

Rfam 3D and R-scape updates

Nancy Ontiveros

nancyontiveros@ebi.ac.uk

Benasque, 16 August 2022



Session overview

- Rfam introduction
- Rfam SEEDs
- 3D updates
- R-scape improvements
- Wish list, possible directions and get involved



Rfam

The non-coding RNA families database

4,094 families

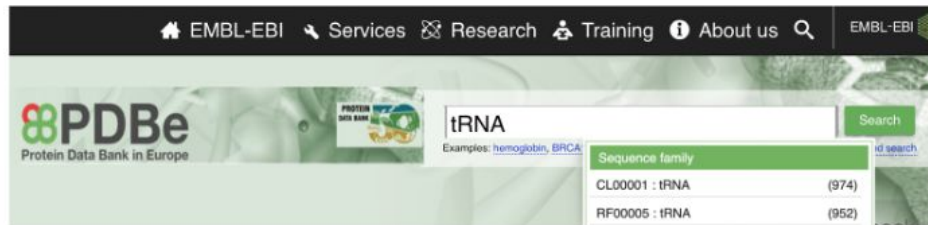
rfam.org

Rfam main propose

Main use of Rfam families is to be a reference to annotate genome datasets, ncRNA families can be found using **Rfam** and **Infernal**

Other uses of Rfam

- **Training sets:** Rfam is also used for algorithm development.
- Browse ncRNA families and know more about them
- Rfam provide identifiers for ncRNA families **Rfam ID**, for example in PDBe



The screenshot shows the PDBe website interface. At the top, there is a navigation bar with links for EMBL-EBI, Services, Research, Training, and About us. Below the navigation bar, the PDBe logo is visible on the left. In the center, there is a search bar with the text 'tRNA' entered. To the right of the search bar, there is a 'Search' button. Below the search bar, there is a table with the following data:

Sequence family	
CL00001 : tRNA	(974)
RF00005 : tRNA	(952)

Where do Rfam families come from?

- Sequences are usually taken from the literature
- From direct submissions from our users/experts
 - **Virus families** - Manja Marz, Kevin Lamkiewicz and Sandra Triebel
 - **xRNAs in Potato virus** - Quentin Vicens
 - **Bacteroidetes families** - Lars Barquis
 - **Hovlinc** - Fei Qi
- Or from expert databases
 - **ZWD** - Zasha Weinberg (ncRNAs from metagenomics)
 - **miRBase** - Sam Griffiths-Jones (micro RNAs)

