CoBold: a method for identifying different functional classes of transient RNA structure features that can impact RNA structure formation *in vivo*

Adrián López Martín[1],[†], Mohamed Mounir[1],[†] and Irmtraud M. Meyer [ID][1],[2],[*]

[1]Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Hannoversche Str. 28, 10115 Berlin, Germany and [2]Freie Universität Berlin, Department of Biology, Chemistry and Pharmacy, Institute of Chemistry and Biochemistry, Thielallee 63, 14195 Berlin, Germany

ABSTRACT

RNA structure formation *in vivo* happens co-transcriptionally while the transcript is being made. The corresponding co-transcriptional folding pathway typically involves transient RNA structure features that are not part of the final, functional RNA structure. These transient features can play important functional roles of their own and also influence the formation of the final RNA structure *in vivo*. We here present CoBold, a computational method for identifying different functional classes of transient RNA structure features that can either aid or hinder the formation of a known reference RNA structure. Our method takes as input either a single RNA or a corresponding multiple-sequence alignment as well as a known reference RNA secondary structure and identifies different classes of transient RNA structure features that could aid or prevent the formation of the given RNA structure. We make CoBold available via a web-server which includes dedicated data visualisation.

INTRODUCTION

Any RNA transcript in a living cell not only encodes information on its encoded final functional products (e.g. proteins or non-coding RNAs), but also how it is supposed to be regulated in its specific cellular contexts *in vivo*. While RNA structure has long been acknowledged as a means to assign a final, functional role to a transcript (e.g. highly structured transcripts such as transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs)), there is now mounting evidence that more transient RNA structures are also involved in regulating key events of gene expression on transcriptome level such as alternative splicing (1), RNA editing (1,2), RNA localisation (2–4) and RNA degradation (5).

These transient RNA features appear and dis-appear on time-scales that are similar to key biological processes such as transcription, splicing, A-to-I editing and RNA structure formation. To complicate matters further, these processes are known to happen co-transcriptionally and in the same cellular compartment, i.e. while the transcript is being made *in vivo*. So far, it has thus been difficult to experimentally investigate and dis-entangle the intricate interplay of these key processes of the transcriptome *in vivo* which are key to defining the functional products in any living cell. Transcriptome-wide, hypothesis-driven studies *in silico* can greatly help in that regard. Based on a computational study of several large transcriptome data sets, we managed to show with significant statistical evidence that co-transcriptional A-to-I editing can regulate alternative splicing via induced changes to local RNA structure elements near splice sites (1).

Overall and based on a wide range of experimental and computational evidence, we know by now that RNA structure formation *in vivo* is highly dependent on the cellular context *in vivo* as well as the kinetics of RNA structure formation *in vivo* (6). One and the same RNA transcript can readily encode more than a single functional RNA structure. Similarly to the concept of alternative splicing, these RNA structure features can be differentially expressed depending on the particulars of the transcript at a specific time and location *in vivo*, a concept we recently introduced as *alternative RNA structure expression*. Which encoded RNA structure is expressed at any given point of the transcript's cellular life *in vivo* depends on subtleties such as *trans* interaction partners (e.g. proteins, ligands, other RNAs), ion concentrations, the temperature and the kinetics of RNA structure formation.

Most computational RNA secondary structure prediction methods assume the one-sequence-one-structure dogma by predicting a single RNA secondary structure (from now on simply called RNA structure) for any given input RNA; well-known examples include the widely-used

---

[*]To whom correspondence should be addressed. Tel: +49 30 9406 3292; Fax: +49 30 9406 3291; Email: irmtraud.meyer@cantab.net
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

methods MFOLD and RNAFOLD (7–12). In reality, however, biologically relevant transcripts can encode between zero and multiple functionally relevant RNA structures which are differentially expressed depending on the particulars of the cellular context *in vivo*. Most RNA structure prediction methods aim to predict the RNA structure that corresponds to the (pseudo-knot free) RNA secondary structure configuration with the minimum overall free energy (so-called minimum-free energy (MFE) approach) in an *in vitro* setting. This so-called MFE RNA structure, however, may not be expressed in any biologically setting *in vivo*.

One conceptual alternative to predicting RNA structure that are functionally relevant *in vivo* is to try to identify the RNA structure(s) that have been conserved during adequately chosen times of evolution. This approach typically, but not always (9), requires as input a multiple-sequence alignment (MSA) of orthologous or homologous RNAs from several, evolutionarily related species. This strategy has the significant conceptual advantage of not having to understand *why* RNA structure features have been conserved provided they *have been* conserved. This so-called comparative approach to RNA structure prediction currently yields the state-of-art in terms of prediction accuracy (13). These methods typically employ mathematically principled approaches such as stochastic context-free grammars (SCFGs) which can explicitly capture different RNA structure formation hypotheses (rather than capturing only the single MFE strategy) and are capable of estimating the reliability of their own predictions (14,15).In this fashion, a recently-developed method know as R-scape (16) employs a G-test statistics in order to assess the statistical significance in terms of E-value for evolutionary-conserved base-pairs within RNA structure containing multiple sequence alignments.

Even these state-of-the-art methods, however, typically ignore the kinetic process of co-transcriptional RNA structure formation. With comparative methods, one can hope that the evolutionarily conserved RNA structure features that are detected and combined into the predicted RNA structures still correspond to the relevant final, functional RNA structure that matter *in vivo*. When considering non-comparative methods that are guided by the MFE strategy such as MFOLD and RNAFOLD, however, it is less clear how they could be modified to take the overall effects of co-transcriptional RNA structure formation into account. Their predictive performance has long been know to decrease markedly for input sequences longer than 200 nucleotides (nt) for which kinetic effects of co-transcriptional are expected to matter most (17). We showed with the RNA structure prediction program CO-FOLD (18) that it is technically and conceptually possible to incorporate overall effects of co-transcriptional into MFOLD, thereby significantly increasing its prediction accuracy, especially for long input sequences. Since then, however, we also found compelling evidence that some transient RNA structure features can be just as highly conserved during evolution as the features of the final, functional RNA structure (19–21). This unfortunately implies that even the state-of-the-art comparative methods may become confused by these highly conserved transient RNA structures features which they may erroneously
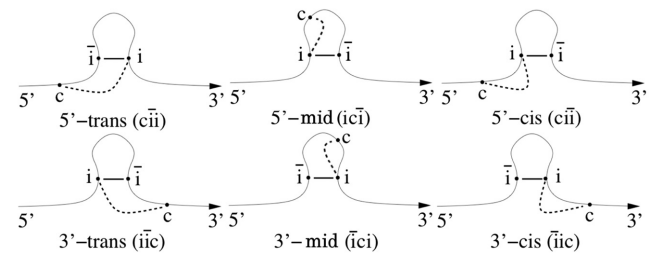

Figure 1. Different functional classes of transient helices that can emerge during co-transcriptional folding and interfere with the formation of the final RNA structure: 5′-TRANS, 5′-MID, 5′-CIS, 3′-TRANS, 3′-MID and 3′-CIS. The base-pair between nucleotides i and $\bar{i}$ is part of a helix of the known RNA structure. The base-pair between c and i is part of a transient helix that clashes with the helix involving base-pair i and $\bar{i}$ of the known structure. The nomenclature for the different types of transient helices depends on the location of the competing base-pair with respect to the base-pair with which it competes (TRANS, MID or CIS) and the location of the location of the transient helix w.r.t. the known helix it clashes with (upstream (5′) or downstream (3′)).
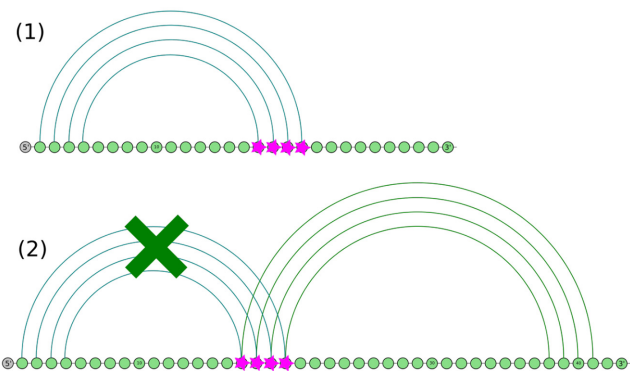
Figure 2. Co-transcriptional role of a transient helix of type 5′-CIS which can easily yield to the formation of the helix it is clashing with during co-transcriptional RNA structure formation by temporarily engaging some of its basepairs from its 5′ side and thereby preventing the formation of other structure features from pairing with them.

try to combine into the predicted final, functional RNA structures.

An early computational data study by us from 2013 (19) showed that known, structured RNA genes not only encode these final RNA structures themselves, but also information on how to form them co-transcriptionally *in vivo*. For this, we proposed a classification of transient helices (so-called clashing helices) based on their compatibility with helices of the known, functionally relevant RNA structure. In that context, we defined helices as consecutive stretches of base-pairs. We here extend the original classification of these clashing, transient helices to the functional classes shown in Figure 1. As the corresponding Table 1 shows, different classes can play different functional roles by either aiding (Figure 2) or preventing (Figure 3) the formation of the know reference RNA structure during co-transcriptional folding *in vivo*. Our original study from 2004 (21) discovered statistically significant 5′-to-3′ asymmetries by quantifying the overall effect that these clashing, transient helices have on co-transcriptional RNA structure formation *in vivo*, as the evolutionary prevalence of 3′ clashing helices was much
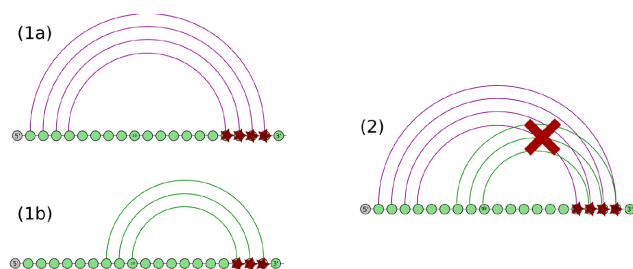
Figure 3. Co-transcriptional role of a transient helix of type 5′-TRANS. A transient helix of type 5′-TRANS can easily prevent the co-transcriptional formation of the helix it is clashing with by getting the RNA structure formation kinetically trapped during co-transcriptional folding as it directly competes with the 5′ side of the original helix for the 3′ side at the time of the formation of the helix.

