`RNAclust.pl` Documentation

Jan Engelhardt[1], Steffen Heyne[2], Sebastian Will[2], Kristin Reiche[3]

[1] *Bioinformatics Group, Department of Computer Science, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*
[2] *Bioinformatics Group, Institute of Computer Science, University of Freiburg, Freiburg, Germany*
[3] *RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology, Perlickstraß 1, D-04103 Leipzig, Germany*

July 29, 2010

Abstract

`RNAclust.pl` is a perl script summarizing all the single steps required for clustering of structured RNA motifs, i.e. identifying groups of RNA sequences sharing a secondary structure motif. It requires as input a multiple FASTA file. In the first step for each input sequence the *base pair probability matrix* of its secondary structure distribution is calculated (using RNAfold from the Vienna RNA package). Secondly, for each pair of base pair probability matrices a *sequence-structure alignment* is calculated using `LocARNA`. Lastly, a hierarchical cluster-tree (in NEWICK format) is derived by *WPGMA clustering* of the pairwise alignment distances.

The calculation of all pairwise sequence-structure alignments is the bottleneck of this pipeline, although comparable fast in case `LocARNA` is used. Hence, `RNAclust.pl` provides the possibility to distribute the calculation of all $\frac{N(N-1)}{2}$ pairwise alignments, with N being the number of input RNA sequences, between different CPUs on one machine (see `--cpu` option). Furthermore, by using the `--start` and `--end` options the calculation of the pairwise alignments can be distributed among different machines.

As the post-processing of a large tree is problematic, you may use `RNAclust.pl --rnasoup` in order to derive those subtrees which are

likely to define a distinct structural motif. By using `--rnasoup` it will run automatically during the clustering process.

1 Invocation

`RNAclust.pl` may be invoked in two different modes. The first mode assumes that all `LocARNA` pairwise alignments are computed on one machine. The second mode distributes the computation of the pairwise `LocARNA` alignments among different machines, thus saving computation time.

1.1 Pairwise alignments on one machine

This mode of `RNAclust.pl` is suitable for a small number of input RNA sequences (usually less than 1000). Fig. 1 outlines the invocation of `RNAclust.pl` in case all `LocARNA` alignments are calculated on one machine.

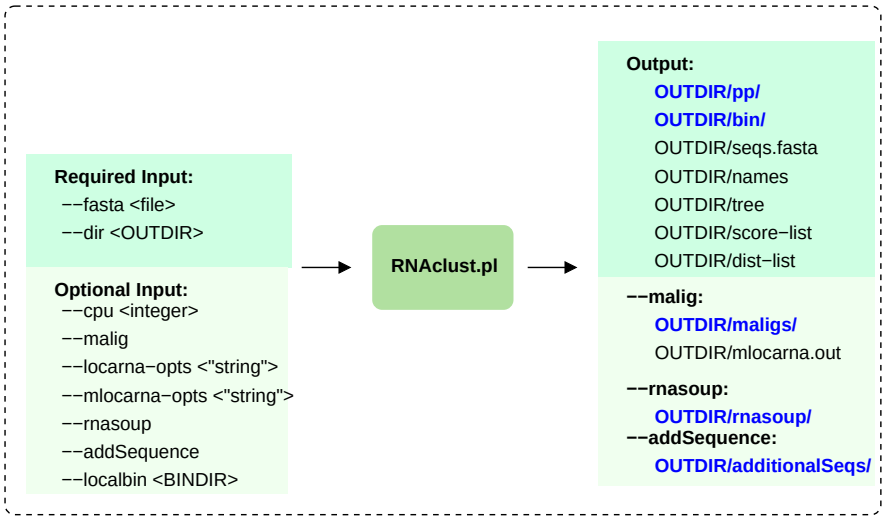

Figure 1: Output of RNAclust computed on one machine.

If particular parameters should be passed to `LocARNA` (computes pairwise alignments in order to built distance matrix) and/or `mlocarna` (computes multiple alignments for each subtree in final cluster-tree), please use options `--locarna-opts` and `--mlocarna-opts`, respectively.

1.2 Pairwise alignments distributed among different machines

This mode is recommended in case the input file contains a large number of RNA sequences. The calculation of pairwise alignments is distributed on different machines. This is realized by invoking `RNAclust.pl` with different parameter settings (Fig. 2).

