# Annotation of the local context of the RNA secondary structure improves the classification and prediction of A-minors

Anna A. Shalybkova[1], Darya S. Mikhailova[2], Ivan V. Kulakovskiy[3,4,5], Liliia I. Fakhranurova[6,7], Eugene F. Baulin[2,8,*]

[1] Department of Chemistry, Lomonosov Moscow State University, Moscow, 119991, Russia
[2] Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, 141701, Russia
[3] Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, 119991, Russia
[4] Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, 119991, Russia
[5] Institute of Protein Research, Russian Academy of Sciences, Pushchino, 142290, Russia
[6] Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences, Pushchino, Moscow Region, 142290, Russia
[7] Shemiakin and Ovchinnikov Institute of Bioorganic Chemistry of the Russian Academy of Sciences, Moscow, 117997, Russia
[8] Institute of Mathematical Problems of Biology RAS - the Branch of Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Pushchino, Moscow Region, 142290, Russia

* To whom correspondence should be addressed. Tel: +74967318504; Fax: +74967318500;

Email: baulin@lpm.org.ru

The authors wish it to be known that, in their opinion, the first 2 authors should be regarded as joint First Authors

**SHORT TITLE** A novel structural classification of RNA A-minors

**KEYWORDS** A-minor, RNA secondary structure, RNA tertiary motif, across-bulged motif, machine learning

## ABSTRACT

Non-coding RNAs play a crucial role in various cellular processes in living organisms, and RNA functions heavily depend on molecule structures composed of stems, loops, and various tertiary motifs. Among those, the most frequent are A-minor interactions, which are often involved in the formation of more complex motifs such as kink-turns and pseudoknots. We present a novel classification of A-minors in terms of RNA secondary structure where each nucleotide of an A-minor is attributed to the stem or loop, and each pair of nucleotides is attributed to their relative position within the secondary structure. By analyzing classes of A-minors in known RNA structures, we found that the largest classes are mostly homogeneous and preferably localize with known A-minor co-motifs, e.g. tetraloop-tetraloop receptor and coaxial stacking. Detailed analysis of local A-minors within internal loops revealed a novel recurrent RNA tertiary motif, the across-bulged motif. Interestingly, the motif resembles the previously known GAAA/11nt motif but with the local adenines performing the role of the GAAA-tetraloop. By using machine learning, we show that particular classes of local A-minors can be predicted from sequence and secondary structure. The proposed classification is the first step toward automatic annotation of not only A-minors and their co-motifs but various types of RNA tertiary motifs as well.

## INTRODUCTION

The ubiquity and importance of non-coding RNAs in living organisms are now widely accepted (Costa 2010). Their functions include gene expression regulation (Hollands et al. 2012, Breaker 2018), RNA modification (Kiss 2001), intron splicing (Dvinge et al. 2019), and transposon control (Aravin et al. 2007). The spatial structure of non-coding RNAs significantly determines their functions (Montange and Batey 2008). It’s known that the RNA structure has modular organization. It is composed of secondary structure elements (stems and loops) and their tertiary interactions forming the so-called RNA tertiary motifs, broadly defined as the structural “building blocks” that are often recurrent and hold their configuration in different structural environments (Leontis et al. 2006). In contrast to RNA secondary structure elements, the definition of a tertiary motif is not strict and covers diverse modules, e.g. coaxial stacking (8 and more nucleotides forming two stacked stems (Walter et al. 1994)) or dinucleotide platforms (base pairs between consecutive nucleotides (Mladek et al. 2012)).

The A-minor interaction is one of the most abundant types of the RNA tertiary interactions, with the total count comparable to that of non-canonical base pairs (Nissen et al. 2001). A-minor involves the insertion of the Sugar edges of adenines into the minor groove of Watson-Crick helices, preferentially at C-G base pairs, where the hydrogen bonds are formed between the adenine and the base pair (Nissen et al. 2001). Adenine can be replaced with other bases but the two most common and most stable (Sponer et al. 2007) types of A-minors are highly specific for adenine bases (types I and II, see (Nissen et al. 2001)). A-minors have been found in various types of non-coding RNAs including ribosomal RNA, ribozymes, riboswitches, and others (Xin et al. 2008). 23S and 5S ribosomal RNA contain almost 200 A-minors (Nissen et al. 2001). Particularly, codon-anticodon helices are recognized by ribosomes through intermolecular A-minors, which led to the conclusion that “the ribosome is a ribozyme” (Lescoute and Westhof 2006). A recent study (Torabi et al. 2021) reported a novel subclass of WC/H A-minors engaging Watson-Crick and/or Hoogsteen edges of an adenine instead of its Sugar edge.

A-minors tend to form clusters (Nissen et al. 2001). In (Nissen et al. 2001) authors describe a motif called A-patch that is formed by a stack of adenines involved in A-minors. In (Hamdani and Firdaus-Raih 2019) such A-patches of 2 stacked adenines and 2 consecutive base pairs have been introduced as sextuples, a novel RNA tertiary motif composed of six bases interconnected with hydrogen bonds. Usually, “A-minor interaction” or just "A-minor" are used to refer to an individual nucleotide triple (one adenine and one base pair), and the term “A-minor motif” to describe a clustering of two and more A-minor interactions (Hendrix et al. 2005). However, this principle is not conventional, and suffixes “interaction” and “motif” are often being interchanged, see e.g. (Nissen et al. 2001, Sheth et al. 2013). Hereinafter, we will use the terms "A-minor interaction" and "A-minor" to refer to an individual nucleotide triple, and "A-minor motif" to refer to a maximum local group of stacked A-minors, i.e. either to an isolated A-minor or to a cluster of two or more A-minors (see Materials and Methods for strict definitions).

A-minors play a crucial role in RNA structure stabilization and molecular recognition often serving as components of more complex motifs. A-minors stabilize coaxial stacking in multiple junctions (Lescoute and Westhof 2006, Geary et al. 2011) and are involved in ribose-zipper motifs (Tamura and Holbrook 2002). According to (Xin et al. 2008), these two types of motifs are the most common co-motifs for A-minor interactions. The kink-turn motif, a kink in the phosphodiester backbone that causes a sharp turn in the RNA helix, is generally stabilized by at least one A-minor interaction (Klein et al. 2001, Razga et al. 2005). GNRA tetraloop-receptor interaction, probably the most studied so far RNA tertiary motif, also employs A-minors (Geary 2008, Fiore and Nesbitt 2013, Wu et al. 2012). Another recurrent RNA motif, UAA/GAN internal loop forming interstrand adenine stack, binds long-range RNA regions via 2 or more A-minor interactions (Lee et al. 2006). ABAB-pseudoknots (also known as H-knots) are often stabilized by the triple helix that is usually formed by adenines of the 3'-closest loop and the minor groove of the 5'-closest stem (Aalberts and Hodas 2005). Thus, the A-minor is among the most important tertiary motifs of non-coding RNAs.

