LocARNA-P: Accurate Boundary Prediction and Improved Detection of Structured RNAs

Sebastian Will[1,2], Tejal Joshi[3], Ivo L. Hofacker[4], Peter F. Stadler[4,5,6,7,8], Rolf Backofen[1,9,*]

1 Chair for Bioinformatics, Institute of Computer Science, Albert-Ludwigs-Universität, Georges-Koehler-Allee, Geb. 106, D-79110 Freiburg, Germany

2 Computation and Biology Group, CSAIL, MIT, 77 Massachusetts Ave, Cambridge MA 02139, USA

3 Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Building 208, DK-2800 Kongens Lyngby, Denmark

4 Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria

5 Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center of Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany

6 Max-Planck-Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany

7 Fraunhofer Institute for Cell Therapy and Immunology, Perlickstraße 1, D-04103 Leipzig, Germany

8 Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501

9 Center for Biological Signaling Studies (BIOSS), University of Freiburg, Albertstr. 19, 79104 Freiburg, Germany

∗ E-mail: backofen@informatik.uni-freiburg.de

Abstract

Current genomic screens for non-coding RNAs (ncRNAs) predict a large number of genomic regions containing potential structural ncRNAs. The analysis of this data requires highly accurate prediction of ncRNA boundaries and discrimination of promising candidate ncRNAs from weak predictions. Exist-

ing methods struggle with these goals because such comparative analysis is based on multiple sequence alignments of orthologous regions and does not revise these alignments based on sequence and structural similarity. To overcome this limitation, we systematically fulfill both requirements by efficiently computing the reliabilities of sequence-structure alignments. The reliability profiles of alignments provide a versatile tool for the manual and automatic analysis of ncRNAs. In particular, we improve the boundary prediction of the widely used non-coding RNA gene finder `RNAz` by a factor of three from a median deviation of 47 to 13 nucleotides. Post-processing `RNAz` predictions with `LocARNA-P` allows much stronger discrimination between true and false-positive `RNAz` predictions than `RNAz`'s own evaluation. This improved accuracy, in this scenario from an AUC of 0.71 to 0.87, significantly reduces the cost of successive analysis steps. The ready-to-use software tool `LocARNA-P` produces high-quality multiple sequence-structure alignments together with associated reliability information and prediction of accurate boundaries. We provide detailed results, a web server for `LocARNA`/`LocARNA-P`, and the software package with detailed documentation and a pipeline for refining screens for structural ncRNA at `http://www.bioinf.uni-freiburg.de/Supplements/LocARNA-P/`.

1 Background

Starting with the discovery of microRNAs (24, 26, 27) and the advent of genome-wide transcriptomics (44, 6, 2), it has become obvious that RNA's crucial role in living cells extends far beyond being a mere template for protein biosynthesis. Indeed, the majority of transcripts might have primarily regulatory functions (32). Elucidating the functional roles of many newly discovered non-coding RNAs (ncRNAs) has thus become a central research interest in molecular biology.

The function of many ncRNAs is determined by their secondary structure rather than their sequence. Such structural ncRNAs can therefore be detected by their stable and evolutionary conserved secondary structures. Recent advances in computational RNomics originated numerous approaches for this purpose (38, 49, 50, 37, 45, 48, 3, 54, 8). Among these methods, `EvoFold` (37) and `RNAz` (50, 51, 16) are efficient enough to be applied to genome-wide surveys in mammals (37, 51) and other metazoan clades (34, 35).

The fast approaches `EvoFold` and `RNAz`, don't revise the given whole-genome alignment at all. The idea of revising the alignment for RNA prediction, pioneered by `MSARi` (8), is also realized in the EM-based approach `CMfinder`(54), which extends the idea from local sequence motif finders such as MEME to the problem of finding local RNA structure motifs. Due the high computational demands of structurally revising the alignment, `CMfinder` has not been applied to complete eukaryotic genomes, but e.g. to a 1% fraction of the human genome, the ENCODE region, in (47).

Whereas `EvoFold` applies stochastic context-free grammars (SCFGs), an approach pioneered by `qrna` (38), `RNAz` is based on the evaluation of folding thermodynamics and covariance. Both approaches classify input alignments either as unstructured or as possessing a common RNA secondary structure; in the latter case, the methods predict a consensus structure of the aligned sequences.

Mainly motivated by efficiency reasons, these approaches rely on multiple sequence alignments that are constructed without taking structural similarity into account. However, because RNA structure is often more conserved than sequence, sequence similarity can be weak even within well-established RNA families. Thus, many ncRNAs cannot be aligned well by pure sequence-alignment techniques, which fail for structured RNAs at pairwise sequence identities below 60% (12). Various algorithmic approaches have been introduced to determine structural similarities and to derive consensus structure patterns for structural RNAs with low sequence identity (42, 19, 41, 13, 18, 31, 4, 21, 52).

The first practical approaches for multiple structural alignment, such as `RNAforester` (19) and `MARNA` (42), depend on predicted or known secondary structures. In practice, however, these approaches are limited by the low accuracy of non-comparative structure prediction. Sankoff's algorithm (41) provides a general solution to the problem of simultaneously computing an alignment and the common secondary structure of two aligned sequences. In its full form, the problem requires $O(n^6)$ CPU time and $O(n^4)$ memory, where n is the length of given RNA sequences. This complexity is prohibitive for most practical problems. There are two variants of the Sankoff algorithms. Programs such as `FoldAlign` (13, 18), `dynalign` (31), and `Stemloc-AMA` (4) implement an energy model for RNA that is evaluated during the alignment computation. In contrast, `PMcomp` (21) and `LocARNA` (52) use a full-featured energy model in their pre-computation step. Therefore, they determine a matrix of base pair probabilities using McCaskill's algorithm (33) for each input sequence. During the alignment process, base pair probabilities are used to assess the similarity of the secondary structures. This strategy saves time during the alignment and overall. Nevertheless the probabilities guide the simultaneous alignment and folding precisely in accordance with the RNA energy model.

As its potentially most important application, multiple sequence-structure alignment suggests itself for overcoming the principle weakness of de-novo prediction of structural ncRNA that relies on pure sequence alignment. However, there are still two caveats. First, many approaches need to employ a sliding window technique because the boundaries of the ncRNAs are not known in advance. This technique can result in poor structure models due to an inaccurate folding context. Second, these approaches are too demanding on computational resources to be used for complete genomic screens.

To overcome these limitations, we propose a new pipeline for structural ncRNA gene finding that employs fast, sequence alignment-based ncRNA finders like `RNAz` as a first filter. The found hits are then extended by genomic context and further analyzed using a novel multiple sequence-structure alignment approach.

Our new alignment method provides a reliability measure for sequence-structure alignment that can be employed for several important tasks in this pipeline, namely for 1.) detecting clusters of structural ncRNAs predicted as putative ncRNA-containing regions by the ncRNA gene finder; 2.) determining accurate ncRNA boundaries using alignment reliabilities based on sequence and structural similarity, and 3.) improving the predictive power of ncRNA gene finding.

The novel method `LocARNA-P` computes a sequence-structure alignment and associated local (e.g. column-wise) and global reliability measures. `LocARNA-P` employs the highly accurate scoring model of LocARNA. For determining the reliabilities for large sets of long RNAs, it is crucial that LocARNA-P preserves the low time and space complexity of LocARNA for its more involved task.

The low complexity of LocARNA and LocARNA-P results from their use of sparsity at the structure level. LocARNA (52) introduced this use of sparsity to Sankoff-style approaches. The same idea is found in FoldAlignM (46) and was later picked up by RAF (10). The approach of RAF is interesting because it combines sparsity on the structure and sequence level (this combination first seen in Stemloc (4)) with a lightweight scoring scheme that significantly improves its efficiency over other Sankoff-style methods (10). With LocARNA-P, instead of improving efficiency, we introduce the calculation of match probabilities and reliabilities as a new quality to Sankoff-style alignment.

Computing pairwise match probabilities for sequence-structure alignments has been discussed previously by Hofacker and Stadler (20) and Harmanci *et al.* (17). However, LocARNA-P improves these results in several ways.

- First, the time and space complexity of LocARNA-P is significantly lower when compared to previous approaches, making the tool applicable to large-scale analysis.
- Second, the match probabilities improve the accuracy of multiple alignment and provide column reliabilities for multiple alignments. In contrast, prior attempts at sequence-structure match probabilities did not handle alignments with more than two sequences.
- Third, our method has been applied to genomic-scale data and benchmarked, whereas prior work presented only a few examples.

Although our evaluation focuses on *de novo* prediction of structural ncRNA, the same method can improve experimental ncRNA detection by deep sequencing in two respects. First, assembling correct ncRNA transcripts from short sequence reads is generally nontrivial (e.g., see Langenberger *et al.* (25)). Second, ncRNAs are often transcribed as precursors, which are hard to detect using deep sequencing. In both cases, one can use the same pipeline that we propose for refining de-novo ncRNA screens. In this case, however, one analyzes genomic regions covered by short reads, after obtaining orthologous genomic regions in related organisms from a whole genome alignment. Furthermore, since deep-sequencing does not distinguish structural and non-structural ncRNAs and RNA motifs, this analysis will yield qualitatively new information about RNA structure in and surrounding the identified transcripts.

For evaluating our approach, we predicted the gene boundaries on a data set of 287 RNAz hits in fly (39) that coincide with FlyBase structural ncRNA annotations of *Drosophila melanogaster*. In this data set, we improved the boundary prediction of RNAz by a factor of about three, reducing the median deviation between annotated and predicted boundaries from 47 nucleotides to only 13 nucleotides. For the purpose of this paper, we refrained from predicting boundaries for the unannotated loci, which would not strengthen our evaluation. Our boundary predictions reveal additional information about the genomic context of the ncRNAs. For instance, the 3' or 5' flanking regions that are conserved in sequence and structures are detected. Visualization of the reliability profiles that underlie our automatic predictions supports their interpretation by an expert. We study exemplarily and, for tRNAs, systematically, predictions in flanking regions and observe that they yield true signals in the majority of cases. Reliability profiles and boundary prediction produce a powerful measure for discriminating false and true positives in an ncRNA screen. We show that this measure significantly improves specificity and sensitivity over the currently used ncRNA probability estimate of RNAz. Due to the large number of ncRNA candidate predictions from a genomic screen, our approach is highly relevant because it identifies a subset of best structural ncRNA candidates for subsequent expensive experimental analysis.

