In silico methods for co-transcriptional RNA secondary structure prediction and for investigating alternative RNA structure expression

Meyer, I.M.

Methods
2017 MAY 01 ; 120: 3-16
2017 APR 20 (first published online: final publication)
doi: 10.1016/j.ymeth.2017.04.009

Publisher: Elsevier

In silico methods for co-transcriptional RNA secondary structure prediction

Irmtraud M. Meyer [*]
Laboratory of Bioinformatics of RNA Structure and Transcriptome Regulation,
Berlin Institute for Medical Systems Biology,
Max Delbrueck Center for Molecular Medicine,
Robert-Rössle-Str. 10,
13125 Berlin-Buch,
Germany
and
Institute of Chemistry and Biochemistry,
Free University,
Thielallee 63,
14195 Berlin,
Germany

24th March 2017

Abstract

RNA transcripts are the primary products of active genes in any living organism, including many viruses. Their cellular destiny not only depends on primary sequence signals, but can also be determined by RNA structure. Recent experimental evidence shows that many transcripts can be assigned more than a single functional RNA structure throughout their cellular life and that structure formation happens co-transcriptionally, i.e. as the transcript is synthesised in the cell. Moreover, functional RNA structures are not limited to non-coding transcripts, but can also feature in coding transcripts.

The picture that now emerges is that RNA structures constitute an additional layer of information that can be encoded in any RNA transcript (and on top of other layers of information such as protein-context) in order to exert a wide range of functional roles. Moreover, different encoded RNA structures can be expressed at different stages of a transcript's life in order to alter the transcript's behaviour depending on its actual cellular context. Similar to the concept of alternative splicing for protein-coding

[*] To whom correspondence should be addressed. Tel: +49–30–9406–3292; Fax: +49–30–9406–3291, Email: irmtraud.meyer@cantab.net

genes, where a single transcript can yield different proteins depending on cellular context, it is thus appropriate to propose the notion of *alternative RNA structure expression* for any given transcript.

This review introduces several computational strategies that my group developed to detect different aspects of RNA structure expression *in vivo*. Two aspects are of particular interest to us: (1) RNA secondary structure features that emerge during co-transcriptional folding and (2) functional RNA structure features that are expressed at different times of a transcript's life and potentially mutually exclusive.

Keywords:

RNA secondary structures, RNA structure prediction, co-transcriptional folding, RNA folding pathways, prediction algorithms

1 Introduction

RNA sequences, or transcripts, are the primary products of all DNA genomes. They and their functional products define the active state of a cell in a living organism. Knowing them and understanding the mechanisms regulating their expression is thus key to understanding cellular life and how it regulates (or mis-regulates) itself.

Introducing the notion of alternative RNA structure expression

The need for taking co-transcriptional RNA structure formation into account Almost all existing methods for RNA secondary structure predictions ignore co-transcriptional structure formation and the influence it may have on the predicted structures.

2 Computational strategies for investigating co-transcriptional RNA structure formation *in vivo*

In the following, the term "RNA secondary structure" will refer to a single set of mutually compatible RNA structure features (*e.g.* helices) that can be made or expressed by the underlying RNA sequence *at the same time.* During co-transcriptional folding, a single RNA thus undergoes a time-ordered series of different RNA secondary structures and a transcript with a known riboswitch thus has at least two functional RNA secondary structures. The term "final RNA secondary structure" will refer to the RNA secondary structure at the end of a single co-transcriptional folding pathway (note that there may be several potential pathways for any given transcript). The term "helix" will be used to denote an RNA structure feature corresponding to a contiguous stretch of base-paired nucleotides, *i.e.* without internal loops and bulges.

2.1 Goal 1: Predicting conserved transient and final RNA secondary structure features of co-transcriptional folding pathways

Computational methods for predicting entire co-transcriptional folding pathways have started to emerge since the mid-1980s. These methods take a single sequence as input and predict the folding kinetics of the emerging sequence as function of simulated time (RNA folding pathway prediction methods). Their raw output typically consists of a list of structural configurations encountered during the simulated folding pathway. These methods typically work in a non-deterministic way [1, 2, 3, 4, 5, 6, 7, 8], although exceptions exist [9], see table 1. Different simulations for the same input sequence thus typically result in different predicted folding pathways.

Folding pathway prediction methods model structural changes during the kinetic folding pathway typically on the level of entire helices rather than individual base-pairs. On extending the RNA sequence, corresponding randomised structural changes are proposed (*i.e.* new helices are created or existing ones destroyed). These changes are accepted with a probability which corresponds to the theoretical rate of the corresponding . As the errors of these folding pathway prediction methods are , stochastic simulation methods tend to have a length limitation of around 200 nt. This constraint precludes the analysis of many naturally occurring transcripts.

A more recent addition from 2008, Kinwalker [9], significantly extends the length limitation from 200 nt to around 1000 nt by employing a deterministic strategy to model distinct secondary structure configurations of the co-transcriptional folding pathway. For this, it combines the repeated execution of a deterministic free energy minimisation (MFE) algorithm with heuristic considerations that judge the kinetic feasibility and speed of potential structural transitions between two deterministically calculated structural configuration. For a given input sequence, Kinwalker returns as output exactly one predicted folding trajectory and the corresponding series of encountered RNA secondary structures, whereas simulation-based methods typically yield different folding pathways for multiple executions of the program for the same input sequence, see table 2.

Conceptual limitations of folding pathway prediction programs:

- the length of input sequence limited to 200 nt or 1000 nt, in case of Kinwalker

- the transcription speed is assumed to be constant

- any *cis*-interaction partners (proteins, ligands, RNAs) as well as their potential effects on co-transcriptional folding are ignored.

Program	algorithm	pseudo	max seq.
		knots	length (nt)
Kinwalker	deterministic (MFE structures)	no	1000
Kinéfold	stochastic simulation (helices)	yes	200
RNAKinetics	stochastic simulation (helices)	no	200

Table 1: **Key features of the three investigated folding pathway prediction methods.** For all three methods, a transcription rate r can be specified by the user for each individual input sequence. We use a simulation time of $t = 2 \cdot L/r$ equal to twice the transcription time for an input sequence of length L to allow the program to converge. To determine the number of simulated trajectories for each input sequence of length L, we use a quadratic function of L as recommended by the authors of Kinéfold and RNAKinetics.

Program	raw output
Kinwalker	list of struct. configurations over simulated time
[9]	(one trajectory per input sequence)
Kinéfold	list of struct. configurations over simulated time
[5, 6, 7]	(multiple trajectories per input sequence)
RNAKinetics	aggregated data across all simulated trajectories
[1, 2, 3]	(for each helix, probabilities over simulated time points)

Table 2: **Type of predictions generated as raw output by the three folding pathway prediction methods.**

We recently presented the first systematic performance benchmark for folding pathway prediction methods for a non-redundant dataset of 32 sequences from six functional RNA families and showed that homologous RNA sequences not only fold into similar RNA structures, but also use similar co-transcriptional folding pathways [10]. For this, we investigated and compared the performance of three folding pathway prediction methods, RNAKinetics [1, 2, 3], Kinwalker [9] and Kinéfold [5, 6, 7]. These methods are freely available and representative of the different underlying algorithms being employed in the field, see table 1 and table 2 for an overview of their key features.

The six functional RNA families which constitute our test set were selected to comprise known final RNA secondary structure as well as known transient structural features, as we were keen to explore how well these could be predicted by the different folding pathway prediction methods. The six functional RNA families are: (1) Levivirus maturation gene (Levivirus), (2) bacterial ribonuclease P type A (RNase P Type A), (3) Hepatitis delta virus ribozyme (HDV ribozyme), (4) Bacterial signal recognition particle 4.5S RNA (SRP 4.5S RNA), (5) Tryptophan operon leader (Trp operon) and (6) S-adenosylmethionine riboswitch (SAM riboswitch). We first compiled high-quality multiple-sequence alignment for each of the six families and then extracted a total of 32 non-redundant sequences that constitute of test set, see table 3 for an overview of different quality measures.

Step 1:

In order to predict evolutionarily conserved transient and final RNA secondary structure features encountered during co-transcriptional folding pathways, we first need to establish a high-quality alignment of multiple homologous RNA sequences. This alignment will later be used to map and compare the RNA structure features predicted for individual sequences. From this multiple-sequence alignment, we first extract representative sequences to be used for individual analysis with one of the folding pathway prediction methods.

