PAPER: 177
TITLE: GraphClust: alignment-free structural clustering of local RNA secondary structures
AUTHORS: Steffen Heyne, Fabrizio Costa, Dominic Rose and Rolf Backofen

Below we list all comments of the reviewers. We inserted our answers after a comment or question.

---------------------- REVIEW 1 --------------------

The authors present an interesting an useful means of clustering RNA secondary structures using graph kernels. The avoidance of high-quality intermediate alignments finally overcomes a computational bottleneck that so far has hampered the analysis of very large sets of RNA structure predictions. I don't have much to complain about this in general well-written paper, except:

- the priority claim of having the first alignment free method ... is most likely exaggerated: there have been several attempts for alignment-free structure comparison, including methods employing graph descriptors or spectra (Barash, Gan&Schlick), work by Dai 2008, compression based (Liu) a probably several other attempts, admittedly none of them particularly practical or thoroughly evaluated.

ANSWER: The reviewer is right, also in stating that the other papers do not apply the approach on large scale datasets. Hence, we we have reformulated the claim in the Introduction, as:

"In this paper, we propose an efficient alignment-free approach for clustering very large sets of RNA sequences according to sequence and structure information. Although alignment-free sequence-comparison methods have recently become popular in the analysis of large data sets and there are alignment-free tools for comparison of RNA with respect to sequence and structure (e.g.~\cite{Schlick2003, Liu2006}), we are not aware of any alignment-free method capable to perform RNA sequence-structure comparisons on hundred of thousands sequences."

- the terminology "RNA folding hypothesis" is awkward and should be avoided. Some formulations exceed this reviewer's parser capacities altogether: "The first step in order to obtain a vector based representation of a folding hypothesis is to materialize one or more representative structures ..." sound very fancy but I guess just means "sample a few secondary structures that are plausible folds of the input sequence".

ANSWER: The phrase "RNA folding hypothesis" has been replaced with "structure" and the text simplified.

- what does the 2x etc mean in section 4.2? the number of recovered RNAs (not very abundant) or the number of distinct clusters into which they fall?

ANSWER: It's the number of distinct clusters. We accordingly rephrased the particular sentence for a better understanding.

- Why should the emergence of novel structural classes be related to the whole genome duplication?

ANSWER: We do not expect whole genome duplications to yield special novel RNA classes that are absent in non-duplicated genomes. As stated in the Discussion section, in this paper we have only analyzed single genomes, i.e. we have not performed cross-species comparisons, as of yet. Thus, resulting structural cluster are expected to contain paralogous genes. In general, such clusters are easier to identify after duplication events due to a probably increased copy number of RNA genes.

---------------------- REVIEW 2 ---------------------

The authors introduce an efficient pipeline to cluster RNA sequences that does not require preliminary sequence alignments, thus saving significant amounts of time. This research is strongly motivated by our need for high-throughput methods for ncRNA annotation.

Overall, the paper is well written and the work presented is a technical tour de force. The authors demonstrate that, with much less time requirement, the performances of their program are similar to those obtained by a state-of-the-art method (locARNA) already developed in the group.

The presentation of the method is very technical and sometimes hard to read. I believe that some material could be moved to supplementary material, while the author could focus on the big picture.

ANSWER: We have carefully revised sections 2 (approach) and 3 (methods) to improve the readability of the manuscript. We have moved parts (e.g. details on pre-processing or applied parameters) to the supplemental material.

Phase 4 of the pipeline appears to be the bottleneck of the method. More emphasis on this model would help to understand the challenges of the problem. In particular, more (experimental) analysis of the impact of the size of the sample set on the performances of the method would be interesting (?).

ANSWER: We added supplementary Figure S1 to show the runtime required for the important phases of the pipline. We also made additional test for the Rfam and small ncRNA benchmark set with a sample on 50% and 25% during phase 4 (see Table S5). We observe the same overall quality interms of F measure and rand index. However the final number of clusters was reduced slightly. For practical purposes this would mean to run 1-2 additional iterations of the pipline.

The clustering produces different clusters on EvoFam sequences and EvoFold hits. Is there any reason for that?

ANSWER: There are several reasons. First, the input sets to build EvoFam families from EvoFold hits was not exactly the same. For example, the EvoFam authors, contrary to us, discarded all EvoFold hits at protein-coding genes. Next, the repeat masking likely yields different results due to changes in the repeat annotation over time and non-standardized repeat-masking protocols. Furthermore, the EvoFam authors relied on a 41-way genomic vertebrate alignment. Yet, we did not use this rich source of information which allows to select specific, highly conserved and therefore likely functional EvoFold predictions.

Are the remaining cluster (i.e. not annotated by Rfam) relevant?

ANSWER: We are convinced that our novel clusters are biologically relevant (at least to some extent, we of course also generate false positives, i.e. we may bring together structural artifacts from pseudogenes etc.). To further illustrate the relevance of our clusters, we provide a new supplemental figure (S2) in which we relate two of our clusters. In the first cluster we recovered a previously described EvoFam family. The second cluster is of fairly similar quality. According to the predicted secondary structure it may even be regarded as having a higher quality, since it contains several compensatory mutations that support the clustering. This second cluster was, for whatever reason, not reported by EvoFam. Thus, as exemplified in Fig. S3, our approach can not only assist in revising existing clusters, it especially has the potential to identify new ones. We conclude, that our approach identifies relevant candidates and opens the road for subsequent studies, e.g. wet-lab experiments further analyzing these structures.

