

Supplement to LocARNA-P: Accurate Boundary Prediction and Improved Detection of Structural RNAs

Extensive supplementary information including figures and raw data of our experiments is available online as web supplement at

<http://www.bioinf.uni-freiburg.de/Supplements/LocARNA-P/>.

The site furthermore provides the LocARNA software package, which comprises LocARNA-P and scripts for the ncRNA screen refinement pipeline, and points to documentation of the tools.

S1 Methods

S1.1 Inside and Outside Algorithm in the Context of Local Folding

Very large alignments, as computed for Figures 1B and C of the main text, can be computed in reasonable time only because LocARNA-P profits from local folding. Local folding imposes a restriction of the span ($j - i$) of base pairs (i, j) to be less or equal L . The idea of local folding is similarly used by the tool RNAfold, which computes base pair probabilities suited as input for LocARNA. The inside algorithm directly profits from a reduced span of base pairs without changing the algorithm. Its time complexity is reduced from $O(n^4)$ to $O(n^2L^2)$ in terms of the sequence length n and the maximal span L . For this result, consider that for each of the $O(n^2)$ base pair matches, corresponding to Z^D -entries, $O(L^2)$ Z^M -entries are computed. Because we consider only significant base pairs, each entry is computed in constant time.

The outside algorithm does not immediately profit from a limited base pair span, because the number of computed Z^M -entries is not straightforwardly bounded by $O(L^2)$ as the number of Z^M -entries in the case of the inside. For obtaining such a bound, we need to modify the outside recursion making use of the following lemma.

Lemma 1 *Assume that all significant base pairs in P_A and P_B have span less or equal L . For $j - i > L$ or $l - k > L$, $Z_{i,j;kl}^M = Z_{1j;1l}^M \cdot Z_{kn;l m}^M$.*

This is easily seen, because no base pairs matches that bridge the hole of alignments outside $i..j$ and $k..l$ are considered under the conditions of the lemma. Maintaining matrix slices $Z_{1..j}^M$ and $Z_{n..m}^M$, only $O(L^2)$ many entries $Z_{i,j;kl}^M$ with $j - i \leq L$ and $l - k \leq L$ need to be computed in order to compute an Z^D -entry. Computing one such $Z_{i,j;kl}^M$ entry either recurses to other such entries with $j - i \leq L$ and $l - k \leq L$ or to at most constantly many entries $Z_{i,j;kl}^M$ that satisfy the condition of the lemma and can therefore be computed without further recursion in constant time. With respective modifications, the outside algorithm runs in $O(n^2L^2)$ time. Recall that computing base pair match probabilities is in $O(n^2)$ and therefore dominated. Finally, also the computation of base match probabilities is $O(n^2L^2)$, because for each of the $O(n^2)$ base pair matches $(g, h) \sim (k, l)$ of Eq. 6, the necessary matrix slices are recomputed in $O(L^2)$ and there are $O(L^2)$ contributions to each match probability.

S1.2 Multiple Alignment and Probabilistic Consistency Transformation

Based on pairwise maximum expected accuracy (MEA) alignment, we construct a multiple alignment \mathcal{A} of sequences S_1, \dots, S_K following a progressive alignment strategy.

For the pairwise alignments, we maximize the expected accuracy that is defined as weighted sum of base match and arc match probabilities. MEA maximizes the expected overlap of the found alignment with suboptimal alignments according to the Boltzmann distribution. We weight base matches against arc matches because structural matches contribute more to the accuracy than sequence matches.

Finally, we use a progressive approach based on the pairwise MEA alignments to build a multiple alignment $\vec{\mathcal{A}}$ of alignment columns together with a consensus secondary structure \mathcal{S} on the columns of $\vec{\mathcal{A}}$. A large improvement is obtained by re-estimating the probabilities of base match and arc matches by probabilistic consistency transformation. This method is known from the approach Probcons by Do *et al.* (9), where it is used in the simpler case of transforming base match probabilities for sequence alignment.

Due to this transformation, the probabilities are re-calculated as

$$P'(i \sim k | S_a, S_b) = \frac{1}{K} \sum_{1 \leq c \leq K} \sum_{1 \leq k' \leq |S_c|} P'(i \sim k' | S_a, S_c) \cdot P'(k \sim k' | S_b, S_c)$$

for base matches and

$$P'((i, j) \sim (k, l) | S_a, S_b) = \frac{1}{K} \sum_{1 \leq c \leq K} \sum_{1 \leq k' < l' \leq |S_c|} P'((i, j) \sim (k', l') | S_a, S_c) \cdot P'((k, l) \sim (k', l') | S_b, S_c)$$

for arc matches.

Do *et al.* suggest to iterate this transformation several times. This iterative re-estimation leads to an extinction of probabilities, however, because in general $\sum_k P(i \sim k | S_a, S_b) < 1$ since position i has a non-zero probability of being deleted. Therefore, we apply the reestimation only once.