General recommendations for high-quality alignments to be used for subsequent analysis with folding pathway prediction methods are:

- The primary sequences in the alignment have to extend up to the transcription start site on the 5' end. This is key for simulating co-transcriptional folding pathways.
- The sequences in the alignment should have an average pairwise primary sequence identity that is neither too low nor too high (ideally, between 50% and 85%) to best prepare for comparative analysis.
- The selected representative sequences have to have a good fit to the known RNA structure features and the alignment quality within known structured regions ought to be high. Manual adjustments of the multiple-sequence alignment can, for example, be made with the help of 4SALE [11].

- The quality of any sub-alignment between known structured regions has to be sufficiently high, for example using MUSCLE [12] which is guided by primary sequence conservation only.

Step 2:

Once a high-quality alignment has been established, several so-called representative sequences are extracted from it. The goal is to identify a sub-set of sequences that are all (1) good representatives of the reference RNA secondary structure encoded in the multiple-sequence alignment and (2) non-redundant in terms of pairwise primary sequence identity. This can be achieved as follows:

- First, order all sequences in the alignment based on the quality of the structural fit to the known reference structure, *e.g.* the number of consensus base-pairs fitting the base-pairs of the known RNA secondary structure and the number and type of gaps within structured regions (in particular, one-sided versus two-sided gaps). From the sub-set of resulting sequences with a good structural fit, select the top-scoring sequence as first representative sequence.

- Continue selecting representative sequences from the structure-fit-ranked list of sequences as long as the pairwise primary sequence identity with respect to every already selected representative sequences is sufficiently low. A decent lower threshold value for the maximum pairwise sequence identity (max. PSI) between any two representative sequences from the same multiple-sequence alignment is 50%. Values used in our study [10] range from 55% to 85% depending on the overall level of primary sequence conservation in each alignment and the desired target number of representative sequences, see table 3 for details.

The above procedure yields a non-redundant set of representative sequences for any given alignment. These representative sequences are used as input sequences for the subsequent analysis with folding pathway prediction methods to detect evolutionarily conserved transient and final RNA secondary structure features.

Step 3:

Once the set of representative sequences has been determined for a given alignment of homologous transcripts, each *individual* representative sequence is used as input to one of the three folding pathway prediction methods. As table 1 and table 2 explain, the three methods Kinwalker, Kinéfold and RNAKinetics differ considerably in their underlying algorithms and type of raw output predicted.

All three folding pathway prediction methods allow the user to specify a (constant) transcription speed for each individual input sequence. This parameter is key for influencing the corresponding simulated co-transcriptional folding

alignment	N	R	L	cons.	max. PSI	gaps	TTL	BPs	canon. BPs	cov.
Levivirus	7	4	158	0.645	0.80	0.154	1.181	54	0.939	0.302
RNase P Type A	24	7	391	0.698	0.70	0.084	4.279	122	0.966	0.441
HDV ribozyme	10	3	152	0.819	0.85	0.070	0.499	61	0.941	0.007
SRP 4.5S RNA	10	5	141	0.726	0.80	0.024	1.240	38	0.976	0.340
trp operon	10	5	107	0.694	0.80	0.068	1.253	29	0.969	0.195
SAM riboswitch	15	7	215	0.532	0.55	0.227	4.310	66	0.895	0.276

Table 3: **Overview of different features and quality measures for the six alignments [10].** N refers to the number of sequences in the alignment, R to the respective number of representative sequences, L to the alignment length (nt). The conservation (cons.) indicates the average pairwise primary sequence conservation, max. PSI the maximum pairwise sequence identity (used for selecting representative sequences from the respective multiple-sequence alignment), gaps the fraction of gaps and TTL the total tree length. BPs specifies the number of base-pairs and canon. BPs the fraction of canonical base-pairs within all base-paired alignment columns. The covariation (cov.) ranges from -2 to +2 and measures the relative frequency of compensatory mutations that retain the base-pairing ability within pairs of base-paired alignment columns. Its value is 0 when there is no variation in paired alignment columns, positive when they comprise compensatory mutations that retain the base-pairing ability and negative when they contain invalid base-pairs.

pathways as a change of transcription speed can have a major impact on co-transcriptional RNA secondary structure formation. This is because RNA structure formation can happen on the same time scale as transcription [13]. There is ample experimental support that a change of transcription speed can yield different RNA structure outcomes [14, 15, 16, 17, 18] and that it need not be constant along the entire transcript (transcriptional pausing sites) [19, 20, 21]. The latter is an effect that none of the three folding pathway prediction methods can currently model.

For our data sets, we choose different values for the transcription speed depending on the evolutionary domain of each representative sequence [22], see table 4.

For Kinéfold and RNAKinetics, we use a simulation time of $t = 2 \cdot L/r$ equal to twice the transcription time for an input sequence of length L to allow the program to converge. To determine the number of simulated trajectories for each input sequence of length L, we use a quadratic function of L as recommended by the authors of Kinéfold and RNAKinetics.

Step 4:

The previous step generates a range of raw output data for all individual representative sequences. For each alignment, the predictions of each method for

alignment	r [nt/s]
Levivirus	30.0
RNase P Type A	22.5
HDV ribozyme	20.0
SRP 4.5S RNA	22.5
trp operon	22.5
SAM riboswitch	75.0

Table 4: **Overview of different values of transcription speed r for each RNA family.** These values are used as parameters for the three folding pathway prediction methods. Note that for Levivirus, the speed corresponds to the speed of replication of the positive RNA. Depending on the polymerase, the transcription speed can range from 200 nt/s for phages, 20–80 nt/s for bacteria to 10–20 nt/s for human pol II [22].

all representative sequences are then aggregated and converted into scores for individual, predicted base-pairs. This is done as follows:

- Kinwalker As this method works in a deterministic manner and predicts a single co-transcriptional folding pathway for each representative sequence, each predicted base-pair is assigned a score which is equal to the fraction of representative sequences for which this base-pair was predicted as part of the structural features encountered during the predicted co-transcriptional folding pathway. Based on our MCC-optimised procedure for known structural features, see figure 2 in [10], we recommend a cut-off value of 0.43 for these base-pair specific scores, *i.e.* any predicted base-pair with a score below 0.43 is discarded.

- Kinéfold This method employs stochastic simulations to predict individual co-transcriptional folding pathways. For each representative sequence, we sample N folding pathways which depends quadratically on the length L of the input sequence. For any predicted base-pair, we first calculate the fraction of pathways that feature this base-pair. We then average this fraction across all representative sequences of the same alignment to arrive at the final score assigned to the predicted base-pair. The MCC-optimised cut-off value we recommend for base-pairs predicted with Kinéfold is 0.755, see figure 2 in [10] for details.

- RNAKinetics The raw output of this methods already corresponds to a summary of helices predicted during any of the simulated co-transcriptional folding pathways. As for Kinéfold, the number N of simulated pathways depends quadratically on the length L of the input sequence. For each predicted helix, a probability value is specified for points in simulation time. For a predicted base-pair, we first pick the maximum probability over time as probability for that base-pair and then assign the average value of these probabilities for all representative sequences as final score. The

MCC-optimised cut-off value we recommend for base-pairs predicted with RNAKinetics is 0.0082, see figure 2 in [10].

Prediction accuracy for known transient and final RNA secondary structures

As table 5 shows, Kinwalker and Kinéfold have an equally high Matthews' correlation coefficient (MCC) of 0.676 and 0.656, respectively, for all known transient and final structural features. The MCC is a combined measure of the sensitivity and the specificity of the prediction, see the caption of table 5 for its definition. For Kinwalker, the high MCC value is due to equally high values for the true positive rate (TPR) and the positive predictive value (PPV), whereas these values are more unbalanced for Kinéfold with a significantly higher PPV (0.885) than TPR (0.501). RNAKinetics has a markedly lower overall MCC value of 0.263 which can be primarily attributed to significantly lower values for the PPV, both for known transient and known final RNA structure features. These low values cannot be rescued by the comparatively high values for the TPR. Typically, the TPR for known transient RNA structure features is significantly lower than for known features of the final RNA secondary structure, RNAKinetics being the exception with a high TPR of 0.722 for known transient features which is even high than the TPR of 0.652 for known final structural features. As we showed in the original paper, see figure 2 in [10], the performance values shown in table 5 are not sensitive to the precise choice of cutoff values.

