Structure Local Multiple Alignment of RNA

Wolfgang Otto*[1] and Sebastian Will*[2] and Rolf Backofen[2]

[1] Bioinformatics, University Leipzig, D-04107 Leipzig
wolfgang@bioinf.uni-leipzig.de

[2] Bioinformatics, Albert-Ludwigs-University Freiburg, D-79110 Freiburg
{*will,backofen*}*@informatik.uni-freiburg.de*

Abstract: Today, RNA is well known to perform important regulatory and catalytic function due to its distinguished structure. Consequently, state-of-the-art RNA multiple alignment algorithms consider structure as well as sequence information. However, existing tools neglect the important aspect of locality. Notably, locality in RNA occurs as similarity of subsequences as well as similarity of only substructures. We present a novel approach for multiple alignment of RNAs that deals with both kinds of locality. The approach extends LocARNA by structural locality for computing all-against-all pairwise, structural local alignments. The final construction of the multiple alignments from the pairwise ones is delegated to T-Coffee. The paper systematically investigates structural locality in known RNA families. Benchmarking multiple alignment tools on structural local families shows the need for algorithmic support of this locality. The improvement in accuracy in special cases is achieved while staying competitive with state-of-the-art alignment tools across the whole Bralibase. LocARNA and its T-Coffee extended variant LocARNATE are freely available at `http://www.bioinf.uni-freiburg.de/Software/LocARNA/`.

1 Introduction

The recent discovery of the ubiquity and vast importance of regulatory and catalytic RNA in biological systems has radically changed our view on RNA [Cou02, Bar04, FW05]. This motivated a series of algorithmic developments in the area of multiple RNA alignment. RNA comparisons are challenging since both structure and sequence information have to be taken into account in order to successfully align RNAs with low sequence identities; pure sequence alignment is failing below of about 60% sequence identity. Spearheading this development are tools based on simultaneous alignment and folding like FoldAlignM [THG07], LARA [BKR07], and LocARNA [WRH$^+$07]. However, these approaches neglect an important aspect of locality.

For RNA, one distinguishes two kinds of locality. First, similarity of RNAs can occur restricted to only corresponding subsequences; this form of locality is well known for sequence alignment. Even this locality is rarely supported by multiple alignment algorithms,

*Both authors contributed equally.

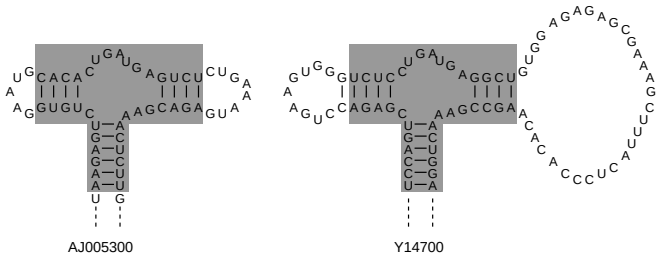


Figure 1: Two similar local substructures. Both hammerhead ribozymes $AJ005300$ and $Y14700$ differ globally. Nevertheless, they share a common functional motif (highlighted), which is structural local.

which thereby assume that the input sequences are accurately excised from their genomic context.

This assumption however does not suffice in face of the second kind of locality. Namely, RNA shows structural locality in the case where only substructures of several RNAs are similar, cf. Figure 1. Such corresponding local substructures can consist of several subsequences that are unconnected to each other at the sequence level. Then, these subsequences are connected only via the structure of the RNAs. An analogous view is that a local substructure consists of a subsequence, where certain subsequences are excluded (therefore called exclusions in the paper). For the simpler case of RNA alignment with fixed input structures, the algorithmic challenge posed by this kind of locality is solved in $O(n^5)$ [BW04].

Contribution In the paper, we show that structural locality plays an important role for RNA similarity and occurs in a number of known RNA families. To our knowledge this feature is for the first time analyzed across the Rfam, a database of known RNA family alignments[GJMM$^+$05].

Responding to this observation, we present the tool LocARNATE, which handles structural and sequence locality in the computation of multiple alignments of RNAs. To our knowledge it is the first multiple alignment approach that supports structural locality of RNAs. The paper describes the extension of the pairwise alignment algorithm of LocARNA[WRH$^+$07] by structural locality without increasing its theoretical complexity. This serves as a basis for the construction of multiple alignments, which is done here using T-Coffee[NHH00]. T-Coffee is chosen since it can do a consistency extension of the information from pairwise LocARNA alignments. Compared to a purely pogressive alignment strategy it is thereby able to avoid many of the typical mistakes. At the same time it respects the high-quality pairwise relation of the sequences derived by LocARNA.