This extinction of probabilities can be avoided when deletion probabilities are explicitly considered in the correct way. One introduces $P(i \sim -_k | S_a, S_b)$ as the probability that i is deleted, where the first matched $i' < i$ is matched to k . However, this increases the cost of the consistency transformation considerably. We therefore omitted this correction (as has been done for similar reasons in Probcons (9)).

Iterative Refinement LocARNA-P improves progressive multiple alignment further by the technique of iterative refinement. Therefore, we iteratively split a multiple alignment into sub-alignments of two bi-partitions of the sequences S_1, \dots, S_K and realign them. Over the iterations, one evaluates the alignments and further refines the best alignment. We make two significant choices in implementing this scheme. We iterate in several rounds; in each round the alignment is split and realigned for all bipartitions of sequences that are induced by the tree. Secondly, we use the reliability score for evaluating the alignment.

S1.3 Finding Optimal Parameters for Boundary Prediction

As described in Section PREDICTING BOUNDARIES FROM A STAR PROFILE of the main text, we want to fit a two-step function to a reliability profile $f : \{1, \dots, n\} \rightarrow \mathbb{R}$. Hence, we want to find values a and b , where the larger value represents the signal level and the smaller value the background. As described, we can solve for given constants a and b , the recursion equations

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

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

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

The partition function variant of Eq. 3 of the main text is

$$\begin{aligned} Z^A(i) &= \exp(-\beta(f(i) - a)^2) \cdot (Z^A(i-1) + Z^B(i-1) \cdot \exp(-\beta\Delta)) \\ Z^B(i) &= \exp(-\beta(f(i) - b)^2) \cdot (Z^A(i-1) \cdot \exp(-\beta\Delta) + Z^B(i-1)) \end{aligned} \quad (2)$$

with initialization $Z^A(0) = 1$ and $Z^B(0) = 1$.

The four partial derivatives of Eq. 2 in directions of a and b are

$$\begin{aligned} \frac{\partial}{\partial a} Z^A(i) &= \exp(-\beta(f(i) - a)^2) \cdot \left(\frac{\partial}{\partial a} Z^A(i-1) + \frac{\partial}{\partial a} Z^B(i-1) \exp(-\beta\Delta) \right) \\ &\quad + 2\beta(f(i) - a) \exp(-\beta(f(i) - a)^2) \cdot (Z^A(i-1) + Z^B(i-1) \exp(-\beta\Delta)) \\ \frac{\partial}{\partial b} Z^A(i) &= \exp(-\beta(f(i) - a)^2) \cdot \left(\frac{\partial}{\partial b} Z^A(i-1) + \frac{\partial}{\partial b} Z^B(i-1) \exp(-\beta\Delta) \right) \\ \frac{\partial}{\partial a} Z^B(i) &= \exp(-\beta(f(i) - b)^2) \cdot \left(\frac{\partial}{\partial a} Z^B(i-1) + \frac{\partial}{\partial a} Z^A(i-1) \exp(-\beta\Delta) \right) \\ \frac{\partial}{\partial b} Z^B(i) &= \exp(-\beta(f(i) - b)^2) \cdot \left(\frac{\partial}{\partial b} Z^B(i-1) + \frac{\partial}{\partial b} Z^A(i-1) \exp(-\beta\Delta) \right) \\ &\quad + 2\beta(f(i) - b) \exp(-\beta(f(i) - b)^2) \cdot (Z^B(i-1) + Z^A(i-1) \exp(-\beta\Delta)). \end{aligned}$$

The four partial derivatives $\frac{\partial}{\partial a} Z^A(n)$, $\frac{\partial}{\partial b} Z^A(n)$, $\frac{\partial}{\partial a} Z^B(n)$, and $\frac{\partial}{\partial b} Z^B(n)$ can be computed by dynamic programming from the above recursion equations. Thus, we can efficiently determine the gradient vector

$$v = \begin{pmatrix} \frac{\partial}{\partial a} Z^A(n) + \frac{\partial}{\partial a} Z^B(n) \\ \frac{\partial}{\partial b} Z^A(n) + \frac{\partial}{\partial b} Z^B(n) \end{pmatrix}$$

for optimizing the partition function $Z^A(n) + Z^B(n)$.

After finding optimal a and b , we continue as in the case of given constants by solving Eq. 3 of the main text by dynamic programming and performing traceback.

S1.4 Computing the Reliability Score of an Alignment

The reliability score, defined in Section COLUMN-WISE STARS, BOUNDARY PREDICTION, AND GLOBAL STAR SCORES of the main text, is computed efficiently by dynamic programming. The algorithm evaluates the following Nussinov-style recursion equation for $1 \leq q \leq q' \leq |\mathcal{A}|$:

$$N(q, q - 1) = 0$$

$$N(q, q') = \max(N(q, q' - 1) + \text{seqSTAR}_{\mathcal{A}}(q'), \max_{q \leq k < q'} N(q, k - 1) + \omega \text{bpSTAR}_{\mathcal{A}}(q, q') + N(k + 1, q' - 1)).$$

Finally, $\text{STARS}_{\mathcal{A}} = |\mathcal{A}|^{-1}N(1, |\mathcal{A}|)$.