Our MCC-optimised cutoff values, see table table 5, can thus be viewed as robust, general recommendation unless a dedicated data set of known structures is available for training.

Summary

Using the computational pipeline described above, individual folding pathway prediction methods can thus be used in combination with a comparative analysis strategy to reliably predict evolutionarily conserved RNA secondary structure features of co-transcriptional folding pathways. Due to conceptual constraints of the folding pathway prediction methods, this strategy is currently limited to transcripts of 200 nt length (1000 nt in case of Kinwalker).

Known transient structure features can be predicted with roughly the same accuracy as structural features of the final RNA secondary structure. Furthermore, the prediction accuracy of the computational analysis pipeline is robust with respect to the recommended, MCC-optimised cutoff values. When dedicated data sets with known RNA secondary structures are available, there is potential to further improve the prediction accuracy by deriving dedicated cutoff values. If desired, the specificity of the analysis can be further increased by combining the predictions of more than a single folding pathway prediction method to significantly increase the overall PPV while only slightly lowering the TPR, see table 3 in [10] for details.

Program		Known transient		Known final		All known		
	Cutoff	TPR	PPV	TPR	PPV	TPR	PPV	MCC
Kinwalker	0.430	0.428	0.318	0.762	0.648	0.693	0.667	0.676
Kinéfold	0.755	0.183	0.378	0.586	0.874	0.501	0.885	0.656
RNAKinetics	0.0082	0.722	0.191	0.652	0.210	0.678	0.231	0.263

Table 5: **Performance figures for known RNA structure features for the three folding pathway predictionmethods [10].** Average true positive rate (TPR) and positive predictive value (PPV) for known transient and known final RNA secondary structure features using the three folding pathway prediction programs at MCC-optimised cutoff values optimised over known features (Cutoff) for the three folding pathway prediction methods Kinwalker, Kinéfold, and RNAKinetics. Matthews' correlation coefficient (MCC) is a measure of both sensitivity and specificity and is defined as $MCC = (TP \cdot TN - FP \cdot FN)/\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}$. TPR is a measurement of sensitivity on base-pair level, and is defined as $TPR = TP/(TP + FN)$. PPV is a measurement of specificity on base-pair level, and is defined as $PPV = 1 - (FP/(TP + FP))$. See also figure 5 for the ROC curve showing the TPR as function of the FPR for all three folding pathway prediction methods.

2.2 Goal 2: Predicting novel, conserved RNA secondary structure features of co-transcriptional folding pathways

The computional pipeline described above for goal 1 gives a sense of the expected prediction accuracy for known transient and final RNA secondary structure features. One limitation of the above comparative analysis pipeline, however, is that it can only be applied to rather short transcripts. In order to detect RNA structure features that may play important functional roles during any time of the RNA transcripts' life in its cellular environment in transcripts of arbitrary length, we require a technically and conceptually different approach.

We devised a computational prediction method called Transat [23] in order to address this situation. This method works in a comparative way by taking a multiple-sequence alignment and a corresponding evolutionary tree linking the sequences in the multiple-sequence alignment as input. It predicts as output a set of evolutionarily conserved helices with calculated log-likelihood values and corresponding estimated p-values. Transat thus implicitly captures the hypothesis that structural features that have been conserved during certain evolutionary times (as specified by the input tree) are likely to be of functional importance.

In devising Transat, we deliberate avoided conceiving a method for predicting *RNA secondary structures*, *i.e.* sets of helices that could *all* be present *at one point in time*. Instead, Transat only aims to identify *individual*, evolutionarily conserved helices. By deliberately ignoring the grouping of predicted helices into distinct RNA secondary structures and leaving the interpretation of the raw Transat predictions to the user, however, we gain the ability to:

- identify pseudo-knotted configurations of helices that most RNA secondary structure prediction methods cannot identify
- not having to model the details the cellular environment (such as ion concentrations, temperature, potential *trans*-interaction partners, transcription speed and potential transcription pausing sites) and the influence they may have on *in vivo* RNA structure formation. If any RNA structure feature is conserved during evolution, Transat will be able to identify it *irrespective of the cellular circumstances and molecular mechanisms that lead to their conservation.*
- find mutually exclusive helices such as those involved in the two distinct structural configurations of riboswitches
- predict potential transient helices that may not be part of the final RNA secondary structure configuration, but may have distinct functional roles at some time during the RNA transcripts' life in the cell, *e.g.* by functioning as local RNA secondary structure features that regulate alternative splicing via RNA editing [24] or as helices interacting with *trans*-acting molecules during co-transcriptional folding in the cell [25].
- generate predictions without having to assume that any input sequence of the input multiple-sequence alignment folds into a global RNA secondary structure spanning the entire transcript. The latter is usually the implicit assumption of any RNA secondary structure prediction method that is based on the principle of free energy minimisation. These so-called MFE methods cannot be used to identify regions devoid of conserved RNA structure features as they will aim to predict additional base-pairs in order to lower the overall free energy of the predicted structure.

2.3 How Transat works:

Transat is a probabilistic method, both in terms of the underlying algorithms and the predictions being made as output. It takes as input a multiple-sequence alignment of homologous RNA sequences and a corresponding evolutionary tree. This tree relates the sequences in the input alignment quantitatively in terms of topology and branch distances.

In its first step, Transat calculates potential helices for each *individual*, ungapped RNA sequence in the multiple-sequence alignment using a fast dynamic programming procedure that depends quadratically on the sequence length.

In a second step, these sequence-specific helices are mapped onto the input multiple-sequence alignment without altering the multiple-sequence alignment itself. This results in a list of candidate helices *along the alignment* which can each be uniquely specified by a corresponding pair of base-paired outer alignment columns x^1 and y^1 as well as a helix length L. Each candidate helix along the alignment can thus be viewed as a list of 5'-alignment columns, $x = (x^1, x^2, \ldots, x^L)$, that constitute the "left" arm of the helix, and a corresponding list of consecutive 3'-alignment columns, $y = (y^L, \ldots, y^2, y^1)$, that constitute the "right" arm of the helix. A candidate helix of length L thus forms base-pairs between pairs of alignment columns starting with (x^1, y^1) as the outer base-pair up to (x^L, y^L) as the inner-most base-pair.

In the third step, TRANSAT calculates a log-likelihood score $\Lambda(h)$ for each candidate helix h along the alignment. For this, it computes the likelihood $P(x, y|\theta_{\text{paired}})$ that the "left" alignment columns x are indeed base-paired with the corresponding "right" alignment columns y as well as the likelihood $P(x, y|\theta_{\text{unpaired}})$ that they are all unpaired. The first hypothesis is captured by a probabilistic evolutionary model which spells out how pairs of base-paired nucleotides evolve as function of evolutionary time. The second hypothesis corresponds to a different probabilistic evolutionary model which specifies how unpaired nucleotides evolve over time, see [23] for more details. Both likelihood values are calculated using the Felsenstein algorithm [26]. For a given candidate helix h along the alignment, this algorithm takes the nucleotides observed in the actual alignment columns of the multiple-sequence alignment (*i.e.* at time t = now) and evolves them back in time along the input tree using the evolutionary model of the respective hypothesis. The log-likelihood ratio assigned to each candidate helix h along the alignment can then be expressed as

$$\Lambda(h) = \log(P(x, y|\theta_{\text{paired}})/P(x, y|\theta_{\text{unpaired}})) \cdot 1/L$$

One key difference with respect to the usual way that these probabilistic models of evolution are employed in the context of RNA secondary structure prediction is that we interpret *one-sided gaps* within candidate pairs of alignment columns not as missing data, but explicitly penalise them by treating them as non-consensus base-pairs. This correctly captures the observation that helices evolve over time by acquiring or losing *entire base-pairs* and significantly contributes to TRANSAT's ability to correctly distinguish base-paired form unpaired alignment columns.

