stacking" and "orthogonal interactions"[22]. Our search indicates that stacking contacts of Asp/Glu side-chains to nucleobase rings are rare (17 non-redundant events, 9 of them being of the "face-to-face" type). In some instances, these Asp/Glu residues form pseudo-pairs with nucleobases[21] or salt-bridges with positively charged amino acids. Yet, the binding geometry observed in these examples along with their poor statistical significance does not support the occurrence of "orthogonal" anion-π interactions involving Asp/Glu side chains and nucleobases (see Discussion section), especially if these numbers are compared with the ≈1200 non-redundant Asp/Glu side chains directly hydrogen bonded to nucleobase edges.

Non-redundant contacts between nucleobase rings and free anions are rare and limited to two sulfates, one sulfonate, two nitrates, one citrate and one acetate anion contact. In these few occurrences, anionic molecules establish strong hydrogen bond contacts with other "assisting" partners leading to incidental stacking interactions. Again, the statistical significance of contacts involving Asp/Glu carboxylate groups and free anions is too small to support the occurrence of anion-π (other than phosphate-π) interactions, in nucleic acid systems and their participation in nucleic acid/protein recognition phenomena.

Do cation-π interactions exist in nucleic acid systems?

To check the occurrence of cation-π contacts involving nucleobases **(Fig. 2)**, we strived to isolate them in PDB structures. A limited number of possible non-redundant cation-π contacts were identified such as 3 K^+, one Na^+ and 18 Lys ammonium groups. Some of the contacts involving K^+ or Na^+ ions were inferred from low-resolution structures where the ion attribution is at best ambiguous. For example, in the structure with PDB code 3CUL, an ion marked as K^+ is at 3.4 Å from a (C)N4 amino group and is more likely a water molecule or a Cl^- ion[20] **(Supplementary Figure S4)**.

Forty-one "non-parallel" Arg-nucleobase contacts were also identified. They are characterized by an Arg-to-nucleobase plane angle > 30° and are further subdivided into 34 "oblique" contacts (30° < angle < 60°) and 7 "orthogonal" (angle > 60°) ones. However, among these few, no unambiguous "orthogonal" contact associated with N-H…π bonds could be observed. In total, 535 non-redundant Arg nucleobase contacts were identified and 95% of them are close to an ideal face-to-face geometry. These numbers are small when compared to the ≈2900 (Arg) and ≈1500 (Lys) non-redundant side chains hydrogen bonded to nucleobase edges. Consequently, if we exclude face-to-face contacts involving the highly hydrophobic Arg side chains (see Discussion section), the PDB searches confirm that cation-π contacts are not present in nucleic acids.

Anion-π interactions in proteins, do they exist?

As a final step of our exploration of all potential interactions involving aromatic rings and ionic species in biomolecules, we examined the possible occurrence of anion-π contacts in proteins. We searched for chloride, sulfate, phosphate and Asp/Glu carboxylate groups contacting amino acid aromatic planes at a distance ≤ 3.5 Å. Like for cation-π interactions in nucleic acids, we found only a few borderline anion-π contacts in the PDB. Hence, we were unable to identify a convincing set of anion-π interactions in proteins.

We also checked the literature where a few reports evaluated the ability of amino acids to engage into anion-π interactions[9,23-25]. The most significant of these studies gathered only a handful of anion-π examples, used too permissive criteria to identify them and, more importantly, misinterpreted a locally wrong contact in a PDB crystal structure **(Supplementary Discussion S1 & Fig. S3)**[23]. Such misfortune related to PDB structures is indeed possible since even high-resolution crystal structures can embed local inaccuracies or much worse[26]. Based on these observations and our PDB searches, we safely conclude that anion-π interactions are not present in proteins.

DISCUSSION

The key result of this study is the uncovering of the previously largely unnoticed difference in electrostatic potential between protein and nucleic acid aromatic groups. As a consequence of this difference, the ion-π interaction profiles are opposite in these two biomolecular systems accounting thus for the intriguing absence of cation-π interactions in nucleic acids. Furthermore, this finding helps to understand conflicting reports in the literature related to anion-π interactions described in proteins[9,23-25]. Finally, the occurrence of recurrent anion-π interactions in nucleic acids is brought to light and is shown to take exclusively the form of intramolecular phosphate-π contacts.

Most of these phosphate-π interactions are similar to those occurring in the sharp turns found in tRNA anticodon and TΨC tetraloop-like motifs[5,6,17] starting with U or more rarely C/C^+ nucleotides[18] or in GNRA tetraloops[16,17] (**Fig. 4 & Supplementary Fig. S1**). They are consequently also found in large RNA structures such as ribosomes. For example, 44 phosphate-π interactions are found in the 50S ribosomal subunit of *Haloarcula marismortui* (PDB code: 4V9F, resolution 2.4 Å), 30 being associated with tetraloops and 14 with larger loops and other long-range contacts. Hence, the widespread occurrence of phosphate-π interactions in decisive RNA molecules leaves no doubt about their significance in nucleic acid systems.

Interestingly, the largest proportion of these phosphate-π interactions is part of two well-conserved interaction modules (**Fig. 4a/b**). The most represented involves two successive phosphate groups, the first forming the stacking contact and the second a hydrogen bond with the Watson-Crick edge of the stacked nucleobase. These "PO_4 assisted" motifs appear essentially in association with imino group bearing nucleobases such as U/Ψ/C^+/G and only parenthetically with A/C nucleobases. As an outcome, the small reported percentage of phosphate-π interactions involving A/C nucleobases seems unrelated to a specific electronic configuration of the stacked nucleobase reflected by a more neutral ESP profile (**Fig. 2**), but rather to the absence of an imino hydrogen-bonding group. In that respect, putting into regard the observed clustering of OP2 atoms and the most electropositive regions of the U/G nucleobases (**Supplementary Fig. S2**) could have enticed us to wrongfully suggest that a strong attractive force associated with the U/G nucleobase type would govern the clustering of the phosphate groups.

Consequently, the latter observation and the fact that "assisted" interactions embedded in tetraloop motifs largely dominate the pool of phosphate-π contacts prompt us to suggest that these interactions are not driving the formation of these particular loops. This is further corroborated by their accumulation on the 3' nucleobase side (**Fig. 3b**) and the absence of additional anionic groups engaging in ion-π interactions.

These conclusions are in line with those of a former study suggesting that, in the absence of polarization effects induced by base protonation or metal binding, direct phosphate-π interactions are quite weak and tolerated rather than significantly stabilizing[17]. Indeed, in all surveyed structures, we failed to identify recurrent polarization effects that could strengthen the phosphate-π interactions. It is also worth noting that for tetraloops starting with a charged C^+ nucleobase[18], the direct phosphate-π interaction is significantly reinforced[17]. However, this charge difference does not affect the structure of the loop that is comparable to those starting with a neutral U nucleobase[18] suggesting that even strong phosphate-π interactions are not instrumental in the formation of these motifs (**Supplementary Fig. S1**).

Estimating the strength of anion-π interactions is indeed a difficult exercise since the characterization of such interactions in the liquid or crystal phase is complicated by the coexistence of multiple binding motifs involving the solvent as well. From a purely structural point of view, Dunitz and Gavezzotti advocated persistently that short interatomic distances in crystal do not necessarily reflect attractive interactions but could also be associated with an increase in potential energy resulting eventually in a repulsive force[27]. Surely, we cannot associate the estimated average OP2 to nucleobase plane distance (≈3.1±0.2 Å) to an energetically strong phosphate-π interaction **(Fig. 3)**. However, we can certainly exclude both a direct strong repulsive contribution that would impede the formation of these important motifs and a strong attractive component that would compete with the formation of regular hydrogen bonds.

Yet, besides the classically invoked dispersion forces or electrostatic substituent effects, other forces promoting the formation of anion-π interactions in RNA may need to be considered. For model systems, it has been shown that the driving force for aromatic association derives essentially from solvation/desolvation effects that are especially strong in polar solvents. In other words, the hydrophobic effect is a key player in biomolecular systems[13]. As such, nucleobases in water tend to stack in order to minimize their solvent exposure while forming simultaneously various types of hydrogen-bonded interactions. In sharp turns (tetraloop-like motifs), the phosphate group “caps” the nucleobase[5,28] and minimizes its solvent exposure. Such nucleobase capping should contribute positively to the global free-energy balance.

Concomitantly, the formation of this capping motif implies a partial dehydration of the intervening OP2 atom that loses at least two out of three of its hydrogen bonded water molecules. Such desolvation contributions tend to be unfavorable to the entire folding process. However, through MD simulations, we found that the water molecules bound to the stacked phosphate in the three analyzed loop motifs are strongly bound as suggested by their significantly longer residency times. Thus it can be hypothesized that these water molecules minimize the desolvation contribution of the phosphate OP2 atom. Further investigations are needed to estimate the free energy contribution of these complex hydration patterns.

Indeed, binding energies are strongly dependent on environmental conditions. Although in the gas phase, anion-π binding strengths are considerable (≈130 kJ mol^{-1})[29], measurements in non-aqueous solvent established that the interaction of anions with single aromatic systems are very weak (binding free energies < 4 kJ mol^{-1})[30]. In hydrated biological systems, even weaker binding free energies can be anticipated. Hence, hydrophobic effects have clearly to be considered as key elements in understanding the right balance of interatomic forces at play in motifs comprising phosphate-π interactions.

Interestingly, capping interactions are not taken over by other charged groups like those from negatively charged Asp/Glu or positively charged Lys side chains that display a strong preference for forming salt bridges or interacting with nucleobase edges[21]. But the relatively frequent stacking of Arg side chains (or guanidinium groups) with aromatic rings deserves closer attention. For nucleic acids, the Arg to nucleobase stacking is often considered as belonging to the cation-π type and as such described as important for the stabilization of protein nucleic acid complexes[31-36] or the binding of nucleoside di/triphosphates[3]. Yet, the Arg side chains also frequently stack with the aromatic rings of the Phe, Tyr and Thr amino acids[37,38]. Further and despite the apparent electrostatic repulsion between them, guanidinium groups are known to form counter intuitive ion pair stacks[37,38].

Nucleobases are also involved in a large variety of stacking interactions. In addition to their above mentioned face-to-face contacts with neutral or charged nucleobases or aromatic amino acids, negatively charged phosphate groups and positively charged Arg side chains, non-planar sugar rings like those found in aminoglycoside antibiotics do stack over nucleobases[39].

This diversity of contacts supports the assumption that, as for the poorly hydrated Arg groups[40], that the stacking of these planar systems is dominated by hydrophobic effects rather than direct and strong cation-π interactions.

All these data support our introductory comment that “π interaction” terms have to be considered with caution. Indeed, these expressions can be quite misleading if they induce one to think that dispersion or other direct forces involving π-orbitals are driving the association process as often inferred from *in vacuum* experiments or quantum mechanical calculations neglecting solvent effects. Recently, through the use of molecular torsional balances, it has been demonstrated that the dispersion term is a small component of the aromatic stacking interaction in opposition to their dominant role *in vacuo*[41]. We suggest that the anion-π interactions we described are not directly related to the aromatic character of a given nucleobase but rather to its large hydrophobic surface. Anion-π interactions that should certainly be distinguished from cation-π interactions correspond thus to another component of the large array of hydrophobic forces at play in biomolecular systems[42].

Finally and given the current broad interest in “anion-π“ interactions, we wondered about their possible use in the design of biomolecular systems, drugs and drug binding sites. Since anions have no stacking affinity for aromatic amino acid rings as emphasized by our data, sulfate and phosphate receptors[43,44] were found to adopt different anion binding strategies that obey complex rules going beyond simple charge compensation considerations. Indeed, in proteins, about one third of all phosphate binding sites where found lacking metal ions or protonated amino acids at hydrogen bonding distance[44]. Such observations led to the development of new classes of neutral phosphate and sulfate receptors relying on optimizing their hydrogen-bonding abilities[43,44]. In order to develop new anion binding strategies, amino acids grafted with appropriate substituents (e.g. fluorinated rings) could be incorporated in the design of anion receptors to create suitable hydrophobic cavities[45].

Similarly, since nucleobases seem tolerant towards anions, RNA molecules could be designed for binding anionic or zwitterionic drugs embedding sulfate or phosphate groups. Although anions would generally prefer to bind to the Watson-Crick edge of nucleobases[20], two examples of biomolecular systems showing anion-π interactions with pyrimidine like fragments have been described. Flavin receptors were found to accommodate various anion types over their uridine-like pyrimidinedione ring[46] and anions like CN^- were even shown to act as efficient inhibitors of urate oxidase systems[47]. Synthetic receptors are currently also integrating anion-π interactions. A cationic SO_4^{2-} binding host molecule was characterized[48] where the sulfate establishes anion-π contacts with diazine rings. There, the trapped SO_4^{2-} retained a water molecule from its hydration shell that contributes to reducing its desolvation penalty as seen in the tetraloop systems described here (**Fig. 4c-f**) pointing towards more general binding schemes.

METHODS

Electrostatic potential map calculations. The geometries of the methylated Phe/Tyr/Thr aromatic rings and nucleobases were optimized using the Hartree-Fock method with the 6-31G** basis set. Based on these, electrostatic potential (ESP) maps[2,15] of three aromatic amino acid rings and six nucleobases **(Fig. 2)** were generated by mapping the 6-31G** electrostatic potentials onto surfaces of constant molecular electron density (0.002 $e^-/Å^3$) using the program SPARTAN (Wavefunction, Irwine, CA). Ψ stands for pseudouridine, a naturally modified uridine frequently occurring in RNA systems and especially in the conserved TΨC loops of tRNAs[5].

PDB survey. The Protein Data Bank (PDB) was searched for all nucleic acid x-ray diffraction structures (at resolutions ≤ 3.0 Å) displaying a nucleobase stacked with: *(i)* a negatively charged phosphate group; *(ii)* a negatively charged Asp or Glu side-chain; or *(iii)* free anions such as halides, nitrates, carbonates, acetates, phosphates or sulfates (we included in our survey forty-eight large ribosomal RNAs whose atom count overrides the limit imposed by the PDB format; given size considerations, only the first biological unit was considered). As of August 2015, the PDB contains ≈*5,000* nucleic acid crystal structures including complexes with proteins over a total of ≈*90,000* biomolecular structures at resolutions ≤ 3.0 Å.

Anion-π contacts in nucleic acids, as identified by the DSSR analysis tool[28], involve at least one atom belonging to a negatively charged group at less than 3.5 Å above or below a nucleobase area defined by the ring atoms. This 3.5 Å contact distance contrasts with the more permissive 4.5 Å distance used elsewhere[23] (see **Supplementary Discussion S1**). However, since we showed that the OP2 to nucleobase plane distance reaches a minimum around 3.5 Å (**Fig. 3c**), we believe that a 3.5 Å cut-off is sufficient for defining anion-π interactions involving oxyanions. Next to anion-π contacts, we similarly searched for all possible contacts involving nucleobases and monoatomic cations, ammonium (Lys) and guanidinium (Arg) groups. For characterizing amino acid to nucleobase stacking contacts, we used the SNAP analysis tool that is part of the 3DNA package[28]. In order to assess their correctness, all interactions involving positively charged as well as all rare interaction types involving negatively charged groups were visualized along with their electron densities (when available from the PDB). The Relibase+ tool[49] was used to search the PDB for anion-π contacts in proteins. When available, the electron densities associated with these contacts were visualized.

Residues having at least one atom with a B-factor exceeding 79 Å^2 were excluded from the search that, nonetheless, takes into account ≈*370,000* nucleotides. The search includes also residues generated by applying crystallographic symmetry operations. Non-redundant nucleotides displaying phosphate-π contacts were tagged as follows. If two nucleotides from different structures share a same residue number, chain code and trinucleotide sequence as well as sugar puckers, backbone dihedral and χ angle codes, they are considered as similar and the one with the best resolution is marked as non-redundant. Alike, if in a same structure two nucleotides share a same residue number and trinucleotide sequence (with different chain codes) as well as sugar puckers, backbone dihedral and χ angle codes, they are considered as similar and the first is marked as non-redundant. The former criteria are used to filter similar structures and the later for filtering structures with multiple related biological units. Note that it is impossible to completely eliminate redundancy from a dataset without eliminating at the same time significant data. Here, we provide an upper limit of a truly "non-redundant" set that marks already close to 3/4 of all nucleotides as redundant.

Molecular dynamics (MD) simulations. A total of 60 ns of MD simulations were performed for a tRNA^{Phe} structure (PDB code: 1EHZ; resolution: 1.9 Å) and a sarcin-ricin ribosomal fragment containing a GNRA tetraloop (PDB code: 4NLF; resolution: 1.0 Å). More information on the simulation protocols is provided in the **supplementary material**.

ACKNOWLEDGEMENTS

P.A. wishes to thank Prof. Eric Westhof for ongoing support as well as Dr Quentin Vicens for careful reading of the manuscript and Dr Xiang-Jun Lu and Prof. Neena Grover for helpful discussions. This work has been published under the framework of the LABEX: ANR-10-LABX0036_NETRNA and benefits from a funding from the state managed by the French National Research Agency as part of the program "Investments for the future". F.L. gratefully acknowledges financial support from the Polish Ministry of Higher Education and Science (Mobility Plus programme, Project No. 1103/MOB/2013/0). We are grateful to the Direction Informatique (Pole HPC) of the University of Strasbourg for backing this work by providing scientific support and access to computing resources.

AUTHOR CONTRIBUTIONS

P.A. and L.D. conceived the project. P.A. and L.D. designed the ESP calculations and PDB surveys and analysis. P.A., L.D. and F.L. designed the MD simulations and analysis. All of the coauthors wrote and edited the manuscript.

References

1. Salonen, L.M., Ellermann, M. & Diederich, F. Aromatic rings in chemical and biological recognition: energetics and structures. *Angew. Chem. Int. Ed. Engl.* **50**, 4808-4842 (2011).
2. Ma, J.C. & Dougherty, D.A. The cation-π interaction. *Chem. Rev.* **97**, 1303-1324 (1997).
3. Mahadevi, A.S. & Sastry, G.N. Cation-π interaction: its role and relevance in chemistry, biology, and material science. *Chem. Rev.* **113**, 2100-2138 (2013).
4. Krygowski, T.M., Szatylowicz, H., Stasyuk, O.A., Dominikowska, J. & Palusiak, M. Aromaticity from the viewpoint of molecular geometry: application to planar systems. *Chem. Rev.* **114**, 6383-6422 (2014).
5. Quigley, G.J. & Rich, A. Structural domains of transfer RNA molecules. *Science* **194**, 796-806 (1976).
6. Auffinger, P. & Westhof, E. An extended structural signature for the tRNA anticodon loop. *RNA* **7**, 334-341 (2001).
7. Gamez, P., Mooibroek, T.J., Teat, S.J. & Reedijk, J. Anion binding involving π-acidic heteroaromatic rings. *Acc. Chem. Res.* **40**, 435-444 (2007).
8. Frontera, A., Gamez, P., Mascal, M., Mooibroek, T.J. & Reedijk, J. Putting anion-π interactions into perspective. *Angew. Chem. Int. Ed. Engl.* **50**, 9564-9583 (2011).
9. Robertazzi, A., Krull, F., Knapp, E.W. & Gamez, P. Recent advances in anion–π interactions. *CrystEngComm* **13**, 3293-3300 (2011).
10. Chifotides, H.T. & Dunbar, K.R. Anion-π interactions in supramolecular architectures. *Acc. Chem. Res.* **46**, 894-906 (2013).
11. Wheeler, S.E. & Bloom, J.W. Toward a more complete understanding of noncovalent interactions involving aromatic rings. *J. Phys. Chem. A* **118**, 6133-6147 (2014).
12. Wheeler, S.E. & Bloom, J.W. Anion-π interactions and positive electrostatic potentials of N-heterocycles arise from the positions of the nuclei, not changes in the π-electron distribution. *Chem. Commun.* **50**, 11118-11121 (2014).
13. Martinez, C.R. & Iverson, B.L. Rethinking the term ‘‘pi-stacking’’. *Chem. Sci.* **3**(2012).
14. Bloom, J.W. & Wheeler, S.E. Taking the aromaticity out of aromatic interactions. *Angew. Chem. Int. Ed. Engl.* **50**, 7847-7849 (2011).
15. Mecozzi, S., West, A.P., Jr. & Dougherty, D.A. Cation-π interactions in aromatics of biological and medicinal interest: electrostatic potential surfaces as a useful qualitative guide. *Proc. Natl. Acad. Sci. USA* **93**, 10566-10571 (1996).
16. Correll, C.C. & Swinger, K. Common and distinctive features of GNRA tetraloops based on a GUAA tetraloop structure at 1.4 Å resolution. *RNA* **9**, 355-363 (2003).
17. Egli, M. & Sarkhel, S. Lone pair-aromatic interactions: to stabilize or not to stabilize. *Acc. Chem. Res.* **40**, 197-205 (2007).
18. Gottstein-Schmidtke, S.R. et al. Building a stable RNA U-turn with a protonated cytidine. *RNA* **20**, 1163-1172 (2014).
19. Auffinger, P. & Westhof, E. Water and ion binding around r(UpA)$_{12}$ and d(TpA)$_{12}$ oligomers - Comparison with RNA and DNA (CpG)$_{12}$ duplexes. *J. Mol. Biol.* **305**, 1057-1072 (2001).
20. Auffinger, P., Bielecki, L. & Westhof, E. Anion binding to nucleic acids. *Structure* **12**, 379-388 (2004).
21. Kondo, J. & Westhof, E. Classification of pseudo pairs between nucleotide bases and amino acids by analysis of nucleotide-protein complexes. *Nucleic Acids Res.* **39**, 8628-8637 (2011).
22. Paulini, R., Muller, K. & Diederich, F. Orthogonal multipolar interactions in structural chemistry and biology. *Angew. Chem. Int. Ed. Engl.* **44**, 1788-1805 (2005).
23. Chakravarty, S., Sheng, Z.Z., Iverson, B. & Moore, B. "η6"-type anion-π in biomolecular recognition. *FEBS Lett.* **586**, 4180-4185 (2012).
24. Jenkins, D.D., Harris, J.B., Howell, E.E., Hinde, R.J. & Baudry, J. STAAR: statistical analysis of aromatic rings. *J. Comput. Chem.* **34**, 518-522 (2013).
25. Breberina, L.M., Milcic, M.K., Nikolic, M.R. & Stojanovic, S.D. Contribution of anion-π interactions to the stability of Sm/LSm proteins. *J. Biol. Inorg. Chem.* **20**, 475-485 (2015).
26. Dauter, Z., Wlodawer, A., Minor, W., Jaskolski, M. & Rupp, B. Avoidable errors in deposited macromolecular structures: an impediment to efficient data mining. *IUCrJ* **1**, 179-193 (2014).
27. Dunitz, J.D. & Gavezzotti, A. How molecules stick together in organic crystals: weak intermolecular interactions. *Chem. Soc. Rev.* **38**, 2622-2633 (2009).

28. Lu, X.J., Bussemaker, H.J. & Olson, W.K. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.* (2015).
29. Zhang, J., Zhou, B., Sun, Z.R. & Wang, X.B. Photoelectron spectroscopy and theoretical studies of anion-π interactions: binding strength and anion specificity. *Phys Chem Chem Phys* **17**, 3131-3141 (2015).
30. Ballester, P. Experimental quantification of anion-π interactions in solution using neutral host-guest model systems. *Acc. Chem. Res.* **46**, 874-884 (2013).
31. Rooman, M., Lievin, J., Buisine, E. & Wintjens, R. Cation-π/H-bond stair motifs at protein-DNA interfaces. *J. Mol. Biol.* **319**, 67-76 (2002).
32. Gromiha, M.M., Santosh, C. & Suwa, M. Influence of cation–π interactions in protein–DNA complexes. *Polymer* **45**, 633-639 (2004).
33. Sathyapriya, R. & Vishveshwara, S. Interaction of DNA with clusters of amino acids in proteins. *Nucleic Acids Res.* **32**, 4109-4118 (2004).
34. Zhu, W. et al. The multiplicity, strength, and nature of the interaction of nucleobases with alkaline and alkaline earth metal cations: a Density Functional Theory investigation. *J. Phys. Chem. A* **108**, 4008-4018 (2004).
35. Imai, Y.N., Inoue, Y. & Yamamoto, Y. Propensities of polar and aromatic amino acids in noncanonical interactions: nonbonded contacts analysis of protein-ligand complexes in crystal structures. *J. Med. Chem.* **50**, 1189-1196 (2007).
36. Borozan, S.Z., Dimitrijevic, B.P. & Stojanovic, S.D. Cation-π interactions in high resolution protein-RNA complex crystal structures. *Comput. Biol. Chem.* **47**, 105-112 (2013).
37. Kubickova, A. et al. Guanidinium cations pair with positively charged arginine side chains in water. *J. Phys. Chem. Lett.* **2**, 1387-1389 (2011).
38. Vazdar, M., Vymetal, J., Heyda, J., CVondrasek, J. & Jungwirth, P. Like-charge guanidinium pairing from molecular dynamics and Ab Initio calculations. *J. Phys. Chem. A* **115**, 11193-11201 (2011).
39. Kondo, J., Koganei, M. & Kasahara, T. Crystal structure and specific binding mode of sisomicin to the bacterial ribosomal decoding site. *ACS Med. Chem. Lett.* **3**, 741-744 (2012).
40. Mason, P.E., Dempsey, C.E., Neilson, G.W., Kline, S.R. & Brady, J.W. Preferential interactions of guanidinum ions with aromatic groups over aliphatic groups. *J. Am. Chem. Soc.* **131**, 16689-16696 (2009).
41. Hwang, J.W. et al. How important are dispersion interactions to the strength of aromatic stacking interactions in solution? *Chem. Sci.* **6**, 4358-4364 (2015).
42. Pace, C.N., Scholtz, J.M. & Grimsley, G.R. Forces stabilizing proteins. *FEBS letters* **588**, 2177-2184 (2014).
43. Ravikumar, I. & Ghosh, P. Recognition and separation of sulfate anions. *Chem. Soc. Rev.* **41**, 3077-3098 (2012).
44. Hirsch, A.K., Fischer, F.R. & Diederich, F. Phosphate recognition in structural biology. *Angew. Chem. Int. Ed. Engl.* **46**, 338-352 (2007).
45. Salwiczek, M., Nyakatura, E.K., Gerling, U.I.M., Ye, S.J. & Koksch, B. Fluorinated amino acids: compatibility with native protein structures and effects on protein-protein interactions. *Chem. Soc. Rev.* **41**, 2135-2171 (2012).
46. Estarellas, C., Frontera, A., Quinonero, D. & Deya, P.M. Anion-π interactions in flavoproteins. *Chem. Asian J.* **6**, 2316-2318 (2011).
47. Estarellas, C., Frontera, A., Quinonero, D. & Deya, P.M. Relevant anion-π interactions in biological systems: the case of urate oxidase. *Angew. Chem. Int. Ed. Engl.* **50**, 415-418 (2011).
48. Galstyan, A., Sanz Miguel, P.J. & Lippert, B. Electrostatics plus O-π interactions rather than "directed" hydrogen bonding keep SO4(2-) in a triangular Pt3Pd3-tris(2,2'-bipyrazine) host. *Chemistry* **16**, 5577-5580 (2010).
49. Hendlich, M., Bergner, A., Gunther, J. & Klebe, G. Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *J. Mol. Biol.* **326**, 607-620 (2003).

Table 1 | Occurrences of non-redundant phosphate-π contacts and average OP1/2 to nucleobase plane distances (August 2015 PDB survey). All analyzed structures have resolutions ≤ 3.0 Å. Occurrences associated with the entire structural set are given in parenthesis (see Method section). The average OP1/2 to nucleobase plane distances are estimated from the non-redundant data sets and are given along with standard deviations.

Nucleobase	DNA		RNA		Total	Dist. (Å)
	OP1	OP2	OP1	OP2		
A	4 (4)	/	31 (65)	49 (97)	84 (166)	3.13 ± 0.17
C/C$^+$	4 (4)	1 (1)	23 (91)	37 (182)	65 (278)	3.07 ± 0.23
G	3 (3)	1 (1)	58 (172)	417 (2232)	479 (2408)	3.13 ± 0.17
T/U	11 (12)	1 (1)	73 (311)	318 (1418)	403 (1742)	3.09 ± 0.19
Ψ	/	/	/	15 (23)	15 (23)	2.94 ± 0.17
Total	22 (23)	3 (3)	185 (639)	836 (3952)	1046 (4617)	3.11 ± 0.18

The chosen contact criteria states that the OP1/2 to nucleobase plane distance has to be ≤ 3.5 Å. When both OP1 and OP2 atoms of the same PO_4 group are at contact distance, only the OP atom with the shortest contact distance is counted.

Table 2 | Occurrence of non-redundant "assisted" phosphate-π contacts involving OP2 atoms in RNA (August 2015 PDB survey). All analyzed structures have resolutions ≤ 3.0 Å. The OP2 to nucleobase plane distance is, as in **Table 1**, ≤ 3.5 Å. Occurrences associated with the complete structural set are given in parenthesis (see Method section).

Nucleo-base	PO_4 assisted contacts[a]	O2' assisted contacts[b]	PO_4 & O2' assisted contacts[c]	Other assisted contacts[d]	Non-assisted contacts[e]	Total
A	27 (48)	19 (46)	19 (46)	10 (16)	7 (11)	82 (167)
C/C^+	39 (177)	7 (17)	2 (42)	3 (7)	9 (37)	60 (280)
G	412 (2182)	1 (1)	17 (88)	24 (58)	20 (65)	474 (2394)
T/U/Ψ	271 (1234)	54 (268)	29 (146)	19 (45)	32 (57)	405 (1750)
Total	749 (3641)	81 (332)	67 (322)	56 (126)	68 (170)	1021 (4591)

[a] Involve a single PO_4 "assisted" phosphate-π contact (**Fig. 4a**) of intra- or intermolecular type.

[b] Involve a single O2' "assisted" phosphate-π contact (**Fig. 4b**) of intra- or intermolecular type.

[c] Involve a PO_4 and a O2' "assisted" phosphate-π contact of intra- or intermolecular type.

[d] The stacked PO_4 group is hydrogen bonded to a distant amino-acid and/or a nucleobase or involved in a rare intranucleotide stacking.

[e] The stacked PO_4 group is not involved in any additional contacts involving nucleotides and/or amino acids.

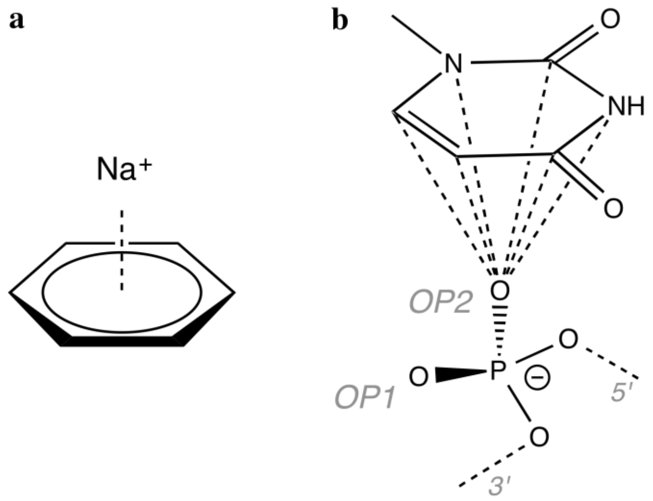


Figure 1 | Cation-π versus anion-π interactions. (**a**) Schematic of a cation-π interaction involving a benzene ring and a Na$^+$ ion. (**b**) Schematic of an anion-π interaction taking the form of a "capping" phosphate-π□contact as seen in tRNA anticodon and TΨC loops[6,17].

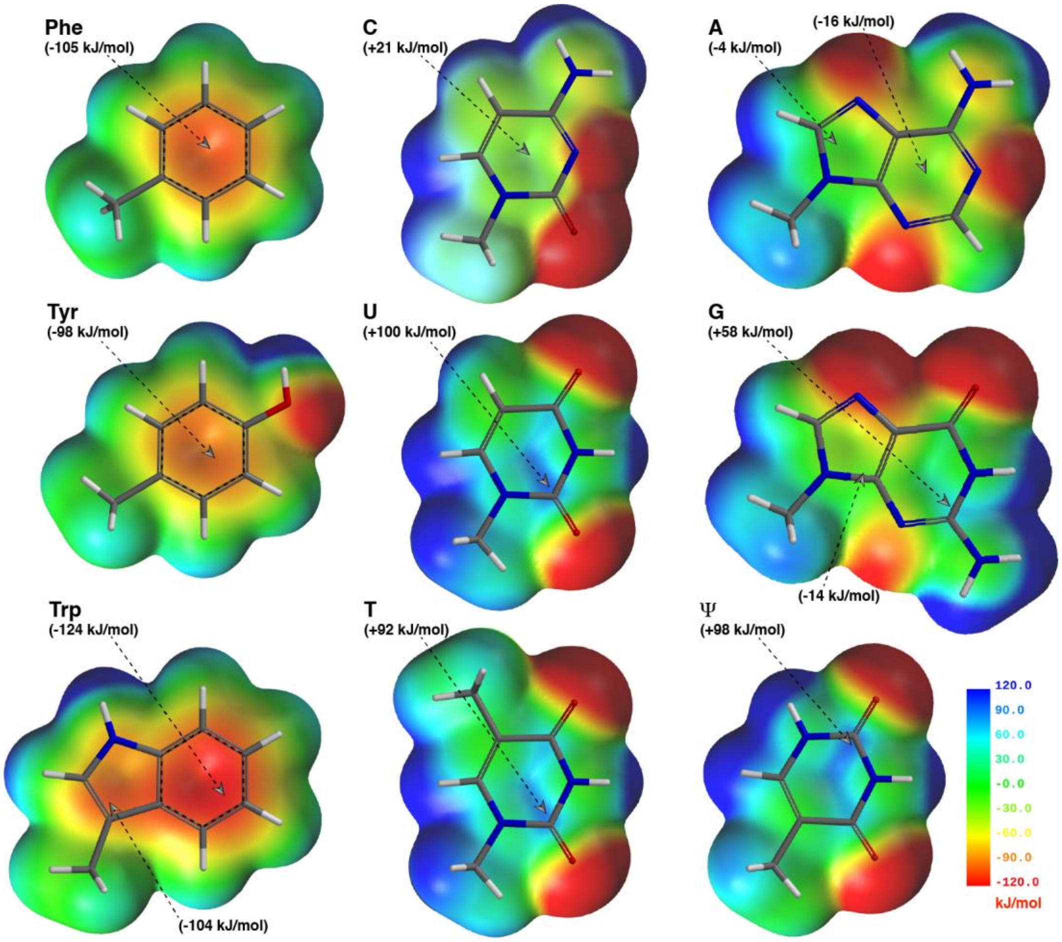


Figure 2 | Phe/Tyr/Trp amino acid versus C/U/T/A/G/Ψ nucleobase *Ab initio* electrostatic potential (ESP) maps. The potential energy values are limited to the ±120 kJ mol^{-1} range to emphasize the variation in ESP associated with substituent effects (note that some regions of the ESP maps, especially those associated with the heterocyclic nitrogen atoms and the heterocyclic substituents, lie beyond the ±120 kJ mol^{-1} range). Arrows mark the location of the highest (nucleobase) and lowest (amino acid) ESP values atop each of the five and six membered rings.

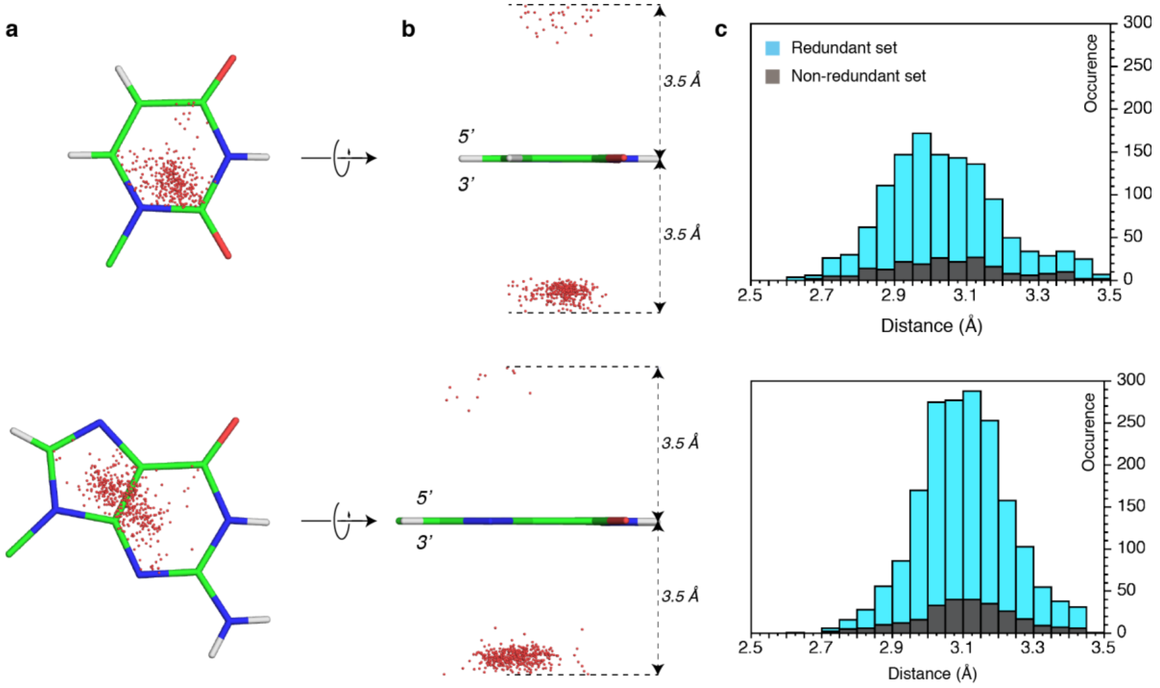


Figure 3 | Phosphate OP2 contacts to U and G nucleobases. (**a**) Top view (3' side) of the U and G nucleobases showing the positions of OP2 atoms above and below the nucleobase plane with a contact distance ≤ 3.5 Å (data are extracted from a non-redundant set). (**b**) 90° rotation of (**a**); the 3.5 Å boundaries are marked by arrows and dashed lines. (**c**) Histogram of the OP2 to nucleobase plane distances drawn from a set including all redundant (blue) and non-redundant contacts (grey; see Method section). The histograms cumulate the 3' and 5' side contacts.

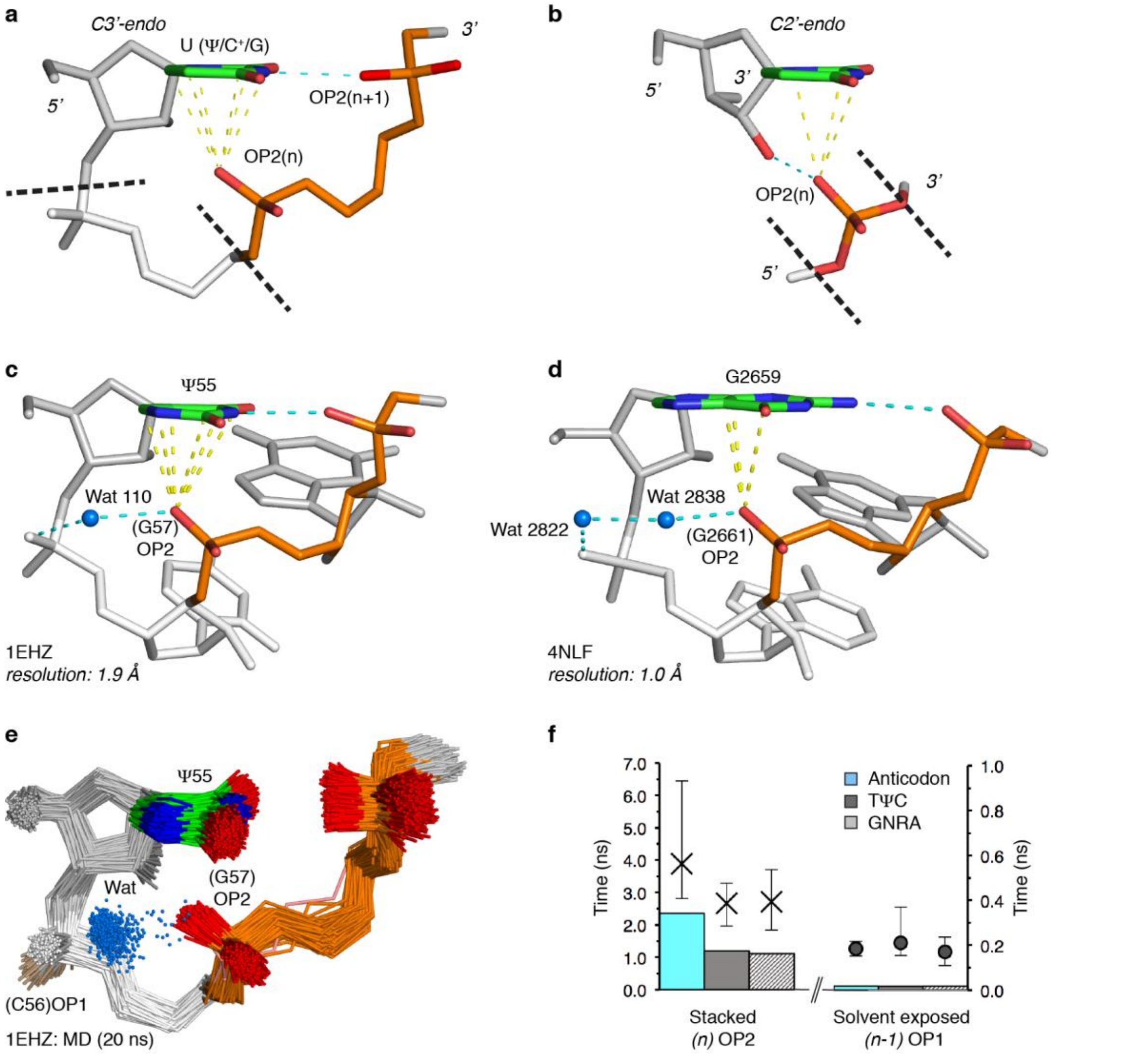


Figure 4 | Schematic representation of "assisted" phosphate-π contacts and MD simulation data related to PO4 hydration. (**a**) Phosphate-π interaction of the "PO4 assisted" type. The ribose adopts a *C3'-endo* pucker. The phosphate interaction module is drawn in orange (OP1/2 in red). Black dashed separators indicate that it is found in tetraloop-like but also in longer intra- and intermolecular motifs. The starting nucleobase is a U and can be replaced by a Ψ/C$^+$/G nucleobase, all bearing an imino group on their Watson-Crick edge. (**b**) Phosphate-π interaction of the "O2' assisted" type. The ribose adopts a *C2'-endo* pucker. The stacked phosphate interaction module is drawn in orange (OP1/2 in red). Black dashed separators indicate that it can be found in intra- and intermolecular motifs. The starting nucleobase is a Ψ. (**c**) Ψ55-OP2(G57) contact in the TΨC loop of a tRNA^{Phe} system (OP2-to-plane distance: 2.90 Å). A water molecule (in blue) bridges the OP2 atoms of nucleotides C56 and G57. (**d**) G2659-OP2(G2661) contact in the GNRA tetraloop of a sarcin-ricin system (OP2-to-plane distance: 3.11 Å). Two water molecules (in blue) bridge the OP2 atoms of nucleotides A2660 and G2661. In (**c**) and (**d**), all anionic oxygen to nucleobase atom distances ≤ 3.5 Å are materialized by dashed yellow lines.

Other hydrogen bonds are drawn in cyan. (**e**) Superimposition of 1000 structures extracted from a 20 ns MD trajectory of the tRNA TΨC loop. The water molecules hydrogen bonded to the stacked OP2 atom with d(Ow…OP2) ≤ 2.9 Å and (Ow-H…OP2) angle ≤ 135° are shown in blue. For clarity, a representation similar to (**a**) that hides the three last sugars and nucleobases of the loop has been used. The solvent exposed (C56)OP1 is drawn in wheat color. (**f**) Histogram of average hydrogen bond (HB) times for a sample of water molecules accounting for 80% of the hydration of the stacked *(n)*OP2 and solvent exposed *(n-1)*OP1 atoms (**see supplementary material**). The markers correspond to the average lifetime of the six longest bound waters; the longest and shortest lived waters are indicated by the top and bottom thin lines. The data are collected over three 20 ns MD trajectories of the anticodon, TΨC and GNRA loops.

Supporting Online Material

CATION-π VERSUS ANION-π INTERACTIONS — BIOMOLECULES CAN'T HAVE BOTH

Luigi D'Ascenzo[1], Filip Leonarski[1,2] & Pascal Auffinger[1*]

[1]Architecture et Réactivité de l'ARN, Université de Strasbourg, Institut de Biologie Moléculaire et Cellulaire du CNRS, 67084 Strasbourg, France

[2]Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

For PA, please forward correspondences to:
Telephone: (33) 388 41 70 49
FAX: (33) 388 60 22 18
e-mail: p.auffinger@ibmc-cnrs.unistra.fr

Methods S1 | Molecular Dynamics (MD) simulation

1) *Molecular dynamics protocols*

Molecular dynamics (MD) simulations were performed using the Amber 14 simulation package[1] and the Amber ff14SB force field (ff99+bsc0+chiOL)[2-4] with refined van der Waals parameters for phosphate oxygen atoms[5]. Parameters for tRNA modified nucleotides[6] were adapted for consistency with ff14SB and the grafted van der Waals parameters of the phosphate oxygen atoms. In order to prevent non-physical ion clustering, we used the Smith and Dang parameters for K^+ and Cl^- ions[7,8].

For the $tRNA^{Phe}$ simulations, the 1EHZ X-ray crystal structure solved at 1.93 Å resolution was used[9]. Wybutosine 37 was changed to 1-methylguanine; other modified bases were kept as in the original structure. The Mg^{2+}, Mn^{2+} and water molecules present in the crystallographic structure were ignored. The tRNA was put in a SPC/E water box, with box edges being at least 12 Å apart from the solute. The final simulation box contains one tRNA molecule, 17698 water molecules, 149 K^+ and 76 Cl^- ions to provide a ≈0.25 M ionic strength. The ion positions were randomized by using the CPPTRAJ program[10].

For the sarcin-ricin loop simulations, the 4NLF X-ray crystal structure solved at 1.00 Å resolution was used[11]. The 2'-trifluoromethylthio-2'-deoxycytosine residue 2667 was changed to cytosine. The SO_4^{2-} and water molecules present in the crystallographic structure were ignored. The sarcin-ricin loop was put in a SPC/E water box, with box edges being at least 15 Å apart from the solute. The final simulation box contains one sarcin-ricin loop, 7862 water molecules, 57 K^+ and 31 Cl^- ions to provide a ≈0.20 M ionic strength. The ion positions were randomized by using the CPPTRAJ program[10].

After equilibration, three 20 ns MD simulations were calculated for each RNA molecule with a 2 fs time step in a NPT ensemble. Berendsen thermostat[12] was used to keep temperature constant at 300 K.

2) *Estimation of water hydrogen bonded (HB) times*

Hydrogen bonded (HB) times of water molecules attached to specific OP1/2 atoms were calculated for each simulation[13-16]. They correspond to the total time a given water molecule is hydrogen bonded to the same solute atom over the entire simulation. Hydrogen bonding criteria are taken as $d(A...D) \leq 2.9$ Å and $(A...H\text{-}D) \geq 135°$; where A and D represent hydrogen bond acceptor and donor atoms, respectively. A stricter than usual distance criterion (2.9 Å instead of 3.2 or 3.5 Å) has been used with the intention to reduce the number of contacts established by transient (short HB times) water molecules.

To get to meaningful conclusions, these HB times require however filtering as their distributions are not homogenous. For the waters with long HB times attached to the stacked OP2 atoms, we removed 20% of waters with the shortest HB times (see **Fig. 4f**) to calculate the average values. On the other hand, for the waters with short HB times attached to the *(n-1)*OP1 atoms, we removed the 20% of waters with the longest HB times (see **Fig. 4f**) to calculate average values. As an example, including these outliers in the calculation of average HB times for the *(n-1)*OP1 atoms result in a doubling of the calculated values (from ≈20 to ≈40 ps). Although relatively short, present simulations demonstrate clearly that the water molecules bound to the stacked OP2 have residency times at least two orders of magnitude greater than those bound to a solvent exposed OP1 atoms. Extending the length of the simulations would result in better sampling and more accurate average HB times for the longest bound water molecules, but would not change present conclusions.

Discussion S1 | Misidentification of ion-π interactions in protein and nucleic acid systems

1) Electron density map visualization is required to assess "rare" non-covalent interactions

Crystallographic structures retrieved from the PDB correspond to a modeled view of crystallographic data. For a more complete evaluation of a crystallographic structure and for avoiding interpretation errors, it is recommended to visualize the associated electron density maps that are deposited in specialized databases such as the Electron Density Server (EDS)[17]. This should be mandatory for assessing the existence of non-covalent interactions for which various error types are recurrently identified in crystallographic structures[18-21].

Here, we highlight one of those. A paper discussing the occurrence of anion-π interactions in biological systems took as supportive example a perpendicular interaction involving a tyrosine ring and a glutamate residue (see Fig. 1b of Chakravarty et al.[22]; associated PDB code: 2R8O[23]). Yet, an examination of the $2F_o$-F_c EDS maps reveals that the densities associated with the Tyr residue are no longer visible at 1.5 σ **(Fig. S3a/b)** while the densities of the surrounding residues including water molecules are still visible at 3.0 σ. Hence, this solvent exposed tyrosine is poorly defined with respect to neighboring residues. In addition, this Tyr residue is defined as an outlier by the PDB validation report. Consequently, these electron density maps do not support the existence of an anion-π interaction. An examination of associated B-factors would not have been conclusive since the authors of the crystal structure report similar values for all the residues in this region that do not exceed 25 Å^2. As a comparison, we show an example of a well defined cation-π interaction were the densities of both the Lys and Tyr residues are clearly visible at 2.0 σ **(Fig. S3c/d)**

In a second example from an RNA system (PDB code: 3CUL[24]), an "ion" marked as K^+ is at 3.4 Å from a (C)N4 amino group and is much more likely a Cl^- ion or a water molecule **(Fig. S4)**. These two examples illustrate how cautiously structural studies should be conducted when dealing with rare occurrences of "new" non-covalent interactions since even high-resolution crystal structures are not void of major local data over-interpretations.

2) Use of "abnormally" large anion-to-ring cutoffs and/or insufficient geometrical characterization of the formed contacts

The use of large cutoffs (≤ 4.5 Å) for defining ion-π contacts has been often reported. These cutoffs go beyond already large cutoffs based on anion to carbon atom distances below the sum of van der Waals radii + 0.8 Å that are consequently in the 4.0 Å range for oxyanions[25,26]. In this report we show that the oxygen to ring distance reaches a minimum around 3.5 Å. Therefore, a 3.5 Å cut-off should be sufficient for oxyanions. The use of cutoffs exceeding 3.5 Å leads to the characterization of contacts that are generally not of the anion-π type as exemplified in following examples (note that cutoffs for F^-, Cl^-, Br^- or I^- have to be scaled appropriately):

- Large cutoffs were used to assess the contribution of cation-π interactions to the stability of protein-DNA complexes[27]. Since the authors based their analysis on calculated interaction energies, no binding distances were reported and consequently, this study difficulty supports the existence of cation-π interactions in protein-DNA complexes (interaction energies were calculated without taking into account solvation effects).
- In a related study by Gromiha et al., the criteria used by the authors prompted them to define cation-π interactions between an Arg residue and a G nucleobase that do not overlap in the crystal structure (see Fig. 1 of Gromiha et al.[28]). In line with our conclusions, these authors stressed the absence of cation-π contacts between Arg and nucleobases below 3.5 Å. Thus Arg residues are at best only stacked over nucleobases. With such large cutoffs, there is no evidence for bonding interactions.
- In a study by Robertazzi et al.[29], a cutoff of 5 Å between aromatic ring centroids and anions is used. These authors rationalized their choice by referring to the weaker resolutions of biomolecular structures compared to small chemical systems and advocate a large tolerance in cutoff choices. We believe that those criteria are too broad to be associated with anion-π interactions.
- Charkrabarty et al.[22] referred to a study[30] using a 4.5 Å cutoff between anions and aromatic ring centers where the authors stated that a 4.5 Å distance corresponds to "*the upper end of those observed in organic salt crystal structures*". Breberina et al.[31] used even larger cutoffs (≤ 7.0 Å) to evaluate anion-π interactions in a specific protein family.
- Wetmore et al. used an amino acid to nucleobase cutoff below 5 Å[32]. This resulted in the consideration of twice the number of pairs they would have analyzed with a cutoff of 3.5 Å (as estimated from their Figs. S9 and S10[32]).

It should also be noted that many authors analyzing biomolecular as well as smaller chemical systems use cutoffs based on the distance of the ion to the center of the aromatic ring[22]. We prefer using a more appropriate atom-to-nucleobase plane distance criterion. These distances are easily and accurately calculated for any nucleobase by the DSSR analysis tool[33].

3) Poor statistical significance for alleged anion-π interactions in proteins

Although the occurrence of interactions termed anion-π in supramolecular systems is no longer under debate, it seems bold, given current knowledge, to extrapolate them to protein systems. The few examples (≈5) highlighted in the study by Chakravarty et al.[22] (see Fig. 1b of the latter reference and **Fig. S3**) are really not convincing enough to support the existence of such interactions. However, these authors identified correctly anion-π interactions in nucleic acids.

A survey of contacts between aromatic amino acids and free anions such as Cl^-, Br^-, F^-, NO_3^- , ClO_4^- and PO_4^{n-} identified few "anion-π" contacts[29]. These contacts involved 22 "strong interactions" with Cl^- and six with phosphate groups, a number not large enough (in our opinion) to build a case given the hundreds of thousands of aromatic amino acids present in PDB structures. Moreover, the possibility that phosphate groups appear as HPO_4^{2-} or $H_2PO_4^-$ species and could form O-H…π interactions is not taken into consideration by the authors. Thus, these studies that identify only a few contacts are not supportive for the occurrence of anion-π interactions in proteins (see also the study by Gamez[34]).

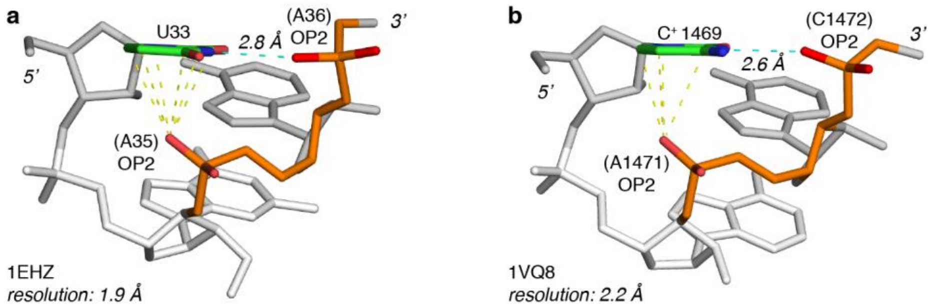


Figure S1 | Anticodon loop like structures starting with a neutral (U) and a charged (C^+) nucleobase are comparable. (**a**) U33-OP2(A36) contact in the anticodon loop of a $tRNA^{Phe}$ system (OP2-to-plane distance: 2.85 Å). (**b**) C^+1469-OP2(A1471) contact in an anticodon loop like structure extracted from the large subunit of the *Haloarcula marismortui* ribosomal system (OP2-to-plane distance: 2.97 Å). In (**a**) and (**b**), all anionic oxygen to nucleobase atom distances ≤ 3.5 Å are represented by dashed yellow lines. Hydrogen bonds are drawn in cyan. The phosphate interacting module is shown in orange (OP1/2 in red).

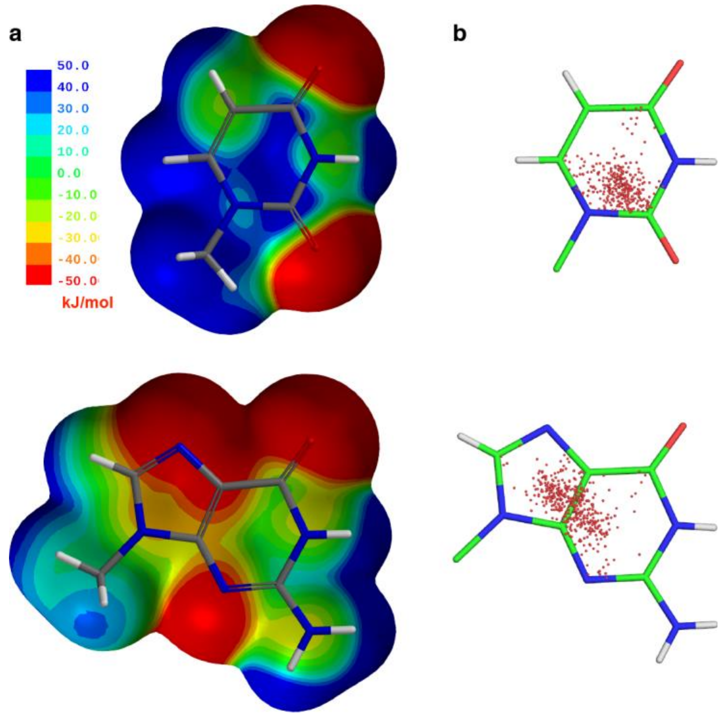


Figure S2 | Incidental correlation between the nucleobase electrostatic potential and the cluster of OP2 positions. (**a**) The U and G electrostatic potential surfaces are plotted with a ±50 kJ mol^{-1} scale (band spread of 10 kJ mol^{-1}). (**b** - similar to **Fig. 3b**). Top view (3' side) of the U and G nucleobases showing the positions of OP2 atoms positioned above and below the nucleobase plane with a ≤ 3.5 Å contact distance (data are extracted from a non-redundant set).

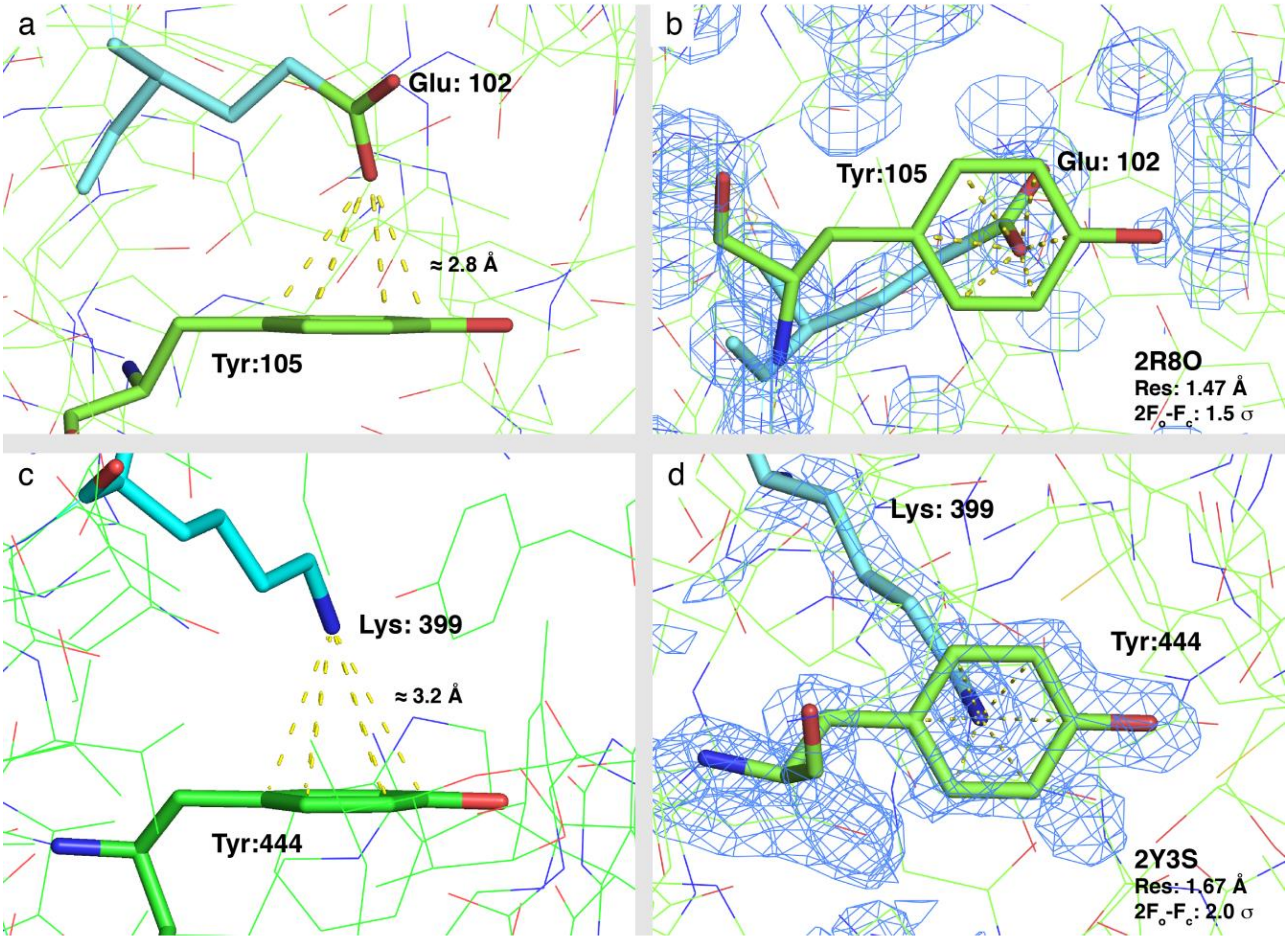


Figure S3 | Example of an anion-π misidentification and a correct identification of a cation-π interaction in protein X-ray crystal structures. Anionic oxygen and nitrogen to Tyr atom distances ≤ 3.5 Å are represented by dashed yellow lines. (**a**) This example has been extracted from a crystal structure[23] that is cited in a publication by Chakravarty et al.[22] emphasizing the existence of anion-π interactions in proteins. (**b**) The 2F$_o$-F$_c$ map demonstrates clearly that this Tyr residue is not associated with a strong electron density and consequently that the it is not forming an anion-π interaction. The Tyr residue is solvent exposed and probably adopting multiple conformations in the crystal. (**c**) In this example[35], the Tyr residue is involved in a cation-π interaction. (**d**) The 2F$_o$-F$_c$ map clearly shows that the Tyr and Lys residues are nicely fitted into appropriate densities in opposition to what is seen in (**b**).

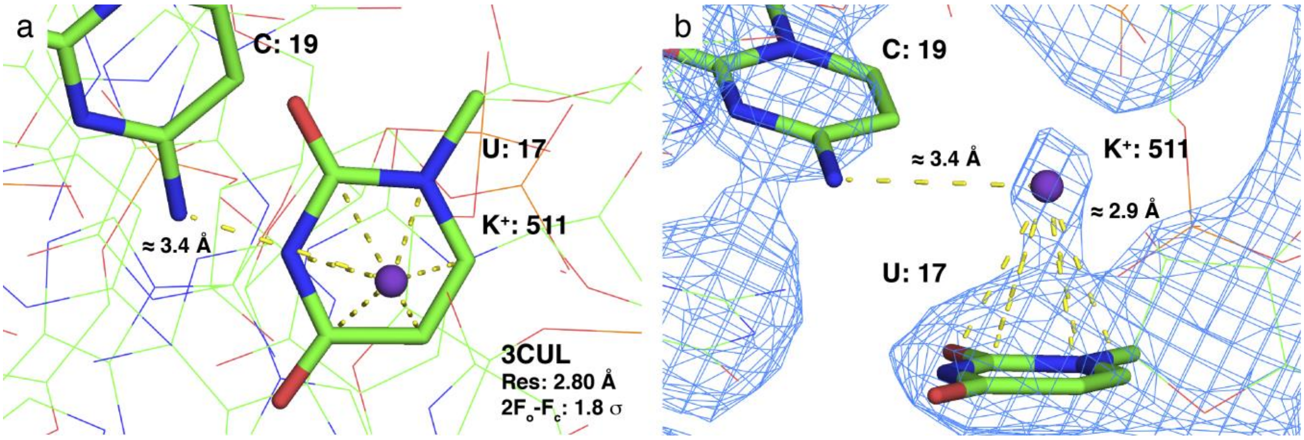


Figure S4 | Example of cation-π misidentification in a nucleic acid X-ray crystal structure. Hydrogen bonds and K^+ to nucleobase contacts ≤ 3.5 Å are represented by dashed yellow lines. (**a**) This example as been extracted from a crystal structure with PDB code 3CUL[24]. (**b**) The $2F_o$-F_c map demonstrates clearly that this K^+ ion occupies an anion[36,37] or water binding site associated with the amino group of C19. Three other dubious monovalent cation-π contacts are found in the PDB: *(i)* K^+: 703 (PDB code: 1DUL); *(ii)* K^+: 1246 (PDB code: 2UUB); *(iii)* Na^+: 403 (PDB code: 3OPI).

References

1. Case, D.A. et al. AMBER 15. (University of California, San Francisco, 2015).

2. Cheatham, T.E., Cieplak, P. & Kollman, P.A. A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.* **16**, 845-862 (1999).

3. Perez, A. et al. Refinement of the amber force field for nucleic acids. Improving the description of {alpha}/{gamma} conformers. *Biophys. J.* **92**, 3817-3829 (2007).

4. Zgarbova, M. et al. Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J. Chem. Theory. Comput.* **7**, 2886-2902 (2011).

5. Steinbrecher, T., Latzer, J. & Case, D.A. Revised AMBER parameters for bioorganic phosphates. *J. Chem. Theory Comput.* **8**, 4405-4412 (2012).

6. Aduri, R. et al. AMBER force-field parameters for the naturally occurring modified nucleosides in RNA. *J. Chem. Theory Comput.* **3**, 1464-1475 (2007).

7. Smith, D.E. & Dang, L.X. Computer simulations of NaCl association in polarizable water. *J. Chem. Phys.* **100**, 3757-3765 (1994).

8. Auffinger, P., Cheatham, T.E. & Vaiana, A.C. Spontaneous formation of KCl aggregates in biomolecular simulations: a force field issue? *J. Chem. Theory Comput.* **3**, 1851-1859 (2007).

9. Shi, H. & Moore, P.B. The crystal structure of yeast phenylalanine tRNA at 1.93Å resolution: a classic structure revisited. *RNA* **6**, 1091-1105 (2000).

10. Roe, D.R. & Cheatham, T.E. PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **9**, 3084-3095 (2013).

11. Kosutic, M. et al. Surprising base pairing and structural properties of 2'-trifluoromethylthio-modified ribonucleic acids. *J. Am. Chem. Soc.* **136**, 6656-6663 (2014).

12. Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F. & DiNola, A. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684-3690 (1984).

13. Auffinger, P. & Westhof, E. RNA hydration: three nanoseconds of multiple molecular dynamics simulations of the solvated tRNA^{Asp} anticodon hairpin. *J. Mol. Biol.* **269**, 326-341 (1997).

14. Auffinger, P. & Westhof, E. Water and ion binding around RNA and DNA (C,G)-oligomers. *J. Mol. Biol.* **300**, 1113-1131 (2000).

15. Auffinger, P. & Westhof, E. Water and ion binding around r(UpA)$_{12}$ and d(TpA)$_{12}$ oligomers - Comparison with RNA and DNA (CpG)$_{12}$ duplexes. *J. Mol. Biol.* **305**, 1057-1072 (2001).

16. Auffinger, P. & Westhof, E. Melting of the solvent structure around a RNA duplex: a molecular dynamics simulation study. *Biophys. Chem.* **95**, 203-210 (2002).

17. Kleywegt, G.J. et al. The Uppsala electron-density server. *Acta Cryst.* **D60**, 2240-2249 (2004).

18. Williams, L.D. Between objectivity and whim: nucleic acid structural biology. *Top. Curr. Chem.* **253**, 77-88 (2005).

19. Wlodawer, A., Minor, W., Dauter, Z. & Jaskolski, M. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS J.* **275**, 1-21 (2008).

20. Wlodawer, A., Minor, W., Dauter, Z. & Jaskolski, M. Protein crystallography for aspiring crystallographers or how to avoid pitfalls and traps in macromolecular structure determination. *FEBS J.* **280**, 5705-5736 (2013).

21. Dauter, Z., Wlodawer, A., Minor, W., Jaskolski, M. & Rupp, B. Avoidable errors in deposited macromolecular structures: an impediment to efficient data mining. *IUCrJ* **1**, 179-193 (2014).

22. Chakravarty, S., Sheng, Z.Z., Iverson, B. & Moore, B. "η6"-type anion-π in biomolecular recognition. *FEBS Lett.* **586**, 4180-4185 (2012).

23. Ishida, H., Huang, H., Yamniuk, A.P., Takaya, Y. & Vogel, H.J. The solution structures of two soybean calmodulin isoforms provide a structural basis for their selective target activation properties. *J. Biol. Chem.* **283**, 14619-14628 (2008).

24. Xiao, H., Murakami, H., Suga, H. & Ferre-D'Amare, A.R. Structural basis of specific tRNA aminoacylation by a small in vitro selected ribozyme. *Nature* **454**, 358-361 (2008).

25. Frontera, A., Gamez, P., Mascal, M., Mooibroek, T.J. & Reedijk, J. Putting anion-π interactions into perspective. *Angew. Chem. Int. Ed. Engl.* **50**, 9564-9583 (2011).

26. Chifotides, H.T. & Dunbar, K.R. Anion-π interactions in supramolecular architectures. *Acc. Chem. Res.* **46**, 894-906 (2013).

27. Wintjens, R., Lievin, J., Rooman, M. & Buisine, E. Contribution of cation-π interactions to the stability of protein-DNA complexes. *J. Mol. Biol.* **302**, 395-410 (2000).

28. Gromiha, M.M., Santosh, C. & Suwa, M. Influence of cation–π interactions in protein–DNA complexes. *Polymer* **45**, 633-639 (2004).

29. Robertazzi, A., Krull, F., Knapp, E.W. & Gamez, P. Recent advances in anion–π interactions. *CrystEngComm* **13**, 3293-3300 (2011).

30. Albrecht, M., Wessel, C., de Groot, M., Rissanen, K. & Luchow, A. Structural versatility of anion-π interactions in halide salts with pentafluorophenyl substituted cations. *J. Am. Chem. Soc.* **130**, 4600-4601 (2008).

31. Breberina, L.M., Milcic, M.K., Nikolic, M.R. & Stojanovic, S.D. Contribution of anion-π interactions to the stability of Sm/LSm proteins. *J. Biol. Inorg. Chem.* **20**, 475-485 (2015).

32. Wilson, K.A. et al. Landscape of π-π and sugar-π contacts in DNA-protein interactions. *J. Biomol. Struct. Dyn.*, 1-14 (2015).

33. Lu, X.J., Bussemaker, H.J. & Olson, W.K. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.* (2015).

34. Gamez, P. The anion–π interaction: naissance and establishment of a peculiar supramolecular bond. *Inorg. Chem. Front.* **1**, 35-43 (2014).

35. Carlson, J.C. et al. Tirandamycin biosynthesis is mediated by co-dependent oxidative enzymes. *Nature chemistry* **3**, 628-633 (2011).

36. Auffinger, P., Bielecki, L. & Westhof, E. Anion binding to nucleic acids. *Structure* **12**, 379-388 (2004).

37. D'Ascenzo, L. & Auffinger, P. Anions in nucleic acid crystallography. in *Nucleic Acid Crystallography: Methods and Protocols*, Vol. 1320 (ed. Ennifar, E.) 337-351 (Springer Science+Business Media, New-York, 2016).

3.2.2 Further remarks and outlook

There are two issues that need to be solved before the paper would be finished and ready for submission. Both are related to features observed in protein systems that are commonly accepted, but based on questionable data. Firstly, we showed numerous examples of claims of anion-π interactions in protein that do not seem convincing for diverse methodological or structural reasons (**Discussion S1 – Paper 1**). Moreover, there are emerging studies on how theoretical models implicating important roles for dispersion forces in biomolecular systems should be interpreted with caution in solvent-accessible systems (Yang et al. 2013; Hwang et al. 2015). The debate on the relevance of anion-π interactions is still open and as stated in a recent review on the subject: "*There is still a gap between theoretical studies and experimental evidence for those weak intermolecular interactions. Although computational studies, gas-phase experiments, and crystallographic results support the attractive nature of anion−π interactions, numerous investigations suggest that anion−π bonding very often is too weak to compete against other noncovalent interactions such as hydrogen bonding, electrostatic attraction, solvent effects, or dipole interactions in solution*" (Giese et al. 2015).

The second issue is related with the stacking of Arg guanidine groups with nucleobases, commonly considered as cation-π interaction. As discussed in the paper, the tendency of guanidine to stack with other aromatic cycles and to form even stacked Gua-Gua parallel pairs (Kubickova et al. 2011; Vazdar et al. 2011) infer that its cationic nature is not relevant for the stacking. Therefore, the usage of cation-π to refer to Arg-nucleobase stacking should be avoided.

3.3 O4’-π interactions in nucleic acids

The ribose O4’ atom is, after the OP2 phosphate atom, the second relevant backbone oxygen atom involved in stacking (or capping) interactions with nucleobases. Analogously to phosphate-π, O4’ atoms cap solvent exposed hydrophobic aromatic sides, participating in hydrophobicity-driven interactions. However, contrary to the phosphate oxygen, O4’ ribose oxygen is not charged and instead of anion-π it is involved in an oxygen-π interactions (identified as O4’-π). From an extensive PDB survey of O4’-π interactions in nucleic acids, more than 6,000 non-redundant (~16,000 total) O4’-π stacking have been identified. A stacking contact has been tagged as redundant when at least one of the two involved nucleotides is tagged as redundant (for redundancy criteria see **Methods**). The average O4’-plane distance for these occurrences is 3.1 ± 0.2 Å, a value similar to OP2-plane distance, but with a larger deviation due to the high heterogeneity of O4’-π contacts (**Fig.3.3**).

A remarkable observation is the relatively large number of O4’-π stacking contact with a distance ~ 2.9 Å, shorter than almost all phosphate-π stacking contacts. This surprising observation is probably related to the occurrence of numerous O4’ stacking within dinucleotide steps, where the local geometry constraints the ribose-nucleobase distance, but could also have further reasons connected to

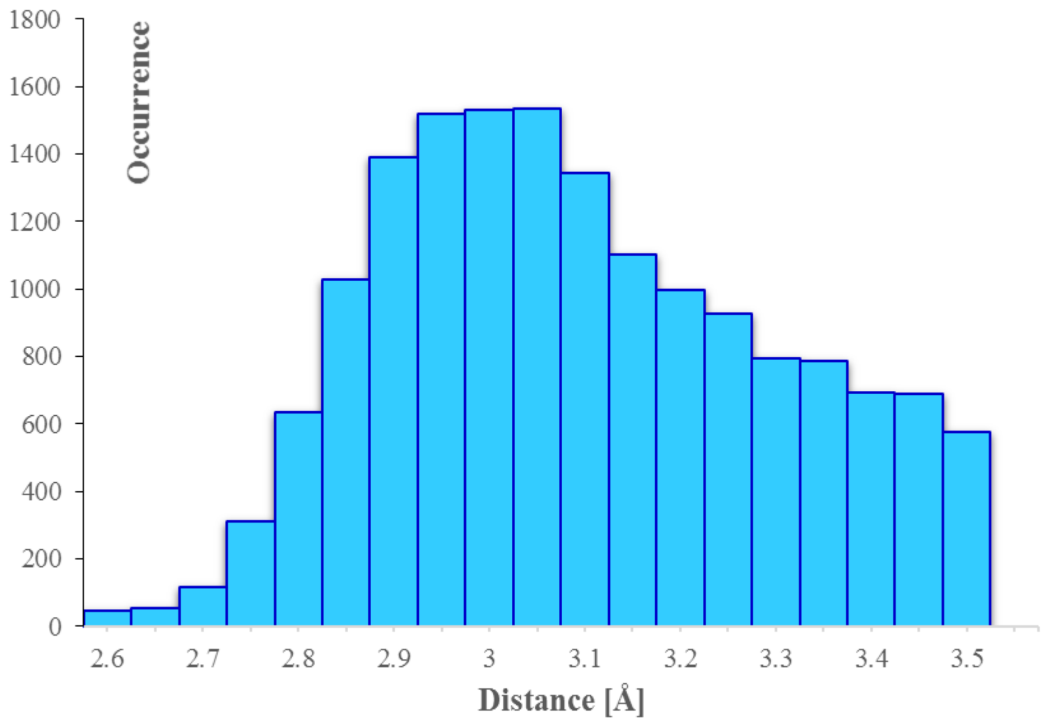


Figure 3.3. Histogram of O4’-nucleobase stacking distances. The values are more scattered compared to phosphate-nucleobase distances (**Fig. 3** in **Paper 1**), but a cutoff of 3.5 Å was still applied in order to avoid borderline or ill-defined stacking interactions. The mean distance value is 3.1 ± 0.2 Å.

the neutral state of the ribose versus the charged phosphate. Overall, the distribution of distances on the histogram and the large deviation infers that a diverse ensemble of families exist and that they are heterogeneous.

To define these families and characterize them, two criteria have been used. The first is based on the glycosidic bond χ of stacked nucleobase defining a base conformation; *anti* is largely preferred in nucleic acids, but locally it is possible to find rare *syn* conformations. Although rare, *syn* conformations have been found in active sites of functional RNAs (Sokoloski et al. 2011). The second criterion is the number of residues of distance between the nucleotides bearing the O4' and the stacked aromatic plane; four groups of contacts are defined when the distance is d = 0 (*consecutive*), $1 \leq d \leq 3$ (*short*), d > 3 (*long*) or when one of the two nucleotides belongs to a symmetry-generated molecule (*lattice*). An overview of the results found with these criteria can be found in **Table 3.1** and **Table 3.2**.

Table 3.1. O4'-π stacking interactions in nucleic acids, organized by category. Data from PDB structures of RNA and DNA at resolution ≤ 3.0 Å, July 2016 survey. Non-redundant total for each category are presented together with numbers without redundancy considerations (given in parenthesis).

Category	DNA		RNA	
	syn	*anti*	*syn*	*anti*
Consecutive	439 (686)	75 (114)	465 (1090)	945 (2762)
Short	6 (7)	54 (70)	66 (149)	786 (2613)
Long	7 (7)	84 (124)	133 (336)	2172 (6806)
Lattice	3 (3)	284 (496)	9 (12)	395 (765)
Total	455 (703)	497 (804)	673 (1587)	4298 (12946)

Table 3.2. O4'-π stacking interactions in nucleic acids, organized by stacked nucleobase. Data from PDB structures of RNA and DNA at resolution ≤ 3.0 Å, July 2016 survey. Non-redundant total for each category are presented together with numbers without redundancy considerations (given in parenthesis).

Nucleobase	DNA		RNA	
	syn	*anti*	*syn*	*anti*
A	52 (68)	88 (149)	372 (911)	2395 (7857)
G	395 (627)	225 (378)	240 (587)	814 (2217)
C	1 (1)	77 (111)	19 (22)	609 (1524)
U/T	7 (7)	104 (166)	42 (67)	480 (1348)
Total	455 (703)	497 (804)	673 (1587)	4298 (12946)

In DNA *consecutive* contacts involving purines in *syn* are the most abundant occurrence, due to the O4’-π stacking present in CpG steps of Z-DNA helices (Wang et al. 1981). *Lattice* contact are also numerous, almost equally distributed among all four nucleobases in *anti*, due to their involvement in helical fragments. The RNA:DNA ratio of occurrence of O4’-π interactions is ~ 10:1 and RNA stacking contacts are much more diverse. The large number of *long* range contacts involving adenines is a clear indication of the participation of O4’-π contacts in long-range interactions such as A-minor motifs, abundant in large ribosomal structures (Nissen et al. 2001). In fact, adenines in *syn* or *anti* conformer cap O4’ atoms more than all the other bases combined. Regarding *syn* stacked nucleobases, a clear abundance of *consecutive* occurrences is reminiscent of the Z-DNA among DNA structures and is in fact related with the remarkable occurrence of Z-DNA dinucleotide steps in RNA structures. More details on the O4’-π stacking families identified in RNA and DNA follow.

3.3.1 Lattice O4’-π in DNA and RNA

The *lattice* O4’-π interactions involve the capping of a nucleobase by the ribose of a symmetry-generated molecule, or vice versa. No clear preference for a particular nucleotide emerged from the survey and the stacked nucleobase is almost exclusively in *anti*. These contacts are ubiquitous in all types of DNA structures and are often found to cap otherwise exposed nucleobases on the helix edges. Their abundance infers remarkable structural roles, such as to protect exposed hydrophobic nucleobase sides and doing so promoting crystal packing. A stabilization effect promoted by crystal packing could be especially useful during the crystallization of short nucleic acids or otherwise unstable fragments. O4’-π contact of the *lattice* type are mainly found in helical 3’ ends, probably due to the presence of an additional phosphate group on the 5’ end that can hinder the approach of symmetric molecules (**Fig 3.4**).

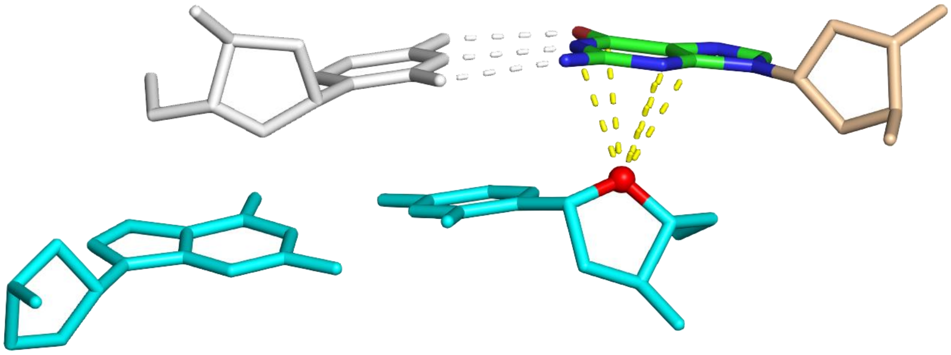


Figure 3.4. A *lattice* O4’-π stacking interaction in DNA. The stacking contact involves the ribose sugar of a symmetry-generated molecule, represented in cyan. Interatomic distances < 3.5 Å are represented by dashed lines. The atom-based colored guanine is residue 12 of chain B in PDB: 5CL8 res.: 1.4 Å; (Mullins et al. 2015).

3.3.2 *O4'-π stacking in DNA*

Among the ~1500 occurrences of O4' stacking in DNA, the ratio between *syn* and *anti* is close to 1:1, contrary to RNA where *anti* prevails ~10:1 over *syn*. In DNA, *syn* nucleobases are mostly involved in consecutive stacking, while *anti* residues tend to cap with symmetry-generated molecules. Guanines are most recurrent in DNA interactions, while adenines are most frequent in RNA. These two characteristic of DNA O4'-π stacking are the result of their involvement in the left-handed Z-DNA structures, where the large slide value between C=G stacked pairs has been already reported to induce the capping of guanine by cytosine riboses (Wang et al. 1981). The stacking is possible because the guanine of Z-DNA CpG steps assumes the rare *syn* conformer (**Fig. 3.5**; Rich 2004).

However, dinucleotide steps adopting a Z-DNA conformation are not limited to CpG but can also include TpA steps, with A in *syn* and the same local topology (**Fig 3.6**; (Wang et al. 1984).

