Towards Long-Range RNA Structure Prediction in Eukaryotic Genes

Dmitri D. Pervouchine [1,2,3] (iD)

[1]   Skolkovo Institute for Science and Technology, Ulitsa Nobelya 3, Moscow 121205, Russia;
d.pervouchine@skoltech.ru; Tel.: +7-495-280-1481 (ext. 3925)

[2]   The Faculty of Bioengineering and Bioinformatics, Moscow State University 1-73, Moscow 119899, Russia

[3]   Faculty of Computer Science, Higher School of Economics, Kochnovskiy Proyezd 3, Moscow 125319, Russia

Abstract: The ability to form an intramolecular structure plays a fundamental role in eukaryotic RNA biogenesis. Proximate regions in the primary transcripts fold into a local secondary structure, which is then hierarchically assembled into a tertiary structure that is stabilized by RNA-binding proteins and long-range intramolecular base pairings. While the local RNA structure can be predicted reasonably well for short sequences, long-range structure at the scale of eukaryotic genes remains problematic from the computational standpoint. The aim of this review is to list functional examples of long-range RNA structures, to summarize current comparative methods of structure prediction, and to highlight their advances and limitations in the context of long-range RNA structures. Most comparative methods implement the "first-align-then-fold" principle, i.e., they operate on multiple sequence alignments, while functional RNA structures often reside in non-conserved parts of the primary transcripts. The opposite "first-fold-then-align" approach is currently explored to a much lesser extent. Developing novel methods in both directions will improve the performance of comparative RNA structure analysis and help discover novel long-range structures, their higher-order organization, and RNA–RNA interactions across the transcriptome.

1. Introduction

Eukaryotic RNA processing is remarkably complex. Nascent pre-mRNA transcripts are spliced, edited, capped, cleaved, and polyadenylated [1]. All these events occur co-transcriptionally and are tightly coupled: Splicing affects cleavage and polyadenylation and viceversa [2,3], RNA editing can disrupt or create binding sites for splicing factors [4,5], etc., but more importantly, as RNA is being synthesized, it becomes coated by an army of RNA-binding proteins and folds into complex intramolecular structures.

The structure of RNA molecules is believed to comprise two levels: the secondary structure, which is formed by proximate regions in the primary sequence, and the tertiary structure, which also includes long-range interactions [6]. That is, the secondary structure is local, i.e., it forms between nearby sequences during Pol II elongation; in contrast, the tertiary structure is global, i.e., it builds from pre-formed helical domains of the local structure. Controversially, both terms refer to the secondary level of structure organization, in the sense that they both constitute residue interactions that are stabilized by stacking energies. The main difference between local and global structure is therefore in the number of nucleotides separating the interacting parts. Notably, long-range interactions between assembled helical domains tend to produce more pseudoknots than do local secondary structures [7]. Throughout this review, I use the term "long-range RNA structure" in the sense that

refers to complementary intramolecular interactions of distant RNA regions rather than to the timing of structure formation, its topology, or 3D organization.

In vivo, the structure of a pre-mRNA is critically important for its processing. Native RNAs are folded co-transcriptionally with the aid of RNA binding proteins (chaperones) or by forming structural intermediates that help to avoid traps in dysfunctional conformations [6,8]. To date, these dynamic interactions are poorly characterized, and assays of RNA–protein interactions and transcription kinetics are just starting to emerge [9,10]. RNA-binding proteins and elongation kinetics introduce large uncertainty to the parameters of the models that are commonly used for RNA folding and represent the major source of discrepancies between computational models of long RNAs [11,12]. As a result, our ability to predict eukaryotic pre-mRNA structure is biased towards local structures, while the prediction of long-range RNA structure remains problematic from the computational point of view.

This review consists of two parts. The first part discusses the existing examples of functional long-range structures in eukaryotic RNAs and outlines molecular mechanisms related to their function. The second part targets bioinformatics readership. It summarizes the current state of the art in the field of comparative RNA structure prediction, with its advances and limitations, and discusses the perspectives and directions where it could next develop. The second part is not designed to be a complete review of all RNA structure predictions methods; thus I cite only the selected computational works that contain the most references to other papers in the field.

2. Instances of Long-Range RNA Structure

Functional long-range base pairings in RNAs are known throughout the tree of life [13]. They are particularly well-studied in viruses [14], including tobacco mosaic virus [15], hepatitis B and C viruses [16,17], Dengue virus [18], and human immunodeficiency virus [19]. Over the past several years, there has been an increasing number of reports on functional long-range structures in eukaryotic RNAs [20]. Table 1 provides a short list of these structures. Their functionality is mainly associated with pre-mRNA processing, usually with splicing, and more rarely with translation [21,22]. Several molecular mechanisms have been proposed for the function of these structures, with different degrees of experimental evidence [20].