In the fourth and last step of TRANSAT, each candidate helix h along the alignment is assigned an estimated p-value. This quantifies the statistical significance of the corresponding log-likelihood value $\Lambda(h)$. This step addresses the following issue. An input multiple-sequence alignment consisting primarily of `G` and `C` nucleotides has a generally higher propensity to form spurious helices than a multiple-sequence alignment with a different nucleotide, di-nucleotide and gap composition. In order to be able to distinguish the different propensities of different input multiple-sequence alignment to form spurious helices, we thus have to assign estimated p-values to the calculated log-likelihood values Λ. These p-values then allow the user to rank predicted helices for a given alignment and, more importantly, allow the ranking of helices deriving from different alignments. P-values are estimated by first re-aligning the original multiple-sequence alignment using T-Coffee [27]. The resulting alignment is then carefully randomised by swapping entire alignment column within bins of similar primary sequence conservation and gap composition [28]. Each randomised multiple-sequence alignment is then assumed to no longer contain any real helices. By default, Transat generates 500 randomised versions for a given input multiple-sequence alignment. The log-likelihood values of all candidate helices "detected" in these randomised multiple-sequence alignment are then aggregated to form a null-distribution of log-likelihood scores from which the p-values of the log-likelihood values for the original multiple-sequence alignment are derived, see pages 5 in [23] for more details.

2.4 Key features of Transat and summary:

First, Transat's prediction accuracy is terms of sensitivity, PPV and FPR is almost independent of the length of the input alignment, see figure 5 in [23]. For an individual sequence, the number of potential bi-secondary RNA secondary structures grows exponentially with the sequence length [29]. For any non-comparative structure prediction method, we thus expect a marked decrease in PPV as function of increasing sequence length. As Transat works in a comparative way by taking a multiple-sequence alignment rather than a single RNA sequence as input, it alleviates this problem as there is *a priori* no reason to expect the number of evolutionarily conserved helices to grow quadratically with the sequence length.

That said, the performance of any method used for automatically generating input alignments for Transat may very well (and strongly) depend on the length of the sequences and the alignment. In this case, one also expects a corresponding decrease in the prediction accuracy of Transat as Transat keeps input alignments fixed. Transat has been devised to tolerate some amount of alignment errors, see [23] and step one and two above, but it technically cannot fix any errors in the input alignment.

Second, Transat works best for input alignment that correspond to a total tree length (TTL) of 2 or more, see figure 1. Below a TTL of 1, the primary sequences in the input multiple-sequence alignments tend to be too closely related both in terms primary sequence identity and lack of co-variation within base-paired regions, resulting in a sub-optimal performance of Transat.

In practice, there is no doubt also an upper limit for the TTL beyond which we expect a decrease in predictive accuracy. This can be attributed to either (1) too many RNA structure variations between the sequences to render the comparative approach meaningful or (2) too significant deviations in terms of primary sequence identity to enable the assembly of a trustworthy input alignment for Transat. Both potential complications have to be considered on a case-by-case basis as they strongly depend on the cellular constraints on the primary RNA sequence and on the RNA secondary structure of the specific RNA family being studied. Cases of RNA structure variation in the setting of viral sequences are, for example, shown in [30].

Third, Transat can successfully identify known transient helices and mutually exclusive helices as shown for the *hok* and *trp*-attenuator data sets for which more than a single functional RNA secondary structure is known, see figure 2 and page 13 in [23] for more details. As each predicted helix is assigned an estimated p-value, the user of Transat can easily focus on those with the highest statistical significance.

Fourth, Transat can be employed to also identify regions of the input multiple-sequence alignment that are *devoid of any conserved RNA structure features*, see *e.g.* Rfam alignments RF00018 (which is known to be bound by multiple copies of the CsrA protein that binds single-stranded regions) and RF00023 (which contains an open reading frame (ORF) that has to remain single-stranded), see figure 4. This cannot be readily achieved using MFE-based methods.

Fifth, Transat can be used to identify potential novel, evolutionarily conserved transient features of the co-transcriptional folding pathways of homologous RNAs, see table 6 for a comparison of key alignment features and figure figure 5 for the predictive performance for folding pathway prediction methods Kinéfold, RNAKinetics and Kinwalker for new transient features identified by Transat.

Kinéfold (blue) and RNAKinetics (green) have a similar prediction performance and manage to detect 76.5% (RNAKinetics) and 67% (Kinéfold) of the new transient features with a false positive rate (FPR) of 7% (both). The outlier is Kinwalker (red) which can detect only 28.1% of the new features with a FPR of 4%. The performance plot shows that Kinéfold and RNAKinetics are capable of detecting truly novel transient features identified by Transat provided the default cutoff values (that were determined based on known features, see table 5) are relaxed and higher FPRs tolerated.

Finally, despite predicting only individual, conserved helices, these Transat predictions can often be readily interpreted to suggest a time-ordered potential co-transcriptional folding pathway, see the example of the Cripavirus internal ribosomal entry site in figure 3.

2.5 Goal 3: Detecting final RNA structures formed during co-transcriptional folding

Any of the three folding pathway prediction methods described for goal 1 can be employed in order to predict the final RNA secondary structures that are formed as the result of co-transcriptional folding pathways, see the text, tables and figures of section 2.1 above. All three methods work in a non-comparative way and take a single RNA as input. One major drawback of these methods, however, is that they can only handle rather short input sequences, up to 200 nt in case of RNAKinetics and Kinéfold and up to 1000 nt in case of Kinwalker. This is primarily due to conceptual reasons: As any of these methods simulate/calculate an actual co-transcriptional folding pathway, any errors made in the early stages of the prediction (*i.e.* while the RNA is still relatively short) are magnified as

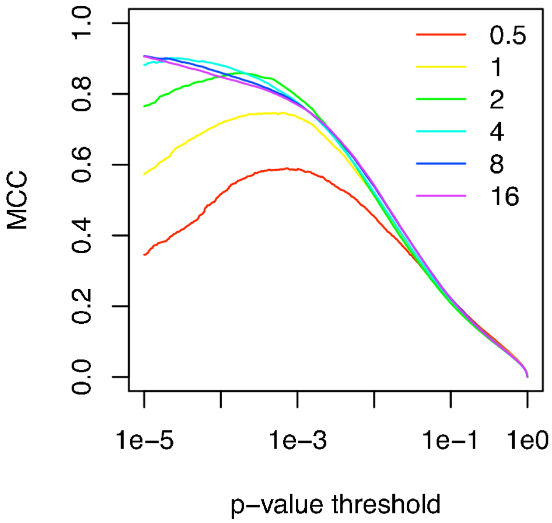

Figure 1: Transat **performance for helices as function of the total tree lengths (TTLs) for a test set of 990 artificially generated, structure encoding multiple-sequence alignments [23].** Different values of the TTL ranging between 0.5 and 16 are indicated by different colours, see the legend. The quality of helices predicted by Transat are measured in terms of MCC as function of the p-value threshold, see caption of table 5 for the MCC definition. The peak MCC-values are highest for TTLs of around 2 or more.

the transcript is elongated and the prediction/simulation progresses. This limitation prevents their use on many naturally occurring RNA transcripts.

We thus set out to devise a new prediction method CoFold [33] with the following goals. The new method should:

- take as input a single RNA,
- predict as output a single RNA secondary structure,
- take some effects of co-transcriptional folding explicitly into account, however, *without* predicting or simulating an actual co-transcriptional folding pathway,
- be also guided to some degree by free energy minimisation and, hopefully,
- outperform the prediction accuracy of existing MFE methods that do not take co-transcriptional effects into account, *especially for longer sequences* (> 200 nt) for which the prediction accuracy of non-comparative MFE methods tends to significantly decrease [34].

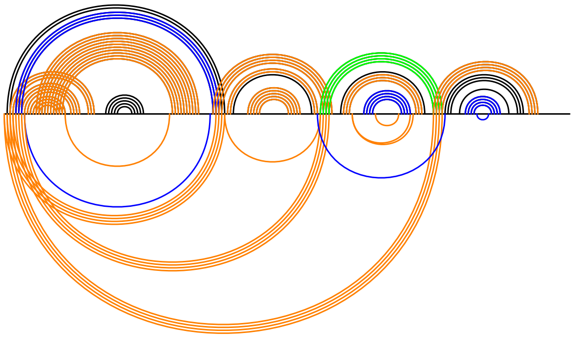

Figure 2: **Conserved helices predicted by Transat for the *hok* data set for a p-value threshold value of right 10^{-3} [23].** The horizontal axis corresponds to the *hok* alignment that was used as input to Transat. Each arc corresponds to a single base-pair linking the corresponding pair of columns in the input alignment. Arcs above the x-axis correspond to known base-pairs, those below to newly predicted base-pairs (false positives). Known base-pairs that are missing from the prediction are shown as black arcs (false negatives). All predicted base-pairs are shown as arcs in a colour that corresponds to their respective, estimated p-values: $< 10^{-5}$ green, $< 10^{-4}$ blue and $< 10^{-3}$ orange. All helices of the known structure are predicted well. In addition, Transat predicts three new helices with statistically significant p-values that may guide the co-transcriptional formation of the final RNA secondary structure. Figure made with R-chie [31, 32].