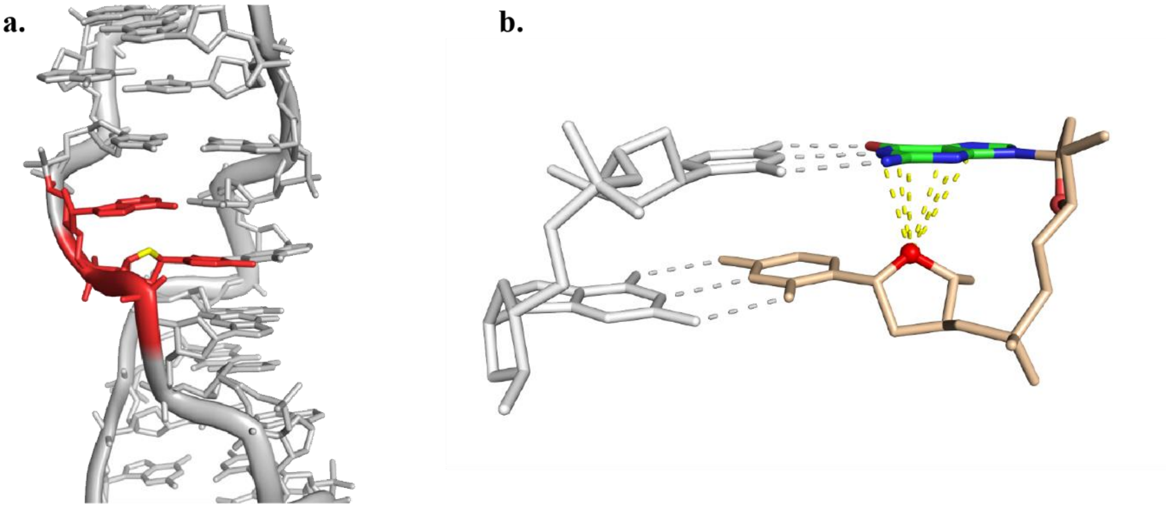


Figure 3.5. O4'-π stacking in a Z-DNA CpG step. (a) A CpG step (red) in the context of a Z-DNA helix (PDB: 1ICK res.: 0.95 Å; Dauter and Adamiak 2001) **(b)** Close-up showing the stacking between the cytosine ribose and the guanine in *syn*. Interatomic distances < 3.5 Å are represented by dashed lines. The atom-based colored guanine is residue 14 of chain B.

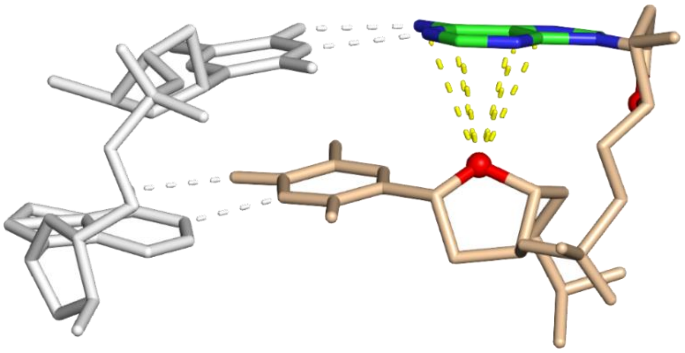


Figure 3.6. O4’-π stacking in a Z-DNA TpA step. Adenine adopts, analogously to G in CpG steps, a *syn* conformation. Interatomic distances < 3.5 Å are represented by dashed lines. The atom-based colored adenine is residue 4 of chain A in PDB: 1VTW res.: 1.2 Å; (Wang et al. 1984).

3.3.3 *O4’-π stacking in RNA*

The most remarkable group among the ~14,000 O4’-π contacts in RNA has been found in dinucleotide steps (*consecutive*) analogous to Z-DNA steps, involving a *syn* stacked nucleobase. These ~1000 contacts have been identified firstly as part of the CpG steps characterizing the 3’ end of UN<u>CG</u> tetraloops (**Fig. 3.7**), then found ubiquitously inside double and single stranded RNAs.

We named these steps collectively “Z-DNA like” or “Z-like” fragments and highlighted their implication in RNA folding, RNA/protein recognition and immune response (**Paper 2**).

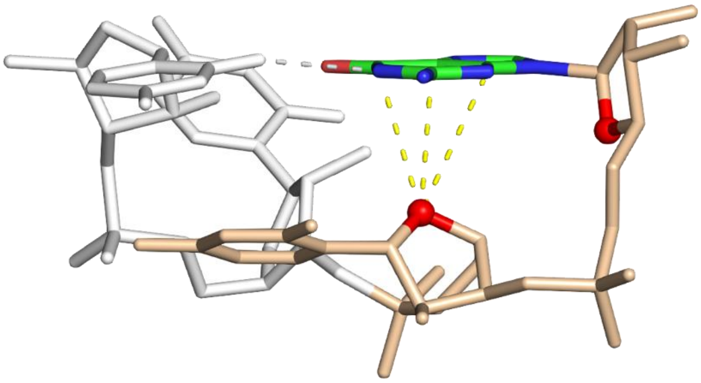


Figure 3.7. O4’-π stacking in a UU<u>CG</u> tetraloop. The dinucleotide step highlighted in wheat is analogous to Z-DNA CpG steps. Interatomic distances < 3.5 Å are represented by dashed lines. The atom-based colored guanine is residue 19 of chain B in PDB: 2HW8 res.: 2.1 Å; (Tishchenko et al. 2006).

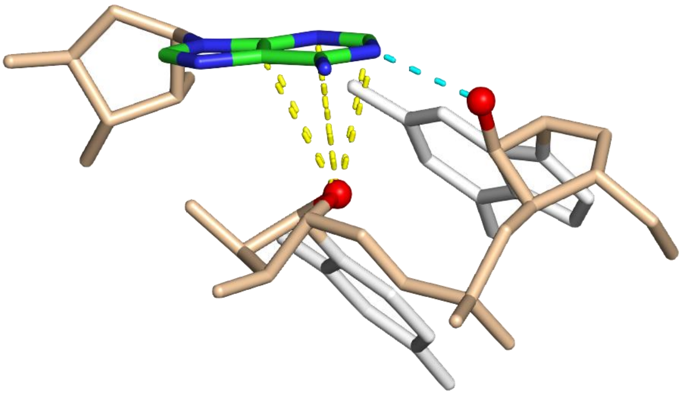


Figure 3.8. Long range O4’-π assisted stacking in the ribosome. The stacked adenine interacts with the O2’ hydroxyl of a base *n* and is stacked with the ribose of base *n-1*. Interatomic distances < 3.5 Å are represented by dashed lines. The atom-based colored adenine is residue 498 of chain BA in PDB: 4YBB; res.: 2.1 Å; (Noeske et al. 2015).

However, the largest amount of O4’-π contacts in RNA are part of *long* or – less frequently – *short* interactions, where a nucleobase (often adenine) interacts with its Watson-Crick side with a O2’ atom of a residue *n* and is stacked with O4’ ribose of residue *n-1*, forming an “assisted” O4’ stacking (**Fig. 3.8**).

Among these assisted stacking contacts there are contacts where the adenine participates in class I A-minor motifs (Nissen et al. 2001), thus binding the O2’ in the minor groove of a G=C pair, while being stacked over the ribose of the nucleotide preceding the pair within the helix (**Fig. 3.9**).

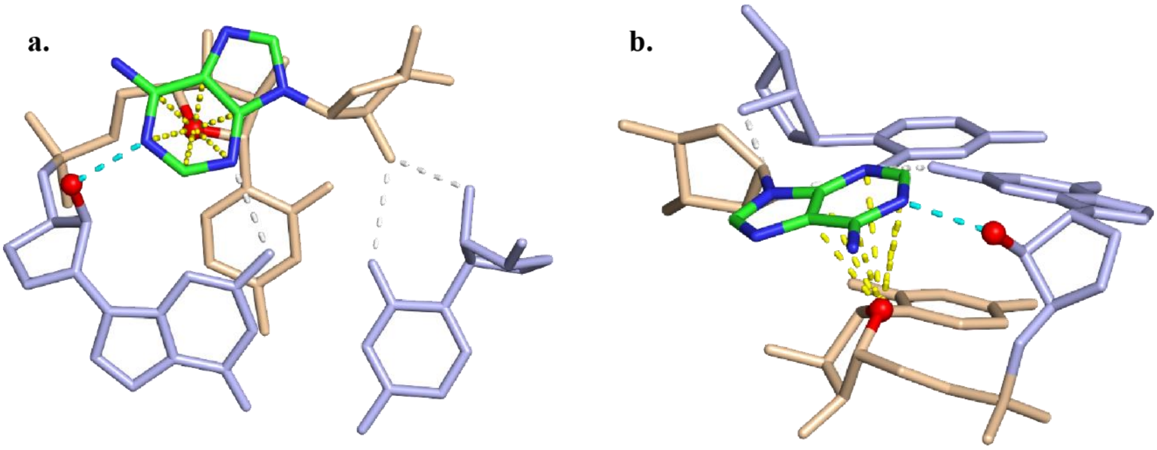


Figure 3.9. O4’-π stacking involved in type I A-minor contact inside the ribosome. (a) Perpendicular and **(b)** side view of the O4’-π stacking, assisted by hydrogen bonds between the adenine and the minor groove of the G=C pair (light blue). Interatomic distances < 3.5 Å are represented by dashed lines. The atom-based colored adenine is residue 574 of chain BA in PDB: 4YBB; res.: 2.1 Å; (Noeske et al. 2015).

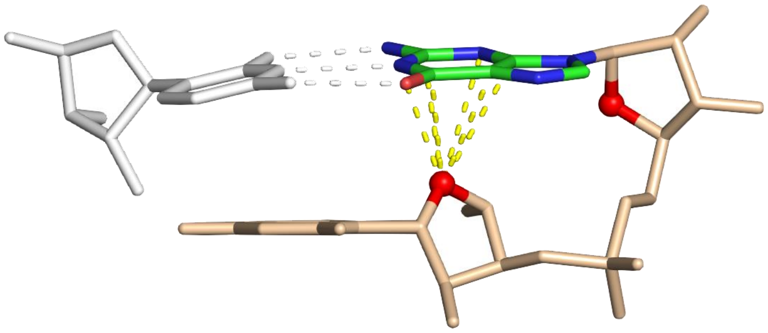


Figure 3.10. O4’-π stacking in in a UpG step. The guanine in *anti* is stacked with the uridine ribose and making a long-range base pair with a cytosine. Interatomic distances < 3.5 Å are represented by dashed lines. The atom-based colored guanine is residue 757 of chain DA in PDB 4YBB; res.: 2.1 Å (Noeske et al. 2015).

The adenine residues involved in A-minor motifs do not systematically show O4’-π contacts and sometimes the capping interaction is present even in other A-minor classes. Overall, their abundance in ribosomal structures and their relevance for their architecture is a strong link to the fact that adenines are by far the most common nucleobase stacked with O4’ atoms in RNA.

A final family of O4’-π interactions include *consecutive* contacts where a guanine in *anti* is stacked over the ribose of the previous nucleotide and simultaneously makes a long range Watson-Crick base pair with a cytosine (**Fig. 3.10**). The local structure of this dinucleotide step is reminiscent of the UpG step belonging to CUUG tetraloops especially found in ribosomes, among the most relevant RNA tetraloop families (Jucker and Pardi 1995b).

Data on phosphate-π stacking inside GNRA tetraloops and O4’-π stacking involved in both UNCG and CUUG tetraloops infer that the stacking of backbone oxygen with nucleobase is a determinant structural feature of tetraloops. Even if these capping interactions do not determine the fold by themselves, they are significant to modulate it. This *secondary* structural role for O4’-π stacking is even more important for long range interactions such as lattice contacts or A-minor motifs.

3.3.4 Future outlook of O4’-π stacking interactions in RNA

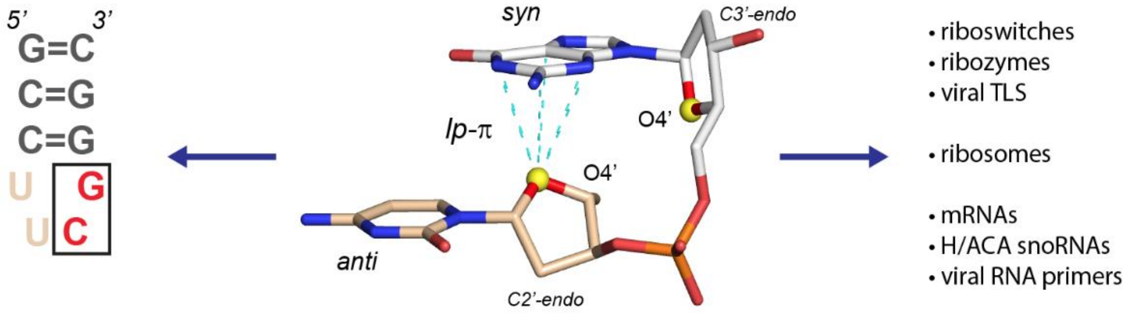
The results presented here on O4’-π interactions in nucleic acids are a starting point for future more in-depth analysis and classifications. Although several other families could emerge using different or more precise structural criteria, I am confident that the group presented here are the most significant for the structure and functions of nucleic acids. One of the most intriguing questions that stay unanswered is the occurrence of O4’-plane distances around 2.9 Å, a very short value for stacking contacts. This feature alone could prompt novel speculations about the strength and the role of stacking interactions in biomolecules and can help to define an “interaction hierarchy” that would describe also *in vivo* conditions. Nevertheless, O4’-π contact constitute a member of the rare non-covalent interactions structuring biomolecules that needs to be better rationalized to be described in MD force fields and other theoretical techniques.

3.4 “Z-DNA like” motifs in RNA

The occurrence of CpG steps within UNCG tetraloops analogous to Z-DNA is just a starting point of a more interesting consideration. In fact, these steps are ubiquitous in RNA, can take the form of NpN dinucleotide steps and are found in remarkable structural and functional locations.

3.4.1 Paper 2. *“Z-DNA like” fragments in RNA: a novel structural motif with implications for folding and RNA/protein recognition and immunology* (Nucleic Acid Res, 2016)

Graphical abstract



The occurrence of d(CpG) dinucleotide steps in left-handed DNA helices is familiar to us since the pioneering work of Alexander Rich who solved the first Z-DNA crystal structure. However, it is largely unrecognized that almost all dinucleotide sequences can adopt “left-handed” conformations in DNA and RNA in double as well as single stranded contexts. As such, we found that “Z-like” steps (Z-steps) are part of UNCG tetraloops but also of small and large RNAs including riboswitches, ribozymes and ribosomes. They are also occasionally involved in long-range base pairs significant to maintain functional folds, such as in a viral tRNA-like structure. The Z-steps involve the coexistence of several rare conformational features: (i) a C2’-*endo* ribose pucker, (ii) a *syn* nucleotide and (iii) a O4’-π stacking between the nucleobase in *syn* and a ribose O4’ atom. We characterized this O4’-π stacking interaction and obtained oxygen-base plane distances ~2.9 Å, which infer a bonding character. This stacking, together with the other unusual structural features, is individually known to induce a conformational stress in the RNA backbone that is associated with a slow kinetic of formation. With the characteristics combined, the stress imposed to the backbone might be significantly larger and account for the low occurrence of the Z-steps (≈0.1% of all dinucleotide steps). As an outcome, Z-steps are probably associated with slow kinetics and once formed could lock specific folds through unique backbone kinks and turns or long-range contacts. Besides their structural role, we found that several regulatory proteins recognize and bind Z-steps in single stranded RNA in a sequence-specific fashion. Even more interestingly, various interferon-induced proteins involved in immunology response recognize these steps through non-sequence specific contacts with the backbone, either in double stranded Z-RNA but also in single stranded Z-steps. Altogether, Z-like steps are rare but specific multifunctional motifs that are significant for RNA/DNA structure and function, are key to some immunologic response mechanisms and represent a further unforeseen example of the stunning diversity of motifs present in the assembly of RNA systems.

5944–5956 *Nucleic Acids Research, 2016, Vol. 44, No. 12* *Published online 5 May 2016*
doi: 10.1093/nar/gkw388

'Z-DNA like' fragments in RNA: a recurring structural motif with implications for folding, RNA/protein recognition and immune response

Luigi D'Ascenzo[1], Filip Leonarski[1,2], Quentin Vicens[1,*] and Pascal Auffinger[1,*]

[1]Architecture et Réactivité de l'ARN, Université de Strasbourg, Institut de Biologie Moléculaire et Cellulaire du CNRS, Strasbourg 67084, France and [2]Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

Received March 07, 2016; Revised April 10, 2016; Accepted April 28, 2016

ABSTRACT

Since the work of Alexander Rich, who solved the first Z-DNA crystal structure, we have known that d(CpG) steps can adopt a particular structure that leads to forming left-handed helices. However, it is still largely unrecognized that other sequences can adopt 'left-handed' conformations in DNA and RNA, in double as well as single stranded contexts. These 'Z-like' steps involve the coexistence of several rare structural features: a *C2'-endo* puckering, a *syn* nucleotide and a lone pair–π stacking between a ribose O4' atom and a nucleobase. This particular arrangement induces a conformational stress in the RNA backbone, which limits the occurrence of Z-like steps to ≈0.1% of all dinucleotide steps in the PDB. Here, we report over 600 instances of Z-like steps, which are located within r(UNCG) tetraloops but also in small and large RNAs including riboswitches, ribozymes and ribosomes. Given their complexity, Z-like steps are probably associated with slow folding kinetics and once formed could lock a fold through the formation of unique long-range contacts. Proteins involved in immunologic response also specifically recognize/induce these peculiar folds. Thus, characterizing the conformational features of these motifs could be a key to understanding the immune response at a structural level.

INTRODUCTION

Diversity in shape between RNA and DNA is striking. Although DNA can adopt A and B helical forms, RNA double strands are never of the B-form, due to the ribose preference for a *C3'-endo* over a *C2'-endo* pucker. Yet, both DNA and RNA can adopt a left-handed Z-form in which *C2'-endo* and *C3'-endo* alternate along a CpG sequence (1).

Historically, Z-DNA was crystallized before A-DNA, B-DNA and Z-RNA (2–4). Its structural properties and in particular the repeated *5'*-pyrimidine-purine-*3'* dinucleotide step along the helix with a purine in *syn* (Figure 1)*,* were described in detail (1,2,5–9). The most frequently crystallized Z-DNA dinucleotide step is CpG, but other Z-DNA steps have been described (1,10,11). However, because the *in vitro* formation of both Z-DNA and Z-RNA usually requires a high ionic strength or specific nucleotide modifications, it was assumed for a long time that Z-forms were mere structural artifacts (3,4).

Although both Z-DNA and Z-RNA have been known to be immunogenic since the 1980s, their biological role was questioned (4). We now know of four families of Z-DNA binding proteins that are all involved in the innate immune response such as the interferon induced form of the RNA editing enzyme ADAR1, the innate immune system receptor DLM-1, the fish kinase PKZ and the pox-virus inhibitor of interferon response EL3 (12,13). These proteins were found to recognize Z-DNA in a conformation-specific manner since most of the contacts with the protein involve backbone atoms (13,14). Moreover, evidence that some of these protein domains interact with Z-RNA *in vitro* have been gathered, which raised issues related to the *in vivo* role of this RNA form (12,15–17).

Here, we report the unanticipated occurrence of Z-like dinucleotide steps at key locations in single stranded RNA regions following a first identification in CUG-regulator binding proteins (18). We also highlight how r(U/ApA) steps are found more frequently than r(CpG) steps. Since our goal is to better characterize rare conformational features in RNA, we examine in detail the structure of what we refer to as a 'Z-like' motif, in particular within the context of r(UNCG) tetraloops where it has never been described although it is an essential component of this fold. We find that the Z-like motif contains a ribose-base or lone pair–π (lp–π) stacking that consists in the close contact of the *5'*-ribose O4' atom with the six-membered *3'*-guanine ring as observed in Z-DNA (Figure 1). This ribose-base

[*]To whom correspondence should be addressed. Tel: +33 388 41 70 49; Fax: +33 388 60 22 18; Email: p.auffinger@ibmc-cnrs.unistra.fr
Correspondence may also be addressed to Quentin Vicens. Email: q.vicens@ibmc-cnrs.unistra.fr

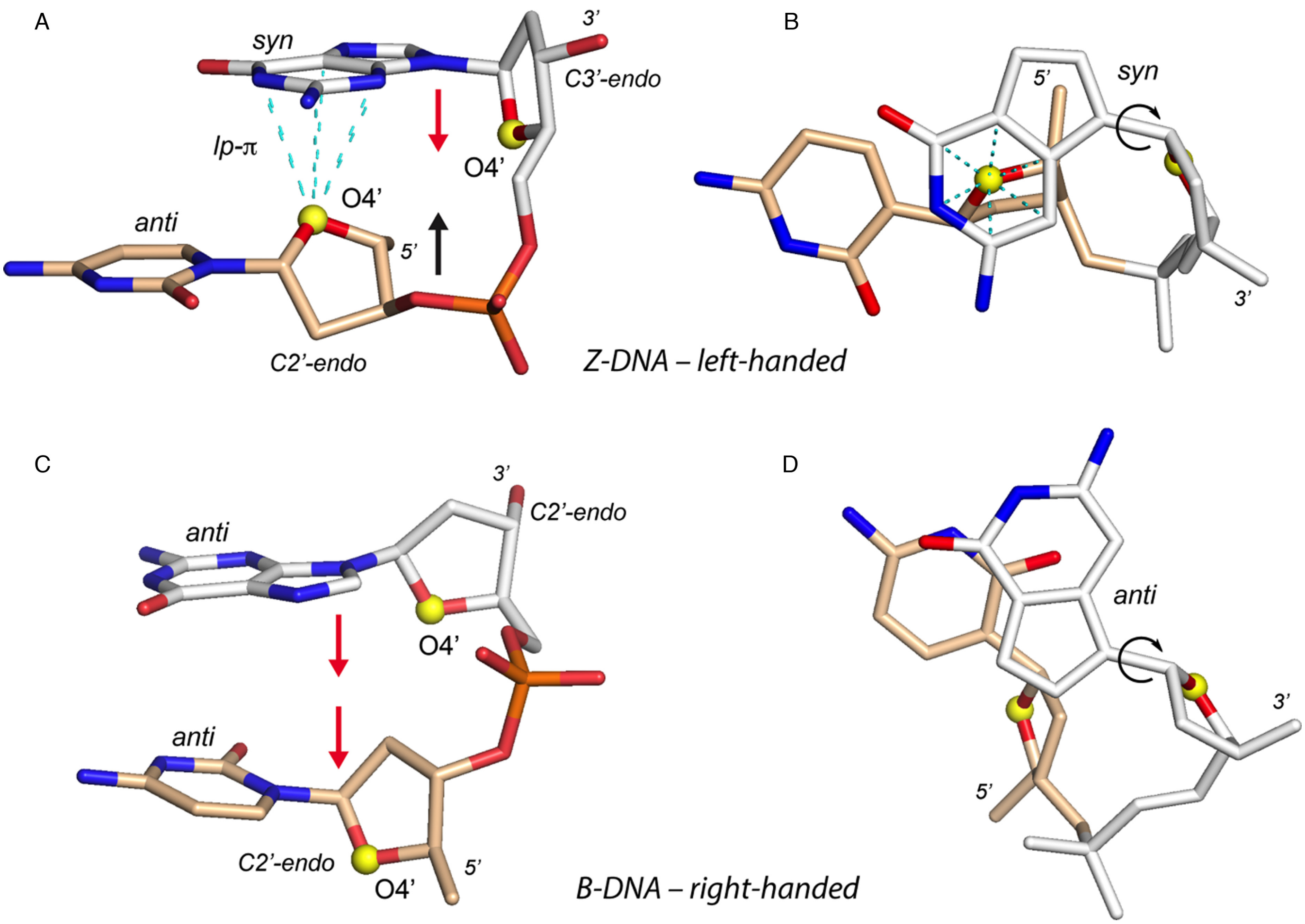


Figure 1. Structural features of Z-DNA and B-DNA CpG steps (similar rules apply for all d/r(NpN) steps). *3′*- and *5′*-nucleotides are coloured in white and wheat, respectively; O4' atoms are shown as yellow spheres. (**A**) d(CpG) step extracted from a Z-DNA crystal structure (PDB: 3P4J; res: 0.55 Å). Note the characteristic antiparallel orientation of the ribose rings marked by red and black arrows. (**B**) Orthogonal view of (**A**) emphasizing the *syn* orientation of the the 3′-residue. In (**A**) and (**B**), the dashed cyan lines correspond to interatomic contact distances (≤3.5 Å) involving the *5′*-O4' atom and the *3′*-six-membered ring atoms. These contacts define a 'capping' or 'lp–π' interaction. (**C**) d(CpG) step extracted from a B-DNA crystal structure (PDB: 1EN3; res: 0.99 Å). Two red arrows mark the parallel orientation of the ribose rings. (**D**) Orthogonal view of (**C**) emphasizing the *3′*-residue *anti* orientation. In (**B**) and (**D**), *3′*-glycosidic '*syn*/anti' bonds marked by circular arrows are similarly oriented and all atoms except the N/O atoms on the Watson–Crick edges and the O4' atoms are in white or wheat colours.

stacking has been mentioned in the first studies of Z-DNA crystallographic structures (2,6,7) but its implications were only investigated several years later (19–21) and never addressed in RNA systems. More generally, such lp–π interactions are currently subject to strong interest in the chemical field where they are considered as a significant and largely unexplored non-covalent interaction type (22–24). Finally, we describe double and single strand Z-like conformations in RNA/protein systems and, among those, in viral RNA that are specifically recognized in a conformation-dependent manner by specialized proteins from the immune system.

MATERIALS AND METHODS

The Protein Data Bank (PDB) was searched for 'Z-DNA like' dinucleotide steps (hereafter named 'Z-like' steps) in DNA and RNA crystallographic structures with resolutions ≤3.0 Å. Z-like steps were characterized using the following criteria: (i) the *3′*- and *5′*-nucleotide adopt a *syn* and *anti* orientation, respectively; (ii) the *5′*-O4' ribose atom is at ≤3.5 Å from the *3′*-nucleobase plane with its projection on the base plane circumscribed in the polygon defined by the ring atoms (Figure 1A and B); these criteria define a lone pair–π interaction that is associated with an atypical antiparallel orientation of the ribose rings with facing *3′*- and *5′*-O4' atoms (6). In order to exclude a few borderline cases, we explicitly imposed antiparallel orientation of the ribose rings in this survey.

Since some Z-like steps have ribose puckers in the north, east or south quadrant, we did not rely for their identification on the classical and more restrictive Z-DNA *C2'-endo* to *C3'-endo* ribose pucker sequence. Thus, we avoid issues related to the difficulty to resolve precisely ribose puckers in experimental structures (25,26). The 3DNA/DSSR analysis tool was used to identify 'lp–π' capping contacts (Figure 1A and B) as well as to calculate backbone torsion angles, hydrogen bond contacts, ribose puckers and to characterize *syn*/*anti* conformations (27).

Z-like steps having atoms with *B-factors* ≥79 Å^2 were excluded from our statistics as well as disordered or terminal steps found at the *3′*- or *5′*-end of the structures unless otherwise specified. Crystallographic structures with resolutions >3.0 Å as well as some cryo-EM and NMR structures were also inspected, although they were not considered for statistics. As of February 2016, the PDB contains ≈5100 nucleic acid crystal structures including complexes with proteins over a total of ≈96 000 biomolecular structures (resolution ≤ 3.0 Å). Images of 3D structures were generated with PyMOL (Schrödinger, L.L.C.; http://www.pymol.org).

Non-redundant Z-like steps were tagged as follows. If two steps from different structures share identical residue numbers, chain codes and tetranucleotide sequences (including a residue before and after the dinucleotide step) as well as ribose puckers, backbone dihedral angle sequences (following a g+, g-, t categorization) and *syn*/*anti* conformations, they are considered as similar and the one with the best resolution is marked as non-redundant. Alike, if in a same structure two Z-like steps share same residue numbers and tetranucleotide sequences (with different chain codes) as well as ribose puckers, backbone dihedral angle sequences and *syn*/*anti* conformations, they are considered as similar and the one corresponding to the first biological unit is marked as non-redundant. The former criteria are used to filter similar PDB structures and the latter to filter structures with multiple related biological units. Note that it is impossible to completely eliminate redundancy from a dataset without eliminating at the same time significant data. Here, we provide an upper limit for a truly 'non-redundant' set.

RESULTS

Z-like steps are found in both DNA and RNA and are not limited to CpG steps

CpG steps are highly represented in Z-DNA left-handed duplexes (Table 1). Besides, a few Z-like TpG/ApA, CpA and GpG/ApC/TpT steps were identified in quadruplex loops, in a rare DNA tetraloop (28) and in single stranded DNA. Also some telomere end-binding proteins recognize Z-like GpG steps (29).

In RNA, over 600 Z-like steps are found in the PDB. Compared with the total number of dinucleotide steps in the database (>600 000), they correspond to rare events, but their presence at key locations in a limited number of RNA families, including ribosomal RNA (see below) makes them noteworthy. Their sequence variety is much greater than in DNA (Table 1). All steps containing a *3′*-purine are represented and ApA steps are more frequent than CpG steps contrasting with the dominance of the latter in DNA. Strikingly, in both nucleic acids, *5′*-pyrimidine steps remain uncommon. Up to now, as only one crystal structure of a Z-RNA helix has been solved (16), most of the identified Z-like steps are located within non-helical regions. However, since a large diversity of sequences are found in Z-DNA helices, the crystallization of further Z-RNA duplexes might reveal a similar sequence diversity.

Ribose-base stacking or 'lp–π' (lone pair–π) interactions define the structure of Z-like steps

A striking characteristic of Z-like steps in DNA relates to the stacking of the O4' atom of the *5′*-deoxyribose ring with the six-membered ring of the *3′*-residue, a contact that was to the best of our knowledge never described in NMR or crystallographic structures of RNA systems (Figure 1A and B). This contact is promoted by the large slide of the two bases and by a specific sequence of backbone dihedral angles that leads to an antiparallel arrangement of the ribose rings with facing O4' atoms (6). This arrangement contrasts with the same strand parallel ribose alignment in B- and A-DNA helical structures (Figure 1C). In addition, this stacking interaction is similar to the stacking of phosphate groups over uridines observed in tRNA anticodon loops (30). Such ribose-base stacking interactions are sometimes called 'lp–π' (lone pair–π) stacking (19–22) and where associated with a rare shift of proton signals in Z-DNA NMR spectra (31).

In order to better characterize these stacking interactions, we calculated the O4' to base plane contact distance (Figure 2). A strong decay towards 3.5 Å in the associated histogram suggests that the O4' atoms are clustered close to the aromatic nucleobase, and that contacts above this limit should no longer be considered as stacking interactions. The average contact distance for Z-DNA and RNA is 2.9 ± 0.1 and 3.0 ± 0.2 Å, respectively. Interestingly, in these histograms the Z-DNA peak is much sharper than the RNA peak. This might be linked to two factors: (i) Z-DNA structures are generally of a much better resolution (average 1.8 Å for DNA versus 2.6 Å for RNA), and (ii) the structural context of Z-like steps is more diverse in RNA than in DNA, leading to more dispersed positions of the O4' atoms over the purine rings (Figure 2B and C). Furthermore, the r(NpG) step histogram displays two peaks. The first is associated with UN$\underline{\text{CG}}$ loops (≈2.9 Å) for which the CpG step is very similar to that in Z-DNA (see below). The second is associated with more diverse RNA turns (≈3.1 Å). For r(NpA) steps, the distance distribution is broader as it is associated with a much larger structural diversity in small and large RNAs. The r(NpA) step distance distribution is like that of the second r(NpG) peak, further stressing that both are associated with a broad structural context extending that of Z-helices and UNCG tetraloops.

Next to 'lp–π' interactions, ribose puckers are very specific in Z-like steps with much stronger constraints on the *5′*- than on the *3′*-nucleotide (Table 2). The former puckers are mainly in the south quadrant (≈91%) while north dominates for the latter (≈62%) followed by south and east. Puckers in the west quadrant remain exceptional. Overall, the N-S (or *C3′-endo-C2′-endo*) pucker configuration prevails in both DNA and RNA.

Z-like steps in r(UN$\underline{\text{CG}}$) loops are similar to those found in left-handed helices

As stated above, Z-like CpG steps are constitutive of r(UN$\underline{\text{CG}}$) tetraloops and structurally similar to those found in Z-DNA and Z-RNA (Figure 3) (32). Despite their high thermodynamic stability (33,34), these tetraloops are rare

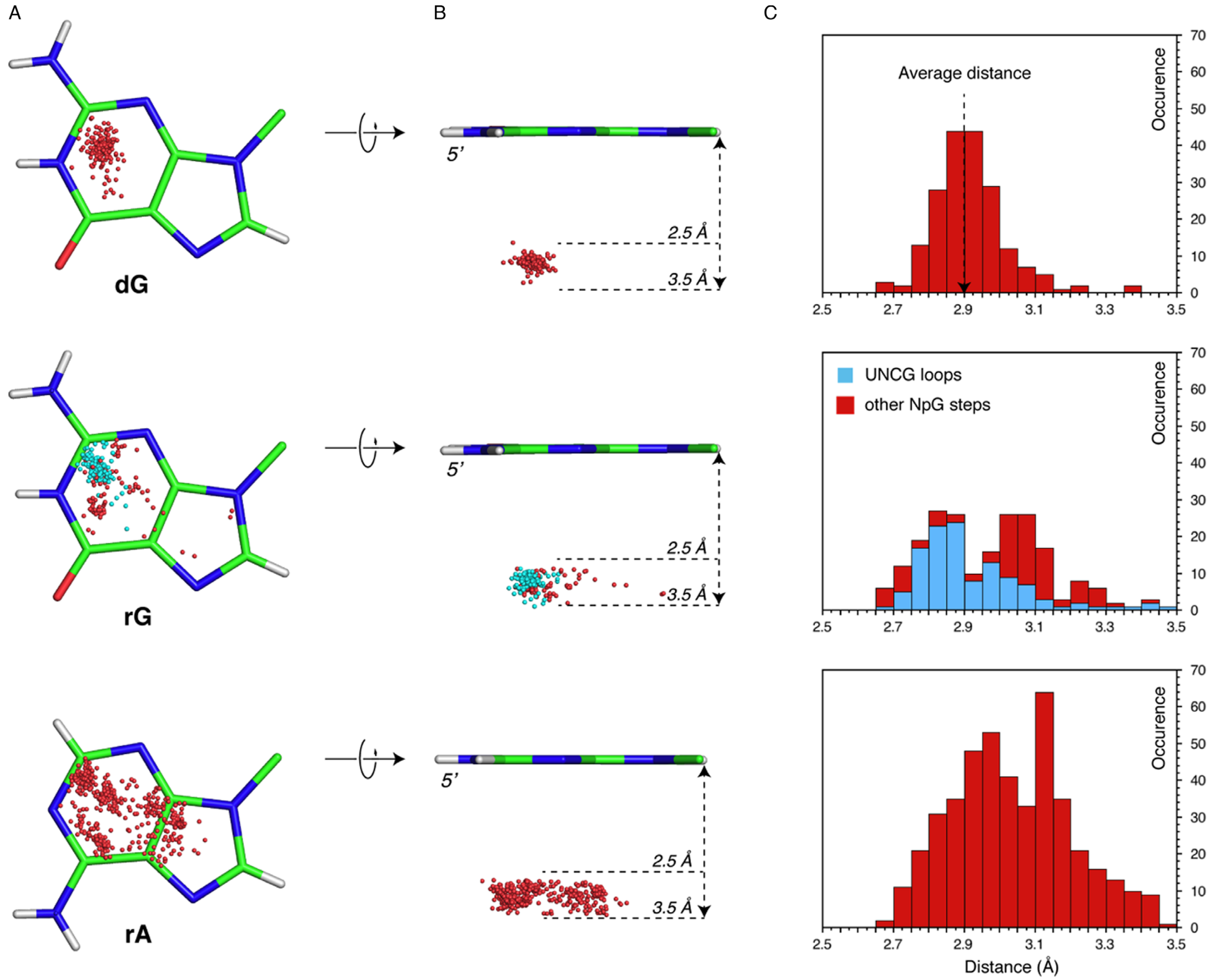


Figure 2. Ribose O4' stacking ('lp–π') to dG, rG and rA nucleobases. (**A**) Top view (5′-side) of the dG, rG and rA nucleobases showing the positions of all O4' atoms above the nucleobase plane with a contact distance ≤ 3.5 Å. For rG, the O4' atoms belonging to r(UNCG) loops are in cyan instead of red. (**B**) 90° rotation of (**A**); the 2.5-3.5 Å boundaries are marked by arrows and dashed lines. (**C**) Histogram of the O4' to nucleobase plane distances drawn from structural sets including all contacts. For rG, O4' positions and related distances belonging to r(UNCG) loops are in cyan instead of red.

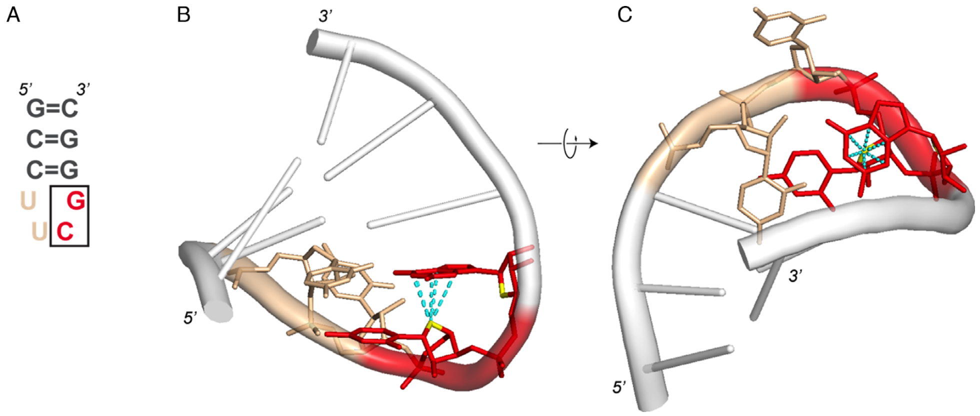


Figure 3. Z-like step in a r(UUCG) tetraloop. In all panels and subsequent figures, r(CpG) steps are shown in red; the ribose O4' atoms are shown in yellow; the cyan dashed lines are defined in Figure 1. In this figure, r(UpU) steps are shown in wheat. (**A**) 2D structure of a r(UUCG) tetraloop. The Z-like r(CpG) step is boxed. (**B**) 3D structure of a r(UUCG) tetraloop (PDB: 1F7Y; res: 2.8 Å). (**C**) 90° rotation of (**B**).

Table 1. Number of non-redundant Z-like steps found in PDB crystallographic RNA and DNA structures (resolution ≤ 3.0 Å; atomic *B-factors* ≤ 79 Å^2; February 2016 PDB survey) over a total >600 000 nucleotide steps

Dinucleotide steps	DNA[a]	RNA[a]
3′-guanine		
CpG	124 (187)	42 (125)
(T/U)pG	9 (9)	13 (21)
ApG	–	14 (57)
GpG	8 (11)	1 (6)
3′-adenine		
CpA	3 (4)	18 (25)
(T/U)pA	13 (15)	42 (127)
ApA	1 (1)	75 (279)
GpA	–	13 (13)
3′-cytosine		
CpC[b]	–	1 (1)
(T/U)pC	–	5 (5)
ApC	–	1 (1)
GpC	–	–
3′-thymine/uridine		
Cp(T/U)	–	2 (2)
(T/U)p(T/U)	1 (1)	1 (1)
Ap(T/U)	–	3 (4)
Gp(T/U)	–	–
Total:	159 (228)	231 (667)

[a] The total number of Z-like steps found in the PDB and with no redundancy considerations is given in parenthesis.

[b] The *3′*-cytosine of this CpC step displays a 50% *syn* and *anti* occupancy (70).

Steps with disordered backbones, usually found in high-resolution Z-DNA structures, were not taken into account. However, modified nucleotides in Z-DNA were considered. Note that these statistics reflect only the step distribution in structures deposited to the PDB and not the in vivo distribution of these steps. The high number of non-redundant CpG steps in Z-DNA is partly related to the incorporation of modified nucleotides in our structural sample and should be considered with caution.

Table 2. Ribose puckers for the *3′*- and *5′*-nucleotides in non-redundant Z-like steps found in PDB crystallographic RNA and DNA structures (resolution ≤ 3.0 Å; *B-factors* ≤ 79 Å^2)

Pucker	DNA[a]	RNA[a]
3′-nucleotide		
North *[C3'-endo]*	87 (132) *[82 (127)]*	132 (419) *[131 (418)]*
South *[C2'-endo]*	27 (34) *[21 (27)]*	34 (87) *[34 (87)]*
West	1 (1)	–
East	44 (61)	63 (160)
5′-nucleotide		
North *[C3'-endo]*	– *[–]*	12 (13) *[8 (9)]*
South *[C2'-endo]*	137 (202) *[130 (193)]*	198 (611) *[190 (599)]*
West	2 (2)	1 (1)
East	20 (24)	18 (41)

[a] The total number of ribose puckers in each category with no redundancy considerations is given in parenthesis.

Steps with disordered backbones, usually found in high-resolution Z-DNA structures, were not taken into account. Modified nucleotides were considered.

in natural RNA systems where almost all known occurrences are gathered in ribosomal structures (see below). r(UNCG) loops were also artificially grafted to RNA structures to serve as stem capping motifs for stabilization and crystallographic purposes, e.g. the r(UUCG) tetraloops in an RNA-protein complex (32), the group II intron (35) and the RNaseP structure (36). All these structures exhibit tetraloop CpG steps whose structure is consistent with those reported in high-resolution NMR structures (37). The lp–π stacking interaction over more than hundred r(UNCG) instances is associated with an average distance of ≈2.9 ± 0.1 Å that is, as mentioned above, very similar to that calculated for Z-DNA CpG steps (Figure 2A and B). Despite the fact that UNCG loops can have any nucleotide at the second position, we could only identify cUUCGg and cUACGg loops in our PDB set of structures.

Z-like steps appear at key locations in small RNAs…

Z-like steps were also identified in a variety of small RNA structures, including: (i) purine riboswitches (38), (ii) classI/II preQ1 riboswitches (39,40), (iii) thiamine pyrophosphate (TPP) riboswitches (41), (iv) lysine riboswitches (42), (v) streptomycin aptamers (43) and (vi) hepatitis delta virus (HDV) ribozymes (44–48) but are absent in other structures like tRNA or group I introns (Figure 4). The diversity of dinucleotide sequences involved in these Z-like steps is unexpected. Contrary to what could be inferred from the dominance of CpG steps in Z-DNA, Z-like steps in RNA form in a large variety of contexts and in the absence of high salt conditions. For instance, >30 purine riboswitches associated with ≈18 different ligands were crystallized, all of them displaying a conserved UpA, ApA or CpA Z-like step (38). In these riboswitches, the *3′*-

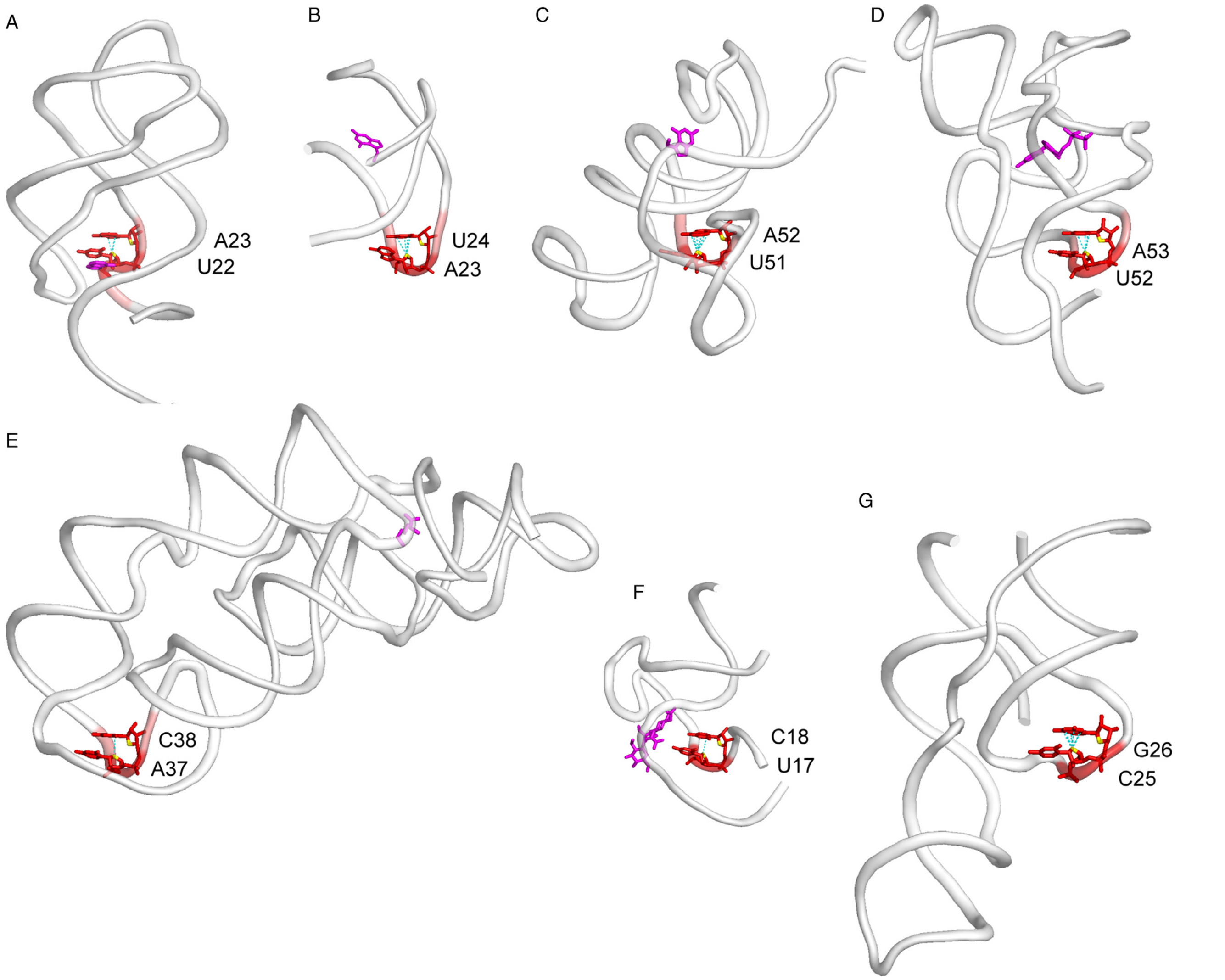


Figure 4. Z-like steps observed in crystallographic structures of small RNA systems. When present, ligands are shown in magenta. (**A**) Purine riboswitch (PDB: 4FE5; res: 1.3 Å); (**B**) class I preQ$_1$ riboswitch (PDB: 3K1V; res: 2.2 Å); (**C**) class II preQ$_1$ riboswitch (PDB: 4JF2; res: 2.3 Å); (**D**) TPP riboswitch (PDB: 2GDI; res: 2.1 Å); (**E**) lysine riboswitch (PDB: 3D0U; res: 2.8 Å); (**F**) streptomycin aptamer (PDB: 1NTA; res: 2.9 Å); and (**G**) HDV ribozyme (PDB: 3NKB; res: 1.9 Å).

nucleotide stacks with the ligand (Figure 4A), while the conserved *5′*-adenine participates in an important base triple. If base triple disrupting mutations of this adenine are detrimental to the structure and activity of the riboswitch, it has been shown that mutations of the *3′*-U to C or A are tolerated and preserve the Z-like step structure that is consequently partly sequence independent (49). As such, Z-like steps must shape in a very specific manner the ligand-binding pocket of purine riboswitches.

In other RNA systems, Z-like steps occur in turns similar to those found in UNCG tetraloops (Figure 3B), where the bottom *5′*-nucleobase often has a solvent-exposed face. There, Z-like steps are involved in junctions or joining regions where at least one of the nucleotides is pairing with distant residues (e.g. the $A_{53}\bullet A_{ZERAZRZ84}$ base pair in the TPP riboswitch). Together, these observations suggest that Z-like steps occur at key locations where they promote specific turns that are strategic for creating precise and not otherwise possible RNA folds (see Discussion).

... are involved in long-range contacts ...

An important long-range interaction involving a Z-like step but no specific turn occurs in a viral tRNA-like structure (TLS; Figure 5A) (50,51). In this RNA, a previously unidentified Z-like ApG step is embedded within a terminal *5′*-UUAG sequence, which was historically not recognized as part of the minimal TLS (the two uridines are not visible in the crystallographic structure). However, the presence of this nucleotide sequence has been proven important to stabilize the global tRNA-like TLS fold via a long-range interaction and is required for aminoacylation. The involved base pair is a *cis*-WC $G_2{=}C_{74}$ pair with a G in *syn*. The existence of this single base pair to hold the global fold allows TLS to have a structural and functional flexibility exploited for viral activity, a functional plasticity present in almost all

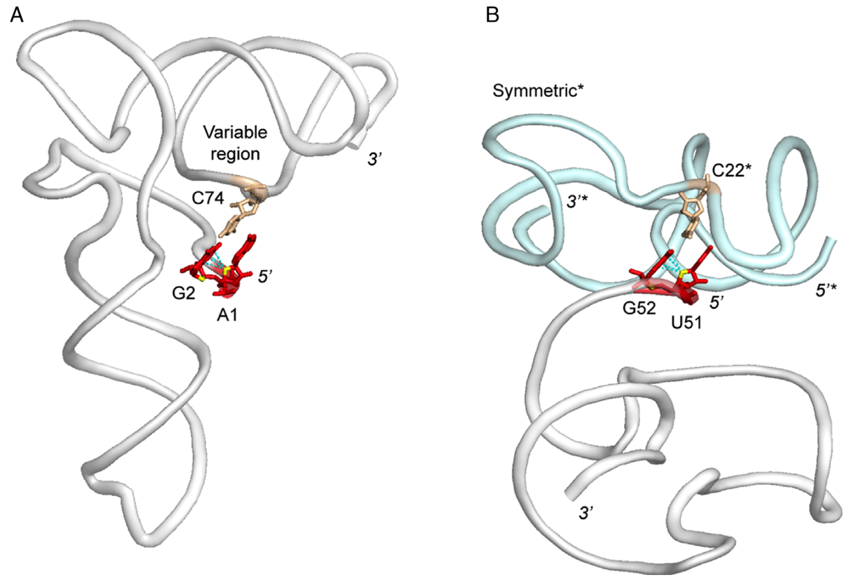


Figure 5. Z-like steps establishing long-range contacts in a tRNA like system (TLS) and a fluoride riboswitch (symmetry contact). The nucleotides forming a Watson–Crick pair with the 5′-nucleotide of the Z-like step are shown in wheat. (**A**) TLS structure (PDB: 4P5J; res: 2.0 Å). (**B**) Fluoride riboswitch (PDB: 4ENC; res: 2.3). The symmetry related molecule is shown in light blue; asterisks mark annotations for this molecule.

tRNA and tRNA related systems (52). It has been proposed that a loss of this interaction is what enables TLS to more readily unfold to allow viral replication. Indeed, a loss of the TLS structure is observed when the 5′-fragment containing the Z-like step is truncated. A similar Z-like step associated with a *cis*-WC G=C pair is present in the crystal structure of a fluoride riboswitch (53). There, it involves a symmetry-related molecule, suggesting that such long-range interactions, although uncommon, are modular elements that are of importance in the fold of specialized RNA molecules (Figure 5B).

... and are also present in ribosomal structures

In addition to being found in small RNA systems, Z-like steps are also present in all available ribosome structures, an indication that they might occur in ribosomal RNA of all organisms. Several of them are clustered in the conserved core of the ribosomal large subunit (LSU; Figure 6), while others are found in non-conserved peripheral regions of the large and small subunits (SSU). Given the complexity of these ribosomal structures, we report only a few conserved occurrences of Z-like steps in the LSU of *E. coli* (54), *S. cerevisiae* (55) and *H. sapiens* (56), as deduced from their 3D structure and from sequence conservation data based on phylogeny (57). Other Z-like steps found in non-conserved regions including those in the SSU are poorly resolved in available crystallographic and cryo-EM structures and therefore will not be discussed here.

Overall, these ribosomal Z-like steps are similar to those found in small RNAs where they allow for distant pairing between nucleobases. Some of them, like the ones within conserved UUCG tetraloops, are additionally contacting proteins that interact specifically with their Z-like CpG step, pointing out that Z-like steps can be part of RNA-protein recognition schemes.

Specific recognition of Z-like steps by regulatory and RNA modification proteins

Single stranded segments integrate Z-like steps that are directly recognized by specialized RNA binding proteins such as the iron regulatory protein 1 (IRP1) (58,59), CUG-binding proteins (18) and proteins associated with H/ACA box snoRNA (60–62). Hereafter, we will briefly address the variety of recognition patterns in which they are involved (Figure 7).

Iron-responsive elements (IREs) are short mRNA stem-loops recognized by the iron regulatory protein 1 (IRP1) at two sites separated by ≈30 Å. One of these sites involves a conserved ApGpU triloop where ApG forms a Z-like step (Figure 7A) (58). These bulged-out A and G nucleotides point toward the protein and are associated with a sharp turn in the RNA backbone. This step contacts five different amino-acids and is specifically sandwiched by two leucine side chains that provide van der Waals contacts to the exposed aromatic surfaces of the A and G residues. Although a subsequent crystal structure with a different IRE element displays the same 2D motif (59), the available NMR structures of this element are not showing a Z-like step. Therefore, this conformation might be protein-induced and/or protein-stabilized (63).

CUG-binding proteins regulate multiple aspects of nuclear and cytosplamic mRNA processing. They are known to preferentially target UGU-rich mRNA elements to accomplish their mRNA processing functions. The UpG step adopts a left-handed Z-RNA conformation (Figure 7B) where the *syn* guanine is recognized through specific Hoogsteen edge-protein backbone interactions (18). The similarities between UpG steps as found in this complexes and CpG steps in structures of a complex of Z-RNA with the ADAR1 Zα protein were described. Interestingly, the U_3-G_4 and U_8-G_9 steps in the GUUGUUUUGUUU sequence in complex

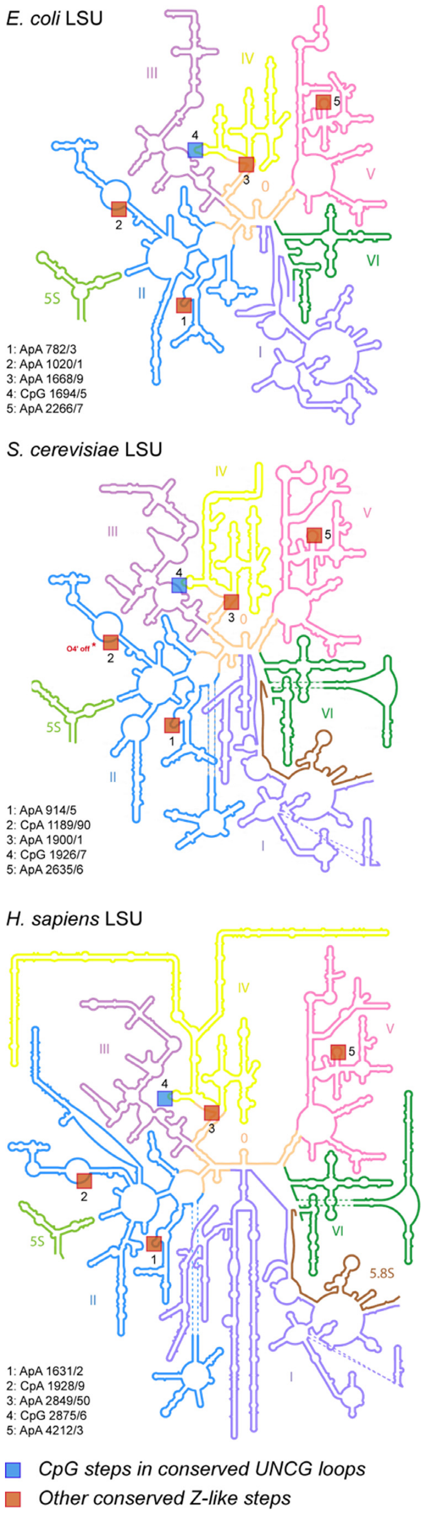


Figure 6. Conserved Z-like steps in 2D structures of three large (LSU) ribosomal subunits from *E. coli*, *S. cerevisiae* and *H. sapiens*. The 2D representations (derived from 3D structures) were adapted from images stored at http://apollo.chemistry.gatech.edu/RibosomeGallery (57). The Z-like steps were inferred from the X-ray *E. coli* (4YBB; res: 2.1 ; chain: DA) (54), *S. cerevisiae* (4U4R, res: 2.8 ; chain: 1) (55) and cryo-EM *H. sapiens* (4UG0, res: 3.6 ; chain: L5) (56) structures. Note that in yeast, for step '2', the O4' atom is shifted by 0.1 Å and therefore not exactly stacked over the adenine ring, illustrating the difficulties to work with large structures of medium to low resolution that often embed local inaccuracies.

with two RNA recognition motifs (bound RRM1 and tandem RRM1/2) share the same Z-like step structural features.

H/ACA ribonucleoprotein particles are a family of pseudouridine synthases that use guide RNAs to specific modification sites (60–62,64). They also participate in eukaryotic ribosomal RNA processing and are a component of vertebrate telomerases. H/ACA RNAs fold into repeats of a consensus hairpin structure that comprise an internal loop and an ACA signature that harbours a *3′*-tail. This three single stranded $A_{58}pC_{59}pA_{60}$ signature forms two consecutive CpA Z-like steps when including the terminal C_{57} stem nucleotide (Figure 7C). The PUA (PseudoUridine synthase and Archaeosine transglycosylase) domain of Cbf5 recognizes very specifically this Z-like step repeat. Although the two Z-like steps are easily characterized by visual inspection, it appears that they represent also borderline steps regarding their lp–π geometry; the C_{57} ribose is almost parallel to the A_{58} nucleobase and the lp–π contact distance of the $C_{59}pA_{60}$ step exceeds 4.0 Å. Although borderline, these consecutive Z-like steps seem stable since they are recurrently observed in H/ACA box systems.

A further occurrence of Z-like steps in RNA/protein complexes involves a bacteria-to-phage 'immune response', i.e. the bacterial phage-resistance system ToxIN involving the protein ToxN that is inhibited *in vivo* by a specific ToxI antitoxin RNA (65). A crystal structure of the complex shows a Z-like 'ApG' step shortly upwards the ToxI *3′*-end that binds to the ToxN groove 1 (Figure 7D). The backbone turn associated to the Z-like step allows the two following adenines to directly point towards the protein recognition pocket and establish several specific interactions. There, amino acid side chains interact with the Watson–Crick edges of the Z-like nucleotides like those in the RNA junctions and turns described above.

Z-like steps in immunology-related viral RNAs

The last and most intriguing aspect of Z-like steps, already mentioned for the bacteriophage system, is related to their role in the immune response. Adenosine deaminase (ADAR1) proteins embed a Zα domain able to recognize Z-DNA as well as Z-RNA duplexes (12,16,66,67). The two available crystal structures of a Zα domain in complex with Z-DNA and Z-RNA CpG hexamers display similar characteristics. In these complexes, the central CpG step is essentially recognized through specific amino acid-phosphate group contacts. Additionally, in the Z-DNA complex, a single weak van der Waals contact of the CH…π type between Tyr177 and a guanine is observed. When the RNA replaces the DNA substrate, this contact disappears, suggesting that the Zα domain could recognize any Z-RNA motif (the authors did however not exclude that crystal-packing effects may have slightly altered the structure of this RNA complex). Indeed, three other d(CA$\underline{\text{CG}}$TG)$_2$, d(CG$\underline{\text{TA}}$CG)$_2$ and d(CG$\underline{\text{GC}}$CG)$_2$, Z-DNA substrates were co-crystalized with the Zα domain, stressing that a Z-like TpA step can be recognized similarly to a CpG step (11). Hence, these protein domains and their analogs (68) could recognize Z-DNA as well as Z-RNA in a non-sequence spe-

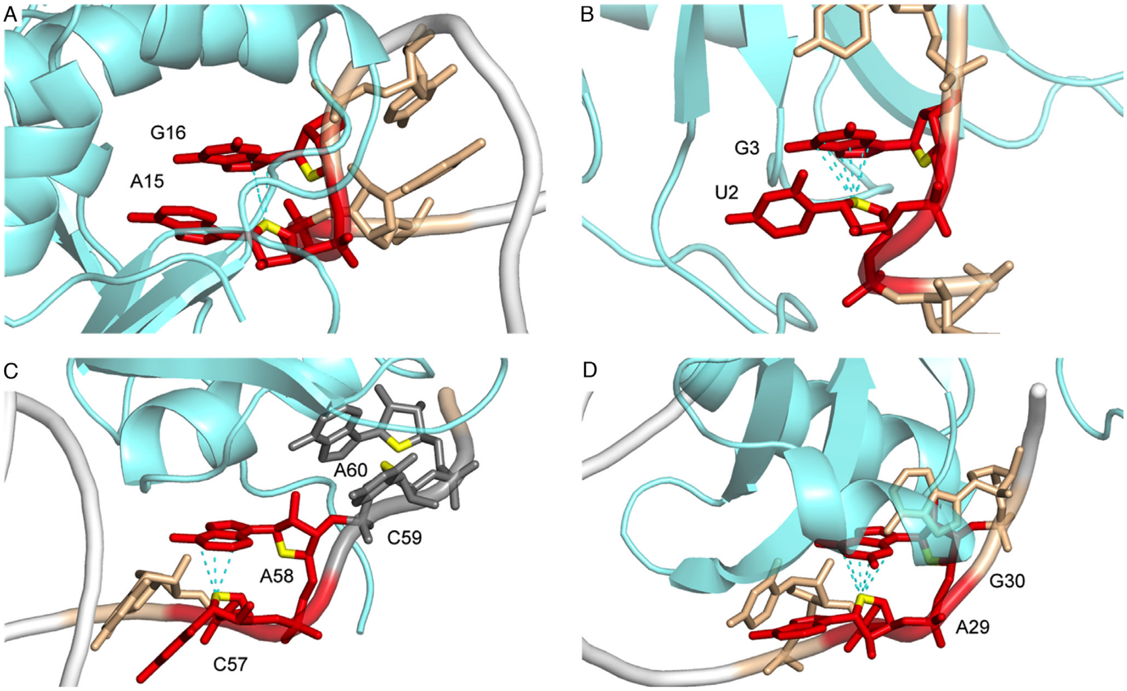


Figure 7. Examples of Z-like steps recognized by proteins. (**A**) IREs mRNA in complex with IRP1 (PDB: 3SNP; res: 2.8 Å). (**B**) CUG-binding protein in complex with UGU-rich mRNA (PDB: 3NMR; res: 1.9 Å). (**C**) H/ACA ribonucleoprotein particle (PDB: 3HAX; res: 2.1 Å). This structure displays two consecutive and borderline Z-like steps. The $C_{59}pA_{60}$ step that displays a lp–π 'contact' > 4.0 Å is shown in grey with O4' atoms in yellow. (**D**) ToxI RNA-ToxN protein complex (PDB: 2XDB; res: 2.6).

cific manner, a process that may be associated with the immune response (69).

Z-like steps were further identified in single stranded RNA *5′*-triphosphate groups (*5′*-PPP-RNA), a signature of viral RNA, when recognized by interferon-induced proteins with tetratricopeptide repeats of the IFIT5 family. Three crystal structures of the human IFIT5 protein in complex with *5'*-PPP-$N_1N_2N_3N_4$ (N = C, U or A) ligands (Figure 8) (70–72) reveal that the ligand is recognized in a non-sequence- but conformation- and modification-specific manner. Indeed, no nucleobase-to-protein contacts are observed in these structures but only contacts to the ribose-phosphate backbone and the binding pocket of this protein does not seem large enough to accommodate the usual capping modifications found in eukaryotic RNAs.

In each of these ligands, the N_1pN_2 step adopts a Z-like conformation inducing the formation of important contacts between the protein and the RNA backbone. The N_1 and N_2 bases do not establish specific hydrogen bonds with protein residues, and there is ample space adjacent to the pyrimidine edges, suggesting that the binding pocket can easily accommodate the larger purine nucleobases as seen in the oligo-A complex. Interestingly, the position of the RNA backbone in the oligo-A complex favors the formation of a lp–π interaction involving the five- instead of the more usual six membered ring (Figure 8B). This particular Z-like step arrangement allows the accommodation of all-purine as well as rare all-pyrimidine sequences and eventually combinations of them without the need to adjust backbone conformation. As such, it allows the incorporation of an all–C sequence for which it was difficult to precisely identify the *syn/anti* nucleobase conformation (Figure 8C and D). Given available structures, it seems very likely that the C_2 base is in *syn*. Yet, these data also imply that it is much more difficult to identify *syn* pyrimidines than *syn* purines in crystallographic structures. Consequently, their number might be slightly underestimated in the PDB (26). Indeed, evidence was given very early that pyrimidines could adopt syn conformations in solution and be associated with Z-steps (4,10).

DISCUSSION

In both DNA and RNA, we observe that not only CpG steps but also almost any dinucleotide sequence can adopt similar Z-like structures, with a preference for those with purines on the *3′*-side. A difference between Z-like steps in RNA and DNA is that in RNA these steps are usually found in single-stranded regions such as loops and junctions, where they contribute to creating specific backbone kinks and turns. However, we were unable to identify recurring interaction patterns, which point to a great diversity of Z-like step usages. In Z-dinucleotides, we noted that the Watson-Crick sites of the two nucleobases point in the same direction. Hence, in most instances the two nucleobases are alternatively or simultaneously forming hydrogen bonds with distant residues. Significantly, a Z-like step has remained unnoticed in the UNCG tetraloop family, although the similarity between Z-DNA and UNCG backbones was mentioned elsewhere (73). This observation highlights the difficulties of circumventing the complex interaction patterns present in even simple and well-studied RNA motifs.

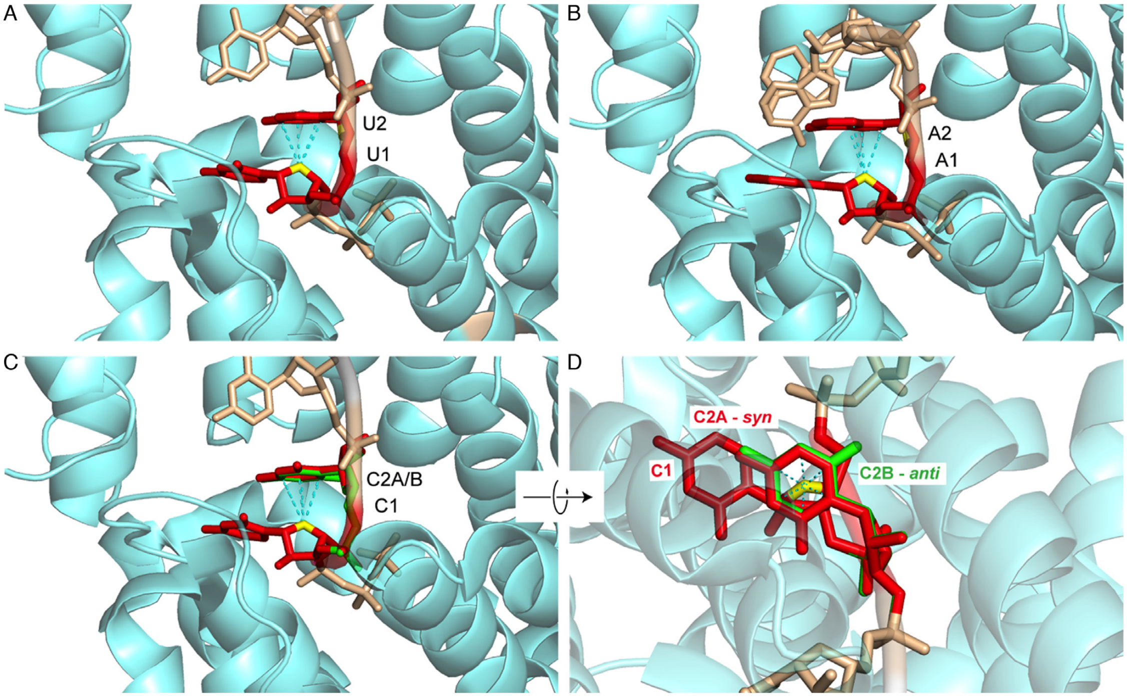


Figure 8. Z-like conformations adopted by the N_1pN_2 step of a 5′-PPP-$N_1N_2N_3N_4$ viral RNA primer in complex with the human interferon induced IFIT5 protein. The 5′-PPP group and the N_3pN_4 residues are shown in wheat and the protein backbone in aquamarine. (**A**) 5′-PPP-UUUU (PDB: 4HOS; res: 2.0 Å). (**B**) 5′-PPP-AAAA (PDB: 4HOT; res: 2.5 Å). Note the rare lp–π stacking involving the five membered ring of A_2. (**C**) 5′-PPP-CCCC (PDB: 4HOR; res: 1.9 Å). (**D**) 90° rotation of (**C**). In (**C**) and (**D**), the nucleobase C_2 adopts a *syn* (A; red) and an *anti* (B; green) alternate conformation illustrating the difficulty of unambiguously assigning *syn* conformations for pyrimidines.

Dissecting the unusual structure of a Z-like step reveals that it is the result of several uncommon events. Both nucleotides are required to adopt specific conformations (*C2'-endo* on the 5′-side, *syn* on the 3′-side) that each occurs at a 2% and 12% frequency, respectively (as estimated from PDB crystal structures). In that respect, the frequency of Z-like steps that combines both conformational features drops to a low ≈0.1%. Moreover, the *syn* nucleobase is fully flipped over the ribose of the *anti* nucleobase, which among most but not all Z-like steps results in a specific stacking interaction involving the O4' of the first ribose where the distance between that O4' atom and the nucleobase ring is shorter than the average stacking interaction distances between aromatic rings (2.9 Å versus 3.4 Å), due to electrostatic and dispersion effects probably dominated by solvent induced-forces (74). This interaction of the lp–π type has been shown through quantum mechanical calculations to be rather weak and is probably an incidental event rather than a potential folding driver (19,21). If this interaction had been promoting folding, it would have been observed much more frequently. In short, although a Z-like step is made of rare and energetically costly conformations, its structure is induced and stabilized by its surrounding during folding.

We wish to posit that the low frequency and the incidental nature of Z-like steps are compatible with a precise structural function. Since *syn* and *C2'-endo* conformations are linked with slow dynamics, they probably need assistance from other elements to overcome the structural stress imposed to the backbone (26,75). We therefore propose that the combination of rare conformations within a Z-like step would create regions that retain their fold once formed. Like other structural elements involved in long-range 3D contacts such as GNRA loops interacting with their receptors (76), Z-like steps could act as conformational locks at strategic locations in RNAs, with the particularity of involving rare nucleotide conformations that need specific structural contexts for their formation. Here, the case of the purine riboswitch is particularly interesting, as the Z-like UpA is part of the ligand-binding site (U22 is directly contacted by the ligand). The Z-like motif is part of the junction J1/2. It stabilizes the final bound structure through long-range interactions with J2/3—J3/1 and is next to the the entry site of the ligand (77,78). Additionally, the preorganized state of the binding pocket that is known to involve J1/2 and precede ligand binding (79) could be in part attributed to the presence of this ‘Z-lock’. We hope that these considerations will encourage studies of the dynamics of formation of Z-like steps, as those may reveal cues to understanding the ligand binding process.

Among all the intramolecular RNA motifs that were described here, the one found in the TLS structure seems the most peculiar since the active fold of the structure requires the formation of a single long-range Watson-Crick base pair involving a Z-like step (50). The study of other folds such as those associated with the purine riboswitch ligand binding pocket or the ion sensitivity reported for the HDV ribozyme (44–48,80–82) and the lysine riboswitch (42) systems may also offer insights about the structural role of these Z-like steps. Exploring Z-like steps may thus suggest how other motifs may turn out to exert their main function through locking 3D folds. It might also be worthwhile to understand how these rare motifs form in the perspective of using them to create specific folds in synthetic biological systems (83).

Regarding interactions with proteins, Z-like steps are found in Z-RNA and Z-DNA duplexes and the viral 5′-PPP-RNA recognition system (70) that are recognized specifically through conformation-dependent interactions. These interactions involve solely backbone atoms, allowing various sequences to be accommodated (11). For example, a family of proteins recognize double stranded Z-DNA and Z-RNA via a common winged helix-turn-helix domain called Zα. In other instances, like the IRE-RNA system (58,59), the CUG binding complexes (18), the H/ACA (62) or the bacteria-to-phage 'immune response' system (65), different recognition patterns implicating the nucleobases and the backbone atoms are at play, which suggest sequence-dependent recognition patterns. For these systems however, no unique RNA/protein recognition pattern could be found and it appears certain that proteins are able to induce Z-conformations in single stranded RNA by using various mechanisms.

As noted by Alexander Rich et al., both Z-RNA and Z-DNA are highly antigenic and are stimulated by a particular structural context. The first Z-DNA binding protein was found to be a Z-RNA binding protein called double stranded RNA adenosine deaminase (or ADAR1) (4) and numerous Z-DNA specific antibodies are found in human autoimmune diseases such as systemic lupus erythematosus (3,84). These proteins recognize short double stranded Z-RNA steps by essentially binding in a non-sequence specific manner to the sugar-phosphate backbone. Our exploration of Z-RNA motifs therefore also uncovers strong ties to the immune response. Deeper investigations of the structural characteristics, occurrence, and associated RNA/protein recognition features of Z-like steps could represent a pathway to further our understanding of the immune response.

In particular, elevation of CpG but also UpA frequencies in influenza A viruses used as an RNA genome model system have been recently involved in the attenuation of the viral pathogenicity and in the simultaneous increase in host response to infection (85). We hypothesize here that some of the involved mechanisms could be related to a certain tendency of these pyrimidine-purine sequences to promote Z-conformations.

CONCLUDING REMARKS

Classical sequence and structure analysis of this new RNA Z-motif is limited due to the low number of RNA crystal structures in the PDB. In that respect, we expect that the current upsurge of medium and high resolution RNA and RNP structures will increase the number and significance of Z-dinucleotide motifs in structural databases, so that a more advanced characterization of their structural and recognition properties will be within reach. With respect to the comments of one referee that wondered about the significance of these rare motifs, we stand by Alexander Rich, who worked hard to convince fellow researchers that Z-DNA was of biological significance and we share his view that Z-DNA and by extension Z-motifs had to be present *in cellulo* since evolution is opportunistic (4). After all, as demonstrated by this survey, Z-conformations are readily accessible to most dinucleotide sequences, when placed in the appropriate environment. Interestingly, some of these Z-steps are recognized by pattern recognition receptors (PRRs) to distinguish between self and non-self. Thus, the ability of some single or double stranded RNA sequences to be twisted into Z-form steps might be a structural key in some immune system diseases (72). Hence, although Z-steps may currently appear as a 'black swan' (86) in the RNA world, we trust that this new motif will find its place in the still incomplete RNA motif library.

ACKNOWLEDGEMENTS

P.A. and Q.V. wish to thank Prof. Eric Westhof for ongoing support and helpful discussions, as well as Prof. Rob Batey and Prof. Richard Giegé for useful comments on the manuscript. This article is dedicated to the memory of Alexander Rich.

FUNDING

French 'Ministère de la recherche et de l'enseignement' (to L.D.); Polish Ministry of Higher Education and Science (Mobility Plus programme) [1103/MOB/2013/0 to F.L.]. Funding for open access charge: CNRS.
Conflict of interest statement. None declared.

REFERENCES

1. Ho,P.S. and Mooers,B.H.M. (1997) Z-DNA crystallography. *Biopolymers*, **44**, 65–90.
2. Wang,A.H.J., Quigley,G.J., Kolpak,F.J., Crawford,J.L., Vanboom,J.H., Vandermarel,G. and Rich,A. (1979) Molecular-structure of a left-handed double helical DNA fragment at atomic resolution. *Nature*, **282**, 680–686.
3. Rich,A. and Zhang,S. (2003) Timeline: Z-DNA: the long road to biological function. *Nat. Rev. Genet.*, **4**, 566–572.
4. Rich,A. (2004) The excitement of discovery. *Annu. Rev. Biochem.*, **73**, 1–37.
5. Drew,H., Takano,T., Tanaka,S., Itakura,K. and Dickerson,R.E. (1980) High-salt d(CpGpCpG), a left-handed Z' DNA double helix. *Nature*, **286**, 567–573.
6. Sundaralingam,M. and Westhof,E. (1981) Structural motifs of the nucleotidyl unit and the handedness of polynucleotide helices. *Int. J. Quantum Chem.*, 287–306.
7. Wang,A.H.J., Quigley,G.J., Kolpak,F.J., Vandermarel,G., Vanboom,J.H. and Rich,A. (1981) Left-handed double helical DNA - Variations in the backbone conformation. *Science*, **211**, 171–176.
8. Gessner,R.V., Frederick,C.A., Quigley,G.J., Rich,A. and Wang,A.H.J. (1989) The molecular structure of the left-handed Z-DNA double helix at 1.0 Å atomic resolution. Geometry, conformation, and ionic interactions of d(CGCGCG). *J. Biol. Chem.*, **264**, 7912–7935.
9. Fuertes,M.A., Cepeda,V., Alonso,C. and Perez,J.M. (2006) Molecular mechanisms for the B-Z transition in the example of poly[d(G-C)•d(G-C)] polymers. A critical review. *Chem. Rev.*, **106**, 2045–2064.
10. Feigon,J., Wang,A.H.J., Vandermarel,G.A., Vanboom,J.H. and Rich,A. (1985) Z-DNA forms without an alternating purine-pyrimidine sequence in solution. *Science*, **230**, 82–84.
11. Ha,S.C., Choi,J., Hwang,H.Y., Rich,A., Kim,Y.G. and Kim,K.K. (2009) The structures of non-CG-repeat Z-DNAs co-crystallized with the Z-DNA-binding domain, hZ(ADAR1). *Nucleic Acids Res.*, **37**, 629–637.
12. Savva,Y.A., Rieder,L.E. and Reenan,R.A. (2012) The ADAR protein family. *Genome Biol.*, **13**, 252.

13. Kim,D., Hur,J., Park,K., Bae,S., Shin,D., Ha,S.C., Hwang,H.Y., Hohng,S., Lee,J.H., Lee,S. *et al.* (2014) Distinct Z-DNA binding mode of a PKR-like protein kinase containing a Z-DNA binding domain (PKZ). *Nucleic Acids Res.*, **42**, 5937–5948.
14. Kim,D., Lee,Y.H., Hwang,H.Y., Kim,K.K. and Park,H.J. (2010) Z-DNA binding proteins as targets for structure-based virtual screening. *Curr. Drug Targets*, **11**, 335–344.
15. Koeris,M., Funke,L., Shrestha,J., Rich,A. and Maas,S. (2005) Modulation of ADAR1 editing activity by Z-RNA in vitro. *Nucleic Acids Res.*, **33**, 5362–5370.
16. Placido,D., Brown,B.A., Lowenhaupt,K., Rich,A. and Athanasiadis,A. (2007) A left-handed RNA double helix bound by the Zα domain of the RNA-editing enzyme ADAR1. *Structure*, **15**, 395–404.
17. de Rosa,M., Zacarias,S. and Athanasiadis,A. (2013) Structural basis for Z-DNA binding and stabilization by the zebrafish Z-DNA dependent protein kinase PKZ. *Nucleic Acids Res.*, **41**, 9924–9933.
18. Teplova,M., Song,J., Gaw,H.Y., Teplov,A. and Patel,D.J. (2010) Structural insights into RNA recognition by the alternate-splicing regulator CUG-binding protein 1. *Structure*, **18**, 1364–1377.
19. Egli,M. and Gessner,R.V. (1995) Stereoelectronic effects of deoxyribose O4' on DNA conformation. *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 180–184.
20. Sponer,J., Gabb,H.A., Leszczynski,J. and Hobza,P. (1997) Base-base and deoxyribose-base stacking interactions in B-DNA and Z-DNA: a quantum-chemical study. *Biophys. J.*, **73**, 76–87.
21. Egli,M. and Sarkhel,S. (2007) Lone pair-aromatic interactions: to stabilize or not to stabilize. *Acc. Chem. Res.*, **40**, 197–205.
22. Mooibroek,T.J., Gamez,P. and Reedijk,J. (2008) Lone pair–π interactions: a new supramolecular bond? *CrystEngComm*, **10**, 1501–1515.
23. Gadre,R.S. and Kumar,A. (2015) In: Scheiner,S (ed). *Noncovalent Forces*. Springer, pp. 391–418.
24. Salonen,L.M., Ellermann,M. and Diederich,F. (2011) Aromatic rings in chemical and biological recognition: energetics and structures. *Angew. Chem. Int. Ed. Engl.*, **50**, 4808–4842.
25. Keating,K.S. and Pyle,A.M. (2010) Semiautomated model building for RNA crystallography using a directed rotameric approach. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 8177–8182.
26. Sokoloski,J.E., Godfrey,S.A., Dombrowski,S.E. and Bevilacqua,P.C. (2011) Prevalence of *syn* nucleobases in the active sites of functional RNAs. *RNA*, **17**, 1775–1787.
27. Lu,X.J., Bussemaker,H.J. and Olson,W.K. (2015) DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.*, **43**, e142.
28. Hickman,A.B., James,J.A., Barabas,O., Pasternak,C., Ton-Hoang,B., Chandler,M., Sommer,S. and Dyda,F. (2010) DNA recognition and the precleavage state during single-stranded DNA transposition in *D.radiodurans*. *EMBO J.*, **29**, 3840–3852.
29. Theobald,D.L. and Schultz,S.C. (2003) Nucleotide shuffling and ssDNA recognition in *Oxytrichanova* telomere end-binding protein complexes. *EMBO J.*, **22**, 4314–4324.
30. Auffinger,P. and Westhof,E. (2001) An extended structural signature for the tRNA anticodon loop. *RNA*, **7**, 334–341.
31. Davis,P.W., Adamiak,R.W. and Tinoco,I. (1990) Z-RNA: The solution NMR structure of r(CGCCCG). *Biopolymers*, **29**, 109–122.
32. Ennifar,E., Nikulin,A., Tishchenko,S., Serganov,A., Nevskaya,N., Garber,M., Ehresmann,B., Ehresmann,C., Nikonov,S. and Dumas,P. (2000) The crystal structure of UUCG tetraloop. *J. Mol. Biol.*, **304**, 35–42.
33. Antao,V.P. and Tinoco,I. (1992) Thermodynamic parameters for loop formation in RNA and DNA hairpin tetraloops. *Nucleic Acids Res.*, **20**, 819–824.
34. Sheehy,J.P., Davis,A.R. and Znosko,B.M. (2010) Thermodynamic characterization of naturally occurring RNA tetraloops. *RNA*, **16**, 417–429.
35. Marcia,M. and Pyle,A.M. (2012) Visualizing group II intron catalysis through the stages of splicing. *Cell*, **151**, 497–507.
36. Krasilnikov,A.S., Yang,X., Pan,T. and Mondragon,A. (2003) Crystal structure of the specificity domain of ribonuclease P. *Nature*, **421**, 760–764.
37. Nozinovic,S., Furtig,B., Jonker,H.R., Richter,C. and Schwalbe,H. (2010) High-resolution NMR structure of an RNA model system: the 14-mer cUUCGg tetraloop hairpin RNA. *Nucleic Acids Res.*, **38**, 683–694.
38. Porter,E.B., Marcano-Velazquez,J.G. and Batey,R.T. (2014) The purine riboswitch as a model system for exploring RNA biology and chemistry. *Biochim. Biophys. Acta*, **1839**, 919–930.
39. Klein,D.J., Edwards,T.E. and Ferre-D'Amare,A.R. (2009) Cocrystal structure of a class I preQ1 riboswitch reveals a pseudoknot recognizing an essential hypermodified nucleobase. *Nat. Struct. Mol. Biol.*, **16**, 343–344.
40. Liberman,J.A., Salim,M., Krucinska,J. and Wedekind,J.E. (2013) Structure of a class II preQ1 riboswitch reveals ligand recognition by a new fold. *Nat. Chem. Biol.*, **9**, 353–355.
41. Serganov,A., Polonskaia,A., Phan,A.T., Breaker,R.R. and Patel,D.J. (2006) Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature*, **441**, 1167–1171.
42. Garst,A.D., Heroux,A., Rambo,R.P. and Batey,R.T. (2008) Crystal structure of the lysine riboswitch regulatory mRNA element. *J. Biol. Chem.*, **283**, 22347–22351.
43. Tereshko,V., Skripkin,E. and Patel,D.J. (2003) Encapsulating streptomycin within a small 40-mer RNA. *Chem. Biol.*, **10**, 175–187.
44. Chen,J.H., Yajima,R., Chadalavada,D.M., Chase,E., Bevilacqua,P.C. and Golden,B.L. (2010) A 1.9 Å crystal structure of the HDV ribozyme precleavage suggests both Lewis acid and general acid mechanisms contribute to phosphodiester cleavage. *Biochemistry*, **49**, 6508–6518.
45. Veeraraghavan,N., Ganguly,A., Chen,J.H., Bevilacqua,P.C., Hammes-Schiffer,S. and Golden,B.L. (2011) Metal binding motif in the active site of the HDV ribozyme binds divalent and monovalent ions. *Biochemistry*, **50**, 2672–2682.
46. Chen,J., Ganguly,A., Miswan,Z., Hammes-Schiffer,S., Bevilacqua,P.C. and Golden,B.L. (2013) Identification of the catalytic Mg^{2+} ion in the hepatitis delta virus ribozyme. *Biochemistry*, **52**, 557–567.
47. Golden,B.L., Hammes-Schiffer,S., Carey,P.R. and Bevilacqua,P.C. (2013) In: Russel,R (ed). *Biophysics of RNA Folding*. Springer, NY, pp. 135–167.
48. Kapral,G.J., Jain,S., Noeske,J., Doudna,J.A., Richardson,D.C. and Richardson,J.S. (2014) New tools provide a second look at HDV ribozyme structure, dynamics and cleavage. *Nucleic Acids Res.*, **42**, 12833–12846.
49. Gilbert,S.D., Love,C.E., Edwards,A.L. and Batey,R.T. (2007) Mutational analysis of the purine riboswitch aptamer domain. *Biochemistry*, **46**, 13297–13309.
50. Colussi,T.M., Costantino,D.A., Hammond,J.A., Ruehle,G.M., Nix,J.C. and Kieft,J.S. (2014) The structural basis of transfer RNA mimicry and conformational plasticity by a viral RNA. *Nature*, **511**, 366–369.
51. Akiyama,B.M., Eiler,D. and Kieft,J.S. (2016) Structured RNAs that evade or confound exonucleases: function follows form. *Curr. Opin. Struct. Biol.*, **36**, 40–47.
52. Giegé,R., Juhling,F., Putz,J., Stadler,P., Sauter,C. and Florentz,C. (2012) Structure of transfer RNAs: similarity and variability. *WIREs RNA*, **3**, 37–61.
53. Ren,A., Rajashankar,K.R. and Patel,D.J. (2012) Fluoride ion encapsulation by Mg^{2+} ions and phosphates in a fluoride riboswitch. *Nature*, **486**, 85–89.
54. Noeske,J., Wasserman,M.R., Terry,D.S., Altman,R.B., Blanchard,S.C. and Cate,J.H. (2015) High-resolution structure of the *Escherichia coli* ribosome. *Nat. Struct. Mol. Biol.*, **22**, 336–341.
55. Garreau de Loubresse,N., Prokhorova,I., Holtkamp,W., Rodnina,M.V., Yusupova,G. and Yusupov,M. (2014) Structural basis for the inhibition of the eukaryotic ribosome. *Nature*, **513**, 517–522.
56. Khatter,H., Myasnikov,A.G., Natchiar,S.K. and Klaholz,B.P. (2015) Structure of the human 80S ribosome. *Nature*, **520**, 640–645.
57. Petrov,A.S., Bernier,C.R., Hershkovits,E., Xue,Y., Waterbury,C.C., Hsiao,C., Stepanov,V.G., Gaucher,E.A., Grover,M.A., Harvey,S.C. *et al.* (2013) Secondary structure and domain architecture of the 23S and 5S rRNAs. *Nucleic Acids Res.*, **41**, 7522–7535.
58. Walden,W.E., Selezneva,A.I., Dupuy,J., Volbeda,A., Fontecilla-Camps,J.C., Theil,E.C. and Volz,K. (2006) Structure of dual function iron regulatory protein 1 complexed with ferritin IRE-RNA. *Science*, **314**, 1903–1908.
59. Walden,W.E., Selezneva,A. and Volz,K. (2012) Accommodating variety in iron-responsive elements: crystal structure of transferrin

receptor 1 B IRE bound to iron regulatory protein 1. *FEBS Lett.*, **586**, 32–35.