2 Results

2.1 Reliability plots

When visualized in the form of a reliability plot, column-wise reliabilities provide a very intuitive view of the local reliability of the alignment. For the ease of interpretation, we project the reliability profile to one particular reference sequence of interest. This proved useful in all studied applications, particularly when the annotation is known or is to be generated for the particular sequence.

Figure 1(c) demonstrates how reliability profiles can support the manual curation of ncRNA alignments. We show sequence and structure reliability along an automatic alignment of nine 7SK ncRNAs generated by `LocARNA-P`. The reliability plot is projected to the RNA of *Xenopus laevis* and complemented by a mountain plot of the consensus structure. The consensus structure, which fits the predicted structure reliabilities well, was obtained from a large hand-curated alignment of 7SK ncRNAs. The general shape of the reliability profile is in agreement with the experiences from hand curating the alignment, where the 5' and 3' ends of the sequences align very well and columns between positions 150 and 250 are extremely variable (14, 30).

2.2 Locating structural ncRNAs using Reliability Profiles

Given a reliability profile projected to the sequence of the reference genome, we computationally determine the location of potential ncRNAs. Therefore, we fit a two-step function of some reliability values a and b such that a value of a indicates a predicted ncRNA locus. Extending the idea of least squares fitting, the quality of a fit is the sum of square deviations plus a penalty Δ for each switch between the values a and b. For a given a and b, the optimal fit is calculated by an exact DP-approach. Instead of fitting all profiles with the same a and b values, we determine optimal values of a and b for each reliability profile using gradient descent optimization. This approach is detailed in the Methods section.

Figure 1(a) shows the reliability plot for the microRNA-cluster from position 90800800 to 90801699 of human chromosome 13. In particular, the structural component of the reliability profile (dark blue) correlates well with the annotated microRNAs, as indicated by the red line. Fitting the two-step function results in a good prediction of the microRNA locations (green line). A larger example from 5 sequences with lengths about 5000 is given in Figure 1(b). The figure profiles the gene *gas5*, whose introns contain 10 C/D-box snoRNAs in human (43). Identifying the C/D-box snoRNAs in this large genomic context is challenging due to their weak conservation signal for both sequence and structure. However,

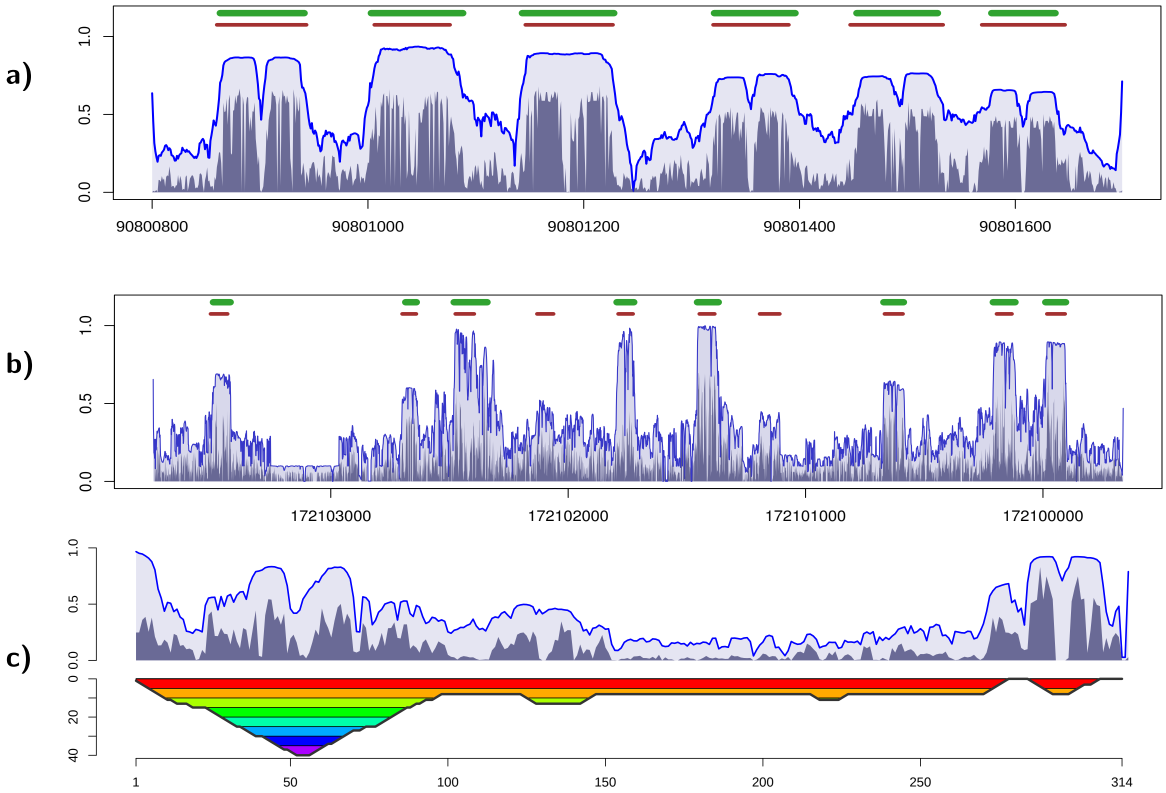

Figure 1: Reliability profile plots with annotations. In each profile plot, the dark blue regions indicate structure reliability, the light-blue regions represent sequence reliability, and the blue line shows the combined column-reliability. In (a) and (b), the known annotation is shown in red. The green lines result from automatic prediction based on the reliability profile. a) plot for the `LocARNA-P`-alignment of the miRNA cluster hg18, chr13, positions 90800800 to 90801699, projected to the human sequence. One can see that the known microRNAs are easily detected using our method b) Reliability plot for the `LocARNA-P`-alignment of gene *gas5*, which hosts ten C/D-box snoRNAs. This is a challenging example because C/D-box snoRNA possess only weak (non-stable) structure and poor sequence conservation. Consequently, CD-box snoRNAs proved to be especially hard to discover using de novo structural ncRNA-predictors like `RNAz`. c) plot of an alignment of 9 ncRNAs from the 7SK ncRNA family projected to the *X. laevis* sequence. The profile is annotated with a mountain plot of the consensus structure. Note how the flanks of the mountain plot and peaks of structure reliability are in good agreement.

we correctly predicted 8 of the 10 snoRNAs. We like to emphasize that `LocARNA-P` supports the computation of these very large instances due to optimally exploiting local folding. For details see the Supplemental Material.

2.3 Accurate Boundaries of structural ncRNA

A common problem in the *de novo* prediction of ncRNA is that only approximate locations of structural RNAs can be identified. This problem is shared even by experimental approaches for ncRNA detection such as tiling arrays and short read sequencing. We show that the reliability profile plot combined with automated detection of high-reliability regions yields accurate boundaries of structural RNA.

In order to verify this claim, we generated a data set of true positive predictions of a recent `RNAz` (50) screen (39) in *Drosophila melanogaster*, which is based on a PECAN alignment of the 12 *Drosophila* genomes (7). In this screen, 120 nucleotide long alignment slices of the whole-genome alignment, called *windows*, at every 40 nucleotides are evaluated with `RNAz`. Each set of overlapping windows with `RNAz`$P \geq 0.5$ in either orientation is combined into a *locus*. As true positives, we selected 287 out of the about predicted 42,000 loci that overlap with at least one of the FlyBase-annotated structural non-coding RNAs in *Drosophila melanogaster*. For each of the loci, we selected all sequences that have at most 25% gaps in the whole-genome alignment slice of the locus region. This filter criterion was proposed by (39) to remove weakly aligned sequences. To enable prediction of ncRNA boundaries that exceed the RNAz prediction and to add background signal, each sequence was extended by genomic context. While large context increases the computational cost of the subsequent re-alignment, its size should significantly exceed the expected deviation between true ncRNA boundaries and `RNAz` prediction, which can be estimated from the annotation (cf. Fig. 2(a)). Thus, we added 100 nucleotides up and down-stream, as long as we stay in the same syntenic block. For only 9 of the 287 loci, we had to be content with a shorter "available" context. These extended locus alignments consist of at average 8.5 sequences where the sequences have an average length of about 325 and a maximal length of 560 nucleotides.

For each locus, we re-aligned its extended sequences in both orientations and calculated according alignment reliabilities, both performed simultaneously by `LocARNA-P`. This resulted in a reliability profile per locus, which we projected to the *Drosophila melanogaster* sequence. For predicting boundaries by fitting the two-step function to the profile, we constrained the fit to predict exactly one range. The predicted boundaries were then compared to the boundaries of both the annotated ncRNA and the `RNAz` locus region.

We compare our predictions with the annotation in FlyBase for the assembly used by the `RNAz` screen. Notably, we make a single exception to this rule for microRNAs. Since we expect to identify their structural precursors instead of the (unstructured) mature miRNA, we compared our predictions to the pre-miRNA annotations from mirbase.

