Roman R Stocsits

RNAsalsa - A Manual

manual version 1.4.2 - February, 2009

This manual documents version 0.8.1 of the `RNAsalsa` software.

Abstract

The software package `RNAsalsa` serves as a framework to perform improved RNA structure predictions as well as alignments of structural RNA sequences. It utilizes prior knowledge about structural patterns, adapted constraint directed thermodynamic folding algorithms, and comparative evidence methods. It automatically and simultaneously generates both highly corrected individual secondary structure predictions within a set of homologous RNA genes and a consensus structure for the set, and it takes then sequence and structure information into account as part of the alignment's scoring function. `RNAsalsa` uses structure information for adjusting and refining the sequence alignment and vice versa.

Phylogeny reconstruction tasks are the main group of applications for `RNAsalsa`; methods can be extended to RNA secondary structure, incorporating both the `RNAsalsa` derived (and mostly improved) models for the secondary structure elements of the RNA and also the alignment based on that extended structure information. Thus, it is possible to make the slower evolution of the structural features a reproducible source of information.

The current version of `RNAsalsa` is command line based and available either as source code, or as a pre-compiled executable for various common operating systems.

1 Installation

The software may be downloaded from

`www.rnasalsa.zfmk.de` or
`www.bioinf.uni-leipzig.de/Software/RNAsalsa`

For installing from source code unpack and compile the downloaded `.tar.gz` or `.zip` file. For instance, in the case of Linux based operating systems, just type

```
tar zxvf RNAsalsa_xxx.tar.gz
cd RNAsalsa_xxx
make
```

That's all. Note, that you need an adequate compiler and the `make` tool pre-installed. Downloadable pre-compiled binaries for Win32 and MacOS are ready for usage.

2 The input requirements ...

1. a known structure constraint that can serve as an external source of structure information and as a fundament of the structure building process.

2. an input alignment of your data. The sequence underlying the structure constraint must be part of the alignment.

... and in detail:

- The input alignment is a standard `FASTA` file. Please note that all sequences within the data set must have an unique initial identifier. Thus, two sequences named '`> Animal 1 (the cute one)`' and '`> Animal 2 (the ugly one)`' will give an error, because the determining initial identifier '`Animal`' is the same for both. Additionally, taxon names must not contain any characters that are not covered by the alphanumeric ASCII table.

- The constraint file is a text file consisting of exactly 3 lines: the name, the nucleotide sequence, and the structure encoding string in dot-bracket format. The dot-bracket conventions mean that each base pair is indicated by a left-handed bracket and a corresponding right-handed bracket, a dot is not obligatory structural but might be, and 'x' means that the considered nucleotide must never be paired. All structure encoding characters lie exactly under the sequence position they encode (in other words, there cannot be any line breaks within either the sequence or the structure string).

3 The structural constraint file

An arbitrary example of a functional constraint text file looks like the following:

```
my_constraint_sequence
AGGCCUAUGCAAACCGUUUGCGGGACGGCU
...((....(((....)).)..))..xx..
```

Please note that the file must consist of **exactly 3 lines**, regardless of the length of the sequence, and that the lines with sequence and structure information must match each other. There must be not any further line breaks or empty lines. Further, the sequence which defines the structure constraint, must be part of your initial input FASTA alignment.

4 The set of switches and options

Typing just 'RNAsalsa' or 'RNAsalsa.exe' (on MS Windows systems) without any further input will display a short help:

```
Usage: RNAsalsa [-h] [-v] [-X] [-p] [-s1] [-s2] [-s3] -c FILE -i FILE
   or: RNAsalsa [-h] [-v] [-s3] -a FILE

-c FILE  Necessary switch and input file containing the constraint.
             1st line name, 2nd line sequence, 3rd line structure.
-i FILE  Necessary switch and input alignment file.
             Input alignment in CLUSTALW or FASTA format.
-p      Switch OFF all PostScript output of structures (default is ON).
-s1      Stringency setting for initial constraint weakening.
-s2      Stringency setting for merging of pairwise consensus structures.
-s3      Stringency setting for building the final consensus structure.
             Stringency values lie between 0.0 (lowest) and 1.0 (highest).
-X       All pairwise alignments from input, no internal sequence alignments.
-a FILE  Re-read existing RNAsalsa output, do multiple alignment only.
             Folding output file format must equal 'SALSA_fold_results.txt'.
-h       Show this help message and exit.
-v       Show version information and exit.
```

5 The process of RNAsalsa – some more details

A typical RNAsalsa run using the default instruction set is started by

```
RNAsalsa -i <input_alignment> -c <constraint_file>
```

RNAsalsa generates graphical PostScript output of all individual structure predictions and the adapted/extended constraint for each sequence in the data set, and the consensus structure based on the final alignment. By adding the switch -p to the RNAsalsa command line you can inhibit the generation of those graphics.

The switches s1, s2, and s3 are very important control parameters with high impact on the RNAsalsa run and your results. Therefore, the setting of those -s switches should be done always with caution and might need some testing by trial and error. Virtually, in most cases it will be possible to improve the results by specific adaption of the -s settings for a certain data set (compared to the defaults settings). There are no common rules for the stringency settings, their meaning and impact will always depend on the analyzed data. The default settings for the three switches s1, s2, and s3 are 0.6.

