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

Sebastian Will[1,2⋆], Christina Schmiedl[1,⋆], Milad Miladi[1], Mathias Möhl[1] and Rolf Backofen[1,3,4,5⋆⋆]

[1]Bioinformatics, Department of Computer Science, University of Freiburg, Germany
[2]Bioinformatics, Department of Computer Science, University of Leipzig, Germany
[3]Centre for Biological Signalling Studies (BIOSS), University of Freiburg, Germany
[4]Centre for Biological Systems Analysis (ZBSA), University of Freiburg, Germany
[5]Centre for Non-coding RNA in Technology and Health, Bagsvaerd, Denmark

Motivation: There is increasing evidence of pervasive transcription, resulting in hundreds of thousands of ncRNAs of unknown function. Standard computational analysis tasks for inferring functional annotations like clustering require fast and accurate RNA comparisons based on sequence and structure similarity. The gold standard for the latter is Sankoff's algorithm [3], which simultaneously aligns and folds RNAs. Because of its extreme time complexity of $O(n^6)$, numerous faster "Sankoff-style" approaches have been suggested. Several such approaches introduce heuristics based on sequence alignment, which compromises the alignment quality for RNAs with sequence identities below 60% [1]. Avoiding such heuristics, as e.g. in LocARNA [4], has been assumed to prohibit time complexities better than $O(n^4)$, which strongly limits large-scale applications.

Results: Breaking this barrier, we introduce SPARSE (Sparse Prediction and Alignment of RNAs using Structure Ensembles), a novel *quadratic time* Sankoff-style approach that does not rely on sequence-based heuristics but employs structural properties of RNA ensembles; its $O(n^2)$ complexity matches the one of sequence alignment. The approach is based on a novel lightweight Sankoff-style alignment model, for which we introduce the algorithm PARSE. For the first time it transfers the Sankoff-model completely to a lightweight energy model; thus, it is more expressive than all previous lightweight methods, which inherit the PM-comp model [2]. In comparison to LocARNA and similar approaches, the novel model enables much stronger sparsification based on the RNA structure ensemble; consequently, SPARSE aligns and folds RNAs with similar alignment and better folding quality in significantly less time. Finally, SPARSE aligns ncRNAs from the challenging low sequence identity region more accurately than tools relying on sequence-based heuristics.

Conclusion: Our results indicate that a complete lightweight Sankoff-style model with stronger sparsification can increase the performance and accuracy of RNA alignment, where the potential of the model points far beyond the studied

⋆ Joint first authors
⋆⋆ Corresponding author; e-mail: backofen@informatik.uni-freiburg.de

prototype. Not falling back on sequence comparison, SPARSE suggests itself for large scale similarity assessment of RNAs with moderate to very low sequence identity.

References

1. Paul P. Gardner, Andreas Wilm, and Stefan Washietl. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res*, 33(8):2433–9, 2005.
2. I. L. Hofacker, S. H. Bernhart, and P. F. Stadler. Alignment of RNA base pairing probability matrices. *Bioinformatics*, 20(14):2222–7, 2004.
3. David Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J Appl Math*, 45(5):810–825, 1985.
4. 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 Comput Biol*, 3(4):e65, 2007.