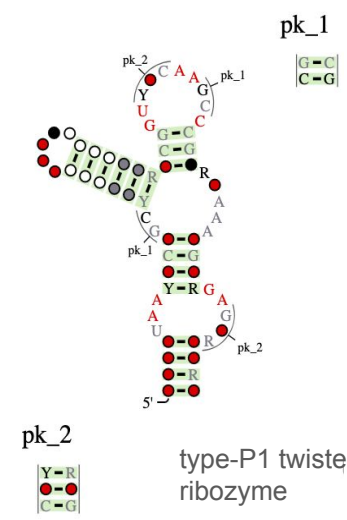
All SEED elements are important in an Rfam family

```
# STOCKHOLM 1.0

#=GF AC RF03160
#=GF ID twister-P1
#=GF DE type-P1 twister ribozyme

URS0000689C5_6183/1-59 CCUG-UAA-CUCCUCGG AUAAA CUGCUGGUCCAAGC-CGAG AUAAA GGAG-GAGGG UUGG
URS000067DD8_6183/1-73 GGGC-UAA-CGCCCGCUGUAGCUC GUAAA GAGUUAUCUGCCGUCCAAGC-CGG GUAAA GGAG-GAGGGUUG-GGCA
URS000068BB1_12908/1-59 AAUU-UAA-UGCAACUGU AUAAA ACAGCAGUGGCAAGU-CCGU AUAAA UGCA-GAGAC AAC
URS000067361_12908/1-59 AAUU-UAA-UGCAACUGU AUAAA ACAGCGUGGCAAGU-CCGU AUAAA UGCA-GAGAC AAC
URS000068EBB_12908/1-52 CCCU-UAA-UGCAGC GUAAA GCGGUGACAAGC-CCGU GUAAA UGCA-GAGUCA AGGG
URS000066DB9_12908/19-74 UCCU-UAA-UGCAGUCC AUAAA GGAACGGUCACAAGC-CGU AUAAA UGCA-GAGUG AGGA
URS000066AC7B_12908/1-53 AUGU-UAA-UGCAGCC GUAAA GCGGUUAACAAGC-CGU GUAAA UGCA-GAGUA ACAU
URS00006919D_12908/1-69 UUGU-UAA-UGUAGCCUAUAUUAU AUAAA AAUUAUAUAGCGGUACAAGC-CG AUAAA UACA-GAGUU GUAA
URS000066C991_12908/1-73 AGAU-UAA-UGUAGCCAUUGUAU-AG AUAAA UUGAUACAUGAGAGAGUUAAGC-CUCU AUAAA UACA-GAAGA AUGU
URS000066C302_7029/1-66 UUUU-UAA-CCAAGCAAAC AAUAA GUUGACAGUCCUAAGC-CUGU AAUAA UUGG-GAAGG AAAA
URS000066D3CD_12908/1-59 ACCG-UAA-UGCAGCUAC GAAAA GUAGCCAGUCCAAGC-CUGG GAAAA UGCA-GAGGG CGGA
URS000066C5F3_12908/18-76 ACCG-UAA-UGCAGCUAC GAAAA GUAGCCAGUCCAAGC-CUGG GAAAA UGCA-GAGGG CGGA
URS00006678DB_12908/21-80 CCGG-UAA-UGCAGCUAC AAGAAA GUAGCCAGUCCAAGC-CUGG AAGAAA UGCA-GAGGG CGGA
URS00006690B_12908/1-60 AGCG-UAA-UGCAGCUAC GUAAA GUAGCCAGUCCUAAGC-CUGG GUAAA UGCA-GAAGG AAC
URS0000669CC_12908/1-60 AGCG-UAA-UGCAGCUAC GUAAA GUAGCCAGUCCUAAGC-CUGG GUAAA UGCA-GAAGG GAC
URS0000668C07_12908/1-60 AGCG-UAA-UGCAGCUAC AUAAA GUAGCCAGUCCUAAGC-CUGG AUAAA UGCA-GAAGG CGAA
URS0000668B2A_12908/17-75 CCGG-UAA-UGCAGCUAC GUAAA GUAGCCAGUCCUAAGC-CUGG GUAAA UGCA-GAAGG CGAA
URS0000671B3_12908/1-60 AGCG-UAA-UGCAGCUAC GAAAA GUAGCCAGUCCUAAGC-CUGG AUAAA UGCA-GAAGG CGAA
URS000066A88F_12908/31-99 CCGG-UAA-UGCAGCUAC GAAAA GUAGCCAGUCCUAAGC-CUGG GAAAA UGCA-GAGUC AGGC
URS000066D268_12908/1-71 CCUC-UAA-UGCAGCUCCGCGU AA GUAGCCAGUCCAAGC-CGGA AAA UGCA-GAGGG GAGA
