*Bioinformatics*
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

# The locality dilemma of Sankoff-like RNA alignments

**Teresa Müller [1], Milad Miladi [1], Frank Hutter [2], Ivo Hofacker [3], Sebastian Will [3,4] and Rolf Backofen [1,5,*]**

[1]Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany,
[2]Machine Learning Lab, Department of Computer Science, University of Freiburg, Germany,
[3]Theoretical Biochemistry Group (TBI), Institute for Theoretical Chemistry, University of Vienna, Austria,
[4] Bioinformatics Group AMIBio, LIX - Laboratoire d'Informatique d'École Polytechnique, IPP, France and
[5]Signalling Research Centres BIOSS and CIBSS, University of Freiburg, Germany.

*To whom correspondence should be addressed.

## Abstract

**Motivation:** Elucidating the functions of non-coding RNAs by homology has been strongly limited due to fundamental computational and modeling issues. While existing simultaneous alignment and folding (SA&F) algorithms successfully align homologous RNAs with precisely known boundaries (global SA&F), the more pressing problem of identifying new classes of homologous RNAs in the genome (local SA&F) is intrinsically more difficult and much less understood. Typically, the length of local alignments is strongly overestimated and alignment boundaries are dramatically mispredicted. We hypothesize that local SA&F approaches are compromised this way due to a score bias, which is caused by the contribution of RNA structure similarity to their overall alignment score.

**Results:** In the light of this hypothesis, we study pairwise local SA&F for the first time systematically—based on a novel local RNA alignment benchmark set and quality measure. First, we vary the relative influence of structure similarity compared to sequence similarity. Putting more emphasis on the structure component leads to overestimating the length of local alignments. This clearly shows the bias of current scores and strongly hints at the structure component as its origin. Second, we study the interplay of several important scoring parameters by learning parameters for local and global SA&F. The divergence of these optimized parameter sets underlines the fundamental obstacles for local SA&F. Thirdly, by introducing a position-wise correction term in local SA&F, we constructively solve its principal issues.

**Contact:** backofen@informatik.uni-freiburg.de

## 1 Introduction

High-throughput sequencing experiments made blatantly evident that the majority of the eukaryotic genome is transcribed. Hundreds of thousands of non-coding RNAs (ncRNAs) and untranslated regions were reported. Revealing the subset of functional RNAs, and their specific functions, remains a major challenge. One important means to contribute to this taks is to identify classes of homologous genes or at least conserved domains. However, aligning homologous RNAs to determine conservation is a notoriously hard problem if the function of these RNAs is carried out by their emergent structure. Such 'structural' RNAs often do not show clear sequence conservation, but have the potential to fold into conserved homologous structures. For this reason, sequence alignment tools fail to align RNAs once the sequence identity is below 60% (Gardner *et al.*, 2005), which is the case for many homologous functional RNAs (Nawrocki and Eddy, 2013). This could explain, potentially many, cases of ncRNAs without any known homologues, not even in closely related species. Thus with some plausibility, the observed absence of homologs could often be a pure artifact, caused by insufficiencies of the commonly used tools.

To overcome such limitations, sequence-structure-based alignment methods like Dynalign, Foldalign and LocARNA (Mathews and Turner, 2002; Torarinsson *et al.*, 2007; Will *et al.*, 2007) have been used routinely to align non-coding RNAs. Most of these tools are derived in one way or the other from Sankoff's approach (Sankoff, 1985), which simultaneously aligns and folds sequence and structure. Sankoff-based alignment approaches can be characterized by a scoring system that adds a dominantly positive structure contribution to a sequence alignment. It has been repeatedly shown that they produce high quality alignments for RNAs that share a common conserved structure (Gardner *et al.*, 2005; Puton *et al.*, 2013). Moreover, they outperform pure sequence-based approaches, especially when aligning regions with lower sequence identity.

 

picture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picture



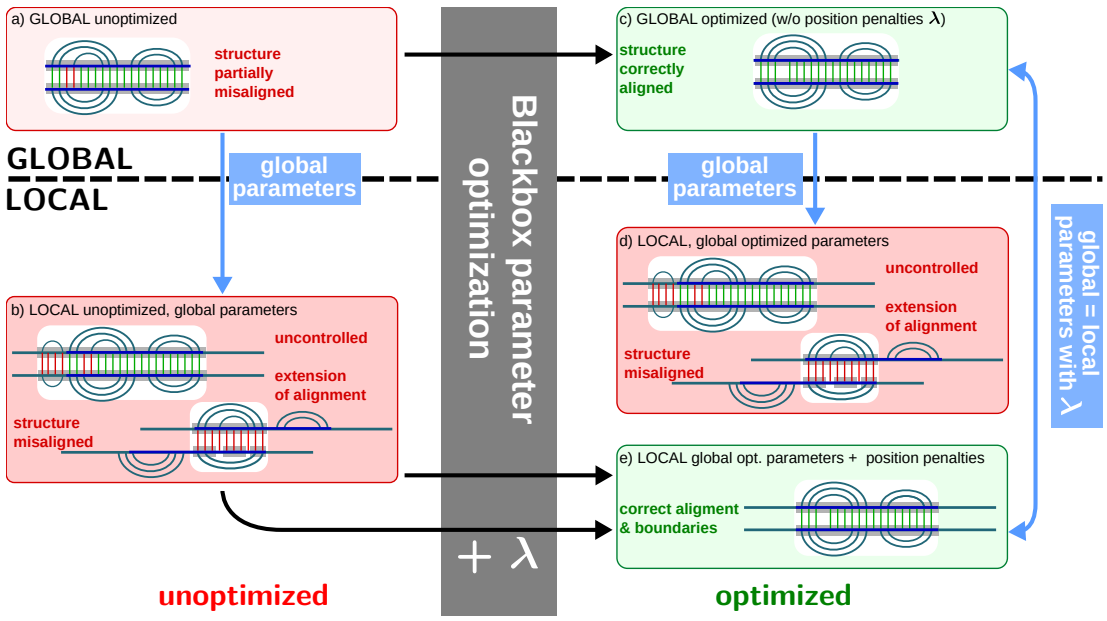


**Fig. 1.** A schematic overview of the trouble with locality for SA&F methods. The left side highlights the general issues of global and local alignments using default parameters. Without optimal parameters, global aligners (Box (a)) will partial misalign structure (red bars). When predicting local alignments (Box (b)), correctly aligning structured motifs gets even more difficult. Additionally, often local alignments are either extended over the motif's boundaries or random structures are aligned. These issues can be improved only partially by using optimized alignment parameters for the global and local predictions (right of the figure). Box (c) demonstrates that the optimized parameter set for the global alignment improves the alignment quality, whereas simply training parameters is insufficient to produce accurate local alignments (Box (d)). Only after introducing a position-wise penalty (Box (e)), global optimized parameters improve the accuracy of the local alignment predictions.

Furthermore, it has been shown that the existing methods perform well for "globally" aligning two or several RNAs that are *already known* to be homologous. However, the more important application of sequence-structure alignment is to reveal so far unknown ncRNA homologues. Since the exact genetic loci are generally unknown in this scenario *global* alignment is of little use. One must rather apply (some form of) *local* alignment. We emphasize that the existing benchmarks don't assess the performance of the existing tools in this second scenario. As well note that clustering-based ncRNA annotation approaches (Miladi *et al.*, 2019, 2017) implicitly detect ncRNAs homologues and could directly profit from improved local alignment.

