Scalable and accessible clustering of ncRNAs based on sequence and secondary structures

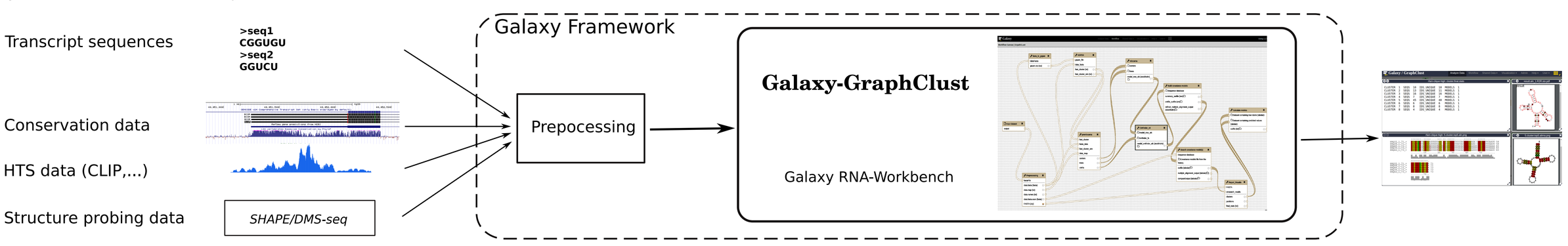
Milad Miladi, Eteri Sokhoyan, Torsten Houwaart, Rolf Backofen and Björn Grüning

Bioinformatics Group, University of Freiburg, Germany

Overview

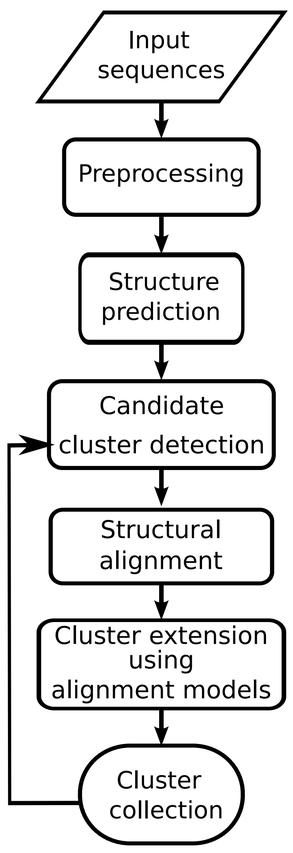
There are many ncRNAs and regulatory elements whose function is still unknown. RNA sequences with putative but unknown functionality can appear for example in genome-wide screens or experiments such as RNA-seq. Clustering of RNA sequences is currently one of the prevalent approaches for detecting and annotating the function of putative ncRNAs and regulatory elements.
Here we present Galaxy-GraphClust, a web-based tool suite for large-scale structural clustering of RNAs based on sequence and structural similarity that is provided via the Galaxy framework.

Galaxy-GraphClust is a realization of GraphClust [1] inside Galaxy framework that drastically simplifies the task of clustering large amounts of RNAs by making it possible to:
a) interactively perform the clustering of RNAs via a web interface,
b) support computations on different back-ends ranging from personal computers to large scale computer clusters,
c) integrate the clustering workflow with high-throughput sequencing (HTS) analysis.



Method

Galaxy-GraphClust can efficiently cluster thousands of RNA sequences according to their sequential and structural information. The method is based on an inverse index utilizing a fast graph kernel-based hashing approach. Individual sequences are folded and the resulting structures are mapped into a high-dimensional vector space. A locality-sensitive hashing technique is then used to identify candidate clusters in linear time. The candidate clusters are then refined in parallel using RNA alignment domain tools. Finally a covariance model is built from each cluster and used to extend clusters by scanning the entire dataset. Alignment of each cluster is visualized and accompanied by consensus secondary structure of the motif with various annotations. Motif positions and scores in the input sequences are reported for further downstream analysis.



Utilizing structure probing data

RNA structure probing experiments such as SHAPE are appealing techniques for determining the base pairing state of each nucleotide. Galaxy-GraphClust is capable of incorporating probing data to assist structure prediction and improve secondary structure graph features. For demonstrating the applicability and as a use case we used the simulated data from ProbeAlign benchmark [2]. Rfam sequences and corresponding SHAPE data of families with at least 10 sequences were extracted. Adjusted Rand Index used to evaluate the clustering quality performance.

