Supplementary Information

Text S1. Generalized description of arbitrary RNA secondary structure**.** List of definitions*.

*Baulin, E., Yacovlev, V., Khachko, D., Spirin, S., & Roytberg, M. (2016). URS DataBase: universe of RNA structures and their motifs. *Database*, 2016.
Available online at http://urs.lpm.org.ru/struct.py?where=3#def

1. Base Pairs, Stems and Links

We consider an RNA molecule as a sequence of nucleotides i.e. as a sequence of letters in the alphabet *{A, C, G, U}*. Nucleotides in a molecule are indexed from 5' - to 3'-end with integers from *1* to *L*; here *L* is the sequence's length.

A **Base Pair** is a pair of nucleotides *(i, j)*, where $i < j$, which forms hydrogen bonds. We consider not only pairs of complementary nucleotides (*A-U* and *G-C* pairs, also known as Watson-Crick pairs) and *G-U* pairs (Wobble pairs), but also non-canonical pairs.

We say that a base pair *(p, q)* is a base pair of **level** 0 if it does not have conflicts with any base pair *(m, n)* such that $m < p$. All base pairs in pseudoknot-free secondary structures has level 0.

A base pair *(p, q)* has **level** *K* if there are pairs $(m_0 , n_0),..., (m_{K-1} , n_{K-1})$ such that for all $i = 0,..., K-1$

- $m_i < p$;

- (m_i , n_i) has level *i*;

- (m_i , n_i) has a conflict with *(p, q)*.

A **Stem** (Standard Stem) is a longest sequence of base pairs of the form *(i, j), (i+1, j-1),..., (i+k, j-k)* such that

1) $k \geq 1$;

2) $i+k < j-k$;

3) All pairs $(i + x, j - x)$, where $x = 0, ..., k$, form base pairs, and all of them are Watson-Crick pairs (WC pairs) or Wobble pairs (WB pairs).

Remark. The URS database contains information of other types of stems, closed stems and free stems. All definitions below (wings, threads, etc.) are related to standard stems. However, they can be applied to stems of arbitrary type.

Pair *(i, j)* is called an **external pair** of the stem or a **face**. Pair *(i + k, j - k)* is called an **internal pair** of the stem.

For a stem (of any type) *(i, j), (i+1, j-1),..., (i+k, j-k)* the fragment *[i, i + k]* of an RNA chain is called a **left wing** of the stem, and the fragment *[j - k, j]* is called a **right wing**.

A **Thread** (or unpaired region) is a fragment *[i, j]*, such that

1) There is no base pair *(k, t)*, such that $i \le k \le j$ or $i \le t \le j$.

2) There are base pairs containing nucleotides *i-1* and *j+1*.

For technical reasons we allow threads of zero length between two wings of stems; the zero length thread is denoted by *[i + 1, i]*, where *i* is the index of the last nucleotide of the previous wing.

A **Tower** is a set of *N* stems of any type such that their wings are located on the chain in the following order: *1L,2L,...,NL,NR,...,2R,1R*, where *iL* is the left wing of *i*-th stem and *iR* is the right wing of *i*-th stem.

Base pairs *(m, n)* and *(p, q)* have a **conflict** if $m < p < n < q$ or $p < m < q < n$. A base pair has a **conflict with a stem** if it has a conflict with a base pair from the stem.

A **Link** is a base pair that does not belong to any stem. A link is **fully coordinated** (or **coordinated**) if it does not have conflicts. A link is **stem-coordinated** (or **weakly coordinated**) if it does not have conflicts with stems but may have conflicts with other links. A link is **stem-independent** (or **independent**) if it has a conflict with a stem.

2. Elementary Closed Regions, Pseudoknots, Signatures and Descriptions

An **Elementary Closed Region** (ECR) is a minimal region *[i, j]* where i < j, such that:

1) There is no base pairs *(k, l)* such that $(i \le k \le j;\ l > j)$ or $(k < i;\ i \le l \le j)$;

2) There is no *l* such that $i < l < j$ and both regions *[i,...,l]* and *[l+1,...,j]* satisfy the condition 1);

3) There are base pairs *(i, k)* and *(l, j)*; possibly, $k = j$ and $i = l$.

A pair of positions *(i,j)* is called a **face** of the ECR *[i,j]*. Note, that if the positions *i* and *j* are paired and belong to a stem then the face of the ECR coincides with the face of the stem.

An ECR *[k, l]* is a **sub-ECR** of an ECR *[i, j]* if $i < k < l < j$ and there are no other ECR *[m, n]* such that $i < m < k < l < n < j$.

An ECR is a **pseudoknot** (or **pseudoknotted**) if base pairs from its stems have conflicts. Otherwise ECR is called **pseudoknot-free** or **classical**.

