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

Anna A. Shalybkova1, Darya S. Mikhailova2, Ivan V. Kulakovskiy3,4,5, Liliia I. Fakhranurova6,7, Eugene F. Baulin2,8,\*

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](mailto: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 A-minors from the known RNA structures, we found that the derived classes correspond well to the known A-minor functions. 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 reliably predicted from sequence and secondary structure. We suggest the proposed classification will be useful for automatic annotation of not only A-minors but various types of RNA tertiary motifs.

# 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 a modular organization and is composed of so-called RNA tertiary motifs, the structural “building blocks” that are often recurrent and hold their configuration in different structural environments (Leontis et al. 2006). Unlike elements of RNA secondary structure (the stems and loops), the tertiary motifs generally do not have strict definitions and include diverse elements, e.g. coaxial stacking (8 and more nucleotides forming two stacked stems (Walter et al. 1994)) and dinucleotide platform (a base pair between consecutive nucleotides (Mladek et al. 2012)).

The A-minor motif is the most abundant type of RNA tertiary motifs (Nissen et al. 2001). A-minor involves the insertion of the sugar edges of adenines into the minor groove of 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-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 stretch of stacked 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 the MATERIALS AND METHODS section 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. 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 relationship between the RNA secondary structure context of A-minors and the diversity of their functions 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 a motif can reliably define its function.

# 

# RESULTS

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

The prepared dataset included 2431 A-minor interactions that form 1504 A-minor motifs, 626 of the motifs being 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 itself 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 cases of A-minor clusters of size (2, 2). A small number of the 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.

The results show that whereas the majority of A-minors form clusters, there is still a great number of isolated A-minors suggesting they are not weak temporary interactions but rather significant tertiary motifs.

## Structural classes of A-minors agree well with their functions

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 ?-S-S-LC-LC-SM, i.e. the base pair nucleotides are from the same stem and the adenine is from an adjacent loop of some type) or a distant adenine (52%, classes of form ?-S-S-LR-LR-SM). 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 corresponding 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 functions (Table 1).

Particularly, A-minors of the class HC-S-S-LR-LR-SM stabilize GAAA-11nt and GNRA-like/minor-groove motifs with only a few exceptions such as A-minor *BA.A.1029.|BA.U.19.-BA.A.868.* from *6ERI* PDB entry that stabilizes an H-knot. 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 functional groups: 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 functional groups of coaxial stacking stabilization and H-knot stabilization respectively.

Thus, the structural classes according to the proposed classification reliably define the function of A-minors. We suggest it can be successfully used for automatic functional annotations.

## 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 usually contains a bulged out base. All 34 of such cases belong to 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.

The number of adenines involved in A-minors of across-bulged motifs varies from 1 to 3 among different organisms. The bulged bases take part in cross-strand base stacking or form A-minors, G-minors, and other N-minors, or RNA-protein interactions. The majority of across-bulged motifs also include a base-triple (see Figure 4A).

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 only architecture of A-minor clusters

We examined all 389 A-minor clusters of size (2, 2). The majority (94%, 366 cases) of such clusters are of A-patch architecture (Nissen et al. 2001), i.e. are formed by a stretch of stacked 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 rest 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.

## Local A-minor motifs can be reliably predicted with machine learning

We applied a machine learning approach to predict A-stems for a given RNA chain using the sequence and the RNA secondary structure information. The cross-validation on the dataset of A-stems showed only 25% PR BPE that was far from acceptable quality (see Figure 6A and Supplementary Figure S1). The feature importances of the *relss* and *localss* groups combined reached the value of 51.1%. Thus, the features representing the local RNA secondary structure context were as significant as those representing the sequence context.

We found three local types (*HP-LC, IC-LC, IP-LC*) of A-stems that could be predicted with notably higher quality (50% PR BPE, see Figure 6B and Supplementary Figure S2). The feature importances of the *relss* and *localss* groups combined reached the value of 48.4%. Of note, the positive rate for the subset of the three local A-stem types was 5.44% which is much greater than 0.19% of the entire dataset.

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

# DISCUSSION

In this work, we proposed strict definitions to describe A-minors, one of the most important types of RNA tertiary structure motifs. The definitions of the A-minor interaction, the A-minor motif, and the A-minor cluster were used to annotate the motifs in experimentally determined RNA spatial structures from 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 for the most frequent classes of A-minors the class reliably defines the A-minor function. Such correspondence could be used to automatically annotate a motif's function 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 often contains a bulged base 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. Interestingly, 11nt-loop has its own two consecutive adenines and we hypothesize there is an on-off switch mechanism depending on the presence of the GAAA hairpin. However, we were unable to find suitable examples in any RNA structure from PDB.

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 formation of N-minors is also suggested to be included in consideration for evolutionary analyses dealing with point mutations.

We showed that a group of three local A-minor classes can be predicted using the machine learning algorithm with reasonably acceptable quality. We found none of the long-range classes of A-minors that could be predicted with the quality close to that of the local ones. We conclude that the prediction of long-range motifs is a more difficult task than in the case of local motifs due to a rapidly growing combinatorial complexity: the numbers of potential and real local motifs both grow proportionally to the length of RNA chain *N*, but the number of potential long-range motifs grows proportionally to *N2* whereas the number of the real ones is proportional to *N*.

