

*Laboratory of Integrative Bioinformatics
Bioinformatics Core
Centro de Genómica y Bioinformática
Universidad Mayor*

structRNAfinder tutorial

Version 1.0

Contact:

Vinicius Maracaja-Coutinho, PhD
Bioinformatics Core
Centro de Genómica y Bioinformática
Universidad Mayor

vinicius.coutinho@umayor.cl
<http://integrativebioinformatics.me>

November 2014
Santiago, Chile

Example of usage

Brief Overview & Datasets

We present here a tutorial for new users of *StructRNAfinder*. The aim of this document is to show an application and a step-by-step usage of the tool. For that, it is necessary to download a publicly available dataset of an *E. coli* transcriptome, corresponding to 154 non-coding RNAs sequences (Sætrom et. al., 2005). Note that, as explained below, the covariance models from Rfam are downloaded automatically by *StructRNAfinder*. (see documentation in install section)

***E. coli* dataset:**

http://nar.oxfordjournals.org/content/suppl/2005/06/07/33.10.3263.DC1/all_rnas_ecoli.fasta.txt

Note: *The estimated computing time for predicting structured RNAs in this transcriptome, using 4 standard processors, is approximately 3 minutes.*

Running *StructRNAfinder*

It is important to stand out that *StructRNAfinder* uses by default the covariance models (CM) extracted from Rfam database (Burge et al., 2012), and it is designed for that CMs. However, it is possible to use any user generated covariance model, where the main difference is the kind of report delivered. Using a covariance model from Rfam, *structRNAfinder* will generate a more complete final report, containing information related to *taxonomy* and *additional annotations* for each RNA family. However, using an user generated covariance model is possible to perform different kind of analysis, i.e. comparative RNomics using a CM generated in one specie and comparing it to sequences from another specie. Here, we will focus exclusively on the usage of the standard Rfam CMs.

Running *StructRNAfinder* with default parameters is very easy. It is only necessary to define the input sequences in FASTA format (i.e. *-i e_coli_trans.fasta*), in which the predictions will be performed; the reference covariance model database (i.e. *-c Rfam.cm*), in this case the models automatically downloaded from Rfam (v12) when you installed the tool; and the number of processors used in the analysis (i.e. *-c 3*). In this case, we use the *cmscan* algorithm from *Infernal* for the structural similarities search.

Note that here we are not specifying the “-r” option. *StructRNAfinder* can generate two different reports. The default is considering only the best hit for each sequence, i.e. one structure per sequence. Using the “-r” option *structRNAfinder* reports all significant hits for the sequence.

```
structRNAfinder -i e_coli_trans.fasta -d Rfam.cm -c 3
```

Note: The headers in the input fasta should have less than 20 characters, if not these will be automatically trimmed to that length.

StructRNAfinder Reports

After running *StructRNAfinder*, the following html reports are generated into the *html* folder:

A - index.html. Table with a summary containing information related to each annotated structured RNA. From fields 1 to 7 we have the information generated by *Infernal* (Nawrocki and Eddy, 2013), while fields 8 and 9 we have the information generated by *RNAfold* (Lorenz *et al.*, 2011):

- 1.- **Sequence:** corresponds to the name of the given sequence in the fasta input (if it have more than 20 characters on the header, it is trimmed automatically). In case of using the “-r” option, i.e. selecting all significant hits in a given sequence, these are enumerated according to each sequence name, suffixing an “_”. For example: *seq1_1*, *seq1_2*, *seq1_3*, etc.
- 2.- **RNA family:** the non-coding RNA family obtained using the annotation available for each *Rfam* covariance model.
- 3.- **ID:** the non coding RNA family identifier.
- 4.- **From_seq:** start coordinate of the hit in the sequence.
- 5.- **To_seq:** end coordinate of the hit in the sequence.
- 6.- **Score:** score obtained by the method used in *Infernal*
- 7.- **Evalue:** *e-value* reported by *Infernal*.
- 8.- **Energy:** obtained in the secondary structure prediction by *RNAfold*.
- 9.- **Structure:** image of the secondary structure obtained through *RNAfold*.