60. Li,L. and Ye,K. (2006) Crystal structure of an H/ACA box ribonucleoprotein particle. *Nature*, **443**, 302–307.
61. Liang,B., Xue,S., Terns,R.M., Terns,M.P. and Li,H. (2007) Substrate RNA positioning in the archaeal H/ACA ribonucleoprotein complex. *Nat. Struct. Mol. Biol.*, **14**, 1189–1195.
62. Duan,J., Li,L., Lu,J., Wang,W. and Ye,K. (2009) Structural mechanism of substrate RNA recruitment in H/ACA RNA-guided pseudouridine synthase. *Mol. Cell.*, **34**, 427–439.
63. McCallum,S.A. and Pardi,A. (2003) Refined solution structure of the iron-responsive element RNA using residual dipolar couplings. *J. Mol. Biol.*, **326**, 1037–1050.
64. Zhou,J., Liang,B. and Li,H. (2011) Structural and functional evidence of high specificity of Cbf5 for ACA trinucleotide. *RNA*, **17**, 244–250.
65. Blower,T.R., Pei,X.Y., Short,F.L., Fineran,P.C., Humphreys,D.P., Luisi,B.F. and Salmond,G.P.C. (2011) A processed noncoding RNA regulates an altruistic bacterial antiviral system. *Nat. Struct. Mol. Biol.*, **18**, 185–246.
66. Herbert,A., Alfken,J., Kim,Y.G., Mian,I.S., Nishikura,K. and Rich,A. (1997) A Z-DNA binding domain present in the human editing enzyme, double-stranded RNA adenosine deaminase. *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 8421–8426.
67. Schwartz,T., Rould,M.A., Lowenhaupt,K., Herbert,A. and Rich,A. (1999) Crystal structure of the Zα domain of the human editing enzyme ADAR1 bound to left-handed Z-DNA. *Science*, **284**, 1841–1845.
68. Schwartz,T., Behlke,J., Lowenhaupt,K., Heinemann,U. and Rich,A. (2001) Structure of the DLM-1-Z-DNA complex reveals a conserved family of Z-DNA-binding proteins. *Nat. Struct. Biol.*, **8**, 761–765.
69. Brown,B.A. 2nd, Lowenhaupt,K., Wilbert,C.M., Hanlon,E.B. and Rich,A. (2000) The Zα domain of the editing enzyme dsRNA adenosine deaminase binds left-handed Z-RNA as well as Z-DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 13532–13536.
70. Abbas,Y.M., Pichlmair,A., Gorna,M.W., Superti-Furga,G. and Nagar,B. (2013) Structural basis for viral 5'-PPP-RNA recognition by human IFIT proteins. *Nature*, **494**, 60–64.
71. Vladimer,G.I., Gorna,M.W. and Superti-Furga,G. (2014) IFITs: emerging roles as key anti-viral proteins. *Front. Immunol.*, **5**.94
72. Hull,C.M., Anmangandla,A. and Bevilacqua,P.C. (2016) Bacterial riboswitches and ribozymes potently activate the human innate immune sensor PKR. *ACS Chem. Biol.*, **11**, 1118–1127.
73. Richardson,J.S., Schneider,B., Murray,L.W., Kapral,G.J., Immormino,R.M., Headd,J.J., Richardson,D.C., Ham,D., Hershkovits,E., Williams,L.D. *et al.* (2008) RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA*, **14**, 465–481.
74. Yang,L.X., Adam,C., Nichol,G.S. and Cockroft,S.L. (2013) How much do van der Waals dispersion forces contribute to molecular recognition in solution? *Nat. Chem.*, **5**, 1006–1010.
75. Mortimer,S.A. and Weeks,K.M. (2009) C2 '-endo nucleotides as molecular timers suggested by the folding of an RNA domain. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 15622–15627.
76. Fiore,J.L. and Nesbitt,D.J. (2013) An RNA folding motif: GNRA tetraloop-receptor interactions. *Q. Rev. Biophys.*, **46**, 223–264.
77. Gilbert,S.D., Stoddard,C.D., Wise,S.J. and Batey,R.T. (2006) Thermodynamic and kinetic characterization of ligand binding to the purine riboswitch aptamer domain. *J. Mol. Biol.*, **359**, 754–768.
78. Buck,J., Furtig,B., Noeske,J., Wohnert,J. and Schwalbe,H. (2007) Time-resolved NMR methods resolving ligand-induced RNA folding at atomic resolution. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 15699–15704.
79. Greenleaf,W.J., Frieda,K.L., Foster,D.A., Woodside,M.T. and Block,S.M. (2008) Direct observation of hierarchical folding in single riboswitch aptamers. *Science*, **319**, 630–633.
80. Leontis,N.B. and Westhof,E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA*, **7**, 499–512.
81. Webb,C.H., Riccitelli,N.J., Ruminski,D.J. and Luptak,A. (2009) Widespread occurrence of self-cleaving ribozymes. *Science*, **326**, 953.
82. Veeraraghavan,N., Ganguly,A., Golden,B.L., Bevilacqua,P.C. and Hammes-Schiffer,S. (2011) Mechanistic strategies in the HDV ribozyme: chelated and diffuse metal ion interactions and active site protonation. *J. Phys. Chem. B*, **115**, 8346–8357.
83. Grabow,W. and Jaeger,L. (2013) RNA modularity for synthetic biology. *F1000Prime Rep.* **5**, 46.
84. Lenert,P. (2010) Nucleic acid sensing receptors in systemic lupus erythematosus: development of novel DNA- and/or RNA-like analogues for treating lupus. *Clin. Exp. Immunol.*, **161**, 208–222.
85. Gaunt,E., Wise,H.M., Zhang,H., Lee,L.N., Atkinson,N.J., Nicol,M.Q., Highton,A.J., Klenerman,P., Beard,P.M., Dutia,B.M. *et al.* (2016) Elevation of CpG frequencies in influenza A genome attenuates pathogenicity but enhances host response to infection. *Elife*, **5**, e12735.
86. Taleb,N.N. (2010) *The Black Swan: The Impact of the Highly Improbable*. The Random House Publishing Group, NY.

3.4.2 Further remarks and outlook

The most biologically intriguing aspect of Z-steps is their involvement in the immune response, with the recognition of their unique backbone configurations by proteins participating in the interferon-induced immunology pathway. The immunogenicity of Z-DNA and Z-RNA is very well documented (Rich and Zhang 2003), however in the case of Z-steps it may imply that single-stranded RNA adopting Z-conformations could be significant for the discrimination between self and non-self. Further studies on the binding and recognition of Z-steps such as those found in UN<u>CG</u> tetraloops and immunology-related proteins have the potential to shed light on a still unknown structural aspect of antigen recognition. If Z-steps were immunogenic, we would have an interpretation for the observation that UNCG tetraloops, although thermodynamically very stable, are rare in PDB structures (GNRA:UNCG occurrence ratio among all RNAs is ~10:1).

Considering the protein recognition of Z-steps, it would be interesting to determine whether the event of protein binding happens after the RNA adopts a Z-conformation, or if the protein itself induces the conformation upon binding. In the Z-RNA structure (PDB: 2GXB; res.: 2.3 Å) a protein known to bind and stabilize Z-DNA, ADAR1 (dsRNA adenosine deaminase), was used to induce the conformational shift from A-RNA to Z-RNA of a duplex (Placido et al. 2007). This consideration opens perspectives on the role of Z-steps as transient and induced conformations, needed to accomplish specific functions upon induction by specific proteins.

As stated in the paper, we did not manage to find conserved structural patterns associated with Z-steps, except a general tendency to occur in junction regions and induce RNA backbone kinks and turns. These characteristics are conserved despite the diversity of RNA molecules, but a future expansion of available PDB structures could lead to the identification of general patterns of occurrence of Z-steps in RNAs. On their role as nucleation sites for RNA folding, further studies on assemblies of complex macromolecules such as ribosomes (Shajani et al. 2011) can highlight the importance of Z-conformations as local hinges formed in the first phases of ribosomal biogenesis.

4. Section II. Analysis of tetraloop folds and their tertiary interactions

4.1 Essential considerations on tetraloops

The phosphate-π and O4'-π interactions described in **Section I** have been found to participate in the structural signature of the most relevant tetraloop families, namely GNRA, UNCG and CUUG. Phosphate-π stacking occurs on the 5' side of the loop and is part of the U-turn signature, allowing the remaining three bases to be stacked with Watson-Crick edges exposed for tertiary interactions. On the other hand, O4'-π stacking occurs on 3' side of the loop, leaving only the 2nd residue to be exposed and limiting the possibility of base-base stacking inside the loop. This difference, generally overlooked by the numerous structural studies on tetraloops, yields structural fold elements that can be used to propose a classification scheme for tetraloops, based on tertiary interactions instead of the actual sequence-based method. Through an extensive PDB survey of hairpin loops, two major "turn" families have been identified, based on the U-turn and Z-turn. They can be used to describe almost all types of RNA tetraloop motifs. Moreover, several loop instances have sequences expected to belong to another family. These unexpected results show that the expectation of "one sequence-one fold" is not always respected and that structural descriptors based on backbone oxygen stacking interactions better grasp the plasticity of tetraloops.

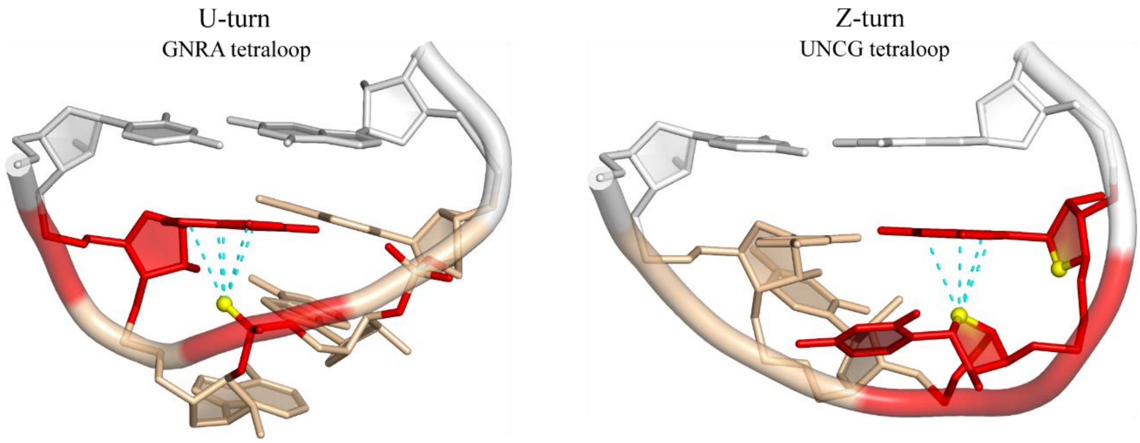
Another surprising outcome is that UNCG tetraloops can be involved in long range interactions inside complex architectures such as ribosomes. Up to now, they were thought to be isolated motifs in respect to long-range RNA-RNA interactions (Hall 2015).

4.2 U-turns and Z-turns: only two folds for tetraloops

The complex available ensemble of tetraloops families (**Table 1.1**) can be reduced to just two basic folds with two variants, based on stacking interactions involving backbone oxygen: U-turn and Z-turn. U-turns are sub-motifs found in anticodon loops and GNRAs, while Z-turns are found in UNCG and CUUG loops, among others.

4.2.1 *Paper 3. Revisiting GNRA and UNCG folds: U-turns versus Z-turns in RNA hairpin loops* (RNA, 2016)

Graphical abstract



GNRA and UNCG tetraloops are among the most common studied RNA structural motifs and have been characterized in terms of sequence preference, fold and interacting protein/RNA partners. We show here that they are the most relevant examples of a global classification scheme, based on structural considerations, that is dominated by just two folds: U-turns and Z-turns. In making this, we integrate the classical definition of the structural signature of U-turns with the stacking between the 1st nucleobase and the 3rd phosphate within the loop. Conversely, a Z-turn constitutes a novel definition for tetranucleotide folds, which is characterized by the stacking between the ribose of the 3rd residue and the 4th nucleobase involved in an O4'-π interaction.

Intriguingly, this binary classification indicated also that some sequences expected to fold into U-turns are instead shaped as Z-turns, and vice versa. Thus, the correlation between a given tetranucleotide sequence and a fold does is not always correct. We highlight this concept with selected examples of GNRA and UNCG dimorphism taken from structural data, considering tretraloops and tetranucleotide sequences embedded in larger loops (≤ 8 nts). A noteworthy example is the anticodon loop, which has to adopt a U-turn for its function, but can hold a UNCG sequence. NMR data pointed out the adoption of a Z-turn by an isolated anticodon loop, in which nucleotide modifications could play a role in dimorphism. Another widely known family of tetraloops, CUUG, adopt a variant of Z-turn, which we called Z_{anti}-turn.

Overall, the existence of only two turns for tetranucleotide folds in RNA hairpins and of their interconversion offers insights into how complex RNAs such as lncRNA may adopt discrete but transient and therefore hard-to-predict structures, not systematically deducible by their sequence.

Revisiting GNRA and UNCG folds: U-turns versus Z-turns in RNA hairpin loops

Luigi D'Ascenzo[1], Filip Leonarski[1,2], Quentin Vicens[1,*] and Pascal Auffinger[1,*]

[1] Université de Strasbourg, CNRS, Architecture et Réactivité de l'ARN, UPR 9002, F-67000 Strasbourg, France

[2] Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

* To whom correspondence should be addressed. Tel: +33 388417049; Fax: +33 388602218;

Email: p.auffinger@ibmc-cnrs.unistra.fr

Correspondence may also be addressed to Quentin Vicens.

Email: q.vicens@ibmc-cnrs.unistra.fr

Running title:

U-turn versus Z-turn in RNA hairpin loops

Keywords:

RNA folding; RNA motif; Structure prediction; Tetraloop; tRNA anticodon;

Journal: **RNA**

Section: **"Letter to the editor"**

ABSTRACT

When thinking about RNA three-dimensional structures, coming across GNRA and UNCG tetraloops is perceived as a boon since their folds have been extensively described. Nevertheless, analyzing loop conformations within RNA and RNP structures led us to uncover several instances of GNRA and UNCG loops that do not fold as expected. We noticed that when a GNRA does not assume its “natural” fold, it adopts the one we typically associate with a UNCG sequence. The same folding interconversion may occur for loops with UNCG sequences, for instance within tRNA anticodon loops. Hence, we show that some structured tetranucleotide sequences starting with G or U can adopt either of these folds. The underlying structural basis that defines these two fold types is the mutually exclusive stacking of a backbone oxygen on either the first (in GNRA) or the last nucleobase (in UNCG), generating an oxygen-π contact. We thereby propose to refrain from using sequences to distinguish between loop conformations. Instead, we suggest to use descriptors such as U-turn (for ‘GNRA-type’ folds) and a newly described Z-turn (for ‘UNCG-type’ folds). Because tetraloops adopt for the largest part only two (inter)convertible turns, we are better able to interpret from a structural perspective loop interchangeability occurring in ribosomes and viral RNA. In this respect, we propose a general view on the inclination for a given sequence to adopt (or not) a specific fold. We also suggest how long non-coding RNAs may adopt discrete but transient structures, which are therefore hard to predict.

RNA architecture is modular and hierarchical, which implies that secondary structural elements such as double stranded helices, hairpins and single stranded loops are linked by tertiary interactions that guide the assembly process (Hendrix et al. 2005; Cruz and Westhof 2009; Butcher and Pyle 2011). The majority of hairpin stems are capped by GNRA or UNCG tetranucleotide sequences —where N is any base and R is a purine (Cheong et al. 2015; Hall 2015). These tetranucleotide loops adopt distinctive folds that involve extensive and well-described networks of hydrogen bonds and stacking interactions (Cheong et al. 1990; Heus and Pardi 1991; Allain and Varani 1995; Jucker and Pardi 1995a; Jucker et al. 1996; Ennifar et al. 2000; Correll and Swinger 2003; Nozinovic et al. 2010). For GNRA and UNCG loops, it is generally assumed that the sequence commands a unique fold. Hence, upon considering sequence alignments and secondary structures of RNA families for which no 3D structures are available, we presume that we understand how these tetraloops will fold.

Here, we present structural evidence that challenges these expectations by identifying GNRA sequences that adopt a UNCG fold and vice-versa, both in tetraloops closed by a Watson-Crick base pair and in tetraloop-like motifs embedded in larger ribosomal and tRNA loops (Auffinger and Westhof 2001). Although this loop dimorphism remains rare within the pool of RNAs for which we currently possess 3D data, it led us to question some basic assumptions we make about RNA folding and structure prediction.

To better characterize these interconversions, we propose a more general structure-based tetraloop and tetraloop-like identification scheme that involves on one side the classical and well-described U-turn (Gutell et al. 2000) and, on the other, a newly defined "Z-turn", which is based on the UNCG tetraloop fold and the Z-RNA CpG step it encompasses (D'Ascenzo et al. 2016). We establish that these two turns and variants thereof are key to the tetraloop and tetraloop-like folding landscape, but also to most turns in RNAs. Atypical and infrequent tetranucleotide folds that do not conform to these rules will be described in more details elsewhere. Here, before pursuing, we need first to (re)define U-turns and Z-turns as they appear in structured tetranucleotide folds in hairpins (see also method section).

U-turn and U_{SH}-turn signatures

A U-turn is a tetranucleotide motif that was first identified in tRNA anticodon and T-loops (Quigley and Rich 1976; Gutell et al. 2000; Auffinger and Westhof 2001; Klosterman et al. 2004) and has since been characterized in a large variety of structural motifs starting with a uridine or a pseudo-uridine. In that respect, U-turns were sometimes called uridine-turns or π-turns (Kim and Sussman 1976; Jucker and Pardi 1995a). U-turns were also associated with "G-starting" motifs such as GNRA tetraloops (**Figure 1A**), or more recently in tetranucleotide motifs involving a protonated cytosine like a uC$^+$UAAu loop (Gottstein-Schmidtke et al. 2014). In short, a U-turn involves a hydrogen bond between the 1st nucleobase —generally with a U/G/C$^+$ imino or amino nitrogen atom— and an OP atom of the 4th nucleotide. This base-phosphate hydrogen bond is of the "5/4/3BPh" type according to a recent classification (Zirbel et al. 2009). It ensues that the 1-4 G•A *trans*-Sugar/Watson-Crick pair (*t*-SW) occurring in GNRA loops should not be considered as a U-turn determinant although it is essential for interactions with GNRA receptors (Fiore and Nesbitt 2013).

As an important outcome, the characteristic 1-4 nucleobase-phosphate (or nucleobase-OP) hydrogen bond imposes the formation of an oxygen-π or phosphate-π stacking contact between the 1st nucleobase and an OP atom of the 3rd nucleotide. A PDB survey led to an average 1-3 OP-π stacking distance of 3.0±0.2 Å, with a maximum distance of 3.5 Å. This oxygen-π contact, which is a further characteristic of U-turns, has rarely been described (Egli and Sarkhel 2007).

It emerges that these two features, namely the 1-4 nucleobase-OP hydrogen bond and the 1-3 OP-π stacking contacts, are sufficient to unambiguously characterize a U-turn. The latter criterion allows to further distinguish between regular and partially degenerated or unfolded U-turns, which correspond to loops with no oxygen-π stacking contact and are most often found at RNA-protein interfaces. However, such occurrences are rare (see next section).

A U-turn variant has been identified for UNAC sequences (**Figure 1B)**. These loops were found to mimic GNRA tetraloops since their backbone conformations are similar (Zhao et al. 2012). The 1-4 interaction involves a U•C *trans*-Sugar/Hoogsteen (*t*-SH) pair instead of a hydrogen bond involving the OP atom of the 4th nucleotide as in more typical U-turns. Yet, in the examples we collected, the OP-π contact between the 1st nucleobase and an OP atom of the 3rd nucleotide is conserved. In the following, we call this U-turn variant a "U_{SH}-turn" because of the consistent presence of a 1-4 *t*-SH pair.

Note that the cGANCg tetraloop in group IIC introns has a backbone that is similar to that of a U-turn and a 1-4 G•A *t*-SW pair (Keating et al. 2008). Although rare, these GANC loops are examples of structured tetraloops with no oxygen-π contact. For all U-turns, it is important to note that the last three nucleobases are stacked in a manner that their exposed Watson-Crick edges can establish specific tertiary contacts such as for example within anticodon-codon associations or with cognate receptors (Fiore and Nesbitt 2013; Tanaka et al. 2013).

Z-turn and Z_{anti}-turn signatures

UNCG tetraloops are not based on a U-turn but on a newly defined "Z-turn": they embed a *trans*-Sugar/Watson-Crick (*t*-SW) interaction between the 1st and 4th nucleobase, associated with a *C2'-endo* pucker of the 3rd residue, and a *syn* conformation of the 4th residue. In addition, the 3rd and 4th ribose rings adopt an uncommon head-to-tail orientation (**Figure 1C**). This particular combination of rare structural features is characteristic of Z-DNA/RNA motifs and implies an O4'-π stacking contact (Egli and Sarkhel 2007; D'Ascenzo et al. 2016). The 3-4 oxygen-π stacking contact in Z-turns is comparable with the 1-3 oxygen-π stacking contact in U-turns. Furthermore, the average stacking distance (3.1±0.2 Å) and the maximum distance (3.5 Å) are similar in both turns. Thus, we can assume that to define a Z-turn as found in UNCG loops, we can rely on both the 1-4 base pair essentially of the *t*-SW type as described below, and the 3-4 O4'-π stacking contact.

Such a definition is not based on the *syn* conformation of the 4th nucleotide and therefore allows to consider rare motifs where the O4' stacking involves bases in *anti*, such as found in some CUUG folds (**Figure 1D**) (Jucker and Pardi 1995b). Hence, as for U-turns, we can define two Z-turn subcategories: the main Z-turn or Z_{syn}-turn —with the 4th nucleobase in *syn*— and the less frequent "Z_{anti}-turn" variant —with the 4th nucleobase in *anti*. Most Z_{anti}-turns are not associated with a *t*-SW 1-4 pair but with a *cis*-Watson-Crick/Watson-Crick (*c*-WW) pair. As such, these Z_{anti}-turns are also known as di-loops. Interestingly, the characteristic *C2'-endo* sugar pucker of UNCG tetraloops seems to be conserved in all Z-turn types.

U-turns and Z-turns dominate the tetranucleotide folding landscape in RNA hairpins

In our unified definition of the two central U-turns and Z-turns in RNA hairpins, each turn is distinguished by the presence of either a 1-3 or a 3-4 oxygen-π contact (Egli and Sarkhel 2007). With the above-defined criteria, we searched the PDB for occurrences of these two turns and their variants in crystal and NMR structures, among tetranucleotide sequences embedded in RNA hairpin loops (**Table 1**). As expected, U-turns in tetranucleotide sequences starting with G, U or C^+ are the most frequent, followed by Z-turns in UNCG tetraloops. U_{SH}-turns are less frequent and are associated with UNAC sequences. Z_{anti}-turns are slightly more frequent and diverse, and comprise essentially CNNG sequences. The "Uncategorized" motifs are of the partially unfolded U-turn type —where the 1-4 interaction is present, but not the OP-π stacking contact. They correspond also to folds that are too rare and/or disordered to allow for their assignment to any clearly-defined category, or to partially unfolded conformations induced by proteins. The rare GANC tetranucleotide loop has only been identified in group IIC introns based on structural and phylogenetic evidence and has only been reported when bound to its cognate receptor (Keating et al. 2008). Thus, our early assumption that the largest part of tetranucleotide folds in hairpins is based on a U-turn or a Z-turn comprising an oxygen-π stacking contact is supported by this survey. Consequently, we can assume that most GNRA and UNCG tetranucleotide fold predictions based on sequence alignments are correct (**Table 1**).

However, these data also indicate that some sequences expected to form a U-turn are associated with a Z-turn and vice-versa. Thus, the sequence of a tetraloop does not systematically dictate its fold. For instance, we identified a GCAAu sequence that adopts a Z_{anti}-turn (**Figure 2**). Further, one GUGA sequence of the GNRA type adopting a Z-turn was observed in a RNA-protein complex (**Figure 3A**). NMR structures of anticodon loops containing the U_{33}NCG sequence were found to adopt a Z-turn under specific conditions, in agreement with their sequence but not with the expected anticodon-codon binding scheme (see below). These examples are more thoroughly described in the following sections. A detailed report describing the structural features of tetranucleotide folds will be provided elsewhere, the main purpose of this account being to establish the interchangeability between U-turns and Z-turns.

GNRA and GNYA dimorphism

Loop dimorphism came upon us serendipitously. We found that it deserved special attention, as we realized that it impacted our ability to derive three-dimensional structures from secondary structures. Upon looking at GNRA and GNYA loops, we noted that the phylogenetically conserved cGUGAg loop that caps helix 93 in domain V of all large ribosomal subunits adopts the expected U-turn. However, the same cGUGAg loop located within a 21 nucleotide long ribosomal fragment in complex with a pseudouridine synthase adopts an unexpected Z-turn, which is made possible through the formation of a 1-4 G•A *t*-SW pair (Czudnochowski et al. 2014) (**Figure 3A**). Whether the Z-turn is induced by the pseudouridine synthase or by crystal constraints is unclear. However, it is tempting to speculate that some RNA binding proteins and modification enzymes could recognize and/or induce Z-turns in GNRA sequences.

Loop dimorphism was also observed in larger motifs containing GNRA sequences, such as the phylogenetically conserved seven-nucleotide uGAAAgg loop that caps helix 35a in the domain II of large ribosomal subunits (Hsiao et al. 2006; Nasalean et al. 2009; D'Ascenzo et al. 2016). In every X-ray and cryo-EM structure of a ribosome available to date (including mitochondrial ribosomes), this uGAAAgg —or uGACAgg in *Homo sapiens* mitochondrial ribosomes (PDB code: 4WT8; resolution: 3.4 Å) (Amunts et al. 2015)— adopts a Z-turn (**Figure 3B**). Although it is imaginable that this GAAA sequence would not be folding like a regular GAAA tetraloop due to the larger size of the loop, we would probably have had difficulties in anticipating its Z-turn fold. However, to us, the most surprising example of a GNRA Z-turn —more precisely a Z_{anti}-turn— is a GCAAu pentaloop observed in X-ray structures of *Haloarcula marismortui* large subunits where it caps helix 12 within domain I. This GCAA Z_{anti}-turn shares a 1-4 *t*-SH G•A pair with a GNRA U-turn (see **Figures 1A** and **2**).

Further evidence of an exchange between U-turns and Z-turns originates from a combination of crystallographic and NMR data, which revealed that GNYA tetraloops —where Y is any pyrimidine— could fold like GNRA and adopt a U-turn since they can potentially form a 1-4 G•A *t*-SH pair (Melchers et al. 2006). But such loops are rare in X-ray structures. Up to now, besides the uGACAg located in the above-mentioned 4WT8 cryo-EM *Homo sapiens* mitochondrial ribosome, only one X-ray occurrence of a uGACAc in *Deinoccocus radiodurans* (**Figure 3C**) has been reported, where the tetranucleotide sequence adopts a U-turn (**Table 1**). Yet, NMR experiments illustrated that a cGUUAg loop (Ihle et al. 2005) and a uGCUAg loop (Melchers et al. 2006) can adopt a Z-turn rather than the anticipated U-turn (PDB codes: 1Z30 and 2EVY).

Overall, although such dimorphism is not frequent among structured RNAs (**Table 1**), it might be relevant when deriving the structures of non-coding RNA that may adopt several transient folds in order to achieve their functions within a large diversity of environments (Cech and Steitz 2014). It would therefore be interesting to explore how such conformational changes occur *in vivo*, especially since an *anti* to *syn* conversion could not easily be fathomed without stem unwinding.

UNCG dimorphism: U-turns or Z-turns in tRNA anticodon loops?

It is generally well appreciated that longer loops —from pentaloops to larger motifs— can embed tetranucleotide sequences that adopt U-turns (Hsiao et al. 2006). One of the most biologically relevant systems to incorporate this fold is the seven-nucleotide long tRNA anticodon loop. In the context of protein synthesis, any U_{33}NNN sequence will adopt a U-turn (Auffinger and Westhof 2001) so that the three anticodon bases are able to associate with the three complementary bases of the codon on the messenger RNA (mRNA). But would a U_{33}NCG anticodon sequence naturally adopt that classical U-turn conformation required for translation instead of the more cogent Z-turn? Do such anticodon loops manage to switch from U-turns to Z-turns and, if yes, which environmental context would direct such a structural transition, or impose one over the other fold?

In that respect, it could be envisaged that nucleotide modifications play a role in facilitating or preventing U_{33}NCG anticodon loops from adopting a Z-turn. NMR experiments were performed on four variants of $tRNA^{Arg1,2}$ stem-loops possessing a U_{33}ACG sequence and containing diverse combinations of RNA modifications such as A_{34}/I and C_{32}/S^2C —PDB codes: 2KRP/Q/V/W (Cantara et al. 2012). This study revealed that all modified and non-modified anticodon loops adopt a Z-turn, although the absence of a natural m^2A_{37} post-transcriptional modification could have biased the outcome. In any case, it seems fair to state that the extent of nucleotide modifications modulates the conformational plasticity of the $tRNA^{Arg1,2}$ anticodon loop in order to secure the essential U-turn conformation (Sundaram et al. 2000). But, in its unmodified state, the loop could also adopt a Z-turn and be recognized by specific proteins, as in the above-mentioned 4LGT pseudouridine synthase complex (**Figure 3A**).

To summarize, these U_{33}ACG anticodon sequences can successively adopt at least three distinct folds. They journey from a Z-turn in their free state, through a “degenerated” fold when bound to their cognate tRNA synthetases —see for example $tRNA^{Arg}$ with a U_{33}ICG anticodon; PDB code: 1F7U (Delagoutte et al. 2000)— to end with a classical U-turn when interacting with mRNA codons. RNA modifications —or their absence— may determine how anticodon loops fold, thereby altering or suppressing the tRNA codon-reading capacity.

Could Z-turns of U_{33}NCG anticodon loop sequences be associated with a specific biological function? Would a Z-turn be necessary for the recognition of modification sites by tRNA synthases? In that case, could Z-turns within anticodon loops also occur when other NpG steps replace CpG within the U_{33}NCG sequence? After all, it has been established that almost all dinucleotide sequences can adopt Z-RNA conformations (see **Figures 3A/B** for GpA and ApA Z-steps) and therefore be part of Z-turns (D'Ascenzo et al. 2016). Indeed, a NMR structure of a UCAGu pantaloop with an ApG Z-step has been reported —PDB code: 1Q75 (Theimer et al. 2003). If that hypothesis holds true, 16 out of the 64 anticodon sequences ending with a G —thereby comprising the four U_{33}NCG sequences— could potentially adopt a Z-turn. Our understanding of translation regulation, of decoding rules and of the role of modified bases in tRNAs could be expanded by these findings (Grosjean and Westhof 2016).

Are other folds possible for U_{33}NNN sequences? A different UGAA fold has been reported in the NMR structure of an RNA hairpin —PDB code: 1AFX (Butcher et al. 1997). However, we did not consider this fold since no 1-4 interaction was present and since this loop has not been reported elsewhere. We already described UNAC sequences (Zhao et al. 2012) that can adopt the alternative U_{SH}-turn variant, where the fold is made possible by the presence of a C36 nucleotide forming a 1-4 U•C *t*-SH pair (**Figure 1B**). We also identified a UUUAa pentanucleotide sequence in a ribosome structure that adopts the Z_{anti}-turn variant and that is closed by a 1-4 U-A *c*-WW pair (**Figure 3D**). Thus, U_{33}NNN anticodon loops can theoretically adopt any of the four folds we described, depending on the nature of nucleotide 36 and the associated structural context. Although most of these folds are rarely found in experimental structures, they can transiently appear in the folding pathways of these loops depending on sequence and modification levels.

Which turns for CNNN and ANNN sequences?

Similarly, we wondered whether CNNN sequences adopt a unique fold specific to their sequence or multiple conformations. When the C nucleotide is protonated, typical U-turns can be formed as shown by NMR and in ribosomes —see C_{1469}AACu in *Haloarcula marismortui* (Gottstein-Schmidtke et al.

2014). It was inferred from NMR and thermodynamic measurements (Proctor et al. 2002) as well as X-ray crystallography (**Figure 3E**) that CNNG sequences can form either Z-turns —PDB code: 1ROQ— (Du et al. 2003; Oberstrass et al. 2006; Schwalbe et al. 2008), or Z_{anti}-turns. For the latter, the 1-4 C=G *c*-WW pair was significantly buckled, probably due to constraints imposed by the "di-loop" fold —PDB code: 1RNG (Jucker and Pardi 1995b). Interestingly, the cCAAGg loop that caps helix 14 of the small subunits of eukaryotic ribosomes (**Figure 3E**) takes the place of a UACG loop in bacterial ribosomes, both forming a Z-turn. Besides UNNC, CNNC sequences could potentially form U_{SH}-turns, although the latter have not yet been observed (**Figure 3F**). Again, these loops starting with a C residue display an unanticipated plasticity, suggesting that the fold they adopt is largely context-dependent.

Tetranucleotide sequences starting with an adenine are almost non-existent, at least in crystallographic structures (**Table 1**). If they exist, they do not seem to display a significant and/or stable 1-4 contact as reported for the other loops described here. Hence, especially when the loop interacts with a protein, it is difficult to refer to these tetranucleotides as being "structured". However, we do not exclude the possibility that additional motifs might emerge in newly deposited crystal or NMR structures. For instance, since a UUUAa pentaloop with a Z_{anti}-turn implying a 1-4 U-A *c*-WW pair was observed, an ANNUn pentaloop with a similar turn and a 1-4 A-U pair cannot be dismissed. Such possibilities have been reported by NMR for uGUUC and CUUGu pentaloops adopting Z_{anti}-turns with a 1-4 G=C or C=G *c*-WW pair —PDB code: 2L6I (Lee et al. 2011).

Phylogenetic considerations on tetranucleotide loops in RNA

Phylogenetic data on 16S rRNA suggested early on that helix 6 (positions 83-86 in *Escherichia coli* 16S rRNA) is capped either by a CUUG (45%), a UUCG (36%) or a GCAA (13%) tetraloop (Woese et al. 1990; Konings and Gutell 1995). Thus, it could be concluded that this stem can be capped either by a Z-turn or by a U-turn. According to our present study, these three sequences can also adopt a Z-turn. Such loop polymorphism might complicate the interpretation of biochemical data, for example when highly conserved GAAA tetraloops in 16S rRNA are substituted by a UACG sequence (Sahu et al. 2012). In addition, the fact that this loop is unstructured in the 4YBB *Escherichia coli* crystal structure (resolution: 2.1 Å) might interrogate classical phylogenetic data interpretations. Indeed, in the seven UNCG tetranucleotide sequences deduced from the 16S *Escherichia coli* 2D structure, only three adopt a canonical Z-turn and the other sequences appear in disordered regions with, however, a G nucleotide in *syn* for four of them. The reasons as to why these loops appear as disordered are not yet understood.

Thus, sequence interchangeability might be hiding structural similarity. As noted above, the Z-turn GAAA loop capping helix 35a in the 50S of *Haloarcula marismortui* could exchange with YNMG sequences. Further, convincing evidence of sequence exchange that lead to similar folds have been reported in studies of viral RNA hairpins (Melchers et al. 2006; Liu et al. 2009; Zoll et al. 2011; Clabbers et al. 2014; Prostova et al. 2015).

Sequence-Structure relationships

It is our hope that the data we gathered (summarized in **Figure 4**) will help to interpret tetranucleotide sequence variations from a structural perspective, as they inform on the prevalence of a sequence to adopt (or not) a given fold. For example, GNNA sequences with a 1-4 G•A base pair can adopt a classical GNRA U-turn fold but also a Z-turn and even a Z_{anti}-turn, but not a U_{SH}-turn. Similarly, UNNG sequences can adopt U-turns and Z-turns, but not the two other less frequent variants. Finally, the GNNG and GNNU sequences are only found in the U-turn category. This classification reflects our current understanding of tetranucleotide turns and might be completed or refined with the advent of new non-coding RNA structures.

Final thoughts about folds and structure prediction

We report that tetraloop and tetranucleotide folds are not systematically determined by their sequence, possibly because of subtle changes in their environment and in the sequence of connected residues. A logical implication of this observation is that, for any given RNA sequence for which the 3D structure is not available, we are unable to ascertain with 100% confidence how the hairpins it contains will fold. With prior knowledge acquired on ribozymes (Schultes and Bartel 2000; Woodson 2015) and riboswitches (Garst et al. 2011; Batey 2015), we became aware that the same RNA sequence can adopt distinct folds in order to carry out specific functions. The structural analysis we present here reveals that only two folds dominate the tetranucleotide landscape. Consequently, predicting whether GNRA, UNCG or related sequences within any non-coding RNA will adopt a U-turn involving a phosphate-π stacking contact or a Z-turn with a O4'-π stacking ceases to be a straightforward exercise. Without additional stereochemical rules, the structure adopted by such tetranucleotide sequences might remain complex to predict and more structural information on these essential folds needs to be accumulated. It could therefore be informative to see how current 3D structure prediction methods would perform when confronted to such non-compliant pieces of the RNA puzzle (Miao et al. 2015).

Efforts to fold these tetranucleotide sequences by molecular dynamics simulations are currently only partially successful, although significant progress has been made into that direction (Kührova et al. 2013; Haldar et al. 2015; Miner et al. 2016). Such modelling attempts have now to face new challenges: finding not only one, but two or more folds, while grasping their relationship with the environment. Recently, some simple procedures based on diffusion maps and Markov models found the alternative Z-turn fold of a GAAA loop (Bottaro et al. 2016). Such methods are however currently limited to small fragments —four nucleotides and no closing base pair in that instance. Although this represents an essential first step in assessing folding pathways, it will certainly be much more challenging to predict the occurrence of such folds or turns embedded in the core of complex RNP particles like ribosomes.

Tetraloop fold variability probably only makes for the tip of the iceberg in the folding adaptability that characterizes regulatory RNAs. Regardless of how daunting they may seem, scenarios of folding plasticity at the local level are both attractive and relevant for molecules that comprise several thousands of nucleotides and that are thought to be mostly devoid of well-defined 3D structures (Gardini and Shiekhattar 2015; Rivas et al. 2016). We could envision how this plasticity of the most basic RNA folds would be well-suited to regulatory RNAs that are obligatory opportunists, by *nature*. The race is on toward "overturning more rules" about RNA structure and folding (Cech and Steitz 2014).

Methods

We searched the PDB (October 2016; X-ray data; resolution ≤ 3.0 Å) for tetranucleotide sequences in RNA hairpins that involve a 1-4 nucleobase-nucleotide interaction and an oxygen-π contact as defined below. For that purpose, we used the DSSR program (Lu et al. 2015). DSSR was also used to isolate tetranucleotide sequences embedded in loops comprising not more than 8 residues. For characterizing 1-3 and 3-4 oxygen-π contact, we specified in DSSR a 3.5 Å cutoff between the OP/O4' oxygen atom and the nucleobase plane. In addition, the projection of the OP/O4' oxygen on the base plane had to lie within the surface of the nucleobase aromatic cycles. A polygon-offset of 0.5 Å was used to take into account crystallographic inaccuracies. We also specified an interbase-angle ≤ 45° to discard severely distorted 1-4 base pairs. Finally, we specified that no atom belonging to the tetranucleotide sequence should have a *B-factor* above 79 Å^2. We visualized most of the structures, with a focus on those that appeared as borderline. In the insets of **Figures 1A/C**, the *d*(OP/O4'…π) histograms were calculated based on all oxygen-π contacts identified in RNA structures from the PDB and, therefore, not only on those found in tetraloop folds. To check for tetranucleotides with 1-4 interactions in NMR structures, we used the RNA FRABASE 2.0 database (Popenda et al. 2010).

For **Table 1**, we specified a redundancy criteria based on sequence and structural parameters (D'Ascenzo et al. 2016). If residues from two different tetranucleotide sequences (including the residues before and after the sequence) shared the same residue numbers, chain codes, ribose puckers, backbone dihedral angle sequences (we used the g+, g-, t categorization) and *syn/anti* conformations, they were considered as similar and the one with the best resolution was labelled as non-redundant. In case of matching resolutions, the nucleotide sequence with the lowest average *B-factor* was selected. Alike, if in a same structure two sequences shared the same residue numbers (with different chain codes) as well as ribose puckers, backbone dihedral angle sequences and *syn/anti* conformations, they were considered as similar and the one corresponding to the first biological unit was marked as non-redundant. To further limit redundancy in the largest ribosomal structures, we restricted our analysis to a single biological assembly. For more details, see reference (Leonarski et al. 2016). Note that it is impossible to eliminate redundancy from such a complex structural ensemble without eliminating at the same time significant data. Here, we provide an upper limit for a truly "non-redundant" tetranucleotide fold set.

ACKNOWLEDGEMENT

PA and QV wish to thank Prof. Eric Westhof for ongoing support and helpful discussions, as well as Prof. Neocles Leontis and Prof. Richard Giegé for useful comments on the manuscript. QV also acknowledges Dr. Yaser Hashem for support.

FUNDING

French ‘Ministère de la recherche et de l’enseignement’ (to L.D.); Polish Ministry of Higher Education and Science (Mobility Plus programme) [1103/MOB/2013/0 to F.L.]; LabEx: ANR-10-LABX-0036_NETRNA (to Q.V.). Funding for open access charge: CNRS.

Conflict of interest statement. None declared.

REFERENCES

Allain FH-T, Varani G. 1995. Structure of the P1 helix from group I self-splicing introns. *J. Mol. Biol.* **250**: 333-353.

Amunts A, Brown A, Toots J, Scheres SH, Ramakrishnan V. 2015. Ribosome. The structure of the human mitochondrial ribosome. *Science* **348**: 95-98.

Auffinger P, Westhof E. 2001. An extended structural signature for the tRNA anticodon loop. *RNA* **7**: 334-341.

Batey RT. 2015. Riboswitches: still a lot of undiscovered country. *RNA* **21**: 560-563.

Bottaro S, Gil-Ley A, Bussi G. 2016. RNA folding pathways in stop motion. *Nucleic Acids Res.* **44**: 5883-5891.

Butcher SE, Dieckmann T, Feigon J. 1997. Solution structure of the conserved 16S-like ribosomal RNA UGAA tetraloop. *J. Mol. Biol.* **268**: 348-358.

Butcher SE, Pyle AM. 2011. The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks. *Acc. Chem. Res.* **44**: 1302-1311.

Cantara WA, Bilbille Y, Kim J, Kaiser R, Leszczynska G, Malkiewicz A, Agris PF. 2012. Modifications modulate anticodon loop dynamics and codon recognition of E. coli tRNA(Arg1,2). *J. Mol. Biol.* **416**: 579-597.

Cech TR, Steitz JA. 2014. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* **157**: 77-94.

Cheong C, Varani G, Tinoco I. 1990. Solution structure of an unusually stable RNA hairpin, 5'GGAC(UUCG)GUCC. *Nature* **346**: 680-681.

Cheong H, Kim N, Cheong C. 2015. RNA structure: tetraloops. In *ELS*. John Wiley & Sons, ltd: Chichester.

Clabbers MTB, Olsthoorn RCL, Gultyaev AP. 2014. Tospovirus ambisense genomic RNA segments use almost complete repertoire of stable tetraloops in the intergenic region. *Bioinformatics* **30**: 1800-1804.

Correll CC, Swinger K. 2003. Common and distinctive features of GNRA tetraloops based on a GUAA tetraloop structure at 1.4 Å resolution. *RNA* **9**: 355-363.

Cruz JA, Westhof E. 2009. The dynamic landscapes of RNA architecture. *Cell* **136**: 604-609.

Czudnochowski N, Ashley GW, Santi DV, Alian A, Finer-Moore J, Stroud RM. 2014. The mechanism of pseudouridine synthases from a covalent complex with RNA, and alternate specificity for U2605 versus U2604 between close homologs. *Nucleic Acids Res.* **42**: 2037-2048.

D'Ascenzo L, Leonarski F, Vicens Q, Auffinger P. 2016. 'Z-DNA like' fragments in RNA: a recurring structural motif with implications for folding, RNA/protein recognition and immune response. *Nucleic Acids Res.* **44**: 5944-5956.

Delagoutte B, Moras D, Cavarelli J. 2000. Transfer-RNA aminoacylation by arginyl-transfer-RNA synthetase - Induced conformations during substrates binding. *EMBO J.* **19**: 5599-5610.

Du Z, Yu J, Andino R, James TL. 2003. Extending the family of UNCG-like tetraloop motifs: NMR structure of a CACG tetraloop from coxsackievirus B3. *Biochemistry* **42**: 4373-4383.

Egli M, Sarkhel S. 2007. Lone pair-aromatic interactions: to stabilize or not to stabilize. *Acc. Chem. Res.* **40**: 197-205.

Ennifar E, Nikulin A, Tishchenko S, Serganov A, Nevskaya N, Garber M, Ehresmann B, Ehresmann C, Nikonov S, Dumas P. 2000. The crystal structure of UUCG tetraloop. *J. Mol. Biol.* **304**: 35-42.

Fiore JL, Nesbitt DJ. 2013. An RNA folding motif: GNRA tetraloop-receptor interactions. *Q. Rev. Biophys.* **46**: 223-264.

Gardini A, Shiekhattar R. 2015. The many faces of long noncoding RNAs. *FEBS J.* **282**: 1647-1657.

Garst AD, Edwards AL, Batey RT. 2011. Riboswitches: structures and mechanisms. *Cold Spring Harb. Perspect. Biol.* **3**: a003533.

Gottstein-Schmidtke SR, Duchardt-Ferner E, Groher F, Weigand JE, Gottstein D, Suess B, Wohnert J. 2014. Building a stable RNA U-turn with a protonated cytidine. *RNA* **20**: 1163-1172.

Grosjean H, Westhof E. 2016. An integrated, structure- and energy-based view of the genetic code. *Nucleic Acids Res.* **44**: 8020-8040.

Gutell RR, Cannone JJ, Konings D, Gautheret D. 2000. Predicting U-turns in ribosomal RNA with comparative sequence analysis. *J. Mol. Biol.* **300**: 791-803.

Haldar S, Kuhrova P, Banas P, Spiwok V, Sponer J, Hobza P, Otyepka M. 2015. Insights into stability and folding of GNRA and UNCG tetraloops revealed by microsecond molecular dynamics and well-tempered metadynamics. *J. Chem. Theory and Comput.* **11**: 3866-3877.

Hall KB. 2015. Mighty tiny. *RNA* **21**: 630-631.

Hendrix DK, Brenner SE, Holbrook SR. 2005. RNA structural motifs: building blocks of a modular biomolecule. *Q. Rev. Biophys.* **38**: 221-243.

Heus HA, Pardi A. 1991. Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. *Science* **253**: 191-194.

Hsiao C, Mohan S, Hershkovitz E, Tannenbaum A, Williams LD. 2006. Single nucleotide RNA choreography. *Nucleic Acids Res.* **34**: 1481-1491.

Ihle Y, Ohlenschlager O, Hafner S, Duchardt E, Zacharias M, Seitz S, Zell R, Ramachandran R, Gorlach M. 2005. A novel cGUUAg tetraloop structure with a conserved yYNMGg-type backbone conformation from cloverleaf 1 of bovine enterovirus 1 RNA. *Nucleic Acids Res.* **33**: 2003-2011.

Jucker FM, Heus HA, Yop PF, Moors HHM, Pardi A. 1996. A network of heterogeneous hydrogen bonds in GNRA tetraloops. *J. Mol. Biol.* **264**: 968-980.

Jucker FM, Pardi A. 1995a. GNRA tetraloops make a U-turn. *RNA* **1**: 219-222.

-. 1995b. Solution structure of the CUUG hairpin loop: a novel RNA tetraloop motif. *Biochemistry* **34**: 14416-14427.

Keating KS, Toor N, Pyle AM. 2008. The GANC tetraloop: a novel motif in the group IIC intron structure. *J. Mol. Biol.* **383**: 475-481.

Kim SH, Sussman JL. 1976. π-turn is a conformational pattern in RNA loops and bends. *Nature* **260**: 645-646.

Klosterman PS, Hendrix DK, Tamura M, Holbrook SR, Brenner SE. 2004. Three-dimensional motifs from the SCOR, structural classification of RNA database: extruded strands, base triples, tetraloops and U-turns. *Nucleic Acids Res.* **32**: 2342-2352.

Konings DAM, Gutell R. 1995. A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S like rRNAs. *RNA* **1**: 559-574.

Kührova P, Banas P, Best RB, Sponer J, Otyepka M. 2013. Computer folding of RNA tetraloops? Are we there yet? *J. Chem. Theory Comput.* **9**: 2115-2125.

Lee CW, Li L, Giedroc DP. 2011. The solution structure of coronaviral stem-loop 2 (SL2) reveals a canonical CUYG tetraloop fold. *FEBS Lett.* **585**: 1049-1053.

Leonarski F, D'Ascenzo L, Auffinger P. 2016. Mg^{2+} ions: do they bind to nucleobase nitrogens? *Nucleic Acids Res.* **in press**.

Leontis NB, Westhof E. 2001. Geometric nomenclature and classification of RNA base pairs. *RNA* **7**: 499-512.

Liu PH, Li LC, Keane SC, Yang D, Leibowitz JL, Giedroc DP. 2009. Mouse hepatitis virus stem-loop 2 adopts a uYNMG(U)a-like tetraloop structure that is highly functionally tolerant of base substitutions. *J. Virol.* **83**: 12084-12093.

Lu XJ, Bussemaker HJ, Olson WK. 2015. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.* **43**: e142.

Melchers WJG, Zoll J, Tessari M, Bakhmutov DV, Gmyl AP, Agol VI, Heus HA. 2006. A GCUA tetranucleotide loop found in the poliovirus oriL by in vivo SELEX (un)expectedly forms a YNMG-like structure: Extending the YNMG family with GYYA. *RNA* **12**: 1671-1682.

Miao Z, Adamiak RW, Blanchet MF, Boniecki M, Bujnicki JM, Chen SJ, Cheng C, Chojnowski G, Chou FC, Cordero P et al. 2015. RNA-Puzzles Round II: assessment of RNA structure prediction programs applied to three large RNA structures. *RNA* **21**: 1066-1084.

Miner JC, Chen AA, Garcia AE. 2016. Free-energy landscape of a hyperstable RNA tetraloop. *Proc. Natl. Acad. Sci. USA* **113**: 6665-6670.

Nasalean L, Stombaugh J, Zirbel CL, Leontis NB. 2009. RNA 3D structural motifs: definition, identification, annotation, and database searching. In *Non-Protein Coding RNAs*, (ed. NG Walter, SA Woodson, RT Batey), pp. 1-26. Springer, Berlin Heidelberg.

Nozinovic S, Furtig B, Jonker HR, Richter C, Schwalbe H. 2010. High-resolution NMR structure of an RNA model system: the 14-mer cUUCGg tetraloop hairpin RNA. *Nucleic Acids Res.* **38**: 683-694.

Oberstrass FC, Lee A, Stefl R, Janis M, Chanfreau G, Allain FH. 2006. Shape-specific recognition in the structure of the Vts1p SAM domain with RNA. *Nat. Struct. Mol. Biol.* **13**: 160-167.

Popenda M, Szachniuk M, Blazewicz M, Wasik S, Burke EK, Blazewicz J, Adamiak RW. 2010. RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. *BMC Bioinformatics* **11**: 231.

Proctor DJ, Schaak JE, Bevilacqua JM, Falzone CJ, Bevilacqua PC. 2002. Isolation and characterization of a family of stable RNA tetraloops with the motif YNMG that participate in tertiary interactions. *Biochemistry* **41**: 12062-12075.

Prostova MA, Gmyl AP, Bakhmutov DV, Shishova AA, Khitrina EV, Kolesnikova MS, Serebryakova MV, Isaeva OV, Agol VI. 2015. Mutational robustness and resilience of a replicative cis-element of RNA virus: Promiscuity, limitations, relevance. *RNA Biol.* **12**: 1338-1354.

Quigley GJ, Rich A. 1976. Structural domains of transfer RNA molecules. *Science* **194**: 796-806.

Rivas E, Clements J, Eddy SR. 2016. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat Methods*: doi: 10.1038/nmeth.4066.

Sahu B, Khade PK, Joseph S. 2012. Functional replacement of two highly conserved tetraloops in the bacterial ribosome. *Biochemistry* **51**: 7618-7626.

Schultes EA, Bartel DP. 2000. One sequence, two ribozymes: Implications for the emergence of new ribozyme folds. *Science* **289**: 448-452.

Schwalbe M, Ohlenschlager O, Marchanka A, Ramachandran R, Hafner S, Heise T, Gorlach M. 2008. Solution structure of stem-loop alpha of the hepatitis B virus post-transcriptional regulatory element. *Nucleic Acids Res.* **36**: 1681-1689.

Sundaram M, Durant PC, Davis DR. 2000. Hypermodified nucleosides in the anticodon of tRNALys stabilize a canonical U-turn structure. *Biochemistry* **39**: 12575-12584.

Tanaka T, Furuta H, Ikawa Y. 2013. Natural selection and structural polymorphism of RNA 3D structures involving GNRA loops and their receptor motifs. In *RNA Nanotechnology and Terapeuthics*, (ed. P Guo), pp. 109-120. CRC Press.

Theimer CA, Finger LD, Feigon J. 2003. YNMG tetraloop formation by a dyskeratosis congenita mutation in human telomerase RNA. *RNA* **9**: 1446-1455.

Woese CR, Winker S, Gutell RR. 1990. Architecture of ribosomal RNA: Constraints on the sequence of "tetra-loops". *Proc. Natl. Acad. Sci. USA* **87**: 8467-8471.

Woodson SA. 2015. RNA folding retrospective: lessons from ribozymes big and small. *RNA* **21**: 502-503.

Zhao Q, Huang HC, Nagaswamy U, Xia Y, Gao X, Fox GE. 2012. UNAC tetraloops: to what extent do they mimic GNRA tetraloops? *Biopolymers* **97**: 617-628.

Zirbel CL, Sponer JE, Sponer J, Stombaugh J, Leontis NB. 2009. Classification and energetics of the base-phosphate interactions in RNA. *Nucleic Acids Res.* **37**: 4898-4918.

Zoll J, Hahn MM, Gielen P, Heus HA, Melchers WJ, van Kuppeveld FJ. 2011. Unusual loop-sequence flexibility of the proximal RNA replication element in EMCV. *PLoS One* **6**: e24818.

Table 1. Number of U-turns, Z-turns and their variants (as defined in the text and in **Figure 1**) associated with tetranucleotide sequences involving a 1-4 nucleobase-nucleotide contact and occurring in RNA hairpin loops not longer than eight residues. These data were derived from a survey of X-ray structures from the PDB (October 2016; resolution ≤ 3.0 Å). The estimated number of non-redundant occurrences is given in brackets. Tetranucleotide sequences having at least one atom with a *B-factor* > 79 Å^2 were excluded. "NMR" in the table refers to folds for which only NMR structures are available; the corresponding PDB codes are given in parenthesis. These structures are not included in the total.

Sequence	U-turn	U$_{SH}$-turn	Z-turn	Z$_{anti}$-turn	Uncategorized
GNRA					
in tetraloops	1353 [416]	—	2 [2]	—	38 [21] [a]
in larger hairpins	515 [151]	—	93 [20]	68 [29] [d]	45 [29] [a]
GNRG					
in tetraloops	47 [17]	—	—	—	—
in larger hairpins	106 [29]	—	—	—	5 [4] [a]
GNNY					
in tetraloops	1 [1]	—	—	1 [1]	4 [3] [b]
in larger hairpins	18 [11]	—	—	3 [3]	75 [29] [c]
GNYA					
in tetraloops	—	—	NMR (1Z30)	—	—
in larger hairpins	1 [1]	—	NMR (2EVY)	—	1 [1] [c]
GNYG					
in tetraloops	—	—	—	—	—
in larger hairpins	12 [7]	—	—	—	1 [1] [c]
UNCG					
in tetraloops	—	—	147 [43]	—	5 [4] [a]
in larger hairpins	6 [4]	—	NMR (1TXS)	—	—
UNNN (not UNCG)					
in tetraloops	—	46 [13]	NMR (2MQT/V)	—	—
in larger hairpins	706 [252]	—	—	2 [1]	55 [24] [d]
C(+)AAC					
in tetraloops	—	—	—	—	—
in larger hairpins	57 [2]	—	—	—	—
CNNG					
in tetraloops	—	—	6 [4]	NMR (1RNG)	—
in larger hairpins	—	—	NMR (1ROQ/2L6I)	74 [52]	—
CNN(notG)					
in tetraloops	—	—	—	—	3 [2] [a]
in larger hairpins	—	—	—	—	24 [9] [d]
ANNN					
in tetraloops	—	—	—	—	12 [6] [d]
in larger hairpins	—	—	—	—	111 [51] [d]
Total	2822 [891]	46 [13]	248 [69]	148 [86]	379 [184]

[a] Mostly U/Z-turn-like, but with non-standard geometry (oxygen-π stacking or hydrogen bond distances above 3.5 Å);
[b] GANC loops in group IIC introns;
[c] Mostly tetraloop folds in hairpins that are not inducing turns (will be discussed elsewhere);
[d] Mostly unstructured;

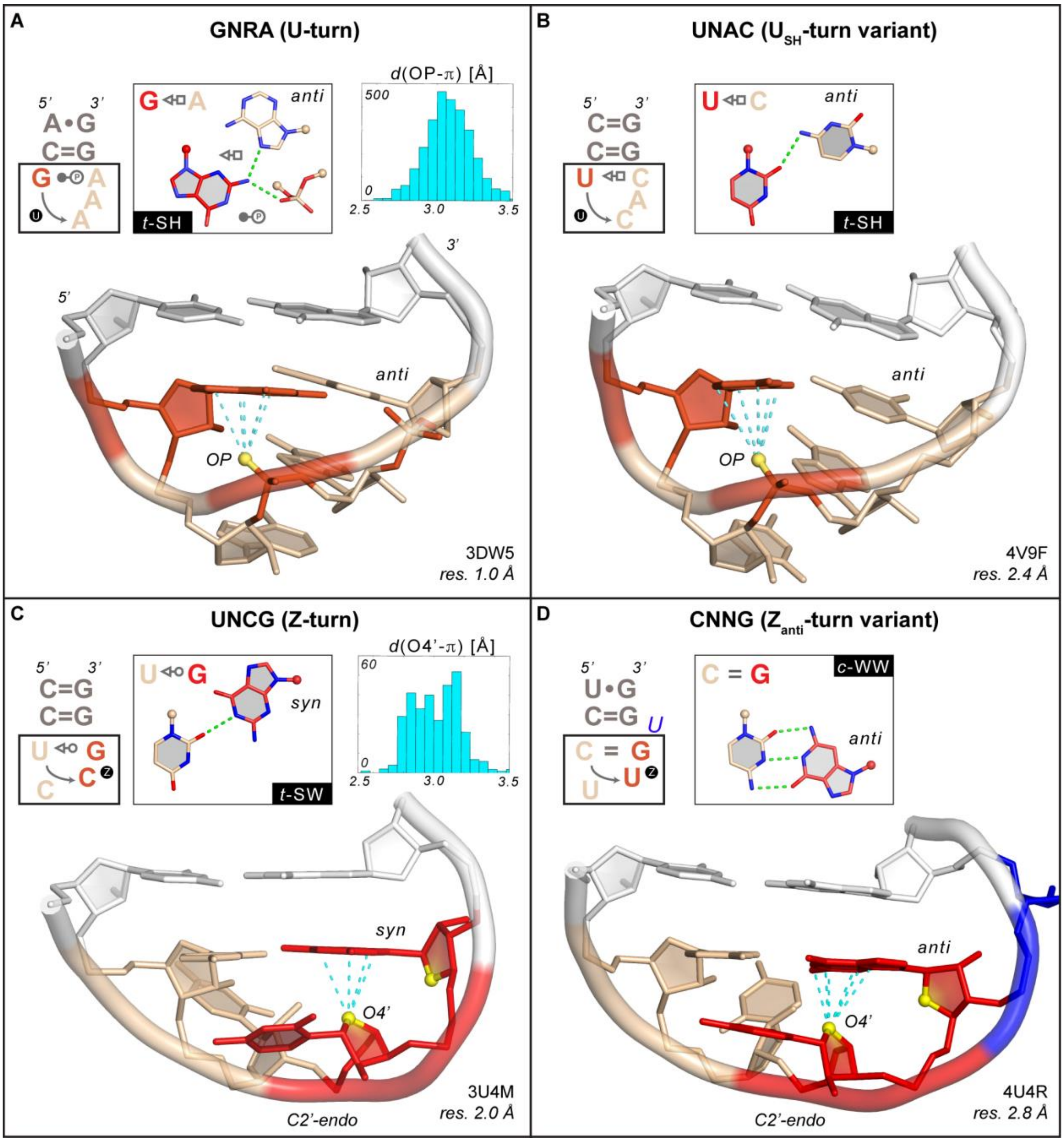


Figure 1. Examples of a GNRA "U-turn" and a UNCG "Z-turn" along with their U$_{SH}$-turn and Z$_{anti}$-turn variants (1-4 base pairs and relevant nucleobase-phosphate hydrogen bonds are shown in the insets). In all panels, the cyan lines mark contact distances between the OP/O4' atoms —emphasized as yellow spheres— and the stacked nucleobase that are associated with oxygen-π contacts ≤ 3.5 Å (see method section and insets of panels A/C). For clarity, all non-relevant OP atoms were hidden. The C=G closing base pairs are shown in white. For all secondary structures, symbols according to the Leontis and Westhof nomenclature were used (Leontis and Westhof 2001; Nasalean et al. 2009).
(**A**) G$_{2659}$AAA tetraloop (chain A) adopting a classical U-turn (symbolized by a circled "U"). The 1st G and the phosphate of the 3rd nucleotide involved in an OP-π contact are marked in red as well as the oxygen atoms of the phosphate involved in the 1-4 base-phosphate hydrogen bond. The three stacked A nucleotides are colored in wheat.

(**B**) U_{253}CAC tetraloop (chain 0) adopting the rare U_{SH}-turn variant (symbolized by a circled “U”). The 1st U and the phosphate of the 3rd nucleotide are marked in red. The three stacked CAC nucleobases and part of their backbone are colored in wheat.

(**C**) U_{2144}CCG tetraloop (chain B) adopting a Z-turn (symbolized by a circled “Z”). The CpG step forming a Z-RNA motif is shown in red. The two ribose O4’ atoms of the CpG step are shown in yellow to mark the characteristic head-to-tail orientation of the sugars. The 4th nucleotide adopts a *syn* conformation. The UpC step is colored in wheat.

(**D**) C_{3194}UUGu pentaloop (chain 1) adopting a rare Z_{anti}-turn variant (symbolized by a circled “Z”). The UpG step forming a Z-RNA motif, with the G adopting an *anti* instead of a *syn* conformation, is shown in red. The two ribose O4’ atoms of the CpG step are shown in yellow to mark the characteristic head-to-tail orientation of the sugars. The CpU step is colored in wheat and the bulged “u” in blue.

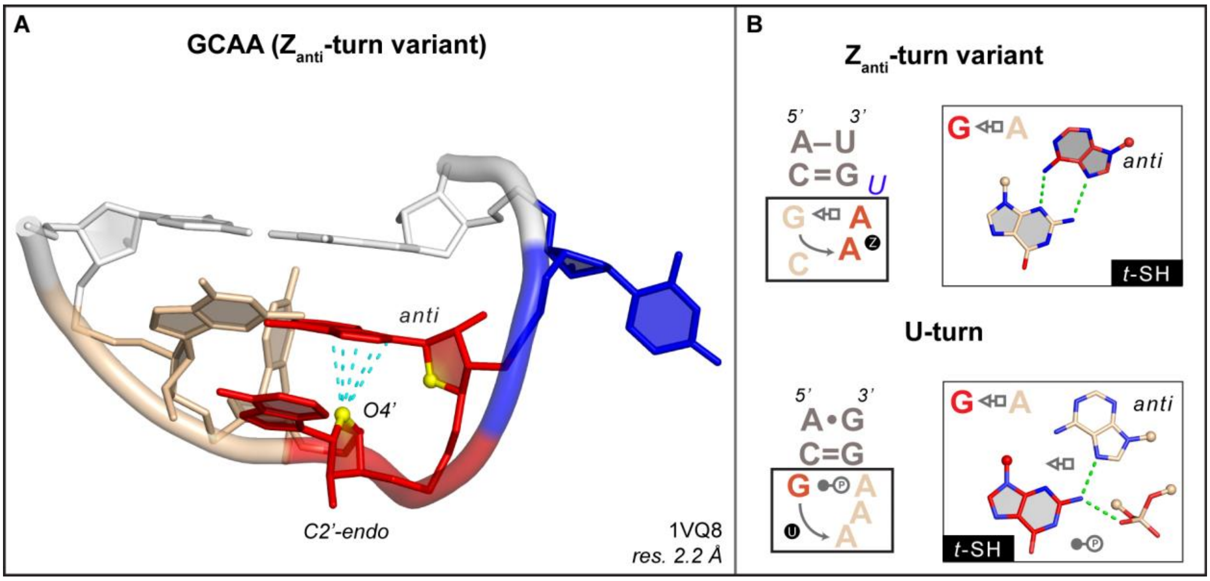


Figure 2. G_{196}CAAu sequence (chain 0) adopting a rare Z_{anti}-turn variant.

(**A**) The ApA step forming a Z-RNA motif, with the A adopting an *anti* instead of a *syn* conformation, is shown in red. The two ribose O4' atoms of the CpG step are shown in yellow to mark the characteristic head-to-tail orientation of the sugars. The GpC step is colored in wheat, the bulged U in blue and the closing base pair in white.

(**B**) Comparison of the secondary structures and of the associated 1-4 G•A *t*-SH pairs for the Z_{anti}- and the U-turns, to emphasize their differences. See also **Figure 1A** for the GAAA U-turn.

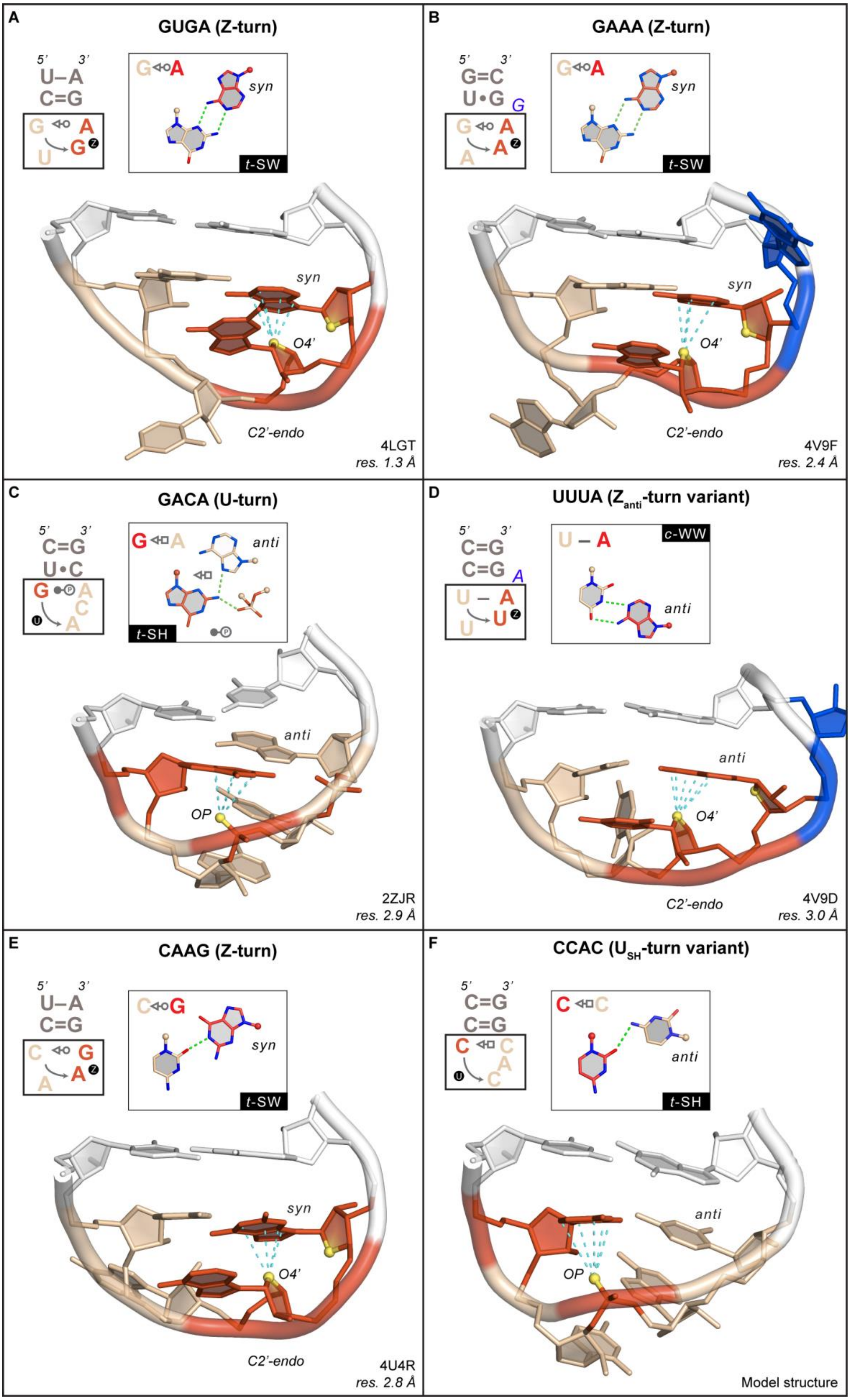

A
GUGA (Z-turn)
5′ 3′
U–A
C=G
G A
U G
G A
syn
t-SW
syn
O4′
C2′-endo
4LGT
res. 1.3 Å
B
GAAA (Z-turn)
5′ 3′
G=C
U•G G
G A
A A
G A
syn
t-SW
syn
O4′
C2′-endo
4V9F
res. 2.4 Å
C
GACA (U-turn)
5′ 3′
C=G
U•C
G A
C
A
G A
anti
t-SH
anti
OP
2ZJR
res. 2.9 Å
D
UUUA (Z$_{anti}$-turn variant)
5′ 3′
C=G
C=G A
U – A
U U
U – A
c-WW
anti
anti
O4′
C2′-endo
4V9D
res. 3.0 Å
E
CAAG (Z-turn)
5′ 3′
U–A
C=G
C G
A A
C G
syn
t-SW
syn
O4′
C2′-endo
4U4R
res. 2.8 Å
F
CCAC (U$_{SH}$-turn variant)
5′ 3′
C=G
C=G
C C
A
C
C C
anti
t-SH
anti
OP
Model structure

Figure 3. Examples of tetranucleotide sequences adopting unanticipated folds (1-4 base pairs are shown in the insets). In all panels, the cyan lines mark contact distances between the OP/O4' atoms —in yellow— and the stacked nucleobase that are associated with oxygen-π contacts ≤ 3.5 Å (see method). For clarity, all non-essential OP atoms were hidden. All closing base pairs are shown in white. All turns are symbolized by a circled "U" or "Z" as in **Figure 1**.

(**A**) G_{2595}UGA sequence (chain E) adopting a Z-turn. The Z-RNA GpA step is shown in red. The O4' atoms of the two GpA ribose's are shown in yellow to mark the characteristic head-to-tail orientation of the sugars. The GpU step is colored in wheat.

(**B**) G_{873}AAAg sequence (chain 0) embedded in a seven-nucleotide loop and adopting a Z-turn. The ApA step that forms a Z-RNA motif is shown in red. The O4' atoms of the two ApA ribose's are shown in yellow to mark the characteristic head-to-tail orientation of the sugars. The GpA step is colored in wheat; the bulged "g" nucleotide is shown in blue.

(**C**) G_{2796}ACA sequence (chain X) adopting a classical U-turn. The 1st G and the phosphate of the 3rd nucleotide involved in an OP-π contact are marked in red as well as the oxygen atoms of the phosphate involved in the 1-4 base-phosphate contact. The stacked ACA nucleotides are colored in wheat.

(**D**) U_{2595}UUAa sequence (chain DA) adopting a Z_{anti}-turn. The UpA step that forms a Z-RNA motif is shown in red. The O4' atoms of the two UpA ribose's are shown in yellow to mark the characteristic head-to-tail orientation of the sugars. The UpU step is colored in wheat; the bulged "a" nucleotide is shown in blue.

(**E**) C_{415}AAG sequence (chain 2) adopting a Z-turn. The ApG step that forms a Z-RNA motif is shown in red. The O4' atoms of the two ApG ribose's are shown in yellow to mark the characteristic head-to-tail orientation of the sugars. The CpA step is colored in wheat.

(**F**) Model structure of a CCAC sequence adopting a U_{SH}-turn. The 1st C and the phosphate of the 3rd nucleotide involved in an OP-π contact are marked in red. The three stacked CAC nucleotides are colored in wheat.

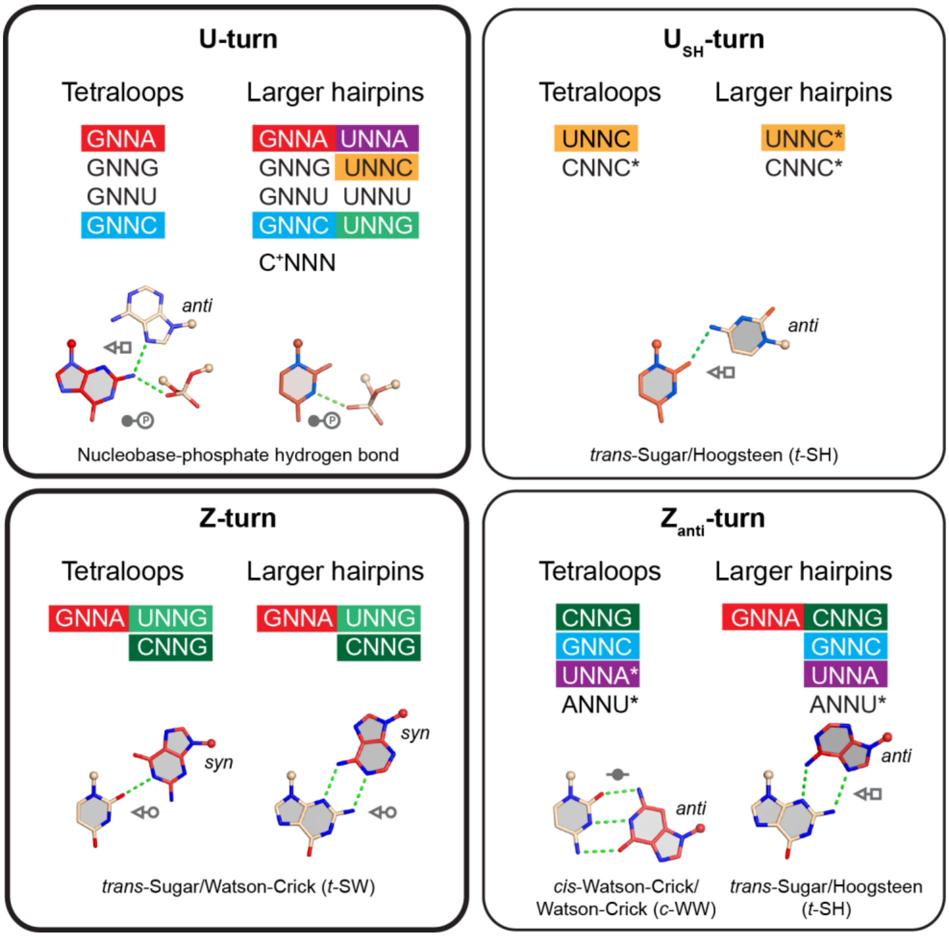


Figure 4. Graphical representation of the sequence-structure relationships for the four —two main and two minor— tetranucleotide turns that we characterized in RNA hairpins. The nucleobase in red is associated with a 1-3 or 3-4 oxygen-π stacking contact. The folds associated with sequences marked by an asterisk are theoretically possible but have not yet been observed in experimental structures. Here, we consider only the 1st and 4th nucleotides. Sequence-structure relationships associated with the 2nd and 3rd nucleotides will be discussed elsewhere.

4.3 UNCG receptors in ribosomes

In opposition to GNRA tetraloops, UNCG have always been considered as “loners” and not involved in specific RNA/protein or RNA/RNA interactions (Hall 2015). It was therefore unexpected to find during the surveys on tetraloops occurrences of *UNCG receptors* in the conserved ribosomal core. In particular, the Z-turn characteristic of these loops allows the 2nd loop residue to be exposed and pointing towards the helix minor groove; this base undergoes in specific cases long-range tertiary interactions with a tetraloop receptor. These receptors show the general characteristic to be composed by a complex assembly of helices and/or strands, again in opposition to the “simpler” GNRA receptors (Fiore and Nesbitt 2013). Thus, the occurrence of UNCG receptors can be observed only in highly complex RNA architectures, such as the two occurrences observed inside prokaryotic and eukaryotic ribosomes with loops capping helices 62 and 35a (**Fig 4.1**).

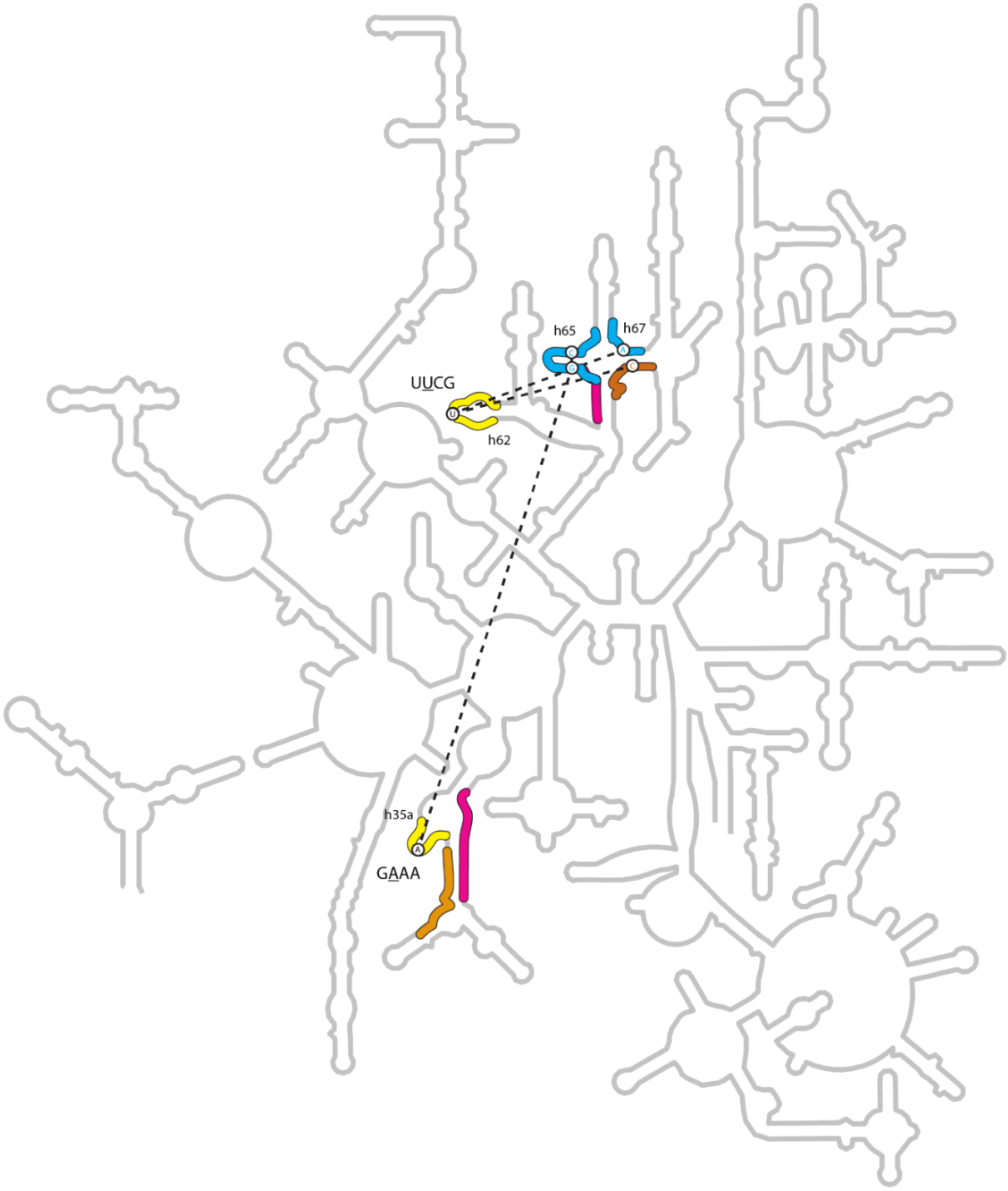


Figure 4.1. UNCG receptors in *E. coli* LSU. A UUCG loop capping helix 62 (yellow) and a GAAA loop capping helix 35a (yellow) both adopt a Z-turn. The 2nd loop base (highlighted with a circle) in both cases interacts with a receptor formed by the assembly of multiple helices (highlighted with colors). Dashed lines represent long-range interactions.

