# A COMPARISON OF RNA HOMOLOGY-DETECTING SOFTWARE

Justin Slotman

Bioinformatics Masters Thesis

December 2008

# The problem of RNA secondary structure prediction

- Primary structure does not necessarily imply secondary structure
- Secondary structure better conserved than primary sequence for RNA
- Common secondary structures can show that two RNAs are related, where sequence alignment failed

# Covariance Models are one approach

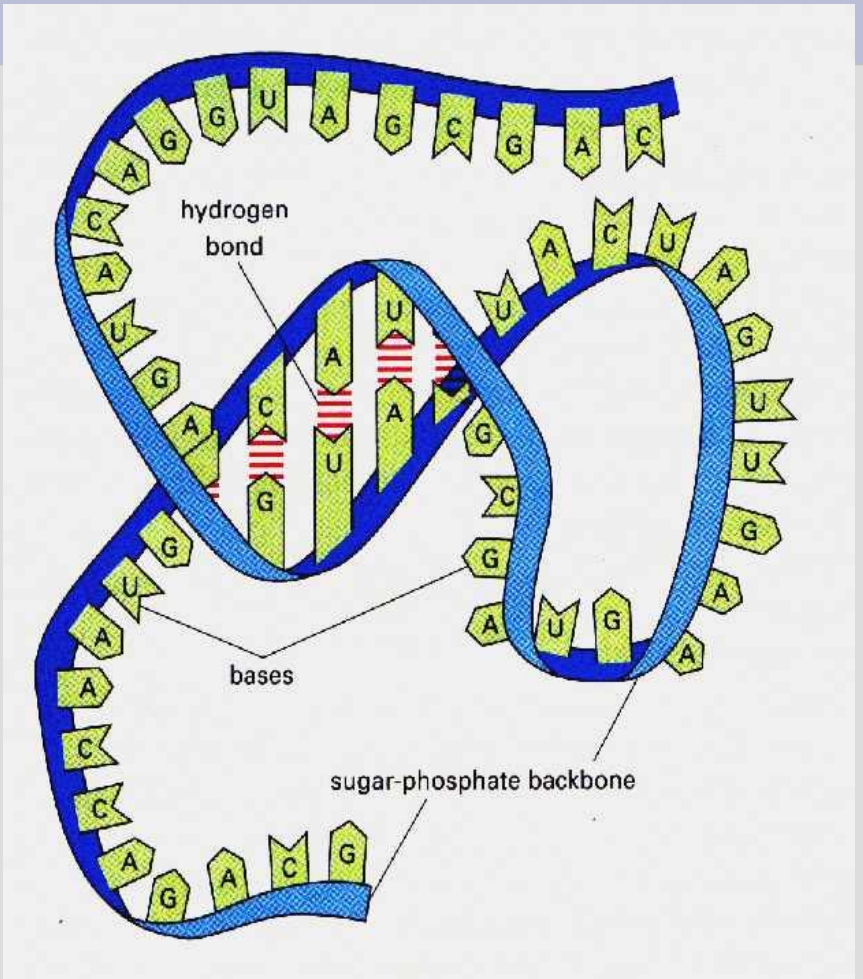
- Probabilistic model
- Describes secondary structure and primary sequence
- Can be used for secondary structure prediction, multiple sequence alignment, database similarity searching
- Intended to find RNAs where sequence alignments alone would not work as well

# Application of Covariance Models to RNA

I. Background of topic, both from biology and computer science perspective
II. Survey of software using CMs
III. Databases used
IV. Methods & results

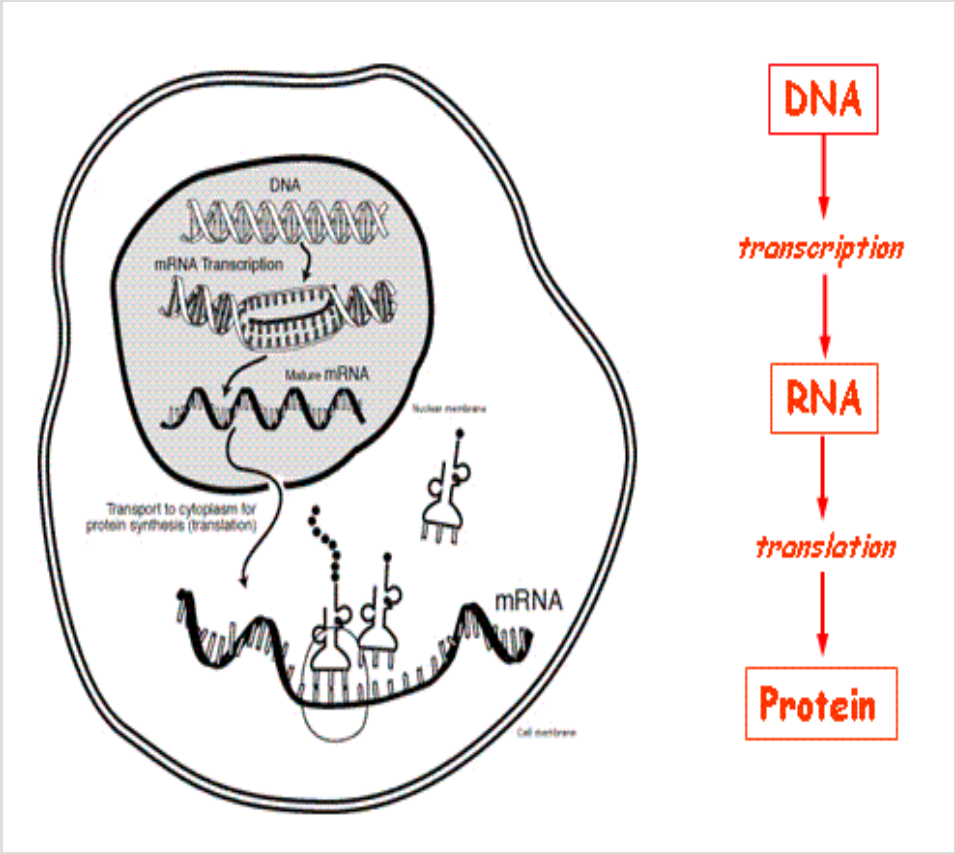
# RNA background

- RNA: Once thought to be mere messenger molecule, but now known to be both an information carrier and an enzymatically active molecule
- Some have suggested it is the original biological molecule (the "RNA world" theory)




http://tigger.uic.edu/classes/phys/phys461/phys450/ANJUM04/RNA_sstrand.jpg

# The Central Dogma

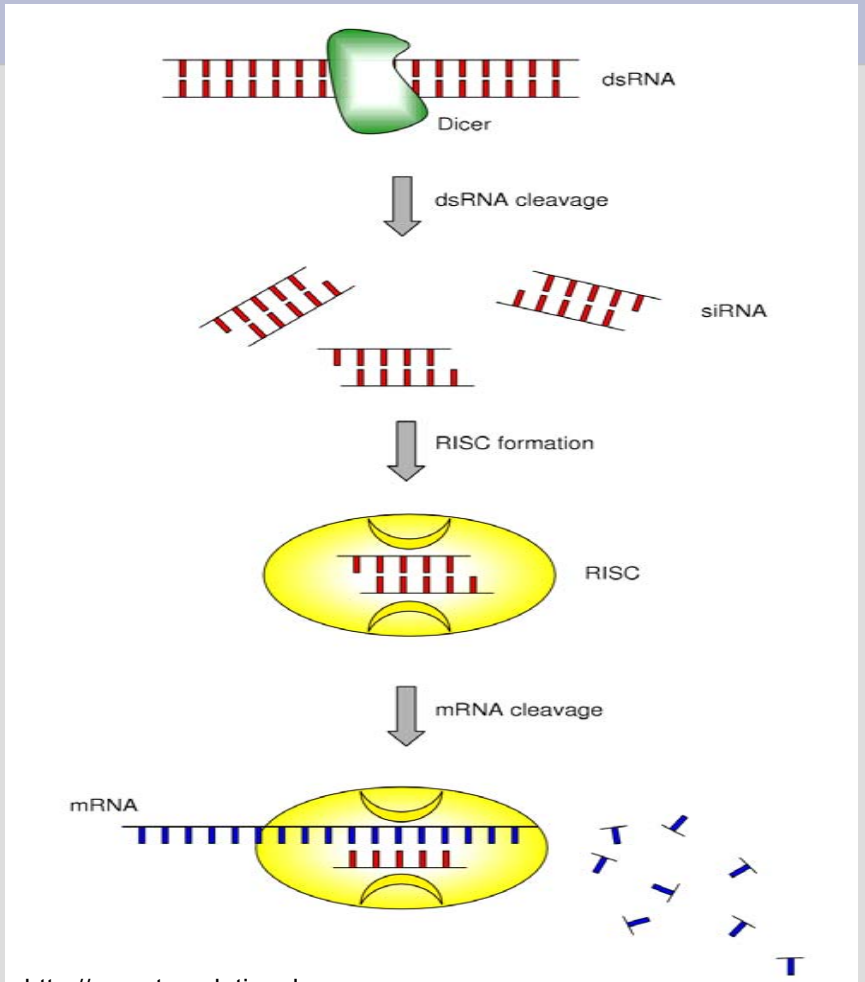



http://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/central_dogma.html

- DNA—main information carrying molecule, copies itself in the process of replication
- DNA is "copied" onto mRNA, the process of transcription
- RNA is then used to create protein, the process of translation
- No information flows from protein to DNA
- Translation assisted by tRNA and rRNA

# But the Central Dogma might be a little too simple ...

- RNA has a much more active role in all facets of cell life than previously realized, in regulation, gene expression, etc.
- Epigenetics: heritable traits that do not involve a change in DNA sequence
- Some epigenetic heredity is due to RNA interference, by methylating certain DNA sites, or activating/degrading certain RNAs and proteins

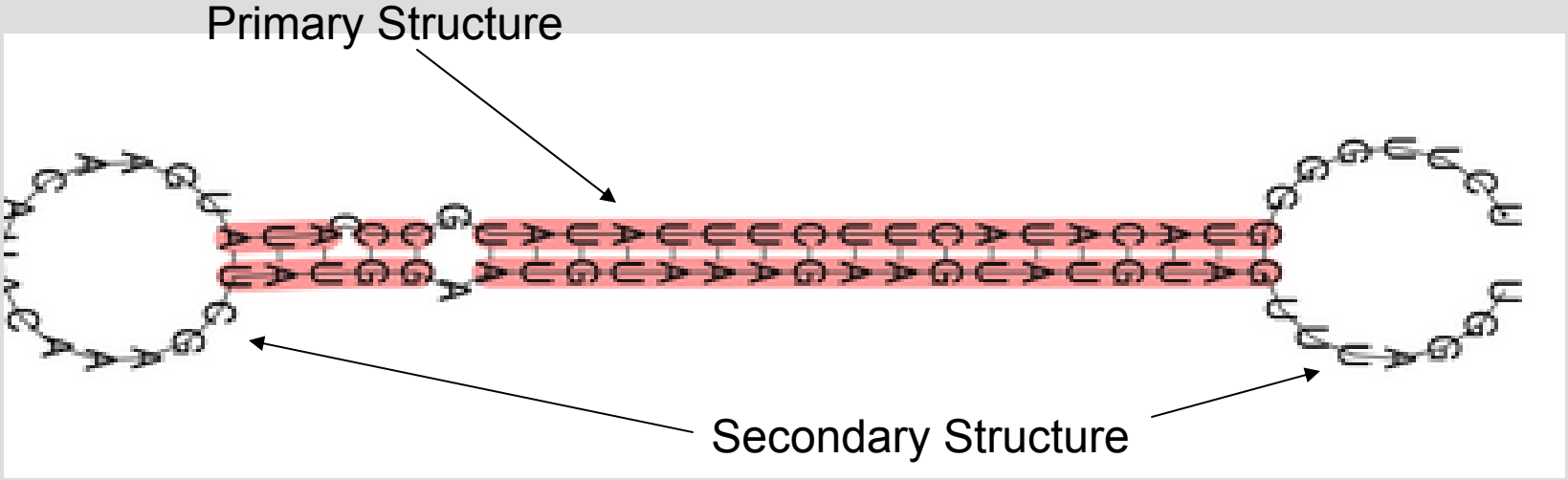



http://www.translational-medicine.com/content/2/1/39/figure/F1?highres=y

# RNA structure background

- Primary Structure
- Refers to sequence of nucleic acid residues

- Secondary Structure
- Refers to a number of small subunits RNA tends to fold into (like stem-loops)




The miRNA *mir-1*.