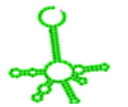






	Infernal						RNAfold	
Sequence	RNA family	Id	From_seq	To_seq	Score	Evalue	Score	Struct
Other								
Ec_rtT	RtT	RF00391	40	171	133.5	2.5e-30	-51.00	
sraF_tpk1	yybP-ykoY	RF00080	4	124	50.1	1.8e-11	-45.50	
antisense								
dicF	DicF	RF00039	2	53	63.2	2.7e-14	-16.10	
Ec_micF	MicF	RF00033	81	174	98.2	5.6e-20	-18.90	
Others-gene								
ffs	Bacteria_small_SRP	RF00169	33	129	65.7	2.3e-16	-32.00	
ssrS	6S	RF00013	1	183	123.2	1.3e-28	-80.10	
gcvB	GcvB	RF00022	1	205	178.5	2.1e-49	-61.00	

Figure 1 - HTML report generated by *StructRNAfinder*. Here we have all information obtained and extracted from the predicted RNA structures. From 1 to 7 we have the information generated by *Infernal*. Fields 8 and 9 are the one generated by *RNAfold*.

Clicking on the sequence ID in the summary table, the user access the complete annotation of this particular predicted RNA. Here, is possible to obtain detailed results retrieved from *Infernal* and *RNAfold*, as well as a full description of the RNA family (Figure 2).

Figure 2 - Additional information obtained by clicking on a particular predicted RNA on the report. Here, users can have access to the Rfam family description, scores, e-value, genomic coordinates, the covariance models and alignments, as well as figures for the predicted structure and for target reference.

C - *summary.html*. Summary page with the general information about hits found.

The summary page contains statistics related to the annotated RNAs and can be accessed by clicking on the summary link on the left menu on the index.html page. Here, we have a report with a table containing the total numbers of the annotated sequences, as well a pie chart with the distribution of each hits according to each represented noncoding RNA family. Also, it is possible to retrieve the additional files generated by *StructRNAfinder* (for additional information see section 5.3 from the Documentation).