The switch -s1 becomes operative during the initial adaptation (weakening) of the constraint. If a certain region of the input alignment is covered by initial structural constraints, then 60 percent of the alignment must be in a condition to fulfill the constraint (in case the default setting of 0.6 is not changed). Note that base pairs are always handled as an entity, and therefore one position always influences the corresponding second position.

The switch `-s2` defines the stringency in a similar way when `RNAsalsa` merges one certain sequence's subset within the set of all possible pairwise alignment foldings of the data to one definite structure model. That model will then serve as an individual constraint for the subsequent thermodynamical folding.

The switch `-s3` finally is operative as a stringency factor also in a similar way during the final calculation of the consensus structure of the complete data set based on the final structure guided alignment.

For all `-s` switches is valid that they can be set to anything between 0 and 1 in 100 steps (2 post decimal positions). If the value is set below (or equals) 0.5 then it might happen that conflicts between equally scored base pairs cannot be solved. In that case `RNAsalsa` posts a warning and the resulting `RNAsalsa` calculations need suspiciousness. Delicate tasks should be repeated with changed `-s` settings.

However, the `-s` values should be in almost all cases higher than 0.5 (special applications might be an exception).

When the parameter `-X` is set, then all pairwise alignments for each sequence are extracted from the input alignment and no internal alignments are done. In some cases this may improve the results, but note that any initial alignment deficiencies cannot be overruled anymore and will lead to systematic errors throughout the complete `RNAsalsa` run.

Sometimes it may be necessary to repeat the final structure guided alignment and the generation of the consensus, e.g. with different `-s3` settings. The switch `-a` allows that by re-reading the folding output of a former run and restarting those final steps.

The switches `-h` and `-v` give a short help and the version information.

6 What else is worth knowing?

The answers to forthcoming *Frequently Asked Questions*.

1. All alignments during an `RNAsalsa` run are calculated by dynamic programming and use affine gap penalties.
2. Thermodynamic foldings are minimum free energy driven.
3. The folding algorithm is taken from the `Vienna RNA package (RNAfold)`, also is the `PostScript` output routine.
4. Constraints are fulfilled as long as they are thermodynamically possible.
5. Memory consumption is highest during the optimization and merging steps that lead to the individual constraints.
6. The internal pairwise folding steps need the most time.
7. For bug reports, eulogies, and for the purpose of communicating funny experiences please contact: `rs@uni-bonn.de` or `RNAsalsa@gmail.com`

7 The output of RNAsalsa

RNAsalsa produces various (that is to say lots of) output files:

- `SALSA_structaln_sequ.aln`

  is the final structure guided multiple alignment in `clustalw` format.

- `SALSA_structaln_struct.aln`

  is the corresponding alignment file for the same sequences containing the structure (dot-bracket) strings instead of letters.

- `SALSA_structaln_comb_typeA.fas`

  is the corresponding alignment file for the same sequences containing **both** the sequence and the structure (dot-bracket) strings one superimposed on the other.

- `SALSA_structaln_comb_typeB.fas`

  is again the same corresponding alignment file containing both the sequence and the structure strings one superimposed on the other. The difference is an additional ID line above the structure string; this might improve the data compatibility with other software in some cases.

- `SALSA_used_constr.txt`

  is a compilation of all adapted individual constraints that were used to intialize and guide the folding process.

- `SALSA_fold_results.txt`

  is the compilation of all individual thermodynamic foldings in dot-bracket and the calculated minimum free energies.

- `SALSA_consensus.txt`

  contains the consensus structure string for the multiple `RNAsalsa` alignment.

- `SALSA_guide_tree.txt`

  gives an overview about the guide tree that was used internally for the multiple structure based alignment.

- `SALSA_consensus_ss.ps`

  is a PostScript representation of the consensus structure.

- `CONS_*_ss.ps`

  is the file name template for a lot of files containing graphical representations of the used individual constraints.

- `STRUC_*_ss.ps`

  is the file name template for all graphical representations of the individual folding results.

- `SALSA_weakened_constr.txt`

  is the temporary result after the initial constraint has underwent first adaptive steps, e.g. after the switch `-s1` has taken effect. This file can be helpful for troubleshooting, especially to check if the constraint information is lost due to a faulty setup of the `RNAsalsa` run. Such a faulty or unsuitable setup often leads to an empty constraint string, namely all the brackets in the initial constraint are lost, and only dots (this means **no** constraint) remain for further folding processes.

8 Appendix

alignment parameters			stringency setting defaults	
match score	10		switch -s1	0.6
mismatch score	0		switch -s2	0.6
gap opening penalty	-11		switch -s3	0.6
gap extension penalty	-3			

Table 1: Some parameter values for matches, mismatches and gap penalties in `RNAsalsa` alignments. Default stringency values for secondary structure adoption may be adapted by the user as well for individual constraints as for the final consensus structure: Optional switch $s1$: minimum frequency of base pairing occurrence in the first constraint adaptation. Optional switch $s2$: the stringency setting for the majority voting procedure to obtain an individual constraint by the fusion of pairwise alignment folding results. Optional switch $s3$: stringency settings for the final consensus structure extraction process.