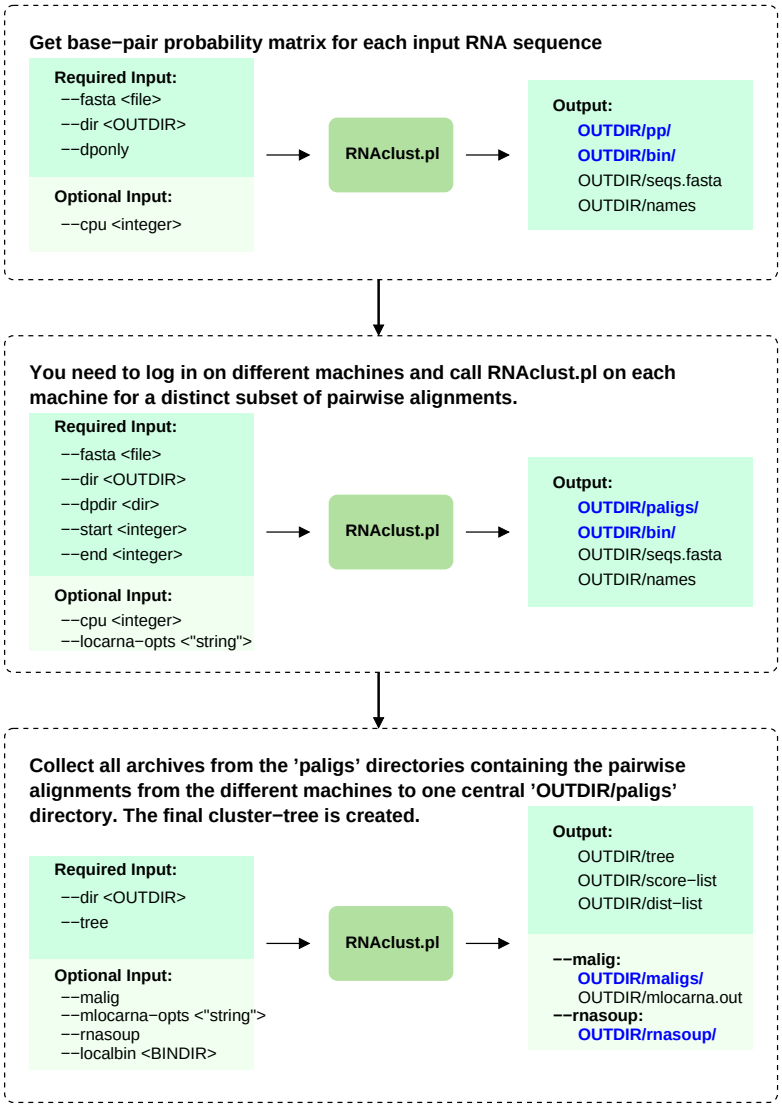

Figure 2: Output of RNAclust computed on different machines.

The first invocation creates for each input sequence the base pair probability matrix. This step is realized on one single machine. The second invocation takes as input the base pair probability matrices and creates subsets of pairwise alignments on different machines. The third invocation requires as input all subsets of pairwise alignments in order to calculate the final cluster-tree, again on one single machine.
You may also skip the first call of `RNAclust.pl` and create the base pair probability matrices for all input sequences on all machines. However, this increases the computation time in case many input sequences are given.

1.3 Identifying the number of clusters

A hierarchical cluster-tree as computed by WPGMA reflects the structural similarities between the RNA input sequences. Usually this tree is large and it is hard to identify relevant groups of RNAs sharing similar secondary structures. For this purpose we provide the `--rnasoup`/`--rnasoup-only` options. The optimal partition (i.e. the optimal number of clusters/groups) is identified by using a decision rule that has been introduced by Duda and Heart [1]. Instead of evaluating the squared error of the pairwise distances we evaluate the squared error from the minimum free energies of the single sequences to the minimum free energy of the consensus secondary structure [3].

1.3.1 Theoretical Background

`RNAclust.pl --rnasoup` retrieves the optimal number of clusters by using a modification of the Duda and Heart rule [1]. Instead of evaluating the squared error of the pairwise distances `RNAclust.pl --rnasoup` evaluates the squared error from the minimum free energies of the single sequences (E_i) to the minimum free energy of the consensus secondary structure (E_{cons_j}) [3]. If for an internal node C with children C_1 and C_2 the increase of the squared error is unexpectedly large the hypotheses that C forms one group is discarded and the substrees C_1 and C_2 are reported as unique RNA groups at significance level k.

The squared error for the hypothesis that C forms one group is defined as

$$J_e(1) = \sum_{i=1}^{N} (E_i - E_{cons})^2 \ . \tag{1}$$

The squared error for the hypothesis that C should rather be splitted into two groups defined by its children is given as

$$J_e(2) = \sum_{j=1}^{2} \sum_{i=1}^{N_j} (E_i - E_{cons_j})^2 \ . \tag{2}$$

The null hypothesis that C is one group is rejected in case the ratio of $J_e(2)$ and $J_e(1)$ is smaller than a predefined critical value:

$$\frac{J_e(2)}{J_e(1)} < 1 - \frac{2}{\pi} - k\sqrt{\frac{2 - \frac{16}{\pi^2}}{N}} \ . \tag{3}$$

k reflects the significance of the decision. The larger k the larger difference of squared error must be before the null hypothesis is rejected. I.e. with small k the rule tends to spot small differences and reports groups containing rather few sequences, while large k reports larger groups.

