

Help:

RNAlieN is a tool for automatic construction of RNA family models from a single sequence. The sections of this help document cover the use of the webservice and the commandline tool, as well as details of the pipeline used in the backend. If you look for source code or a installation guide for the tool please refer to the Tool subpage. Please note that constructions can take up to 24h, if you want to construct multiple families please use the tool instead of the webservice. This manual is included with the Tool as manual.pdf. ()

Table of contents:

[Webservice](#)[Commandline Tool](#)[Benchmark](#)[Pipeline](#)[Retrieve Taxonomy id](#)

Webservice:

Input:

RNAlieN Webservice accepts a fasta file with a single sequence as obligatory input. Example GlmS fasta file content:

```
>AARQ02000011.1/391-585
AAUUGAAUAGAAGCGCCAGAACUGAUUGGGACGAAAUGCUUGAAGGUGAAAUCCCUGAA
AAGUAUCGAUCAGUUGACGAGGAGGAGAUUAAUCGAAGUUUCGGCGGGAGUCUCCCGGCU
GUGCAUGCAGUCGUUAAGUCUUACUUACAAUCAUUUGGGUGACCAAGUGGACAGAGUAG
```

UAAUGAAACAUGCUU

Optionally the organism (*Listeria monocytogenes* FSL N3-165) where the sequence originates from can be passed as an NCBI Taxonomy id(393124), see (Retrieve Taxonomy id).

Output:

RNAlieN collects potential family members in multiple iterations. While the construction process continues, the progress and intermediate results are summarized in a table. Each iteration goes through multiple states (loading, sequence search, sequence retrieval, candidate alignment, candidate filtering, query selection, model calibration, done).

Progress:

Iteration	Upper taxonomic threshold id	Aligned Sequences	Alignment Link	Covariance Model Link	Status
0	205922	0	loading	loading	done
1	205922	0	stockholm-format	covariance-model	done
2	294	3	stockholm-format	covariance-model	done
3	136843	8	stockholm-format	covariance-model	model calibration

Once the construction is finished all result files are available in the result table. The organisms of found sequences are visualized in the taxonomic overview.

Results:

Log	Fasta	Stockholm Alignment	Covariance Model	Rnaz Output	cmstat Output	Zip Archive
---------------------	-----------------------	-------------------------------------	----------------------------------	-----------------------------	-------------------------------	-----------------------------

The evaluation results for the constructed covariance model and the stockholm alignment used to build it are summarized in the evaluation table.

Evaluation Results

CMstat statistics for result.cm		RNAz statistics for result alignment:	
Sequence Number	1604	Mean pairwise identity	70.09
Effective Sequences	15.75	Shannon entropy	0.52075
Consensus length	336	GC content	0.5253
Expected maximum hit-length	742	Mean single sequence minimum free energy	-121.74
Basepairs	67	Consensus minimum free energy	-79.69
Bifurcations	5	Energy contribution	-73.3
Modeltype	"cm"	Covariance contribution	-6.39
Relative Entropy CM	0.59	Combinations pair	1.62
Relative Entropy HMM	0.46	Mean z-score	-0.4
		Structure conservation index	0.65
		Background model	dinucleotide
		Decision model	structural RNA alignment quality
		SVM decision value	-0.64
		SVM class propability	0.149028
		Prediction	OTHER

The taxonomic tree of the organisms the included sequences originated from can be found in the zoom and collapsable taxonomy tree

The diagram illustrates the taxonomic distribution of 16S rDNA sequences. The sequences are categorized into three main groups: Bacteria, Eukaryota, and unclassified sequences. The Bacteria group is further divided into 20 subgroups, while Eukaryota is divided into two subgroups. The unclassified sequences group is divided into two subgroups. The diagram shows that the majority of sequences are from Bacteria, with a smaller proportion from Eukaryota and unclassified sequences.