4.3.1 Receptor I involves the UUCG capping helix 62

The UUCG tetraloop capping helix 62 is the only UNCG conserved among all species of the ribosomal structures deposed in the PDB (except for mitochondrial ribosomes, **Table 4.1**). Helix 62

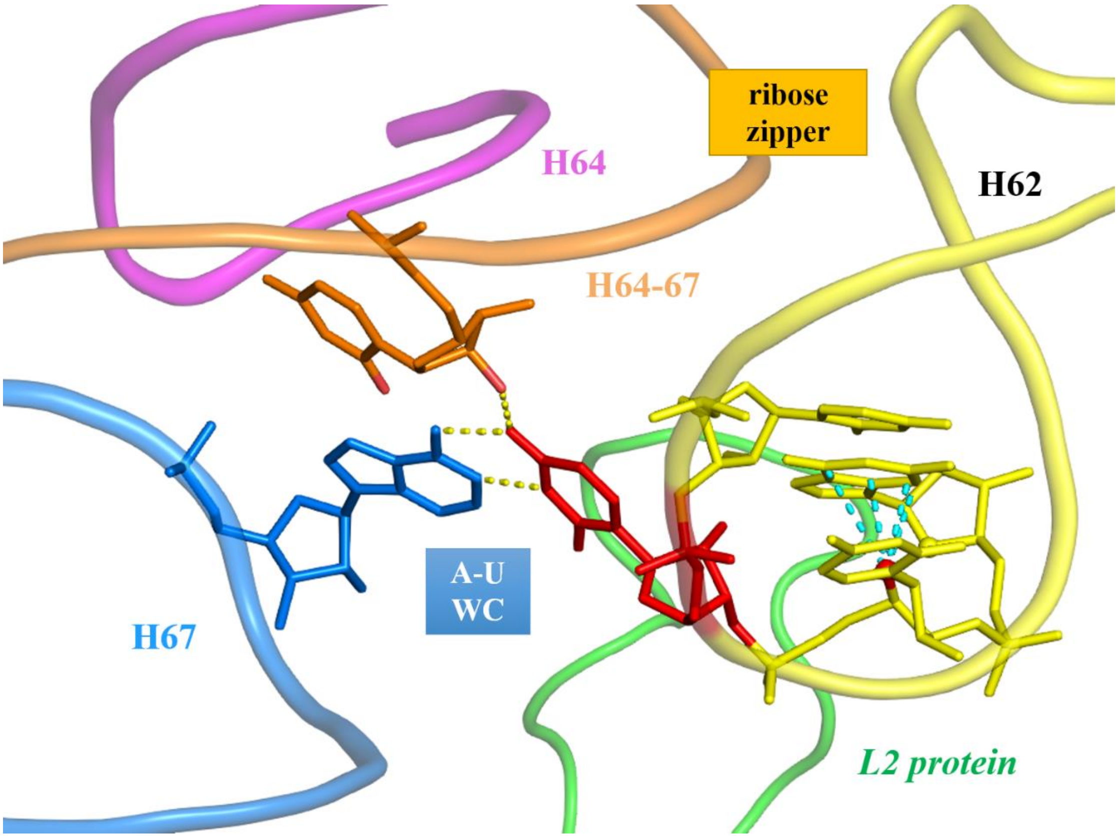


Figure 4.2. 3D representation of a UUCG receptor in *E. coli* LSU. Helix 62 (yellow) is capped by a UUCG loop in which the 2nd residue (uridine, red) makes long-range interactions with two nucleotides of helix 67 (light blue and orange). One of the contacts involves a U-A canonical Watson-Crick base pair. Ribosomal L2 protein (green) is also found in the vicinity of the receptor. Interatomic contacts with distances <3.5 Å are shown with dashed lines. Color codes for strands are consistent with **Fig. 4.1**.

is located in the subdomain IV of the LSU and makes tertiary interactions with helix 67. It is situated on the interface between subunits, being close to helix 44 of the SSU and part of the intersubunit bridge B5 (Liu and Fredrick 2016). The 2nd residue of the UUCG makes a canonical Watson-Crick base pair with a bulged adenine localized in the junction between helices 66 and 67 (**Fig. 4.2**).

The Watson-Crick U-A pair is part of a second order pseudoknot (Antczak et al. 2014), which complicates the local topology of the receptor complex. The interaction network is completed by a hydrogen bond between the loop uridine O4 and the hydroxyl O2’ of a residue belonging to another region of helix 67. In addition, there are several interactions involving atoms from the minor groove of helix 62 and the backbone of the helix 67. Based on phylogenetic data (Petrov et al. 2013) the residues of UUCG and the bulged adenine making the U-A pair are highly conserved in the ribosomes of more than 120 species. Conversely, the residue to which the O2’ belongs appears less conserved. Several amino acids of the ribosomal L2 protein interact with the UUCG loop, such in the case of *E. coli* ribosome where an Arg makes a hydrogen bond with the N7 of the 4th loop residue. It is possible to speculate on the fact that during the ribosome assembly helix 67 adopts its final fold first, then the helix 62 carrying the loop docks over its surface being locked by a strong network of interactions. Finally, L2 protein locks in place the helical segments forming the tetraloop receptor. An interesting observation in this perspective is that helix 62 is missing in the porcine mitochondrial ribosome (Greber et al. 2015), while helix 67 is similarly at the surface of the ribosome. However, the adenine involved in the U-A pair is replaced by a uridine and it is not bulged from the helix.

4.3.2 Receptor II involves the GAAA capping helix 35a

Helix 35a within the domain I of the LSU is capped in prokaryotic and eukaryotic species by a GAAA loop, adopting an uncharacteristic Z-turn (**Paper 3**). Helix 35a is situated in the vicinity of the ribosomal peptidyl transferase center, thus being highly phylogenetically conserved (Petrov et al. 2013) also in the mitochondrial ribosomes (**Table 4.1**). Typical of the fold adopted by the loop, the 2nd loop residue (A) points towards the loop minor groove, participating in a type I A-minor motif with a canonical C=G Watson-Crick pair located on helix 65 (**Fig. 4.3**).

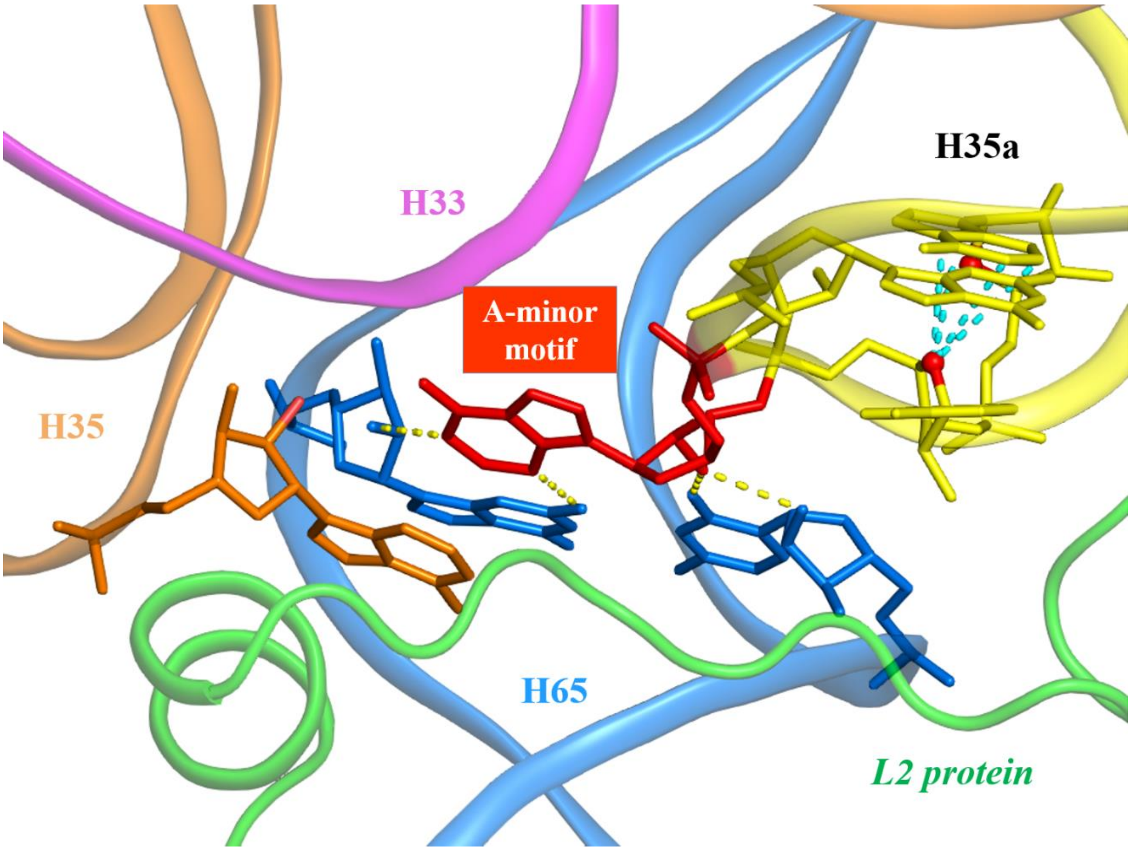


Figure 4.3. 3D representation of a GAAA (UNCG-fold) receptor in *E. coli* LSU. Helix 35a (yellow) is capped by a GAAA loop in which the 2nd residue (adenine, red) is involved in a class I A-minor motif with a C=G Watson-Crick pair on helix 65 (light blues). A stacking interactions between the loop adenine and an adenine from helix 35 (orange) is also observed. Ribosomal L2 protein (green) is found in the vicinity of the receptor. Interatomic contacts with distances <3.5 Å are shown with dashed lines. Color codes for strands are consistent with **Fig. 4.1**.

Table 4.1 Data of UNCG receptors among different species ribosomes. Data on ribosomal structure are sorted by species and present information extracted from the indicated PDB file. The observed sequences for tetraloops interacting with receptor I (UNCG capping helix 62) and Receptor II (GNNN capping helix 35a) are reported, together with number of residue/chain of the loop residues and of the receptor.

PDB code	Res (Å)	h62 UNCG seq.	Resi./chain	Receptor I	h35a GNNN seq.	Resi./chain	Receptor II
H. marismortui (50S)							
4V9F	2.4	UUCG	1770-73/0	A(1885/0)	GAAA	873-876/0	C(1844/0)=G(1832/0)
E. coli (50S)							
4YBB	2.1	UUCG	1692-95/DA	A(1829/DA)	GAAA	780-783/DA	C(1788/DA)=G(1776/DA)
T. termophilus (50S)							
4Y4O	2.3	UUCG	1692-95/2A	A(1829/2A)	GAAA	780-783/2A	C(1788/2A)=G(1776/2A)
S. cerevisiae (60S)							
4U4R	2.8	UUCG	1924-27/1	A(2188/1)	GAAA	912-915/1	C(2146/1)=G(2134/1)
H sapiens (60S)							
4UG0	3.6	UUCG	2873-76/L5	A(3692/L5)	GAAA	1629-1632/L5	C(3650/L5)=G(3638/L5)
O. cuniculus (60S)							
3JAH	3.4	UUCG	2873-76/5	A(3692/5)	GAAA	1629-1632/5	C(3650/5)=G(3638/5)
S. scrofa (60S)							
3J7P	3.5	UUCG	2873-76/5	A(3692/5)	GAAA	1629-1632/5	C(3650/5)=G(3638/5)
P. falciparum (60S)							
3J79	3.2	UUCG	2192-95/A	A(2481/A)	GAAA	1031-1034/A	C(2439/A)=G(2427/A)
T. aestivum (60S)							
4V7E	5.5	UUCG	1920-23/Aa	A(2183/Aa)	GAAA	915-918/Aa	C(2140/Aa)=G(2128/Aa)
K. lactis (60S)							
4V91	3.7	UUCG	1924-27/1	A(2188/1)	GAAA	912-915/1	C(2146/1)=G(2134/1)
D. melanogaster (60S)							
4V6W	6	UUCG	2240-43/A5	A(2566/A5)	GAAA	1112-1115/A5	C(2513/A5)=G(2525/A5)
T. brucei (60S)							
4V8M	5.6	UUCG	109-12/BB	A(456/BB)	GAAA	1011-1014/BA	C(414/BB)=G(402/BB)
Mitochondrial S. cerevisiae							
3J6B	3.2	No h62	-	A(1736/A)	GAAA	671-674/A	C(1695/A)=G(1683/A)
Mitochondrial S. scrofa							
5AJ4	3.8	No h62	-	A(969/BA)	GAAA	326-329/BA	C(838/BA)=G(826/BA)
Mitochondrial H. sapiens							
3J7Y	3.4	No h62	-	-	GACA	1990-1994/A	C(2508/A)=G(2496/A)

The receptor topology is completed by other accessory interactions. An adenine from helix 35 is stacked with the loop adenine, while various hydrogen bond contacts are observed between atoms on the helix 35a minor groove and helix 65. Analogously with the previous receptor, L2 protein is found in the vicinity. A similar folding scheme can be proposed, with helix 65 “docking” helix 35a and the L2 protein coming last to lock the final RNA architecture.

4.3.3 What are the features of a UNCG receptor?

It is remarkable to notice that in both cases the UNCG receptors are constituted by a complex ensemble of strands, coming from different ribosomal regions. It appears that this architectural complexity proper of UNCG receptors can be found only in highly structured RNA, such as ribosomes. At the same time, the specificity of these receptors involve a larger amount of elements than for a simpler GNRA receptor, making the description of UNCG receptor less general and bound to local topology characteristics. Anyhow, the concept that UNCG tetraloop as isolated motif not involved in specific RNA-RNA interactions should be rectified to include the possibilities presented here. With forthcoming structural data on complex RNA systems it would be hopefully possible to systematically define the nature of UNCG receptors.

4.3.4 More on complex tetraloops assemblies inside the ribosome

During the research of tetraloop motifs interacting in the ribosomes, a remarkable example of loop assembly emerged. It is found in the domain I of prokaryotic and eukaryotic LSU and is formed by three U-turns and an unstructured U-turn, forming two facing pairs with bottom residues stacked (**Fig. 4.4**).

The interaction between pairs of loops is reminiscent of the D- and T-loop interdigitation found in tRNA structure, which forms the characteristic elbow. The elbow allows the tRNA to be docked by recognizing another exposed base pair found on the docking platform of various RNAs. This *universal RNA structural motif* (Lehmann et al. 2013) is involved in the docking of tRNA by T-box riboswitch

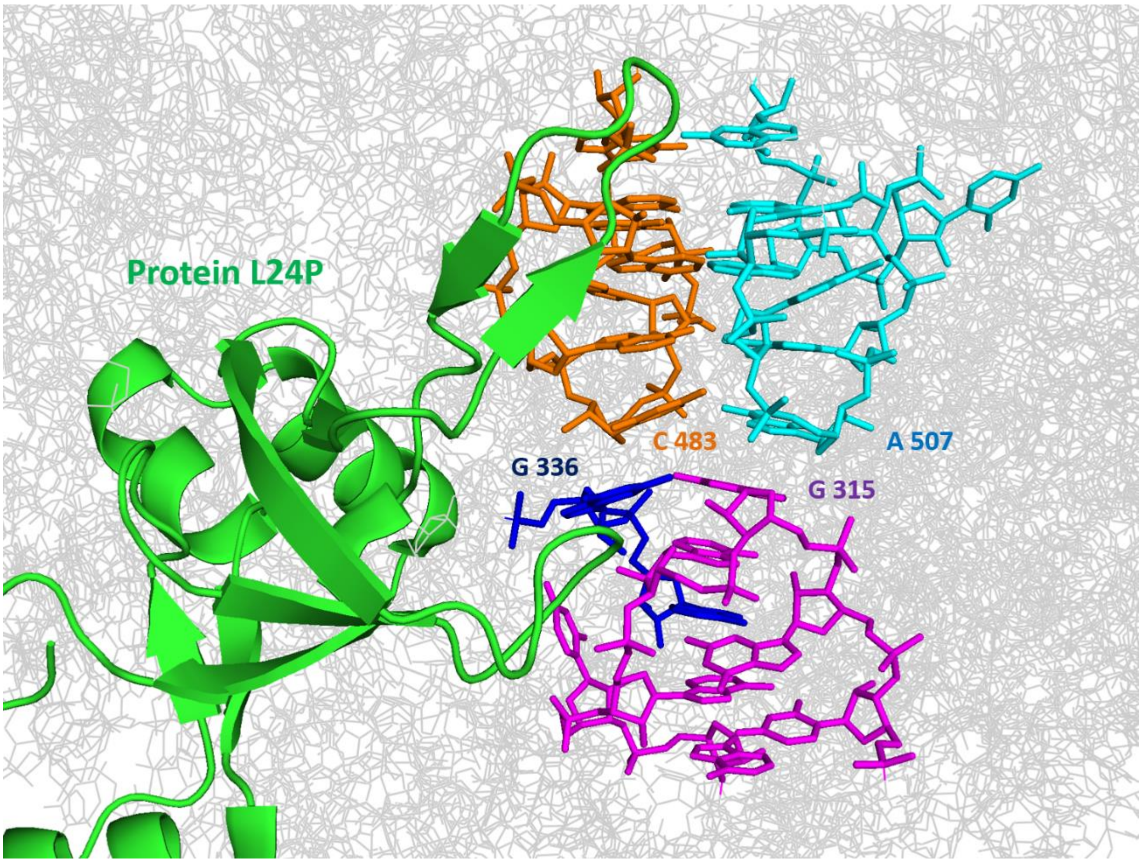


Figure 4.4. An assembly of four tetraloops inside the *H. marismortui* ribosome. The highlighted tetraloops belong to helix 19 (magenta), helix 18 (blue), helix 23 (orange and cyan). Ribosomal L24 protein is showed with green cartoons. The information on residue numbers are from *H. marismortui* LSU (PDB: 4V9F; res.:2.4 Å; (Gabdulkhakov et al. 2013).

(Zhang and Ferre-D'Amare 2013), RNAseP (Reiter et al. 2010) and the L1 Stalk of the ribosomal LSU (Tishchenko et al. 2012). This occurrence of four loops in contact with one another inside the ribosome therefore constitutes a fourth example of this general RNA-RNA recognition based on intermolecular base–base stacking. In this particular example, the close proximity of the L24 protein could suggest a direct role of this protein in the formation and stabilization of this unique arrangement during ribosome biogenesis.

5. Section III.

Ions interacting with nucleic acids and other environmental effects

5.1 Essential considerations on environmental effects on RNA

The intrinsic plasticity of biomolecules, required for their structure and functions, makes them particularly sensible to modulation by environmental factors. Knowing how solvent atoms and physical conditions influence nucleic acids is on the same importance level of knowing biomolecular structure, in the perspective of understanding their biological roles. I worked on the interaction patterns between solvent molecules and biomolecules, with a focus on charged species interacting with nucleic acids. The polyanionic nature of nucleic acids makes them very good binder of cationic molecules, but locally also anionic species can enter the first hydration shells. Among anionic species, also Asp and Glu side chains have to be considered; surprisingly, they show in various contexts to assume an unexpected neutral state, both in pairing with nucleobases or in forming carboxyl(ate) pairs. These rare contacts are significant for protein structures and involve characteristic very short hydrogen bond distances (~2.6 Å).

Analysing X-ray structural data from PDB and CSD databases, it is possible to characterize the binding of diverse solvent molecules with nucleic acid. However, one of the major issues of X-ray crystallography is the identification and assignation of solvent density to the correct species, due to the general high mobility of solvent paired with the relative small number of electrons of ionic species (especially metal ions such as Mg^{2+} or Na^{+}). In fact, during our database surveys, a large number of solvent misattributions or local issues appeared, especially for Mg^{2+} ions putatively binding purine imine nitrogen atoms. In order to assess these problems and to augment the reliability of structural data, we identified the main issues and proposed solutions to avoid such pitfalls in future endeavours.

Further, MD simulations were performed to study the dynamics of RNA first solvent shell under different temperature conditions. The goal has been to study RNA structure temperature effect on motifs such as tetraloops, by analysing the properties of solvent molecules that are part of nucleic acid structure. I will detail some preliminary results on this still ongoing project.

5.2 Metal ions interacting with nucleic acids

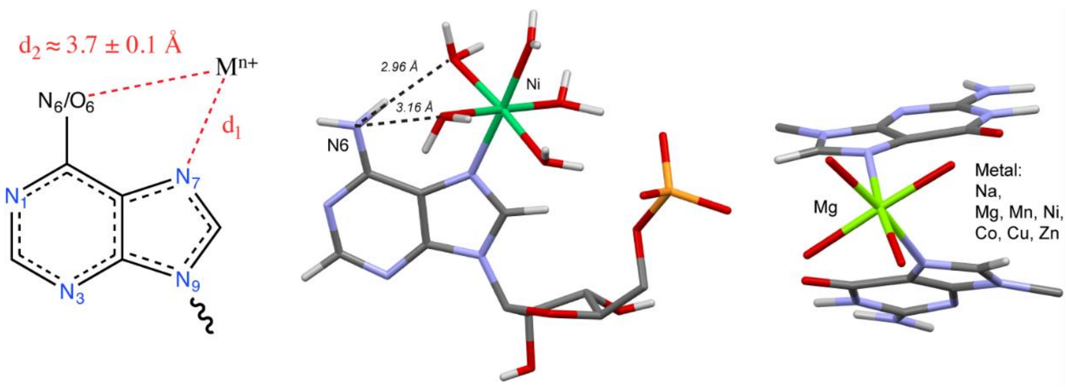
The most relevant metal ion interacting with nucleic acids, and specifically RNA, is Mg^{2+}. In addition, monovalent cations such as Na^{+} and K^{+} also have been showed to possess remarkable structural and functional roles for proteins and nucleic acids (Page and Di Cera 2006; Lambert et al. 2009; Pechlaner and Sigel 2012). I will detail in the first place the result of a review on Na^{+} and K^{+} binding to nucleic acids and then present our results on Mg^{2+} binding to nucleic acids. The subject about Mg^{2+} is so broad that its binding modes to different atoms (imine nitrogens, phosphate oxygens, carbonyl oxygens…) have to be analyzed in different studies, converging to a final comprehensive set of data for Mg^{2+} binding to nucleotides.

5.2.1 Review 1. *Sodium and potassium interactions with nucleic acids* (Met. Ions Life Sci., 2016)

In comparison to Mg^{2+}, it is less appreciated that monovalent cations (specifically Na^+ and K^+) are important for nucleic acid structure and functions. In particular, results on the involvement of monovalent cations in folding and catalytic activity are still emerging. We review the current state of detection techniques for K^+ and Na^+ ions in nucleic acid structures ranging from X-ray crystallography to nuclear magnetic resonance and MD simulations. Moreover, we raise awareness on the common pitfalls encountered while dealing with monovalent cations with all these methods. Subsequently, we present an analysis on specific and non-specific cation binding to nucleic acids, discussed through various relevant examples. Together with phosphate contacts, monovalent cations are often found to bind nucleic acid grooves forming specific coordination patterns with nucleobases. A well-known example of monovalent binding to nucleobases are quadruplexes, which structure changes if a smaller Na^+ or a larger K^+ is coordinated by guanine quartets. The same differences are exploited by a protein K^+ channel, suggesting the existence of a universal recognition motif for dehydrated K^+ ions. The same recognition could be also involved in RNA cation-induced conformational switches observed in fluorescent aptamers such as spinach-based sensors. Additionally, Na^+ and K^+ bind complex RNA folds and are associated with functional modulation in introns, riboswitches, ribozymes and ribosomes, which activity is altered or even inhibited by Na^+ ions. Overall, we stress that the nature of buffers used in biophysical or biochemical experiments too often contains Na^+ instead of the more biologically relevant K^+, yielding results that are sometimes misleading when compared to more physiological conditions.

5.2.2 Paper 4. *Binding of metals to purine N7 nitrogen atoms and implications for nucleic acids: a CSD survey* (Inorg. Chim. Acta, 2016)

Graphical abstract



Purine N7 atoms are considered to be the best nucleobase metal binding sites, but the properties of this binding are still not well formalized. We describe the occurrences and coordination geometry of alkali, alkaline earth and biologically relevant transition metals to purine N7 atoms, with an exhaustive survey of the Cambridge Structural Database. Three metal binding geometries to purine N7 were identified: *(i)* a tetracoordinated metal interacting with both N7 and N6/O6, specific to Cu or Zn; *(ii)* a hexacoordinated metal with direct N7 interaction and water mediated with N6, specific to Ni or Co and with an adaptable water coordination sphere; *(iii)* a hexacoordinated metal interacting with two purine N7 atoms. The occurrences of Mg^{2+} binding with N7 atoms are rare, being limited to 2 examples, inferring a weak affinity for this site. This suggests a low relevance for Mg^{2+} purine N7 binding sites in most RNA and DNA contexts. Consequently, the search was extended to small molecules imine sites and water molecules, in order to characterize the binding properties of metal ions and extract data to help with the current issues in the assignment of ions in large biomolecular systems. In particular, imidazolates were found to be valuable mimic of 6-oxopurines, but with a shorter distance between the imino nitrogen and the carbonyl oxygen that would allow metal ions to simultaneously bind them. This allowed us to use imidazolates as "affinity balance", based on the coordination distance difference between metal-N_{imino} and metal-$O_{carboxyl}$ coordination. Again, Mg^{2+} have been found a clear preference for oxygen, supporting the hypotheses that their binding to N7 purine atoms in biomolecules should not be considered of primary importance.

Inorganica Chimica Acta 452 (2016) 82–89

Contents lists available at ScienceDirect

Inorganica Chimica Acta

journal homepage: www.elsevier.com/locate/ica

Binding of metals to purine N7 nitrogen atoms and implications for nucleic acids: A CSD survey

Filip Leonarski [a,b], Luigi D'Ascenzo [a], Pascal Auffinger [a,*]

[a] Architecture et Réactivité de l'ARN, Université de Strasbourg, Institut de Biologie Moléculaire et Cellulaire du CNRS, 67084 Strasbourg, France
[b] Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

ARTICLE INFO

Article history:
Received 21 February 2016
Received in revised form 31 March 2016
Accepted 1 April 2016
Available online 9 April 2016

Keywords:
Imine nitrogen
Imidazolate
Chemical balance
Metal ions
Metal substitution

ABSTRACT

Understanding the structure and function of RNA and DNA systems depends partly on our comprehension of the binding features of metal ions to nucleobases. Such knowledge is important for an unambiguous assignment of ionic species to solvent electronic densities in crystallographic structures. Since the purine N7 atom is considered to be the best nucleobase metal binding site, we focus herein on describing the occurrence and coordination geometries of direct binding of alkali, alkaline earth and biologically relevant transition metals to this site. Further, we compare binding of such metals to purine N7 atoms, as well as imine sites occurring in small molecules such as imidazolates and water molecules. We analyze also the structure of the coordination shell of penta- and tetrahydrated metal ions bound to one or two purine N7 atoms. These structures can be used to validate proposed Mg^{2+} and other metal binding sites in large PDB structures where such assignments are often difficult to make. This survey suggests that Mg^{2+} ions bind with weak affinities to nucleobase N7 atoms. Although Mg^{2+} ions are essential to nucleic acid systems, purine N7 binding sites are, in most contexts, probably not of primary importance in RNA and DNA.

1. Introduction

The binding of metal ions to nucleic acids and proteins, despite the large number of studies and books devoted to the subject, is still a very active and important field of research [1,2]. It is well appreciated that nucleotides and amino acids interact with all biologically relevant alkali, alkaline earth and transition metals participating in the metallome [3,4] through nucleobases, sugar-phosphate backbones, amino acid side chains and peptide backbones. Their binding affinities depend on the type of the involved metals and binding site atoms that are sometimes categorized as hard and soft [5,6]. It has been proposed to separate metal ions into oxygen seeking, sulfur/nitrogen seeking and borderline or intermediate classes [7]. As such, it is assumed that hard metals such as alkaline and alkaline earth ions (including Mg^{2+}) are likely to associate with the anionic oxygen atoms from phosphate (nucleic acids) or carboxylate groups (proteins) while softer metals (Mn^{2+}, Ni^{2+}, Cu^{2+}, Zn^{2+} or Cd^{2+} as well as Ag^{+} and Tl^{+}) prefer to interact with histidine/nucleobase imine nitrogen atoms [8]. Accordingly, it has been suggested to group Na^{+}, Mg^{2+}, K^{+} and Ca^{2+} as an oxygen class, Mn^{2+}, Fe^{2+} and Co^{2+} as an imidazole class and Cu^{2+}, Ni^{2+} and Zn^{2+} as a sulfur class [9]. Although useful, such classifications have to be considered with caution. Indeed, the hard Mg^{2+} ion is found in chlorophyll where it interacts in a pentacoordinated manner with exactly five nitrogen atoms, four belonging to the chlorin group and one to an additional histidine ring, instead of its ordinary oxygen atom ligands. However, such an "*out-law*" complex requires assistance of chelatase enzymes for its formation [3].

In the general case, the stabilities of complexes formed by divalent metal ions in biologically relevant conditions has been predicted to follow the order [3,10]:

$$Mg^{2+} < Mn^{2+} < Fe^{2+} < Co^{2+} < Ni^{2+} < Cu^{2+} > Zn^{2+}$$

That is similar to the covalent contribution of metals [8]:

$$K^{+} < Na^{+} < Ca^{2+} < Mg^{2+} < Mn^{2+} < Fe^{2+} < Co^{2+} \sim Ni^{2+} \sim Cu^{2+} > Zn^{2+}$$

In this respect, it can be recalled that the concentrations of unbound metal ions in the cytosol range from millimolar (Na^{+}, K^{+}, Mg^{2+}) to micro- (Mn^{2+}, Fe^{2+}, Ca^{2+}), nano- (Co^{2+}, Ni^{2+}), femto- (Zn^{2+}) and attomolar (Cu^{+}/Cu^{2+}) [11]. While proteins are exposed to almost all kinds of biogenic metal ions including those considered as toxic in higher organisms (such as Cd^{2+}, [12]), nucleic

* Corresponding author. Tel.: +33 388 41 70 49; fax: +33 388 60 22 18.
E-mail address: p.auffinger@ibmc-cnrs.unistra.fr (P. Auffinger).

http://dx.doi.org/10.1016/j.ica.2016.04.005
0020-1693/© 2016 Elsevier B.V. All rights reserved.

acids are *in vivo* almost exclusively surrounded by K^+ and Mg^{2+} and possibly by Na^+ and Ca^{2+} ions [13,14]. When other ions are found in the vicinity of nucleic acids, they are usually chelated like Zn^{2+} in zinc finger motifs.

For nucleobases, it has become common knowledge that the best direct metal binding site is the purine N7 nitrogen [15,16] and that direct binding to imine N1/N3 nitrogen atoms is much less frequent and occurs only under specific conditions. The stabilities of single nucleoside/metal complexes, determined in solution by the affinity of N7 atoms, are the weakest for Mg^{2+} and Ca^{2+} and highest for ions such as Mn^{2+}, Zn^{2+}, Cd^{2+} and Cu^{2+} in that order. Further, adenine has weaker affinities for these metals than guanine and even negative affinities for Ca^{2+} and Mg^{2+}. The same order of affinities is reported for nucleotide macrochelate formation and all other possible mono-, di- and triphosphate combinations as well as for some dinucleotides [1]. It has also been reported that the binding affinities of Mg^{2+} and Mn^{2+} to thiophosphates is not very different and much weaker than that of Zn^{2+} and Cd^{2+}.

In crystallographic structures deposited in the PDB, numerous examples of direct binding of transition metals to N7 atoms have been reported [13,14] next to a very large number of N7-Mg^{2+} binding events. The latter are supposed to play a role, for instance, in the catalytic mechanism of the hammerhead ribozyme [17,18]. Yet, these N7-Mg^{2+} binding events are rather surprising given the above mentioned preference of these ions for anionic oxygen atoms belonging to phosphate groups [7,8]. To get a better view on the solvent structure of these large systems in experimental studies, various metal ions are used as substitutes in the identification of biologically relevant binding sites. For example, Tl^+, Rb^+ and Cs^+ have been used as heavy atom replacements for detecting Na^+ and K^+ binding sites, while Mn^{2+}, Zn^{2+} or Cd^{2+} are used as substitutes for Mg^{2+} ions [19]. Some other metals, like Cd^{2+}, are also used as probes to study the effect of metal ions on nucleic acids [20–23].

Hence, to clarify issues related to the structural characterization and the role of metal ions in nucleic acid structures from the PDB, it is important to gather reliable data and statistics on the preferred coordination modes of these metals to nucleobases and especially the purine N7 sites. For that purpose, we surveyed the Cambridge Structural Database (CSD) for metal binding to similar sites. We concentrated on the binding of hexacoordinated metal ions that are probably the most biologically relevant. Present data complement those already reported for protein systems [8,9,24–26] and represent an addition to existing web services providing access to structural databases for metal binding sites [27,28].

2. Material and methods

The Cambridge Structural Database (CSD Version 5.37, February 2016) [29–31] was searched to characterize metal to nitrogen atom coordination distances. We considered purine nucleobase-like fragments and an imidazole ring fragment as found in both purines or histidine amino acid side chains (Fig. 1). Note that for imidazole and more specifically histidine rings, the two nitrogen atoms are often reported as equivalent [8]. We considered also an imine fragment common to the above-mentioned motifs where the nitrogen is strictly bound to two carbon atoms. Besides, we extracted metal–water coordination distances from the CSD. We integrated in our search the following transition metals from the first and second row (Mn, Fe, Co, Ni, Cu, Zn, Cd) as well as alkali and alkaline earth metals. We further included Tl that is sometimes used as a K^+ ion substitute [13,14]. However, we did not consider beryllium (Be) since it is not present in its ionic forms in the PDB. A metal to nitrogen coordination distance was selected based on the existence of a coordination bond defined by the CSD. To eliminate non-specific coordination, we excluded metals that are located at more than ±1.0 Å from the plane defined by the nitrogen-containing fragment. All searches were performed with the *ConQuest* [32] software using filters so that disordered and error-containing structures were excluded. The searches were restricted to structures with *R*-factor values $\leqslant 0.05$ unless otherwise stated. The Mercury program was used for analyses and visualization of all these structures [33].

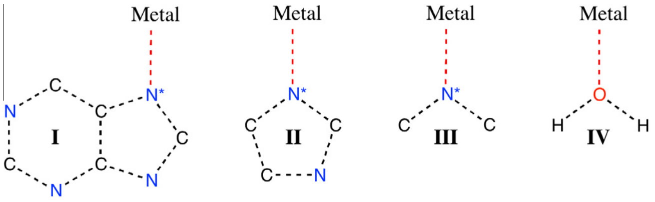


Fig. 1. CSD search fragments used for characterizing metal binding to various purine (**I**), imidazole ring (**II**), imine like fragments (**III**) and water (**IV**). The black dashed lines indicate that any bond type, as defined by the CSD, can be considered. The red dashed lines indicate that we searched for a direct coordination between the metal and the N/O atoms as defined in the CSD structures. The "*" next to a nitrogen indicates that only sp^2 atoms are taken into account. All fragments are planar. Water hydrogen atoms where explicitly included in fragment **IV**. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Unfortunately structures deposited to the CSD, even those of very high resolution, are not free of errors that persist despite significant structure validation efforts [8,24,34]. Such structures are difficult to eliminate from a search ensemble. Some of those, associated with unreasonably short or long coordination distances were eliminated after visual inspection that unavoidably includes a certain level of arbitrariness. On the other hand, differences in coordination lengths might be attributable to specific solid-state interactions involving particular ligands. It can be noted that some of the metal to nitrogen coordination distance histograms show more than one peak that are most probably associated with different metal oxidation states or Jahn–Teller effects as in the case of Cu^{2+}, low-spin Co^{2+}, Ni^{3+}, high-spin Cr^{2+} and Mn^{3+} [35]. Moreover, CSD oxidation states of transition metals are sometimes ill-defined presumably due to typographical or other mistakes [36,37]. Thus, associating an oxidation state with a given metal is generally not straightforward. When possible, we present for each element arguments that could lead to such an assignment. As always, a critical eye is required even when working with high-resolution crystallographic data.

3. Results and discussion

3.1. Statistical and structural overview of metal binding to imine nitrogen atoms and water

As expected, among all metals the coordination distances with N/O atoms vary the most for the alkali and alkaline earth ions indicating that these distances can be used as a part of the ion identification process (Table 1). The differences appear less significant for the investigated transition metals. Cd is the largest metal ion with distances around ≈2.3 Å followed by Mn with distances around ≈2.2 Å and Mg with coordination distances below ≈2.1 Å. As such, the Mn coordination distances exceed systematically those of Mg by ≈0.1 Å.

There is no significant difference in coordination distances of investigated metals to N/O atoms associated with fragments **III** and **IV**, respectively. Those differences lie within statistical uncertainties. Thus, given the precision of the collected data associated

Table 1
Average metal to nitrogen coordination distances derived from the CSD (version 5.37, update February 2016), obtained by analyzing biologically relevant fragments (Fig. 1). The number of hits is given in brackets. Standard deviations are provided when the number of hits is above ten. For some elements, more than one peak could be identified in distance histograms and average values with standard deviations are given for each of them. The searches were restricted to structures with *R*-factor values ⩽0.05 unless otherwise stated. Disordered, error containing, polymeric and powder structures were excluded from the search.

Metals[a]	Fragment **I** Purine: N7[b]	Fragment **II** Imidazole N[c]	Fragment **III** All imine N[d]	Fragment **IV** Water[e]
Alkali metals and thallium[f]				
Li-Lithium (Li^+)	2.10 [3][g]	2.09 ± 0.03 [20]	2.05 ± 0.07 [682]	1.96 ± 0.06 [562]
Na-Sodium (Na^+)	2.60 [3][g]	2.44 [3]	2.45 ± 0.05 [135]	2.41 ± 0.08 [3342]
K-Potassium (K^+)	–[g]	2.82 [68][g]	2.85 ± 0.06 [81]	2.86 ± 0.13 [2222]
Rb-Rubidium (Rb^+)	–[g]	2.94 [1][g]	3.04 ± 0.12 [10]	3.04 ± 0.17 [107]
Cs-Cesium (Cs^+)	–[g]	–[g]	3.11 [2]	3.24 ± 0.14 [326]
Tl-Thallium (I,III)	–[g]	2.49 [8]	2.42 ± 0.23 [221]	2.88 ± 0.38 [31]
(Thallium peaks)	–[g]	(2.32, 2.78)	(2.27 ± 0.08, 2.68 ± 0.14)	
Alkali earth metals				
Mg-Magnesium (Mg^{2+})	2.23 [2]	2.19 ± 0.02 [37]	2.10 ± 0.07 [825]	2.06 ± 0.03 [1362]
Ca-Calcium (Ca^{2+})	–[g]	2.52 ± 0.07 [29]	2.48 ± 0.11 [328]	2.41 ± 0.06 [855]
Sr-Strontium (Sr^{2+})	–[g]	2.66 [9]	2.65 ± 0.11 [141]	2.61 ± 0.06 [293]
Ba-Barium (Ba^{2+})	2.85 [1][g]	2.9 [5]	2.89 ± 0.09 [134]	2.83 ± 0.09 [553]
Transition metals				
Mn-Manganese (I–V)	2.32 [2][g]	2.23 ± 0.04 [345]	2.19 ± 0.11 [6294]	2.19 ± 0.05 [2524]
(Manganese peaks)			(2.03 ± 0.05, 2.27 ± 0.05)	
Fe-Iron (I–V)	2.16 [3]	2.09 ± 0.09 [749]	2.07 ± 0.11 [9337]	2.10 ± 0.05 [995]
(Iron peaks)		(1.98 ± 0.03, 2.15 ± 0.05)	(1.97 ± 0.04, 2.15 ± 0.07)	
Co-Cobalt (I–IV)	2.05 ± 0.09 [11]	2.05 ± 0.08 [1317]	2.05 ± 0.11 [9710]	2.10 ± 0.03 [3946]
(Cobalt peaks)	(1.96, 2.10)	(1.93 ± 0.02, 2.13 ± 0.04)	(1.94 ± 0.04, 2.13 ± 0.05)	
Ni-Nickel (I–IV)	2.10 ± 0.05 [10]	2.07 ± 0.07 [1122]	2.01 ± 0.10 [12198]	2.08 ± 0.04 [4199]
(Nickel peaks)		(1.91 ± 0.02, 2.07 ± 0.07)	(1.89 ± 0.04, 2.07 ± 0.05)	
Cu-Copper (I–III)	2.00 ± 0.03 [43]	1.99 ± 0.03 [2238]	2.00 ± 0.05 [26031]	2.24 ± 0.23 [5315]
(Copper peaks)				(1.97 ± 0.02, 2.37 ± 0.17)
Zn-Zinc (I, II)	2.05 ± 0.5 [27]	2.05 ± 0.06 [927]	2.08 ± 0.07 [11509]	2.09 ± 0.05 [2715]
Cd-Cadmium (Cd^{2+})	2.33 ± 0.05 [25]	2.29 ± 0.05 [780]	2.34 ± 0.06 [4666]	2.31 ± 0.04 [1458]

[a] When appropriate, oxidation states as mentioned in the CSD are given in parenthesis.
[b] Statistics for the imine purine N7 atoms of fragment **I**.
[c] Cumulated statistics for the two imidazole nitrogen atoms found in fragment **II**. These statistics include those related to fragment **I**.
[d] Cumulated statistics for the imine atom found in fragment **III**. These statistics include those related to fragments **I** and **II**.
[e] Statistics for fragment **IV**.
[f] Thallium in its Tl^+ form is often considered as a K^+ substitute and as such has been added to this table.
[g] No restrictions were applied to these searches.

with the large diversity of structural fragments considered here, it is difficult to infer simple rules regarding N/O coordination distances.

Non-biologically-relevant contexts or environments occurring frequently in the CSD may affect our statistics as well. Even very precise quantum mechanical calculations provide context dependent coordination distances. For instance, the Mg^{2+}…Ow coordination distance of a single water molecule has been calculated by quantum mechanical techniques to be in the 1.92–1.96 Å range while for $Mg[H_2O]_6^{2+}$ the same distance lie in the 2.08–2.10 Å range and therefore closer to crystallographic derived values [38].

Some metals display more than one optimal coordination distance. Thallium, which is considered as a good mimic for K^+ in crystallographic investigations [39], displays a ≈2.7 Å coordination distance to water and a ≈2.5 Å coordination distance to imine nitrogen atoms. Yet, Tl binds poorly to oxygen atoms, therefore the statistics for Tl–O are not very reliable. Two coordination peaks appear in the Tl–N histograms that could have as origin a different thallium oxidation state (≈2.3 Å for Tl^{3+} and ≈2.7 Å for Tl^+). For Mn, two peaks at ≈2.0 and ≈2.2 Å are distinguishable in the imine nitrogen histograms. The short coordination distance could be related to Mn atoms in a rare +3 oxidation state and associated with a Jahn–Teller effect [35]. The +2 oxidation state is certainly related to the more common 2.2–2.3 Å coordination distances. Note that the Mn^{2+}–water coordination distance (≈2.19 Å) is larger that the Mg^{2+}–water coordination distance (≈2.06 Å). For Fe, Co and Ni, the two peaks related to the imine containing fragments are separated by 0.1–0.2 Å. We assume that the shortest and longest coordination distances can be attributed to the +3 and +2 oxidation states, respectively. For Zn, two peaks associated with a large spread are also observed. All transition metal to water coordination distance histograms are single peaked except the one related to Cu that is the result of a well-documented Jahn–Teller effect [24]. This effect is not observed when nitrogen ligands are involved.

3.2. Low Mg^{2+} binding occurrence to N7 atoms correlates with nitrogen metal affinities

The occurrence of Mg^{2+} binding events to imine nitrogen atoms is relatively low in the CSD, especially for nucleobases. Only one instance of a Mg^{2+} ion bound to the N7 atoms of two in-plane theophylline like purines [40] and another showing a Mg^{2+} binding to two stacked guanine N7 atoms (high *R*-factor) have been reported [41]. This observation correlates with the low binding level of Mg^{2+} ions to histidines in proteins [8] and has to be compared with the higher occurrence of other transition metals next to N7 atoms (Table 1). Only one occurrence of Mn^{2+} binding to N7 (high *R*-factor) has been reported [42] along with two pentahydrated Mn^{2+} ion containing structures for which no coordinates were deposited [43]. These examples will be described below. It has to be noted that the apparent low occurrence of complexes with Mg^{2+} and Mn^{2+} should be viewed with caution since it may perhaps only reflect the fact that these compounds crystallize less easily.

Regarding other transition metals, their higher affinity for nitrogen ligands correlates with a much larger number of contacts to fragment **III** compared to fragment **IV** (water). The reverse is observed for alkali and alkali-earth metals including Mg^{2+} where the more frequent binder is oxygen. In that respect, it has been reported from quantum mechanical investigations that for Mg^{2+} the O6 inner shell binding mode is favored over the N7 binding mode [44]. It is worth mentioning that Tl^{+} shows an opposite trend and is, like transition metals, more frequently bound to nitrogen atoms (Table 1).

3.3. Binding characteristics of metals and water to purine N7 atoms

The CSD embeds a large diversity of chemical compounds where tetra- and pentacoordinated metals coordinate to various atoms such as sulfur, chloride, other metals,... Since such coordination patterns are rare in nucleic acids, we restrict our survey to the more biologically relevant hexacoordinated metals. We found that, when binding to purine N7 atoms, metal ions are mostly in plane with the base and at a 3.7 ± 0.1 Å distance from the purine N6/O6 atoms (Fig. 2). This distance along with the coordination distance to the N7 atom could be used to distinguish these metals from water molecules or NH_4^+ ions in lower resolution structures from the PDB.

Indeed, the hydrogen bond distance for water molecules to N7 atoms is, as expected, close to 2.8 ± 0.1 Å. For 6-aminopurines, a water molecule hydrogen bonded to both N6/N7 can be observed with N6...Ow distances of 3.1 ± 0.1 Å. Besides, other in plane water molecules are observed with N6...Ow distances in the 3.5–4.5 Å range. For 6-oxopurines, the average O6...Ow distance histogram has a first peak at ≈3.6 Å and a second close to 4.0 Å. No "in plane" water molecules at hydrogen bond distance of both N7/O6 atoms are observed.

Metal–N6/O6 distances around 3.0 Å were only observed for tetracoordinated transition metals such as Cu or Zn and might only be observed in structures of the PDB under specific crystallization conditions. Such coordination schemes seem unlikely for Mg^{2+} or Mn^{2+} ions for which direct coordination to both N7/O6 atoms has not been reported (see below).

Metals can also bind to the N7 atom of 6-aminopurines with similar distances as in 6-oxopurines. It is probable that this N7 purine site will more difficultly accommodate larger ions such as Ca^{2+}, Sr^{2+} or even K^{+}. Thus, this 6-aminopurine site is probably more selective for smaller ions and it has even been argued if small ions can bind to it. Gas phase quantum mechanical calculations (performed in the absence of water) suggest that Mg^{2+} might not bind to the N7 atom of adenine [45]. Calculations taking into account first-shell water molecules reach opposite conclusions [16,46].

3.4. Pentahydrated metal ions binding to purine N7 atoms

Eight structures of pentahydrated metals (Co, Ni and Cd) bound to the N7 atom of purine like fragments were deposited to the CSD as well as four other fragments (with Mn, Ni and Fe) for which no coordinates were archived (Table 2). Seven of these metals are bound to guanine or inosine fragments and one to adenine. The metal ion positions are consistent with the ion coordination distances noted in Table 1. In all instances, the metal–N6/O6 distance is 3.7 ± 0.1 Å (Fig. 2). Thus, from this limited set of examples, divalent transition metals appear to bind similarly to 6-aminopurines and 6-oxopurines. Hence, this metal–N6/O6 distance can be considered as a reliable criterion for characterizing metal binding to N7 atoms in the lower resolution structures of the PDB.

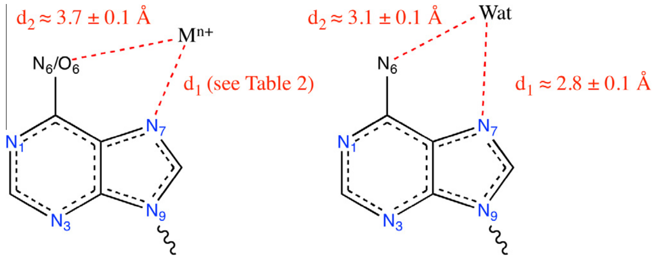


Fig. 2. Definition of two characteristic distances for metal binding to purine N7 atoms. (Left) The average d2 distance is derived from an ensemble of hexacoordinated metal ions binding to the N7 atom. (Right) d1 and d2 distances for water molecules hydrogen bonded to both N6/N7 atoms of 6-aminopurines.

Table 2
CSD structures of pentahydrated ions bound to a purine N7 atom (Fig. 3).

Purine	Ion	d1[a]	d2[b]	Hydrogens	*R*-factor [%]	CSD code
Pentahydrated metal						
G	Cd	2.37	3.83	No	6.0	AGOPCD
Inosine	Ni	2.11	3.74	Yes	7.5	ANIMPH01
G	Co	2.13	3.71	Yes	3.4	BIPVIF01
Inosine	Co	2.15	3.70	Yes	4.3	DIDSOY
Inosine	Co	2.12	3.68	Yes	2.8	FIZHUR
Inosine	Ni	2.06	3.64	Yes	3.2	FIZJAZ10
Inosine	Co	2.16	3.79	Yes	5.1	IMPCOH
A	Ni	2.07	3.74	Yes	2.4	ZZZAAF01
G[c]	Mn					FAMNIQ01
G[c]	Mn					QQQGLY
G[c]	Ni					GUOSNI
G[c]	Fe					FAMNEM

[a] Distance between the metal ion and the N7 atom (Fig. 2).
[b] Distance between the N7 and the N6/O6 atoms (Fig. 2).
[c] No coordinates were deposited to the CSD.

The placement of the five coordinated water molecules is similar in all structures suggesting a regular hydration pattern for ions bound to purine N7 atoms (Fig. 3). Overall, the five water molecules form along with the N7 atom an octahedral coordination scheme. The closest water molecule to the N6/O6 atoms is at a 2.8 ± 0.1 Å hydrogen bond distance. The second closest water molecule to the N6/O6 atoms is at a 3.4 ± 0.2 Å non-hydrogen bonding distance indicating an asymmetrical arrangement with respect to the purine plane. Both these distances are in agreement with N6–Ow distances close to 2.8 and 3.3 Å derived from quantum mechanical calculations [16] The water closest to the N6/O6 atoms is either a hydrogen bond acceptor or donor. Hence, in this case also, these two N6/O6-water distances can be considered as useful criteria for validating metal binding to N7 atoms in PDB structures.

3.5. Water orientation in the metal coordination sphere is adaptable

The positions of the hydrogen atoms belonging to coordinated water molecules are also in agreement with quantum mechanical calculations [16,38] and first principle molecular dynamics calculations of the hydration of Mg^{2+} ions in aqueous solution [47–49]. These calculations as well as high-resolution CSD data, suggest that water molecules tend to asymmetrically coordinate Mg^{2+} ions through one of the oxygen atom lone pairs, an outcome that could probably not have been derived from simple electrostatic gas phase considerations. In the pentahydrated metal bound purines (Table 2), the angle associated with the metal ion and the bisector of the coordinated water molecules is 146 ± 17°. These angles are 152 ± 14° (CSD code: YOHJAI) and 171 ± 2° (CSD code: CIRVAA01) in two neutron diffraction structures of hexahydrated Mg^{2+} ions. However, in DFT calculations of hexahydrated Mg^{2+} gas phase clusters, the water molecules coordinate rather symmetrically to Mg^{2+} [19,38,48]. These differences suggest a strong influence of the environment on the orientation of the water molecules. Such an unexpected adaptability of the water molecule orientation seems

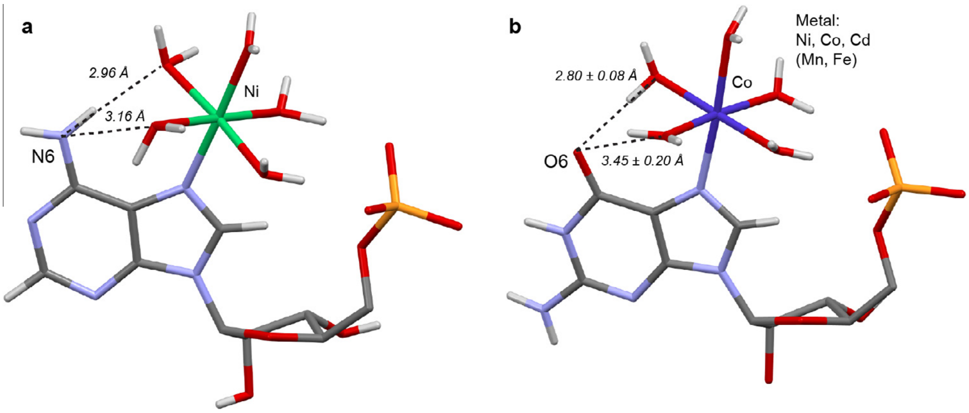


Fig. 3. CSD structures of pentahydrated metals binding to purine N7 atoms and N6/O6–Ow distances. For clarity, the sugar carbon bound hydrogens are not shown. (a) Structure derived from a complex showing a Ni ion bound to an adenine (CSD code: ZZZAAF01). (b) Structure derived from a complex showing a Co ion bound to a guanine (CSD code: BIPVIF01). The average distances are derived from the structures listed in Table 2.

required to accommodate metal binding to both 6-aminopurines and 6-oxopurines.

In the adenine/Ni complex, the purine amino group remains essentially planar despite the presence of the hydrated metal ion, in opposition to quantum mechanical calculations that advocate its strong pyramidalisation [16]. Yet, one has to exert caution regarding the hydrogen positions inferred from crystallographic structures that might be sometimes affected by refinement options and might not always be reliable [34]. In this respect, it can be noted that guanine amino groups that are distant from metal ions are sometimes non planar in CSD structures. Such hydrogen atom positions are certainly also very sensitive to their environment.

As a word of caution, it should be considered that in biomolecular systems metals could bear other ligands than water to complete their solvation shell such as for example Cl coordinated to Zn as observed in a Z-DNA structure [50]. This is also important for metals like Pt, Pd or Ag. Ligands like OH^- are also probable and were considered as bridging compounds in bimetallic complexes [51,52].

3.6. Hydrated metal ions coordinated to two purine N7 atoms

Besides the above-mentioned pentahydrated metal ions coordinating to a single N7 atom, we found only two binding patterns in the CSD that involve tetrahydrated metals and two purine N7 atoms. The first comprises a planar and the second a stacked arrangement of the two purines (Fig. 4). In one instance, an amino group is found in position 6 (Table 3). All these structures are similar and the purines are organized in a head-to-tail manner. In biomolecular systems, such arrangements would correspond to metal mediated base pairs. As for pentahydrated metals (Fig. 2), the orientation of the metal-bound water molecules differs. Metal coordinated water molecule close to position 6 are involved in a hydrogen bond with O6 but not N6 atoms. In the latter, the closest hydrogen atoms of water are pointing away from the amino group.

Table 3
CSD structures of tetrahydrated ions bound to two N7 atoms in a planar purine arrangement (Fig. 4a).

Purine	Ion	d1[a]	d3[b]	Hydrogens	*R*-factor [%]	CSD code
Tetrahydrated metal – planar arrangement						
A	Cu	2.0/2.0	4.0	Yes	3.1	AMADCU
G	Cu	2.0/2.1	4.1	Yes	4.8	BAHMAY
Theophylline	Mg	2.2/2.2	4.6	Yes	4.9	CUCZEH
Theophylline	Cd	2.3/2.3	4.6	Yes	4.4	DIFRAL
G	Ni	2.2/2.2	4.3	Yes	3.8	HOPBOD
2-Amino purine	Co	2.1/2.1	4.3	Yes	9.5	HOZBEF
Xanthine	Zn	2.2/2.2	4.4	No	5.6	JIXFEB
Xanthine	Ni	2.2/2.2	4.3	Yes	3.0	LIZMOW
G	Cd	2.3/2.3	4.6	No	3.5	NARREE
G	Cd	2.3/2.3	4.6	No	4.7	NARRII
Inosine	Cu	2.0/2.0	4.5	Yes	4.0	TAHYPC

[a] Distance between the metal ion and the N7 atoms (Fig. 2).
[b] Distance between the two N7 atoms.

The second pattern involves stacked head-to-tail purines that occur in the CSD structure of the cyclic diguanylic acid or cyclic d-GMP complex with Mg^{2+} [41], a molecule that is recognized as a second messenger used in signal transduction in a wide variety of bacteria [42]. In these patterns, the two purines are highly tilted (37 ± 5°; Table 4). They display large *R*-factor values (8.8 ± 1.9) that

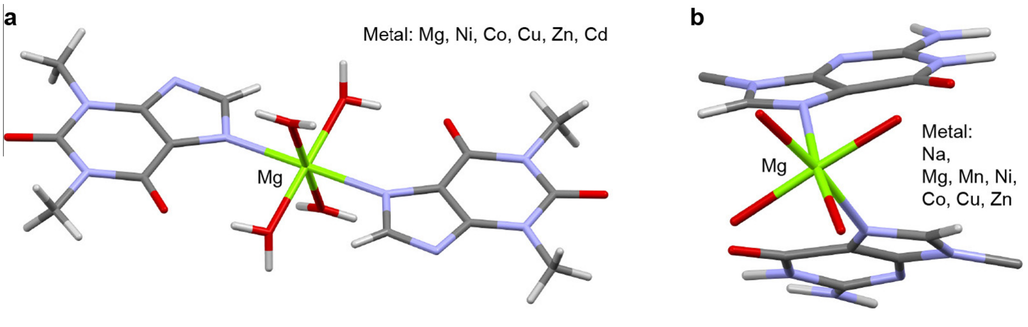


Fig. 4. CSD structures of tetrahydrated metals binding to two purine N7 atoms. (a) This structure is derived from a complex showing a Mg^{2+} ion bound to two theophyllines in a planar arrangement (CSD code: CUCZEH). (b) This structure is derived from a complex showing a Mg^{2+} ion bound to two guanines in a stacked arrangement (CSD code: SUKHUB).

Table 4
CSD structures of tetrahydrated ions bound to two N7 atoms in a stacked purine arrangement (Fig. 4b).

Purine	Ion	d1[a]	d3[b]	Angle[c]	Hydrogens	*R*-factor [%]	CSD code
Tetrahydrated metal – stacked arrangement							
Inosine	Co	2.2/2.2	2.9	41	No	9.0	BEXRAX10
G	Co	2.2/2.2	2.9	42	No	10.0	BEXREB10
G	Zn	2.2/2.2	3.0	36	No	6.7	DAZTED
G	Zn	2.1/2.1	3.0	33	No	6.1	DAZTIH
G	Cu	2.0/2.0	2.8	44	Yes	4.7	ESIWOT
Inosine	Cu	2.0/2.0	2.8	42	No	9.3	GANXOI
G	Ni	2.1/2.1	2.9	39	No	13.1	GAVDIQ
G	Na	2.4/2.6	3.3	36	Yes	7.9	GUOPNA12
G	Mn	2.3/2.4	3.0	30	No	9.6	QOCVIP
G	Co	2.2/2.3	2.9	35	No	11.2	SIWWIE10
G	Mg	2.3/2.3	2.9	33	No	9.4	SUKHUB

[a] Distance between the metal ion and the N7 atoms (Fig. 2).
[b] Distance between the two N7 atoms.
[c] Tilt angle between the two purine planes.

suggest the occurrence of structural stress in the crystals. As expected, the two metal–N7 distances are close. Besides, short N7–N7 distances (2.9 ± 0.1 Å) are observed (the latter are similar to the distance between nearby water molecules in the first metal hydration shell). These distances are also much shorter than the stacking distance between two nucleobases (≈3.4 Å). Such distances and angles should be regarded as characteristic for metal coordination to two N7 atoms belonging to stacked purines. Indeed, such tilts are rare in biomolecular structures and might be characteristic of metal binding if they are associated with a short N7–N7 distance. In these parallel and stacked arrangements as well as in the binding of pentahydrated ions to N7 atoms, all metals occupy the same binding spots and suggest their ability to replace each other in larger structures.

Interestingly, for the stacked arrangement (Fig. 4b), the N7 binding site can also be occupied by a Na^+ ion (CSD code: GUOPNA12) [53]. The tilt of the two purines (36°) is similar to those observed in other complexes (Table 4). All other features of this arrangement are similar too, with the exception of the dissymmetric Na^+–N7 distances (2.4 and 2.6 Å). The N7–N7 distance is also larger (3.3 Å) and closer to normal stacking distances. This represents an example where a monovalent ion can occupy a site that is generally attributed to divalent ions in biomolecular systems, a feature that should be kept in mind during the refinement and analysis of the solvent structure of the large nucleic acid systems deposited to the PDB.

3.7. Simultaneous metal binding to N6/O6 and N7 atoms – the case of (imid)azolates

The simultaneous binding of Mg^{2+} or a transition metal to purine O6/N7 atoms is sometimes considered in quantum mechanical calculations [54] and it has marginally been inferred that binding to O6 over N7 atoms is preferred [44,55,56]. In the CSD, such events are not observed for purines but only for the closely related imidazolate and azolate compounds (over 500 occurrences with 13 different metals including Na^+, Ca^{2+}, Sr^{2+} and Ba^{2+}). The main difference between purines and (imid)azolates relates to the presence of a carboxylate group in the latter and a slightly different binding site geometry that leads to a ≈0.4 Å change between the N/O coordinating atoms (Fig. 5). In (imid)azolates, the two N/O atoms are at the appropriate ≈2.7 Å distance for completing the coordination sphere of a transition metal ion such as Mg^{2+} whereas purines with a ≈3.1 Å distance and a different orientation of the coordinating groups are not. This 2.7 Å distance correlates with the coordination distance d(Ow…Ow) of water molecules in the first hydration shell of first row transition metals that is around ≈2.9 Å (Table 5). However, this (imid)azolate site can also accommodate larger ions such as Na^+, Ca^{2+} and Cd^{2+} for which the d(Ow…Ow) distance extends to ≈3.2 Å for Cd^{2+} and even ≈3.5 for Na^+. Thus, another plausible explanation is related to the fact that the carboxyl group has a higher affinity than the carbonyl group for these metals. Very likely, we observe here a combination of both effects.

Larger ions such as Ca^{2+} [57] and Ba^{2+} [58] were found to coordinate simultaneously to N7/O6 atoms in PDB structures. Therefore, it is also likely that ions such as Na^+ or K^+ (that has, like Ba^{2+}, a ≈2.8 Å coordination distance) could bind to the N7/O6 atoms of a guanine in large nucleic acid structures. Yet, it seems that O6 atoms are better binding sites for alkali ions as they are involved in maintaining guanine quartet structures occurring for instance in telomeres. Alkali ions were also reported to interact with thymine O2 atoms in a structure of d(ApT) minihelix [59]. Note that for 6-aminopurines, no tautomeric forms involving the deprotonation of the adenine and associated with a direct metal–N6 contact, as reported elsewhere for metals such as platinum [15,60], were observed.

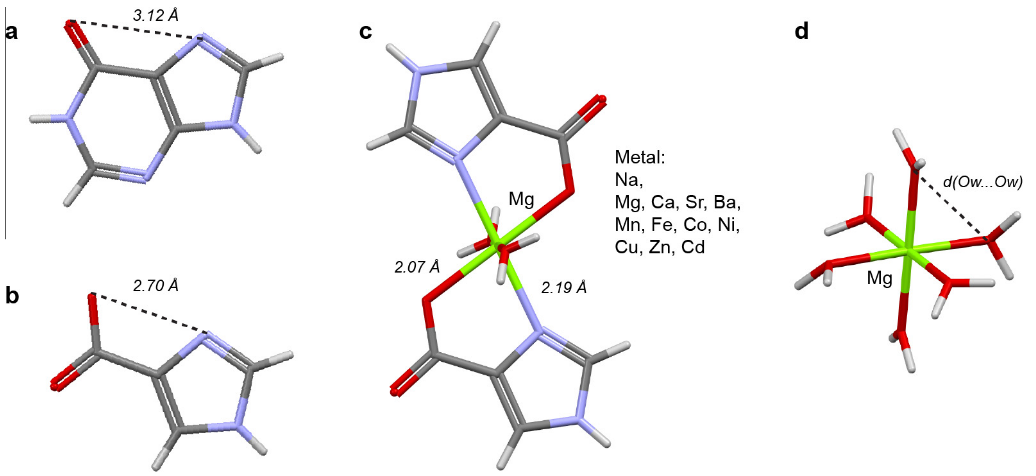


Fig. 5. Difference between 6-oxopurine and imidazolate like fragments. (a and b) Comparison between the imide nitrogen and carbonyl oxygen atom distances in inosine (CSD code: FIZHUR) and imidazolate (CSD code: DEZNIG). (c) Complex between an imidazolate and a Mg^{2+} ion (CSD code: DEZNIG). (d) Neutron diffraction structure of a hexahydrated Mg^{2+} complex (CSD code: YOHJAI).

Table 5
Distance d(Ow···Ow) between two close water molecules in the first hydration shell of hexahydrated metal ions derived from a CSD search (Fig. 5d). The number of hits is given in brackets. Standard deviations are provided when the number of hits is above ten. The searches were restricted to structures with *R*-factor values ⩽0.05. Disordered, error containing, polymeric and powder structures were excluded from the search.

Metals	d(Ow···Ow)
Na	3.47 [2]
Mg	2.92 ± 0.06 [132]
Ca	3.22 [5]
Mn	3.08 ± 0.08 [53]
Fe	2.99 ± 0.06 [29]
Co	2.95 ± 0.06 [138]
Ni	2.91 ± 0.06 [130]
Cu	2.87 [3]
Zn	2.95 ± 0.07 [61]
Cd	3.21 ± 0.10 [21]

3.8. Are (imid)azolates a reliable N/O metal affinity balance?

As described above, imidazolates and the related azolates are specific classes of molecular fragments where an imine nitrogen and an anionic carboxylate oxygen atom bind simultaneously to a metal ion in a close to perfect geometry. Thus, we thought that these compounds could reflect the difference in affinity of a metal for the N versus O atoms [61] and could provide information similar to those provided by a large family of "molecular balances" that involve, among others, the use of rotameric folding molecules to quantify non-covalent interactions [62]. We are calling this tentatively a "metal affinity balance". Hence, we compared the metal–N to metal–O coordination distances in these compounds (Table 6). Despite the shortage of data leading to poor statistics, the hard alkaline earth metals including Mg^{2+} seem to prefer binding to oxygen over nitrogen. On the other hand, the softer cations such as Cu, Zn or Cd are associated with shorter metal–N distances reflecting a higher affinity for nitrogen. This last result is somewhat surprising since the imine nitrogen atom is in competition with an "anionic" carboxylate oxygen atom. The apparent preference of Mg^{2+} for oxygen atoms is probably also at play in large nucleic acid systems suggesting that N7 atoms are at best secondary interaction sites populated only under specific conditions and that contact distances of Mg^{2+} to imine nitrogen atoms can be stretched from the optimal 2.1 Å coordination distance to a less frequent 2.2 Å or even larger coordination distance in specific contexts.

Table 6
Metal–N and metal–O coordination distances in imidazolate (Fig. 5) and azolate compounds. A negative and positive "delta" value suggest a higher affinity of the metal for oxygen and nitrogen, respectively. The number of hits is given in brackets. Standard deviations are provided when the number of hits is above ten. No restrictions were applied to this search.

Metals	d(metal...N)	d(metal...O)	Delta
Alkali earth metals			
Mg	2.20 [2]	2.09 [2]	–0.11
Ca	2.54 [2]	2.44 [2]	–0.10
Sr	2.74 ± 0.05 [15]	2.66 ± 0.06 [15]	–0.08
Ba	2.97 ± 0.13 [18]	2.86 ± 0.09 [18]	–0.11
Transition metals			
Mn	2.23 ± 0.08 [63]	2.21 ± 0.08 [63]	–0.02
Fe	2.16 ± 0.06 [22]	2.15 ± 0.07 [22]	–0.01
Co	2.03 ± 0.12 [190]	2.02 ± 0.11 [190]	–0.01
Ni	2.06 ± 0.05 [196]	2.10 ± 0.06 [196]	+0.04
Cu	1.98 ± 0.08 [87]	2.12 ± 0.21 [87]	+0.14
Zn	2.08 ± 0.06 [130]	2.18 ± 0.08 [130]	+0.10
Cd	2.27 ± 0.04 [263]	2.40 ± 0.07 [263]	+0.13

3.9. Metal ion substitutions in small and large structures

We described here several metal binding patterns associated with a large diversity of metals. These binding patterns suggest that transition metal ions, that produce easily recognizable electron densities, can be used as a probe for inferring the binding of ions such as Mg^{2+}, Na^{+} or K^{+} that display weak and/or non-characteristic electron densities. However, such an assertion should be taken with some caution especially regarding their relevance for *in vitro* as well as crystallographic studies. For instance, the crystal structures of the Mg^{2+} and Mn^{2+} cyclic d-GMP complexes exhibit both the same arrangement of stacked purines linked to the metal through their N7 atoms [42]. However, their spectroscopic properties were found to be very different in solution. While the Mg^{2+} ion did not produce any signal change with respect to the metal free conditions, the Mn^{2+} ion affected significantly the spectroscopic properties of this molecule. Similarly, very small spectroscopic effects of Mg^{2+} ions over Zn^{2+} or Cd^{2+} ions on the hammerhead ribozyme were reported [63]. On the crystallographic side, a systematic study conducted on the binding of metals to a RNA duplex, revealed strong differences in the association of eleven monoatomic ions and two hexamines to the structure. This study raised interrogations related to the use of metal substitutions in large structures [64]. Indeed, a conformational change induced by the presence of a Mn^{2+} ion was also reported for a signal recognition particle (SRP) where Mg^{2+} was substituted by Mn^{2+} [65]. In these structures, Mn^{2+} changed the conformation of a nucleotide by binding to a site that allowed to form inner sphere coordination with a N7 atom and a non bridging phosphate oxygen atom from a neighboring residue. Here also, the exact position of a metal is strongly dependent upon the nature of the metal and its environment.

4. Summary and perspectives

Present data extracted from the CSD should help warrant correct identification of metal binding sites in nucleic acids and other biopolymers, a process that is often complicated by the lower resolution of the structures deposited to the PDB. This is especially true for ions like Mg^{2+} that are isoelectronic with Na^{+}/NH_4^{+} ions and water molecules [66]. For instance, we showed that N7–ion distances could be used for preliminary metal identification but also for distinguishing metals from water molecules or hydrogen bonding ions such as NH_4^{+}. The N6/O6–ion distances around 3.7 Å and the metal bound water molecule orientation are likewise characteristic of metal binding to purine N7 atoms. Although simultaneous binding of transition metals and Mg^{2+} ions to N7/O6 atoms can be excluded, larger ions such as K^{+} or Ba^{2+} are able to coordinate to both atoms. However in 6-aminopurines, only the smaller divalent metals seem to be able to bind to N7 atoms.

Such data should also help to better conduct and understand biochemical substitution experiments whose results are sometimes difficult to interpret. Besides, this survey provides data allowing to improve parameterization of classical and polarizable molecular dynamics force fields, the accuracy of which being essential for providing a good understanding of structure–dynamics–function relationships of biomolecular systems. Designing a correct parameterization for these force fields is still very challenging and depends largely on our ability to interpret experimental data [18,67].

Acknowledgements

P.A. wishes to thank Prof. Eric Westhof for ongoing support as well as Dr Quentin Vicens for careful reading of the manuscript and for helpful discussions. L.D. received a PhD fellowship from the French "Ministère de la recherche et de l'enseignement".

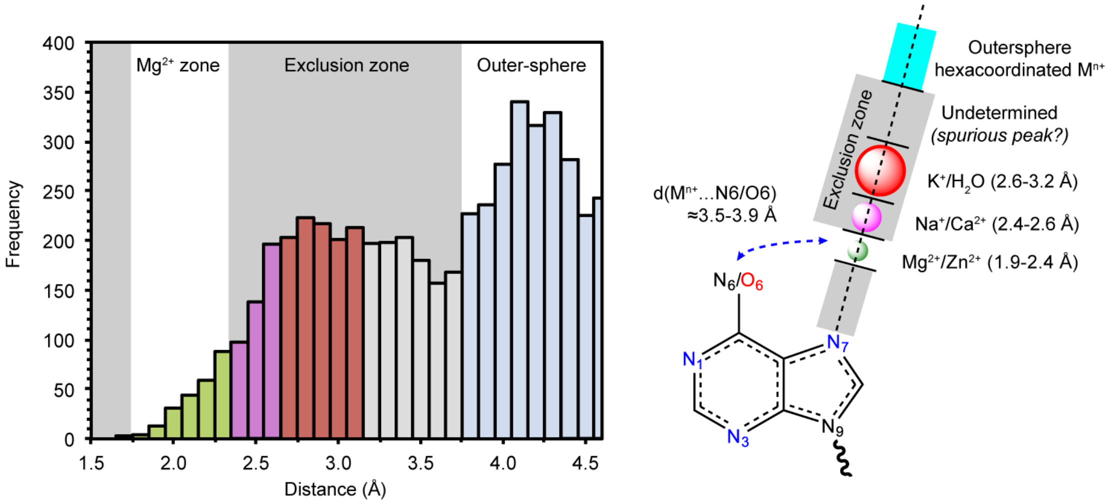
F.L. gratefully acknowledges financial support from the Polish Ministry of Higher Education and Science (Mobility Plus programme, Project No. 1103/MOB/2013/0).

References

[1] R.K.O. Sigel, H. Sigel, Metal–ion interactions with nucleic acids and their constituents, in: J. Reedjik, K. Poeppelmeier (Eds.), Comprehensive Inorganic Chemistry II, Elsevier, Oxford, 2013, pp. 623–660.
[2] T. Dudev, C. Lim, Chem. Rev. 114 (2014) 538.
[3] R.J.P. Williams, Coord. Chem. Rev. 216 (2001) 583.
[4] H.A. Hill, P.J. Sadler, J. Biol. Inorg. Chem. 21 (2016) 5.
[5] D.R. Garmer, N. Gresh, J. Am. Chem. Soc. 116 (1994) 3556.
[6] T. Dudev, Y.L. Lin, M. Dudev, C. Lim, J. Am. Chem. Soc. 125 (2003) 3168.
[7] E. Nieboer, D.H.S. Richardson, Environ. Pollut. Ser. B 1 (1980) 3.
[8] M.J. Harding, M.W. Nowicki, M.D. Walkinshaw, Cryst. Rev. 16 (2010) 247.
[9] H. Zheng, M. Chruszcz, P. Lasota, L. Lebioda, W. Minor, J. Inorg. Biochem. 102 (2008) 1765.
[10] H. Irving, R.J.P. Williams, Nature 162 (1948) 746.
[11] A.W. Foster, D. Osman, N.J. Robinson, J. Biol. Chem. 289 (2014) 28095.
[12] T.W. Lane, M.A. Saito, G.N. George, I.J. Pickering, R.C. Prince, F.M.M. Morel, Nature 435 (2015) 42.
[13] P. Auffinger, N. Grover, E. Westhof, Met. Ions Life Sci. 9 (2011) 1.
[14] P. Auffinger, L. D'Ascenzo, E. Ennifar, Met. Ions Life Sci. 16 (2016) 167.
[15] B. Lippert, Coord. Chem. Rev. 200–202 (2000) 487.
[16] J. Sponer, M. Sabat, L. Gorb, J. Leszczynski, B. Lippert, P. Hozba, J. Chem. Phys. B 104 (2000) 7535.
[17] A. Mir, J. Chen, K. Robinson, E. Lendy, J. Goodman, D. Neau, B.L. Golden, Biochemistry 54 (2015) 6369.
[18] M.T. Panteva, G.M. Giambasu, D.M. York, J. Phys. Chem. B 119 (2015) 15460.
[19] C.W. Bock, A.K. Katz, G.D. Markham, J.P. Glusker, J. Am. Chem. Soc. 121 (1999) 7360.
[20] D. Donghi, J. Schnabl, Met. Ions Life Sci. 9 (2011) 197.
[21] M.C. Erat, R.K. Sigel, Met. Ions Life Sci. 9 (2011) 37.
[22] M. Pechlaner, R.K. Sigel, Met. Ions Life Sci. 10 (2012) 1.
[23] R.K. Sigel, M. Skilandat, A. Sigel, B.P. Operschall, H. Sigel, Met. Ions Life Sci. 11 (2013) 191.
[24] M.H. Harding, Acta Crystallogr., D 55 (1999) 1432.
[25] M.H. Harding, Acta Crystallogr., D 57 (2001) 401.
[26] M.H. Harding, Acta Crystallogr., D 56 (2000) 857.
[27] H. Zheng, M.D. Chordia, D.R. Cooper, M. Chruszcz, P. Muller, G.M. Sheldrick, W. Minor, Nat. Protoc. 9 (2014) 156.
[28] H. Zheng, I.G. Shabalin, K.B. Handing, J.M. Bujnicki, W. Minor, Nucleic Acids Res. 43 (2015) 3789.
[29] F.H. Allen, Acta Crystallogr., B 58 (2002) 380.
[30] J. Chisholm, E. Pidcock, J. van de Streek, L. Infantes, S. Motherwell, F.H. Allen, CrystEngComm 8 (2006) 11.
[31] C.R. Groom, F.H. Allen, Angew. Chem. Int. Ed. Engl. 53 (2014) 662.
[32] I.J. Bruno, J.C. Cole, P.R. Edgington, M. Kessler, C.F. Macrae, P. McCabe, J. Pearson, R. Taylor, Acta Crystallogr., B 58 (2002) 389.
[33] C.F. Macrae, I.J. Bruno, J.A. Chisholm, P.R. Edgington, P. McCabe, E. Pidcock, L. Rodriguez-Monge, R. Taylor, J. van de Streek, P.A. Wood, J. Appl. Crystallogr. 41 (2008) 466.
[34] A.L. Spek, Acta Crystallogr., D 65 (2009) 148.
[35] C. Murray, B. Gildea, H. Muller-Bunz, C.J. Harding, G.G. Morgan, Dalton Trans. 41 (2012) 14487.
[36] G.P. Shields, P.R. Raithby, F.H. Allen, W.D. Motherwell, Acta Crystallogr., B 56 (2000) 455.
[37] S.E. Harris, A.G. Orpen, J.J. Bruno, R. Taylor, J. Chem. Inf. Model. 45 (2005) 1727.
[38] G.D. Markham, J.P. Glusker, C.W. Bock, J. Phys. Chem. B 106 (2002) 5118.
[39] S. Basu, S.A. Strobel, Methods 23 (2001) 264.
[40] Y. Shi, B. Lou, Acta Crystallogr., E 71 (2015) 321.
[41] M. Egli, R.V. Gessner, L.D. Williams, G.J. Quigley, G.A. Vandermarel, J.H. Vanboom, A. Rich, C.A. Frederick, Proc. Natl. Acad. Sci. U.S.A. 87 (1990) 3235.
[42] V. Stelitano, A. Brandt, S. Fernicola, S. Franceschini, G. Giardina, A. Pica, S. Rinaldo, F. Sica, F. Cutruzzola, Nucleic Acids Res. 41 (2013) e79.
[43] P. De Meester, D.M. Goodgame, T.J. Jones, A.C. Skapski, Biochem. J. 139 (1974) 791.
[44] J.E. Sponer, V. Sychrovsky, P. Hobza, J. Sponer, Phys. Chem. Chem. Phys. 6 (2004) 2772.
[45] N. Russo, M. Toscano, A. Grand, J. Phys. Chem. A 107 (2003) 11533.
[46] L. Rulisek, J. Sponer, J. Phys. Chem. B 107 (2003) 1913.
[47] J.M. Martinez, R.R. Pappalardo, E.S. Marcos, J. Am. Chem. Soc. 121 (1999) 3175.
[48] F.C. Lightstone, E. Schwegler, R.Q. Hood, F. Gygi, G. Galli, Chem. Phys. Lett. (2001) 549.
[49] A. Bhattacharjee, A.B. Pribil, B.R. Randolf, B.M. Rode, T.S. Hofer, Chem. Phys. Lett. 536 (2012) 39.
[50] P. Drozdzal, M. Gilski, R. Kierzek, L. Lomozik, M. Jaskolski, Acta Crystallogr., D 69 (2013) 1180.
[51] J.P. Glusker, A.K. Katz, C.W. Bock, Struct. Chem. 12 (2001) 323.
[52] C. Grauffel, C. Lim, J. Phys. Chem. B (2015).
[53] M. Dracinsky, M. Sala, P. Hodgkinson, CrystEngComm 16 (2014) 6756.
[54] A.S. Petrov, G. Lamm, G.R. Pack, J. Phys. Chem. B 106 (2002) 3294.
[55] A. Famulari, F. Moroni, M. Sironi, M. Raimondi, Comput. Chem. 24 (2000) 341.
[56] I. Solt, I. Simon, A.G. Csaszar, M. Fuxreiter, J. Phys. Chem. B 111 (2007) 6272.
[57] T.K. Chiu, R.E. Dickerson, J. Mol. Biol. 301 (2000) 915.
[58] M.M. Perbandt, M.M. Vallazza, C.C. Lippmann, C.C. Betzel, V.V. Erdmann, Acta Crystallogr., D 57 (2001) 219.
[59] N.C. Seeman, J.M. Rosenberg, F.L. Suddath, J.J. Kim, A. Rich, J. Mol. Biol. 104 (1976) 109.
[60] M. Roitzsch, B. Lippert, J. Am. Chem. Soc. 126 (2004) 2421.
[61] A. Nimmermark, L. Ohrstrom, J. Reedijk, Z. Kristallogr. 228 (2013) 311.
[62] I.K. Mati, S.L. Cockroft, Chem. Soc. Rev. 39 (2010) 4195.
[63] G. Wang, B.L. Gaffney, R.A. Jones, J. Am. Chem. Soc. 126 (2004) 8908.
[64] E. Ennifar, P. Walter, P. Dumas, Nucleic Acids Res. 31 (2003) 2671.
[65] R.T. Batey, J.A. Doudna, Biochemistry 41 (2002) 11703.
[66] L.D. Williams, Top. Curr. Chem. 253 (2005) 77.
[67] R.P.P. Neves, S.F. Sousa, P.A. Fernandes, M.J. Ramos, J. Chem. Theory Comput. 9 (2013) 2718.