The classic RNA structure probing method based on the difference in reactivities of single-stranded and double-stranded residues, even in its modern high-throughput incarnation [23], is not quite useful at long distances because it can detect whether a nucleotide is paired, but it cannot tell to which other nucleotide. While local interaction partners can be guessed from the nearby sequence, too many options arise for complementarity at long ranges. A method of photochemical cross-linking with psoralens was developed in 1979 to localize structural interactions in eukaryotic RNAs [24]. Although this method was recently implemented in high-throughput [25–28], it is not yet in common use. The most convincing assays for RNA functionality are based on double mutants, i.e., a mutation that disrupts the RNA helix and leads to the loss of function, followed by a compensatory mutation that restores base pairing and regains the function. This method is significantly more laborious because it requires introducing point mutations in constructs or to the genome, and is limited to the cases when the single-mutant state is not lethal.

Among the eukaryotic genes with functional long-range base pairings, the most known are genes with mutually exclusive exons (MXEs), of which the most fascinating example is Down's syndrome cell adhesion molecule *Dscam* in *Drosophila* (see [29] for review). The history of *Dscam* started in 2005 when it was found that its exon 6 cluster, which consists of 48 variable exons, contains competing long-range RNA base pairings that form in a mutually exclusive way [30]. It was proposed that competing RNA structure exposes a group of exons in a loop and thereby ensures that one and only one exon is included in the mature transcript. Later, a similar splicing pattern was found also in exon 4, exon 9, and exon 17 clusters of *Dscam* [31–33]. However, the details of the molecular mechanism remained incomplete until many more structures were discovered in this gene, including locus control region [34] and another set of long-range structures [35]. The same principle for mutually exclusive

splicing as in *Dscam* was observed in other genes, including competing long-range RNA structures in *14-3-3ζ* gene [31], bidirectional pairing control of alternative exon 4 inclusion *srp* pre-mRNAs [36], and multiple competing base pairings in *MRP1* gene [37] in *Drosophila* (see [29] for more details).

While mutually exclusive exon choice is a peculiar splicing pattern, long-range RNA structure is important for coordination of other types of alternative splicing events. To name a few, the *Nmnat* gene controls the inclusion of its alternative exon coupled with alternative polyadenylation by a pair of complementary intronic sequences in *Drosophila* [38]. Human splicing factor 1 (*SF1*) contains a long-range RNA structure in a constitutive intron preventing intron retention that leads to a lethal frameshift [20]. A cluster of six exons in human *DST* gene undergoes mutually inclusive splicing, a scenario opposite to that of mutually exclusive exons, in which either all exons in the array, or none of them are included in the mRNA. This pattern is likely due to a pair of complementary sequences, which flank the exon cluster and lead to its exclusion by forming an RNA helix that exposes the entire cluster in a loop [39].

Several mechanisms were proposed to explain the impact of long-range RNA structure on splicing [29,40]. Among them, the two major scenarios are the hindrance of a stretch of the pre-mRNA in a loop and spatial approximation of distant regulatory elements (in fact, the former causes the latter). Two mammalian genes, a kinesin superfamily member *KIF21A* and an actin regulator *ENAH*, each contain a distal intronic site that is bound by Rbfox1 and Rbfox2 in the mouse brain and Rbfox2 in human 293T cells. However, these sites act as splicing enhancers only when brought in proximity of the target exon via the formation of a long-range RNA bridge, a duplex which spans over 10Kb (hence the name) [41]. Splicing of the catalytic subunit of the human telomerase gene *TERT* also depends on the long-range RNA pairing between repeat clusters, which approximates exons 6 and 9 and suppresses exons 7 and 8, thereby promoting the so-called "minus beta" splicing [42].

Table 1. Functional long-range RNA structures in *Drosophila* and Human.

Species	Gene	Function	Length *	Spread *	References
Drosophila	*Dscam*	Exon 4 cluster	13	4500	[31]
	Dscam	Exon 6 cluster	16	11,000	[30,32,34]
	Dscam	Exon 9 cluster	16	14,000	[31]
	Dscam	Exon 17 cluster	15	1000	[33]
	Mhc	Exon 7 cluster	14	2500	[31]
	Mhc	Exon 9 cluster	14	1600	[31]
	Mhc	Exon 11 cluster	15	2600	[31]
	Nmnat	Exon 5 and polyA site	14	400	[38]
	Atrophin	Exon 10	16	350	[38]
	srp	Exon 4 cluster	21	450	[36]
	14-3-3ζ	Exon 5 cluster	22	1200	[31]
Human	*SF1*	Exon 10	17	100	[20]
	ENAH	Exon 11a	18	1800	[41]
	DST	Exons 47-52	15	10,000	[39]
	SMN2	Exon 7	8 + 7 + 8	280	[43,44]
	PLP1	Exon 3	10 + 5	600	[45]
	TERT	Exons 7 and 8	Repeat	6500	[42]
	NEAT1	Paraspeckle formation	N/A	10,000	[46]

(*) Length: approximate number of base pairs in complementary regions; Spread: loop size, i.e., sequence distance between complementary parts; N/A: not applicable.