There exists anecdotal evidence in the RNA community that the current pairwise sequence-structure alignment tools have difficulties to characterize homologous RNAs based on purely pairwise comparisons. For example, one study claimed that BLAST (Altschul *et al.*, 1990) still works better for this task than sequence-structure alignment (Menzel *et al.*, 2009). Another piece of evidence was provided by Heyer (2000), who showed that the Erdös-Rény law of a logarithmic growth of (local) alignment length (resp. score) for random sequences holds for many sequence analysis problems, except for sequence-structure alignment with a high structural scoring contribution. However, to the best of our knowledge, as of yet the relation of global Sankoff-like alignment to local Sankoff-like alignment has never been studied systematically. Consequently, it is insufficiently understood, how to ensure both correct alignments and a sub-linear growth with sequence length at the same time.

Here, we approach this open question utilizing the popular sequence-alignment tool LocARNA (Will *et al.*, 2007). LocARNA inherits the scoring system of PMcomp (Hofacker *et al.*, 2004), which strongly reduces the computational load compared to the Sankoff's original algorithm for simultaneous alignment and folding. Due to several algorithmic advancements, LocARNA moreover drastically reduces the run time over PMcomp (even reducing the theoretical complexity by a quadratic factor in sequence length over the original Sankoff algorithm). PMcomp's scoring system itself turned out to be a highly successful advancement in RNA alignment, it is implemented e.g. in LocARNA, Sparse (Will *et al.*, 2015), and FoldalignM (Torarinsson *et al.*, 2007). LocARNA's high performance and popularity makes it well suited to serve as representative of simultaneous folding and alignment algorithms in this work.

Recall that, as we explained before the existing benchmarks are insufficient to study the different aspects of the local alignment problem. Therefore, as a prerequisite we design a novel *local* alignment benchmark. On this basis, we examine the effect of modifications to the LocARNA algorithm on local pairwise alignment. First, we simply tweak the balance factor (structure weight) between the sequence and the structure similarity component of LocARNA's alignment evaluation ('alignment score'). Here we are interested in the effect on the quality of the local alignment, the correct detection of motif boundaries and the length of the reported alignments for random sequences.

In a second step, we investigated the interplay of different other parameters of LocARNA's score. Performing parameter optimization, we showed that the optimal scoring parameters for local alignment drastically differ from the optimal parameters for global alignment. This is the effect of optimization attempting to compensate for the local alignment detrimental effects of the structure contribution bias. In this way, the strongly deviating trained parameter sets indicate deficiencies of the scoring scheme. As well, note that using different parameter sets for local and global alignment is not suitable to resolve the existing problems with local RNA alignment. Using different schemes would moreover be highly inconvenient: tools used in post-processing would have to be adapted to the different scoring schemes (often enough it is unclear how to achieve this). A popular example for such a tool is RNAz (Gruber *et al.*, 2010) (resp. RNAcode (Washietl *et al.*, 2011)), which determines a conserved structure (resp. a conserved non-coding RNA) from calculated alignments.

picture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picture

Finally, we go beyond studying different parametrizations of LocARNA by introducing a new position-wise penalty. This additional score component is designed to counteract the negative properties of a high structure score and allows for using the same scoring parameters. A graphical summary of our investigations and its results is given in Fig. 1.

Our results underline that the original scoring system *rewards* matching structures on top of the sequence score without being able to properly balance the two components. This compromises the ability to properly align RNAs locally. We show that the proposed position-wise penalty can alleviate these problems, which strongly supports our hypothesis. While we performed our deep analysis for one specific PMcomp-like system, the results indicate issues of any system for scoring RNA alignments that is composed from a sequence and structure component. This includes Sankoff's original simultaneous folding and alignment algorithm.

## 2 MATERIALS AND METHODS

### Local alignment scores and growth of random alignments

The prerequisite for local alignment, as e.g. stated in the seminal work of Karlin & Altschul (Karlin and Altschul, 1990) about the statistical significance of local sequence alignment, is that the expected *global* score for random sequences is negative. Otherwise, "the maximal segment would tend to be the whole sequence" (see Karlin & Altschul (Karlin and Altschul, 1990), page 2265). Log-odds scores, which are commonly used for sequence alignments, automatically satisfy this property. The reason is simply that the expected scores for independent sequences is the negative Kullbach-Leibler (KL) divergence between distribution for alignment edges in the case of homologous sequences versus the distribution of alignment edges for independent sequences (Altschul, 1991). In more detail, the expected global score for independent sequences $E$ is given by

$$E_{\text{random}} = \sum_{a,b} q_a q_b \log(\frac{p_{ab}}{q_a q_b}) = -\sum_{a,b} q_a q_b \log(\frac{q_a q_b}{p_{ab}}) \qquad (1)$$

where $p_{ab}$ is the probability for seeing an alignment edge $a, b$ in homologous sequences, and $q_a$ (resp. $q_b$) is the background probability of $a$ (resp. $b$). This is equivalent to the negative KL divergence between the two distributions $(p_{ab})_{a,b}$ and $(q_a q_b)_{a,b}$. As the KL divergence is positive, the expected score should be negative. The side effect is that for homologous sequences, the expected global score is $E_{\text{evol}} = \sum_{a,b} p_{ab} \log(\frac{p_{ab}}{q_a q_b})$. $E_{\text{evol}}$ is the KL divergence between $(q_a q_b)_{a,b}$ and $(p_{ab})_{a,b}$ and therefore positive.

The relation between negative expected global score for random sequences and the length of maximal matched segments has not been assessed for the two-dimensional case of alignment, not to speak of sequence structure alignment. However, for the one-dimensional case of maximal segments in a single sequence, it was shown (Karlin *et al.*, 1990) that the local scores for a single sequence (i.e., where a segment is defined by consecutive hits in a single sequence) grows with the logarithm of the sequence length $n$ if the expected global score is negative, and with $\sqrt{n}$ if the expected score is 0.

To profit from established thermodynamic energy models, e.g. with empiricial or independently trained parameterization, common Sankoff-like sequence-structure scores consist of two components, namely a sequence alignment score and a structure contribution, which is positive for matched structures (see next section). In consequence, their score cannot be directly written as log-odds score as there is no background distribution for the structural which is used to weight this contribution. Heyer (2000) showed by experimental analysis that having $\sqrt{n}$ as a normalization factor, which would indicate a global expected score of 0, is probably too low for a scoring with a high structural contribution, and too high for a low structural contribution. This would indicate a logarithmic growth trend only for a low structure scoring. However, the effect on the actual alignments has not been investigated yet. Thus, we consider in the paper the actual global score for random sequences, the effect on the alignment quality as well as the effect on the detection of alignment boundaries (i.e., the correct detection of the actual motif).

