

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

structRNAfinder

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

Content

1. Overview.....	3
1.1. Background	3
1.2. Terminology	3
1.3. Comments, criticisms and suggestions	4
1.4. Summary of available tools	4
2. Installation of structRNAfinder and required softwares	5
3. Usage	6
3.1. Before Running	6
3.1.1. Pipeline.....	6
3.1.2. CASE A: Using covariance models from Rfam.....	7
3.1.3. CASE B: Using a different covariance model.	17
3.1.4. Generated Files	18
3.2. Using structRNAfinder.....	19
4. References	21

1. Overview

1.1. Background

Noncoding RNAs (ncRNAs) are key regulators of many different biological processes in organisms from all kingdom of life (Amaral *et al.*, 2013; Toffano-Nioche *et al.*, 2013; Oliveira *et al.*, 2011). Its exploration is becoming routine in genomics and transcriptomics projects. As in proteins, their structure is directly associated to function. Thus, different classes of ncRNAs can be identified and functionally annotated based on specific characteristics derived from its predicted secondary structure.

Nowadays, there is large number of tools for RNA secondary structure predictions available (Machado-Lima *et al.*, 2008), most of them are focused exclusively on the identification of particular classes of ncRNAs (i.e. tRNAs, miRNAs, snoRNAs). Thus, for a full computational annotation of the repertoire of the noncoding RNA content in a genome or transcriptome, is necessary to use a variety of programs, and the plethora of input/output files in different formats and conventions, makes it difficult and complicate for professionals with little programming skills. Here, we describe *structRNAfinder*, an integrated and automated toolkit for ncRNA annotation based on secondary structure inference.

1.2. Terminology

- **noncoding RNAs (ncRNAs):** a functional RNA molecule that is not translated in a protein.
- **contigs:** assembled sequences from sequencing reads.
- **covariance models:** it is like a sequence profile, but scoring a combination of the consensus sequence and RNA secondary structure.
- **mature sequence:** the sequence region correspondent to the structure hits processed from longer sequences.

1.3. Comments, criticisms and suggestions

Please, feel free to contact us in case of comments, criticisms and suggestions at: **vinicius.coutinho@umayor.cl** or by accessing directly our GitHub page at: <https://github.com/viniciusmaracaja/structRNAfinder>.

1.4. Summary of available tools

Infernal software: tool for searching for RNA structures in nucleotide sequences (Nawrocki and Eddy, 2013)

RNAfold software: tool for secondary structure predictions in RNA or DNA sequences, based on the folding and calculation of the free energy viable to form the structure (Lorenz *et al.*, 2011).

structRNAfinder: it is the main script the of *StructRNAfinder*. It performs all the analysis by calling other subscripts related to each comparison.

SRF_Infernal2table: it generates a tab delimited file with the information related to the selected hits for secondary structures, based on covariance models comparisons against the primary sequences performed by Infernal.

SRF_extractMature: it extracts the mature sequence of the best hits filtered previously. The output is generated in two files in Fasta format, one with the complete sequence and another with the mature sequences of selected hits.

SRF_taxonomy: it uses the Krona tools (Ondov *et al.*, 2011) to generate dynamic graphics with the taxonomic assignation for all secondary structures based on each RNA family taxonomic distribution reported on Rfam (Griffiths-Jones, 2003).

SRF_generateHtml: it integrates all the results previously obtained with additional information extracted from Rfam database for each hit in a friendly output in html format.

SRF_generateHtmlOther: it integrates all the results previously obtained to generate the a friendly output in html format. This program is only used for other databases than Rfam.

SRF_generateJS: it generates the JavaScript files necessary to produce the summary chart on the html final report.

2. Installation of structRNAfinder and required softwares

The installation procedure is performed by a shell script that contains all instructions to install the required softwares, as well to download the Rfam databases. To install structRNAfinder and the requirement please type into a terminal:

```
sudo sh install.sh
```

The script needs the superuser privileges to be executed. At the time of execution it will be asking if the user wants to install or download any particular required software or database. The order for the installation is the follow:

1.- *Would you like to install/re-install Vienna? (y/n) [n]*

If you only press enter without put a letter the answer not (n) is taken by default.

In this part, the script downloads the source code of the Vienna package available at <http://www.tbi.univie.ac.at/RNA/download.php?id=viennarna-2.1.7>, and compiles it using the installation recommendation by the authors.

2.- *Would you like to install/re-install Infernal? (y/n) [n]*

If you only press enter without put a letter the answer not (n) is taken by default.

Here, the precompiled binary files are downloaded from <http://selab.janelia.org/software/infernal/infernal-1.1-linux-intel-gcc.tar.gz>, extracted into a temporary folder and then copied into the Linux default binary folder (/usr/local/bin).

3.- *Would you like to download/re-download Rfam.cm? (y/n) [y]*

If you only press enter without put a letter the answer yes (y) is taken by default.

Here, the Rfam database is downloaded from <ftp://selab.janelia.org/pub/rfam/rfam-11.0/Rfam.cm.gz>. Then, the necessary files to run cmscan using the cmpress program (Infernal recommendation) are generated. By default, the path to extract the database is the home folder of the current user. It can be changed by the user.