Figure 2(a) shows the deviation of the boundaries determined by `LocARNA-P` from the annotated boundaries in a notched box plot. We measure this deviation as a sum of differences between the predicted and annotated 3'-end and 5'-end. Non-overlapping notches indicate a significant difference in the median because a notch represents the approximate 95% confidence interval of the median (5). For understanding the dependency on the strand-orientation, we show medians for analyzing the plus and minus strand or even the annotated strand, finding no significant differences. However, there is a significant difference between the `RNAz` boundaries, and the boundaries detected by `LocARNA-P`. The median for `RNAz` is 47, whereas the median for our method is between 10 and 13 (depending on the strand orientation). This indicates that significant improvements of the boundary prediction, as shown in Figure 2(b), are common. We emphasize that this improvement is even more important for practical applications because RNA folding is well known to be very context-sensitive.

We investigated cases where the `LocARNA-P` prediction differs from the given annotation to a greater extent. Some of these cases are plainly due to incomplete or incorrect annotation. For example, for *snoRNA U3* (FlyBase id *snoRNA:U3:54Aa*) and *smnRNA:331* only partial genes are annotated. In the case of *SnoRNA:3*, the annotation is incorrect for the 2004 assembly used for the `RNAz`-screen (39). In the current assembly, however, the annotation matches the predicted signal.

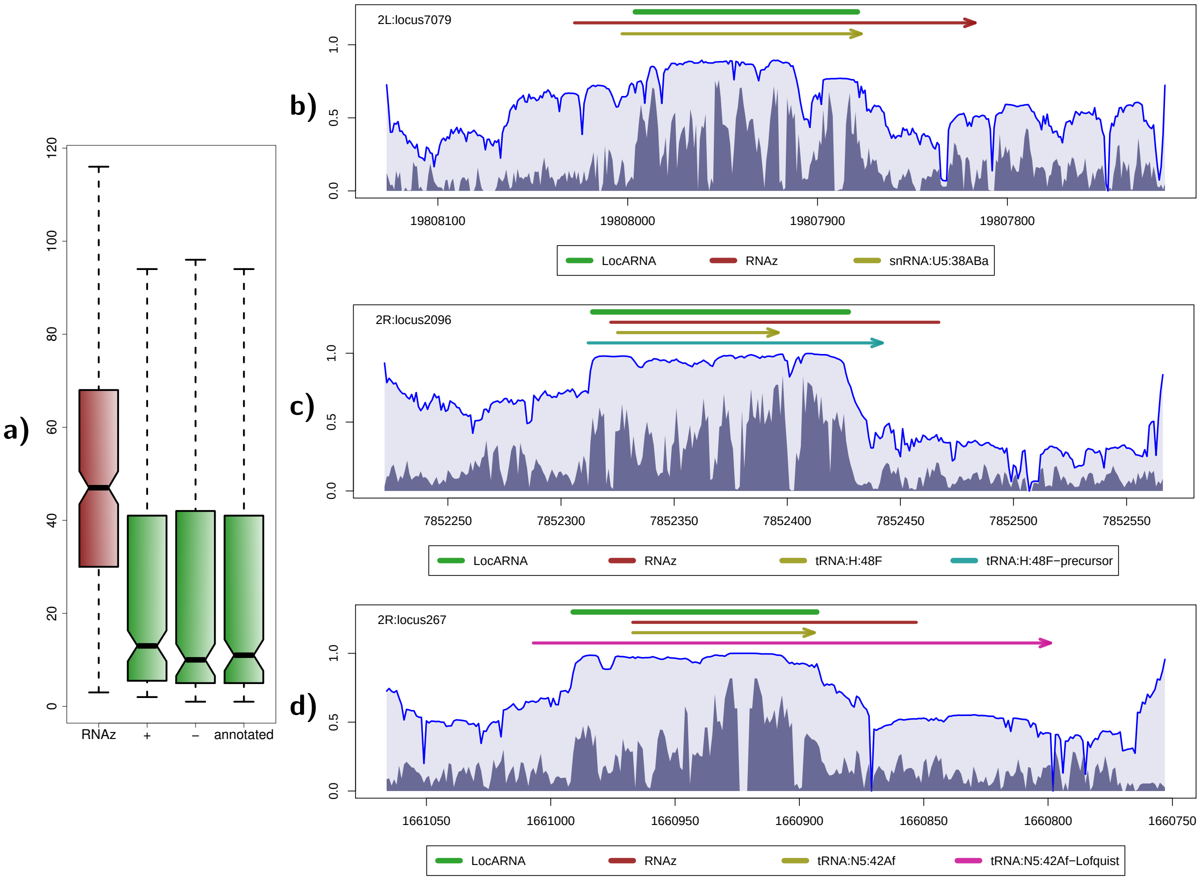

Figure 2: Accurate ncRNA boundaries for *Drosophilids* `RNAz` screen. a) Deviation from annotated boundaries. We compare the deviation of `RNAz` with the deviation of the boundaries as determined with our method, predicting for + strands, - strands, or the strands as annotated. When the notches around the medians do not overlap, there is strong evidence that the medians differ. Panels b), c), and d) show reliability plots with annotated regions, `RNAz` predictions (red) and `LocARNA-P` predictions (green). b) `LocARNA-P` precisely locates the snoRNA:U5:38ABa annotated in FlyBase. c) For tRNA:H:48F our prediction is well correlated with the precursor (cyan line) as described by Frendewey *et al.* (11) (FlyBase annotation). d) In the case of tRNA:N5:42Af, the magenta line shows the tRNA precursor, including the flanking region given by Lofquist and Sharp (29). Here, `RNAz` indicates a 3' extension, whereas `LocARNA-P` indicates the structure in the 5' part of the precursor. As shown by Lofquist and Sharp (29), the 5'-flanking regions of the tRNA5Asn genes differentially arrest RNA polymerase III.

In many cases, however, the predicted extended signals may correspond to precursors with conserved structure, as in the case of miRNAs. For tRNAs, we analyzed this source of incongruence between prediction and annotation in more detail. The tRNAs are known to undergo processing after being transcribed as precursors. The *annotated* tRNA "genes" are always the mature tRNAs. In contrast, the precursor is in almost all cases unknown. There is no agreement in the literature as to the exact extend of the precursor. Morl and Marchfelder (36) estimate a length of only 5-15 nucleotides for the 3'-trailer, while recent deep sequencing data show that this length exceeds often 20 nucleotides (e.g., Lee *et al.* (28)). Consistent with these findings, it is not surprising that our method very often predicts a signal that not only covers the complete mature tRNA, but also extends in both the 5' and 3' direction, indicating that the putative precursors may also form structures outside the range of the mature products of functional importance. In the two examples given in Figures 2(c) and (d) (respective FlyBase ids tRNA:H:48F and tRNA:N5:42Af), we compare the predictions to precursors described in the literature. In the case of tRNA:N5:42Af, `RNAz` predicts a 3' extension, whereas `LocARNA-P` unveils a signal in the 5' flanking region. The latter is consistent with the observation (29) that the 5'-flanking regions of the tRNA5Asn genes differentially arrest RNA polymerase III.

This disagreement between `RNAz` and `LocARNA-P` concerning the 5' and 3' flanking region motivated us to look at the length distributions of 5' and 3' flanking regions of tRNAs as predicted by `LocARNA-P`. If these extensions were only due to random fluctuations, then one would assume the same distribution for both 5' and 3' regions. However, Figure 3 shows that the distributions are significantly different. Whereas the predicted 3'-ends coincide well with the mature tRNA, `LocARNA-P` tends to detect an additional structure signal in the 5' region. The non-randomness of this signal strongly suggests that `LocARNA-P` detects a true signal for structural conservation in the 5' part of the tRNA precursors.

2.4 Improving Discrimination Power of ncRNA Screens

All current predictors of structural RNA suffer from a high false discovery rate. In many cases, e.g. for experimental analysis, one is interested in selecting a small set of high-

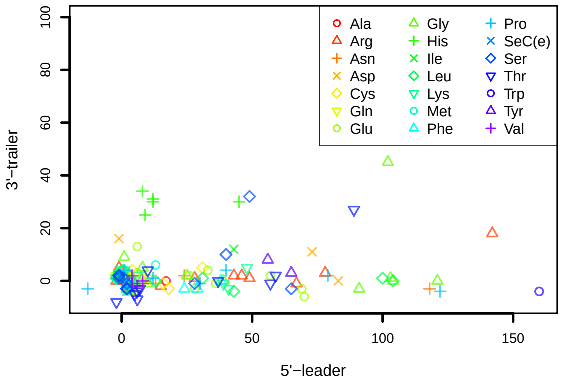

Figure 3: Distribution of predicted lengths of 5' and 3' flanking regions for tRNAs. The figure omits four outliers with 3'-trailers longer than 100. The length distribution suggests that `LocARNA-P` detects a true signal in the 5'-leader of tRNAs precursors.

confidence predictions. In an `RNAz` screen, the most straightforward and common method for this purpose is to rely on `RNAz`'s own evaluation and increase the threshold for positive predictions. Note that `RNAz` evaluates a locus by the maximal ncRNA class probability "`RNAz` max. P" of the contributing windows, since `RNAz` originally predicts probabilities that each single window contains "structural RNA" and then combines overlapping windows with $P \geq 0.5$ into a "locus".

We propose an alternative strategy that re-scores each `RNAz` prediction based on its `LocARNA-P` reliability profile and boundary prediction. We compared the resulting *`LocARNA-P` discriminator* to the currently used `RNAz` max. P discriminator for discriminating `RNAz` loci, which themselves are hits of the `RNAz` screen. To avoid confusion, we emphasize that this differs from estimating the false discovery rates of either tool `RNAz` or `LocARNA-P`. The resulting Figure 4 shows that the novel strategy retains significantly more true positives for a given improvement in specificity.