http://rfam.sanger.ac.uk/family?acc=RF00103

# Typical RNA secondary structures

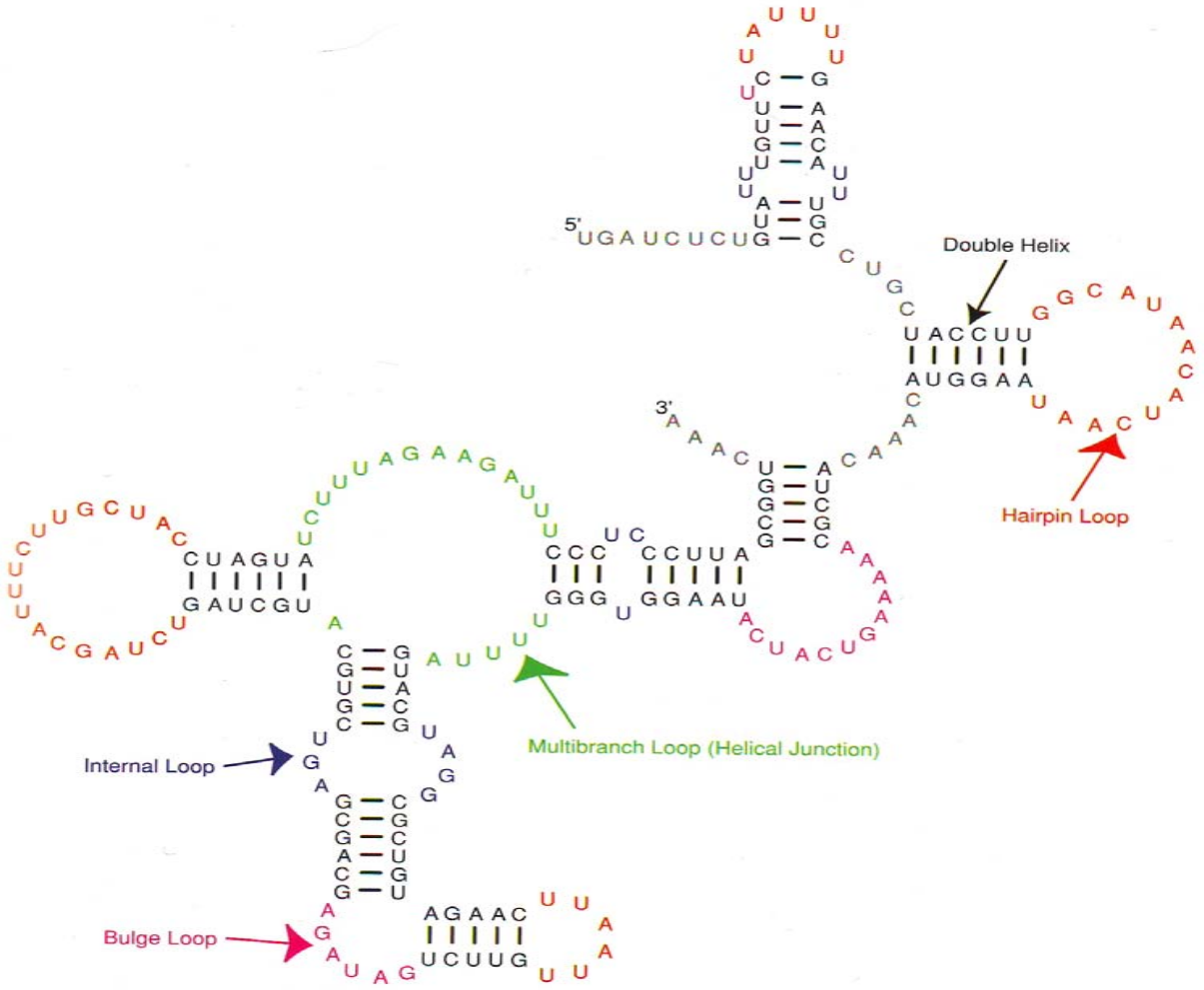



**FIGURE 6.2** The RNA secondary structure of the 3′ UTR from the *D. sucinea* R2 element (Lathe & Eickbush, 1997; Mathews et al., 1997). Base pairs in nonhelical regions, known as loops, are colored by type of loop.

From Baxevanis' Bioinformatics, 3rd edition.

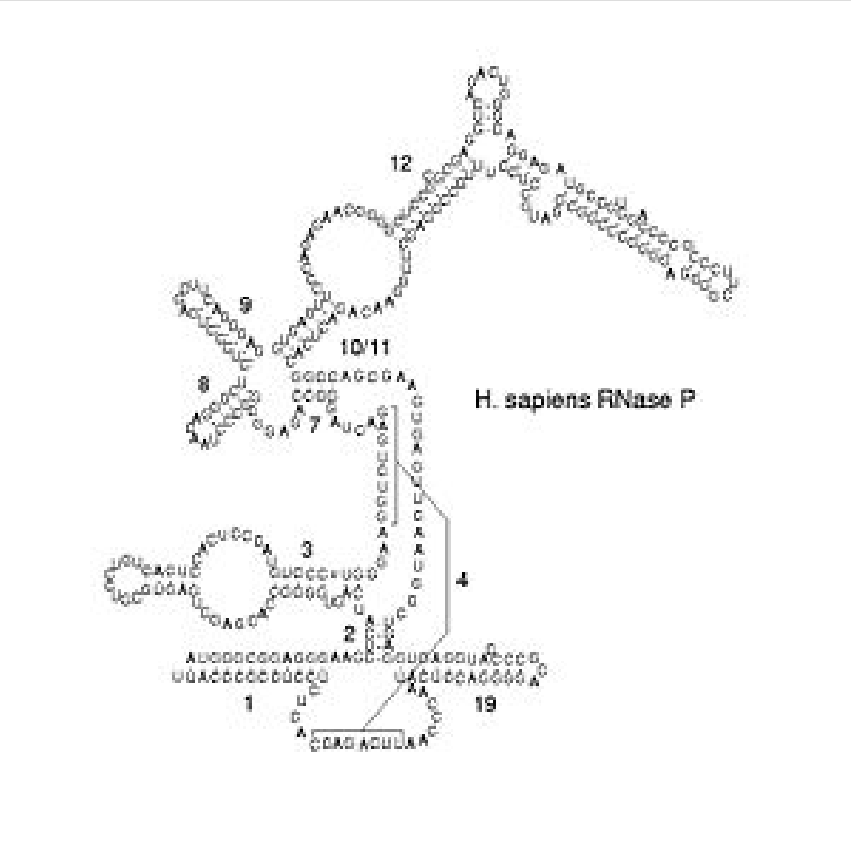
# Types of RNA

- Messenger RNA
- Transfer RNA
- Ribosomal RNA
- Non-coding RNAs
  - Ribozymes & riboswitches
  - Cis-regulatory elements
  - Micro RNAs
  - siRNAs and shRNAs
  - snRNAs and snoRNAs
  - Telomerase RNA

# ncRNAs

- ncRNAs: Functional, not information carriers
- In essence, all RNA that isn't messenger
- Other than the well known transfer and ribosomal RNAs, it was once dismissed as “junk”
- Now known to have critical regulatory functions (as cis-regulatory elements or gene expression regulating miRNAs)

# Ribozymes & Riboswitches

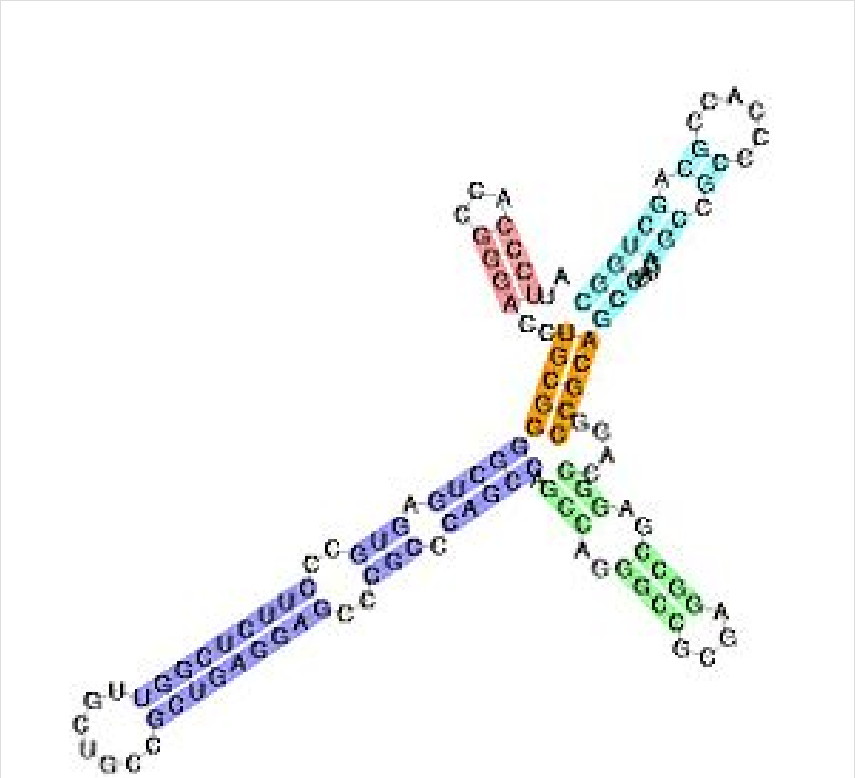



http://rfam.sanger.ac.uk/family?acc=RF00009

- Ribozymes: RNAs that function like enzymes (rnase P, self-cleaving RNA, possibly ribosomes)
- Riboswitches: untranslated segments attached to mRNA that let an mRNA self-regulate itself; common in bacteria

# Cis-regulatory elements

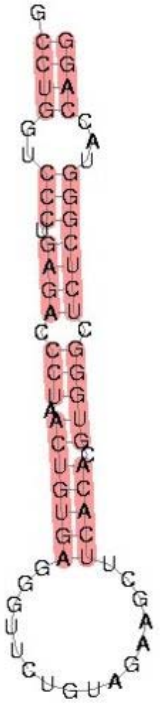
- Cis regulation: gene produces functional RNA that regulates genes on the same strand of DNA (as opposed to trans regulation, which acts on distant strands)
- Attach to binding site and influence transcription
- Sequences in the tens to hundreds
- Others influence RNA replication



**Apolipoprotein B (apoB) 5' UTR cis-regulatory element**

http://rfam.sanger.ac.uk/family?acc=RF00463

# miRNAs



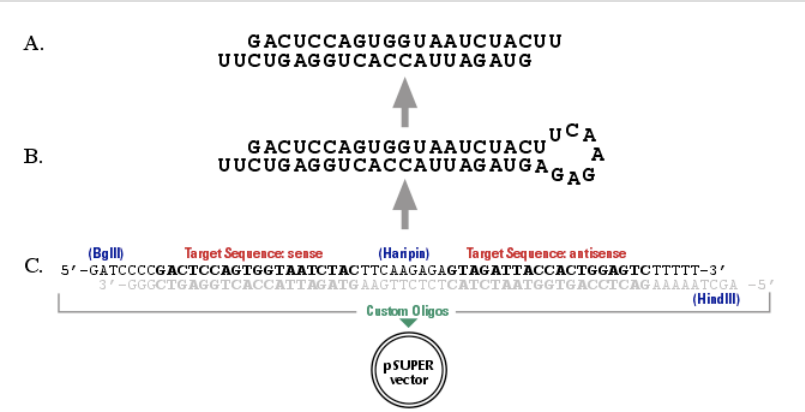
Lin-4 microRNA precursor.

http://rfam.sanger.ac.uk/family?acc=RF00052

- miRNAs: micro RNAs, usually 21-23 nucleotides in length
- Formed from precursors about 50-80 nucleotides in length
- Regulate gene expression by binding to mRNAs
- Also specify mRNA cleavage sites (another regulatory function, the degradation of mRNA)
- May also methylate complementary genomic sites