This rule allows to directly incorporate biological features for ncRNAs into the decision process. NcRNA families are defined by structural similarities and hence the free energies of the single secondary structures should be similar to the free energy of the consensus structure (following the idea of the structure conservation index [2, 4]). The consensus structure as calculated by `RNAalifold` is the secondary structure all aligned sequences simultaneously fold into. In case they do not share the same secondary structure no base pairs are reported in their consensus structure.

A test of this adapted Duda rule on a `LocARNA`-based cluster-tree of 3901 `RFAM`-sequences [5] resulted in a Matthews correlation coefficient (MCC) of 0.8 for $0.8 \leq k \leq 1.2$.

1.4 Semi-automatic group finding

Beside the full automatic approach followed by `RNAclust.pl --rnasoup` one may favourite a method where the user is able to infer in the process of group-

finding. `RNAclust.pl --rnasoup` outputs a complete partition of the input cluster-tree at different significance levels that can be easily analysed with a viewer written by Jan Engelhardt (`http://www.bioinf.uni-leipzig.de/~jane/software/soupviewer/manual.php`). The viewer provides information about the structural conservation, the secondary structure plot as well as the multiple alignment for each internal node. This viewer enables the user to refine the outcome of `RNAclust.pl`.

1.5 Contents of the output directory

`pp/`	Directory containing base pair probability matrices for each input RNA sequence. Naming convention: Increasing integer numbers corresponding to position in input FASTA file.
`bin/`	Directory containing local binaries of all needed tools. Has the advantage that script is independent of network access in case of long-time runs.
`paligs/`	Directory containing the pairwise `LocARNA` alignments in zipped archives.
`names`	File containing only the names of the input sequences.
`seqs.fasta`	Local copy of the input FASTA file.
`score-list`	Pairwise `LocARNA` score list. The first two columns are the indices of the input sequences which correspond to the sequence at position x in `seqs.fasta` and `names`.
`dist-list`	Pairwise distance list, retrieved from the `LocARNA` score: $distance(i,j) = max(0, q - score(i,j))$, where q is the 99%-quantile of all pairwise scores.
`tree`	Final cluster-tree.

Output in case `--malig` option is used

`maligs/`	Directory containing for each internal node of the cluster-tree a multiple alignment created by `mlocarna`.
`mlocarna.out`	Output of `mlocarna`

Output in case `--rnasoup` option is used

`rnasoup/`	Directory containing the following files:
`tree_xml`	An XML-file which represents the hierarchical cluster tree but contains additional informations. The structure conservation index, the minimum free enerygy as well as the `RNAclust.pl --rnasoup` group predictions are stored there.
`LOG`	Contains some status messages of `RNAclust.pl --rnasoup` including error messages if there are any.
`mfe_consensus.txt`	The STDOUT messages of mlocarna are saved here.

Output in case `--addSequence` option is used

`additionalSeqs/`	Directory containing the same data like the normal output directory but for the new calculated tree consisting of an original tree and some additional added sequences.
`addSeqs.fasta`	Local copy of the additional sequences in FASTA format.
`allSeqs.fasta`	Local copy of the original input Fasta file and the additional sequences.

1.6 Required third-party software

`RNAalifold`	Part of the `Vienna RNA Package`; Computes the minimum free energy consensus secondary structure of an alignment. `http://www.tbi.univie.ac.at/~ivo/RNA/`
`RNAfold`	Part of the `Vienna RNA Package`; Computes the minimum free energy secondary structure of a single RNA sequence. `http://www.tbi.univie.ac.at/~ivo/RNA/`
`njplot`	Tree viewer. Not required but might be useful. `http://pbil.univ-lyon1.fr/software/njplot.html`
`LocARNA`	LocARNA is a tool for producing fast and high-quality pairwise and multiple alignment of RNA sequences. LocARNA 1.5.2 or higher is required. `http://www.bioinf.uni-freiburg.de/Software/LocARNA/`

2 Bug Reports

Please send any bugs you encounter to `jane@bioinf.uni-leipzig.de` or `kristin.reiche@izi.fraunhofer.de`.

References

[1] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2001.

[2] A. R. Gruber, S. H. Bernhart, I. L. Hofacker, and S. Washietl. Strategies for measuring evolutionary conservation of RNA secondary structures. *BMC Bioinformatics*, Feb 26:9:122, 2008.

[3] B. Kaczkowski, E. Torarinsson, K. Reiche, J. H. Havgaard, P. F. Stadler, and J. Gorodkin. Structural profiles of human mirna families from pairwise clustering. *Bioinformatics*, 25(3):291–294, Feb 2009.

[4] S. Washietl, I. L. Hofacker, and P. F. Stadler. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A*, 102(7):2454–2459, Feb 2005.

[5] S. Will, K. Reiche, I. L. Hofacker, P. F. Stadler, and R. Backofen. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol*, 3(4):e65, Apr 2007.