EnsembleRNA

September 20, 2016

Title	Visualize the structural ensemble for a given RNA
Version	1.0.0
Author	Chanin Tolson Woods
URL	http://ribosnitch-ensemblerna.rhcloud.com
URL2	http://ribosnitch.bio.unc.edu/software

Description

EnsembleRNA is a package for the visualization and comparison of RNA structural ensembles. This package creates a stable map of conformational space for a given RNA and its mutants. The map explores the most diverse conformational space and generates the structures using established Boltzmann-weighted suboptimal sampling algorithms. Using vector representation based on arc diagram nested loop patterns, EnsembleRNA projects clusters of structures from the map into two dimensions using metric multidimensional scaling. Individual RNA ensembles are visualized in this space by varying the size of the structure clusters in a bubble plot.

The sequence from the FASTA file is the reference used to create the map of conformational space, unless otherwise specified. To compare two reference structures, the same sequence or set of structures must be used to create the map of conformational space. Larger RNAs may require more sequences for a stable visualization. Selective 2'-hydroxyl acylation and primer extension (SHAPE) data can be included to guide the prediction of the reference ensemble.

Note: EnsembleRNA is written for use on a Linux/Unix-type operating system. Use of EnsembleRNA in any instance requires the installation of the numpy, jinja2, ipython, mpld3, matplotlib, scipy, and sklearn modules for Python. Also required is the RNAstructure package from the Mathews lab.

Depends	Python 2.7 or Python 3.5
License	GPL (>=3)
Imports	numpy, jinja2, ipython, mpld3, matplotlib, and sklearn
Required	RNAstructure

Table of Contents

Imports and Requirements

For use on a Linux/Unix operating system

1) python (recommended version 2.7 or 3.5)

2) numpy (recommended 1.11.0)
    pip install numpy

3) scipy (recommended version 0.17.1)
    pip install scipy

4) sklearn (recommended version 0.0)
    pip install sklearn

5) jinja2 (recommended version 2.8)
    pip install jinja2

6) ipython (recommended version 4.2.0)
    pip install ipython

7) mpld3 (recommended version 0.2)
    pip install mpld3

8) matplotlib (recommended version 1.5.1)
    pip install matplotlib

9) RNAstructure (recommended version 5.8)
    http://rna.urmc.rochester.edu/RNAstructureWeb/
    Download command-line applications for your platform
    Extract to /usr/local/bin (or directory of your choice)
    Add following 2 lines to ~/.bash_profile (path may be different)
        export PATH=$PATH:/usr/local/bin/RNAstructure/exe
        export DATAPATH=/usr/local/bin/RNAstructure/data_tables

Installation

1) Download requirements listed above

2) Download EnsembleRNA package

3) Place package in /usr/local/bin (or directory of your choice)

4) tar -zxvf ensemblerna (extract)

5) cd ensemblerna (enter extracted directory)

6) sudo python setup.py install (install ensemblerna as python module)

7) ensemblerna -h (test installation in any directory)

Usage

ensemblerna <fasta file> <output directory> [options]

Options

General

-h, --help	show this help message and exit
-v, --version	show program's version number and exit

Inputs

-sh --shape	Includes shape data in the reference ensemble prediction. Ignored if -d flag is used (Default is None)
-d --db	Dot-bracket structures for reference ensemble (Default is None)
-m --map	Sequence to create the map of conformational space. Ignored if -md flag is used (Default is reference fasta file)
-md --mapdb	Dot-bracket structures for the map of conformational space. A previously created map can be used to project new ensembles onto the same space (Default is None)
-s --size	Number of mutants for the map of conformational space. Higher numbers increase structural diversity. Ignored if –md flag is used (Default is 10)
-p --plotmap	Plot the map T/F (Default is T)
-r --range	Range of nucleotides to visualize. Predicted structures will include the full length of the input RNA but only the given range will be plotted (Default is 1 to sequence length)
-pi --plotinteractive	Plot the interactive file T/F (Default is T)
-th --threadmax	Maximum number of threads for multi-threading. (Default is 1)
-i --ignorestems	Ignore stems with fewer than i base pairs. (Default is 3)
-n --num	Number of Boltzmann sampled structures to produce for the visualization (Default is 1000)