### Details of the LocARNA alignment score

*Sankoff-style alignment scores.* Sankoff-style alignment algorithms optimize a score consisting of a sequence and structure score. The structure score depends on the structures of the RNAs, which are predicted simultaneously with the sequence alignment such that the combined score becomes optimal. In this way, Sankoff-style algorithms find and align the RNAs at the same time; the task is therefore called simultaneous alignment and folding (SA&F). In Sankoff's original algorithm, the score is composed of an edit distance and the energies of the predicted structures; for computing optimal alignments, the combined score is minimized (see Supplementary section 1). Since computing optimal local alignments typically requires similarity scores, which are able to distinguish similar subsequences from dissimilar ones, reformulations based on a similarity score have been suggested. For example, Foldalign introduced a similarity score based on sequence energies and negated energies; PMcomp suggested to compute the structure similarity from log odds of base pair probabilities.

*Alignment and scoring by LocARNA.* Following the idea of PMcomp, LocARNA scores structure based on precomputed base pair probability matrices. This allows LocARNA to compute alignments with a lot less overhead compared to the original Sankoff algorithm, while maintaining good accuracy. Moreover, by exploiting the sparsity of the structure space, LocARNA compute pairwise RNA alignment in only $O(n^4)$ time—compared to the $O(n^6)$ time complexity of the Sankoff algorithm.

To determine an optimal pairwise alignment, it optimizes an RNA alignment score by dynamic programming. This score is the sum of two components (Eqs. 2 and 4). The *sequence score component* evaluates the similarity at all sequence positions $A_s$ that are not involved in matching base pairs:

$$\sum_{(i,k)\in A_s} \sigma(i,k) - \gamma N_{gap} - \beta N^o_{gap} \qquad (2)$$

Here, $\sigma(i,k)$ denotes RIBOSUM base match similarity of $a_i$ and $b_k$, and $N^o_{gap}$ and $N_{gap}$ respectively count gap openings and gap extensions—each respectively penalized by $\beta$ or $\gamma$.

For the second component, LocARNA scores each base pairs $(i,j)$ of sequence $x$ in the predicted consensus structure, by

$$\Psi^x_{ij} = \log(p_{ij}/p_0)/\log(1/p_0), \qquad (3)$$

i.e. essentially by a log-odd score of the base pair's probability $p_{ij}$ against a background probability $p_0$ in $x$. The complete *structure score component* is a sum of contributions for each match $(ij;kl)$ of base pairs $(i,j)$ and $(k,l)$ in the consensus structure $S$:

$$\sum_{(ij;kl)\in S} \omega(\Psi^a_{ij} + \Psi^b_{kl}) + \tau\sigma'(i,j,k,l), \qquad (4)$$

where $\sigma'$ yields RIBOSUM base pair similarities. Observe how the parameter structure weight $\omega$ controls the contribution of the *structure score component* $\Psi^a_{ij} + \Psi^b_{kl}$ to the remaining *sequence score component* of the LocARNA score; the parameter $\tau$ , called *tau factor*, controls the contribution of sequence similarity at the ends of predicted base pairs; as rationale of this parameter, there are contradicting motivations to penalize or even favor mutations at the ends of base pairs—compensatory mutations

provide support for the conservation of the base pair, but discourage aligning its ends (see Supplementary section 1).

Our extension of a *position-wise penalty* is achieved by subtracting a term $\lambda$ for each scored position. Thus, the gap term $\gamma N_{gap}$ has to be replaced by $-(\gamma + \lambda)N_{gap}$. Each sequence match has now to be scored by $\sigma(i,k) - 2\lambda$. Finally, for each base pair match, a penalty of $-4\lambda$ has to be added, as four positions are scored (Supplementary formula 7).

*Sequence only score.* For computational efficiency, LocARNA computes the structure scoring part only if the base pair probability $p_{ij}$ of nucleotide $i$ and nucleotide $j$ is over a defined threshold $p$. By setting the threshold $p$ to its highest possible value (1), LocARNA's objective function will only use its sequence part and not align any base pair for the alignment prediction.

## Alignment quality measures

*Comparison to a reference alignment* . A standard approach to measure the alignment quality, given a reference alignment, is the *Sum-of-Pairs Score (SPS)* (Thompson *et al.*, 1999), which is defined as: SPS $= \frac{\text{\# correctly predicted columns}}{|\text{reference}|}$. A perfect prediction has SPS of one, whereas an SPS of zero indicates a completely mispredicted alignment. To extend this measurement for local alignment, we have to take both the length of the reference alignment as well as the length of the predicted motif into account. Thus, we define a new measurement *maxSPS* as follows: maxSPS $= \frac{\text{\# correctly predicted columns}}{\max(|\text{reference}|,|\text{prediction}|)}$ A upper boundary for the maxSPS value is the SPS of the same local alignment. Therefore a maxSPS value is always less than or equal to the correspond SPS value (see Supplementary section 2 and Fig. S1).

*Structure prediction quality.* The *Matthews correlation coefficient* (MCC) evaluates the quality of the predicted structures from simultaneously aligned and folded alignments (Gorodkin *et al.*, 2001)(see Supplementary section 2 for details).

*Quality of alignment boundaries.* The alignment quality can be measured by computing how many nucleotides of the local motif are predicted and how many nucleotides of the context are not part of the predicted alignment. This allows us to introduce *sensitivity* as measure how well the structured RNA motif is found and *specificity* assesses how well the alignment avoids extensions into the context. True positive TP or true negative TN values are all nucleotides that are part or not part of the predicted and the reference alignment. False positive FP values are all nucleotides that are part of the predicted but not of the reference alignment and false negative FN values are all nucleotides which are not part of the predicted but part of the reference alignment. Sensitivity and specificity are defined as usually given TP, FP, TN and FN (see Supplementary Fig. S2).

## Data

*Artificial data set.* A data set of random sequences to investigate the expected alignment score and alignment length of sequence structure algorithms was generated. We created an alignment input data set of 7000 FASTA files each having two randomly generated RNA sequences. The random sequences were generated by controlling the features alignment length (100 nt) and GC-content (average 50%). The *average pairwise sequence identity* (APSI) is computed using the ALISTAT tool from the HMMER package (Version 3.2.1) (Wheeler and Eddy, 2013). The APSI distribution is on average 40% (see Supplementary Fig. S7). The python function from *random* library was used to generate random RNA sequences. On average the data set has a uniform distribution of the four bases with an average GC-content of 50% (see Supplementary Fig. S6).

*BRAliBase.* The BRAliBase 2.1 is a well established alignment benchmark data set (Wilm *et al.*, 2006). It contains in total 18,990 ncRNAs alignments from 36 different Rfam families. For each alignment a raw file, containing the unaligned input sequences and a reference file, providing the corresponding mostly hand-curated reference alignments, are provided. Each file is annotated with the (APSI) and the *structure conservation index* (SCI) information. We computed pairwise alignments and therefore only used the pairwise subset k2 of the BRAliBase. In total k2 has 8976 entries from 36 different ncRNA families. For length distribution see Supplementary Fig. S8. The average SCI of BRAliBase is 0.93. All alignments below SCI 0.6 were excluded from the data set. Therefore BRAliBase fits well to our benchmark tasks.

The shuffled ncRNAs data set was produced by applying *FASTA-shuffle-letters* from the MEME suite (Bailey *et al.*, 2009) to all k2 sequences, to calculate the expected normalized scores.