4.- *Would you like to install/re-install Bio::Graphics? (y/n) [y]*

If you only press enter without put a letter the answer yes (y) is taken by default.

Finally, the Perl libraries necessary to generate the images are installed. In this part, the script install the libraries libgd-perl and Bio::Graphics using cpan tool. If you already have it installed, or you want to install manually, please answer “n” (not).

3. Usage

3.1. Before Running

StructRNAfinder is designed to analyze data originated from transcriptome and genome projects. So, as a input it is required the sequences in Fasta format, as well a covariance model (CM format) related to the secondary structures used for comparison.

3.1.1. Pipeline

StructRNAfinder integrates two third-part softwares and other in-house scripts that allows to perform the noncoding RNA annotations based on secondary structure inference. This tool generates a friendly output of the obtained results of each step.

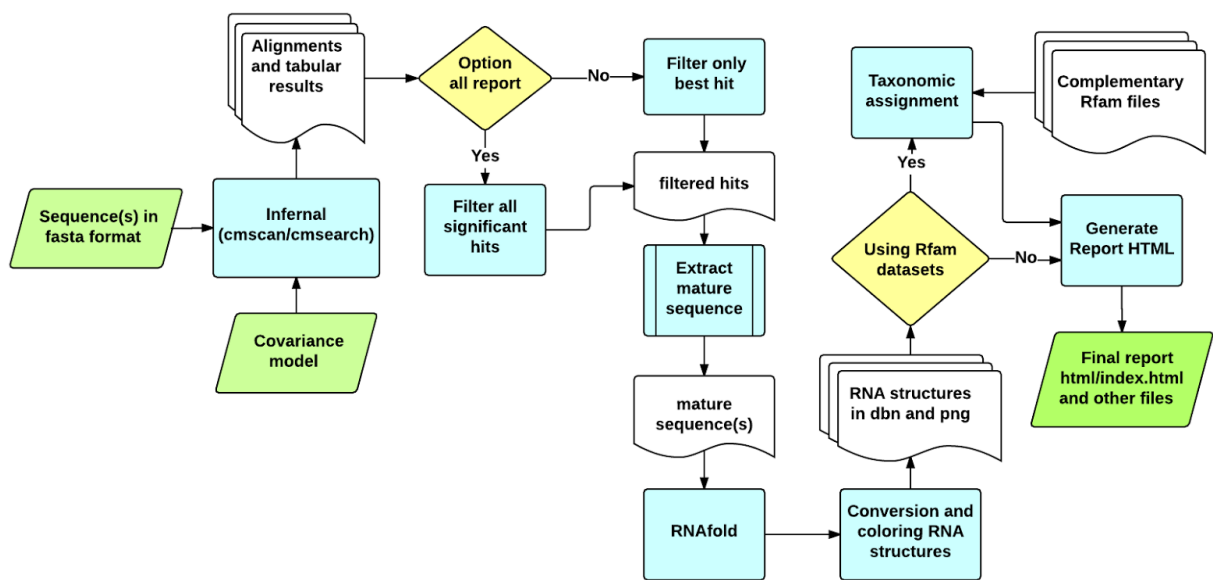


Fig. 1. Workflow explaining all steps implemented and performed by *StructRNAfinder*.

This program is optimized to use the covariance models available in the Rfam database. This mean that the generated output uses other complementary files, annotations and information extracted from that database. However, it do not exclude the possibility of using other covariance models as input, since the tool main pipeline is

conserved, the unique difference is on the step of generating the html final reports. This option is useful for comparative genomics analysis, i.e. using a sequence from one organism and CMs from another, as well implementing CMs originated by other softwares than Infernal.

At the following, we have two cases of use based on both methods: using or not CMs from Rfam database.

3.1.2.CASE A: Using covariance models from Rfam

Step 1: Running Infernal

With the aim to find structural homologies between the RNA/DNA sequence and the covariance models, StructRNAfinder makes use of Infernal. Infernal is a standard software used for this purpose. This program implements two methods to perform this task:

cmsearch: Searching covariance model against a sequence database.

cmscan: Searching sequences against a covariance model database.

The selection between one and another depends on user interest. So that, if you want more information you can explore the Infernal user's guide available in <http://selab.janelia.org/software/infernal/Userguide.pdf>.

Sorting and analyzing the obtained results is performed based on the classification proposed by Rfam, in which the different noncoding RNA families are divided in:

ncRNAs genes (Gene) - genes that produce functional RNAs instead of proteins:

Gene; antisense

Gene; antitoxin

Gene; CRISPR

Gene; lncRNA

Gene; miRNA

Gene; ribozyme

Gene; rRNA

Gene; snRNA

Gene; snoRNA

Gene; sRNA

Gene; tRNA

Gene; (it is classified as other if it does not have other classification).

Regulatory elements (Cis-reg) - RNAs that regulates the expression of closed located genes:

Cis-reg; frameshift_element

Cis-reg; IRES

Cis-reg; leader

Cis-reg; riboswitch

Cis-reg; thermoregulator

Cis-reg; (it is classified as other if it does not has other classification).

self-splicing intron: this group can catalyze their own removal from host transcripts (mRNA, tRNA and rRNA precursors) in a wide range organisms.