S2 Results

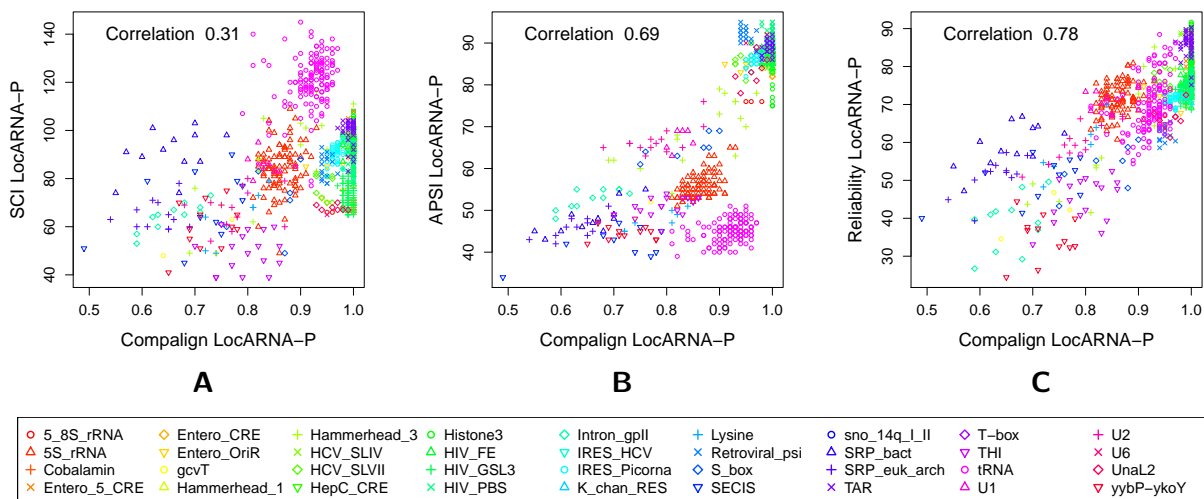
S2.1 Reliabilities of Alignments as Quality Measurement

S2.1.1 Correlation of Different Alignment Features with Alignment Quality

LocARNA already provided an alignment score, which is used to determine the optimal alignment for a pair of sequences. However, note that raw alignment scores are usually not suited for our purpose since they were not designed to be used as estimates of the quality of the output alignment. Rather, they are used to distinguish between different alignments of the *same* sequences in order to predict the best alignment. Thus there is no reason why the numerical values should be comparable between different sets of input sequences. Many sequence-structure alignment programs, such as Lara, PMcomp or LocARNA do not compute an alignment score for their multiple alignments at all. FoldAlignM is one of the few programs that returns such alignment scores. In this particular case there is a weak correlation between alignment quality and alignment score (cf. Supplementary Figure 3).

Column reliabilities, on the other hand, provide a theoretically well-founded basis for computing a reliability score of a complete LocARNA-P alignment that assesses how trustworthy the generated alignment is. The *reliability score of an alignment \mathcal{A} and a consensus structure* is defined as the sum of sequence reliabilities $\text{seqSTAR}_{\mathcal{A}}(q)$ for single-stranded positions q of the consensus structure, and base pair reliabilities $\text{bpSTAR}_{\mathcal{A}}(q, q')$ for consensus base pairs (q, q') . The *reliability score of the alignment \mathcal{A}* is then defined as the maximum reliability score of the alignment over all possible consensus structures normalized by alignment length. It can be calculated efficiently using a Nussinov-style algorithm (cf. Section COLUMN-WISE STARS, BOUNDARY PREDICTION, AND GLOBAL STAR SCORES of the main text and Section COMPUTING THE RELIABILITY SCORE OF AN ALIGNMENT).

We study the ability of several features to predict the quality of LocARNA-P alignments by computing LocARNA-P alignments for a benchmark data set of 10-fold alignments. The benchmark data set is drawn from Rfam, such that hand-curated reference alignments are available and it covers the variety of ncRNA families (55) (details given in Section BENCHMARK SET USED FOR CORRELATION ANALYSIS). For each alignment, we obtain the **compalign** score that compares the generated alignment with the reference alignment. This score is a good measure of the true alignment quality. The features are compared against this quality measure and correlation is measured as Pearson correlation coefficient.