*LocalBRAliBase.* A novel local alignment benchmark set is generated by placing all ncRNAs of the BRAliBase into its shuffled genomic context. The genomic context was derived from the European Nucleotide Archive (ENA) hosted by the European Molecular Biology Laboratory (EMBL) (Hussein *et al.*, 2018). Using the accession number, of each ncRNA, the according nucleotide sequence was downloaded in FASTA format from ENA. For every ncRNA, the start and end positions inside the downloaded nucleotide sequence are known. Using this position the ncRNA and its genomic context can be located. A fixed size (100 nt) of the genomic context was extracted equally up- and downstream of the ncRNA. In the case of limited context on one side, the missing nucleotides were extracted from the other side (context). If a full extraction of the context failed or the nucleotide sequence could not be found by its accession number the entry is excluded from the final LocalBRAliBase.

For the context elongation of 200 nt, 2,750 ncRNAs had not enough context available or could not be found by their accession number in ENA and therefore were excluded from the final data set. The flanking regions were dinucleotide shuffled using the tool uShuffle (Jiang *et al.*, 2008).

## Optimization setup

SMAC (Sequential Model-based Algorithm Configuration) is a black box optimization tool that identifies (sub-)optimal parameter combinations for configuring arbitrary algorithms (Hutter *et al.*, 2011) (see Supplementary section 3). The relationship between parameters and the desired algorithm result (alignment) is learned by optimizing a cost-function or quality score function. A python wrapper is handling the parameter settings and the input instances and applies them to our cost-function (scoring-function). The objective function is written in Perl and computes the LocARNA alignment for a given instance and parameter setting and the quality of this computed alignment.

The alignment quality measures used for the optimization of LocARNA parameters in the global mode is the geometric mean of SPS and MCC. The MCC is in most cases between 0 and 1, but it can also take a value down to -1. If the MCC was negative its value was set to zero. For the local alignment mode we used the maxSPS-value metric as described before. For details please refer to the Supplementary section 2.

The BRAliBase data set was used to infer the global alignment optimal parameter configuration and LocalBRAliBase, with a shuffled genomic context of length 200 nucleotides, was used for the local alignment mode.

Both data sets are filtered so that the number of instances per families are comparable amongst families and the data set is uniformly distributed, leaving the global data set with 2090 instances and the local data set with 1370 training instances. To identified the robustness of the parameter optimization and to exclude potential over fitting a 10-Fold cross validation was performed (see Supplementary Fig. S3).

The four LocARNA parameters gap extension ($\gamma$), gap opening ($\beta$), structure weight ($\omega$) and a tau factor ($\tau$) were subject to optimization within the relevant ranges (i.e., $\gamma : [-1000, 0], \beta : [-1500, 0], \omega : [0, 1000], \tau : [0, 100]$)

## 3 RESULTS AND DISCUSSION

We performed several experiments to elucidate our guiding question from different angles, namely whether there is a general difficulty of using Sankoff-like scores for the simultaneous *local* alignment and folding of RNAs. Since we are interested in shading light on the general phenomenon, we perform our experiments using the state-of-the-art RNA alignment tool LocARNA, as a representative of Sankoff-like approaches, or more specifically their PMcomp-like variety, which turned out particularly successful over the last decade.

Since we conjectured a potential positive structure scoring bias, which results from the non-negative structure score contribution, we start by directly quantifying the dimension of this bias in our first experiment. In our second experiment, we show the direct effect on the length of local alignments. While, as our first two studies show, overestimation of the alignment boundaries can be avoided by reducing the structure weight, we anticipated a conflicting negative effect on the alignment accuracy. The counterplay of these effects was investigated in a third study.

Furthermore, we were interested in optimal parameter settings for local and global alignment. Ideally, one could find a common optimal set of parameters for local and global alignment. As we discuss in more detail elsewhere, this has advantages in downstream analysis. Our final experiment attempts to directly rescue local alignment from the conjectured bias. A successful rescue could provide most direct evidence for the conjectured phenomenon; as well, such results could be directly useful to improve local RNA alignment.

### Assessing the quality of local alignment requires a novel, specifically tailored benchmark

In order to perform these studies, we designed a novel benchmark LocalBRAliBase and a local alignment quality measurement (maxSPS) for assessing local RNA alignment methods and their parametrization.

To best of our knowledge, there is not a comparably well established benchmark set for *local* alignments evaluation, like BRAliBase for *global* alignments. Therefore, we generated a local alignment benchmark set by taking the ncRNA's of the BRAliBase as local alignment motifs and embedded this motifs into their shuffled genomic contexts, also referenced as flanking regions. The flanking regions were shuffled, because the real genomic context of an ncRNA can be of high similarity. Specially for ncRNA inputs having lower sequence identity, it could be biologically meaningful to align the context. Hence we clear possible similarity of the context but not its nucleotide frequency.

To get insights on how sequences with lower sequence identity or high structure conservation behave differently in comparison to the overall set, we filtered the LocalBRAliBase. The set of sequences with lower sequence identity was constructed by filtering for sequences APSI smaller than 70 (APSI $< 70$). It is a more challenging task to find a local motif of low sequence identity in its genomic context. On the other hand for local motifs that are highly structured, accounting for structure should help to find more exact alignments. To have a set of sequences that are highly structured we filtered the LocalBRAliBase for sequences with a SCI over 100 (SCI $> 100$). The LocalBRAliBase construction resulted in 6226 entries for k2. The filtering of this k2 instances resulted in 3019 instances for APSI smaller 70 and 2165 instances for SCI greater 100.

The LocalBRAliBase data set was used to investigate boundary detection and the local alignment quality. However, to evaluate the quality of a local alignment it is still important to validate the correctly predicted alignment edges. The maxSPS is therefore a novel way of scoring local alignment by penalizing context extensions but also local alignments which are to short (methods). It extends the idea of the SPS measure by penalizing extensions of local alignments into the context (in contrast to the SPS score of the local alignment). Thus, the measure properly assesses the quality of local alignments, taking into account that good local alignments are similar to the reference in terms of boundaries *and* alignment columns. Note that the latter implicitly penalizes too short predictions. (see methods and Supplementary Fig. S1)

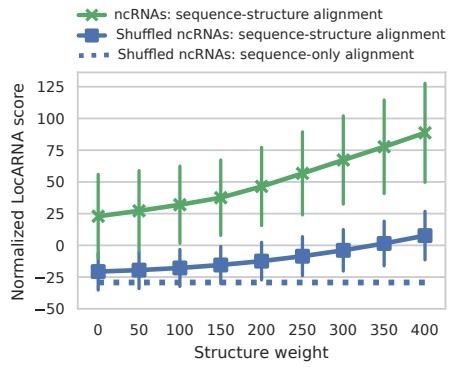


**Fig. 2.** The structure contribution leads to a positive scoring bias. The average normalized LocARNA scores for increasing structure weights (on the x-axis) for the three date sets: Sequence-structure alignments of ncRNAs from k2-BRAliBase (green, X), sequence-structure alignments of the shuffled ncRNA from k2–BRAliBase (blue, box), and sequence-only alignments of shuffled ncRNA from k2–BRAliBase (blue, dashed line). Sequence-only alignments were obtained by turning off structure matching in LocARNA's algorithm. The vertical bars represent the standard deviations. With increasing structure weights, the score value gap between sequence-structure and sequence-only alignments of the shuffled sequences increases; remarkably, even for the structure weight 0, sequence-structure alignments have a higher average score than the sequence-only alignments. Comparing alignments scores of true ncRNAs to the alignments scores of shuffled ncRNA shows that LocARNA is able to distinguish between shuffled and real ncRNAs.

