Isostericity and tautomerism of base pairs in nucleic acids

Eric Westhof

Architecture et Réactivité de l’ARN, Université de Strasbourg, Institut de Biologie Moléculaire et Cellulaire, CNRS, 15 rue René Descartes, 67084 Strasbourg, France

Review

Isostericity and tautomerism of base pairs in nucleic acids

Isostericity and tautomerism of base pairs in nucleic acids

The natural bases of nucleic acids form a great variety of base pairs with at least two hydrogen bonds between them. They are classified in twelve main families, with the Watson–Crick family being one of them. In a given family, some of the base pairs are isosteric between them, meaning that the positions and the distances between the C1 carbon atoms are very similar. The isostericity of Watson–Crick/Watson–Crick pairs between the complementary bases forms the basis of RNA helices and of the resulting RNA secondary structure. Several defined suites of non-Watson–Crick base pairs assemble into RNA modules that form recurrent, rather regular, building blocks of the tertiary architecture of folded RNAs. RNA modules are intrinsic to RNA architecture and therefore disconnected from a biological function specifically attached to a RNA sequence. RNA modules occur in all kingdoms of life and in structured RNAs with diverse functions. Because of chemical and geometrical constraints, isostericity between non-Watson–Crick pairs is restricted and this leads to higher sequence conservation in RNA modules with, consequently, greater difficulties in extracting 3D information from sequence analysis. Nucleic acid helices have to be recognised in several biological processes like replication or translational decoding. In polymerases and the ribosomal decoding site, the recognition occurs on the minor groove sides of the helical fragments. With the use of alternative conformations, protonated or tautomeric forms of the bases, some base pairs with Watson–Crick-like geometries can form and be stabilized. Several of these pairs with Watson–Crick-like geometries extend the concept of isostericity beyond the number of isosteric pairs formed between complementary bases. These observations set therefore limits and constraints to geometric selection in molecular recognition of complementary Watson–Crick pairs for fidelity in replication and translation processes.

1. Introduction

A distinguishing feature of RNA molecules is the formation of hydrogen-bonded pairs between the bases along the polymer. These pairs can be intramolecular, meaning a folding back on itself of the polymer, or intermolecular between two identical or different single-stranded RNA molecules. Base–base interactions present in nucleic acids comprise generally between one and three H-bonds, although base–base oppositions without any H-bond can be observed in crystal structures. The base pairs involving at least two “standard” H-bonds, can be ordered into twelve families where each family is a 4 × 4 matrix between the usual four bases [1,2]. Half of the twelve families present the ribose moieties in cis (on the same side of the line of approach of the H-bonds), and the other half in trans (on opposite sides of the line of approach). The common Watson–Crick pairs belong to one of these families, the cis Watson–Crick/Watson–Crick family. The other eleven families gather the non-Watson–Crick pairs that appear generally in folded RNAs. In some of those twelve families, the 4 × 4 matrix is partially filled because only some base–base contacts are able to lead to the formation of two “standard” H-bonds with proper geometry and distances. Very generally, each type of base pair is linked to a specific relative orientation of the sugar-phosphate backbone. Thus, in the cis Watson–Crick/Watson–Crick the strands are antiparallel with the standard conformation of each nucleotide. The cis Watson–Crick pairs form the secondary structure and all the other eleven families are critical for the formation of RNA modules, the building blocks of the tertiary structure.

RNA architecture can thus be viewed as the hierarchical assembly of preformed double-stranded helices defined by cis Watson–Crick base pairs and RNA modules maintained by cis and trans non-Watson–Crick base pairs. RNA modules are recurrent ensemble of ordered non-Watson–Crick base pairs [2,3]. Such RNA modules are often a characteristic of structured non-coding RNAs with specific biological functions, although no specific biological function can be assigned to any RNA module. It is, therefore, important to be able to recognise such genomic elements within genomes [3].
The natural bases of nucleic acids have a strong preference for one tautomeric form, a central chemical fact that guarantees fidelity in their hydrogen bonding potential. Unsurprisingly, the very large majority of hydrogen-bonded base–base interactions underlying secondary and tertiary structures of RNA can be explained with the standard chemical forms of the nucleic acid bases. It is, however, regularly observed that the C and A bases are observed in their protonated forms, at their N3 and N1 positions, respectively. The energy cost associated with the protonation in a neutral environment is largely compensated by the energy gain in the stability of the final architecture where hydrogen bonding and stacking are both fully satisfied. Recently, the occurrence of base pairs implying a tautomeric form of a nucleic acid base has been observed in key biological structures (for a review, see [4]).