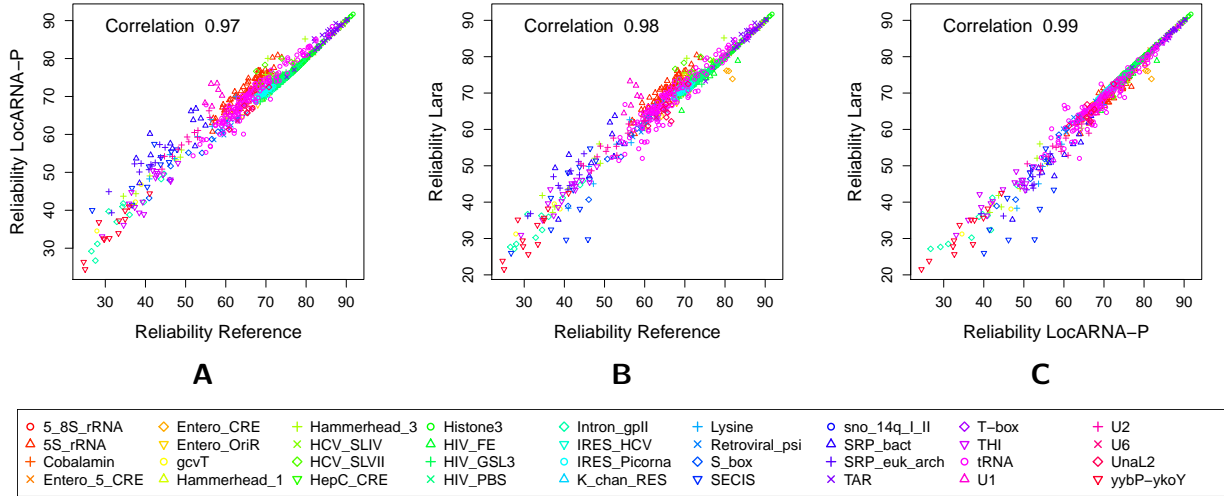
Supplementary Figure 1. Correlation plots for features over the benchmark data set. We distinguish members of different families as described in the legend. Alignment quality (as measured by compalign) versus A) SCI B) APSI and C) reliability. Note that the comparison to SCI in (A) shows very weak correlation (correlation coefficient 0.31.)

Other features of an alignment are *average pairwise sequence identity* (APSI) and *structure conservation index* (SCI); the latter compares the free energy of the consensus structure with the average of minimum free energy of individual sequences. Those measures have been used as additional features for assessing alignment quality (e.g., see Rose *et al.* (41)).

Supplementary Figure 1A shows that the SCI is not correlated well to the quality of LocARNA-P alignments (correlation 0.31). APSI shows better correlation of 0.69 (see Suppl. Fig. 1B). However, this correlation is not surprising since sequences with high APSI are much easier to align than sequences with low APSI. Interestingly, the correlation of APSI with the alignment reliability score is comparably weak (correlation 0.52.)

We calculated the reliability scores for all LocARNA-P-alignments of the benchmark set. Supplementary Figure 1C shows the reliability score against alignment quality with correlation coefficient of 0.78.

Inspired by this, we study the reliability score also for alignments that are not produced by LocARNA-P, but by other methods (like Lara), or even for the reference alignments. For this purpose, we calculate base match and base pair match probabilities using LocARNA-P. Then, we calculate the column-wise reliabilities by equations 1 and 2 of the main text, using the given alignment instead of the LocARNA-P alignment, and combine these column-wise reliabilities to a reliability score as described above. We have found a very strong correlation between the reliability of the best alignment produced by LocARNA-P and reliabilities of both, the reference alignment from Rfam and the best alignment produced by Lara (see Supplementary Figure 2A, B and C). This indicates that LocARNA-P probabilities approximate true alignment match probabilities very well and yield an excellent general model of sequence structure alignment.



Supplementary Figure 2. Correlation over the benchmark set between reference reliability and reliability of best alignment produced by LocARNA-P (A) and Lara (B), as well as between LocARNA-P and Lara (C), respectively

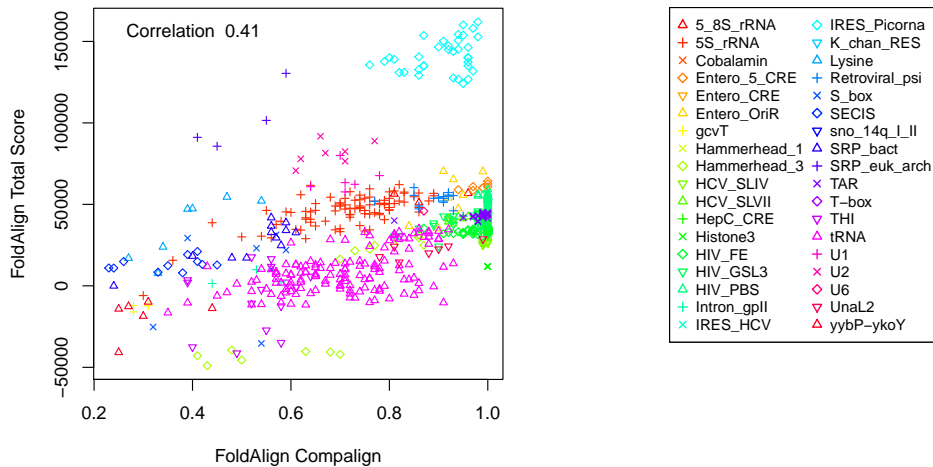
Therefore, the reliability score calculated by LocARNA-P can be utilized as quality measure *even* for alignment that are not produced by LocARNA-P.