# siRNAs & shRNAS

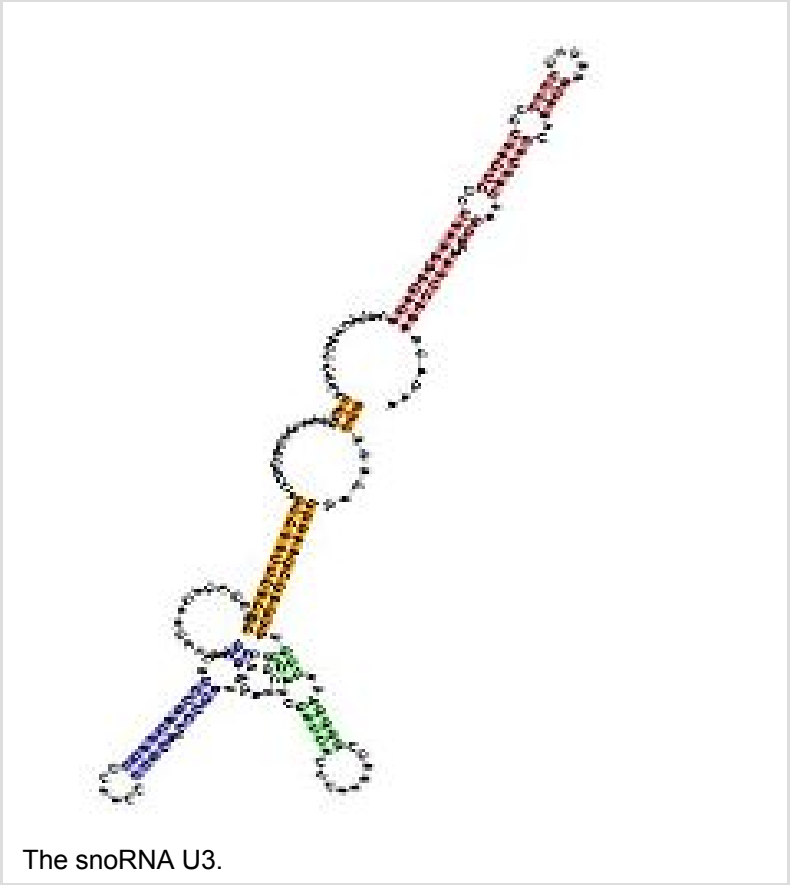
- Small interfering RNAs: function in RNA interference
- Formed from precursors, small hairpin RNAs
- Industry appears to be very interested in these




http://www.oligoengine.com/products/pSUPER.html

# snRNA and snoRNA

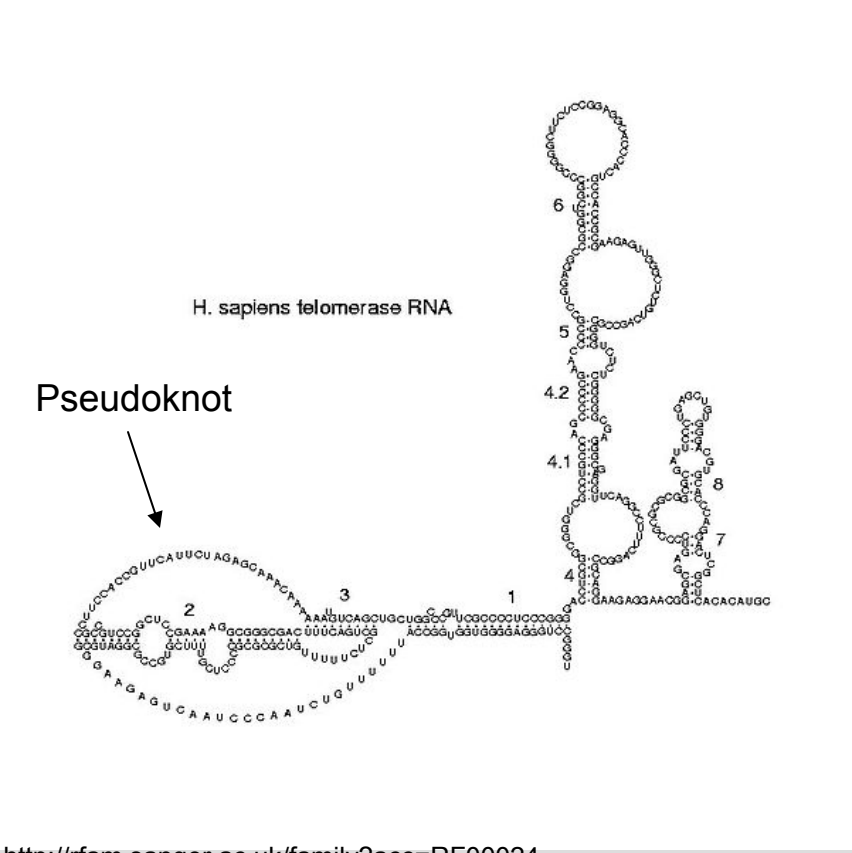
- snRNA: small nuclear RNA
- Active as regulators, splicing agents, telomere maintenance
- Major snRNA class are snoRNAs, small nucleolar RNAs
- Aid in the nucleolus' main function: ribosome creation
- Form RNA-protein complexes (snoRNPs)
- Act via methylation and pseudouridylation (the isomerisation of uridine)



The snoRNA U3.

http://rfam.sanger.ac.uk/family?acc=RF00012

# Pseudoknots & Telomerase RNA




http://rfam.sanger.ac.uk/family?acc=RF00024

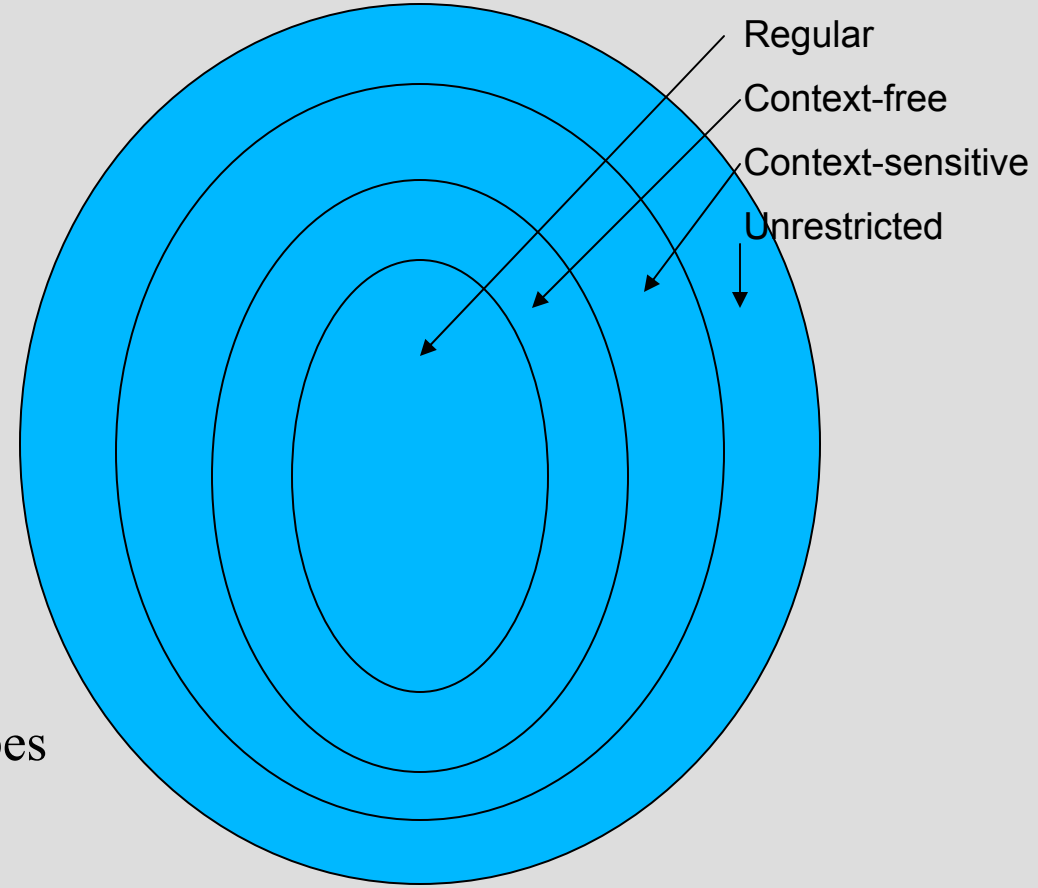
- Pseudoknot: type of tertiary structure
- Tertiary structure: units of RNA secondary structure that are formed by hydrogen bonding and can be categorized into classes or “domains”
- Base pairing with pseudoknots does not follow typical grammatical rules; as a consequence pseudoknots are very difficult to predict
- Found in telomerase RNA, which helps to maintain telomeres

# Covariance Models

- Algorithm: Sequence of instructions that must be performed to solve a well-formulated problem (how computer programs accomplish their work)
- Dynamic programming: type of algorithm that breaks problems into smaller problems (can lead to huge complexity)
- DP is used quite a bit to solve RNA secondary structure prediction problems

# Covariance Models Background: Grammars

- Grammar: In computer science terms, a set that describes the possible words or statements in a language
- Chomsky hierarchy of grammars:
  - Regular
  - Context-free
  - Context-sensitive
  - Unrestricted (phase structure)
- Automata: An abstract computational device that describes individual grammars



# Covariance Models Background: Grammars

- Regular grammars: Generate sequence from left to right, and are thus useful for modeling primary sequence
- Context-free grammars: Originally devised to describe natural languages, they have rules that allow the grammars to make correlations between ends of sentences—useful for RNA, where sequence differences may not imply secondary structure differences

# Covariance Models Background: Grammars

- Context-sensitive grammars: Grammars that have additional rules involving nonterminal character replacements that differentiate them from context-free
- Stochastic grammars: Probabilistic grammars where characters are given scores based on consensus of how a grammar is thought to work; every Chomsky hierarchy grammar can have a stochastic form

# Covariance Models Background: Stochastic Grammars

- Useful for biological analysis, since there are numerous grammatical exceptions in DNA/RNA; a probabilistic model can account for exceptions
- Example: sequence profiles that contain enough specificity to find distantly related family member
- Hidden Markov Model profiles are a widely used type of stochastic grammar

# Covariance Models Background: Stochastic Grammars & CMs

- Covariance models are another type of stochastic grammar based profile
- Unlike HMMs, they can be used to predict secondary structure
- Are the “SCFG analogue of profile HMMs”
- Specify a repetitive tree-like SCFG architecture
- Detailed, complex probabilistic models

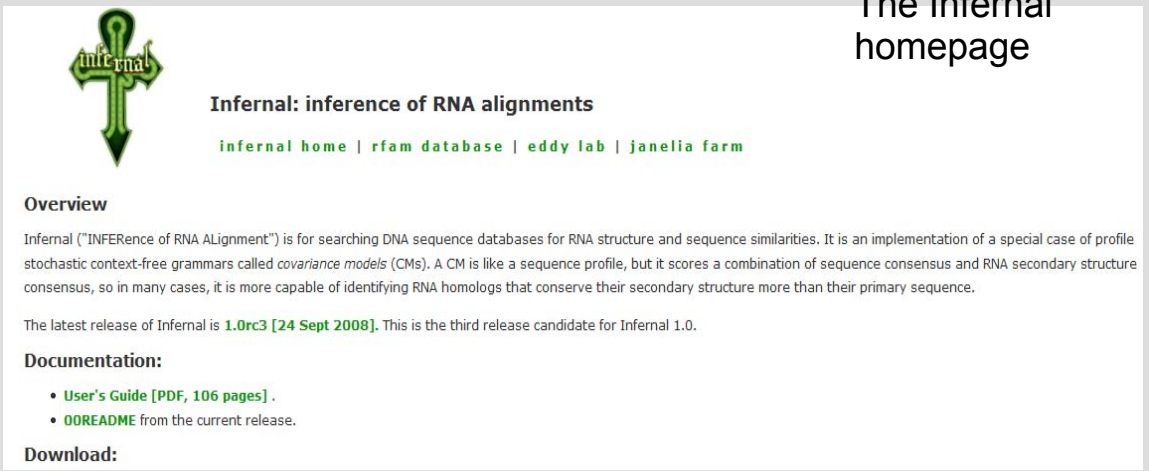
# Software & Databases Used

- CM-using software:
  - The Infernal suite
  - CMfinder
- CM-using software:
  - CARNAC
  - miRNAminer
  - BLAT
- Databases used:
  - miRBase
  - Rfam
  - RNA strand
  - UCSC Genome Browser
  - ENSEMBL
  - NCBI Genome