URS0000665BCB_12908/1-62 ACUG-UAA-UGCAGCUCC AA GGGGAGCGGUUUAAGC-CUU AA UGCA-GAACGA CAGG
URS000066A582_12908/23-73 CUUG-UAA-UGCAGC GUAAA ACAGUGACAAGC-CUGU GUAAA UGCA-GAGU CAAA
URS0000669FD5_12908/1-54 GUUG-UAA-UGCAGC GUAAAAA GCGGUCACAAGC-CGG GUAAAAA UGCA-GAGUG CAAC
URS000066ACCD_12908/1-54 CUCU-UAA-GGCACC AUAAA GUCAGUGACAAGC-CUGU AUAAA UGCU-GAGUCA AGAG
URS000066A1B6_12908/24-75 UCUU-UAA-UGCUAC AUAAA ACAGUGACAAGU-CUG AUAAA UGCA-GAGUCA AAGA
URS000066C5EF_12908/17-67 UCUU-UAA-UGCUAC AAGA CCGUUAACAAGC-CGG AAGA UGCA-GAGGA CAGA
URS0000665F92_12908/17-72 ACUG-UAA-UGCAGC AUUGAGAAAA CCGUUAACAAGC-CGG AUUGAGAAAA UGCA-GAGGA CAGA
URS000066A42F_12908/21-75 UUAG-UAA-UGUGGC UUGAAAAA ACAGUUGACAAGC-CUGU UUGAAAAA CACA-GAGCA CUA
URS0000668F1F_12908/25-79 UCUU-UAA-UGUGAC AUUGAAAAA GGAGUUGUAAGU-CUCC AUUGAAAAA CACA-GAACA CAGA
URS000066C8BE_12908/30-81 CUUU-UAA-UGCCAC AUAAA CCGGUUGCAAGU-CGG AUAAA GGC-GAGCA GAAA
URS000066CC91_12908/1-60 UCAU-UAA-UGCGAUAC GUAAA GUUAAGGUUAAGU-CUU GUAAA CGCA-GAUAG AUGA
URS000066A3F9_12908/24-83 CUUU-UAA-UACAGUGGU GUAAAU GCCAAAGGUCACAAGC-CUU GUAAAU UGUA-GAGUG AAGU
URS00006691F4_12908/1-55 GUUG-CAA-CUCUAC AUGAGAAA CCAGUUGCAAGU-CUGG AUGAGAAA AGAG-GAGCA CAAC
URS000066B302_12908/1-54 GUUG-CAA-CUCUAC AUGAGAAA CCAGUUGCAAGU-CUGG AUGAGAAA AGAG-GAGCA CAAC
URS0000667BFD_12908/1-55 GUGC-UAA-CCAUAC AUGAGAAA CCGUUAACAAGC-CGG AUGAGAAA AUGG-GAGGG CGAC
URS0000668282_12908/1-61 GUUG-UAA-ACAAGCACU AGAAA AGUGAAGGUCCAAGC-CUU AGAAA GUGU-GAGGC CAAC
URS0000667FC0_12908/1-63 GGUG-UAA-CACGGCUACGG GUAAAC CCGUUAAGGUUAAGU-CUU GUAAAC UGUG-GAUGA CACC
URS000066AA51_12908/1-63 GGUG-UAA-CACGGCUACGG GGAAC CCGUUAAGGUUAAGU-CUU GGAAC UGUG-GAUGA CACC
URS00006695F0_12908/23-89 GAAA-UAA-UGUCUACAGACU AUGAA AGUCUGCCGUUAACAAGU-CGGG AUGAA AGCA-GAGGA UUUC
URS000066AF4C_6183/1-72 CUCU-CAA-CUCCGCUUAGCUCC GUAAA GGGUUAUCUGCCGUCCAAGC-CGG GUAAA GGAG-GAGGU CGGG
URS000066C384_12908/1-78 UUUU-UAA-CCAGCCACUAGCAUUGACA AGAAAA UGUC-GUGUUUGCCGUCUCCAAGC-CGG AGAAAA UGGG-GAGGU UUUU
(((.....(((BB<<<<<<.....>>>>>>AAA bb.>>>.....))))..aaa.....)))
uuuu.UAA.uGCaGcGaguaucu.....AuAAA.....gauacuGcCGGUCCAAGC.CCGG..AuAAA.....uGCa.GAGGG...aaaa
//
```