S2.1.2 Benchmark set used for Correlation Analysis

For studying the ability of several features to predict the quality of sequence-structure alignments, we choose a subset of Rfam as our benchmark set. For the benchmark set, one draws k -way alignments from Rfam, such that we get sets of k sequences each with known hand-curated reference alignments. Such a set covers the variety of families in the Rfam. This protocol was already performed for the benchmark sets of Bralibase 2.1 (55). Therefore, we choose the Bralibase benchmark set k10 of 10-way alignments as starting point for our benchmark set. k10 consists of 845 sets of 10 related RNA sequences from 36 RNA families. When inspecting the set, we noticed that all instances from the family IRES.HCV deviate from the other benchmark instances by their large length diversity. Therefore, we exclude IRES.HCV to obtain a more homogeneous set.

For the correlation analysis, we determine several features for each benchmark instance and its LocARNA-P alignment. Notably, the `compalign` score is computed, which compares the generated alignment with the reference alignment. This score is a good measure of the true alignment quality since we can believe in the reference alignment. Correlation is measured as Pearson correlation coefficient.

For the family IRES.HCV, which we didn't include in our benchmark set, we observe the following. The quality of alignments in this family is not as well correlated to reliabilities as for all other RNA families. Whereas for our benchmark set, we observe a correlation of the reliability



Supplementary Figure 3. FoldAlign score versus alignment quality (as measured by compalign). In the plot, we distinguish members of different RNA families by colors and symbols as given in the legend on the right.

score and the alignment quality of 0.78, doing the same analysis for the k10 benchmark set (including IRES_HCV instances) yields correlations 0.57. The reason for the different behavior is that while the family shows a very high APSI between 84% and 95% making them easy to align, the family members largely vary in sequence lengths. This leads to the insertion of many gaps and therefore weakens the reliability score. Recall that gaps do not contribute to the reliability score as explained on Section COLUMN-WISE STARS, BOUNDARY PREDICTION, AND GLOBAL STAR SCORES of the main text.

S2.1.3 Correlation of Alignment Scores and Alignment Quality

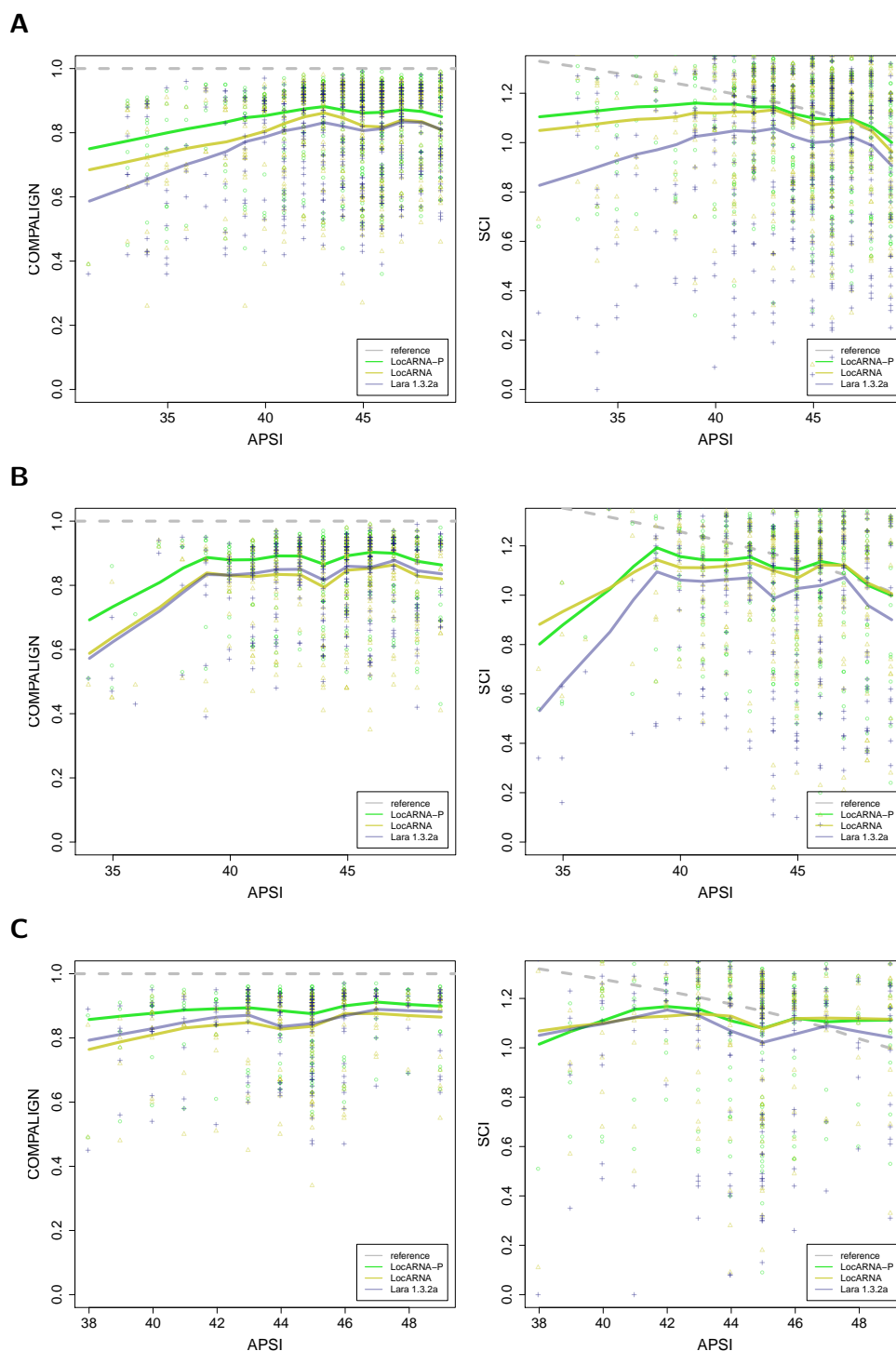
FoldAlignM is one of the few programs that yield an alignment score, whereas programs such as PMcomp and Lara as well as LocARNA do not return a score for their multiple alignments. Therefore, we investigate the correlation of the FoldAlignM score and PMcomp score of the FoldAlignM alignments to reference alignments for the previously described benchmark set of 10-fold alignments from Rfam.