# The Infernal suite

- Cmalign
- Cmbuild
- Cmcalibrate
- Cmemit
- Cmscore
- Cmsearch
- cmstat

The Infernal homepage

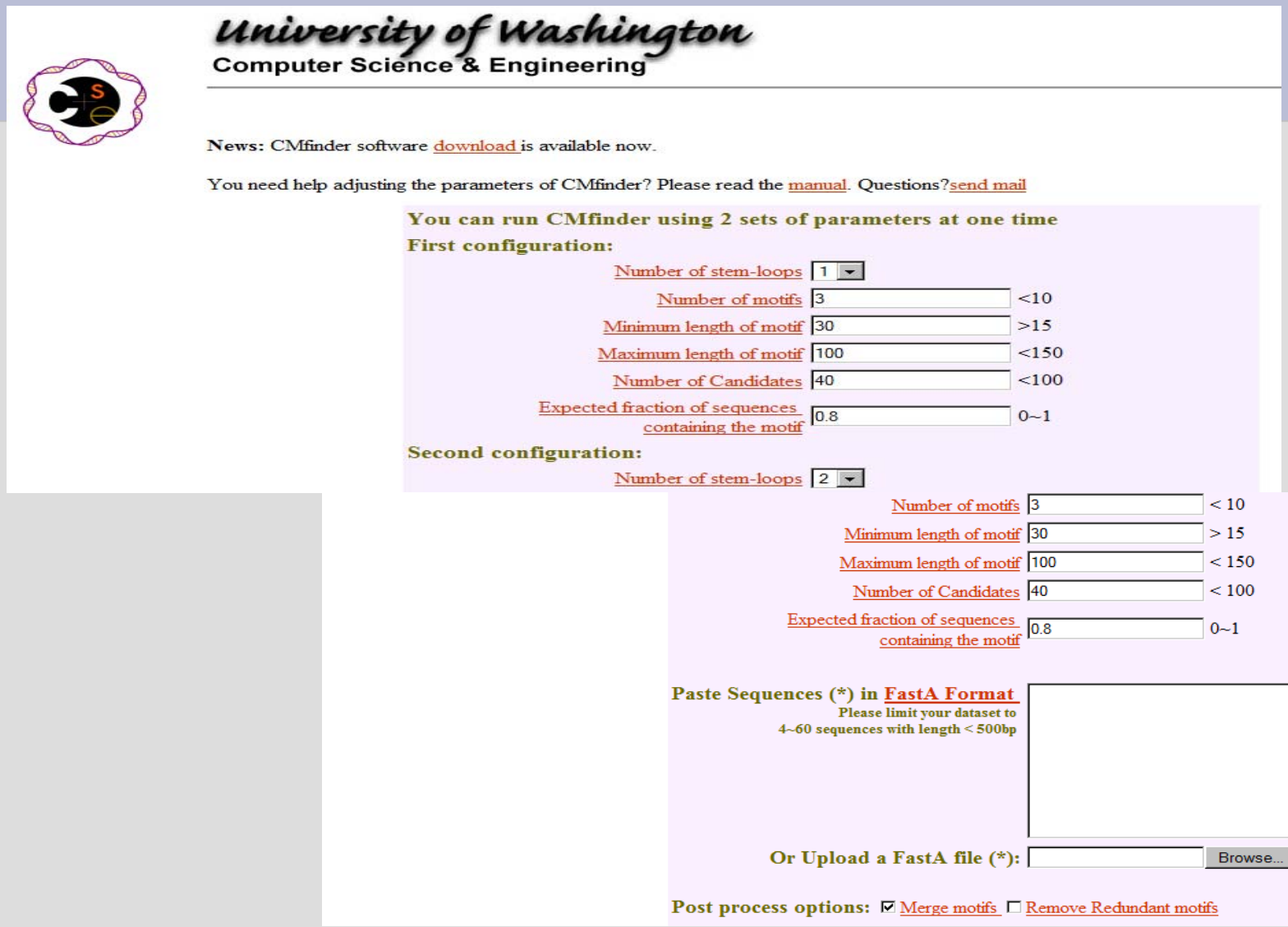



http://infernal.janelia.org/

```
Fred@Fred_HP_Dorf ~/infernal-1.0rc3
$ cd lab

Fred@Fred_HP_Dorf ~/infernal-1.0rc3/lab
$ cmsearch -g let_7_full.cm cow2.fa
# cmsearch :: search a sequence database with an RNA CM
# INFERNAL 1.0rc3 (September 2008)
# Copyright (C) 2008 HHMI Janelia Farm Research Campus
# Freely distributed under the GNU General Public License (GPLv3)
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# command:    cmsearch -g let_7_full.cm cow2.fa
# date:       Sun Nov 23 02:09:59 2008
# num seqs:   1
# dbsize(Mb): 0.004140
#
# Pre-search info for CM 1: let_7_full.sto-1
#
#                                 cutoffs          predictions
#                          ------------------  ------------------
# rnd  mod  alg  cfg   beta     E value   bit sc     surv     run time
# ---  ---  ---  ---  -----  ----------  -------  -------  -----------
   1   cm  ins  glc  1e-15       1.000     1.59   0.0196  00:00:07.19

CM: let_7_full.sto-1
>15
```

Infernal in Cygwin

# CMfinder



University of Washington
Computer Science & Engineering

**News:** CMfinder software download is available now.

You need help adjusting the parameters of CMfinder? Please read the manual. Questions?send mail

**You can run CMfinder using 2 sets of parameters at one time**

**First configuration:**

| | | |
|---|---|---|
| Number of stem-loops | 1 | |
| Number of motifs | 3 | <10 |
| Minimum length of motif | 30 | >15 |
| Maximum length of motif | 100 | <150 |
| Number of Candidates | 40 | <100 |
| Expected fraction of sequences containing the motif | 0.8 | 0~1 |

**Second configuration:**

| | | |
|---|---|---|
| Number of stem-loops | 2 | |
| Number of motifs | 3 | < 10 |
| Minimum length of motif | 30 | > 15 |
| Maximum length of motif | 100 | < 150 |
| Number of Candidates | 40 | < 100 |
| Expected fraction of sequences containing the motif | 0.8 | 0~1 |

**Paste Sequences (*) in FastA Format**
Please limit your dataset to
4~60 sequences with length < 500bp

**Or Upload a FastA file (*):** Browse...

**Post process options:** ☑ Merge motifs ☐ Remove Redundant motifs

http://wingless.cs.washington.edu/htbin-post/unrestricted/CMfinderWeb/CMfinderInput.pl

# miRNAminer


**miRNAminer is a web-based tool used for homologous miRNA gene search in several species. Given a search query, candidate homologs are identified using BLAST search and then tested for their known miRNA properties, such as secondary structure, energy, alignment and conservation, in order to assess their fidelity. Default parameters are stringent, though these can be relaxed. miRNAminer can be used prior to depositing novel miRNAs in public databases (such as miRbase).**

**NEW: See also miRviewer, a global view of homologous miRNA genes in many species.**

**Updates | Discovered miRNAs | Help**

-21.0 max RNA fold energy (delta G)

70 minimal length of precursor sequence (nt)

180 maximal length of precursor sequence (nt)

1 number or results to report

☑ Require seed conservation in mature miRNA (nt 2-8)

Target genomes (hold Ctrl and click to select more than one):

- Human (homo sapiens)
- Mouse (mus musculus)
- Rat (rattus norvegicus)
- Chimp (pan troglodytes)
- Dog (canis familiaris)

MicroRNA precursor sequence **(required)**:

AAUGAUGGAGGUGCAGGCGUUUCCUGGGGAUUAAUGACCAGCUGGGAAGAACCAGUGGCCCUCGGCUCUGCCUCCCAGCCAGCCAUUAACUCCAAGGAAAUGUCUUUUGCUGAGGU

Mature microRNA sequence **(required)**:

CCAGCUGGGAAGAACCAGUGGC

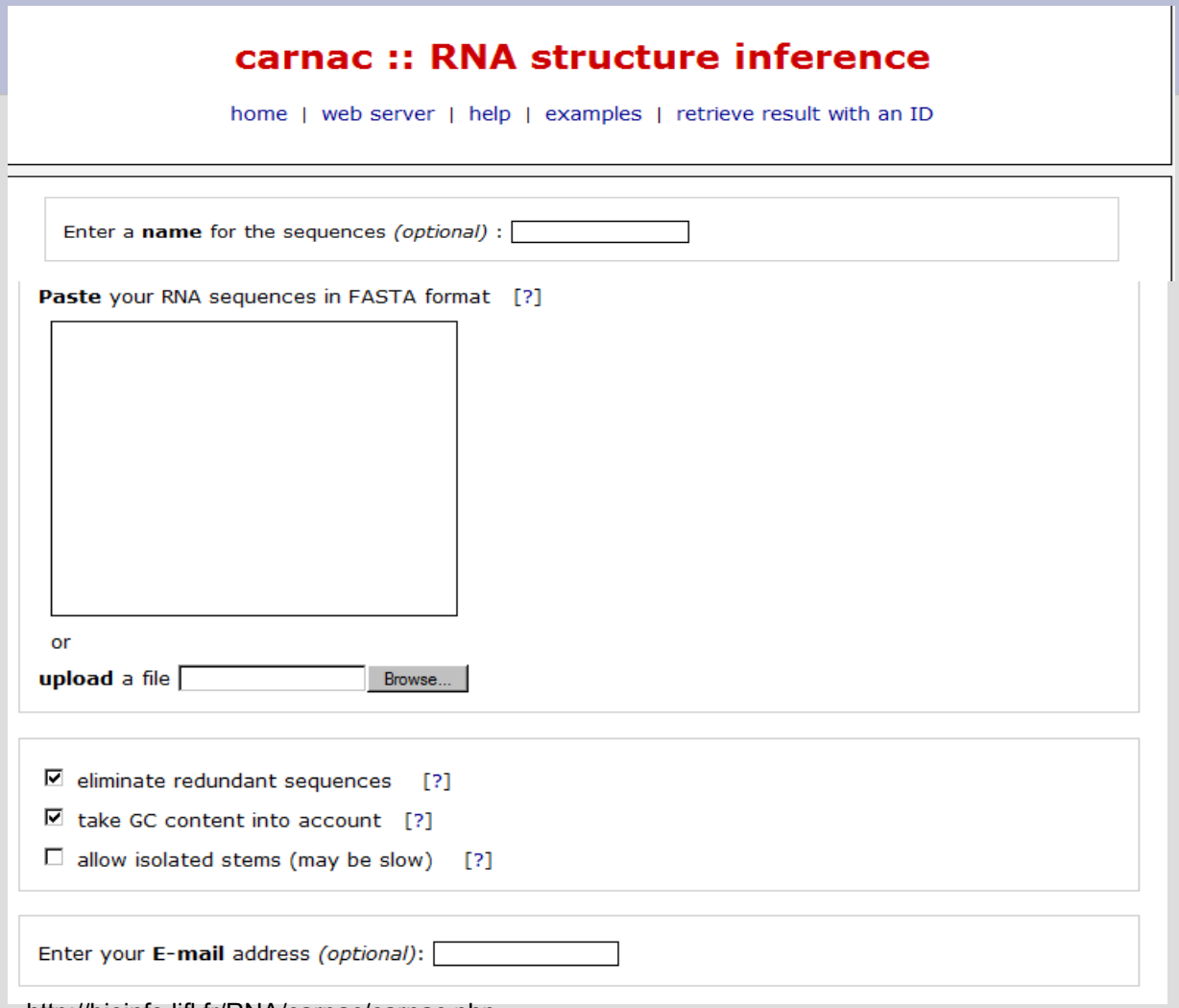
Your email address-miRNAminer will send results to that email (optional)

**When you're ready, press Submit and wait for the result screen (search time depends on the number of requested results and the number of searched genom**

Reset Submit

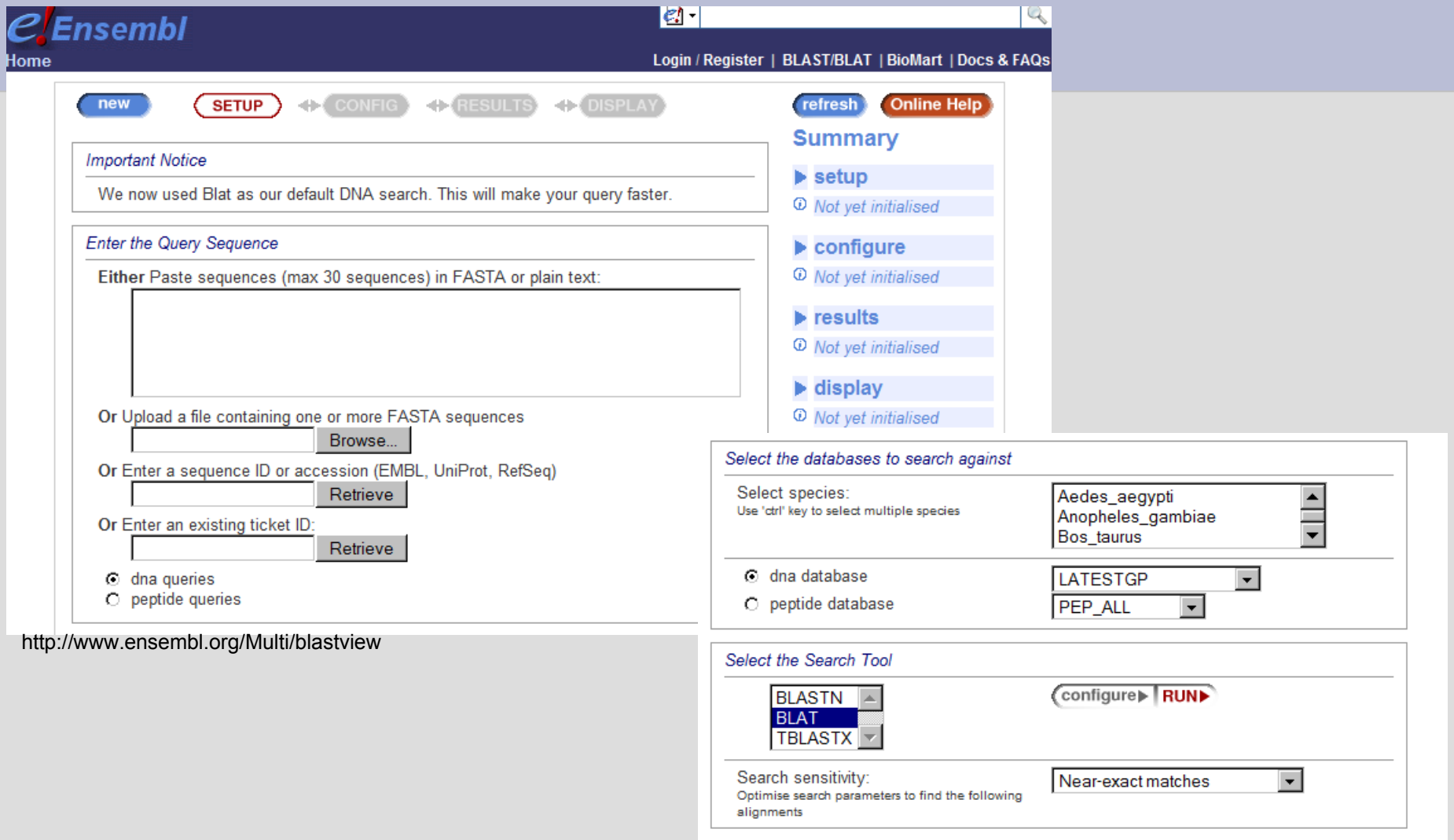
http://groups.csail.mit.edu/pag/mirnaminer/