Table 1. Functional classification of the six types of transient helices shown in Figure 1. Transient helices of type 5′-CIS and 3′-TRANS are expected to have a beneficial effect on the co-transcriptional formation of the known reference RNA structure as they can easily yield to the helix with which they clash. In contrast to this, transient helices of type 5′-TRANS and 3′-CIS can potentially prevent the formation of the helix with which they clash as they may get the known RNA structure kinetically trapped during co-transcriptional folding. All other types of transient helices can be expected to play a neutral functional role w.r.t. the co-transcriptional of the reference RNA structure. For more information, see also Figures 2 and 3

	TRANS	MID	CIS
5′	Disruptive	Neutral	Beneficial
3′	Beneficial	Neutral	Disruptive

lower than the 5′ one. This asymmetry arises from the different effects that the different types of helix have on the co-transcriptional formation of the final structure. Potentially beneficial helices of type 5′-CIS and 3′-TRANS tend to be more conserved that helices of type 3′-CIS and 5′-TRANS that could prevent the co-transcriptional formation of the known RNA structure.

The experimental identification of transient RNA structure features remains a complex and time-consuming task that is typically strongly tailored to the specific requirements and features of a single transcript of interest. Existing approaches can be subdivided into *in vitro* and *in vivo* methods. *In vitro* strategies allow for the tight control of key features such as the temperature, transcription rate, concentration of certain ions and select interaction partners and are thus often preferred to facilitate reproducibility. These *in vitro* methods, however, cannot capture some of the (known or unknown) complexities of the cellular environment *in vivo* that may critically influence the RNA structure formation *in vivo* (22). In his review paper from 2000 (23), Higgs already noted marked discrepancies in determined RNA structures that are due to differences in *how* and in which *molecular context* the molecule of interest was probed experimentally (e.g. *in vivo* versus *in vitro*, NMR versus X-ray). It should thus not come as a surprise that these differences will probably extend to the experimental determination of *transient* RNA structure features which can be expected to be even more sensitive to the particular features of the cellular environment *in vivo*.

The first experimental protocol for the high-throughput probing of co-transcriptional RNA structure features on a transcriptome-wide scale *in vivo* and *in vitro*, called SPET-SEQ, was published by Incarnato *et al.* only in 2017 (24). This method is capable of investigating transient RNA structures at single-nucleotide resolution and for individual transcription intermediates due to the fast-acting reagent DMS. SPET-seq's main limitation is that it primarily enables the RNA structure investigation of the 3′-ends of transcripts. Even with this limitation, however, the research team gained valuable and unique insight into the RNA structure formation of mature cytosolic RNAs in *E. coli*.

Strobel *et al.* (25) recently succeeded in extending SHAPE-probing to a high-throughput platform that allows for the simultaneous testing of several variants of the same riboswitch during co-transcriptional structure formation *in vitro*. The formation of the ZTP riboswitch that they investigate is known to be kinetically controlled and had to be carefully resolved time-wise in order to be able to determine which folding intermediate is involved in binding the key ligand (alarmone ribonucleotide ZMP) during transcriptional elongation. They achieve this by stalling the RNA polymerase at different positions along the DNA, by SHAPE-probing the thus resulting transcripts of different lengths and by combining this with a high-throughput protocol for reading out the resulting transcript-specific sequences of SHAPE-probing signals.

Their method, however, probes a transcript of less than 200 nucleotides length. The influence of folding kinetics during co-transcriptional folding is typically expected to be a major influence for transcript larger than that (26) which constitutes the majority of primary transcripts for many organisms. The experimental determination of co-transcriptional folding intermediates and their functional role(s) *in vivo* thus remains a real challenge, see for example the long (and alternatively spliced) Xist gene that plays a key role in dosage compensation in mammals and beyond (27).

One orthogonal strategy to the high-throughput, SHAPE-based probing of multiple co-transcriptionally stalled versions of the same emerging transcript is single-molecular (SM) microscopy which investigates one particular molecular at a time, e.g. the recent review (28). This type of microscopy allows for the investigation of even minor effects of ligand binding and its impact on global RNA structure formation.

In same cases, well-established methods such as mutations and genetic selection can be employed to gather evidence for the existence and location of meta-stable structures that are co-transcriptionally formed *in vivo* in case they have distinct functional roles whose absence or presence can be readily assessed *in vivo*, *e.g.* as recently performed by Masachis *et al.* to uncover a new layer of toxin-antitoxin regulation in *Helicobacter pylori* (29).

Large-scale experiments of histone transcripts with a mutated and slower version of the nominal polymerase II combined with chemical RNA structure probing of emerging poly-A+ transcripts reveal that the correct transcription speed is key to their correct pre-mRNA maturation (30). More specifically, Saldi *et al.* show that the slower transcription yields polyadenylated histone mRNAs that extend past

the stem-loop processing site, thereby establishing a causal link between transcription speed, co-transcriptional structure formation and correct mRNA maturation for this specific class of genes.

Beyond this particular class of pol II described genes, co-transcriptional events may play a key role in regulating correct splicing, see (31) for a recent review by the Neugebauer group and (1) for a transcriptome-wide, computational study by us that establishes a causal link between local RNA structure features, RNA editing, alternative splicing and Alu repeats (see also (3)).

Here, we present a new computational method, called COBOLD (in German, a 'Kobold' denotes a mischievous, but mostly well-meaning little devil that play tricks on people) that is capable of identifying candidate transient helices of all functional classes shown in Figure 1 and Table 1 and their relationship to a known reference RNA structure. Depending on the input evidence supplied by the user (i.e. either a single input sequence or a corresponding multiple-sequence alignment (MSA)), our method can readily identify new RNA structure features that can impact the co-transcriptional formation of the reference RNA structure, either in a positive or negative way. When used in a comparative way, i.e. with an input MSA, COBOLD is capable of assigning estimated reliability values (i.e. *P*-values) that allow for the filtering and ranking of the predicted transient features. We investigate the merits of COBOLD for five biological cases where functional RNA structures and transient RNA structure features are known and show that our method can be used to swiftly gain valuable insight into RNA structure formation *in vivo* that could otherwise be only obtained via dedicated experiments. Our hope for the future is that COBOLD will help identifying more experimentally confirmed cases of functional transient RNA structure features which can then in turn be used to improve computational prediction methods such as COBOLD. We make COBOLD accessible via two dedicated web-servers at www.e-rna.org where users can run COBOLD analyses and also visualize the COBOLD predictions in a number of ways.

COBOLD is the first method of its kind and will hopefully aid in identifying transient features that are difficult to investigate experimentally while also playing key functional roles in diverse biological systems *in vivo* ranging from a viral, bacterial to eukaryotic.

MATERIALS AND METHODS

We here describe our new computational method COBOLD which can either be used in a non-comparative or comparative way depending on the input evidence that the user can supply for the RNA transcript of interest.

Non comparative part of COBOLD

Without any evidence on orthologous transcripts from evolutionary related species, COBOLD takes the sequence of a single RNA transcript as input and calls a highly modified version of COFOLD (18). In its original implementation, the RNA structure predicted by COFOLD is derived from a 2D matrix which stores the entries of the underlying dynamic-programming algorithm via a traceback-procedure which

Figure 4. Two-dimensional dynamic programming matrix underlying the algorithm of the non-comparative part of COBOLD. Both axes correspond to an input sequence of length L. Each $T_i = (1, i)$ entry in this matrix corresponds to the optimal RNA structure S_i that is co-transcriptional formed for a sub-sequence L_i, $L_i \le L$, of the original input sequence. By starting the traceback procedure at field $T_i = (1, i)$ of this matrix, we can retrieve S_i by subsequently following the structural information stored in each subsequent node, which also points to the following one in the path. At the end, the RNA secondary structure for sub-sequence L_i is obtained.

starts at the left top corner of this matrix depicted in Figure 4. In order to identify relevant transient RNA structure features that could clash with the reference RNA structure, we modified the original traceback procedure of COFOLD so that it is now capable of collecting information on all relevant transient RNA structures of the different functional classes shown in Figure 1 in a *single* traversal of the optimisation matrix shown in Figure 4.

The key idea behind the new trace-back procedure is to realise that this matrix not only contains information on the optimal co-transcriptionally folded RNA structure for the fully extended input sequence of length L, but also for all sub-sequences $L_i := (x_1, x_2, \ldots, x_i)$ of length $L_i \le L$, see Figure 4. The new trace-back procedure traverses all optimal co-transcriptionally folded RNA structures S_i for all sub-sequences L_i and thereby gathers information on all different classes of transient helices within these S_i that could impact the formation of the known reference RNA structure for length L. The minimum length required for a transient helix is a global parameter in COBOLD that can be adjusted by the user (default value is 2 bp).