Our theoretical results are supported 1.) by benchmarks using selected RNA sequences from the Rfam that show distinguished structural locality as well as 2.) by non-biased Bralibase 2.1 benchmarks. The Bralibase 2.1 is a compilation of true, hand-curated alignments for the purpose of assessing the accuracy of RNA alignment tools. [WMS06]

2 Preliminaries

An *(RNA) sequence* S is a word of $\Sigma = \{A, C, G, U\}$. We denote by A_i the ith symbol in A, by $A_{i..j}$ the subsequence from position i to j, and by $|A|$ the length of A. An *(RNA) structure* P *for* S is a set of *base pairs* (or *arcs*) $(i, j) \in \{1 \dots n\} \times \{1 \dots n\}$, $i < j$. A structure P is called *crossing* iff $\exists (i, i'), (j, j') \in P : i < j < i' < j'$. Otherwise it is called *non-crossing* or *nested*. In the paper, we assume that RNA structures are non-crossing. We define a partial ordering $\prec$ on pairs of natural numbers by $(i, i') \prec (j, j')$ iff $j < i < i' < j'$. Obviously, $\prec$ orders the base pairs of a structure P according to their nesting.

A *pairwise alignment* $\mathcal{A}$ *of two sequences* A *and* B is a subset of $[1..|A|] \cup \{-\} \times [1..|B|] \cup \{-\}$, where for all pairs $(i, j), (i', j') \in \mathcal{A}$ holds 1.) $i \leq i' \Rightarrow j \leq j'$ 2.) $i = i' \neq - \Rightarrow j = j'$, and 3.) $j = j' \neq - \Rightarrow i = i'$. We define the projections $\pi_1\mathcal{A} = \{i \neq - \mid \exists j : (i, j) \in \mathcal{A}\}$ and $\pi_2\mathcal{A} = \{j \neq - \mid \exists i : (i, j) \in \mathcal{A}\}$. An alignment $\mathcal{A}$ of A and B is called *global*, iff $\pi_1(\mathcal{A}) = [1..|A|]$ and $\pi_2\mathcal{A} = [1..|B|]$. A *sequence local motif of a sequence* A is a range $[i..j]$ for some $1 \leq i, j \leq |A|$. An alignment $\mathcal{A}$ of A and B is called *sequence local* iff $\pi_1\mathcal{A}$ is a sequence local motif for A and $\pi_2\mathcal{A}$ is a sequence local motif for B.

A *consensus structure* P *for an alignment* $\mathcal{A}$ *of* A *and* B is a pair (P_A, P_B) of a structure P_A for A and a structure P_B for B, such that 1.) for all $(i, j), (i', j') \in \mathcal{A}$ holds $(i, i') \in P_A$ iff $(j, j') \in P_B$, 2.) P_A contains only positions in $\pi_1\mathcal{A}$, and 3.) P_B contains only positions in $\pi_2\mathcal{A}$.

3 Locality

Structural Locality in Pairwise Alignments We distinguish sequence and structural locality. Adopting a graph theoretic view, sequence local motifs of a sequence A are sets of connected vertices in a graph $G_{\text{seq}} = (V, E)$, where $V = [1..|A|]$ and $E = \{(i, i+1) | 1 \leq i < |A|\}$. For a structure P of A, we define a *structural local motif for* A *and* P as a set of connected vertices in the *structure graph* $G_{struct} = (V, E \cup P)$ *of* A *and* P. By this definition, structural local motifs correspond to "substructures", where the connection of bases can be either due to the backbone or due to bonds between base pairs.

An alignment $\mathcal{A}$ of two RNA sequences A and B is *structural local for consensus structure* (P_A, P_B) iff $\pi_1\mathcal{A}$ is a structural local motif for A and P_A as well as $\pi_2\mathcal{A}$ is a structural local motif for B and P_B.

To emphasize the orthogonality of sequence locality and structural locality, we require a (purely, i.e. sequence global) structural local motif for A to contain 1 and $|A|$, otherwise we may speak of a *sequence and structural local motif*. This extends to alignments.