# CARNAC




http://bioinfo.lifl.fr/RNA/carnac/carnac.php

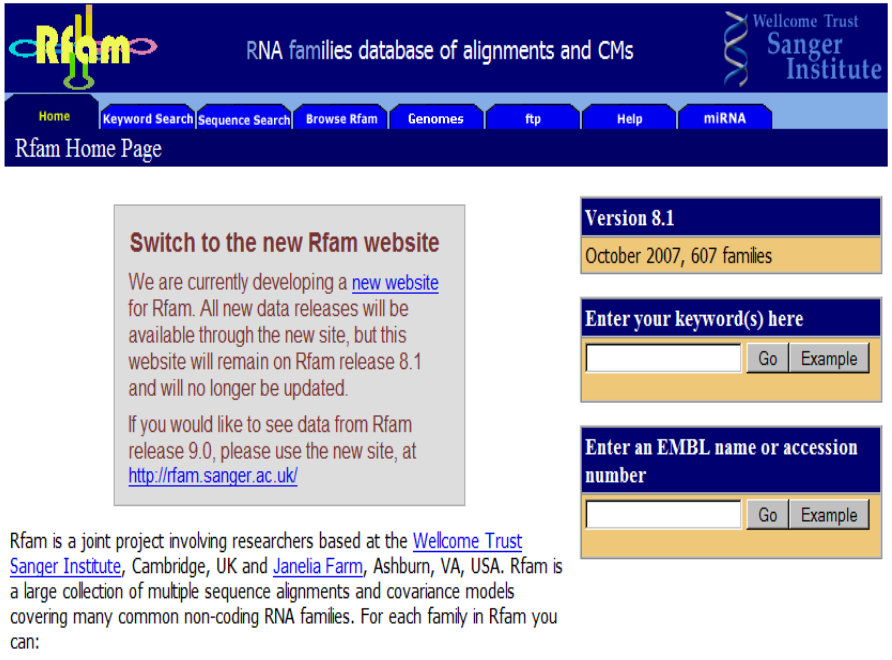
# BLAT




http://www.ensembl.org/Multi/blastview

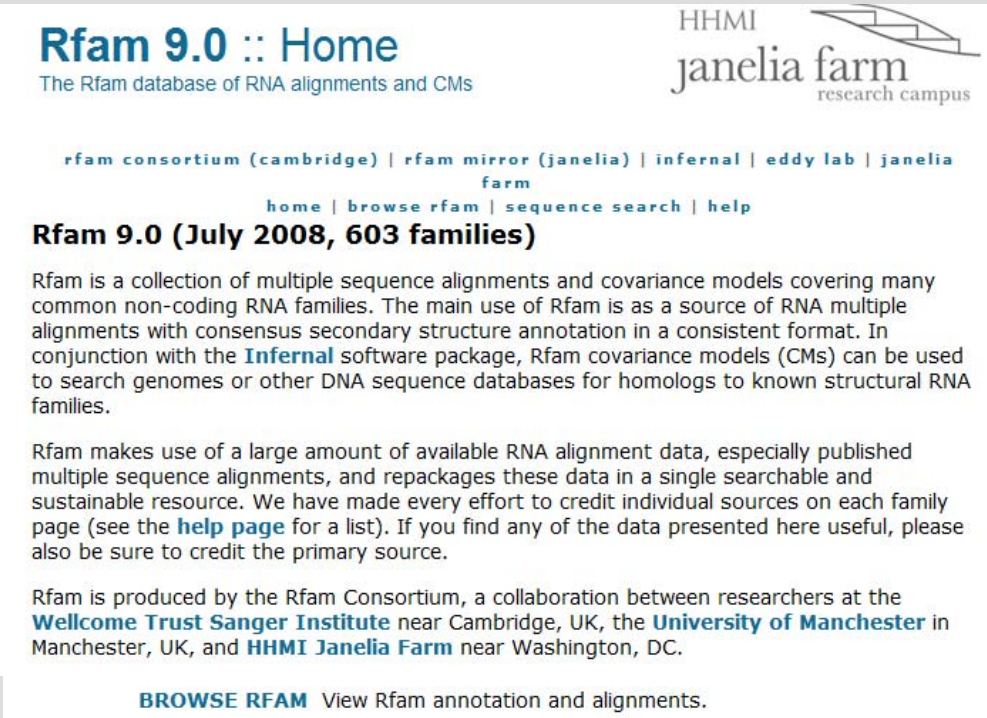
# Rfam

Rfam 8.0



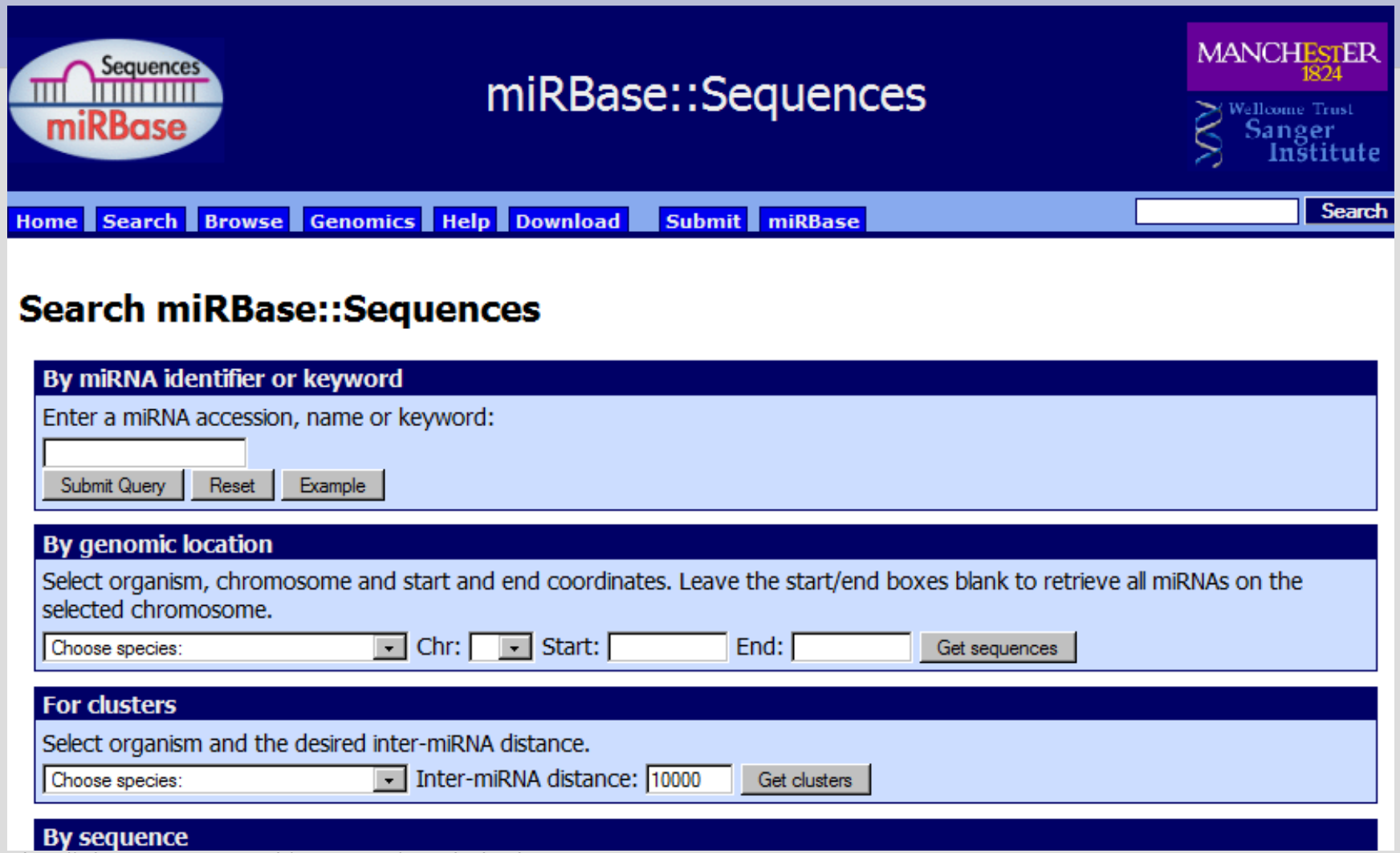

http://www.sanger.ac.uk/Software/Rfam/

Rfam 9.0




http://rfam.janelia.org/

# miRBase




http://microrna.sanger.ac.uk/sequences/search.shtml

# RNA Strand

**RNA STRAND v2.0 - The RNA secondary STRucture and statistical ANalysis Database**

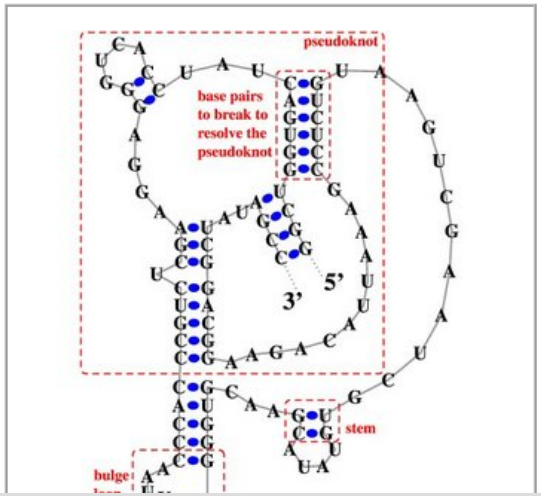
[ Home | Search | Analyse | Submit structures | News | Help ]

RNA STRAND contains ***known RNA secondary structures*** of any type and organism. The ultimate goal of this database is to incorporate a comprehensive collection of known RNA secondary structures, and to provide the scientific community with simple yet powerful ways of **analysing, searching and updating** the proposed database.

Current holdings: 4666 secondary structures in total.