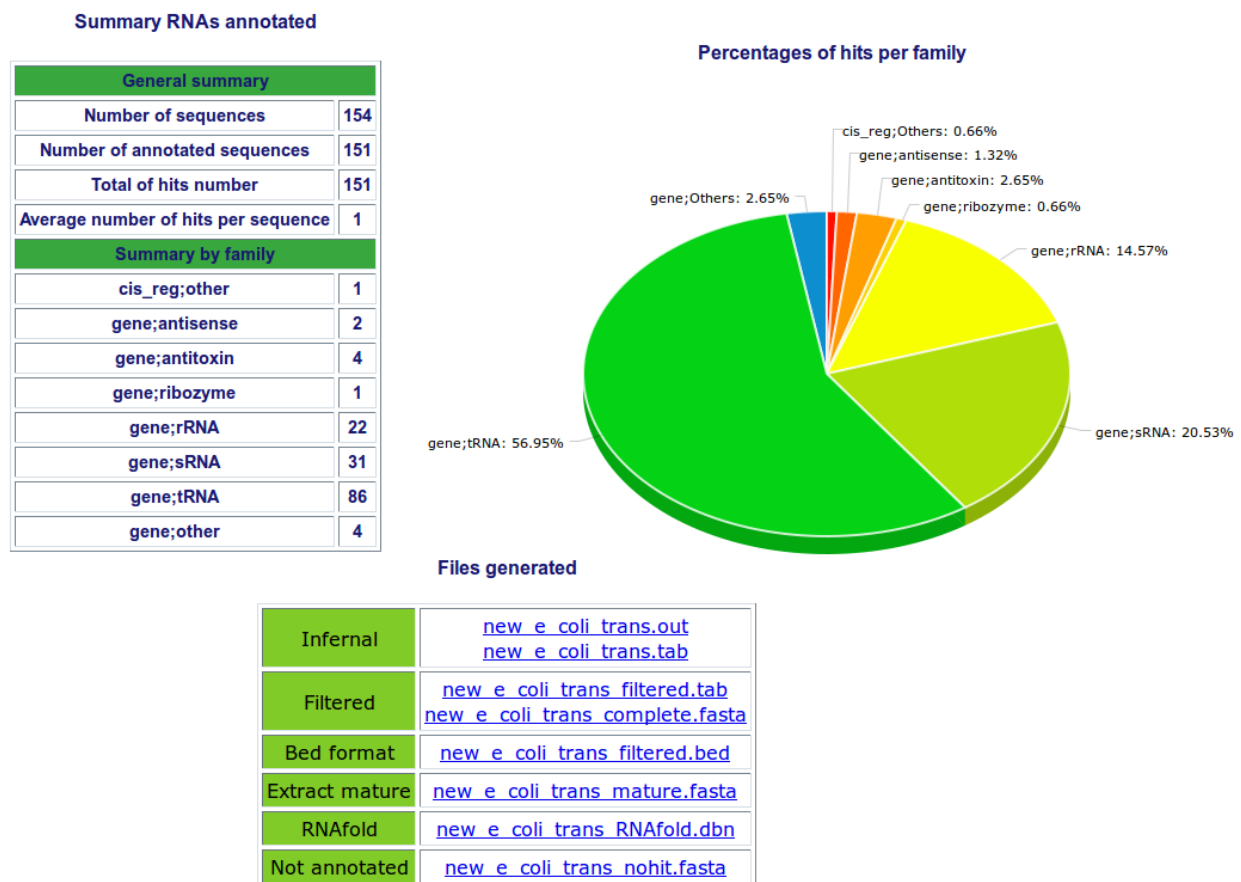


Figure 3 - Summary information of all hits found. The first table is divided in two sections: *General summary* (reporting all hits found) and *Summary by family* (number of hits for each family). The pie graph represent the portions of each family based on the total hits. Bellow is shown a table with direct links to the files generated during the course of the tool.

D - *loci.html*. Page containing representative figures of the position of each structure alignment on the sequence

A visual representation of region where the noncoding RNA was predicted is available on this page. Here, blue lines represent the hits in positive strand and the light blue in negative strand. This page is more useful in the case of the -r option is specified.

