RAPPORT DE RECHERCHE

L R I

AUTOMATED MOTIF DISCOVERING IN RNA MOLECULES

DJELLOUL M / DENISE A

Unité Mixte de Recherche 8623
CNRS-Université Paris Sud – LRI

03/2008

Rapport de Recherche N° 1490

CNRS – Université de Paris Sud
Centre d'Orsay
LABORATOIRE DE RECHERCHE EN INFORMATIQUE
Bâtiment 490
91405 ORSAY Cedex (France)

Automated motif discovering in RNA molecules

Mahassine Djelloul[1] and Alain Denise[1,2]
[1]LRI, Université Paris-Sud 11 and CNRS
[2]IGM, Université Paris-Sud 11 and CNRS
email:{Mahassine.Djelloul, Alain.Denise}@lri.fr

Abstract

We used a novel graph-based approach to identify recurrent RNA tertiary motifs embedded within secondary structure. We catalogued all the secondary structural elements of the RNA molecule and clustered them using an innovative graph similarity measure. We applied our method to three widely studied structures: *H.m* 50S, *E.coli* 50S and *T.th* 16S. We identified 10 known motifs without any prior knowledge of their shapes or positions. We additionally identified four putative new motifs.

1 Introduction

RNA adopts complex three dimensional (3D) folds to perform biological functions in the cell. This molecular packing is the tertiary structure. Structural studies have revealed that RNA tertiary structure is modular and composed of conserved building blocks called *motifs*, the formation of which is sequence-dependent [2, 25, 36, 9, 11]. Thus, the identification and classification of RNA structural motifs based on both sequence and structure information is of value for RNA folding prediction and modelling.

A number of representations of RNA tertiary structure at different levels of detail have been generated and used to develop automated methods for identifying motifs within RNA molecules. The first basic representations were Cartesian coordinates of the atoms or backbone torsion angles found in 3D structures (X-ray or NMR) [28, 30, 10, 7, 35, 12]. Further studies used these representations to develop graph-theoretical representations [8, 1]. In 2001, a descriptive base-pairing nomenclature was proposed by Leontis and Westhof (LW) to systematically annotate and classify non-WC basepairs [20, 19, 16, 37, 14]. In a LW nomenclature-based representation, the tertiary structure is viewed as a (topological) general graph with vertices representing bases labelled by their sequence letter and residue number, and the edges are the interactions between bases labelled by their type of bond. This high-level and unambiguous representation of sequence and structure information will allow improved understanding of sequence-structure relations.

Motif recognition in structural genomics requires two problems to be addressed:

1. Given a description of a *known* motif, identify this motif in target structures,
2. Given a structure, identify *unknown* motifs within it.

Using graph theory, the problem of identifying a known pattern in a target graph reduces to (i) searching for isomorphic occurrences of the pattern. This, known as subgraph isomorphism, is NP-complete in general graphs (i.e graphs without any restriction on any graph parameter) [34], or (ii) finding similar occurrences of the pattern. Practically, this consists of identifying a maximum common subgraph of two input graphs and calculating a score of similarity based on that common substructure. If the similarity score fulfils certain pre-set conditions, the two graphs are considered similar. However, the maximum common subgraph (MCS) problem is NP-hard, APX-hard and W[1]-hard [13] and such an approach is not feasible except for very small graphs such as those in chemoinformatics [15] in which data objects to be identified are chemical compounds described by planar graphs of small size (up to 15 nodes).

The identification of unknown motifs is made more difficult by the fact that the pattern is equally unknown. Thus, different approaches have been proposed. In particular, one study [35] used a previous work on *RNA worms* [7] to identify recurrent backbone conformations. However, and as pointed out by the authors, these motifs displayed no apparent secondary or primary structure signature and are thus unsuitable for prediction or modelling of RNA. Other studies used the Cartesian coordinates or a derived graph model to search for new patterns in RNA structures [8, 28]. Neither approach, however, addressed the problem of identifying occurrences with inserted bases or basepairs. Indeed, occurrences of a same motif are not always identical but rather display very similar features [22]. The variations observed may be due to natural changes induced by evolution or experimental errors in data collection.

In this paper, we propose a new method for identifying and classifying similar occurrences of *a priori* unknown RNA motifs using the (topological) graph of the tertiary structure. RNA structural motifs are defined as "small, recurrent, directed and ordered stacked arrays of *isosteric* non-WC basepairs that intersperse the secondary structural elements and fold into essentially identical three dimensional structures" [21]. Two non-canonical basepairs are said *isosteric* if they belong to the same geometric family and can substitute each other without distorting the fundamental three-dimensional structure of the motif [21].

In the next section, we introduce our proposal approach for discovering putative RNA motifs.

2 Materials and Methods

2.1 Data

We downloaded crystal structures from the NDB database [3]. We used the annotation program Rnaview [37] to produce the corresponding RNA graph (see details below). We considered 14 types of interactions: the phosphodiester (backbone) link, the canonical WC pairing GC and AU (to which the wobble pairing GU is commonly added) and the 12 non-WC basepairs defined in the Leontis and Westhof (LW) nomenclature [20, 19]. This classification is based on the observation that a non-canonical interaction involves three dictinct edges: the Watson-Crick edge, the Hoogsteen edge and the Sugar edge. The bases interact in either of two orientations with respect to the glycosidic bonds, *cis* or *trans* relative to the hydrogen bonds.

2.2 Methods

2.2.1 Overview

We used a graph-based representation of the RNA tertiary structure with vertices representing the nucleotides labelled by their sequence letter (and their residue number in the sequence), and edges representing the observed interactions between the nucleotides, labelled by the type of chemical bond. These bonds are:

- phosphodiester bonds (backbone) linking nucleotides adjacent in the sequence,
- the WC or canonical pairings (GC, AU) and the wobble pairing GU forming the skeleton of the secondary structure,
- the 12 non-WC (non-canonical) basepairs defined by LW nomenclature.

We considered wobble pairings to be canonical. Backbone links are directed from $5'$ to $3'$ and non-canonical pairings with different interacting edges are directed according to the rule WC > Hoogsteen > Sugar-edge. The rest of the interactions are symmetrical.

We undertook the following three steps:

1. identify all *secondary structural elements* of the RNA tertiary structure;
2. calculate a similarity measure for each pair of structural elements;
3. cluster the structural elements according to the similarity measure.

These steps are detailed below.

1. Identifying secondary structural elements

A previous study [18] identifying RNA motifs described *local* RNA motifs as ”often bracketed” by secondary structural elements. Based on these observations, we took the following approaches: we firstly only considered backbone and canonical interactions (not including pseudoknots). Then, using a classical tree representation of the secondary structure [31, 26], we extracted the structural elements corresponding to the bulges, internal, junction, and terminal loops modelled by graphs given by their vertices (the nucleotides) and their edges (the flanking canonical basepairs). Then, for each secondary structural element, and given that we were looking for local motifs, we restored all non-canonical edges between each of its vertices.