The classification of pseudoknots used in URSDB is based on the notion of **signature**. The classification is close to topological classification [Andersen JE, Penner RC, Reidys CM, Waterman MS. Topological classification and enumeration of RNA structures by genus. *J Math Biol. 2013 Nov;67(5)*]. The main difference between the classifications is that our classification takes into account only stems.

Consider all stems of an ECR and index them with latin letters according to positions of their wings from 5'- to 3'-end. The left wing of the stem will be denoted with a small letter, e.g. **a**, the right wing will be denoted with a capital letter, e.g. **A**, and the stem will be denoted with two letters, e.g. **aA**.

A **full signature** of an ECR is a sequence of its wing letters given according to the wings positions on the chain from 5'- to 3'-end.

Example 1. See Fig.D1a,b. Let ECR *[10, 70]* contains three stems, *([10, 15]; [65, 70])*, *([20, 25]; [45, 50])*, *([30, 35]; [55, 60])*, here *[10, 15]* and *[65, 70]* are wings of the stem *([10, 15]; [65, 70])*, etc. Then the stem *([10, 15]; [65, 70])* is **aA** stem, stem *([20, 25]; [45, 50])* is **bB** stem, and stem *([30, 35]; [55, 60])* is **cC** stem. The full signature of the ECR is **abcBCA**. A fragment *[20, 60]* is a sub-ECR of the initial ECR.

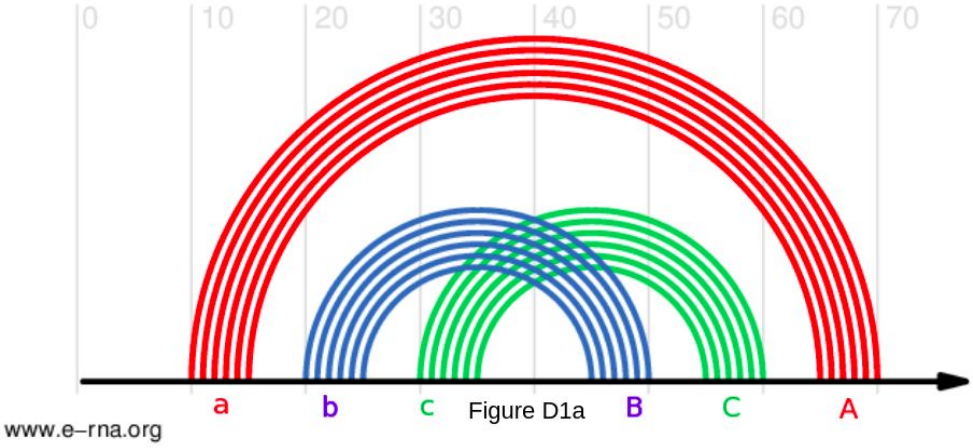

Figure D1a

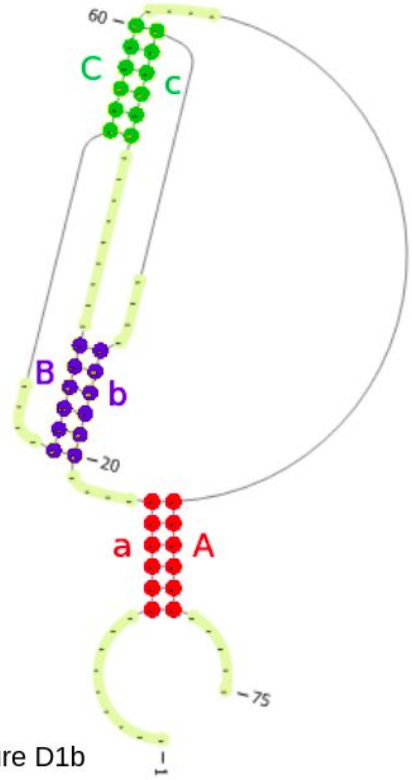

Figure D1b

Figure D1a. Positions of wings within the ECR from the Example 1; the stem **aA** is given in red; the stem **bB** is given in blue, and the stem **cC** is given in green. A fragment *[20, 60]* is a sub-ECR of the initial ECR.
Figure D1b. Schematic representation of the secondary structure of the stem from the Example 1.

Example 2. See Fig.D2a,b. Let ECR contains four stems, *([10, 15]; [70, 75])*, *([20, 25]; [50, 55])*, *([30, 35]; [40, 45])*, *([60, 65]; [80, 85])*, here *[10, 15]* and *[70, 75]* are wings of the stem *([10, 15]; [70, 75])*, etc. Then the stem *([10, 15]; [70, 75])* is **aA** stem, stem *([20, 25]; [50, 55])* is **bB** stem, stem *([30, 35]; [40, 45])* is **cC** stem, and stem *([60, 65]; [80, 85])* is **dD** stem. The full signature of the ECR is **abcCBdAD**. The fragment *[20, 55]* is a sub-ECR of the initial ECR, and the fragment *[30, 45]* is a sub-ECR of the ECR *[20, 55]*.

