© 2000 Oxford University Press Nucleic Acids Research, 2000, Vol. 28, No. 1 375–376

Database of non-canonical base pairs found in known RNA structures

Uma Nagaswamy, Neil Voss, Zhengdong Zhang and George E. Fox*

Department of Biology and Biochemistry, University of Houston, Houston, TX 77204-5934, USA

Received September 2, 1999; Revised and Accepted October 13, 1999

ABSTRACT

Atomic resolution RNA structures are being published at an increasing rate. It is common to find a modest number of non-canonical base pairs in these structures in addition to the usual Watson–Crick pairs. This database summarizes the occurrence of these rare base pairs in accordance with standard nomenclature. The database, http://prion.bchs.uh.edu/ , contains information such as sequence context, sugar pucker conformation, *anti*/*syn* base conformations, chemical shift, pK_a values, melting temperature and free energy. Of the 29 anticipated pairs with two or more hydrogen bonds, 20 have been encountered to date. In addition, four unexpected pairs with two hydrogen bonds have been reported bringing the total to 24. Single hydrogen bond versions of five of the expected geometries have been encountered among the single hydrogen bond interactions. In addition, 18 different types of base triplets have been encountered, each of which involves three to six hydrogen bonds. The vast majority of the rare base pairs are antiparallel with the bases in the *anti* configuration relative to the ribose. The most common are the GU wobble, the Sheared GA pair, the Reverse Hoogsteen pair and the GA imino pair.

INTRODUCTION

RNA structure was once envisioned largely as a collection of stems comprised of Watson–Crick base pairs and the single-stranded loops defined by these stems. The first structure of tRNAs soon revealed otherwise (1). Comparative techniques were developed to reliably identify many G:U wobble interactions and later some non-canonical interactions (mainly A:G mismatches) in sets of aligned RNA sequences (2–4). Nevertheless, it was not technically feasible to elucidate significant numbers of structures until recently. Advances in methodology for studying nucleic acids at high resolution in general, and in NMR in particular, have now dramatically changed this outlook. The recent outpouring of structures for various small RNAs and RNA fragments has, as expected, revealed the presence of many unusual base–base interactions. Several combinations of these non-standard interactions have been found in multiple structures and have been designated as motifs. Nevertheless, the individual non-canonical pair may occur in other situations where the motif is not formed. It therefore will be useful to have a database, which provides the user with easy access to information on any non-canonical base pair.

Beginning with Donohue's early analysis (5), it has been recognized that many alternative interactions between bases exist in which two hydrogen bonds are possible. A recent tabulation by Burkard *et al*. (6) lists 29 possible pairs of this type including the standard pairs. There are two major types of pairs, normal and flipped. In what are considered to be normal pairs, the strands are antiparallel and the base is in the *anti* orientation relative to the ribose. In the flipped arrangements either the strands must be parallel or the base must be switched to the *syn* configuration. Within each category one can further distinguish purine–purine, purine–pyrimidine and pyrimidine–pyrimidine types. In constructing the database, we attempted to classify each base pair as belonging to one of these types. Thus, if one of the hydrogen bonds was water mediated that fact was ignored in deciding the proper classification but noted in the detailed data. In addition to many of the anticipated pairs, several unexpected pairs have been encountered. The best known of these is an AC wobble pair that has been found multiple times. This pair is able to occur because the N1 of the A is protonated, which allows it to act as a hydrogen bond donor. This pair is otherwise a normal purine–pyrimidine pair and is classified as such. Three other unexpected pairs resulting in two or more hydrogen bonds have each been reported once. In addition, a number of non-standard interactions have been encountered in which only one hydrogen bond was formed. Typically these could be classified as a variant of one or two of the primary types. Finally, to date, at least 18 different interactions involving three bases have been encountered. The core of these base triples is typically one of the standard pairs, most commonly either a Watson–Crick or a Reverse Hoogsteen. The third base can interact in a surprising variety of ways to add from one to three additional hydrogen bonds.