Group	Subgroup
Bacteria	Actinobacteria
	Ascomycota
	Bacteroidetes
	Chloroflexi
	Cyanobacteria
	Firmicutes
	Gemmatimonadetes
	Planctomycetes
	Proteobacteria
	Spirochaetes
	Thaumarchaeota
	Thermotomarchaeota
	Thymomonadetes
	Thymomonadetes
	Thymomonadetes
	Thymomonadetes
	Thymomonadetes
	Thymomonadetes
	Thymomonadetes
	Thymomonadetes
Eukaryota	Opisthokonta
	Viridiplantae
unclassified sequences	cellular organisms
	unclassified sequences

Included Sequences

Taxonomy Id	Included In Iteration	Entry Header
136843	1	gb CP008896.1 c4840158-4839764 <i>Pseudomonas fluorescens</i> strain UK4, complete genome
86264	1	gii559137252.1-396 TPA: <i>Pseudomonas brassicacearum</i> subsp. <i>brassicacearum</i> NFM421 tmRNA gene Psmom_brass_NFM421
294	1	gb 559137242.1-394 TPA: <i>Pseudomonas fluorescens</i> NZI7 tmRNA gene Psmom_prot_e_Pf5
380021	2	gii610523943:c953216-952821 <i>Pseudomonas protegens</i> Cab57 DNA, complete genomecmS_3_396 '+'
200451	2	gb CP004045.1 :3880333-3880731 <i>Pseudomonas poae</i> RE ⁺ 1-1-14, complete genomecmS_3_399 '+'
136843	2	gb CP011507.1 c4107953-4107555 <i>Pseudomonas trivialis</i> strain IHBB745, complete genomecmS_3_399 '+'
29442	2	gii559137212.1-236 TPA: <i>Pseudomonas tolaasii</i> NCPPB 2192 tmRNA gene Psmom_fluor_SBW25cmS_3_236 '+'
294	2	gii559137242.1-394 TPA: <i>Pseudomonas fluorescens</i> NZI7 tmRNA gene Psmom_prot_e_Pf5cmS_1_394 '+'
1232139		gii652789639:3292883-3293275 <i>Pseudomonas pseudoalcaligenes</i> genome assembly Ppseudo_Pac, chromosome : lcmS_2_393 '+'
587851	3	gb AY911521.1 :1-348 <i>Pseudomonas chlororaphis</i> tmRNA (<i>ssrA</i>) gene, partial sequencecmS_1_346 '+'
587753	3	gb CP009290.1 :5735684-5736076 <i>Pseudomonas chlororaphis</i> subsp. <i>aurantiaca</i> strain JD37, complete genomecmS_2_393 '+'
578833	3	gb CP007441.1 c964480-964104 <i>Pseudomonas stutzeri</i> strain 28a24, complete genomecmS_2_377 '+'
312306	3	gii559136889.1-392 TPA: <i>Pseudomonas entomophila</i> L48 tmRNA gene Psmom_entom_L48cmS_1_392 '+'
251722	3	gii559137038.1-387 TPA: <i>Pseudomonas syringae</i> pv. <i>aesculi</i> str. 0893_23 tmRNA gene Psmom_syrin_1448AcmS_1_387 '+'
208964	3	gb CP008749.1 :c901873-901516 <i>Pseudomonas aeruginosa</i> PAO1H20 genomecmS_2_354 '+'
136845	3	gb AY911522.1 :1-349 <i>Pseudomonas putida</i> tmRNA (<i>ssrA</i>) gene, partial sequencecmS_1_347 '+'
136842	3	gb CP011110.1 :5753906-5754298 <i>Pseudomonas chlororaphis</i> strain PCL1606, complete genomecmS_2_393 '+'
136841	3	gb CP012679.1 :4779561-4779918 <i>Pseudomonas aeruginosa</i> strain PA1RG, complete genomecmS_2_354 '+'
103796	3	gb CP011972.1 :5087900-5088288 <i>Pseudomonas syringae</i> pv. <i>actinidiae</i> ICMP 18884, complete genomecmS_3_389 '+'
86264	3	gb CP002585.1 :c970721-970325 <i>Pseudomonas brassicacearum</i> subsp. <i>brassicacearum</i> NFM421, complete genomecmS_3_397 '+'

Input:

Parameter	Switch	Required	Description	Default value
inputFastaFilePath	-i	yes	Path to input fasta file	
outputPath	-o	yes	Path to output directory	

inputTaxId	-t	no	NCBI taxonomy ID number of input RNA organism	
inputnSCICutoff	-z	no	Only candidate sequences with a nSCI higher than this value are accepted.	1.0
inputEvalCutoff	-e	no	Evalue cutoff for cmsearch filtering.	0.001
inputBlastDatabase	-b	no	Specify name of blast database to use.	nt
coverageFilter	-a	no	Filter blast hits by coverage of at least 80%.	True
singleHitperTax	-s	no	Only the best blast hit per taxonomic entry is considered.	True
threads	-c	no	Number of available cpu slots/cores.	1
taxonomyRestriction	-r	no	Restrict search space to taxonomic kingdom (bacteria,archaea,eukarya).	
sessionIdentifier	-d	no	Optional session id that is used instead of automatically generated one.	
Help	-?	no	Print this help	

Example call for RNAlieN:

```
RNAlieN -i /home/user/newrna.fa -c 5 -t 562 -o /home/user/temp/ -d construction1
```

This command starts RNAlieN with the input fasta file /home/user/newrna.fa, 5 cores, the taxonomy id set to E. coli. RNAlieN will create a directory called construction1 in /home/user/temp/. On how to obtain the taxonomy id for the organism the sequence originates from see Retrieve Taxonomy id.

Output:

RNAlieN creates a output folder which contains all files used in the construction and log files. The most relevant output files and the logfile are directly deposited in the main folder. Each step of model expansion has an own subfolder. Following is a description of the output folder contents, of the different iteration directories that are created during model construction (empty, initial model construction, model expansion), the blast log directory that is contained in each iteration directory and the Log file content.

+RNAlieN output folder

```
|
|--Log file: Summary for construction process, tool versions, iteration info, evaluation results (see Log
file)
|--result.cm: Result covariance model
|--result.stockholm: Result stockholm alignment
|--result.fa: Result fasta
|--result.csv
|--evaluation: contains RNAz and cmstat output for result-files
|--log: contains non-iteration specific log files
|--1 iteration directories (see description below)
|--2 ..
|--3 ..
```

+iterationdirectory: initial model construction

```
|
|--model.cm: Result covariance model
```

```

|--model.stockholm: Result stockholm alignment
|--model.fa: Result fasta
|--log: Raw and processed blast hits and accepted /rejected candidates (see iteration log directory)
|--input.fa: The input fasta sequence |--input.fold: RNAfold output of the input fasta sequence |--1.fa:
fasta file for first candidate |--1.alifold: RNAalifold file for the input sequence and the first candidate
|--1.fold: RNAfold output for the first candidate
|--2.fa: Fasta file for the second candidate
|--..

```

+iterationdirectory: modelexpansion

```

|
|--model.cm: Result covariance model
|--model.stockholm: Result stockholm alignment
|--model.fa: Result fasta
|--log: Directory that contains raw and processed blast hits (see blastdirectory)
|--1.fa: fasta file for first candidate
|--1.cmsearch
|--2.fa
|--2.cmsearch
|--..

```

+iteration log directory

```

|
|--1_1blastOutput: Raw blast output for first query, indicated by leading 1_
|--1_2blastHits: Parsed blast output
|--1_3blastHitsFilteredByLength: Blasthits filtered by exceeding 3* query length
|--1_3ablastHitsFilteredByLength: Blasthits filtered by having >80% coverage
|--1_4blastHitsFilteredByParentTaxId: Only one blasthit per parent taxid
|--1_5filteredBlastResult: Only one blasthit per taxid
|--1_6requestedSequenceElements: Blasthit derived sequences requested from Entrez
|--1_10afullSequencesWithSimilars: Sequences retrieved from Entrez
|--1_10fullSequences: fullSequencesWithSimilars filtered for only containing unique sequences
|--2_1blastOutput: Raw blast output for second query, indicated by leading 2_
|--..
|--11candidates: All query specific sequences merged
|--12candidatesFilteredByCollected: Filter for sequences not identical with collected
|--13selectedCandidates: Sequences selected either by nSCI or cmsearch for inclusion in model by
set evalve cutoff
|--14rejectedCandidates: Sequences that were rejected
|--15potentialCandidates: Sequences that are within a 10^3 interval of the set evalve cutoff,
will be reevaluated at end of modelconstruction

