Mathematical Biology and Bioinformatics
2019. V. 14. № S. P. t37– t43. doi: 10.17537/2019.14.t37

The translation into English of the original article published in Russian
Baulin E. et al. Mathematical Biology and Bioinformatics. 2018. V. 13. № 2. P. 526–533
doi: 10.17537/2018.13.533

========== TRANSLATIONS OF PUBLISHED ARTICLES ==========

УДК: 577.214.5

Diverse RNA pseudoknots exist for short stems only

Baulin E.*[1,2], Korinevskaya A.[3], Tikhonova P.[3], Roytberg M.[1,2,3]

[1]*Laboratory of Applied Mathematics, 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*
[2]*Faculty of Innovations and High Technology, Moscow Institute of Physics and Technology (State University), Dolgoprudny, Moscow Region 141700, Russia*
[3]*Faculty of Computer Science, National Research University Higher School of Economics, Moscow 101000, Russia*

Abstract. RNA secondary structure prediction including pseudoknotted structures of arbitrary types is a well-known NP-hard problem of computational biology. By limiting the possible types of pseudoknots the problem can be solved in polynomial time. According to the empirical thermodynamic parameters, the formation of a stem starts to decrease free energy of the structure only after the formation of the third stack of base pairs. Thus, the short stems may be unstable and provide a limited contribution to the overall free energy of a folded RNA molecule. Therefore, detailed analysis of stems in pseudoknots could facilitate reducing pseudoknots complexity. In this paper, we show that the pseudoknots from experimentally determined RNA spatial structures are primarily formed by short stems of 2–3 base pairs. The short stems tend to avoid hairpins and prefer internal loops that indicates that they could be energetically insignificant. An exclusion of short stems reduces the diversity of pseudoknots to two basic types which are H-knots (signature abAB) and kissing loops (signature abAcBC). The only exception is a pseudoknot formed by 12–13 stems that was found in group II intron molecule from *Oceanobacillus iheyensis* only in the presence of exon segment IBS1. In the absence of IBS1 the pseudoknot is reduced to kissing loops type.

Key words: *pseudoknot, short stem, RNA secondary structure, pseudoknot signature, base pair, stem, group II intron.*

INTRODUCTION

Ribonucleic acid (RNA) is one of the most important classes of biopolymers in the living organisms. Along with DNA and proteins RNA molecules are vital for many cell processes (see review [1]). In the last two decades a plenty of functional noncoding RNA molecule types has been discovered that led to a new vision of RNA role in a cell. Apart from well-known messenger, transfer and ribosomal RNAs there are microRNA, transfer-messenger RNA, small nuclear RNA, long noncoding RNA, etc., that are involved in translation [2], gene regulation [3], RNA processing [4] and other processes.

Since the function of an RNA molecule strongly depends on its spatial structure it seems very important to analyze its structural patterns including all levels of organization: sequence,

*baulin@lpm.org.ru

secondary structure and tertiary structure. Such element as pseudoknot is of particular interest since it's the most complex type of RNA motifs. At the time there is no common opinion whether pseudoknots should be treated as a part of secondary or tertiary structure of RNA [5]. Pseudoknots are required for many processes such as tertiary folding and activity of ribozymes [6, 7], splicing [8], telomerase functioning [9].

RNA secondary structure prediction is a classical problem of RNA bioinformatics. A number of algorithms can solve this problem in polynomial time (see review [10]) but predicting either classical RNA structures (i.e. excluding pseudoknots) [11, 12] or only few pseudoknot types [13, 14]. In [15] it has been shown that RNA secondary structure prediction including pseudoknots of arbitrary types is an NP-hard problem.

Here we report that the pseudoknots from experimentally determined RNA spatial structures are primarily formed by short stems of 2 or 3 base pairs. According to [16–18] the folding of a stem starts to decrease the free energy of a structure after the formation of the third stack of base pairs only. Thus short stems are expected to be energetically insignificant and it becomes promising to investigate the role of short stems in RNA pseudoknotted structures. In this work we have analyzed all pseudoknots from experimentally determined RNA spatial structures concerning the lengths of their constituent stems. We have shown that having short stems excluded from the consideration the diversity of pseudoknots is reduced to two basic types. We also report new evidence for energetic insignificance of short stems.

