R-scape User's Guide

RNA Significant Covariation Above Phylogenetic Expectation

http://selab.org/
Version 0.1; FEB 2016

Elena Rivas
elenarivas@fas.harvard.edu
Department of Molecular and Celullar Biology
Harvard University
16 Divinity Avenue
Cambridge MA 02138 USA
http://eddylab.org/

R-scape

Contents

1 Introduction

R-scape (RNA Significant Covariation Above Phylogenetic Expectation) is a program that given a multiple sequence alignment (MSA) of RNA sequences, it finds the pairs of positions that show a pattern of significant covariation. Each covariation score has an E-value associated to it. E-values are determined using a null model of covariation due to phylogeny but independent of any structural constraints.

How to avoid reading this manual

- Follow the quick installation instructions on page 4.
- Go to the tutorial section on page 6, which walks you through some examples of using R-scape on real data.

Everything else, you can read later.

How do I cite R-scape?

Rivas, E. *et al.*, *"A statistical test of RNA base pair covariation applied to proposed lncRNA structures"*, March 2016, submitted.

2 Installation

Quick installation instructions

Download `R-scape_version.tar.gz` from `http://selab.org/`, or from `hhps://hithub.com/EddyRivasLab/R-scape`; unpack it, configure, and make:

```
> tar xf R-scape_version.tar.gz
> cd R-scape_version
> ./configure
> make
> make install
```

The newly compiled binary (`R-scape`) is in the `R-scape_version/bin` directory. You can run it from there, as in this example:

```
> bin/R-scape tutorial/updated_Arisong.sto
```

That's it. You can keep reading if you want to know more about customizing a R-scape installation, or you can skip ahead to the next chapter, the tutorial.

System requirements

Operating system: R-scape is designed to run on POSIX-compatible platforms, including UNIX, Linux and MacOS/X. The POSIX standard essentially includes all operating systems except Microsoft Windows. We have tested most extensively on Linux and MacOS/X because these are the machines we develop on.

Compiler: The source code is C conforming to POSIX and ANSI C99 standards. It should compile with any ANSI C99 compliant compiler, including the GNU C compiler `gcc`. We test the code using the `gcc` compiler.

Libraries and other installation requirements: R-scape includes three software libraries: the Easel library package (`http://selab.org/`), and two independent programs FastTree (Price et al., 2010) (for building phylogenetic trees) and R2R (Weinberg and Breaker, 2011) (for drawing consensus RNA structures). All will automatically compile during R-scape's installation process. By default, R-scape does not require any additional libraries to be installed by you, other than standard ANSI C99 libraries that should already be present on a system that can compile C code.

Makefile targets

`all` Builds everything. Same as just saying `make`.

`install` Installs the binaries (`R-scape`, `FastTree`, `r2r`).

By default, programs are installed in `R-scape_version/bin`. You can customize the location of the binaries by replacing

```
> ./configure
```

with

```
> ./configure --prefix=/the/directory/you/want
```

The newly compiled binaries are now in the `/the/directory/you/want/bin` directory.

`clean` Removes all files generated by compilation (by `make`). Configuration (files generated by `./configure`) is preserved.

distclean Removes all files generated by configuration (by ./configure) and by compilation (by make).

Why is the output of 'make' so clean?

Because we're hiding what's really going on with the compilation with a wrapper. If you want to see what the command lines really look like, pass a V=1 option (V for "verbose") to make, as in:

```
> make V=1
```

What gets installed by 'make install', and where?

The top-level configure file has a variable RSCAPE_HOME that specifies the directory where make install will install thingg: RSCAPE_HOME/bin.

By default RSCAPE_HOME is assigned to the current directory R-scape_version.

The best way to change this default is when you use ./configure, and the most important variable to consider changing is --prefix. For example, if you want to install R-scape in a directory hierarchy all of its own, you might want to do something like:

```
> ./configure --prefix=/usr/local/rscape
```

That would keep R-scape out of your system-wide directories like /usr/local/bin, which might be desirable. Of course, if you do it that way, you'd also want to add /usr/local/rscape/bin to your $PATH.