As we did before for features of the LocARNA-P alignments, we study the ability of the FoldAlignM score to predict alignment quality by computing FoldAlignM alignments for the benchmark set and determining the compalign score for measuring the alignment quality.

Supplementary Figure 3 shows the correlation between FoldAlignM's score and the alignment quality (correlation coefficient 0.41). Since FoldAlignM didn't yield single alignments for all of the benchmark instances, we excluded those 31 instances from the FoldAlignM comparison.

S2.2 Assessing Multiple Alignment Performance by the Bralibase 2.1 Benchmark

We measure the performance of *LocARNA-P* using the Bralibase 2.1 benchmark. The benchmark consists of a series of multiple alignment problem instances with known reference alignments. The sequences and their reference alignments are selected from the Rfam such that the reference alignments can be trusted (55). The benchmark distinguishes data sets k2, k3, k5, k7, k10, k15 respectively for pairwise, 3-way, 5-way, 7-way, 10-way, and 15-way alignments. For benchmarking an alignment program, it is suggested to compute all multiple alignment for each data set and evaluate the alignments by their `compalign` score to the reference and structure conservation index (SCI). For comparing the performance of several algorithms, one plots the evaluations versus the average pairwise sequence identity of the reference alignment (APSI). We compared the predecessor tool *LocARNA*, *Lara* (1) and the new method *LocARNA-P*. We used the most recent released version of *Lara* (1.3.2a). For the *LocARNA-P* alignments, we performed consistency-transformation and two rounds of iterative refinement. A more comprehensive Bralibase 2.1 comparison of alignment tools was given by Bauer *et al.* (1) and is therefore not repeated here. Supplementary Figure 4 shows the outcome on the benchmark sets k5, k7, and k10 of 5-, 7-, and 10-way alignments. The complete benchmark results are available online in the web supplement to the paper. Our results show a significant performance improvement over the non-probabilistic tool *LocARNA* and better alignment accuracy than *Lara* on the Bralibase 2.1 benchmark. The latter is clearly significant for up to 10-way multiple alignments. For the k15 benchmark set, the slight advantage of *LocARNA-P* over *Lara* should be judged critically. For 15-way alignments, there are only few multiple alignment instances in the important low APSI range. Even worse, there is almost no diversity, since the instances are mainly tRNA alignments.



Supplementary Figure 4. Bralibase 2.1 benchmark for A) 5-way, B) 7-way, and C) 10-way alignments. Alignment accuracy is measured by comparison to the reference alignment (compalign score, left) and structure conservation index (SCI, right). The measures are plotted vs. average pairwise sequence identity (x-axis). The figures show that LocARNA-P significantly improves alignment accuracy for the hard low sequence identity instances over Lara, which in (1) performed best in this benchmark among a series of competitors, and its non-probabilistic older sibling LocARNA.