Still, the RNA secondary structure context of A-minors has been understudied. The only example is the analysis of ribosomal RNAs (Xin et al. 2008) where authors showed that 67% of the adenines involved in A-minor interactions are located in single-stranded regions forming tertiary motifs in hairpins, internal, or junction loops.

In this work, we present a novel classification of A-minors in terms of RNA secondary structure. Each nucleotide of an A-minor was attributed to the corresponding stem or loop and each pair of nucleotides was attributed to their relative position (from the same stem or loop, from adjacent stem and loop, or from distant elements). The results of the classification of A-minors from known RNA structures suggest that the structural context of an A-minor can mostly define its co-motifs.

## RESULTS

### Analysis of A-minors from the representative set of RNA structures shows that 64% of them are clustered

To perform a comprehensive analysis of A-minors, we assembled a dataset of 2431 A-minors annotated in the representative set of RNA spatial structures (Leontis and Zirbel 2012) from the Protein Data Bank (PDB, (Berman et al. 2002)), see "Annotation of A-minors in known RNA structures" of Materials and Methods.

The dataset included 2431 A-minor interactions that form 1504 A-minor motifs including 626 A-minor clusters. About 12% of all the A-minors were intermolecular, i.e. formed by 2 or 3 different RNA chains. Intermolecular A-minors were found to be isolated more often than the intramolecular ones (52% and 34% respectively, see Figure 1). A-minors of Type II were isolated only in 21% of cases, whereas Type I and X were isolated in a significantly larger number of cases, 37%, and 41% respectively. The overall number of clustered A-minors made up 64% (1553 out of 2431) and only 878 A-minors (36%) were isolated. However, it can be seen (Figure 1) that A-minor clusters themselves composed only 42% of all the A-minor motifs.

As shown in Figure 2, the majority of all A-minor motifs (878 out of 1504) are isolated A-minor interactions. The next biggest class contains 389 A-minor clusters of size (2, 2), i.e. clusters made of two adenines and two base pairs. A small number of clusters having more base pairs than adenines (19 cases) represent uncertain cases with some adenines located between two consecutive base pairs. The opposite case of more adenines than base pairs is quite common and is represented by 155 motifs, and the largest motifs include up to 5 base pairs and 7 adenines.

Thus, the majority of A-minors form clusters, but there are many isolated A-minors, especially among the intermolecular interactions.

### Largest structural classes of A-minors are in a good match with their known co-motifs

To identify the structural environment of A-minors, we introduced a new classification of A-minors in terms of their secondary structure context (see "Classification of A-minors" in Materials and Methods) and applied it to 2431 A-minors of our primary dataset.

All the 2431 A-minors belong to 99 different structural classes according to the proposed classification. 73 classes are represented by at least 2 A-minors and only 10 classes are represented by at least 50 cases. 2110 A-minors (87%) involve a base-pair belonging to some stem and either an adjacent adenine (35%, classes of form L-S-S-LC-LC-SM, i.e. the base pair belongs to a stem and the adenine belongs to an adjacent loop L of any type) or a distant adenine (52%, classes of form L-S-S-LR-LR-SM, i.e. the base pair belongs to a stem and the adenine belongs to a distant loop L of any type). The distribution of the classes is very diverse as no single class covers the major fraction of the cases: the largest class HC-S-S-LR-LR-SM makes up only 22.5% and the 10 most frequent classes make up 73.8% of the total number of A-minor interactions.

The 10 most frequent classes were analyzed with respect to their tendency to be clustered (see Figure 3). It was found that the share of clustered A-minors significantly varies among the classes being from 42.6% for adjacent bulged adenines (class BC-S-S-LC-LC-SM) to 92.6% for distant adenines from pseudoknotted internal loops (class IP-S-S-LR-LR-SM). We also considered A-minor motifs that contain A-minors of the most frequent classes. The share of clusters among them was also notably different ranging from about 35% to almost 84%.

Out of the 10 most frequent A-minor classes, only 5 largest classes frequently occur in non-ribosomal non-coding RNAs. The classes are in good concordance with known A-minor co-motifs, see Table 1 and the next three paragraphs for details.

Particularly, 68% of A-minors of the class HC-S-S-LR-LR-SM stabilize GAAA-11nt, GNRA-like/minor-groove, and other tetraloop-receptor motifs. 52% of A-minors of this class involve an adenine within GNRA-sequence.

Intramolecular IC-S-S-LR-LR-SM A-minor interactions have been found within UAA/GAN and UAA/GAN-like internal loops with cross-strand stacked adenines. The class also includes intermolecular A-minors between SSU RNA and codon-anticodon helix with adenines being looped out and not forming cross-strand stacks.

IC-S-S-LC-LC-SM A-minors are divided almost equally into two subgroups: kink-turn stabilization and the formation of a new motif, the across-bulged motif, that mimics tetraloop-receptor interaction (discussed below). A-minors of classes JC-S-S-LC-LC-SM and

HP-S-S-LC-LC-SM are in perfect correspondence with coaxial stacking stabilization and pseudoknot stabilization respectively.

Thus, the most common structural classes according to the proposed classification reflect well the widely known co-motifs of A-minors. We suggest the classification can be successfully used for automatic annotations as it significantly narrows down the possible A-minor co-motifs.

### The reference type I A-minors confirm general composition of A-minor classes

To verify the results, we replicated the analysis using a smaller set of 103 canonical A-minors of type I present both in our primary dataset and in the CaRNAval database (Reinharz et al. 2018), see "Reference set of A-minors" in Materials and Methods.

Out of 103 A-minors, 67 (65%) are L-S-S-LR-LR-SM (long-range A-minors) and 35 (34%) are L-S-S-LC-LC-SM (local A-minors), with the single exception of loop-loop HC-JC-JC-LR-LR-SM A-minor. Of note, according to the CaRNAval database, as many as 15 of canonical A-minors belong to the loop-loop class, but 14 of them actually include a base pair from a stem that had been discarded during the pseudoknot removal stage of the CaRNAval pipeline.

The overall share of clustered A-minors in the set of 103 type I A-minors is 74% (76 A-minors). The three largest structural classes are HC-S-S-LR-LR-SM (34 A-minors, 65% of them are clustered), IC-S-S-LR-LR-SM (21 A-minors, all of them are clustered), and JC-S-S-LC-LC-SM (10 A-minors, 70% of them are clustered). The three classes together comprise 63% of the subset.

All in all, the reference set of 103 type I A-minors confirmed main observations made on our primary dataset of 2431 A-minors.

## Analysis of IC-S-S-LC-LC-SM A-minors reveals a new recurrent motif

Analysis of IC-S-S-LC-LC-SM A-minor interactions (147 A-minors forming 100 A-minor motifs) revealed that along with 66 A-minors (56 motifs) involved in kink-turn stabilization there is a subgroup of 68 A-minors (34 motifs) involved in previously undescribed but recurrent motif (Supplementary Table S1). We named it the across-bulged motif as the thread opposite to A-minor adenines contains bulged out bases. All 34 such cases were found to be homologs of 7 unique motifs, 5 of which belong to ribosomal RNAs, and 1 is found in a riboswitch and in a single-guide RNA (Table 2). The structure of the ribosome from *Spinacia oleracea* (PDB entry 6ERI) was chosen to illustrate the ribosomal motifs as it was the only structure containing cases of all 5 unique motifs.