To remove pseudoknots, we used *secrna*, a program developed by Y. Ponty [29] which inputs an RNA pseudoknotted structure and returns its corresponding secondary structure without pseudoknots. The interested reader is referred to [32] for a survey on the related computational methods.

2. Computing a similarity measure between two structural elements

The similarity measure between two structural elements involves computing a *largest extensible common non-canonical subgraph.* The following definitions and notations will be useful to explain this notion. The size of a graph G is defined by the number of its edges. The

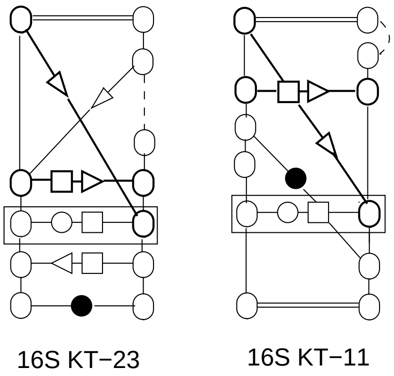


Figure 1: Two structural elements with their LECNS (in bold) of size 2. There is a larger common non-canonical subgraph (size 3) comprising the framed basepair, but it is not extensible. Dashed backbone indicates free nucleotides.

non-canonical size of G, denoted $||G||$, is the number of its non-canonical edges. A graph containing only non-canonical edges is *non-canonical*. A *common non-canonical subgraph* of two graphs G_1 and G_2 is a non-canonical graph H that occurs in both G_1 and G_2.

The *completion* of a non-canonical subgraph H in a graph G is the graph obtained by adding to H all canonical and backbone edges of G with at least one end in H. A common non-canonical subgraph of two graphs G_1 and G_2 is *extensible* if its completions in G_1 and in G_2, respectively, are isomorphic. Now, the *largest extensible common non-canonical subgraph* (LECNS) of G_1 and G_2 is an extensible common non-canonical subgraph of G_1 and G_2 whose size is maximal. Figure 1 illustrates the notion of LECNS.

We implemented an algorithm for computing the LECNS of two given structural elements. Our algorithm makes use of Valiente's graph isomorphism algorithm [34]. To identify the sequence signature of a motif, only the labels of the edges were considered relevant for the mapping.

The similarity between two graphs G_1 and G_2, denoted $sim(G_1, G_2)$, is defined by :

$$sim(G_1, G_2) = \begin{cases} \dfrac{||LECNS(G_1, G_2)||}{max(||G_1||, ||G_2||)} & \text{if } ||LECNS(G_1, G_2)|| > 1 \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

We considered a single common non-canonical edge not to be a relevant motif, and thus included the condition $||LECNS(G_1, G_2)|| > 1$ in the formula. The following properties hold:

- $0 \leq sim(G_1, G_2) \leq 1$,
- $sim(G_1, G_2) = sim(G_2, G_1)$,
- $sim(G_1, G_2) = 1 \Rightarrow$ the completions of the largest non-canonical subgraphs of G_1 and G_2 are isomorphic,
- $sim(G_1, G_2) = 0 \Rightarrow G_1$ and G_2 have no common non-canonical subgraph of size > 1.

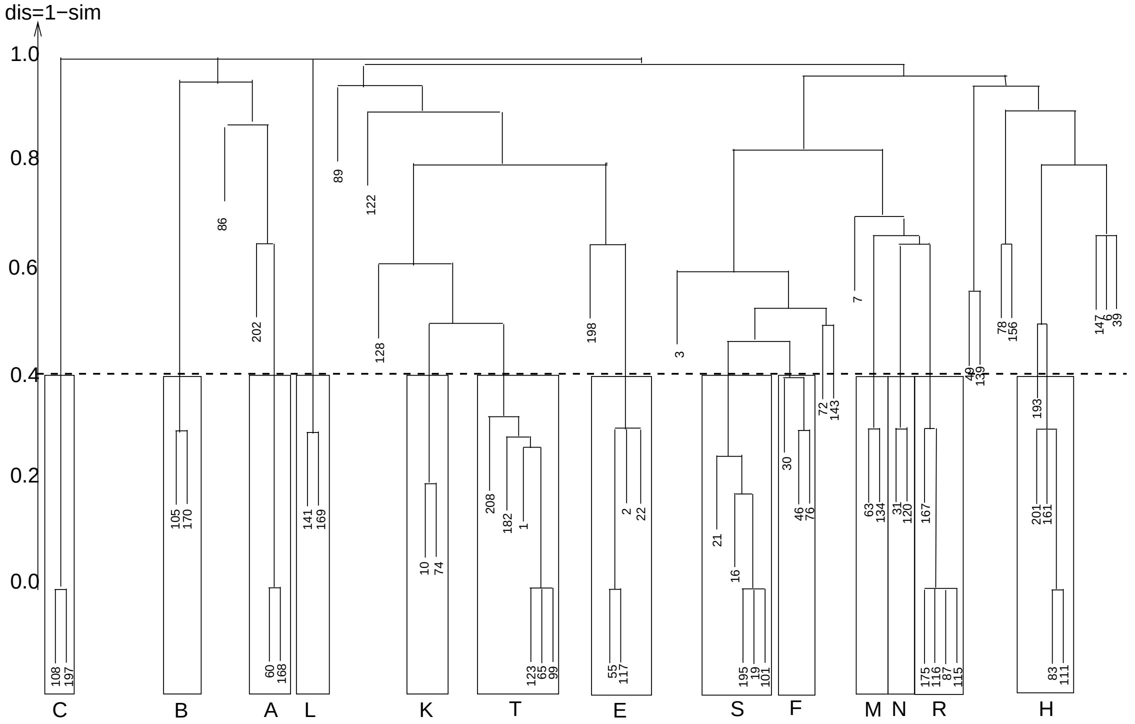


Figure 2: Dendrogram of hierarchical clustering of *H.m* 23S RNA produced with *hclust*. The structural elements are numbered from 1 to 209 (see Catalogue, section 3.1). Rectangular boxes correspond to clusters obtained using the 0.6 similarity threshold. Structural elements clustered with a null similarity value are not shown. See supplementary material.

3. Clustering structural elements

We clustered the structural elements in three steps:

Step 1. We performed a classical hierarchical clustering with average linkage (UPGMA algorithm) analysis based on the measure of similarity defined above. We used the *hclust* function of the *R* Project for statistical computing (http://www.r-project.org/). The resulting dendrogram is presented in Figure 2. (Note that since *hclust* requires a dissimilarity measure, we set $dis(G_1, G_2) = 1 - sim(G_1, G_2)$).