Overall, the non-comparative part of COBOLD thus has the same time-and-memory requirements as the original COFOLD—namely $\mathcal{O}(L^3)$ time and $\mathcal{O}(L^2)$ memory for calculating the matrix shown in Figure 4—and requires only $\mathcal{O}(L^2)$ time for the modified traceback procedure for an input sequence of length L. The non-comparative part of COBOLD thus corresponds to a computationally efficient

tool for identifying different functional classes of potentially relevant transient RNA structure features that could impact a given reference RNA structure.

Comparative part of COBOLD

If the user supplies a multiple-sequence alignment (MSA) containing the transcript of interest as well as a corresponding reference RNA structure as input to COBOLD, it becomes conceptually possible to computationally detect and rank transient helices based on their degree of evolutionary conservation.

In order to detect potential transient helices, COBOLD utilises TRANSAT (33), a fully probabilistic method contributed by us earlier that is capable of detecting conserved helices of functional RNA structures, including transient, pseudo-knotted and alternative structures. TRANSAT takes as input a MSA of homologous sequences and a corresponding evolutionary tree (if the tree is not known, it can be readily inferred by TRANSAT via a maximum-likelihood method). Briefly, TRANSAT first calculates all possible candidate helices for each un-gapped sequence in the input alignment. These features are then mapped onto the input MSA and the corresponding base-paired alignment columns considered as candidate helices. For each candidate helix within the MSA, TRANSAT then calculates a so-called log-likelihood score which employs two dedicated probabilistic models of evolution in order to quantitatively compare the likelihood of two competing hypothesis: one, that the pair of alignment columns evolved as base-pairs, two, that the same set of alignment columns evolved independently. Finally (and optionally), TRANSAT estimates *P*-values for the the predicted helices in order to take into account that different input MSAs have different overall propensities to form spurious helices. This propensity depends on a number of complex features of the input alignment such as the dinucleotide composition within the sequences and the gap-pattern. For this, it generates a sufficiently large number of (carefully) randomised versions of the original input alignment in order to determine a corresponding null-distribution of log-likelihood scores. This distribution is then used to assign *P*-values to the log-likelihood scores of the helices detected in the original input MSA.

For COBOLD, two features of TRANSAT are particularly noteworthy. First, the fact that TRANSAT is a fully probabilistic method that does not employ the MFE strategy and that is capable of assigning reliability values to the detected helices. Second, that its performance has been shown to be independent of the length of the input alignment. Moreover, it is computationally efficient, requiring only $\mathcal{O}(N \cdot L^2)$ time for an input alignment of N sequences and length L. Given a known reference RNA structure and an input alignment of decent quality, it is thus possible to assess the presence and potential impact of conserved transient helices irrespective of sequence length.

As COBOLD takes a given reference RNA structure as input, the conserved helices identified by TRANSAT can be readily assessed with respect to the given reference structure and classified into the categories of clashing helices specified in Figure 1. In the current implementation of COBOLD, users can choose to report only the statistically most significant transient helix that clashes with any helix of the reference RNA structure. It is also possible to impose an overall threshold value for the statistical significant of any predicted helices, via a *P*-value threshold of 10^{-2}. COBOLD can report the detected helices sorted by functional class, see Figure 1. It is thus for example easy to identify the most significant transient helix of type 5′-CIS that could prevent the formation of a key helix of the reference structure.

The comparative mode of COBOLD allows users to readily explore different evolutionary time-scales on which transient helices have been conserved by utilising different input MSAs capturing different species and evolutionary time scales. This type of analysis can provide valuable additional insight into the potential functional roles of select transient helices.

Finally, it is possible to use the predictions of the comparative part of COBOLD as filter to post-process the predictions of the non-comparative part of COBOLD, e.g. to focus on well-conserved transient helices or transient helices with a significant statistical significance.

Web-server of COBOLD for visualisation

The predictions by COBOLD can offer unique insight into the complex interplay of transient helices and distinct structural features of the reference RNA structure. In order to explore these relationships in a visual and more intuitive manner while also highlighting the underlying quantitative information, we have built two dedicated COBOLD web-servers at www.e-rna.org for running COBOLD analyses and for visualising and manipulating COBOLD predictions. The visualization server is technically based on R SHINY(http://www.rstudio.com/shiny/) (34) and utilises our own R4RNA R-package for RNA structure visualisation (35).

The visualisation tool allows to directly compare the predicted transient helices to the given reference structure. This is achieved via a so-called double-arc plot. Any base-pair is shown as arc (or semi-circle) linking the two base-paired sequence position along the horizontal line which can represent the input sequence or the input alignment. The double-arc plot shows the features of the reference RNA structure on top of the horizontal line and the relevant predicted features below, see e.g. Figure 5. The user can then click on key structural features of the reference RNA structure to highlight the relevant predicted transient features. When visualising predictions of the comparative mode of COBOLD, the user can impose a *P*-value threshold.

When considering predictions of both modes in COBOLD, both sets of predictions can be combined, for example in order to focus on helices with a high degree of evolutionary conservation. Moreover, users can focus on transient helices of a distinct functional class, e.g. those that are potentially disruptive (e.g. 5′-CIS) or particularly beneficial (e.g. 3′-TRANS). And finally, the web-server allows for the download of the resulting high-quality figures in pdf and png format that can be readily integrated into publications.
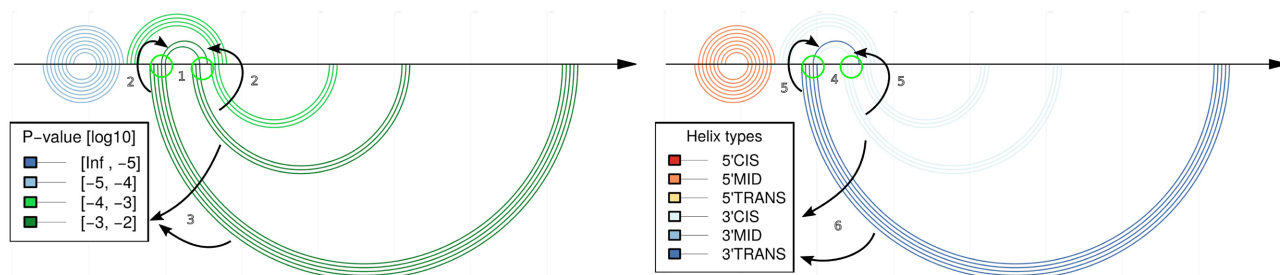
Figure 5. Example figure of an arch plot: The horizontal line represents the input MSA. Each arc or semi-circle corresponds to a single base-pair connecting the corresponding two sequence positions. Arcs on top represent the reference RNA structure, in this case the permissive conformation of the Levivirus. Arcs below the horizontal line depict the transient helices predicted by the comparative mode of CoBold. For each shown transient helix, it is easy to identify the corresponding helix or helices it is clashing with (in this case indicated by the arrows) as these share base-paired positions along the horizontal line, see the feature highlighted by the circles and arrows (1–2, 4–5). The left part of the figure shows the predicted helices and the reference helix/helices they are clashing with in the known structure colored by statistical significance (i.e. *P*-value) (5) whereas the right part of the figure shows them colored by functional class (6). In this case, the most prominent feature is thus the predicted transient helix comprising four base-pairs of high statistical significance (see dark green color and legend, as indicated by the arrow on the bottom left) and of beneficial type (i.e. 3′-TRANS) (see dark blue color and legend on the bottom right) which may aid in the formation of the 3′ helix of the reference RNA structure conformation.

RESULTS

We focus the evaluation of CoBold on five select data sets whose transient features and known RNA structures have been experimentally confirmed and whose functional roles have been well established (25,36). These examples showcase how functional features and regulatory feedback loops can be encoded into sequences via structural features such as helices that are mutually exclusive. The overview below includes information on the respective accession numbers of the reference sequences for which the reference RNA structures are known as well as the IDs of the multiple sequence alignments employed.

- HDV ribozyme (Figure 6):
  The Hepatitis Delta Virus (HDV) is a single-stranded (ss) RNA virus that aggravates symptoms of human hepatitis. This particular ribozyme regulates the self-cleaving process required for viral replication. This is achieved via two different structural configurations that either allow or prevent replication depending on the environmental conditions. In the active state, the region on the sequence where the self-cleaving happens is exposed, so HDV can undergo the process and replicate. During poor conditions for replication, however, the self-cleaving region is sequestered by a hairpin, thereby preventing replication. Notwithstanding, if optimal replication conditions are achieved once more, the 5′ region of the sequestering hairpin is itself sequestered, so the self-cleaving region is exposed again and the replication can occur. The HDV ribozyme sequence has a length of 152 nt. In this particular case, we employed the alignment RF00094 from RFAM (37), with M28267.1 635-775 as reference sequence.

- 5′-UTR of Leviviridae Levivirus (Figure 7):
  The Levivirus is a ssRNA bacteriophage, i.e. a virus that kills certain bacteria. Its 5′ UTR controls the translation of its proteins by regulating the binding of the ribosome and the replicase, which compete for the same region. If the Levivirus is not in need of protein translation, the ribosomal binding site is sequestered by an upstream complementary region and trapped in a helix. However, if protein translation is needed, this complementary region is paired with a different region further upstream in the sequence, thereby releasing the originally mentioned region and enabling ribosome binding. This particular sequence has a length of 158 nt. For this analysis we used a custom alignment, choosing GQ153927.1 1-132 as reference sequence.