3 Tutorial

Here's a tutorial walk-through of some how to use R-scape. This should suffice to get you started.

Modes of R-scape

MSA input with annotated consensus secondary structure

R-scape Reports all pairs which covariation scores have E-values smaller or equal to a target E-value.
Draws the given consensus structure annotated with the significantly covarying base pairs.

MSA input without an annotated secondary structure

R-scape Reports all pairs which covariation scores have E-values smaller or equal to a target E-value.
Builds the best consensus structure that includes all significantly covarying pairs,
the maximum-covariation optimal consensus structure.
Draws the *maximum-covariation optimal consensus structure* annotated with the significantly covarying base pairs.

In the Tutorial section, I'll show examples of running each R-scape, using examples in the `tutorial/` subdirectory of the distribution.

Files used in the tutorial

The subdirectory `/tutorial` in the R-scape distribution contains the files used in the tutorial.

The tutorial provides several examples of RNA structural alignments, all in Stockholm format:

updated_Arisong.sto Structural alignment of the ciliate Arisong RNA. This alignment is an updated version of the one published in (Jung et al., 2011).

ar14.sto Structural alignment of the α-proteobacteria ncRNA ar14. This alignment is an updated version of the one published in (del Val et al., 2012).

RF00005.sto Rfam v12.0 (Nawrocki et al., 2015) seed alignment of tRNA.

RF00001-noss.sto Rfam v12.0 seed alignment of 5S rRNA, after removing the consensus secondary structure.

Running R-scape on one alignment file

To run R-scape with default parameters on alignment file `tutorial/updated_Arisong.sto` use:

```
> bin/R-scape tutorial/updated_Arisong.sto
```

Default parameters are:

Target E-value: default is 0.05. R-scape reports pairs which covariation score has E-value smaller or equal to the target value. The target E-value can be changed with option **-E <x>**, $x >= 0$.

Pairwise percent identity: Sequences with more than 97% similarity to each other are removed. Pairwise % identity is defined as the ratio of identical positions divided by the minimum length of the two sequences. The maximum pairwise percentage identity in the alignment can be changed with option **-I <x>**, $0 < x <= 1$.

`Gaps in columns` Columns with more than 50% gaps are removed. The gap threshold for removing columns can be modified using option **`--gapthresh <x>`** , $0 < x <= 1$.

`Covariation statistic` The default covariation statistic is the product-average corrected G-Test (equivalent to option **`--GTp`**).

`Covariation Class` R-scape uses the 16 component covariation statistic (C16), unless the number of sequences in the alignment is ≤ 8 or the length of the alignment is ≤ 50, in which case it uses the two-class covariation statistic (C2). A particular covariation class can be selected using either **`--C16`** or **`--C2`**.

The threshold for the minimum number of sequences can be changed with option `--nseqthresh <n>`. The threshold for the minimun alignment length can be changed with option `--alenthresh <n>`.

`Null alignments:` R-scape in order to estimate E-values produces 20 null alignments, unless the product of the number of sequences by the length of the alignment $<$ 10,000 in which case the number of null alignments is 50; or $<$ 1,000 in which case it is 100. The number of null alignments can be controlled with option **`--nshuffle <n>`**.

A full list of the R-scape options is fund by using

```
> R-scape -h
```

Tabular output per input file

The output file **`tutorial/updated_Arisong.out`** looks like this:

```
# MSA updated_Arisong_1 nseq 69 (95) alen 65 (150) avgid 65.16 (64.97) nbpairs 20 (20)
# Method Target_E-val cov_at_target_E-val [cov_min,conv_max] [FP | TP True Found | Sen PPV F]
# GTp    0.05         41.31               [-9.74,89.08]     [2 | 9 20 11 | 45.00 81.82 58.06]
               93              104      43.65   6.01567e-06
*              94              110      43.17   4.29923e-05
*              96              108      65.95   0
*              98              106      89.08   0
...
```

The output file is a tabular list of significant pairs ordered by the first positions:

`First column` indicates whether the significant pair is part of the given structure (*), or not. If the pair is not in the structure, we distinguish whether the pair is compatible with the given structure ($\sim$) o not, in which case it is a blank.