A threshold value was needed to obtain distinct clusters from the tree. This involved defining the minimal similarity value required within a single cluster. Thus, we took the known motifs of *H. m* 23S (E-loop, Sarcin-Ricin, C-loop, K-turn) as a reference [18, 22]. The value giving optimal clustering of these motifs was 0.6 (Figure 2). In particular, it distinguished a perturbed sarcin-ricin occurrence (Helix 23S Junction G475) in *H.m* 23S (fig.4 of [18]) from a variant of 23S E-loop motif (23S G720) (fig.15 of [18]). We checked that all similar members within the same cluster had the same backbone orientation. Structural elements with a different backbone orientation from the other cluster members were not retained. The structural element 2 was thereby excluded from the cluster E (Figure 2).

This first step clustered 41 of the 209 structural elements in *H. m* 23S. We identified 13 clusters, nine of which corresponded to known RNA motifs. Notably, although this threshold value was set using one reference structure *H. m* 23S, it also proved optimal for the other structures.

Step 2. Once the clusters had been generated, we extracted a representative common subgraph, called the *non-canonical core*, for each cluster and used it to identify a consensus structure for the cluster. The *non-canonical core* of a cluster is the largest extensible non-canonical subgraph common to more than 50% of the total number of members in the cluster. We checked whether the structural environment surrounding the non-canonical core shares common features at the level of the secondary structure. Clusters L, M and N did not have such common features. Each of these clusters contained an internal loop and a junction loop from which no consensus structure could be derived. The clustering of these structural elements based solely on graph-similarity criteria could not be explained biologically; thus, the corresponding clusters were not considered to be relevant potential motifs.

Step 3. We used the non-canonical core of clusters retained for further analysis to perform graph-based comparisons with given structural elements. Thus, structural elements not belonging to any cluster but containing this core and consistent with the consensus structure were detected and added to their ”natural” cluster. Indeed, the similarity threshold value of 0.6 was a good indicator of pairwise similarity when the non-canonical edges of the motif contributed to more than 3/5 of the non-canonical sizes of the two input graphs. Most structural elements (i.e. clustered at step 1) filled this criterion. Those that did not, like the sarcin-ricin element (see structural element 3 in Figure 2), had a pairwise similarity value with each member of their expected cluster below the threshold because the number of the non-canonical edges of the motif in these structural elements contributed to less than 3/5 of their non-canonical size.

We thus clustered eight additional structural elements including the sarcin element S3 (see Appendix).

3 Results and Discussion

We validated the identified motifs in two ways:

- by verifying that the known RNA motifs (C-loops, K-turns, Sarcin-Ricins, E-loops) were correctly clustered;
- by calculating the RMSD between all members within a cluster.

To compare our results with previous findings [18, 22], we used the same ribosomal crystal structures: *H. marismortui* 50S (pdb 1s72), *E.coli* 50S (pdb 2aw4) and *T. thermophilus* 16S (pdb 1j5e).

3.1 The catalogue

The database is available at `http://www.lri.fr/∼md/RNA/CATALOGUE/catalogue.htm`. We listed all secondary structural elements for each chain in each structure. We gave the following data for each structural element:

1. an identifier: a sequential number corresponding to its rank in the tree representation,

2. the set of its non-canonical labels. These are codes used for the names of the interactions between nucleotides. The correspondance between the codes and the names of the interactions are summarised in a table on the home page of the url cited above,

3. a descriptor: the detailed list of its nucleotides and all interactions between them,

4. a 2D view of its corresponding graph produced with Graphviz (http://www.graphviz.org/). This layout is unclear for some structural elements; in these cases, it might be helpful to refer back to the descriptor. The colours used are black for backbone, red for WC basepairs and blue for non-canonical interactions,

5. a 3D view: a pdb file that isolates the structural element in the molecule.

3.2 Clustering

The clustering results are given for *H.m* 23S, *E.coli* 23S and *T.th* 16S (Figure 3 and Table 1). No clusters were formed in the 5S chain of either *H.m* or *E.coli*. Figure 3 shows the 2D diagram of the consensus structure of each motif found (ie. a structure observed in more than half the number of occurrences). For each motif, Table 1 lists the molecule it was observed in, the number of occurrences found and the reference of any corresponding known motif. Occurrences of modified known motifs that were not clustered with their expected families are mentioned in the last column of the table. Further details for each motif are given in the Appendix.

Known motifs

C-loop (Family C)

Two of three occurrences of the C-loop motif (C-96 and C-50) were clustered into family (C) for *H.m* 23S and *E.coli* 23S. The C-38 C-loop motif was not clustered into this family because the completion of its largest common non-canonical subgraph was not isomorphic to the completion of the same non-canonical subgraph in the reference C-96 motif. Moreover, the U2721-A2761 pairing in C-96 is canonical whereas its mapped basepair C963-A1005 in C-38 is a non-canonical *cis* WC/WC.

K-turn (Family K)

This motif was observed in *H.m* 23S and *T.th* 16S. In *H.m* 23S, KT-7 and KT-38 were grouped together in cluster (K). The *trans* Sugar-edge/Sugar-edge base-pairing in KT-46 and KT-58 (id 99 and 123) were not included in the annotation program output; therefore, they were not considered similar to the reference KT-7 occurrence and were clustered into family (T). KT-15 did not match the definition of a motif embedded within a secondary structural element. Indeed, a canonical pairing, A248-U265, ”cuts” the internal loop into two bulges (id 23 and 24). In the latter, the reported *cis* Sugar-edge/Sugar-edge basepair G249-U265 was not output by Rnaview. Finally, in KT-42 (internal loop 89) two non-canonical basepairs forming the non-canonical core of a typical K-turn were not output by Rnaview, and thus this structural element was not considered similar to a typical K-turn. Composite K-turns do not correspond to any secondary structural element and thus were not identified by our method.