Step 2: Filtering significant hits and extracting the mature sequence

This step changes depending on user criteria. Here, we can obtain one hit per sequences (the best hit), or all good hits for that sequence (-r option). The “-r” option generates a complete report containing all significant hits for each sequence. The significant hits for each region is then selected, and its mature sequence extracted based on the position of the matching region from the structure covariance model on the sequence. Then, that region is compared to the complete target sequence and covariance model size, in order to expand it in order to achieve a mature sequence with the same length reported to the RNA family covariance model used on the comparison. Here, we can see two different possibilities (Fig. 2):

A: Expand the mature sequence. This is produced when the matching region is smaller than the covariance model and the target sequence. In this case, the mature sequence is expanded to equalize, if it is possible, the target length.

B: Keep the mature sequence length, since the match is completely inserted into the target sequence.

This procedure has the finality to obtain an optimum length to predict the secondary structure related to the RNA family covariance model in the next step.

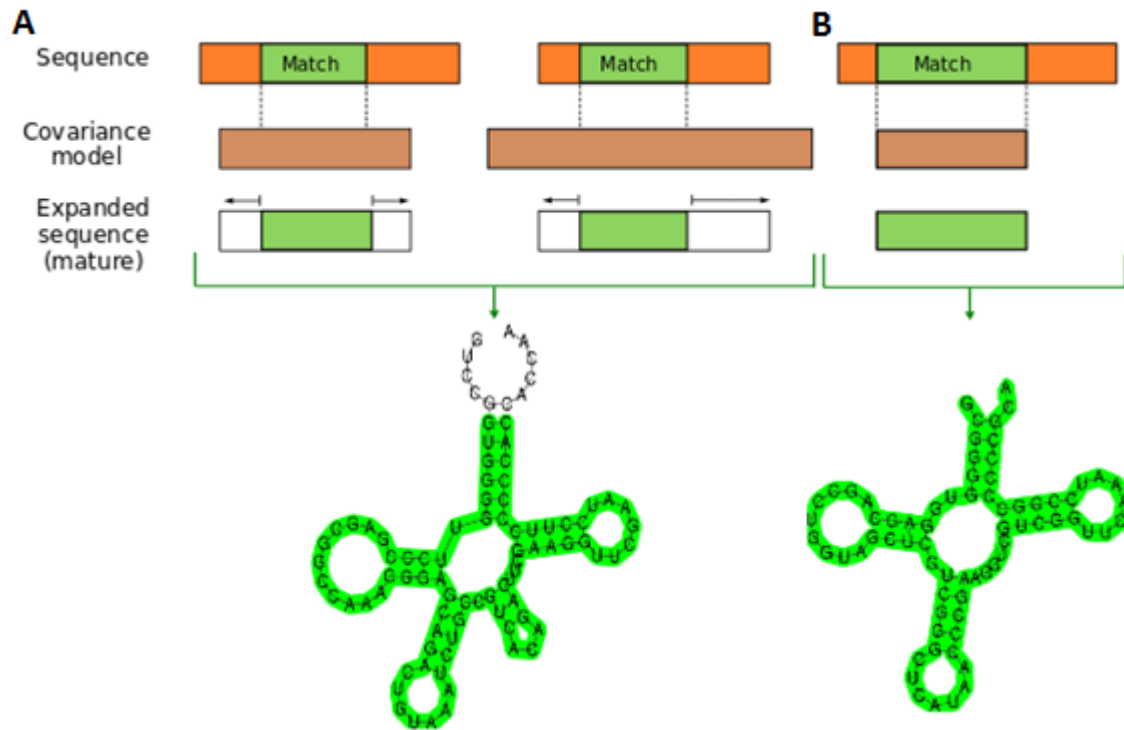


Fig. 2: Different matching cases illustrations.

Step 3: Running RNAfold

After obtained the mature sequence, the next step is to predict the secondary structure with the most reliable minimum energy. This is performed using the RNAfold program, part of the Vienna package (Lorenz *et al.*, 2011). It uses the mature sequence previously extracted from the original file and generates the predicted secondary structure. The output is a file with the structure and the energy of each sequence, as well the postscript (ps) files for each calculated structures. To more information about RNAfold, please visit:

<https://www.tbi.univie.ac.at/RNA/RNAfold.html>

With the aim of show the information in a more clear and comprehensive way, all postscript files are converted in a common image format (png). The match region is colored in green, in a way that the both expanded and matching regions between the query and target can be observed (Figure 2).

Step 4: Taxonomy assignation

From the repositories of Rfam database we extracted the taxonomic information for each assigned RNA families. It is summarised in the folder “/share/structRNAfinder” from the installation path of structRNAfinder. When running the tool, it cross-reference the predicted RNAs with that list, generating a new file (Rfam_specie.tab) with the taxonomic information that will be used in the final html reports.

Step 5: Generating html reports

With all the information collected through the tool workflow, it generates reports in html format. These are described in following:

index.html → this page contain a summary table of all annotated RNAs. It has the main data obtained through Infernal and RNAfold.