# 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 in good agreement with the known functions of A-minor motifs. We also showed that the local A-minors from internal loops can not only stabilize kink-turn motifs but also form previously not described 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. The results of the classification were also used to formulate the A-minor prediction problem. Using a machine-learning algorithm we showed that the particular local classes of A-minors can be reliably predicted using the sequence and secondary structure information. Thus, we show that the proposed classification can be successfully used both to automatically annotate functions of A-minor motifs by their structural context and to predict A-minors of particular local classes. Also, the use of the classification is not limited to A-minors as it may be applied to any other kind of RNA tertiary motifs.

# MATERIALS AND METHODS

## Description of RNA secondary structure elements

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). 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 unpaired regions (*threads*) confined by a stem. Each loop is ascribed 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.

## 

## 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 base pair, where the adenine faces the minor groove of the base pair and forms H-bonds. A-minors with the O2’ atom of the adenine involved in H-bonds belong to the geometric types I and II (Nissen et al. 2001). Otherwise, A-minors are designated with the type X (eXtended, see (Lu et al. 2015)). 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 = { vi = (Ai, Li, Ri) }* is the set of A-minor interactions annotated with DSSR and *E = { eij = (vi, vj) }* is the set of edges between them. *(vi, vj) ϵ E* if either *Ni = Nj* or *Ni* and *Nj* are stacked, where *Ni ϵ {Ai, Li, Ri}* and *Nj ϵ {Aj, Lj, Rj}*. 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 *Ai* and *Aj* or two different base pairs *(Li, Ri)* and *(Lj, Rj)*. 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

All A-minors were classified with respect to the RNA secondary structure elements containing 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 (from one stem or one loop, from adjacent stem and loop, from distant elements).

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 to each other within the secondary structure. Overall, the A-minor is classified as having type HC-IC-HPIC-LR-LR-SM.

## 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 *NidNj* relationships, where *Nk* is *A*, *L*, or *R* of the corresponding A-minor *vk* and *d* *ϵ {“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).

## Machine learning model

We will formulate the A-minor prediction problem as the binary classification problem. To choose an object of classification we examined the annotated A-minors from the known RNA structures. We found that the annotation of A-minors contains a notable number of false negatives and false positives. The majority of such cases are due to adenines located evenly between two consecutive base pairs such that it is unclear with which particular base pair the adenine forms the A-minor interaction. Considering that more than 60% of all annotated A-minors belong to A-minor clusters and that nearly 90% of all A-minors include a base pair that belongs to some stem, we decided to lower the resolution and choose a pair of a stem and a stretch of unpaired adenines as the object of the classification (we called it an *A-stem*). We considered all A-stems and trained a model to predict if a particular A-stem forms A-minors (see Figure 10).

The representative set of structures consisted of 130 RNA chains containing A-minor interactions. From these chains, we have manually selected a non-redundant set of 44 RNA chains (see Supplementary Table S4) to avoid possible over-training. The resulting set of A-stems included 347 positive A-stems and 183298 negative A-stems (0,19% positive rate, the set can be found at <https://github.com/febos/urs_aminors>).

Each A-stem was annotated with 288 features belonging to four groups: *localseq* (local sequence context), *relseq* (distance features in terms of sequence), *localss* (local secondary structure context), and *relss* (distance features in terms of secondary structure). A detailed description of all the features is provided in Supplementary Table S5 and at <https://github.com/febos/urs_aminors>.

A-stems were classified in terms of RNA secondary structure in a similar manner as A-minors. Thus, an A-stem containing A-minor interactions of type *IC-S-S-LC-LC-SM* has been attributed to type *IC-LC*.

To solve the A-stems prediction problem we used the RandomForest algorithm (Liaw and Wiener 2002) from the scikit-learn python library (Pedregosa et al. 2011). The experiments have been carried out using a group K-fold cross-validation method (Rodriguez et al. 2009) with *K = 10*. The following parameters have been chosen via the grid search method: *n\_estimators = 100, max\_depth = 50, min\_samples\_leaf = 5, max\_features = 50, class\_weight='balanced', criterion='entropy'.*

To assess the quality of the binary classification results we used the precision-recall break-even point metrics (PR BPE), i.e. the value of precision at such threshold that the equality of precision and recall values is achieved.

**SUPPLEMENTARY INFORMATION**

Suplemental\_Text\_S1.pdf - **Supplementary Text S1**. A generalized description of arbitrary RNA secondary structure. List of definitions.

Supplemental\_Figure\_S1.pdf - **Supplementary Figure S1**. Cross-validation results for the entire dataset of A-stems. The dataset included 183298 negatives and 347 positives (0.19% positive rate). (A) Normalized distributions of predicted probabilities for an A-stem to belong to positives. (B) Precision-Recall curve. (C) ROC curve (AUC = 0.86). (D) Precision-Recall metrics with threshold = 0.2. Each circle denotes the precision-recall values for a given RNA chain. The size of a circle is proportional to the number of A-minors.