References

1. Bauer, M., Klau, G. W., and Reinert, K. (2007). Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization. *BMC Bioinformatics*, **8**, 271.
2. Bertone, P., Stoc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., Rinn, J. L., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M., and Snyder, M. (2004). Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242–2246.
3. Bompfünnewerer, A. F., Backofen, R., Bernhart, S. H., Hertel, J., Hofacker, I. L., Stadler, P. F., and Will, S. (2008). Variations on RNA folding and alignment: lessons from Benasque. *Journal of Mathematical Biology*, **56**(1-2), 129–144.
4. Bradley, R. K., Pachter, L., and Holmes, I. (2008). Specific alignment of structured RNA: stochastic grammars and sequence annealing. *Bioinformatics*, **24**(23), 2677–83.
5. Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. Wadsworth.
6. Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammanna, H., Helt, G., Sementchenko, V., Piccolboni, A., Bekiranov, S., Bailey, D. K., Ganesh, M., Ghosh, S., Bell, I., Gerhard, D. S., and Gingeras, T. R. (2005). Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, **308**, 1149–1154.
7. Clark, A. G., Eisen, M. B., Smith, D. E., and MacCallum, I. (2007). Evolution of genes and genomes on the drosophila phylogeny. *Nature*, **450**(7167), 203–18.
8. Coventry, A., Kleitman, D. J., and Berger, B. (2004). MSARI: multiple sequence alignments for statistical detection of RNA secondary structure. *Proc. Natl. Acad. Sci. USA*, **101**(33), 12102–7.
9. Do, C. B., Mahabhashyam, M. S. P., Brudno, M., and Batzoglou, S. (2005). ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res*, **15**(2), 330–40.
10. Do, C. B., Foo, C.-S., and Batzoglou, S. (2008). A max-margin model for efficient simultaneous alignment and folding of RNA sequences. *Bioinformatics*, **24**(13), i68–76.
11. Friendewey, D., Dingermann, T., Cooley, L., and Soll, D. (1985). Processing of precursor tRNAs in *Drosophila*. Processing of the 3' end involves an endonucleolytic cleavage and occurs after 5' end maturation. *Journal of Biological Chemistry*, **260**(1), 449–54.
12. Gardner, P. P., Wilm, A., and Washietl, S. (2005). A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Research*, **33**(8), 2433–9.
13. Gorodkin, J., Heyer, L., and Stormo, G. (1997). Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res*, **25**(18), 3724–32.

14. Gruber, A. R., Kilgus, C., Mosig, A., Hofacker, I. L., Hennig, W., and Stadler, P. F. (2008a). Arthropod 7SK RNA. *Mol Biol Evol*, **25**(9), 1923–30.
15. Gruber, A. R., Bernhart, S. H., Hofacker, I. L., and Washietl, S. (2008b). Strategies for measuring evolutionary conservation of RNA secondary structures. *BMC Bioinformatics*, **9**, 122.
16. Gruber, A. R., Findeiss, S., Washietl, S., Hofacker, I. L., and Stadler, P. F. (2010). RNAZ 2.0: Improved noncoding RNA detection. In *PSB10*, volume 15, pages 69–79.
17. Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., and Altman, S. (1983). The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, **35**(3 Pt 2), 849–57.
18. Harmanci, A. O., Sharma, G., and Mathews, D. H. (2008). PARTS: probabilistic alignment for RNA joint secondary structure prediction. *Nucleic Acids Research*, **36**(7), 2406–17.
19. Havgaard, J. H., Lyngso, R. B., Stormo, G. D., and Gorodkin, J. (2005). Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*, **21**(9), 1815–24.
20. Heyne, S., Will, S., Beckstette, M., and Backofen, R. (2009). Lightweight comparison of RNAs based on exact sequence-structure matches. *Bioinformatics*, **25**(16), 2095–2102.
21. Höchsmann, M., Töller, T., Giegerich, R., and Kurtz, S. (2003). Local similarity in RNA secondary structures. In *Proceedings of Computational Systems Bioinformatics (CSB 2003)*, volume 2, pages 159–168. IEEE Computer Society.
22. Hofacker, I. L. and Stadler, P. F. (2004). The partition function variant of sankoff’s algorithm. In *Computational Science - ICCS 2004, Part IV*, Lecture Notes in Computer Science LNCS 3039, pages 728–735, Heidelberg. Springer Verlag.
23. Hofacker, I. L., Bernhart, S. H., and Stadler, P. F. (2004). Alignment of RNA base pairing probability matrices. *Bioinformatics*, **20**(14), 2222–7.
24. Klein, R. J. and Eddy, S. R. (2003). RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics*, **4**(1), 44.
25. Knudsen, B. and Hein, J. (2003). Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Research*, **31**(13), 3423–8.
26. Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–857.
27. Langenberger, D., Bermudez-Santana, C., Hertel, J., Hoffmann, S., Khaitovich, P., and Stadler, P. F. (2009). Evidence for human microRNA-offset RNAs in small RNA sequencing data. *Bioinformatics*, **25**(18), 2298–301.
28. Lau, N. C., Lim, L. P., Weinstein, E. G., and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in *caenorhabditis elegans*. *Science*, **294**, 858–862.