Search RNA STRAND ID

| | |
|---|---|
| Search | Search for RNA STRAND entries, supports multiple search criteria |
| Analyse | Analyse one or a group of RNA secondary structures |
| Submit | Submit new RNA secondary structures to RNA STRAND |




**Structural feature occurences in RNA STRAND**

| #RNAs | #Occurences | Structural motif |
|---|---|---|
| 2333 | 6746 | Pseudoknots |
| 3582 | 17537 | Multibranched loops |
| 2992 | 35650 | Internal loops |
| 2898 | 31392 | Bulge loops |
| 4575 | 43442 | Hairpin loops |
| 2296 | 48730 | Non-canonical base pairs |

**Most common RNA types in RNA STRAND**

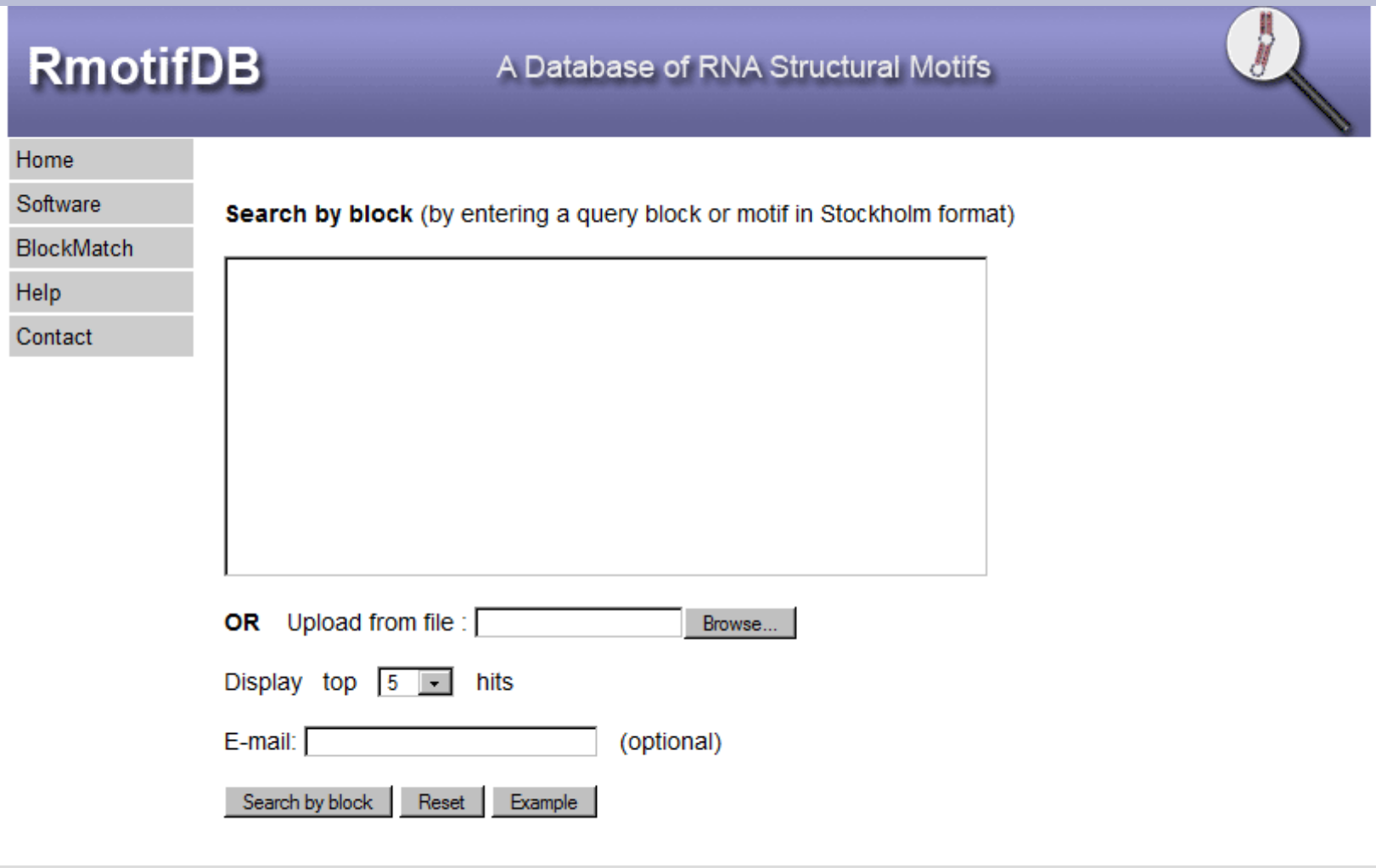
| # RNAs | RNA type |
|---|---|
| 726 | Transfer Messenger RNA |
| 723 | 16S Ribosomal RNA |
| 707 | Transfer RNA |
| 470 | Ribonuclease P RNA |
| 450 | Synthetic RNA |

**Provenance of RNA STRAND structures**

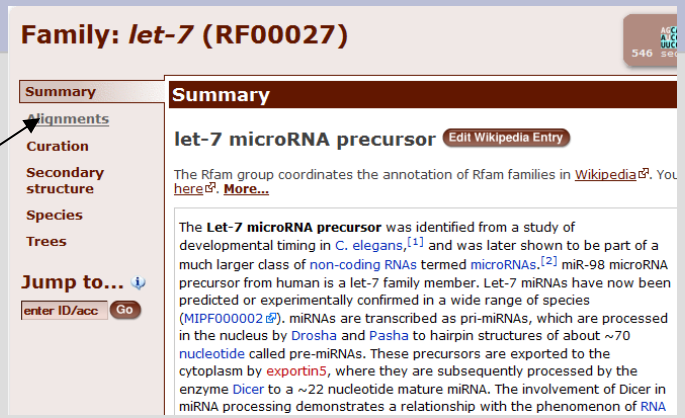
| #RNAs | Source and link to source |
|---|---|
| 1059 | RCSB Protein Data Bank |
| 1056 | Gutell Lab CRW Site |
| 726 | tmRNA Database |
| 622 | Sprinzl tRNA Database |
| 454 | RNase P Database |
| 383 | SRP Database |
| 313 | Rfam Database |
| 53 | Nucleic Acid Database |

http://www.rnasoft.ca/strand/

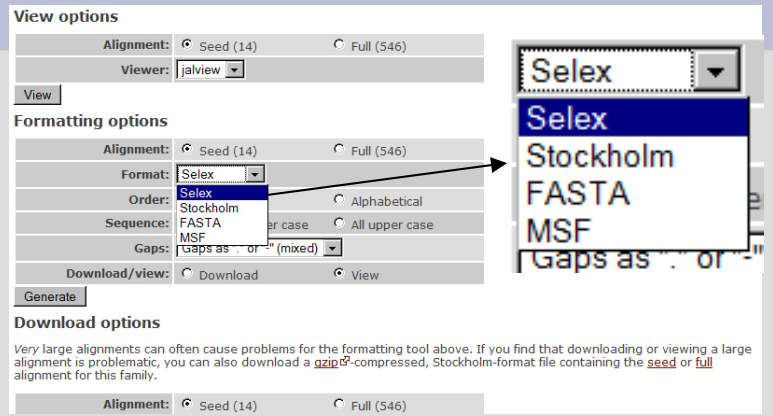
# RmotifDB


RmotifDB
A Database of RNA Structural Motifs
Home
Software
BlockMatch
Help
Contact
Search by block (by entering a query block or motif in Stockholm format)
OR Upload from file :
Browse...
Display top 5 hits
E-mail: (optional)
Search by block
Reset
Example

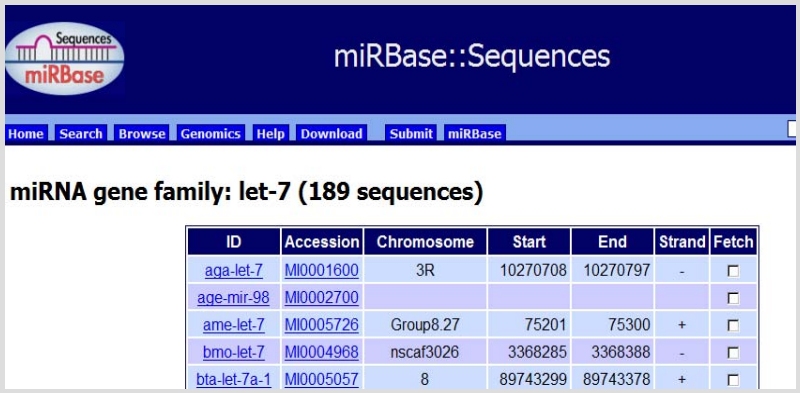
# Data Collection: microRNAs



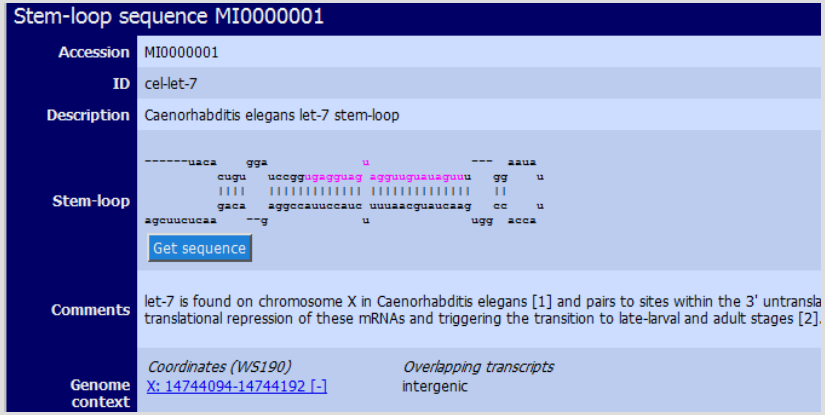

http://rfam.sanger.ac.uk/family?acc=RF00027




http://rfam.sanger.ac.uk/family?acc=RF00027




http://microrna.sanger.ac.uk/cgi-bin/sequences/mirna_summary.pl?fam=MIPF0000002




http://microrna.sanger.ac.uk/cgi-bin/sequences/mirna_entry.pl?acc=MI0000001

# Data Collection & Analysis: miRNAs

```
Information about the quality of your homolog miRNA:
Match found on chromosome 11 from 121522480 to 121522549 Strand(-)

E-value :3.73299E-5
Sequence    :UGAGGUAGUAGGUUGUAUAGUUUAGAAUUACAUCAAGGGAGAUAACUGUACAGCCUCCUAGCUUUCCUUG
RNA fold    :.(((.(((.(((((((((((((((.........(((......)))))))))))))))))).))).)))......
Fold energy :-22.0 kcal/mol
Pairing     :62.86 %
Length      :70 nt

Alignment with precursor [identity=0.73] and mature(^) [identity=1]:
query mature           ^^^^^^^^^^^^^^^^^^^^^^^^
query precursor     17 UGAGGUAGUAGGUUGUAUAGUUUGGAAUAUUACCACCGGUG--AACUAUGCAAUUUUCUACCUUACC  83
.                      ||||||||||||||||||||||.||||...|.||..||.|  ||||.|.||...|.|||.|||.||
result precursor     1 UGAGGUAGUAGGUUGUAUAGUUUAGAAUUACAUCAAGGGAGAUAACUGUACAGCCUCCUAGCUUUCC  67
result mature          ^^^^^^^^^^^^^^^^^^^^^^^^

View the miRNA homolog on Ensembl ContigView, link below:
ENSEMBL: http://www.ensembl.org/Homo_sapiens/contigview?panel_zoom=on;l=11%3A121522480-121522549;h=

View the miRNA homolog on UCSC browser, link below:
UCSC: http://genome.ucsc.edu/cgi-bin/hgTracks?org=human&position=chr11:121522480-121522549&miRNA=pack
```

ENSEMBL

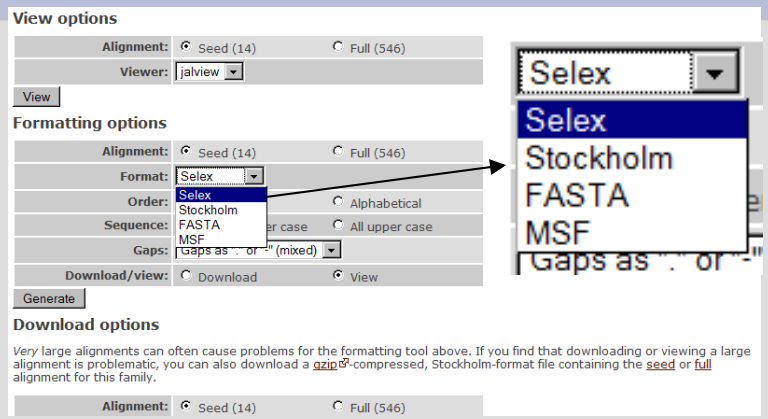
NCBI

UCSC

```
>ref|NW_001471609.1|Gga26_WGA299_2:319615-339685 Gallus gallus chromosome 26 genomic
contig, reference assembly (based on Gallus_gallus-2.1)
CAGGAGTCCCTCTTGTGTGTGTCAGAGAGCCCCATGTCCCTCTCCATGTGCTGACACTGAGCTCCTTGCA
GAGCTGGGACACGGAGCTGGAGGCTTTTGCCCAGGCCTATGCAGAGAAGTGCATCTGGGACCACAACAAG
GAGAGGGGCCGACGGGGGGAAAACCTCTTTGCTATGGCCCCAATGCTGGATCTGGAATTTGCTGTGGAGG
ACTGGAATGCGGAGGAGAAATTCTACAACCTGACGACTTCCACGTGTGTCTCTGGGCAGATGTGTGGCCA
CTACACCCAGGTACCAACCTGCTGGGGCAGAGGGGAAGTTTGGTGGGGAAGGAGCTGTGTCAGAGCCCTG
GGTCCTCCCAGAGTCTTTGCAAAGAGATGGGGAATCTGTGCTGGGCACCAGCCAGGAATCACTGATACAG
```

Fasta sequence data.

