Working with RNA

Yazdan Asgari

2019

RNA in Bioinformatics

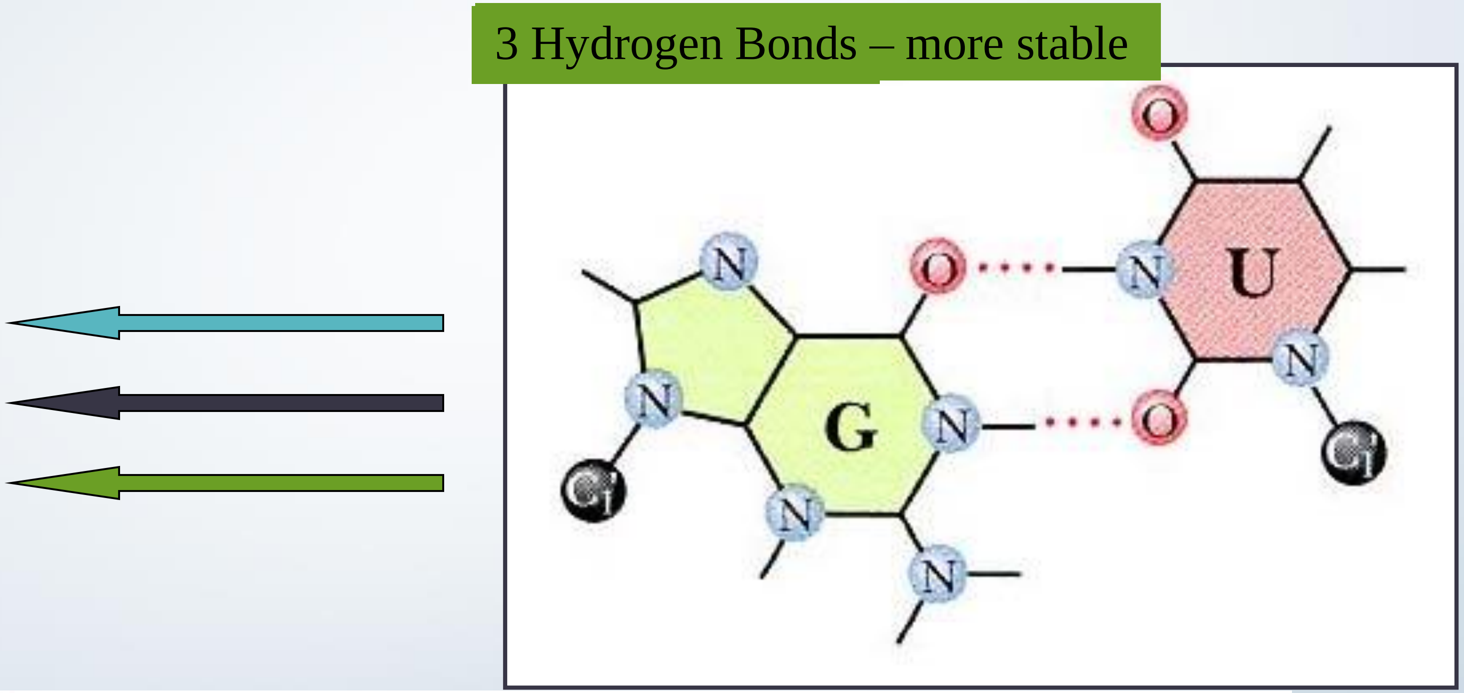
- RNA is the “poor cousin” of bioinformatics
- The NCBI has no RNA section
- The RNA world is receiving ever-increasing attention