For our experiment, we select a positive data set consisting of the 287 annotated `RNAz` loci in fly determined for the previous experiment. For the negative set, we generated 250 `RNAz` decoy alignments that consist of windows with `RNAz` P-score ≥ 0.5 by shuffling. For shuffling, we apply a greedy strategy based on the tool `rnazRandomizeAln.pl` of `RNAz`. The details are described in Methods. We preferred this strategy over a generate and test

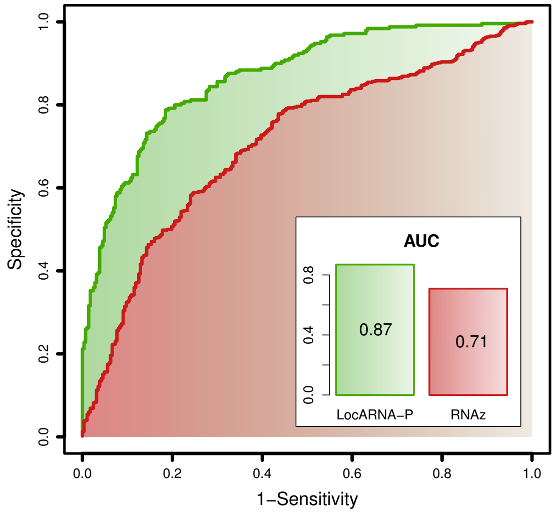

Figure 4: Discriminating ncRNAs. ROC curves for discriminating `RNAz` loci, which are positives of an `RNAz` screen, by `RNAz` itself (using the max. P value) and after re-scoring with `LocARNA-P` by the reliability score. By post-processing *de-novo* ncRNA screens, `LocARNA-P` significantly improves the discriminative power over `RNAz`.

approach, which is computationally expensive since shuffling a large locus consisting of several windows will rarely result in a `RNAz` decoy locus.

The red ROC curve of Figure 4 displays the effect of using a higher RNAz threshold between $0.5 < p_{\text{threshold}} < 1.0$, as commonly used to reduce the false discovery rate. The green curve shows the characteristic of our reliability based discriminator. For the negative set, we furthermore extended the decoy loci alignments by their shuffled original alignment context and obtained the context-extended sequences as described in the previous section for the positive locus alignments. For positive and negative examples, we computed the `LocARNA-P` reliability profile of the corresponding locus. From these profiles, we predicted boundaries and determined the average column reliability inside and outside of the predicted boundaries. The *reliability-based discriminator* is defined as the difference between average inside and outside reliability.

The discriminative power of the two measures, given as the area under the curve (AUC), is estimated at 0.71 for `RNAz` max. P compared to 0.87 for the new reliability-based measure.

Correlation of reliability The reliability of alignments, as defined in this paper[1], is a novel feature that has not been used in ncRNA-screens before. Thus, we compared reliability to features that have been previously applied to measure the quality of sequence-structure alignment within the prediction of ncRNAs. To this end, we investigated how well certain features, including the average pairwise sequence identity (APSI), the structural conservation index (SCI (50)), and our new reliability measure, correlate to alignment quality on a benchmark set of 10-fold (reference) alignments from the Rfam database. For each benchmark alignment, we re-aligned the sequences using `LocARNA-P` and compared the produced alignment with the reference alignment using the `compalign` score, which refers to a sum-of-pairs score (SPS) introduced in this specific form with Bralibase 2.1 (53). We observed that the SCI does not correlate well with the quality of the alignments as measured by the `compalign` score. APSI shows better correlation (0.69), which is expected because sequences with high APSI are much easier to align than sequences with low APSI. However, the highest correlation (0.78) is achieved by the reliability score. To rule out the possibility that this correlation is observed only in `LocARNA-P`-generated alignments, we also calculated the reliability scores for alignments that were produced but by a second sequence-structure alignment method (`Lara` (1)). We found a very strong correlation (0.99) between the reliabilities for the alignments of the different methods. This finding indicates that `LocARNA-P` reliabilities yield a very good general model of sequence structure alignment.

3 Discussion

Finding structurally conserved regions is one of the main tasks in the analysis of non-coding RNA. Approaches using sequence alignments as input heavily rely on alignment quality and are thus strongly limited by the low availability of high-quality alignments. Sankoff-style methods for the simultaneous alignment and folding of the homologous RNA sequences overcome this limitation and are thus considered the gold standard for that purpose. However, the biological interpretation of such alignments poses major problems

[1] Note that the term reliability has been used in the related context of RNA structure prediction before, albeit obviously with a very different definition (23).

because straightforward resampling methods, which are routine in assessing the significance of pairwise sequence alignments, are precluded by their extensive resource consumption.

By defining sequence-structure reliability, we introduced a novel measure of the local and global quality of sequence-structure alignments. This allows an analysis of the local quality of computed alignments and, in particular, improves the prediction of ncRNAs. Column-wise reliabilities that capture the confidence in specific alignment columns are calculated from match probabilities. Computing base and base pair match reliabilities allows structured regions to be distinguished from unstructured regions of the alignment in the form of a reliability profile that allows visual inspection and interpretation.

A score based on reliabilities turned out to be highly correlated with the alignment quality of sequence-structure alignments, where *quality* is understood in terms of similarity to reference alignments measured by the `compalign` score. The correlation is independent of whether these alignments were generated by `LocARNA-P` or other tools. This result shows that reliability captures general properties of correct sequence-structure alignments. Remarkably, the structural conservation index (SCI), reported in Gruber *et al.* (15) as the best method for detecting conserved secondary structure in sequence alignments, is a much worse measure of the alignment quality of sequence-structure alignments.

Furthermore, the reliability profiles can even be used to improve the computational prediction of ncRNA transcripts.

We evaluated the two most important tasks of such an analysis. Albeit we performed this study for an RNAz-screen, the suggested refinement would as well work for other RNA predictors and would furthermore support the prediction of structured RNA in transcripts assembled from RNA-seq data. First, we determined accurate transcript boundaries from the projected profile. This is of particular importance because an incorrect boundary prediction compromises all subsequent analysis steps that require a model of the secondary structure. We show that `LocARNA-P` improves `RNAz` predicted boundaries by an average factor of about three. Second, based on the profile combined with the predicted boundaries, we computed a new discriminator for ncRNAs. Applied in the post-processing step of an `RNAz` screen, this discriminator is significantly stronger in distinguishing true `RNAz` hits from false-positive predictions than the max. P discriminator that is currently proposed by `RNAz` for this purpose. This improvement is of particular relevance, because it reduces the number of ncRNA candidates for subsequent, more expensive, analysis steps.

4 Methods

4.1 Sequence-Structure Reliability

We follow the common pattern of introducing reliability for pairwise alignment, and then extending the idea to multiple alignments. Initially, we consider two sequences, A and B, with corresponding base pair probability matrices, P^A and P^B, respectively. The matrices are usually calculated from the respective sequence by McCaskill's partition function approach (33). We are going to compute a high-quality sequence-structure alignment of A and B together with additional information on the confidence in the individual alignment columns and the predicted consensus structure. These column-wise reliabilities facilitate the interpretation of the sequence-structure alignment and allow for further automated analysis. A sequence-structure alignment is a pair consisting of a sequence alignment $\mathcal{A}$ of A and B and a secondary structure $\mathcal{S}$ of $\mathcal{A}$ that maximizes a scoring function composed of sequence similarity and structure similarity. $\mathcal{A}$ consists of a set of base matches written as $i \sim k$, where i is a position in A, and k a position in B. The consensus secondary structure $\mathcal{S}$ for an alignment $\mathcal{A}$ consists of a set of arc matches $(i,j) \sim (k,l)$, where $i \sim k \in \mathcal{A}$ and $j \sim l \in \mathcal{A}$ are two matches in $\mathcal{A}$, (i,j) is a base pair on sequence A and (k,l) is a base pair on sequence B.

Our scoring function assigns a similarity value to a pair $(\mathcal{A}, \mathcal{S})$. It combines a log-odds score for the probabilities of matched base pairs with a Ribosum-like scoring of sequential matches (22) and uses affine gap cost. It provides substantial improvements over the original scoring function of `LocARNA`, which has been applied in (52). The scoring function is described in detail in the Appendix.

4.2 Match Probabilities

We are going to calculate reliabilities based on match probabilities. Probability-based reliabilities have been introduced by `Probcons` (9) for sequence alignment using profile HMMs, we extend it in the context of sequence-structure alignment. We are going to define probabilities of single base matches and arc matches in sequence-structure alignments. Therefore, we define probabilities of pairs $(\mathcal{A}, \mathcal{S})$ of alignment and consensus structure. Such probabilities are defined under the assumption of a Boltzmann distribution over pairs $(\mathcal{A}, \mathcal{S})$ that is based on the scoring of `LocARNA`.

Computing match probabilities via a statistical mechanics model has been successfully introduced for *sequence* alignment by `Probalign` (40). However, the analogous approach was not considered for multiple *sequence-structure* alignment. By assuming a Boltzmann distribution, our approach differs from methods that obtain probabilities from generative models such as hidden Markov models (HMMs) or stochastic context free grammars (SCFGs). Such methods produce structures with probabilities determined by given transition probabilities. The main advantage of the non-generative approach taken here is that the underlying similarity scores have a very intuitive semantic.

The *probability of a pair of alignment and consensus structure* $(\mathcal{A}, \mathcal{S})$, written $\Pr[(\mathcal{A}, \mathcal{S})|A, B]$, is calculated by dividing the Boltzmann weight $\exp(-\beta\,\mathrm{Sc}(\mathcal{A}, \mathcal{S}))$ by Z_{AB}, where Sc denotes the scoring function of the previous subsection, the *inverse temperature* β controls the distribution of probabilities and the *partition function* Z_{AB} is the sum over the Boltzmann weights of all pairs $(\mathcal{A}, \mathcal{S})$. Once the probability of an alignment and consensus structure pair is defined, we can define base match and arc match probabilities. The *probability of an arc match* $(i, j) \sim (k, l)$, where $(i, j) \in P^A$ and $(k, l) \in P^B$, is defined as the sum of all probabilities of pairs $(\mathcal{A}, \mathcal{S})$ that contain this match. Similarly, the *probability* $\Pr[i \sim k|A, B]$ *of a base match* $i \sim k$ is defined as the sum of the probabilities of all alignment consensus structure pairs matching the two bases A_i and B_k. For later use, we introduce an exclusive base match probability $\Pr[i \sim_s k|A, B]$ of a match $i \sim k$ that is not part of a structural match.