METHODS

Terminology and formal definitions

We consider 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 only pairs of complementary nucleotides (A–U and G–C pairs, also known as Watson – Crick pairs) and G–U pairs (Wobble pairs).

A *Stem* is a 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.

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 (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 stem is called a *short stem* if $k < 3$. 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 any base pair from the stem.

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 $(i \leq k \leq j; l > j)$ or $(k < i; i \leq l \leq j)$;

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 work follows our original formal definitions described in [19] and at http://urs.lpm.org.ru/struct.py?where=3#def. For classical RNA structures the definitions coincide with the commonly accepted [20].

Pseudoknot signatures

We use a classification of pseudoknots based on the notion of signature. The classification is close to topological classification from [21]. The main difference is that we take into account only stems excluding isolated base pairs. The definition of signature coincides with the one from [22]. Similar definitions also occur in works [10, 23–25].

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 the sequence of its wing letters given according to the wings positions on the chain from 5'- to 3'-end, see Fig. 1.

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, see Fig. 1.

Stems *xX, yY, ...* are connected within an upper signature if both the word *xy...* and inverted word *...YX* 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, see Fig. 1.

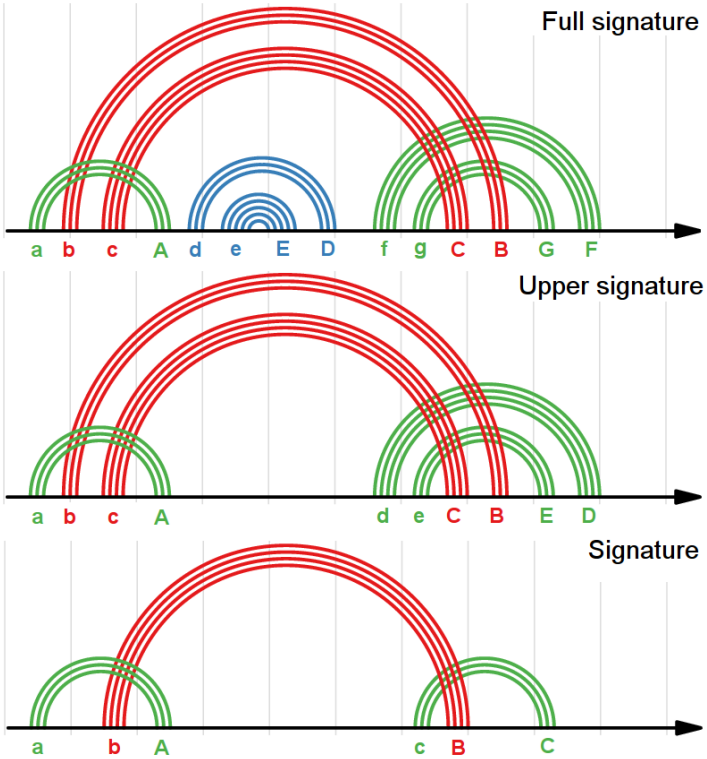

Fig. 1. Signature of the pseudoknotted ECR. The ECR contains seven stems; each stem is labeled with a letter (see the text). The word abcAdeEDfgCBGF composed of such letters is the full signature of the pseudoknot. The nested stems named dD and eE are removed. The letters for the remaining stems are reassigned. The word abcAdeCBED is the upper signature of the pseudoknot. We combine each family of parallel stems into one arc. The letters are reassigned. The word abAcBC is the signature of the pseudoknot. The figure is prepared with R-chie web-server (www.e-rna.org/r-chie/).

Input data

We used the URSDB database [19] as a source of experimentally determined structures as the only database of RNA structures and RNA-protein complexes containing PDB entries [26] with annotated signatures of pseudoknots. A non-redundant set of RNA-containing PDB entries has been selected using the BGSU RNA Site ([27], version 2.156_all of November 24th 2017). The set included 2300 structures formed by 2974 RNA molecules from 1987 PDB entries. Since URSDB counts interchain stems we have recalculated all pseudoknot signatures excluding such stems. Next, all signatures have been recalculated in three separate ways - excluding stems of length 2, excluding stems of length 3 and excluding stems of both lengths 2 and 3. To verify that the non-redundant set contains all possible pseudoknot types we repeated the process for the complete set of RNA-containing PDB entries from URSDB.