`Second and third columns` are the two positions of the pair, $i \leq j$ respectively. Positions are relative to the input alignment.

`Forth column` is the covariation score

`Fifth column` is the E-value. Significant positions have E-values $<< 1$.

The output file also includes two comment lines per alignment in the file:

`First comment line` describes properties of the alignment: number of sequence (nseq), alignment length (alen), average percentage identity (avgid), and number of base pairs (nbpairs). Values in parenthesis correspond to the alignment as is given. Values not in parenthesis correspond to the analyzed alignment after the default filters have been applied.

`Second comment line` describes properties of the R-scape search: the covariation method (GTp), the E-value threshold (0.05), the score at that E-value (42.3), the range of scores for all pairs in the alignments (from -9.7 to 89.1), the number of covarying not base pairs (2), the number of covarying base pairs (9), the number of base pairs (20), and the total number of covarying pairs (11). Lastly we provide the sensitivity (SEN=45.0=9/20), positive predictive value (PPV=81.8=9/11), and F-measure (F=58.1 = 2 * SEN * PPV / (SEN+PPV)).

Outputs per alignment

Two files are produced per alignment in the input file:

File **`tutorial/updated_Arisong_1.R2R.sto`** is a Stockholm formatted alignment that includes the input alignment annotated with the consensus structure. This Stockholm file also includes the additional annotation required to use the drawing program R2R.

It is possible that the resulting drawing will show parts of the secondary structure occluded from each other (especially for long RNAs). Using this file, one can customize a different drawing of the structure using the R2R documentation, provided in `lib/R2R/R2R-manual.pdf`.

File **`tutorial/updated_Arisong_1.his`** looks like this:

```
more tutorial/updated_Arisong.his
89.046301       0.000480769
68.879658       0.000961538
65.816370       0.00144231
56.115959       0.00192308
...
&
41.054795       9.61538e-06
40.799521       3.84615e-05
40.288973       6.73077e-05
...
&
49.478836       2.13504e-20
49.223562       4.27009e-20
48.968288       1.06752e-19
...
&
```

The first column is a covariation score (x). The second column is the survival function $P(X > x)$, that is the frequency of pairs having score larger than x. The file includes three histograms separated by a “&” line. The three histogram correspond to:

`First histogram` the given alignment, all possible pairs.

`Second histogram` the aggregation of all null alignments, all possible pairs.

`Third histogram` the expected null histogram according to the tail Gamma fit.

Graphical outputs per alignment

Three plots are produced per alignment in the input file:

updated_Arisong_1

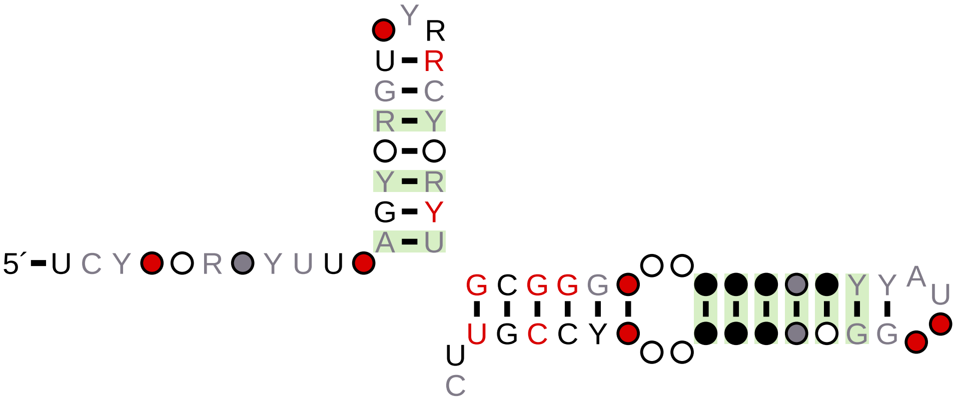

Figure 1: **tutorial/updated_Arisong_1.R2R.sto.{pdf,svg}: annotated consensus secondary structure.** Base pairs with covariation scores equal or below the target E-value (0.05 as default) are depicted in green. By default only positions in the alignment with more than 50% occupancy are depicted (unless they form a base pair). Option `--r2rall` forces the depiction of all positions in the alignment.

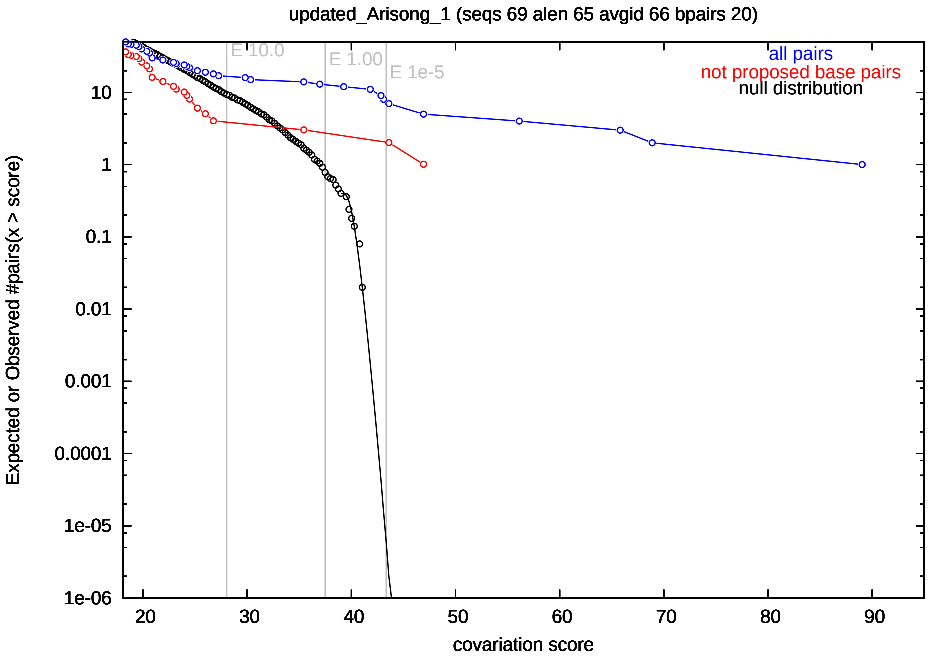

Figure 2: **tutorial/updated_Arisong_1.his.{ps,svg}: covariation scores histogram.** The histogram of scores for all pairs in the given alignment is depicted in blue. The histogram for the null alignments is depicted in black. A black line indicates to fit to a truncated Gamma distribution of the tail of the null distribution. In red, we plot the histogram of scores for the pairs in the given alignment excluding those proposed as base pairs.