For this, we used the widely used, non-comparative MFE methods Mfold [35] and RNAfold [36] as a starting point. Both methods take a single RNA sequence as input and predict as output an RNA secondary structure. Both are guided by free energy minimisation alone and do not take any effects of co-transcriptional folding into account.

2.6 How CoFold works:

Transat employs a modified version of the Zuker-Stiegler algorithm [36]. employed by Mfold and RNAfold to calculate the thermodynamically most favourable, pseudo-knot-free RNA secondary structure for any given input RNA. For this, the algorithm (1) decomposes the overall free energy of any possible (pseudo-knot free) RNA secondary structure into a sum of free-energy contributions from various structural lego-like building blocks (such as stacking interactions between pairs of adjacent base pairs, unpaired nucleotides and hairpin loops) and (2) employs a dynamic programming procedure to derive the RNA secondary structure that minimises the overall free energy for a given input RNA sequence. The latter takes $O(L^3)$ time, where L denotes the length of the input sequence. RNA structures predicted by Mfold and RNAfold are

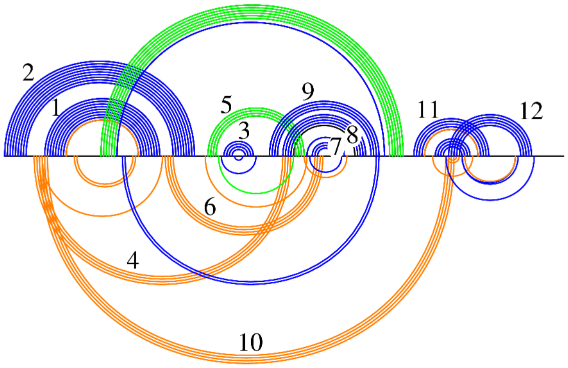

Figure 3: Transat **predictions for the Cripavirus internal ribosomal entry site (IRES),** Rfam **family RF00458, for a p-value threshold of** 10^{-3} **[23].** The helices that render the known structure pseudo-knotted are correctly predicted by Transat. In addition, Transat predicts a few distinct helices that clash with helices of the known, final structure and that may help to guide the overall co-transcriptional structure formation, see the numbers next to the respective helices. The numbering of helices is not part of the Transat predictions, but an interpretation by the user. New helix 4 may yield to final helix 8, new helix 6 to final helix 7 and new helix 10 to final helix 12. These novel helices may thus serve as guiding transient helices during the co-transcriptional formation of the known, final RNA structure. Figure made with R-chie [31, 32].

correspondingly called minimum-free energy (MFE) structures and the methods themselves referred to as MFE methods.

Our modification of the Zuker-Stiegler algorithm consists of

- altering the default free-energy contribution associated with any base-pair by a scaling function $\gamma(d)$ whose value only depends on the distance d of the two base-paired nucleotides along the input sequence.

Technically, this is achieved via a scaling function which is defined as

$$\gamma(d) := \alpha \cdot (e^{-\frac{d}{\tau}} - 1) + 1$$

This is an exponential decay function with two free parameters $\alpha \in]0, 1]$ and $\tau > 0$ [nt^{-1}]. As $\lim_{d\to\infty} \gamma(d) = 1 - \alpha$, the value of α determines the range of values of γ: $\gamma(d) \in]1 - \alpha, 1]$. For example, setting $\alpha = 0.3$ modifies energy values to vary within 70% to 100% of their original values. Parameter τ defines the steepness of the exponential decay as function of the distance d between base-pairing nucleotides. More precisely, τ is the nucleotide distance at which the max-value of $\gamma(d = 0) = 1$ is lowered by $\alpha(1 - 1/e)$. Nucleotides close to each other along the input sequence are thus more likely to base-pair than

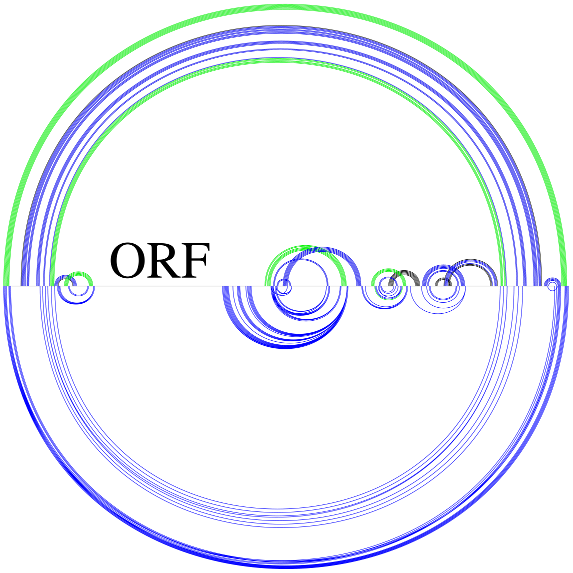

Figure 4: Transat **predictions for a bacterial tmRNA (**Rfam **family RF00023 with alignment length 655 bp) for a p-value threshold value of** 10^{-4} **[23].** Transat captures the helices of the known pseudo-knotted RNA secondary structure well, see arcs above the horizontal line. The known open reading frame (ORF) is predicted to be completely devoid of statistically significant helices. This indicates that this region has to remain single-stranded for the tmRNA to function properly in the cell. The arcs below the horizontal line correspond to new base-pairs predicted by Transat. Arcs with a p-value $< 10^{-5}$ are shown in green green, those with p-values $< 10^{-4}$ in blue. Figure made with R-chie [31, 32].

nucleotides further apart as a higher value of d implies a correspondingly lower scaling factor $\gamma(d)$.

This is what we expect as one overall effect of co-transcriptional folding: , nucleotides close to each other along the sequence have less difficulty "finding each other" than nucleotides further apart. We expect this effect to be more pronounced the faster the transcription speed is. This secondary effect can actually be captured via different values of α and τ, as was shown for a sub-set of viral sequences [33].

2.7 Key features of CoFold:

One goal in devising CoFold was to come up with an MFE method that also performs well for longer sequences, in particular those longer than 200 nt. We thus compiled a corresponding data set of 248 sequences with known functional

	cons.	gaps	canon. BPs	cov.
New transient	0.915	0.009	0.954	0.036
Known transient	0.767	0.023	0.909	0.100
Known final	0.758	0.021	0.963	0.305

Table 6: **Comparison of alignment features for known and novel RNA structure features of co-transcriptional folding pathways of homologous RNAs [10]:** novel transient features ("New transient"), known transient features ("Known transient") and known features of the final RNA secondary structure ("Known final"). As before, see caption of table 3, the conservation (cons.) indicates the average pairwise primary sequence conservation and gaps the fraction of gaps in the alignments. Canon. BPs specifies the fraction of canonical base-pairs within all base-paired alignment columns. Values for the covariation (cov.) range from -2 to +2 and measure the relative frequency of compensatory mutations that retain the base-pairing potential within pairs of base-paired alignment columns. The covariation is 0 when there is no variation in paired alignment columns, positive when they comprise compensatory mutations that retain the base-pairing ability and negative when they contain invalid base-pairs. The features for new transient helices are based on the six statistically most significant helices predicted by Transat for each alignment in [10] that have less than 50% overlap with the known structural features. New transient helices are more highly conserved than known transient and final helices, both in terms of primary sequence conservation (see cons. and gaps) and valid base-pairs (see canon. BPs) which is in line with their lower value of covariation.