RESULTS

RNA structures of the non-redundant set included 12793 stems of different lengths. Short stems made up to more than half of all stems that are part of pseudoknots: 1171 stems of 2303. We also discovered that short stems tend to avoid hairpins (p-value < $2.2 \cdot 10^{-16}$, Fisher's exact test) and slightly prefer internal loops instead (p-value = $6.29 \cdot 10^{-9}$, Fisher's exact test), see Table 1. Short stems adjacent to internal loops could be stabilized by noncanonical base pairs inside the loop and by neighboring stems, but the short stems adjacent to hairpins lack additional sources of stabilization. This supports the previous data [16–18] that short stems are energetically unstable.

Table 1. Colocalization of stems and adjacent loops of different types

Adjacent loop	Short stems (2-3 bp)	Longer stems (4 bp and more)
Hairpin	2242	3010
Non-hairpin	3895	3646
Internal loop	2066	1923
Non-internal	4071	4733

Table 2. Diversity of pseudoknot signatures with and without short stems (based on 1987 PDB entries of the non-redundant set)

Signature	Number of pseudoknots including short stems	Number of pseudoknots excluding short stems
abAB	289	74
abAcBC	67	43
abcdBCAD	2	0
abcdCADB	2	0
abAcdBDeCE	3	0
abcdCABeDE	1	0
abAcdCeBEfDF	1	0
abAcdeBEfDFC	9	0
abcdCeAEfDFB	1	0
abcdCeBEAfDF	1	0
abAcdCefDFgBGE	1	0
abAcdeDfgFEChBHiGI	17	0
abAcdeDfghiHFECjkGKlBLJmIM	2	0
Total	**396**	**117**

Mathematical Biology and Bioinformatics. 2019. V. 14. № S. doi: 10.17537/2019.14.t37

In the non-redundant set of RNA structures we found 396 pseudoknots with 13 different signatures, see Table 2. The most frequent types were H-knots (signature **abAB**, 289 instances) and kissing loops (signature **abAcBC**, 67 instances). After excluding the short stems all recalculated pseudoknot signatures belonged to two basic types mentioned above (see Table 2).

DISCUSSION

We also analyzed pseudoknot signatures presented in the complete set of RNA-containing PDB entries from URSDB. By doing so we found 6986 pseudoknots in total and 1276 pseudoknots after short stems exclusion. Interestingly 6 of the latter were neither H-knots nor kissing loops and kept their signature **abcdCefAFDEB** after short stems exclusion.

All 6 pseudoknots of signature **abcdCefAFDEB** were found in the instances of group II intron molecule from *Oceanobacillus iheyensis*. In total we counted 26 structures of this molecule in URSDB composing equivalence class NR_all_35054.3 [27]. The pseudoknots **abcdCefAFDEB** required the specific exon segment called Intron Binding Site 1 (IBS1, sequence AUAA). Apart from 6 pseudoknots found earlier (from 5 PDB entries: 4ds6, 4faq, 4fau, 4y1n, 4y1o) we found 3 more (PDB entries 4far, 3eoh, 3eog) where the similar pseudoknots were formed but in these cases IBS1 was not ligated with the intron and the interchain stems of pseudoknots were removed from the consideration thus changing the pseudoknot signature.

The pseudoknot **abcdCefAFDEB** requires IBS1, thus it probably exists only before or during splicing. It should be noted that IBS1 is absent in the representative structure of NR_all_35054.3 class (PDB entry 5j01) that explains the absence of signature **abcdCefAFDEB** in Table 2.