For the later algorithmic treatment an alternative view of structural locality is required. Obviously, a structural local motif M for A and P (i.e. actually any motif $M \subseteq [1..|A|]$) is of the form $M = [i_1..i'_1] \cup \dots \cup [i_k..i'_k]$, i.e. it corresponds to a series of subsequences of A. The ranges $[i'_p + 1..i_{p+1} - 1]$ $(1 \leq p < k)$ are called *exclusions of* M, since we get M

by excluding them from the range $[i_1..i'_k]$. For an exclusion $[x..x']$ of a motif $M \subseteq [1..|A|]$ there is a base pair $(i, i') \in P, \{i, i'\} \in M$ where $(x, x') \prec (i, i')$. Denote the according to $\prec$ minimal such (i, i') as *bridge of* (x, x'). The following lemma gives an alternative characterization of structural locality, which will be used by our algorithm. An analogous statement is proven in [BW04].

Lemma 1 *A motif $M \subseteq [1..|A|]$ is structural local for A and P iff there is a bridge for each exclusion of M and each base pair in P is the bridge of at most one exclusion in M.*

Structural Locality in Multiple Alignments In contrast to our pairwise alignment definition, a multiple alignment, e.g. from Rfam, is usually given as a sequence of alignment columns. Thus it does not make explicit, which bases are locally aligned and which parts of the alignment are excluded from the structural local alignment due to their dissimilarity. However, structural locality can still be observed in such alignments.

For this purpose, multiple alignments are decomposed into their pairwise subalignments. Then, we assess structural locality by the presence of type I or type II exclusions in the pairwise alignments, which are defined as follows.

In a pairwise alignment $\mathcal{A}$, a *type I exclusion of length l and error rate e* is a subalignment (i.e. a continuous window) of l columns where 1.) in one sequence all columns contain a gap with the exception of at most $l \cdot e$ columns and 2.) no base in the l columns forms a base pair to any other base in the alignment.

A *type II exclusion in $\mathcal{A}$ of length l and error rate e* is a continuous window of l columns where 1.) more than $l \cdot e$ columns in one of the two sequences form a base pair with another base inside the window and 2.) for the other sequence, no bases inside of the window contribute to base pairs. Hence, type II exclusions correspond to the exclusion of substructures.

4 Structural Local Alignment

Based on the previous definitions, we will provide evidence for the ubiquity of structural locality in the results section. Here, we develop a structural local multiple alignment approach. The general workflow of the method is depicted in Figure 2.

Pairwise RNA Alignment We start our description by reviewing global and sequence-local pairwise alignment. [WRH+07] We compute an alignment $\mathcal{A}$ and a consensus structure $P = (P_A, P_B)$ of the given RNA sequences A and B that together maximize the score

$$\text{score}(\mathcal{A}, P) = \sum_{\substack{(i,k)\in P_A,(j,l)\in P_B \\ (i,j)\in\mathcal{A},(k,l)\in\mathcal{A}}} \tau(i,j,k,l) + \sum_{(i,j)\in\mathcal{A}_s} \sigma(A_i, B_j) - N_{\text{gap}}\gamma,$$

where N_{gap} denotes the number of gaps in $\mathcal{A}$ and $\tau(i,j,k,l)$ is the score contribution for matching the arcs (i,k) and (j,l). In LocARNA, $\tau(i,j,k,l)$ depends on the ensemble probabilities of the two arcs, as computed by McCaskill's algorithm [McC90], which is implemented in the Vienna RNA Package [HFS+94]. This kind of scoring by base pair probabilities was introduced for the tool PMcomp/PMmulti [HBS04] as a much simplified scoring for Sankoff-style simultaneous alignment and folding [San85]. In LocARNA, very improbable arcs (below a given threshold) are forbidden in P, which significantly reduces the algorithmic complexity, making the approach applicable in practice. For details see [WRH+07].