```

Log file

Benchmark

RNAfamilies constructed for benchmark are available as archive:

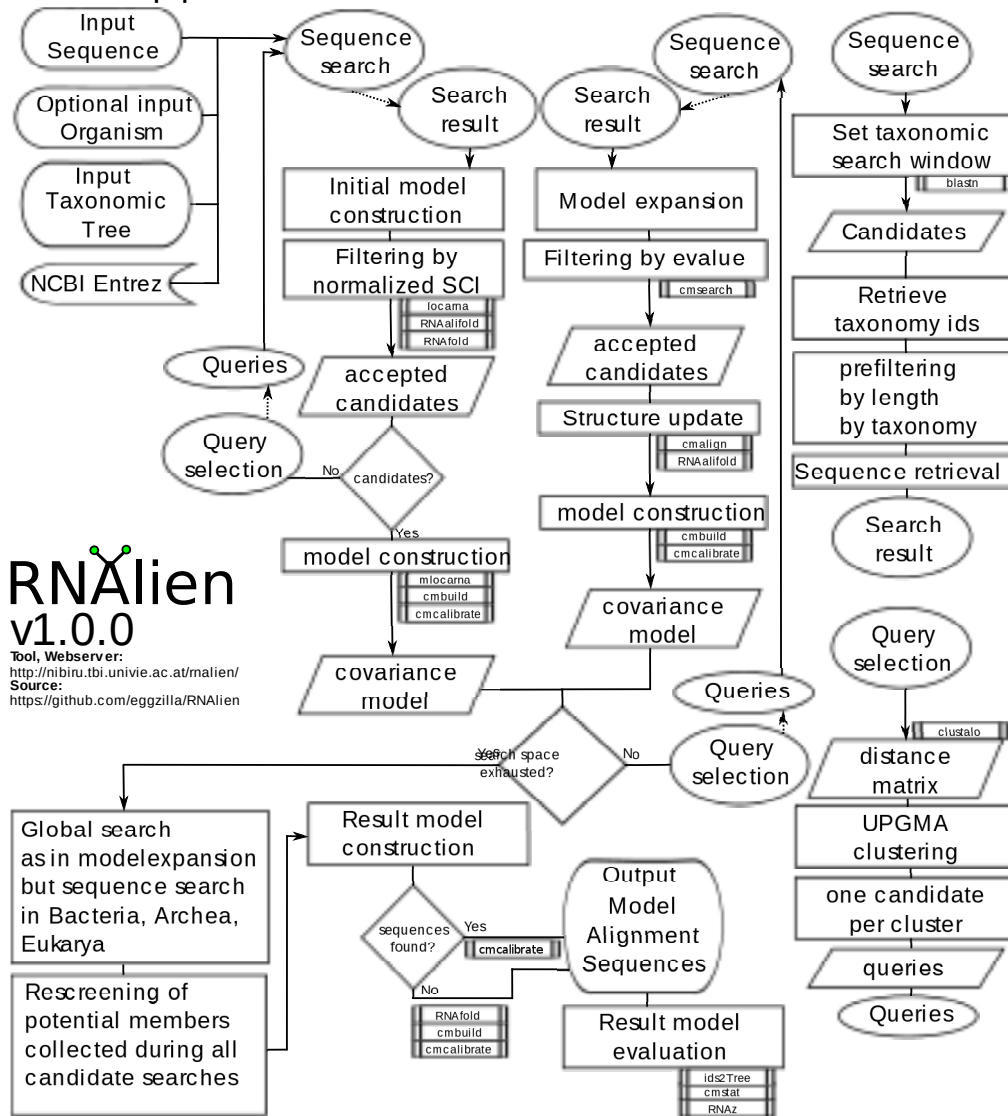
Rfam id	Family name	Archive link (.tar.gz)
RF00001	5S rRNA	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/1.tar.gz)
RF00002	5S rRNA	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/2.tar.gz)
RF00003	U1	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/3.tar.gz)
RF00004	U2	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/4.tar.gz)
RF00005	tRNA	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/5.tar.gz)
RF00008	Hammerhead_3	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/6.tar.gz)
RF00010	RNaseP_bact_a	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/7.tar.gz)
RF00011	RNaseP_bact_b	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/8.tar.gz)
RF00017	Metazoa_SRP	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/9.tar.gz)
RF00023	tmRNA	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/10.tar.gz)
RF00026	U6	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/11.tar.gz)
RF00028	Intron_gpl	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/12.tar.gz)
RF00029	Intron_gpll	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/13.tar.gz)
RF00032	Histone3	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/14.tar.gz)
RF00037	IRE_I	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/15.tar.gz)
RF00044	Phage_pRNA	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/16.tar.gz)
RF00050	FMN	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/17.tar.gz)
RF00059	TPP	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/18.tar.gz)
RF00114	S15	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/19.tar.gz)
RF00162	SAM	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/20.tar.gz)
RF00164	s2m	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/21.tar.gz)
RF00167	Purine	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/22.tar.gz)
RF00168	Lysine	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/23.tar.gz)
RF00169	Bacteria_small_SRP	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/24.tar.gz)
RF00174	Cobalamin	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/25.tar.gz)
RF00175	HIV-1_DIS	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/26.tar.gz)
RF00177	SSU_rRNA_bacteria	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/27.tar.gz)
RF00207	K10_TLS	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/28.tar.gz)
RF00209	IRES_Pesti	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/29.tar.gz)
RF00234	glmS	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/30.tar.gz)
RF00374	Gammaretro_CES	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/31.tar.gz)
RF00380	ykoK	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/32.tar.gz)
RF00458	IRES_Cripavirus	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/33.tar.gz)
RF00480	HIV_FE	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/34.tar.gz)
RF00500	TCV_H5	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/35.tar.gz)
RF00504	Glycine	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/36.tar.gz)
RF00843	mir-228	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/37.tar.gz)
RF00871	mir-689	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/38.tar.gz)
RF01051	c-di-GMP-I	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/39.tar.gz)
RF01054	preQ1-II	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/40.tar.gz)
RF01073	GP_knot1	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/41.tar.gz)
RF01118	PK-G12rRNA	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/42.tar.gz)
RF01380	HIV-1_SD	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/44.tar.gz)
RF01510	MFR	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/45.tar.gz)
RF01689	AdoCbl-variant	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/46.tar.gz)
RF01734	crcB	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/47.tar.gz)
RF01786	c-di-GMP-II	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/48.tar.gz)
RF01831	THF	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/49.tar.gz)
RF01852	tRNA-Sec	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/50.tar.gz)
RF01856	Protozoa_SRP	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/51.tar.gz)
RF01857	Archaea_SRP	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/52.tar.gz)