4.3 Efficient Calculation of Match Probabilities

The match probabilities are efficiently calculated by `LocARNA-P` using dynamic programming for computing partition functions inside and outside of subsequence pairs $A_i \dots A_j$ and $B_k \dots B_l$. Finally, these partition functions are combined for obtaining probabilities. The use of inside and outside algorithms for this purpose is well known from stochastic context-free grammars. However, a naive application of this algorithm results in a very high time complexity of $O(n^6)$ and space complexity of $O(n^4)$, where n is the length of the input sequence. This rapid growth of space and time requirements with the input size would limit the algorithm to only small instances. As described in detail in the Methods section, we were able to calculate the match probabilities in a much lower complexity of $O(n^4)$ time and $O(n^2)$ space, which is essential for the applicability of the approach in practice.

4.4 Column Reliabilities

Based on the pairwise match probabilities, we define column reliabilities for a given or generated multiple alignment $\mathcal{A}$ of K sequences $S_1, \dots, S_K$. The *sequence reliability* $\mathrm{seqrel}_{\mathcal{A}}(q)$ *of a column* q and the *base pair reliability* $\mathrm{bprel}_{\mathcal{A}}(q, q')$ *of a pair of columns* q *and* q' are defined as the sum (over all pairwise alignments) of the base match probabilities associated with column q, and the arc match probabilities for columns q and q', respectively:

$$\mathrm{seqrel}_{\mathcal{A}}(q) = \frac{1}{\binom{K}{2}} \sum_{1 \le a < b \le K} \Pr[\bar{\mathcal{A}}_a(q) \sim_s \bar{\mathcal{A}}_b(q) | S_a, S_b] \tag{1}$$

$$\mathrm{bprel}_{\mathcal{A}}(q, q') = \tag{2}$$
$$\frac{1}{\binom{K}{2}} \sum_{1 \le a < b \le K} \Pr[(\bar{\mathcal{A}}_a(q), \bar{\mathcal{A}}_a(q')) \sim (\bar{\mathcal{A}}_b(q), \bar{\mathcal{A}}_b(q')) | S_a, S_b],$$

where $\bar{\mathcal{A}}_a(q)$ is defined as the position in sequence S_a associated with column q, unless there is no such position. Note that we implicitly ignore terms in the summations where the functions $\bar{\mathcal{A}}_a$ or $\bar{\mathcal{A}}_b$ are undefined. Finally, in addition to the column-wise sequence reliability, we define a *column-wise structure reliability* indicating how reliably the column is aligned and part of a base pair in the consensus structure: $\mathrm{strrel}_{\mathcal{A}}(q) = \sum_{q'<q} \mathrm{bprel}_{\mathcal{A}}(q', q) + \sum_{q<q'} \mathrm{bprel}_{\mathcal{A}}(q, q')$.

Finally, the sum $\mathrm{seqrel}_{\mathcal{A}}(q) + \omega\, \mathrm{strrel}_{\mathcal{A}}(q)$ defines an overall reliability for column q, where we apply a factor ω to control the weight of structure. These values will be called reliability profile of an alignment.

4.5 Predicting Boundaries from a Reliability Profile

The reliability profiles are now used to determine regions of conserved secondary structure. Formally, let $f : \{1, \dots, n\} \to \mathbb{R}$ denote a reliability profile of length n, i.e. $f(i) = \mathrm{seqrel}_{\mathcal{A}}(q) + \omega\, \mathrm{strrel}_{\mathcal{A}}(q)$. We fit a two-step function g to f, such that g approximates f as good as possible. Therefore we determine constants a and b, such that $\sum_{i=1}^{n}(f(i) - g(i))^2 + \delta(f(i-1), f(i))\Delta$ is minimal for all $g : 1, \dots, n \to \{a, b\}$, where $\delta(x, x) = 0$ and $\delta(x, y) = 1$ for $x \neq y$, $\Delta \in \mathbb{R}$ is a penalty for switching between the values of g, and $g(0) := a$. Basically, we perform a least square distance approximation of f extended by a penalty term. The larger value of a and b represents the signal level, whereas the smaller value the background.

For given constants a and b, an optimal function g can be computed by dynamic programming. Therefore, we solve the recursion equations

$$\begin{aligned} A(i) &= (f(i) - a)^2 + \min(A(i-1), B(i-1) + \Delta) \\ B(i) &= (f(i) - b)^2 + \min(A(i-1) + \Delta, B(i-1)) \end{aligned} \tag{3}$$

with initialization $A(0) = 0$ and $B(0) = 0$ for $A(n)$ and $B(n)$ and obtain g by traceback.

For finding optimal constants a and b, we formulate a partition function version of these equations. We choose to optimize the partition function $Z^A(n) + Z^B(n)$ instead of the cost $A(n) + B(n)$, since the partition functions allows computing partial derivations. This allows finding constants that minimize $Z^A(n) + Z^B(n)$ by gradient descent optimization. For sufficiently high β, such constants will also minimize the cost $A(n)+B(n)$. More details are given in the supplementary information.

4.6 Computing the Reliability Score of an Alignment

Given is an alignment $\mathcal{A}$. We build on the previous definitions of $\mathrm{seqrel}_{\mathcal{A}}$ and $\mathrm{bprel}_{\mathcal{A}}$, which refer to match probabilities computed by `LocARNA-P` from the sequences of the alignment $\mathcal{A}$.

The *reliability score* $\mathrm{relsc}_{\mathcal{A}}$ *of the multiple alignment* $\mathcal{A}$ is $\max_{\mathcal{S}}(\mathrm{relsc}_{\mathcal{A}}(\mathcal{S}))$ divided by the length of $\mathcal{A}$, where $\mathcal{S}$ denotes a consensus structure for the alignment $\mathcal{A}$ and $\mathrm{relsc}_{\mathcal{A}}(\mathcal{S})$ is the sum of reliabilities $\omega\,\mathrm{bprel}_{\mathcal{A}}(q, q')$ over all column pairs (q, q') in $\mathcal{S}$ and $\mathrm{seqrel}_{\mathcal{A}}(q)$ over all columns q that are not paired in $\mathcal{S}$. We apply a fixed weight of structure $\omega \in \mathbb{R}$ in order to weigh structure reliability against sequence reliability. At a weight of $\omega = 2$, sequence and structure have the same influence, since each structure reliability contribution consumes two alignment columns. In `LocARNA-P`, we use a default of $\omega = 3$ to emphasize the structural component.

The reliability score is computed efficiently by a Nussinov-style dynamic programming algorithm. Details are given in the Supplemental Material.

4.7 Generating Decoy Locus Alignments

Rose *et al.* (39) define an *`RNAz` locus alignment* as a slice of the 12-flies PECAN whole genome alignment that is covered by window alignments of at most 120 columns with `RNAz` probability of $P \geq 0.5$ for either the + or − strand. A *decoy locus alignment* is covered by windows with `RNAz` probability $P \geq 0.5$ and has identical length, base composition, and gap pattern and similar conservation pattern to an existing `RNAz` locus alignment. However, a decoy locus alignment is not contained in any genome alignment and therefore cannot be true positive.

We generate such decoys from true `RNAz` locus alignments by gentle shuffling as described in Rose *et al.* (39). Gentle shuffling randomly permutes alignment columns, but exchanges only columns with identical gap pattern and similar sequence conservation. We apply gentle shuffling to successive non-overlapping windows in the locus alignment, but do not shuffle overlapping windows to maintain the locality of the base composition and conservation pattern. To speed up the approach for larger loci, we apply a greedy strategy and allow several attempts per window (at most 25 times) until RNAz evaluates to $P \geq 0.5$. Each window of the resulting alignment is tested for coverage by RNAz $P \geq 0.5$ windows in exactly the way of Rose *et al.* Finally, decoy context is generated by gentle shuffling of original context in the whole genome alignment.

4.8 Benchmarks

We measure the performance of LocARNA-P using the Bralibase 2.1 (53) benchmark set. We have shown (Suppl. Fig. 4) that there is a significant improvement in the quality of multiple alignment compared to competing methods. Furthermore, we assessed the properties of the reliability score in detail. We found that this score is a better predictor of alignment quality than previously used measures such as average sequence similarity and structure conservations index (Suppl. Fig. 1). For details, we refer to supplementary material.

Availability and Supplemental Material

At http://www.bioinf.uni-freiburg.de/Supplements/LocARNA-P/, LocARNA-P can be downloaded as part of the LocARNA software package, which is open source software and released under the GNU general public license. Documentation and scripts for running the refinement of an RNAz screen are provided. Furthermore, supplemental material is available for this article. Additional results, figures and data files can be viewed and downloaded at the same url.

Acknowledgments

This work was supported by the German Research Foundation (grants BA 2168/3-1 as well as BA 2168/2-2, STA 850/7-1 [under the auspices of SPP-1258 “Sensory and Regulatory RNAs in Prokaryotes”], the Excellence Initiative of the German Federal and State Governments (EXC 294 to R.B) and WI 3628/1-1), ESF grant no. 14494 (administered by the Sächsische AufbauBank), the German Federal Ministry of Education and Research (BMBF grant 0313921 FRISYS), and the Austrian Gen-AU project “regulatory ncRNAs”.

We thank the anonymous reviewers of a previous manuscript for their valuable comments.