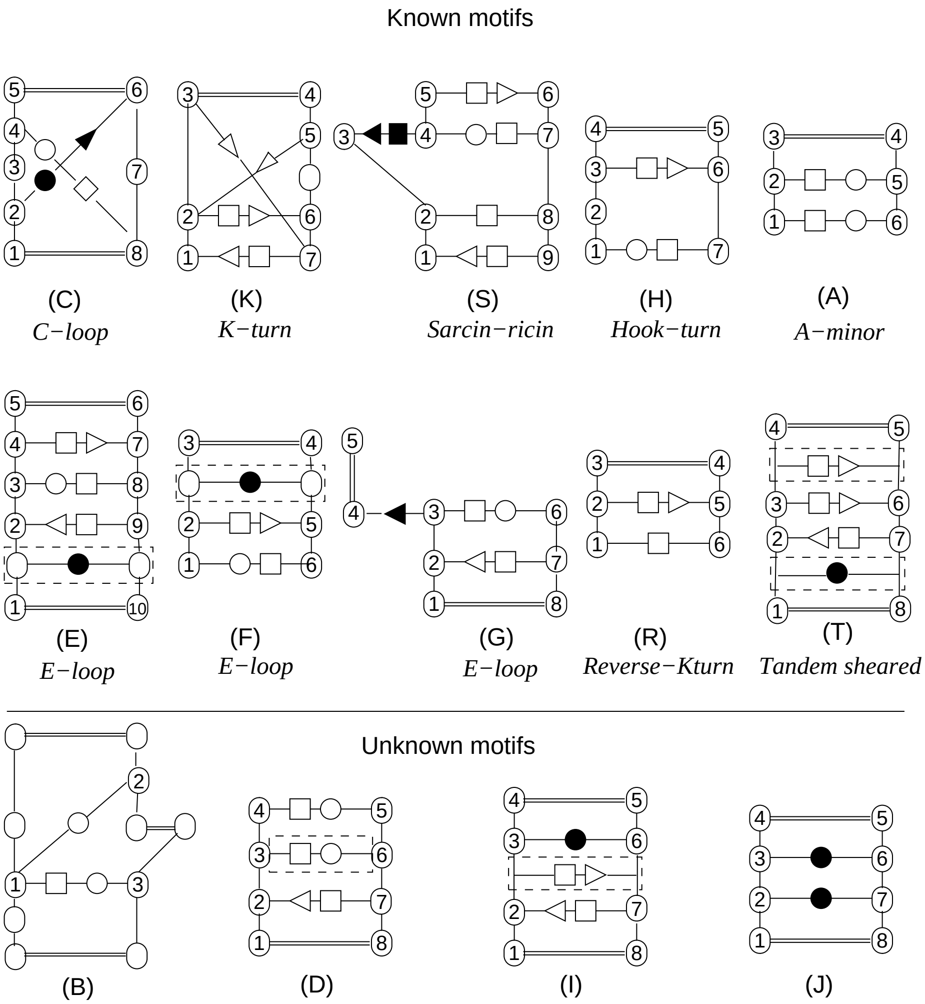


Figure 3: Recurrent motifs found in ribosomal structures. For further details on each motif, see Appendix.

Motifs	Molecule	PDB file	Occur.	Known/Unknown
(C)	*H.m*23S	1s72	2	C-loop [22]
	*E.coli*23S	2aw4	2	C-loop [22]
(K)	*H.m*23S	1s72	2	Kturns KT-7, KT-38 [22]
(S)	*H.m*23S	1s72	6	Sarcin-ricin [18]
	*E.coli*23S	2aw4	5	Sarcin-ricin [18]
	*T.th*16S	1j5e	2	Sarcin-ricin [18]
(H)	*H.m*23S	1s72	5	Hook-turn [33]
	*E.coli*23S	2aw4	6	Hook-turn [33]
(A)	*H.m*23S	1s72	3	A-minor [23]
(E)	*H.m*23S	1s72	3	23S E-loop [18]
	*T.th*16S	1j5e	4	23S E-loop [18]
(F)	*E.coli*23S	2aw4	5	23S E-loop comprising sarcin G2664 [18]
	*H.m*23S	1s72	5	23S E-loop comprising composite sarcin G911 [18]
(G)	*E.coli*23S	2aw4	2	23S E-loop [18]
(R)	*H.m*23S	1s72	7	Reverse-Kturn [17]
	*E.coli*23S	2aw4	6	Reverse-Kturn [17]
(T)	*E.coli*23S	2aw4	8	Tandem sheared
	*H.m*23S	1s72	6	Tandem sheared comprising KT-46, KT-58 [22]
	*T.th*16S	1j5e	2	Tandem sheared
(B)	*H.m*23S	1s72	2	Unknown
(D)	*E.coli*23S	2aw4	2	Unknown
(I)	*T.th*16S	1j5e	2	Unknown
(J)	*T.th*16S	1j5e	2	Unknown

Table 1: List of the clusters formed in *H.m 23S*, *E.coli 23S* and *T.th 16S*.

In *T.th* 16S, neither known ocurrence, KT-11 or KT-23, were similar according to our similarity measure (Figure 1) and hence did not form a cluster.

Sarcin-ricin (Family S)

In *T. th* 16S, both known occurrences of the sarcin-ricin motif were clustered into family (S). Six known local occurrences of this motif observed in *H.m* 23S, were also clustered into this family. One composite occurrence, Helix36 Junction G911, was not recognised as a sarcin-ricin motif. The *trans* Hoogsteen/Hoogsteen basepair A913-G1071, which is part of the non-canonical core of a typical sarcin was not output by Rnaview. Additionally, the discontinued backbone between residues G1071 and G1292 prevented mapping the completions of the subgraphs corresponding to the non-canonical core. This F72 occurrence was clustered with two other occurrences of sarcin-like motifs, F76 and F30, into the 23S-Eloop family (F).
Five of six occurrences observed in *E. coli* 23S were clustered together in family (S). G2664 was not recognised as a sarcin motif because A2654-C2666 was output by Rnaview as a *trans* Hoogsteen/WC and not a *trans* Hoogsteen/Hoogsteen, as in the sarcin core. This F199 occurrence was clustered with E-loop family (F).

E-loop (Families E, F, G)

The bacterial E-loop motif consists of two isosteric submotifs related by 180° rotation [18],:
- *trans* Hoogsteen/Sugar-edge,
- *trans* WC/Hoogsteen or *trans* Sugar-edge/Hoogsteen,
- *cis* bifurcated or *trans* Sugar-edge/Hoogsteen.

Some examples of 23S rRNA E-loops were also shown (see fig. 15 of [18]). Family (E) is similar to a 23S rRNA E-loop variant, which has a *trans* WC/Hoogsteen rather than a *trans* Sugar-edge/Hoogsteen at the second basepair of the submotif. E22 and E35 motifs(see Appendix), together with families (F) and (G), despite lacking one sheared basepair, still qualify as another variant of the 23S E-loop (see fig. 15 of [18]). Sarcin-like motifs F72, F76 and F30 may also be classified as bulged-G motifs [5] .

Hook-turn (Family H)

The H161 motif of family (H) was identified as a hook-turn (see fig.5 of [33]). In addition to the significant number of occurrences observed in both *H.m* 23S and *E.coli* 23S, this family is conspicuous in that the sequence signature of the non-canonical core is strikingly conserved (see Appendix). Furthermore, all occurrences of this motif seem to occur at corresponding positions in both structures.

A-minor (Family A)

A close examination of the three family (A) occurrences revealed that A60 is an A-minor motif, similar to that previously reported in [23] .
This motif is termed A-minor because it involves the insertion of the smooth minor groove edges of adenine residues into the minor groove of neighbouring helices, preferentially at C-G basepairs. This motif plays an important role in stabilising the tertiary structure of RNA [27].