Long-range RNA structures contribute to human disease, including neurological disorders and other pathologies [47]. In particular, a long-distance RNA structure that consists of three adjacent intronic RNA stems is a critical regulator of splicing in Survival Motor Neuron 2 (*SMN2*) exon 7, the skipping of which is linked to spinal muscular atrophy, a hereditary infant disease leading to early death [43,44]. Alternative splicing of human *PLP1*, a gene responsible for X-linked leukodystrophy Pelizaeus–Merzbacher disease, is also regulated by a long-distance interaction between two highly

conserved complementary intronic elements [45]. Antisense oligonucleotides represent a prominent strategy for targeting such structured RNAs therapeutically, and some of them are already approved for clinical use [47,48]. In this regard, the identification of long-range RNA structures implicated in human disease becomes exceptionally important.

Several recent computational and experimental studies independently concluded that long-range base pairings in eukaryotic RNAs are abundant [20,25–28,39]. On the other hand, the examples from the short catalogue listed here (Table 1) demonstrate that long-range intramolecular base pairings are crucial for pre-mRNA processing. They must therefore represent only the tip of the iceberg, and efficient computational methods are needed to discover many more structures that are still hidden in eukaryotic genomes.

3. Predicting Long-Range RNA Structure

The universe of RNA structure prediction tools can be broadly divided into methods predicting intramolecular vs. intermolecular structure, on one hand, and methods based on single-sequence (de novo) vs. comparative sequence analysis on the other hand. The majority of de novo intramolecular methods implement dynamic programming for free energy minimization. Dynamic programming is effective only for unknotted structures, and its use for long RNA folding is limited because long-range interactions become shunted by local nested base pairings [39].

Although long-range RNA structure is intramolecular, it could be considered intermolecular from the prediction standpoint. Though intermolecular methods are generally as complex, many of them model RNA–RNA interactions as disassembly of the local structure followed by intermolecular hybridization, i.e., as interaction of pre-formed helical domains (see [49,50] for review and benchmark). Another possibility is to account for the contribution of pseudoknots by considering individual helices instead of base pairs, but this approach must be combined with phylogenetics [51]. At the scale of eukaryotic genomes RNA–RNA interaction prediction becomes challenging both in terms of performance and specificity because the amount of random complementarity grows with the length.

When single-sequence analysis fails, comparative methods provide a powerful alternative. The advantage of comparative methods is twofold. First, they confine the search space to evolutionarily conserved regions, which at least partly reduces the complexity and improves specificity. Second, at least hypothetically, they gain statistical power through observing compensatory changes in covarying positions [52–56]. These ideas, which stem from covariance models [57], have been remarkably successful in the discovery of riboswitches [58]. Among examples presented in this review, many functional long-range RNA structures were, in fact, first discovered in multiple sequence alignments and later confirmed experimentally.

In eukaryotic genomes, however, the comparative formulation becomes intricate when it meets large distances and complex organization of the genes. The conservation rates in exons and introns are fundamentally different, introns may not always be aligned, or the alignment may not be unique. While some methods use phylogenetic substitution models to fold protein-coding sequences [55,56], much less is known about comparative folding of introns and untranslated regions. Here, a reasonable solution is to combine multiple sequence alignment with RNA folding, which is the famous "simultaneous folding and aligning problem" that was first formulated in 1984 by Sankoff [59]. The Sankoff algorithm is computationally expensive, and its rigorous implementation for two sequences has the time and memory complexity $\mathcal{O}(n^6)$ and $\mathcal{O}(n^4)$, respectively, where n is the length of each sequence [60]. One can use it to realign an existing multiple alignment, but the depth of this realignment is limited [61]. In application to long-range intramolecular RNA structure, Sankoff method is far beyond computational capacity for $n \simeq 10,000$ (see Table 1). It could be adapted for simultaneous alignment and intermolecular structure prediction for two pairs of RNA sequences with time and memory complexity of $\mathcal{O}(n^4)$, which is still impractical for most human introns.

Eukaryotic long-range RNA structures listed in Table 1 have a number of characteristic properties. First, they evolve under negative selection, although the rate of conservation of a structural element

depends on the time in evolution when it was first acquired. After the last common ancestor, the evolution of *Dscam* went differently in *Drosophila* and in *Chelicerata* [35], while the regulatory sequence in *Nmnat* remained remarkably conserved [38]. Most of the listed structures contain uninterrupted helices of at least 12 nucleotides, many of which are surrounded by more diffuse base pairings. This is likely due to the free energy constraints to maintain long-range interaction, although the intervening sequences could also be structured since there is no apparent correlation between the stem length and the loop size. Finally, almost all examples in the table are located in syntenic regions, e.g., in introns separating orthologous exons, possibly reflecting locations related to their function and evolution.