Appendix: The LocARNA-P Dynamic Programming Algorithm

In this appendix, we explain dynamic programming algorithm of `LocARNA-P` and the necessary foundations due to `LocARNA` in formal detail. Recall that in Methods, we defined probabilities for matches in the alignment of two RNA sequences A and B with associated base pair probability matrices P^A and P^B, respectively. Suitable matrices are usually obtained from the respective RNA sequence using McCaskill's algorithm (`RNAfold -p`). The probabilities are defined on the basis of the alignment score of `LocARNA`, by assuming Boltzmann-distribution of alignment consensus structure pairs. This allows building on the highly accurate and established `LocARNA`-score. After describing the `LocARNA`-score and the algorithm of `LocARNA`, we present the algorithm of `LocARNA-P`, which efficiently computes these probabilities. The use of these probabilities in a probabilistic consistency transformation for progressive multiple alignment and iterative alignment refinement is discussed in the Supplemental Material.

A.1 Preliminaries: RNA Alignment by `LocARNA`

`LocARNA` is a Sankoff-style algorithm, which simultaneously folds and aligns RNA sequences. The original Sankoff algorithm (41) provides a general solution to the problem of simultaneously computing an alignment and a common secondary structure of the two aligned sequences. Without heuristic restrictions, the problem requires $O(n^6)$ CPU time and $O(n^4)$ memory, where n is length of the RNA sequences to be aligned. In contrast to Sankoff-style methods like `FoldAlign` (13, 18) and `dynalign` (31), `PMcomp` (21) and `LocARNA` employ structure models of the RNAs, which are reasonably obtained using McCaskill's algorithm (33) on the basis of a full-featured energy model.

Alignment Score. Define the *single-stranded part of the alignment, denoted by $\mathcal{A}_s$,* by: if $i \sim k \in \mathcal{A}_s$ then there is no pair $j \sim l$ such that $(i,j) \sim (k,l) \in \mathcal{S}$ or $(j,i) \sim (k,l) \in \mathcal{S}$.

LocARNA determines the pair $(\mathcal{A}, \mathcal{S})$ that maximizes the score function

$$\mathrm{Sc}(\mathcal{A}, \mathcal{S}) = \sum_{(i,j)\sim(k,l)\in\mathcal{S}} \tau(i,j;k,l) + \sum_{i\sim k\in\mathcal{A}_s} \sigma(i,k) - N_{\mathrm{gap}}\gamma, \tag{4}$$

where $\tau(i,j;k,l)$ is the score for matching the arcs (i,j) and (k,l), $\sigma(i,k)$ is the similarity score for a (mis)match of positions i and k in A and B respectively, γ is the gap score parameter and N_{gap} is the number of insertions and deletions in the alignment $\mathcal{A}$. Although we define and henceforth discuss only linear gap cost to ease presentation, the actual LocARNA-score features affine gap cost, which is supported by LocARNA, as well as by our implementation of LocARNA-P, with very moderate space and time overhead.

We use arc-match scores $\tau(i,j;k,l) := \Psi^A_{ij} + \Psi^B_{kl}$, where Ψ^A_{ij} and Ψ^B_{kl} are base pair scores that are derived from the base pairing probability matrices of the two individual sequences. More precisely, we define

$$\Psi^X_{ij} = \begin{cases} \log \frac{P^X_{ij}}{p^X_0} \Big/ \log \frac{1}{p^X_0} & \text{if } P^X_{ij} \geq p^* \\ -\infty & \text{otherwise,} \end{cases} \tag{5}$$

where P^X_{ij} is equilibrium pairing probability for sequence $X \in \{A, B\}$ as computed by McCaskill's algorithm (33), p^X_0 is the expected probability for a pairing to occur at random in sequence X and p^* is the cut-off probability, below which the arcs are ignored. Formally, this is expressed by assigning $-\infty$ as weight in this case. We call base pairs with probability greater or equal p^* *significant*. The term $\log P^X_{ij}/p^X_0$ is the log-odds score for having a specific base pairing against the null model of a random pairing, and $\log 1/p^X_0$ is a normalization factor that transforms the weights to a maximum of 1. This normalization is introduced to ease balancing the sequence score against the structure score.

LocARNA-P uses exactly the same scoring function as LocARNA. However, it does not maximize the score according to this function, but computes match probabilities based on this scoring function. How match probabilities relate to the scoring function is detailed in the corresponding paragraph in Methods.

Efficient Alignment Using Base Pair Probabilities. `LocARNA` maximizes its score by efficiently evaluating a recursion equation using dynamic programming. The essential improvement of `LocARNA` over `PMcomp` is through the consideration of only significant base pairs in predicted structures. As we argued earlier (cf. Will *et al.* (52)), by filtering we keep only $O(n)$ significant base pairs in each sequence and only $O(1)$ that share a given right end. Consequently, `LocARNA` improves the time complexity of `PMcomp` from $O(n^6)$ to $O(n^4)$ and, even more importantly, the space complexity from $O(n^4)$ to $O(n^2)$. The favorable time and space complexity of `LocARNA` is retained when extending the approach for the computation of match probabilities in `LocARNA-P`. Because of this structural analogy of the algorithms, we review the recursion structure of `LocARNA` in detail.

Both `PMcomp` and `LocARNA` define two 4-dimensional matrices M and D that are filled recursively. $M_{i\,j;k\,l}$ is defined as the maximal score of an alignment of subsequences $A_{i..j}$ and $B_{k..l}$. $D_{i\,j;k\,l}$ is the best score of an alignment of $A_{i..j}$ and $B_{k..l}$ with additional condition that the base pairs (i, j) and (k, l) are matched. The `LocARNA`/`PMcomp` recursion can be written in the form $M_{i\,i-1;k\,k-1} = 0$,

$$M_{i\,j;k\,l} = \max \begin{cases} M_{i\,j-1;k\,l-1} + \sigma(j, l) \\ M_{i\,j-1;k\,l} + \gamma \\ M_{i\,j;k\,l-1} + \gamma \\ \max_{j'l'} M_{i\,j'-1;k\,l'-1} + D_{j'\,j;l'\,l} \end{cases}$$

$$D_{i\,j;k\,l} = M_{i+1\,j+1;k-1\,l-1} + \tau(i, j; k, l).$$

In contrast to `PMcomp`, `LocARNA` evaluates this recursion keeping only $O(n^2)$ entries in memory at any time. Due to the restriction to significant base pairs, the fourth case of the M recursion runs over only $O(1)$ pairs of significant base pairs and consequently the total algorithm has $O(n^4)$ time complexity.

Regarding the space complexity, we first observe that the D-entries are needed only for matches $(i, j) \sim (k, l)$ of significant base pairs., i.e. only $O(n^2)$ many. Thus, the D-matrix can be easily represented by a 2D-matrix indexed by base pairs. Second, due to the special structure of the M-recursion, which fixes the left subsequence ends i and k, we can compute all entries $D_{i\,\cdot;k\,\cdot}$ recursing only to entries $M_{i+1\,\cdot;k+1\,\cdot}$.[2] Thus, a single $O(n^2)$ sized M-matrix is sufficient for the computation of all $D_{i\,\cdot;k\,\cdot}$, since the matrix can be reused for all left ends i and k.

An extension that explicitly incorporates base pair stacking without increasing complexity is described by Bompfunewerer *et al.* (3).

A.2 Partition Function Version of `LocARNA`

The calculation of alignment match probabilities by `LocARNA-P` is based on calculating partition functions. Recall that the probability of a pair of alignment and consensus structure $(\mathcal{A}, \mathcal{S})$ is given by

$$Pr[(\mathcal{A},\mathcal{S})|A,B] = \exp(-\beta\,\mathrm{Sc}(\mathcal{A},\mathcal{S}))\, Z_{AB}^{-1}$$

where the *partition function Z_{AB} for sequences A and B* is defined as

$$Z_{AB} := \sum_{(\mathcal{A},\mathcal{S}) \text{ of } A,B} \exp(-\beta\,\mathrm{Sc}(\mathcal{A},\mathcal{S}))$$

and β is a parameter that controls the distribution, called the inverse temperature.

Calculating match probabilities in `LocARNA-P` consists of three phases, which are comparable to the algorithm of Hofacker and Stadler (20), but go beyond this algorithm in terms of complexity. First, an inside dynamic programming algorithm computes inside partition functions. This part of the `LocARNA-P` algorithm has the same recursion structure as the `LocARNA` algorithm. Second, a corresponding outside algorithm calculates outside partition functions. We devise a dynamic programming algorithm that computes these values in the given complexity envelope. Finally, we show how to obtain the single base and base pair match probabilities. Again, this phase remains within the complexity bounds.

[2] We introduce notation using index $\cdot$ as wildcard. For example, $M_{i+1\,\cdot;k+1\,\cdot}$ refers to the matrix slice of entries $M_{i+1\,j;k+1\,l}$, where $i+1 \leq j \leq n$ and $k+1 \leq l \leq m$. We freely use analogous notation in the following.

A.2.1 Inside Algorithm

We define two four-dimensional matrices

$$Z^M_{i\,j;k\,l} = \sum \left\{ \exp(-\beta\,\mathrm{Sc}(\mathcal{A},\mathcal{S})) \;\middle|\; \begin{array}{l} \mathcal{A} \text{ alignment of} \\ A_{i..j} \text{ and } B_{k..l}, \\ \mathcal{S} \text{ consensus secondary} \\ \text{structure for } \mathcal{A} \end{array} \right\}$$

and

$$Z^D_{i\,j;k\,l} = \sum \left\{ \exp(-\beta\,\mathrm{Sc}(\mathcal{A},\mathcal{S})) \;\middle|\; \begin{array}{l} \mathcal{A} \text{ alignment of} \\ A_{i..j} \text{ and } B_{k..l}, \\ \mathcal{S} \text{ consensus secondary} \\ \text{structure for } \mathcal{A}, \\ \text{where } (i,j) \sim (k,l) \in \mathcal{S} \end{array} \right\}.$$

Note that $Z^D_{i\,j;k\,l}$ is valid (and later has to be computed) only for significant[3] base pairs (i,j) and (k,l).