HOME

+ ANNOTATED RNAs

+ SUMMARY

TAXONOMY

CONTACT

Laboratory of Integrative Bioinformatics

Bioinformatics Core

Centro de Genômica e Bioinformática

Universidade Federal do Rio de Janeiro

		Infernal						RNAfold	
Sequence	RNA family	Id	From_seq	To_seq	Score	Evalue	Energy	Struct	
Others-cis									
Ec_rIT	RiT	RF00391	40	171	135.3	6.7e-32	-43.90		
Others-gene									
ffs	Bacteria_small_SRP	RF00169	33	129	65.5	1.4e-15	-26.70		
ssrS	6S	RF00013	1	183	125.0	2e-28	-75.60		
gcvB	GcvB	RF00022	1	205	176.8	5.6e-44	-55.50		
ssrA	tmRNA	RF00023	106	466	221.6	9.2e-66	-111.10		
antisense									
dicF	DicF	RF00039	2	53	75.8	1.8e-18	-10.70		
Ec_micF	MicF	RF00033	81	174	114.1	7.5e-20	-9.30		
antitoxin									
QUAD1a_RyeC	SIB_RNA	RF00113	1	116	114.7	6.2e-26	-51.00		

Fig. 3. Example of the main page (index.html), where the lists of all hits are showed.

It should be noted that the dynamically “Annotated RNAs” menu is create while executing StructRNAfinder. It changes depending on the analyzed sample and annotated RNA families.

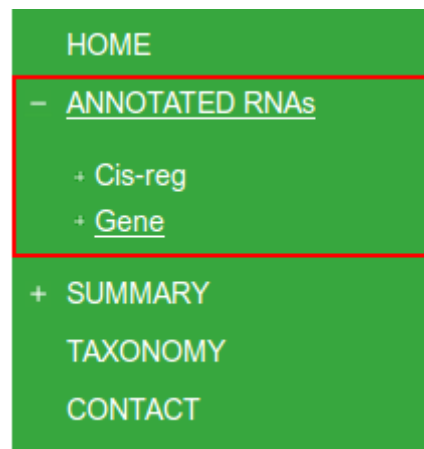


Fig.4. Menu generated automatically when executing StructRNAfinder.

sequenceName.html → The folder called “tables” has a file in html format for each annotated RNAs. In this table is available a detailed information extracted from Infernal, RNAfold, and the complete annotation obtained from Rfam database for each assigned structure.

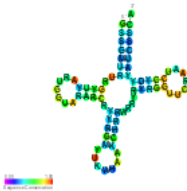

argQ		
cmscan-Rfam		
Family : tRNA	Id: RF00005	Type: Gene; tRNA;
Dominio: Eukaryota; Bacteria; Viruses; Archaea; unclassified	Ontology: GO:0030533 triplet codon-amino acid adaptor activity	
Description: Transfer RNA (tRNA) molecules are approximately 80 nucleotides in length. Their secondary structure includes four short double-helical elements and three loops (D, anti-codon, and T loops). Further hydrogen bonds mediate the characteristic L-shaped molecular structure. tRNAs have two regions of fundamental functional importance: the anti-codon, which is responsible for specific mRNA codon recognition, and the 3' end, to which the tRNAs corresponding amino acid is attached (by aminoacyl-tRNA synthetases). tRNAs cope with the degeneracy of the genetic code in two manners: having more than one tRNA (with a specific anti-codon) for a particular amino acid; and 'wobble' base-pairing, i.e. permitting non-standard base-pairing at the 3rd anti-codon position.		
Score: 66.4	E-value: 1.2e-15	
From sequence: 1	To sequence: 74	
Alignments: <pre> ((((((, <<<< >>>>, <<<< >>>>, , , , , <<<< >>>>))))))): NC RF00005 1 GgagauaUAGCucAgU.GGU.AgaGCguogGacUuaaAAuCogaagg.cgcggGUUCgAaUCCcgcuauucCa 71 G:A:::UA:CUCAG GG AGAG:++::G:CU AA:C:::GG CG::GGUUCGAUCC::C:::U:CA argQ 1 GCAUCCGUAGCUCAGCUGGAUAGAGUACUCGGCUACGAACCGAGCGGUCGAGGUUCGAAUCCUCCGGAUGCA 74 *****8888*****pp</pre>		
RNAfold		
<p>>argQ</p> <p>CAUCCGUAGCUCAGCUGGAUAGAGUACUCGGCUACGAACCGAGCGGUCGG AGGUUCGAAUCCUCCGGAUGCA ..(((((((...)))))...(((((.....))))).-))((((.....))))))... (-16.30)</p>		
Length sequence: 73	Length match: 72	

Fig. 5. Example of the table with a complete information for each annotated structure. The user can access it by clicking on the name of the sequence in the main page..

summary.html → The summary item in the menu is composed by two sections (Fig 6), Annotated RNAs and *Loci* distribution.