In sum, functional long-range RNA structures have the following characteristic properties:

1.　most long-range RNA structures are well-conserved;
2.　the core of a long-range RNA structure is a long, nearly-perfect complementary match;
3.　elements of long-range RNA structures are located in syntenic regions.

The first property justifies the so-called "first-align-then-fold" limit of Sankoff's method (Figure 1), in which a set of orthologous sequences is first aligned and the alignment is then folded. It is the most frequent approach in comparative methods [49]. By construction, it disregards the hypothetical cases of sequences that have diverged beyond recognition, but their structure has remained unchanged. Its sensitivity is limited by the quality of the input alignment, particularly by the uncertainty of aligning mutually exclusive exons that arise from genomic duplications, or by misalignment of conserved structural elements that are too short relative to the size of non-conserved background [34,62]. Apart from these special circumstances, "first-align-then-fold" is a simple, fast, and powerful approach that is used in many current comparative methods, including comparative RNA–RNA interaction prediction [52,63] and probabilistic sampling [64].
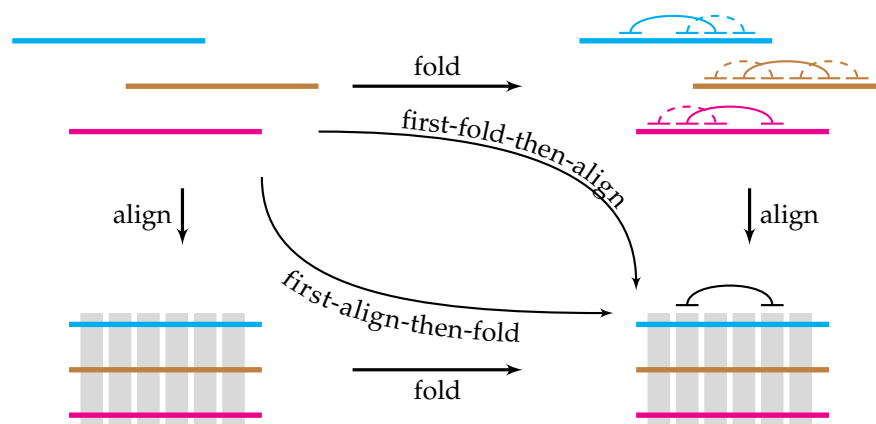
Figure 1. A "commutative diagram" of the alignment and folding tradeoff. Top left: unaligned RNA sequences. Bottom left: their structure-agnostic alignment; conserved regions are shown in gray. Top right: sparse folding identifies candidate helices shown as arcs. Bottom right: conserved helices are matched by structure-aware alignment or identified in a multiple sequence alignment.

The advantage of folding a multiple sequence alignment compared to folding its consensus sequence is the covariation statistics. In general, it is possible to gain statistical power from observing covarying positions only when sequences mutate, e.g., in rapidly evolving viral genomes [54]. However, the examples from mammalian and insect genes (Table 1) show little or no variation, suggesting that functional structures evolve under strong negative selection [38]. In addition, compensatory patterns can arise not only to maintain base-pairing interactions, but also as a result of synchronized mutations that preserve binding of a common interaction partner in antisense genomic orientation [53]. An example of this is *RP11-439A17.4* long non-coding RNA, which is located in antisense to *HIST2H2BA* gene and overlaps a transcription factor binding site, which also occurs in almost all human histone

genes in sense orientation, resulting in a seeming compensatory pattern [39]. Thus, the comparative approach is less efficient in the case of extreme conservation.

The second and the third property suggest the opposite, "first-fold-then-align" route (Figure 1) which is explored to a much lesser extent. Indeed, the number of folds for a single sequence taken to the power of their combinations in the multiple alignment does not appear feasible at first glance. A pioneer work on local secondary structure folded and aligned enterovirus mRNAs using phylogenetic comparison of potential stems followed by the consistency analysis of structure graph [65]. These ideas were extended to long-range structures [39], where a dramatic reduction of the fold space was achieved by considering sparse structures, i.e., ones that consist of long, nearly complementary matches. This, and the assumption that the interacting parts are located in syntenic regions, decrease the number of ways in which the helices can be aligned to the point where whole-transcriptome analyses become possible [39]. However, while this approach is quite sensitive, it does so at the expense of a high false positive rate and cannot deal with repeats and other low complexity regions.