2. Isostericity

Geometrically, isostericity between base pairs means that the positions and distances between the C1’ carbon atoms are very similar [5]. The isostericity of the complementary Watson–Crick pairs (G/C and A/U) in the cis Watson–Crick/Watson–Crick family forms the basis of antiparallel RNA helices and the resulting RNA secondary structure (see Fig. 1). In base pairs forming usual double-stranded helices, the C1’ distances are around 10.5 Å. In order for a given base pair to fit within a regular helix, such a distance of 10.5 Å should be maintained between the carbon C1’ atoms. The sets of isosteric pairs in each of the twelve families have been previously analysed and described [5]. Here, some isosteric pairs across families are analysed.

It is worthwhile to note that the non-complementary Watson–Crick/Watson–Crick pairs (G/U and A/C) are isosteric between them with the trans configuration (Fig. 2) (which, with standard conformations of the nucleotides, would imply a parallel arrangement of the two strands and not, as in standard DNA and RNA helices, an antiparallel orientation). In parallel helices, because of the superposition of the approximate twofold axis with the helical axis, the two grooves would be identical in width and depth, unlike in B-DNA or RNA helices. In the trans configuration, the complementary Watson–Crick/Watson–Crick pairs (A/U and G/C) in their usual tautomeric forms are not isosteric (Fig. 3). However, note that with a tautomeric change of the C to the imino form or of the G to the enol form a trans Watson–Crick/Watson–Crick G = C pair isosteric to the trans Watson–Crick/Watson–Crick A–U pair would be formed.

When evaluating the possibilities of forming isosteric base pairs, the conformation of the base with respect to the sugar, syn or anti, should be considered. In the standard conformations, the nucleotides in RNA helices have a sugar in C3’-endo and the base is in the anti conformation with the Watson–Crick edge pointing away from the sugar-phosphate backbone. Also, in the standard conformation, the O5’ atom is above the sugar in the gauche–plus conformation, with a favourable contact to a C–H bond of the base (either C6-H of Y or C8-H of R). In order to enumerate the number of base pairs isosteric to the standard complementary ones with
similar positions in space of the C1’ carbon atoms and C1’ distances around 10.5 Å, one needs to consider all parameters in all possible combinations (protonation, tautomerism, syn/anti equilibrium). Furthermore, all three of these parameters are strongly dependent on base modifications or alterations, since some base modifications change the tautomeric ratio of bases or the syn/anti equilibrium of nucleotides. In addition, any of those changes will alter the stacking patterns between base pairs and, therefore, the overall stability and the preferred base pair.

Here, three modes of formation of base pairs isosteric to the usual Watson–Crick pairs and, thus, with Watson–Crick-like geometries, are described. The first mode involves tautomerism of one of the bases. The second mode involves a change in hydrogen bonding edge from the Watson–Crick to the Hoogsteen edge of one of the bases. The third mode involves modified bases, generally chemical modifications of a U. The first and third modes generally transform wobble geometry (with the pyrimidine base moved towards the major groove side of the base pair) into a Watson–Crick-like geometry. The first two modes are observed with natural bases, although modified bases, especially pseudouridine or inosine, do also participate in base pair formation with Watson–Crick-like geometry. In the last two modes, protonation of one of the bases can also occur. Here, mainly base pairs with two H-bonds are considered because, in those pairs, internal geometrical constraints restrict the number of possibilities, which is not the case with one or no H-bond.

Several of the descriptions and conclusions presented here are applicable in various molecular processes and RNA structures. However, here, the description will be restricted to the situation in the decoding A site of the ribosome (for a previous overview, see [4]).

3. Processes generating base pairs with Watson–Crick-like geometries

3.1. Tautomerism

Some tautomers of the standard bases can form pairs that are isosteric with the usual Watson–Crick pairs.

The four natural nucleic acid bases (A, G, C, U) are characterised by their highly preferred tautomeric form, a chemical fact which is so central for precise and regular recognition. The minor tautomers are estimated to be present only in ratios around 1 for 10⁴ standard states [6–9]. The simple position exchange of the amino and keto groups in G or in C (giving iso-G or iso-C, respectively) yields bases with highly ambivalent tautomerism, iso-G, or with too facile deamination reactions, iso-C [10]. In 1976, Topal and Fresco published two ground breaking articles on base pairing recognition in replication [11] and in translation [12]. They widened the concept of complementarity and analysed with great insight the consequences of base tautomerism in both processes. Indeed, with a keto-enol tautomerism on either base, both the C/A/A/C and U/C/G/U pairs display exactly the same dimensions as the standard complementary pairs C = G /G = C or U–A/A–U (Fig. 4) [4]. This is unlike the situation in the non-isosteric wobble pairs UoG/GoU where the pyrimidine is displaced in the major groove creating a small cavity on the minor groove side.