Reverse-Kturn (Family R)

Family (R) was identified as a reverse-Kturn (see fig. 2 of [17]). Of note, R175 did not superimpose well with other occurrences of this motif (RMSD $>$ 4 Å).

Tandem sheared (Familiy T)
Family (T) is the well known tandem sheared GA motif. Three occurrences of this motif, T53, T131 and T3, in *E. coli* 23S and two, T65 and T1, in *H.m* 23S may also be 23S E-loops. The clustering of these occurrences with tandem sheared motifs is not inconsistent since both families share a common non-canonical core.

Putative new motifs

These clusters (B, D, I, J) do not contain, as far as we know, known motifs. B170 was identified as a three-way junction belonging to family B (see fig. 7 of [24]).

4 Conclusion

The present work describes the first automated method for cataloguing all secondary structural elements of an RNA molecule and extracting similar occurrences of structural motifs on the basis of a graph of the tertiary structure. Using an innovative graph similarity measure, we identified numerous occurrences of structural motifs despite the presence of base and basepair insertions in some of these motifs. Such information regarding variation in basepairing and position of insertions and deletions will allow the analysis and prediction of the 3D structure of RNA motifs based on sequence signature in homologous RNA molecules and the structure-based alignment of homologous sequences.

Our method relies on the LECNS algorithm, which identifies the largest common non-canonical subgraph of any two graphs, and hence determines the non-canonical core of an RNA motif. The results showed that this algorithm successfully detects theoretical structural similarities within the graph model of the tertiary structure. However, the detection of composite occurrences made of discontinuous strands is still limited even at this high level of representation. A large proportion of the motifs found correspond to known structural motifs. Further expert examination of the putative new motifs will be required to confirm whether they represent real structural motifs.

With an expected increase in the number of available crystal structures, such an automated method which accelerates the identification and classification of recurrent RNA motifs will be useful in assessing their abundance in an RNA structure or in compiled databases such as the RNAJunction database [4]. We believe this will advance our understanding of the mechanism by which these motifs mediate the folding process of RNA and perform their biological roles in the cell.

Supplementary material

Dendrograms of hiearchical clustering of *T. th* 16s, *H. m* 23S and *E. coli* 23S are available at `http://www.lri.fr/∼md/RNA/Dendrograms/dendrogram.htm`.

Acknowledgements

We thank Eric Westhof for helpful discussions and his critical review of the manuscript. We also thank Dominique Barth, Franck Quessette and Sandrine Vial for fruitful discussions. We are grateful to Y. Ponty for providing *secrna*, and to J. Allali, R. Rivière and F. Lemoine for helping with implementation details.

MD acknowledges grant support from the Direction for International Affairs of the University Paris-Sud 11.

References

[1] P.J. Artymiuk, R.V. Spriggs, and P. Willett. Graph theoretic methods for the analysis of structural relationships in biological macromolecules: Research articles. *J. Am. Soc. Inf. Sci. Technol.*, 56(5):518–528, 2005.

[2] R.T. Batey, R.P. Rambo, and J.A. Doudna. Tertiary motifs in RNA structure and folding. *Angew. Chem. Int. Ed.*, 32:2326–2343, 1999.

[3] H.M. Berman, W.K. Olson, D.L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S.-H. Hsieh, A. R. Srinivasan, and B. Schneider. The Nucleic Acid Database: A Comprehensive Relational Database of Three-Dimensional Structures of Nucleic Acids. *Biophys.J.*, 63:751–759, 1992.

[4] E. Bindewald, R. Hayes, Y.G. Yingling, W. Kasprzak, and B.A. Shapiro. RNAJunction: a database of RNA junctions and kissing loops for three-dimensional structural analysis and nonodesign. *Nucleic Acids Research*, 36:392–397, 2008.

[5] C.C. Corell, J. Beneken, M.J. Plantinga, M. Lubbers, and Y-L. Chan. The common and the distinctive features of the bulged-G motif based on a 1.04 Åresolution RNA structure. *Nucleic Acids Research*, 31:6806–6818, 2003.

[6] W.L. DeLano. The pymol molecular graphics system. *DeLano Scientific, Palo Alto, CA, USA. http://www.pymol.org*, 2002.

[7] C.M. Duarte, L.M. Wadley, and A.M. Pyle. RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space. *Nucleic Acids Research*, 31(16):4755–4761, 2003.

[8] A. Harrison, D.R. South, P. Willett, and P.J. Artymiuk. Representation, searching and discovery of patterns of bases in complex RNA structures. *Journal of Computer-Aided Molecular Design*, 17(8):537–549, 2003.

[9] D.K. Hendrix, S.E. Brenner, and S.R. Holbrook. RNA structural motifs : building blocks of a modular molecule . *Quarterly Reviews of Biophysics*, 38:221–243, 2005.

[10] E. Hershkovitz, E. Tannenbaum, S.B. Howerton, A. Sheth, A. Tannenbaum, and L.D. Williams. Automated identification of RNA conformational motifs: theory and application to the HM LSU 23S rRNA. *Nucleic Acids Research*, 31:6249–6257, 2003.

[11] S.R. Holbrook. RNA structure: the long and the short of it . *Current Opinion in Structural Biology*, 15:302–308, 2005.

[12] H.C. Huang, U. Nagaswamy, and G.E. Fox. The application of cluster analysis in the intercomparison of loop structures in RNA. *RNA*, 11(4):412–423, 2005.

[13] X. Huang, J. Lai, and S. F. Jennings. Maximum common subgraph: some upper bound and lower bound results. *BMC Bioinformatics*, 7(Suppl 4):S6, 2006.

[14] F. Jossinet and E. Westhof. Sequence to Structure (S2S): display, manipulate and interconnect RNA data from sequence to structure. *Bioinformatics*, 21(15):3320–3321, 2005.

[15] Si Quang Le, Tu Bao Ho, and T.T Hang Phan. A novel graph-based similarity measure for 2D chemical structures. *Genome Informatics*, 15(2):82–91, 2004.

[16] S. Lemieux and F. Major. RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire. *Nucleic Acids Research*, 30(19):4250–4263, 2002.

[17] N.B. Leontis, A. Lescoute, and E. Westhof. The building blocks and motifs of RNA architecture. *Current Opinion in Structural Biology*, 16:1–9, 2006.

[18] N.B. Leontis, J. Stombaugh, and E. Westhof. Motif prediction in ribosomal RNAs - lessons and prospects for automated motif prediction in homologous RNA molecules. *Biochimie*, 84:961–973, 2002.

[19] N.B. Leontis, J. Stombaugh, and E. Westhof. Survey and summary. The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Research*, 30:3497–3531, 2002.