RNAstructure

-maxd --maxdistance	The maximum number of bases between the two nucleotides in a pair (Default is no restriction)
-t --temperature	Temperature at which the calculation takes place in Kelvin (Default is 310.15 K)
-si --SHAPEintercept	The intercept used with SHAPE restraints. Ignored if -d flag is used (Default is -0.6 kcal/mol)
-sm --SHAPEslope	The slope used with SHAPE restraints. Ignored if -d flag is used. (Default is 1.8 kcal/mol)

Output

For both reference and map of conformational space

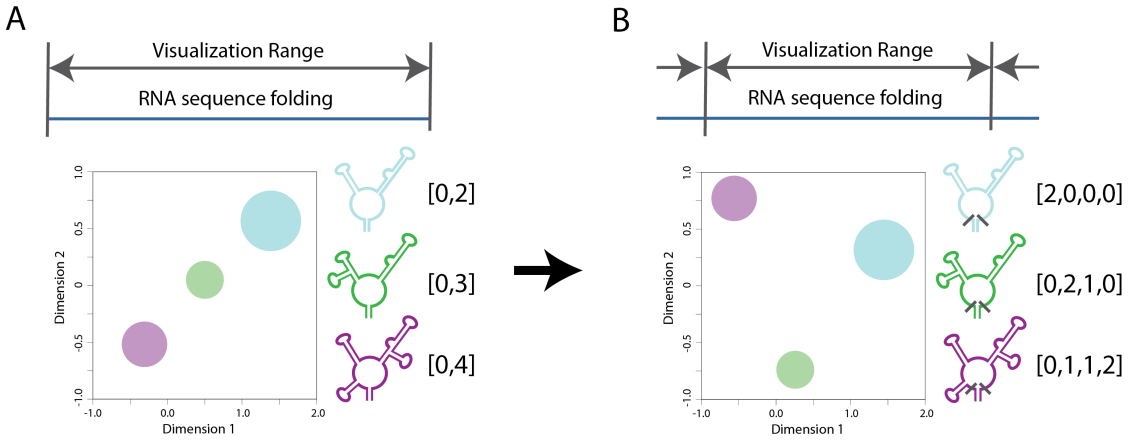
.csv CSV file with cluster number, cluster size, and representative structure
.db Dot-Bracket file with structures
.pdf PDF file with visualization plot
.png PNG file with visualization plot

Interactive visualization

.html HTML file with interactive plotting

Troubleshooting

Diagonal line visualization



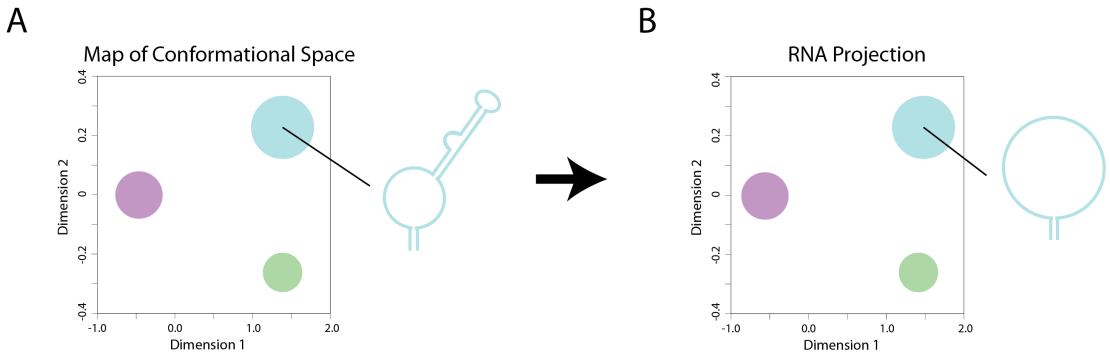
Problem If all bubbles lie on a diagonal line, there is correlation between dimensions 1 and 2 (A). By default, EnsembleRNA defines RNA structure based on the outermost stacks and loops (the most abstracted representation). In this case, the structures are similar from the level of the outermost stack, but interesting differences may exist for loops within that outer stack.

Solution To address this problem, the full-length of the RNA can be folded, while the visualization is focused on a shorter range that excludes the outer stack (B). This change reveals the more subtle differences between structures.