- Number of sequence
- Alignment
- Secondary Structure reference (SS_cons)
- Reference sequence (RF)



Ongoing projects to update Rfam families

3D Integration



Biotechnology and
Biological Sciences
Research Council

126 Rfam families map 3D
information, 30 of them are already
reviewed and updated

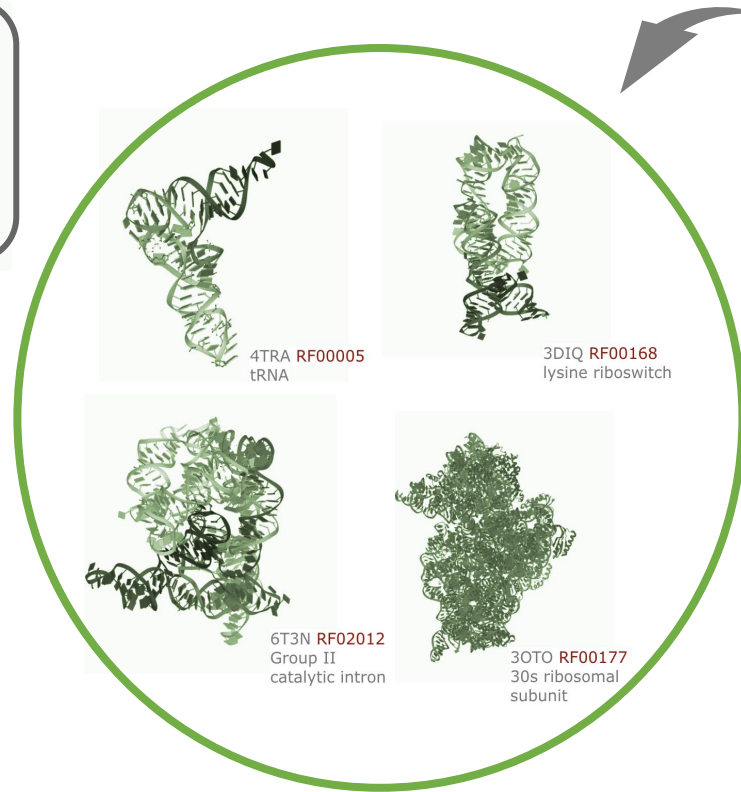
Update with R-scape



Dr. Elena Rivas

We analysed all Rfam families and
30 families have been selected
and updated with R-scape model