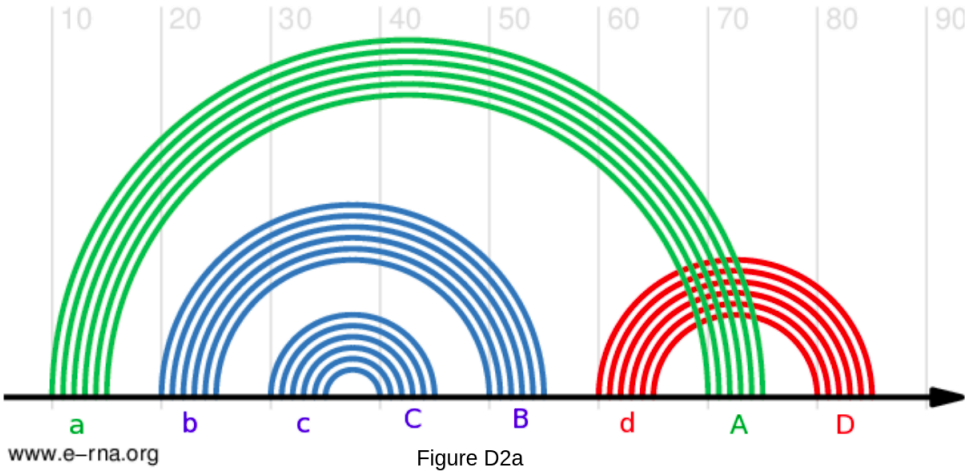

Figure D2a

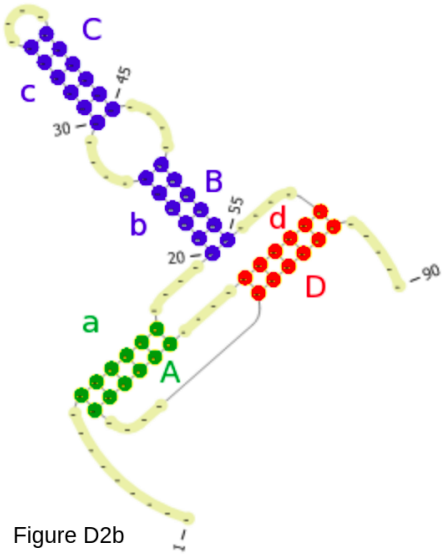

Figure D2b

Figure D2a. Positions of wings within the ECR from the Example 2; the stem **aA** is given in green; the stems **bB** and **cC** are given in blue, and the stem **dD** is given in red. A fragment *[20, 55]* is a sub-ECR of the initial ECR, and a fragment *[30, 45]* is a sub-ECR of the ECR *[20, 55]*.
Figure D2b. Schematic representation of the secondary structure of the stem from the Example 2.

Example 3. See Fig.D3a,b. Let ECR contains six stems, *([2, 7]; [90, 95])*, *([10, 15]; [80, 85])*, *([20, 25]; [50, 55])*, *([30, 35]; [40, 45])*, *([60, 65]; [120, 125])*, *([70, 75]; [110, 115])*, here *[2, 7]* and *[90, 95]* are wings of the stem *([2, 7]; [90, 95])*, etc. Then the stem *([2, 7]; [90, 95])* is **aA** stem, stem *([10, 15]; [80, 85])* is **bB** stem, stem *([20, 25]; [50, 55])* is **cC** stem, stem *([30, 35]; [40, 45])* is **dD** stem, stem *([60, 65]; [120, 125])* is **eE** stem, and stem *([70, 75]; [110, 115])* is **fF** stem. The full signature of the ECR is **abcdDCefBAFE**. The fragment *[20, 55]* is a sub-ECR of the initial ECR, and the fragment *[30, 45]* is a sub-ECR of the ECR *[20, 55]*.