Fig. 4. Standard Watson–Crick pairs and related tautomeric pairs. Standard complementary U–A (left) and C = G (right) Watson–Crick pairs and the isosteric C ~ A and U ~ G pairs formed using the tautomers imino (for C and A – left) and enol (for U and G – right). The symbol ~ has been used throughout for noting a pair involving a tautomer. The environments of the major (deep) and minor (shallow) grooves of the standard Watson–Crick pairs are the same in both types of the tautomeric forms (imino and enol) for the C ~ A and U ~ G pairs.
In their article on “Base pairing and fidelity in codon-anticodon interaction”, Topal & Fresco [12] discuss the base pairing schemes, some of which involving tautomerism, that possess the dimensions and shapes close to the complementary Watson–Crick pairs so that they can be accommodated or pass through the sieve formed by the steric and geometric constraints imposed by the ribosome. They stress the point that, while formation of unfavoured tautomers would not occur once the nucleotides are within the ribosomal cavity (mainly because of water exclusion), unfavoured tautomers that are formed before being closed up of the ribosomal cavity, and according to solution equilibria imposed by chemical potentials, would be locked in. Tautomeric and isosteric U~G/G~U base pairs were recently observed in crystal structures of the ribosome with mRNA bound in presence of near-cognate tRNAs [13,14] (see Fig. 5).

**3.2. Non-Watson–Crick pairs**

Some cis Watson–Crick/Hoogsteen base pairs are isosteric to usual Watson–Crick pairs.

In the cis Watson–Crick/Hoogsteen base pairs, the Hoogsteen edge of one base faces the Watson–Crick edge of the other with both sugars on the same side of the line of approach along the hydrogen bonds. With such a configuration, the sugar-phosphate backbones would run in a parallel fashion with all other stereo-chemical parameters (syn/anti equilibrium and sugar-phosphate torsion angles) being identical. In order to accommodate such a base pair type within an antiparallel helix, the easiest way is to flip the base from the anti to the syn conformation with a possible reorientation of the hydroxyl OS′ atom from the gauche-plus to the trans conformer. An example is the cis Watson–Crick/Hoogsteen base pair between A and U base pair (noted Asyn–U). The distance between the C1′ atoms is 10.5 Å, but in order to fit within an antiparallel helix, the base presenting its Hoogsteen edge, U in this case, should be in the syn orientation. Such a pair is mediated via a C5-H...N1 H-bond between the uracil and the adenine. Note, however, that with a pseudouridine, instead of a uridine, forming a Asyn–Vsyn base pair, a much stronger H-bond N1-H...N1 between the pseudouridine and the adenine is formed. Pseudouridines in the anticodon do not constitute a common situation in the codon/anticodon minihelix, but they do occur in specific situations for modulating decoding (for review see [15]).

A special situation occurs with the pairs involving two purines (R-R). The C1′...C1′ distances are slightly longer, around 11 Å for Gsyn-Asyn and 11.6 Å for the non-isosteric Gsyn-Gsyn. However, with a protonated adenine, the resulting A+syn-Gsyn base pair has a C1′...C1′ distance around 10.7 Å. A similar distance occurs in the syn base pair containing the modified base Inosine present at position 34 in the tRNA anticodon. Some of those of base pairs have recently been observed crystallographically in ribosomal subunits [16]. In that structure [16], a water-mediated base pair Gsyn-Gsyn has been observed and could be extrapolated to a Gsyn-Gsyn mediated by a water molecule (Gsyn-Gsyn).

In the context of the codon/anticodon minihelix within the decoding A-site of the ribosome, such base pairs should be easily accommodated. The base in syn can, in principle, belong to either the nucleotide on the messenger side or to the nucleotide in the anticodon loop. When on the messenger side, the most favourable position would be the first nucleotide codon because it follows the sharp kink in the 5′-phosphate along the path of the messenger and the base can rotate more freely [17,18]. When in the anticodon loop, the most favourable position would be also the first anticodon nucleotide (position 34), which also follows the sharp turn at residue 33 that positions base 34 at the most open tip of the anticodon loop. Because of the anti to syn conformational changes and the occurrence of base modifications, stacking interactions between base pairs have also to be considered, possibly altering the preferences solely based on the sugar-phosphate backbone. Extreme stacking situations occur in GoU pairs where the 5′-GoU′-pair has a good intrastrand stacking with the next pair, while the 5′-Uog3′ pair does not stack well with the next pair but shows interstrand stacking [19]. Some possibilities and occurrence of unusual base pairs in codon/anticodon interactions are given in Table 1.