### The expected background alignment score is shifted due to spurious structure contribution

A general bias on unrelated sequences, as it could be caused by the structure component of the LocARNA score, should be visible by a shift of expected LocARNA scores towards positive scores, if one puts more weight on the structure component. Note that in LocARNA, this works directly by increasing the parameter "structure weight". We quantified such effects based on the ncRNAs from BRAliBase. For each pair of ncRNAs, we shuffled both sequences using dinucleotide shuffling (see Methods). Then, for a series of different structure weighs (ranging from 0 to 400, in increments of 50), we aligned the pairs of ncRNAs as well as the pairs of shuffled ncRNAs with LocARNA. Finally, we compared the distribution of the scores, normalized to the sum of sequence lengths (Fig. 2). The normalized score reflects the average score contribution per nucleotide and allows comparing alignment scores from sequences of different lengths.

We observe that the normalized score achieved when aligning ncRNAs is always positive and increases with an increased structure weight. Moreover, for the shuffled ncRNAs, the score increases with increasing structure weight, although these RNAs cannot contain any conserved structure. This supports the idea of a general positive shift due to the structure component even for unrelated RNAs. We also compared the score to a sequence-only score, which is (by Equation $(1)$) expected to be negative. Here, we completely turn off structure alignment (i.e. switch to pure sequence alignment) by setting LocARNA's threshold for the base pair probability to 1; notably there is a subtle difference to setting the structure weight to zero. Only in the latter case, structure could still have an effect on the alignment: dissimilar nucleotides at the ends of two base pairs would rather be scored as structure match than two negative sequence matches, since the structure match has score zero (due to zero structure weight).

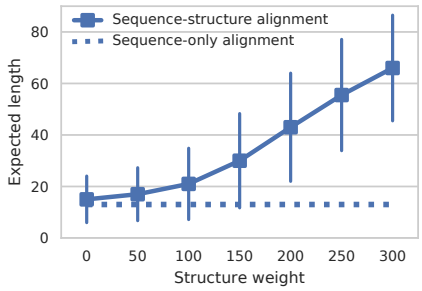


**Fig. 3.** The expected lengths of subsequences in local alignments strongly grow with increasing structure weight for artificially generated random sequence pairs (blue, box). In contrast, local sequence-only alignments produce much shorter alignments, without having any influence of the structure weight (dashed line). The expected lengths (y-axis) are estimated as average lengths of the predicted local LocARNA alignments over different structure weights (X-axis) using randomly generated RNA sequences of length 100 (see Methods). The sequence-only scoring predicts a local average alignment of length around 15 nts, regardless of the structure weight. A structure weight of 300 already leads to local sequence-structure alignments covering about 60% of the input sequences.

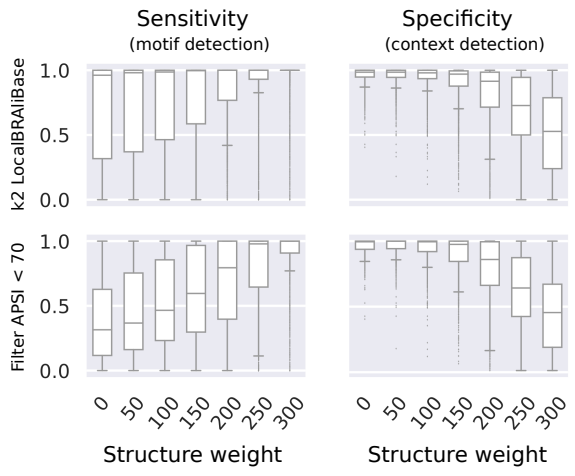


**Fig. 4.** Using a higher structure weight improves the local alignment quality but also leads to an extension of the alignment into the context. The local alignment performance metrics (sensitivity and specificity) are shown for the LocalBRAliBase data set (top) and the subset filtered for APSI less than 70% (bottom). There is a trade-off between covering the motif and restricting the alignment to the motif boundaries. A high structure weight helps to predict the complete local motif (top left), but it also leads to a extension into the context (top right). For a low structure weight the situation is vice versa. For sequences with lower APSI, the effect of not finding the complete or just parts of the motif is stronger (bottom left vs. top left). However, the ability to detect the context for a respective structure weight is less different comparing lower APSI sequences to the complete k2 LocalBRAliBase (right).

Note also that a part of the observed shift for the shuffled ncRNA sequences could be explained by their higher energy compared to the unshuffled ncRNAs, since higher energy is correlated with lower scores (see Supplementary Figs. S11 and S12), which contributes to this gap. To make sure that not any database biases lead to wrong conclusions we repeated the experiment with an artificial data set (see Supplementary Fig. S4). Having no bias like different GC-Content or different sequence length leads to a even slightly more positive scores.

For structure weight 0, we are more or less back to the sequence score, except that for base pair the alignment could use the Ribosum score (Klein and Eddy, 2003) instead of the sequence score for both ends of the base pair. Nevertheless, even the Ribosum score is negative in expectation since it is a log odds score as well. While this positive contribution fact of the structure score was already clear, no one until now investigated the relative contribution of structure and sequence part of the score.

The main problem here is that one cannot make a simple theoretical analysis as given for sequence alignment by Equation (1) as the structure score in LocARNA (and many others) depends on base probabilities which use information stemming from the whole sequence. This way, the different alignment edges are not independent anymore.

## The local alignment length of random sequences increases with growing structure weight

The previous experiment (Fig. 2) shows that the score clearly distinguishes positive data (pairs of ncRNAs) from negative data (pairs for shuffled ncRNAs); so far, this is in line with previous results that show the good quality of the LocARNA score for global alignment.

Still, this leaves the questions, whether the observed bias has effects on the quality of local alignments and how strong such effects are. On the one hand, the required negative expected scores, to avoid linear growth of local alignment with its input sequence lengths, seems to be met for commonly used structure weights, e.g. the default weight 200. On the other hand, this does not rule out detrimental overestimation of boundaries due to a score bias.

To quantify this, we first determined the expected length of local alignments between random RNA pairs; this was determined for LocARNA in default setting as well as for sequence only setting. This length indicates the expected overestimation of real ncRNA alignments, even if their genomic context is completely unrelated. At the current default setting (structure weight of 200) almost half of the genomic context could be added to the aligned local motif.

