

A method for automated discovering of RNA tertiary motifs

Mahassine Djelloul¹ and Alain Denise^{1,2}

¹ LRI, Université Paris-Sud 11 and CNRS,

² IGM, Université Paris-Sud 11 and CNRS

{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 [14,20]. 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). Further studies used these representations to develop graph-theoretical representations (see [8] for a review). In 2001, a descriptive base-pairing nomenclature was proposed by Leontis and Westhof (LW) to systematically annotate and classify non-WC basepairs [10,7]. In a LW nomenclature-based representation, the tertiary structure is viewed as a 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: (a) Given a description of a *known* motif, identify this motif in target structures, or (b) 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 [18], or (ii) finding similar occurrences of the pattern. Practically, this consists of identifying a maximum common subgraph of two input graphs. However, the maximum common subgraph (MCS) problem is NP-hard, APX-hard and W[1]-hard [5] and such an approach is not feasible except for very small graphs such as those in chemoinformatics [6]. 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 [19] used a previous work on *RNA worms* [3] 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 [4,15]. 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 [12].

In this paper, we propose a new method for identifying and classifying similar occurrences of *a priori* unknown RNA motifs using the 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” [11]. Two non-canonical basepairs are said *isosteric* if they belong to the same geometric family and can substitute each other without distorting the fundamental 3D structure of the motif.

2 Materials and Methods

We downloaded crystal structures from the NDB database [1]. We used the annotation program Rnview [21] to produce the corresponding 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. We considered 14 types of interactions: the phosphodiester bond, the canonical WC pairing GC and AU (to which the wobble pairing GU is added), and the 12 non-WC basepairs defined in the Leontis and Westhof (LW) nomenclature [10]. 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. Identifying secondary structural elements

Using a classical tree representation of the secondary structure [16], 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.

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 *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 [18]. 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:

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

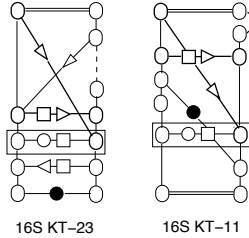


Figure 1. Two structural elements with their LECNS (in bold) of size 2. There is a larger common non-canonical subgraph (size 3) comprising the framed basepair, but it is not extensible. Dashed backbone indicates free nucleotides.

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.

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. The resulting dendrogram is presented in Figure 2. A threshold value was needed to obtain distinct clusters from the tree. This

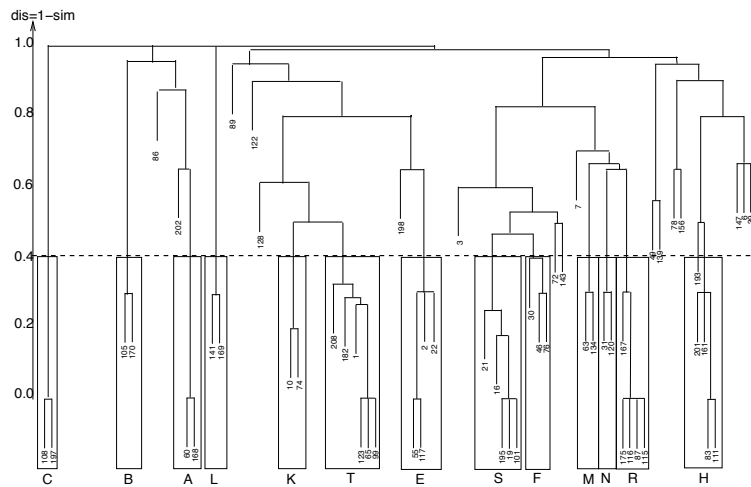


Figure 2. Dendrogram of hierarchical clustering of *H.m* 23S RNA produced with *hclust* (R Project for statistical computing (<http://www.r-project.org/>)). The structural elements are numbered from 1 to 209. Rectangular boxes correspond to clusters obtained using the 0.6 similarity threshold. Structural elements clustered with a null similarity value are not shown.