For completeness, the pairs just discussed above should not be confused with those present in triple helices poly(A-U):poly(U) where the cis Watson–Crick/Hoogsteen base pair between poly(U) and the A strand of the poly(A-U) (A=U) is used instead with both bases in the anti conformation [20]. Although a suggestion was made about the possibility to use such base pairs in
Table 1
Some examples of possible use of non-Watson–Crick pairs with Watson–Crick-like geometries in specific cases of tRNA/mRNA interactions involving either a pseudouridine (Ψ') or an inosine (I). The symbols (see [1] for more details) are the following: a black (cis pair) square indicates that the Hoogsteen edge is used, a black (cis pair) circle indicates that the Watson–Crick is used, a W letter in-between means that base pair is water-mediated, a ~ symbol means that a tautomeric form of one of the base is required. The paired bases in the anticodon and codon that are considered are underlined.

| tRNA residue 34 | Type of base pair | mRNA residue + 3 | tRNA anticodon | Codon | Refs. tRNA | Refs. crystal |
|-----------------|-------------------|-----------------|----------------|-------|------------|--------------|
| Ψsyn | [black] | A | Ile ΨΨΨ | ΨΨΨ | [34] | By analogy [16] |
| Gsyn | ●● | G | Tyr ΨΨΨ | ΨΨΨ | [35] |
| Gsyn | ● | A' | Tyr ΨΨΨ | ΨΨΨ | [35] |
| G | ● | Ψsyn | Lys ΨΨΨ | ΨΨΨ | [35] |
| G | ● | Asyn | Tyr ΨΨΨ | ΨΨΨ | [35] |
| G | ● | I | Tyr ΨΨΨ | ΨΨΨ | [16] |
| I | ● | Ψsyn | Tyr ΨΨΨ | ΨΨΨ | [16] |
| Ψ' | Watson–Crick | A | Ile ΨΨΨ | ΨΨΨ | [34] |
| Ψ' | Watson–Crick-like U–G | G | | | |
| C' (Lysidine, Agmatine) | Watson–Crick-like C–A | A | | | |
| A (36) | Watson–Crick | Ψ' (+1) | Ser IGA | ΨΨΨ | [26–28] |
| A (36) | Ψsyn (+1) | Ser IGA | ΨΨΨ | [37] |

U' or C' indicate, respectively, a modified uridine or a modified cytosine.

double-stranded DNA helices [21], it was quickly dismissed [22]. It is worth remembering that in order to obtain isostericity between the complementary bases when forming the cis Watson–Crick/Hoogsteen pairs, U●●A and C●●G, a Watson–Crick shift to the imino form of the C or a protonation of the C is required. It can be remarked that such a (U●●A) pair presents a much shorter C1'...C1' distance around 7.5 Å [23]. Recently, a●●.syn base pairs have been observed transiently in DNA [24].

3.3. Base modifications

Modified bases can lead to pairs isoteric with the usual Watson–Crick pairs.

Several recent crystal structures of the 30S particles in presence of short messengers and of anticodon hairpins with modifications at position 34 have shown that the third base pair of the codon/anticodon complex adopts a Watson–Crick-like geometry [28,29,31,33,36].

The U at position 34 in the anticodon loop of tRNAs is the most frequently and diversely modified nucleoside observed in tRNAs [29]. Position 34 is also called the wobble position because it can pair with the third codon nucleotide in some instances forming wobble base pairs, especially GuoU. Wobble base pairs, unlike the Watson–Crick pairs, are not isoteric upon reversal [19]. This observation is particularly relevant in the case of one of the nucleotides fixed by the molecular environment as it is in the decoding A site of the ribosome. In the decoding site, the nucleotide (+3) is fixed by contacts to the ribosome and the lateral movement of the C1' between the U34 (of a U34oG(+3)) and G34 (of a G34oU(+3)) is around 2 Å. It has therefore been suggested that a G34oU(+3) can be accommodated in the decoding site but a U34oG(+3) not. Modifications of the U34 (U34Ψ') promote the formation of U34Ψ'–G(+3) with Watson–Crick-like geometry that can fit within the tight ribosomal decoding site. The modifications of U34Ψ' can be extensive from CS-alkylation and/or 2-thiolation to C34-alkylation and/or 2-thiolation to the taurine derivatives in mammalian mitochondrial tRNAs [25].