Supplemental\_Figure\_S2.pdf - **Supplementary Figure S2**. Cross-validation results for the A-stem dataset of types IC-LC, HP-LC, & IP-LC. The dataset included 1043 negatives and 60 positives (5.44% positive rate). (A) Normalized distributions of predicted probabilities to be positive. (B) Precision-Recall curve. (C) ROC curve (AUC = 0.88). (D) Precision-Recall metrics for each RNA chain separately with threshold = 0.2. The size of a circle is proportional to the number of A-minors.

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**. List of 2431 A-minor interactions.

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 A-stems. List of descriptions

**ACKNOWLEDGEMENTS**

Not applicable.

**FUNDING**

Not applicable.

# 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, 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

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

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

**TABLES**

Table 1. Recurrent A-minor motifs and their functions.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Description** | **#A-minors** | **#A-minor**  **motifs** | **Function / Tertiary motif** |
| HC-S-S-  LR-LR-SM | **Base-pair** belongs to a stem  **Adenine** belongs to a classical hairpin  **Long-range interaction** | 547 | 374 | Tetraloop-receptor recognition (Geary et al. 2008, Fiore and Nesbitt 2013, Wu et al. 2012) |
| IC-S-S-  LR-LR-SM | **Base-pair** belongs to a stem  **Adenine** belongs to a classical internal loop  **Long-range interaction** | 352 | 180 | UAA/GAN internal loop motif (Lee et al. 2006); Decoding of the codon:anticodon base pairs (Lescoute and Westhof 2006) |
| JC-S-S-  LC-LC-SM | **Base-pair** belongs to a stem  **Adenine** belongs to a classical junction  **Local interaction** | 276 | 201 | Coaxial stacking stabilization (Lescoute and Westhof 2006, Geary et al. 2011) |
| IC-S-S-  LC-LC-SM | **Base-pair** belongs to a stem  **Adenine** belongs to a classical internal loop  **Local interaction** | 147 | 100 | Kink-turn stabilization (Klein et al. 2001, Razga et al. 2005);  Across-bulged motif |
| HP-S-S-  LC-LC-SM | **Base-pair** belongs to a stem  **Adenine** belongs to a pseudoknotted hairpin  **Local interaction** | 112 | 51 | H-knot stabilization (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. A-minor column represents one A-minor interaction per motif.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Unique**  **motif** | **PDB:**  **CHAIN** | **A-minor** | | | **Adenine(s)**  **thread** | **Thread with**  **bulged bases** | **Bulged bases**  **involved in** | **Molecule** |
| **Adenine** | **Base pair** | |
| I | 4yaz:R | A78 | C7 | G77 | **-AAa-** | **-Ca--** |  | Cyclic di-GMP-I  riboswitch |
| II | 6dtd:C | A30 | G11 | U28 | **-A---** | **-C---** | RNA-protein interaction | Cas13b sgRNA |
| III | 6eri:Ax | A58 | A29 | U55 | **uAAa-** | **gAAcc** | A-minor | 5S rRNA |
| IV | 6eri:BA | A1095 | C1078 | G1093 | **aA-cc** | **uU-g-** | Cross-strand  base stacking | 16S rRNA |
| V | 6eri:BA | A345 | C342 | G363 | **cAAu-** | **-G-a-** | G-minor | 16S rRNA |
| VI | 6eri:AA | A1343 | G1340 | C1354 | **cAAg-** | **cA---** | Cross-strand  base stacking | 23S rRNA |
| VII | 6eri:AA | A876 | G874 | C921 | **cA-cu** | **aAA--** | A-minor | 23S rRNA |
| VIII | 6eri:AA | U2699 | C2697 | G2745 | **cUCu-** | **aA---** | Cross-strand  base stacking | 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 rightmost column. Pie radii are proportional to the natural logarithm of the absolute numbers.

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

Figure 3. 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 (B) motif VIII, a representative case from 6ERI entry (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, A-minors: S2.A.996.|S2.C.674.-S2.A2M.1031., S2.A.997.|S2.G.673.-S2.C.1032.) (B) A-patch formed by a cross-strand adenine stack, LSU rRNA (PDB ID: 5TBW, A-minors: 1.A.2696.|1.C.2630.-1.G.2648., 1.A.2758.|1.U.2629.-1.A.2649.)

Figure 6. Precision-Recall curves representing the cross-validation results on (A) the entire dataset of A-stems (183298 negatives and 347 positives, 0.19% positive rate), and (B) the subset of three local types of A-stems: IC-LC, HP-LC, & IP-LC (1043 negatives and 60 positives, 5.44% positive rate). The line of precision-recall break-even points is shown as blue dots.

Figure 7. AUA A-patch from LSU rRNA, PDB ID: 5TBW. A-minors: 1.A.3106.|1.C.2893.-1.G.2908. (in orange), 1.A.3129.|1.A.2892.-1.U.2909. (in gray). (A) Top view. (B) Side view. Uracil 1.U.3105. 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 an A-stem 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.