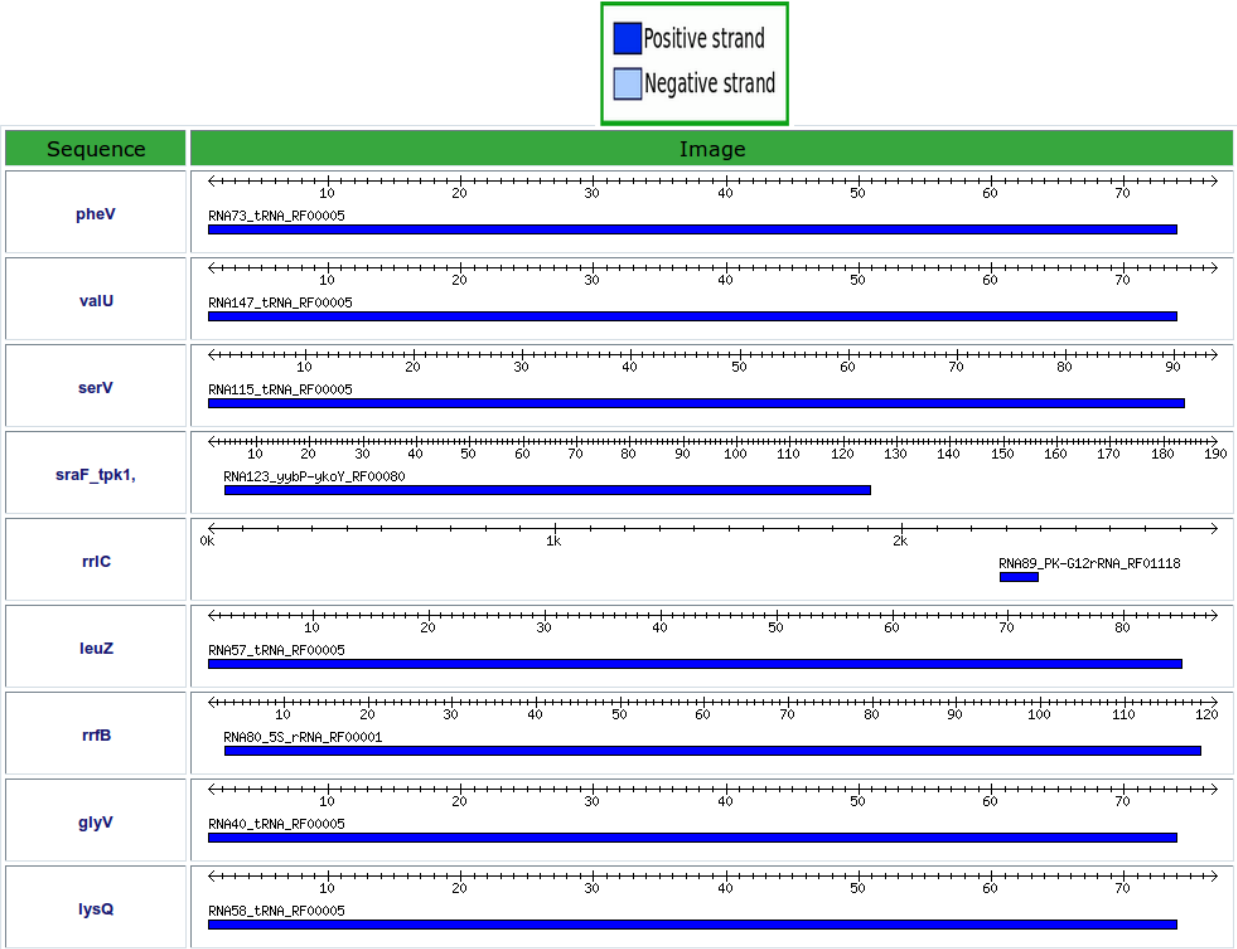


Figure 3 - General overview of the representative figures of the specific *loci* on the sequences containing a particular predicted structured RNA. Note that it is possible to have more than one structure per sequence.

E - *taxonomy.html*. Taxonomic overview of each RNA family found.

This page provides a complete evolutionary annotation for each RNA family assigned. The tool uses the *Krona* package (Ondov *et al.*, 2011) to generate interactive graphics for the visualization of the abundance of each RNA classes that belongs to different taxonomical groups. It is useful for studies on RNA families acquired over evolution, i.e. RNAs belonging to different evolutionary taxonomic groups.