An analogous situation occurs with A34oC(+3) versus C34oA(+3). In tRNAs, A34 is enzymatically deaminated into inosine leading to a Watson–Crick geometry for I34oC(+3), a wobble geometry for I34oU(+3) and, as suggested, to a cis Watson–Crick/Hoogsteen I34●●A(+3)Ψsyn with a Watson–Crick-like geometry. An equivalent type of base pair was recently observed [18]. In bacteria, C34 is modified at position 2 into lysidine in tRNA(Ile) and, in archaea, C34 is modified also at position 2 into agmatine. In both lysidine and agmatine, the cytosine rings are protonated at N1, promoting a tautomeric form and leading to Watson–Crick-like geometries with A [26–28]. Also, 5-formylcytidine (5F) at position-34 in human mitochondrial tRNAMet (anticodon f5CAU) allows the decoding as well of AUG as of AUA by forming a tautomeric pair [29].

Interestingly, the YΨΨΨ pairs between two Us have a single H-bond (O4(U)–N3(C)) with a C1'...C1' distance between 2.8 and 10.6 Å [30] (an analogous pair between two Cs can be formed between N4(C1) and N3(C2). Modified, as well as, unmodified Us can form such base pairs (see [31]). However, pairs between C and U have been observed with a single H-bond (N4(C)...O4(U)) and a water molecule linking the two N3 atoms leading to larger C1'...C1' distances around 12 Å [32].

4. Conclusions

Crystallographic observations showed that the ribosomal grip around the triplet codon/anticodon sterically fits best with the dimensions and volume of a standard RNA helix with the recognition processes occurring in the shallow minor groove side of the three base pairs [13,33]. Consequently, base pairs with Watson–Crick-like geometries and dimensions should be accommodated within the decoding site [4,14]. Several chemical processes can promote the formation of such Watson–Crick-like base pairs, base protonation, base tautomerism, and base modifications, all contribute or concur together with anti-syn conformational changes and/or H-bonding edge variations. Here, some of the principles and possibilities for forming base pairs with Watson–Crick-like geometries are described. Possible base pairs can be suggested when the relative internal geometry of the chemical groups forming H-bonds can serve as a constraining guide (i.e. when two H-bonds are formed). With a single (or no) H-bond, the number of possibilities increase and only external steric constraints can restrict the choice. In any case, definite choices can only be made with extensive and detailed crystallographic studies.

Acknowledgments

The author is grateful to Valérie Fritsch for the molecular drawings and to Gula Yusupova and Marat Yusupov discussions on ribosomal decoding. This work has been published under the framework of the LABEX: ANR-10-LABX 0036_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

[1] Leontis, N.B. and Westhof, E. (2001) Geometric nomenclature and classification of RNA base pairs. RNA 7, 499–512.
[2] Lescoute, A., Leontis, N.B., Massine, C. and Westhof, E. (2005) Recurrent structural RNA motifs. Nucleic Acids Res. 33, 2395–2409.
[3] Cruz, J.A. and Westhof, E. (2011) Sequence-based identification of 3D structural modules in RNA with RMIDetect. Nat. Methods 8, 513–521.
[4] Westhof, M. Yusupov, C. Yusupova, Recognition of Watson–Crick base pairs: constraints and limits due to geometric selection and tautomerism, F1000 Prime Reports, 6 (2014) 19.
[5] Stombaugh, J., Zirbel, C.L., Westhof, E. and Leontis, N.B. (2005) Recurrent structural RNA motifs. Nucleic Acids Res. 33, 2294–2312.
[6] Dreyfus, M., Bensaude, O., Dodin, G. and Dubois, J.E. (1976) Tautomerism in cytosine and 3-methylcytosine. A thermodynamic and kinetic study. J. Am. Chem. Soc. 98, 6338–6340.
[7] Katritzky, A.R. and Waring, A.J. (1962) Tautomeric axines. Part I. The tautomerism of 1-methyluracil and 5-hydroxymethyluracil. J. Chem. Soc., 1540–1544.
[8] Switzer, C.Y., Moroney, S.E. and Benner, S.A. (1993) Enzymatic recognition of Watson–Crick base pairs: constraints and limits due to geometric selection and tautomerism, F1000 Prime Reports, 6 (2014) 19.
[9] Medeiros, G.C. and Thomas, G.J. (1971) On the tautomeric structure of inosine. Biochim. Biophys. Acta 238, 1–4.
[10] Arnott, S. and Selsing, E. (1974) Structures for the polynucleotide complexes poly(dA) with poly(dT) and poly(dT) with poly(dA) with poly(dT). J. Mol. Biol. 88, 509–521.
[11] Langridge, R. and Rich, A. (1960) X-ray diffraction studies of synthetic polynucleotides. Acta Crystallogr. 13, 1052.