When we destroy sequence and structure homology in the instances of LocalBRAliBase by shuffling (methods), we still observe longer alignments for increasing structure weight (Supplementary Fig. S5). In this benchmark, the input sequences have relatively large and diverse lengths of typically about 250-300 nt composed of 200 nt context and the ncRNA sequence. Thus, instead of directly comparing sequences from the shuffled benchmark, we generate an artificial data set of RNAs of length 100, having the same nucleotide distribution (Methods). As can be seen in Fig. 3, sequence-only score does produce (on average) an alignment of length 10 or roughly 10% of the input. Given a typical length of 50-100 (see length distribution of BRAliBase ncRNAs in Fig. S8) for most of the ncRNAs, we would expect that the predicted boundaries are extended by 10-20 nts over the real boundaries. For high structure weight, however, more than 60% of the random sequences are covered by a local alignment, which implies that we do not have any chance to detect the real boundaries of ncRNAs even in shuffled genomic context.

## Emphasizing structure improves alignment accuracy but seriously overestimates boundaries of local alignments

Sequence-structure alignment tools were introduced to overcome the limitations of sequence-based approaches in the detection of conserved structures (Gardner *et al.*, 2005). Thus, a high structure weight would be desirable for a correct detection of conserved structures. However, the length of local alignment of random sequences already suggest that we usually cannot find the true transcript boundaries using a high structure weight as it would likely extend a real ncRNA into the genomic context. While sequence alignment does not share these problems, resorting to sequence-only alignment (or setting LocARNA's structure weight very low; see Fig. 4) does not resolve the dilemma. Not picking up structure similarity can lead to bad alignment quality or even a complete miss of the local alignment, specially for sequences with low sequence identity (see Fig. 4 bottom).

picture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picture



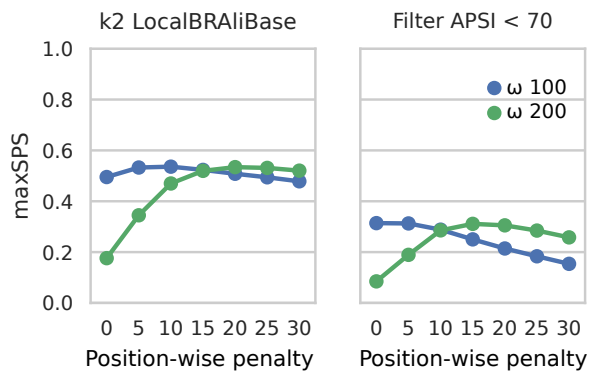


**Fig. 5.** Position-wise penalties can improve the average alignment quality of local alignments (as measured by maxSPS). For the entire k2 LocalBRAliBase (left) and a subset of low sequence similarity (APSI $< 70$, right), we plot the average maxSPS for different position-wise penalties $\lambda$ (x-axis). We show our results for two different structure weights, the default value of $\omega = 200$ and $\omega = 100$, which are close to our optimization results for global and local alignments. The plots demonstrate that the position-wise penalty can compensate for the bias of using a higher structure weight. A position-wise penalty of about 15 allows using the high structure weight of 200, an optimal value for the global alignment. Hence, the penalty allows applying a single set of optimal parameters for global and local alignment—as we argue in the text, this is beneficial on its own.

To investigate this further, we determined the agreement of the positions covered by the local alignment with the actual transcript boundaries using the LocalBRAliBase. Note that we do not investigate here how good the actual alignment is. Finding the correct transcript boundaries, however, is the necessary step to get a good alignment as we could re-align the detected transcripts using global alignment in a following step. For the purpose of measuring the agreement of transcript boundaries and regions detected by local alignment, we investigated the sensitivity and specificity of the predicted alignment areas. The sensitivity measures the fraction of positions that are correctly predicted according to a reference alignment (with optimal value 1), whereas the specificity measures how much of the context is not aligned (with optimal value 1). A example can be found in Fig. S2. Taking sensitivity and specificity into account will reveal alignments which are extended into the context, but also too short alignments and is therefore a helpful boundary detection measurement.

As Fig. 4 shows, a high structure weight does detect the full transcript (sensitivity close to 1) but tend to align into the context (indicated by a lower specificity). The initial arguments of this section, that a structure weight is needed to detect the ncRNA is clearly proved by looking at the sequences with lower sequence identity. With structure weight 100 on average less than half of the ncRNA is found. Sequence-based scoring or not adding a positive structure contribution, on the other hand, does fail to detect the correct transcript. See structure weight 0.

## Global and local alignment require fundamentally different parameter sets

In the previous experiment, we varied only the single parameter weighting the structure. While we see that this does not allow to simultaneously meet both objectives, accurate alignment and accurate boundaries in local alignment, we cannot rule out yet that such reconciliation is possible due to the complex interplay of the structure weight with the other essential parameters, which control the affine gap cost (gap opening and extension) and the importance of sequence similarity at structural matches (tau factor).

For directly studying such potential effects, we apply machine learning techniques as a tool to illuminate this apparent contradiction from the perspective of optimal parameter settings. This allows us to simultaneously optimize these parameters for there suitability to reproduce reference alignments from a benchmark set. In this way, we determined independent optimal parameter sets for global alignment as well as for local alignment (Methods).

While we learn parameters for global alignment directly form the k2 data set of BRAliBase, we employed the derived local benchmark set from LocalBRAliBase for optimizing the parameters for local alignment (Methods). The optimized parameter sets are shown in Table 1 together with LocARNA's current default values for comparison. Interestingly, already our optimization results for global alignment on the k2 data set differ quite substantially from LocARNA's default values, since such parameters were never systematically optimized before. While the structure weight was apparently chosen well for aligning k2 sequences, the optimal gap opening cost is considerable higher, but compensated by much lower gap extension costs. We assume that such shift of gap extension costs to opening costs is also beneficial for aligning most other RNAs with LocARNA. Moreover, on k2 it turned out advantageous to more strongly consider sequence similarity at the ends of matched base pairs (higher tau factor), where originally we had set the influence of sequence to zero to avoid penalizing compensatory mutations.

More importantly for our study, comparing the optimized parameters for local and global alignments, we see most striking differences. Apparently, even allowing variations in the relevant other parameters, a common ideal parameter set that works equally well for local and global alignment cannot be found.

As discussed before a high structure contribution in local alignments leads to alignment extensions into the context and will produce a high number of wrong alignment edges. The structure weight is nearly halved, as is the tau-factor. As the tau-factor is another part of the structure scoring, it is conceivable the tau-factor and structure weight are just correlated.

Furthermore, the gap costs are significantly increased. Higher gap costs help to avoid aligning similar regions that are spread of a longer region and therefore most likely do not belong to the same local motif. Therefore, we hypothesize that the optimized parameters for local alignment limit the bias due to the structure contribution as to avoid strong overestimation of boundaries. Note that there is no corresponding pressure for global alignment. It seems that almost the only resolution is to lower structure weight (and tau factor). At the same time, as the optimization for global alignment (and the results from our previous experiment) indicate, this tends to lower the alignment accuracy of the locally aligned subsequences.

## Position-wise penalty as an alternative to reduced structure weight

Previous results suggest that even the optimized parameters for local alignment are forced into an unfortunate compromise between sensitivity and specificity that cannot be satisfyingly resolved. Moreover, the large deviation between optimized parameters for the local and global alignments do cause additional problems for tools like RNAz (Gruber *et al.*, 2010) that rely on statistics over local and global alignments, since for such tools, different parameterization lead to different statistics distributions.

In this work, we are first of all interested to systematically study the potential issues of using Sankoff-like scoring in local alignment. However, we suggest a simple modification of the score, which has the potential to compensate such issues. This can yield final direct evidence, as well as suggests more sophisticated solutions to the problem.