5.2.3 Paper 5. Mg^{2+} ions: do they bind to nucleobase nitrogens? (Nucleic Acid Res., 2016)

Graphical abstract



An extensive survey of PDB nucleic acid structures was performed to analyze the binding of Mg^{2+} (and other metal ions) to nucleic acid N1/N3/N7 imine nitrogen atoms. A significant number of PDB structures at resolution > 3 Å contains mono- and/or divalent ions, which are associated with a risk for misinterpretation embedded in the lower experimental definition of the structural data. As such, Mg^{2+} is found mostly within low-resolution ribosomal structures and binds only rarely to purine imine nitrogen atoms. In fact, nucleic acid N1/N3/N7 atoms constitute *secondary* coordination sites, while anionic oxygen are primary binding sites. The rare occurrences of Mg^{2+} binding to N7 atoms at resolution ≤ 3 Å are often characterized by unrealistically long coordination distances or odd geometries, which suggests a misinterpretation with other ions such as Na^{+} or K^{+}, or even water molecules. On the other hand, the binding of Mg^{2+} with N1/N3 imine nitrogen atoms is not reliably observed. Almost all direct bindings of Mg^{2+} to N7 with distances < 2.4 Å are suspicious, especially for the low-affinity site on the Hoogsteen edge of adenine. Similarly, Mg^{2+} coordination to two N7 of head-to-tail pairs of stacked guanines has been found in various prokaryotic ribosomes, but the low B-factors associated with significant electron density excess of the ions suggest they would be more realistically modeled as larger Zn^{2+} ions. Another double binding motif, not identified in the previous CSD survey, involves both N7 and O6 of two stacked guanines; in this case the putative Mg^{2+} ion is coordinated in a way reminiscent of the DNA/RNA quadruplexes and is likely a monovalent or transition metal ion, possibly Na^{+}. Magnesium ions binding to imine nitrogen atoms at distances between 2.6 and 3.2 Å lie in the Mg^{2+}…N distance exclusion zone. These distances implicate Na^{+} (~ 2.4 Å) or K^{+}/H_2O/NH_4^{+} (~ 2.8 Å) coordination; considering that these ions are misattributed to Mg^{2+}, the challenge becomes to discern between K^{+}, H_2O and NH_4^{+}.

Further, we call attention on the sometimes excessive application of crystallographic distance restraints to build hydration spheres during structural refinement. This practice, although useful to correctly position octahedral coordination spheres, has been found to suffer from incorrect coordination distances (2.18 Å) which can conceal the presence of Na^+ ions. Similarly, ion replacement strategies have to be handled with great care and especially the temptation to propose new coordination topologies should be resisted, considering the structural and physico-chemical modifications associated with a different ion nature.

Comparing the data from the present study with MgRNA, a repository of classified Mg^{2+} binding sites in RNAs (Zheng et al. 2015), we notice that this database provides the false impression of relevance of Mg^{2+} to N7 contacts, a conclusion that our data do not support. In that respect, we discuss some of the issues we identified in the protocols that were used that might have led to an overestimation of the number of Mg^{2+} binding sites to RNA.

Lastly, we propose several rules to facilitate ion assignment procedures, regarding the placement of ions next to N7 atoms. These rules can easily be adapted to the binding of ions to other nucleic acid sites.

Nucleic Acids Research Advance Access published December 6, 2016

Nucleic Acids Research, 2016
doi: 10.1093/nar/gkw1175

Mg^{2+} ions: do they bind to nucleobase nitrogens?

Filip Leonarski[1,2], Luigi D'Ascenzo[1] and Pascal Auffinger[1,*]

[1]Université de Strasbourg, CNRS, Architecture et Réactivité de l'ARN, UPR9002, F-67000 Strasbourg, France and [2]Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

Received August 21, 2016; Revised November 10, 2016; Editorial Decision November 11, 2016; Accepted November 14, 2016

ABSTRACT

Given the many roles proposed for Mg^{2+} in nucleic acids, it is essential to accurately determine their binding modes. Here, we surveyed the PDB to classify Mg^{2+} inner-sphere binding patterns to nucleobase imine N1/N3/N7 atoms. Among those, purine N7 atoms are considered to be the best nucleobase binding sites for divalent metals. Further, Mg^{2+} coordination to N7 has been implied in several ribozyme catalytic mechanisms. We report that Mg^{2+} assigned near imine nitrogens derive mostly from poor interpretations of electron density patterns and are most often misidentified Na^+, K^+, NH_4^+ ions, water molecules or spurious density peaks. Consequently, apart from few documented exceptions, Mg^{2+} ions do not bind to N7 atoms. Without much of a surprise, Mn^{2+}, Zn^{2+} and Cd^{2+}, which have a higher affinity for nitrogens, may contact N7 atoms when present in crystallization buffers. In this respect, we describe for the first time a potential Zn^{2+} ribosomal binding site involving two purine N7 atoms. Further, we provide a set of guidelines to help in the assignment of Mg^{2+} in crystallographic, cryo-EM, NMR and model building practices and discuss implications of our findings related to ion substitution experiments.

INTRODUCTION

Magnesium has unique physicochemical properties (1,2) and is recognized as the most important divalent ion for RNA folding, structure and function (3–8). Next to monovalent cations and polyamines (9–11), the main Mg^{2+} function is to counterbalance the high concentration of charged phosphate groups present in nucleic acids, but also to assist folding and function through specific binding modes. As such, it is critical to precisely characterize these binding modes.

A recent PDB survey, available through the MgRNA web site (12), which followed earlier efforts to understand Mg^{2+} binding to RNA (13–16), established a classification of these binding sites. Based on these data, for inner-sphere binding, it was found that Mg^{2+} coordination to phosphate and carbonyl groups dominate followed by a still significant number of coordination patterns to imine sites comprising principally purine N7 and less often N1/N3 atoms. Likewise, other studies relay the opinion that N7 positions make for significant nucleobase metal binding sites (17–19). These views contrast with the understanding, based on the pioneering work of RPJ Williams, that alkali earth metals—including Mg^{2+}—poorly bind to imine atoms, unlike transition metals such as Mn^{2+}, Zn^{2+} or Cd^{2+} (8,20–25). These facts cast doubt on the involvement of nitrogen-bound Mg^{2+} ions in catalytic mechanisms, as previously proposed for hammerhead and pistol ribozymes (18,26–33).

In general, the assignment of ions and other solvent molecules in crystallographic structures is a complex undertaking which seems to lead to harmless attribution errors. After all, ion-binding sites are often believed to play a mere structural role. However, Mg^{2+} are sometimes also identified at key locations where misidentifications can dramatically alter our perception of how biomolecular systems perform their tasks. Many of these errors have been described and are related to the fact that Mg^{2+} is isoelectronic with water and Na^+ (34–37). Yet, despite this awareness and other studies reporting recurrent crystallographic misinterpretations (23,35,38–45), errors are still present in many recently deposited PDB structures while older ones are rarely amended (40,46–51).

Identifying errors in structures is a difficult, tedious, but essential undertaking since, if not corrected, these errors will persist in databases and silently affect the outcome of later studies. Further, they can contaminate the results of database surveys (43,52). For example, an RNA polymerase structure with 485 Mg^{2+} and 5 476 waters (PDB code: 1IW7; resolution: 2.6 Å) was released by the PDB in 2002 (cited by (38)) and a *T. thermophilus* 70S ribosome structure, with ≈1 330 Mg^{2+} per assembly, was released in 2014 (PDB code: 4V83; resolution: 3.5 Å). Given their medium to low-resolution range, these structures contain an excessive number of Mg^{2+} and water molecules, since it was suggested that ions assigned to solvent electron densities at resolutions lower than 2.5 Å are not particularly reliable (23,53).

Here, we critically re-examine inner-sphere binding of Mg^{2+} to imine N1/N3/N7 atoms in RNA, DNA and

[*]To whom correspondence should be addressed. Tel: +33 388 41 70 49; Fax: +33 388 60 22 18; Email: p.auffinger@ibmc-cnrs.unistra.fr

purine containing metabolites (PDB; May 2016; resolution ≤ 3.0 Å). We limited our investigations to a small subset of all potential Mg^{2+} binding sites in nucleic acids, since analyzing with the same level of details all potential Mg^{2+} binding modes would have been too lengthy. For the same reasons, we did not analyze outer-sphere binding involving water mediated Mg^{2+} to N7 contacts, which is often considered as significant (12,14,54).

Based on our survey, we established that reliable inner-sphere binding occurrences to imine nitrogen atoms are much less frequent than assumed up to now. We notably reduced the number of Mg^{2+} to nitrogen binding types described in earlier classifications (12). Concomitantly, we characterized a large array of misattribution errors and identified some of the underlying reasons that led to them. Not the least of those is a tendency of experimentalists to want to see ions in their density patterns. This causes an overall bias in the database, because experimentalists have systematically interpreted ambiguous information in a given direction, that is toward unjustified or weakly justified identification of solvent peaks as ions and especially Mg^{2+}. These findings call for a more thorough examination of all ion binding sites found in newly deposited crystal structures, and for a re-examination of the Mg^{2+} assignment process for nucleic acid structures. They also advocate for the more systematic use of anomalous diffraction data to identify heavy ions. Such comprehensive and detailed studies are required to move forward on a subject that received already so much attention. As recently claimed (8), we are still at the beginning of understanding the complex interrelationships that link metals to nucleic acid systems.

We conclude this study by providing a set of rules to facilitate ion binding pattern identification. For example, particular care should be taken to respect the octahedral coordination of Mg^{2+} ions and to avoid the overuse of crystallographic restraints that may lead to a confusion between Mg^{2+} and Na^+ since the main criterion allowing to distinguish them, namely their respectively 2.07 and 2.40 Å coordination distances, is altered. Therefore, we suggest that a significant number of the electron density patterns attributed to Mg^{2+} are generated by other solvent species such as Na^+, K^+, $NH_4{}^+$, polyamines or water. In support to these considerations, we present examples where it is indeed the case and stress the necessity to critically examine solvent density maps with a thorough knowledge of all the types of solvent particles present in purification and crystallization buffers (55,56).

MATERIALS AND METHODS

PDB survey

All ≈ 5 500 nucleic acid crystal structures deposited to the Protein Data Bank (PDB; May 2016; resolution ≤ 3.0 Å) were searched for Mg^{2+} binding to purine and pyrimidine imine N1/N3/N7 atoms (or N_b atoms as defined in reference (12)). To determine cut-off distances for the identification of Mg^{2+} bound to imine nitrogens, we relied on a histogram derived from a CSD search (CSD: Cambridge Structural Database, Version 5.37, February 2016) that identified precise Mg^{2+} to water coordination distances as well as ion exclusion zones (Figure 1). Note that the CSD (57) is a repository for small molecule crystallographic structure that were solved with much better accuracy and, in general, at much higher resolution than those from the PDB. These data parallel those derived from quantum mechanical calculations (58), other PDB surveys (23,59) and first principles molecular dynamics simulations of Mg^{2+} in aqueous solution (60–62) that all suggest that: (i) the $Mg^{2+}\ldots OH_2$ coordination distance is slightly below 2.1 Å; (ii) no water oxygens are found within a $d(Mg^{2+}\ldots Ow)$ ≈2.2–3.8 Å 'exclusion zone'; (iii) the second coordination shell starts at a 3.8 Å distance from Mg^{2+} and peaks at 4.2 Å. However, since we mostly deal with medium to low-resolution crystallographic structures (3.0 Å ≥ resolution ≥ 2.0 Å), we used more relaxed criteria to identify solvent species around imine nitrogens. Further, we need to consider that, although the most appropriate $Mg^{2+}\ldots O$ coordination distance is in the 2.06–2.08 Å range, the default value in the libraries used by the PHENIX (63) and REFMAC (64) refinement programs for $d(Mg^{2+}\ldots Ow)$ is 2.18 Å. In some instances, this overestimated coordination distance induces serious stereochemical approximations (see below). Bearing in mind that we focus on Mg^{2+} to nitrogen distances, we have also to consider that some authors estimate that the $Mg^{2+}\ldots N$ distance is slightly longer (≈2.2 Å) than the $Mg^{2+}\ldots O$ distance in agreement with quantum mechanical calculations and PDB/CSD surveys (12,2125). Thus, to a first approximation, our procedures place Mg^{2+} with $d(Mg^{2+}\ldots N) \leq 2.4$ Å in the pool of possible direct binders, while those with distances in the 2.4–3.8 Å exclusion zone were inspected for misidentification.

Since CSD surveys established that divalent ions directly interacting with a purine or imidazole nitrogen lone pair are located in the C–N=C plane (25,65), we applied a 1.0 Å cut-off on the distance between the ion and the nucleobase plane. This criterion applies to divalent ions and not to the less strongly bound alkali (Na^+, K^+) and the larger alkali earth ions (Ca^{2+}, Sr^{2+}) that display a greater propensity to lie out-of-plane. The searches included also contacts generated by applying crystallographic symmetry operations.

In the ≤ 3.0 Å resolution range, ions with *B*-factors ≥ 79 $Å^2$ were excluded from our statistics since such high *B*-factors do not warrant unequivocal binding site characterizations. Further, we excluded ions with *B*-factors ≤ 1.0 $Å^2$ that are definitely not reliable for Mg^{2+} and hint to the presence of a more electron rich atom (see below). Only Mg^{2+} with occupancy of 1.0 were considered unless otherwise specified. Finally, for all Mg^{2+} ions close to imine nitrogens that we identified as suspect, the $F_o - F_c$ and $2F_o - F_c$ electron density maps deposited to the Uppsala Electron Density Server (EDS) were visualized (66). When these maps were not available—typically for large ribosomal structures—we calculated them with phenix.maps by using the structure factors retrieved from the PDB (63).

Non-redundant Mg^{2+} binding sites were identified as follows. If two nucleotides from different structures involved in a similar Mg^{2+} binding event shared the same residue numbers, chain codes, trinucleotide sequences, ribose puckers, backbone dihedral angle sequences (we used the g+, g−, t categorization) and *syn*/*anti* conformations, they were considered as similar and the one with the best resolution was marked as non-redundant. In case of matching resolu-

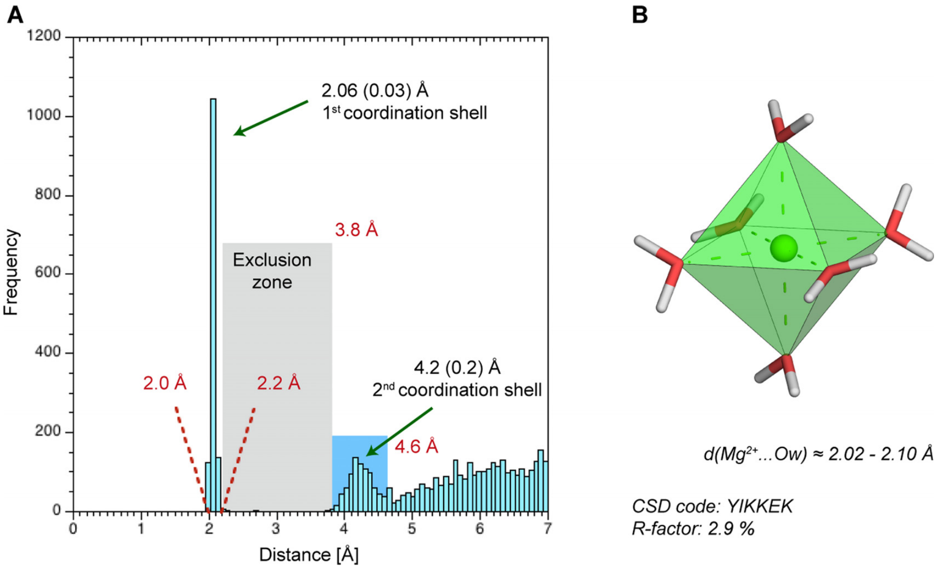


Figure 1. The Mg^{2+} first hydration shell is strictly defined as deduced from high-resolution crystal structures. (**A**) $d(Mg^{2+}\ldots Ow)$ histogram derived from the CSD (version 5.37, update February 2016; *R*-factors ≤ 5%). No disordered, error containing, polymeric or powder structures were included. Standard deviations for the average $Mg^{2+}\ldots Ow$ coordination distances are given in parenthesis. The water exclusion zone and the second Mg^{2+} coordination shell are marked by a grey and a light blue rectangle, respectively. (**B**) An ultra high-resolution $Mg[H_2O]_6{}^{2+}$ CSD x-ray structure examplifyes the strict octahedral water arrangement around Mg^{2+}(125).

tions, the nucleotide with the lowest *B*-factor was selected. Likewise, if in the same structure two nucleotides involved in a similar Mg^{2+} binding event shared the same residue numbers and trinucleotide sequences (with different chain codes) as well as ribose puckers, backbone dihedral angle sequences and *syn*/*anti* conformations, they were considered as similar and the one corresponding to the first biological unit was marked as non-redundant. To further limit redundancy in the largest ribosomal structures, we restricted our analysis to a single biological assembly when more than one was present (see Supplementary Material for selection criteria).

Two non-redundant sets were calculated with a 2.4 and a 3.5 Å $d(Mg^{2+}\ldots N1/N3/N7)$ distance cutoff, respectively (Table 1). Note that it is impossible to completely eliminate redundancy from such a complex structural ensemble without eliminating at the same time relevant data. Here, we provide an upper limit for a truly 'non-redundant' set. Redundancy issues are further complicated by some systematic assignment errors such as the nucleotide misidentification identified in the first *H. marismortui* 50S structures that leads to the characterization of two distinct structural ensembles (Supplementary Table S1 and Figure S1).

RESULTS AND DISCUSSION

Mg^{2+} to imine N1/N3/N7 contacts are rare

As of May 2016, ≈56 000 Mg^{2+} ions are assigned in ≈1 000 nucleic acid crystallographic structures from the PDB (resolution ≤ 3.0 Å). This corresponds to a ratio of roughly one Mg^{2+} per eight nucleotides. Comparatively, under the same resolution threshold, ≈25 000 Mg^{2+} are found in ≈8 500 proteins. With no resolution limit, the number of Mg^{2+} rises to ≈100 000 in nucleic acids and only ≈30 000 in proteins. In nucleic acids, the largest number of ions comes from low-resolution ribosomal structures that do not allow reliable identification of light solvent particles.

Out of the ≈56 000 Mg^{2+} found in nucleic acid structures, ≈1 000 (≈1.8%) display partial occupancies, 59 are associated with occupancies above 1.0, ≈4 100 (≈7%) display *B*-factors ≥ 79 $Å^2$ and ≈480 (< 1%) display *B*-factors ≤ 1.0 $Å^2$. We excluded these ions from statistics in Table 1. In the remaining pool, around 3 900 (≈7%) ions display $d(Mg^{2+}\ldots N)$ ≤ 3.5 Å. If we consider the more stringent $d(Mg^{2+}\ldots N)$ ≤ 2.4 Å criterion that is more in line with the coordination distance derived from the CSD (Figure 1), only 293 (≈0.5%) Mg^{2+} are contacting imine nitrogens. Most of these Mg^{2+} are close to 108 adenine and 69 guanine N7 atoms, with only 20 close to N1/N3 positions. This number drops to 126 (≈0.3%) if we consider only non-redundant Mg^{2+} binding sites. These values are to be compared with the ≈8 500 (≈15%) Mg^{2+} in direct contact (≤ 2.4 Å) with phosphate oxygens that are considered to be the primary nucleic acid binding sites for Mg^{2+}.

$d(Mg^{2+}\ldots N)$ histograms reveal unrealistic coordination distances

As stated in the Methods section, the coordination geometry of Mg^{2+} to water and other ligands is strictly defined. Ideally, the $d(Mg^{2+}\ldots N)$ and the $d(Mg^{2+}\ldots Ow)$ PDB and CSD histograms should display a similar profile. However, in the former (Figure 2), we could not identify a clear peak around 2.1 Å. Furthermore, the exclusion zone identified in Figure 1 is significantly populated in the PDB data, suggesting ion misidentifications. Mg^{2+} assignments with coordination distances in the 2.4–2.6 Å range may correspond to Na^+ that are frequently present in crystallization buffers—for example as sodium cacodylate—and have coordination distances to water around 2.4 Å, as shown in a

Table 1. Number of non-redundant $Mg^{2+}\ldots$N1/N3/N7 contacts in structures from the PDB (resolution ≤ 3.0 Å)

	$d(Mg^{2+}\ldots N1)$		$d(Mg^{2+}\ldots N3)$		$d(Mg^{2+}\ldots N7)$	
	≤ 3.5 Å	≤ 2.4 Å	≤ 3.5 Å	≤ 2.4 Å	≤ 3.5 Å	≤ 2.4 Å
DNA						
DA	—	—	—	—	3 (3)	—
DG	*NR*	*NR*	1 (1)	—	23 (23)	8 (8)
DC	*NR*	*NR*	1 (1)	—	*NR*	*NR*
RNA						
A	116 (243)	5 (5)	131 (198)	2 (6)	245 (783)	24 (108)
G	*NR*	*NR*	84 (121)	5 (6)	1324 (2386)	84 (191)
C	*NR*	*NR*	122 (192)	3 (3)	*NR*	*NR*
Total:	116 (243)	5 (5)	339 (513)	10 (15)	1595 (3195)	116 (307)

The total number of occurrences is given in parenthesis ('*NR*' stands for 'Non-Relevant'). Ions with *B*-factors ≤ 1.0 $Å^2$ and ≥ 79 $Å^2$ were not counted.

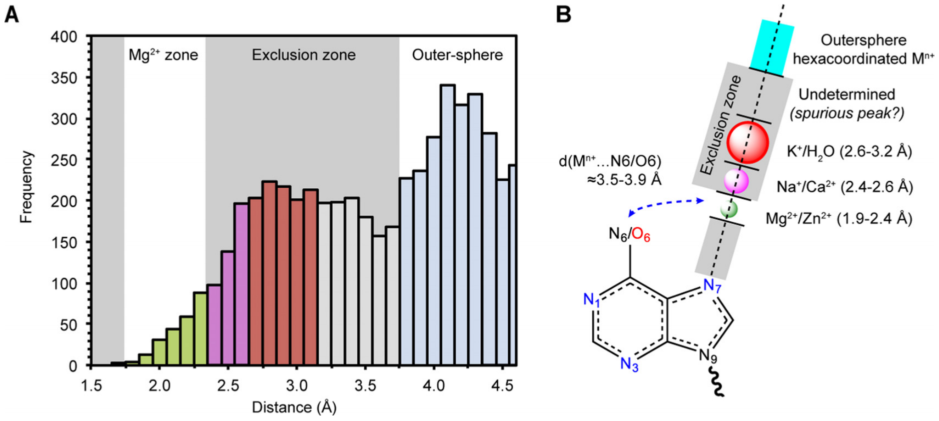


Figure 2. Mg^{2+} coordination to purine N7 atoms derived from PDB structures. (**A**) $d(Mg^{2+}\ldots N7)$ histogram (derived from the PDB; May 2016; resolution ≤ 3.0 Å). Ions with *B*-factors ≤ 1.0 $Å^2$ and ≥ 79 $Å^2$ were excluded. The different ion binding zones are colored according to Figures 1 and 2B. (**B**) Scheme showing the different ion binding zones in front of the purine N7 atom (the oxygen and nitrogen atoms able to associate with a cation are shown in red and blue, respectively). The $d(M^{n+}\ldots N6/O6)$ expected distance range is also indicated. Note that these cutoff distances are indicative and simply suggest that the potential ion assignments close to the mentioned limits should be considered with greater care.

hammerhead ribozyme structure (PDB code: 3ZP8; resolution: 1.55 Å) (67). Note that Ca^{2+} with similar coordination distances are mentioned in the crystallization conditions of some *T. thermophilus* 30S structures (see Supplementary Material). Mg^{2+} assignments with coordination distances in the 2.6–3.2 Å range may correspond to K^+, NH_4^+ or water, all with coordination distances around 2.8 Å. Mg^{2+} assignments in the 3.2–3.8 Å range may be related to the presence of anions (68,69), solvent molecules pertaining to the purification and crystallization buffers, contaminants or may be related to crystallographic artifacts (40,70). In accordance with CSD data, the broad peak around 4.2 Å is attributable to Mg^{2+} interacting with imine sites through their first hydration shell. Interestingly, the abnormalities in the $d(Mg^{2+}\ldots N7)$ histogram become more apparent when compared to the $d(Mn^{2+}\ldots N7)$ histograms that show a clear first shell peak in the 2.1–2.6 Å range (Figure 2 and Supplementary Figure S2).

A second binding criterion, derived from CSD searches (25), specifies that when a transition metal or Mg^{2+} binds to a purine N7 atom, the allowed $d(M^{2+}\ldots N6/O6)$ should be in the 3.5–3.9 Å range (Figure 2B). Out of 111 non-redundant ions with $d(Mg^{2+}\ldots N7) \leq 2.4$ Å, only 62 ions satisfy this criterion. Thus, we infer that the majority of Mg^{2+} close to imine nitrogens are misidentified (Supplementary Figure S3).

With a few exceptions, all direct binding occurrences with $d(Mg^{2+}\ldots N7) \leq 2.4$ Å are suspect

Mg^{2+} singly bound to adenine N7 atoms is not observed. Direct ion binding to the N7 position of adenine is complicated by the presence of the nucleobase amino group that imposes steric and electrostatic constraints (25). Although pentahydrated ion-to-N7 binding has been observed at high-resolution in an adenine/nickel complex (CSD code: ZZZAAF01; Supplementary Figure S4), only few partially hydrated and no $Mg(H_2O)_5^{2+}$ were located close to (A)N7 in the PDB. In a hammerhead ribozyme (29), the conformation of a nucleobase that involves a $d(Mg^{2+}\ldots N6/N7) \leq 2.0$ Å coordination was probably incorrectly modeled (71) (Figure 3A). Elsewhere in the same structure, completing the hydration sphere of an adenine-bound Mg^{2+} resulted in severe clashes with adjacent nucleotides (Supplementary Figure S5). In a glmS ribozyme (PDB code: 2GCV; resolution: 2.1 Å), the *B*-factor is higher for the metal ion than for the attached nucleobase and waters, which hints at refinement issues combined to ion misidentification (Figure 3B). Moreover, a Mg^{2+} has been placed at 3.1 Å from a cytosine N4 atom (most probably a Cl^- ion (68,69); Supplementary

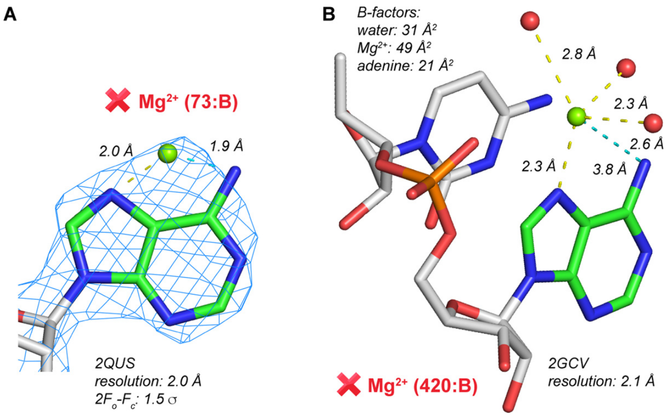


Figure 3. Unrealistic binding of Mg^{2+} to adenine N7 atoms. Mg^{2+} to N6 coordination distances are shown in cyan. The red cross is used to mark misidentified Mg^{2+} ions. (**A**) Ill-placed Mg^{2+} according to electron density patterns and coordination distances. (**B**) The Mg^{2+} hydration sphere is incomplete with erratic coordination distances. Fixing the hydration sphere of this ion with a proper hexacoordinated geometry would result in clashes similar to those shown in Supplementary Figure S5.

Figure S6A), suggesting to critically reexamine all other solvent assignments proposed for this glmS ribozyme. Finally, a suspicious Mg^{2+} bound to a N7 atom—more probably Na^+—is found in a ribosome where $d(Mg^{2+}\ldots N7)$ and $d(Mg^{2+}\ldots Ow)$ are ≈2.4 and ≈2.6 Å, respectively (Supplementary Figure S6B).

This quasi-absence of reliable $Mg(H_2O)_5{}^{2+}$ to (A)N7 contacts suggests a low Mg^{2+} affinity for this site (25). However, in rare instances, secondary Mg^{2+} contacts to N7 atoms complemented by primary contacts to anionic phosphate or amino acid oxygens may be associated with the formation of appropriate but rare Mg^{2+} binding pockets (see below).

Mg^{2+} singly bound to guanine N7 atoms: is it more probable? Binding of divalent metals to (G)N7 atoms has been reported more frequently in both the CSD and the PDB as a probable result of the larger electronegativity of guanine versus adenine Hoogsteen edges (25). In the PDB, 41 non-redundant Mg^{2+} bind solely to (G)N7 atoms (no other direct contact to DNA/RNA atoms). Out of those, only 20 Mg^{2+} comprising three $Mg(H_2O)_5{}^{2+}$ satisfy the $3.5 \leq d(M^{2+}\ldots O6) \leq 3.9$ Å criterion (Figure 2). Two of these $Mg(H_2O)_5{}^{2+}$ are present in the same synthase/tRNA structure (PDB code: 4YCO; resolution: 2.1 Å) and have been modeled based on octahedral densities displaying merged water peaks. $d(Mg^{2+}\ldots Ow) = 2.18$ Å restraints were used during refinement (Figure 4A). As a result, Na^+ can be fitted with a similar level of confidence into these density patterns (see below).

The remaining 21 ions with outlier $d(Mg^{2+}\ldots O6)$ distances were assigned without proper care for stereochemistry and hydration patterns. In a few instances, a complete hydration shell was modeled. However, without well-defined solvent density patterns, these hydration shells display poor geometry. In a twister ribozyme structure (PDB code: 5DUN; resolution: 2.6 Å), five waters were fitted in a density pattern lacking octahedral symmetry and the Mg^{2+} *B*-factor is larger than that of the bound nucleobase (Figure 4B). In a group II intron (PDB code: 4E8N; resolution: 3.0 Å), a Mg^{2+} is placed in front of an N7 atom and displays a non-octahedral coordination (Figure 4C).

Mg^{2+} binding to (G)N7 in DNA is rare. Ion coordination to the N7 of a terminal guanine was identified in five Z-DNA hexamers with resolution ≈1.0 Å (Figure 4D) while in a few structures (PDB codes: 4HIG, 4HIF, 1D39), Mn^{2+}, Zn^{2+} and Cu^{2+} replace Mg^{2+} (72,73). All these structures belong to the $P2_12_12_1$, space group. However, these hexamers crystallize in the $P2_12_12_1$ space group in both a Mg^{2+} and a spermine form (74–76) suggesting that Mg^{2+} is not specifically required to stabilize the crystal. Further, under high $MgCl_2$ and $CaCl_2$ concentrations (500 mM), this hexamer crystallizes in a $P32$ space group and, surprisingly, no direct ion binding to N7 was reported (77) as in another Z-DNA structure with much lower divalent ion concentrations (30 mM), stressing the difficulty to predict such ion binding patterns (78).

Pentahydrated Mg^{2+} coordination to (G)N7 was reported in only one B-DNA structure (PDB code: 1DCR; resolution: 1.6 Å). It shows clearly identifiable solvent density peaks as well as coordination distances and geometry consistent with Mg^{2+} binding. Similarly, Co^{2+} and Zn^{2+} bind to a terminal guanine of two DNA hexamers (PDB codes: 1FD5, 1P26; resolutions: 1.10, 2.92 Å) (79). Again, the terminal position is favored since steric hindrance prevents pentahydrated metals to bind to N7 within a B-DNA helical context (79,80) (Supplementary Figure S4). It is surprising that $Mg(H_2O)_5{}^{2+}$ binds to terminal Z- and B-DNA but not to RNA nucleobases.

M^{n+} bound to two N7 atoms of stacked purines: Mg^{2+}, Zn^{2+} or a monovalent cation? Next to pentahydrated ion binding sites, we searched for an atypical pattern involving the coordination of a divalent cation to two N7 atoms belonging to purines arranged in a stacked head-to-tail manner (Figure 5A), a pattern that we identified earlier in the CSD (25) and was also described by others (8,12,14,81). We identified 45 cases with $d(Mg^{2+}\ldots N7) \leq 2.4$ Å, all associated with four prokaryotic ribosome sites (Table 2 and Figure 5B). Sites I, III and IV were observed in important structural elements—three-way junction for site I and bulges for sites III and IV— while site II is constitutive of ribosomal helix 52. Hence, we analyzed more systematically the 289 potential binding sites in the 126 prokaryotic ribosomes we surveyed (Supplementary Table S1). We excluded 18 instances where both purine *B-factors* are larger than 79 Å^2 leading thus to a total of 271 binding sites. Among them, Mg^{2+} was assigned in 243 instances and Sr^{2+} in 28 instances since this ion was used in crystallization buffers of *H. marismortui* large ribosomal subunits (see below).

However, for the largest number of sites, Mg^{2+} attribution is inappropriate since only less than one out of every six ion satisfies the $d(Mg^{2+}\ldots N7) \leq 2.4$ Å criterion (Table 2 and Figure 5B). Therefore, these sites appear to be occupied by waters or non-Mg^{2+} ions for which the coordination distance to N7 is > 2.4 Å. These data illustrate the difficulty of thoroughly analyzing the current pool of ribosomal structures for which we do not only have to deal with various resolution levels but also with a large gamut of crystallization protocols, refinement procedures and interpretation habits

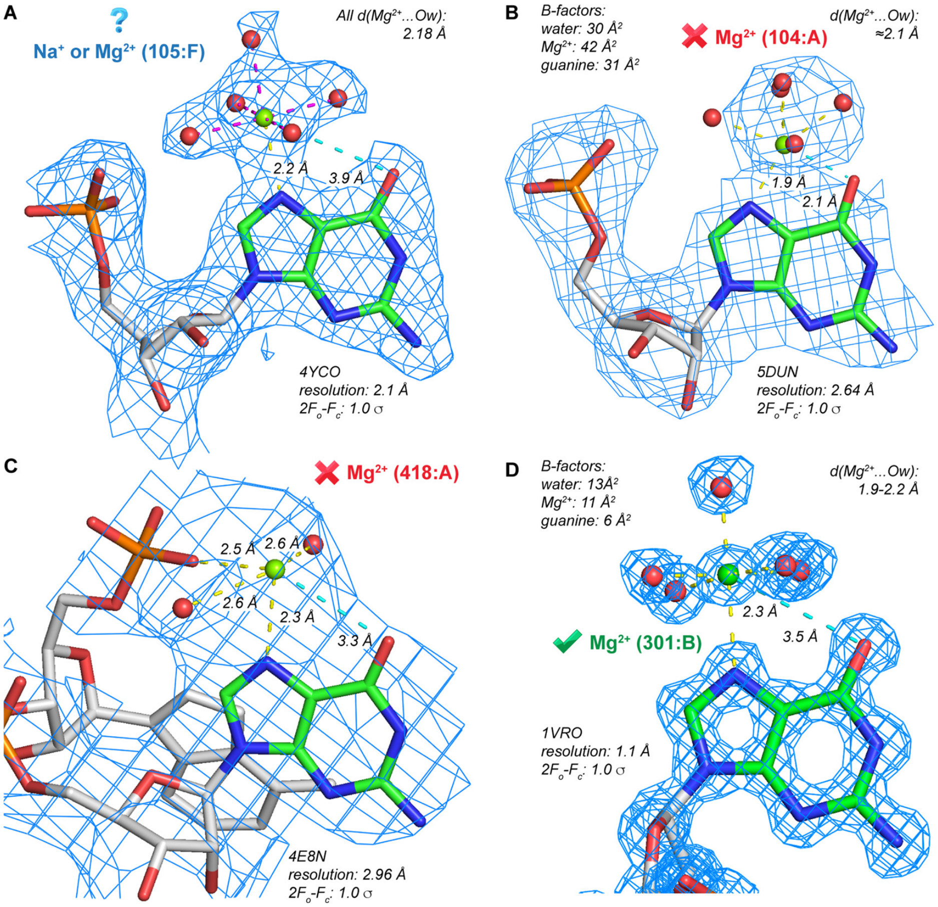


Figure 4. Mg^{2+} close to guanine N7 atoms in PDB structures. The cyan question mark, red cross and green mark are used to identify: (i) sites where either Na^+ or Mg^{2+} match the electron density, (ii) a misidentified and (iii) a correctly placed ion. (**A**) The $d(Mg^{2+}\ldots Ow)$ coordination distances, shown in magenta, were irrealistically modeled to 2.18 Å. A larger Na^+ could equally fit into this density pattern. (**B**) Mg^{2+} is distant from the electron density center, leading to underestimated $d(Mg^{2+}\ldots N7)$ and $d(Mg^{2+}\ldots O6)$ distances. (**C**) Incomplete and poorly defined Mg^{2+} coordination shell. The coordination distances suggest the presence of Na^+ or water but not Mg^{2+}. (**D**) A reliable but rare pentahydrated coordination pattern with separate densities for water and Mg^{2+}. Similar patterns are reported in a set of high-resolution Z-DNA crystal structure (PDB codes: 1VRO, 292D, 2DCG, 2ELG, 336D).

(37,82–83). Going through such a demanding process is at least necessary for some sites. Such coordination distance spreads could certainly not have been foreseen otherwise.

In the following, we focus on the 43 ions with $d(Mg^{2+}\ldots N7) \leq 2.4$ Å found at site I and IV. Although their coordination distances seem appropriate, we noticed several inconsistencies that led us to question their identity. The first is associated with ion *B-factors* that show a propensity to be lower than those of the attached purines (Table 2 and Figure 5C): in one instance, for site I, the Mg^{2+} *B*-factor was set to zero (PDB code: 4U20); in 10 instances, the ion *B*-factors were set to ≤ 1.0 Å^2. Such low *B*-factors usually appear when the atom at the origin of the observed density has more electrons than the one used in the model. We identified additional strategies used by experimentalists to absorb excess density in the F_o-F_c maps at site IV: first, an occupancy of 1.3 was assigned to Mg^{2+} (PDB code: 1N32; resolution: 3.0 Å); second, two Mg^{2+} separated by 1.7 Å and each with 1.0 occupancy were modeled (PDB code: 4B3M; resolution: 2.9 Å). Both of these scenarios are physically impossible. Consequently, we pondered about which other ion could explain such density patterns and realized that, by analogy with zinc-fingers where Zn^{2+} binds to two histidine residues, this double N7 motif represents also an appropriate binding site for Zn^{2+}, a metal that should not be perceived as a trace element given high Zn^{2+} intracellular concentrations (84). We realized also that Zn^{2+} has been identified in the ribosomal proteins (85–87) of almost all *T. thermophilus*, *E. coli* and *S. cerevisae* structures (Supplementary Table S1). Further, although Zn^{2+} displays a tetrahedral coordination in zinc-fingers, this ion binds sometimes to (G)N7 with an octahedral coordination (see PDB code: 4HIF; resolution: 0.85 Å) (73). In order to validate our hypothesis of Zn^{2+} replacing Mg^{2+}, we examined the peak height of the electron densities associated with some of these ions. In 11 instances out of 43, the density of the ion remains visible at sigma levels above those corresponding to neighboring phosphorus atoms, therefore strongly suggesting the presence of a transition metal (Figure 5D). When we attempted to re-refine some of these structures by replacing Mg^{2+} by Zn^{2+}, reasonable ion *B*-factors and no abnormal positive or negative peaks in the F_o-F_c density maps were obtained.

As mentioned above, 28 structures of the large *H. marismortui* ribosomal subunit where crystallized in the pres-

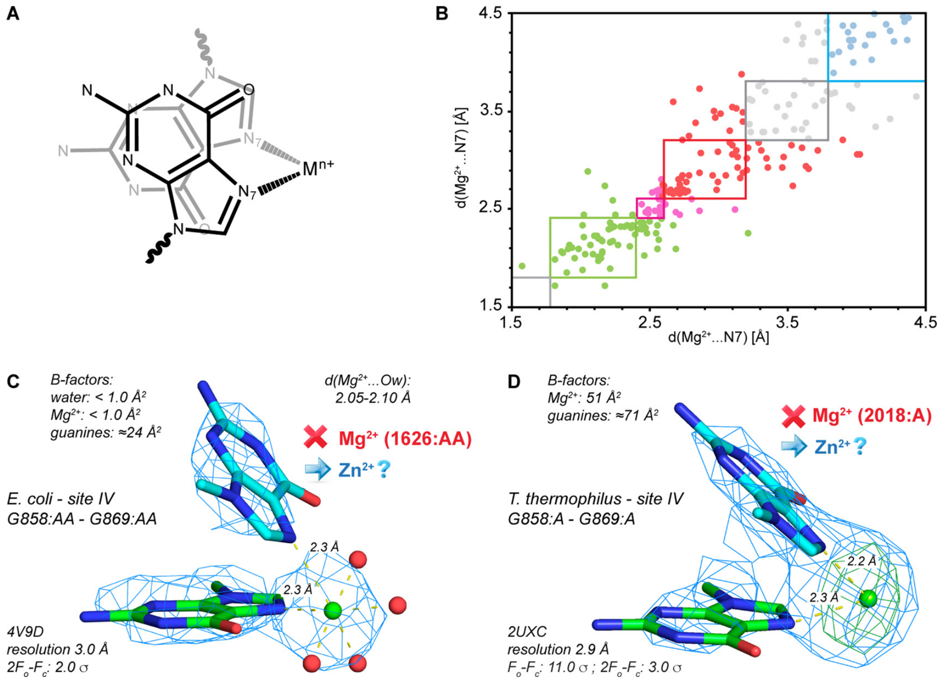


Figure 5. M^{n+} bound to a double N7 motif. (**A**) Schematical representation of this ion binding pattern with two guanines. Guanine-adenine combinations were also identified (Table 2). (**B**) 2D diagram showing Mg^{2+} distances with respect to each of the bound N7 atoms. The Mg^{2+} dots are colored according to their distance to the closest N7 atom (see Figure 2). The colored rectangular boxes frame the ions with respect to both coordination distances. (**C**) A Mg^{2+} placed close to site IV in an *E. coli* structure (Table 2). Though the coordination distance is correct, this ion assignment is ambiguous since ion and water *B*-factors < 1.0 Å^2 are not consistent with those of the guanines. These facts hint to the presence of a more electron dense ion such as Zn^{2+}. (**D**) Mg^{2+} placed close to site IV in a *T. thermophilus* structure. This ion assignement is ambiguous since, although the coordination distances are in agreement with those of Mg^{2+}, the $F_o - F_c$ density (in orange) points to the presence of a more electron dense ion, possibly Zn^{2+}.

Table 2. Occurrence of Mg^{2+} and Sr^{2+} ions bound to two N7 atoms of stacked purines in ribosomal structures (resolution ≤ 3.0 Å; see Supplementary Table S1 and Figure 5)

Site	Org.[a]	Res.[b]	$d(Mg^{2+}\ldots N7) \leq$ 2.4 Å[c]	$d(Mg^{2+}\ldots N7) >$ 2.4 Å[d]	Second-shell Mg^{2+} [e]	Sr^{2+} [f]	Empty site [g]	Total [h]
Large ribosomal subunit (LSU)								
I.	*HM*	G:824-G:854	27 (18)	1 (1)	—	28 (0)	1	57
	TT	G:733-A:761	4 (3)	37 (27)	—	*NR*	1	42
	EC	G:733-A:761	5 (4)	7 (5)	—	*NR*	—	12
	DR	G:733-A:761	—	—	—	*NR*	—	—
II.[i]	*TT*	G:1358-G:1371	—	13 (6)	26 (4)	*NR*	3	42
	EC	G:1358-G:1371	—	12 (8)	—	*NR*	—	12
	DR	G:1358-G:1371	—	—	—[j]	*NR*	1	1
Small ribosomal subunit (SSU)								
III.[i]	*TT*	G:581-G:758	2 (2)	44 (17)	2 (1)	*NR*	4	52
IV.	*TT*	G:858-G:869	6 (6)	35 (26)	—	*NR*	—	41
	EC	G:858-G:869	1 (1)	11 (10)	—	*NR*	—	12
Total**:**			45 (33)	160 (99)	28 (5)	28 (0)	10	271

The number of ions with an inappropriate *B*-factor—lower than that of at least one of its bound N7 atoms—is given in parenthesis. A site is counted only when the *B*-factors of any of the two purines are below 79 Å^2; no *B*-factor criterion was applied to the ion.

[a] Organisms in which these motifs occur; HH, TT, EC and DR stand for *H. marismortui*, *T. thermophilus*, *E. coli* and *D. radiodurans*, respectively.

[b] The residue numberings are those found in the most representative PDB structure for each organism as noted in Supplementary Table S1, namely 4V9F, 4Y4O, 4YBB, and 5DM6 for HM, TT, EC and DR, respectively.

[c] Both $d(Mg^{2+}\ldots N7)$ distances have to be below 2.4 Å.

[d] One of both $d(Mg^{2+}\ldots N7)$ distances has to be in the 2.4–3.8 Å range.

[e] Both $d(Mg^{2+}\ldots N7)$ distances have to be in the 3.8–4.6 Å range.

[f] Sr^{2+} ions are only present in some HH–LSU structures (see Supplementary Table S1). Sr^{2+} ions are considered if $d(Mg^{2+}\ldots N7) \leq 3.0$ Å.

[g] No ions with the criteria defined above are found at these sites.

[h] Total number of identified sites.

[i] Site II. in HM and site III. in EC have no double N7 ion binding motif.

[j] Without the *B*-factor < 79 Å^2 criterion, two second shell Mg^{2+} ions are reported in DR structures (PDB codes: 5DM6, 5DM7).

ence of Sr^{2+}. In site I, Mg^{2+} is systematically replaced by Sr^{2+} that accounts better for the observed electron density and results in ion *B-factors* larger than those of the nucleobases (Table 2). However, a close inspection of the crystallographic data suggests that Sr^{2+} is incompatible with the observed electron density: *B-factors* are twice those of the nucleobase with some $d(Sr^{2+}\ldots N7)$ distances as short as 2.2 Å, while the CSD estimated Sr^{2+} coordination distance is ≈2.6 Å and the preferred ligands are oxygens (23,88). Moreover, all alkali earth metals, including Sr^{2+}, are poor N7 binders (25). Rightfully, the authors of these structures did not envisage the binding of Cd^{2+}, an ion that is present in all *H. marismortui* structures (Supplementary Table S1) and has an excellent affinity for nitrogens but has also 10 and 18 more electrons than Sr^{2+} and Zn^{2+}, respectively.

We are aware that the data we gathered are not sufficient to unambiguously identify the ions present at these locations. However, the possibility that site I and IV bind Zn^{2+} is strongly supported by our analysis and should be further investigated. Based on EXAFS experiments, it has been proposed that the *E. coli* 70S ribosomes tightly bind to 8 equivalents of Zn^{2+} (87,89). The authors of this study suggested that, next to zing-finger motifs, another strong Zn^{2+} binding site was associated with ribosomal RNA but were unable to characterize it. Therefore, we hypothesize that these double N7 binding sites, that are poor binding sites for alkali earth ions, are the best ribosomal locations for Zn^{2+} and other transition metals and bind eventually monovalent cations when transition metals are not present. Definite answers regarding the identity of these ions will have to wait for multi-wavelength anomalous diffraction measurements (36,56,90).

M^{n+} bound to N7/O6 atoms of stacked guanines: Mg^{2+} or a monovalent cation? Contrary to the double N7 binding site described above (Figure 5), a pattern where both N7 and O6 atoms belonging to stacked purines coordinate Mg^{2+} has not been identified in the CSD. In the PDB, such a motif with $d(Mg^{2+}\ldots N7/O6) \leq 2.4$ Å is found 8 times and is associated with ApG and GpG steps (Figure 6A). This motif has been first described in a P4–P6 group I intron structure (Figure 6B) and is since cited as a good example of a well-defined Mg^{2+} binding pocket (91–93). However, even at 2.25 Å resolution (PDB code: 1HR2), the ion density is merged to that of the attached waters prohibiting unequivocal Mg^{2+}/Na^{+} identification. In support to this assumption, a water has replaced Mg^{2+} in a related P4-P6 structure (PDB code: 2R8S; resolution: 1.95 Å) marking a poor divalent binding site. Moreover, this binding pattern is reminiscent of that of monovalent ions to carbonyl groups in DNA/RNA quadruplexes where an ion bridges two O6 atoms of 'stacked' guanines. Indeed, a 'semi-quadruplex' binding pattern with Mg^{2+} bound to the carbonyl groups of a GpU step has been identified in the same group I intron fragment (Figure 6C). This Mg^{2+} is more probably Na^{+} given $d(M^{n+}\ldots Ow)$ in the 2.3–2.5 Å range.

Thus, we hypothesize that this site is not occupied by Mg^{2+} but rather by monovalent cations or transition metals as suggested by the binding of a hexacoordinated cobalt ion to a B-DNA structure (PDB code: 4R4A, resolution: 1.45 Å). Such sites could also be occupied by Mn^{2+}, questioning results from ion substitution experiments. To summarize, the significance of this site is limited since the binding of Mg^{2+} to consecutive purines has only been reported in 8 instances, although every ribosome contains on average > 200 similar purine-purine steps.

N7: a secondary Mg^{2+} binding site next to primary anionic oxygens. As discussed above, Mg^{2+} to N7 binding is rare and existing assignments are often questionable. However, in some instances, N7 may correspond to a secondary Mg^{2+} coordination site when the ion is primarily bound to anionic phosphate or carboxylate oxygens. We gathered evidence from CSD structures that when such multiple binding occurs, the distance to the anionic oxygen over the nitrogen atom is systematically shorter by 0.1-0.2 Å (25).

In the PDB, we identified 80 sites with $d(Mg^{2+}\ldots N7/OP) \leq 2.4$ Å for which 71 satisfy the $d(Mg^{2+}\ldots OP) < d(Mg^{2+}\ldots N7)$ criterion. Among those, 51 and 20 sites involve an N7 atom belonging to an adenine and guanine nucleobase, respectively. Yet, a large number of them are redundant. For example, 48 out of 51 adenines are located in a loop capping helix 11 of the large ribosomal subunit and involve an N7 atom and two phosphate groups (Figure 7A). This site is present in all four ribosome families (Supplementary Table S1). Elsewhere in ribosomes, we identified only 8 weak non-redundant sites. These sites are at best occupied by Mg^{2+} satisfying our stereochemical criteria in five instances in the 134 surveyed ribosome structures. Hence, Mg^{2+} directly bound to phosphate groups are rarely establishing direct contacts to N7 atoms given the paucity of appropriate structural contexts in RNA and DNA.

Furthermore, it is important to consider that solvation conditions in the surveyed ensemble of ribosomal structures are very heterogeneous. While the highest populated site (Supplementary Figure S7A) points to Mg^{2+} in high-resolution *H. marismortui* structures (see PDB code: 1VQ8; resolution: 2.2 Å), in two other high-resolution structures from *E. coli* (PDB code: 4YBB; resolution: 2.1 Å) and *T. thermophilus* (PDB code: 4Y4O; resolution: 2.3 Å), $d(Mg^{2+}\ldots N7)$ distances are more consistent with the presence of Na^{+} (Supplementary Figure S7B and C). Further studies are necessary to isolate the factors that favor the binding of one or the other ion to this location.

Besides ribosomal structures, a binding pattern involving a single phosphate group with $d(Mg^{2+}\ldots N7/OP) \leq 2.4$ Å was reported in only two instances. In a c-di-GMP riboswitch (PDB code: 3Q3Z; resolution: 2.51 Å), the bound OP atom is in an equatorial (*cis*) position with respect to the N7 atom (Supplementary Figure S8A) while it is opposite (*trans*) to the N7 atom in a DNA quadruplex (Figure 7B and Supplementary Figure S8B). Interestingly, these patterns involve crystal contacts that are part of the ion coordination shell. Lastly, in a CRISPR-Cas RNA complex (Figure 7C), a single hexacoordinated Mg^{2+} bound to a Glu carboxylate ligand in *trans* has been identified (94). This ion is found in a tight binding pocket at the RNA/protein interface. As shown elsewhere, one of the Mg^{2+} first shell water forms a hydrogen bond with an oxygen of the carboxylate group (95).

The ion placement should be checked carefully when $d(Mg^{2+}\ldots OP) > d(Mg^{2+}\ldots N7)$. For example, see Sup-

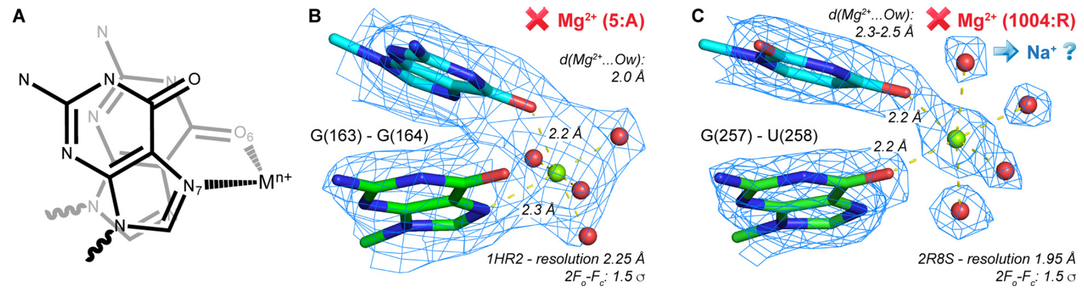


Figure 6. M^{n+} bound to N7/O6 atoms of RpG steps. (**A**) Schematical representation of this ion binding pattern. (**B**) Mg^{2+} binding as reported in a group I intron structure. (**C**) Probable Na^+ binding observed in a group I intron structure of slightly better resolution.

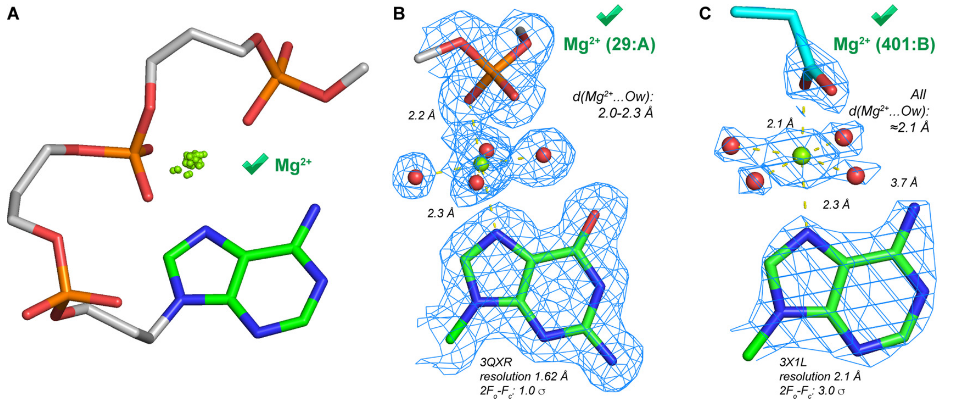


Figure 7. Mg^{2+} bound to N7 and anionic oxygens. (**A**) Overlap of 51 Mg^{2+} found in helix 11 of large ribosomal subunits with $d(Mg^{2+}\ldots N7/OP) \leq 2.4$ Å. Loop configuration is taken from a *H. marismortui* structure (PDB code: 4V9F; resolution 2.4 Å). All structures were superimposed on the adenine base. (**B**) Mg^{2+} bound to a (G)N7 and a phosphate group in B-DNA (crystal contact); separate density peaks for ion and water allow for more reliable ion identification. (**C**) Mg^{2+} bound to a (G)N7 and a glutamate carboxyl group in a RNA/protein complex.

plementary Figure S8C where $d(Mg^{2+}\ldots N7) \approx 2.3$ Å and $d(Mg^{2+}\ldots OP) \approx 2.4$ Å (PDB code: 462D; resolution: 2.3 Å). Here, the coordination distances strongly suggest the presence of Na^+. Note that the distances to phosphate oxygens (when not restrained) are much more reliable and accurate than those to nitrogens. Yet, these examples of simultaneous binding to N7 and anionic oxygens remain exceptional.

Suspicious Mg^{2+} binding occurrences in the 2.6-3.2 Å range: Na^+, K^+, NH_4^+ or water?

As stressed by Table 1 and Figure 2, Mg^{2+} are often placed in the 2.4–3.2 Å exclusion zone that corresponds to the coordination distance range for $Na^+/K^+/NH_4^+$ and water. Na^+ ions were clearly identified in several structures with resolution ≤ 2.0 Å. For example, in a hammerhead ribozyme (PDB code: 3ZP8; resolution: 1.55 Å) (67), two out of sixteen Na^+ are bound to (G)N7 and one is bound to (A)N7. The associated octahedral coordination patterns are similar to those for Mg^{2+} in Z-DNA (Figure 4D) with, however, $d(Na^+\ldots N7) \approx 2.4$–$2.6$ Å. These Na^+ are associated with density patterns showing clearly identifiable metal-bound water molecules. One of these residues (G10; Figure 8A) is often linked to the hammerhead ribozyme catalytic mechanism (32,96). This residue is also associated with direct Mn^{2+} binding but was never unambiguously shown to be in direct contact with Mg^{2+}. Further, 12 examples of Na^+ to N7 contacts, where Na^+ displays an octahedral coordination, are found in 8 structures with resolutions ≤ 2.0 Å (PDB codes: 2R1S, 2R20, 2R21, 2R22, 3ND4, 3ZP8, 3DIL, 3PNC). Interestingly, besides Z-DNA structures, no Mg^{2+} to N7 contacts with separate water densities were identified. It is possible that the binding of these hydrated Na^+ is induced by the crystallization buffers since a 1.7 M sodium malonate or NaCl buffer were used to crystallize a hammerhead ribozyme and a *H. marismortui* large ribosomal subunit, respectively (PDB codes: 3ZP8, 1S72; resolutions: 1.55, 2.4 Å). On the contrary, the authors of a lysine riboswitch structure containing 29 well-resolved Na^+ ions (PDB code: 3DIL; resolution: 1.9 Å) mentioned the use of a ≈0.1 M sodium citrate buffer (97). These data shake the common idea that Na^+ octahedral coordination is difficult to observe due to a weaker stability of its hydration shell compared to hydrated Mg^{2+}. Such a belief might have led to misidentifications in *H. marismortui* where octahedral densities with Na^+ coordination distances in the 2.4-2.6 Å range were attributed to Mg^{2+} and where Na^+ labels were used for species with coordination distances in the 2.8–3.2 Å range that are more typical for K^+ ions (14). Indeed, the octahedral Mg^{2+}/Na^+ coordination geometries are difficult to distinguish when the refinement protocols involve distance restraints (Figure 8B).

Next to Na^+, binding of K^+ to nucleic acid N7 atoms is rarely observed. We identified only 79 instances with $d(K^+\ldots N7)$ in the 2.6–3.2 Å range; 10 of those are found in structures with resolution ≤ 2.0 Å (PDB codes: 5EW4,

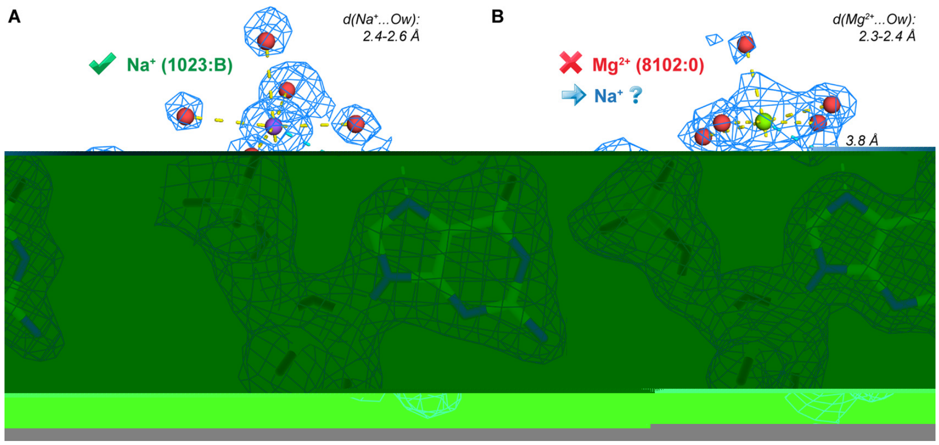


Figure 8. Na$^+$ coordination. (**A**) Pentahydrated Na$^+$ bound to (G10)N7 in a hammerhead ribozyme structure. (**B**) Based on coordination distances in the 2.3–2.5 Å range, this pentahydrated Mg^{2+} is probably a Na$^+$. Note that some water molecules display isolated density blobs even at a 2.4 Å resolution.

1HQ1, 5EW7, 1DUL, 4WO2, 4CN5, 4YAZ). The detection of such ions is complicated by their weak binding affinity and a less nicely defined non-octahedral coordination shell involving preferentially eight ligands with coordination distances similar to those of water molecules (≈2.8 Å). Therefore, K$^+$ is difficult to distinguish from water especially in case of mixed water/K$^+$ occupancy. When K$^+$ is mentioned in the crystallization conditions, anomalous diffraction experiments should systematically be conducted to detect its presence (36,56,98,99).

NH_4^+ ions are common in crystallization buffers due to the recurrent use of $(NH_4)_2SO_4$. Hence, many water molecules close to nucleotides could correspond to hidden NH_4^+ ions and this hypothesis should be seriously considered (14,50). When it occurs, binding of NH_4^+ resembles that of water although, instead of being surrounded by two donor and two acceptor atoms, these ions should be surrounded by four acceptor atoms. Such differences are very subtle and NH_4^+ was assigned in only four structures at resolutions ≤ 2.5 Å (100).

Mg^{2+} replacing co-solvent molecules like polyamines: can this happen? When resolution is insufficient and/or data treatment inappropriate, co-solvent molecules like polyamines might remain hidden. In a refinement using low temperature data, isolated 'water' peaks converted into a 'tube' of electron density and resulted in the correct placement of a spermine molecule. It was inferred that at room temperature, the methylene groups were thermally disordered while the more ordered amino groups, which are stabilized through direct hydrogen bonds, appeared as spheres of electron density (35).

In a fragment of a human signal recognition particle (101), we identified several Mg^{2+} at 2.8 Å from N7 atoms (Figure 9A) and none in the appropriate 2.1–2.4 Å coordination range. A closer examination of the $2F_o-F_c$ maps revealed a tube of density that could be interpreted as resulting from the presence of a polyamine. Tentatively, we placed a spermine molecule into this density and suggest that this model, supported by the presence of spermine in the crystallization buffer, constitutes a reasonable working hypothesis (Figure 9B). In a combined x-ray/neutron Z-DNA diffraction structure in complex with a spermine molecule (PDB codes: 1WOE, 1V9G; resolutions: 1.5, 1.8 Å), ammonium

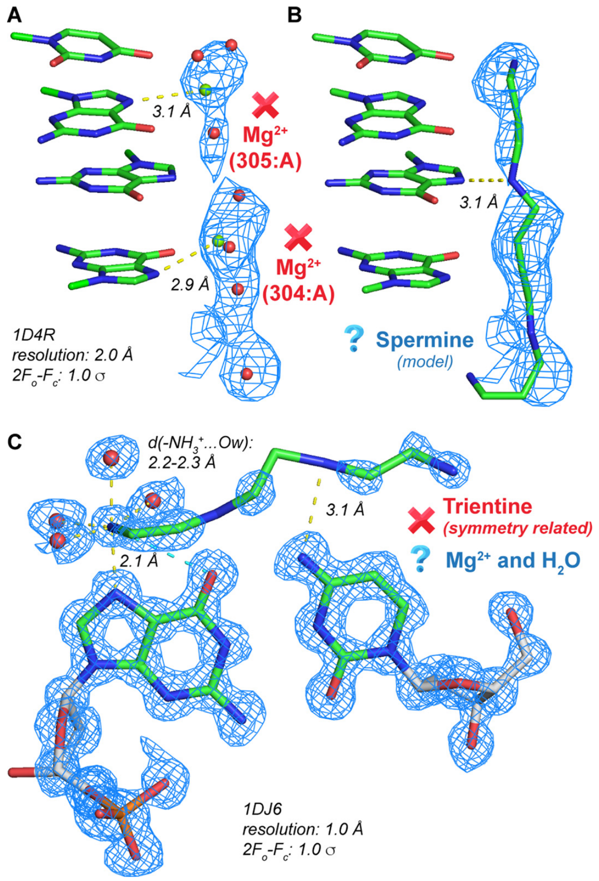


Figure 9. Polyamine misattributions. (**A**) Mg^{2+} ions with inapropriate coordination distances are close to N7 atoms in a human SRP helix 6 structure. (**B**) A spermine molecule—spermine is mentioned in the crystallization conditions—has tentativelly been fitted into the electron density in place of the original Mg^{2+} and water molecules. (**C**) A misplaced symmetry related polyamine lined up on the major groove of a Z-DNA G=C pair. Note the coordination pattern of the hydrated –NH_3^+ head that fits a pentahydrated Mg^{2+} (see Figure 4D).

groups are at hydrogen bond distance to both guanine N7 and phosphate groups (78) (see also PDB code: 4HIG, 2F8W; resolutions: 0.8, 1.2 Å) indicating that N7 sites are good docking spots for ammonium groups.

We traced also the opposite type of misidentification, namely a polyamine positioned at Mg^{2+} binding sites. In a Z-DNA hexamer (102), a symmetry related polyamine is unusually lined up on the major groove of a terminal G=C pair. Here, a $–NH_3^+$ group is at 2.1 Å from a N7 atom (Figure 9C) and the coordination is similar to that shown Figure 4D. Further, the Mg^{2+} ion placed in this structure is at 2.6 Å from the closest oxygen and its coordination shell is not octahedral. This observation stresses that odd solvent density interpretations occur even in high-resolution structures.

Mixed Mg^{2+} and monovalent cation/water occupancies: are they meaningful? It has been reported that some DNA major groove hexahydrated Mg^{2+} binding sites are not fully occupied but that a monovalent cation can partially occupy such a site (35). The latter event was identified through anomalous diffraction experiments involving Tl^+ ions (103). Furthermore, K^+ or NH_4^+ could overlap with inner-sphere water molecules of a hexacoordinated Mg^{2+} (104,105). However, it is less likely that water overlays with Mg^{2+} in direct contact to a N7 atom. Yet, this was reported in a Z-DNA structure (PDB code: 1ICK; resolution: 0.95 Å) where a 0.24 occupancy water and a 0.76 occupancy Mg^{2+} share the same position (106). This site is similar to the Z-DNA Mg^{2+} binding site described for Z-DNA (Figure 4D) and illustrates the outcomes of unusual protocols employed to satisfy crystallographic constraints. Although such quirks are rare, they are present in high-resolution crystal structures as mentioned above, a fact that should not be ignored when surveying structural databases.

Direct Mg^{2+} binding to imine N1/N3 nitrogens is unlikely

Mg^{2+} binding to N1 and N3 sites is observed in only nine non-redundant instances and appears at best marginal (Table 1). None of these Mg^{2+} binding occurrences passes visual scrutiny. All Mg^{2+} have one or more contacts in the 2.5–3.0 Å exclusion zone. Titration experiments suggest that Mg^{2+} and Ca^{2+} close to these imine atoms form outer sphere complexes (107). Hence, these inner-sphere contacts represent obvious misattributions.

For instance, an ill-placed Mg^{2+} ion in front of an (A)N1 atom is found in a *T. thermophilus* ribosome structure (108) (see the density extension on the adenine Watson–Crick edge; Figure 10A). Here, the 2.9 Å distance to an amino group of an arginine amino acid suggests the binding of a water molecule. Similarly, in a thiamine riboswitch, a direct binding of Mg^{2+} to N3 with a tetrahedral coordination and long Mg^{2+} to ligand distances suggests the presence of a water molecule rather than an ion (Figure 10B).

Mg^{2+} do not bind to the N7 atom of purine containing metabolites

We checked if Mg^{2+} to N7 binding could be associated with purine containing metabolites, such as ATP, by using the Relibase+ program to search the PDB (109). In the ≤ 3.0 Å resolution range, close to 25 000 such metabolites were identified but only four binding sites with $d(Mg^{2+}\ldots N7)$ in the 3.0–3.5 Å range, and none with $d(Mg^{2+}\ldots N7) \leq 3.0$ Å. In the best resolution structure (PDB code: 4H2U; resolution: 2.1 Å), Mg^{2+} is at a 2.8–3.0 Å distance from all its ligands including a phosphate oxygen and a positively charged Arg side chain. This unambiguous result strongly illustrates the poor Mg^{2+} binding potential of N7 atoms.

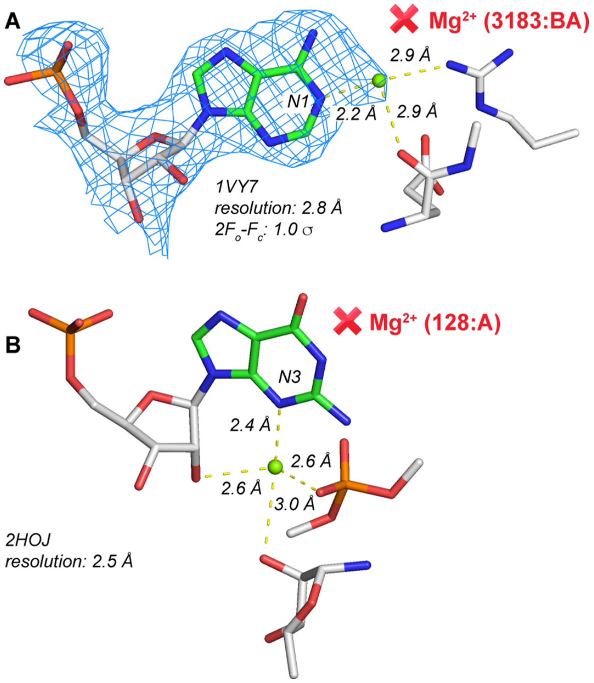


Figure 10. Missassigned Mg^{2+} ions close to imine N1/N3 atoms. (**A**) This figure illustrates the pitfalls of placing ions into poorly defined density patterns. See, for example, the unrealistic Mg^{2+} to arginine contact. (**B**) The tetrahedral coordination inferred from the solvent density at the N3 site and $d(Mg^{2+}\ldots N/O)$ in the 2.4–3.0 Å range suggest the presence of a water molecule and excludes that of Mg^{2+}.

Resolutions > 3.0 Å

If serious identification issues arise at resolutions better that 3.0 Å (≤ 3.0 Å), such issues are certainly much more severe at lower resolutions. In that respect, it is important to note that the PDB contains a significant number of structures with resolutions lower than 3.0 Å. Many of these structures comprise Mg^{2+}, Na^+ and even NH_4^+ ions (110). For instance, we counted in this resolution range 84 nucleic acid structures containing Na^+ (including 12 ribosomes) while 43 Mg^{2+} and 2 K^+ containing structures at resolutions lower than 4.0 Å as well as 63 cryo-EM structures containing Mg^{2+} with resolutions > 3.0 Å were deposited to the PDB. Although crystallography is making significant progress, we believe that assigning light mono- and divalent ions at such resolutions can be detrimental to the crystallographic process and problematic in the development of data mining tools since misinterpretation odds are too high (43,52). In this resolution range (> 3.0 Å), Na^+, Mg^{2+} and their hydration shell are essentially modeled (see below) and, consequently, should be excluded from database surveys.

About the use of coordination distance restraints and modeled hydration spheres

As noted above, the default refinement procedures often involve the use of restraints to place water molecules coordinated to Mg^{2+}. Therefore, their positions are approximated or sometimes entirely modeled. The most obvious example comes from water with $d(Mg^{2+}\ldots Ow) = 2.18$ Å. Such water molecules represent a large part of those that are bound to Mg^{2+} in the PDB (Figure 11). The use of restraints might help to position properly the octahedral coordination shell. However, it has several important drawbacks. The first is that this coordination distance is not appropriate for Mg^{2+} since it is intermediate between the 2.07 Å coordination distance expected for $d(Mg^{2+}\ldots Ow)$ and the ≈2.40 Å coordination distance expected for $d(Na^{+}\ldots Ow)$. Thus, the use of restraints to model the Mg^{2+} hydration shell might make impossible the unambiguous assignment of the electron density peak to Mg^{2+} or Na^{+}. This is especially true when water and ion densities are merged. In those cases, since we identified well defined coordination shells for Mg^{2+} and Na^{+} in structures with resolution ≤ 2.0 Å, the possibility of Na^{+} coordination should at least be considered during the refinement process.

The use and implications of crystallographic restraints have already been noted elsewhere as well as the less frequent but more appropriate use of ≈2.1 Å restraints (12). Finally, we note that restraints are mainly used for Mg^{2+} and rarely for other ions such as Na^{+} and Mn^{2+}, as deduced from the $d(Na^{+}/Mn^{2+}\ldots Ow)$ histograms (Figure 11 and Supplementary Figure S9). For these ions and at least in nucleic acid structures, restraints do not seem necessary.

Because of the issues mentioned here, it appears worthwhile to tag modeled water molecules associated with the systematic use of restraints especially at resolutions > 3.0 Å. Occupancies could be set to zero as is already done by some authors for polyatomic ligands. Specific identifiers could be added to the more accommodating mmCIF files. But, as noted elsewhere, non-crystallographers visualizing a biomolecular system might not be aware of the presence of modeled waters (40). Therefore, we suggest that visualization programs should include an option to turn 'on' the modeled part of the structure that should remain hidden when the structure is first opened. Turning on the visualization of the modeled part of the structure should require a voluntary action. Lastly, the use of restraints should be systematically mentioned in PDB headers and validation reports.

Ion substitution experiments

To identify Na^{+} or Mg^{2+} when the resolution is insufficient, replacement strategies are used (14,111,112). However, they do not supersede direct evidence obtained from high-resolution structures. This is especially true when Mn^{2+} ions are used as substitutes since the affinity of Mn^{2+} over Mg^{2+} for N7 is higher (20,25). Further, although rarely described, Mn^{2+}/Mg^{2+} substitutions can induce significant structural changes; see for instance a crystallographic study of a signal recognition particle (104). There, Mn^{2+} changed the conformation of a nucleotide by linking the N7 to a phosphate oxygen from a neighboring residue. For another RNA structure for which soaking with 13 different metals was performed, a similar conformational change induced by the binding of Mn^{2+}, Zn^{2+} and Co^{2+} with respect to the native Mg^{2+} structure was reported, resulting in a direct N7 to M^{n+} contact (113). Indeed, Mn^{2+} are not perfect substitutes for Mg^{2+} and replacement of $MgCl_2$ by $MnCl_2$ in *H. marismortui* crystallization buffers resulted systematically into twinned crystals (14).

Such ion-induced conformational changes might be more frequent than expected in crystallographic structures with insufficient resolution or in spectroscopic experiments such as electron paramagnetic resonance (EPR) and NMR (8). Drawbacks of substitution experiments might even become worse when larger transition metals like Zn^{2+} or Cd^{2+} are used. Indeed, strong binding of transition metals to N7 sites significantly affects the nucleobase chemical properties (21). It has been shown that Cd^{2+} binding to (G)N7 leads to an acidification of the N1 imino group that can consequently deprotonate at physiological pH and affect the interpretation of biochemical experiments (24).

Thus, in the best resolution structure of $tRNA^{Phe}$ (PDB code: 1EHZ; resolution: 1.93 Å), anomalous data derived from crystals soaked with $MnCl_2$ and $CoCl_2$ were used to identify ion binding sites. In this structure, all four sites close to N7 atoms were associated with Mn^{2+} or Co^{2+} and unrealistically short 2.0 Å metal-water restraints were used. In the parent $tRNA^{Phe}$ structure that was obtained without the use of soaking procedures (PDB code: 1EVV; resolution: 2.0 Å), no N7-bound ions were reported (114).

Further, numerous soaking experiments were performed on the hammerhead ribozyme showing consistently the presence of a transition metal bound to (G10)N7 and, in a high-resolution structure, the presence of Na^{+} (32,96). Despite significant efforts, no direct evidence of Mg^{2+} binding to this site has been reported (Figure 8A).

Hence, binding sites presenting new coordination topologies should not be proposed based uniquely on substitution experiments combined or not with anomalous diffraction data unless a similar binding site backing up the proposed topology has unambiguously been identified in unrelated high-resolution structures (115–117). Given the affinity of Mn^{2+} for N7, this transition metal could eventually replace Mg^{2+} but more probably also Na^{+} commending great care in the interpretation of ion substitution experiments. This is not purely speculative since we are aware of at least one example of substitution of Na^{+} by Mn^{2+} in a protein crystal structure (118).

Comparison with the MgRNA database

The abundance of poorly modeled or incorrectly identified Mg^{2+} ions in nucleic acids has already been noted and was taken into account in an attempt to build an exhaustive and comprehensive classification of Mg^{2+} binding sites, including 41 inner-sphere coordination patterns among which eight are associated with nitrogen sites (12). The MgRNA database defines a set of rules to separate good from bad ion assignments (52). Based on a complex combination of geometrical and crystallographic criteria derived from those used by the CheckMyMetal web server (37), benchmark sets for each binding pattern were defined, embracing 15% of the

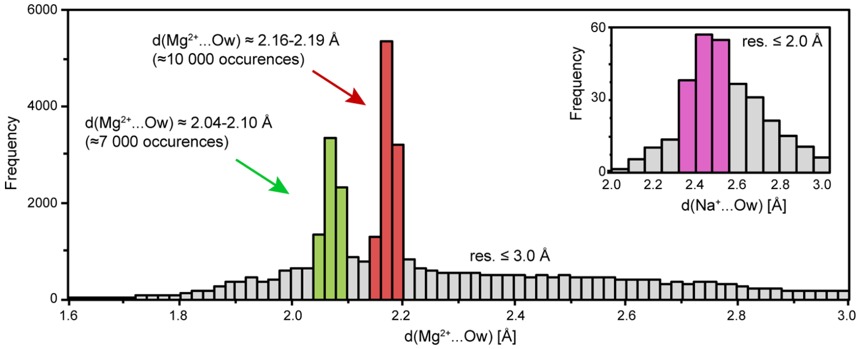


Figure 11. d(Mg^{2+}...Ow) histogram for nucleic acid crystal structures (PDB; May 2016; resolution ≤ 3.0 Å) that emphasize the systematic use of crystallographic restraints around 2.07 and 2.18 Å. The d(Na^+...Ow) histogram peaks around 2.4 Å and no peaks associated with crystallographic restraints are apparent (see insert; resolution ≤ 2.0 Å). Supplementary Figure S9 displays d(Mn^{2+}/Na^+...Ow) histograms where peaks due to the use of restraints are also absent (resolution ≤ 3.0 Å).

full dataset—for details, see (12). However, we believe that these criteria were not restrictive enough to exclude all dubious coordination patterns, resulting in a still considerable overestimation of the Mg^{2+} to N7 binding.

In this section, we identify shortcomings in the criteria defined by MgRNA that make further investigations necessary and describe methods to improve them. We have to stress that the numbers provided by MgRNA are not directly comparable to ours since we use resolutions ≤ 3.0 Å on a May 2016 dataset while no resolution limits are applied in MgRNA on the September 2014 dataset. Here, we adopt the MgRNA nomenclature where O_{ph} corresponds to phosphate oxygens (OP1/OP2); O_r to O2'/O4'/O3'/O5' ribose oxygens; O_b to nucleobase oxygens and N_b to nucleobase nitrogens.

In MgRNA, 284 Mg^{2+} to N1/N3/N7 (N_b) contacts were identified and placed in the benchmark dataset (Figure 4). The authors chose a representative of this category in the 2QOU ribosome (resolution: 3.93 Å) that shows a perfectly modeled hydration shell with d(Mg^{2+}...N7) = 2.18 Å and d(Mg^{2+}...Ow) = 2.08 Å, raising once more the issue of modeled water molecules (see above). Cleary, such a structure including a modeled hydration shell is not representative. Overall, the N_b benchmark dataset contains 77 structures with resolution ≤ 3.0 Å and 92 structures with resolutions > 3.0 Å. The lowest resolution structure is 4V5Y (resolution: 4.45 Å). Among the 77 structures with resolution ≤ 3.0 Å, 127 Mg^{2+} were identified but only 12 of them satisfy the d(Mg^{2+}...N7) ≤ 2.4 Å criterion. The largest d(Mg^{2+}...N7) is 3.17 Å (PDB code: 4PEA; resolution: 2.95 Å) and the average d(Mg^{2+}...N7) of 2.64±0.40 Å is too long for Mg^{2+} to N7 contacts. Based on these data, only 12 out of the 284 sites satisfy our criteria although their coordination shell is far from being strictly octahedral.

For the double N7 site discussed above ($2N_b$), called 'purine N7-seat' in MgRNA (Figure 5), 158 occurrences constitute the benchmark set. The representative site is extracted from an *E. coli* ribosome structure (PDB code: 2I2V; resolution: 3.22 Å), that used d(Mg^{2+}...N7) ≈2.08 Å restraints. Monovalent cation or transition metal binding was not considered. This double N7 binding site was tagged as a novel Mg^{2+}-binding motif, although it has been mentioned elsewhere (8,14).

We already addressed the direct binding of an ion to N7 and a nucleobase oxygen atom ($O_b.N_b$) (Figure 6). The representative site was taken from 1HR2 (resolution: 2.25 Å). To us, it is unlikely that Mg^{2+} binds to this site, which is most probably involved in the binding of a monovalent cation, possibly Na^+. Further, only 26 instances are found in the MgRNA benchmark dataset stressing its limited relevance.

Occurrence of simultaneous binding to N7 and two base or sugar atoms were categorized into three binding types ($2O_b.N_b$, $O_b.2N_b$, $2O_r.N_b$) derived from only five crystal structures with resolution in the 3.3-3.9 Å range. No structure with sufficient resolution to interpret solvent binding details is available to support the genuineness of these modeled sites. Therefore, they should not be labeled as Mg^{2+} binding sites.

The representative $O_{ph}.N_b$ site (PDB code: 3R8S; resolution: 3.0 Å) displays good coordination distances to water, phosphate oxygen and N7 atoms. However, a closer examination revealed that all the $Mg(H_2O)_4{}^{2+}$ but also all the neighboring nucleotide *B*-factors display an unrealistic 0.01 $Å^2$ value. In 1VQ8 (resolution: 2.2 Å), d(Mg^{2+}...N7) is stretched to 3.12 Å. Indeed, all these occurrences in the benchmark dataset have not been identified by us mainly because of inappropriate resolution and/or d(Mg^{2+}...N7) > 2.4 Å. Thus, this Mg^{2+}-binding motif should be excluded from the MgRNA classification.

The benchmark set for the last MgRNA site, labelled *cis*-$2O_{ph}.N_b$, comprises 118 occurrences. The representative site is extracted from the 1VS6 ribosome structure with a 3.46 Å resolution and d(Mg^{2+}...Ow) = 2.08 Å. Overall, for this dataset, the level of redundancy is high. The 118 sites, exclusively identified in ribosomes, are found at two locations involving (A)N7 or (G)N7. We identified 49 out of 118 sites in structures with resolution ≤ 3.0 Å. Among those with (A)N7, 18 have d(Mg^{2+}...N7) ≤ 2.4 Å and 15 have d(Mg^{2+}...N7) > 2.4 Å. The remaining 16 occurrences involve (G)N7 with d(Mg^{2+}...N7) > 2.4 Å. One of these MgRNA sites, called '10-member ring with purine N7', has been identified in the present study (Figure 7). Again, the significance of this site is low and there is not enough evidence to suggest its presence other than in rare and highly specific ribosomal pockets. This site can probably also accommodate Na^+ ions (Supplementary Figure S7).

Consequently, MgRNA still contains a large number of misidentified Mg^{2+} and was not successful in creating reliable benchmark datasets. We identified several factors that led to such issues. First, the process could be improved if structures with resolutions > 3.0 Å were excluded (38,40). Second, strict enforcement of $d(Mg^{2+}\ldots N/O)$ cutoffs would lead to a significant reduction of false positive. Although uncertainties in the coordination distances are difficult to estimate (23,37), it seems problematic to accept distances >2.4 Å to validate new Mg^{2+} binding sites. Third, coordination distance issues involving restrained water molecules have to be identified more systematically (12,23,37). In that respect, indicators based on the bond valence theory should be considered with caution for resolutions > 2.0 Å (119). Such indicators should not be used when restraints on coordination bonds are present. As an outcome, when stricter criteria are used, the eight binding sites described by MgRNA, reduce to two for which we found a limited number of convincing occurrences, namely N_b and *cis*-$2O_{ph}.N_b$.