Is there any structural similarities between the clusters found?

ANSWER: Our post-processing phase in which we apply infernal's cmsearch is actually intended to resolve remaining structural similarities and to obtain a final structural partition of the input data. Nevertheless, structural similarities between identified clusters can easily be visualized and analyzed by a subsequent hierarchical clustering, e.g. using RNAclust, a LocaRNA-based clustering approach, previously developed in our group. For example, when applied to GraphClust-derived clusters from the EvoFam dataset, we can easily detect novel miRNA candidates, see the novel supplemental Figure S3.

The secondary structure encoding is neat. How simpler models would affect the clustering performances?

ANSWER: Due to lack of space, we presented in the paper only the model that achieves optimal results in its discriminative capacity as measured on a small developmental set derived from some RFam families. Simpler models, which are obtained limiting the size and complexity of the fragments used in the graph kernel, achieve lower performance.

Other comments:
* What is the influence (if any) of the window size on the performances?

ANSWER: The window size influences the locality of the structural features that the method is aware of. We chose two window sizes: one of 30 nt and the other of 150nt in order to capture both local hairpins and larger multi-loop structures. Non reported tests confirm that the method presented is rather stable w.r.t. the window size parameter.

* Fig. 1 is not clear and does not help as much as it should to understand the encoding. Overall, I would suggest to improve the description of the structure encoding.

ANSWER: We revised the figure and the caption for a better understanding.

* Fig.2 is too small and not readable.

ANSWER: Figures have been enlarged and the caption text improved.

* font is too small in Section 2.3 & 3, which makes the text hard to read.

ANSWER: This is due to the Bioinformatics style sheet which dictates small font sizes for Method sections. * Why not using Rfam 10 which features a better annotation?

ANSWER: To annotate novel clusters we in fact (as stated in the methods section) already use Rfam v10.1. We deliberately did not update the Rfam benchmark set for reasons of comparability. Using the original unmodified benchmark set from our LocaRNA paper ensures fair benchmarking.

* Section 4.2: how many cluster are predicted if we increase the CSI value?

ANSWER: Supplemental Figure S2 depicts SCI/MPI density heat-maps of GraphClust-generated clusters for the various datasets used in this contribution. From this figure alone it is already clear that we obtain several clusters with SCIs exceeding our SCI threshold of 0.5. Overall (for all datasets listed in Tab.1), we obtain 185 cluster with a SCI > 0.9, for example. If requested, other thresholds than 0.5 can be added to Table 1 or SCI-histograms for each dataset can be added to the supplement.

Typos:
* page 4, 1st paragraph: p' is missing in the definition
* Table 1, footnote: Plase --> Please
* page 8, last paragraph of discussion: pipleine --> pipeline

ANSWER: Thanks for reporting these typos. These, together with other minor errors, have been fixed.

---------------------- REVIEW 3 --------------------
Clustering of local RNA structure is an important problem for finding structural motifs. Current approaches are based on sequence-structure alignment that makes them time consuming. Here the key idea is to decrease the load of alignment. Local RNA structures are represented by Graph kernel, and by using an efficient method for nearest neighbor search candidate clusters are determined. Afterward, in order to incorporate conservation an alignment-based method (LocaRNA) is called to refine the clusters. After removing the found clusters, clustering and refinement steps are repeated. The method works as a nice improvement over previous approach "LocaRNA", by giving LocaRNA a rich candidates of clusters. It results in a slightly better accuracy, but huge speedup. Since the main advantage of the current approach is its speedup, more comprehensive comparison of time and space requirements should be added to the paper. Also, it is would be informative if the time and space requirements are measured for each steps in pipeline for different datasets.

ANSWER: We added supplementary Figure S1 to show the runtime of the different phases of our pipeline for all datasets. We also added a short paragraph to describe the space and memory requirements. Excepts phase 4, a memory limit of 3.5 GB was sufficient. The memory requirement in phase 4 depends directly on the size of the sparse vector. For example, the Rfam benchmark set has a vector of 500MB. Datasets up to 30.000 sequence fragements (~150nt in length) are therefore possible to process on a normal machine (with 4 GB RAM).

Finally, I think it is not fair to call the approach alignment free, as the pipeline uses LocaRNA which is based on alignment.

ANSWER: Please see the answer to reviewer 4 on this issue.

---------------------- REVIEW 4 --------------------

Overall this is a strong paper which was well liked by all reviewers. Please respond to all reviewer comments in the paper.

ANSWER: We acknowledge the helpful comments of the reviewers to improve the quality of the manuscript. We carefully addressed their questions and suggestions in the manuscript.

E.g. it is mentioned that it is not fair to claim that this is the first alignment free method for the problem investigated. In fact it is questionable that the method is alignment free. Please modify the text appropriately.

ANSWER: We have better clarified the intended meaning in the text (in the Introduction and Approach sections), i.e. candidate clusters are computed via a method that has linear complexity which does not use alignment techniques. This allows the structural clustering of hundreds of thousands of sequences, which cannot be achieved with other methods. Finally, we post-process these clusters in parallel: for each of the much smaller (typically 10-20 sequences) candidate clusters we employ an alignment based method to filter inconsistent sequences. Since the post-processing phase is just a refinement step we call the main method "alignment free".