- SAM riboswitch (Figure 8):
  This particular riboswitch is a structure widely shared among bacteria which allows for a switch in gene expression depending on the presence of a certain coenzyme (in this case, *S*-adenosylmethionine (SAM)). With SAM binding, the riboswitch adopts an RNA structure that stops transcription. This structure comprises two helices, one called anti-anti-terminator and one called terminator (in 5′ to 3′ order). Without SAM binding, this RNA structure is partly re-arranged. In this rearrangement, the two anti–anti-terminator and terminator helices are destroyed and converted into a single new helix that pairs the 3′ arm of the former anti–anti-terminator helix and the 5′ arm of the former terminator helix into a new anti-terminator helix. This anti-terminator helix allows for transcription to proceed. The SAM riboswitch has a length of 215 nt. In this particular study we employed a custom alignment, with AL009126.3 1258276-1258464 as reference sequence.

- Tryphtophan (Trp) operon leader (Figure 9):
  This structure regulates the transcription of the trp operon in *E. coli*. Similarly to the above examples, it senses if the optimal conditions for transcription are given or not. In this case, however, the region is responsive to the transcription speed itself. While transcription is being conducted by the ribosome as its nominal speed, a terminator hairpin forms which stops transcription. If conditions are not optimal for transcription and the ribosome stalls, an anti-terminator structure emerges that sequesters the terminator which in turns allows the ribosome to transcribe the whole sequence, instead of just a partial one. The sequence of the trp operon leader has a length of 152 nt. For the purpose
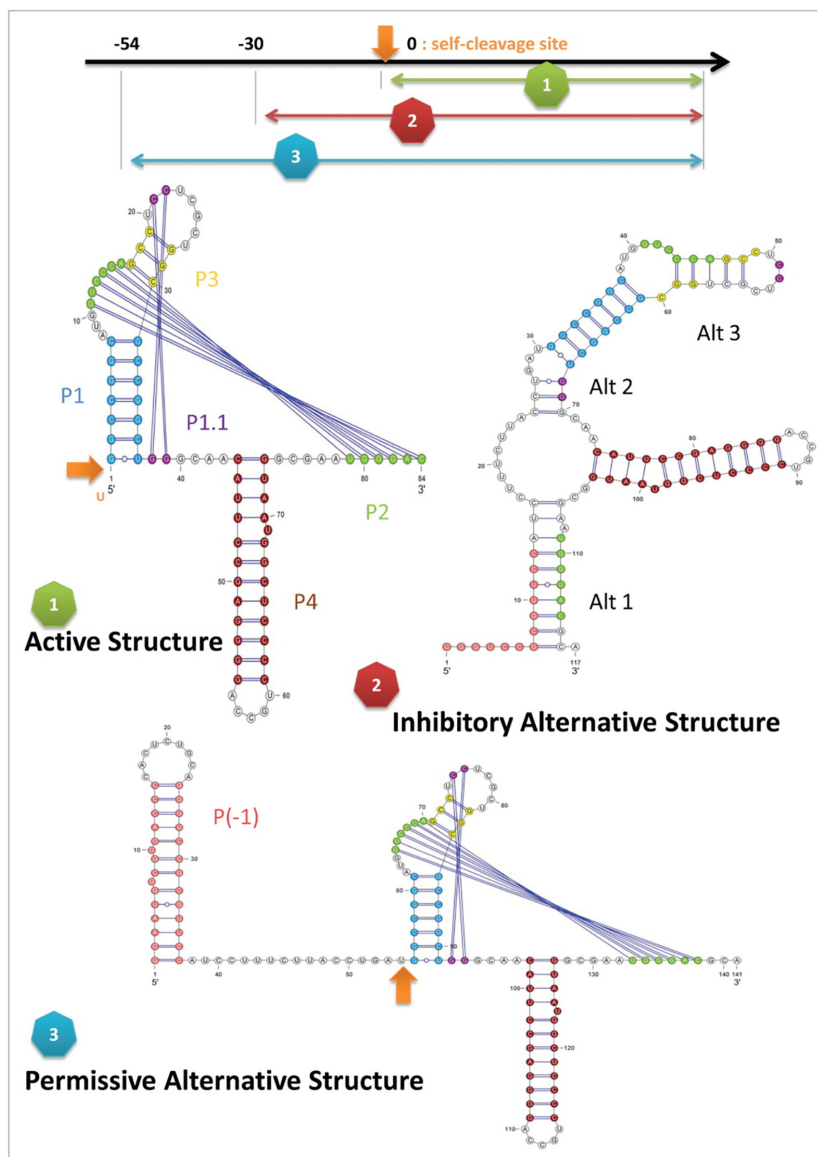
Figure 6. Three different functional conformations the HDV ribozyme: the active structure, the inhibitory structure and the permissive alternative structure. The self-cleaving site is indicated by the orange arrow, being exposed in the active and permissive structure and sequestered in the inhibitory one. Picture from (36).

of this analysis we used RFAM alignment RF00513 with AE005174.2 2263095-2263188 as reference.

- ZTP riboswitch (Figure 10):

This structure switch regulates transcription in bacteria via the presence or absence of the purine bio-synthetic intermediate 5-aminoimidazole-4-carbozamide riboside 5′-triphosphate (ZTP) molecule. Here, we focus on the riboswitch appearing in the bacterium *Clostridium beijerinckii* whose structure is commonly designated as *Clostridium beijerinckii pfl*. This riboswitch's structure was recently studied in great detail experimentally (25) and its cofolding path was extracted by analysing the SHAPE values of its different sub-sequences. We wanted to check if our proposed method is able to recover that particular structure that was extracted experimentally with a significant amount of effort and time. The ZTP ri-

boswitch we use has a length of 172 nt. In this study, we employed an improved version of the RF01750 alignment from (25) with CP000721.1 1211931-1212029 as reference sequence. Our improvement consists of removing all-gap columns from the alignment, as well as sequences that do not contribute to the overall evolutionary information contained in the alignment.