RF01998	group-II-D1D4-1	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/53.tar.gz)
RF02001	group-II-D1D4-3	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/54.tar.gz)
RF02095	mir-2985-2	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/55.tar.gz)
RF02253	IRE_II	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/56.tar.gz)
RF02519	ToxI	archive (http://nibiru.tbi.univie.ac.at/rnalien_tmp/benchmark/57.tar.gz)

Values for the specificity and recall plots were computed with the RNAlienStatistics executable and the `rnalienstatistics.pl` (<https://github.com/eggzilla/RNALien/blob/master/scripts/alienresultstatistics.pl>) script included in RNAlien.

RNAalien pipeline

RNAlie pipeline



Retrieve Taxonomy id

To obtain the Taxonomy id go to NCBI Taxonomy (<http://www.ncbi.nlm.nih.gov/taxonomy>). Enter the organisms name in the Taxonomy field (e.g Escherichia coli) and click search. You are then redirected to a Summary page (<http://www.ncbi.nlm.nih.gov/taxonomy/?term=Escherichia+coli>). Click the organism name. Depending on how specific your search was you can either select from a list of organisms (<http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=562>) or you are directed to a organism page Escherichia coli str. K-12 substr. MG1655star (<http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=879462&lvl=3&lin=f&keep=1&srchmode=1&unlock>). You can find the taxonomy id right below the organisms name (Escherichia coli str. K-12 substr. MG1655star Taxonomy ID: 879462).

TBI - 2015