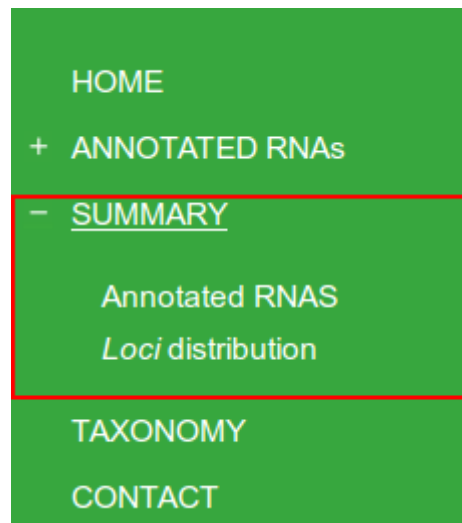


Fig. 6. Left-side menu. Here, it is possible to access further information related to the whole annotation. By clicking on Annotated RNAs, it will be provided a summary of all annotations performed. By clicking on *Loci* distribution, it is possible to have a global distribution of each predicted and annotated RNA along the sequences.

Annotated RNAs → This page contains a table with the general data obtained for all hits, followed by a pie chart with the percentage of each annotated ncRNA family. This graphic shows a general overview of the annotation process. A second table is available, where the user can have access to all files generated by structRNAfinder in FASTA and BED formats.

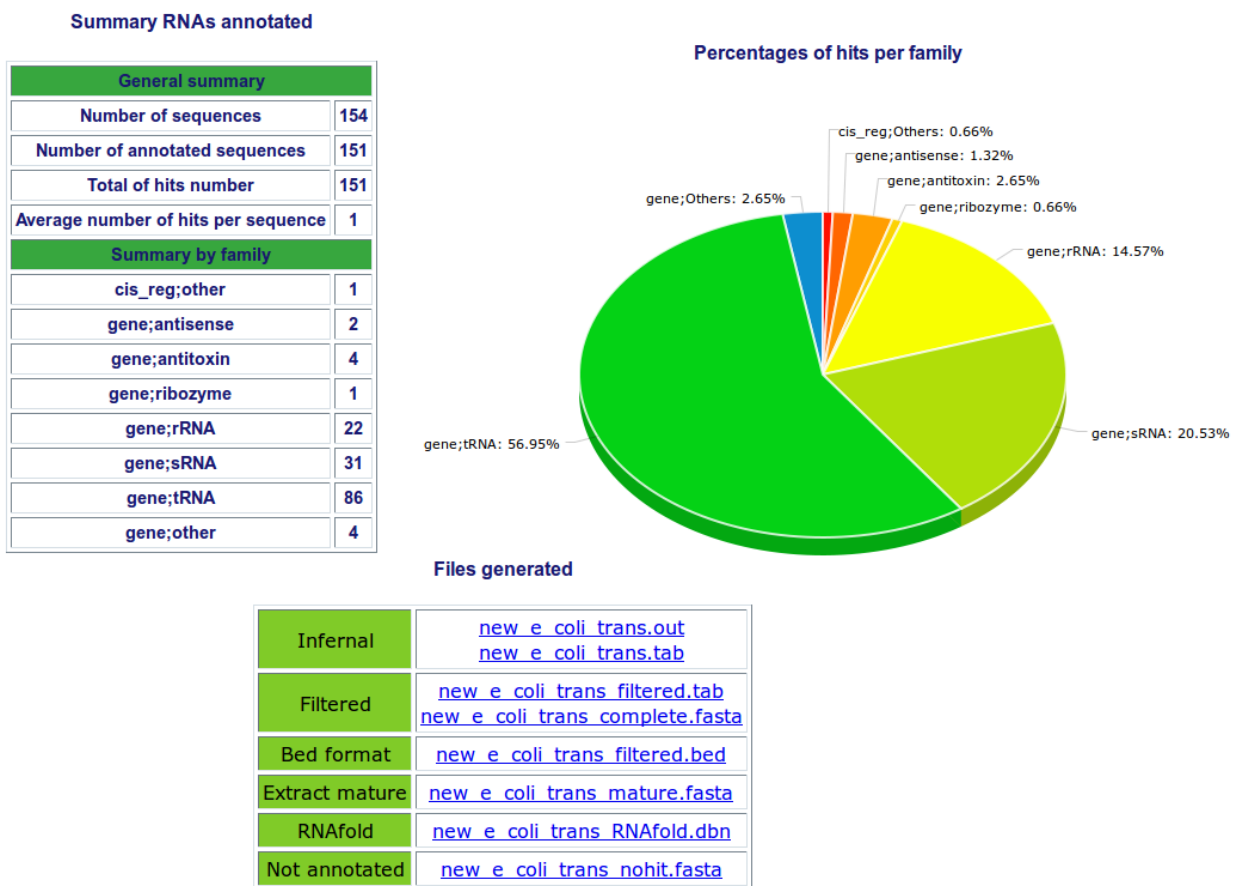


Fig. 7. Summary information of all hits found.

ncRNAs *loci* distributions → In this section, the tool presents an image with the general distribution of all matches for each sequence. It is useful for a visual distribution of clustered RNAs originated from the same precursors, or to have a general overview of the distribution of each RNA along a complete genome analyzed. The blue color blocks represents the hits found in the positive strand, while the sky-blue blocks represents the hits found in the negative strand.