The score is efficiently maximized by a dynamic programming algorithm. First define a helper function

$$\mathrm{h}(M,k,l) = \max \begin{cases} M(k-1,l-1) + \sigma(A_j, B_l) \\ M(k-1,l) + \gamma \\ M(k,l-1) + \gamma \\ \max_{k'l'} M(k'-1,l'-1) + D_{i\,j\,k'\,l'}. \end{cases}$$

The DP algorithm is now specified by the recursion

$$\begin{aligned} M_{i\,j}(k,l) &= \mathrm{h}(M_{i\,j},k,l) \\ D_{i\,j\,k\,l} &= M_{i\,j}(k-1,l-1) + \tau(i,j,k,l). \end{aligned}$$

Initialisation is simply by $M_{i\,j}(k,i) = M_{i\,j}(i,k) = k\gamma$. As given, the recursion computes the global alignment score. For the case of sequence local alignment, where we search the best alignment of subsequences, we modify the recursion for $i = 0$ and $j = 0$ by

$$M_{0\,0}(k,l) = \max(0, \mathrm{h}(M_{0\,0},k,l))$$

with initialization $M_{0\,0}(k,0) = M_{0\,0}(0,k) = 0$.

Pairwise Structural Local RNA Alignment Due to Lemma 1, certain exclusions are allowed in structural local alignments. Algorithmically, this distinguishes structural local alignments from sequence local or global alignments. The score is extended by adding one exclusion cost ϵ per exclusion. According to Lemma 1 (raised from motifs to alignments in a straightforward way), each exclusion in a local alignment has a bridge in the consensus structure and no two exclusions share the same bridge. This is enforced by counting the number of exclusions below each arc match in both sequences. For this purpose, we distinguish eight states, corresponding to eight different matrices. State NN means there is no exclusion for the arc match starting at (i,j). State XN means there is exactly one exclusion for this arc match in the first sequence, state NX is analogous for the second sequence, and state XX means there is exactly one exclusion in each of the sequences. In addition we introduce states for alignments that have exclusions immediately at the right

end of the first or the second sequence, which can therefore be extended. At the same time we keep track of the number of exclusions in the other sequence. This results in states ON,NO,OX,XO. The recursions are now given as follows. For $i > 0$ or $j > 0$,

$$
\begin{aligned}
M_{i\,j}^{NN}(k\,l) &= \mathrm{h}(M_{i\,j}^{NN}, k, l) \\
M_{i\,j}^{NX}(k\,l) &= \max(\mathrm{h}(M_{i\,j}^{NX}, k, l), M_{i\,j}^{ON}(k-1\,l) + \epsilon) \\
M_{i\,j}^{XN}(k\,l) &= \max(\mathrm{h}(M_{i\,j}^{XN}, k, l), M_{i\,j}^{ON}(k\,l-1) + \epsilon) \\
M_{i\,j}^{XX}(k\,l) &= \max(\mathrm{h}(M_{i\,j}^{XX}, k, l), M_{i\,j}^{ON}(k-1\,l) + \epsilon, M_{i\,j}^{NO}(k\,l-1) + \epsilon) \\
M_{i\,j}^{ON}(k\,l) &= \max(M_{i\,j}^{ON}(k-1\,l), M_{i\,j}^{NN}(k\,l)) \\
M_{i\,j}^{OX}(k\,l) &= \max(M_{i\,j}^{OX}(k-1\,l), M_{i\,j}^{NX}(k\,l)) \\
M_{i\,j}^{NO}(k\,l) &= \max(M_{i\,j}^{NO}(k\,l-1), M_{i\,j}^{NN}(k\,l)) \\
M_{i\,j}^{XO}(k\,l) &= \max(M_{i\,j}^{XO}(k\,l-1), M_{i\,j}^{XN}(k\,l)).
\end{aligned}
$$

Now, the scores for alignments enclosed by arc matches are read of these matrices as

$$D_{i\,j\,k\,l} = \max_{s \in \{NN,NX,XN,XX\}} M_{i\,j}^{s}(k-1\,l-1) + \tau(i,j,k,l).$$

Finally, the complete alignment score is obtained by the same recursion as for the global or purely sequence local case by evaluating $M_{0\,0}(k,l) = \mathrm{h}(M_{0\,0}, k, l)$ or $M_{0\,0}(k,l) = \max(0, \mathrm{h}(M_{0\,0}, k, l))$, respectively.

Note that the time complexity of $O(|A|^2|B|^2)$ and the space complexity of $O(|A||B|)$, both complexities given under the assumption of a fixed probability threshold, is not increased by supporting structural locality. In a practical implementation, the space for storing the M matrices can be limited to grow by a factor of only 4, since for the states NO,ON,OX,XO it is sufficient to store only matrix lines (ON,OX) or even single values (NO,XO) for evaluating the recursion.