29. Lee, R. and Ambros, V. (2001). An extensive class of small RNAs in *Caenorhabditis elegans*. *Science*, **294**, 862–864.
30. Lee, Y. S., Shibata, Y., Malhotra, A., and Dutta, A. (2009). A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev*, **23**(22), 2639–49.
31. Lofquist, A. and Sharp, S. (1986). The 5'-flanking sequences of *Drosophila melanogaster* tRNA^{Asn} genes differentially arrest RNA polymerase III. *Journal of Biological Chemistry*, **261**(31), 14600–6.
32. Marz, M., Donath, A., Verstaete, N., Nguyen, V. T., Stadler, P. F., and Bensaude, O. (2009). Evolution of 7SK RNA and its protein partners in metazoa. *Mol. Biol. Evol.*, **26**, 2821–2830.
33. Mathews, D. H. and Turner, D. H. (2002). Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *Journal of Molecular Biology*, **317**(2), 191–203.
34. Mattick, J. S., Taft, R. J., and Faulkner, G. J. (2009). A global view of genomic information - moving beyond the gene and the master regulator. *Trends in Genetics*.
35. McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**(6-7), 1105–19.
36. Missal, K., Rose, D., and Stadler, P. F. (2005). Non-coding RNAs in *Ciona intestinalis*. *Bioinformatics*, **21 Suppl 2**, ii77–ii78.
37. Missal, K., Zhu, X., Rose, D., Deng, W., Skogerbo, G., Chen, R., and Stadler, P. F. (2006). Prediction of structured non-coding RNAs in the genomes of the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *J Exp Zool B Mol Dev Evol*, **306**(4), 379–92.
38. Morl, M. and Marchfelder, A. (2001). The final cut. The importance of tRNA 3'-processing. *EMBO Rep*, **2**(1), 17–20.
39. Pedersen, J. S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E. S., Kent, J., Miller, W., and Haussler, D. (2006). Identification and Classification of Conserved RNA Secondary Structures in the Human Genome. *PLoS Comput Biol*, **2**(4), e33.
40. Rivas, E. and Eddy, S. R. (2001). Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, **2**(1), 8.
41. Rose, D., Hackermuller, J., Washietl, S., Reiche, K., Hertel, J., Findeiss, S., Stadler, P. F., and Prohaska, S. J. (2007). Computational RNomics of drosophilids. *BMC Genomics*, **8**, 406.
42. Roshan, U. and Livesay, D. R. (2006). Probalalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics*, **22**(22), 2715–21.

43. Sankoff, D. (1985). Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, **45**(5), 810–825.
44. Siebert, S. and Backofen, R. (2005). MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics*, **21**(16), 3352–9.
45. Smith, C. M. and Steitz, J. A. (1998). Classification of gas5 as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5'-terminal oligopyrimidine gene family reveals common features of snoRNA host genes. *Mol Cell Biol*, **18**(12), 6897–909.
46. The FANTOM Consortium (2005). The transcriptional landscape of the mammalian genome. *Science*, **309**(5740), 1559–63.
47. Torarinsson, E., Sawera, M., Havgaard, J. H., Fredholm, M., and Gorodkin, J. (2006). Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res*, **16**(7), 885–9.
48. Torarinsson, E., Havgaard, J. H., and Gorodkin, J. (2007). Multiple structural alignment and clustering of RNA sequences. *Bioinformatics*, **23**(8), 926–32.
49. Torarinsson, E., Yao, Z., Wiklund, E. D., Bramsen, J. B., Hansen, C., Kjems, J., Tommerup, N., Ruzzo, W. L., and Gorodkin, J. (2008). Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions. *Genome Res*, **18**(2), 242–51.
50. Uzilov, A. V., Keegan, J. M., and Mathews, D. H. (2006). Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. *BMC Bioinformatics*, **7**(1), 173.
51. Washietl, S. and Hofacker, I. L. (2004). Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *Journal of Molecular Biology*, **342**(1), 19–30.
52. Washietl, S., Hofacker, I. L., and Stadler, P. F. (2005a). Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA*, **102**(7), 2454–9.
53. Washietl, S., Hofacker, I. L., Lukasser, M., Huttenhofer, A., and Stadler, P. F. (2005b). Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat Biotechnol*, **23**(11), 1383–90.
54. Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F., and Backofen, R. (2007). Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. *PLOS Computational Biology*, **3**(4), e65.
55. Wilm, A., Mainz, I., and Steger, G. (2006). An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms Mol Biol*, **1**, 19.
56. Yao, Z., Weinberg, Z., and Ruzzo, W. L. (2006). CMfinder – a covariance model based RNA motif finding algorithm. *Bioinformatics*, **22**(4), 445–52.