In the following, we focus on two of the above examples to highlight that COBOLD can not only help in identifying these structural features for a given reference RNA structure, but also aid in their functional interpretation. Please refer to the supplementary information (SI) for a detailed discussion of the other systems (Supplementary Figures S3–S5), where their corresponding results (Supplementary Figures S6–S25) will be discussed in a similar manner as we will
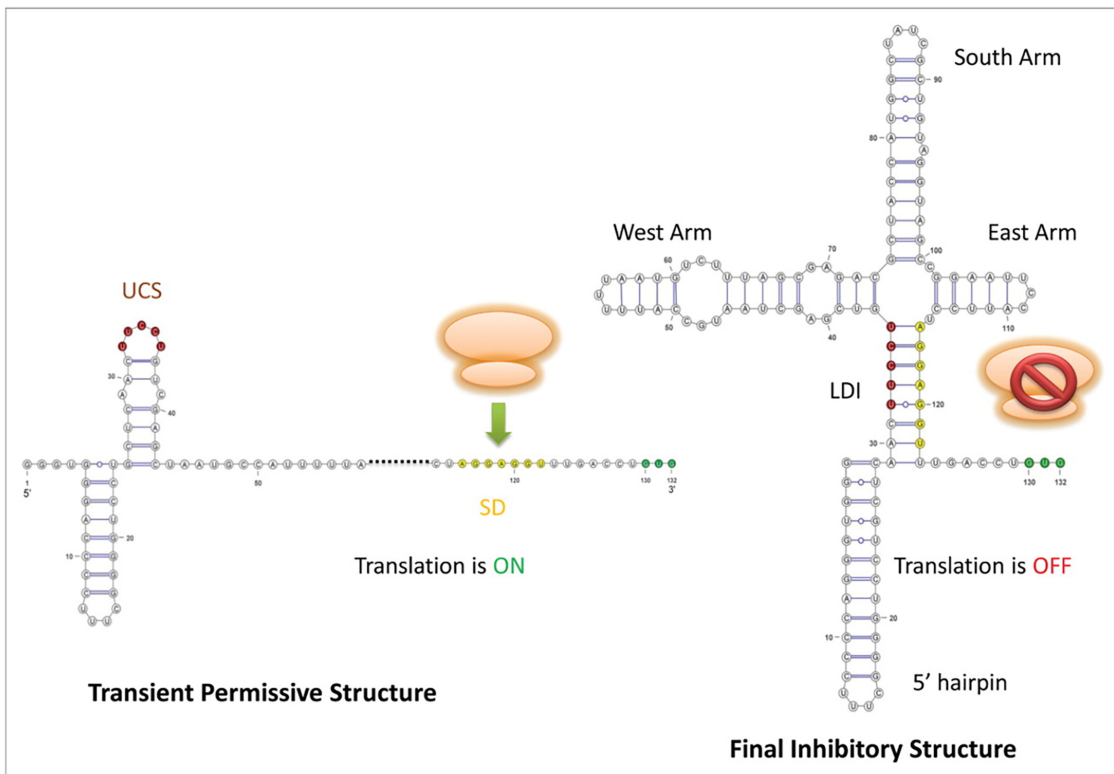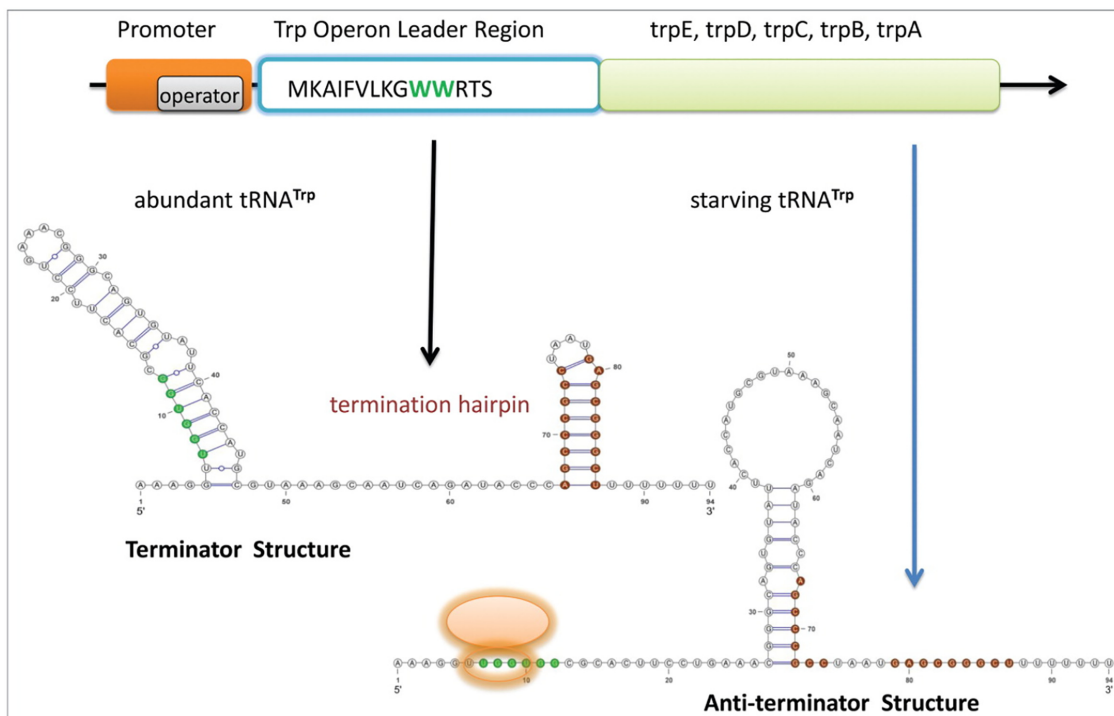
Figure 7. Two functional conformations of the 5′ UTR of the Leviviridae Levivirus: the inhibitory conformation and the permissive alternative conformation. The ribosome binding site (SD) site is indicated by the yellow nucleotides, being exposed in the permissive structure and sequestered in the inhibitory one. Picture from (36).
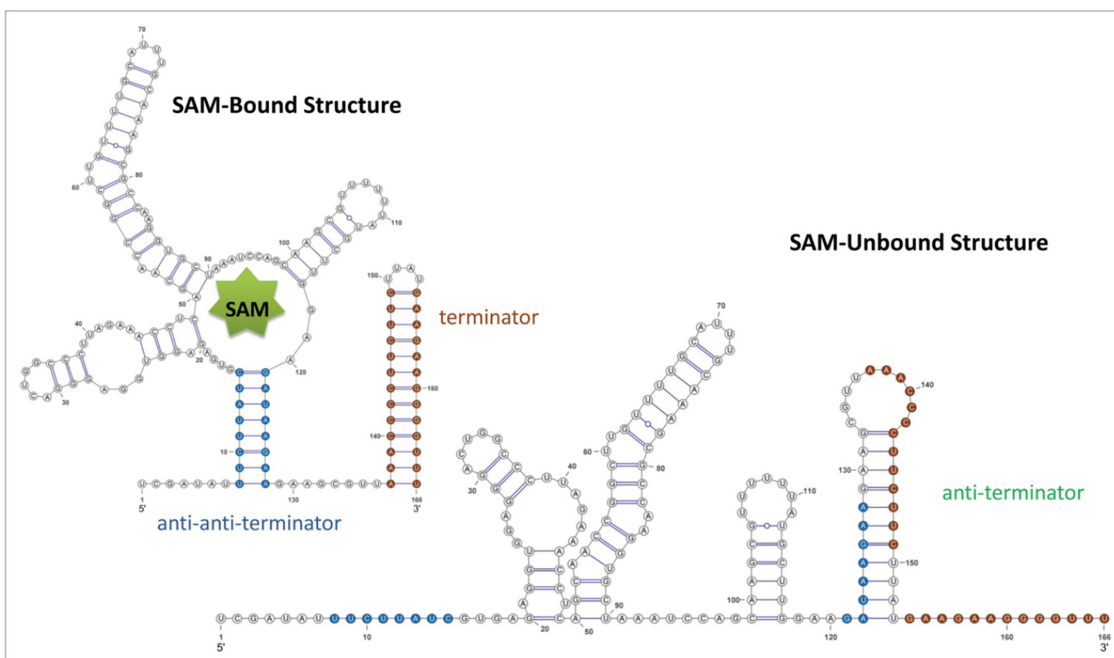

Figure 8. Two functional conformations of SAM-responsive riboswitch: the bound structure and the unbound structure. The terminator, anti-terminator and anti-anti-terminator hairpins are indicated in brown, green and blue, respectively. It can be seen how the terminator is sequestered in the unbound state and free at the bound one. Picture from (36).

Figure 9. Two functional conformations of the tryphtophan (TRP) operon leader: the terminator structure and the anti-terminator structure. The ribosome binding site and anti-terminator hairpin are indicated in green and brown respectively. It can be seen how the binding site is sequestered in the terminator state and free and accessible at the anti-terminator one, where the terminator is sequestered by the anti-terminator. Picture from (36).
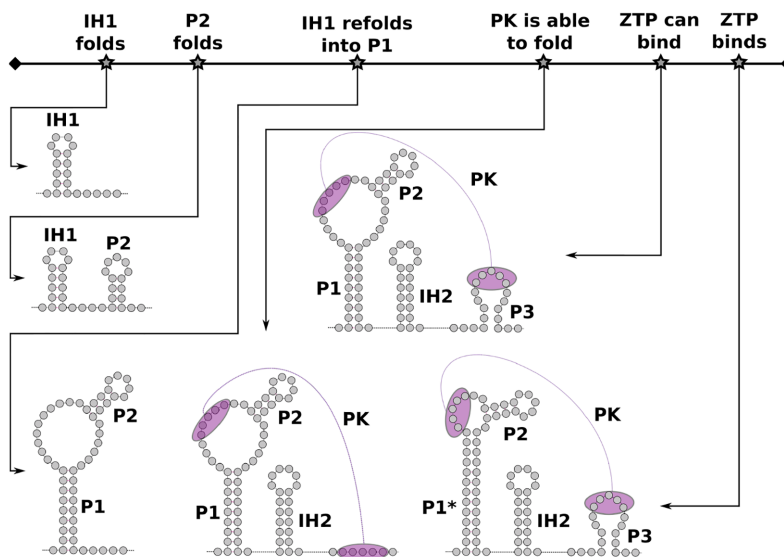
Figure 10. Co-transcriptional folding pathway of the ZTP riboswitch: Picture showing the cofolding path as well as the different conformations for the ZTP riboswitch for the ZTP-bound and the ZTP-unbound state. The time-line at the top outlines the relevant structural steps happening during transcription with the corresponding RNA structures shown below. Figure adapted from (25).

do in the following text with the two selected ones (Supplementary Figures S1 and S2). For each case, we proceed as follows:

(1) As input to CoBOLD, we use the reference RNA sequence, a corresponding MSA and evolutionary tree as well as a reference RNA structure.

(2) Using the comparative mode of CoBOLD, we detect all evolutionarily conserved helices that are clashing with the reference RNA structure, rank them in terms of significance (i.e. by *P*-value) and sort them by helix type, see Figure 1.

(3) Using the non-comparative mode of CoBOLD, we detect all co-transcriptionally emerging transient helices

that are clashing with the reference RNA structure, rank them in terms of stability and sort them by helix type. The stability of a helix is measured in terms of its minimum-free energy.

(4) Depending of the desired analysis and its stringency (e.g. in terms of *P*-value threshold imposed by the user), we either consider the predictions of the non-comparative and comparative mode separately or, alternatively, compute the intersection of helices predicted by both modes in COBOLD.

(5) When the different clashing helices have been predicted, we interpret their relevance and potential functional role in terms of their type (see Table 1), as they could either prevent or perturb the formation of the reference structure or, on the contrary, aid its formation. In addition, we can quantitatively assess the extent of their evolutionary conservation and, thus, likely functional relevance via their estimated *P*-values, for cases when the user can supply an input alignment for the sequence of interest.

Detecting evolutionary-conserved clashing helices

The 5′ UTR of the Leviviridae Levivirus has two distinct functional conformations, the inhibitory and permissive conformation, see Figures 11 and 14. They share some structural features, but contain contain mutually exclusive helices. Using the comparative mode of COBOLD, we can identify a conserved transient helix that is clashing with the region located near the UCS, which is a requirement for the formation of the cruciform structure of the inhibitory conformation. This particular helix is threatening both functional conformations of the Levivirus. When we consider only the most significant clashing helices (see Figures 13 and 15), we find that these, well-conserved helices are typically of the beneficial type (3′-CIS) or neutral (5′-TRANS or 3′-TRANS) and may thus explicitly aid in the formation of the known functional configuration. The helix clashing with the 5′ hairpin is the only one that clashes with a structure shared by both conformations. It is officially classified as having a neutral effect, but—due to its significance in terms of *P*-value—may play a beneficial role by protecting the formation of that specific helix. There are only some neutral helices competing with some nucleotides in the cruciform, mostly at the west and south arms.

For the other four functional families, the observations are mostly similar to the ones for the Levivirus. In particular, when focusing on the most significant helices, most of the predicted transient helices are of the beneficial or neutral type.

