APPENDIX

SPARSE: Quadratic Time Simultaneous Alignment and Folding of RNAs Without Sequence-Based Heuristics

Sebastian Will, Christina Otto, Milad Miladi, Mathias Möhl and Rolf Backofen

1 SUBSEQUENCE SCORE AND PROOF OF LEMMA 1

For the proof of Lemma 1 (Main Text), we need to define the score of subsequences.

DEFINITION 1 (Subsequence Score). *Given sequences A and B, with according functions σ, Ψ^A and Ψ^B. The* score of a structure alignment triple $(S,T,\mathcal{A})$ of subsequences $A[\hat{i}..\hat{j}]$ of A and $B[\hat{k}..\hat{l}]$ of B *is defined analogously to the score of entire sequences* (1) *as*

$$\begin{aligned}\text{score}[\hat{i}..\hat{j}\,\hat{k}..\hat{l}](S,T,\mathcal{A}) :=& \\ \sum_{(i,j)\in S} \Psi^A_{i+\hat{i}-1\,j+\hat{i}-1} &+ \sum_{(k,l)\in T} \Psi^B_{k+\hat{k}-1\,l+\hat{k}-1} \\ + \sum_{(i,k)\in\mathcal{A}} \sigma(i+\hat{i}-1, k+\hat{k}-1) &+ N_{indel}\gamma, \qquad (1)\end{aligned}$$

where N_{indel} is the number of indels in the alignment of the subsequences.

In the definition, note that since $(S,T,\mathcal{A})$ is a structure alignment triple, S is a structure of $A[\hat{i}..\hat{j}]$, S is a structure of $B[\hat{k}..\hat{l}]$, and $\mathcal{A}$ is an alignment of the subsequences. Essentially, we map the positions of the subsequences in intervals $[1..(\hat{j}-\hat{i}+1)]$ and $[1..(\hat{l}-\hat{k}+1)]$ back to the coordinates of the entire sequences when accessing values of σ, Ψ^A and Ψ^B.

PROOF OF LEMMA 1. $M^{ab}(i,k) \geq M^{ab}(i^*,k^*)+(i-i^*)\gamma+(k-k^*)\gamma$ holds, since each optimal sparse structure alignment triple of the subsequences $A[a^L+1\ldots i^*]$ and $B[b^L+1\ldots k^*]$ is a sparse structure alignment triple of the subsequences $A[a^L+1\ldots i]$ and $B[b^L+1\ldots k]$ with score $M^{ab}(i^*,k^*)+(i-i^*)\gamma+(k-k^*)\gamma$.

For "$\leq$", let $(S,T,\mathcal{A})$ be an, in the sparse structure and alignment space, optimal structure alignment triple of the subsequences $A[a^L+1\ldots i]$ and $B[b^L+1\ldots k]$. Then

$$\text{score}[a^{\text{L}}+1..a^{\text{R}}-1\,b^{\text{L}}+1..b^{\text{R}}-1](S,T,\mathcal{A}) = M^{ab}(i,k).$$

Let us restrict $(S,T,\mathcal{A})$ to the subsequences $A[a^L+1\ldots i^*]$ and $B[b^L+1\ldots k^*]$ by removing any base pair $(j,j')\in S$ with $\{j,j'\}\not\subseteq\{a^L+1\ldots i^*\}$, any base pair $(l,l')\in T$ with $\{l,l'\}\not\subseteq\{b^L+1\ldots k^*\}$, and any match $(j,l)\in\mathcal{A}$ with $j\notin\{a^L+1\ldots i^*\}$ or $l\notin\{b^L+1\ldots k^*\}$. We call this restriction $(S^r,T^r,\mathcal{A}^r)$. One can show that $(S,T,\mathcal{A})=(S^r,T^r,\mathcal{A}^r)$.

Subproof: Assume that $S^r\neq S$. Since $S^r\subseteq S$, there is a base pair $(j,j')\in S\setminus S^r$ with $i^*<j'\leq i$. Since $(S,T,\mathcal{A})$ is in the sparse space, this implies $(j,j')\in P$ and $\Pr^{\text{loop}}_a[(j,j')|A]\geq\theta_3$. Consequently, j' is represented, which contradicts the maximality of i^*. Analogous arguments show $T^r=T$ and $\mathcal{A}^r=\mathcal{A}$.

Finally, $(S,T,\mathcal{A})=(S^r,T^r,\mathcal{A}^r)$ implies

$$\begin{aligned}\text{score}[a^{\text{L}}+1..a^{\text{R}}-1\,b^{\text{L}}+1..b^{\text{R}}-1](S,T,\mathcal{A}) =& \\ \text{score}[a^{\text{L}}+1..i^*\,b^{\text{L}}+1..j^*](S^r,T^r,\mathcal{A}^r)& \\ +(i-i^*)\gamma+(k-k^*)\gamma.&\end{aligned}$$

In turn, $M^{ab}(i,k)\leq M^{ab}(i^*,k^*)+(i-i^*)\gamma+(k-k^*)\gamma$.

2 FURTHER EVALUATION RESULTS

Figure 1 compares the alignment quality vs. sequence identity behavior of LocARNA, SPARSE, and RAF for three-way multiple alignment instances (benchmark set k3 of Bralibase 2.1.)