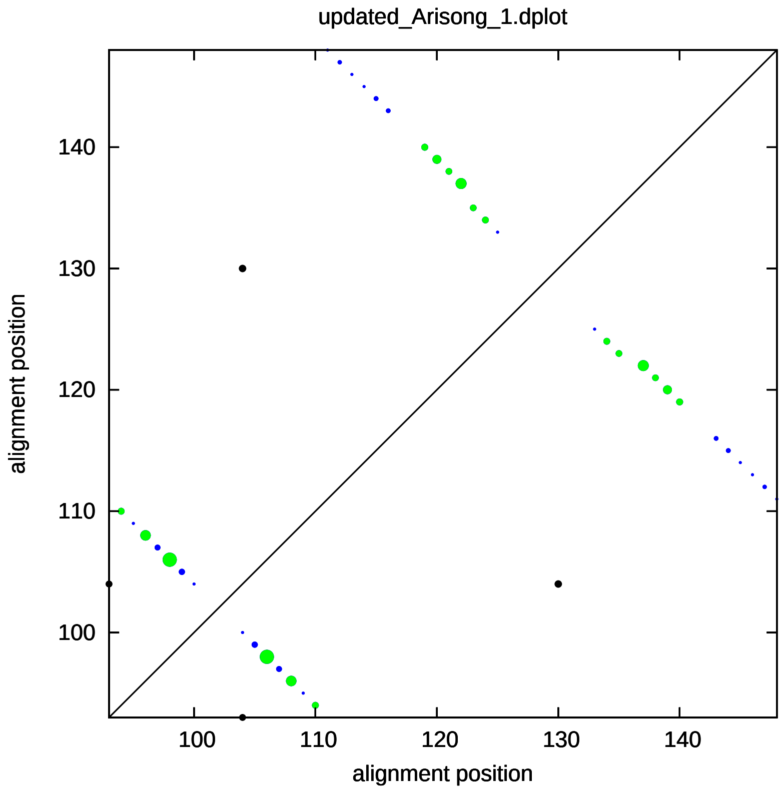

Figure 3: **`tutorial/updated_Arisong_1.dplot.`{`ps,svg`}: dotplot.** Dot size is proportional to the covariation score. In blue we depict the consensus base pairs; in green, the consensus base pairs that show significant covariation; in orange (none shown in this plot), we depict other pairs that have significant covariation, are not part of the consensus secondary structure but are compatible with it; in black we depict other significant pairs. Position are relative to the input alignment

Other tabular outputs

R-scape produces two more tabular outputs per input file that are more relevant for benchmarking purposes, those are:

File **`tutorial/updated_Arisong.sum`** looks lie:

```
more tutorial/updated_Arisong.sum
#target_E-val   MSA                    nseq    alen    avgid    method  TP      True    Found   SEN     PPV
0.050000        updated_Arisong_1      69      65      65.16    GTp     9       20      20      45.00   45.00
```

This file produces a one line output per alignment in the file.

`Column 1` Target E-value.

`Column 2` Alignment name.

`Column 3` Number of sequence in the analyzed alignment.

`Column 4` Number of columns analyzed.

`Column 5` Average percentage identity in the analyzed alignment.

`Column 6` Covariation statistic.

`Column 7` Number of significant base pairs (tf).

`Column 8` Number of base pairs (true).

`Column 9` Number of significant pairs (found).

`Column 10` Sensitivity = tf/true.

`Column 11` Positive predictive value = tf/found.

File **`tutorial/updated_Arisong.roc`** looks like:

```
more tutorial/updated_Arisong.roc
# MSA nseq 69 alen 65 avgid 65.163163 nbpairs 20 (20)
# Method: GTp
#cov_score            FP              TP              Found              True         Negatives         Sen     PPV     F       E-
89.04630               0               1                  1                20              2060         5.00    100.00  9.52    0
88.79103               0               1                  1                20              2060         5.00    100.00  9.52    0
88.53575               0               1                  1                20              2060         5.00    100.00  9.52    0
...
```

This file produces a tabular output for each alignment and for as a function of the covariation score. The values in the file are described by the commented line.

4 Outputs

For each alignment file rnafile.sto, R-scape produces the following output files:

rnafile.out Tabular output with the significant pairs, with their score and E-value.

rnafile.sorted.out Tabular output sorted from highest to lowest E-value.

rnafile.roc Tabular output that provides statistics for each score value.

rnafile.sum Tabular output with a line summary statistics per alignment in the file.