[20] N.B. Leontis and E. Westhof. Geometric nomenclature and classification of RNA base pairs. *RNA*, 7:499–512, 2001.

[21] N.B. Leontis and E. Westhof. Analysis of RNA motifs. *Current Opinion in Structural Biology*, 13:300–308, 2003.

[22] A. Lescoute, N.B Leontis, C. Massire, and E. Westhof. Recurrent structural RNA motifs, Isostericity matrices and sequence alignments. *Nucleic Acids Research*, 33:2395–2409, 2005.

[23] A. Lescoute and E. Westhof. The A-minor motifs in the decoding recognition process. *Biochimie*, 88:993–999, 2006.

[24] A. Lescoute and E. Westhof. Topology of three-way junctions in folded RNAs. *RNA*, 12:83–93, 2006.

[25] P.B. Moore. Structural motifs in RNA . *Annu. Rev. Biochem.*, 68:287–300, 1999.

[26] M.Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46:591–621, 1984.

[27] P. Nissen, J.A. Ippolito, N. Ban, P.B. Moore, and T.A. Steitz. RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *PNAS*, 98:4899–4903, 2001.

[28] D. Oranit, R. Nussinov, and H. Wolfson. ARTS: alignment of RNA tertiary structures. *Bioinformatics*, 21(suppl 2):ii47–53, 2005.

[29] Y. Ponty. *Modélisation de séquences génomiques structurées, génération aléatoire et application*. PhD thesis, Université Paris-Sud 11, November 2006.

[30] M. Sarver, C.L. Zirbel, Jesse Stombaugh, Ali Mokdad, and Neocles B. Leontis. FR3D: Finding local and composite recurrent structural motifs in RNA 3D structures. *Journal of Mathematical Biology*, 56(1-2):215–252, 2008.

[31] B.A. Shapiro. An algorithm for comparing multiple RNA secondary structures. *Computer Applications in the Biosciences*, 4(3):387–393, 1988.

[32] S. Smit, K. Rother, J. Heringa, and R. Knight. From knotted to nested RNA structures: A variety of computational methods for pseudoknot removal. *RNA*, 14(3):410–416, 2008.

[33] S. Szep, J. Wang, and P.B. Moore. The crystal structure of a 26-nucleotide RNA containing a hook-turn. *RNA*, 9:44–51, 2003.

[34] G. Valiente. *Algorithms on Trees and Graphs*. Springer-Verlag, Berlin, 2002.

[35] L.M. Wadley and A.M. Pyle. The identification of novel RNA structural motifs using COMPADRES: an automated approach to structural discovery. *Nucleic Acids Research*, 32:6650–6659, 2004.

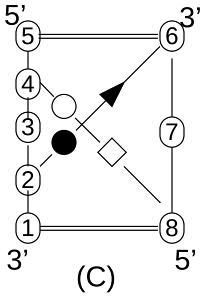
[36] E. Westhof and P. Auffinger. RNA Tertiary structure. *Encyclopedia of Analytical Chemistry. R.A. Meyers (Ed)*, pages 5222–5232, 2000.

[37] H. Yang, F. Jossinet, N. Leontis, L. Chen, J. Westbrook, H.M. Berman, and E. Westhof. Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Research*, 31:3450–3460, 2003.

Appendix: Clusters of structural elements.

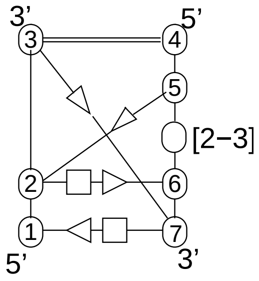
Further details on the motifs described in Figure 3 and Table 1 are provided. For each occurrence of a motif, we give:

- the 2D diagram of its consensus structure. The variable length of the single strands is indicated between square brackets. Observed basepair insertions are framed in dashed boxes;
- its sequence. Indicated in **bold** are conserved bases, but not their corresponding pairing, as represented in the consensus structure;
- the catalogued secondary structural element in which the occurrence was observed,
- the RMSD value calculated with Pymol [6] by aligning the non-canonial core with that of a reference occurrence. For known motifs, the (known) reference occurrence was used; otherwise, the reference occurrence was chosen to minimise the sum of the pairwise RMSDs,
- if the occurrence corresponds to a previously reported motif, its name is also given (between brackets).



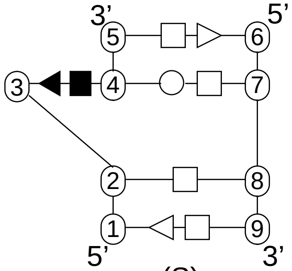
(C)

PDB	Inst.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	Catalogue	RMSD (ref. C197)
1s72	C197	2721_U	2720_C	2719_A	2718_C	2717_C	2763_G	2762_C	2761_A	Internal L. (197)	0.00 (C-96)
	C108	1429_U	1428_C	1427_A	1426_C	1425_G	1439_C	1438_G	1437_A	Internal L. (108)	0.54 (C-50)
2aw4	C98	1323_C	1322_A	1321_A	1320_C	1319_C	1333_G	1332_G	1331_G	Internal L. (98)	0.68
	C201	2684_U	2683_C	2682_A	2681_C	2680_U	2727_A	2726_A	2725_A	Internal L. (201)	0.62



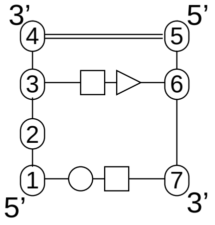
(K)

PDB	Inst.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	Catalogue	RMSD (ref. K10)
1s72	K10	79_G	80_A	81_G	93_C	94_G	97_G	98_A	Internal L.(10)	0.00 (KT-7)
	K74	938_G	939_A	940_G	1026_C	1027_G	1031_G	1032_A	Internal L.(74)	0.85 (KT-38)



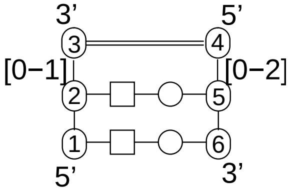
(S)