We also performed a search for group II intron molecules from other organisms and found three such structures - two from *Lactococcus lactis* (5g2x, 5g2y; class NR_all_28269.1) and one from *Pylaiella littoralis* (4r0d; class NR_all_05993.1). Structures from 5g2x and 4r0d also contained IBS1 and formed complex pseudoknots. However, these pseudoknots contained short stems and without them had simpler signature **abAcBC**. Therefore, in all available RNA-containing PDB entries there is one and only example of a complex pseudoknot that does not contain short stems.

CONCLUSION

We analyzed short stems in the formation of pseudoknotted RNA structures. We have shown that short stems tend to avoid hairpins and prefer internal loops that indicates that they could be energetically insignificant. By excluding short stems, all of the pseudoknots from the non-redundant set of RNA structures belonged to two basic types with signatures **abAB** and **abAcBC**. From this point it might be possible to design an algorithm for RNA secondary structure prediction based on two iterations: (a) prediction a structure with no short stems allowing only two mentioned types of pseudoknots and (b) addition of energetically favorable and sterically possible short stems.

Analysis of the complete set of available RNA structures revealed the only example of a complex pseudoknot that did not contain short stems. This pseudoknot was found in group II intron molecule from *Oceanobacillus iheyensis* only in the presence of exon segment IBS1.

FUNDING

This work was supported by the Russian Foundation for Basic Research [grant number 16-04-01640].

ACKNOWLEDGEMENTS

We thank Ivan V. Kulakovskiy for fruitful discussions.

REFERENCES

1. Marzluff W.B. Twenty years of RNA: reflections on post-transcriptional regulation. *RNA (New York, NY)*. 2015. V. 21. № 4. P. 687–689. doi: 10.1261/rna.050997.115.
2. Eiring A.M., Harb J.G., Neviani P., Garton Ch., Oaks J.J., Spizzo R., Liu Sh., Schwind S., Santhanam R., Hickey Ch.J. et al. miR-328 functions as an RNA decoy to modulate hnRNP E2 regulation of mRNA translation in leukemic blasts. *Cell*. 2010. V. 140. № 5. P. 652–665. doi: 10.1016/j.cell.2010.01.007.
3. Bartel D.P. MicroRNAs: target recognition and regulatory functions. *Cell*. 2009. V. 136. № 2. P. 215–233. doi: 10.1016/j.cell.2009.01.002.
4. Kapranov P., Cheng J., Dike S., Nix D.A., Duttagupta R., Willingham A.T., Stadler P.F., Hertel J., Hackermüller J., Hofacker I.L. et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*. 2007. V. 316. № 5830. P. 1484–1488. doi: 10.1126/science.1138341.
5. Shapiro B.A., Yingling Ya.G., Kasprzak W., Bindewald E. Bridging the gap in RNA structure prediction. *Current Opinion in Structural Biology*. 2007. V. 17. № 2. P. 157–165. doi: 10.1016/j.sbi.2007.03.001.
6. Rastogi T., Beattie T.L., Olive J.E., Collins R.A. A long-range pseudoknot is required for activity of the Neurospora VS ribozyme. *The EMBO Journal*. 1996. V. 15. № 11. P. 2820. doi: 10.1002/j.1460-2075.1996.tb00642.x.
7. Ke A., Zhou K., Ding F., Cate J.H., Doudna J.A. A conformational switch controls hepatitis delta virus ribozyme catalysis. *Nature*. 2004. V. 429. № 6988. P. 201–205. doi: 10.1038/nature02522.
8. Adams P.L., Stahley M.R., Kosek A.B., Wang J., Strobel S.A. Crystal structure of a self-splicing group I intron with both exons. *Nature*. 2004. V. 430. № 6995. P. 45–50. doi: 10.1038/nature02642.
9. Theimer C.A., Blois C.A., Feigon J. Structure of the human telomerase RNA pseudoknot reveals conserved tertiary interactions essential for function. *Molecular Cell*. 2005. V. 17. № 5. P. 671–682. doi: 10.1016/j.molcel.2005.01.017.
10. Condon A., Davy B., Rastegari B., Zhao Sh., Tarrant F. Classifying RNA pseudoknotted structures. *Theoretical Computer Science*. 2004. V. 320. № 1. P. 35–50. doi: 10.1016/j.tcs.2004.03.042.
11. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*. 2003. V. 31. № 13. P. 3406–3415. doi: 10.1093/nar/gkg595.
12. Reeder J., Höchsmann M., Rehmsmeier M., Voss B., Giegerich R. Beyond Mfold: recent advances in RNA bioinformatics. *Journal of Biotechnology*. 2006. V. 124. № 1. P. 41–55. doi: 10.1016/j.jbiotec.2006.01.034.
13. Rivas E., Eddy S.R. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of Molecular Biology*. 1999. V. 285. № 5. P. 2053–2068. doi: 10.1006/jmbi.1998.2436.
14. Lyngsø R.B., Pedersen C.N.S. Pseudoknots in RNA secondary structures. In: *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology. ACM*. 2000. P. 201–209. doi: 10.1145/332306.332551.
15. Lyngsø R.B., Pedersen C.N.S. RNA pseudoknot prediction in energy-based models. *Journal of Computational Biology*. 2000. V. 7. № 3–4. P. 409–427. doi: 10.1089/106652700750050862.
16. Tan Z., Zhang W., Shi Ya., Wang F. RNA folding: structure prediction, folding kinetics and ion electrostatics. *Advance in Structural Bioinformatics*. Springer Netherlands, 2015. P. 143–183. doi: 10.1007/978-94-017-9245-5_11.
17. Xia T., SantaLucia J. Jr., Burkard M.E., Kierzek R., Schroeder S.J., Jiao X., Cox Ch., Turner D.H. Thermodynamic parameters for an expanded nearest-neighbor model for