We consider a classical internal loop (IC) to be a canonical across-bulged motif, when: (1) its one thread contains at least one adenine that forms an A-minor with the preceding stem; (2) the flanking residues of the adenine's thread form a base triple either with the 5'-end or the 3'-end base of the opposite thread; (3) at least one of the opposite thread's bases is bulged out.

The number of adenines involved in A-minors of across-bulged motifs varies from 1 to 2 among different organisms. The bulged bases take part in the cross-strand base stacking or form A-minors, G-minors, and other N-minors, base-phosphate interactions (Zirbel et al. 2009), or RNA-protein interactions. Usually, the across-bulged motifs also include a base-triple (see Figure 4A), but in motifs I and II one base of the base triple is missing that leaves it with a base pair instead (Table 2).

We assumed that there could be across-bulged motifs without A-minors and inspected internal loops of similar sizes (4-2 and 4-3) within the 6ERI PDB entry. Indeed, we found such a motif with two pyrimidines instead of adenines (motif VIII, see Table 2 and Figure 4B).

The spatial structure of the across-bulged motif was found to resemble the structure of the well-known GAAA-11nt motif (see Figure 4C). In the case of the GAAA-11nt motif, the A-minors are formed with a GAAA-tetraloop, but within the across-bulged motif, the A-minors are formed with the local adenines of the internal loop.

### A-patch is the primary architecture of A-minor clusters

We examined all 389 A-minor clusters that consist of two adenines and two base pairs (i.e. of size (2, 2), see "A-minor interaction and A-minor motif definitions" in Materials and Methods). The majority (94%, 366 cases) of such clusters are of A-patch architecture (Nissen et al. 2001), i.e. are formed by a stack of adenines involved in A-minors with stacked base pairs. In 304 cases (83%), an A-patch includes a stack of consecutive adenines (see Figure 5a), and in 62 cases (17%) it is formed by a cross-strand adenine stack (see Figure 5b).

Both architectures include the same top 3 structural classes - HC-S-S-LR-LR-SM, IC-S-S-LR-LR-SM, and JC-S-S-LC-LC-SM, but in different proportions: 32%-15%, 15%-29%, and 12%-18% respectively. 271 out of 304 A-patches with consecutive adenine stack include consecutive within a stem base-pairs, 22 cases include non-consecutive stacked base-pairs and in 11 cases adenines of an A-patch are not consecutive but 1 nucleotide apart from each other in sequence. 38 out of 62 A-patches with a cross-strand adenine stack are formed by adenines from the same loop, and 24 cases include adenines from two distant RNA secondary structure elements.

Out of 177 A-minor clusters of larger sizes, 23 (13%) are formed by a single stretch of consecutive stacked adenines, 103 (58%) include cross-strand stacking of adenines, and the remaining 51 clusters (29%) are not of A-patch architecture.

The results suggest that the A-patch is the primary architecture of A-minor clusters. It's also worth noting that a minor but noticeable part of A-patches contains a cross-strand adenine stack that we believe allows the A-minor cluster to achieve greater stability.

**Particular classes of local A-minor motifs can be predicted with machine learning**

We applied the proposed classification of A-minors to assess if A-minor motifs can be predicted from the RNA sequence and secondary structure.

We formulated the problem of computational prediction of A-minors as a binary classification problem with a (stem, A-stretch) pair of a stem and a stretch of unpaired adenines (i.e. free of stem-forming base pairs) as an object of classification. The pairs that form A-minor interactions have been named *A-stems* and treated as positives. We trained a random forest classifier and, considering positive:negative class imbalance of 1:10 to 1:200 depending on the considered classes, used the area under the precision-recall curve (AUPRC) as the primary quality estimate (see "Machine learning framework" in Materials and Methods).

The cross-validation on the entire dataset of (stem, A-stretch) pairs showed very low overall AUPRC (below 0.1). However, we identified two types of local A-stems that can be predicted with notably higher quality (Figure 6). First, there were HP-LC A-stems, i.e. A-minor interactions between a pseudoknotted hairpin and an adjacent stem, which demonstrated the mean AUPRC of 0.73 (st. dev. 0.17) in 4-fold cross-validation. Second, there were IP-LC A-stems, i.e. A-minor interactions between a pseudoknotted internal loop and an adjacent stem, which demonstrated the mean AUPRC score of 0.43 (st. dev. 0.17). Other types of (stem, A-stretch) pairs did not achieve AUPRC scores higher than 0.2.

Although HP-LC and IP-LC (stem, A-stretch) pairs constitute a limited dataset size (21 HP-LC A-stems and 12 IP-LC A-stems with positive:negative class imbalance of 1:10, see Figure 6 caption), yet the results were stable in terms of cross-validation and were obtained with only 5 features, which were independently selected on HP-LC data when predicting IP-LC A-stems and vice versa. The 5 features used to predict HP-LC A-stems were the numbers of (1) nucleotides, (2) adenines, and (3) cytosines between the wings of the stem, (4) the number of nucleotides between the adenines and the left wing of the stem, and (5) the boolean value reflecting whether the thread that follows the left wing belongs to a pseudoknotted hairpin. The 5 features used to predict IP-LC A-stems were (1) the boolean value reflecting whether the wing preceding the adenines is the right wing of the stem, (2) the boolean value reflecting whether the wing that follows the adenines is the wing that follows the right wing of the stem, (3) the number of right wings between the adenines and the right wing of the stem, (4) the boolean value reflecting whether the thread that precedes the adenines is the thread that precedes the right wing of the stem, and (5) the boolean value reflecting whether the adenines' thread is the thread that follows the right wing of the stem (see also Supplementary Table S5).

Thus, we can conclude that the proposed classification can be successfully used to predict particular types of local A-minor motifs.

## DISCUSSION

In this work, we proposed strict definitions of A-minor motifs and A-minor clusters and used them on top of the DSSR annotation of A-minor interactions to identify the motifs in experimentally determined RNA spatial structures of the representative set with 3.0 Aº resolution cutoff. More than 60% of A-minors were found to be located in A-minor clusters. Nearly 90% of the clusters were found to adopt the well-known A-patch architectures.

We proposed the novel classification of tertiary motifs in terms of RNA secondary structure and applied it to analyze A-minors in known RNA structures. We found that the most frequent classes of A-minors correspond well to their most known co-motifs, such as tetraloop-receptor motifs and coaxial stacking. Such correspondence could be used to improve the automatic annotation of A-minor co-motifs using its structural context. It should be also noted that the proposed classification is not limited to A-minors and can be applied to a wide range of RNA tertiary motifs.