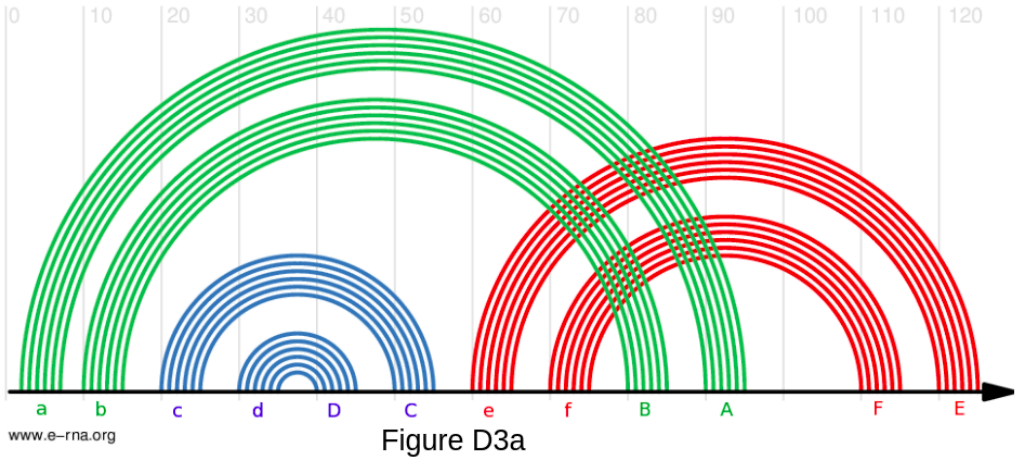

Figure D3a

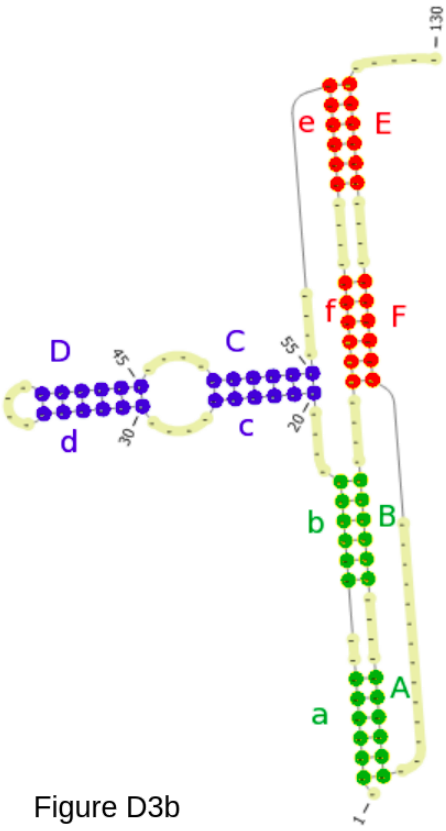

Figure D3b

Figure D3a. Positions of wings within the ECR from the Example 3; the stems **aA** and **bB** are given in green; the stems **cC** and **dD** are given in blue, and the stems **eE** and **fF** are given in red. A fragment *[20, 55]* is a sub-ECR of the initial ECR, and a fragment *[30, 45]* is a sub-ECR of the ECR *[20, 55]*.
Figure D3b. Schematic representation of the secondary structure of the stem from the Example 3.

An **upper signature** of the ECR is a string obtained from its full signature by

1) deletion of fragments corresponding to sub-ECRs;

2) "renaming" of the letters preserving their order to obtain a string containing all letters of a proper beginning of the alphabet.

Example 4.

The upper signature of the ECR from the Example 1 is **aA**; the fragment **bcBC** corresponding to the sub-ECR *[20, 60]* was deleted from the full signature **abcBCA**.

The upper signature of the ECR from Example 2 is **abAB**. Firstly, the fragment **bcCB** corresponding to the sub-ECR *[20, 55]* was deleted from the full signature **abcCBdAD**; the obtained string is **adAD**. Then we replace **d** and **D** with **b** and **B** obtaining **abAB**.

Analogously, the upper signature of the ECR from Example 3 is **abcdBADC**.

Stems **xX**, **yY**, **zZ**,... are connected within an upper signature if both the word **xyz...** and inverted word **...ZYX** are subwords of the upper signature.

A **signature** (or a reduced signature) of the ECR is a string obtained from its upper signature by

1) deletion all letters except **x** and **X** (the first letter of the left part and the last letter of the right part) corresponding to chains of connected stems;

2) "renaming" of the letters preserving their order to obtain a string containing all letters of a proper beginning of the alphabet.

Example 5.

Signatures of ECRs from examples 1 and 2 coincide with their upper signatures. The signature of the ECR from Example 3 is **abAB** and coincides with the signature of ECR from example 2.

Typical signatures

a) H-knot: **abAB**;

b) Kissing Loops: **abAcBC**;

c) Triple knot: **abcABC**.

3. Stems and Loops

Here and below, we assume a fixed chain with a given RNA secondary structure on it. The chain may be considered as an alternating sequence of threads and wings. To preserve generality, we assume that before the first and after the last nucleotide of the chain wings of "external stem" are added.

Each stem is associated with the part of the chain that is **internal** to it - the part between the end of the left wing and the start of the right wing of this stem, in other words, between the nucleotides that create the internal pair of the stem. For a fictitious external stem the internal part is the entire original RNA sequence.

Let *H* be a stem and *(i, j)* its internal pair.

Definition 1. The position of the chain *t* is *internal* to stem *H* (synonym: *lies inside H*), if $i < t < j$. Fragment of chain is *internal* to stem *H* (synonym: *lies inside H*), if all the positions are internal

to stem *H*. stem *H1 lies inside* the stem *H* (*is internal* to *H*), if all the positions of her wings are internal to *H*.