DATABASE DESCRIPTION

The primary objective of this database is to provide rapid access to all the RNA structures in which a particular rare base pair has been found. A secondary objective is to summarize the important properties that have been associated with each non-canonical base pair. A literature survey of RNA structures was performed to obtain information on the non-canonical base

*To whom correspondence should be addressed. Tel: +1 713 743 8363; Fax: +1 713 743 8351; Email: frox@uh.edu

pairs. As of this writing, approximately 250 mismatches have been identified. Each of these is tabulated in accordance with the classification scheme described above (6). Interactions that have only one hydrogen bond or involve three bases have been included as separate categories. Each entry provides access to citation information, the immediate sequence context and a brief common name for the structures in which the base pairs were found. Additional structural information such as melting point, chemical shift, comments, etc. is provided when this information is available. In many cases a structure has been described in a series of several papers from the same research group. When this occurs, we have typically only cited the most recent paper but may have included results from the earlier papers. In other instances, different groups have studied the same structure, usually with different methods (i.e. one group using NMR and the other crystallography). In these instances it is not unheard of for differences in interpretation to exist. In these cases we have created separate database entries for each approach that will be linked together in the summary tables. In addition to several static tables of information, a query system has been set-up, which allows browsing of the database in a variety of ways. Thus, for example, the simple query, GA, will locate data on all types of GA pairs. The database can be accessed from http://prion.bchs.uh.edu/

ANALYSIS OF BASE PAIR OCCURENCE

One of the primary reasons for developing this database was the expectation that as it grows it likely will become clear that certain non-standard interactions will be frequently associated with particular local environments. Including the standard Watson–Crick examples, to date 24 bp involving two or more hydrogen bonds have been encountered. This includes 10 of the 11 expected normal pairs. The only normal pair, which has not been found, is the AC Reverse Hoogsteen pair. In terms of total numbers, the vast majority of the non-canonical pairs encountered are normal in that the base is in the *anti* configuration relative to the sugar and the strands are antiparallel. Some patterns are also apparent in the database. For example, the AU Reverse Hoogsteen base pair is typically either at the 5′ end of a stem with a non-canonical pair above it or is flanked by two non-canonical base pairs. The types of neighboring pairs include the GA imino, the AA N7-amino symmetric, a sheared GA base pair and the GU wobble base pair but never a standard Watson–Crick pair. The protonated AC wobble pair is geometrically similar to the GU wobble pair and therefore frequently incorporated into regular helices. Regardless of type, pairs of the GG type are frequently flanked by at least one non-canonical pair. It is expected that observations of this type may lead to predictive rules about mismatches in the internal and terminal loops of unsolved structures of fragments of naturally occurring RNAs. That is to say, it may be possible to eventually develop predictive rules for the occurrence of at least some of the non-standard pairs.

ACKNOWLEDGEMENTS

This work was supported by an internship awarded to NV by the W. M. Keck Center for Computational Biology and by National Aeronautics and Space Administration Grant NAG5-8140 to G.E.F.

REFERENCES

1. Suddath,F.L., Quigley,G.J., McPherson,A., Sneden,D., Kim,J.J., Kim,S.H. and Rich,A. (1974) *Nature*, **248**, 20–24.
2. Gautheret,D., Konings,D. and Gutell,R.R. (1995) *RNA*, **1**, 807–814.
3. Gautheret,D., Konings,D. and Gutell,R.R. (1994) *J. Mol. Biol.*, **242**, 1–8.
4. Gutell,R.R., Power,A., Hertz,G.Z., Putz,E.J. and Stormo,G.D. (1992) *Nucleic Acids Res.*, **20**, 5785–5795.
5. Donohue,J. (1956) *Proc. Natl Acad. Sci. USA*, **42**, 60–65.
6. Burkard,M.E., Turner,D.H. and Tinoco,I.,Jr (1999) In Gesteland,R.F., Cech,T.R. and Atkins,J.F. (eds), *The RNA World -Second Edition.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 675–680.