# Data Collection & Analysis: The Infernal process




cmbuild

```
Fred@Fred_HP_Dorf ~/infernal-1.0rc3/lab
$ cmbuild
Incorrect number of command line arguments.
Usage: cmbuild [-options] <cmfile output> <alignment file>
```

cmcalibrate (lengthy!)

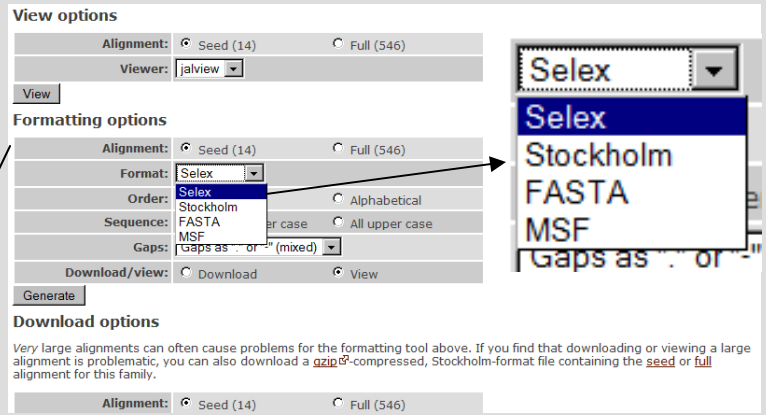
```
# Forecasting time for 2 processor(s) to calibrate CM 1: let_7_seed.sto-1
#
# stage     mod  cfg  alg  expL (Mb)   filN predicted time
# --------  ---  ---  ---  ---------  -----  --------------
  exp tail  hmm  glc  vit      17.01      -       00:03:07
  exp tail  hmm  glc  fwd      17.01      -       00:04:50
  exp tail   cm  glc  cyk       1.50      -       00:18:16
  exp tail   cm  glc  ins       1.50      -       00:24:04
  filter      -  glc    -          -  10000       00:13:00
  exp tail  hmm  loc  vit      17.01      -       00:03:05
  exp tail  hmm  loc  fwd      17.01      -       00:05:45
  exp tail   cm  loc  cyk       1.50      -       00:15:02
  exp tail   cm  loc  ins       1.50      -       00:39:53
  filter      -  loc    -          -  10000       00:18:24
# --------  ---  ---  ---  ---------  -----  --------------
# all         -    -    -          -      -       02:25:31
//
```

cmsearch

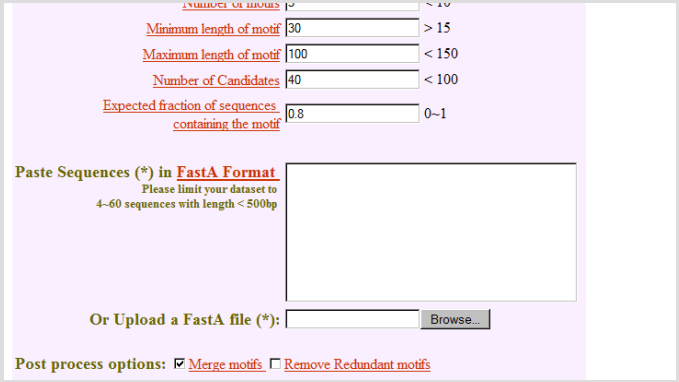
```
Fred@Fred_HP_Dorf ~/infernal-1.0rc3/lab
$ cmsearch let_7_seed.cm cow2.fa
# cmsearch :: search a sequence database with an RNA CM
# INFERNAL 1.0rc3 (September 2008)
# Copyright (C) 2008 HHMI Janelia Farm Research Campus
# Freely distributed under the GNU General Public License (GPLv3)
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# command:    cmsearch let_7_seed.cm cow2.fa
```

```
Plus strand results:

uery = 1 - 83, Target = 1019 - 1091
core = 34.88, E = 3.232e-09, P = 4.882e-12, GC =  42

        <<<<<<-<<<<<<<<<<<<<<<<<<<<<<<<~~~~~~>>>>>>>>>>>>>>>>>>>>>>>>>>
     1 ccaGGaUgAGGuAGuAGguuGuauaGUu*[28]*aACuauaCaacCUaCUaCCUuuCCug 82
       :C:GG:  AGGUAG A  UUG AUAGUU       AACUAU CAA  U CUACCU :CC:G
  1019 UCCGGU-AAGGUAGAAAAUUGCAUAGUU*[19]*AACUAUACAACCUACUACCUCACCGG 1090

       >
    83 g 83
       :
  1091 A 1091
```

# Data Collection & Analysis: CMfinder & CARNAC



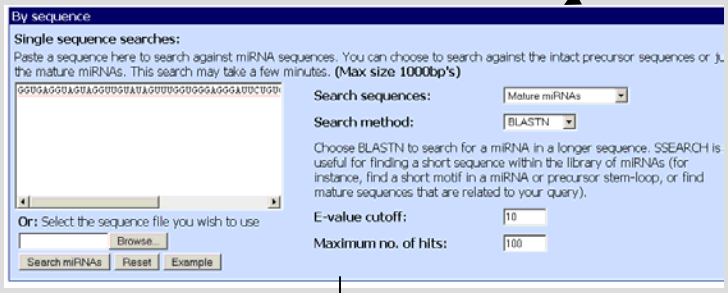

CMfinder

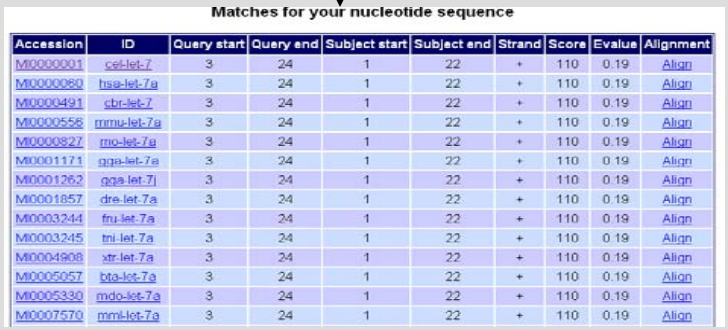



CARNAC

# Infernal & miRNAminer results

```
 Plus strand results:

Query = 1 - 83, Target = 1019 - 1091
Score = 34.88, E = 3.232e-09, P = 4.882e-12, GC =  42

        <<<<<<-<<<<<<<<<<<<<<<<<<<<~~~~~>>>>>>>>>>>>>>>>>>>>>>>>
      1 ccaGGaUgAGGuAGuAGguuGuauaGUu*[28]*aACuauaCaacCUaCUaCCUuuCCug 82
        :C:GG:  AGGUAG A  UUG AUAGUU       AACUAU CAA  U CUACCU :CC:G
   1019 UCCGGU-AAGGUAGAAAAUUGCAUAGUU*[19]*AACUAUACAACCUACUACCUCACCGG 1090

        >
     83 g 83
        :
   1091 A 1091
```

Matches for your nucleotide sequence

| Accession | ID | Query start | Query end | Subject start | Subject end | Strand | Score | Evalue | Alignment |
|---|---|---|---|---|---|---|---|---|---|
| MI0000001 | cel-let-7 | 3 | 24 | 1 | 22 | + | 110 | 0.19 | Align |
| MI0000060 | hsa-let-7a | 3 | 24 | 1 | 22 | + | 110 | 0.19 | Align |
| MI0000491 | cbr-let-7 | 3 | 24 | 1 | 22 | + | 110 | 0.19 | Align |
| MI0000556 | mmu-let-7a | 3 | 24 | 1 | 22 | + | 110 | 0.19 | Align |
| MI0000827 | rno-let-7a | 3 | 24 | 1 | 22 | + | 110 | 0.19 | Align |
| MI0001171 | gga-let-7a | 3 | 24 | 1 | 22 | + | 110 | 0.19 | Align |
| MI0001262 | gga-let-7j | 3 | 24 | 1 | 22 | + | 110 | 0.19 | Align |
| MI0001857 | dre-let-7a | 3 | 24 | 1 | 22 | + | 110 | 0.19 | Align |
| MI0003244 | fru-let-7a | 3 | 24 | 1 | 22 | + | 110 | 0.19 | Align |
| MI0003245 | tni-let-7a | 3 | 24 | 1 | 22 | + | 110 | 0.19 | Align |
| MI0004908 | xtr-let-7a | 3 | 24 | 1 | 22 | + | 110 | 0.19 | Align |
| MI0005057 | bta-let-7a | 3 | 24 | 1 | 22 | + | 110 | 0.19 | Align |
| MI0005330 | mdo-let-7a | 3 | 24 | 1 | 22 | + | 110 | 0.19 | Align |
| MI0007570 | mml-let-7a | 3 | 24 | 1 | 22 | + | 110 | 0.19 | Align |

# CARNAC & CMfinder results

Enter a **name** for the sequences *(optional)* : hairpin

**Paste** your RNA sequences in FASTA format [?]

```
>D00721/306-255
CAACAGCGAAGCGCGCCAGGGAAACACACCAUGUGUGGUAUAU
>M21212/157-106
CAACAGCGAAGCGGAACGGCGAAACACACCUUGUGUGGUAUAU
>M14879/224-175
AAACAGAGAAGUCAACCAGAGAAACACACGUUGUGGUAUAUUA
>D00685/306-255
CAACAGCGAAGCGCGCCAGGGAAACACACCAUGUGUGGUAUAU
>M17439/226-177
AAACAGAGAAGUCAACCAGAGAAACACACGUUGUGGUAUAUUA
```

**Results for job Nov270332372008**

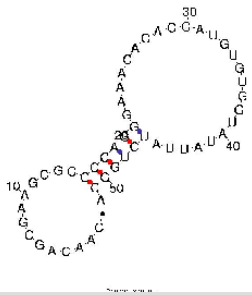
- Sequence 1 : D00721 306-255
  CT file , PS file , JPEG file, bracket notation, list of stems
- Sequence 2 : M21212 157-106
  CT file , PS file , JPEG file, bracket notation, list of stems
- Sequence 3 : M14879 224-175
  CT file , PS file , JPEG file, bracket notation, list of stems

PS and JPEG files are generated with NAVIEW.

**Visualize all foldings with RNAfamily**

press here

RNAfamily is a Java applet that displays several RNA foldings at a glance. You can download the Java Virtual Machine here.



**Paste Sequences (*) in FastA Format**

Please limit your dataset to 4~60 sequences with length < 500bp

```
>D00721/306-255
CAACAGCGAAGCGCGCCAGGGAAACACACCAUGUGU
>M21212/157-106
CAACAGCGAAGCGGAACGGCGAAACACACCUUGUGU
>M14879/224-175
AAACAGAGAAGUCAACCAGAGAAACACACGUUGUGG
>D00685/306-255
CAACAGCGAAGCGCGCCAGGGAAACACACCAUGUGU
```

**Or Upload a FastA file (*):** Browse...

**University of Washington**
**Computer Science & Engineering**

**CMfinder 1.0: Results**

Results will be sent to justin_slotman@yahoo.com. Please wait while we are processing your request...

Merge motifs ...