Rfam is updating families using 3D structures

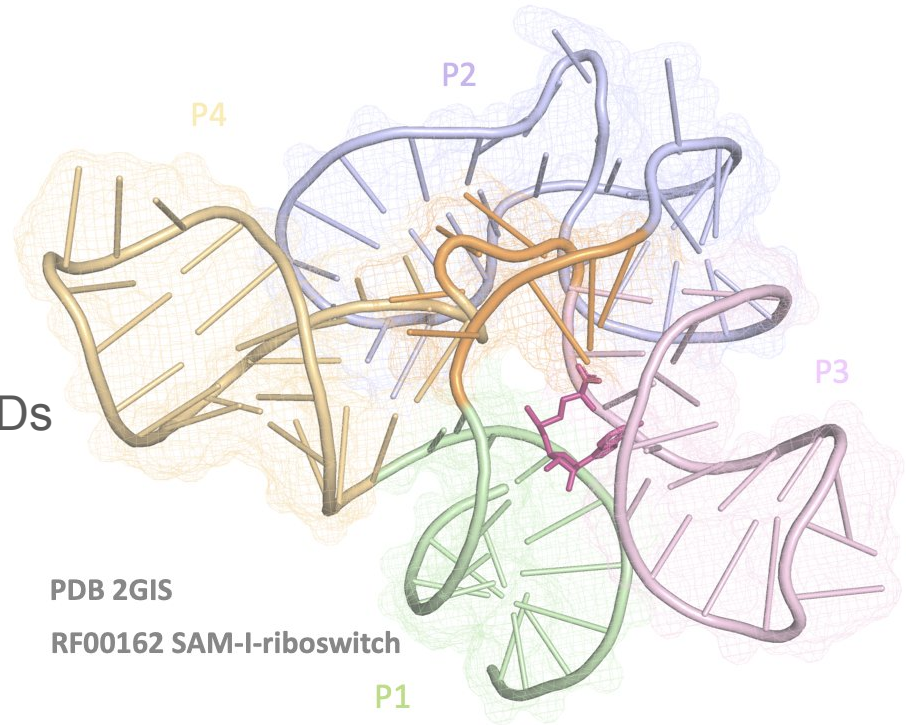


Rfam families
were created
prior to 3D
structure
determination

Steps to improve Rfam alignments with 3D



1. Align PDB sequences in SEEDs
2. Review secondary structure
3. Update family

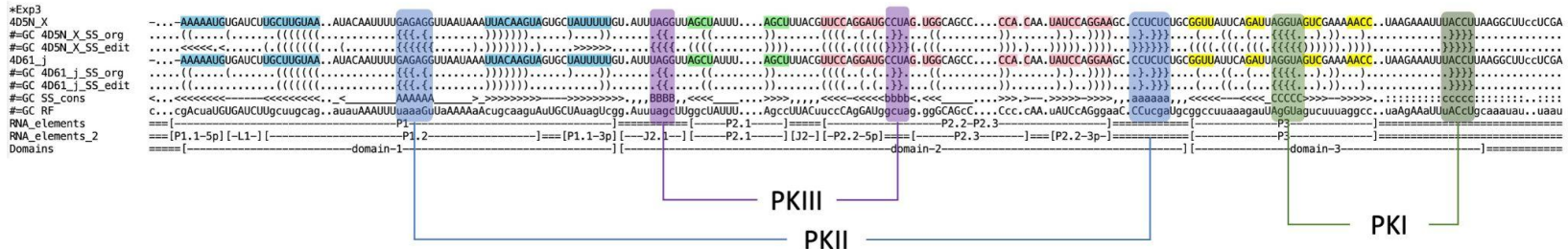
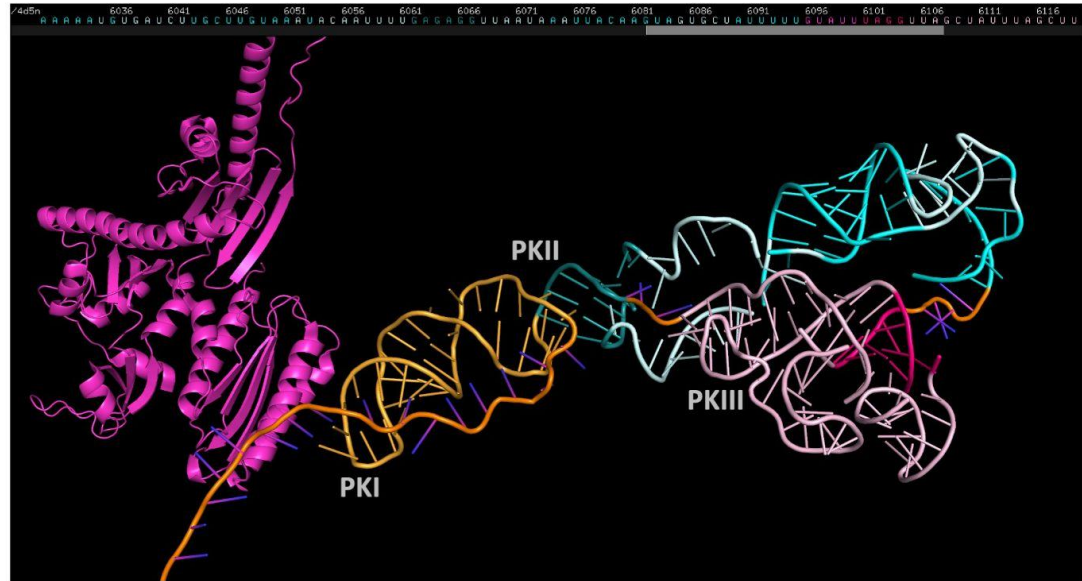


Step 2. Review secondary structure from 3D

Family:
IRES_Cripavirus
(RF00458)