Definition 2. The position of the chain *t* **belongs to** the stem *H*, if it is internal to the *H* and there is no stem *H1*, lying inside *H*, such that $x < t < y$, where *(x, y)* is the external pair (face) of *H1*.

Definition 3. **Loop** of the stem *H* is the set of all positions that belong to stem *H*.

Each position that is not included into a bond belongs to at least one loop - normal or external. If a position of a thread (wing) belongs to a loop, then the entire thread (wing) belongs to this loop.

If the structure is pseudoknot-free, each loop in terms of Definition 3 is a loop in terms of the Nearest Neighbour Model and vice versa. In addition, each thread belongs to exactly one loop (possibly external), and no wing belongs to any loop. For pseudoknotted structures both of these properties do not hold.

4. Loop Structure

Definition 4. Let *H* be a stem and *(u, v)* its internal pair. Region *[i, j]* is called *H-related ECR* (in general case **stem-related ECR** or **S-ECR**) if

1) *[i, j]* lies inside *H*;

2) There is no such pair *(k, t)* that $(i \le k \le j < t < v)$ or $(u < k < i \le t \le j)$;

3) There are pairs *(i, k)* and *(t, j)*, where $k \le j$; $i \le t$;

4) There is no *l* such that $i < l < j$ and both regions *[i,...,l]* and *[l+1,...,j]* satisfy the conditions 1) - 3);

5) There is no other than *[i, j]* region *[i', j']* such that $i' \le i < j \le j'$ and the region *[i', j']* satisfies the conditions 1) - 4).

Pair *(i, j)* is called the *face* of the stem-related ECR.

Statement 1. Let $Z = [f, g]$ be an H-related ECR; *(u, v)* is an internal pair of stem *H*. Then:

1) The entire section *Z* lies inside *H*.

2) A wing lies entirely inside *Z* or lies entirely outside *Z*.

3) *Z* starts with a left wing of a stem *H1*, lying inside *H*, and ends with a right wing of a stem *H2*, lying inside *H*.

4) If *H1 = H2* is one and the same stem, face *(s, t)* of region *Z* is the face of a stem. Otherwise, *s* is the beginning of the left wing of *H1* and *t* is the end of the right wing of *H2*.

Proof - follows from the Definition 4 and the fact that the wings do not overlap.

Definition 5. Let *Z* be a stem-related ECR. Region *Z* is called **simple** if its face is the face of a stem and called **pseudoknotted** otherwise. Pseudoknotted S-ECRs are also called **blocks**.

Statement 2. Let *H* be a stem and *(u, v)* its internal pair. Then:

1) There are no two H-related ECRs that overlap.

2) Let position *t* lie inside *H*. The position *t* does not belong to *H* if and only if *t* lies inside H-related ECR *Z* (i.e., lies inside *Z*, but not in its face).

Proof - follows from Definitions 1, 2 and 4.

Definition 6. Let *H* be a stem and *(u, v)* its internal pair. Let $(s_1, t_1), ..., (s_n, t_n)$ - faces of all of the H-related ECRs; $s_1 < t_1 < ... < s_n < t_n$. For convenience let $t_0 = u$; $s_{n+1} = v$. Suppose *k* - integer; $1 \le k \le n+1$. Then the *k*-th **side** of the loop of *H* is a fragment $[t_{k-1} + 1, s_k - 1]$.

If $s_k = t_{k-1} + 1$, *k*-th side of the loop of *H* has a length of zero.

Statement 3. Let *H* be a stem and *(u, v)* its internal pair. Let $(s_1, t_1),..., (s_n, t_n)$ - faces of all of the H-related ECRs; $s_1 < t_1 < ... < s_n < t_n$. For convenience let $t_0 = u$; $s_{n+1} = v$. Then the loop of *H* is the union of all the faces of H-related ECRs and located among them sides.

Proof - follows from the Statement 2.

Statement 4. Let *H* be a stem and *(u, v)* its internal pair and a position *x* belongs to side *(t, s)* of the loop of *H*. Then:

1) The position *x* is not involved in any bond or belongs to wing of a stem *H'*, the other wing of which lies outside *H*.

2) If *x* belongs to a thread (wing), then this entire thread (wing) belongs to the same side of the loop of *H*.

Proof - follows from definition of side and the fact that wings do not overlap.

Statements 3 and 4 describe the possible structures of loops. Note that in case of pseudoknot-free structures, all closed regions are simple and each side consists of a single thread. Therefore, we give the following definition:

Definition 7. A loop is called **classical** if it does not contain wings and faces of blocks. A loop is called **isolated** if it does not contain wings and called **pseudoknotted** otherwise.