The diagram in Figure 1 becomes commutative, i.e., predictions by the two methods coincide, when all three properties are met. Here, one unexplored and a potentially productive approach would be to match, within a certain distance limit, all standalone conserved regions that do not have protein-coding constraints and score them by how abruptly the conservation outside of them ends. It is a frequently observed pattern that non-complementary background in long-range RNA structures is washed out immediately outside of the complementary region. Another potentially useful direction is to combine comparative approaches with experimental methods that give global mapping of RNA duplexes to narrow down the space of potential complementary interactions [25–28].

4. Concluding Remarks

Intramolecular structure of eukaryotic RNAs is not limited to hairpins and can span thousands of bases. Recent high throughput experimental assays confirmed that distant interactions in the human transcriptome are very abundant. However, the computational identification of long-range RNA structure remains problematic because the interacting parts are separated by long distances.

The principles of computational identification of RNA structure by comparative methods span between two extremes. On one side are the so-called "first-align-then-fold" methods, which essentially look for complementary regions in multiple sequence alignments. They are powerful for well-conserved sequences, but hardly applicable to non-conserved regions in eukaryotic genes that often harbor functional RNA structure elements (Table 1). On the other side are more complex "first-fold-then-align" methods, which are applicable to non-conserved regions, but have a high false positive rate.

Characteristic features of long-range RNA structures that are outlined in this review demonstrate that, with additional assumptions, both types of methods are computationally tractable at the scale of eukaryotic genes. Further development in these both directions will expand the capabilities of comparative RNA structure analysis and lead to a discovery of many novel long-range RNA structures, their higher-order organization, and RNA–RNA interactions in the human transcriptome.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Saldi, T.; Cortazar, M.A.; Sheridan, R.M.; Bentley, D.L. Coupling of RNA Polymerase II transcription elongation with pre-mRNA Splicing. *J. Mol. Biol.* **2016**, *428*, 2623–2635. [CrossRef] [PubMed]
2. Kaida, D. The reciprocal regulation between splicing and 3'-end processing. *Wiley Interdiscip. Rev. RNA* **2016**, *7*, 499–511. [CrossRef] [PubMed]

3. Lepennetier, G.; Catania, F. Exploring the impact of cleavage and polyadenylation factors on pre-mRNA splicing across Eukaryotes. *G3 Genes Genomes Genet.* **2017**, *7*, 2107–2114. [CrossRef] [PubMed]

4. Laurencikiene, J.; Kallman, A.M.; Fong, N.; Bentley, D.L.; Ohman, M. RNA editing and alternative splicing: The importance of co-transcriptional coordination. *EMBO Rep.* **2006**, *7*, 303–307. [CrossRef] [PubMed]

5. Solomon, O.; Oren, S.; Safran, M.; Deshet-Unger, N.; Akiva, P.; Jacob-Hirsch, J.; Cesarkas, K.; Kabesa, R.; Amariglio, N.; Unger, R.; et al. Global regulation of alternative splicing by adenosine deaminase acting on RNA (ADAR). *RNA* **2013**, *19*, 591–604. [CrossRef] [PubMed]

6. Schroeder, R.; Barta, A.; Semrad, K. Strategies for RNA folding and assembly. *Nat. Rev. Mol. Cell Biol.* **2004**, *5*, 908–919. [CrossRef] [PubMed]

7. Brion, P.; Westhof, E. Hierarchy and dynamics of RNA folding. *Annu. Rev. Biophys. Biomol. Struct.* **1997**, *26*, 113–137. [CrossRef] [PubMed]

8. Herschlag, D. RNA chaperones and the RNA folding problem. *J. Biol. Chem.* **1995**, *270*, 20871–20874. [CrossRef] [PubMed]

9. Van Nostrand, E.L.; Pratt, G.A.; Shishkin, A.A.; Gelboin-Burkhart, C.; Fang, M.Y.; Sundararaman, B.; Blue, S.M.; Nguyen, T.B.; Surka, C.; Elkins, K.; et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods* **2016**, *13*, 508–514. [CrossRef] [PubMed]

10. Paulsen, M.T.; Veloso, A.; Prasad, J.; Bedi, K.; Ljungman, E.A.; Magnuson, B.; Wilson, T.E.; Ljungman, M. Use of Bru-Seq and BruChase-Seq for genome-wide assessment of the synthesis and stability of RNA. *Methods* **2014**, *67*, 45–54. [CrossRef] [PubMed]

11. Morgan, S.; Higgs, P. Evidence for kinetic effects in the folding of large RNA molecules. *J. Chem. Phys.* **1996**, *105*, 7152. [CrossRef]

12. Lai, D.; Proctor, J.R.; Meyer, I.M. On the importance of cotranscriptional RNA structure formation. *RNA* **2013**, *19*, 1461–1473. [CrossRef] [PubMed]