RNA secondary structure with a large fraction of long sequences (average sequence length 776 nt, minimum length 110 nt, maximum length 3578 nt), see table 8 for details. Based on our detailed examinations [33], we conclude that:

- Capturing one overall effect of co-transcriptional folding via the scaling function $\gamma(d)$ leads to a significantly improved prediction performance of CoFold w.r.t. the state-of-the-art MFE method RNAfold, especially for sequences longer than 1000 nt, see table 7. Compared to RNAfold, CoFold predicts 7% more base-pairs with 6% higher specificity, corresponding to an 6% increase in MCC. A further increase in PPV, TPR and MCC of 4% can be achieved with CoFold when switching from the default energy parameters by Mathews [37] to those derived by Andronescu [38] via a joint computational tweaking of 363 free energy parameters, compare the respective performance figures of CoFold and CoFold-A in table 7. When zooming into the sub-set of 23S RNAs of the data set (av. length 3069 nt, min. length 2882 nt, max. length 3578 nt), CoFold and CoFold-A increase the MCC of RNAfold on average by 8% and 12%, respectively, which is considerable. Figure 6 shows a comparative illustration of the two RNA secondary structures predicted by RNAfold

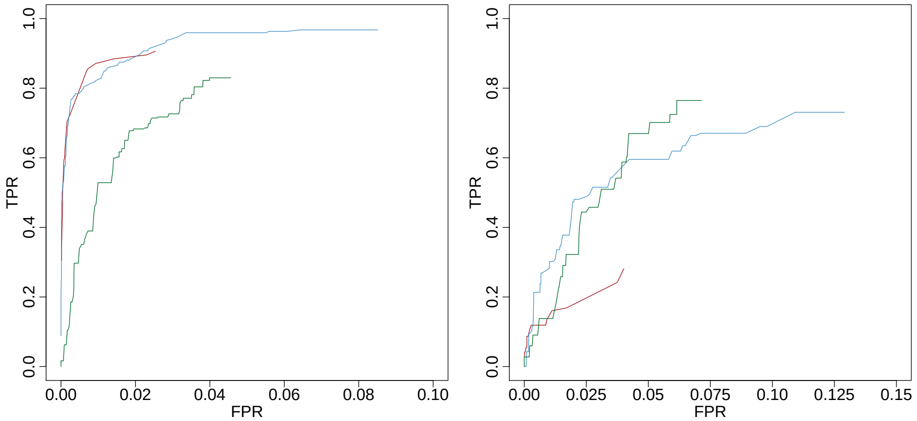

Figure 5: **ROC curves showing the predictive performance of the three folding pathway prediction methods: Kinéfold (blue line), RNAKinetics (green line) and Kinwalker (red line), once for the known transient and final RNA structure features (left figure) and once for the new potential, transient helices identified by Transat (right figure) [10].** The true positive rate (TPR) is shown on the vertical axis and the false positive rate (FPR) on the horizontal axis. Note that the axes use different scales. The new transient features in the right figure correspond to the six statistically most significant helices predicted by Transat for each of the six multiple-sequence alignments. Each candidate helix has to have less than 50 % overlap with any of the known structural features.

and CoFold-A for the 23S RNA of gamma-proteobacteria *Pseudomonas aeruginosa* of length 2893 nt.

- The free energy differences between the energies of the RNA secondary structures predicted by CoFold and CoFold-A and those predicted by RNAfold do *not* correlate with an improved prediction accuracy. Moreover, CoFold improves the prediction accuracy without significantly modifying the corresponding free energy of the respective RNA secondary structure predicted by RNAfold for the same RNA. The improvements in predictive power can thus be attributed to capturing effects of co-transcriptional folding. This is a conceptually important insight as it confirms the earlier hypothesis by Morgan and Higgs from 1996. They conjectured that the observed discrepancies between the free energies of evolutionarily conserved RNA secondary structure and those predicted by MFE methods for longer sequences "cannot simply be put down to errors in the free energy parameters used in the model", but are likely due to effects of kinetic structure formation [34].

	TPR	FPR	PPV	MCC
RNAfold	0.4630	0.000176	0.3974	0.4281
RNAfold-A	0.5202	0.000160	0.4476	0.4817
CoFold	0.5283	0.000159	0.4579	0.4910
CoFold-A	0.5780	0.000145	0.5006	0.5370

Table 7: **Prediction accuracy of** CoFold**,** CoFold-A**,** RNAfold **and** RNAfold-A **for base pairs in the sub-set of long sequences (length $>$ 1000 nt) [33].** Performance specified in terms of true positive rate (TPR), false positive rate (FPR), positive predictive value (PPV) and Matthews correlation coefficient (MCC), see the caption of table 5 for definitions.

- The new algorithm underlying CoFold effectively only depends on *one* free parameter as the two free parameters α and τ of the scaling function γ turn out to be strongly correlated. This is evident when investigating the optimal prediction accuracy in terms of average MCC for different combinations of α and τ values, see figure 1 in [33]. The observed correlation between α and τ can be well described by a linear function $\alpha = a{\cdot}\tau{+}b$ with slope $a = 6.1{\cdot}10^{-4}{\pm}2{\cdot}10^{-5}$ and intercept $b = 0.105{\pm}0.016$ ($R^2 = 98.4\%$). Cross-validation experiments validate the robustness of parameter training and the overall approach. For optimal average MCC values, we determine $\alpha = 0.50$ and $\tau = 640$. These are the parameters we recommend for general use with CoFold and CoFold-A.

3 Availability

Transat, CoFold, CoFold-A and the RNA structure visualisation program R-chie are freely available for use via our web-server at `www.e-rna.org`. This is also where you can download the respective software for local use.

4 Summary and outlook

In summary, it quite remarkable that it is possible to detect conserved RNA secondary structure features of co-transcriptional folding pathways based on “naked” RNA sequences alone. Much of it is due to the power of the comparative approach

As we showed for goal 1, computational methods for predicting actual co-transcriptional folding pathways can be combined with a comparative analysis to successfully identify transient and final RNA secondary structure features that have been conserved during evolution. Our main result, namely that homologous RNA transcripts not only assume similar final RNA secondary structures, but also reach their respective target structure via similar co-transcriptional folding pathways, opens the possibility to study these conserved transient structure

clade	short seqs. (≤ 1000 nt)	long seqs. (> 1000 nt)	combined data set
Bacteria	54	15	69
Eukaryotes	97	15	112
Virus	20	0	20
Archea	16	17	33
Chloroplast	0	14	14
sum	187	61	248
sequence length (nt)			
average	247	2397	776
minimum	110	1245	110
maximum	628	3578	3578

Table 8: **Key features of the data set used for evaluating** CoFold **and** CoFold-A **[33].** The set consists of 248 sequences with known RNA secondary structures. One special focus is the sub-set of 61 long sequences (length > 1000 nt; average 2397 nt, min length 1245 nt, max length 3578 nt) which consists of 16S and 23S rRNAs only (27 sequences of type 16S RNAs and 34 of type 23S RNAs). These sequences were extracted from multiple-sequence alignments from the Comparative RNA Web site (CRW) [39]. The 187 sequences corresponding to the sub-set of comparatively short sequences (length < 1000 nt) were derived from 21 biologically diverse Rfam families [40]. When compiling both sub-sets, great care was taken to ensure the resulting sequences are nonredundant (maximum pairwise sequence identity of 85%), have an RNA secondary structure that is well supported in terms of evolutionary evidence and are diverse in terms of represented clades.