The actual alignment is produced from the alignment matrices by traceback. In order to maintain the good space complexity, the M-matrices are recomputed on demand during the traceback phase; notably this does not increase the total complexity.

Finally note that, although the recursions are given for linear gap cost only, the extension to affine gap cost can be done in the way of Gotoh without increasing the complexity. The needed additional space is only linear in the lesser sequence length.

Multiple Alignment Using T-Coffee For constructing a (structural local) multiple alignment of sequences $A^{(1)}, \ldots, A^{(m)}$, we compute all pairwise (structural local) alignments as described above. From the pairwise alignments, we compile a library of alignment edges $(L_{kl})_{1 \le k,l \le m}$. L_{kl} contains an edge (i,j) with an alignment score dependent weight (between 1000 and 2000) iff in the pairwise alignment of $A^{(k)}$ and $A^{(l)}$, $A_i^{(k)}$ is aligned to $A_j^{(l)}$. All other edges get a weight of zero. This library is fed as primary library to T-Coffee. From this, T-Coffee computes an extended library by increasing the edge weights of pairwise edges that transitively fit to alignment edges to third sequences. The multiple

alignment is finally computed in a progressive fashion much like CLUSTALW, however using the extended library for scoring base similarity.

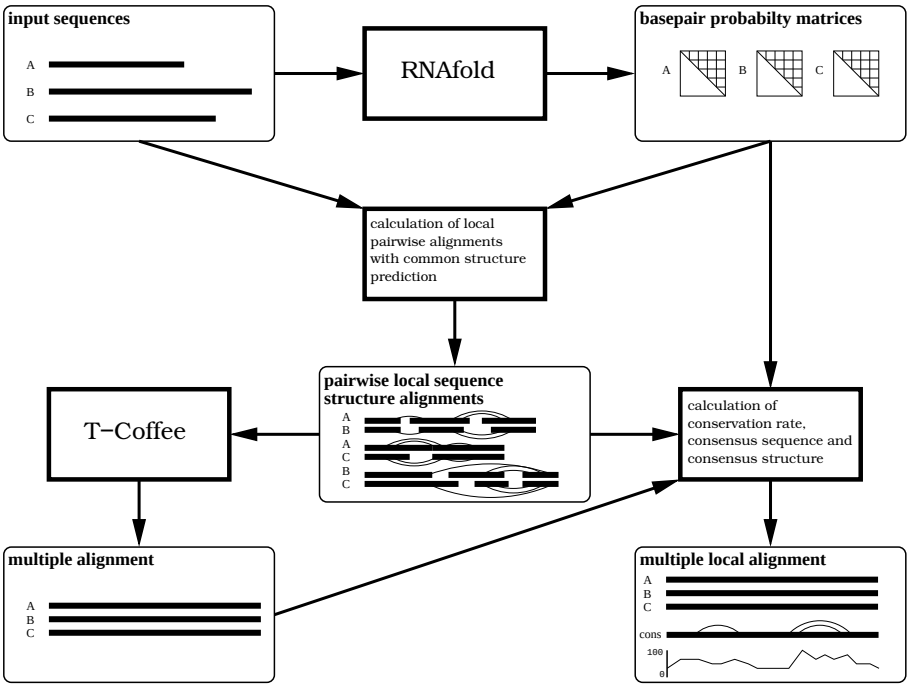


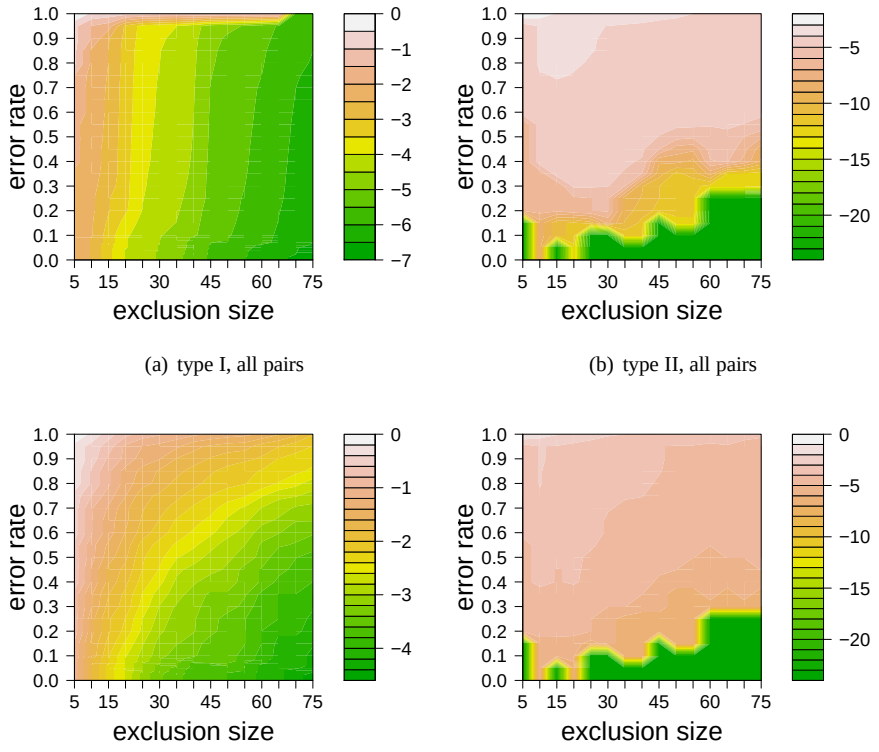
Figure 2: General workflow of the multiple alignment algorithm of LocARNATE

Local Motif of a Multiple Alignment Once a multiple alignment is constructed out of the (structural local) pairwise alignments, we can determine the structural local columns of this multiple alignment. This is done by assigning to each column a sum-of-pairs score over its pairwise alignment edges. There, each edge contributes with a weight of 1 if it got a non-zero weight in T-Coffee's primary library. As result, one gets a profile that reports a degree of locality for each column. Applying a fixed threshold, one finally extracts the local motif (subset of local columns) described by the alignment.

5 Results

Structural Locality in RNA Families In order to assess the demand for structural locality aware alignment, we analyze the occurrence of structural locality in the Rfam. We identify two reasons for structural locality. In alignments of two RNAs, type I exclusions of length l are subsequences of alignment columns where one of alignment strings consists of almost only gaps (with an error rate of e). Type II exclusions are subsequences, where