PDB	Inst.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	Catalogue	RMSD (ref.S195)
1s72	S195	2690_U	2691_A	2692_G	2693_U	2694_A	2701_G	2702_A	2703_A	2704_C	Internal L.(195)	0.00 (G2701)
	S101	1368_U	1369_A	1370_G	1371_U	1372_A	2053_G	2054_A	2055_A	2056_C	Internal L.(101)	0.27 (G2053)
	S19	211_U	212_A	213_G	214_U	215_A	225_G	226_A	227_A	228_C	Internal L.(19)	0.43 (G225)
	S16	173_C	174_A	175_G	176_U	177_A	159_G	160_A	161_A	162_C	Internal L.(16)	0.45 (G159)
	S21		380_A	381_G	382_U	383_A	406_G	407_A	408_A		Junction L.(21)	0.44 (G406)
	S3		463_A	464_G	465_U	466_A	475_G	476_A	477_A		Junction L.(3)	0.55 (G475)
2aw4	S91	**1264_A**	1265_A	1266_G	1267_U	1268_A	2012_G	2013_A	2014_A	**2015_A**	Internal L.(91)	0.41 (G2012)
	S21	240_C	241_A	**242_G**	**243_U**	244_A	254_G	255_A	256_A	257_C	Internal L.(21)	0.43 (G254)
	S18	203_A	204_A	205_G	206_U	207_A	189_G	190_A	191_A	192_C	Internal L.(18)	0.56 (G189)
	S23		371_A	372_G	373_U	374_A	400_G	401_A	402_A		Junction L.(23)	0.58 (G400)
	S5		457_A	458_G	459_U	460_A	469_G	470_A	471_A		Junction L.(5)	0.57 (G469)
1j5e	S63	888_G	889_A	890_G	891_U	892_A	906_G	907_A	908_A	909_A	Internal L.(63)	0.34 (G906)
	S68	**1345_U**	1346_A	1347_G	1348_U	1349_A	1373_G	1374_A	1375_A	**1376_U**	Junction L.(68)	0.55 (G1373)



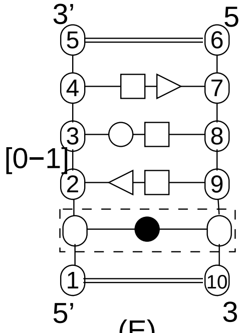
(H)

PDB	Inst.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	Catalogue	RMSD (ref.H83)
1s72	H83	1096_U	1097_A	1098_A	1099_G	1257_C	1258_G	1259_A	Internal L.(83)	0.00
	H111	1457_U	1458_A	1459_A	1460_G	1483_C	1484_G	1485_A	Internal L.(111)	0.76
	H201	2774_U	2775_A	2776_A	2777_G	2797_C	2798_G	2799_A	Internal L.(201)	0.33
	H161	2242_U	2243_C	2244_A	2245_C	2256_G	2257_G	2258_A	Junction L.(161)	1.61
	H193	2673_U	2674_G	2675_A	2676_C	2809_G	2810_G	2811_A	Internal L.(193)	1.61
2aw4	H73	999_U	1000_A	1001_A	1002_G	1153_C	1154_G	1155_A	Internal L.(73)	0.42
	H101	1352_U	1353_A	1354_A	1355_G	1376_C	1377_G	1378_A	Internal L.(101)	0.69
	H106	1578_U	1579_A	1580_A	1581_G	1417_C	1418_G	1419_A	Internal L.(106)	0.31
	H205	2739_U	2740_A	2741_A	2742_G	2762_C	2763_G	2764_A	Internal L.(205)	0.52
	H161bis	2197_U	2198_A	2199_A	2200_C	2223_G	2224_G	2225_A	Junction L. (161)	1.60
	H196	2637_U	2638_G	2639_A	2640_G	2774_C	2775_G	2776_A	Internal L.(196)	1.66



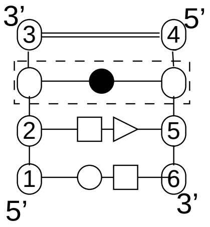
(A)

PDB	Inst.	(1)	(2)	(3)	(4)	(5)	(6)	Catalogue	RMSD (ref. A202)
1s72	A60	766_A	767_A	769_C	892_G	895_A	896_C	Internal L.(60)	0.74
	A168	2429_A	2430_A	2432_C	2459_G	2460_A	2461_U	Junction L.(168)	1.09
	A202	2783_A	2784_A			2788_A	2789_U	Terminal L. (202)	0.00



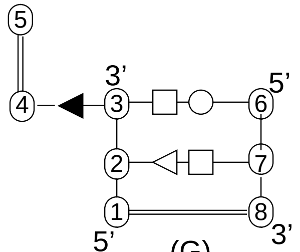
(E)

PDB	Inst.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	Catalogue	RMSD (ref. E43)
1s72	E55	705_C	706_G	707_C	708_A	709_G	719_C	720_G	721_A	722_G	723_G	Internal L.(55)	0.62
	E117	1542_G	1543_G	1544_U	1545_C	1546_G	1639_U	1640_C	1641_A	1642_A	1643_C	Internal L.(117)	0.78
	E22	267_G	269_G	270_U					241_A	242_A	244_C	Junction L.(22)	–
1j5e	E43	580_U	581_G	582_U	583_A	584_G	757_U	758_G	759_A	760_G	761_G	Internal L.(43)	0.00
	E58	799_G	800_G	801_U	802_A	803_G	779_C	780_A	781_A	782_A	783_C	Internal L.(58)	0.53
	E35	483_C	**484_G**	486_U	487_A	488_C	446_G	447_G	448_A		450_G	Internal L.(35)	–
	E54	683_G	685_G	686_U	687_A			703_G	704_A	705_G	707_C	Internal L.(54)	1.69



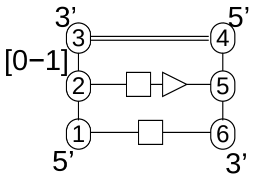
(F)

PDB	Inst.	(1)	(2)	(3)	(4)	(5)	(6)	Catalogue	RMSD (ref. F199)
2aw4	F145	1865_U	1866_A	1867_G	1874_C	1875_G	1876_A	Internal L.(145)	0.92
	F65	860_U	861_A	862_G	915_C	916_G	917_A	Internal L.(65)	0.32
	F130	1716_U	1717_A	1718_G	1742_C	1743_G	1744_A	Internal L.(130)	0.54
	F199	2656_U	2657_A	2658_C	2663_G	2664_G	2665_A	Internal L.(199)	0.00 (sarcin G2664)
	F59	1188_U	1189_A	1190_G	817_C	818_G	819_A	Junction L.(59)	0.35
1s72	F46	589_U	590_A	591_A	567_U	568_G	569_A	Internal L.(46)	0.30
	F76	954_U	955_A	956_G	1011_C	1012_A	1013_A	Internal L.(76)	0.43
	F30	359_U	360_A	362_G	290_C	292_G	293_A	Internal L.(30)	0.37
	F72	1293_U	1294_A	1295_G	910_C	911_G	912_A	Junction L.(72)	0.33 (composite sarcin G911)
	F143	1972_U	1973_A	1974_G	2008_U	2009_G	2010_A	Junction L.(143)	0.54



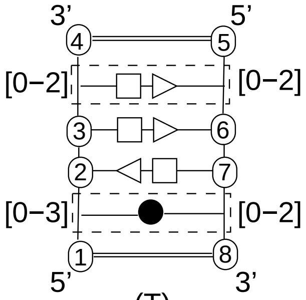
(G)