An Stockholm alignment file can include multiple alignments. R-scape produces the following output files, one for each individual alignment in the input Stockholm file:

Alignment with consensus secondary structure

If the given alignment does have a consensus secondary structure (#=GF SS_cons markup), the following files are produced

rnafile_msaname.R2R.sto Stockholm file annotated by a modified version of the R2R program. This file includes the information necessary to draw the consensus structure, and to annotate the significantly covarying base pairs.

rnafile_msaname.R2R.sto.{pdf,svg} Drawing of the R-scape-annotated consensus secondary structure.

rnafile_msaname.his A two column histogram file for the covariation scores.

rnafile_msaname.his.ps Plot of the score histogram. Drawing of this file requires that program **gnuplot** is installed somewhere in the ${PATH}, of that the environmental variable GNUPLOT pointing to a gnuplot executable is defined.

rnafile_msaname.dplot.{ps,svg} Dot plot of the consensus secondary structure annotated according to covariation. Drawing of this file requires that program **gnuplot** is installed somewhere in the ${PATH}, of that the environmental variable GNUPLOT pointing to a gnuplot executable is defined.

For each alignment, **msaname** is given by <ACC>_<ID>, the combination of the accession #=GF AC <AC> and name #=GF ID <ID> in the Stockholm-format markups (or one of two if the other in not defined). If none of those fields are defined, **msaname** is a number describing the order in the file of the given alignment.

Alignment without consensus secondary structure

Alternatively, if the alignment does not have a consensus secondary structure (or if it does and the option R-scape --cyk is used) R-scape produces the following additional files describing the maximal-covariation optimal secondary structure:

rnafile_msaname.cyk.R2R.sto

rnafile_msaname.cyk.R2R.sto.{pdf,svg}

rnafile_msaname.cyk.his

rnafile_msaname.cyk.his.{ps.svg}

rnafile_msaname.cyk.dplot.{ps,svg}

These files are formatted identically to those for describing the given consensus structure.

5 Options

The whole list of options can be found using

```
> R-scape -h.
```

Some important options are:

Covariation statistic options

-E <x>

Target E-value is $x \geq 0$.

--GT, --MI, --MIr, --MIg, --CHI, --OMES, --RAF, --RAFS,

We favor the G-test covariation statistic, but a total of eight covariation statistics are currently implemented in R-scape. For each covariation statistic (GT, for instance), R-scape can also calculate its average product correction (GTp) ad its average sum corrections (GTa). Each option `--GT` stands for three independent ones: `--GT, --GTp, --GTa`.

The R-scape default is `--GTp`.

Details of the definition and provenance of the different covariation statistics can be found in the R-scape manuscript: Rivas, E. & Eddy S. E., *"A statistical test of RNA base pair covariation applied to proposed lncRNA structures"*. In a nutshell, given two alignment columns i, j,

$$\begin{array}{lrcl}
\text{G-test:(Woolf, 1957)} & \mathrm{GT}(i,j) & = & 2\sum_{a,b} \mathrm{Obs}_{ij}^{ab} \log \frac{\mathrm{Obs}_{ij}^{ab}}{\mathrm{Exp}_{ij}^{ab}}, \\
\text{Pearson's chi-square:} & \mathrm{CHI}(i,j) & = & \sum_{a,b} \frac{\left(\mathrm{Obs}_{ij}^{ab} - \mathrm{Exp}_{ij}^{ab}\right)^2}{\mathrm{Exp}_{ij}^{ab}}, \\
\text{Mutual information:(Shannon, 1948; Gutell et al., 1994)} & \mathrm{MI}(i,j) & = & \sum_{a,b} P_{ij}^{ab} \log \frac{P_{ij}^{ab}}{p_i^a\, p_j^b}, \\
\text{MI normalized:(Martin et al., 2005)} & \mathrm{MIr}(i,j) & = & \frac{\mathrm{MI}(i,j)}{H(i,j)} = \frac{\mathrm{MI}(i,j)}{-\sum_{a,b} P_{ij}^{ab} \log P_{ij}^{ab}}, \\
\text{MI with gap penalty:(Lindgreen et al., 2006)} & \mathrm{MIg}(i,j) & = & \mathrm{MI}(i,j) - \frac{N_{ij}^G}{N}, \\
\text{Obs-Minus-Exp-Squared:(Fodor and Aldrich, 2004)} & \mathrm{OMES}(i,j) & = & \sum_{a,b} \frac{\left(\mathrm{Obs}_{ij}^{ab} - \mathrm{Exp}_{ij}^{ab}\right)^2}{N_{ij}}, \\
\text{RNAalifold (RAF):(Hofacker et al., 2002)} & \mathrm{RAF}(i,j) & = & \mathrm{B}_{i,j}, \\
\text{RNAalifold Stacking (RAFS):(Lindgreen et al., 2006)} & \mathrm{RAFS}(i,j) & = & \frac{1}{4}\left(\mathrm{B}_{i-1,j+1} + 2\,\mathrm{B}_{i,j} + \mathrm{B}_{i+1,j-1}\right).
\end{array}$$