13. Edlind, T.D.; Cooley, T.E.; Richards, S.H.; Ihler, G.M. Long range base-pairing in the leftward transcription unit of bacteriophage lambda. Characterization by electron microscopy and computer-aided sequence analysis. *J. Mol. Biol.* **1984**, *179*, 351–365. [CrossRef]

14. Nicholson, B.L.; White, K.A. Functional long-range RNA-RNA interactions in positive-strand RNA viruses. *Nat. Rev. Microbiol.* **2014**, *12*, 493–504. [CrossRef] [PubMed]

15. Archer, E.J.; Simpson, M.A.; Watts, N.J.; O'Kane, R.; Wang, B.; Erie, D.A.; McPherson, A.; Weeks, K.M. Long-range architecture in a viral RNA genome. *Biochemistry* **2013**, *52*, 3182–3190. [CrossRef] [PubMed]

16. Shin, M.K.; Kim, J.H.; Ryu, D.K.; Ryu, W.S. Circularization of an RNA template via long-range base pairing is critical for hepadnaviral reverse transcription. *Virology* **2008**, *371*, 362–373. [CrossRef] [PubMed]

17. Fricke, M.; Dunnes, N.; Zayas, M.; Bartenschlager, R.; Niepmann, M.; Marz, M. Conserved RNA secondary structures and long-range interactions in hepatitis C viruses. *RNA* **2015**, *21*, 1219–1232. [CrossRef] [PubMed]

18. De Borba, L.; Villordo, S.M.; Iglesias, N.G.; Filomatori, C.V.; Gebhard, L.G.; Gamarnik, A.V. Overlapping local and long-range RNA-RNA interactions modulate dengue virus genome cyclization and replication. *J. Virol.* **2015**, *89*, 3430–3437. [CrossRef] [PubMed]

19. Ooms, M.; Abbink, T.E.; Pham, C.; Berkhout, B. Circularization of the HIV-1 RNA genome. *Nucleic Acids Res.* **2007**, *35*, 5253–5261. [CrossRef] [PubMed]

20. Pervouchine, D.D.; Khrameeva, E.E.; Pichugina, M.Y.; Nikolaienko, O.V.; Gelfand, M.S.; Rubtsov, P.M.; Mironov, A.A. Evidence for widespread association of mammalian splicing and conserved long-range RNA structures. *RNA* **2012**, *18*, 1–15. [CrossRef] [PubMed]

21. Tajima, Y.; Iwakawa, H.O.; Kaido, M.; Mise, K.; Okuno, T. A long-distance RNA-RNA interaction plays an important role in programmed −1 ribosomal frameshifting in the translation of p88 replicase protein of *Red clover necrotic mosaic virus*. *Virology* **2011**, *417*, 169–178. [CrossRef] [PubMed]

22. Ruegsegger, U.; Leber, J.H.; Walter, P. Block of *HAC1* mRNA translation by long-range base pairing is released by cytoplasmic splicing upon induction of the unfolded protein response. *Cell* **2001**, *107*, 103–114. [CrossRef]

23. Watters, K.E.; Lucks, J.B. Mapping RNA structure in vitro with SHAPE chemistry and next-generation sequencing (SHAPE-Seq). *Methods Mol. Biol.* **2016**, *1490*, 135–162. [PubMed]

24. Shen, C.K.; Hearst, J.E. A technique for relating long-range base pairing on single-stranded DNA and eukaryotic RNA processing. *Anal. Biochem.* **1979**, *95*, 108–116. [CrossRef]

25. Ramani, V.; Qiu, R.; Shendure, J. High-throughput determination of RNA structure by proximity ligation. *Nat. Biotechnol.* **2015**, *33*, 980–984. [CrossRef] [PubMed]

26. Aw, J.G.; Shen, Y.; Wilm, A.; Sun, M.; Lim, X.N.; Boon, K.L.; Tapsin, S.; Chan, Y.S.; Tan, C.P.; Sim, A.Y.; et al. In vivo mapping of eukaryotic RNA interactomes reveals principles of higher-order organization and regulation. *Mol. Cell* **2016**, *62*, 603–617. [CrossRef] [PubMed]

27. Lu, Z.; Zhang, Q.C.; Lee, B.; Flynn, R.A.; Smith, M.A.; Robinson, J.T.; Davidovich, C.; Gooding, A.R.; Goodrich, K.J.; Mattick, J.S.; et al. RNA duplex map in living cells reveals higher-order transcriptome structure. *Cell* **2016**, *165*, 1267–1279. [CrossRef] [PubMed]

28. Sharma, E.; Sterne-Weiler, T.; O'Hanlon, D.; Blencowe, B.J. Global mapping of human RNA-RNA interactions. *Mol. Cell* **2016**, *62*, 618–626. [CrossRef] [PubMed]