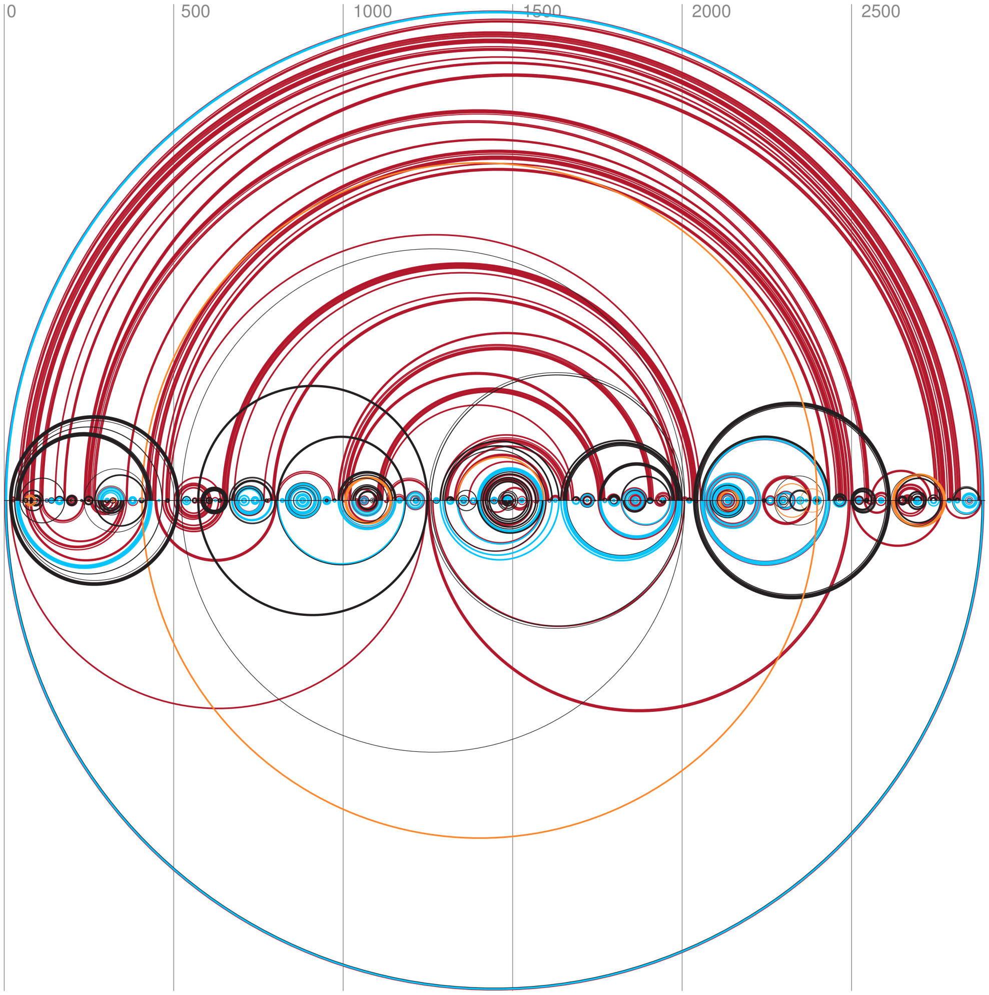

Figure 6: **Comparison of the two RNA secondary structure predicted by RNAFOLD (top) and COFOLD-A (bottom) for the 23S RNA of the gamma-proteobacteria *Pseudomonas aeruginosa* of 2893 nt length [33].** Black arcs indicate base-pairs that are part of the known reference structure, but missing in the prediction (false negatives), red arcs correspond to incorrectly predicted base-pairs (false positives) and blue arcs to correctly predicted base-pairs (true positives). Neither RNAFOLD nor COFOLD-A can handle pseudo-knotted RNA secondary structures. Base-pairs that render the RNA secondary structure pseudo-knotted are indicated by orange arcs. The multitude of red arcs on top of the horizontal line imply that RNAFOLD predicts many incorrect base-pairs spanning 100 nt or more. These mostly disappear with COFOLD-A, see arcs shown below the horizontal line. Using COFOLD-A rather than RNAFOLD, the MCC rises significantly from 43% to 58% (+15%) which can be attributed to a simultaneous increase of the true positive rate (45% to 61%) and the positive predictive value (41% to 56%). The false positive rate of 0.01% is equally low for both prediction programs. Arc-plot made with R-CHIE [31, 32].

features in more detail. This could, for example, by achieved in more fine-grained computational studies that examine the time-wise ordering of transient structure features to investigate:

- **(1)** how conserved transient structure features guide the co-transcriptional formation of the final RNA secondary structure, in particular
    - **(1a)** if and how they prevent mis-folded structure intermediates and/or
    - **(1b)** whether or not they serve as anchors for potential *trans*-acting interaction partners. The latter project will require a decent number of experimentally confirmed test cases of known *trans*-interactions, *e.g.* with proteins or other RNA transcripts, to compile a test set and a potential training set for purely computational analyses, see [41] for an example of how known *trans*-interactions can be captured using folding pathway prediction methods.

As the existing folding pathway prediction methods can only be applied to rather short transcripts ($<$ 200 nt length, or $<$ 1000 nt in case of KINWALKER), there is also ample scope to improve the existing folding pathway prediction methods or, alternatively, to invent a conceptually new:

- **(2)** The primary predictive power of the computational pipeline presented above derives from employing a comparative approach. This power could, for example, be better harnessed by devising the first *comparative* method for folding pathway prediction. Rather than taking a single RNA as input, it would take an (ideally, un-aligned) set of homologous RNAs as input and predict/simulate the corresponding co-transcriptional folding pathways in a comparative way. This approach is not only likely to result in a superior prediction accuracy, but could also remove the conceptual limitation of handling only rather short transcripts.

  There already exists a Markov-chain Monte-Carlo-based method SIMULFOLD [42] for co-estimating homologous RNA secondary structures, multiple-sequence alignments and evolutionary trees. This method works without considering co-transcriptional structure formation. This alignment-free method could inspire a similar, comparative method that could also capture some aspects of co-transcriptional RNA structure formation.

We described how TRANSAT can be used to successfully identify conserved helices in given multiple-sequence alignments of homologous transcripts that most methods for RNA secondary structure prediction miss due to technical constraints. Examples include mutually exclusive helices such as those in riboswitches, transient helices that are not part of the final reference structure or helices involved in pseudo-knotted configurations. Moreover, TRANSAT does not force helices into a single, global RNA secondary structure which almost all methods for RNA secondary structure implicitly do. One conceptual bottle-neck of TRANSAT (as well as most comparative RNA secondary structure prediction methods) is that it requires a high-quality input alignment in order to perform well.

- **(3)** One idea to further improve Transat is thus to convert it into an alignment-free method while keeping (A) the probabilistic nature of the underlying algorithm and its predictions and (B) the time-and-memory efficiency of the algorithm. As Transat internally operates on helices (rather than individual nucleotides or base-pairs) once the candidate helices for individual sequences have been predicted, see step 1 of its description in 2.3, this can be viewed as a realistic goal. CARNAC [43] was the first to apply the alignment-free idea in the context of RNA secondary structure prediction, albeit outside a probabilistic framework and without knowing the theoretical time-and-memory requirements of the underlying, heuristic algorithm.

CoFold was started as a conceptual idea to see if the state-of-the-art method RNAfold could be improved by capturing a single, basic overall effect that co-transcriptional folding has on structure formation, namely the reachability of potential pairing partners along the RNA sequence. This works surprisingly well. There is thus ample scope for further improvements:

- **(4)** One obvious effect of co-transcriptional folding that is not yet captured is the 5'-to-3' directionality. This is one of the key features of co-transcriptional folding. Many state-of-the-art methods for RNA secondary structure use a dynamic programming algorithm to generate their predictions and CoFold is no exception. These dynamic programming algorithms, however, have inherent no notion of left and right, 5' and 3', but are symmetric w.r.t. both sequence ends. This makes it conceptually challenging to incorporate 5'-to-3' biases into them.

Recent technological advances allow us to investigate RNA structures *in vivo* with unprecedented detail and on a transcriptome-wide scale [44, 45, 46, 47].

Acknowledgements

I thank past and present group members and collaborators for the privilege of working with them.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] A. Mironov, L. Dyakonova, A. Kister, A kinetic approach to the prediction of RNA secondary structures, Journal of Biomolecular Structure & Dynamics 2 (5) (1985) 953–962.

[2] A. Mironov, V. Lebedev, A kinetic model of RNA folding, Biosystems 30 (1-3) (1993) 49–56.

[3] L. Danilova, D. Pervouchine, A. Favorov, A. Mironov, RNAkinetics: a web server that models secondary structure kinetics of an elongating RNA, Journal of Bioinformatics and Computational Biology 4 (2) (2006) 589–596.

[4] C. Flamm, W. Fontana, I. L. Hofacker, P. Schuster, RNA folding at elementary step resolution, RNA 6 (3) (2000) 325–38.

[5] H. Isambert, E. D. Siggia, Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme, Proceedings of the National Academy of Science of the USA 97 (12) (2000) 6515–20.

[6] A. Xayaphoummine, T. Bucher, F. Thalmann, H. Isambert, Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations, Proceedings of the National Academy of Science of the USA 100 (26) (2003) 15310–5.

[7] A. Xayaphoummine, T. Bucher, H. Isambert, Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots., Nucleic Acids Res 33 (Web Server issue) (2005) W605–10.

[8] A. Gultyaev, F. von Batenburg, C. Pleij, The computer-simulation of RNA folding pathways using a genetic algorithm, Journal of Molecular Biology 250 (1) (1995) 37–51.