where a, b are (non-gap) residues; N is the total number of aligned sequences; Obs_{ij}^{ab} is the observed count of $a : b$ pairs in columns i, j (only counting when both a,b are residues); N_{ij} is the total number of residue pairs in columns i, j (only counting when both a,b are residues); P_{ij}^{ab} is the observed frequency of pair $a : b$ in columns i, j ($P_{ij}^{ab} = \frac{Obs_{ij}^{ab}}{N_{ij}}$); $\mathrm{Exp}_{ij}^{ab} = N_{ij} p_i^a p_j^b$ is the expected frequency of pair $a : b$ assuming i, j are independent, where p_i^a are the marginal frequencies of a residues in column i (averaged to all other positions) ($p_i^a = \frac{1}{L-1}\sum_{j \neq i}\sum_b P_{ij}^{ab}$); $N_{ij}^G = N - N_{ij}$ is the number of pairs involving at least one gap symbol; the definition of $\mathrm{B}_{i,j}$ used in the RAF and RAFS statistics is involved, a concise definition can be found elsewhere (Lindgreen et al., 2006).

The background corrections (Dunn et al., 2007) for a given covariation statistic above $\mathrm{COV}(i,j)$ are,

$$\begin{array}{lrcl}
\text{Average product correction} & \mathrm{COVp}(i,j) & = & \mathrm{COV}(i,j) - \frac{\mathrm{COV}(i)\mathrm{COV}(j)}{\mathrm{COV}}. \\
\text{Average sum correction} & \mathrm{COVa}(i,j) & = & \mathrm{COV}(i,j) - \left(\mathrm{COV}(i) + \mathrm{COV}(j) - \mathrm{COV}\right),
\end{array}$$

--C2, --C16

For all the covariation statistics (except RAF and RAFS), one can do a 16-component (C16) or a two-component (C2) calculation, depending on whether it uses the 16 possible pair combinations, or those are group in two classes depending on whether they form a Watson-Crick pair (6 cases, including U:G and G:U), or whether they do not (10 cases).

R-scape's default is the 16 component covariation statistic, unless the number of sequences in the alignment is $\leq$ 8 or the length of the alignment is $\leq$ 50, in which case it uses the two-class covariation statistic.

Search options

--cyk

An optimal secondary structure is computed that includes all significant base pairs. The files for this maximum-covariation optimal structure all include the suffix `.cyk`.

--window <n>

R-scape can be run in a window scanning version for long alignments. The window size is $n > 0$.

--slide <n>

In scanning mode, this options sets the number of positions to move from window to window, $n > 0$.

Input alignment options

-I <x>

Only sequences with less than $0 < x \leq 1$ fraction identity are considered in the analysis.

--gapthresh <x>

Only columns with less than $0 < x \leq 1$ fraction of gaps are considered in the analysis.

--submsa <n>

Analyzes a random subset of the input alignment.