The matrix entries are recursively computed by following equations (cf. Fig. 5(a)):

$$\begin{aligned} Z^M_{i\,i-1;k\,k-1} &= 1 \\ Z^M_{i\,j;k\,l} &= \sum \begin{cases} Z^M_{i\,j-1;k\,l-1} \cdot \exp(-\beta\sigma(j,l)) \\ Z^M_{i\,j-1;k\,l} \cdot \exp(-\beta\gamma) \\ Z^M_{i\,j;k\,l-1} \cdot \exp(-\beta\gamma) \\ \sum_{j'l'} Z^M_{i\,j'-1;k\,l'-1} \cdot Z^D_{j'\,j;l'\,l} \end{cases} \\ Z^D_{i\,j;k\,l} &= Z^M_{i+1\,j+1;k-1\,l-1} \cdot \exp(-\beta\tau(i,j;k,l)). \end{aligned}$$

These equations are a direct translation of the `LocARNA` recursion to its partition function variant. The translation is straightforward because the decomposition of the `LocARNA`

[3] i.e. significant according to respective base pair probability matrices P^A and P^B

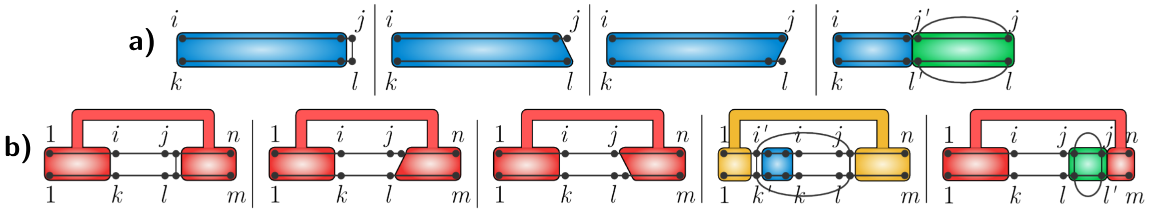

Figure 5: Inside and outside decomposition by the recursions. The blue regions correspond to the matrix Z^M and the green region to Z^D. The red regions correspond to entries in Z'^M; the yellow region represents an entry of Z'^D. a) Inside. b) Outside.

recursion is unambiguous.

The total partition function is obtained as $Z_{AB} = Z^M_{1\,n;1\,m}$.

A good space and time complexity is achieved using the same ideas as in the original `LocARNA` recursion for maximizing the score. When evaluating the recursion for $Z^M_{1\,n;1\,m}$, we compute and store the entries $Z^D_{i\,j;k\,l}$ for significant base pairs (i, j) and (k, l). A computation order of increasing $j - i$ avoids dependency conflicts. The entries require $O(n^2)$ space due to the number of significant base pairs. Note that the matrix Z^D is conveniently implemented as a 2-dimensional array that is indexed with base pairs. One entry $Z^D_{i\,j;k\,l}$ depends only on entries of the matrix slice $Z^M_{i\,\cdot;k\,\cdot}$ and other values in Z^D. Therefore, efficient computation requires only $O(n^2)$ additional space for the matrix slice. The matrix slice is implemented as a two-dimensional array, which is reused for the computation of each Z^D-entry. Time complexity is only $O(n^4)$, since computing one entry in Z^M is performed in time $O(1/p^*) = O(1)$, when only significant base pairs are considered.

The outside algorithm needs to access Z^D, hence this matrix is kept in memory throughout.

A.2.2 Outside Algorithm

The outside algorithm computes partition functions of alignments outside of subsequences $A_{i..j}$ and $B_{k..l}$ and corresponding consensus structures. An *alignment of A and B outside i..j and k..l* is an alignment of A and B containing only edges $i' \sim k'$, where $i' < i$ and $k' < k$ or $j < i'$ and $l < k'$.

We define

$$Z'^{M}_{i\,j;k\,l} = \sum \left\{ \exp(-\beta\,\mathrm{Sc}(\mathcal{A},\mathcal{S})) \;\middle|\; \begin{array}{l} \mathcal{A} \text{ alignment of } A \text{ and } B \\ \text{outside } i..j \text{ and } k..l, \\ \mathcal{S} \text{ consensus secondary} \\ \text{structure for } \mathcal{A}. \end{array} \right\}$$

$Z'^{D}_{i\,j;k\,l}$ is valid only for i,j,k, and l, where significant base pairs (i,j) and (k,l) exist for P^A and P^B respectively. Then, it is defined as $Z'^{D}_{i\,j;k\,l} := Z'^{M}_{i\,j;k\,l}$ and is understood as the partition function outside of the match of base pairs (i,j) and (k,l). Note that we introduce the extra matrix Z'^{D} for preparing the space optimization.

The matrix entries are recursively computed after initialization $Z'^{M}_{i\,n;k\,m} = Z^{M}_{1\,i-1;1\,l-1}$ by

$$Z'^{M}_{i\,j;k\,l} = \sum \begin{cases} Z'^{M}_{i\,j+1;k\,l+1} \cdot \exp(-\beta\sigma(j,l)) \\ Z'^{M}_{i\,j+1;k\,l} \cdot \exp(-\beta\gamma) \\ Z'^{M}_{i\,j;k\,l+1} \cdot \exp(-\beta\gamma) \\ \sum_{i'<i,k'<k} Z'^{D}_{i'\,j+1;k'\,l+1} \cdot Z^{M}_{i'+1\,i-1;k'+1\,k-1} \\ \qquad\qquad \cdot \exp(-\beta\tau(i',j+1;k',l+1)) \\ \sum_{j'>j,l'>l} Z'^{M}_{i\,j';k\,l'} \cdot Z^{D}_{j+1\,j';l+1\,l'} \end{cases}$$

and $Z'^{D}_{i\,j;k\,l} = Z'^{M}_{i\,j;k\,l}$. An illustration of the underlying decomposition is given in Figure 5(b).

So far, the recursion follows the lines of Hofacker and Stadler (20). However, we restructure the evaluation of these recursions in `LocARNA-P` in order to maintain the complexity bounds. For initialization we use the inside matrix slice $Z^{M}_{1\,\cdot;1\,\cdot}$, which can be recomputed in $O(n^2)$ time.[4]

Then, we compute all entries $Z'^{D}_{i\,j;k\,l}$ for significant base pairs (i,j) and (k,l) in the order from outside to inside, i.e. for decreasing distances $j-i$. As in the case of the inside partition functions, Z'^{D} is implemented as a two-dimensional array of size $O(n^2)$. For

[4] In our implementation, we skip this recomputation, since the matrix is still available from the last step of the inside algorithm.

obtaining all entries $Z'^D_{i\,\cdot;k\,\cdot}$, we fill a matrix slice $Z'^M_{i\,\cdot;k\,\cdot}$. During this computation for fixed i and k, we recurse to four different kind of matrix entries. First and second, we recurse to entries of matrices Z'^D and Z^D. Both are maintained in $O(n^2)$ space and dependencies are resolved due to computation order. The same holds for the third kind of entries in the matrix slice $Z'^M_{i\,\cdot;k\,\cdot}$, where dependencies are resolved by computation of entries $Z'^M_{i\,j;k\,l}$ in the order of decreasing j and l. However, there is a fourth kind of entry, namely those of the form $Z^M_{\cdot\,i;\cdot\,k}$. This matrix slice is recomputed in $O(n^2)$ time each time before we start filling a matrix slice $Z'^M_{i\,\cdot;k\,\cdot}$. Clearly, this slice adds another space of $O(n^2)$. The space for the matrix slices $Z'^M_{i\,\cdot;k\,\cdot}$ and $Z^M_{\cdot\,i;\cdot\,k}$ is reused for each left ends i and k of significant base pairs.

Despite of the necessary recomputation of slices $Z^M_{\cdot\,i;\cdot\,k}$ the time complexity is $O(n^4)$. Here, we argue again that summations run only over pairs of significant arcs and consequently take constant time.

A.2.3 Calculation of Alignment Edge Probabilities

The probability of a structural alignment edge is easily computed as

$$P((i,j)\sim(k,l)|A,B) = \frac{1}{Z_{AB}} \cdot Z^D_{i\,j;k\,l} \cdot Z'^D_{i\,j;k\,l}$$

from the efficiently computed matrices Z^D and Z'^D.

Computing the probabilities of base matches requires a case distinction on the "immediately enclosing" arc match $(i,j)\sim(k,l)$ of a base match $x\sim y$. $P(x\sim y|A,B)$ is efficiently computed as

$$\frac{\exp(-\beta\sigma(x,y))}{Z_{AB}} \begin{pmatrix} \displaystyle\sum_{(i,j)\sim(k,l)} \begin{pmatrix} Z'^D_{i\,j;k\,l} \\ \cdot\exp(-\beta\tau(i,j;k,l)) \\ \cdot Z^M_{i+1\,x-1;k+1\,y-1} \\ \cdot Z^M_{x+1\,j-1;y+1\,l-1} \end{pmatrix} \\ +Z^M_{1\,x-1;1\,y-1}\cdot Z^M_{x+1\,n;y+1\,m} \end{pmatrix}. \tag{6}$$

Note that we need to cover the case of no enclosing arc match explicitly.

The quadratic space envelope requires recomputation of $Z^M_{i+1\,\cdot;k+1\,\cdot}$ and $Z^M_{\cdot\,k-1;\cdot\,l-1}$, for each $(i,j) \sim (k,l)$. Given Z^D, all $Z^M_{i+1\,\cdot;k+1\,\cdot}$ can be clearly recomputed in $O(n^2)$. Note that also all $Z^M_{\cdot\,k-1;\cdot\,l-1}$ can be recomputed in $O(n^2)$ by a right-reducing variant of the given left-reducing recursion for the matrix Z^M.