A stem is called **pseudoknotted** if its loop is pseudoknotted.

Let us apply the classification of loops on Matthews-Turner to the introduced generalization based on the number of faces included in the loop. Note that in this case, the faces can be both faces of the stems (in other words - simple, closed regions) and faces of the blocks (complex, closed regions).

Definition 8. A loop will be called **hairpin** if it does not contain faces and therefore has a single side. A loop will be called **internal** if it contains exactly one face and therefore has two sides. A loop will be called **multiple junction** if it contains more than one face and therefore more than two sides.

An internal loop is called **bulge** if one of its sides has a length of zero.

This classification covers both normal and external (belonging to the "external" stem) loops.

Supplementary Figures

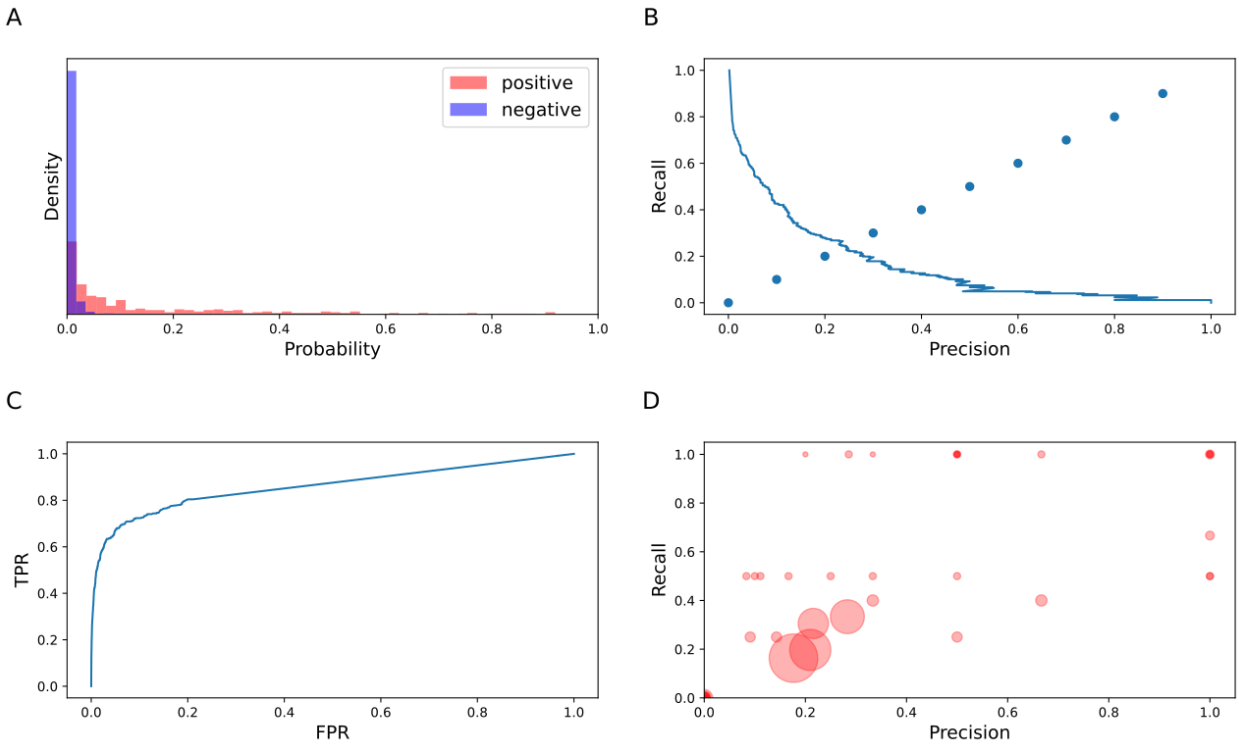

Figure S1. Cross-validation results for the entire dataset of A-stems. The dataset included 183298 negatives and 347 positives (0.19% positive rate). (A) Normalized distributions of predicted probabilities for an A-stem to belong to positives. (B) Precision-Recall curve. (C) ROC curve (AUC = 0.86). (D) 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.

