

# RNAclust.pl Documentation

Kristin Reiche

*Fraunhofer Institute for Cell Therapy und Immunology,  
Perlickstr. 1, D-04103 Leipzig, Germany*

November 17, 2008

## 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 RNAsoup (<http://www.bioinf.uni-leipzig.de/~kristin/Software/RNAsoup/>) in order to derive those subtrees which are likely to define a distinct structural motif.

# 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 one 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.

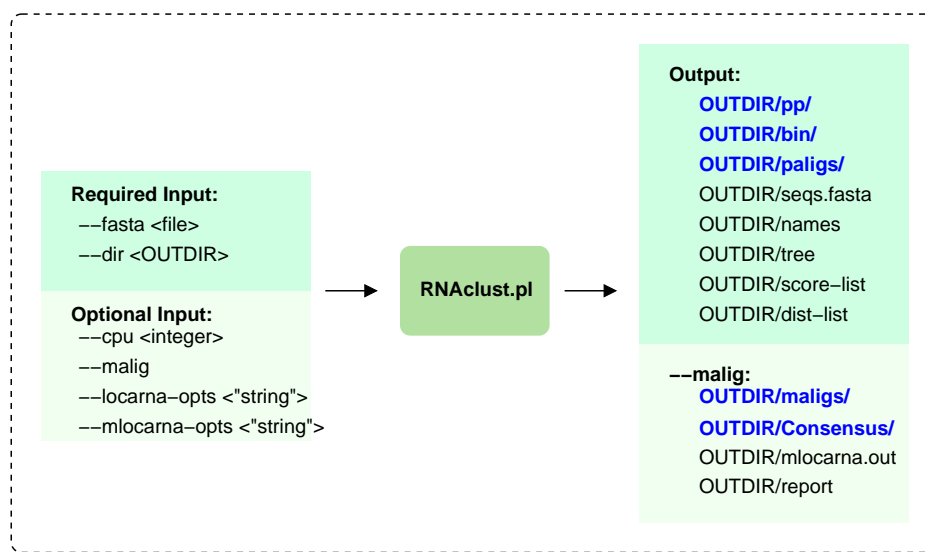


Figure 1:

If particular parameters should be passed to `LocARNA` (computes pairwise alignments in order to build 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).