Detailed annotation of IC-S-S-LC-LC-SM A-minors revealed a novel recurrent motif of RNA tertiary structure, the across-bulged motif, that contains bulged bases in the opposite to the A-minors strand of the internal loop. The spatial structure of the across-bulged motif was found to resemble the structure of the well-known GAAA-11nt motif.

In a number of cases of the across-bulged motif, the bulged base also interacts with the minor groove of another base pair, forming G-minor or A-minor interactions. We were also able to find a case of the across-bulged motif with two pyrimidines instead of two adenines that form U-minor and C-minor interactions. Thus, although A-minor motifs prefer adenines, other bases can form analogous motifs. In the analysis of A-minor clusters, we also found A-patches of size (2, 2) that actually included another base stacked between the two adenines (see Figure 7). These findings suggest the need for annotation of other N-minors along with A-minors by the commonly used annotation software like the DSSR program used in the current work. The pipeline in (Reinharz et al. 2018) is a good example of such annotation of A-minors not restricted to adenine bases. The formation of N-minors is also suggested to be included in consideration for evolutionary analyses dealing with point mutations.

With a machine learning framework, we showed that HP-LC A-stems can be predicted *in silico* from the RNA sequence and secondary structure with an acceptable quality (0.73 AUPRC).

Other than HP-LC A-stems, only IP-LC A-stems allowed reaching AUPRC over 0.2. We consider the following explanation. In the case of HC-LC, IC-LC, and other local (stem, A-stretch) pairs with an A-stretch belonging to a classical loop, the information of relative features is very limited and doesn't describe the environment outside the loop and its adjacent stems. In the case of HC-LR, IC-LR, and other long-range (stem, A-stretch) pairs, there is an opposite effect, the relative features cover a large amount of sequence and secondary structure volume which is irrelevant for a given pair. However, in the case of local (stem, A-stretch) pairs with an A-stretch from a pseudoknotted loop, the features describing relative distances reflect the relevant local context that allows distinguishing an A-stem from a non-interacting (stem, A-stretch) pair.

## CONCLUSION

In this work, we proposed a novel classification to describe A-minors in terms of RNA secondary structure. The classification was applied to A-minors annotated in the known RNA 3D-structures. The dataset consisted of more than 2400 interactions forming more than 1500 motifs. The majority of A-minors formed clusters of the typical size of 2-3 interactions. The analysis of the largest annotated classes showed that they are highly homogeneous and in good agreement with the known co-motifs of A-minors. We also showed that the local A-minors from internal loops can not only stabilize kink-turn motifs but also form a novel recurrent RNA tertiary motif, the across-bulged motif. The across-bulged motif was found to mimic the well-known GAAA-11nt motif but with local adenines forming A-minors. Using a machine-learning framework we showed that the particular local classes of A-minors with adenines from pseudoknotted loops can be predicted using the sequence and secondary structure information. Thus, we show that the proposed classification can be successfully used both to automatically annotate co-motifs of A-minor motifs by their structural context and to predict A-minors of particular local classes.

## MATERIALS AND METHODS

### A-minor interaction and A-minor motif definitions

Our definition of the A-minor interaction follows the one from the DSSR program (Lu et al. 2015) that is widely used for RNA motifs annotation.

The *A-minor interaction* (*A-minor*) is the nucleotide triple of an adenine and a cWW base pair (cis-Watson-Crick/Watson-Crick base pair according to the Leontis-Westhof classification (Leontis and Westhof 2001)), where the adenine faces the minor groove of the base pair and forms H-bonds. According to the DSSR manual, in canonical A-minors (types I and II), the adenine has its minor groove edge facing the minor groove of a base pair, and the O2' atom of adenine is involved in H-bonds with the pair; in the miscellaneous type X (eXtended), the adenine uses its Watson-Crick edge or major-groove edge to interact with the minor groove of a base pair, without resorting to the O2' atom (see page 38 at http://docs.x3dna.org/dssr-manual.pdf). Furthermore, in type I A-minors both the O2' and the N3 of the adenine are inside the minor groove of the base pair (i.e. lie in between O2' atoms of the base pair), and in the type II version, the O2' of the adenine is outside the near strand O2' atom whereas the N3 of the adenine is inside (Nissen et al. 2001).

The adenine of A-minor is referred to as “*A*”, the nucleotides of the base pair are referred to as “*L*” (if located closer to the 5’-end of the RNA chain) and “*R*” (if located closer to the 3’-end of the RNA chain). An A-minor interaction is called *intramolecular* if all three participating nucleotides belong to the same RNA chain and *intermolecular* otherwise. If “*L*” and “*R*” nucleotides belong to different RNA chains their assignment order is determined by the lexical order of their RNA chain identifiers.

For each entry from the Protein Data Bank (PDB, (Berman et al. 2002)), we constructed an undirected graph $G = (V, E)$, where $V = \{ v_i = (A_i, L_i, R_i) \}$ is the set of A-minor interactions annotated with DSSR and $E = \{ e_{ij} = (v_i, v_j) \}$ is the set of edges between them. $(v_i, v_j) \in E$ if either there are $N_i$ *and* $N_j$ that are the same nucleotide or there are $N_i$ and $N_j$ that are stacked, where $N_i \in \{A_i, L_i, R_i\}$ and $N_j \in \{A_j, L_j, R_j\}$. A connected component within the graph $G$ is called the *A-minor motif*. The A-minor motif is called the *A-minor cluster* if it involves at least two different adenines $A_i$ and $A_j$ or two different base pairs $(L_i, R_i)$ and $(L_j, R_j)$. The size of an A-minor motif is defined by a pair of numbers: the number of adenines and the number of base pairs, e.g. A-minor motif of size (3, 2) involves three adenines and two base pairs. We call the A-minor interaction *clustered* if it belongs to an A-minor cluster.

## Classification of A-minors

To describe the RNA secondary structure, we used the generalization of the Nearest Neighbor Model (NNM (Mathews et al. 1999)) proposed in (Baulin et al. 2016). This approach allows a uniform description of arbitrary secondary RNA structures including pseudoknotted structures of any complexity, thus ensuring that each nucleotide is associated with at least one secondary structure element. The following additional definitions are required for the A-minor classification (the complete set of strict definitions is provided in Supplementary Text S1 and at http://urs.lpm.org.ru/struct.py?where=3#def).

A *stem* is a sequence of at least two consecutive Watson-Crick or Wobble base pairs. Two strands of a stem are called its *left wing* and *right wing*. A *loop* is a set of *threads* (regions that are free of stem-forming base pairs) confined by a stem. Each loop is assigned with one of the following common types: *hairpin* (H), *bulge* (B), *internal loop* (I), or *multiple junction* (J). In addition, each loop is classified in regard to pseudoknots: a loop is called *pseudoknotted* (P) if it is involved in a pseudoknot, *isolated* (I) if it is adjacent to a pseudoknot, and *classical* (C) otherwise (see Figure 8). Pseudoknotted loops may contain both threads and stem wings.

Each nucleotide is being ascribed either with a stem (S) or with a set of loops (concatenation of loops' letter pairs of pattern [HIBJ][ICP] in alphabetical order).