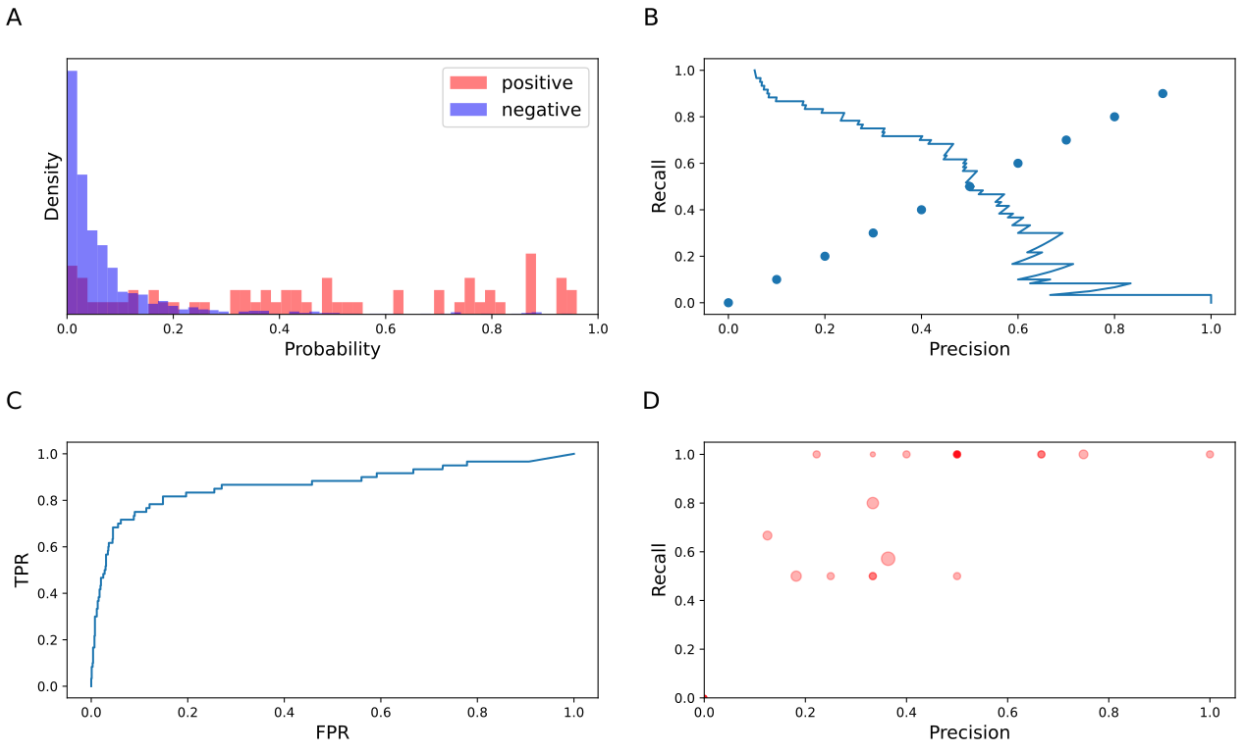

Figure S2. Cross-validation results for the A-stem dataset of types IC-LC, HP-LC, & IP-LC. The dataset included 1043 negatives and 60 positives (5.44% positive rate). (A) Normalized distributions of predicted probabilities to be positive. (B) Precision-Recall curve. (C) ROC curve (AUC = 0.88). (D) Precision-Recall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

Supplementary Tables

Supplementary Table S1. List of 2431 A-minor interactions.
[The table is available as an external xlsx file]

Supplementary Table S2. List of the edges forming 626 A-minor clusters.
[The table is available as an external xlsx file]

Supplementary Table S3. Nonredundant set of 44 RNA chains.
[The table is available as an external xlsx file]

Supplementary Table S4. Features of A-stems. List of descriptions.
[The table is available as an external xlsx file]

Supplementary Table S5.
List of across-bulged motifs found in the representative set of PDB structures.