Cluster Rounds	Model	
	Free-energy	Free-energy + simSHAPE
1	0.54	0.63
2	0.77	0.82
3	0.78	0.81

Table 1: Adjusted Rand Index evaluation on clustering ProbeAlign dataset

Discovering RBP binding site motifs

Structure from clustering eCLIP data:

Rfam structure entry for RF00032:

result

Table 2: Rscape-R2R output of the clustering vs. Rfam structure entry for Histone 3UTR SL family

Stem-loop binding protein (SLBP) is a well characterized RNA binding protein (RBP) that binds to a structural motif in Histone mRNAs. As a use case we extracted target binding sites of the human SLBP eCLIP experiment from *Nostrand et al* [3]. Peaks with two different log2-fold-change thresholds of 4 and 5 were extracted. To diminish the chance of missing the binding motif, the peak regions were extended 60 nucleotides both up- and downstream. The sequences of the resulting 306 and 3171 peak regions were used as input for the GraphClust workflow. These steps can be done within the Galaxy framework. In both cases the cluster with the highest cardinality and reliable alignment matched to the Rfam reference structure for Histone 3' UTR stem-loop. The agreement between small and large input sets demonstrates Galaxy-GraphClust resilience to the noise and scalability.

Accessible interface through Galaxy

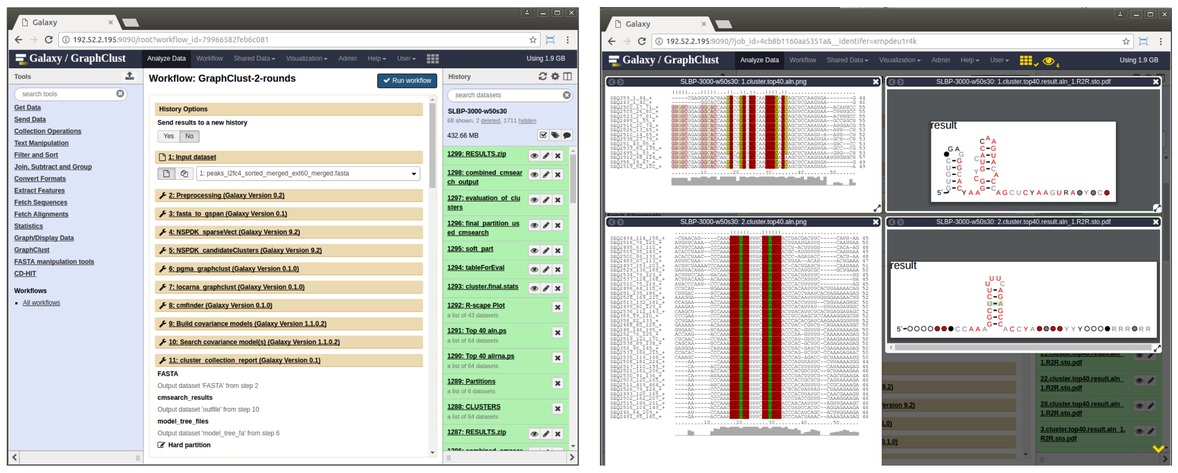


Figure 1: Input and configuration

Table 3: Output example

Conclusion

To the best of our knowledge this is the only available web-based RNA clustering tool that is capable of clustering very large numbers of sequences according to their sequence and structural similarities in an interactive and configurable fashion. The applicability of the tool has been demonstrated for predicting conserved structural motifs under the presence of noisy and unrelated sequences or long surrounding contexts. Furthermore the SLBP use case highlights the benefits of the Galaxy integration, by enabling the combination of an HTS analysis directly with an RNA-clustering workflow.

References

[1] Heyne et al., GraphClust: alignment-free structural clustering of local RNA secondary structures. *Bioinfomatics*, 2012.

[2] Ge et al., ProbeAlign: incorporating high-throughput sequencing-based structure probing information into ncRNA homology search. *BMC Bioinformatics*, 2014,

[3] Nostrand et al., Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP).. *Nature methods*, 2016.

Availability

Repository: https://github.com/BackofenLab/docker-Galaxy-GraphClust
Contact: {miladim, gruening}@cs.uni-freiburg.de

Funding

This work was supported by the Deutsche Forschungsgemeinschaft(DFG) and German Federal Ministry of Education and Research (BMBF).