All A-minors were classified with respect to the RNA secondary structure elements involving their nucleotides. Each nucleotide of an A-minor was attributed to the corresponding stem or loop(s), and each pair of nucleotides was attributed to their relative position: within the same stem or the same loop (same element, SM), from an adjacent stem and loop (local, LC), from distant elements (long-range, LR). Thus, each A-minor belongs to A-B-C-AB-AC-BC structural class, where A-B-C are secondary structure elements of the A, L, and R nucleotides respectively, and AB-AC-BC are relations of the AL, AR, and LR nucleotide pairs.

An example of an A-minor from a lysine riboswitch (PDB code 3D0U, chain A) is presented in Figure 9. Here the adenine A124 belongs to a classical hairpin (HC) adjacent to stem 8. A20 of the noncanonical base pair (A20, G66) belongs to the classical internal loop (IC) confined by stem 2. G66 of the base pair belongs to the same loop and also belongs to a pseudoknotted hairpin (HP) of stem 4 and therefore is assigned with HPIC. As A20 and G66 share a loop, the nucleotide pair A20-G66 is annotated with the relative position SM (from the same element). Pairs A124-A20 and A124-G66 are annotated with LR as the nucleotides are distant from each other within the secondary structure. Overall, the A-minor is classified as having type HC-IC-HPIC-LR-LR-SM.

Of note, the idea of the proposed classification can be used for other RNA motifs and interactions. For example, non-canonical base pairs can be ascribed with an A-B-AB class. For motifs with a rather larger number of bases, a more coarse-grained system can be applied, for instance, switching the focus from nucleotides to threads and wings.

### Annotation of A-minors in known RNA structures

1074 RNA-containing PDB entries from the representative set of RNA structures (version 3.76 with the 3.0 A resolution cutoff (Leontis and Zirbel 2012)) were selected for the analysis. To annotate A-minor interactions, the DSSR program (version v1.8.5-2018nov29 (Lu et al. 2015)) was used. A-minor motifs were annotated using the python library from the URSDB (https://github.com/febos/urslib). The resulting dataset included 2431 A-minors composing 1504 A-minor motifs (see Supplementary Table S2 and Supplementary Table S3).

Each A-minor interaction was annotated with the features related to its geometric parameters, involved H-bonds, the local context of RNA secondary structure including annotations of RNA tetraloop sequences (Klosterman et al. 2004), and the size of the corresponding A-minor motif (see Supplementary Table S2 for the detailed description of all features). Edges within A-minor clusters were annotated with the features of the involved A-minor interactions and base stacking interactions between them along with the edge description in the form of $N_idN_j$ relationships, where $N_k$ is *A*, *L*, or *R* of the corresponding A-minor $v_k$ and $d \in$ *{"e"* - equality*, "n"* - consecution*, "s"* - stacking*, "ns"* - stacking and consecution*}.* For example, the description "***AsA_LeL_ReR***" depicts an edge between two A-minors made of the same base pair and non-consecutive stacked adenines (see Supplementary Table S3 for the detailed description of all features).

## Reference set of A-minors

To validate the DSSR annotation we compared the primary dataset of 2431 A-minors with the A-minors provided in the CaRNAval database (Reinharz et al. 2018). Since the CaRNAval database contains only motifs that include at least two base pairs between two different secondary structure elements, only A-minors of type I are present among three-nucleotide motifs (RIN#2, 194 occurrences, see http://carnaval.lri.fr/all_HEADERS/info_cluster_2.html). These 194 motifs occur in 37 RNA chains, 21 of which are also included in our primary dataset. These RNA chains include 103 A-minors from the CaRNAval database, and all of them are also presented in our dataset (see Supplementary Table S2). In total, for those 21 RNA chains our dataset includes 240, 119, and 210 intramolecular A-minors of type I, type II, and type X respectively.

Such a significant difference (more than 2-fold in the case of type I A-minors) between the datasets arises from the following issues. First, the CaRNAval pipeline is more rigorous with respect to base pairs, i.e. requires the adenine to form two strict Sugar-Edge/Sugar-Edge base pairs (Leontis and Westhof 2001) for A-minor to be annotated, whereas the A-minor definition used in the DSSR does not require the adenine to form base pairs. Furthermore, out of 2431 A-minors of our full dataset, DSSR annotated no base pairs formed by the adenine for 683 cases, only one base pair for 1609 cases (including base pairs of intermediate Leontis-Westhof types), and two base pairs for 139 A-minors. Thus it is clear that the DSSR definition of A-minor cannot be reduced to a pure base triple, i.e. three bases interconnected by a set of base pairs, and following the DSSR definition allows to avoid discarding numerous intermediate cases. Second, unlike CaRNAval, the DSSR annotation includes intermolecular A-minors and A-minors containing modified nucleotides and nucleotides with missing atoms. Third, the CaRNAval pipeline does not consider the A-minors formed within a secondary structure element

(X-X-X-SM-SM-SM classes according to the proposed classification, where X is any stem or any set of loops).

Of note, unlike DSSR, the CaRNAval pipeline does not restrict an A-minor to be formed by an adenine base, but the only such annotated example of a type I A-minor formed by a guanine base does not belong to the considered set of 21 RNA chains.

**Machine learning framework**

We formulate the task of computational prediction of A-minors as a binary classification problem. To choose an object of classification we examined the annotated A-minors from the known RNA structures. First, DSSR annotates a considerable number of intermediate cases of A-minors, where, for example, an adenine is located evenly between two consecutive base pairs such that it is unclear with which particular base pair the adenine forms the A-minor interaction. Second, more than 60% of all annotated A-minors belong to A-minor clusters and nearly 90% of all A-minors include a base pair that belongs to some stem. Considering these facts, to avoid fitting the model to the technical features of the annotation rather than to principal features of A-minor interactions, we used a more coarse-grained approach and chose a *(stem, A-stretch)* pair of a stem and a stretch of unpaired adenines (i.e. adenines free of stem-forming base pairs) as the target object of the classification. Thus we considered all possible (stem, A-stretch) pairs and trained a model to predict if a particular (stem, A-stretch) pair forms an *A-stem*, i.e. forms at least one A-minor interaction (see Figure 10).

The representative set of structures consisted of 130 RNA chains containing A-minor interactions. To reduce redundancy we manually excluded homologous RNA chains from different organisms. The resulting 44 RNA chains included exactly one structure of each type of RNA molecule present in the representative set. The resulting set of (stem, A-stretch) pairs included 347 A-stems and 183298 non-interacting (stem, A-stretch) pairs (0,19% positive rate, the complete data is available at https://github.com/febos/urs_aminors).