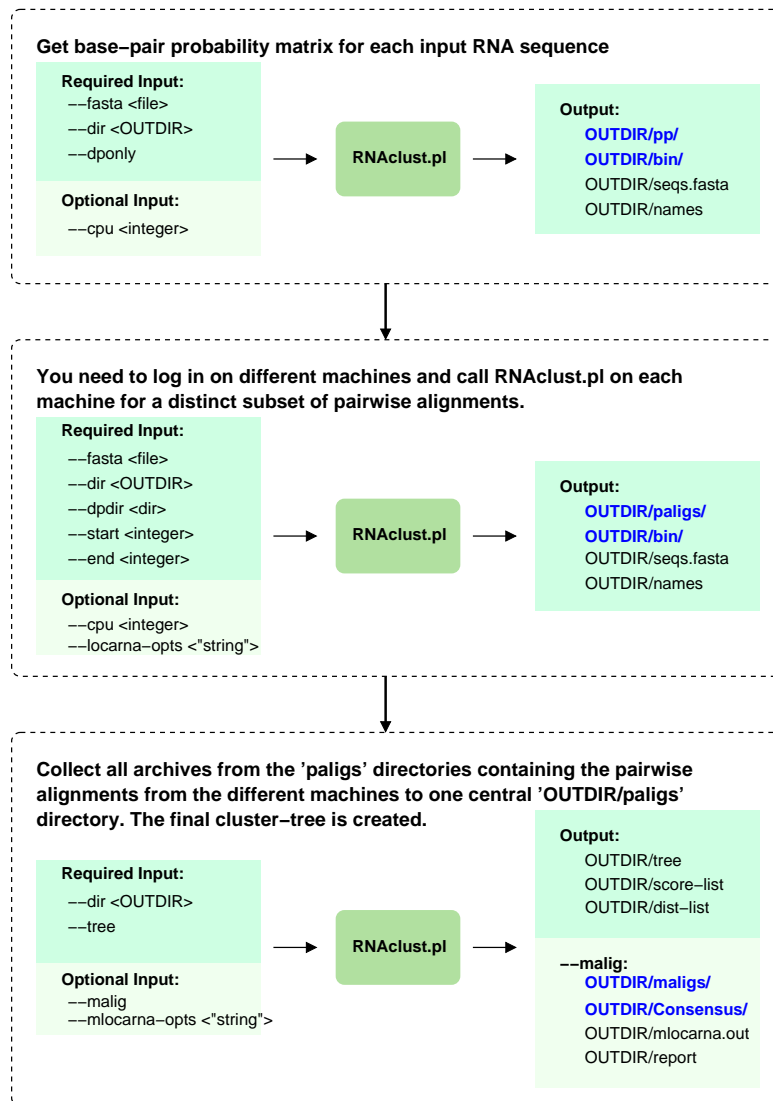


Figure 2:

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 Contents of the output directory

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

#### Output in case `--malig` option is used

<code>maligs/</code>	Directory containing for each internal node of the cluster-tree a multiple alignment created by <code>mlocarna</code> .
----------------------	---

<code>Consensus/</code>	Directory containing for each internal node the consensus secondary structure plot.
<code>mlocarna.out</code>	Output of <code>mlocarna</code>
<code>report</code>	Report which states for each subtree of the cluster-tree (in the following order) the number of leaves, the <code>mlocarna</code> score, the mean pairwise identity, the structure conservation index, the number of loops observed in the consensus secondary structure, the minimum free energy of the consensus secondary structure, and the length of the consensus structure. This report may help by identifying interesting subtrees. However, we recommend to use <code>RNAsoup</code> in order to retrieve subtrees from the cluster-tree which are likely to share the same secondary structure motif.

## 1.4 Required third-party software

<code>RNAalifold</code>	Part of the Vienna RNA Package; Computes the minimum free energy consensus secondary structure of an alignment. <a href="http://www.tbi.univie.ac.at/~ivo/RNA/">http://www.tbi.univie.ac.at/~ivo/RNA/</a>
<code>RNAfold</code>	Part of the Vienna RNA Package; Computes the minimum free energy secondary structure of a single RNA sequence. <a href="http://www.tbi.univie.ac.at/~ivo/RNA/">http://www.tbi.univie.ac.at/~ivo/RNA/</a>
<code>njplot</code>	Tree viewer. Not required but might be useful. <a href="http://pbil.univ-lyon1.fr/software/njplot.html">http://pbil.univ-lyon1.fr/software/njplot.html</a>
<code>RNAz</code>	Evaluates if a multiple alignment contains a conserved secondary structure. Used for the final report for all possible subtrees. <a href="http://www.tbi.univie.ac.at/~wash/RNAz/">http://www.tbi.univie.ac.at/~wash/RNAz/</a>

## 1.5 Post-processing of the cluster-tree

A hierarchical cluster-tree 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 `RNAsoup` (<http://www.bioinf.uni-leipzig.de/>

`~kristin/Software/RNAsoup/`), which identifies subtrees which are likely to contain RNA sequences of similar secondary structures. **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.

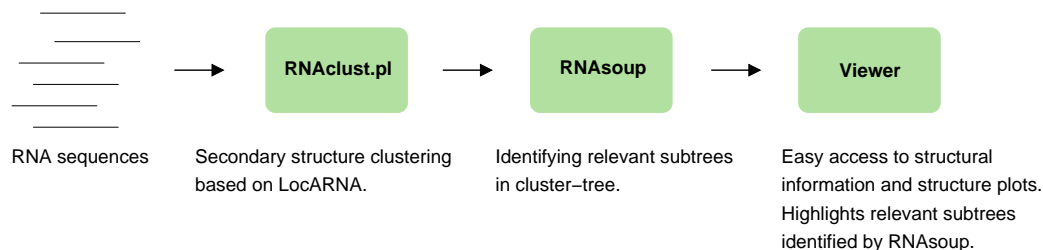


Figure 3:

## 2 Bug Reports

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