Mg^{2+} ion assignment and validation checklist

In order to facilitate the ion assignment process, we defined a set of rules regarding the placement of ions in solvent electron densities next to N7 atoms that can easily be extended to the binding of ions to other sites (Table 3). In that perspective, we would like to stress a few points that we consider of importance. First, numerous competing ionic species might be present in crystallization buffers, sometimes as contaminants and should be taken into account (70). For instance, in our survey, it was not immediately apparent that Zn^{2+} could bind to a specific ribosome site (Figure 5) especially since Zn^{2+} is not mentioned in the crystallographic conditions. Therefore, it is important to integrate excess electron density that can reveal the presence of transition metals or electron rich K^+. Second, an ion *B*-factor lower than those of the bound nucleobase or water molecules or an ion occupancy significantly higher than 1.0 should hint to the presence of an electron rich atom. In such instances, anomalous diffraction data should be collected at the appropriate wavelengths. It has to be noted that a large excess of unassigned electron density might affect not only the position of the ion and its hydration shell, but can also wrongfully force nucleobases to come closer to the excess electron density center leading to unreliable coordination distances (43,117). On the other hand, weak electron density patterns manifested by high *B*-factors or negative F_o-F_c peaks suggest 'wishful' ion attributions. Anions such as Cl^-, $SO_4{}^{2-}$ or even cacodylate are also often disregarded (68,69). If the identity of an ion is inferred from binding sites observed in a different structure, the original data should be carefully checked including the electron density peak height, *B-factor* value, coordination number, bond distances and angles as well as the $2F_o-F_c$ and F_o-F_c maps in order to avoid replicating errors. Finally, when no reasonable solution emerges, protonated and tautomeric forms of the coordinated nucleobase or the surrounding residues should be considered (120).

In case of doubt, and especially in the $d(Mg^{2+}\ldots N7)$ $\approx$3.2–3.8 Å exclusion range (Figures 1 and 2), density patterns should not be assigned to Mg^{2+}. Such patterns are probably related to the presence of other ionic or molecular species present in the crystallization buffer or as contaminants. X-ray data are also prone to experimental errors that might result in weak/spurious electron density peaks (121,122). In those instances, density assignment is counterproductive even if it reduces R_{work} and R_{free} values. It can here be reminded that the PDB allows to use the UNK code for placing an atom at positions where atom identity is uncertain.

We have to stress that the chosen cutoff distances are merely indicative. They are less stringent for Mg^2 binding to nitrogen than to oxygen. These distances will be refined in further studies when more high-quality data become available. A rule of thumb is that, at least in nucleic acids, $d(Mg^{2+}\ldots N7) \leq 2.4$ Å and $d(Mg^{2+}\ldots N7) > d(Mg^{2+}\ldots O)$. Caution should be exerted when restraints are used, especially with the 2.18 Å default value, suggesting that the data do not allow to differentiate Mg^{2+} from Na^+ and that further refinement without restraints should be conducted.

Further reasons can lead to bad ion assignments among which we list: (i) the possibility that ions were placed automatically or without great care into density blobs in order to lower the R_{free} value; (ii) existing stereochemical knowledge was ignored; (iii) wishful thinking; (iv) the replication of errors already present in PDB structures and (v) overestimation of the amount of information that can be extracted from low-resolution structures. Hence, it is suggested to exclude structures displaying obvious ion identification errors from database surveys, at least as far as ion placement is concerned (40).

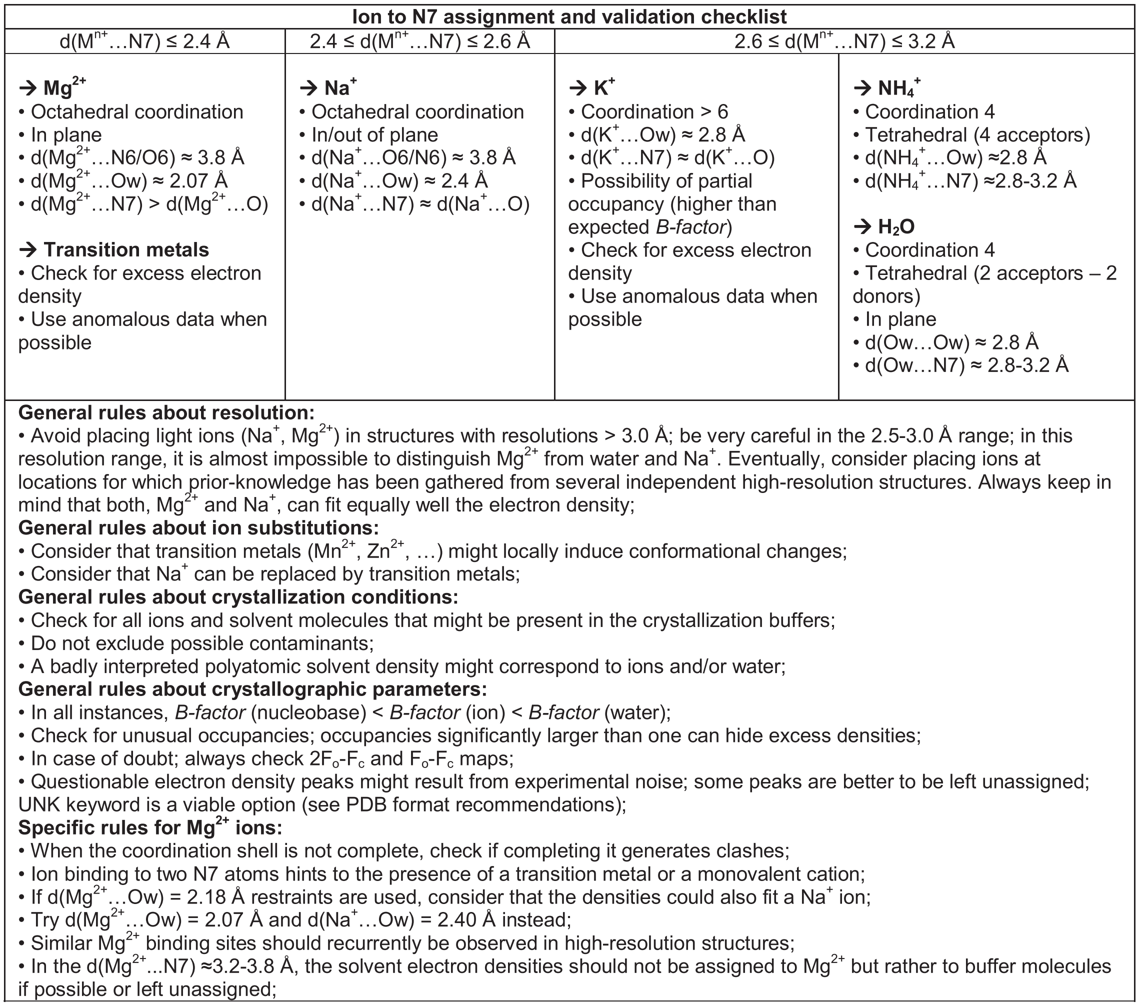
CONCLUSION AND PERSPECTIVES

Based on the data we gathered, we conclude that nearly all the Mg^{2+} to N7 contacts reported in PDB structures need to be reexamined and propose a Mg^{2+} assignment checklist to facilitate this endeavor. Indeed, non-ambiguous examples of Mg^{2+} binding to N7 are excessively rare and are limited to a few occurrences where Mg^{2+} binding seems to result from unique crystallization conditions and/or is associated with primary contacts with anionic oxygens. Additionally, we noted that none of the 25 000 purine metabolites from the PDB establish Mg^{2+} to N7 contacts, an additional strong evidence that Mg^{2+} does rarely bind to purine N7 sites. Consequently, we conclude that almost all Mg^{2+} assignments to solvent density in front of N7 atoms, as found in PDB structures, are incorrect. This outcome significantly diverges from that presented by the MgRNA survey that identified 8 binding modes involving imine nitrogens in opposition to barely two by us (12).

Interestingly, we characterized a potential Zn^{2+} binding site in prokaryotic ribosomal structures that involves two head-to-tail stacked purines in the core of a three-way junction, a finding that opens a new window on the complexity of the metal/nucleic acid ecosystem (123,124). However, we were unable to establish if these transition metal binding sites are populated *in vivo* or if they may only be found under specific *in crystallo* conditions.

From a purely methodological point of view, the most interesting outcome of this study resides in the recognition of

Table 3. Ion to N7 assignment and validation checklist



Ion to N7 assignment and validation checklist			
$d(M^{n+}…N7) \leq 2.4$ Å	$2.4 \leq d(M^{n+}…N7) \leq 2.6$ Å	$2.6 \leq d(M^{n+}…N7) \leq 3.2$ Å	
→ Mg^{2+} • Octahedral coordination • In plane • $d(Mg^{2+}…N6/O6) \approx 3.8$ Å • $d(Mg^{2+}…Ow) \approx 2.07$ Å • $d(Mg^{2+}…N7) > d(Mg^{2+}…O)$ **→ Transition metals** • Check for excess electron density • Use anomalous data when possible	**→ Na^+** • Octahedral coordination • In/out of plane • $d(Na^+…O6/N6) \approx 3.8$ Å • $d(Na^+…Ow) \approx 2.4$ Å • $d(Na^+…N7) \approx d(Na^+…O)$	**→ K^+** • Coordination > 6 • $d(K^+…Ow) \approx 2.8$ Å • $d(K^+…N7) \approx d(K^+…O)$ • Possibility of partial occupancy (higher than expected *B-factor*) • Check for excess electron density • Use anomalous data when possible	**→ NH_4^+** • Coordination 4 • Tetrahedral (4 acceptors) • $d(NH_4^+…Ow) \approx 2.8$ Å • $d(NH_4^+…N7) \approx 2.8\text{-}3.2$ Å **→ H_2O** • Coordination 4 • Tetrahedral (2 acceptors – 2 donors) • In plane • $d(Ow…Ow) \approx 2.8$ Å • $d(Ow…N7) \approx 2.8\text{-}3.2$ Å
General rules about resolution: • Avoid placing light ions (Na^+, Mg^{2+}) in structures with resolutions > 3.0 Å; be very careful in the 2.5-3.0 Å range; in this resolution range, it is almost impossible to distinguish Mg^{2+} from water and Na^+. Eventually, consider placing ions at locations for which prior-knowledge has been gathered from several independent high-resolution structures. Always keep in mind that both, Mg^{2+} and Na^+, can fit equally well the electron density; **General rules about ion substitutions:** • Consider that transition metals (Mn^{2+}, Zn^{2+}, …) might locally induce conformational changes; • Consider that Na^+ can be replaced by transition metals; **General rules about crystallization conditions:** • Check for all ions and solvent molecules that might be present in the crystallization buffers; • Do not exclude possible contaminants; • A badly interpreted polyatomic solvent density might correspond to ions and/or water; **General rules about crystallographic parameters:** • In all instances, *B-factor* (nucleobase) < *B-factor* (ion) < *B-factor* (water); • Check for unusual occupancies; occupancies significantly larger than one can hide excess densities; • In case of doubt; always check $2F_o\text{-}F_c$ and $F_o\text{-}F_c$ maps; • Questionable electron density peaks might result from experimental noise; some peaks are better to be left unassigned; UNK keyword is a viable option (see PDB format recommendations); **Specific rules for Mg^{2+} ions:** • When the coordination shell is not complete, check if completing it generates clashes; • Ion binding to two N7 atoms hints to the presence of a transition metal or a monovalent cation; • If $d(Mg^{2+}…Ow) = 2.18$ Å restraints are used, consider that the densities could also fit a Na^+ ion; • Try $d(Mg^{2+}…Ow) = 2.07$ Å and $d(Na^+…Ow) = 2.40$ Å instead; • Similar Mg^{2+} binding sites should recurrently be observed in high-resolution structures; • In the $d(Mg^{2+}...N7) \approx 3.2\text{-}3.8$ Å, the solvent electron densities should not be assigned to Mg^{2+} but rather to buffer molecules if possible or left unassigned;			

specific and frequent octahedral Na^+ coordination modes identified in structures with resolution ≤ 2.0 Å. Those are very difficult to distinguish from octahedral Mg^{2+} coordination modes, especially when 2.18 Å coordination distance restraints are used during the refinement process. Undeniably, such restraints often combined with poor resolutions do not allow to distinguish Na^+ from Mg^{2+} based on their respective 2.40 and 2.07 Å coordination distances. We suggest that the use of restraints on Mg^{2+} coordination distances is probably at the origin of the large number of Mg^{2+} misidentifications in nucleic acids and that Na^+ binding should always be considered as an alternative.

To conclude, we strongly believe that careful visual examination of crystallographic data is needed in order to create a reliable 'prior knowledge' dataset before developing or using automatic assignment protocols, pattern detection algorithms or machine learning tools (12,47). Further, 'prior-knowledge' should only be based on Mg^{2+} binding motifs unambiguously characterized in multiple unrelated high-resolution structures and not on circumstantial evidences as it is too often the case. Currently, automatic re-refinement workflows such as PDB-REDO cannot resolve solvent attribution issues that remain one of the last major bottlenecks in the interpretation of crystallographic data (37,44,46,48,49). This work should provide a more solid experimental ground for the development of molecular dynamics force-fields that sometimes rely on the erroneous assumption that N7 is an appropriate Mg^{2+} binding site (12,31).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

P.A. wishes to thank Prof. Eric Westhof for ongoing support and helpful discussions, as well as Prof. Neocles Leontis, Prof. Richard Giegé, Dr Eric Ennifar and Dr Quentin Vicens for careful reading of the manuscript and for beneficial discussions. The authors wish also to thank Prof. Jamie Cate and Dr Albert Weixlbaumer for constructive discussions on the difficult 'Zn^{2+}' topic as well as Prof. Alexander Serganov on more general crystallographic issues. We thank the referees for their generous comments.

FUNDING

French ‘Ministère de la recherche et de l’enseignement’ (to L.D.); Polish Ministry of Higher Education and Science (Mobility Plus programme) [1103/MOB/2013/0 to F.L.]. Funding for open access charge: French National Research Agency [ANR-15-CE11-0021-02].

Conflict of interest statement. None declared.

REFERENCES

1. Erat,M.C. and Sigel,R.K. (2011) Methods to detect and characterize metal ion binding sites in RNA. *Met. Ions Life Sci.*, **9**, 37–100.
2. Pechlaner,M. and Sigel,R.K. (2012) Characterization of metal ion-nucleic acid interactions in solution. *Met. Ions Life Sci.*, **10**, 1–42.
3. Pyle,A.M. (2002) Metal ions in the structure and function of RNA. *J. Biol. Inorg. Chem.*, **7**, 679–690.
4. Woodson,S.A. (2005) Metal ions and RNA folding: a highly charged topic with a dynamic future. *Curr. Opin. Chem. Biol.*, **9**, 104–109.
5. Draper,D.E. (2008) RNA folding: thermodynamic and molecular descriptions of the roles of ions. *Biophys. J.*, **95**, 5489–5495.
6. Auffinger,P., Grover,N. and Westhof,E. (2011) Metal ion binding to RNA. *Met. Ions Life Sci.*, **9**, 1–35.
7. Erat,M.C., Coles,J., Finazzo,C., Knobloch,B. and Sigel,R.K. (2012) Accurate analysis of Mg^{2+} binding to RNA: From classical methods to a novel iterative calculation procedure. *Coord. Chem. Rev.*, **256**, 279–288.
8. Sigel,R.K. and Sigel,H. (2013) Metal-ion interactions with nucleic acids and their constituents. In: Reedjik,J and Poeppelmeier,K (eds). *Comprehensive Inorganic Chemistry II*. Elsevier, Oxford, Vol. **3**, pp. 623–660.
9. Auffinger,P., D’Ascenzo,L. and Ennifar,E. (2016) Sodium and potassium interactions with nucleic acids. *Met. Ions Life Sci.*, **16**, 167–201.
10. Miller-Fleming,L., Olin-Sandoval,V., Campbell,K. and Ralser,M. (2015) Remaining mysteries of molecular biology: The role of polyamines in the cell. *J. Mol. Biol.*, **427**, 3389–3406.
11. Trachman,R.J. 3rd and Draper,D.E. (2013) Comparison of interactions of diamine and Mg^{2+} with RNA tertiary structures: similar versus differential effects on the stabilities of diverse RNA folds. *Biochemistry*, **52**, 5911–5919.
12. Zheng,H., Shabalin,I.G., Handing,K.B., Bujnicki,J.M. and Minor,W. (2015) Magnesium-binding architectures in RNA crystal structures: validation, binding preferences, classification and motif detection. *Nucleic Acids Res.*, **43**, 3789–3801.
13. Banatao,D.R., Altman,R.B. and Klein,T.E. (2003) Microenvironment analysis and identification of magnesium binding sites in RNA. *Nucleic Acids Res.*, **31**, 4450–4460.
14. Klein,D.J., Moore,P.B. and Steitz,T.A. (2004) The contribution of metal ions to the structural stability of the large ribosomal subunit. *RNA*, **10**, 1366–1379.
15. Stefan,L.R., Zhang,R., Levitan,A.G., Hendrix,D.K., Brenner,S.E. and Holbrook,S.R. (2006) MeRNA: a database of metal ion binding sites in RNA structures. *Nucleic Acids Res.*, **34**, D131–D134.
16. Schnabl,J., Suter,P. and Sigel,R.K. (2012) MINAS—a database of metal ions in nucleic acidS. *Nucleic Acids Res.*, **40**, D434–D438.
17. Lippert,B. (2000) Multiplicity of metal ion binding patterns to nucleobases. *Coord. Chem. Rev.*, **200-202**, 487–516.
18. Erat,M.C., Kovacs,H. and Sigel,R.K. (2010) Metal ion-N7 coordination in a ribozyme branch domain by NMR. *J. Inorg. Biochem.*, **104**, 611–613.
19. Bartova,S., Pechlaner,M., Donghi,D. and Sigel,R.K. (2016) Studying metal ion binding properties of a three-way junction RNA by heteronuclear NMR. *J. Biol. Inorg. Chem.*, **21**, 319–328.
20. Bock,C.W., Katz,A.K., Markham,G.D. and Glusker,J.P. (1999) Manganese as a replacement for magnesium and zinc: functional comparison of the divalent ions. *J. Am. Chem. Soc.*, **121**, 7360–7372.
21. Sponer,J., Sabat,M., Gorb,L., Leszczynski,J., Lippert,B. and Hobza,P. (2000) The effect of metal binding to the N7 site of purine nucleotides on their structure, energy, and involvement in base pairing. *J. Chem. Phys. B*, **104**, 7535–7544.
22. Wang,G., Gaffney,B.L. and Jones,R.A. (2004) Differential binding of Mg^{2+}, Zn^{2+}, and Cd^{2+} at two sites in a hammerhead ribozyme motif, determined by ^{15}N NMR. *J. Am. Chem. Soc.*, **126**, 8908–8909.
23. Harding,M.J., Nowicki,M.W. and Walkinshaw,M.D. (2010) Metals in protein structures: a review of their principal features. *Cryst. Rev.*, **16**, 247–302.
24. Sigel,R.K., Skilandat,M., Sigel,A., Operschall,B.P. and Sigel,H. (2013) Complex formation of cadmium with sugar residues, nucleobases, phosphates, nucleotides, and nucleic acids. *Met. Ions Life Sci.*, **11**, 191–274.
25. Leonarski,F., D’Ascenzo,L. and Auffinger,P. (2016) Binding of metals to purine N7 nitrogen atoms and implications for nucleic acids: a CSD survey. *Inorg. Chim. Acta.*, **452**, 82-89.
26. Scott,W.G., Finch,J.T. and Klug,A. (1995) The crystal structure of an all-RNA hammerhead ribozyme: a proposed mechanism for RNA catalytic cleavage. *Cell*, **81**, 991–1002.
27. Sigel,R.K. and Pyle,A.M. (2007) Alternative roles for metal ions in enzyme catalysis and the implications for ribozyme chemistry. *Chem. Rev.*, **107**, 97–113.
28. Ward,W.L., Plakos,K. and DeRose,V.J. (2014) Nucleic acid catalysis: metals, nucleobases, and other cofactors. *Chem. Rev.*, **114**, 4318–4342.
29. Chi,Y.I., Martick,M., Lares,M., Kim,R., Scott,W.G. and Kim,S.H. (2008) Capturing hammerhead ribozyme structures in action by modulating general base catalysis. *PLoS Biol.*, **6**, e234.
30. Mir,A., Chen,J., Robinson,K., Lendy,E., Goodman,J., Neau,D. and Golden,B.L. (2015) Two divalent metal ions and conformational changes play roles in the hammerhead ribozyme cleavage reaction. *Biochemistry*, **54**, 6369–6381.
31. Panteva,M.T., Giambasu,G.M. and York,D.M. (2015) Force field for Mg^{2+}, Mn^{2+}, Zn^{2+}, and Cd^{2+} ions that have balanced interactions with nucleic acids. *J. Phys. Chem. B*, **119**, 15460–15470.
32. Mir,A. and Golden,B.L. (2016) Two active site divalent ions in the crystal structure of the hammerhead ribozyme bound to a transition state analogue. *Biochemistry*, **55**, 633–636.
33. Ren,A., Vusurovic,N., Gebetsberger,J., Gao,P., Juen,M., Kreutz,C., Micura,R. and Patel,D.J. (2016) Pistol ribozyme adopts a pseudoknot fold facilitating site-specific in-line cleavage. *Nat. Chem. Biol.*, **12**, 702–708.
34. Nayal,M. and Di Cera,E. (1996) Valence Screening of water in protein crystals reveals potential Na^{+} binding sites. *J. Mol. Biol.*, **256**, 228–234.
35. Williams,L.D. (2005) Between objectivity and whim: nucleic acid structural biology. *Top. Curr. Chem.*, **253**, 77–88.
36. Echols,N., Morshed,N., Afonine,P.V., McCoy,A.J., Miller,M.D., Read,R.J., Richardson,J.S., Terwilliger,T.C. and Adams,P.D. (2014) Automated identification of elemental ions in macromolecular crystal structures. *Acta Cryst.*, **D70**, 1104–1114.
37. Zheng,H., Chordia,M.D., Cooper,D.R., Chruszcz,M., Muller,P., Sheldrick,G.M. and Minor,W. (2014) Validation of metal-binding sites in macromolecular structures with the CheckMyMetal web server. *Nat. Protoc.*, **9**, 156–170.
38. Wlodawer,A., Minor,W., Dauter,Z. and Jaskolski,M. (2008) Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS J.*, **275**, 1–21.
39. Zheng,H., Chruszcz,M., Lasota,P., Lebioda,L. and Minor,W. (2008) Data mining of metal ion environments present in protein structures. *J. Inorg. Biochem.*, **102**, 1765–1776.
40. Cooper,D.R., Porebski,P.J., Chruszcz,M. and Minor,W. (2011) X-ray crystallography: assessment and validation of protein-small molecule complexes for drug discovery. *Expert Opin. Drug Dis.*, **6**, 771–782.
41. Pozharski,E., Weichenberger,C.X. and Rupp,B. (2013) Techniques, tools and best practices for ligand electron-density analysis and results from their application to deposited crystal structures. *Acta Cryst.*, **D69**, 150–167.
42. Wlodawer,A., Minor,W., Dauter,Z. and Jaskolski,M. (2013) Protein crystallography for aspiring crystallographers or how to avoid pitfalls and traps in macromolecular structure determination. *FEBS J.*, **280**, 5705–5736.
43. Dauter,Z., Wlodawer,A., Minor,W., Jaskolski,M. and Rupp,B. (2014) Avoidable errors in deposited macromolecular structures: an impediment to efficient data mining. *IUCrJ*, **1**, 179–193.

44. Weichenberger,C.X., Afonine,P.V., Kantardjieff,K. and Rupp,B. (2015) The solvent component of macromolecular crystals. *Acta Cryst.*, **D71**, 1023–1038.
45. Raczynska,J.E., Wlodawer,A. and Jaskolski,M. (2016) Prior knowledge or freedom of interpretation? A critical look at a recently published classification of 'novel' Zn binding sites. *Proteins*, **84**, 770–776.
46. Joosten,R.P., Womack,T., Vriend,G. and Bricogne,G. (2009) Re-refinement from deposited X-ray data can deliver improved models for most PDB entries. *Acta Cryst.*, **D65**, 176–185.
47. Read,R.J., Adams,P.D., Arendall,W.B. 3rd, Brunger,A.T., Emsley,P., Joosten,R.P., Kleywegt,G.J., Krissinel,E.B., Lutteke,T., Otwinowski,Z. *et al.* (2011) A new generation of crystallographic validation tools for the protein data bank. *Structure*, **19**, 1395–1412.
48. Joosten,R.P., Joosten,K., Murshudov,G.N. and Perrakis,A. (2012) PDB_REDO: constructive validation, more than just looking for errors. *Acta Cryst.*, **D68**, 484–496.
49. Joosten,R.P., Long,F., Murshudov,G.N. and Perrakis,A. (2014) The PDB_REDO server for macromolecular structure model optimization. *IUCrJ*, **1**, 213–220.
50. Touw,W.G., Joosten,R.P. and Vriend,G. (2016) New biological insights from better structure models. *J. Mol. Biol.*, **428**, 1375–1393.
51. van Beusekom,B., Perrakis,A. and Joosten,R.P. (2016) Data mining of macromolecular structures. *Methods Mol. Biol.*, **1415**, 107–138.
52. Minor,W., Dauter,Z., Helliwell,J.R., Jaskolski,M. and Wlodawer,A. (2016) Safeguarding structural data repositories against bad apples. *Structure*, **24**, 216–220.
53. Harding,M.M. and Hsin,K.Y. (2014) Mespeus–a database of metal interactions with proteins. *Methods Mol. Biol.*, **1091**, 333–342.
54. Sigel,R.K. and Sigel,H. (2010) A stability concept for metal ion coordination to single-stranded nucleic acids and affinities of individual sites. *Acc. Chem. Res.*, **43**, 974–984.
55. Lamb,A.L., Kappock,T.J. and Silvaggi,N.R. (2015) You are lost without a map: Navigating the sea of protein structures. *Biochim. Biophys. Acta*, **1854**, 258–268.
56. Mueller-Dieckmann,C., Panjikar,S., Schmidt,A., Mueller,S., Kuper,J., Geerlof,A., Wilmanns,M., Singh,R.K., Tucker,P.A. and Weiss,M.S. (2007) On the routine use of soft X-rays in macromolecular crystallography. Part IV. Efficient determination of anomalous substructures in biomacromolecules using longer X-ray wavelengths. *Acta Cryst.*, **D63**, 366–380.
57. Groom,C.R. and Allen,F.H. (2014) The Cambridge Structural Database in retrospect and prospect. *Angew. Chem. Int. Ed. Engl.*, **53**, 662–671.
58. Markham,G.D., Glusker,J.P. and Bock,C.W. (2002) The arrangement of first and second-sphere water molecules in divalent magnesium complexes: results from molecular orbital and density functional theory and from structural crystallography. *J. Phys. Chem. B*, **106**, 5118–5134.
59. Kuppuraj,G., Dudev,M. and Lim,C. (2009) Factors governing metal-ligand distances and coordination geometries of metal complexes. *J. Phys. Chem. B*, **113**, 2952–2960.
60. Martinez,J.M., Pappalardo,R.R. and Marcos,E.S. (1999) First-principles ion-water interaction potentials for highly charge monoatomic cations. Computer simulations of Al^{3+}, Mg^{2+}, and Be^{2+} in water. *J. Am. Chem. Soc.*, **121**, 3175–3184.
61. Lightstone,F.C., Schwegler,E., Hood,R.Q., Gygi,F. and Galli,G. (2001) A first principle molecular dynamics simulation of the hydrated magnesium ion. *Chem. Phys. Lett.*, 549–555.
62. Bhattacharjee,A., Pribil,A.B., Randolf,B.R., Rode,B.M. and Hofer,T.S. (2012) Hydration of Mg^{2+} and its influence on the water hydrogen bonding network via ab initio QMCF MD. *Chem. Phys. Lett.*, **536**, 39–44.
63. Adams,P.D., Afonine,P.V., Bunkoczi,G., Chen,V.B., Davis,I.W., Echols,N., Headd,J.J., Hung,L.W., Kapral,G.J., Grosse-Kunstleve,R.W. *et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Cryst.*, **D66**, 213–221.
64. Murshudov,G.N., Skubak,P., Lebedev,A.A., Pannu,N.S., Steiner,R.A., Nicholls,R.A., Winn,M.D., Long,F. and Vagin,A.A. (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta Cryst.*, **D67**, 355–367.
65. Harding,M.H. (1999) The geometry of metal-ligand interactions relevant to proteins. *Acta Cryst.*, **D55**, 1432–1443.
66. Kleywegt,G.J., Harris,M.R., Zou,J.Y., Taylor,T.C., Wahlby,A. and Jones,T.A. (2004) The Uppsala electron-density server. *Acta Cryst.*, **D60**, 2240–2249.
67. Anderson,M., Schultz,E.P., Martick,M. and Scott,W.G. (2013) Active-site monovalent cations revealed in a 1.55-A-resolution hammerhead ribozyme structure. *J. Mol. Biol.*, **425**, 3790–3798.
68. Auffinger,P., Bielecki,L. and Westhof,E. (2004) Anion binding to nucleic acids. *Structure*, **12**, 379–388.
69. D'Ascenzo,L. and Auffinger,P. (2016) Anions in nucleic acid crystallography. *Methods Mol. Biol.*, **1320**, 337–351.
70. Giegé,R. (2013) A historical perspective on protein crystallization from 1840 to the present day. *FEBS J.*, **280**, 6456–6497.
71. DePristo,M.A., De Bakker,P.I. and Blundell,T.L. (2004) Heterogeneity and inaccuracy in protein structures solved by x-ray crystallography. *Structure*, **12**, 831–838.
72. Kagawa,T.F., Geierstanger,B.H., Wang,A.H.J. and Ho,P.S. (1991) Covalent modification of guanine bases in double-stranded DNA. The 1.2 Å Z-DNA structure of d(CGCGCG) in the presence of $CuCl_2$. *J. Biomol. Chem.*, **266**, 20175–20184.
73. Drozdzal,P., Gilski,M., Kierzek,R., Lomozik,L. and Jaskolski,M. (2013) Ultrahigh-resolution crystal structures of Z-DNA in complex with Mn^{2+} and Zn^{2+} ions. *Acta Cryst.*, **D69**, 1180–1190.
74. Gessner,R.V., Frederick,C.A., Quigley,G.J., Rich,A. and Wang,A.H.J. (1989) The molecular structure of the left-handed Z-DNA double helix at 1.0 Å atomic resolution. Geometry, conformation, and ionic interactions of d(CGCGCG). *J. Biol. Chem.*, **264**, 7912–7935.
75. Ho,P.S. and Mooers,B.H.M. (1997) Z-DNA crystallography. *Biopolymers*, **44**, 65–90.
76. Brzezinski,K., Brzuszkiewicz,A., Dauter,M., Kubicki,M., Jaskolski,M. and Dauter,Z. (2011) High regularity of Z-DNA revealed by ultra high-resolution crystal structure at 0.55 Å. *Nucleic Acids Res.*, **39**, 6238–6248.
77. Chatake,T. and Sunami,T. (2013) Direct interactions between Z-DNA and alkaline earth cations, discovered in the presence of high concentrations of $MgCl_2$ and $CaCl_2$. *J. Inorg. Biochem.*, **124**, 15–25.
78. Chatake,T., Tanaka,I., Umino,H., Arai,S. and Niimura,N. (2005) The hydration structure of a Z-DNA hexameric duplex determined by a neutron diffraction technique. *Acta Cryst.*, **D61**, 1088–1098.
79. Yang,X.L., Robinson,H., Gao,Y.G. and Wang,A.H. (2000) Binding of a macrocyclic bisacridine and ametantrone to CGTACG involves similar unusual intercalation platforms. *Biochemistry*, **39**, 10950–10957.
80. Labiuk,S.L., Delbaere,L.T. and Lee,J.S. (2003) Cobalt(II), nickel(II) and zinc(II) do not bind to intra-helical N(7) guanine positions in the B-form crystal structure of d(GGCGCC). *J. Biol. Inorg. Chem.*, **8**, 715–720.
81. Egli,M., Gessner,R.V., Williams,L.D., Quigley,G.J., Vandermarel,G.A., Vanboom,J.H., Rich,A. and Frederick,C.A. (1990) Atomic-resolution structure of the cellulose synthase regulator cyclic diguanylic acid. *Proc. Natl. Acad. Sci. U.S.A.*, **87**, 3235–3239.
82. McLellan,T.J., Marr,E.S., Wondrack,L.M., Subashi,T.A., Aeed,P.A., Han,S., Xu,Z., Wang,I.K. and Maguire,B.A. (2009) A systematic study of 50S ribosomal subunit purification enabling robust crystallization. *Acta Cryst.*, **D65**, 1270–1282.
83. Khatter,H., Myasnikov,A.G., Mastio,L., Billas,I.M., Birck,C., Stella,S. and Klaholz,B.P. (2014) Purification, characterization and crystallization of the human 80S ribosome. *Nucleic Acids Res.*, **42**, e49.
84. Maret,W. (2013) Zinc biochemistry: from a single zinc enzyme to a key element of life. *Adv. Nutr.*, **4**, 82–91.
85. Laity,J.H., Lee,B.M. and Wright,P.E. (2001) Zinc finger proteins: new insights into structural and functional diversity. *Curr. Opin. Struct. Biol.*, **11**, 39–46.
86. Dresios,J., Chan,Y.L. and Wool,I.G. (2005) Ribosomal zinc finger proteins: the structure and the function of yeast YL37a. In: UIuchi,S and Kuldell,N (eds). *Zinc Finger Proteins: From Atomic Contact to Cellular Function*. Springer, Boston, pp. 91–98.
87. Hensley,M.P., Tierney,D.L. and Crowder,M.W. (2011) Zn(II) binding to Escherichia coli 70S ribosomes. *Biochemistry*, **50**, 9937–9939.

88. Zhang,J. and Ferre-D'Amare,A.R. (2014) Dramatic improvement of crystals of large RNAs by cation replacement and dehydration. *Structure*, **22**, 1363–1371.
89. Gunasekera,T., Easton,J.A., Sugerbaker,S.A., Klingbeil,L. and Crowder,M.W. (2009) Zn(II) homeostasis in *E. coli*. In: Long,EC and Baldwin,MJ (eds). *Bioinorganic Chemistry*. American Chemical Society, pp. 81–95.
90. Thorn,A. and Sheldrick,G.M. (2011) ANODE: anomalous and heavy-atom density calculation. *J. Appl. Cryst.*, **44**, 1285–1287.
91. Juneau,K., Podell,E., Harrington,D.J. and Cech,T.R. (2001) Structural basis of the enhanced stability of a mutant ribozyme domain and a detailed view of RNA-solvent interactions. *Structure*, **9**, 221–231.
92. Freisinger,E. and Sigel,R.K. (2007) From nucleotides to ribozymes - A comparison of their metal ion binding properties. *Coord. Chem. Rev.*, **251**, 1834–1851.
93. Frederiksen,J.K., Li,N.S., Das,R., Herschlag,D. and Piccirilli,J.A. (2012) Metal-ion rescue revisited: biochemical detection of site-bound metal ions important for RNA folding. *RNA*, **18**, 1123–1141.
94. Osawa,T., Inanaga,H., Sato,C. and Numata,T. (2015) Crystal structure of the CRISPR-Cas RNA silencing Cmr complex bound to a target analog. *Mol. Cell.*, **58**, 418–430.
95. Dudev,T. and Lim,C. (2004) Monodentate versus bidentate carboxylate binding in magnesium and calcium proteins: what are the basic principles? *J. Phys. Chem. B*, **108**, 4546–4557.
96. Martick,M., Lee,T.S., York,D.M. and Scott,W.G. (2008) Solvent structure and hammerhead ribozyme catalysis. *Chem. Biol.*, **15**, 332–342.
97. Serganov,A., Huang,L. and Patel,D.J. (2008) Structural insights into amino acid binding and gene control by a lysine riboswitch. *Nature*, **455**, 1263–1268.
98. Tereshko,V., Wilds,C.J., Minasov,G., Prakash,T.P., Maier,M.A., Howard,A., Wawrzak,Z., Manoharan,M. and Egli,M. (2001) Detection of alkali metal ions in DNA crystals using state-of-the-art X-ray diffraction experiments. *Nucleic Acids Res.*, **29**, 1208–1215.
99. Ennifar,E., Walter,P. and Dumas,P. (2003) A crystallographic study of the binding of 13 metal ions to two related RNA duplexes. *Nucleic Acids Res.*, **31**, 2671–2682.
100. Safaee,N., Noronha,A.M., Rodionov,D., Kozlov,G., Wilds,C.J., Sheldrick,G.M. and Gehring,K. (2013) Structure of the parallel duplex of poly(A) RNA: evaluation of a 50 year-old prediction. *Angew. Chem. Int. Ed. Engl.*, **52**, 10370–10373.
101. Wild,K., Weichenrieder,O., Leonard,G.A. and Cusack,S. (1999) The 2 Å structure of helix 6 of the human signal recognition particle RNA. *Structure Fold. Des.*, **7**, 1345–1352.
102. Ohishi,H., Suzuki,K., Ohtsuchi,M., Hakoshima,T. and Rich,A. (2002) The crystal structure of N(1)-[2-(2-amino-ethylamino)-ethyl]-ethane-1,2-diamine (polyamines) binding to the minor groove of d(CGCGCG)(2), hexamer at room temperature. *FEBS Lett.*, **523**, 29–34.
103. Howerton,S.B., Sines,C.C., VanDerveer,D. and Williams,L.D. (2001) Locating monovalent cations in the grooves of B-DNA. *Biochemistry*, **40**, 10023–10031.
104. Batey,R.T. and Doudna,J.A. (2002) Structural and energetics of metal ions essential to SRP signal recognition domain assembly. *Biochemistry*, **41**, 11703–11710.
105. Auffinger,P., Bielecki,L. and Westhof,E. (2004) Symmetric K^+ and Mg^{2+} ion binding sites in the 5S rRNA loop E inferred from molecular dynamics simulations. *J. Mol. Biol.*, **335**, 555–571.
106. Dauter,Z. and Adamiak,D.A. (2001) Anomalous signal of phosphorus used for phasing DNA oligomer: importance of data redundancy. *Acta Cryst.*, **D57**, 990–995.
107. Knobloch,B., Linert,W. and Sigel,H. (2005) Metal ion-binding properties of (N3)-deprotonated uridine, thymidine, and related pyrimidine nucleosides in aqueous solution. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 7459–7464.
108. Polikanov,Y.S., Steitz,T.A. and Innis,C.A. (2014) A proton wire to couple aminoacyl-tRNA accommodation and peptide-bond formation on the ribosome. *Nat. Struct. Mol. Biol.*, **21**, 787–793.
109. Hendlich,M., Bergner,A., Gunther,J. and Klebe,G. (2003) Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *J. Mol. Biol.*, **326**, 607–620.
110. Marcia,M. and Pyle,A.M. (2012) Visualizing group II intron catalysis through the stages of splicing. *Cell*, **151**, 497–507.
111. Huang,L., Serganov,A. and Patel,D.J. (2010) Structural insights into ligand recognition by a sensing domain of the cooperative glycine riboswitch. *Mol. Cell.*, **40**, 774–786.
112. Bowman,S.E., Bridwell-Rabb,J. and Drennan,C.L. (2016) Metalloprotein crystallography: more than a structure. *Acc. Chem. Res.*, **49**, 695–702.
113. Ennifar,E., Walter,P. and Dumas,P. (2010) Cation-dependent cleavage of the duplex form of the subtype-B HIV-1 RNA dimerization initiation site. *Nucleic Acids Res.*, **38**, 5807–5816.
114. Jovine,L., Djordjevic,S. and Rhodes,D. (2000) The crystal structure of yeast phenylalanine tRNA at 2.0 Å resolution: cleavage by Mg^{2+} in 15-year old crystals. *J. Mol. Biol.*, **301**, 401–414.
115. Robart,A.R., Chan,R.T., Peters,J.K., Rajashankar,K.R. and Toor,N. (2014) Crystal structure of a eukaryotic group II intron lariat. *Nature*, **514**, 193–197.
116. Marcia,M. and Pyle,A.M. (2014) Principles of ion recognition in RNA: insights from the group II intron structures. *RNA*, **20**, 516–527.
117. Wang,J. (2010) Inclusion of weak high-resolution X-ray data for improvement of a group II intron structure. *Acta Cryst.*, **D66**, 988–1000.
118. Kawamura,T., Kobayashi,T. and Watanabe,N. (2015) Analysis of the HindIII-catalyzed reaction by time-resolved crystallography. *Acta Cryst.*, **D71**, 256–265.
119. Muller,P., Kopke,S. and Sheldrick,G.M. (2003) Is the bond-valence method able to identify metal atoms in protein structures? *Acta Cryst.*, **D59**, 32–37.
120. Borbulevych,O., Martin,R.I., Tickle,I.J. and Westerhoff,L.M. (2016) XModeScore: a novel method for accurate protonation/tautomer-state determination using quantum-mechanically driven macromolecular X-ray crystallographic refinement. *Acta Cryst.*, **D72**, 586–598.
121. Borek,D., Minor,W. and Otwinowski,Z. (2003) Measurement errors and their consequences in protein crystallography. *Acta Cryst.*, **D59**, 2031–2038.
122. Diederichs,K. (2016) Crystallographic data and model quality. In: Ennifar,E (ed). *Nucleic Acid Crystallography: Methods and Protocols*. Springer, NY, pp. 147-173.
123. Williams,R.J.P. (2001) Chemical selection of elements by cells. *Coord. Chem. Rev.*, **216**, 583–595.
124. Hill,H.A. and Sadler,P.J. (2016) Bringing inorganic chemistry to life with inspiration from R. J. P. Williams. *J. Biol. Inorg. Chem.*, **21**, 5–12.
125. Hennings,E., Schmidt,H. and Voigt,W. (2013) Crystal structures of hydrates of simple inorganic salts. I. Water-rich magnesium halide hydrates $MgCl_2.8H_2O$, $MgCl_2.12H_2O$, $MgBr_2.6H_2O$, $MgBr_2.9H_2O$, $MgI_2.8H_2O$ and $MgI_2.9H_2O$. *Acta Cryst.*, **C69**, 1292–1300.

Supplementary Material

Mg^{2+} ions: do they bind to nucleobase nitrogens?

Filip Leonarski[1,2], Luigi D'Ascenzo[1] and Pascal Auffinger[1,*]

[1] Architecture et Réactivité de l'ARN, Université de Strasbourg, Institut de Biologie Moléculaire et Cellulaire du CNRS, Strasbourg, 67084, France

[2] Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

* To whom correspondence should be addressed. Tel: +33 388 41 70 49; Fax: +33 388 60 22 18; Email: p.auffinger@ibmc-cnrs.unistra.fr.

Table of content

About ribosomal PDB structures (resolution ≤ 3.0 Å) and assembly choice

Currently (Mai 2016), the PDB contains 134 X-ray structures of ribosomes with resolution ≤ 3.0 Å (**Table S1**) that comprise: *(i)* 12 *Thermus thermophilus* small subunits (30S), *(ii)* 3 *Deinococcus radiodurans* large subunits (50S), *(iii)* 57 *Haloarcula marismortui* large subunits (50S), *(iv)* 39 *Thermus thermophilus* 70S ribosomes, *(v)* 15 *Escherichia coli* 70S ribosomes and *(vi)* 7 *Saccharomyces cerevisiae* 80S ribosomes.

For the 61 ribosomal structures that are considered as large by the PDB (70S and 80S) and only available in mmCIF files, only one biological assembly was analyzed (see **Table S1**), the choice of which was based on the following consideration: the biological assembly with the lowest average *B-factor* value is retained. However, for the *T. thermophilus* 70S structures, numbering inconsistencies between the two or four assemblies that resulted from in-house PDB annotations were found. Hence, when such discrepancies occur, we chose to use the structure that has a numbering consistent with the 2D structures found at http://apollo.chemistry.gatech.edu/RibosomeGallery (1). Appropriate annotations are made in **Table S1** and **Table 2**.

T. thermophilus (30S)

The best resolution structure is **2VQE** (2.50 Å). All structures comprise Zn^{2+} and Mg^{2+} ions. K^{+} has been attributed in eight of them. For some of these structures, Ca^{2+}, Na^{+}, NH_4^{+}, Cl^{-} and acetate ions may also be present ($CaCl_2$, KCl, NH_4Cl, magnesium acetate, sodium cacodylate, …) according to the crystallization conditions noted in the PDB (see for example **1FJG**). The publication describing the purification and crystallization protocols gathered does not mention the use of $CaCl_2$ (2). Proteins S4 and S14 contain each a zinc-finger domain (3). Thus, each of these structures contain two unambiguously identified Zn^{2+}.

D. radiodurans (50S)

The best resolution structure is **5DM6** (2.90 Å). Mg^{2+} is the only ion that has been assigned in these 3 structures. Zn^{2+} is not present since the cysteine residues that should be involved in a zinc finger motif form disulfide bridges. Na^{+}, NH_4^{+}, spermidine and Cl^{-} may also be present according to the crystallization conditions noted in the PDB. Further information on purification and crystallization protocols can be found in reference (4).

H. marismortui (50S)

Currently, the best resolution structures are **1VQO** and **1VQ8** (2.20 Å). Oddly, all 58 structures contain Cd^{2+} and do not mention Zn^{2+}. Cadmium was introduced in the crystallization buffers as $CdCl_2$ for improving the crystallization conditions (5,6). The composition of the crystallization buffer for **1FFK** was: 1.2 M KCl, 0.5 M NH_4Cl, 100 mM KCH_3COO, 30 mM $MgCl_2$, 7% polyethylene glycol (PEG) 6000, 15 mM tris, 15 mM MES, and 1 mM $CdCl_2$ (pH 7.1), with transfer of crystals in the following buffer: 12% PEG 6000, 22% ethylene glycol, 1.7 M NaCl, 0.5 M NH_4Cl, 100 mM KCH_3COO, 30 mM $MgCl_2$, and 1 mM $CdCl_2$ (pH 6.2) at 4°C (6). The four proteins L24e, L37e, L37ae, and L44e, contain zing finger motifs. Four of these domains, where Cd^{2+} replaces Zn^{2+}, were identified (7). Further, one additional weak and probably suspicious Cd^{2+} binding site has been reported.

This group of *H. marismortui* structures is very heterogeneous. In early structures, residue 1329 (chain 0) is wrongly assigned to an adenine as shown by $2F_o$-F_c and F_o-F_c maps in **1FFK** (**Figure S1**). This error has been subsequently corrected in **4V9F** that is a re-refinement of **1FFK** and in the structures marked as such in **Table S1**. The best resolution structures with the corrected sequence are thus, **4V9F**, **1YHQ** and **3CC2** (2.40 Å). Modified ribosomal nucleotides are assigned in **4V9F**.

Twenty-eight of these structures contain Sr^{2+}. The use of $SrCl_2$ seems to be related to the crystallization of antibiotic/ribosome complexes. In one study, it is mentioned that crystals were soaked with antibiotic in a buffer containing among other: 1.7 M NaCl, 0.5 M NH_4Cl, 1nM $CdCl_2$, 100mM KCH_3COO, 6.5 mM CH_3COOH and with either 30 mM $MgCl_2$, or 21 mM $MgCl_2$ and 30 mM $SrCl_2$, or no $MgCl_2$ and 100 mM $SrCl_2$ (8). These conditions suggest that these ribosomes, when crystallized following such soaking protocols, might include very high NaCl and NH_4Cl concentrations. The best resolution structure with Sr^{2+} and the corrected sequence is **3CCM** (2.55 Å).

It should be noted that *H. marismortui* structures contain single ribosomal RNA chains, which in some structures are labeled “chain 0” and “chain A” in others. Here, these chains were considered equivalent.

T. thermophilus (70S)

The best resolution structure is **4Y4O** (2.30 Å) that contain also a full array of modified ribosomal nucleotides. Zn^{2+} is present in all 40 structures.

E. coli

The best resolution structure is **4YBB** (2.10 Å). It is also currently the best resolution of all ribosomal structures and contains a full array of modified ribosomal nucleotides. Zn^{2+} is present in all *E. coli* structures. For **4YBB**, a strong discrepancy between the *B-factors* of both assemblies present in the crystal unit has to be noted (average *B-factors* around ≈67 and ≈123 $Å^2$, respectively) signifying that the analysis should be limited to the first biological assembly. Besides Mg^{2+} and Zn^{2+}, the **4YBB** structure contains also acetate, 1,2-ethanediol, di(hydroxyethyl)ether, putrescine, spermidine, MPD, TRIS, tetraethylene- and pentaethyleneglycol. Purification and crystallization protocols can be found in reference (9,10).

S. cerevisiae (80S)

The best resolution structure is **4U4R** (2.80 Å). Osmium hexamine is present in all seven structures as well as eight Zn^{2+} ions per assembly. Purification and crystallization protocols for eukaryotic ribosomes can be found in references (11-13).

Table S1: List of 134 ribosomal structures in the PDB (resolution ≤ 3.0 Å). The best resolution structure in each category is shown in blue. In each category, structures are ordered according to PDB deposition date (first criterion), best resolution (second criterion) and best *R-free* value (third criterion).

PDB code	Deposited	Res. (R-Free)	Residue count	Ions	Ref.
***T. Thermophilus* (30S)** *— 12 structures*					
1FJG	08/2000	3.00 (0.26)	4068	Zn^{2+}, Mg^{2+}	(14)
1N32	10/2002	3.00 (0.27)	4077	Zn^{2+}, Mg^{2+}	(15)
1XMQ	10/2004	3.00 (0.24)	4078	Zn^{2+}, Mg^{2+}	(16)
2UUB	03/2007	2.80 (0.24)	4086	Zn^{2+}, Mg^{2+}, K^{+}	(17)
2UUA	03/2007	2.90 (0.25)	4086	Zn^{2+}, Mg^{2+}, K^{+}	(17)
2UXC	03/2007	2.90 (0.26)	4086	Zn^{2+}, Mg^{2+}, K^{+}	(18)
2VQE	03/2008	2.50 (0.28)	4086	Zn^{2+}, Mg^{2+}, K^{+}	(19)
2VQF	03/2008	2.90 (0.26)	4086	Zn^{2+}, Mg^{2+}, K^{+}	(19)
3T1Y	07/2011	2.80 (0.27)	4065	Zn^{2+}, Mg^{2+}	(20)
4B3M	07/2012	2.90 (0.25)	4072	Zn^{2+}, Mg^{2+}, K^{+}	(21)
4B3T	07/2012	3.00 (0.24)	4072	Zn^{2+}, Mg^{2+}, K^{+}	(21)
4B3R	07/2012	3.00 (0.25)	4072	Zn^{2+}, Mg^{2+}, K^{+}	(21)
***D. radiodurans* (50S)** *— 3 structures*					
2ZJR[j]	03/2008	2.91 (0.31)	6562	Mg^{2+}	(22)
5DM6	09/2015	2.90 (0.27)	6490	Mg^{2+}	(23)
5DM7[j]	09/2015	3.00 (0.33)	6490	Mg^{2+}	(23)
H. marismortui* (50S)** *— with wrong adenine at position 1329 (see* ***Figure S1*) — 32 structures*					
1FFK	07/2000	2.40 (0.26)	6825	Cd^{2+}, Mg^{2+}, K^{+}	(6)
1JJ2	07/2001	2.40 (0.22)	7279	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(24)
1K8A	10/2001	3.00 (0.26)	7279	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(25)
1K9M	10/2001	3.00 (0.26)	7279	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(25)
1KD1	11/2001	3.00 (0.27)	7279	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(25)
1M90	07/2002	2.80 (0.22)	7282	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(26)
1N8R[k]	11/2002	3.00 (0.24)	7279	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(27)
1NJI	12/2002	3.00 (0.21)	7279	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(27)
1QVG	08/2003	2.90 (0.26)	7288	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(28)
1Q81	08/2003	2.95 (0.26)	7281	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(26)
1Q82	08/2003	2.98 (0.25)	7281	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(26)
1Q86	08/2003	3.00 (0.26)	7285	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(26)
1S72	01/2004	2.40 (0.22)	7464	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(7)
1VQO[a]	12/2004	2.20 (0.25)	7478	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(29)
1VQ8[a]	12/2004	2.20 (0.25)	7479	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(29)
1VQP	12/2004	2.25 (0.25)	7483	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(29)
1VQK	12/2004	2.30 (0.25)	7480	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(29)
1VQL	12/2004	2.30 (0.25)	7482	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(29)
1VQM	12/2004	2.30 (0.25)	7482	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(29)
1VQN	12/2004	2.40 (0.25)	7484	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(30)
1VQ9	12/2004	2.40 (0.26)	7480	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(29)
1VQ7	12/2004	2.50 (0.24)	7482	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(30)
1VQ5	12/2004	2.60 (0.24)	7482	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(29)
1VQ4	12/2004	2.70 (0.23)	7482	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(29)
1VQ6	12/2004	2.70 (0.23)	7483	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(30)
2OTL	02/2007	2.70 (0.25)	7462	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(31)
2OTJ	02/2007	2.90 (0.24)	7463	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(31)
2QEX	06/2007	2.90 (0.24)	7320	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(32)
2QA4	06/2007	3.00 (0.29)	7486	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(33)
3CPW	04/2008	2.70 (0.23)	7334	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(34)
3CXC	04/2008	3.00 (0.23)	7282	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(35)
3OW2	09/2010	2.70 (0.25)	6801	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	—
H. marismortui* (50S)** *— with correct guanine at position 1329 (see* ***Figure S1*) — 26 structures*					
1YHQ	01/2005	2.40 (0.23)	7480	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(36)
1YIJ	01/2005	2.60 (0.22)	7481	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(36)
1YI2	01/2005	2.65 (0.21)	7481	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(36)

1YIT	01/2005	2.80 (0.22)	7478	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(36)
1YJ9	01/2005	2.80 (0.24)	7478	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(36)
1YJW	01/2005	2.90 (0.22)	7481	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(36)
1YJN	01/2005	3.00 (0.23)	7481	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(36)
3CC2	02/2008	2.40 (0.23)	7517	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(8)
3CCM	02/2008	2.55 (0.24)	7517	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(8)
3CC7	02/2008	2.70 (0.23)	7517	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(8)
3CCJ	02/2008	2.70 (0.23)	7517	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(8)
3CC4	02/2008	2.70 (0.24)	7517	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(8)
3CCE	02/2008	2.75 (0.23)	7517	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(8)
3CD6	02/2008	2.75 (0.24)	7520	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(8)
3CCU	02/2008	2.80 (0.22)	7517	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(8)
3CCL	02/2008	2.90 (0.22)	7517	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(8)
3CCV	02/2008	2.90 (0.22)	7517	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(8)
3CCQ	02/2008	2.90 (0.23)	7517	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(8)
3CCS	02/2008	2.95 (0.24)	7517	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(8)
3CCR	02/2008	3.00 (0.25)	7517	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(8)
3CMA	03/2008	2.80 (0.24)	7522	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(37)
3CME	03/2008	2.95 (0.26)	7522	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(37)
3G6E	02/2009	2.70 (0.23)	7217	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(38)
3G71	02/2009	2.85 (0.23)	7217	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(38)
3I56	07/2009	2.90 (0.24)	7217	Cd^{2+}, Sr^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(38)
4V9F[b]	02/2012	2.40 (0.21)	7583[c,d]	Cd^{2+}, Mg^{2+},K^{+}, Na^{+}, Cl^{-}	(39)
***T. thermophilus* (70S)** — *42 structures*					
4V51	07/2006	2.80 (0.31)	21886[c,e]	Zn^{2+}, Mg^{2+}	(40)
4V67	20/2008	3.00 (0.32)	22654[c,e]	Zn^{2+}, Mg^{2+}	(41)
4V7L	11/2009	3.00 (0.27)	22204[c,e]	Zn^{2+}, Mg^{2+}	(42)
4V7Y	08/2010	3.00 (0.27)	21368[c,e]	Zn^{2+}, Mg^{2+}, K^{+}	(43)
4V7W	08/2010	3.00 (0.28)	21368[c,e]	Zn^{2+}, Mg^{2+}, K^{+}	(43)
4V7X	08/2010	3.00 (0.28)	21368[c,e]	Zn^{2+}, Mg^{2+}, K^{+}	(43)
4V8D	12/2011	3.00 (0.24)	21658[c,e]	Zn^{2+}, Mg^{2+}	(44)
4V8B	12/2011	3.00 (0.27)	21634[c,e]	Zn^{2+}, Mg^{2+}	(44)
4V8I	12/2011	2.70 (0.25)	21484[c,e]	Zn^{2+}, Mg^{2+}	(43)
4V8G	12/2011	3.00 (0.25)	21368[c,e]	Zn^{2+}, Mg^{2+}	(43)
4V9H	03/2013	2.86 (0.25)	11728[c,d]	Mg^{2+}	(45)
4LNT	07/2013	2.94 (0.26)	21492[c,e]	Zn^{2+}, Mg^{2+}	(46)
4V9R	12/2013	3.00 (0.26)	21468[c,e]	Zn^{2+}, Mg^{2+}, K^{+}	(47)
4V90	02/2014	2.95 (0.24)	11801[c,d]	Zn^{2+}, Mg^{2+}	(48)
1VY5	05/2014	2.55 (0.28)	21748[c,e]	Zn^{2+}, Mg^{2+}, K^{+}	(49)
1VY4	05/2014	2.60 (0.26)	21748[c,e]	Zn^{2+}, Mg^{2+}, K^{+}	(49)
1VY7	05/2014	2.80 (0.28)	21602[c,e]	Zn^{2+}, Mg^{2+}, K^{+}	(49)
1VY6	05/2014	2.90 (0.29)	21448[c,e]	Zn^{2+}, Mg^{2+}, K^{+}	(49)
4W2E	06/2014	2.90 (0.30)	11638[c,d]	Zn^{2+}, Mg^{2+}	(50)
4W2G	09/2014	2.55 (0.27)	21748[c,e]	Zn^{2+}, Mg^{2+}, K^{+}	(51)
4W2H	09/2014	2.70 (0.26)	21596[c,e]	Zn^{2+}, Mg^{2+}, K^{+}	(51)
4W2I	09/2014	2.70 (0.26)	21748[c,e]	Zn^{2+}, Mg^{2+}, K^{+}	(52)
4W2F	10/2014	2.40 (0.28)	21748[c,e]	Zn^{2+}, Mg^{2+}, K^{+}	(51)
4WPO	10/2014	2.80 (0.25)	24064[c,e]	Zn^{2+}, Mg^{2+}, K^{+}	(53)
4WQF	10/2014	2.80 (0.26)	23898[c,e]	Zn^{2+}, Mg^{2+}	(53)
4WQU	10/2014	2.80 (0.26)	23918[c,e]	Zn^{2+}, Mg^{2+}	(53)
4WQY	10/2014	2.80 (0.27)	23760[c,e]	Zn^{2+}, Mg^{2+}, K^{+}	(53)
4WSD	10/2014	2.95 (0.24)	21694[c,e]	Zn^{2+}, Mg^{2+}	(54)
4Y4O[b]	02/2015	2.30 (0.25)	21468[c,e]	Zn^{2+}, Mg^{2+}	(55)
4Y4P	02/2015	2.50 (0.28)	21748[c,e]	Zn^{2+}, Mg^{2+}, K^{+}	(56)
4Z3S	03/2015	2.65 (0.27)	21748[c,e]	Zn^{2+}, Mg^{2+}, K^{+}	(57)
4Z8C[j]	04/2015	2.90 (0.25)	21482[c,e]	Zn^{2+}, Mg^{2+}	(58)
5DOY	09/2015	2.60 (0.28)	21754[c,e]	Zn^{2+}, Mg^{2+}, K^{+}	(57)
5E81	10/2015	2.95 (0.24)	21961[c,e]	Zn^{2+}, Mg^{2+}	(59)
5FDV	12/2015	2.80 (0.24)	20986[c,e]	Zn^{2+}, Mg^{2+}	(56)
5F8K	12/2015	2.80 (0.29)	20948[c,e]	Zn^{2+}, Mg^{2+}, K^{+}	(56)
5FDU	12/2015	2.90 (0.23)	20974[c,e]	Zn^{2+}, Mg^{2+}	(56)

5HAU[j]	12/2015	3.00 (0.26)	23832[c,e]	Zn^{2+}, Mg^{2+}	(60)
5HD1[j]	01/2016	2.70 (0.27)	21484[c,e]	Zn^{2+}, Mg^{2+}	(60)
5HCQ[j]	01/2016	2.80 (0.25)	21472[c,e]	Zn^{2+}, Mg^{2+}	(60)
5HCR[j]	01/2016	2.80 (0.27)	21482[c,e]	Zn^{2+}, Mg^{2+}	(60)
5HCP[j]	01/2016	2.89 (0.27)	21474[c,e]	Zn^{2+}, Mg^{2+}	(60)
***E. coli* (70S)** *— 12 structures*					
4V9D	07/2012	3.00 (0.26)	20931[c,e]	Zn^{2+}, Mg^{2+}	(9)
4V9O	05/2013	2.90 (0.27)	45264[c,f]	Zn^{2+}, Mg^{2+}	(10)
4V9P	05/2013	2.90 (0.27)	44972[c,f]	Zn^{2+}, Mg^{2+}	(10)
4U1U	04/2014	2.95 (0.28)	20810[c,e]	Zn^{2+}, Mg^{2+}	(61)
4U27	06/2014	2.80 (0.26)	20808[c,e]	Zn^{2+}, Mg^{2+}	(61)
4U26	06/2014	2.80 (0.27)	20810[c,e]	Zn^{2+}, Mg^{2+}	(61)
4U20	06/2014	2.90 (0.28)	20794[c,e]	Zn^{2+}, Mg^{2+}	(61)
4U24	06/2014	2.90 (0.26)	20794[c,e]	Zn^{2+}, Mg^{2+}	(61)
4U25	06/2014	2.90 (0.26)	20794[c,e]	Zn^{2+}, Mg^{2+}	(61)
4U1V	06/2014	3.00 (0.27)	20808[c,e]	Zn^{2+}, Mg^{2+}	(61)
4WOI	10/2014	3.00 (0.25)	21699[c,e]	Zn^{2+}, Mg^{2+}	(62)
4YBB[b]	02/2015	2.10 (0.23)	20744[c,e,h]	Zn^{2+}, Mg^{2+}	(63)
***S. cerevisiae* (80S)** *— 7 structures* [i]					
4V88	10/2011	3.00 (0.23)	35856[c,g]	Zn^{2+}, Mg^{2+}	(12)
4U4R	07/2014	2.80 (0.25)	35344[c,g]	Zn^{2+}, Mg^{2+}	(64)
4U3U	07/2014	2.90 (0.24)	35344[c,g]	Zn^{2+}, Mg^{2+}	(64)
4U3M	07/2014	3.00 (0.24)	35344[c,g]	Zn^{2+}, Mg^{2+}	(64)
4U4U	07/2014	3.00 (0.26)	35344[c,g]	Zn^{2+}, Mg^{2+}	(64)
4U52	07/2014	3.00 (0.26)	35344[c,g]	Zn^{2+}, Mg^{2+}	(64)
4U4Q	07/2014	3.00 (0.26)	35346[c,g]	Zn^{2+}, Mg^{2+}	(64)

[a] Although these ribosomes have the best 2.20 Å resolution for *H. marismortui*, they embed assignment errors as shown in **Figure S1** and should be considered with caution.

[b] These structures embed ribosomal modifications.

[c] These very large ribosomal structures are only available in mmCIF format in the PDB.

[d] These large structures contain one biological assembly (space group: P1211).

[e] These large structures contain two biological assemblies (space group: P212121).

[f] These large structures contain four biological assemblies (space group: P1211).

[g] These large structures contain two biological assemblies (space group: P1211).

[h] The *B-factors* of the first assembly are significantly higher than those of the second assembly that should consequently not be considered for further structural analysis.

[i] The numbering of these files are consistent with the 2D structures available at http://apollo.chemistry.gatech.edu/RibosomeGallery (1).

[j] The numbering of these files follows sequence and is not aligned towards *E. coli* ribosome.

[k] Structure factors not available.

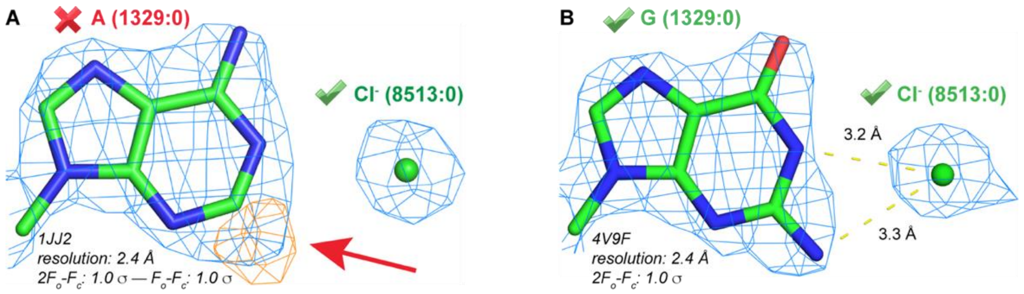
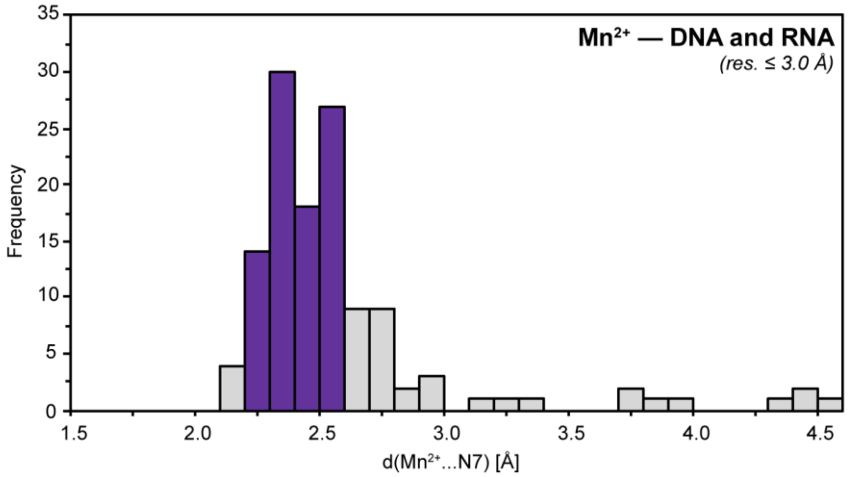


Figure S1. Nucleotide assignment error in *H. marismortui*. (**A**) Although a density for an N2 amino group is clearly visible in the F$_o$-F$_c$ map (red arrow), this nucleotide has been incorrectly assigned to an adenine in **1FFK** and related structures (**Table S1**) at odds with a Cl$^-$ ion bound to its Watson-Crick edge. (**B**) This error has been corrected In **4V9F** and related structures (**Table S1**).



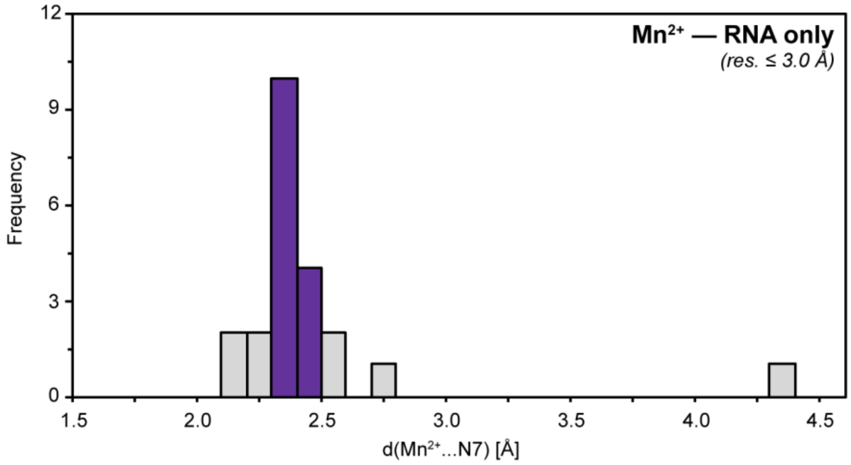


Figure S2. d(Mn^{2+}…N7) histograms derived from the PDB (May 2016; resolution ≤ 3.0 Å). These histograms emphasize the quasi-absence of coordination above 2.6 Å. For the first histogram, the largest number of data commes from nucleosomes (top).

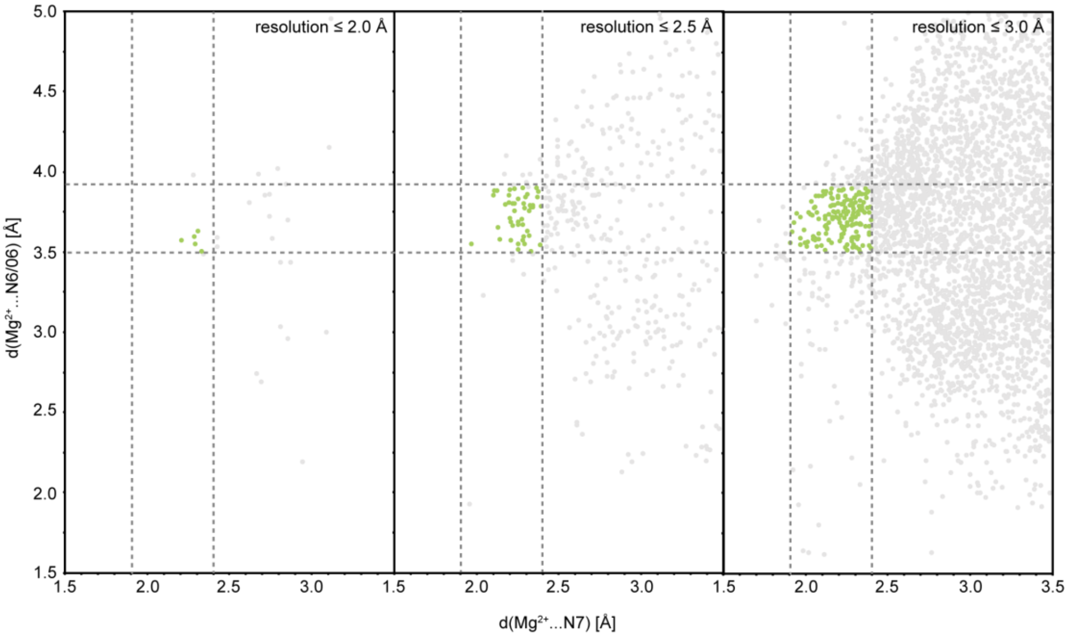


Figure S3. Mg^{2+} to N7 scatter plots for the ≤ 2.0, ≤ 2.5 and ≤ 3.0 Å resolution ranges. The 1.9 ≤ d(Mg^{2+}…N7) ≤ 2.4 Å and the 3.5 ≤ d(Mg^{2+}…N6/06) ≤ 3.9 Å limits are marked by dashed grey lines. The Mg^{2+} ions in a suitable range are shown in green, the others in grey (**Figure 2B**).

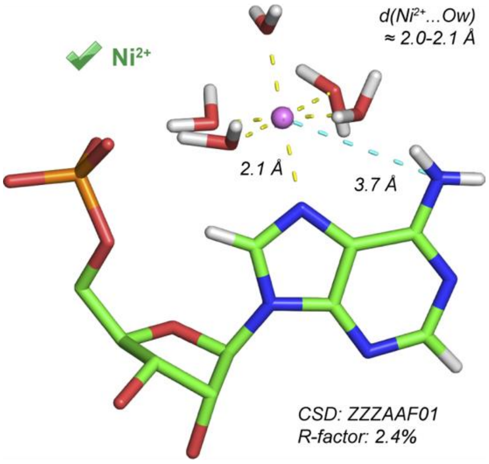


Figure S4. Ni$(H_2O)_5^{2+}$ bound to an adenine (CSD structure). For further data on transition metals binding to N7 of purines, see (65).

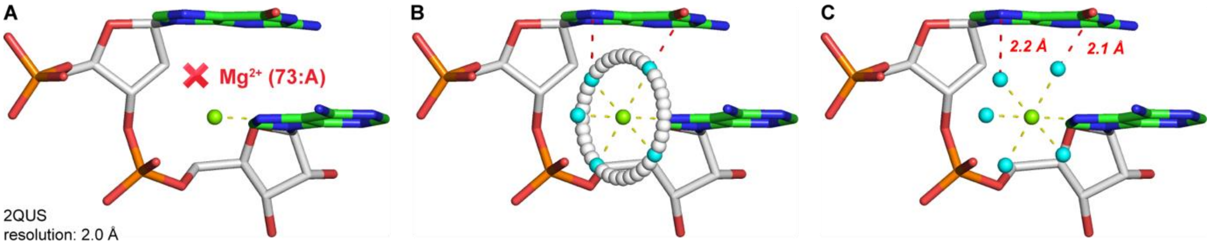


Figure S5. Clashes associated with the modeling of Mg^{2+} hydration shells. (**A**) In the **2QUS** structure, a Mg^{2+} ion has been placed at 2.0 Å from the N7 atom of an adenine. (**B**) A complete hydration shell “set” with a 5° increment has been caculated for this ion. The one with the smallest chashes is shown in cyan in (**B**) and (**C**).

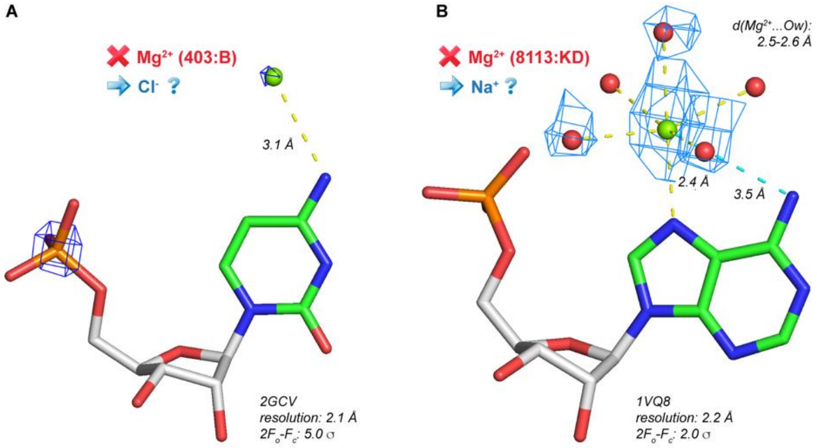


Figure S6. Two Mg^{2+} misidentifications. (**A**) Based on coordination distances, this assigned Mg^{2+} ion should be replaced by a Cl^- ion (66,67). (**B**) This Mg^{2+} ion should be replaced by Na^+, given coordination distances in the 2.4-2.6 Å range.

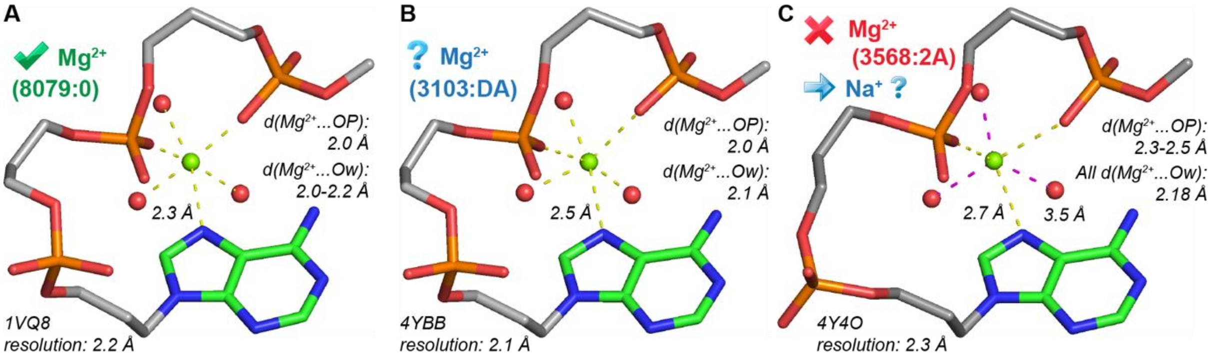


Figure S7. Ion binding to N7 and two OP atoms. (**A**) In the *H. marismortui* structure, the coordination distance are compatible with Mg^{2+}. (**B**) In the *E. coli* structure, the *d*(Mg^{2+}…N7) = 2.5 Å coordination distance is not comptible with Mg^{2+}. The distances to water might have been restrained. (**C**) In the *T. thermophilus* structure, the coordination distances strongly suggest the presence of Na^+ rather than Mg^{2+}. The ion to water distances restrained to 2.18 Å are shown in magenta).

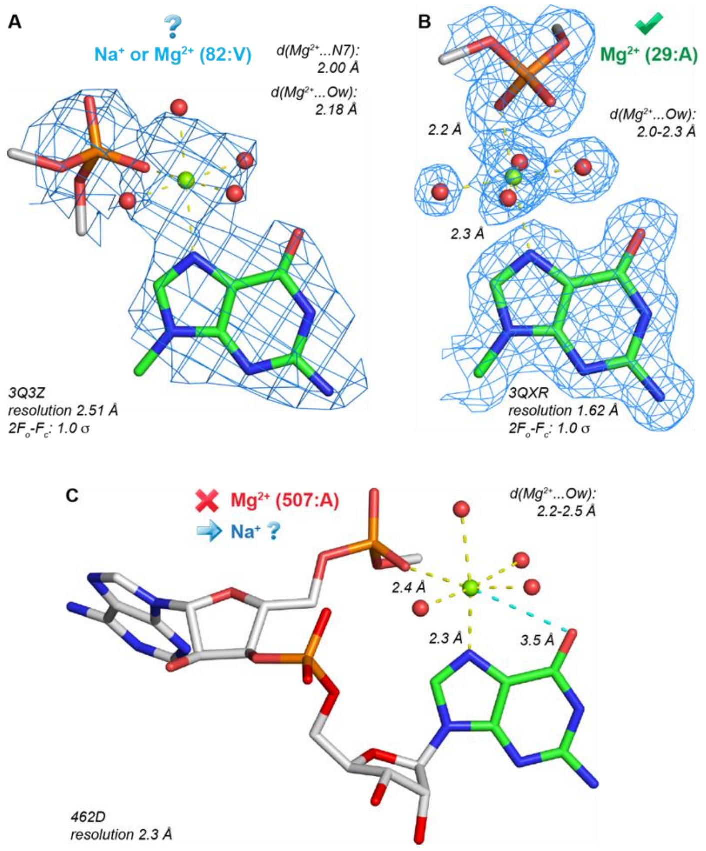


Figure S8. Ion binding to N7 and one OP atoms. (**A**) The OP atom binds in *cis* to Mg^{2+}. However, the use of *d*(Mg^{2+}…Ow) = 2.18 Å restraints prevents its unambiguous distinction from Na^{+}. (**B**) Same as **Figure 7B**. The OP atom binds in *trans*. (**C**) For this Mg^{2+} binding, the coordination distances around 2.4 Å and *d*(Mg^{2+}…OP) > *d*(Mg^{2+}…N7) suggest the binding of Na^{+}.

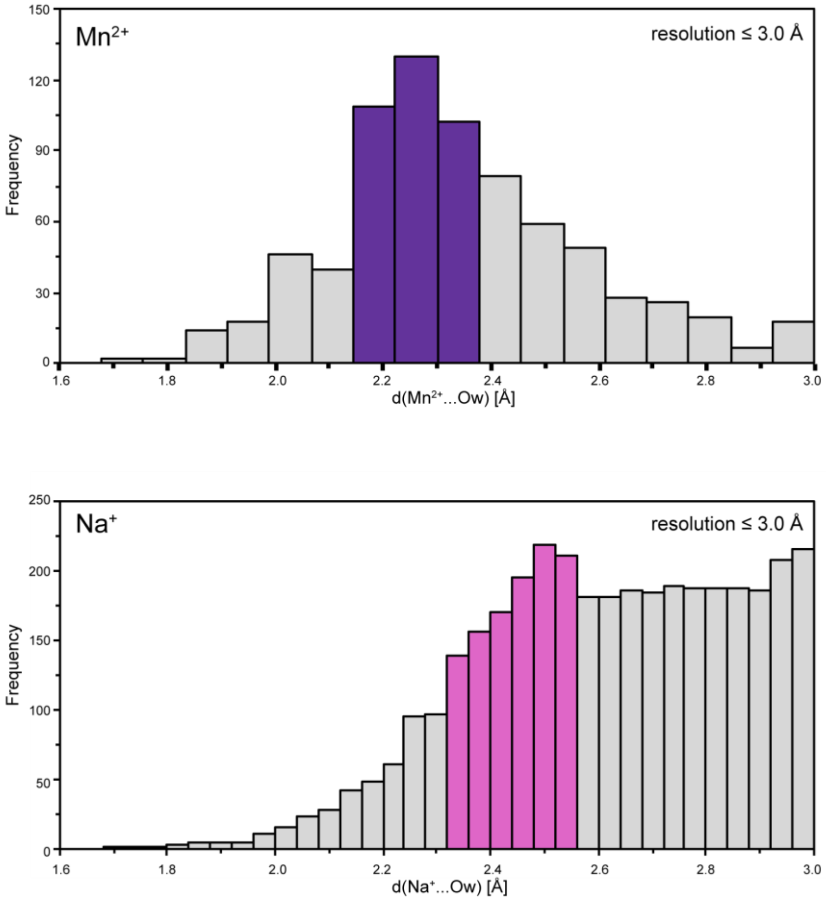


Figure S9. *d*(Mn^{2+}/Na^+…Ow) histograms derived from the PDB emphasizing the absence of systematic restraints for modelling the hydration shell of these metals (May 2016; resolution ≤ 3.0 Å).