The results are now available for download: Compressed archive (.zip) or result directory

Thank you for using CMfinder 1.0

Back to Form

**Index of /CMfinder/data/__0.85288784677**

- Parent Directory
- seq.fasta
- seq.fasta.cm.h1.1
- seq.fasta.cm.h1.2
- seq.fasta.motif.h1.1
- seq.fasta.motif.h1.2
- seq.fasta.summary

*Apache/2.2.8 (Fedora) DAV/2 PHP/5.2.6 Server at wingless.cs.washington.edu Port 80*

# Infernal & CMfinder results

**Index of /CMfinder/data/__0.85288784677**

- Parent Directory
- seq.fasta
- seq.fasta.cm.h1.1
- seq.fasta.cm.h1.2
- seq.fasta.motif.h1.1
- seq.fasta.motif.h1.2
- seq.fasta.summary

*Apache/2.2.8 (Fedora) DAV/2 PHP/5.2.6 Server at wingless.cs.washington.edu Port 80*

```
Plus strand results:

Query = 1 - 83, Target = 1019 - 1091
Score = 34.88, E = 3.232e-09, P = 4.882e-12, GC =  42

        <<<<<<-<<<<<<<<<<<<<<<<<<<<<~~~~~>>>>>>>>>>>>>>>>>>>>>>>>>>>
      1 ccaGGaUgAGGuAGuAGguuGuauaGUu*[28]*aACuauaCaacCUaCUaCCUuuCCug 82
        :C:GG:  AGGUAG A  UUG AUAGUU      AACUAU CAA  U CUACCU :CC:G
   1019 UCCGGU-AAGGUAGAAAAUUGCAUAGUU*[19]*AACUAUACAACCUACUACCUCACCGG 1090

        >
     83 g 83
        :
   1091 A 1091
```

**Search by block** (by entering a query block or motif in Stockholm format)

```
# STOCKHOLM 1.0
#=GF AU    Infernal 0.1

#=GS D00721/306-255 WT    1.00
#=GS M21212/157-106 WT    1.00
#=GS M14879/224-175 WT    1.00
#=GS D00685/306-255 WT    1.00
#=GS M17439/226-177 WT    1.00

#=GS D00721/306-255 DE     15.. 51      61.194061
#=GS M21212/157-106 DE     15.. 51      57.970718
```

**OR** Upload from file : Browse...

Display top 5 hits

# Rfam and BLAT alignments

Formatting options

| | | |
|---|---|---|
| **Alignment:** | Seed (5) | Full (5) |
| **Format:** | FASTA | |
| **Order:** | Tree | Alphabetical |
| **Sequence:** | Inserts lower case | All upper case |
| **Gaps:** | Gaps as "." or "-" (mixed) | |
| **Download/view:** | Download | View |

Generate

```
>D00721/306-255
CAACAGCGAAGCGCGCCAGGGAAACACACCAUGUGUGGUAUAUUAUCUGGCA
>M21212/157-106
CAACAGCGAAGCGGAACGGCGAAACACACCUUGUGUGGUAUAUUACCCGUUG
>M14879/224-175
AAACAGAGAAGUCAACCAGAGAAACACACGUUGUGGUAUAUUACCUGGUA
>D00685/306-255
CAACAGCGAAGCGCGCCAGGGAAACACACCAUGUGUGGUAUAUUAUCUGGCA
>M17439/226-177
AAACAGAGAAGUCAACCAGAGAAACACACGUUGUGGUAUAUUACCUGGUA
```

**BLAT Search Genome**

| Genome: | Assembly: | Query type: | Sort output: | Output type: |
|---|---|---|---|---|
| Human | Mar. 2006 | BLAT's guess | query,score | hyperlink |

```
>D00721/306-255
CAACAGCGAAGCGCGCCAGGGAAACACACCAUGUGUGGUAUAUUAUCUGGCA
>M21212/157-106
CAACAGCGAAGCGGAACGGCGAAACACACCUUGUGUGGUAUAUUACCCGUUG
>M14879/224-175
AAACAGAGAAGUCAACCAGAGAAACACACGUUGUGGUAUAUUACCUGGUA
>D00685/306-255
CAACAGCGAAGCGCGCCAGGGAAACACACCAUGUGUGGUAUAUUAUCUGGCA
>M17439/226-177
AAACAGAGAAGUCAACCAGAGAAACACACGUUGUGGUAUAUUACCUGGUA
```

submit | I'm feeling lucky | clear

**BLAT Search Results**

| ACTIONS | QUERY | SCORE | START | END | QSIZE | IDENTITY | CHRO | STRAND | START | END | SPAN |
|---|---|---|---|---|---|---|---|---|---|---|---|
| browser details | D00685/306-255 | 20 | 24 | 43 | 52 | 100.0% | X | - | 45714471 | 45714490 | 20 |
| browser details | D00721/306-255 | 20 | 24 | 43 | 52 | 100.0% | X | - | 45714471 | 45714490 | 20 |

# Conclusion

- Infernal & miRNAminer:
  - Infernal very sensitive versus miRNAminer
- CARNAC & CMfinder:
  - CARNAC user-friendly
  - CMfinder output difficult to interpret
- Infernal & CMfinder
  - CM-using software head to head
- BLAT and Rfam
  - Rfam full alignments include more results than corresponding BLAT searches

# References - Part 1


- Eddy, S.R. & Durbin, R. (1994). RNA sequence analysis using covariance models. Nucleic Acids Research, Vol. 22, No.11, 2079-2088.
- Eddy, S.R.. (2006). Computational Analysis on RNAs. Cold Spring Harbor Symposium on Quantitative Biology, Vol. 71. 117-128.
- Rivas, E. & Eddy, S.R. (2001). Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics, Vol 2, No. 8. Retrieved October 12, 2008 from the World Wide Web: http://www.biomedcentral.com/1471-2105/2/8
- Durbin, R., Eddy, S.R., Krogh, A., & Mitchison G.. (1998, 2006). Biological Sequence Analysis (11th ed.) Cambridge, Cambridge University Press.
- Gilbert, W. (1986). The RNA World. Nature, Vol. 319. 618.
- Cech, T. (2004). Exploring the New RNA World [Document posted on Web site Nobelprize.org]. Retrieved October 12,2008 from the World Wide Web: http://nobelprize.org/nobel_prizes/chemistry/articles/cech/index.html
- Crick, F. (1958). On Protein Synthesis. Symposia of the Society for Experimental Biology, Vol. 12. 139-163.
- Crick, F. (1970). Central dogma of molecular biology. Nature, Vol. 227. 561-563.
- Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (1983, 2002). Molecular Biology of the Cell (4th ed.) [Book posted on Web site NCBI]. New York, Garland Science. Retrieved October 12,2008 from the World Wide Web: http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=mboc4.TOC&depth=2
- Bird, A. (2007). Perceptions of epigenetics. Nature, Vol. 447. 396-398.

# References - Part 2


- Matzke, M.A. & Birchler J.A. (2005). RNAi-mediated paths in the nucleus. Nature Reviews Genetics, Vol. 6. 24-35.
- Wassenegger, M (2000). RNA-directed DNA methylation. Plant Molecular Biology, Vol. 43. 203-220.
- Aufsatz, W., Mette, M.F., van der Winden, J., Matzke, A.J., & Matzke M.A. (2002). RNA-directed DNA methylation in Arabidopsis. Proceedings of the National Academy of Sciences of the U.S.A, Vol. 99, supplement 4. 16499-16506.
- Tang, W., Luo, X.Y., & Sanmuels, V. (2001). Gene silencing: Double-stranded RNA mediated mRNA degradation and gene inactivation. Cell Research, Vol. 11. 181-186.
- Coffin, J.M., Hughes, S.H., & Varmus, H.E. (1997). Reverse Transcriptase and the Generation of Retroviral DNA. Chapter in Retroviruses [Book posted on Web site NCBI]. Plainview, Cold Spring Harbor Laboratory Press. Retrieved October 12,2008 from the World Wide Web: http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=rv.TOC
- Becker, W.M., Kleinsmith, L.J., & Hardin, J. (2003). The World of the Cell (5th ed.) San Francisco, Benjamin Cummings.
- Baxevanis, A.D., & Ouellette, B.F.F. (2002) Bioinformatics: a practical guide to the analysis of genes and proteins. (3rd ed.) Hoboken, John Wiley & Sons.
- Valencia-Sanchez, M.A., Liu, J., Hannon, G.J., & Parker, R. (2006). Control of translation and mRNA degradation by miRNAs and siRNAs. Genes & Development, Vol. 20. 515-524.
- Berg, J.M., Tymoczko, J.L., Stryer, L., & Clarke, N.D. (2002) Biochemistry (5th ed.) [Book posted on Web site NCBI]. New York, W.H. Freeman & Company. Retrieved October 12,2008 from the World Wide Web: http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=stryer.TOC&depth=2
- Brown, T.A. (2002) Genomes (2nd ed.) [Book posted on Web site NCBI]. Oxford, BIOS Scientific Publishers Ltd. Retrieved October 12,2008 from the World Wide Web: http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=genomes.TOC&depth=2

# References - Part 3


- Prasanth, K.V. & Spector, D.L. (2007) Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. Genes & Development, Vol. 21. 11-42.
- Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., & Altman, S. (1983) The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. Cell, Vol. 21, Issue 3, Part 2. 849-857.
- Tang, J., & Breaker, R.R. (2000). Structural diversity of self-cleaving ribozymes. Proceedings of the National Academy of Sciences of the U.S.A, Vol. 97, no. 11. 5784-5789.
- Miranda-Rios, J. (2007) The THI-box Riboswitch, or How RNA Binds Thiamin Pyrophosphate. Structure, Vol. 15, issue 3. 259-265.
- Gilbert, S.F. (2000) Developmental Biology (6th ed.) [Book posted on Web site NCBI]. Sunderland, Sinauer Associates, Inc. Retrieved October 12,2008 from the World Wide Web: http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=dbio
- Lee R.C., Feinbaum R.L., Ambros V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell, Vol. 75, issue 5. 843-854.
- Yekta, S., Shih I., & Bartel, D.P. (2004) MicroRNA-Directed Cleavage of HOXB8 mRNA. Science, Vol. 304, no. 5670. 594-596.
- Ronemus, M., & Martienssen, R. (2005). RNA interference: Methylation mystery. Nature, Vol. 433. 472-473.
- Paddison, P.J., Caudy, A.A., Bernstein, E., Hannon, G.J., & Conklin, D.S. (2002). Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells. Genes & Development, Vol. 16. 948-958.
- Egloff,s., Van Herreweghe, E., & Kiss, T. (2006). Regulation of Polymerase II Transcription by 7SK snRNA: Two Distinct RNA Elements Direct P-TEFb and HEXIM1 Binding. Molecular and Cellular Biology, Vol. 26, no. 2. 630-642.

# References - Part 4


- Jady, B.E., & Kiss, T. (2001) A small nucleolar guide RNA functions both in 2'-*O*-ribose methylation and pseudouridylation of the U5 spliceosomal RNA. The EMBO Journal, Vol. 20. 541-551.
- Lukowiak, A.A., Narayanan, A., Li, Z.H., Terns, R.M., & Terns, M.P. (2001) The snoRNA domain of vertebrate telomerase RNA functions to localize the RNA within the nucleus. RNA, Vol. 7, no. 12. 1833-1844.
- Rivas, E. & Eddy, S.R. (1999). A dynamic programming algorithm for RNA structure prediction including pseudoknots. Journal of Molecular Biology, Vol. 285, issue 5. 2053-2068.
- Jones, N.C., & Pevzner, P.A. (2004) An Introduction to Bioinformatics Algorithms. Cambridge, MIT Press.
- Hopcraft, J.E., & Ullman, J.D. (1979) Introduction to Automata Theory, Languages and Computation. Addison-Wesley.
- The Eddy Lab. (2008) INFERNAL User's Guide. [Document posted on Web site Infernal: inference of RNA alignments.] Retrieved September 30, 2008 from the World Wide Web: http://infernal.janelia.org
- Yao, Z., Weinberg, Z., & Ruzzo, W.L. (2006) CMfinder—a covariance model based RNA motif finding algorithm. Bioinformatics, Vol. 22., no. 4. 445-452.
- Yao, Z., Weinberg, Z., & Ruzzo, W.L. (2005) CMfinder 1.0 Manual. [Document posted on the Web site University of Washington Computer Science & Engineering.] Retrieved October 12, 2008 from the World Wide Web: http://wingless.cs.washington.edu/CMfinder/manual.htm
- Knudsen, B., & Hein, J. (1999) RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics, Vol. 15., no. 6. 446-454.
- Knudsen, B., & Hein, J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Research, Vol. 31., no. 13. 3423-3428.

# References - Part 5


- Artzi, A., Kiezun, A., & Shomron, S. (2008). MiRNAminer: a tool for homologous microRNA gene search. BMC Bioinformatics, Vol 9, no. 39. Retrieved October 12, 2008 from the World Wide Web: http://www.biomedcentral.com/1471-2105/9/39
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., & Lipman, D.J. (1990) Basic local alignment search tool. Journal of Molecular Biology, Vol. 285, issue 5. 2053-2068.
- Kent, W.J. (2002) BLAT—The BLAST-Like Alignment Tool. Genome Research, Vol. 12, issue 4. 656-664.
- Touzet, H., & Perriquet, O. (2004) CARNAC: folding families of related RNAs. Nucleic Acids Research, Vol. 32, Web Server issue. W142-W145.
- Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Research, Vol. 31, no. 13. 3406-3415.
- Mathews, D.H., Sabina, J., Zuker, M., & Turner, D.H. (1999) Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure. Journal of Molecular Biology, Vol. 288. 911-940.
- Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., & Bateman, A. (2005) Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Research, Vol. 33, Database Issue. D121-D124.
- Griffiths-Jones, S.(2004) The microRNA Registry. Nucleic Acids Research, Vol. 32, Database Issue. D109-D111.
- Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., Enright, A.J. (2006) MiRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Research, Vol. 34, Database Issue. D140-D144.
- Griffiths-Jones, S., Saini, H.K., Bateman, A., Enright, A.J. (2008) MiRBase: tools for microRNA genomics. Nucleic Acids Research, Vol. 36, Database Issue. D154-D158.