29. Jin, Y.; Dong, H.; Shi, Y.; Bian, L. Mutually exclusive alternative splicing of pre-mRNAs. *Wiley Interdiscip. Rev. RNA* **2018**, *9*, e1468. [CrossRef] [PubMed]

30. Graveley, B.R. Mutually exclusive splicing of the insect *Dscam* pre-mRNA directed by competing intronic RNA secondary structures. *Cell* **2005**, *123*, 65–73. [CrossRef] [PubMed]

31. Yang, Y.; Zhan, L.; Zhang, W.; Sun, F.; Wang, W.; Tian, N.; Bi, J.; Wang, H.; Shi, D.; Jiang, Y.; et al. RNA secondary structure in mutually exclusive splicing. *Nat. Struct. Mol. Biol.* **2011**, *18*, 159–168. [CrossRef] [PubMed]

32. May, G.E.; Olson, S.; McManus, C.J.; Graveley, B.R. Competing RNA secondary structures are required for mutually exclusive splicing of the *Dscam* exon 6 cluster. *RNA* **2011**, *17*, 222–229. [CrossRef] [PubMed]

33. Yue, Y.; Li, G.; Yang, Y.; Zhang, W.; Pan, H.; Chen, R.; Shi, F.; Jin, Y. Regulation of *Dscam* exon 17 alternative splicing by steric hindrance in combination with RNA secondary structures. *RNA Biol.* **2013**, *10*, 1822–1833. [CrossRef] [PubMed]

34. Wang, X.; Li, G.; Yang, Y.; Wang, W.; Zhang, W.; Pan, H.; Zhang, P.; Yue, Y.; Lin, H.; Liu, B.; et al. An RNA architectural locus control region involved in *Dscam* mutually exclusive splicing. *Nat. Commun.* **2012**, *3*, 1255. [CrossRef] [PubMed]

35. Yue, Y.; Meng, Y.; Ma, H.; Hou, S.; Cao, G.; Hong, W.; Shi, Y.; Guo, P.; Liu, B.; Shi, F.; et al. A large family of *Dscam* genes with tandemly arrayed 5' cassettes in *Chelicerata*. *Nat. Commun.* **2016**, *7*, 11252. [CrossRef] [PubMed]

36. Yue, Y.; Yang, Y.; Dai, L.; Cao, G.; Chen, R.; Hong, W.; Liu, B.; Shi, Y.; Meng, Y.; Shi, F.; et al. Long-range RNA pairings contribute to mutually exclusive splicing. *RNA* **2016**, *22*, 96–110. [CrossRef] [PubMed]

37. Yue, Y.; Hou, S.; Wang, X.; Zhan, L.; Cao, G.; Li, G.; Shi, Y.; Zhang, P.; Hong, W.; Lin, H.; et al. Role and convergent evolution of competing RNA secondary structures in mutually exclusive splicing. *RNA Biol.* **2017**, *14*, 1399–1410. [CrossRef] [PubMed]

38. Raker, V.A.; Mironov, A.A.; Gelfand, M.S.; Pervouchine, D.D. Modulation of alternative splicing by long-range RNA structures in *Drosophila*. *Nucleic Acids Res.* **2009**, *37*, 4533–4544. [CrossRef] [PubMed]

39. Pervouchine, D.D. IRBIS: A systematic search for conserved complementarity. *RNA* **2014**, *20*, 1519–1531. [CrossRef] [PubMed]

40. Rubtsov, P.M. Role of pre-mRNA secondary structures in the regulation of alternative splicing. *Mol. Biol.* **2016**, *50*, 935–943. [CrossRef]

41. Lovci, M.T.; Ghanem, D.; Marr, H.; Arnold, J.; Gee, S.; Parra, M.; Liang, T.Y.; Stark, T.J.; Gehman, L.T.; Hoon, S.; et al. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat. Struct. Mol. Biol.* **2013**, *20*, 1434–1442. [CrossRef] [PubMed]

42. Wong, M.S.; Shay, J.W.; Wright, W.E. Regulation of human telomerase splicing by RNA:RNA pairing. *Nat. Commun.* **2014**, *5*, 3306. [CrossRef] [PubMed]

43. Singh, N.N.; Lawler, M.N.; Ottesen, E.W.; Upreti, D.; Kaczynski, J.R.; Singh, R.N. An intronic structure enabled by a long-distance interaction serves as a novel target for splicing correction in spinal muscular atrophy. *Nucleic Acids Res.* **2013**, *41*, 8144–8165. [CrossRef] [PubMed]

44. Singh, N.N.; Lee, B.M.; Singh, R.N. Splicing regulation in spinal muscular atrophy by an RNA structure formed by long-distance interactions. *Ann. N. Y. Acad. Sci.* **2015**, *1341*, 176–187. [CrossRef] [PubMed]