PDB entry	A-minor (one per motif)			Adenine(s) thread is	Unique motif	Adenine(s) thread	Thread with bulged bases	Bulged bases involved in	Molecule	Organism
	Adenine	Base pair								
4yaz	R.A.78.	R.C.7.	R.G.77.	3'-thread	I	--AA-a-	----Ca---		Cyclic di-GMP-I riboswitch	geobacter
6dtd	C.A.30.	C.G.11.	C.U.28.	3'-thread	II	--A----	----C----	Protein	Cas13b sgRNA	prevotella buccae
4v9f	9.A.56.	9.U.28.	9.A.54.	3'-thread	III	-uAA-g-	guugCC---		5S rRNA	haloarcula marismortui
5dm6	Y.A.59.	Y.C.32.	Y.G.56.	3'-thread	III	--AAA--	---gAAc--	A-minor	5S rRNA	deinococcus radiodurans
5fdv	1B.A.57.	1B.A.29.	1B.U.55.	3'-thread	III	--AAA--	---gAAc--	A-minor	5S rRNA	thermus thermophilus
5j7l	DB.A.57.	DB.A.29.	DB.U.55.	3'-thread	III	--AAA--	---gUCc--	U-minor & C-minor	5S rRNA	escherichia coli
5t5h	D.A.55.	D.A.27.	D.U.52.	3'-thread	III	-uAA-g-	------c--		5S rRNA	trypanosoma cruzi
5tbw	AS.A.55.	AS.C.28.	AS.G.52.	3'-thread	III	-uAA-g-	------c--		5S rRNA	saccharomyces cerevisiae
6eri	Ax.A.58.	Ax.A.29.	Ax.U.55.	3'-thread	III	-uAA-a-	---gAAcc-	A-minor	5S rRNA	spinacea oleracea
6hma	B.A.56.	B.A.27.	B.U.53.	3'-thread	III	-uAA-g-	---gU-cac		5S rRNA	staphylococcus aureus
6qzp	L7.A.55.	L7.G.27.	L7.C.52.	3'-thread	III	-uAA-g-	------c--		5S rRNA	homo sapiens
4y4o	1a.A.1146.	1a.C.1128.	1a.G.1144.	3'-thread	IV	-cA--cu	---uU-g--	Cross-strand stack	16S rRNA	thermus thermophilus
6eri	BA.A.1095.	BA.C.1078.	BA.G.1093.	3'-thread	IV	-aA--cc	---uU-g--	Cross-strand stack	16S rRNA	spinacea oleracea
4y4o	1a.A.373.	1a.C.370.	1a.G.391.	5'-thread	V	-cAA-u-	----G-a--	G-minor	16S rRNA	thermus thermophilus
5j7l	AA.A.373.	AA.C.370.	AA.G.391.	5'-thread	V	-cAA-u-	----G-a--	G-minor	16S rRNA	escherichia coli
5ngm	Aa.A.381.	Aa.C.378.	Aa.G.399.	5'-thread	V	-cAA-u-	----G-a--	G-minor	16S rRNA	staphylococcus aureus
6eri	BA.A.345.	BA.C.342.	BA.G.363.	5'-thread	V	-cAA-u-	----G-a--	G-minor	16S rRNA	spinacea oleracea
4v88	A6.A.445.	A6.C.442.	A6.G.462.	5'-thread	V	-cAA-u-	----G-a--	G-minor	18S rRNA	saccharomyces cerevisiae
6az1	1.A.488.	1.C.485.	1.G.510.	5'-thread	V	-cAA-u-	----G-a--	G-minor	18S rRNA	leishmania donovani
6qzp	S2.A.493.	S2.C.490.	S2.G.510.	5'-thread	V	ccA--cu	----G-ag-	G-minor	18S rRNA	homo sapiens
4y4o	2A.A.1322.	2A.G.1319.	2A.C.1333.	5'-thread	VI	-cAA-u-	---aG----	Cross-strand stack	23S rRNA	thermus thermophilus
5dm6	X.A.1322.	X.G.1319.	X.C.1333.	5'-thread	VI	-gAA-g-	---cG----	Cross-strand stack	23S rRNA	deinococcus radiodurans
5j7l	DA.A.1322.	DA.C.1319.	DA.G.1333.	5'-thread	VI	-cAA-c-	---gG----	Cross-strand stack	23S rRNA	escherichia coli
6eri	AA.A.1343.	AA.G.1340.	AA.C.1354.	5'-thread	VI	-cAA-g-	---cA----	Cross-strand stack	23S rRNA	spinacea oleracea
6hma	A.A.1359.	A.G.1356.	A.C.1370.	5'-thread	VI	-gAA-g-	---cG----	Cross-strand stack	23S rRNA	staphylococcus aureus
6qzp	L5.A.2418.	L5.U.2415.	L5.A.2429.	5'-thread	VI	-gAA-c-	---gA----	Cross-strand stack	28S rRNA	homo sapiens
4v9f	0.A.961.	0.G.958.	0.C.1008.	5'-thread	VII	cgA--cc	---aAA---	A-minor	23S rRNA	haloarcula marismortui
4y4o	2A.A.866.	2A.G.864.	2A.C.912.	5'-thread	VII	-cA--cu	---aAA---	A-minor	23S rRNA	thermus thermophilus
5dm6	X.A.866.	X.G.864.	X.C.912.	5'-thread	VII	-cA--cu	---aAA---	A-minor	23S rRNA	deinococcus radiodurans
6eri	AA.A.876.	AA.G.874.	AA.C.921.	5'-thread	VII	-cA--cu	---aAA---	A-minor	23S rRNA	spinacea oleracea
6hma	A.A.911.	A.G.909.	A.C.957.	5'-thread	VII	-cA--cu	---aAA---	A-minor	23S rRNA	staphylococcus aureus
5tbw	1.A.1002.	1.A.998.	1.U.1050.	5'-thread	VII	cgAA-u-	---aAA---	A-minor	25S rRNA	saccharomyces cerevisiae
6qzp	L5.A.1742.	L5.A.1738.	L5.U.1790.	5'-thread	VII	cgAA-u-	---aAA---	A-minor	28S rRNA	homo sapiens
6az3	1.A.1053.	1.A.1049.	1.U.1101.	5'-thread	VII	caAA-u-	---aAA---	A-minor	rRNA alpha	leishmania donovani