only one of the RNAs forms structure (again with error rate e). Our statistic of the Rfam seed sequences is shown in Figure 3.



(a) type I, all pairs

(b) type II, all pairs

(c) type I, families

(d) type II, families

Figure 3: Locality in the Rfam. We show the percentage of type I and type II exclusions for all pairs and for single families. Colors indicate frequency varying with exclusion size and allowed error rate.

LocARNATE: A Tool for Local Multiple Alignment Our structural locality aware multiple alignment approach for RNA, which combines an extended version of LocARNA with T-Coffee for constructing consistency based alignments, is implemented using C++ and Perl. It is available as the tool `locarnate` in the LocARNA software package.

Case Study Figure 4 gives an example for the identification of a local motif in a multiple local alignment.

Alignment Accuracy on the Bralibase The alignment accuracy of our approach is compared to two other programs Lara and FoldAlignM using the Bralibase benchmark. The

```
S1_AJ295015.1/58-1        .ACAGAGUCUGACAAA-------CCGUCACUGAAGACGUUCAA.C-UU----------------GCGUU----GAACAGAAACUCUGC
S2_AF170503.1/280-333     .GAAAGGUCUGUGCUU-------AGCACACUGACGA.GUUCCUGAAAU----------------G--------GAAC.GAAACCUUUU
S3_M83545.1/56-3          .CAUAAGUCUGGGCUA-------AGCCCACUGAUGA.GUCGCUGAAAU----------------G--------CGAC.GAAACUUAUG
S4_D00685.1/1-46          .........GCCAGACGU-G-GACCCGGCCUGAUGA.GUCCGAAA-------------------------GGAC.GAAACAGUA.
S8_J02439.1/42-95         .UGUCCGU.AGUGGAUGU-G-UAUCCACUCUGAUGA.GUCCGAAA-------------------------GGAC.GAAACGGAUG
S5_M17439.1/1-48          .........ACCGGAUGU-GCUUUCCGGUCUGAUGA.GUCCGUGA-------------------------GGAC.GAAACAGGAC
S6_AJ536620.1/206-152     CC.ACCGUCGGAAAGUGU-G-CGCUUUCCCUGAUGA.GCCCAAAA-------------------------GGGC.GAAACGGUAC
S7_Y12833.1/339-285       CC.GCUAUAUGGGGAUGU-G-UGUCCCUACUGACGA.GUUCAAAA-------------------------GAAC.GAAAUAGUUA
S10_Y14700.1/133-53       ..UCCAGUCGAGACCUGAAG-UGGGUUUCCUGAUGA.GGCUGUGGAGAGAGCGAAA-GCUUUACUCCCACACAAGCC.GAAACUGGA.
S9_AJ247113.1/134-53      ..UCCAGUCGAGACCUGAAG-UGGGUUUCCUGACGA.GGCUGUGGAGAGAGCAAAUUGCUUUACUCCCGCACAAGCC.GAAACUGGA.
#=GC SS_cons              ..<<<<<<.<<<<<<.......>>>>>>........<<<<....................................>>>>....>>>>>>.
#=GC conservation          .........*******-------*************.********----------------------------****.*********.
```