The features for classification were based on RNA sequence and secondary structure information. Each (stem, A-stretch) pair was annotated with 288 features describing local contexts of the stem and the A-stretch (*local features*, e.g. the base pair content of the stem, the stem's length, types of neighboring loops, lengths of neighboring wings and threads, the length of the A-stretch, base types of A-stretch neighboring residues, etc.), and distances between the stem's wings and between each wing and the A-stretch in terms of sequence and secondary structure (*relative features*, e.g. number of guanines in sequence between the objects, number of right wings, number of bulges, whether the A-stretch is located between the stem's wings, whether the thread adjacent to the stem's left wing is the A-stretch thread, etc.). A detailed description of all features is provided in Supplementary Table S5 and at https://github.com/febos/urs_aminors.

(stem, A-stretch) pairs were classified in terms of RNA secondary structure in a similar manner as A-minors. Thus, a (stem, A-stretch) pair that could form A-minor interactions of type *IC-S-S-LC-LC-SM* has been attributed to *IC-LC* type.

Next, we applied the RandomForest algorithm (Liaw and Wiener 2002) implemented in the scikit-learn Python package (Pedregosa et al. 2011). The experiments have been carried out using 4-fold cross-validation method (Rodriguez et al. 2009).

Each model has been trained on the 5 best features selected using an automatic feature selection tool of scikit-learn (SelectFromModel). The parameter class_weight = 'balanced' was used to account for class disbalance. All the other hyper-parameters were left at default values. To ensure there is no information leak through huge pool of features, the best results for HP-LC and IP-LC classes were validated by performing feature selection using (stem, A-stretch) pairs of a different class: the results for the HP-LC pairs have been obtained using the 5 best features selected for the IP-LC pairs and vice versa. The source code is available at https://github.com/febos/urs_aminors/blob/master/ML_Astems.ipynb). To assess the quality of the binary classification results we used the area under the precision-recall curve (AUPRC).

## SUPPLEMENTARY INFORMATION

Suplemental_Text_S1.pdf - **Supplementary Text S1**. A generalized description of arbitrary RNA secondary structure. List of definitions.

FileS1.jmol - **Supplementary File S1**. A saved state for the Jmol visualization software (http://jmol.sourceforge.net/) with a structural alignment of eight representative across-bulged motifs from Table 2. The structures have been superposed using the SETTER web-server (http://setter.projekty.ms.mff.cuni.cz/). The adenines that form A-minors are shown in blue, the base triple is shown in yellow, the bulged bases are shown in purple.

FigureS1.jpg - **Supplementary Figure S1**. A structural alignment of eight representative across-bulged motifs from Table 2. The structures have been superposed using the SETTER web-server (http://setter.projekty.ms.mff.cuni.cz/). The adenines that form A-minors are shown in blue, the base triple is shown in yellow, the bulged bases are shown in purple.

Supplemental_Table_S1.pdf - **Supplementary Table S1**. List of across-bulged motifs found in the representative set of PDB structures.

Supplemental_Table_S2.xlsx - **Supplementary Table S2**. Lists of 2431 A-minor interactions annotated in this study, 194 A-minor interactions of type I annotated in work (Reinharz et al. 2018), and 103 A-minor interactions of type I from the intersection of the two datasets.

Supplemental_Table_S3.xlsx - **Supplementary Table S3**. List of the edges forming 626 A-minor clusters.

Supplemental_Table_S4.xlsx - **Supplementary Table S4**. Nonredundant set of 44 RNA chains.

Supplemental_Table_S5.xlsx - **Supplementary Table S5**. Features of (stem, A-stretch) pairs. List of descriptions

## ACKNOWLEDGEMENTS

Not applicable.

## FUNDING

ML analysis was supported by RSF grant 20-74-10075 to I.V.K.

## REFERENCES

Aalberts DP, Hodas NO. Asymmetry in RNA pseudoknots: observation and theory. Nucleic Acids Research. 2005 Jan 1;33(7):2210-4. 10.1093/nar/gki508

Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ. Developmentally regulated piRNA clusters implicate MILI in transposon control. Science. 2007 May 4;316(5825):744-7. 10.1126/science.1142612

Baulin E, Yacovlev V, Khachko D, Spirin S, Roytberg M. URS DataBase: universe of RNA structures and their motifs. Database. 2016 Jan 1;2016. 10.1093/database/baw085

Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P. The protein data bank. Acta Crystallographica Section D: Biological Crystallography. 2002 Jun 1;58(6):899-907. 10.1107/S0907444902003451

Breaker RR. Riboswitches and translation control. Cold Spring Harbor perspectives in biology. 2018 Nov 1;10(11):a032797. 10.1101/cshperspect.a032797

Costa FF. Non-coding RNAs: meet thy masters. Bioessays. 2010 Jul;32(7):599-608. 10.1002/bies.200900112

Darty K, Denise A, Ponty Y. VARNA: Interactive drawing and editing of the RNA secondary structure. Bioinformatics. 2009 Aug 1;25(15):1974. 10.1093/bioinformatics/btp250

Dvinge H, Guenthoer J, Porter PL, Bradley RK. RNA components of the spliceosome regulate tissue-and cancer-specific alternative splicing. Genome research. 2019 Oct 1;29(10):1591-604. 10.1101/gr.246678.118

Fiore JL, Nesbitt DJ. An RNA folding motif: GNRA tetraloop–receptor interactions. Quarterly reviews of biophysics. 2013 Aug;46(3):223-64. 10.1017/S0033583513000048

Geary C, Baudrey S, Jaeger L. Comprehensive features of natural and in vitro selected GNRA tetraloop-binding receptors. Nucleic acids research. 2008 Mar 1;36(4):1138-52. 10.1093/nar/gkm1048

Geary C, Chworos A, Jaeger L. Promoting RNA helical stacking via A-minor junctions. Nucleic acids research. 2011 Feb 1;39(3):1066-80. 10.1093/nar/gkq748

Hamdani HY, Firdaus-Raih M. Identification of Structural Motifs Using Networks of Hydrogen-Bonded Base Interactions in RNA Crystallographic Structures. Crystals. 2019 Nov;9(11):550. 10.3390/cryst9110550

Hendrix DK, Brenner SE, Holbrook SR. RNA structural motifs: building blocks of a modular biomolecule. Quarterly reviews of biophysics. 2005 Aug;38(3):221-43. 10.1017/S0033583506004215

Hollands K, Proshkin S, Sklyarova S, Epshtein V, Mironov A, Nudler E, Groisman EA. Riboswitch control of Rho-dependent transcription termination. Proceedings of the National Academy of Sciences. 2012 Apr 3;109(14):5376-81. 10.1073/pnas.1112211109

Kerpedjiev P, Hammer S, Hofacker IL. Forna (force-directed RNA): simple and effective online RNA secondary structure diagrams. Bioinformatics. 2015 Oct 15;31(20):3377-9. 10.1093/bioinformatics/btv372

Kiss T. Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. The EMBO journal. 2001 Jul 16;20(14):3617-22. 10.1093/emboj/20.14.3617

Klein DJ, Schmeing TM, Moore PB, Steitz TA. The kink-turn: a new RNA secondary structure motif. The EMBO journal. 2001 Aug 1;20(15):4214-21. 10.1093/emboj/20.15.4214