For efficient evaluation, one regroups the computation by iterating over all $(i,j) \sim (k,l)$ and accumulating the probability contributions of each arc pair to all $P(x \sim y|A,B)$. In this way, recomputation causes a time complexity of $O(n^4)$ for the computation of all base match probabilities. However, this computation is still an expensive step of the entire algorithm. Considering only pairs of arcs with a match probability larger than or equal to p^* (or some independently chosen cut-off probability) is a reasonable, easily applicable heuristic that reduces the cost of this computation step in practice.

References

1. Bauer, M., Klau, G. W., and Reinert, K. (2007). Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization. *BMC Bioinformatics*, **8**, 271.

2. Bertone, P., Stoc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., Rinn, J. L., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M., and Snyder, M. (2004). Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242–2246.

3. Bompfunewerer, A. F., Backofen, R., Bernhart, S. H., Hertel, J., Hofacker, I. L., Stadler, P. F., and Will, S. (2008). Variations on RNA folding and alignment: lessons from Benasque. *Journal of Mathematical Biology*, **56**(1-2), 129–144.

4. Bradley, R. K., Pachter, L., and Holmes, I. (2008). Specific alignment of structured RNA: stochastic grammars and sequence annealing. *Bioinformatics*, **24**(23), 2677–83.

5. Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. Wadsworth.

6. Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., Sementchenko, V., Piccolboni, A., Bekiranov, S., Bailey, D. K., Ganesh, M., Ghosh, S., Bell, I., Gerhard, D. S., and Gingeras, T. R. (2005). Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, **308**, 1149–1154.

7. Clark, A. G., Eisen, M. B., Smith, D. E., and MacCallum, I. (2007). Evolution of genes and genomes on the drosophila phylogeny. *Nature*, **450**(7167), 203–18.

8. Coventry, A., Kleitman, D. J., and Berger, B. (2004). MSARI: multiple sequence alignments for statistical detection of RNA secondary structure. *Proc. Natl. Acad. Sci. USA*, **101**(33), 12102–7.

9. Do, C. B., Mahabhashyam, M. S. P., Brudno, M., and Batzoglou, S. (2005). ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res*, **15**(2), 330–40.

10. Do, C. B., Foo, C.-S., and Batzoglou, S. (2008). A max-margin model for efficient simultaneous alignment and folding of RNA sequences. *Bioinformatics*, **24**(13), i68–76.

11. Frendewey, D., Dingermann, T., Cooley, L., and Soll, D. (1985). Processing of precursor tRNAs in Drosophila. Processing of the 3' end involves an endonucleolytic cleavage and occurs after 5' end maturation. *Journal of Biological Chemistry*, **260**(1), 449–54.

12. Gardner, P. P., Wilm, A., and Washietl, S. (2005). A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Research*, **33**(8), 2433–9.

13. Gorodkin, J., Heyer, L., and Stormo, G. (1997). Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res*, **25**(18), 3724–32.

14. Gruber, A. R., Kilgus, C., Mosig, A., Hofacker, I. L., Hennig, W., and Stadler, P. F. (2008a). Arthropod 7SK RNA. *Mol Biol Evol*, **25**(9), 1923–30.

15. Gruber, A. R., Bernhart, S. H., Hofacker, I. L., and Washietl, S. (2008b). Strategies for measuring evolutionary conservation of RNA secondary structures. *BMC Bioinformatics*, **9**, 122.

16. Gruber, A. R., Findeiss, S., Washietl, S., Hofacker, I. L., and Stadler, P. F. (2010). RNAZ 2.0: Improved noncoding RNA detection. In *PSB10*, volume 15, pages 69–79.

17. Harmanci, A. O., Sharma, G., and Mathews, D. H. (2008). PARTS: probabilistic alignment for RNA joinT secondary structure prediction. *Nucleic Acids Research*, **36**(7), 2406–17.

18. Havgaard, J. H., Lyngso, R. B., Stormo, G. D., and Gorodkin, J. (2005). Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*, **21**(9), 1815–24.

19. Höchsmann, M., Töller, T., Giegerich, R., and Kurtz, S. (2003). Local similarity in RNA secondary structures. In *Proceedings of Computational Systems Bioinformatics (CSB 2003)*, volume 2, pages 159–168. IEEE Computer Society.

20. Hofacker, I. L. and Stadler, P. F. (2004). The partition function variant of sankoff's algorithm. In *Computational Science - ICCS 2004, Part IV*, Lecture Notes in Computer Science LNCS 3039, pages 728–735, Heidelberg. Springer Verlag.

21. Hofacker, I. L., Bernhart, S. H., and Stadler, P. F. (2004). Alignment of RNA base pairing probability matrices. *Bioinformatics*, **20**(14), 2222–7.

22. Klein, R. J. and Eddy, S. R. (2003). RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics*, **4**(1), 44.

23. Knudsen, B. and Hein, J. (2003). Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Research*, **31**(13), 3423–8.

24. Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–857.

25. Langenberger, D., Bermudez-Santana, C., Hertel, J., Hoffmann, S., Khaitovich, P., and Stadler, P. F. (2009). Evidence for human microRNA-offset RNAs in small RNA sequencing data. *Bioinformatics*, **25**(18), 2298–301.

26. Lau, N. C., Lim, L. P., Weinstein, E. G., and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in *caenorhabditis elegans*. *Science*, **294**, 858–862.

27. Lee, R. and Ambros, V. (2001). An extensive class of small RNAs in *Caenorhabditis elegans*. *Science*, **294**, 862–864.

28. Lee, Y. S., Shibata, Y., Malhotra, A., and Dutta, A. (2009). A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev*, **23**(22), 2639–49.

29. Lofquist, A. and Sharp, S. (1986). The 5'-flanking sequences of Drosophila melanogaster tRNA5Asn genes differentially arrest RNA polymerase III. *Journal of Biological Chemistry*, **261**(31), 14600–6.

30. Marz, M., Donath, A., Verstaete, N., Nguyen, V. T., Stadler, P. F., and Bensaude, O. (2009). Evolution of 7SK RNA and its protein partners in metazoa. *Mol. Biol. Evol.*, **26**, 2821–2830.

31. Mathews, D. H. and Turner, D. H. (2002). Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *Journal of Molecular Biology*, **317**(2), 191–203.

32. Mattick, J. S., Taft, R. J., and Faulkner, G. J. (2009). A global view of genomic information - moving beyond the gene and the master regulator. *Trends in Genetics*.

33. McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**(6-7), 1105–19.

34. Missal, K., Rose, D., and Stadler, P. F. (2005). Non-coding RNAs in Ciona intestinalis. *Bioinformatics*, **21 Suppl 2**, ii77–ii78.

35. Missal, K., Zhu, X., Rose, D., Deng, W., Skogerbo, G., Chen, R., and Stadler, P. F. (2006). Prediction of structured non-coding RNAs in the genomes of the nematodes Caenorhabditis elegans and Caenorhabditis briggsae. *J Exp Zoolog B Mol Dev Evol*, **306**(4), 379–92.

36. Morl, M. and Marchfelder, A. (2001). The final cut. The importance of tRNA 3'-processing. *EMBO Rep*, **2**(1), 17–20.

37. Pedersen, J. S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E. S., Kent, J., Miller, W., and Haussler, D. (2006). Identification and Classification of Conserved RNA Secondary Structures in the Human Genome. *PLoS Comput Biol*, **2**(4), e33.

38. Rivas, E. and Eddy, S. R. (2001). Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, **2**(1), 8.

39. Rose, D., Hackermuller, J., Washietl, S., Reiche, K., Hertel, J., Findeiss, S., Stadler, P. F., and Prohaska, S. J. (2007). Computational RNomics of drosophilids. *BMC Genomics*, **8**, 406.

40. Roshan, U. and Livesay, D. R. (2006). Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics*, **22**(22), 2715–21.

41. Sankoff, D. (1985). Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, **45**(5), 810–825.

42. Siebert, S. and Backofen, R. (2005). MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics*, **21**(16), 3352–9.

43. Smith, C. M. and Steitz, J. A. (1998). Classification of gas5 as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5'-terminal oligopyrimidine gene family reveals common features of snoRNA host genes. *Mol Cell Biol*, **18**(12), 6897–909.

44. The FANTOM Consortium (2005). The transcriptional landscape of the mammalian genome. *Science*, **309**(5740), 1559–63.

45. Torarinsson, E., Sawera, M., Havgaard, J. H., Fredholm, M., and Gorodkin, J. (2006). Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res*, **16**(7), 885–9.

46. Torarinsson, E., Havgaard, J. H., and Gorodkin, J. (2007). Multiple structural alignment and clustering of RNA sequences. *Bioinformatics*, **23**(8), 926–32.

47. Torarinsson, E., Yao, Z., Wiklund, E. D., Bramsen, J. B., Hansen, C., Kjems, J., Tommerup, N., Ruzzo, W. L., and Gorodkin, J. (2008). Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions. *Genome Res*, **18**(2), 242–51.

48. Uzilov, A. V., Keegan, J. M., and Mathews, D. H. (2006). Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. *BMC Bioinformatics*, **7**(1), 173.

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

50. Washietl, S., Hofacker, I. L., and Stadler, P. F. (2005a). Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA*, **102**(7), 2454–9.

51. Washietl, S., Hofacker, I. L., Lukasser, M., Huttenhofer, A., and Stadler, P. F. (2005b). Mapping of conserved RNA secondary structures predicts thousands of functional non-coding RNAs in the human genome. *Nat Biotechnol*, **23**(11), 1383–90.

52. Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F., and Backofen, R. (2007). Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. *PLOS Computational Biology*, **3**(4), e65.

53. Wilm, A., Mainz, I., and Steger, G. (2006). An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms Mol Biol*, **1**, 19.

54. Yao, Z., Weinberg, Z., and Ruzzo, W. L. (2006). CMfinder – a covariance model based RNA motif finding algorithm. *Bioinformatics*, **22**(4), 445–52.