Figure 4: Example of a LocARNATE alignment of hammerhead ribozymes in stockholm-format. The line *#=GC conservation* marks conserved columns with a conservation rate of at least 0.5 by ***** (also highlighted in light gray) and excluded columns by **-** (darkgray). Note that the conserved columns correspond to the functional motive in Figure 1.

Bralibase consists of a collection of hand-curated multiple RNA alignments of 2 to 15 sequences each. We restrict the comparison to the most interesting subset of the Bralibase, namely alignments with less than 50% sequence identity. For the benchmark, one re-aligns the sequences of each such alignment with the candidate alignment tool and compares the result with the true alignment. The comparison is done by `compalignp`, as suggested for the Bralibase 2.1 benchmark [WMS06]. The resulting COMPALIGN score measures how accurately the generated alignment reproduces the given, true alignment - a score of 1.0 is optimal. This benchmark was done in the same way by Bauer et al.[BKR07], where Lara and FoldAlignM passed as the most successful sequence-structure alignment programs. The result of this test is reported in Figure 5.

An immediate, striking observation is that the tools LocARNATE and Lara seem to improve their accuracy with increasing number of sequences. The same effect is not seen for FoldAlignM, which is the only tool in this comparison that does not enjoy the consistency extension of T-Coffee. For 15 sequences, the comparably worse pairwise alignment of Lara is even outweighed by this effect and Lara is again on par with LocARNATE.

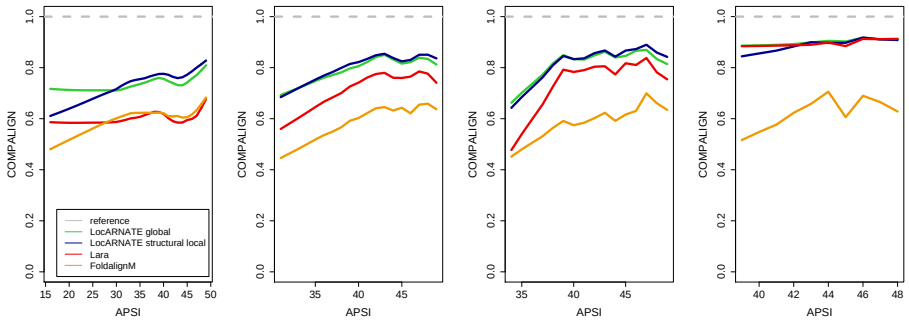


Figure 5: Benchmark on the Bralibase-fragment with $APSI < 50\%$ for alignments with 2,5,7, and 15 sequences (from left to right). The curves show the dependency between sequence identity (APSI) and alignment accuracy (COMPALIGN) for each of the four benchmarked algorithms.

Alignment Accuracy on Selected Rfam Alignments We select multiple subalignments of 7 sequences per alignment from the Rfam seed alignments. A benchmark set EI of 20 alignments with type I exclusions and a benchmark set EII with 10 type II exclusions is chosen. The sets EI (EII) are produced by each time selecting four pairwise alignments that have type I (type II) exclusions with length $l \geq 20$ ($l \geq 10$) and error rate $e \leq 0.25$ ($e \leq 0.6$), respectively. Of the eight sequences, we drop one at random. The, according to the Rfam, true alignment is obtained by projecting the corresponding Rfam family's seed alignment to the selected 7 sequences (deleting all only-gap columns). For each benchmark alignment, we align by LocARNATE with and without support of structural locality, Lara, and FoldAlignM. For each computed alignment, we obtain a COMPALIGN score by comparison with the true alignment. The results are shown in Figure 6.