Klosterman PS, Hendrix DK, Tamura M, Holbrook SR, Brenner SE. Three-dimensional motifs from the SCOR, structural classification of RNA database: extruded strands, base triples, tetraloops and U-turns. Nucleic acids research. 2004 Apr 15;32(8):2342-52. 10.1093/nar/gkh537

Lai D, Proctor JR, Zhu JY, Meyer IM. R-CHIE: a web server and R package for visualizing RNA secondary structures. Nucleic acids research. 2012 Jul 1;40(12):e95-. 10.1093/nar/gks241

Lee JC, Gutell RR, Russell R. The UAA/GAN internal loop motif: a new RNA structural element that forms a cross-strand AAA stack and long-range tertiary interactions. Journal of molecular biology. 2006 Jul 28;360(5):978-88. 10.1016/j.jmb.2006.05.066

Leontis NB, Westhof E. Geometric nomenclature and classification of RNA base pairs. Rna. 2001 Apr 1;7(4):499-512. 10.1017/s1355838201002515

Leontis NB, Lescoute A, Westhof E. The building blocks and motifs of RNA architecture. Current opinion in structural biology. 2006 Jun 1;16(3):279-87. 10.1016/j.sbi.2006.05.009

Leontis NB, Zirbel CL. Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking. In: RNA 3D structure analysis and prediction 2012 (pp. 281-298). Springer, Berlin, Heidelberg. 10.1007/978-3-642-25740-7_13

Lescoute A, Westhof E. The A-minor motifs in the decoding recognition process. Biochimie. 2006 Aug 1;88(8):993-9. 10.1016/j.biochi.2006.05.018

Lescoute A, Westhof E. Topology of three-way junctions in folded RNAs. Rna. 2006 Jan 1;12(1):83-93. 10.1261/rna.2208106

Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002 Dec 3;2(3):18-22.

Lu XJ, Bussemaker HJ, Olson WK. DSSR: an integrated software tool for dissecting the spatial structure of RNA. Nucleic acids research. 2015 Dec 2;43(21):e142. 10.1093/nar/gkv716

Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. Journal of molecular biology. 1999 May 21;288(5):911-40. 10.1006/jmbi.1999.2700

Mládek A, Šponer JE, Kulhánek P, Lu XJ, Olson WK, Šponer J. Understanding the sequence preference of recurrent RNA building blocks using quantum chemistry: the intrastrand RNA dinucleotide platform. Journal of chemical theory and computation. 2012 Jan 10;8(1):335-47. 10.1021/ct200712b

Montange RK, Batey RT. Riboswitches: emerging themes in RNA structure and function. Annu. Rev. Biophys.. 2008 Jun 9;37:117-33. 10.1146/annurev.biophys.37.032807.130000

Nissen P, Ippolito JA, Ban N, Moore PB, Steitz TA. RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. Proceedings of the National Academy of Sciences. 2001 Apr 24;98(9):4899-903. 10.1073/pnas.081082398

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: Machine learning in Python. Journal of machine learning research. 2011;12(Oct):2825-30.

Rázga F, Koča J, Šponer J, Leontis NB. Hinge-like motions in RNA kink-turns: the role of the second A-minor motif and nominally unpaired bases. Biophysical journal. 2005 May 1;88(5):3466-85. 10.1529/biophysj.104.054916

Reinharz V, Soulé A, Westhof E, Waldispühl J, Denise A. Mining for recurrent long-range interactions in RNA structures reveals embedded hierarchies in network families. Nucleic acids research. 2018 May 4;46(8):3841-51. 10.1093/nar/gky197

Rodriguez JD, Perez A, Lozano JA. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE transactions on pattern analysis and machine intelligence. 2009 Dec 1;32(3):569-75. 10.1109/TPAMI.2009.187

Sheth P, Cervantes-Cervantes M, Nagula A, Laing C, Wang JT. Novel features for identifying A-minors in three-dimensional RNA molecules. Computational biology and chemistry. 2013 Dec 1;47:240-5. 10.1016/j.compbiolchem.2013.10.004

Šponer JE, Reblova K, Mokdad A, Sychrovský V, Leszczynski J, Šponer J. Leading RNA tertiary interactions: structures, energies, and water insertion of A-minor and P-interactions. A quantum chemical view. The Journal of Physical Chemistry B. 2007 Aug 2;111(30):9153-64. 10.1021/jp0704261

Tamura M, Holbrook SR. Sequence and structural conservation in RNA ribose zippers. Journal of molecular biology. 2002 Jul 12;320(3):455-74. 10.1016/S0022-2836(02)00515-6

Torabi SF, Vaidya AT, Tycowski KT, DeGregorio SJ, Wang J, Shu MD, Steitz TA, Steitz JA. RNA stabilization by a poly (A) tail 3′-end binding pocket and other modes of poly (A)-RNA interaction. Science. 2021 Feb 5;371(6529). 10.1126/science.abe6523

Walter AE, Turner DH, Kim J, Lyttle MH, Müller P, Mathews DH, Zuker M. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. Proceedings of the National Academy of Sciences. 1994 Sep 27;91(20):9218-22. 10.1073/pnas.91.20.9218

Wu L, Chai D, Fraser ME, Zimmerly S. Structural variation and uniformity among tetraloop-receptor interactions and other loop-helix interactions in RNA crystal structures. PloS one. 2012;7(11). 10.1371/journal.pone.0049225

Xin Y, Laing C, Leontis NB, Schlick T. Annotation of tertiary interactions in RNA structures reveals variations and correlations. Rna. 2008 Dec 1;14(12):2465-77. 10.1261/rna.1249208

Zirbel CL, Šponer JE, Šponer J, Stombaugh J, Leontis NB. Classification and energetics of the base-phosphate interactions in RNA. Nucleic acids research. 2009 Aug 1;37(15):4898-918. 10.1093/nar/gkp468

## TABLES

Table 1. Most common A-minor motifs

| Class | Description | #A-minors | #A-minor motifs | Co-motifs (Percentage of A-minors) |
|---|---|---|---|---|
| HC-S-S-LR-LR-SM | **Adenine** belongs to a classical hairpin<br>**Base-pair** belongs to a stem<br>**Long-range interaction** | 547 | 374 | Tetraloop-receptor motifs (68%) (Geary et al. 2008, Fiore and Nesbitt 2013, Wu et al. 2012) |
| IC-S-S-LR-LR-SM | **Adenine** belongs to a classical internal loop<br>**Base-pair** belongs to a stem | 352 | 180 | UAA/GAN & UAA/GAN-like internal loops (76%) (Lee et al. 2006); codon:anticodon base |

| | | | | |
|---|---|---|---|---|
| | **Long-range interaction** | | | pairs (6%) (Lescoute and Westhof 2006) |
| `JC-S-S-LC-LC-SM` | **Adenine** belongs to a classical junction<br>**Base-pair** belongs to a stem<br>**Local interaction** | 276 | 201 | Coaxial stacking (84%) (Lescoute and Westhof 2006, Geary et al. 2011) |
| `IC-S-S-LC-LC-SM` | **Adenine** belongs to a classical internal loop<br>**Base-pair** belongs to a stem<br>**Local interaction** | 147 | 100 | Kink-turns (45%) (Klein et al. 2001, Razga et al. 2005);<br>Across-bulged motifs (46%) |
| `HP-S-S-LC-LC-SM` | **Adenine** belongs to a pseudoknotted hairpin<br>**Base-pair** belongs to a stem<br>**Local interaction** | 112 | 51 | H-knots (38%) & kissing hairpins (62%) (Aalberts and Hodas 2005) |