Detecting features of the cofolding pathway of the ZTP riboswitch

The co-transcriptionally folding pathway of the ZTP riboswitch was recently studied experimentally in great details (25). When analysing the ZTP riboswitch with the comparative mode of COBOLD and zooming into significant helices with a maximum *P*-value of 10^{-3}, we recover all the relevant regions of the functional structure with only a few false positives, see Figure 12. Only two structural features are missing: two small, independent helices which have

not been evolutionary conserved. These two structural elements, even though they are involved in the formation of the final structure, are the least biologically relevant as they (1) do not fulfil an explicit known biological role and (2) are not directly involved in the formation of the pseudonot and the functioning of the riboswitch. COBOLD thus allows us to recover the key features from the recent, detailed experimental investigation based on sequence-based information from a multiple-sequence alignment alone.

Detecting co-transcriptionally relevant transient helices

We now consider the insight that can be gained from considering the non-comparative predictions of COBOLD. Figures 16 and 17 (left side) showcase the resulting predictions for both known functional conformations of the Levivirus. We can readily identify the known helices that appear during transcription and that interfere with the known reference structure, most of which are evolutionarily well conserved.

The majority of the identified transient helices belong to the neutral or beneficial kind and disruptive helices (5′-CIS) are almost absent. On closer inspection, i.e. when comparing these non-comparative with the comparative results of COBOLD), these disruptive helices turn out to be not evolutionarily conserved and can thus be expected to not play a distinct functional role on the investigated evolutionary time-scale.

The most prevalent transient helices are of type 5′-MID, 3′-CIS and 3′-TRANS for the inhibitory Levivirus confirmation, which actively promote the co-transcriptional formation of the known helices and which are often found between the two known structural conformations.

For the other four families, we observe similar findings. The vast majority of the identified transient helices are beneficial or neutral with respect to the known structural conformations. A fair number of these is located between the mutually exclusive structures, thereby effectively serving as potential stepping-stones on a refolding pathway. One notable exception is the TRP operon leader, for which COBOLD identifies an unusual number of disruptive (5′-TRANS) transient helices that are very well conserved. These helices may potentially serve as highly-sensitive sensors that may prevent the formation of the nominal RNA confirmation unless the correct *in vivo* conditions are met. The TRP operon leader is also unusual in the sense that most of the transient features identified by the non-comparative mode are evolutionary well conserved and in the sense that most of these features are predicted to affect both known reference RNA structures. COBOLD is also capable of identifying the small transient helix that features during the co-transcriptional folding of the long *E. coli* 23S rRNA that was experimentally identified using SPET-SEQ by Incarnato *et al.* in 2017 (24).

DISCUSSION

Exciting recent research from diverse biological systems reveal that RNA structure features are responsible for regulating diverse and important steps in gene expression within the transcriptome. Yet, how these RNA structures form and
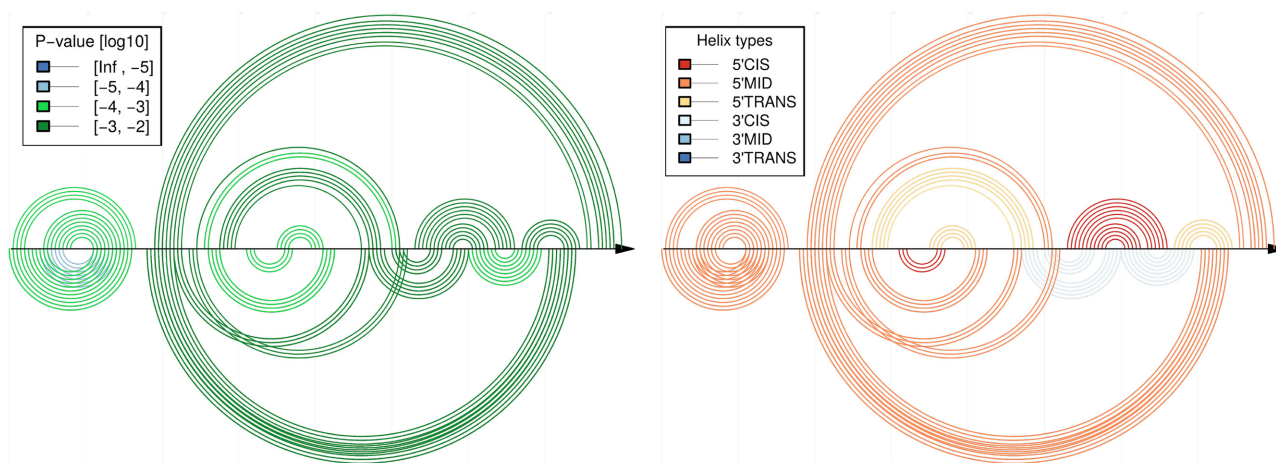
Figure 11. Comparative COBOLD predictions for the inhibitory conformation of the Levivirus with the inhibitory RNA conformation shown on top and the predicted transient helices shown below the horizontal line, once colored in terms of statistical significance (left) and once in terms of type of transient helix, see also the previous figure for more general information on this type of plot.
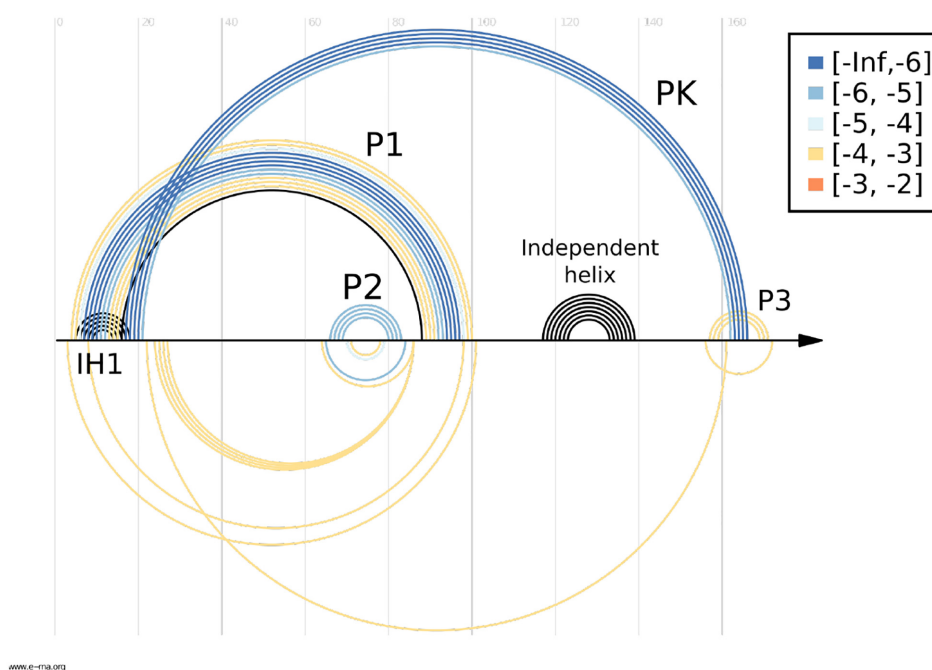

Figure 12. Sensitivity-specificity plot for the comparative COBOLD predictions for the ZTP riboswitch for a *P*-value threshold of 10^{-3}. The top of this figure shows the known RNA structure features. All non-black features correspond the features that have been correctly predicted by COBOLD. These are colored depending on the statistical significance of the corresponding predicted helix. Only the independent helix is missing from the COBOLD predictions. The bottom part of the figure shows only a few predicted base-pairs that are not part of the known RNA structure. The bottom part of the figure thus illustrates the high specificity of COBOLD (which would be perfect if no features were shown below the line), whereas the top part of the figure thus conveniently summarises the high sensitivity of the COBOLD predictions (hence the name sensitivity-specificity plot for this kind of figure). Overall, COBOLD thus readily recovers the experimentally confirmed RNA structure features that matter for the co-transcriptional folding pathway of the ZTP-riboswitch with high sensitivity an specificity.

change in a living biological system *in vivo* remains poorly understood and particularly challenging to study, both experimentally and computationally.