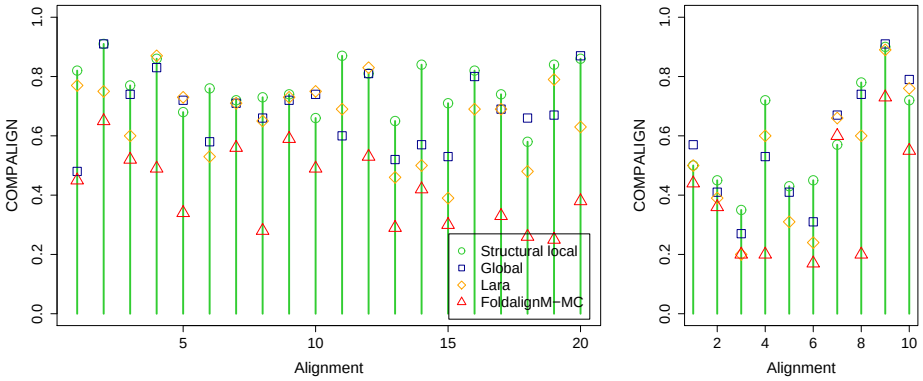


Figure 6: Benchmark on the alignment sets EI(left) and EII(right). Both sets consist of multiple alignments, each of seven sequences. EI contains type I exclusions, EII type II exclusions. The accuracy (COMPALIGN) is plotted for each single alignment and for each of the algorithms.

6 Conclusion

As we show by analysis of the whole Rfam database, structural locality is a wide spread feature of known RNA families. Structural locality is formalized by connectivity in the structure graph and via the notion of exclusions. Some families show strong structural locality, which motivates the development of special algorithmic support of this kind of locality. While current state-of-the art tools are not aware of this locality, we show that structural locality can be integrated into the tool LocARNA without increasing its complexity. By supporting this locality, the alignment accuracy for certain RNA families is increased significantly. We show by extensive benchmarks using the critical fragment of Bralibase 2.1 that the accuracy for families without obvious structural locality is not affected.

Acknowledgement Wolfgang Otto is supported by the Konrad-Adenauer-Stiftung as a scholarship holder. We thank the anonymous reviewers for their valuable comments.

References

[Bar04] David P. Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–97, 2004.

[BKR07] Markus Bauer, Gunnar W. Klau, and Knut Reinert. Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization. *BMC Bioinformatics*, 8:271, 2007.

[BW04] Rolf Backofen and Sebastian Will. Local Sequence-Structure Motifs in RNA. *Journal of Bioinformatics and Computational Biology (JBCB)*, 2(4):681–698, 2004.

[Cou02] Jennifer Couzin. Breakthrough of the year. Small RNAs make big splash. *Science*, 298(5602):2296–7, 2002.

[FW05] Martha J. Fedor and James R. Williamson. The catalytic diversity of RNAs. *Nat Rev Mol Cell Biol*, 6(5):399–412, 2005.

[GJMM+05] Sam Griffiths-Jones, Simon Moxon, Mhairi Marshall, Ajay Khanna, Sean R. Eddy, and Alex Bateman. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research*, 33 Database Issue:D121–4, 2005.

[HBS04] I. L. Hofacker, S. H. Bernhart, and P. F. Stadler. Alignment of RNA base pairing probability matrices. *Bioinformatics*, 20(14):2222–7, 2004.

[HFS+94] Ivo L. Hofacker, Walter Fontana, Peter F. Stadler, Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte Chemie*, 125:167–188, 1994.

[McC90] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–19, 1990.

[NHH00] C. Notredame, D. G. Higgins, and J. Heringa. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302(1):205–17, 2000.

[San85] David Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, 45(5):810–825, 1985.

[THG07] Elfar Torarinsson, Jakob H. Havgaard, and Jan Gorodkin. Multiple structural alignment and clustering of RNA sequences. *Bioinformatics*, 23(8):926–32, 2007.

[WMS06] Andreas Wilm, Indra Mainz, and Gerhard Steger. An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms Mol Biol*, 1:19, 2006.

[WRH+07] Sebastian Will, Kristin Reiche, Ivo L. Hofacker, Peter F. Stadler, and Rolf Backofen. Inferring Non-Coding RNA Families and Classes by Means of Genome-Scale Structure-Based Clustering. *PLOS Computational Biology*, 3(4):e65, 2007.