formation of RNA duplexes with Watson− Crick base pairs. *Biochemistry*. 1998. V. 37. № 42. P. 14719–14735. doi: 10.1021/bi9809425.
18. Mathews D.H., Sabina J., Zuker M., Turner D.H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*. 1999. V. 288. № 5. P. 911–940. doi: 10.1006/jmbi.1999.2700.
19. Baulin E., Yacovlev V., Khachko D., Spirin S., Roytberg M. URS DataBase: universe of RNA structures and their motifs. *Database*. 2016. V. 2016. P. baw085. doi: 10.1093/database/baw085.
20. Zuker M., Mathews D.H., Turner D.H. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In: *RNA Biochemistry and Biotechnology*. Springer Netherlands, 1999. P. 11–43. doi: 10.1007/978-94-011-4485-8_2.
21. Andersen J.E., Penner R.C., Reidys C.M., Waterman M.S. Topological classification and enumeration of RNA structures by genus. *Journal of Mathematical Biology*. 2013. V. 67. № 5. P. 1261–1278. doi: 10.1007/s00285-012-0594-x.
22. Bon M., Vernizzi G., Orland H., Zee A. Topological classification of RNA structures. *Journal of Molecular Biology*. 2008. V. 379. № 4. P. 900–911. doi: 10.1016/j.jmb.2008.04.033.
23. Rødland E.A. Pseudoknots in RNA secondary structures: representation, enumeration, and prevalence. *Journal of Computational Biology*. 2006. V. 13. № 6. P. 1197–1213. doi: 10.1089/cmb.2006.13.1197.
24. Reidys C.M., Huang F.W.D., Andersen J.E., Penner R.C., Stadler P.F., Nebel M.E. Topology and prediction of RNA pseudoknots. *Bioinformatics*. 2011. V. 27. № 8. P. 1076–1085. doi: 10.1093/bioinformatics/btr090.
25. Chiu J.K.H., Chen Y.P.P. Conformational features of topologically classified RNA secondary structures. *PloS one*. 2012. V. 7. № 7. Article No. e39907. doi: 10.1371/journal.pone.0039907.
26. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. The protein data bank. *Nucleic Acids Research*. 2000. V. 28. № 1. P. 235–42. doi: 10.1093/nar/28.1.235.
27. Leontis N.B., Zirbel C.L. Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking. In: *RNA 3D structure analysis and prediction*. Springer Berlin Heidelberg, 2012. P. 281–298. doi: 10.1007/978-3-642-25740-7_13.

Received 16.06.2019.
Published 18.07.2019.