References

1. Petrov, A.S., Bernier, C.R., Hershkovits, E., Xue, Y., Waterbury, C.C., Hsiao, C., Stepanov, V.G., Gaucher, E.A., Grover, M.A., Harvey, S.C. *et al.* (2013) Secondary structure and domain architecture of the 23S and 5S rRNAs. *Nucleic Acids Res.*, **41**, 7522-7535.
2. Clemons, W.M., Jr., Brodersen, D.E., McCutcheon, J.P., May, J.L., Carter, A.P., Morgan-Warren, R.J., Wimberly, B.T. and Ramakrishnan, V. (2001) Crystal structure of the 30 S ribosomal subunit from Thermus thermophilus: purification, crystallization and structure determination. *J. Mol. Biol.*, **310**, 827-843.
3. Wimberly, B.T., Brodersen, D.E., Clemons, W.M., Jr., Morgan-Warren, R.J., Carter, A.P., Vonrhein, C., Hartsch, T. and Ramakrishnan, V. (2000) Structure of the 30S ribosomal subunit. *Nature*, **407**, 327-339.
4. McLellan, T.J., Marr, E.S., Wondrack, L.M., Subashi, T.A., Aeed, P.A., Han, S., Xu, Z., Wang, I.K. and Maguire, B.A. (2009) A systematic study of 50S ribosomal subunit purification enabling robust crystallization. *Acta Cryst.*, **D65**, 1270-1282.
5. von Bohlen, K., Makowski, I., Hansen, H.A., Bartels, H., Berkovitch-Yellin, Z., Zaytzev-Bashan, A., Meyer, S., Paulke, C., Franceschi, F. and Yonath, A. (1991) Characterization and preliminary attempts for derivatization of crystals of large ribosomal subunits from Haloarcula marismortui diffracting to 3 A resolution. *J. Mol. Biol.*, **222**, 11-15.
6. Ban, N., Nissen, P., Hansen, J., Moore, P.B. and Steitz, T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, **289**, 905-920.
7. Klein, D.J., Moore, P.B. and Steitz, T.A. (2004) The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. *J. Mol. Biol.*, **340**, 141-177.
8. Blaha, G., Gurel, G., Schroeder, S.J., Moore, P.B. and Steitz, T.A. (2008) Mutations outside the anisomycin-binding site can make ribosomes drug-resistant. *J. Mol. Biol.*, **379**, 505-519.
9. Dunkle, J.A., Wang, L., Feldman, M.B., Pulk, A., Chen, V.B., Kapral, G.J., Noeske, J., Richardson, J.S., Blanchard, S.C. and Cate, J.H. (2011) Structures of the bacterial ribosome in classical and hybrid states of tRNA binding. *Science*, **332**, 981-984.
10. Pulk, A. and Cate, J.H. (2013) Control of ribosomal subunit rotation by elongation factor G. *Science*, **340**, 1235970.
11. Ben-Shem, A., Jenner, L., Yusupova, G. and Yusupov, M. (2010) Crystal structure of the eukaryotic ribosome. *Science*, **330**, 1203-1209.
12. Ben-Shem, A., Garreau de Loubresse, N., Melnikov, S., Jenner, L., Yusupova, G. and Yusupov, M. (2011) The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science*, **334**, 1524-1529.
13. Khatter, H., Myasnikov, A.G., Mastio, L., Billas, I.M., Birck, C., Stella, S. and Klaholz, B.P. (2014) Purification, characterization and crystallization of the human 80S ribosome. *Nucleic Acids Res.*, **42**, e49.
14. Carter, A.P., Clemons, W.M., Brodersen, D.E., Morgan-Warren, R.J., Wimberly, B.T. and Ramakrishnan, V. (2000) Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature*, **407**, 340-348.
15. Ogle, J.M., Murphy, F.V., Tarry, M.J. and Ramakrishnan, V. (2002) Selection of tRNA by the ribosome requires a transition from an open to a closed form. *Cell*, **111**, 721-732.
16. Murphy, F.V.t., Ramakrishnan, V., Malkiewicz, A. and Agris, P.F. (2004) The role of modifications in codon discrimination by tRNA(Lys)UUU. *Nat. Struct. Mol. Biol.*, **11**, 1186-1191.
17. Weixlbaumer, A., Murphy, F.V.t., Dziergowska, A., Malkiewicz, A., Vendeix, F.A., Agris, P.F. and Ramakrishnan, V. (2007) Mechanism for expanding the decoding capacity of transfer RNAs by modification of uridines. *Nat. Struct. Mol. Biol.*, **14**, 498-502.
18. Dunham, C.M., Selmer, M., Phelps, S.S., Kelley, A.C., Suzuki, T., Joseph, S. and Ramakrishnan, V. (2007) Structures of tRNAs with an expanded anticodon loop in the decoding center of the 30S ribosomal subunit. *RNA*, **13**, 817-823.
19. Kurata, S., Weixlbaumer, A., Ohtsuki, T., Shimazaki, T., Wada, T., Kirino, Y., Takai, K., Watanabe, K., Ramakrishnan, V. and Suzuki, T. (2008) Modified uridines with C5-methylene substituents at the first position of the tRNA anticodon stabilize U.G wobble pairing during decoding. *J. Biol. Chem.*, **283**, 18801-18811.
20. Vendeix, F.A., Murphy, F.V.t., Cantara, W.A., Leszczynska, G., Gustilo, E.M., Sproat, B., Malkiewicz, A. and Agris, P.F. (2012) Human tRNA(Lys3)(UUU) is pre-structured by natural modifications for cognate and wobble codon binding through keto-enol tautomerism. *J. Mol. Biol.*, **416**, 467-485.
21. Perez-Fernandez, D., Shcherbakov, D., Matt, T., Leong, N.C., Kudyba, I., Duscha, S., Boukari, H., Patak, R., Dubbaka, S.R., Lang, K. *et al.* (2014) 4'-O-substitutions determine selectivity of aminoglycoside antibiotics. *Nat. Commun.*, **5**, 3112.
22. Harms, J.M., Wilson, D.N., Schluenzen, F., Connell, S.R., Stachelhaus, T., Zaborowska, Z., Spahn, C.M. and Fucini, P. (2008) Translational regulation via L11: molecular switches on the ribosome turned on and off by thiostrepton and micrococcin. *Mol. Cell.*, **30**, 26-38.
23. Kaminishi, T., Schedlbauer, A., Fabbretti, A., Brandi, L., Ochoa-Lizarralde, B., He, C.G., Milon, P., Connell, S.R., Gualerzi, C.O. and Fucini, P. (2015) Crystallographic characterization of the ribosomal binding site and molecular mechanism of action of Hygromycin A. *Nucleic Acids Res.*, **43**, 10015-10025.
24. Klein, D.J., Schmeing, T.M., Moore, P.B. and Steitz, T.A. (2001) The kink-turn: a new RNA secondary structure motif. *EMBO J..* **20**, 4214-4221.
25. Hansen, J.L., Ippolito, J.A., Ban, N., Nissen, P., Moore, P.B. and Steitz, T.A. (2002) The structures of four macrolide antibiotics bound to the large ribosomal subunit. *Mol. Cell.*, **10**, 117-128.
26. Hansen, J.L., Schmeing, T.M., Moore, P.B. and Steitz, T.A. (2002) Structural insights into peptide bond formation. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 11670-11675.
27. Hansen, J.L., Moore, P.B. and Steitz, T.A. (2003) Structures of five antibiotics bound at the peptidyl transferase center of the large ribosomal subunit. *J. Mol. Biol.*, **330**, 1061-1075.
28. Schmeing, T.M., Moore, P.B. and Steitz, T.A. (2003) Structures of deacylated tRNA mimics bound to the E site of the large ribosomal subunit. *RNA*, **9**, 1345-1352.
29. Schmeing, T.M., Huang, K.S., Kitchen, D.E., Strobel, S.A. and Steitz, T.A. (2005) Structural insights into the roles of water and the 2' hydroxyl of the P site tRNA in the peptidyl transferase reaction. *Mol. Cell.*, **20**, 437-448.
30. Schmeing, T.M., Huang, K.S., Strobel, S.A. and Steitz, T.A. (2005) An induced-fit mechanism to promote peptide bond formation and exclude hydrolysis of peptidyl-tRNA. *Nature*, **438**, 520-524.

31. Schroeder, S.J., Blaha, G., Tirado-Rives, J., Steitz, T.A. and Moore, P.B. (2007) The structures of antibiotics bound to the E site region of the 50 S ribosomal subunit of Haloarcula marismortui: 13-deoxytedanolide and girodazole. *J. Mol. Biol.*, **367**, 1471-1479.
32. Schroeder, S.J., Blaha, G. and Moore, P.B. (2007) Negamycin binds to the wall of the nascent chain exit tunnel of the 50S ribosomal subunit. *Antimicrob. Agents Chemother.*, **51**, 4462-4465.
33. Kavran, J.M. and Steitz, T.A. (2007) Structure of the base of the L7/L12 stalk of the Haloarcula marismortui large ribosomal subunit: analysis of L11 movements. *J. Mol. Biol.*, **371**, 1047-1059.
34. Ippolito, J.A., Kanyo, Z.F., Wang, D., Franceschi, F.J., Moore, P.B., Steitz, T.A. and Duffy, E.M. (2008) Crystal structure of the oxazolidinone antibiotic linezolid bound to the 50S ribosomal subunit. *J. Med. Chem.*, **51**, 3353-3356.
35. Zhou, J., Bhattacharjee, A., Chen, S., Chen, Y., Duffy, E., Farmer, J., Goldberg, J., Hanselmann, R., Ippolito, J.A., Lou, R. *et al.* (2008) Design at the atomic level: design of biaryloxazolidinones as potent orally active antibiotics. *Bioorg. Med. Chem. Lett.*, **18**, 6175-6178.
36. Tu, D., Blaha, G., Moore, P.B. and Steitz, T.A. (2005) Structures of MLSBK antibiotics bound to mutated large ribosomal subunits provide a structural explanation for resistance. *Cell*, **121**, 257-270.
37. Simonovic, M. and Steitz, T.A. (2008) Peptidyl-CCA deacylation on the ribosome promoted by induced fit and the O3 '-hydroxyl group of A76 of the unacylated A-site tRNA. *RNA*, **14**, 2372-2378.
38. Gurel, G., Blaha, G., Moore, P.B. and Steitz, T.A. (2009) U2504 determines the species specificity of the A-site cleft antibiotics: the structures of tiamulin, homoharringtonine, and bruceantin bound to the ribosome. *J. Mol. Biol.*, **389**, 146-156.
39. Gabdulkhakov, A., Nikonov, S. and Garber, M. (2013) Revisiting the Haloarcula marismortui 50S ribosomal subunit model. *Acta Cryst.*, **D69**, 997-1004.
40. Selmer, M., Dunham, C.M., Murphy, F.V., Weixlbaumer, A., Petry, S., Kelley, A.C., Weir, J.R. and Ramakrishnan, V. (2006) Structure of the 70S ribosome complexed with mRNA and tRNA. *Science*, **313**, 1935-1942.
41. Korostelev, A., Asahara, H., Lancaster, L., Laurberg, M., Hirschi, A., Zhu, J., Trakhanov, S., Scott, W.G. and Noller, H.F. (2008) Crystal structure of a translation termination complex formed with release factor RF2. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 19684-19689.
42. Stanley, R.E., Blaha, G., Grodzicki, R.L., Strickler, M.D. and Steitz, T.A. (2010) The structures of the anti-tuberculosis antibiotics viomycin and capreomycin bound to the 70S ribosome. *Nat. Struct. Mol. Biol.*, **17**, 289-293.
43. Polikanov, Y.S., Blaha, G.M. and Steitz, T.A. (2012) How hibernation factors RMF, HPF, and YfiA turn off protein synthesis. *Science*, **336**, 915-918.
44. Demeshkina, N., Jenner, L., Westhof, E., Yusupov, M. and Yusupova, G. (2012) A new understanding of the decoding principle on the ribosome. *Nature*, **484**, 256-259.
45. Tourigny, D.S., Fernandez, I.S., Kelley, A.C. and Ramakrishnan, V. (2013) Elongation factor G bound to the ribosome in an intermediate state of translocation. *Science*, **340**, 1235490.
46. Maehigashi, T., Dunkle, J.A., Miles, S.J. and Dunham, C.M. (2014) Structural insights into +1 frameshifting promoted by expanded or modification-deficient anticodon stem loops. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 12740-12745.
47. Bulkley, D., Brandi, L., Polikanov, Y.S., Fabbretti, A., O'Connor, M., Gualerzi, C.O. and Steitz, T.A. (2014) The antibiotics dityromycin and GE82832 bind protein S12 and block EF-G-catalyzed translocation. *Cell Rep.*, **6**, 357-365.
48. Chen, Y., Feng, S., Kumar, V., Ero, R. and Gao, Y.G. (2013) Structure of EF-G-ribosome complex in a pretranslocation state. *Nat. Struct. Mol. Biol.*, **20**, 1077-1084.
49. Polikanov, Y.S., Steitz, T.A. and Innis, C.A. (2014) A proton wire to couple aminoacyl-tRNA accommodation and peptide-bond formation on the ribosome. *Nat. Struct. Mol. Biol.*, **21**, 787-793.
50. Gagnon, M.G., Lin, J., Bulkley, D. and Steitz, T.A. (2014) Crystal structure of elongation factor 4 bound to a clockwise ratcheted ribosome. *Science*, **345**, 684-687.
51. Polikanov, Y.S., Osterman, I.A., Szal, T., Tashlitsky, V.N., Serebryakova, M.V., Kusochek, P., Bulkley, D., Malanicheva, I.A., Efimenko, T.A., Efremenkova, O.V. *et al.* (2014) Amicoumacin A inhibits translation by stabilizing mRNA interaction with the ribosome. *Mol. Cell*, **56**, 531-540.
52. Polikanov, Y.S., Szal, T., Jiang, F., Gupta, P., Matsuda, R., Shiozuka, M., Steitz, T.A., Vazquez-Laslop, N. and Mankin, A.S. (2014) Negamycin interferes with decoding and translocation by simultaneous interaction with rRNA and tRNA. *Mol. Cell*, **56**, 541-550.
53. Lin, J., Gagnon, M.G., Bulkley, D. and Steitz, T.A. (2015) Conformational changes of elongation factor G on the ribosome during tRNA translocation. *Cell*, **160**, 219-227.
54. Rozov, A., Demeshkina, N., Westhof, E., Yusupov, M. and Yusupova, G. (2015) Structural insights into the translational infidelity mechanism. *Nat. Commun.*, **6**, 7251.
55. Polikanov, Y.S., Melnikov, S.V., Soll, D. and Steitz, T.A. (2015) Structural insights into the role of rRNA modifications in protein synthesis and ribosome assembly. *Nat. Struct. Mol. Biol.*, **22**, 342-344.
56. Seefeldt, A.C., Graf, M., Perebaskine, N., Nguyen, F., Arenz, S., Mardirossian, M., Scocchi, M., Wilson, D.N. and Innis, C.A. (2016) Structure of the mammalian antimicrobial peptide Bac7(1-16) bound within the exit tunnel of a bacterial ribosome. *Nucleic Acids Res.*, **44**, 2429-2438.
57. Polikanov, Y.S., Starosta, A.L., Juette, M.F., Altman, R.B., Terry, D.S., Lu, W., Burnett, B.J., Dinos, G., Reynolds, K.A., Blanchard, S.C. *et al.* (2015) Distinct tRNA accommodation intermediates observed on the ribosome with the antibiotics hygromycin A and A201A. *Mol. Cell*, **58**, 832-844.
58. Roy, R.N., Lomakin, I.B., Gagnon, M.G. and Steitz, T.A. (2015) The mechanism of inhibition of protein synthesis by the proline-rich peptide oncocin. *Nat. Struct. Mol. Biol.*, **22**, 466-469.
59. Rozov, A., Demeshkina, N., Khusainov, I., Westhof, E., Yusupov, M. and Yusupova, G. (2016) Novel base-pairing interactions at the tRNA wobble position crucial for accurate reading of the genetic code. *Nat. Commun.*, **7**, 10457.
60. Gagnon, M.G., Roy, R.N., Lomakin, I.B., Florin, T., Mankin, A.S. and Steitz, T.A. (2016) Structures of proline-rich peptides bound to the ribosome reveal a common mechanism of protein synthesis inhibition. *Nucleic Acids Res.*, **44**, 2439-2450.
61. Noeske, J., Huang, J., Olivier, N.B., Giacobbe, R.A., Zambrowski, M. and Cate, J.H. (2014) Synergy of streptogramin antibiotics occurs independently of their effects on translation. *Antimicrob. Agents Chemother.*, **58**, 5269-5279.
62. Wasserman, M.R., Pulk, A., Zhou, Z., Altman, R.B., Zinder, J.C., Green, K.D., Garneau-Tsodikova, S., Cate, J.H. and Blanchard, S.C. (2015) Chemically related 4,5-linked aminoglycoside antibiotics drive subunit rotation in opposite directions. *Nat. Commun.*, **6**, 7896.

63. Noeske, J., Wasserman, M.R., Terry, D.S., Altman, R.B., Blanchard, S.C. and Cate, J.H. (2015) High-resolution structure of the *Escherichia col*i ribosome. *Nat. Struct. Mol. Biol.*, **22**, 336-341.
64. Garreau de Loubresse, N., Prokhorova, I., Holtkamp, W., Rodnina, M.V., Yusupova, G. and Yusupov, M. (2014) Structural basis for the inhibition of the eukaryotic ribosome. *Nature*, **513**, 517-522.
65. Leonarski, F., D'Ascenzo, L. and Auffinger, P. (2016) Binding of metal ions to purine N7 atoms and implications for nucleic acids: A CSD survey. *Inorg. Chim. Acta*, **452**, 82-89.
66. Auffinger, P., Bielecki, L. and Westhof, E. (2004) Anion binding to nucleic acids. *Structure*, **12**, 379-388.
67. D'Ascenzo, L. and Auffinger, P. (2016) Anions in nucleic acid crystallography. *Methods Mol. Biol.*, **1320**, 337-351.

5.2.4 Further remarks and outlook on Mg^{2+} binding to nucleic acids

The idea behind the CSD (**Paper 4**) and PDB survey (**Paper 5**) was to propose a general picture of Mg^{2+} binding to nucleic acids, considering all possible nucleotide binding sites. However, the numerous issues found during this endeavor prompted us to focus on just few nucleotide positions at a time, highlighting all the structural pitfalls that are associated with the assignment of coordination of Mg^{2+} to these atoms. The ultimate goal remains to write a comprehensive paper about the rules, issues and biological implications of Mg^{2+} binding to nucleotides. In this perspective, these papers are meant to be a part of a larger story. An investigation of Mg^{2+} binding to nucleobase carbonyl oxygen atoms is currently ongoing in our group, with analogous premises, methodology and approach. A final exploration of Mg^{2+} binding to phosphate oxygen atoms will be likely the last chapter of this fascinating but still incomplete tale.

5.3 Asp/Glu side chains and “free” anions interacting with nucleic acids

Besides cationic metals, a diverse ensemble of anionic species can be localized by structural analysis in the nucleic acid first hydration shell, directly interacting with nucleotides (Auffinger et al. 2004). Among these anions, our focus has been on inorganic molecules such as halides, sulphate and negatively charged amino acid side chains (Asp and Glu). A surprising outcome of these surveys is that anions have been found not only to bind to hydrogen-bond donor sites, but also to hydrogen bond acceptor sites such as phosphate oxygen atoms. These occurrences can be explained by considering that anions assume unexpected protonation states under physiological conditions.

Aspartate and glutamate side chains are involved in the recognition networks between proteins and nucleic acids. These two amino acids bear a carboxylate group, negatively charged under physiological conditions. In order to obtain a statistical view of the Asp/Glu binding with nucleic acids, we searched the PDB for crystallographic structures of nucleic acid-protein complexes at resolution ≤ 2.5 Å. A total of 1844 structures were included in the search. The binding events were defined considering the *syn* versus *anti* conformer of carboxyl and carboxylate groups, collectively identified as *carboxyl(ate)* (**Fig. 5.1**).

A variable cutoff for hydrogen bonds was applied, considering the data on short interaction distances between carboxyl(ate) and oxygen atoms (**Paper 6**). A 2.8 Å cutoff was used for interactions involving phosphate groups, while interactions with other nucleotide atoms were searched with a 3.2 Å cutoff. The survey took into account the anionic as well as the neutral form of Asp/Glu side chains, interacting with nucleotides by a single or two hydrogen bonds. The latter case is associated with the formation of nucleobase-carboxyl(ate) pseudo-pairs (Kondo and Westhof 2011). The results sorted by base and inferred protonation state of Asp/Glu side chain are reported in **Table 5.1**.

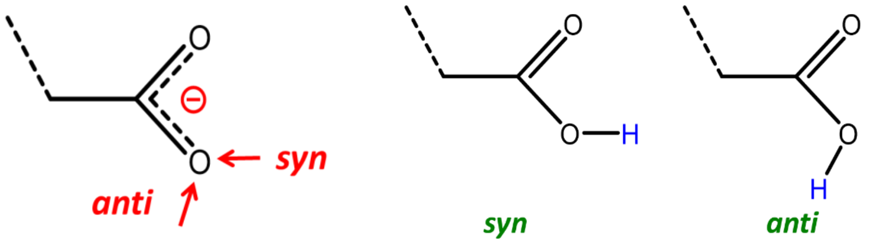


Figure 5.1 Carboxyl(ate) groups *syn* and *anti* conformers. Schematic illustration of the *syn* versus *anti* lone pair on a carboxylate group (red) and the *syn* versus *anti* hydrogen for a carboxyl group (green).

Table 5.1. Asp/Glu interactions with nucleic acids. 1844 PDB structures of protein/nucleic acid complexes at resolution ≤ 2.5 Å were surveyed (August 2014). For all contacts the pseudo-base pair (with two hydrogen bonds), the *syn* and the *anti* (one hydrogen bond each) interaction modes are reported. Variable distance cutoff were used depending on the nature of the nucleotide atom. Number of occurrences refer to non-reundant set.

	Asp/Glu carboxylate			**Asp/Glu carboxyl (protonated)**			**Asp/Glu total**		
	pseudo- pair	*syn*	*anti*	pseudo-pair	*syn*	*anti*	pseudo-pair	*syn*	*anti*
G	51	31	33	-	-	1	51	31	34
A	-	5	4	8	2	2	8	7	6
C	-	45	41	4	2	-	4	47	41
U/T	-	8	18	-	1	3	-	9	21
PO_4	-	-	-	-	**24**	**10**	-	24	10
Total	51	89	96	12	29	16	63	118	112

The most intriguing events of Asp/Glu binding are the ones involving hydrogen bond interactions with anionic phosphate oxygen atoms, thus inferring the occurrence of a protonated carboxyl group. In term of non-redundant occurrences, these contacts are among the most represented. A further structural analysis uncovered a large number of instances showing ill-defined contacts in term of electron density map quality or presence of metal ions in the proximity that blurred the picture. Few trustworthy cases were found in the vicinity of the active site of endonuclease enzymes, showing short interaction distances ~ 2.6 Å. An example of a Glu contacting a phosphate in the complex between a homing endonuclease and its target DNA (Marcaida et al. 2008) is presented in **Fig. 5.2**.

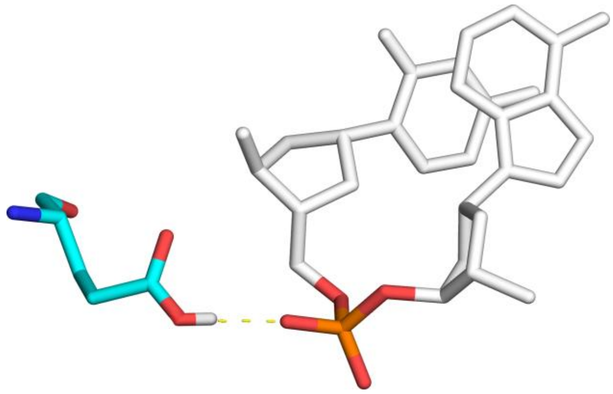
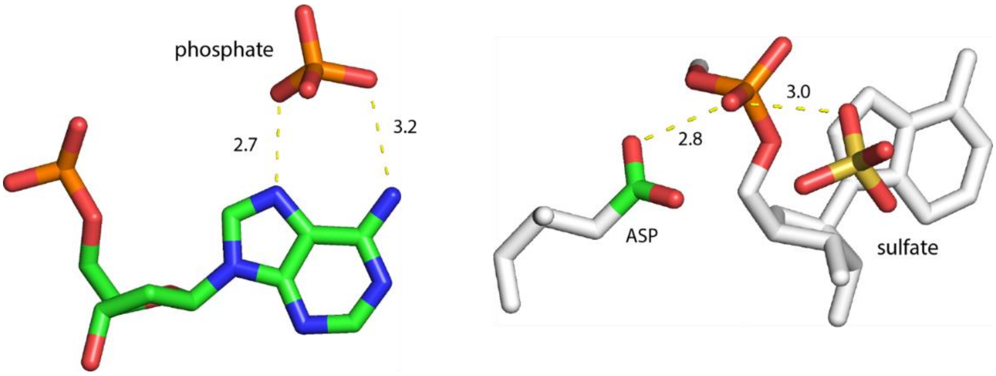


Figure 5.2 Protonated Glu in the active site of a homing endonuclease. Glu 117 (chain G) makes a short hydrogen bond with an anionic phosphate oxygen belonging to the DNA target helix. The inferred protonated state of the carboxyl group is highlighted by adding a hydrogen atom, not visible in the crystallographic structure (PDB: 2VS7; res.: 2.0 Å; Marcaida et al. 2008).

5.3.1 Review 2. *Anions in nucleic acid crystallography* (Methods Mol. Biol. 2016)

Graphical abstract



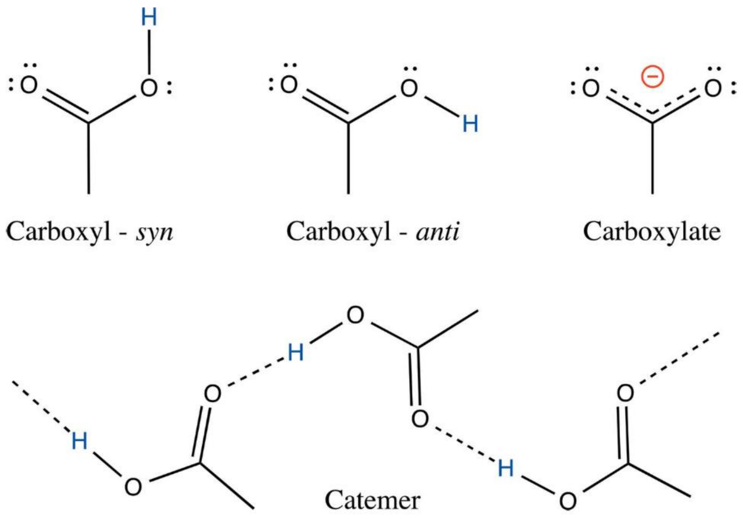
Anions are widely used in crystallization buffers as phasing agents, pH regulators and cation co-salts, but their inclusion in the first hydration shell of negatively charged nucleic acids seems counterintuitive. However, since a 2004 study performed in the group it is appreciated that anions can interact with nucleic acids in certain conditions. We review here the current state of anions binding to nucleic acids as observed in PDB structures, with considerations on their importance in nucleic acid crystallography as well as their biological relevance. Due to various identification and attribution issues, it is likely that many anions have been missed in the available structures (especially in the diffusive solvent bulk) and their occurrence is thus underestimated. To help with their identification and to avoid misinterpretations, we report data about their water coordination distances, as well as their coordination modes with nucleobases. A special attention is paid to the highly represented chloride and sulfate ions. Together with the coordination of anions with electropositive sites, protonation issues emerged for some instances of coordination with electronegative positions. In order to explain these instances, we found the possibility of adenine and cytosine protonation in several examples to be inconsistent and probably due to attribution inaccuracies. On the other hand, neutral or protonated anions are more likely to occur. The main message of the review is that understanding the anion binding properties should help to avoid mislabeling of electron densities and provide insight related to their potential effects in crowded cellular environments.

5.4 Anionic or neutral carboxyl(ate) interactions and their involvement in protein structures

The studies on carboxyl(ate) interactions during the search for protein-nucleic acids binding made us realize the importance of these functional groups for protein structure as well as crystal engineering. This interest culminated in a project of research in the CSD for carboxyl(ate) interacting with themselves and water molecules. The information yielded by this analysis have been extended to protein, where Asp/Glu pair can be found at key structural locations, such as the example on thaumatin that will be discussed later.

5.4.1 Paper 6. *A comprehensive classification and nomenclature of carboxyl-carboxyl(ate) supramolecular motifs and related catemers: implications for biomolecular systems* (Acta Cryst B, 2015)

Graphical abstract



Carboxyl and carboxylate groups can interact between themselves to form cyclic dimers and associate in many different ways through a single interlinking hydrogen bond to form specific supramolecular motifs. Further, they can form catemers that are polymeric-like chains formed by hydrogen bonded carboxylic groups in crystals. Through an exhaustive exploration of the Cambridge Structural Database (CSD) we reduced the apparently infinite number of single hydrogen bond arrangements involving these groups to 17 isolated carboxyl–carboxyl (13) and carboxyl–carboxylate (4) motifs. In addition, we found that only eight distinct catemer motifs involving repetitive combinations of *syn* and *anti* carboxyl groups can be formed. Statistical data related to the occurrence and conformational preferences of these motifs are presented along with data related to the strength of the hydrogen bonds they can form. We show that interaction distances are shorter for hydrogen bonds involving carboxylate groups (~2.5 vs ~2.7 Å), pointing towards a stronger type of hydrogen bond when a charged and/or electron delocalized species is implied. This distance difference between neutral and charged species can be used to infer protonation states in crystallographic biomolecular structures. Additionally, such strong hydrogen bonds are found in proteins where Asp/Glu amino acids form recurrent carboxyl–carboxylate motifs that are part of complex interaction networks playing a role in structure and folding. We consequently present data emphasizing how the exploration of small molecules can help understanding larger and more complex biomolecular systems.

STRUCTURAL SCIENCE
CRYSTAL ENGINEERING
MATERIALS

ISSN 2052-5206

A comprehensive classification and nomenclature of carboxyl–carboxyl(ate) supramolecular motifs and related catemers: implications for biomolecular systems

Luigi D'Ascenzo and Pascal Auffinger*

Architecture et Réactivité de l'ARN, Université de Strasbourg, Institut de Biologie Moléculaire et Cellulaire du CNRS, 67084 Strasbourg, France. *Correspondence e-mail: p.auffinger@ibmc-cnrs.unistra.fr

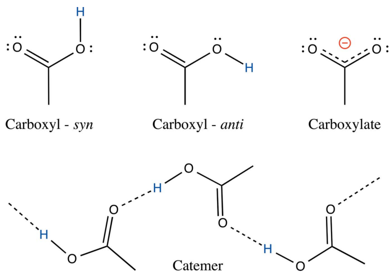
Received 9 November 2014
Accepted 8 February 2015

Edited by A. D. Bond, University of Copenhagen, Denmark

Keywords: supramolecular motifs; biomolecular systems; crystal engineering; pharmaceuticals.

Supporting information: this article has supporting information at journals.iucr.org/b

Carboxyl and carboxylate groups form important supramolecular motifs (synthons). Besides carboxyl cyclic dimers, carboxyl and carboxylate groups can associate through a single hydrogen bond. Carboxylic groups can further form polymeric-like catemer chains within crystals. To date, no exhaustive classification of these motifs has been established. In this work, 17 association types were identified (13 carboxyl–carboxyl and 4 carboxyl–carboxylate motifs) by taking into account the *syn* and *anti* carboxyl conformers, as well as the *syn* and *anti* lone pairs of the O atoms. From these data, a simple rule was derived stating that only eight distinct catemer motifs involving repetitive combinations of *syn* and *anti* carboxyl groups can be formed. Examples extracted from the Cambridge Structural Database (CSD) for all identified dimers and catemers are presented, as well as statistical data related to their occurrence and conformational preferences. The inter-carboxyl(ate) and carboxyl(ate)–water hydrogen-bond properties are described, stressing the occurrence of very short (strong) hydrogen bonds. The precise characterization and classification of these supramolecular motifs should be of interest in crystal engineering, pharmaceutical and also biomolecular sciences, where similar motifs occur in the form of pairs of Asp/Glu amino acids or motifs involving ligands bearing carboxyl(ate) groups. Hence, we present data emphasizing how the analysis of hydrogen-containing small molecules of high resolution can help understand structural aspects of larger and more complex biomolecular systems of lower resolution.



1. Introduction

Carboxyl and carboxylate [written collectively as carboxyl(ate)] groups are found in a large variety of biomolecular compounds and also in drugs and synthetic molecular systems. For the former, the two Asp and Glu amino acids represent ~2% of the ~2 million amino acids found in the Protein Data Bank (PDB, November 2014 release; Berman *et al.*, 2000). For the latter, they assemble to form essential supramolecular synthons recurrently used in crystal engineering (Desiraju, 2007, 2013; Merz & Vasylyeva, 2010) and are present in ~37 000 (~5–6%) of the ~675 000 crystal structures in the Cambridge Structural Database (CSD Version 5.35, November 2013; see Table 1; Allen, 2002; Chisholm *et al.*, 2006; Groom & Allen, 2014).

Despite the fact that carboxyl groups figure among the best investigated hydrogen-bond functionalities (Huggins, 1936; Leiserowitz, 1976; Berkovitch-Yellin & Leiserowitz, 1982; Steiner, 2001, 2002; Das & Desiraju, 2006; Rodríguez-Cuamatzi *et al.*, 2007), no systematic classification of carboxyl–carboxyl motifs is currently available. This is also true, but to a

Table 1
Number of structures in the CSD (Version 5.35, November 2013) containing at least one carboxyl, carboxylate or metal-bound carboxyl(ate) group and number of structures with low *R*-factor values ($R \le 0.05$). Disordered, error-containing, polymeric and powder structures were excluded from the search.

	All	$R \le 0.05$
Carboxyl	14 452	8254
Carboxylate	9283	5446
Metal-bound carboxyl	492	305
Metal-bound carboxylate	13 438	9082
Total	37 665	23 087

lesser extent, for carboxyl–carboxylate interaction modes (Wohlfahrt, 2005; Rodríguez-Cuamatzi *et al.*, 2007; Langkilde *et al.*, 2008). Indeed, the latter interaction types are essential in biology where numerous close contacts between Asp/Glu side chains have been reported (Gandour, 1981; Sawyer & James, 1982; Ramanadham *et al.*, 1993; Flocco & Mowbray, 1995; Torshin *et al.*, 2003; Wohlfahrt, 2005; Langkilde *et al.*, 2008).

For synthetic carboxyl dimers, the most common interaction mode is the centrosymmetric cyclic dimer, but numerous other dimers involving a single interlinking hydrogen bond have been characterized. Interestingly, some of these dimers can form catemers (Fig. 1), defined as infinite one-dimensional patterns involving their carboxyl groups (Leiserowitz, 1976; Berkovitch-Yellin & Leiserowitz, 1982; Kuduva *et al.*, 1999; Beyer & Price, 2000; Das *et al.*, 2005; Das & Desiraju, 2006; DeVita Dufort *et al.*, 2007; Rodríguez-Cuamatzi *et al.*, 2007; Saravanakumar *et al.*, 2009; Sanphui *et al.*, 2013). A complete classification of catemer motifs is also currently missing.

The formation of carboxyl(ate) dimers and further of carboxyl catemer motifs implies the involvement of common *syn* but also less common *anti* conformers, as well as the *syn* and/or *anti* lone pairs of the O atoms (Görbitz & Etter, 1992*a*; Das *et al.*, 2005; Das & Desiraju, 2006; Sanphui *et al.*, 2013; Fig. 1). Theoretical studies have investigated the relative stability of the *syn* and *anti* conformers. It is generally accepted that in the gas phase, the *syn* conformer is favoured over the *anti* conformer by 21.4–28.9 kJ mol^{-1} depending on the theoretical level and basis set used in quantum chemical calculations (Kamitakahara & Pranata, 1995; Sato & Hirata, 1999; Nagy, 2013). In aqueous solution, the estimated relative energy difference between the two conformers is reduced to 7.12 kJ mol^{-1} (Nagy, 2013). A further point of interest involves the relative basicity of the *syn* and *anti* lone pairs of carboxylate groups. Theoretical studies have reported that although the *syn* lone pairs are intrinsically more basic, the basicity difference decreases and even cancels out when environmental effects are taken into consideration (Li & Houk, 1989; Allen & Kirby, 1991; Gao & Pavelites, 1992). In line with these data, a significant number of catemer motifs involving *anti* conformers have been observed in various crystallographic surroundings, supporting the fact that environmental effects are able to reverse anticipated conformational equilibria (Das & Desiraju, 2006). *Anti* conformers have also been considered in drug discovery strategies involving bioisosterism (McKie *et al.*, 2008; Meanwell, 2011; Allen *et al.*, 2012).

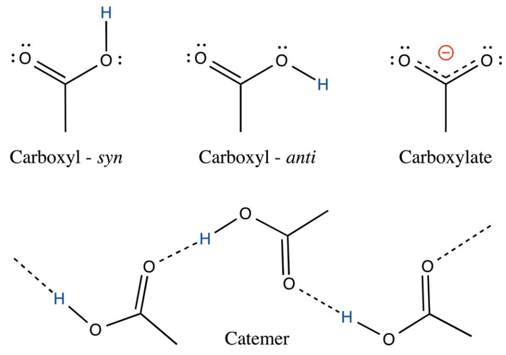


Figure 1
Carboxyl(ate) groups (*syn* and *anti* conformers) and schematic structure of a catemeric chain. The *syn* and *anti* lone pairs of the three carboxyl(ate) O atoms are marked by double dots.

Given the importance of these carboxyl–carboxyl(ate) dimers in both the chemical and biochemical realms, the present study aims at:

(i) providing an exhaustive classification of all possible dimers and catemers involving these groups;

(ii) proposing a systematic nomenclature for them;

(iii) defining recurrent hydrogen-bond properties.

This study should contribute to an improved understanding of the structural diversity observed in small-molecule crystal structures, and provide insights into crystal engineering of new materials (Desiraju, 2007, 2013), including pharmaceutical co-crystals (Blagden *et al.*, 2007). However, the main incentive of the study resides in acquiring reliable statistical data that will help to understand carboxyl(ate) interactions in biomolecular systems. In this respect, analysing small-molecule crystal structures, where H atoms are systematically observed, has a clear edge over exploring biomolecular systems where H atom positions are rarely reported (Ahmed *et al.*, 2007; Fisher *et al.*, 2012).

2. Methods

The Cambridge Structural Database (CSD Version 5.35, November 2013) was searched for structures containing carboxyl–carboxyl(ate) motifs by using explicit H-atom positions. All searches were performed with the *ConQuest* software (Bruno *et al.*, 2002) using filters so that error-containing, polymeric and powder structures were excluded, as well as structures marked as disordered. Although H-atom disorder is common in carboxylic systems, structures where the H atom could not be unambiguously assigned to a single O atom were not considered (Leiserowitz, 1976; Berkovitch-Yellin & Leiserowitz, 1982; Wilson *et al.*, 1996; Das *et al.*, 2005; Thomas *et al.*, 2010; Hursthouse *et al.*, 2011). This criterion leads to exclusion of 12 out of the 23 catemers listed by Das & Desiraju (2006). However, Steiner (2001) reported that statistics were not affected by excluding disordered structures. The searches were also restricted to structures with low *R*-factor values ($R \le 0.05$) unless otherwise specified. Metal-bound carboxyl(ate)

Table 2
Number of structures and fragments containing carboxyl(ate) groups in the CSD.

Only low R-factor structures ($R \leq 0.05$) are considered. Statistics were also collected for the *anti* conformer subgroups that take into account the carboxyl groups that are involved in intra- and intermolecular hydrogen bonds, respectively. Distances are in Å, angles in °.

	No. of structures	No. of fragments	d(C=O)	d(C—O)	θ(O=C—O)	θ(C—C=O)	θ(C—C—O)	θ(C—C—OH)
Carboxyl-*syn*	6852	9295	1.22 ± 0.02	1.31 ± 0.02	124 ± 1	123 ± 2	113 ± 2	111 ± 3
Carboxyl-*anti* (intermolecular)	209	223	1.21 ± 0.01	1.31 ± 0.02	121 ± 2	122 ± 2	117 ± 2	112 ± 4
Carboxyl-*anti* (intramolecular)	760	945	1.22 ± 0.01	1.30 ± 0.02	121 ± 1	120 ± 2	118 ± 2	110 ± 4
Carboxylate	5353	6760	1.25 ± 0.02	–	125 ± 2	117 ± 2	–	–

groups were excluded given their specific structural features (Hocking & Hambley, 2005). Note that the November 2006 CSD release contained less than 2/3 of the structures found in the November 2013 release. Thus, the present searches significantly extend those presented in earlier publications on smaller samples of structures (Kuduva *et al.*, 1999; DeVita Dufort *et al.*, 2007; Langkilde *et al.*, 2008).

Since carboxyl(ate) groups are involved in strong hydrogen bonds (Jeffrey, 1997; Steiner, 2001; Langkilde *et al.*, 2008), a stringent hydrogen-bond cut-off criterion could be used ($O\cdots O \leq 2.8$ Å). The H-atom positions were not considered for analysing hydrogen-bond lengths since their position is systematically unreliable when not derived from neutron diffraction experiments (Vishweshwar *et al.*, 2004; Allen & Bruno, 2010). Neutron diffraction surveys provide an average 1.018 Å (Allen & Bruno, 2010) or even a 1.070 Å value (Vishweshwar *et al.*, 2004) for the carboxyl O—H bond length, compared with an average of 0.87 Å derived from our survey. Hence, we have not used H atoms in the subsequent analysis, except for obviously differentiating carboxyl from carboxylate groups and for defining the *syn*/*anti* character of the former. An incidental advantage of not using H atoms is that our defined criteria can be used in biomolecular systems where H atoms are rarely characterized.

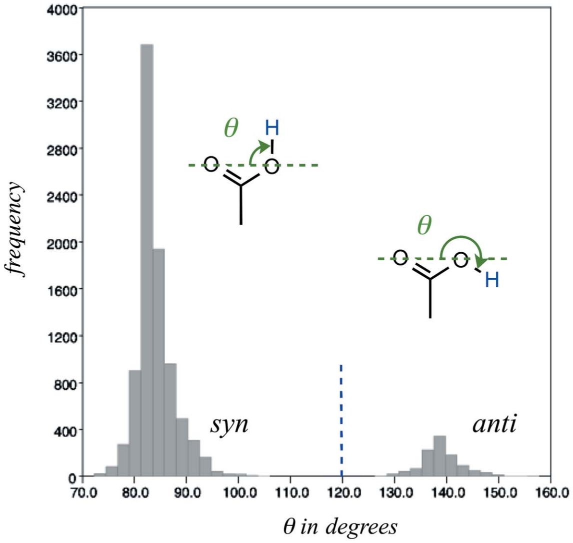


Figure 2
Geometric parameters used for separating the carboxyl *syn* and *anti* conformers. The *syn* conformers are defined by a θ value below 120° (marked by a blue dashed line on the histogram; θ corresponds to the O···O—H angle). The *anti* conformers are defined by a θ value greater than 120°. The histogram has been derived from an ensemble of low R-factor ($R \leq 0.05$) carboxylic acid containing structures.

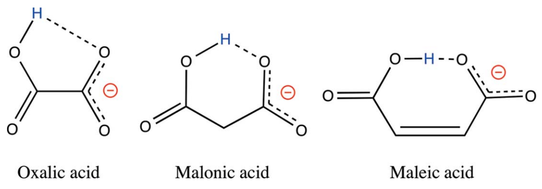
The geometric parameters used to distinguish the *syn* and *anti* conformers of the carboxyl groups and the spatial *syn* and *anti* arrangement of carboxyl–carboxyl(ate) dimers are detailed in §3.1. Specific criteria were used to exclude a few borderline and possibly error-containing structures. For instance, the WEGBUH structure (Ying, 2012) contains a short (2.58 Å) interaction between two O atoms of the carboxylic hydroxyl groups that corresponds rather to a carboxyl–carboxylate than to a carboxyl–carboxyl motif. Similarly, a significant number of structures are excluded where the H atoms are located out of the O=C—O plane by more than 0.4 Å.

The results of the searches were analysed using *Vista* (CCDC, 1994), and all structures were visualized using *Mercury* CSD Version 3.3 (Macrae *et al.*, 2008). Catemer structures were individually examined and classified. The possibility that some of the presented catemer motifs could belong to large rings rather than infinite chains was considered and excluded.

3. Results

3.1. Carboxyl and carboxylate groups

Carboxylic acids bear a proton that is commonly found in the *syn* and more rarely in the *anti* conformation. In order to distinguish between the *syn* and *anti* conformers, we imposed the following criterion on the O···O—H angle (θ) (Fig. 2). The *syn* conformer corresponds to θ angle values between 0 and 120°; the *anti* confirmer to θ angle values between 120 and 180°. The relative proportion of these conformers is roughly 9/1 in favour of *syn*, while negatively charged carboxylate groups represent about 2/3 of the total carboxyl groups (Table 2). The main geometric features of carboxyl(ate) groups are



Oxalic acid Malonic acid Maleic acid

Figure 3
Three dicarboxylic acids with an *anti* carboxyl group involved in an intramolecular hydrogen bond, schematically displayed under the CSD most represented mono-anion dicarboxylic acid form.

similar to those reported in an early study (Leiserowitz, 1976). Our updated values are reported in Table 2. Note that, due to its partial double-bond character, the C═O bond of carboxyl groups is shorter by ~ 0.11 Å than the adjacent C—O(H) hydroxyl bond.

The *anti* conformer population is more heterogeneous than the *syn* population since they are involved in a large diversity of intermolecular but also intramolecular bonds such as those observed in oxalic, malonic, maleic (Fig. 3) as well as phthalic acids. For the three former acids in their most represented mono-anion dicarboxylic acid form, the average *d*(O···O) hydrogen-bond distances are 2.67 ± 0.03 (10 structures), 2.46 ± 0.03 (20 structures) and 2.44 ± 0.03 Å (107 structures), stressing the formation of very short hydrogen bonds. Since the scope of this study is to examine supramolecular motifs, we eliminated from our searches all 'intramolecular' contacts involving an *anti* carboxyl conformer unless otherwise specified. When structures containing intramolecular hydrogen bonds were excluded, the number of fragments containing an *anti* carboxyl conformer decreased from 1168 to 223.

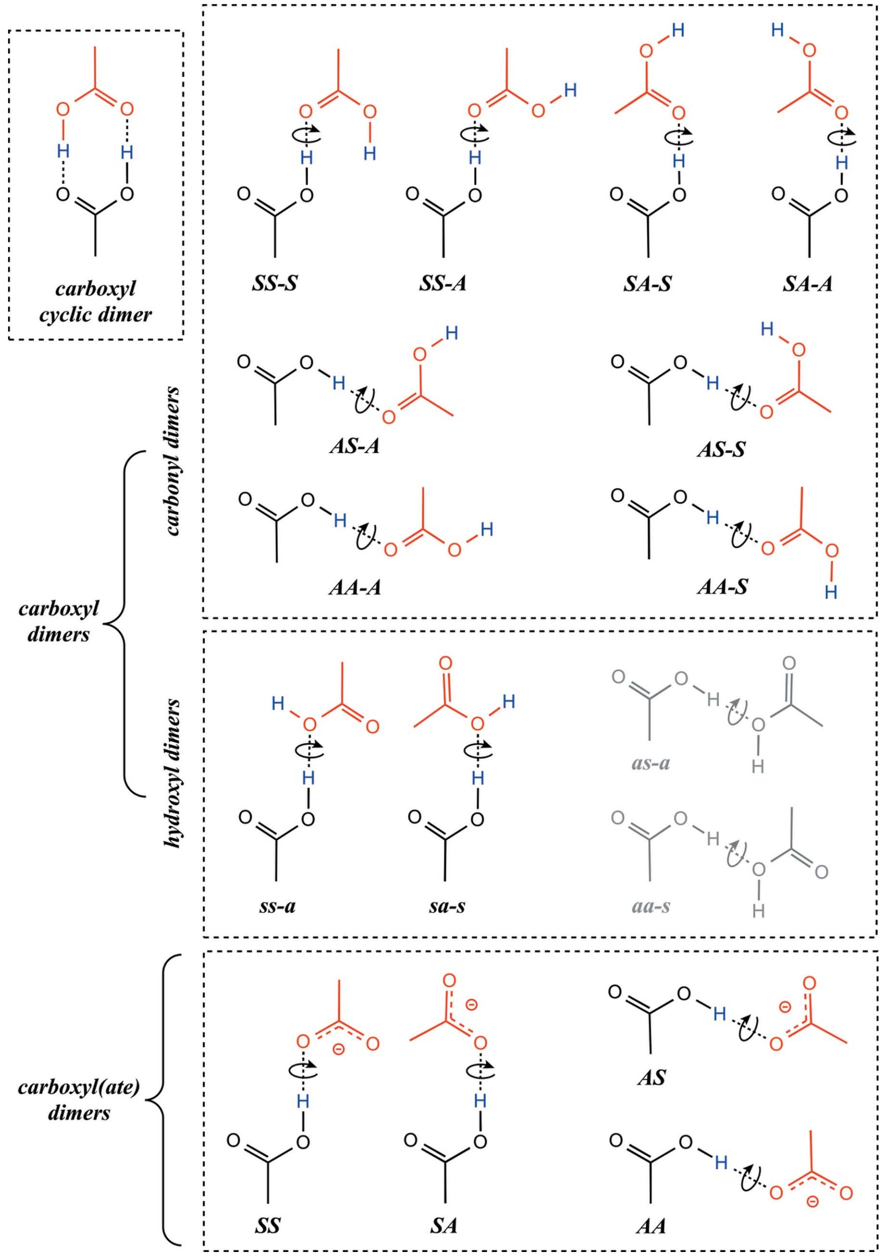


Figure 4
All 17 possible carboxyl–carboxyl(ate) dimers with accompanying nomenclature. The cyclic dimer is represented in the top left box; the eight '*carbonyl dimers*' involving a hydroxyl donor and a carbonyl acceptor group are represented in the top right box; the four '*hydroxyl dimers*' involving a donor and acceptor hydroxyl group are represented in the central box (the two *as-a* and *aa-s* dimers not identified in the CSD are shaded); the four carboxyl–carboxylate dimers are represented in the bottom box.

3.2. Carboxyl–carboxyl(ate) interactions

3.2.1. Nomenclature. An evaluation of carboxyl(ate) interaction modes based on the *syn*/*anti* carboxylic conformers and the *syn*/*anti* carboxyl(ate) lone pairs led to a total of 17 carboxyl–carboxyl(ate) dimers comprising: (i) one cyclic dimer; (ii) 12 carboxyl–carboxyl dimers involving a single hydrogen bond; (iii) 4 carboxyl(ate) dimers. Free rotation around the interlinking hydrogen bond is considered for all except the cyclic dimer (Fig. 4). The formation of three-centred or bifurcated hydrogen bonds was not considered since they do not appear in previous (Görbitz & Etter, 1992*b*) and current CSD surveys as well as in molecular dynamics simulations of formate and acetate ions in water (Payaka *et al.*, 2009, 2010). This simplifies considerably the presented nomenclature.

Sixteen interaction modes involve a single hydrogen bond linking the two units. We propose a three letter nomenclature for carboxyl–carboxyl dimers based on:

(i) the *syn* or *anti* conformer of the first carboxyl group that is by convention always the hydrogen-bond donor group of the dimer;

(ii) the *syn* or *anti* lone pair of the carbonyl hydrogen-bond acceptor group of the dimer;

(iii) the *syn* or *anti* conformer of the dimer hydrogen-bond acceptor group.

The first letter (*S* or *A*) corresponds to the *syn* or *anti* conformer; the second letter (*S* or *A*) to the lone pair involved in the hydrogen bond; the third letter (*S* or *A* separated by a dash from the two others) to the position of the H atom not involved in the hydrogen bond. For the eight dimers involving the participation of a

Table 3
Number of structures and fragments containing carboxyl–carboxyl(ate) dimers in the CSD.

	No. of structures†	No. of fragments†	d(O···O)‡
Carboxyl–carboxyl			
Cyclic dimer	1741 (2984)	1929 (3385)	2.65 ± 0.03
Carbonyl dimer			
SS-S	57 (91)	64 (98)	2.68 ± 0.04
SS-A	57 (80)	62 (88)	2.64 ± 0.05
SA-S	204 (333)	234 (378)	2.67 ± 0.05
SA-A	18 (25)	19 (26)	2.65 ± 0.06
AS-S	4 (6)	4 (6)	2.68 ± 0.05
AS-A	6 (7)	6 (7)	2.64 ± 0.02
AA-S	11 (15)	11 (16)	2.64 ± 0.04
AA-A	3 (3)	3 (3)	2.70 ± 0.04
Hydroxyl dimer			
ss-a	2 (7)	2 (7)	2.71
sa-s	6 (8)	6 (8)	2.76 ± 0.12
as-a	– (–)	– (–)	–
aa-s	– (–)	– (–)	–
Carboxyl–carboxylate			
SS	801 (1199)	947 (1429)	2.53 ± 0.05
SA	319 (492)	357 (554)	2.58 ± 0.05
AS	27 (48)	29 (52)	2.52 ± 0.06
AA	61 (102)	68 (117)	2.54 ± 0.06

† The number of structures and fragments are given for structures with low R-factors ($R \leq 0.05$). The number of structures and fragments derived from the entire CSD (no R-factor restrictions) are given in parentheses. ‡ Average distances (Å) calculated for the $R \leq 0.05$ subset.

carbonyl lone pair in the hydrogen bond ('*carbonyl dimers*'), capital letters are used. Lowercase letters are used for the four dimers involving the hydroxyl lone pair ('*hydroxyl dimers*'). A two capital-letter code suffices for the four carboxyl–carboxylate dimers.

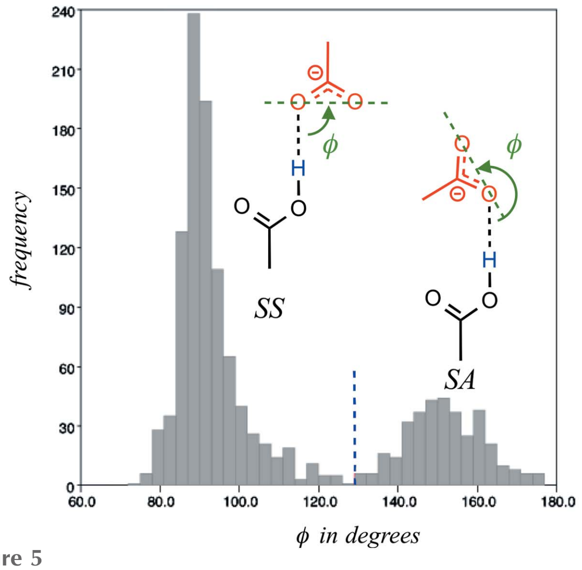


Figure 5
Geometric parameters used for separating carboxyl–carboxylate dimers involving *syn* or *anti* lone pairs. The histogram has been drawn for a sub-ensemble of *SS* and *SA* dimers. The *syn* conformers are defined by a ϕ value below 130° marked by a blue dashed line on the histogram; ϕ corresponds to the O(H)···O···O angle. The *anti* conformers are defined by a ϕ value greater than 130°.

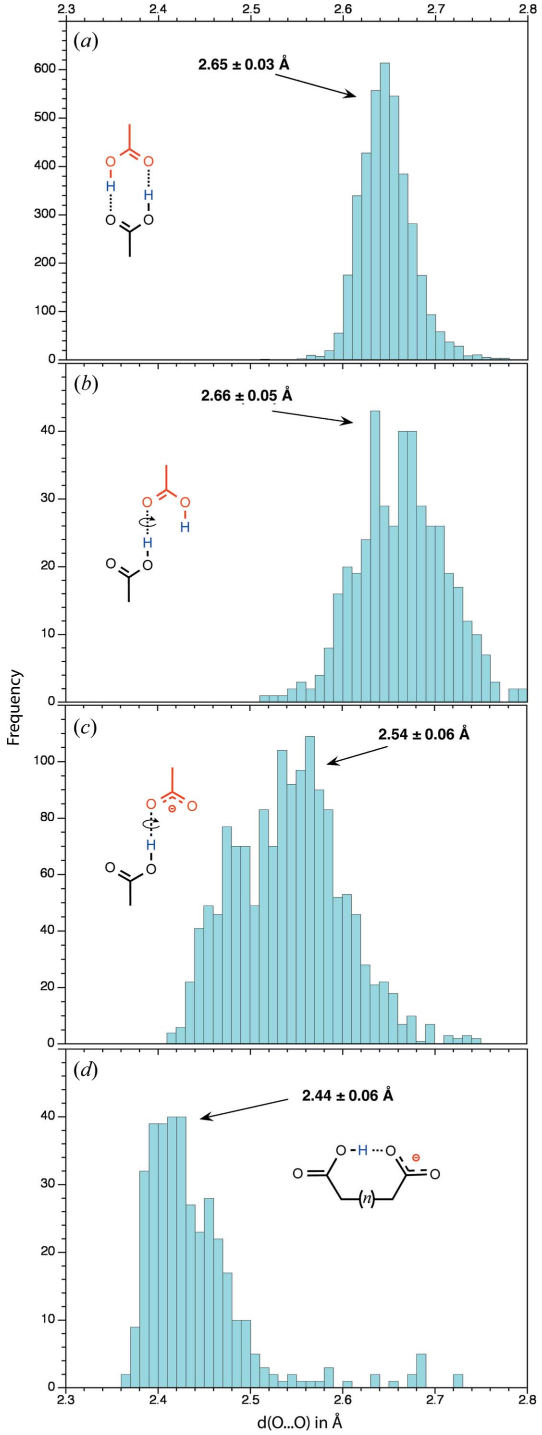


Figure 6
Histograms showing the distance distribution between the two O atoms involved in the interlinking hydrogen bond(s) for carboxyl–carboxyl(ate) dimer structures with low R-factors ($R \leq 0.05$). The arrows mark the average values. (*a*) d(O···O) histogram for the two carboxyl···carboxyl hydrogen bonds of the cyclic dimers. (*b*) d(O···O) histogram for the non-cyclic carboxyl···carboxyl hydrogen bonds. All *syn* and *anti* conformers are taken into account. (*c*) d(O···O) histogram for the carboxyl···carboxylate hydrogen bonds (intramolecular hydrogen bonds are not considered). All *syn* and *anti* conformers are taken into account. (*d*) d(O···O) histogram for the carboxyl···carboxylate intramolecular hydrogen bond found in mono-anion dicarboxylic acids (see for instance Fig. 3).

3.2.2. Geometric classification criteria. As noted above (Fig. 2), simple geometric criteria can be used to filter the carboxyl *syn* and *anti* conformers. It was less obvious how to discriminate dimers based on their *syn* or *anti* lone pair bonding types. After having tried several options, we found that the histograms showing the ϕ angle that corresponds to the O(H)···O···O angle involving the hydrogen-bond donor O atom and the two carboxylate O atoms are the most helpful to achieve such a goal. The histogram drawn for the carboxyl–carboxylate dimers is unambiguous and prompted us to use a 130° cut-off for isolating the *SS* and *AA* from the *SA* and *AS* carboxyl–carboxylate dimers, respectively (Fig. 5). Although a clear partition is difficult to identify on the *SS-S* dimer histogram (data not shown), a visualization of these dimers confirmed the soundness of the defined criteria. As is often the case, borderline conformations are observed and are difficult to eliminate but do not alter the inferred landscape.

3.2.3. Carboxyl–carboxyl interaction modes. *Cyclic dimer*: This dimer is undoubtedly the best represented in the CSD (Table 3). The distance between the O atoms involved in the hydrogen bond is on average close to 2.65 ± 0.03 Å (Fig. 6) and consequently shorter by 0.17 Å than the accepted $H_2O\cdots OH_2$ hydrogen-bond length (2.82 Å). Cyclic dimers are almost perfectly planar.

‘*Carbonyl dimers*’: Eight ‘*carbonyl dimer*’ types were identified (Table 3). The four types involving the *syn* conformer of the donor carboxyl group and among them, the *SA-S* dimers, are well represented. The *synplanar* rotamers are generally not observed except for the *SA-S* dimers where they are as prominent as *antiplanar* rotamers (Fig. 7). Note that *syn-* and *antiplanar* rotamers are defined by inter-dimer dihedral angles with values close to 0 and 180°, respectively (see, for example, Fig. 7*c*). The ACETAC09 acetic acid structure seems to be stabilized by a C—H···O interaction involving the methyl group, an orientation that is not found for chloroacetic acid in the CLACET01 structure and illustrates how weak interactions participate in structural networks.

Not surprisingly, the four dimer types involving the *anti* conformer of the donor carboxyl are rare. Among them, the *AA-S* dimer that involves the *anti* lone pair of a carbonyl group is best represented. However, convincing structures are found for each dimer type (Fig. 8). The hydrogen-bond length distribution is broader than the one given for the cyclic dimers, while the average hydrogen-bond length is roughly the same (2.66 ± 0.05 Å; Fig. 6).

‘*Hydroxyl dimers*’: Although the two carboxyl hydroxyl groups could form hydrogen bonds, this interaction occurs rarely. Only two *ss-a* and six *sa-s* conformers were characterized (Table 3; Fig. 9). None of the two other possible *as-a* and *aa-s* conformers were observed. This points to the fact that the lone pairs of carboxyl —OH groups seem to be much less basic and/or accessible to other carboxyl groups than the lone pairs of more common hydroxyl groups.

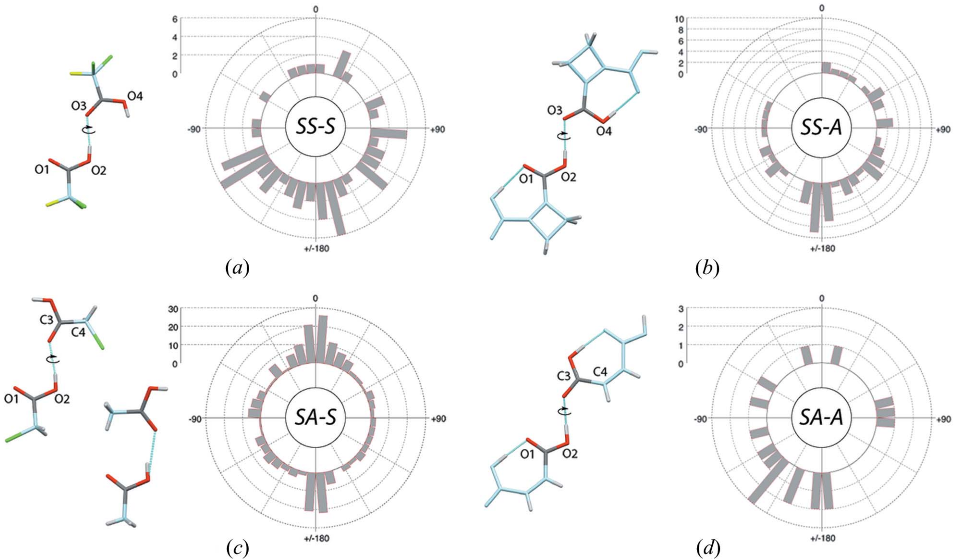


Figure 7
Carboxyl–carboxyl dimers involving a *syn* conformer and the lone pair of a carbonyl group (‘*carbonyl dimer*’) along with their rotamer distribution around the interlinking hydrogen bond for structures with $R \leq 0.05$. The C and O atoms not belonging to the interacting carboxyl groups are shown in light blue, F and Cl atoms are shown in yellow and green, respectively. (*a*) *Antiplanar SS-S* dimer (NAGVUM) and O1—O2—O3—O4 dihedral angle rotamer histogram. (*b*) *Antiplanar SS-A* dimer (CBUCDX01) and O1—O2—O3—O4 dihedral angle rotamer histogram. (*c*) *Antiplanar* and *synplanar SA-S* dimers (CLACET01 and ACETAC09) and O1—O2—C3—C4 dihedral angle rotamer histogram. (*d*) *Antiplanar SA-A* dimer (MALIAC12) and O1—O2—C3—C4 dihedral angle rotamer histogram.

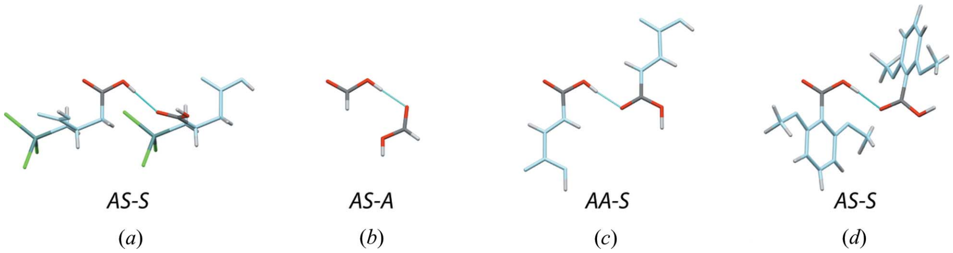


Figure 8
Carboxyl–carboxyl dimers involving an *anti* conformer and the lone pair of a carbonyl group (‘*carbonyl dimer*’). The C and O atoms not belonging to the interacting carboxyl groups are shown in light blue, Cl and Ge atoms are shown in green and dark green, respectively. (*a*) *AS-S* dimer (WOKPOC). (*b*) *AS-A* dimer (NEWXAO). (*c*) *AA-S* dimer involving two fumaric acid molecules (KACNAD). (*d*) *AA-A* dimer (DMOXEA01).

3.2.4. Carboxyl–carboxylate interaction modes. The *SS* dimer,

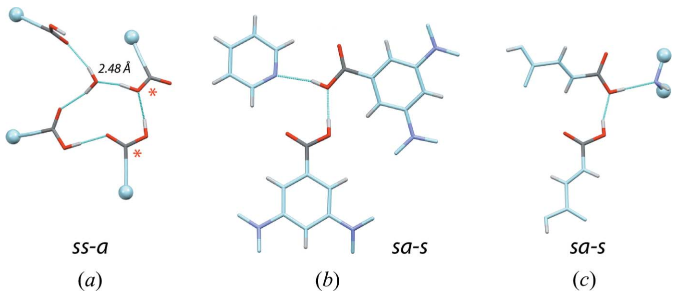


Figure 9
Rare carboxyl–carboxyl dimers involving the lone pair of the hydroxyl group (‘*hydroxyl dimers*’). The C and O atoms not belonging to the interacting carboxyl groups are shown in light blue, N atoms are shown in magenta. The light blue spheres indicate that the molecule has been truncated for visualization purposes. (*a*) *Antiplanar SS-A* dimer (CACTUW; $R = 0.04$). Due to the size of the system, only the interacting fragments are shown. The unusually short carboxyl–O*w* distance is given. The red asterisks mark the carboxyl groups involved in the *ss-a* dimer. (*b*) *Antiplanar sa-s* dimer (CAYJAO; $R = 0.06$). (*c*) *Synplanar sa-s* dimer involving two fumaric acid molecules (EMONAW; $R = 0.11$). The N-containing interacting molecule has been truncated due to its size.

involving a hydrogen bond between a *syn* hydroxyl group and a *syn* carboxylate lone pair, is the most prevalent carboxyl–carboxylate dimer in the CSD (Table 3). The *antiplanar SS* dimer is frequently observed while dimers close to the *synplanar* orientation are much less represented (Fig. 10). Some rare occurrences of the *synplanar* orientation stabilized by intervening groups (such as NH_4^+ in JEDPUE; see Fig. 10) are reported. In those instances, the distances between the O atoms not involved in the hydrogen bond exceed 3.0 Å.

All *SA* rotamers, involving a hydrogen bond between a *syn* hydroxyl group and an *anti* carboxylate lone pair, are nicely represented with some preference for the *antiplanar* orientations. The *AS* and *AA* dimers are less abundant but are still observed in a significant number of structures.

The most distinctive feature of these carboxyl–carboxylate dimers is related to the very short average hydrogen-bond distance between the two O atoms (2.54 ± 0.06 Å), which does not seem to be dependent on the dimer type (Fig. 6). The shortest observed hydrogen bonds (2.43 ± 0.04 Å) belong to intramolecular mono-anion dicarboxylic acids (Figs. 6 and 10).

3.2.5. Carboxyl(ate)–water hydrogen-bond length. The hydrogen-bond length between carboxyl(ate) groups and water molecules is strongly dependent on the acceptor or donor character of the former. When bound to the hydroxyl group, the average $d(\mathrm{C{-}O(H)}\cdots \mathrm{O}w)$ distance is 2.59 ± 0.06 Å (Fig. 11*a*); when bound to a carboxyl(ate) carbonyl group, the average $d(\mathrm{C{=}O}\cdots\mathrm{O}w)$ distance (2.77 ± 0.07 Å) becomes close to water hydrogen-bond distances (Figs. 11*b* and *c*). The shortest reported hydrogen-bond lengths are close to 2.4 Å. Such a short length is found in the CACTUW structure (Vishweshwar *et al.*, 2004), where the $(\mathrm{C{=}O})\mathrm{O{-}H}\cdots\mathrm{O}w$ distance is close to 2.48 Å and involves an *anti* conformer (Fig. 9*a*). Interestingly, only 44 water molecules establish a hydrogen bond with the lone pair of the carboxyl—OH group either in *syn* or *anti* (compared with the ~ 2800 water molecules found around the other groups), confirming its poor acceptor potential. The associated distances are close to 2.80 Å.

3.2.6. Catemers. *Nomenclature*: The dimer nomenclature can be adapted without major modifications to the catemer motifs for which two classes can be defined: (i) the *homo-catemers* involving the formation of a continuous chain of the same dimer and (ii) the *hetero-catemers* involving two alternating dimer types. In the latter case, we impose the convention that the *syn* conformer precedes the *anti* conformer. Thus, the *SS-A·AS-S* code should be used instead of the *AS-S·SS-A* code. In the current CSD release, four *homo-* and four *hetero-catemer* types were identified (Table 4 and Fig. 12).

Catemer formation rule: The *SS-S* and *SA-S homo-catemers* are the most represented followed by the *SS-A·AA-S hetero-catemers*. Three other catemers are poorly repre-

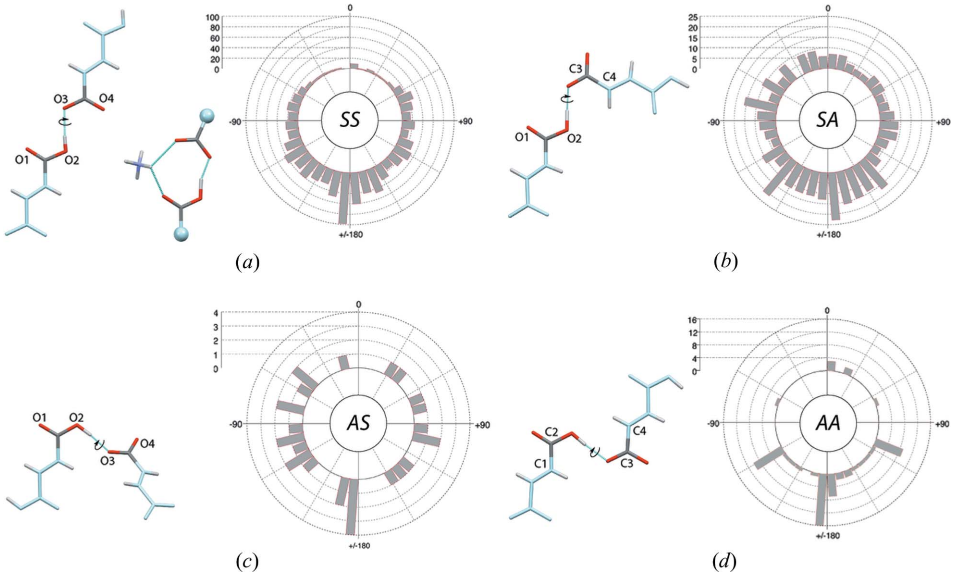


Figure 10
The four carboxyl–carboxylate dimer types and their rotamer distribution around the interlinking hydrogen bond for structures with $R \leq 0.05$. The C and O atoms not belonging to the interacting carboxyl or carboxylate groups are shown in light blue, N atoms are shown in magenta. (*a*) (Left) *Antiplanar SS* dimer involving two fumaric acid molecules (HUSSUJ). (Middle) *Synplanar SS* dimer (JEDPUE). An NH_4^+ molecule links the carboxyl(ate) groups. The light blue spheres indicate that the molecule has been truncated for visualization purposes. (Right) O1—O2—O3—O4 dihedral angle rotamer histogram. (*b*) *Antiplanar SA* dimer involving two fumaric acid molecules (CLEMAS) and O1—O2—C3—C4 dihedral angle rotamer histogram. (*c*) *Antiplanar AS* dimer involving two fumaric acid molecules (SEGSAZ) and O1—O2—O3—O4 dihedral angle rotamer histogram. (*d*) *Antiplanar AA* dimer involving two fumaric acid molecules (BAHLEC) and C1—C2—C3—C4 dihedral angle rotamer histogram.

Table 4
Number of catemer-containing structures in the CSD.

Only low *R*-factor structures ($R \leq 0.05$) are taken into account (see complete list in Table S1). Disordered, error-containing, polymeric and powder structures were excluded from the search.

	No. of structures
Homo-catemer	
SS-S	23
SA-S	67
AS-A	3
AA-A	3
Hetero-catemer	
SS-A·AS-S	1
SS-A·AA-S	17
SA-A·AS-S	1
SA-A·AA-S	7

sented but still present in the CSD. These catemers involve the eight ‘*carbonyl dimers*’ shown Fig. 4.

After closer examination of the catemer nomenclature (Table 4), a simple rule emerged. If the dimer starts with a *syn* or an *anti* conformer it should end with an identical conformer. Thus, the *SS-S*, *SA-S*, *AS-A* and *AA-A* dimers form *homo-catemers* since the first and the last conformers are identical, while the *SS-A*, *SA-A*, *AS-S* and *AA-S* dimers need to associate with a complementary motif and can only form *hetero-catemers*. According to this rule, all eight possible *homo-* and *hetero-catemer* combinations were identified in the CSD, although the *SS-A·AS-S* (ROZHEU; Dawid *et al.*, 2009) and *SA-A·AS-S* catemers (MEKLOE; Das & Desiraju, 2006) were identified in only one instance. Table S1 of the supporting information provides a list of all characterized catemers, which were manually checked to confirm that they are not part of large rings.

4. Discussion

4.1. A systematic classification of carboxyl–carboxyl(ate) dimers . . .

By using simple stereochemical considerations, we have demonstrated that the apparently overwhelming diversity of carboxyl–carboxyl(ate) dimers (Rodríguez-Cuamatzi *et al.*, 2007) can be reduced to 17 supramolecular motifs when considering free rotation around the interlinking hydrogen bond. A hierarchy of motifs emerged that distinguishes first the cyclic dimer (1929 fragment occurrences), followed by the *SS* (947 occurrences), *SA* (357 occurrences) and *SA-S* dimers (234 occurrences) (Table 3). The other dimers are less represented and some are rare, especially those in the ‘*hydroxyl dimer*’ class where the *as-a* and the *aa-s* types are absent from the current CSD release (Fig. 4). This latter observation is in agreement with the fact that strong donor groups such as carboxyl —OH functions are also poor acceptors, as reported in small molecules and biomolecular systems (Ramanadham *et al.*, 1993; Steiner, 2002).

The reasons as to why in certain circumstances, carboxyl groups prefer to form single hydrogen-bonded dimers extending sometimes into polymeric-like catemeric chains rather than cyclic dimers remains a subject of astonishment, although much has been written on this topic including considerations related to the preferential involvement of *syn* and *anti* lone pairs and conformers (Glusker, 1998; Sato & Hirata, 1999; Nagy, 2013).

In order to appreciate better these conformational preferences, statistical models predicting the number of hydrogen bonds that might form between any donor/acceptor pair in a crystal structure have been derived using CSD data (Allen *et al.*, 1999; Galek *et al.*, 2014) along with computational models providing estimates of their intrinsic stability (Dunitz & Gavezzotti, 2012). These studies confirmed the pre-eminence

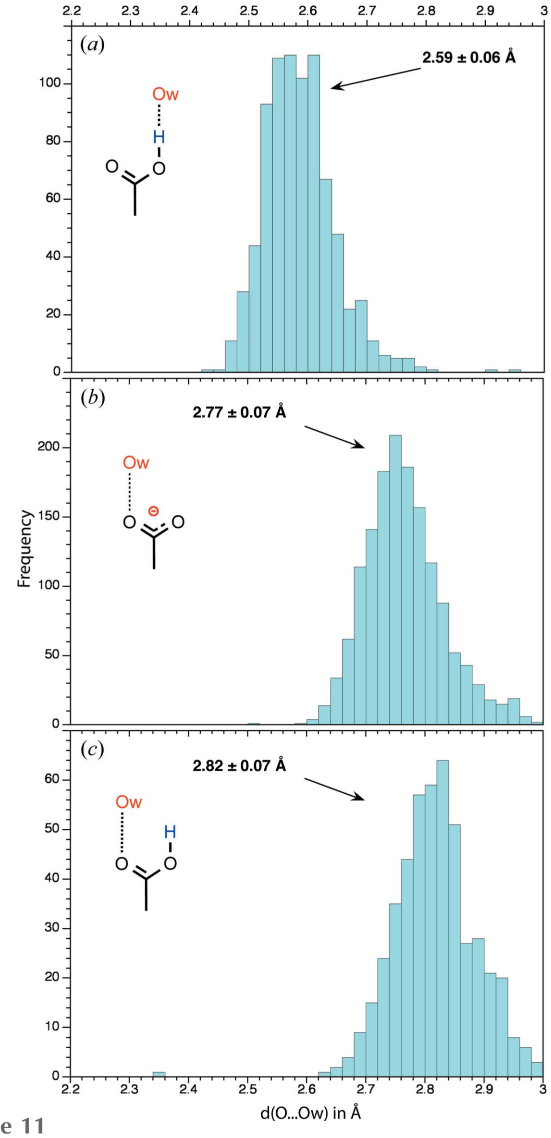


Figure 11
Histograms showing the distance distribution between the two O atoms directly involved in the carboxyl(ate)–water hydrogen bond. For clarity, only water molecules positioned in a 1 Å slice above and below the plane defined by the three heavy atoms of the carboxyl(ate) groups are considered. A cut-off of 2.2 Å for $d(\mathrm{C{=}O\cdots H{-}O}w)$ or $d(\mathrm{C{-}OH\cdots O}w)$ was used. (*a*) $d(\mathrm{C{-}OH\cdots O}w)$ histogram involving carboxyl groups. (*b*) $d(\mathrm{C{=}O\cdots O}w)$ histogram involving carboxylate groups. (*c*) $d(\mathrm{C{=}O\cdots O}w)$ histogram involving carbonyl O atoms of the carboxyl group.

of the cyclic dimer over other motifs. Although such approaches appear promising, they suffer from: (i) drawbacks related to the still noticeable lack of a sufficient number of crystal structures; (ii) the difficulty to take into account environmental effects; (iii) important approximations in the calculation of the interatomic forces at play in such complex systems. In this respect, non-additive contributions are especially difficult to estimate and quantum mechanical calculations confirmed that the energy gap between different motifs is small and lies within the precision limits of the methods (Meot-Ner *et al.*, 1999; Meot-Ner, 2012).

The most important factor to take into account is related to the strong competition of alternate binding motifs. Indeed, in CSD crystal structures, it was established that the probability of formation of dimers was around 30%, the remainder forming hydrogen bonds with a great variety of other acceptors (Steiner, 2001, 2002). Interestingly, unforeseen motifs are still brought to light. To cite only a few of them, new crystal forms of aspirin were recently published (Hursthouse *et al.*, 2011) and a crystallization study of a family of mono-substituted salicylic acid compounds reported an unexpectedly large diversity of motifs (Montis & Hursthouse, 2012). To understand the association rules of these supramolecular synthons and to be able to be truly predictive, we probably still have to expand current databases by orders of magnitude.

4.2. … and associated catemers

For catemers, we designed a simple rule derived from the carboxyl–carboxyl(ate) dimer nomenclature that postulates that only eight catemer motifs can be formed (Fig. 12). As for dimers, a catemer hierarchy exists, with the *SA-S* catemer being the most represented (Table 4). The possible origin of the less frequent formation of catemer motifs over the common cyclic dimer has been addressed by several authors and is of special interest in crystal engineering (Beyer & Price, 2000; Das & Desiraju, 2006; Sanphui *et al.*, 2013). Basically, the same factors involved in the preferential formation of one or the other dimer play a role here, namely steric factors, supporting C—H···O interactions and hydrogen-bond competition with various types of chemical groups in addition to specific stereoelectronic effects. These observations stress that intrinsic or local energetic considerations are not sufficient to describe the formation rules of these motifs (Leiserowitz, 1976; Berkovitch-Yellin & Leiserowitz, 1982; Kuduva *et al.*, 1999; Das & Desiraju, 2006; Hursthouse *et al.*, 2011).

As for dimers, new catemer patterns are still uncovered such as in the 1,2-phenylenedipropynoic acid where two carboxylic groups from the same molecule are involved in the formation of a *SA-A·AS-S* catemeric chain (unfortunately the structure was not deposited in the CSD; Saravanakumar *et al.*, 2009). Furthermore, recent examples of carboxylic acid catemer and dimer synthon polymorphs were reported (Gajda *et al.*, 2009; Sanphui *et al.*, 2013). Overall, we characterized 122 catemers that can be compared with the 73 catemers characterized from a survey of the April 1998 CSD (Kuduva *et al.*, 1999). Note that in this present study, we were able to categorize two particularly rare catemers observed in only one instance each (Table 4). This is fortunate since we believe to have now a complete structural sample of each of the eight possible homo- and hetero-catemer structures.

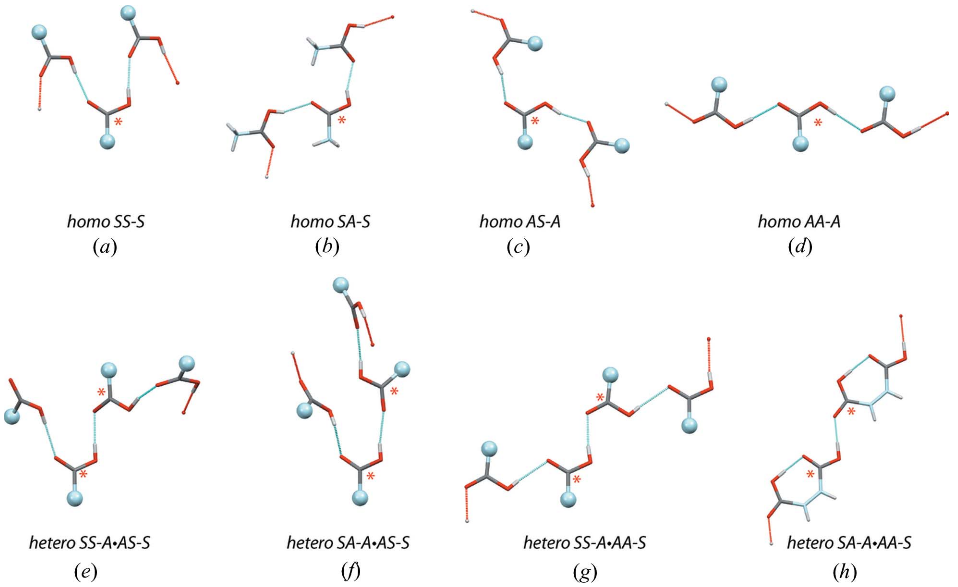


Figure 12
Examples of the eight catemer types identified in the CSD. The C and O atoms not belonging to the interacting carboxyl groups are shown in light blue. The white and red dots mark the position of the connected carboxylic groups in the catemeric chain. The red asterisks mark the carboxyl groups used for naming the catemer. The light blue spheres indicate that the molecule has been truncated for visualization purposes. (*a*) *SS-S* homo-catemer (XONNET); (*b*) *SA-S* homo-catemer (ACETAC07); (*c*) *AS-A* homo-catemer (GIMRAW); (*d*) *AA-A* homo-catemer (DMOXBA01); (*e*) *SS-A·AS-S* hetero-catemer (ROZHEU); (*f*) *SS-A·AA-S* hetero-catemer (WOKPOC); (*g*) *SA-A·AS-S* hetero-catemer (MEKLOE). (*h*) *SA-A·AA-S* hetero-catemer (MALIAC12).

4.3. Short hydrogen bonds

Besides these classification attempts, this study supports findings established in earlier surveys on smaller structural samples that hydrogen bonds involving carboxyl–carboxylate dimers are on the shorter and consequently stronger side of hydrogen bonds (Jeffrey & Saenger, 1991; Jeffrey, 1997; Steiner, 2001, 2002; Vishweshwar *et al.*, 2004; Langkilde *et al.*, 2008). It is beyond the scope of this paper to analyse the reasons as to why such short hydrogen bonds are formed. However, the topic of short or 'strong' hydrogen bonds involving amongst others the carboxyl(ate) groups found in proteins has received great attention especially since they were associated with enzymatic catalytic mechanisms (Perrin & Nielson, 1997; Katz *et al.*, 2002; Gilli & Gilli, 2009; Perrin, 2010; Hosur *et al.*, 2013) involving either the *syn* or *anti* lone pairs (Zimmerman *et al.*, 1991).

The carboxyl–carboxyl hydrogen bonds are generally considered as π-cooperative bonds or bonds belonging to the class of 'resonance-assisted hydrogen bonds' (RAHB; Vishweshwar *et al.*, 2004; Bertolasi *et al.*, 2006; Gilli & Gilli, 2009). In these motifs, the COOH donor is activated by π-cooperative hydrogen bonding (O—H$\cdots$O═C). The carboxyl–carboxylate hydrogen bonds that involve a bond between an acid and its conjugate base fall clearly in a different pool where the stabilizing effect is induced by the presence of the negative charge. These bonds are also called ionic hydrogen bonds (Steiner, 1999; Meot-Ner, 2012) or negatively 'charge-assisted hydrogen bonds' (CAHB; Vishweshwar *et al.*, 2004; Gilli & Gilli, 2009). They are on average $\sim$ 0.1 Å shorter than the RAHB hydrogen bonds (Fig. 6). This is particularly obvious when both groups have similar pK_a values as in protein structures where they play important structural and sometimes catalytic functions (Cleland & Kreevoy, 1994; Hosur *et al.*, 2013).

A third category of hydrogen bonds is found in mono-anion dicarboxylic compounds (Fig. 3). These intramolecular hydrogen bonds can be regarded as very short CAHBs given their average 2.43 Å distance (Fig. 6*d*). Consequently, they also belong to the strongest class of hydrogen bonds among those involving carboxyl(ate) groups. The shortening of the hydrogen bond is attributed to the presence of the electronegative O acceptor atom. They are probably further stabilized by some synergism due to increased π-delocalization facilitated by their intramolecular character (Perrin & Nielson, 1997). These dimers involve both the *anti* conformer and a carbonyl lone pair, supporting the view that the lone pair basicity scale might be essentially contextual. Further, these mono-anion dicarboxylic compounds are involved in the formation of at least two types of hetero-catemeric chains: (i) the *SA-A·AS-S* (Fig. 7*d*) and (ii) *SA-A·AA-S* types (Fig. 13*d*).