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 [9,12]. The value giving optimal clustering of these motifs was 0.6 (Figure 2). Of note, 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, thus they 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.

3 Results and Discussion

The catalogue 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 validated the identified motifs in two ways: (i) by verifying that the known RNA motifs (C-loops, K-turns, Sarcin-Ricins, E-loops) were correctly clustered; (ii) by calculating the RMSD between all members within a cluster. To compare our results with previous findings [9,12], we used the same ribosomal crystal structures: *H. marismortui* 50S (pdb 1s72), *E.coli* 50S (pdb 2aw4) and *T. thermophilus* 16S (pdb 1j5e).

The clustering results are given for *H.m* 23S, *E.coli* 23S and *T.th* 16S (Figure 3). 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). Some examples of the motifs found are given below. The complete list can be found in [2].

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.

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 Rnview. 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 Rnview 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).

Hook-turn (Family H). The H161 motif of family (H) was identified as a hook-turn (see fig.5 of [17]). 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

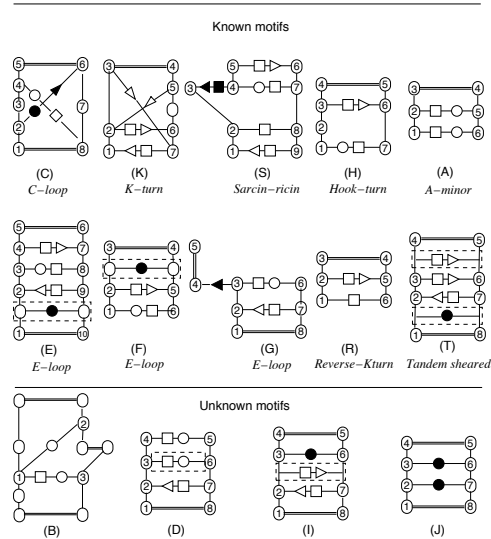


Figure 3. Recurrent motifs found in ribosomal structures.

conserved. Furthermore, all occurrences of this motif seem to occur at corresponding positions in both structures.

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 [13]).

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 base-pairing 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. 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.

Acknowledgements

We thank E. Westhof for helpful discussions and his critical review of the manuscript, Y. Ponty for providing *secrna*, and J. Allali, R. Rivière and F. Lemoine for helping with implementation details. This research was partially supported by the DIGITEO PASAPAS project, and by a grant support to MD from the Direction for International Affairs of the University Paris-Sud 11.

References

- [1] 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.
- [2] M. Djelloul and A. Denise. Automated motif discovering in RNA molecules. Technical Report 1490, LRI, Université Paris-Sud 11, 2008.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] N.B. Leontis, A. Lescoute, and E. Westhof. The building blocks and motifs of RNA architecture. *Curr. Opin. in Struct. Biol.*, 16:1–9, 2006.
- [9] 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.
- [10] 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.
- [11] N.B. Leontis and E. Westhof. Analysis of RNA motifs. *Curr. Opin. in Struct. Biol.*, 13:300–308, 2003.
- [12] 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.
- [13] A. Lescoute and E. Westhof. Topology of three-way junctions in folded RNAs. *RNA*, 12:83–93, 2006.
- [14] P.B. Moore. Structural motifs in RNA. *Annu. Rev. Biochem.*, 68:287–300, 1999.
- [15] D. Oranit, R. Nussinov, and H. Wolfson. ARTS: alignment of RNA tertiary structures. *Bioinformatics*, 21(suppl 2):ii47–53, 2005.
- [16] B.A. Shapiro. An algorithm for comparing multiple RNA secondary structures. *Computer Applications in the Biosciences*, 4(3):387–393, 1988.
- [17] 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.
- [18] G. Valiente. *Algorithms on Trees and Graphs*. Springer-Verlag, Berlin, 2002.
- [19] 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.
- [20] E. Westhof and P. Auffinger. RNA Tertiary structure. *Encyclopedia of Analytical Chemistry*. R.A. Meyers (Ed), pages 5222–5232, 2000.
- [21] 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.