We start with hypothesizing that, on average, there is a length-dependent shift of the score due to the structure contribution. Already Heyer (2000) observed that local alignment cannot work properly, if the expected alignment score grows linearly with alignment length. While the expected length of proper local alignments should grow at most logarithmically with the input lengths, such a dependency would cause

picture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picture

Table 1. Optimized parameters differ strongly for global and local LocARNA alignments. However, after introducing a position-wise penalty in LocARNA, optimal parameters for local alignment (with penalty) turn out to be very close to the trained parameters for global alignment. Compared to LocARNA's default values, the paramter optimization makes gap extensions cheaper and gap opening more expensive. To improve local alignments, the structure-weight is strongly decreased compared to the optimal weight for global alignment (which is nearly the same as the default weight). In the same way, the trained tau-factor for local alignment is smaller than for global alignment. The other parameters are changed by the training for either local or global alignments in the same way (compared to default parameters). Using a position-wise penalty $\lambda$ of 15 for the local alignment prediction, the optimized parameters are almost equal to the optimized parameters for global alignments. All parameter training was performed using the black box optimizer SMAC (see Methods). Example LocARNA calls are given in the supplement.

| Parameter set | Gap extension ($\gamma$) | Gap open ($\beta$) | Structure weight ($\omega$) | Tau factor ($\tau$) |
|---|---|---|---|---|
| Default values | 350 | 500 | 200 | 0 |
| Global optimized | 68 | 807 | 210 | 72 |
| Local optimized with no penalty ($\lambda$= 0) | 136 | 975 | 115 | 38 |
| Local optimized with penalty ($\lambda$= 15) | 82 | 883 | 176 | 71 |

linear growth. This motivates us to subtract a position-wise penalty from the score; in order to compensate the hypothesized linear dependency of score's on the input length due to the structure contribution. Concretely, we penalize each position in the locally aligned subsequences; to require only a single additional parameter, all positions are penalized equally.

We first tested, using a simple grid search for the penalty, the effect of penalty on the alignment quality for structure weights 100 and 200. We have chosen these values as they were roughly the result of the parameter optimization for local and global alignment, respectively.

Our results in Fig. 6 support the expected effect of the penalty on the sensitivity and specificity, namely increasing specificity on the cost of sensitivity. However, as seen in Fig. 5, in both cases the penalty is required to improve the overall alignment quality. The optimal penalty for structure weight 100 is 5-10, for 200 it is 15-20. Similar optimal and near-optimal performance was obtained for penalty 15 in our family-wise analysis of the six largest RNA families with structure weight 200 that is provided in Supplementary Fig. S13. To get further insides on how the penalty is acting we investigated the F1 distribution for the complete k2 LocalBRAliBase and two filtered sets of the LocalBRAliBase, see Supplementary Fig. S9. This shows that overall the penalty allows to improve the alignment quality, even more for higher structure weights (200). Notice that, if the data set is filtered for a specific subset, e.g. sequences with APSI below 70 or with a SCI above 100 the optimal position-wise penalty value is shifting. For sequences with a lower sequence identity a structure weight around 5 seem to be optimal. However, looking at very structured sequences a penalty of 15 gives again an optimal result. Therefore we came to the conclusion that a penalty around 15 should be desirable, which is than also used for the second local parameter optimization (see Table 1 line 4).

This allows the use of identical parameter set for global and local alignment, which is particularly beneficial for tools that rely on statistics over global and local alignments (like RNAz; as discussed before).

Finally comparing the current default LocARNA parameter setting and our now improved, combination of global optimized parameters and the position-wise penalty, shows that we can drastically improve the local alignment prediction of LocARNA. Comparing the medians of optimized and default parameter settings (Fig. 7), it is notable that using the default parameter setting for more than half of the sequences the alignment prediction failed completely. This is substantially improved when using the optimized parameter settings with penalty 15. If the alignment boundaries would be perfectly predicted we could achieve a median of 0.95 (Supplementary Fig. S10), however the current improvement is already astonishing.

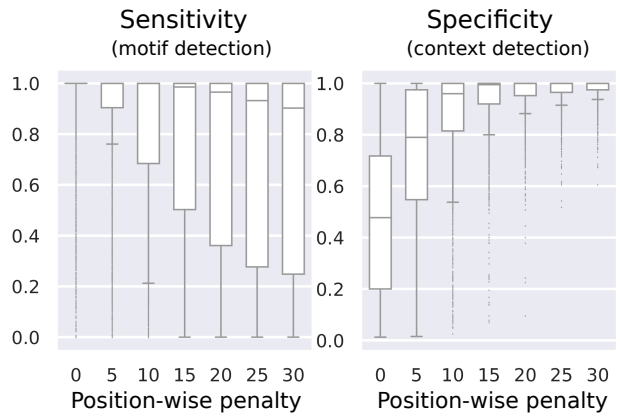


**Fig. 6.** The compensating effect of position-wise penalties. The boundary detection (sensitivity + specificity) is shown for different position-wise penalties using global optimized parameters and structure weight 200. For higher penalties almost nothing of the context will be aligned (right, specificity). On the other hand a high penalty leads partly to not finding the complete local motif (left, sensitivity). This results can be matched to the trade-off issue of finding the correct alignment boundaries, shown in Fig. 4. Using high structure weight for alignment predictions will result in finding the complete motif but it also leads to an extension into the context. Now, for a high structure weight (200), already at penalty 15 most local motifs are still fully covered, but most of the predictions are not extended into the context. This shows how the bias due to the structure contribution can be compensated by using a position-wise penalty.

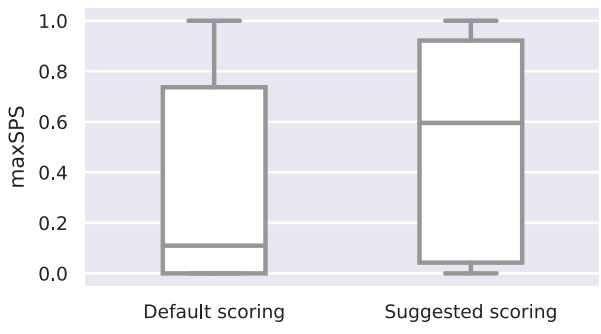


**Fig. 7.** Improvement in local alignment quality by using the suggested parameter setting over the LocARNA default parameter settings. The suggested parameter set is the optimized parameters for global alignment (Table 1) combined with a position-wise penalty of 15. The structure weight has been rounded to 200. We observe a considerable improvement, the median maxSPS is increased from 0.11 to 0.60. A example comparing the default and suggested scoring can be found in Supplementary Fig. S14.

## 4 CONCLUSION

In this work we systematically studied the comparison of non-coding RNAs by local simultaneous alignment and folding (local SA&F), which is highly demanded for the important challenge of identifying homologues from the plethora of identified RNA transcripts with unknown function. Over the years, several approaches for local SA&F have been proposed and there

have been attempts to improve local alignment prediction and to counteract alignment elongation beyond motif boundaries. A prominent example is Foldalign (Havgaard *et al.*, 2005), which limits the maximum length of motifs. However, the fundamental issues of local SA&F as discussed in this work were neither investigated in depth before nor quantified systematically.