We here present a new computational method, called COBOLD, that is capable of identifying different functional classes of transient features that may play distinct roles in guiding or preventing the formation of a known reference RNA structure *in vivo*. COBOLD bases its predictions on sequence information only, which is either supplied by the user in terms of an single RNA structure (non-comparative mode of COBOLD) or a corresponding multiple sequence alignment of evolutionarily related species (comparative mode of COBOLD). COBOLD employs two distinct computational algorithms to generate predictions in a comparative or non-comparative way. The non-comparative mode of COBOLD explicitly takes the overall effects of co-
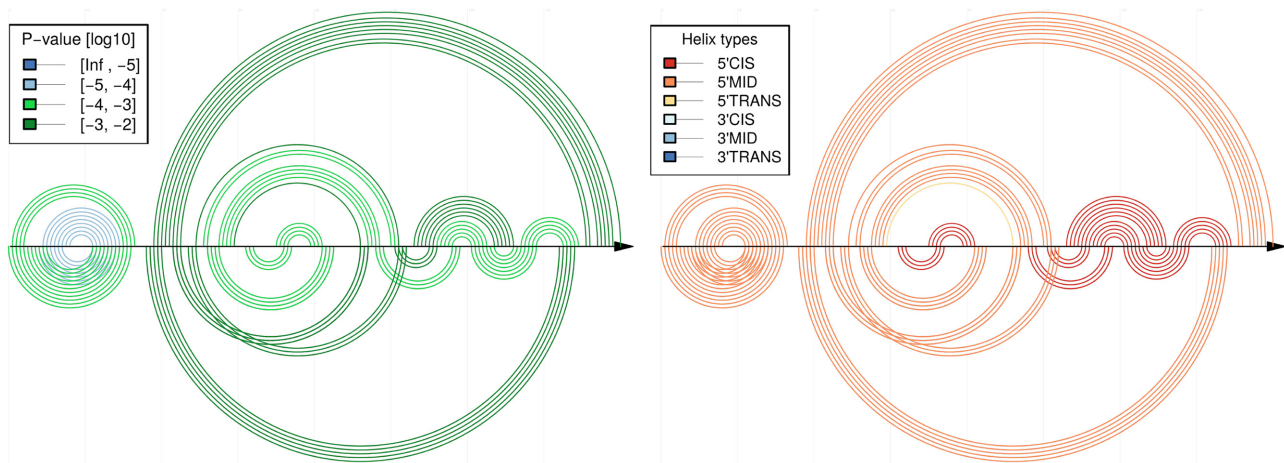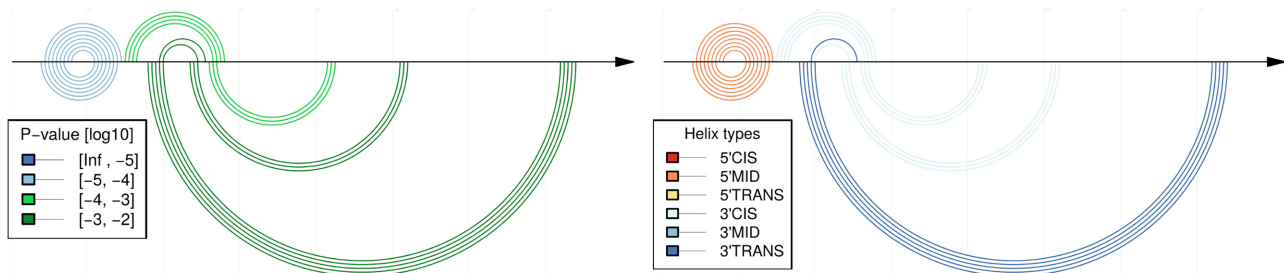
Figure 13. Comparative CoBold predictions of minimal *P*-value for the inhibitory conformation of the Levivirus. Compared to the similar, earlier Figure 11, this figure only shows the statistically most significant transient helix for each helix of the known reference RNA conformation which is shown on top, as usual. It thus represents a reduced version of the predictions in Figure 11 and allows to quickly assess the most relevant features viewed from the perspective of helix in the reference RNA structure.
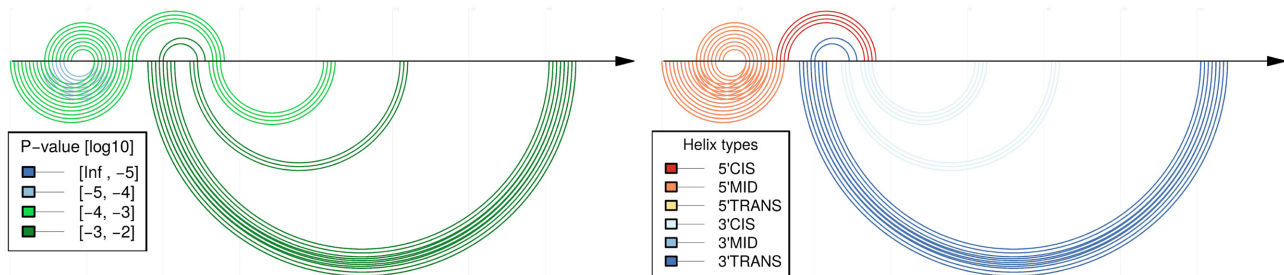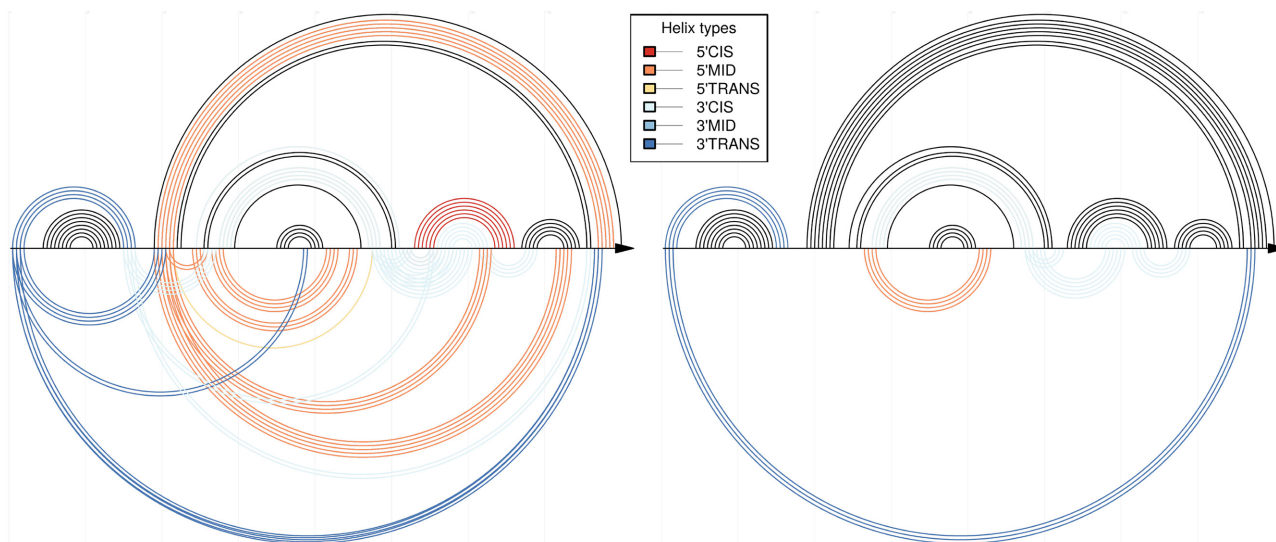

Figure 14. Comparative CoBold predictions for the permissive conformation of the Levivirus, see Figure 11 for the corresponding results for the inhibitory conformation of the Levivirus. As before, the top shows the reference RNA structure features and the bottom the predicted transient helices.
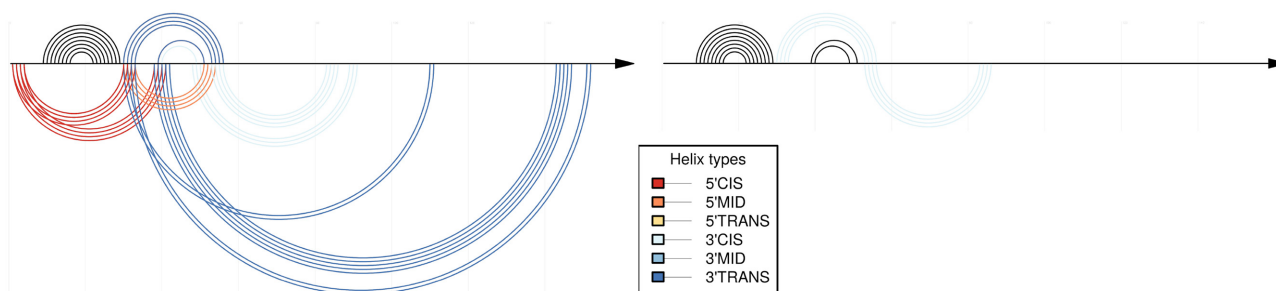

Figure 15. Comparative CoBold predictions of minimal *P*-value for the permissive conformation of the Levivirus. As before for Figures 11 and 13 for the inhibitory conformation, this figure highlights only the statistically most significant transient helix for each helix of the known RNA confirmation when compared to the predictions shown in previous Figure 14. As before, the top shows the reference RNA structure features and the bottom the predicted transient helices.

transcriptional folding into account when generating its predictions and has been shown to generate high-quality RNA structure predictions, especially for long RNA transcripts for which the kinetic effects of co-transcriptional folding are expected to play a prominent role. The comparative mode of CoBold utilises a probabilistic method as well as dedicated, quantitative models of evolution to identify evolutionarily conserved transient helices and is also capable of estimating the reliability of the predicted helices. Provided that the a multiple-sequence alignment of decent quality can be supplied by the user, the comparative mode

of CoBold should provide the most reliable prediction of transient helices with a potential functional role. Compared to the non-comparative mode of CoBold, it should also allow the identification of transient features that rely on *trans* interactions with ligand or other molecules which is something that the non-comparative mode of CoBold is conceptually incapable of modelling. Both modes of CoBold allow the user to generate predictions in a computationally efficient way, both in terms of memory and time, thereby readily enabling the analysis of transcripts that are significantly long than 200 nt. Predictions by CoBold can be eas-

Figure 16. Non-comparative CoBold predictions for the inhibitory conformation of the Levivirus. Compared to Figure 11, this figure shows the predictions of CoBold that are based on the single RNA input sequence alone rather than an input multiple-sequence alignment. The right part of this figure shows the same predictions when also filtered by the comparative predictions and thus highlights only the most evolutionarily conserved transient features that are predicted by the non-comparative mode by CoBold. Note that coloring by *P*-value is not possible when using the non-comparative mode of CoBold, hence only the coloring by type of transient helix. As before, the top shows the reference RNA structure features and the bottom the predicted transient helices.

Figure 17. Non-comparative CoBold predictions for the permissive conformation of the Levivirus. Compared to Figure 14, this figure shows the predictions of CoBold that are based on the single RNA input sequence alone rather than the corresponding input multiple-sequence alignment. The right part of this figure shows the same predictions when also filtered by the comparative predictions which does not yield any evolutionarily conserved features. Note that coloring by *P*-value is not possible when using the non-comparative mode of CoBold, hence only the coloring by type of transient helix. As before, the top shows the reference RNA structure features and the bottom the predicted transient helices.

ily generated and visualised via two dedicated web-servers at www.e-rna.org which allow for an intuitive as well as quantitative in-depth analysis of the predicted transient helices and their detailed impact on features the known reference RNA structure.