45. Taube, J.R.; Sperle, K.; Banser, L.; Seeman, P.; Cavan, B.C.; Garbern, J.Y.; Hobson, G.M. PMD patient mutations reveal a long-distance intronic interaction that regulates *PLP1/DM20* alternative splicing. *Hum. Mol. Genet.* **2014**, *23*, 5464–5478. [CrossRef] [PubMed]

46. Lin, Y.; Schmidt, B.F.; Bruchez, M.P.; McManus, C.J. Structural analyses of NEAT$_1$ lncRNAs suggest long-range RNA interactions that may contribute to paraspeckle architecture. *Nucleic Acids Res.* **2018**, *46*, 3742–3752. [CrossRef] [PubMed]

47. Bernat, V.; Disney, M.D. RNA Structures as mediators of neurological diseases and as drug targets. *Neuron* **2015**, *87*, 28–46. [CrossRef] [PubMed]

48. Singh, N.N.; Howell, M.D.; Androphy, E.J.; Singh, R.N. How the discovery of ISS-N1 led to the first medical therapy for spinal muscular atrophy. *Gene Ther.* **2017**, *24*, 520–526. [CrossRef] [PubMed]

49. Umu, S.U.; Gardner, P.P. A comprehensive benchmark of RNA-RNA interaction prediction tools for all domains of life. *Bioinformatics* **2017**, *33*, 988–996. [CrossRef] [PubMed]

50. Lai, D.; Meyer, I.M. A comprehensive comparison of general RNA-RNA interaction prediction methods. *Nucleic Acids Res.* **2016**, *44*, e61. [CrossRef] [PubMed]

51. Wiebe, N.J.; Meyer, I.M. TRANSAT– method for detecting the conserved helices of functional RNA structures, including transient, pseudo-knotted and alternative structures. *PLoS Comput. Biol.* **2010**, *6*, e1000823. [CrossRef] [PubMed]

52. Seemann, S.E.; Richter, A.S.; Gesell, T.; Backofen, R.; Gorodkin, J. PETcofold: Predicting conserved interactions and structures of two multiple alignments of RNA sequences. *Bioinformatics* **2011**, *27*, 211–219. [CrossRef] [PubMed]

53. Bindewald, E.; Shapiro, B.A. Computational detection of abundant long-range nucleotide covariation in *Drosophila* genomes. *RNA* **2013**, *19*, 1171–1182. [CrossRef] [PubMed]

54. Fricke, M.; Marz, M. Prediction of conserved long-range RNA-RNA interactions in full viral genomes. *Bioinformatics* **2016**, *32*, 2928–2935. [CrossRef] [PubMed]

55. Pedersen, J.S.; Meyer, I.M.; Forsberg, R.; Simmonds, P.; Hein, J. A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic Acids Res.* **2004**, *32*, 4925–4936. [CrossRef] [PubMed]

56. Pedersen, J.S.; Forsberg, R.; Meyer, I.M.; Hein, J. An evolutionary model for protein-coding regions with conserved RNA structure. *Mol. Biol. Evol.* **2004**, *21*, 1913–1922. [CrossRef] [PubMed]

57. Eddy, S.R.; Durbin, R. RNA sequence analysis using covariance models. *Nucleic Acids Res.* **1994**, *22*, 2079–2088. [CrossRef] [PubMed]

58. Sun, E.I.; Rodionov, D.A. Computational analysis of riboswitch-based regulation. *Biochim. Biophys. Acta* **2014**, *1839*, 900–907. [CrossRef] [PubMed]

59. Sankoff, D. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.* **1985**, *45*, 810–825. [CrossRef]

60. Havgaard, J.H.; Torarinsson, E.; Gorodkin, J. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. *PLoS Comput. Biol.* **2007**, *3*, 1896–1908. [CrossRef] [PubMed]

61. Will, S.; Yu, M.; Berger, B. Structure-based whole-genome realignment reveals many novel noncoding RNAs. *Genome Res.* **2013**, *23*, 1018–1027. [CrossRef] [PubMed]

62. Hatje, K.; Kollmar, M. Expansion of the mutually exclusive spliced exome in *Drosophila*. *Nat. Commun.* **2013**, *4*, 2460. [CrossRef] [PubMed]

63. Kato, Y.; Sato, K.; Hamada, M.; Watanabe, Y.; Asai, K.; Akutsu, T. RactIP: Fast and accurate prediction of RNA-RNA interaction using integer programming. *Bioinformatics* **2010**, *26*, i460–i466. [CrossRef] [PubMed]

64. Meyer, I.M.; Miklos, I. SimulFold: Simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. *PLoS Comput. Biol.* **2007**, *3*, e149. [CrossRef] [PubMed]

65. Touzet, H.; Perriquet, O. CARNAC: Folding families of related RNAs. *Nucleic Acids Res.* **2004**, *32*, W142–W145. [CrossRef] [PubMed]