PDB	Inst.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	Catalogue	RMSD (ref. G27)
2aw4	G27	297_G	298_G	299_A	319_G	323_C	339_U	340_A	341_C	Junction L.(27)	0.00
	G140	1967_C	1968_G	1969_A	1972_G	1833_C	1931_U	1932_A	1933_G	Junction L.(140)	1.32



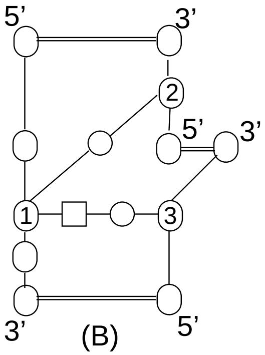
(R)

PDB	Inst.	(1)	(2)	(3)	(4)	(5)	(6)	Catalogue	RMSD (ref. R138)
1s72	R87	1132_A	1133_A	1134_G	1228_C	1229_C	1230_A	Internal L.(87)	0.45
	R115	1527_A	1528_A	1529_G	1662_C	1663_G	1664_A	Internal L.(115)	0.60
	R116	1658_A	1659_A	1660_G	1531_U	1532_G	1533_A	Junctionl L.(116)	0.37
	R175	2397_G	2398_A	2399_G	2389_U	2390_U	2391_C	Terminal L.(175)	4.45
	R167	2307_A	2308_U	2310_G	2298_C	2299_G	2300_A	Terminal L.(167)	0.49
	R120	1572_A	1573_A	1574_C	1622_G	1623_C	1624_A	Internal L.(120)	0.44
	R134	1767_A	1768_C	1769_C	1774_G	1775_A	1776_A	Internal L.(134)	0.63
2aw4	R113	1469_A	1470_A	1471_G	1520_U	1521_G	1522_A	Internal L.(113)	0.37
	R128	1689_A	1690_A	1691_C	1696_G	1697_G	1698_A	Internal L.(128)	0.76
	R59	820_A	821_A	946_C	971_G	972_A	973_A	Junction L.(59)	0.69
	R77	1028_A	1029_A	1030_C	1124_G	1125_G	1126_A	Internal L.(77)	0.53
	R138	1802_A	1803_A	1804_C	1813_G	1814_G	1815_A	Internal L.(138)	0.00
	R107	1571_A	1572_A	1574_C	1424_G	1426_G	1427_A	Junction L.(107)	0.60

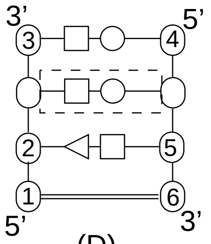


(T)

PDB	Inst.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	Catalogue	RMSD (ref. T3)
2aw4	T53	703_U	704_G	705_A	707_G	724_U	726_A	727_A	728_G	Internal L.(53)	0.73
	T131	1720_U	1721_G	1722_A	1724_G	1736_U	1738_G	1739_A	1740_G	Internal L.(131)	0.52
	T3	510_C	512_G	513_A	516_C	24_G	27_G	28_A	30_G	Internal L.(3)	0.00
	T37	536_G	537_G	538_A	539_G	554_U	555_G	556_A	557_C	Internal L.(37)	0.80
	T176	2350_C	2351_G	2352_A	2353_G	2364_C	2365_G	2366_A	2367_G	Internal L.(176)	0.87
	T185	2466_C	2468_A	2469_A	2470_G	2480_C	2481_G	2482_A	2484_G	Internal L.(185)	0.52
	T89	1208_C	1212_G	1213_A	1215_G	1234_U	1236_G	1237_A	1238_G	Internal L.(89)	1.16
	T144		1857_G	1858_A	1860_G	1882_U	1884_G	1885_A		Internal L.(144)	0.58
1s72	T65	794_U	795_G	796_A	798_G	815_U	817_G	818_A	819_A	Internal L.(65)	1.01
	T99	1312_G	1316_G	1317_A	1319_G	1338_U	1340_G	1341_A	1342_C	Internal L.(99)	0.98 (KT-46)
	T123	1602_C	1605_G	1606_A	1608_G	1587_U	1589_G	1590_A	1592_G	Internal L.(123)	0.60 (KT-58)
	T1	516_A	518_G	519_A	522_U	21_G	24_G	25_A	27_U	Internal L.(1)	0.62
	T208	2873_C	2874_G	2875_A	2876_G	2881_C	2882_G	2883_A	1884_G	Internal L.(208)	1.19
	T182	2501_G	2503_A	2504_A	2505_G	2515_C	2516_G	2517_A	2519_C	Internal L.(182)	0.59
1j5e	T107	1416_G	1417_G	1418_A	1419_G	1481_U	1482_G	1483_A	1484_C	Internal L.(107)	0.74
	T108	1431_C	1432_G	1433_A	1435_G	1466_C	1467_G	1468_A	1469_G	Internal L.(108)	1.26

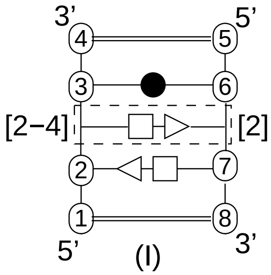


PDB	Inst.	(1)	(2)	(3)	Catalogue	RMSD (ref.B105)
1s72	B170	2369_A	2356_A	2330_U	Junction L.(170)	0.89
	B105	1682_A	1414_A	1696_U	Junction L.(105)	0.00



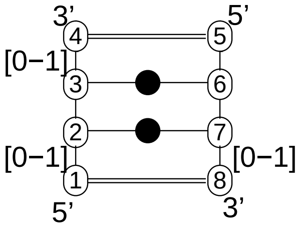
(D)

PDB	Inst.	(1)	(2)	(3)	(4)	(5)	(6)	Catalogue	RMSD (ref. D46)
2aw4	D43	618_G	619_G	621_A	607_U	609_A	610_C	Internal L.(43)	0.00
	D16	158_U	159_G	161_A	165_A	167_A	168_G	Terminal L.(16)	1.57



(I)

PDB	Inst.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	Catalogue	RMSD (ref. I47)
1j5e	I99	1303_C	1304_G	1307_U	1308_U	1329_A	1330_U	1333_A	1334_G	Internal L.(99)	1.85
	I47	605_U	606_G	611_A	612_C	628_G	629_G	632_A	633_G	Internal L.(47)	0.00



(J)

PDB	Inst.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	Catalogue	RMSD (ref. J51)
1j5e	J51	662_G	663_A	664_G	666_G	740_U	741_G	742_G	743_U	Internal L.(51)	0.00
	J53	673_G	675_A	676_A	677_U	713_G	714_G	715_A	717_C	Internal L.(53)	0.79