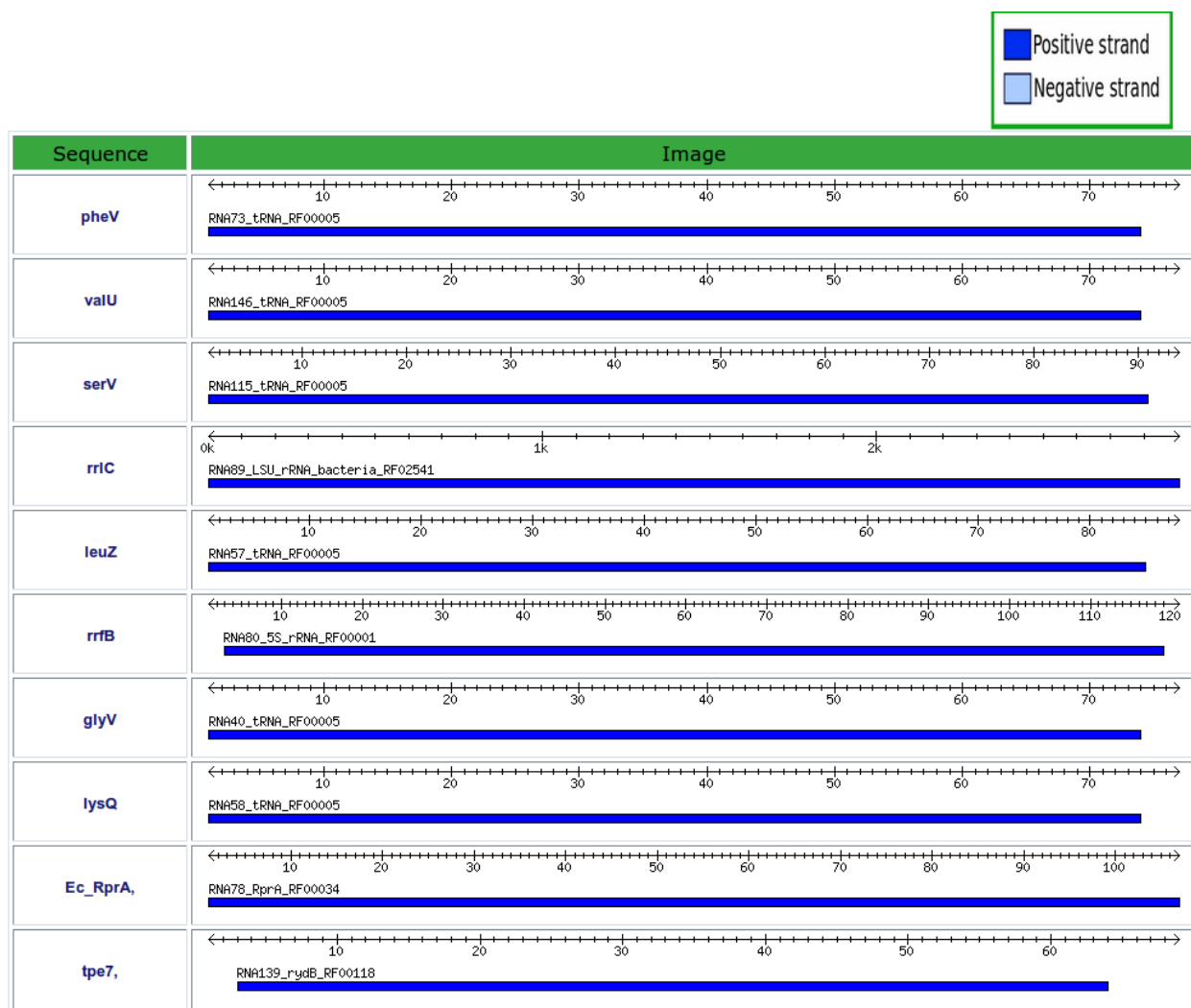


Fig. 8. The location of hits by each sequence is represented with blue line if was found in positive strand and sky blue if was found in negative strand.

taxonomy → Using the Krona library, *StructRNAfinder* generates an interactive graphic that allows the visualization of the taxonomic representation of the RNAs families predicted, based on the information extracted from Rfam.