Rather counterintuitively, the shortest carboxyl(ate)–water hydrogen bonds involve the neutral carboxyl and not the charged carboxylate group (Fig. 11). Such short hydrogen bonds were analysed by density functional theory (Śmiechowski *et al.*, 2011; Brown *et al.*, 2012) and extensively discussed in a small-molecule neutron diffraction study where the authors were able to demonstrate the associated chain of polarization events (Vishweshwar *et al.*, 2004). The latter group observed that not only charge and resonance assistance can lead to very short intermolecular hydrogen bonds [$d(\text{O}\cdots\text{O}) \simeq 2.4$–2.5 Å], but polarization assistance must also be considered in terms of σ-cooperative stabilization (see Fig. 9*a*). These synergistic effects were named 'synthon-assisted hydrogen bonds' or SAHB (Brown *et al.*, 2012). Examples of such multi-centred short hydrogen bonds can also be found in biomolecular systems and might play a significant role at catalytic sites (Cleland & Kreevoy, 1994; Katz *et al.*, 2002).

4.4. Implications for biomolecular systems

Carboxyl dimers that involve simultaneous protonation of two Asp/Glu amino acids have not been reported in biomolecular systems, although carboxyl–carboxylate dimers appear to be relatively frequent in a wide pH range that can extend to 8.0 (Sawyer & James, 1982; Flocco & Mowbray, 1995; Torshin *et al.*, 2003; Wohlfahrt, 2005; Langkilde *et al.*, 2008). The formation of such interactions is surprising since it is generally assumed that given the pK_a of the Asp ($\sim$ 3.9) and Glu ($\sim$ 4.3) residues (Pace *et al.*, 2009), they would be deprotonated at physiological pH. As an outcome, carboxyl(ate) groups can form four different dimer types that extend to 16 when the two Asp/Glu amino-acid types are considered. However, since H-atom positions can rarely be observed in macromolecular systems, *SA* and *AS* dimers cannot be differentiated and this number reduces to nine due to degeneracy.

It was reported that the *SA*/*AS* arrangement is the most common in proteins (62%) followed by *SS* (24%) and *AA* (14%; Wohlfahrt, 2005), in contrast to the present study where the *SS* dimer dominates (Table 3). This originates probably from the better accessibility of the *anti* lone pairs of the Asp/Glu residues that are not shielded by large chemical groups, as is observed in a majority of CSD structures. However, it remains to be determined whether the *SA* or *AS* arrangements is favoured or if they are energetically not differentiable. In other words, if the *anti* conformer is preferred or not over the *syn* conformer or if these preferences are contextual as so often witnessed in all types of chemical systems. Theoretical calculations on model systems favour the *AS* arrangement (Wohlfahrt, 2005), while the present study identifies the *SA* arrangement as being the most frequent (Table 3).

To identify the protonated state of Asp/Glu residues in X-ray structures, efforts based on stereochemical factors have been made. The most obvious consideration relates to the hydrogen-bond proximity of two carboxyl(ate) O atoms, the associated distance being generally well below 2.7 Å (Sawyer & James, 1982; Ramanadham *et al.*, 1993; Flocco & Mowbray, 1995; Torshin *et al.*, 2003; Wohlfahrt, 2005; Langkilde *et al.*, 2008). The carboxyl C—O(H) and C═O bond lengths differ by $\sim$ 0.1 Å (Table 2) and the bond electron densities have also been exploited in the analysis of high-resolution protein structures ($\leq$ 1.3 Å), leading to clear identification of protonated Asp/Glu residues (Ahmed *et al.*, 2007; Fisher *et al.*, 2012). In the absence of good neutron diffraction structures (Ahmed *et al.*, 2007; Hosur *et al.*, 2013), such techniques could help to unscramble the degeneracy issue mentioned above. On a similar line of thought, short side-chain Asp/Glu carboxyl(ate) to O*w* distances could be used to infer protonation states of the residues (Ramanadham *et al.*, 1993).

5. Summary and perspectives

This work illustrates the diversity of supramolecular motifs generated by a single chemical group and offers a comprehensive carboxyl–carboxyl(ate) dimer and catemer nomenclature. As noted above:

(i) 17 possible carboxyl–carboxyl(ate) interaction modes including *syn* and *anti* conformers as well as carbonyl lone pairs were identified;

(ii) among them, the cyclic dimer is the most represented;

(iii) instances of all other possible interaction modes were found in the CSD, except the two *as-a* and *aa-s* '*hydroxyl dimers*';

(iv) based on this classification, eight catemeric types could be uniquely identified;

(v) the *anti* conformers are well represented and form distinguishable supramolecular motifs implying no significant basicity difference between the *syn* and *anti* lone pairs;

(vi) the strongest (intramolecular) hydrogen bonds are observed in mono-anion dicarboxylic compounds and involve simultaneously an *anti* conformer and an *anti* lone pair, supporting the fact that *anti* interactions are by no means weaker than *syn* interactions;

(vii) the shortest hydrogen-bond lengths found in this survey, including those formed with water molecules, are close to 2.36 Å (Fig. 6*d*);

(viii) cooperative effects appear to be important in probably all systems involving carboxyl(ate) groups and should always be considered.

Although significant progress has been achieved in crystal engineering, it seems appropriate to recall a sobering thought by Steiner, who wrote in a paper on hydrogen-bond competition: '*Even though it is true that strong hydrogen-bond donors tend to interact with strong acceptors, this is valid only as a tendency. Weak acceptors also have a certain chance of attracting the strong donor. This weakens the general applicability of rules for predicting hydrogen-bond modes from hierarchies of donor and acceptor strengths and indeed all such rules published are very unreliable in practice*' (Steiner, 2001). Further, Desiraju, witnessing the constant discovery of unforeseen structures, noted that after all: '*it would seem that brute-force method will eventually win*' (Desiraju, 2007), suggesting that many more interaction rules of increasing complexity will be brought to light and that concerted but also serendipitous crystallization experiments are still very much needed to make progress in the field. These considerations on small supramolecular synthons apply fully to biomolecular systems where carboxyl(ate) groups are found to adapt in surprising and still insufficiently documented ways to their local environment.

Acknowledgements

PA wishes to thank Professor Eric Westhof for ongoing support. We thank Professor Richard Giegé for stimulating discussions. This work has been published under the framework of the LABEX: ANR-10-LABX0036_NETRNA and benefits from a funding from the state managed by the French National Research Agency as part of the program 'Investments for the future'.

References

Ahmed, H. U., Blakeley, M. P., Cianci, M., Cruickshank, D. W. J., Hubbard, J. A. & Helliwell, J. R. (2007). *Acta Cryst.* D**63**, 906–922.

Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.

Allen, F. H. & Bruno, I. J. (2010). *Acta Cryst.* B**66**, 380–386.

Allen, F. H., Groom, C. R., Liebeschuetz, J. W., Bardwell, D. A., Olsson, T. S. & Wood, P. A. (2012). *J. Chem. Inf. Model.* **52**, 857–866.

Allen, F. H. & Kirby, A. J. (1991). *J. Am. Chem. Soc.* **113**, 8829–8831.

Allen, F. H., Motherwell, W. D. S., Raithby, P. R., Shields, G. P. & Taylor, R. (1999). *New J. Chem.* **23**, 25–34.

Berkovitch-Yellin, Z. & Leiserowitz, L. (1982). *J. Am. Chem. Soc.* **104**, 4052–4064.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Bertolasi, V., Pretto, L., Gilli, G. & Gilli, P. (2006). *Acta Cryst.* B**62**, 850–863.

Beyer, T. & Price, S. L. (2000). *J. Phys. Chem. B*, **104**, 2647–2655.

Blagden, N., de Matas, M., Gavan, P. T. & York, P. (2007). *Adv. Drug Deliv. Rev.* **59**, 617–630.

Brown, M. A., Vila, F., Sterrer, M., Thürmer, S., Winter, B., Ammann, M., Rehr, J. J. & van Bokhoven, J. A. (2012). *J. Phys. Chem. Lett.* **3**, 1754–1759.

Bruno, I. J., Cole, J. C., Edgington, P. R., Kessler, M., Macrae, C. F., McCabe, P., Pearson, J. & Taylor, R. (2002). *Acta Cryst.* B**58**, 389–397.

CCDC (1994). *Vista.* Cambridge Crystallographic Data Centre, Cambridge, England, http://www.ccdc.cam.ac.uk.

Chisholm, J., Pidcock, E., van de Streek, J., Infantes, L., Motherwell, S. & Allen, F. H. (2006). *CrystEngComm*, **8**, 11–28.

Cleland, W. W. & Kreevoy, M. M. (1994). *Science*, **264**, 1887–1890.

Das, D. & Desiraju, G. R. (2006). *Chem. Asian J.* **1**, 231–244.

Das, D., Desiraju, G. R., Jetti, R. K. R. & Boese, R. (2005). *Acta Cryst.* E**61**, o1588–o1589.

Dawid, U., Pruchnik, F. P. & Starosta, R. (2009). *Dalton Trans.* pp. 3348–3353.

Desiraju, G. R. (2007). *Angew. Chem. Int. Ed.* **46**, 8342–8356.

Desiraju, G. R. (2013). *J. Am. Chem. Soc.* **135**, 9952–9967.

DeVita Dufort, M., Davison, M., Lalancette, R. A. & Thompson, H. W. (2007). *Acta Cryst.* C**63**, o646–o649.

Dunitz, J. D. & Gavezzotti, A. (2012). *Cryst. Growth Des.* **12**, 5873–5877.

Fisher, S. J., Blakeley, M. P., Cianci, M., McSweeney, S. & Helliwell, J. R. (2012). *Acta Cryst.* D**68**, 800–809.

Flocco, M. M. & Mowbray, S. L. (1995). *J. Mol. Biol.* **254**, 96–105.

Gajda, R., Katrusiak, A. & Crassous, J. (2009). *CrystEngComm*, **11**, 2668–2676.

Galek, P. T. A., Chisholm, J. A., Pidcock, E. & Wood, P. A. (2014). *Acta Cryst.* B**70**, 91–105.

Gandour, R. D. (1981). *Bioorg. Chem.* **10**, 169–176.

Gao, J. & Pavelites, J. J. (1992). *J. Am. Chem. Soc.* **114**, 1912–1914.

Gilli, G. & Gilli, P. (2009). *The Nature of the Hydrogen Bond. Outline of a Comprehensive Hydrogen Bond Theory.* Oxford University Press.

Glusker, J. P. (1998). *Top. Curr. Chem.* **198**, 1–56.

Görbitz, C. H. & Etter, M. C. (1992*a*). *J. Am. Chem. Soc.* **114**, 627–631.

Görbitz, C. H. & Etter, M. C. (1992*b*). *J. Chem. Soc. Perkin Trans. 2*, pp. 131–135.

Groom, C. R. & Allen, F. H. (2014). *Angew. Chem. Int. Ed.* **53**, 662–671.

Hocking, R. K. & Hambley, T. W. (2005). *Dalton Trans.* pp. 969–978.

Hosur, M. V., Chitra, R., Hegde, S., Choudhury, R. R., Das, A. & Hosur, R. V. (2013). *Crystallogr. Rev.* **19**, 3–50.

Huggins, M. L. (1936). *J. Org. Chem.* **1**, 407–456.

Hursthouse, M. B., Montis, R. & Tizzard, G. J. (2011). *CrystEngComm*, **13**, 3390–3401.

Jeffrey, G. A. (1997). *An Introduction to Hydrogen Bonding.* New York: Oxford University Press.

Jeffrey, G. A. & Saenger, W. (1991). *Hydrogen Bonding in Biological Structures.* Berlin: Springer-Verlag.

Kamitakahara, A. & Pranata, J. (1995). *Bioorg. Chem.* **23**, 256–262.

Katz, B. A., Spencer, J. R., Elrod, K., Luong, C., Mackman, R. L., Rice, M., Sprengeler, P. A., Allen, D. & Janc, J. (2002). *J. Am. Chem. Soc.* **124**, 11657–11668.

Kuduva, S. S., Craig, D. C., Nangia, A. & Desiraju, G. R. (1999). *J. Am. Chem. Soc.* **121**, 1936–1944.

Langkilde, A., Kristensen, S. M., Lo Leggio, L., Mølgaard, A., Jensen, J. H., Houk, A. R., Navarro Poulsen, J.-C., Kauppinen, S. & Larsen, S. (2008). *Acta Cryst.* D**64**, 851–863.

Leiserowitz, L. (1976). *Acta Cryst.* B**32**, 775–802.

Li, Y. & Houk, K. N. (1989). *J. Am. Chem. Soc.* **111**, 4505–4507.

Macrae, C. F., Bruno, I. J., Chisholm, J. A., Edgington, P. R., McCabe, P., Pidcock, E., Rodriguez-Monge, L., Taylor, R., van de Streek, J. & Wood, P. A. (2008). *J. Appl. Cryst.* **41**, 466–470.

McKie, A. H., Friedland, S. & Hof, F. (2008). *Org. Lett.* **10**, 4653–4655.

Meanwell, N. A. (2011). *J. Med. Chem.* **54**, 2529–2591.

Meot-Ner, M. (2012). *Chem. Rev.* **112**, PR22–PR103.

Meot-Ner, M., Elmore, D. E. & Scheiner, S. (1999). *J. Am. Chem. Soc.* **121**, 7625–7635.

Merz, K. & Vasylyeva, V. (2010). *CrystEngComm*, **12**, 3989–4002.

Montis, R. & Hursthouse, M. B. (2012). *CrystEngComm*, **14**, 5242–5254.

Nagy, P. I. (2013). *Comput. Theor. Chem.* **1022**, 59–69.

Pace, C. N., Grimsley, G. R. & Scholtz, J. M. (2009). *J. Biol. Chem.* **284**, 13285–13289.

Payaka, A., Tongraar, A. & Rode, B. M. (2009). *J. Phys. Chem. A*, **113**, 3291–3298.

Payaka, A., Tongraar, A. & Rode, B. M. (2010). *J. Phys. Chem. A*, **114**, 10443–10453.

Perrin, C. L. (2010). *Acc. Chem. Res.* **43**, 1550–1557.

Perrin, C. L. & Nielson, J. B. (1997). *Annu. Rev. Phys. Chem.* **48**, 511–544.

Ramanadham, M., Jakkal, V. S. & Chidambaram, R. (1993). *FEBS Lett.* **323**, 203–206.

Rodríguez-Cuamatzi, P., Arillo-Flores, O. I., Bernal-Uruchurtu, M. I. & Höpfl, H. (2007). *Supramol. Chem.* **19**, 559–578.

Sanphui, P., Bolla, G. U. D., Mukherjee, A. K. & Nangia, A. (2013). *CrystEngComm*, **15**, 34–38.

Saravanakumar, R., Varghese, B. & Sankararaman, S. (2009). *CrystEngComm*, **11**, 337–346.

Sato, H. & Hirata, F. (1999). *J. Mol. Struct. Theochem*, **461–462**, 113–120.

Sawyer, L. & James, M. N. (1982). *Nature*, **295**, 79–80.

Śmiechowski, M., Gojło, E. & Stangret, J. (2011). *J. Phys. Chem. B*, **115**, 4834–4842.

Steiner, T. (1999). *Chem. Commun.* pp. 2299–2300.

Steiner, T. (2001). *Acta Cryst.* B**57**, 103–106.

Steiner, T. (2002). *Angew. Chem. Int. Ed.* **41**, 48–76.

Thomas, L. H., Blagden, N., Gutmann, M. J., Kallay, A. A., Parkin, A., Seaton, C. C. & Wilson, C. C. (2010). *Cryst. Growth Des.* **10**, 2270–2274.

Torshin, I. Y., Harrison, R. W. & Weber, I. T. (2003). *Protein Eng.* **16**, 201–207.

Vishweshwar, P., Jagadeesh Babu, N., Nangia, A., Mason, S. A., Puschmann, H., Mondal, R. & Howard, J. A. K. (2004). *J. Phys. Chem. A*, **108**, 9406–9416.

Wilson, C. C., Shankland, N. & Florence, A. J. (1996). *Faraday Trans.* **92**, 5051–5057.

Wohlfahrt, G. (2005). *Proteins*, **58**, 396–406.

Ying, S.-M. (2012). *Inorg. Chem. Commun.* **22**, 82–84.

Zimmerman, S. C., Korthals, J. S. & Gramer, K. D. (1991). *Tetrahedron*, **47**, 2649–2660.

STRUCTURAL SCIENCE
CRYSTAL ENGINEERING
MATERIALS

Volume 71 (2015)

Supporting information for article:

A comprehensive classification and nomenclature of carboxyl–carboxyl(ate) supramolecular motifs and related catemers: implications for biomolecular systems

Luigi D'Ascenzo and Pascal Auffinger

Table S1 List of all catemer-containing structures identified in the CSD with $R1 \leq 0.05$.

Homo-catemer	
SS-S	DAYNUN; ENIFAJ; FIKJEO; GUWCOQ; HUKJUT; HUSXAU; IBUHES; IBUHOC; MUCQAD; NAGVUM; SUVYEN; TARTAC; TARTAL; TARTAL01; TARTAL02; TARTAL03; TARTAL04; TORTEA; VAFCUB; WOCHIF; XEDZUC; XONNET; ZOGTUK
SA-S	ACETAC07; ACETAC09; ARUVAK; BELQOZ; BIPCIQ10; BUJYEL; CANSAL; CIJLOW; CIPZIL; CITRAC10; CITRAC11; CLACET01; DIYDIY; DMCPCX; DOTWOY; EVORIQ; EYONAI; FIWQEI; FOHREI; FORMAC01; GELZIG; GOGPEY; GOGPIC; HEKWOJ; HIFWOJ; HUGSEH; HUWLIV; IMEHIS; IROQOV; ISORUD; ISOSAK; ISOSAL; IVEBIV; JUMVIW; KABGOL; KIKLIZ; KOJZEO; KUTMIW; MEKMOF; MIYCON; MEDNOZ01; NUFJUU; OWUSEF; OXALAC05; OXALAC07; PEPPAD; PEZWAS; QAGMOG; QUQHAL; QURQID; RACCEE; SAWBUN; SDPPCX; TEHMAU; TETROL01; TUSPOM; UCAYUU; URUPUT; VEVSIZ; VOHNUC; WANROU; WASJAD01; WINVAR; WOCTUD; WUXHUS; ZAVTOF; ZILBOL
AS-A	GIMRAW; NEWXAO; ROGHOL
AA-A	DMOXDA01; DMOXDA02; HUMGOL
Hetero-catemer	
SS-A•AS-S	ROZHEU
SS-A•AA-S	BCOCDC; CBUDCX; CBUDCX01; CBUDCX02; CBUDCX03; CBUDCX04; COMHFN; COMHFN01; CPRDCA; FIGMAJ; FURDCB01; IDAKAB; JUNCUQ; JUNMAG; MEKLOE; MIGPEX; RAJJUH
SA-A•AS-S	WOKPOC
SA-A•AA-S	CPIBFC; CUKGIZ; KAMKAK; KAKTOS; MALIAC11; MALIAC12; MALIAC13

5.4.2 Further remarks and outlook

One of the most important outcomes of this work is that very short hydrogen bonds should be evaluated as individual interactions and their occurrence, although rare, has to be considered in prediction or computational techniques such as molecular dynamics. A parametrization for a unique interaction between two specific atoms that can lead to distances ~ 2.5 Å is needed for a trustworthy description of the system, wherever they appear. This work infers also that structural analysis has the potential to improve our perception of protonation states inside large crystallographic structures and shed light on biochemical mechanism involving variation of pH/protonation.

Analyses such as the study of interaction distances between carboxyl(ate) and water molecules can show significant patterns regarding the strength of hydrogen bond donor/acceptor and protonation states can be successfully highlighted even when the hydrogen atoms are not directly visible. In this direction, an example of Asp/Glu interaction can be found within the thaumatin (a sweetener protein, PDB: 2VHK, res.: 0.9 Å) core. Considering the very short interaction distance (~ 2.5 Å), even without explicit hydrogen atoms (present only in some ultra-high resolution structures) the protonation states can be reliably inferred. Thus, interaction distances between particular atoms of charged species can be used to infer protonation states even in lower resolution structures, such as the large majority of relevant biomolecules.

5.5 Preliminary data on MD analysis of temperature effects on RNA

The binding of charged species is a significant factor in understanding how the environment can act on nucleic acids by affecting the properties of solvent coordination. Yet, it is just a part of the wide spectrum of environmental effects. In the effort to provide a more complete picture, I used MD simulations on RNA during my PhD. The data presented in this chapter are preliminary results, because, at the time of writing (September 2016), more detailed analysis were still being performed.

The MD data on GNRA tetraloops presented in **Paper 1** highlighted the structural role of bound water molecules for these loops. Given the symbiotic nucleic acid/solvent relationship, it is clear that these hydration patterns are involved in the stabilization of RNA structures. The main focus of further MD studies has been to analyze the RNA first shell solvent structure, by measuring simulated water and ions residency times surrounding the loop, as a measure of the loop stability. In fact, a previous MD study conducted in the group in 2002 on a RNA duplex showed that the dynamics of water molecules and ions located in the RNA first coordination shell strongly depends on the temperature (Auffinger and Westhof 2002). The most spectacular variation, observed between 278 K and 310 K, was a halving of the water molecule highest residence time (from 1 ns to 0.5 ns). The associated invariability of RNA structure suggested that this observation was related to a “premelting” of the solvent surrounding RNA.

Starting from these premises, we planned to conduct similar MD simulation at variable temperatures on tetraloop motifs. The goal was to assess the stability of the loop through temperature-induced melting of the first shell solvent structure, in order to study how base mutations within the loop, in the closing pair or in the stem affects it. However, the results of these simulations were contradictory. Although a destabilization of the tetraloop was observed upon disruption of the Watson-Crick closing pair, we needed further results to isolate the key aspects of this phenomenon. To gather these information and isolate the relevant point to analyze, we shifted our attention to simpler RNA duplex systems. Thus, $(GpC)_{12}$ and $(ApU)_{12}$ duplexes have been studied in sets of 50 ns MD simulations, at five constant temperatures: 277 K, 300 K, 320 K, 340 K and 360 K. More details on the simulations setup can be found in **Chapter 2.5.1**. Each simulation was performed in three independent copies. A total of 30 trajectories were analyzed in term of residence time of water and K^+ ions around nucleotide positions. Residence time values were added for each nucleotide, as well as for each one of the three copies. This yielded an effective sampling time for the binding of water and ions in the order of hundreds of ns. The preliminary results are presented in **Figure 5.3** and **Figure 5.4.**

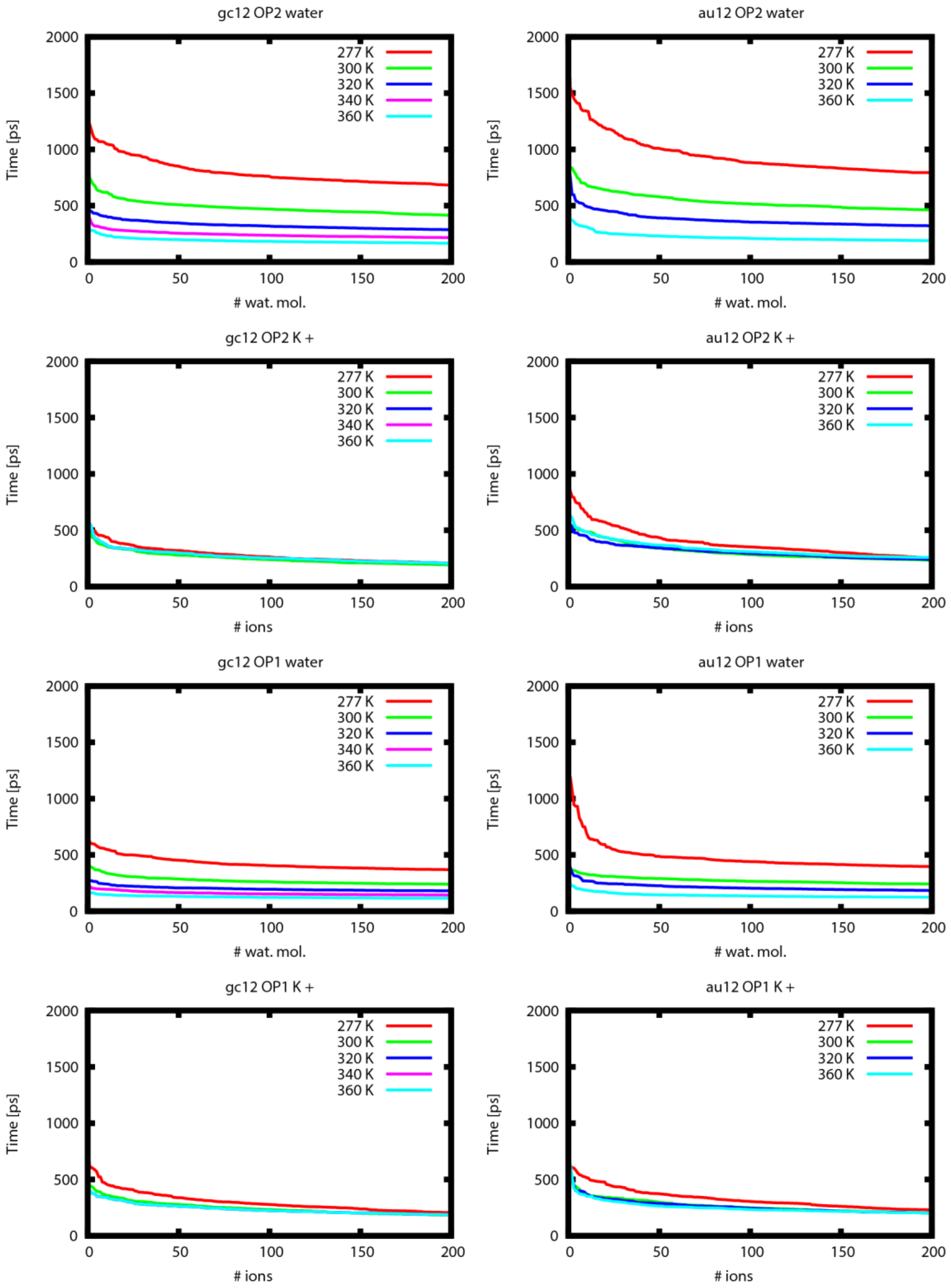


Figure 5.3. **Residence time profiles for water and K$^+$ ions bound on phosphate oxygen atoms.** The number of water/K$^+$ ions bound to OP2/1 (PDB identifier) phosphate oxygen atoms are plotted as a function of their coordination contact times, adding all the contributions from the 24 nucleotides. Plots on the left columns refers to the (GpC)$_{12}$ duplex, while plots on the right column to (ApU)$_{12}$ duplex. On top of each plot, it is indicated the corresponding phosphate oxygen and water/K$^+$ coordinating atom partners. Colored profiles represent different temperatures used in the respective simulations, in accord to the legend in the upper-right section of each plot. Time values are reported in ps.

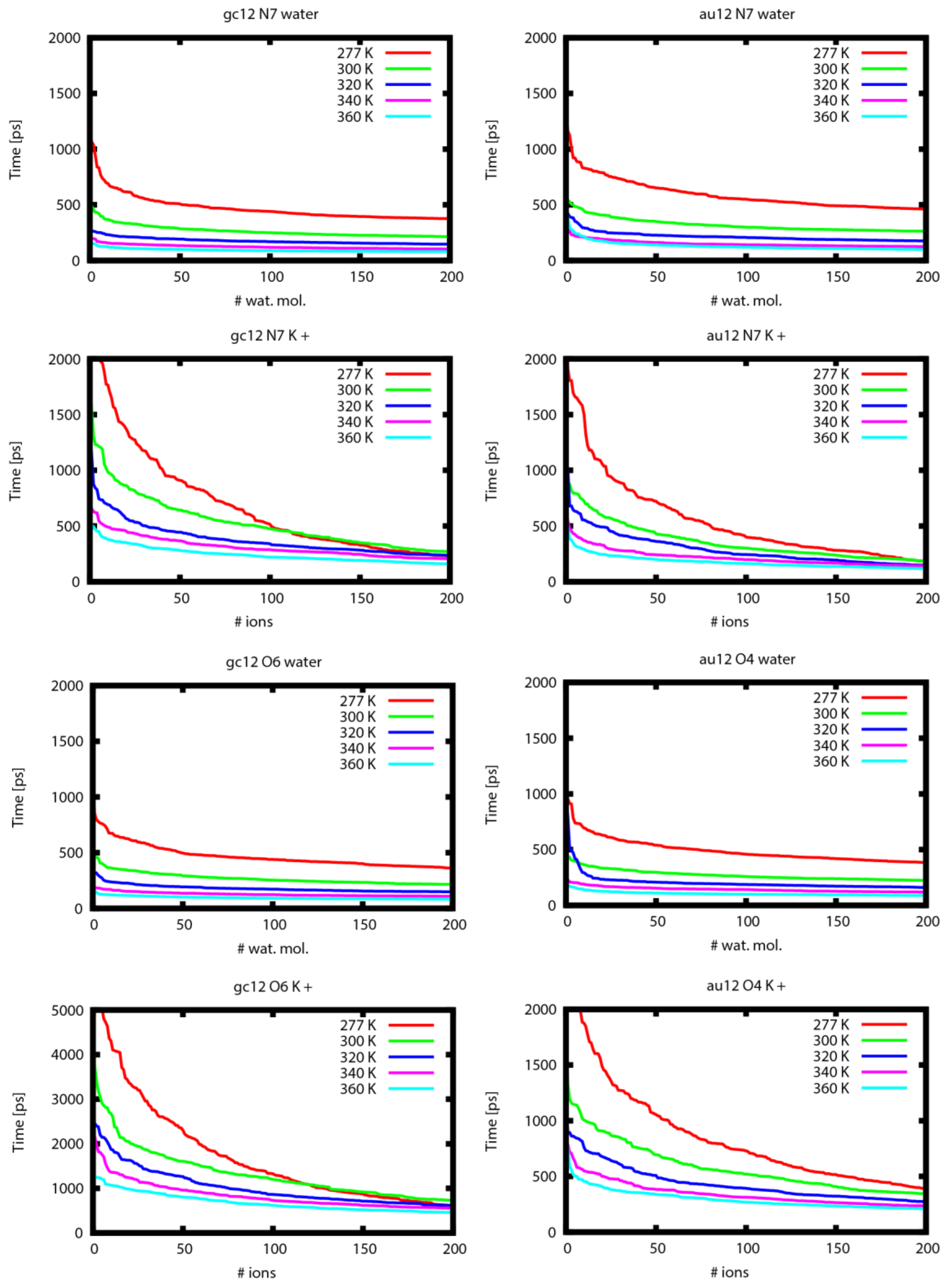


Figure 5.4. Residence time profiles for water and K$^+$ ions bound on base pairs major groove. The number of water/K$^+$ ions bound to atoms of the base pairs major groove atoms are plotted as a function of their coordination contact times, adding contribution for each base pair. Plots on the left columns refers to the (GpC)$_{12}$ duplex, while plots on the right column to (ApU)$_{12}$ duplex. On top of each plot, it is indicated the corresponding nucleobase and water/K$^+$ coordinating atom partners. Colored profiles represent different temperatures used in the respective simulations, in accord to the legend in the upper-right section of each plot. Time values are reported in ps, on a 2,000 ps scale for each graph, except (GpC)$_{12}$ O6 K$^+$ that is plotted with a 5,000 ps scale.

The data on residence time of water with OP2 and OP1 atoms show the most significant temperature dependence. At 277 K both duplexes show water molecules with residence times larger or close to 1 ns, in accord with the data obtained in 2002 over a 2.4 ns simulation. The mild decay of binding profiles for water molecules infers a dynamic coordination scheme, without molecules bound for significantly longer times than others; the only partial exception is the profile of water bound to OP1 atoms of $(ApU)_{12}$ at 277 K. Conversely, profiles of K^+ coordination to phosphate oxygen atoms show no remarkable temperature dependence. Slightly longer (~800 ps versus ~500 ps) residence times at 277 K are observed for the coordination of most long-bound K^+ with phosphate of $(ApU)_{12}$ duplex compared to $(GpC)_{12}$.

Concerning the solvent interactions with major grooves, water molecules show a similar temperature-dependent profiles compared to phosphate binding; the only difference is a slightly more pronounced difference between the longest residence times and the others. No differences appear comparing the carbonyl oxygen of uridine (O4) and guanine (O6). Overall, N7 atoms appear to provide a preferred coordination site for water compared to carbonyl oxygens. Residence times for K^+ bound to N7 and O4/O6 are sensibly longer compared to phosphate atoms, and in the case of guanines O6 exceed even 5 ns. In fact, K^+ ions have already shown a high affinity for the major groove of $(GpC)_{12}$ duplexes, and specifically to guanine O6 atoms (Auffinger and Westhof 2000). On the other hand, K^+ ions binding in $(ApU)_{12}$ duplex major grooves have been found to make ion bridges between base pairs and thus coordinating to U(O4) and A(N7) simultaneously (Auffinger and Westhof 2001b). Further analysis based on these results have to be run in order to clarify the binding of K^+, which has been rarely found to bind purine N7 in crystallographic structures (see **Paper 4** and **Paper 5**).

The preliminary data presented on the temperature dependence of solvent in the RNA first coordination shell are overall in accord with earlier results. The residence time profiles infer a “disorganization” of the solvent structure when raising the temperature, but this effect is not clear and it is difficult to quantify in the current state. More in-depth analysis on the binding of all solvent atoms (and not only the longest residence time profiles) could reveal some of the still hidden results in this perspective, suggesting which nucleotide sites should be focused in order to gain insightful information. The simulation performed established an investigation base in terms of protocols and parameters, that will be utilized in future endeavors, for more complex systems than RNA duplexes. Our final goal will remain to shed light on some of the darkest aspects of how RNA is modulated by its context, at the local atomistic level.

6. Conclusions and perspectives

6.1 Concluding remarks

The results presented in this thesis work complement our knowledge of RNA and RNA/protein systems. The description of rare non-covalent interactions, RNA tetraloop motifs and environmental effects on RNA will help to improve the understanding of biomolecular recognition networks. These insights are especially suited to improve structural determination with techniques such as X-ray crystallography and computational methods (*e.g.* force fields parametrization).

Stacking interactions between aromatic nucleobases and backbone oxygen atoms have been characterized in nucleic acid systems. The main outcome on this subject is that these interactions are probably weak due to their need of assistance by stronger contextual interactions. Nonetheless, they are significant for the fold of nucleic acid architecture, thus structural/functional analysis on these systems cannot overlook them. Our description of stacking interactions has always been based on structural considerations about contact distances and local geometry, without referring to energy values. Although energetic descriptions are useful to infer the relative strengths of biomolecular interactions, they need to be considered with respect to the local context. This, in the case of biomolecular systems that function in complex cellular media, is actually difficult. On the other hand, one can speculate about the fact that structural observations are affected by the same limitations. In this perspective, all the results presented and discussed in this work have to be considered as true for the environment in which structural data where obtained. Relative considerations such as those on the interaction assistance, have more chances to be actually conserved, even *in vivo*.

The results on RNA tetraloops are meant to describe these motifs with simple although effective structural signatures, which also bear also information on their tertiary interactions. Here, the case of UNCG tetraloop receptors is anecdotal. The identification of a fundamental ensemble of structural elements also goes in the direction of redefining tetraloop classification. This is particularly important considering the occurrences of unexpected folds and has the potential to be extremely useful in the forthcoming structural determination of lncRNAs.

The considerations on environmental conditions have been focussing in the first place on the binding modes of charged species to nucleic acids. During this endeavour, a large number of structural inaccuracies in the analyzed structures prompted us to assess the issues of solvent interpretation by structural methods, especially X-ray diffraction. Moving from structural studies, more complex and dynamic environmental effects have been begun to be assessed by MD simulations on RNA systems.

Altogether, these points are linked by the *fil rouge* of the unbreakable relationship structure/function and are meant to be pieces of a colourful and dynamic puzzle that portrays how biomolecules behave inside beings, from their synthesis up to their degradation. A taste of the forthcoming steps I foresaw to construct my personal *scientific puzzle* is presented in the following and last chapter.

6.2 What does the future hold: investigations of crowding conditions by MD

When asked about my scientific future, I like to point towards the study of *bigger biomolecular systems embedded in more complex environments*. For this reason, I include as a perspective to my PhD thesis the overview of a research project oriented towards this direction. The goal of the project is to study by MD how crowding conditions affect RNA and RNA-RNA systems, employing the T-box tRNA riboswitch as model system (Zhang and Ferre-D'Amare 2013; Zhang and Ferre-D'Amare 2015).

Thermodynamic and kinetic properties of biomolecules in intracellular crowded environment are different from those commonly reported for *in vitro* dilute solutions. Yet, given the high complexity of biological systems, molecular details are generally difficult to ascertain. The aim of the project is to develop a molecular sensor, based on the bacterial T-box tRNA riboswitch, to study crowding effects on RNA recognition through full atomistic MD simulations. A particular attention will be given to the stacking between two otherwise solvent exposed base pairs that are part of the tRNA/riboswitch recognition motif (**Fig. 6.1**).

In this unique docking platform, the RNA-RNA recognition mode involves an intermolecular π–π stacking of two base pairs, in a manner analogous to the complex of tetraloops found inside the ribosome (**Fig. 4.4** in **Chapter 4.3.4**). This stacking interaction is essential for the function of the tRNA riboswitch, which is a key element of many bacteria. For its hydrophobic nature and absence of hydrogen bonding it is hypothesized to be very responsive to environmental alterations such as those associated with the presence of crowding agents or with temperature variations. Given these properties, this complex represents a good opportunity to design a molecular sensor based on the

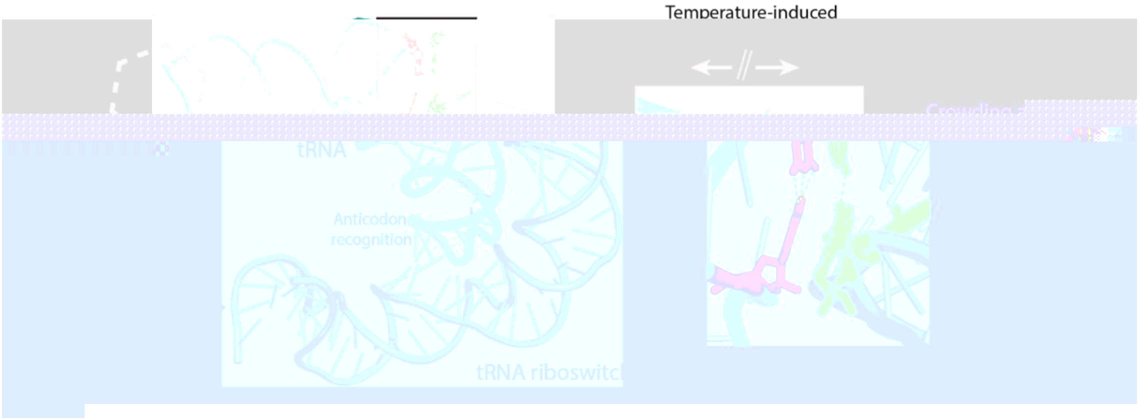


Figure 6.1. **Graphical overview of the study on crowding conditions affecting a tRNA riboswitch**. Schematic representation of thermal disruption at the level of the docking platform, together with a list of potential crowding agents. The tRNA/riboswitch complex is taken from PDB: 4LCK; res.:3.2 Å; (Zhang and Ferre-D'Amare 2013).

above-mentioned stacking interaction. It can provide information related to the effect of a large panel of crowding molecules and crowding conditions on RNA systems as well as on fundamental stacking processes that are significantly modulated by solvation changes.

The tRNA/riboswitch system will be simulated by full atomistic molecular dynamics (MD) in different ionic and crowding conditions by using temperature changes as control variable. In this endeavor, raising temperature would induce a controlled disruption of the docking platform in order to screen the effects of a large number of ionic and crowding conditions. Besides fundamental molecular recognition issues, this model system will allow to specifically study crowding on a dynamic system that mimics a step of the tRNA release from the riboswitch. Consequently, this setup is designed to replicate and evaluate the stabilizing potential of crowding agents ranging from monomers to polymeric biomolecules, the latter having for goal to mimic more closely intracellular conditions.

The first step of the project is to develop MD models for RNA in crowding conditions, given that up to now no full atomistic MD simulations of RNA in crowding conditions have been reported. This will be accomplished using the tRNA as starting model, in nanosecond (ns) to microsecond (μs) range simulations with cosolutes such as ethylene glycol (EG) and short polyethylene glycol (PEG, ≤ 20 EG monomers). After this phase, the goal will be to shift to the whole tRNA/riboswitch complex and analyze the stability of the docking platform versus thermal disruption. MD simulations at different temperatures will be run, ranging from 293 to 350 K. It is estimated that in this temperature range and on the investigated time-scales, the RNA parts of the system will remain stable while the docking platform will open. Indeed, the hydrophobicity-driven stacking interaction is expected to be the most sensitive part of the complex to temperature changes and likely the first interaction to break during temperature changes. Crowders are expected to modulate this temperature sensitivity, a hypothesis that will be verified in the last phase of the project. The last part of the project consists in evaluating the (macro)molecular crowding effects on the tRNA/riboswitch, with simulations up to μs time scales in progressively more complex conditions involving a palette of different crowding agents. Altogether, these simulations will help to understand crowding effects on RNA systems and on a family of noncovalent interactions that is fundamental for main recognition processes. It is my hope that the *in silico* development of this tRNA/riboswitch environmental sensor will also be extended to experimental *in vitro* techniques, that in parallel with data coming from MD simulations could synergistically help to further our understanding on (macro)molecular crowding effects.

7. Bibliography

Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW et al. 2010. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Cryst.* **D66**: 213-221.

Allen FH. 2002. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Cryst.* **B58**: 380-388.

Amunts A, Brown A, Bai XC, Llacer JL, Hussain T, Emsley P, Long F, Murshudov G, Scheres SH, Ramakrishnan V. 2014. Structure of the yeast mitochondrial large ribosomal subunit. *Science* **343**: 1485-1489.

Amunts A, Brown A, Toots J, Scheres SH, Ramakrishnan V. 2015. Ribosome. The structure of the human mitochondrial ribosome. *Science* **348**: 95-98.

Anger AM, Armache JP, Berninghausen O, Habeck M, Subklewe M, Wilson DN, Beckmann R. 2013. Structures of the human and Drosophila 80S ribosome. *Nature* **497**: 80-85.

Antao VP, Tinoco I. 1992. Thermodynamic parameters for loop formation in RNA and DNA hairpin tetraloops. *Nucleic Acids Res.* **20**: 819-824.

Antczak M, Zok T, Popenda M, Lukasiak P, Adamiak RW, Blazewicz J, Szachniuk M. 2014. RNApdbee--a webserver to derive secondary structures from pdb files of knotted and unknotted RNAs. *Nucleic Acids Res* **42**: W368-372.

Arranz-Mascaros P, Bazzicalupi C, Bianchi A, Giorgi C, Godino-Salido ML, Gutierrez-Valero MD, Lopez-Garzon R, Savastano M. 2013. Thermodynamics of anion–π interactions in aqueous solution. *J Am Chem Soc* **135**: 102-105.

Auffinger P, Bielecki L, Westhof E. 2004. Anion binding to nucleic acids. *Structure* **12**: 379-388.

Auffinger P, Grover N, Westhof E. 2011. Metal ion binding to RNA. *Met. Ions Life Sci.* **9**: 1-35.

Auffinger P, Westhof E. 2000. Water and ion binding around RNA and DNA (C,G)-oligomers. *J. Mol. Biol.* **300**: 1113-1131.

Auffinger P, Westhof E. 2001a. An extended structural signature for the tRNA anticodon loop. *RNA* **7**: 334-341.

Auffinger P, Westhof E. 2001b. Water and ion binding around r(UpA)$_{12}$ and d(TpA)$_{12}$ oligomers - Comparison with RNA and DNA (CpG)$_{12}$ duplexes. *J. Mol. Biol.* **305**: 1057-1072.

Auffinger P, Westhof E. 2002. Melting of the solvent structure around a RNA duplex: a molecular dynamics simulation study. *Biophys. Chem.* **95**: 203-210.

Ball P. 2008. Water as an active constituent in cell biology. *Chem. Rev.* **108**: 74-108.

Baltimore D. 1970. Viral RNA-dependent DNA polymerase. *Biotechnology* **24**: 3-5.

Batey RT, Sagar MB, Doudna JA. 2001. Structural and energetic analysis of RNA recognition by a universally conserved protein from the signal recognition particle. *J. Mol. Biol.* **307**: 229-246.

Berendsen HJC, Grigera JR, Straatsma TP. 1987. The missing term in effective pair potential. *J. Phys. Chem.* **97**: 6269-6271.

Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A. 1984. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**: 3684-3690.

Berget SM, Moore C, Sharp PA. 2000. Spliced segments at the 5 ' terminus of adenovirus 2 late mRNA. *Rev. Med. Virol.* **10**: 356-362.

Bergonzo C, Henriksen NM, Roe DR, Cheatham TE, 3rd. 2015. Highly sampled tetranucleotide and tetraloop motifs enable evaluation of common RNA force fields. *RNA*.

Berman HM. 2008. The Protein Data Bank: a historical perspective. *Acta Cryst.* **A64**: 88-95.

Berman HM, Kleywegt GJ, Nakamura H, Markley JL. 2012. The Protein Data Bank at 40: reflecting on the past to prepare for the future. *Structure* **20**: 391-396.

Bottaro S, Gil-Ley A, Bussi G. 2016. RNA folding pathways in stop motion. *Nucleic Acids Res.* **44**: 5883-5891.

Brown A, Amunts A, Bai XC, Sugimoto Y, Edwards PC, Murshudov G, Scheres SH, Ramakrishnan V. 2014. Structure of the large ribosomal subunit from human mitochondria. *Science* **346**: 718-722.

Brown A, Shao S, Murray J, Hegde RS, Ramakrishnan V. 2015. Structural basis for stop codon recognition in eukaryotes. *Nature* **524**: 493-496.

Brown ID, McMahon B. 2002. CIF: the computer language of crystallography. *Acta Cryst.* **B58**: 317-324.

Brownlee GG. 1971. Sequence of 6S RNA of E. coli. *Nat New Biol* **229**: 147-149.

Bruno IJ, Cole JC, Edgington PR, Kessler M, Macrae CF, McCabe P, Pearson J, Taylor R. 2002. New software for searching the Cambridge Structural Database and visualizing crystal structures. *Acta Cryst.* **B58**: 389-397.

Butcher SE, Dieckmann T, Feigon J. 1997. Solution structure of the conserved 16S-like ribosomal RNA UGAA tetraloop. *J. Mol. Biol.* **268**: 348-358.

Butcher SE, Pyle AM. 2011. The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks. *Acc. Chem. Res.* **44**: 1302-1311.

Cate JH, Gooding AR, Podell E, Zhou KH, Golden BL, Kundrot CE, Cech TR, Doudna JA. 1996. Crystal structure of a group I ribozyme domain - Principles of RNA packing. *Science* **273**: 1678-1685.

Cech TR, Steitz JA. 2014. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* **157**: 77-94.

Cheatham TE, Miller JL, Fox T, Darden TA, Kollman PA. 1995. Molecular dynamics simulations on solvated biomolecular systems: the particle mesh Ewald method leads to stable trajectories of DNA, RNA and proteins. *J. Am. Chem. Soc.* **117**: 4193-4194.

Chen AA, Garcia AE. 2013. High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA* **110**: 16820-16825.

Cheong C, Varani G, Tinoco I. 1990. Solution structure of an unusually stable RNA hairpin, 5'GGAC(UUCG)GUCC. *Nature* **346**: 680-681.

Cheong H, Kim N, Cheong C. 2015. RNA structure: tetraloops. In *ELS*. John Wiley & Sons, ltd: Chichester.

Chow LT, Gelinas RE, Broker TR, Roberts RJ. 1977. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* **12**: 1-8.

Clark BF. 2006. The crystal structure of tRNA. *J Biosci* **31**: 453-457.

Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwel JW, Kollman PA. 1995. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**: 5179-5197.

Coulocheri SA, Pigis DG, Papavassiliou KA, Papavassiliou AG. 2007. Hydrogen bonds in protein-DNA complexes: where geometry meets plasticity. *Biochimie* **89**: 1291-1303.

Cruz JA, Westhof E. 2009. The dynamic landscapes of RNA architecture. *Cell* **136**: 604-609.

Dauter Z, Adamiak DA. 2001. Anomalous signal of phosphorus used for phasing DNA oligomer: importance of data redundancy. *Acta Cryst.* **D57**: 990-995.

de Hoog P, Gamez P, Mutikainen I, Turpeinen U, Reedijk J. 2004. An aromatic anion receptor: anion-pi interactions do exist. *Angew Chem Int Ed Engl* **43**: 5815-5817.

Deng NJ, Cieplak P. 2007. Molecular dynamics and free energy study of the conformational equilibria in the UUUU RNA hairpin. *J. Chem. Theory Comput.* **3**: 1435-1450.

Deng NJ, Cieplak P. 2010. Free energy profile of RNA hairpins: a molecular dynamics simulation study. *Biophys. J.* **98**: 627-636.

Desai R, Kilburn D, Lee HT, Woodson SA. 2014. Increased ribozyme activity in crowded solutions. *J. Biol. Chem.* **289**: 2972-2977.

Desiraju GR. 2013. Crystal engineering: from molecule to crystal. *J. Am. Chem. Soc.* **135**: 9952-9967.

Dougherty DA. 1996. Cation-π interactions in chemistry and biology: a new view of benzene, Phe, Tyr, and Trp. *Science* **271**: 163-168.

Dunitz JD. 2015. Intermolecular atom-atom bonds in crystals? *IUCrJ* **2**: 157-158.

Dunitz JD, Gavezzotti A. 2009. How molecules stick together in organic crystals: weak intermolecular interactions. *Chem. Soc. Rev.* **38**: 2622-2633.

Duszczyk MM, Wutz A, Rybin V, Sattler M. 2011. The Xist RNA A-repeat comprises a novel AUCG tetraloop fold and a platform for multimerization. *RNA* **17**: 1973-1982.

Egli M, Gessner RV. 1995. Stereoelectronic effects of deoxyribose O4' on DNA conformation. *Proc. Natl. Acad. Sci. USA* **92**: 180-184.

Egli M, Sarkhel S. 2007. Lone pair-aromatic interactions: to stabilize or not to stabilize. *Acc. Chem. Res.* **40**: 197-205.

Ennifar E, Nikulin A, Tishchenko S, Serganov A, Nevskaya N, Garber M, Ehresmann B, Ehresmann C, Nikonov S, Dumas P. 2000. The crystal structure of UUCG tetraloop. *J. Mol. Biol.* **304**: 35-42.

Erat MC, Coles J, Finazzo C, Knobloch B, Sigel RKO. 2012. Accurate analysis of Mg^{2+} binding to RNA: From classical methods to a novel iterative calculation procedure. *Coord. Chem. Rev.* **256**: 279-288.

Estarellas C, Quinonero D, Deya PM, Frontera A. 2013. Anion-pi Interactions Involving [MXn](m-) Anions: A Comprehensive Theoretical Study. *Chemphyschem* **14**: 145-154.

Eyal Z, Matzov D, Krupkin M, Wekselman I, Paukner S, Zimmerman E, Rozenberg H, Bashan A, Yonath A. 2015. Structural insights into species-specific features of the ribosome from the pathogen Staphylococcus aureus. *Proc Natl Acad Sci U S A* **112**: E5805-5814.

Fernald RD. 2011. Systems biology meets behavior. *Proc. Natl. Acad. Sci. USA* **108**: 17861-17862.

Fernandez IS, Bai XC, Murshudov G, Scheres SH, Ramakrishnan V. 2014. Initiation of translation by cricket paralysis virus IRES requires its translocation in the ribosome. *Cell* **157**: 823-831.

Fica SM, Tuttle N, Novak T, Li NS, Lu J, Koodathingal P, Dai Q, Staley JP, Piccirilli JA. 2013. RNA catalyses nuclear pre-mRNA splicing. *Nature* **503**: 229-234.

Fiore JL, Nesbitt DJ. 2013. An RNA folding motif: GNRA tetraloop-receptor interactions. *Q. Rev. Biophys.* **46**: 223-264.

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. 1998. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**: 806-811.

Fischer N, Neumann P, Konevega AL, Bock LV, Ficner R, Rodnina MV, Stark H. 2015. Structure of the E. coli ribosome-EF-Tu complex at <3 A resolution by Cs-corrected cryo-EM. *Nature* **520**: 567-570.

Fisher SJ, Blakeley MP, Cianci M, McSweeney S, Helliwell JR. 2012. Protonation-state determination in proteins using high-resolution X-ray crystallography: effects of resolution and completeness. *Acta Cryst.* **D68**: 800-809.

Gabdulkhakov A, Nikonov S, Garber M. 2013. Revisiting the Haloarcula marismortui 50S ribosomal subunit model. *Acta Cryst.* **D69**: 997-1004.

Garreau de Loubresse N, Prokhorova I, Holtkamp W, Rodnina MV, Yusupova G, Yusupov M. 2014. Structural basis for the inhibition of the eukaryotic ribosome. *Nature* **513**: 517-522.

Garst AD, Edwards AL, Batey RT. 2011. Riboswitches: structures and mechanisms. *Cold Spring Harb. Perspect. Biol.* **3**: a003533.

Gaudin C, Ghazal G, Yoshizawa S, Elela SA, Fourmy D. 2006. Structure of an AAGU tetraloop and its contribution to substrate selection by yeast RNase III. *J Mol Biol* **363**: 322-331.

Geary C, Baudrey S, Jaeger L. 2008. Comprehensive features of natural and in vitro selected GNRA tetraloop-binding receptors. *Nucleic Acids Res* **36**: 1138-1152.

Giese M, Albrecht M, Plum C, Hintzen D, Valkonen A, Rissanen K. 2012. Pentafluorophenyl salicylamine receptors in anion–π interaction studies. *Supramolecular Chemistry* **24**: 755-761.

Giese M, Albrecht M, Rissanen K. 2015. Anion–π interactions with fluoroarenes. *Chem. Rev.*

Gogala M, Becker T, Beatrix B, Armache JP, Barrio-Garcia C, Berninghausen O, Beckmann R. 2014. Structures of the Sec61 complex engaged in nascent peptide translocation or membrane insertion. *Nature* **506**: 107-110.

Gottstein-Schmidtke SR, Duchardt-Ferner E, Groher F, Weigand JE, Gottstein D, Suess B, Wohnert J. 2014. Building a stable RNA U-turn with a protonated cytidine. *RNA* **20**: 1163-1172.

Greber BJ, Bieri P, Leibundgut M, Leitner A, Aebersold R, Boehringer D, Ban N. 2015. Ribosome. The complete structure of the 55S mammalian mitochondrial ribosome. *Science* **348**: 303-308.

Greber BJ, Gerhardy S, Leitner A, Leibundgut M, Salem M, Boehringer D, Leulliot N, Aebersold R, Panse VG, Ban N. 2016. Insertion of the Biogenesis Factor Rei1 Probes the Ribosomal Tunnel during 60S Maturation. *Cell* **164**: 91-102.

Groom CR, Bruno IJ, Lightfoot MP, Ward SC. 2016. The Cambridge Structural Database. *Acta Cryst.* **B72**: 171-179.

Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S. 1983. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35**: 849-857.

Haldar S, Kuhrova P, Banas P, Spiwok V, Sponer J, Hobza P, Otyepka M. 2015. Insights into stability and folding of GNRA and UNCG tetraloops revealed by microsecond molecular dynamics and well-tempered metadynamics. *J. Chem. Theory and Comput.* **11**: 3866-3877.

Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, Terns RM, Terns MP. 2009. RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex. *Cell* **139**: 945-956.

Hall KB. 2015. Mighty tiny. *RNA* **21**: 630-631.

Hashem Y, Auffinger P. 2009. A short guide to molecular dynamics simulations of RNA systems. *Methods* **47**: 187-197.

Hashem Y, des Georges A, Fu J, Buss SN, Jossinet F, Jobe A, Zhang Q, Liao HY, Grassucci RA, Bajaj C et al. 2013. High-resolution cryo-electron microscopy structure of the Trypanosoma brucei ribosome. *Nature* **494**: 385-389.

Hay BP, Custelcean R. 2009. Anion-p Interactions in Crystal Structures: Commonplace or Extraordinary? *Cryst. Growth & Design* **9**: 2539-2545.

Hendrix DK, Brenner SE, Holbrook SR. 2005. RNA structural motifs: building blocks of a modular biomolecule. *Q. Rev. Biophys.* **38**: 221-243.

Henkin TM. 2014. The T box riboswitch: A novel regulatory RNA that utilizes tRNA as its ligand. *Biochim Biophys Acta* **1839**: 959-963.

Heus HA, Pardi A. 1991. Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. *Science* **253**: 191-194.

Holbrook SR. 2005. RNA structure: the long and the short of it. *Curr. Opin. Struct. Biol.* **15**: 302-308.

Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, Merrill SH, Penswick SH, Zamir JR. 1965. Structure of a ribonucleic acid. *Science* **147**: 1462-1465.

Hospital A, Goni JR, Orozco M, Gelpi JL. 2015. Molecular dynamics simulations: advances and applications. *Adv. Appl. Bioinform. Chem.* **8**: 37-47.

Humphrey W, Dalke A, Schulten K. 1996. VMD: visual molecular dynamics. *J Mol Graph* **14**: 33-38.

Hwang JW, Dial BE, Li P, Kozik ME, Smith MD, Shimizu KD. 2015. How important are dispersion interactions to the strength of aromatic stacking interactions in solution? *Chem. Sci.* **6**: 4358-4364.

Jaeger L, Michel F, Westhof E. 1994. Involvement of a GNRA tetraloop in long-range RNA tertiary interactions. *J. Mol. Biol.* **236**: 1271-1276.

Jossinet F, Ludwig TE, Westhof E. 2010. Assemble: an interactive graphical tool to analyze and build RNA architectures at the 2D and 3D levels. *Bioinformatics* **26**: 2057-2059.

Jucker FM, Heus HA, Yop PF, Moors HHM, Pardi A. 1996. A network of heterogeneous hydrogen bonds in GNRA tetraloops. *J. Mol. Biol.* **264**: 968-980.

Jucker FM, Pardi A. 1995a. GNRA tetraloops make a U-turn. *RNA* **1**: 219-222.

Jucker FM, Pardi A.. 1995b. Solution structure of the CUUG hairpin loop: a novel RNA tetraloop motif. *Biochemistry* **34**: 14416-14427.

Kaminishi T, Schedlbauer A, Fabbretti A, Brandi L, Ochoa-Lizarralde B, He CG, Milon P, Connell SR, Gualerzi CO, Fucini P. 2015. Crystallographic characterization of the ribosomal binding site and molecular mechanism of action of Hygromycin A. *Nucleic Acids Res.* **43**: 10015-10025.

Kapral GJ, Jain S, Noeske J, Doudna JA, Richardson DC, Richardson JS. 2014. New tools provide a second look at HDV ribozyme structure, dynamics and cleavage. *Nucleic Acids Res.* **42**: 12833-12846.

Keating KS, Toor N, Pyle AM. 2008. The GANC tetraloop: a novel motif in the group IIC intron structure. *J. Mol. Biol.* **383**: 475-481.

Khatter H, Myasnikov AG, Natchiar SK, Klaholz BP. 2015. Structure of the human 80S ribosome. *Nature* **520**: 640-645.

Kim SH, Quigley GJ, Suddath FL, McPherson A, Sneden D, Kim JJ, Weinzierl J, Rich A. 1973a. Three-dimensional structure of yeast phenylalanine transfer RNA: folding of the polynucleotide chain. *Science* **179**: 285-288.

Kim SH, Quigley GJ, Suddath FL, McPherson A, Sneden D, Krse JJ, Weinzierl J, Rich A. 1973b. Three-dimensional structure of yeast phenylalanine transfer RNA: folding of the polynucleotide chain. *Science*: 285-288.

Kleywegt GJ, Harris MR, Zou JY, Taylor TC, Wahlby A, Jones TA. 2004. The Uppsala electron-density server. *Acta Cryst.* **D60**: 2240-2249.

Klinge 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.

Klosterman PS, Hendrix DK, Tamura M, Holbrook SR, Brenner SE. 2004. Three-dimensional motifs from the SCOR, structural classification of RNA database: extruded strands, base triples, tetraloops and U-turns. *Nucleic Acids Res.* **32**: 2342-2352.

Koh CS, Brilot AF, Grigorieff N, Korostelev AA. 2014. Taura syndrome virus IRES initiates translation by binding its tRNA-mRNA-like structural element in the ribosomal decoding center. *Proc Natl Acad Sci U S A* **111**: 9139-9144.

Kondo J, Westhof E. 2011. Classification of pseudo pairs between nucleotide bases and amino acids by analysis of nucleotide-protein complexes. *Nucleic Acids Res.* **39**: 8628-8637.

Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR. 1982. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell* **31**: 147-157.

Krygowski TM, Szatylowicz H, Stasyuk OA, Dominikowska J, Palusiak M. 2014. Aromaticity from the viewpoint of molecular geometry: application to planar systems. *Chem Rev* **114**: 6383-6422.

Kubickova A, Krizek T, Coufal P, Wernersson E, Heyda J, Jungwirth P. 2011. Guanidinium cations pair with positively charged arginine side chains in water. *J. Phys. Chem. Lett.* **2**: 1387-1389.

Kuhrova P, Best RB, Bottaro S, Bussi G, Sponer J, Otyepka M, Banas P. 2016. Computer folding of RNA tetraloops: identification of key force field deficiencies. *J. Chem. Theory Comput.*

Kumar V, Chen Y, Ero R, Ahmed T, Tan J, Li Z, Wong AS, Bhushan S, Gao YG. 2015. Structure of BipA in GTP form bound to the ratcheted ribosome. *Proc Natl Acad Sci U S A* **112**: 10944-10949.

Kurata S, Weixlbaumer A, Ohtsuki T, Shimazaki T, Wada T, Kirino Y, Takai K, Watanabe K, Ramakrishnan V, Suzuki T. 2008. Modified uridines with C5-methylene substituents at the first position of the tRNA anticodon stabilize U.G wobble pairing during decoding. *J. Biol. Chem.* **283**: 18801-18811.

Kürova P, Otyepka M, Sponer J, Banas P. 2014. Are Waters around RNA More than Just a Solvent? − An Insight from Molecular Dynamics Simulations. *J. Chem. Theory. Comput.* **10**: 401-411.

Ladner JE, Jack A, Robertus JD, Brown RS, Rhodes D, Clark BF, Klug A. 1975. Structure of yeast phenylalanine transfer RNA at 2.5 A resolution. *Proc Natl Acad Sci U S A* **72**: 4414-4418.

Lambert D, Leipply D, Shiman R, Draper DE. 2009. The influence of monovalent cation size on the stability of RNA tertiary structures. *J. Mol. Biol.* **390**: 791-804.

Lawson CL, Patwardhan A, Baker ML, Hryc C, Garcia ES, Hudson BP, Lagerstedt I, Ludtke SJ, Pintilie G, Sala R et al. 2016. EMDataBank unified data resource for 3DEM. *Nucleic Acids Res.* **44**: D396-403.

Lee RC, Feinbaum RL, Ambros V. 1993. The C-elegans heterochronic gene Lin-4 Encodes small Rnas with antisense complementarity to Lin-14. *Cell* **75**: 843-854.

Lehmann J, Jossinet F, Gautheret D. 2013. A universal RNA structural motif docking the elbow of tRNA in the ribosome, RNAse P and T-box leaders. *Nucleic Acids Res* **41**: 5494-5502.

Leontis NB, Lescoute A, Westhof E. 2006. The building blocks and motifs of RNA architecture. *Curr. Opin. Struct. Biol.* **16**: 279-287.

Leontis NB, Stombaugh J, Westhof E. 2002. The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.* **30**: 3497-3531.

Leontis NB, Westhof E. 2001. Geometric nomenclature and classification of RNA base pairs. *RNA* **7**: 499-512.

Leontis NB, Westhof E. 2003. Analysis of RNA motifs. *Curr. Opin. Struct. Biol.* **13**: 300-308.

Leontis NB, Zirbel CL. 2012. Nonredundant 3D Structure Datasets for RNA Knowledge Extraction and Benchmarking. In *RNA 3D Structure Analysis and Prediction*, (ed. NB Leontis, E Westhof), pp. 281-298.

Levitt M. 2001. The birth of computational structural biology. *Nat Struct Biol* **8**: 392-393.

Liu Q, Fredrick K. 2016. Intersubunit Bridges of the Bacterial Ribosome. *J Mol Biol*.

Lu XJ, Bussemaker HJ, Olson WK. 2015. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.* **43**: e142.

Lu XJ, Olson WK. 2003. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* **31**: 5108-5121.

Ma JC, Dougherty DA. 1997. The cation-π interaction. *Chem. Rev.* **97**: 1303-1324.

Macrae CF, Bruno IJ, Chisholm JA, Edgington PR, McCabe P, Pidcock E, Rodriguez-Monge L, Taylor R, van de Streek J, Wood PA. 2008. Mercury CSD 2.0 - New features for the visualization and investigation of crystal structures. *J. Appl. Cryst.* **41**: 466-470.

Marcaida MJ, Prieto J, Redondo P, Nadra AD, Alibes A, Serrano L, Grizot S, Duchateau P, Paques F, Blanco FJ et al. 2008. Crystal structure of I-DmoI in complex with its target DNA provides new insights into meganuclease engineering. *Proc. Natl. Acad. Sci. USA* **105**: 16888-16893.

Marraffini LA, Sontheimer EJ. 2010. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nature Reviews Genetics* **11**: 181-190.

Masquida B, Beckert B, Jossinet F. 2010. Exploring RNA structure by integrative molecular modelling. *N. Biotechnol.* **27**: 170-183.

McCammon JA, Gelin BR, Karplus M. 1977. Dynamics of folded proteins. *Nature* **267**: 585-590.

Mecozzi S, West AP, Jr., Dougherty DA. 1996. Cation-π interactions in aromatics of biological and medicinal interest: electrostatic potential surfaces as a useful qualitative guide. *Proc. Natl. Acad. Sci. USA* **93**: 10566-10571.

Melchers WJG, Zoll J, Tessari M, Bakhmutov DV, Gmyl AP, Agol VI, Heus HA. 2006. A GCUA tetranucleotide loop found in the poliovirus oriL by in vivo SELEX (un)expectedly forms a YNMG-like structure: Extending the YNMG family with GYYA. *RNA* **12**: 1671-1682.

Mercer TR, Dinger ME, Mattick JS. 2009. Long non-coding RNAs: insights into functions. *Nat. Rev. Gen.* **10**: 155-159.

Michel F, Westhof E. 1990. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J. Mol. Biol.* **216**: 585-610.

Moore PB. 1999. Structural motifs in RNA. *Annu. Rev. Biochem.* **68**: 287-300.

Morris KV, Mattick JS. 2014. The rise of regulatory RNA. *Nat. Rev. Genet.* **15**: 423-437.

Motorin Y, Grosjean H. 2005. Transfer RNA modification. *ELS*.

Mullins EA, Shi R, Parsons ZD, Yuen PK, David SS, Igarashi Y, Eichman BF. 2015. The DNA glycosylase AlkD uses a non-base-flipping mechanism to excise bulky lesions. *Nature* **527**: 254-258.

Murray LJ, Arendall WB, 3rd, Richardson DC, Richardson JS. 2003. RNA backbone is rotameric. *Proc. Natl. Acad. Sci. USA* **100**: 13904-13909.

Nakano S, Miyoshi D, Sugimoto N. 2014. Effects of molecular crowding on the structures, interactions, and functions of nucleic acids. *Chem. Rev.* **114**: 2733-2758.

Naray-szabo G, Ferenczy GG. 1995. Molecular electrostatics. *Chem. Rev.* **95**: 829-847.

Nissen P, Ippolito JA, Ban N, Moore PB, Steitz TA. 2001. RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc. Natl. Acad. Sci. USA* **98**: 4899-4903.

Noeske J, Wasserman MR, Terry DS, Altman RB, Blanchard SC, Cate JH. 2015. High-resolution structure of the *Escherichia col*i ribosome. *Nat. Struct. Mol. Biol.* **22**: 336-341.

Noller HF. 2005. RNA structure: reading the ribosome. *Science* **309**: 1508-1514.

Nozinovic S, Furtig B, Jonker HR, Richter C, Schwalbe H. 2010. High-resolution NMR structure of an RNA model system: the 14-mer cUUCGg tetraloop hairpin RNA. *Nucleic Acids Res.* **38**: 683-694.

Pace CN, Grimsley GR, Scholtz JM. 2009. Protein ionizable groups: pK values and their contribution to protein stability and solubility. *J. Biol. Chem.* **284**: 13285-13289.

Page MJ, Di Cera E. 2006. Role of Na^+ and K^+ in enzyme function. *Physiol. Rev.* **86**: 1049-1092.

Paudel BP, Rueda D. 2014. Molecular crowding accelerates ribozyme docking and catalysis. *J Am Chem Soc* **136**: 16700-16703.

Pechlaner M, Sigel RK. 2012. Characterization of metal ion-nucleic acid interactions in solution. *Met. Ions Life Sci.* **10**: 1-42.

Perez A, Marchan I, Svozil D, Sponer J, Cheatham III TE, Laughton CA, Orozco M. 2007. Refinement of the amber force field for nucleic acids. Improving the description of {alpha}/{gamma} conformers. *Biophys. J.* **92**: 3817-3829.

Persch E, Dumele O, Diederich F. 2015. Molecular recognition in chemical and biological systems. *Angewandte Chemie-International Edition* **54**: 3290-3327.

Petrov AS, Bernier CR, Hershkovits E, Xue Y, Waterbury CC, Hsiao C, Stepanov VG, Gaucher EA, Grover MA, Harvey SC et al. 2013. Secondary structure and domain architecture of the 23S and 5S rRNAs. *Nucleic Acids Res.* **41**: 7522-7535.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**: 1605-1612.

Placido D, Brown BA, Lowenhaupt K, Rich A, Athanasiadis A. 2007. A left-handed RNA double helix bound by the Zα domain of the RNA-editing enzyme ADAR1. *Structure* **15**: 395-404.

Pley HW, Flaherty KM, McKay DB. 1994. Model for an Tertiary Interaction from the Structure of an Intermolecular Complex between a GAAA Tetraloop and an RNA Helix. *Nature* **372**: 111-113.

Polikanov YS, Melnikov SV, Soll D, Steitz TA. 2015. Structural insights into the role of rRNA modifications in protein synthesis and ribosome assembly. *Nat. Struct. Mol. Biol.* **22**: 342-344.

Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**: D501-504.

Quigley GJ, Rich A. 1976. Structural domains of transfer RNA molecules. *Science* **194**: 796-806.

Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G. 2000. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* **403**: 901-906.

Reiter NJ, Osterman A, Torres-Larios A, Swinger KK, Pan T, Mondragon A. 2010. Structure of a bacterial ribonuclease P holoenzyme in complex with tRNA. *Nature* **468**: 784-789.

Rich A. 2004. The excitement of discovery. *Annu. Rev. Biochem.* **73**: 1-37.

Rich A, Zhang S. 2003. Timeline: Z-DNA: the long road to biological function. *Nat. Rev. Genet.* **4**: 566-572.

Richardson JS, Schneider B, Murray LW, Kapral GJ, Immormino RM, Headd JJ, Richardson DC, Ham D, Hershkovits E, Williams LD et al. 2008. RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA* **14**: 465-481.

Rivas E, Clements J, Eddy SR. 2016. Lack of evidence for conserved secondary structure in long noncoding RNAs.

Roe DR, Cheatham TE. 2013. PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **9**: 3084-3095.

Saenger W. 1984. *Principles of nucleic acid structure.* Springer-Verlag, New-York.

Salonen LM, Ellermann M, Diederich F. 2011. Aromatic rings in chemical and biological recognition: energetics and structures. *Angew. Chem. Int. Ed. Engl.* **50**: 4808-4842.

Sarver M, Zirbel CL, Stombaugh J, Mokdad A, Leontis NB. 2008. FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J. Math. Biol.* **56**: 215-252.

Schmeing TM, Huang KS, Kitchen DE, Strobel SA, Steitz TA. 2005. Structural insights into the roles of water and the 2' hydroxyl of the P site tRNA in the peptidyl transferase reaction. *Mol. Cell.* **20**: 437-448.

Schottel BL, Chifotides HT, Dunbar KR. 2008. Anion-pi interactions. *Chem Soc Rev* **37**: 68-83.

Shajani Z, Sykes MT, Williamson JR. 2011. Assembly of bacterial ribosomes. *Annu. Rev. Biochem.* **80**: 501-526.

Shalev-Benami M, Zhang Y, Matzov D, Halfon Y, Zackay A, Rozenberg H, Zimmerman E, Bashan A, Jaffe CL, Yonath A et al. 2016. 2.8-A cryo-EM structure of the large ribosomal subunit from the eukaryotic parasite Leishmania. *Cell Rep.* **16**: 288-294.

Sharp KA. 2016. Unpacking the origins of in-cell crowding. *Proc Natl Acad Sci U S A* **113**: 1684-1685.

Shi H, Moore PB. 2000. The crystal structure of yeast phenylalanine tRNA at 1.93 Å resolution: a classic structure revisited. *RNA* **6**: 1091-1105.

Sohmen D, Chiba S, Shimokawa-Chiba N, Innis CA, Berninghausen O, Beckmann R, Ito K, Wilson DN. 2015. Structure of the Bacillus subtilis 70S ribosome reveals the basis for species-specific stalling. *Nat Commun* **6**: 6941.

Sokoloski JE, Godfrey SA, Dombrowski SE, Bevilacqua PC. 2011. Prevalence of *syn* nucleobases in the active sites of functional RNAs. *RNA* **17**: 1775-1787.

Spalding DA. 1873. Instinct. With original observations on young animals. *Macmillian's Magazine* **27**: 282-293.

Spek AL. 2009. Structure validation in chemical crystallography. *Acta Cryst.* **D65**: 148-155.

Steitz TA, Moore PB. 2003. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. *Trends Biochem Sci* **28**: 411-418.

Sun WJ, Li JH, Liu S, Wu J, Zhou H, Qu LH, Yang JH. 2016. RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data. *Nucleic Acids Res* **44**: D259-265.

Temin HM, Mizutani S. 1970. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature* **226**: 1211-1213.

Thapar R, Denmon AP, Nikonowicz EP. 2014. Recognition modes of RNA tetraloops and tetraloop-like motifs by RNA-binding proteins. *Wiley Interdiscip. Rev. RNA* **5**: 49-67.

Tishchenko S, Gabdulkhakov A, Nevskaya N, Sarskikh A, Kostareva O, Nikonova E, Sycheva A, Moshkovskii S, Garber M, Nikonov S. 2012. High-resolution crystal structure of the isolated ribosomal L1 stalk. *Acta Crystallogr D Biol Crystallogr* **68**: 1051-1057.

Tishchenko S, Nikonova E, Nikulin A, Nevskaya N, Volchkov S, Piendl W, Garber M, Nikonov S. 2006. Structure of the ribosomal protein L1-mRNA complex at 2.1 angstrom resolution: common features of crystal packing of L1-RNA complexes. *Acta Cryst.* **D62**: 1545-1554.

Treger M, Westhof E. 2001. Statistical analysis of atomic contacts at RNA-protein interfaces. *J Mol Recognit* **14**: 199-214.

Tucker BJ, Breaker RR. 2005. Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.* **15**: 342-348.

Tuerk C, Gauss P, Thermes C, Groebe DR, Gayle M, Guild N, Stormo G, d'Aubenton-Carafa Y, Uhlenbeck OC, Tinoco I, Jr. et al. 1988. CUUCGG hairpins: extraordinarily stable RNA secondary structures associated with various biochemical processes. *Proc. Natl. Acad. Sci. USA* **85**: 1364-1368.

Uhlenbeck OC. 1990. Tetraloops and RNA folding. *Nature* **346**: 613-614.

Valegård K, Murray JB, Stockley PG, Stonehouse NJ, van den Worm S, Stockley PG, Lijas L. 1997. The three dimensional structures of two complexes between recombinant MS2 capsids ans RNA operator fragments reveal sequence-specific protein-RNA interactions. *J. Mol. Biol.* **270**: 724-738.

Vazdar M, Vymetal J, Heyda J, CVondrasek J, Jungwirth P. 2011. Like-charge guanidinium pairing from molecular dynamics and Ab Initio calculations. *J. Phys. Chem. A* **115**: 11193-11201.

Voorhees RM, Fernandez IS, Scheres SH, Hegde RS. 2014. Structure of the mammalian ribosome-Sec61 complex to 3.4 A resolution. *Cell* **157**: 1632-1643.

Wagner GP, Pavlicev M, Cheverud JM. 2007. The road to modularity. *Nat. Rev. Genet.* **8**: 921-931.

Wahlestedt C. 2013. Targeting long non-coding RNA to therapeutically upregulate gene expression. *Nat Rev Drug Discov* **12**: 433-446.

Wang AHJ, Hakoshima T, van der Marel G, Boom JH, Rich A. 1984. AT base pairs are Less stable than GC base pairs in Z-DNA: the crystal structure of d(m5CGTAm5CG). *Cell* **37**: 321-331.

Wang AHJ, Quigley GJ, Kolpak FJ, Vandermarel G, Vanboom JH, Rich A. 1981. Left-handed double helical DNA - Variations in the backbone conformation. *Science* **211**: 171-176.

Wang DX, Wang MX. 2013. Anion-π interactions: generality, binding strength, and structure. *J. Am. Chem. Soc.* **135**: 892-897.

Waterhouse PM, Graham HW, Wang MB. 1998. Virus resistance and gene silencing in plants can be induced by simultaneous expression of sense and antisense RNA. *Proc. Natl. Acad. Sci. U S A* **95**: 13959-13964.

Weisser M, Voigts-Hoffmann F, Rabl J, Leibundgut M, Ban N. 2013. The crystal structure of the eukaryotic 40S ribosomal subunit in complex with eIF1 and eIF1A. *Nat Struct Mol Biol* **20**: 1015-1017.

Westhof E. 1988. Water: an integral part of nucleic acid structure. *Annu. Rev. Biophys., Biophys. Chem.* **17**: 125-144.

Westhof E, Auffinger P. 2000. RNA tertiary structure. In *Encyclopedia of analytical chemistry.*, (ed. RA Meyers), pp. 5222-5232. John Wiley & Sons, Ltd, Chichester.

Wheeler SE, Bloom JW. 2014a. Anion-π interactions and positive electrostatic potentials of N-heterocycles arise from the positions of the nuclei, not changes in the π-electron distribution. *Chem. Commun.* **50**: 11118-11121.

Wheeler SE, Bloom JW. 2014b. Toward a more complete understanding of noncovalent interactions involving aromatic rings. *J. Phys. Chem. A* **118**: 6133-6147.

Wheeler SE, Houk KN. 2009. Through-space effects of substituents dominate molecular electrostatic potentials of substituted arenes. *Journal of Chemical Theory and Computation* **5**: 2301-2312.

Wheeler SE, Houk KN. 2010. Are anion/pi interactions actually a case of simple charge-dipole interactions? *J Phys Chem A* **114**: 8658-8664.

Williams RJP. 2001. Chemical selection of elements by cells. *Coord. Chem. Rev.* **216**: 583-595.

Woese CR, Winker S, Gutell RR. 1990. Architecture of ribosomal RNA: Constraints on the sequence of "tetra-loops". *Proc. Natl. Acad. Sci. USA* **87**: 8467-8471.

Wohlfahrt G. 2005. Analysis of pH-dependent elements in proteins: geometry and properties of pairs of hydrogen-bonded carboxylic acid side-chains. *Proteins* **58**: 396-406.

Wong W, Bai XC, Brown A, Fernandez IS, Hanssen E, Condron M, Tan YH, Baum J, Scheres SH. 2014. Cryo-EM structure of the Plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine. *Elife* **3**.

Woodson SA. 2005. Metal ions and RNA folding: a highly charged topic with a dynamic future. *Curr. Opin. Chem. Biol.* **9**: 104-109.

Wu H, Henras A, Chanfreau G, Feigon J. 2004. Structural basis for recognition of the AGNN tetraloop RNA fold by the double-stranded RNA-binding domain of Rnt1p RNase III. *Proc. Natl. Acad. Sci. USA* **101**: 8307-8312.

Yang LX, Adam C, Nichol GS, Cockroft SL. 2013. How much do van der Waals dispersion forces contribute to molecular recognition in solution? *Nat. Chem.* **5**: 1006-1010.

Zanier K, Luyten I, Crombie C, Muller B, Schumperli D, Linge JP, Nilges M, Sattler M. 2002. Structure of the histone mRNA hairpin required for cell cycle regulation of histone gene expression. *RNA* **8**: 29-46.

Zgarbova M, Otyepka M, Sponer J, Mladek A, Banas P, Cheatham TE, Jurecka P. 2011. Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J. Chem. Theory. Comput.* **7**: 2886-2902.

Zhang J, Ferre-D'Amare AR. 2013. Co-crystal structure of a T-box riboswitch stem I domain in complex with its cognate tRNA. *Nature* **500**: 363-366.

Zhang J, Ferre-D'Amare AR. 2015. Structure and mechanism of the T-box riboswitches. *Wiley Interdiscip. Rev. RNA* **6**: 419-433.

Zhao Q, Huang HC, Nagaswamy U, Xia Y, Gao X, Fox GE. 2012. UNAC tetraloops: to what extent do they mimic GNRA tetraloops? *Biopolymers* **97**: 617-628.

Zheng H, Shabalin IG, Handing KB, Bujnicki JM, Minor W. 2015. Magnesium-binding architectures in RNA crystal structures: validation, binding preferences, classification and motif detection. *Nucleic Acids Res.* **43**: 3789-3801.

Zirbel CL, Sponer JE, Sponer J, Stombaugh J, Leontis NB. 2009. Classification and energetics of the base-phosphate interactions in RNA. *Nucleic Acids Res.* **37**: 4898-4918.

UNIVERSITÉ DE STRASBOURG

Luigi D'ASCENZO

Etude des réseaux de reconnaissance biomoléculaire à l'échelle atomique pour les systèmes ARN et ARN/protéines

École Doctorale des Sciences de la Vie et de la Santé
STRASBOURG

Résumé

Mis à part les liaisons hydrogène, d'autres interactions non covalentes participent dans les réseaux de reconnaissance ARN et ARN-protéines. Parmi celles-ci, j'ai étudié les interactions oxygène-π. Cette interaction prend la forme phosphate-π dans les U-turns et O4'-π dans les motifs ARN-Z. Je propose une nouvelle classification des boucles de quatre nucléotides, décrivant les U-turn et les Z-turn à partir d'interactions oxygène-π. De plus, les motifs "Z-like" présents dans tous les ARN, sont aussi reconnus par certaines protéines immunologiques. Pour mieux comprendre les réseaux de reconnaissance biomoléculaire, nous avons examiné les interactions entre cations/anions et ARN. Nous avons trouvé de nombreuses erreurs dans les structures de la PDB et proposé des règles pour améliorer l'attribution d'espèces ioniques. Les résultats de cette thèse amélioreront notre connaissance des réseaux de reconnaissance biomoléculaire et aideront aux techniques de modélisation structurale des ARN.

Mots clés: interactions oxygène-π; tétraboucles ; ARN-Z ; repliement d'ARN ; solvatation ; interactions ARN-Mg^{2+} ; protonation ; analyse de données PDB

Résumé (anglais)

Together with hydrogen bonds, uncommon non-covalent interactions are fundamental for recognition networks in RNA and RNA-protein systems. Among them, I focused on oxygen-π stacking. This interaction takes the form of phosphate-π within U-turns and of ribose O4'-π within "Z-RNA" motifs. In that respect, a novel classification of tetraloops is proposed, defining U-turns and Z-turns based on their oxygen-π stacking properties. Further, "Z-like" motifs are found to pervade small and large RNAs, being also a recognition pattern for immunology-related proteins. To better understand biomolecular recognition networks, we reviewed the binding of metal ions and anions within RNA, finding many examples of Mg^{2+} misattribution in PDB structures. We propose rules to avoid attribution errors. The results of this thesis will improve our knowledge and understanding of biomolecular recognition networks, as well as assist structural determination and structural modelling techniques of RNA systems.

Keywords: oxygen-π interactions; tetraloops; Z-RNA; RNA folding; solvation; RNA-Mg^{2+} interactions; protonation; PDB data mining