The CoBold analysis for five biological families with multiple known RNA structure configurations and known transient features shows that CoBold is able to identify known transient features and their likely impact on the formation of the known RNA structure configuration(s) for a varied arrangement of systems. Most families encode evolutionarily well-conserved transient helices that could explicitly aid the formation of the known RNA structure or serve as stepping-stones on the folding trajectory between two distinct RNA structure conformations. Our analysis of the TRP operon also discovered a number of well-conserved transient features that could easily hinder the formation of the known RNA structure confirmation(s). These features

may serve as fine-tuning regulators *in vivo* that may prevent the formation of the known RNA structure unless the correct cellular conditions are met.

As CoBold provides very detailed insight into the transient helices, their level of evolutionary conservation and their likely impact on distinct features of a known RNA structure confirmation, we hope that the predictions of CoBold will not only aid the experimental identification of novel transient features, but will also significantly facilitate the investigation of their functional roles *in vivo* and help to improve computational prediction methods such as CoBold. Finally, with the exciting, recent advent of nucleotide-based therapeutics for a range of human diseases, we also hope that CoBold can significantly aid in the design of therapeutic agents to engineer co-transcriptional folding pathways *in vivo* or to prevent or encourage the formation of specific RNA features with distinct functional roles (39–41,41).

SUPPLEMENTARY DATA

FUNDING

REFERENCES

1. Mazloomian,A. and Meyer,I.M. (2015) Genome-wide identification and characterisation of tissue-specific RNA editing events in D. melanogaster and their potential role in regulating alternative splicing. *RNA Biol.*, **12**, 1391–1401.
2. Cugusi,S., Li,Y., Jin,P. and Lucchesi,J.C. (2016) The Drosophila helicase MLE targets hairpin structures in genomic transcripts. *Plos Genet.*, **12**, e1005761.
3. Chillon,I. and Pyle,A.M. (2016) Inverted repeat Alu elements in the human lincRNA-p21 adopt a conserved secondary structure that regulates RNA function. *Nucleic Acids Res.*, **44**, 9462–9471.
4. Liu,S.R., Hu,C.G. and Zhang,J.Z. (2016) Regulatory effects of cotranscriptional RNA structure formation and transitions. *WIRES RNA*, **7**, 562–574.
5. Bjork,P. and Wieslander,L. (2015) The Balbiani Ring Story: Synthesis, Assembly, Processing, and Transport of Specific Messenger RNA-Protein Complexes. In: Kornberg,R.D. (ed). *Annual Review Of Biochemistry*. Vol. **84**, pp. 65–92.
6. Hein,P.P., Kolb,K.E., Windgassen,T., Bellecourt,M.J., Darst,S.A., Mooney,R.A. and Landick,R. (2014) RNA polymerase pausing and nascent-RNA structure formation are linked through clamp-domain movement. *Nat. Struct. Mol. Biol.*, **21**, 794–802.
7. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem. Chem. Month.*, **125**, 167–188.
8. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
9. Meyer,I.M. and Miklós,I. (2007) SimulFold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. *PLoS Comput. Biol.*, **3**, e149.
10. Geis,M., Flamm,C., Wolfinger,M.T., Tanzer,A., Hofacker,I.L., Middendorf,M., Mandl,C., Stadler,P.F. and Thurner,C. (2008) Folding kinetics of Large RNAs. *J. Mol. Biol.*, **379**, 160–173.
11. Danilova,L.V., Pervouchine,D.D., Favorov,A.V. and Mironov,A.A. (2006) RNAKINETICS: a web server that models secondary structure kinetics of an elongating RNA. *J. Bioinform. Comput. B.*, **04**, 589–596.
12. Xayaphoummine,A., Bucher,T. and Isambert,H. (2005) Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Res.*, **33**, W605–W610.
13. Gardner,P.P. and Giegerich,R. (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, **5**, 140.
14. Knudsen,B. and Hein,J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.*, **31**, 3423–3428.
15. Pedersen,J.S., Forsberg,R., Meyer,I.M. and Hein,J. (2004) An evolutionary model for protein-coding regions with conserved RNA structure. *Mol. Biol. Evol.*, **21**, 1913–1922.
16. Rivas,E., Clements,J. and Sean,E.R. (2016) A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat. Methods*, **1**, 45–48.
17. Morgan,S.R. and Higgs,P.G. (1998) Barrier heights between ground states in a model of RNA secondary structure. *J. Phys. A: Math. Gen.*, **31**, 3153–3170.
18. Proctor,J.R. and Meyer,I.M. (2013) C o F old : an RNA secondary structure prediction method that takes co-transcriptional folding into account. *Nucleic Acids Res.*, **41**, e102.
19. Zhu,J.Y.A., Steif,A., Proctor,J.R. and Meyer,I.M. (2013) Transient RNA structure features are evolutionarily conserved and can be computationally predicted. *Nucleic Acids Res.*, **41**, 6273–6285.
20. Meyer,I.M. (2005) Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. *Nucleic Acids Res.*, **33**, 6338–6348.
21. Meyer,I.M. and Miklós,I. (2004) Co-transcriptional folding is encoded within RNA genes. *BMC Mol. Biol.*, **10**, 5–10.
22. Lai,D., Proctor,J.R. and Meyer,I.M. (2013) On the importance of cotranscriptional RNA structure formation. *RNA*, **19**, 1461–1473.
23. Higgs,P.G. (2000) RNA secondary structure: physical and computational aspects. *Q. Rev. Biophys.*, **33**, 199–253.
24. Incarnato,D., Morandi,E., Anselmi,F., Simon,L.M., Basile,G. and Oliviero,S. (2017) In vivo probing of nascent RNA structures reveals principles of cotranscriptional folding. *Nucleic Acids Res.*, **45**, 9716–9725.
25. Strobel,E.J., Cheng,L., Berman,K.E., Carlson,P.D. and Lucks,J.B. (2019) A ligand-gated strand displacement mechanism for ZTP riboswitch transcription control. *Nat. Chem. Biol.*, **15**, 1067–1076.
26. Morgan,S. and Higgs,P.G. (1996) Evidence for kinetic effects in the folding of large RNA molecules. *Annu. Rev. Biophys.*, **105**, 7152–7157.
27. Jones,A.N. and Sattler,M. (2019) Challenges and perspectives for structural biology of lncRNAs-the example of the Xist lncRNA A-repeats. *J. Mol. Cell Biol.*, **11**, 845–859.
28. Ray,S., Chauvier,A. and Walter,N.G. (2019) Kinetics coming into focus: single-molecule microscopy of riboswitch dynamics. *RNA Biol.*, **16**, 1077–1085.
29. Masachis,S., Tourasse,N.J., Lays,C., Faucher,M., Chabas,S., Iost,I. and Darfeuille,F. (2019) A genetic selection reveals functional metastable structures embedded in a toxin-encoding mRNA. *eLife*, **8**, e47549.
30. Saldi,T., Fong,N. and Bentley,D.L. (2018) Transcription elongation rate affects nascent histone pre-mRNA folding and 3 ' end processing. *Gene. Dev.*, **32**, 297–308.
31. Herzel,L., Ottoz,D.S.M., Alpert,T. and Neugebauer,K.M. (2017) Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. *Nat. Rev. Mol. Cell. Biol.*, **18**, 637–650.
32. Masachis,S., Tourasse,N.J., Lays,C., Faucher,M., Chabas,S., Iost,I. and Darfeuille,F. (2019) A genetic selection reveals functional metastable structures embedded in a toxin-encoding mRNA. *eLife*, **8**, e47549.
33. Wiebe,N.J.P. and Meyer,I.M. (2010) Transat—a method for detecting the conserved helices of functional RNA structures, including transient, pseudo-knotted and alternative structures. *PLoS Comput. Biol.*, **6**, e1000823.
34. RStudio, Inc (2013) In: Easy web applications in R.
35. Lai,D., Proctor,J.R., Zhu,J.Y.A. and Meyer,I.M. (2012) R- chie : a web server and R package for visualizing RNA secondary structures. *Nucleic Acids Res.*, **40**, e95.
36. Zhu,J.Y.A. and Meyer,I.M. (2015) Four RNA families with functional transient structures. *RNA Biol.*, **12**, 5–20.
37. Kalvari,I., Argasinska,J., Quinones-Olvera,N., Nawrocki,E.P., Rivas,E., Eddy,S.R., Bateman,A., Finn,R.D. and Petrov,A.I. (2017) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.*, **46**, D335–D342.
38. Galloway,K.E., Franco,E. and Smolke,C.D. (2013) Dynamically reshaping signaling networks to program cell fate via genetic controllers. *Science*, **341**, 1235005–1235005.
39. McKeague,M., Wong,R.S. and Smolke,C.D. (2016) Opportunities in the design and application of RNA for gene expression control. *Nucleic Acids Res.*, **44**, 2987–2999.
40. Duval,M., Simonetti,A., Caldelari,I. and Marzi,S. (2015) Multiple ways to regulate translation initiation in bacteria: mechanisms, regulatory circuits, dynamics. *Biochimie*, **114**, 18–29.
41. Pascut,D., Bedogni,G. and Tiribelli,C. (2015) Silencing efficacy prediction: a retrospective study on target mRNA features. *Bioscience Rep.*, **35**, e00185.