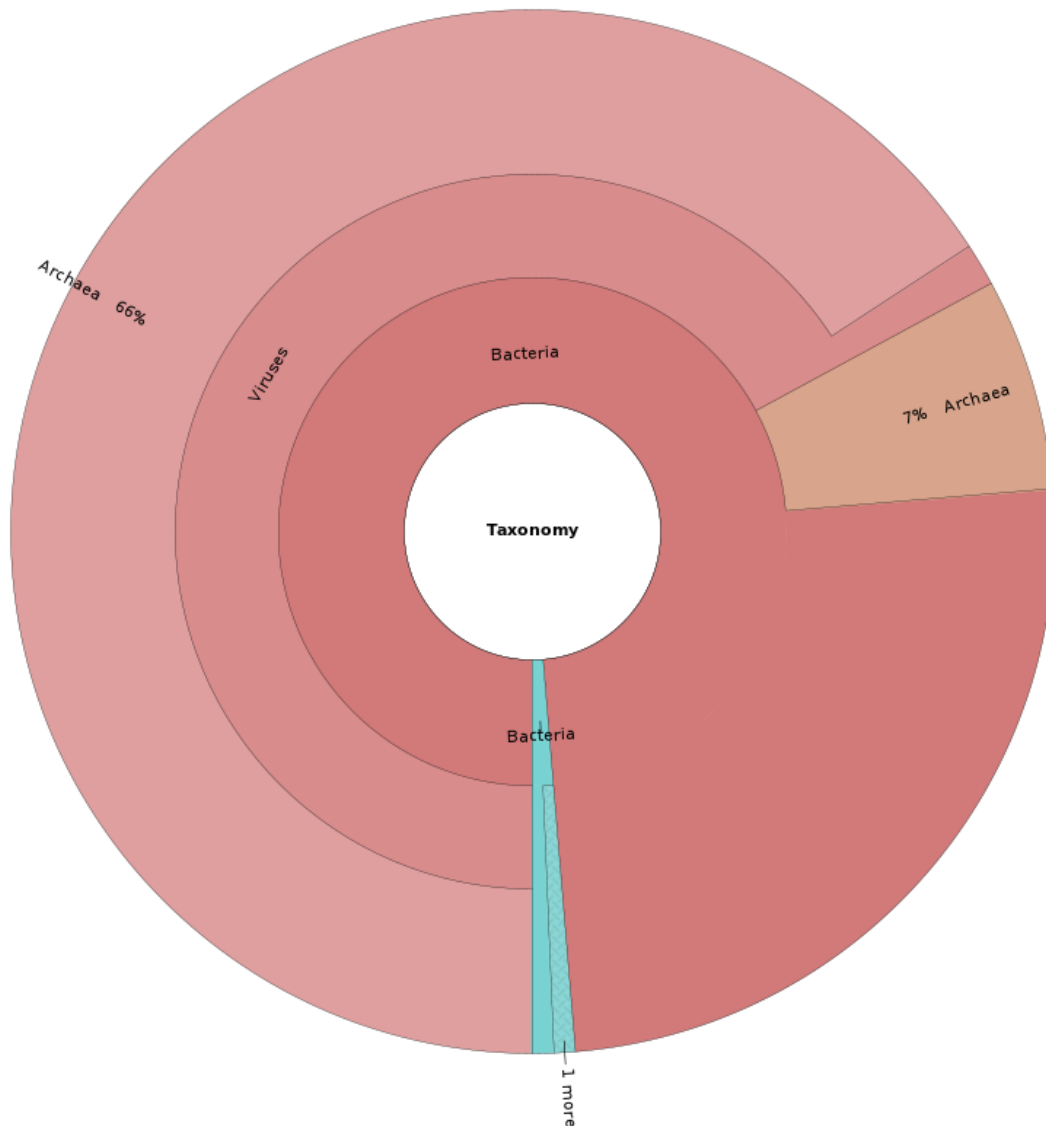


Figure 4 - Interactive graphics with the taxonomy information retrieved from Rfam database information from each RNA family predicted.

Running *structRNAfinder* in genome sequences - multiple significant hits

As mentioned before, *structRNAfinder* can also be used to predict and annotate non-coding RNAs in genomes or clusters of ncRNAs originated from the same RNA sequence. In this case is necessary use the “-r” option when running the tool. Here, it will be reported all significative hits for each sequence. The *cmsearch* algorithm from *Infernal* is used in these structural similarities searches.

Running *structRNAfinder* with other databases

If you want to run *structRNAfinder* using different covariance models than that provided by Rfam, you have to run it on the same way previously described. The only difference is observed on the html report, since it is not possible to retrieve additional information such as RNA family, taxonomy, ontology, etc.

References

- Burge,S.W. *et al.* (2012) Rfam 11.0: 10 years of RNA families. *Nucl. Acids Res.*, **41**, D226–236.
- Lorenz,R. *et al.* (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
- Nawrocki and Eddy. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**(22), 2933–2935.
- Ondov,B.D. *et al.* (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, **12**, 385.
- Pål Sætrom, Ragnhild Sneve, Knut I. Kristiansen, Ola Snøve, Thomas Grünfeld, Torbjørn Rognes, and Erling Seeberg. (2005). Predicting non-coding RNA genes in Escherichia coli with boosted genetic programming. *Nucleic Acids Research*. doi:10.1093/nar/gki644