Table 2. Representatives of across-bulged motifs. Bulged bases and bases involved in A-minor interactions are shown in capital letters.

| **Unique motif** | **PDB: Chain** | **A-minors** | | **All bases** | **Base triple / non-canonical base pair** | **Adenine(s) thread** | **Thread with bulged bases** | **Bulged bases involved in** | **Molecule** |
|---|---|---|---|---|---|---|---|---|---|
| | | **Adenines** | **Base pairs** | | | | | | |
| I | 4yaz:R | A78<br>A79 | (C7,G77)<br>(G8,C76) | U3-G8,<br>C76-A82 | A6,A80<br>cWH | `-AAa` | `-C--a` | | Cyclic di-GMP-I riboswitch |
| II | 6dtd:C | A30 | (G11,U28) | A7-C12<br>G27-A32 | U10,A29<br>cWW | `aA--` | `-C--u` | RNA-protein int. | Cas13b sgRNA |
| III | 6eri:Ax | A57<br>A58 | (A29,U55)<br>(C30,G54) | A22-C30<br>G54-U61 | U56,C28,A59<br>cWW,cSW | `uAAa` | `gAacc` | A-minor with (C4,G117) | 5S rRNA |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| IV | 6eri:BA | A1095 | (C1078,G1093) | A1073-C1079<br>G1092-U1099 | A1094,G1077,C1097<br>cWW,cSW | aAcc | uU--g | cross-strand stack with C1230;<br>RNA-protein int.;<br>5BPh with A1229 | 16S rRNA |
| V | 6eri:BA | A345<br>A346 | (C342,G363)<br>(G343,C362) | C342-G349<br>C358-G363 | C344,A361,U347<br>tWH,cWW | cAAu | -G--a | G-minor with (C60,G326);<br>4BPh with C341 | 16S rRNA |
| VI | 6eri:AA | A1342<br>A1343 | (C1339,G1355)<br>(G1340,C1354) | C1339-G1345<br>C1351-G1355 | C1341,C1352,G1344<br>tWH,cWW | cAAg | cA--- | cross-strand stack with A1645 | 23S rRNA |
| VII | 6eri:AA | A876 | (G874,C921) | A873-U880<br>G916-U922 | C875,A918,U878<br>tWH,cWW | cAcu | aAA-- | A-minors with (C2280,G2294),<br>(C2281,G2293) | 23S rRNA |
| VIII | 6eri:AA | U2699<br>C2700 | (A2696,U2746)<br>(C2697,G2745) | A2696-G2703<br>C2741-U2746 | C2698,A2743,U2701<br>tWH,cWW | cUCu | aA-- | cross-strand stack with A1700 | 23S rRNA |

## FIGURE LEGENDS

Figure 1. Distribution of clustered A-minor interactions by geometric and molecular types. Shares of clustered A-minor motifs are shown in the lower part of the figure. The absolute numbers of the interactions and motifs are shown in grey. Shares of clustered interactions and clusters are shown in orange. The chart is in log-scale. An A-minor motif is counted as inter/intra-molecular if it contains at least one inter/intra-molecular A-minor interaction, therefore the sum of intramolecular and intermolecular A-minor motifs is not equal to the total number of motifs.

Figure 2. A-minor motifs rarely involve more than 3 adenines. Circle areas are proportional to the number of A-minor motifs.

Figure 3. The share of clustered A-minors significantly varies among the 10 most frequent structural classes. The absolute numbers of the interactions and motifs are shown in grey. Shares of clustered interactions and clusters are shown in orange. Both charts are in log-scale. *Counted all motifs containing at least one interaction of the given class. Therefore the sum of the motif numbers is not equal to the total number of motifs.

Figure 4. 3D structure and RNA secondary structure scheme of two across-bulged motifs and a GAAA-11nt motif. (A) motif VI, a representative case from 6ERI entry (see Table 2) (B) motif VIII, a representative case from 6ERI entry (see Table 2) (C) GAAA-11nt motif from 2R8S, chain R, involving A-minor A152|C223-G250. Base pairs of stems and adenines of 11nt-loop are shown in grey. Base-triples are shown in yellow. A-minor adenines are shown in blue. Bulged bases are shown in purple.

Figure 5. Different A-patch architectures of size (2,2). (A) A-patch formed by a stack of consecutive adenines, SSU rRNA (PDB ID: 6QZP, chain S2, A-minors: A996|C674-(A2M)1031, A997|G673-C1032) (B) A-patch formed by a cross-strand adenine stack, LSU rRNA (PDB ID: 5TBW, chain 1, A-minors: A2696|C2630-G2648, A2758|U2629-A2649)

Figure 6. Precision-Recall curves representing the 4-fold cross-validation (cv) results on (A) the dataset of HP-LC (stem, A-stretch) pairs (169 negatives and 21 positives, 11.05% positive rate, 0.7 AUPRC), and (B) the dataset of IP-LC (stem, A-stretch) pairs (116 negatives and 11 positives, 8.66% positive rate, 0.35 AUPRC). The line of precision-recall break-even points is shown as blue dots.

Figure 7. AUA A-patch from LSU rRNA, PDB ID 5TBW, chain 1. A-minors: A3106|C2893-G2908 (in orange), A3129|A2892-U2909 (in gray). (A) Top view. (B) Side view. Uracil U3105 stacked between two adenine bases is shown in purple.

Figure 8. Pseudoknot-related classes of internal loops. (A) An internal loop in the absence of pseudoknots is called Classical. (B) An internal loop adjacent to a pseudoknot is called Isolated. (C) An internal loop involved in a pseudoknot is called Pseudoknotted. The graph has been prepared using R-chie (Lai et al. 2012) and forna (Kerpedjiev et al. 2015).

Figure 9. The secondary structure of the lysine riboswitch from PDB entry 3D0U. The structure is visualized with VARNA (Darty et al. 2009). Loops are assigned with their types, classes, and confining stems. A-minor A124-A20-G66 of type X is emphasized on the structure and presented separately with the 3D structure of its nucleotides in red, green, and blue. Each nucleotide of the A-minor is annotated with the element of RNA secondary structure. Each nucleotide pair is annotated with their relative positions within the secondary structure.

Figure 10. Definition of an A-stem classification problem. If a (stem, A-stretch) pair involves A-minors it belongs to the positive class and to the negative class otherwise. The emphasized stem consists of two base pairs, and the emphasized stretch of unpaired adenines consists of two bases.