--treefile <f>

A phylogenetic tree in Newick format can be given (by default a tree is created from the alignment using the program FastTree (Price et al., 2010)). R-scape checks that the number of taxa and the names of the taxa matches for all alignments analyzed.

Output options

--outmsa <f>

The actual alignment analyzed can be saved in Stockholm format to file ¡f¿.

Plotting options

--nofigures

None of the graphical outputs are produced using this option.

`--r2rall`

Forces R2R to draw all positions in the alignment. By default only those that are more than 50% occupied or are base paired are depicted.

6 Some other topics

How do I cite R-scape?

Pending a publication, the appropriate citation is to the web server, `github.com/EddyRivasLab/R-scape`.

You should also cite what version of the software you used. We archive all old versions, so anyone should be able to obtain the version you used, when exact reproducibility of an analysis is an issue.

The version number is in the header of most output files. To see it quickly, do something like `R-scape -h` to get a help page, and the header will say:

```
# R-scape :: RNA Structural Covariation Above Phylogenetic Expectation
# R-scape 0.1 (FEB 2016)
# Copyright (C) 2016 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

So (from the second line there) this is from R-scape v0.1.

How do I report a bug?

Email us, at `elenarivas@fas.harvard.edu`.

Before we can see what needs fixing, we almost always need to reproduce a bug on one of our machines. This means we want to have a small, reproducible test case that shows us the failure you're seeing. So if you're reporting a bug, please send us:

- A brief description of what went wrong.
- The command line(s) that reproduce the problem.
- Copies of any files we need to run those command lines.
- Information about what kind of hardware you're on, what operating system, and what compiler and version you used, with what configuration arguments.

7 Acknowledgments

We thank Roian S.E. Egnor for suggesting the name R-scape, and the Centro de Ciencias de Benasque Pedro Pascual in Spain, for their hospitality, over numerous and wonderful summers.

References

del Val, C., Romero-Zaliz, R., Torres-Quesada, O., Peregrina, A., Toro, N., and Jiménez-Zurdo, J. I. (2012). A survey of sRNA families in α-proteobacteria. *RNA Biol*, 9:119–129.

Dunn, S. D., Wahl, L. M., and Gloor, G. B. (2007). Mutual information without the influence of phylogeny or entropy dramatically improves residue contact predictions. *Bioinformatics*, 24:333–340.

Fodor, A. A. and Aldrich, R. W. (2004). Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins: Structure, Function, and Bioinformatics*, 56(2):211–221.

Gutell, R. R., Larsen, N., and Woese, C. R. (1994). Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiol. Rev.*, 58:10–26.

Hofacker, I. L., Fekete, M., and Stadler, P. F. (2002). Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, 319:1059–1066.

Jung, S., Swart, E. C., Minx, P. J., Magrini, V., Mardis, E. R., Landweber, L. F., and Eddy, S. R. (2011). Exploiting *Oxytricha trifallax* nanochromosomes to screen for noncoding RNA genes. *Nucl. Acids Res.*, 39:7529–7547.

Lindgreen, S., P.P., G., and Krogh, A. (2006). Measuring covariation in RNA alignments: physical realism improves information measures. *Bioinformatics*, 22 (24):2988–2995.

Martin, L., Gloor, G., Dunn, S., and Wahl, L. (2005). Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, 21:4116–4124.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., and Finn, R. D. (2015). Rfam 12.0: updates to the RNA families database. *Nucl. Acids Res.*, 43:D130–D137.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 - approximately maximum-likelihood trees for large alignments. *PLOS ONE*, 5:e9490.

Shannon, C. (1948). A note on the concept of entropy. *Bell System Tech. J*, 27:379–423.

Weinberg, Z. and Breaker, R. R. (2011). R2R – software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinformatics*, 12:3.

Woolf, B. (1957). The log likelihood ratio test (the G-test). *Annals of Human Genetics*, 21(4):397–409.