Example For a 250 nucleotide RNA, include the entire sequence in the fasta file. Only visualize the range from nucleotide 50 to 200 using the range flag (-r or --range).

ensemblerna <fasta file> <output directory> **-r 50 200**

Selecting medoid structure

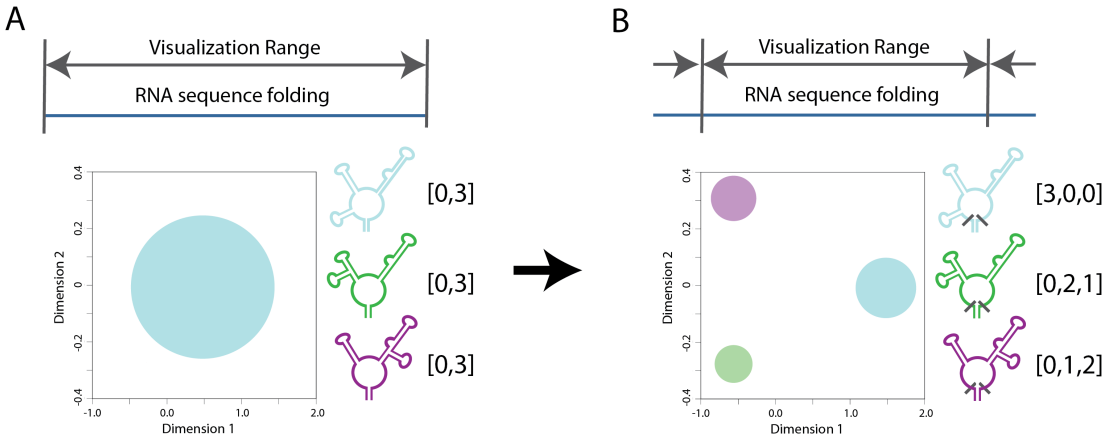


Problem The default medoid structure is chosen from the map of conformational space (A). Using this medoid keeps the representation consistent between different ensembles projected onto the same space. However, the best representative structure for the projected RNA may be different.

Solution A structure selected from the projected RNA ensemble may be more representative (B). Alternatively, the minimum free energy structure from either the map or the projected RNA ensemble can be used.

Example Check the .db file in the output folder.

Single point visualization



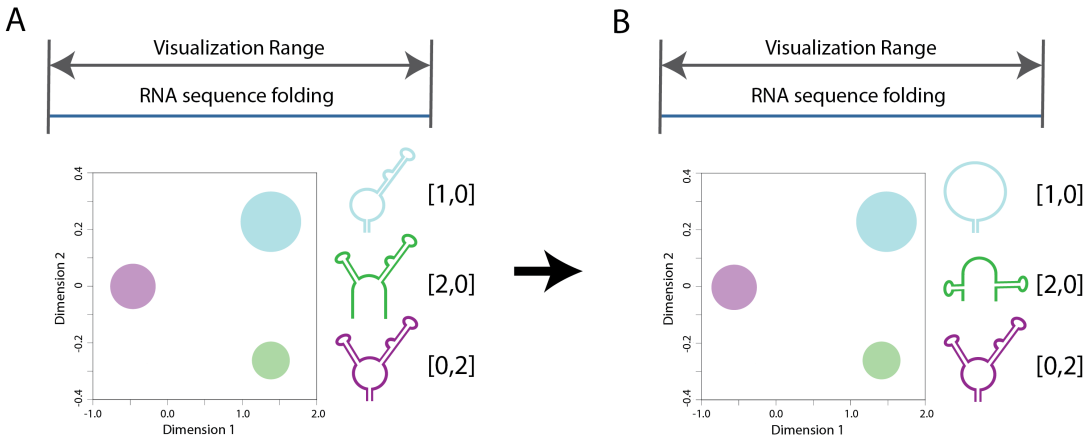
Problem If all structures are placed in a single cluster, the visualization may be too broad (A). By default, EnsembleRNA defines RNA structure based on the outermost stacks and loops (the most abstracted representation). In this case, the structures are the exact same from the level of the outermost stack, but interesting differences may exist for loops within that outer stack.

Solution To address this problem, the full-length of the RNA can be folded, while the visualization is focused on a shorter range that excludes the outer stack (B). This change reveals the more subtle differences between structures.