Description: *Cripavirus* internal ribosome entry site (*IRES*)

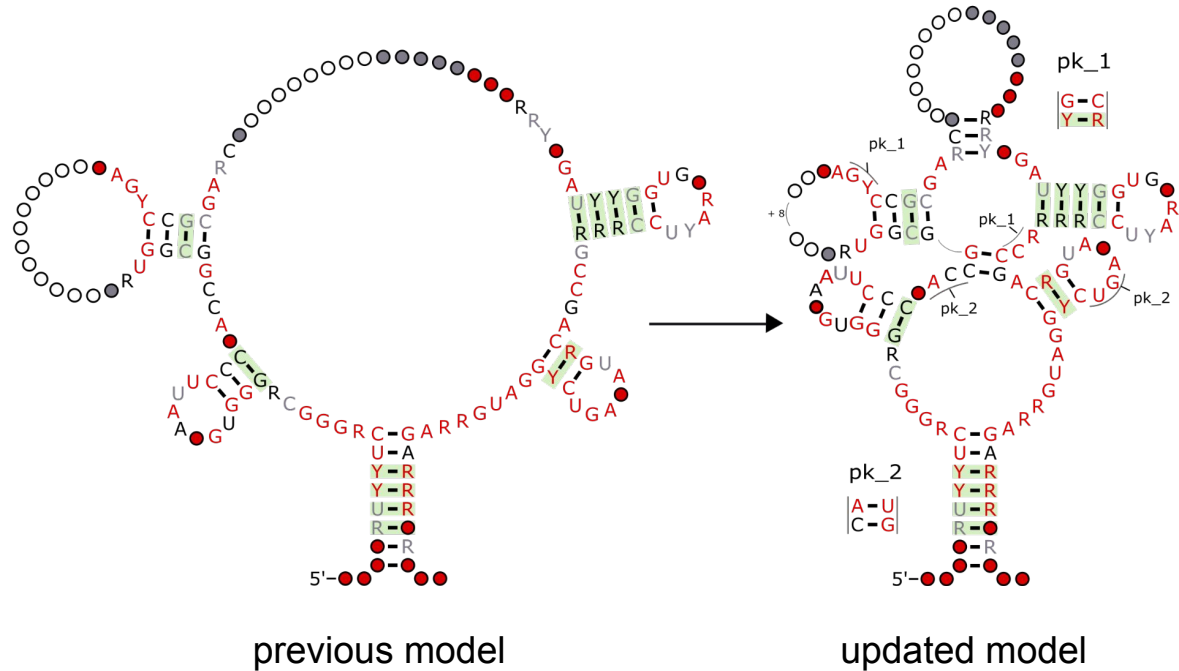
17 16 13
sequ spec struc



Step 3. Update 2D consensus in Rfam

- Add, remove or correct base pairs
- Include missing structures like pseudoknots

RF00050 FMN riboswitch



What kind of issues do we deal with

- Inconsistent secondary structures → manually eliminated
- Inconsistent reference sequence → need a refinement with Infernal
- Modified nucleotides → manually corrected in the secondary structure consensus
- Pseudoknots need to be include manually in the secondary structure consensus
- Chimeric structures that do not reflect the correct secondary structure