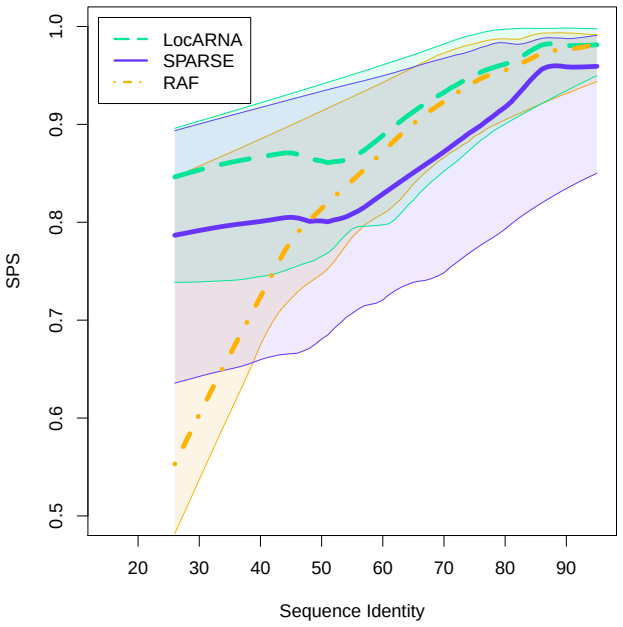


Fig. 1. Alignment quality (measured by sum-of-pairs score SPS) at different sequence identities for three-way alignments (Bralibase 2.1 set k3). The representation is analogous to Figure 4 (Main Text).

3 EXAMPLE ALIGNMENT SPARSE VS. LOCARNA

We compare the alignment and folding of two RNAs from family gcvT by LocARNA and SPARSE (Figure 2.) This example illustrates the benefits of incorporating loop deletions and insertion. Since LocARNA disallows loop deletions and insertions, it cannot predict structure in deleted regions; consequently, it predicts large unstable loops. In contrast, SPARSE supports loop deletions (represented by '_' in the alignment) and thus predicts a large stem in sequence A, which is deleted in sequence B. Thereby, the remaining parts can also form a stable structure with short loop regions. This behavior is reflected by the MCC prediction quality scores,

LocARNA:

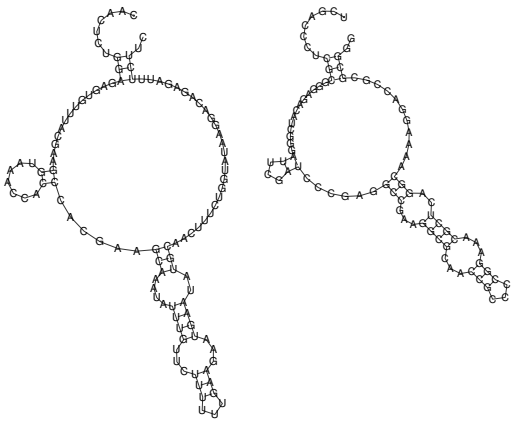
A

```
  -........(((..............((........)).....
A -CAACUCUGGAGAGUGUUUACGAAGGUAAACCACCCACGA
B UCGACCCUCGCGGGAGACAUCGGGAUU----CGAUCCCGA
  .........(((..............((.----..)).....

  .(((....((((...(((...))))...)))))..))))....
A AGCAAAUAUUUGUUCUUUUUUGAAGAAUGAAUAUGCAACU
B GGCCGA-AGGCGCAACCGCCCCGGAAACGCUCAGGCAA--
  .(((..-.((((...(((...))))...)))))..))))..--

  ........................)))..
A UUCUGGUAUAAGGACAGAGAUUUCUUC
B ----------AAGGACCG----CGCGGG
  ----------........----.))))..
```

B



SPARSE:

C

```
  ----.((-(((((.....((((((....-))))))).(((
A ----CAA-CUCUGGAGAGUGUUUACGAAG-GUAAACCACC
B UCGACCCUCGCGGGAGACAUCGGGAUUCGAUCCCGAGGCC
  .....((.(((((.....((((((.....)))))).(((

  ...((((...((((((((((((...))))))))))))).
A CACGAAGCAAAUAUUUGUUCUUUUUUGAAGAAUGAAUAUG
B GAAGGCGCAACCG_____________CCC_____________CGGA
  ...((((...(((_____________..._____________))).

  ...))))..)))........)))))..))......
A CAACUUUCUGGUAUAAGGACAGAGAUUUCUUC
B -AACGCUCAGGCAAAAGGACCGCGCGGG----
  -..)))))..)))........)))))..)).-----
```

D

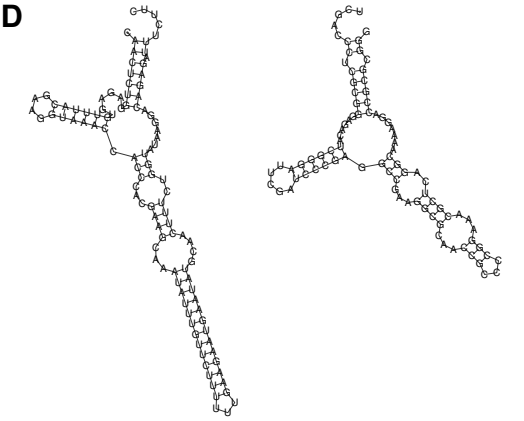


Fig. 2. Alignment and structure prediction results of LocARNA and SPARSE for two example RNAs of the family gcvT. **A** and **C** show the respective alignments computed by the tools; respective subfigures **B** and **D** visualize the simultaneously predicted structures projected on each sequence A and B by RNAplot (Lorenz *et al.*, 2011). **A,B** Since LocARNA cannot predict structure in deleted regions, large unpaired regions are predicted in the multiloop, which destabilize the structures. **C,D** The more flexible model in SPARSE allows loop deletions and insertions (represented by '_' in the alignment) and thus can align stems of varying length; in this example, the stems at the bottom. This results in smaller loops and thus more stable structures.

which assess the similarity of the predicted structures to the Rfam-derived reference structures, ranging from 0 to 1. In this example, LocARNA achieves a MCC of 0.51, which SPARSE improves to 0.94.

REFERENCES

Lorenz, R., Bernhart, S. H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms Mol Biol*, **6**, 26.