Example For a 250 nucleotide RNA, include the entire sequence in the fasta file. Only visualize the range from nucleotide 50 to 200 using the range flag (-r or --range).

ensemblerna <fasta file> <output directory> **-r 50 200**

Outer loop visualization

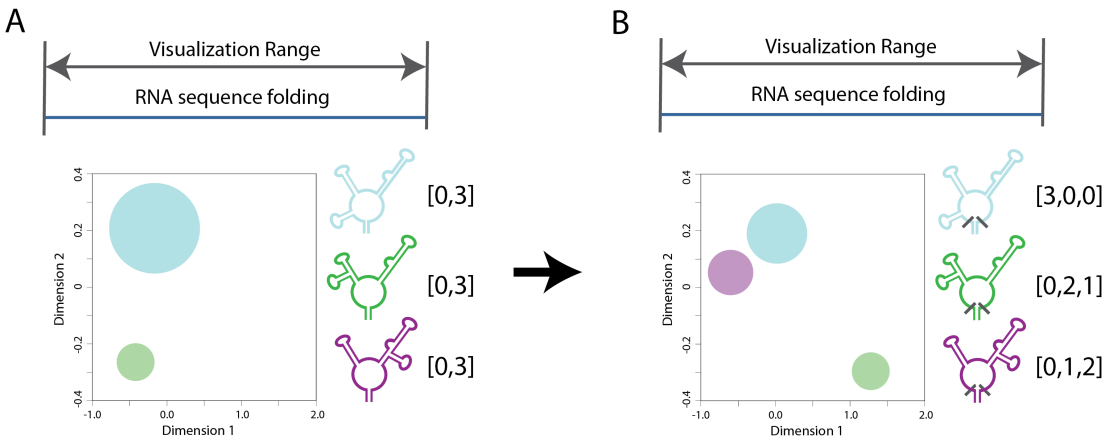


Problem RNA clusters with only outer loops are considered to be more similar to each other than those clusters with an outer stack (A). The clusters with only outer loops are most often very diverse groupings. While some structures may be similar to those with outer stacks, many structures will be quite different.

Solution Our nestedness representation method accounts for this increased diversity in clusters with only outer loops (B). Looking at the cluster medoid may be useful in assessing the similarity of these clusters to those with outer stacks.

Example Check the .csv file or the .html file in the output folder.

Increasing map coverage



Problem The default map size (the number of single point mutants included in the map) is automatically set to 10. This is a reasonably size for RNAs of shorter length (100-200 nucleotides). However, longer RNAs will likely require larger map sizes to sufficiently explore the structural space for an RNA.

Solution Increase the map size in increments until the number of clusters structures converges. At this map size, additional single point mutants will not increase the structural diversity in the map of conformational space. The optimal map size varies by RNA.

Example For an 800 nucleotide RNA, include the entire sequence in the fasta file. Increase the map size from 10 to 100.

ensemblerna <fasta file> <output directory> **-s 100**

Documentation References

Suboptimally sampled structures are generated using the RNAstructure package
http://rna.urmc.rochester.edu/RNAstructureWeb/
Version 5.8 (references 2, 3, 4, and 5)

1. D.H. Mathews. 2004. Using an RNA Secondary Structure Partition Function to Determine Confidence in Base Pairs Predicted by Free Energy Minimization. RNA, 10:1178-1190. (2004).
2. S. Duan, D.H. Mathews, D. H. Turner. 2006. Interpreting Oligonucleotide Microarray Data to Determine RNA Secondary Structure. Application to the 3' End of Bombyx mori R2 DNA. Biochemistry, 45:9819-9832.
3. D.H. Mathews. 2006. Revolutions in RNA Secondary Structure Prediction. Journal of Molecular Biology, 359:526-532.
4. S. Wuchty, W. Fontana, I. L. Hofacker, P. Schuster. 1999. Complete Suboptimal Folding of RNA and the Stability of Secondary Structures. Biopolymers, 49:145-165.
5. Y. Ding and C.E. Lawrence. 2003. A Statistical Sampling Algorithm for RNA Secondary Structure Prediction. Acids Research, 31:7280-7301.
6. E.J. Merino, K. A. Wilkinson, J. L. Coughlan and K. M. Weeks. 2005. RNA structure analysis at single nucleotide resolution by selective 2-hydroxyl acylation and primer extension (shape). J Am Chem Soc, 127:4223-4231.
7. K.E. Deigan, T. W. Li, D. H. Mathews and K. M. Weeks. 2009. Accurate SHAPE-directed RNA Structure Determination. Proceedings of the National Academy of Sciences USA, 106:97-102.
8. K. Pearson. 1901. On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine, 11:559–572.
9. D.B. Carr, R. J. Littlefield, W.L. Nicholson, J.S. Littlefield. 1987. Scatterplot Matrix Techniques for large N. Journal of the American Statistical Association, 389:424-436.
10. J. Ritz, J. Martin and A. Laederach. 2012. Evaluating our ability to predict the structural disruption of RNA by SNPs. BMC Genomics, 13(Suppl 4):S6.