RNAsoup Documentation

Kristin Reiche

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

November 11, 2008

Abstract

RNAsoup (**Spot grOUPs in RNA cluster-tree**) is a post-processing tool of a structural clustering pipeline for **structured RNAs**. It requires as input a binary cluster-tree, the minimum free energy (MFE) of the consensus secondary structure for each internal node, and a FASTA file of the input sequences. It detects the optimal partition (i.e. finding the optimal number of clusters/groups) into distinct subtrees where each subtree contains structurally related RNA sequences. RNAsoup is based on a decision rule introduced by Duda and Heart [1]. Instead of evaluating the squared error of the pairwise distances RNAsoup evaluates the squared error from the minimum free energies of the single sequences to the minimum free energy of the consensus secondary structure.

1 Invocation

A shell script (`runRNAsoup.sh`) is available which prepares the input for RNAsoup, and lastly calls RNAsoup. This separates the computationally expensive step to calculate multiple alignments for each internal node from the fast step to identify the groups in the cluster-tree. Once the alignments are available RNAsoup can easily be invoked for different significance levels with-

out the need of retrieving the alignments and minimum free energies of the consensus secondary structures a second time.

1.1 Invocation of the Shell Script

```
sh runRNAsoup.sh <source-directory> <target-directory>
```

The source-directory must contain:

`seqs.fasta`	Input sequences in FASTA format. Each sequence must be given on one line and not be splitted over several lines.
`tree`	Hierarchical cluster tree in NEWICK format.

The target-directory will contain:

`aligs/`	Directory containing for each internal node of the cluster-tree a multiple alignment (PS and CLUSTALW) as well as the secondary structure plot (PS).
`partitions/`	Directory containing for a predefined set of significance levels k the optimal partition of the cluster-tree.
`partitions/partition*.txt`	Files containing the partitions of the cluster-tree for different significance levels k.
`partitions/tree`	Hierarchical cluster tree with additional information (NEWICK).
`mfe_consensus.txt`	File created by `rnasoup_consMFE.pl`. Reports the alignment and the MFE of the consensus secondary structure.
`mlocarna.out`	Output of `mlocarna`
`LOG`	A log file

The leave names in the input tree must not contain the characters {$(), ; :$**} and the tree must terminate with ';'. A sequence in** `seqs.fasta` **must not be given on separate lines.**

partitions/tree is identical to the input tree, except that the ID of the corresponding multiple alignment found in aligs/ is added for each internal node. If bootstrap values are enabled in the tree-viewer njplot those IDs occur at the branching points of the internal nodes enabling you to find easily the corresponding alignment and secondary structure plot.

Format of files partitions/partition_k*.txt:

```
============ Node 1 ==============                Node ID
No. leaves:  7                                    Number of leaves
consmfe:     -32.27                               Minimum free energy of the
                                                  consensus secondary structure
sci:         0.786251                             Structure conservation index
Locarna:     RNAsoup_out/aligs/intermediate6.aln  Relative path to multiple
                                                  sequence-structure alignment
Leaves:                                           List of leave names
leaf ID10_AC010675.6/79489-79378-ID_mir-395
leaf ID8_AC010675.6/83269-83368-ID_mir-395
leaf ID6_AC005508.1/72384-72483-ID_mir-395
leaf ID7_AC005508.1/71201-71109-ID_mir-395
leaf ID9_AL731607.3/16000-16087-ID_mir-395
leaf ID4_AL606645.2/172471-172383-ID_mir-395
leaf ID5_AL606645.2/171779-171696-ID_mir-395
Left child:          2                            ID of left child
Right child:         9                            ID of right child
==================================
```

1.2 Stand-alone Invocation of RNAsoup

```
RNAsoup [-t tree] [-f fasta] [-m mfe_consensus] [-o outdir] [-k num] [-h] [-v]

-t file     Tree in NEWICK format
-f file     FASTA file of all sequences in tree
-m file     File containing the RNAalifold consensus MFE for each subtree
-o dir      Output directory which is created to store the output
```

```
-k float    Significance level k
-h          Show this help message
-v          Print version information
```

If k is not given `RNAsoup` outputs for a predefined set of significance levels the identified groups (see directory `partitions/`).

1.3 Format of mfe_consensus.txt

Usually you do not need to create `mfe_consensus.txt` by yourself. Use `rnasoup_consMFE.pl` instead. However, here is the format:

```
>path_to_alignment_file
n: number_of_sequences_in_alignment
list_of_sequence_names
mfe: mfe_of_consensus
```

See `examples/RNAsoup_out/mfe_consensus.txt` for an example. The sequence names must be equal to the names in `seqs.fasta` and be given on one line.

1.4 Required third-party software

`mlocarna`	Traverses the tree and builds for each node the multiple alignment progressively `http://www.bioinf.uni-freiburg.de/Software/LocARNA/`
`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/`
`coloraln.pl`	Part of the `Vienna RNA Package`
`njplot`	Tree viewer. Not required but might be useful. `http://pbil.univ-lyon1.fr/software/njplot.html`

2 Semi-automatic group finding

Beside the full automatic approach followed by `RNAsoup` one may favourite a method where the user is able to infer in the process of group-finding. For this purpose a tree viewer (`SoupViewer`) has been developed which highlights the subtrees, which are likely to form a separate group, and, additionally, provides an easy access to secondary structure plots, alignment plots as well as structure conservation information for each subtree. This enables the user in an easy way to refine the outcome of `RNAsoup`.

2.1 Download - `SoupViewer`

`http://www.bioinf.uni-leipzig.de/~jane/software/soupviewer/manual.php`

3 Theoretical Background

`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 `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}). 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, 3]). 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 [4] resulted in a Matthews correlation coefficient (MCC) of 0.8 for $0.8 \leq k \leq 1.2$.

4 Bug Reports

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

[4] 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.