[9] M. Geis, C. Flamm, M. T. Wolfinger, A. Tanzer, I. L. Hofacker, M. Middendorf, C. Mandl, P. F. Stadler, C. Thurner, Folding kinetics of large RNAs., Journal of Molecular Biology 379 (1) (2008) 160–173.

[10] J. Y. A. Zhu, A. Steif, J. R. Proctor, I. M. Meyer, Transient RNA structure features are evolutionarily conserved and can be computationally predicted, Nucleic Acids Research 41 (12) (2013) 6273–6285.

[11] P. Seibel, T. Müller, T. Dandekar, J. Schultz, M. Wolf, 4SALE: A tool for synchronous RNA sequence and secondary structure alignment and editing, BMC Bioinformatics 7 (2006) 498.

[12] R. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Research 32 (5) (2004) 1792–1797.

[13] H. M. Al-Hashimi, N. G. Walter, RNA dynamics: it is about time, Current Opinion in Structural Biology 18 (2008) 321–329.

[14] B. Lewicki, T. Margus, J. Remme, K. Nierhaus, Coupling of rRNA transcription and ribosomal assembly in vivo – formation of active ribosomal-subunits in Escherichia coli requires transcription of RNA genes by host RNA polymerase which cannot be replaced by T7 RNA polymerase, Journal of Molecular Biology 231 (1993) 581–593.

[15] M. Y. Chao, M. Kan, S. Lin-Chao, RNAII transcribed by IPTG-induced T7 RNA polymerase is non-functional as a replication primer for ColE1-type plasmids in Escherichia coli, Nucleic Acids Research 23 (1995) 1691–1695.

[16] T. Pan, X. Fang, T. Sosnick, Pathway modulation, circular permutation and rapid RNA folding under kinetic control, Journal of Molecular Biology 286 (13) (1999) 721–731.

[17] S. Heilman-Miller, S. Woodson, Effect of transcription on folding of the Tetrahymena ribozyme, RNA 9 (6) (2003) 722–733.

[18] S. Heilman-Miller, S. Woodson, Perturbed folding kinetics of circularly permuted RNAs with altered topology, Journal of Molecular Biology 328 (2) (2003) 385–394.

[19] F. Toulme, C. Mosrin-Huaman, I. Artsimovitch, A. Rahmouni, Transcriptional pausing *in vivo*: A nascent RNA hairpin restricts lateral movements of RNA polymerase in both forward and reverse directions, Journal of Molecular Biology 351 (1) (2005) 39–51.

[20] J. Wickiser, W. Winkler, R. Breaker, D. Crothers, The speed of RNA transcription and metabolite binding kinetics operate an FMN riboswitch, Molecular Cell 18 (1) (2005) 49–60.

[21] T. Wong, T. Sosnick, T. Pan, Folding of noncoding RNAs during transcription facilitated by pausing-induced nonnative structures, Proceedings of the National Academy of Science of the USA 104 (46) (2007) 17995–18000.

[22] T. Pan, T. Sosnick, RNA folding during transcription, Annual Review of Biophysics and Biomolecular Structure 35 (2006) 161–175.

[23] N. J. P. Wiebe, I. M. Meyer, Transat — method for detecting the conserved helices of functional RNA structures, including transient, pseudoknotted and alternative structures, PLoS Computational Biology 6 (6) (2010) e1000823.

[24] A. Mazloomian, I. M. Meyer, Genome-wide identification and characterization of tissue-specific RNA editing events in D. melanogaster and their potential role in regulating alternative splicing, RNA Biology 12 (12) (2015) 1391–1401.

[25] D. Lai, J. R. Proctor, I. M. Meyer, On the importance of co-transcriptional RNA structure formation, RNA 19 (2013) 1461–1473.

[26] J. Felsenstein, Evolutionary trees from DNA sequences: a maximum likelihood approach, Journal of Molecular Evolution 17 (6) (1981) 368–376.

[27] C. Notredame, D. Higgins, J. Heringa, T-Coffee: A novel method for fast and accurate multiple sequence alignment, Journal of Molecular Biology 302 (1) (2000) 205–217.

[28] S. Washietl, I. Hofacker, Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics, Journal of Molecular Biology 342 (1) (2004) 19–30.

[29] C. Haslinger, P. F. S. PF, RNA structures with pseudo-knots: graph-theoretical, combinatorial, and statistical properties, Bulletin of Mathematical Biology 61 (3) (1999) 437–467.

[30] J. Pedersen, I. Meyer, R. Forsberg, P. Simmonds, J. Hein, A comparative method for finding and folding RNA secondary structures within protein-coding regions, Nucleic Acids Res. 32 (16) (2004) 4925 – 4936.

[31] D. Lai, J. R. Proctor, J. Y. Zhu, I. M. Meyer, R-CHIE: a web server and R package for visualizing RNA secondary structures, Nucleic Acids Research 40 (12) (2012) e95.

[32] D. Lai, I. M. Meyer, e-RNA: a collection of web servers for comparative RNA structure prediction and visualisation, Nucleic Acids Research 42 (Web Server Issue) (2014) W373–W376.

[33] J. R. Proctor, I. M. Meyer, CoFold: an RNA secondary structure prediction method that takes co-transcriptional folding into account, Nucleic Acids Research 41 (9) (2013) e102.

[34] S. Morgan, P. Higgs, Evidence for kinetic effects in the folding of large RNA molecules, Journal of Chemical Physics 105 (16) (1996) 7152–7157.

[35] M. Zuker, Mfold web server for nucleic acid folding and hybridization prediction, Nucleic Acids Research 31 (13) (2003) 3406–3415.

[36] M. Zuker, P. Stiegler, Optimal computer folding of large RNA sequences using thermodynamic and auxiliary information, Nucleic Acids Research 9 (1981) 133–148.

[37] D. H. Mathews, J. Sabina, M. Zuker, D. H. Turner, Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure, J Mol Biol 288 (5) (1999) 911–940.

[38] M. Andronescu, A. Condon, H. H. Hoos, D. H. Mathews, , K. P. Murphy, Efficient parameter estimation for RNA secondary structure prediction, Bioinformatics 23 (13) (2007) I19–I28.

[39] J. Cannone, S. Subramanian, M. Schnare, J. Collett, L. D'Souza, Y. Du, B. Feng, N. Lin, L. Madabusi, K. Muller, N. Pande, Z. Shang, N. Yu, R. Gutell, The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs, BMC Bioinformatics 3 (2002) 2.

[40] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, A. Bateman, Rfam: annotating non-coding RNAs in complete genomes, Nucleic Acids Research 33 (2005) D121–D124.

[41] R. J. W. Schoemaker, A. P. Gultyaev, Computer simulation of chaperone effects of Archael C/D box sRNA binding on rRNA folding, Nucleic Acids Research 34 (7) (2006) 2015–2026.

[42] I. M. Meyer, I. Miklós, Simulfold: Simultaneously Inferring an RNA Structure Including Pseudo-Knots, a Multiple Sequence Alignment and an Evolutionary Tree Using a Bayesian Markov Chain Monte Carlo Framework, PLoS Computational Biology 3 (8) (2007) e149.

[43] H. Touzet, O. Perriquet, CARNAC: folding families of related RNAs., Nucleic Acids Research 32 (2004) W142–145.

[44] M. Zubradt, P. Gupta, S. Persad, A. M. Lambowitz, J. S. Weissman, S. Rouskin, DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo, Nature 14 (1) (2016) 75–82.

[45] J. G. Aw, Y. Shen, A. Wilm, M. Sun, X. N. Lim, K. L. Boon, S. Tapsin, Y. S. Chan, C. P. Tan, A. Y. Sim, T. Zhang, T. T. Susanto, Z. F. Z2, N. Nagaraj, Y. Wan, In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation, Molecular Cell 62 (4) (2016) 603–617.

[46] Z. Lu, Q. C. Zhang, B. Lee, R. A. Flynn, M. A. Smith, J. T. Robinson, C. Davidovich, A. R. Gooding, K. J. Goodrich, J. S. Mattick, J. P. Mesirov, T. R. Cech, H. Y. Chang, RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure, Cell 165 (5) (2016) 1267–1279.

[47] E. Sharma, T. Sterne-Weiler, D. O'Hanlon, B. J. Blencowe, Global Mapping of Human RNA-RNA Interactions, Molecular Cell 62 (4) (2016) 618–626.