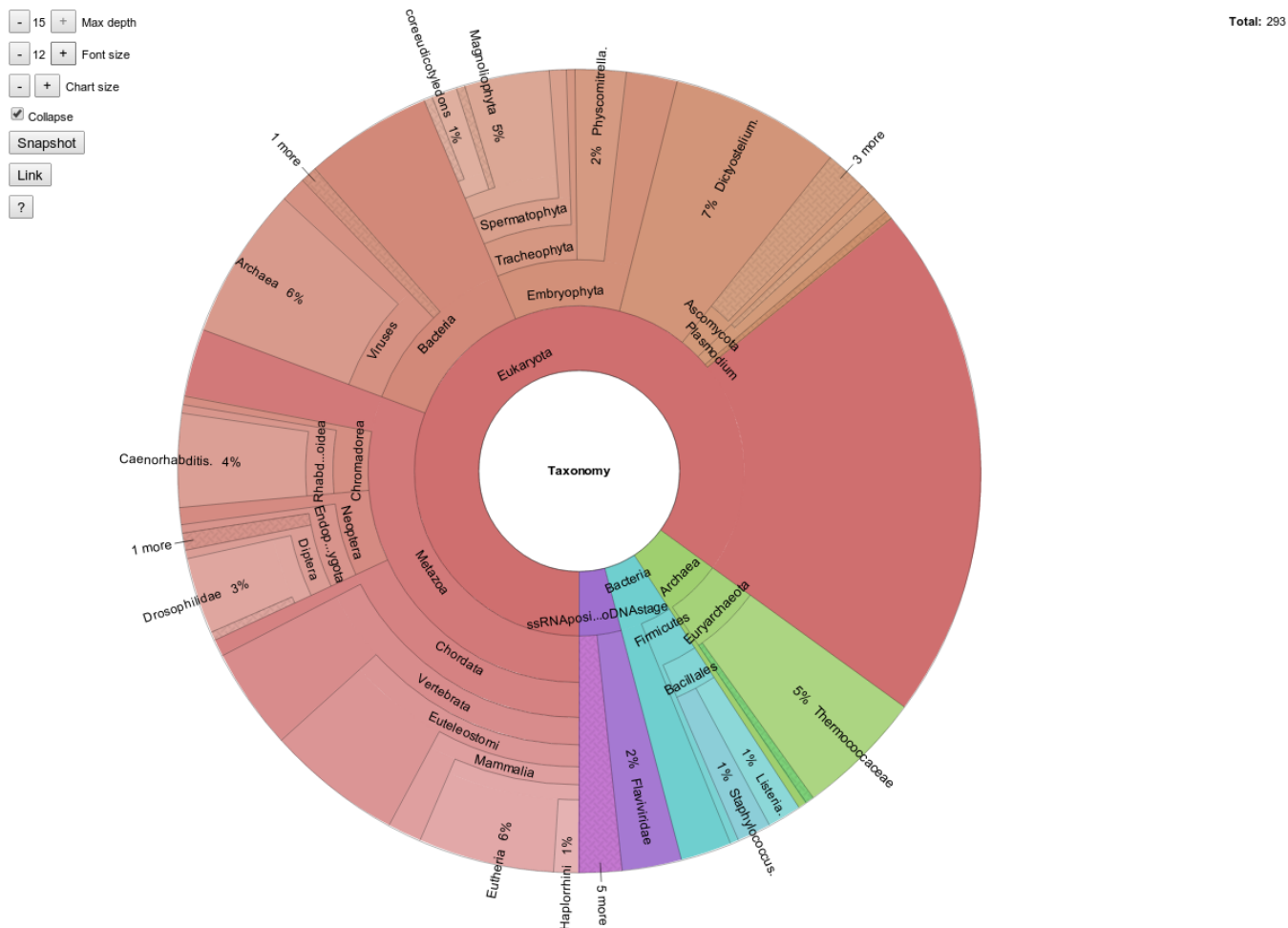


Fig. 9. Interactive graphics with the taxonomy information retrieved from Rfam database for each RNA family predicted.

3.1.3.CASE B: Using a different covariance model.

For the cases in which the user is not using the covariance models from Rfam, the protocol is almost the same. It runs Infernal, filter the significant hits, extracts the mature sequences and predict secondary structures using RNAfold. The main difference is on the final html reports, since it is not possible to obtain the complementary information available in Rfam database. So, it is not feasible to obtain the summary annotation and taxonomic graphics, and the generated tables have only the basic information extracted from the annotation of RNAs, i.e. a comparison between sequences and the matching structure. On the following we have some examples on the final report:

→ **index.html**: In this case, the main table is very similar to the report originated when using Rfam covariance models. The main difference is that the RNA structures are not sorted by families.



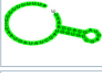


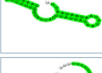





Sequence	Infernal						RNAfold	
	RNA family	Id	From_seq	To_seq	Score	Evalue	Score	Struct
GRMZM2G506611_T01	rlt51	RF01490	142	222	26.4	0.00066	-57.87	
GRMZM2G004842_T01	23S-methyl	RF01065	102	190	28.3	0.00035	-48.40	
GRMZM2G109779_T01	23S-methyl	RF01065	2	90	23.4	0.0039	-54.50	
GRMZM2G029892_T01	Histone3	RF00032	318	273	21.2	0.0017	-9.20	
GRMZM2G546516_T01	Histone3	RF00032	318	274	20.4	0.0031	-9.20	
GRMZM2G578769_T01	Histone3	RF00032	54	99	18.6	0.0034	-9.90	
GRMZM2G587775_T01	Histone3	RF00032	24	69	20.2	0.0012	-7.30	
GRMZM2G519705_T01	rox2	RF01666	79	12	22.1	0.002	-20.90	
GRMZM2G527208_T01	rox2	RF01666	79	12	22.1	0.002	-20.90	
GRMZM2G543836_T01	rox2	RF01666	79	12	22.1	0.002	-20.90	
GRMZM2G373179_T01	Spi-1	RF00232	163	254	21.9	0.0024	-77.50	

Fig. 10. Main table generated using other covariance model. The difference more notorious is that the hits are not classified by families.