Rfam families updated using R-scape

Update with R-scape



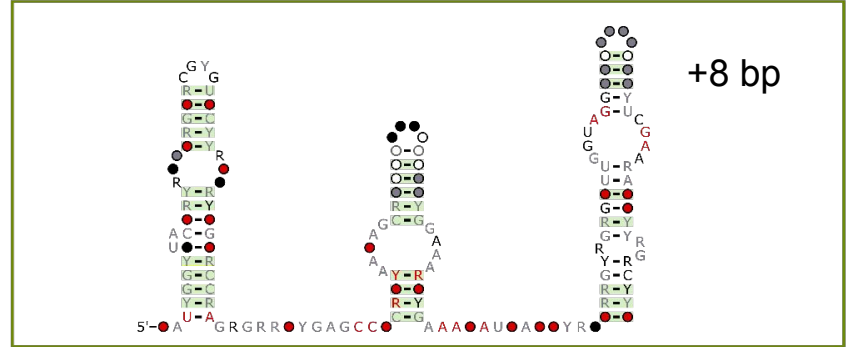
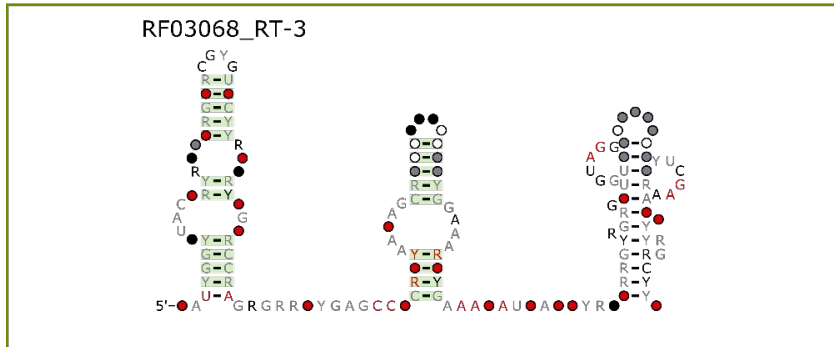
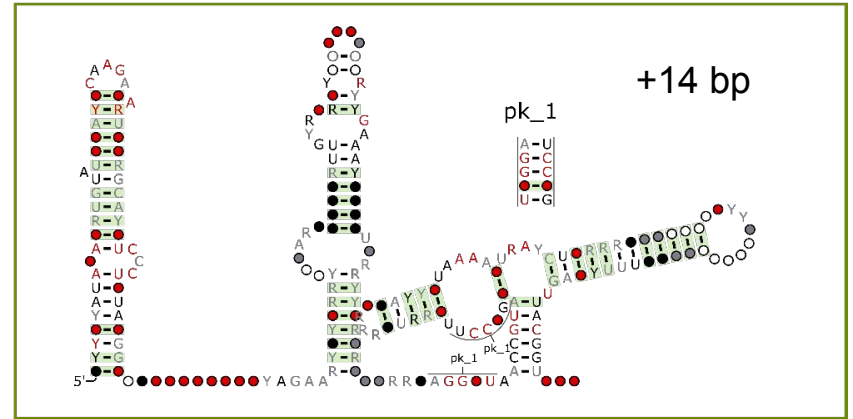
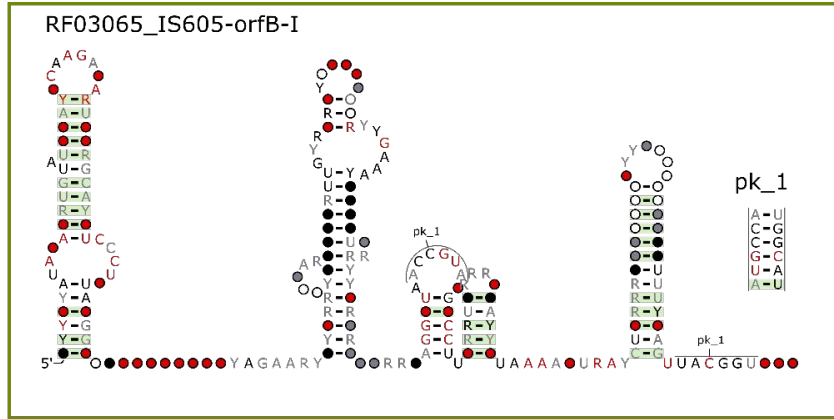
Dr. Elena Rivas