In the first step, we designed a new local alignment benchmark set (LocalBRAliBase) together with an appropriate novel local alignment quality measure (maxSPS). This benchmark set is essential for studying local alignment and constitutes a valuable contribution by itself, since it enables proper assessment of local RNA alignment method performance for the first time.

With this benchmark set at hand, we could empirically confirm our initial hypothesis that the structure prediction component in SA&F introduces a bias to the total similarity score. The intuitive explanation for this bias is that the structure contribution of SA&F is purely positive by definition. The bias was shown to be sufficiently high in current tools to compromise the prediction of correct alignment boundaries. Even worse, we showed that current scoring schemes do not allow to properly balance the overestimation of alignment length (i.e., misprediction of boundaries) against the sensitivity for detecting structural homology. By varying only the relative weight of the structure component in the scoring, we demonstrated this effect clearly: sufficient emphasis on structure is required to identify local motifs, however this compromises the accuracy of boundary prediction at the same time. As wrong parameterization of the scoring could still be the source of the problem, we subsequently used machine learning to optimize the alignment parameters, finding that the divergence between accuracy of boundary and structure detection cannot be resolved by re-parametrization. Surprisingly, we found that a position-wise penalty could completely resolve the problem, yielding the same set of parameters for global and local SA&F alignment. This has the further benefit that downstream tools like RNAz for predicting ncRNAs do not have to differentiate between local and global scoring.

In summary, we clearly showed that the current SA&F scoring schemes are not directly applicable to local SA&F. Moreover, we identified the positive structure contribution to the alignment score as the major bias source for overestimating the alignment boundaries. Finally, we presented a constructive solution of the observed issues through applying a position-wise penalty. By raising awareness in the community, precise identification of the issues, and pointing out viable solutions, this work constitutes an important step towards reliable local SA&F approaches in future studies.

## Availability

The benchmark data, detailed results and scripts are available at https://github.com/BackofenLab/local_alignment. The RNA alignment tool LocARNA, including the modifications proposed in this work, is available at https://github.com/s-will/LocARNA/releases/tag/v2.0.0RC6.

## Acknowledgements

We thank Christina Otto for the fruitful discussion and ideas in the initial phase of this work.

## Funding

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC-2189 - [Project id 390939984] and projects [DFG grants BA 2168/14-1, BA 2168/3-3] and the German Federal Ministry of Education and Research (BMBF) [BMBF grant 031A538A de.NBI].

## References

Altschul, S. F. (1991). Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol*, **219**(3), 555–65.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, **215**(3), 403–10.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). Meme suite: tools for motif discovery and searching. *Nucleic acids research*, **37**(suppl_2), W202–W208.

Gardner, P. P., Wilm, A., and Washietl, S. (2005). A benchmark of multiple sequence alignment programs upon structural rnas. *Nucleic acids research*, **33**(8), 2433–2439.

Gorodkin, J., Stricklin, S. L., and Stormo, G. D. (2001). Discovering common stem–loop motifs in unaligned rna sequences. *Nucleic Acids Research*, **29**(10), 2135–2144.

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.

Havgaard, J. H., Lyngsø, R. B., and Gorodkin, J. (2005). The foldalign web server for pairwise structural rna alignment and mutual motif search. *Nucleic acids research*, **33**(suppl_2), W650–W653.

Heyer, L. J. (2000). A generalized Erdös-Rény law for sequence analysis problems. *Methodology and Computing in Applied Probability*, **2**(3), 309–329.

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

Hussein, A., Cerdeño Tárraga, A., Toribio, A.-L., Milano, A., Alako, B., Amid, C., Smirnov, D., Richards, E., Cochrane, G., Cleland, I., Rajan, J., Martínez-Villacorta, J., Reddy, K., Rosello, M., Silvester, N., Pakseresht, N., Leinonen, R., Holt, S., Vijayaraja, S., Kay, S., Jayathilaka, S., Keane, T., Liu, X., and Harrison, P. W. (2018). The European Nucleotide Archive in 2018. *Nucleic Acids Research*, **47**(D1), D84–D88.

Hutter, F., Hoos, H. H., and Leyton-Brown, K. (2011). Sequential model-based optimization for general algorithm configuration. In *International Conference on Learning and Intelligent Optimization*, pages 507–523. Springer.

Jiang, M., Anderson, J., Gillespie, J., and Mayne, M. (2008). ushuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC bioinformatics*, **9**(1), 192.

Karlin, S. and Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA*, **87**(6), 2264–8.

Karlin, S., Dembo, A., and Kawabata, T. (1990). Statistical composition of high-scoring segments from molecular sequences. *The Annals of Statistics*, **18**.

Klein, R. J. and Eddy, S. R. (2003). Rsearch: finding homologs of single structured rna sequences. *BMC bioinformatics*, **4**(1), 44.

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

Menzel, P., Gorodkin, J., and Stadler, P. F. (2009). The tedious task of finding homologous noncoding RNA genes. *RNA*, **15**(12), 2075–82.

Miladi, M., Junge, A., Costa, F., Seemann, S. E., Havgaard, J. H., Gorodkin, J., and Backofen, R. (2017). RNAscClust: clustering rna sequences using structure conservation and graph based motifs. *Bioinformatics*, **33**(14), 2089–2096.

Miladi, M., Sokhoyan, E., Houwaart, T., Heyne, S., Costa, F., Grüning, B., and Backofen, R. (2019). Graphclust2: Annotation and discovery of structured rnas with scalable and accessible integrative clustering. *GigaScience*, **8**(12). giz150.

Nawrocki, E. P. and Eddy, S. R. (2013). Computational identification of functional rna homologs in metagenomic data. *RNA biology*, **10**(7), 1170–1179.

Puton, T., Kozlowski, L. P., Rother, K. M., and Bujnicki, J. M. (2013). CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction. *Nucleic Acids Res*.

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

Thompson, J. D., Plewniak, F., and Poch, O. (1999). A comprehensive comparison of multiple sequence alignment programs. *Nucleic acids research*, **27**(13), 2682–2690.

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

Washietl, S., Findeiß, S., Müller, S. A., Kalkhof, S., Von Bergen, M., Hofacker, I. L., Stadler, P. F., and Goldman, N. (2011). Rnacode: robust discrimination of coding and noncoding regions in comparative sequence data. *Rna*, **17**(4), 578–594.

Wheeler, T. J. and Eddy, S. R. (2013). nhmmer: Dna homology search with profile hmms. *Bioinformatics*, **29**(19), 2487–2489.

Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F., and Backofen, R. (2007). Inferring noncoding rna families and classes by means of genome-scale structure-based clustering. *PLoS computational biology*, **3**(4), e65.

Will, S., Otto, C., Miladi, M., Möhl, M., and Backofen, R. (2015). SPARSE: quadratic time simultaneous alignment and folding of RNAs without sequence-based heuristics. *Bioinformatics*, **31**(15), 2489–2496.

Wilm, A., Mainz, I., and Steger, G. (2006). An enhanced rna alignment benchmark for sequence alignment programs. *Algorithms for molecular biology*, **1**(1), 19.