GRMZM2G506611_T01	
cmscan-Rfam	
Family : rli51	Id:
Score: 26.4	E-value: 0.00066
From sequence: 142	To sequence: 222
Alignments:	
<pre> v v v v NC ::::::::::{(((~((((((((((((((((((,,,,,*****)))))))))}}))})-))) CS rli51 1 AUAUuacAaaGUuuaaGcCaCcuauaguuUCUAC*[9]**[34]*UCCUUAAcuCUCuCgUuAaaUagcuauagGuGgCuuaaaaCuu 121 A A A::U::A::CCAC U::AG:U C AC AAC U UCUG U A:CU::A GUGG::U::A: : GRMZM2G506611_T01 142 AAAA AUGAGGAUAGACCCCACCUGCAGGUGCAAC*[7]**[4]*-----AACGUUAUCUGAU---AGACCUGCAUGGGGGUCUAUAC 222 *****998..8....4.....3444444444444...45***** pp </pre>	
RNAfold	
<p><GRMZM2G506611_T01 AAUGUUUGCAGCCCCUCACAAAAAUGAGGAUAGACCCCACCUGCAGGUGC AACACACAAUUAACCGUUAUCUGAUAGACCUGCAUGUGGGGUCUAUAC UCUUUUUUGUGAGAGGGGCCU ((((((((((((((-(((((.....)))))).)))))).-.))))).)))))) (-57.87)</p>	
Length sequence : 120	
Length match: 70	

For more information about the generating of covariance models, please visit:
<https://sites.google.com/site/rnainformatics/practical-building-covariance-models>.

3.1.4.Generated Files

StructRNAfinder generates several files during the whole process that can be useful when exploring more deeply the data. Here we have a list of all generated files:

→ Files generated during the homologous search through Infernal:

File.tab : It is divided in different fields. First a header, where is described the problem and the options used. Second, a list containing the best hits founded.

File.out : This file have the previous information plus the alignment founded between the target and query.

For a more detailed description, please visit:

<http://selab.janelia.org/software/infernal/Userguide.pdf>

→ Filtering and extracting the mature sequence

File_filtered.tab: The filtered file from the original file.tab that contain only the list of best hit for each sequence, and its corresponding information. In the case that you selected all report option (-r) this file contain the list of all significant hits for each sequence.

File_complete.fasta: File in fasta format with the best hits filtered, with the complete sequence (not only the matching region). This file is only generated if the “-r” is not defined.

File_mature.fasta: File in fasta format of the filtered hits with the mature sequences extracted from each sequence.

File_filtered.bed: File in bed6 format with the filtered hit.

→ secondary structure files using RNAfold

RNAfold.out: This document contains the sequence and the secondary structures predicted for each sequence.

RNAs.png: In the “images” folder is available png files with figures for the predicted secondary structures for each sequence.

3.2. Using structRNAfinder

Usage: **structRNAfinder** [OPTIONS] -i sequence.fasta -d covarianceModel.cm

To run structRNAfinder, the user must use the following required options: the sequence in fasta format (-i) and the covariance model (-d). Additional options are optional, such as:

-m or --method - method to search for structural RNAs in a sequence dataset, cmscan or cmsearch [default: cmscan]

To find more information on the structural similarities searches available in Infernal please visit:

<http://selab.janelia.org/software/infernal/Userguide.pdf>

-d or --database - covariance models reference database

-n or --otherDB - when using other covariance model reference reference than Rfam [default: false]

To indicate the covariance model to be used in the comparisons use the option “-d” or “-database”. If the models are not the Rfam ones, use the option “-n” or “--otherDB”.

-r or --report - report all significative annotated RNAs. By default only is showed the best hit per sequence [default: False]

To indicate if you want to see all hits for each sequence in the final results. If it is not used, only the best hit for each sequence is shown.

-s or --score - minimum score to each hit [default: 10]

-e or --e-value - maximum e-value to each hit [default: 0.01]

To modify the score and the e-value that are used to filter the significance thresholds in the the analysis performed by Infernal.

-t or --tblout save parseable table of hits to file

-o or --output direct output to file <f>, not stdout

Name of the output files of Infernal. When it is not specify a name, the tool keeps the same name of the files from the input sequences, changing the extension to .tab and .out, respectively.

-c or --cpu number of parallel CPUs to use for multi-threads.

If it is specified the number of threads to be used, by default is 1.

4. References

Amaral,P.P. *et al.* (2013) noncoding RNAs in homeostasis, disease and stress responses: an evolutionary perspective. *Brief. Funct. Genomics*, 12(3), 254–278.

Burge,S.W. *et al.* (2012) Rfam 11.0: 10 years of RNA families. *Nucl. Acids Res.*, **41**, D226–236.

E. P. Nawrocki and S. R. Eddy, Infernal 1.1: 100-fold faster RNA homology searches , *Bioinformatics* 29:2933-2935 (2013).

Lorenz,R. *et al.* (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, 6, 26.

Machado-Lima,A. *et al.* (2008) Computational methods in noncoding RNA research. *J. Math. Biol.*, 56, 15–49.

Ondov BD, Bergman NH, and Phillippy AM. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*. 2011 Sep 30; 12(1):385.