Rfam option (`--Rfam`)

to avoid tr,sc and other base pairs that Rfam cannot use

Family	Rfam Basepairs	R-scape Basepairs	Improvement	Usable	Covariance Rfam vs Rscape	win cov	Usable
RF02033_HEARO	31	55	24	✓	31/55	24	✓
RF03065_IS605-orfB-I	27	41	14	✓	27/41	14	✓
RF02969_DUF3800-I	28	40	12	✓	28/32	4	✓
RF02913_pemK	15	25	10	✓	15/17	2	✓
RF01867_CC2171	6	16	10	✓	5/6	1	✓
RF02221_sRNA-Xcc1	16	25	9	✓	16/17	1	✓
RF03068_RT-3	20	28	8	✓	20/28	8	✓
RF03072_raiA	36	44	8	✓	36/42	6	✓
RF03135_L4-Archaeoglobi	21	29	8	✓	21/25	4	✓
RF03064_RAGATH-18	17	24	7	✓	17/17	0	✗
RF02987_GA-cis	16	22	6	✓	16/17	1	✓
RF03077_RT-2	35	40	5	✓	35/38	3	✓
RF02005_group-II-D1D4-6	47	52	5	✓	47/50	3	✓
RF02944_c4-2	24	29	5	✓	24/25	1	✓
RF02968_DUF3800-IX	18	22	4	✓	18/19	1	✓
RF01688_Actino-pnp	9	12	3	✓	09/12	3	✓
RF03144_eL15-Euryarchaeota	11	14	3	✓	11/14	3	✓
RF01731_TwoAYGGAY	42	45	3	✓	42/44	2	✓
RF01794_sok	15	18	3	✓	15/17	2	✓
RF03158_L31-Actinobacteria	9	12	3	✓	9/11	2	✓
RF02004_group-II-D1D4-5	44	47	3	✓	44/45	1	✓
RF02947_cow-rumen-2	16	19	3	✓	16/17	1	✓
RF00062_HgcC	1	4	3	✓	1/2	1	✓
RF01864_plasmodium_snoR21	0	3	3	✓	0/1	1	✓
RF03046_Pseudomonadales-1	26	29	3	✓	26/27	1	✓
RF03019_RT-16	32	35	3	✓	32/33	1	✓

orfB-I and RT-3 updated with R-scape model





Wish list - what do users want in Rfam

- Include non Watson Crick base pairs
- Include long interactions (viruses)
- Update more frequently
- Integrate chemical probing data (SHAPE data)
- Variants, human ncRNAs and diseases
- Include protein binding sites (crosslinking data)
- Include other structural information (like triplets)

Possible directions for Rfam

Rfam B, a database for all ncRNA families using secondary structure predictions



Get involved!

- **LitScan**, help us to reinforce the known names
- **Wikipedia**, “You are the experts”, help us to improve the summary of your favorite ncRNA
- **Families**, New ncRNA?, why not **submit a family**
 - 👍 if you have a sequence with function,
 - ▶▶ better if you have an alignment,
 - ▶▶ ▶▶ better if you have an alignment and the biochemical tests for the secondary structure

Thank You!

 **Rfam**
team



Past members

Anton I. Petrov
Ioanna Kalvari

EMBL-EBI

Nancy Ontiveros
Emma Cooke
Carlos Ribas
Andrew Green
Blake Sweeney
Alex Bateman

Collaborators

Sam Griffiths-Jones
Eric Nawrocki
Elena Rivas
Sean Eddy
Manja Marz
Kevin Lamkiewicz
Sandra Triebel
Zasha Weinberg



Useful references



Any suggestions are very welcome!, get in contact with us Rfam Help, <https://docs.rfam.org>

Want to know more about Rfam and how to search Rfam families, here our last publications

Nucleic Acids
Research 

[Rfam 14: expanded coverage of metagenomic, viral and microRNA families.](#) Kalvari *et al.* NAR (2020)



in Bioinformatics

[Non-coding RNA analysis using the Rfam database.](#) Kalvari *et al.* Curr. Protoc. Bioinformatics (2018)

Web <http://rfam.org/>

Twitter <https://twitter.com/RfamDB> @RfamDB

Github: <https://github.com/Rfam>

Blog: <https://xfam.wordpress.com/tag/rfam/>