Basics

RNA Basics

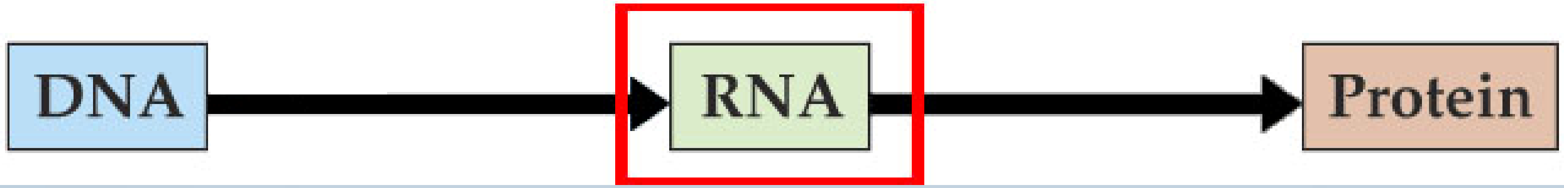
- RNA bases A,C,G,U
- Canonical Base Pairs
  - A-U
  - G-C
  - G-U

    “wobble” pairing



3 Hydrogen Bonds – more stable

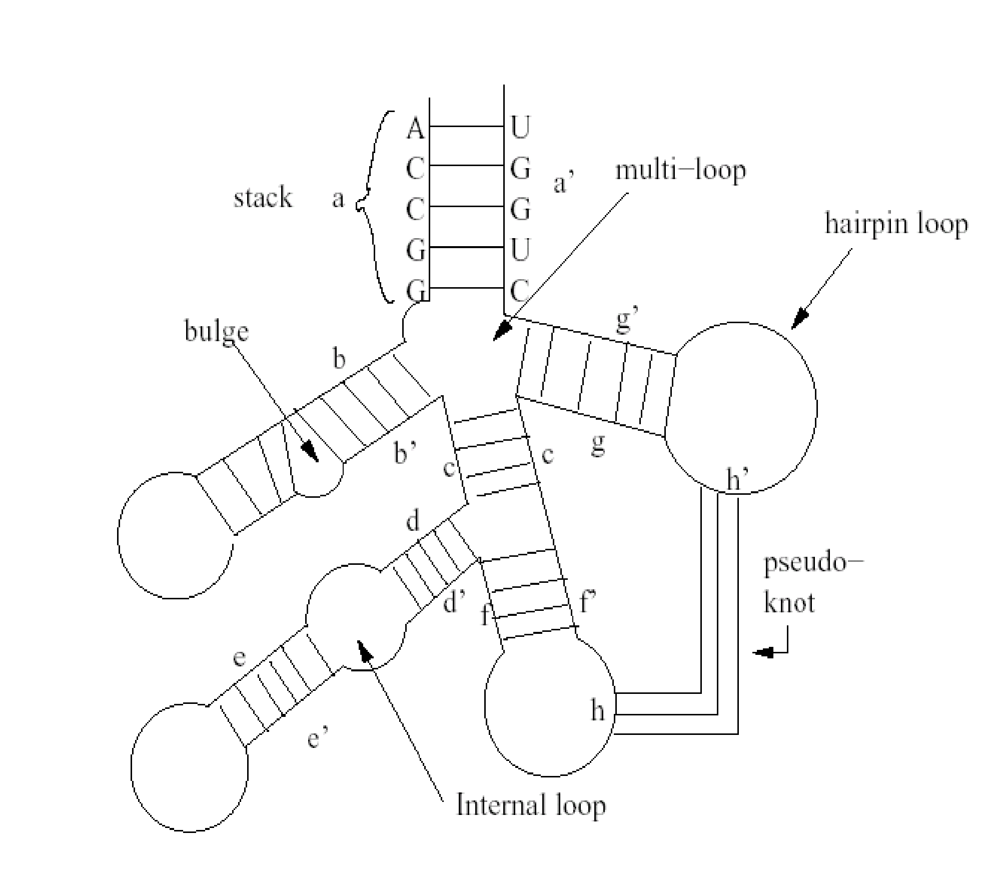
  - Bases can only pair with **one** other base.



RNA Secondary Structure

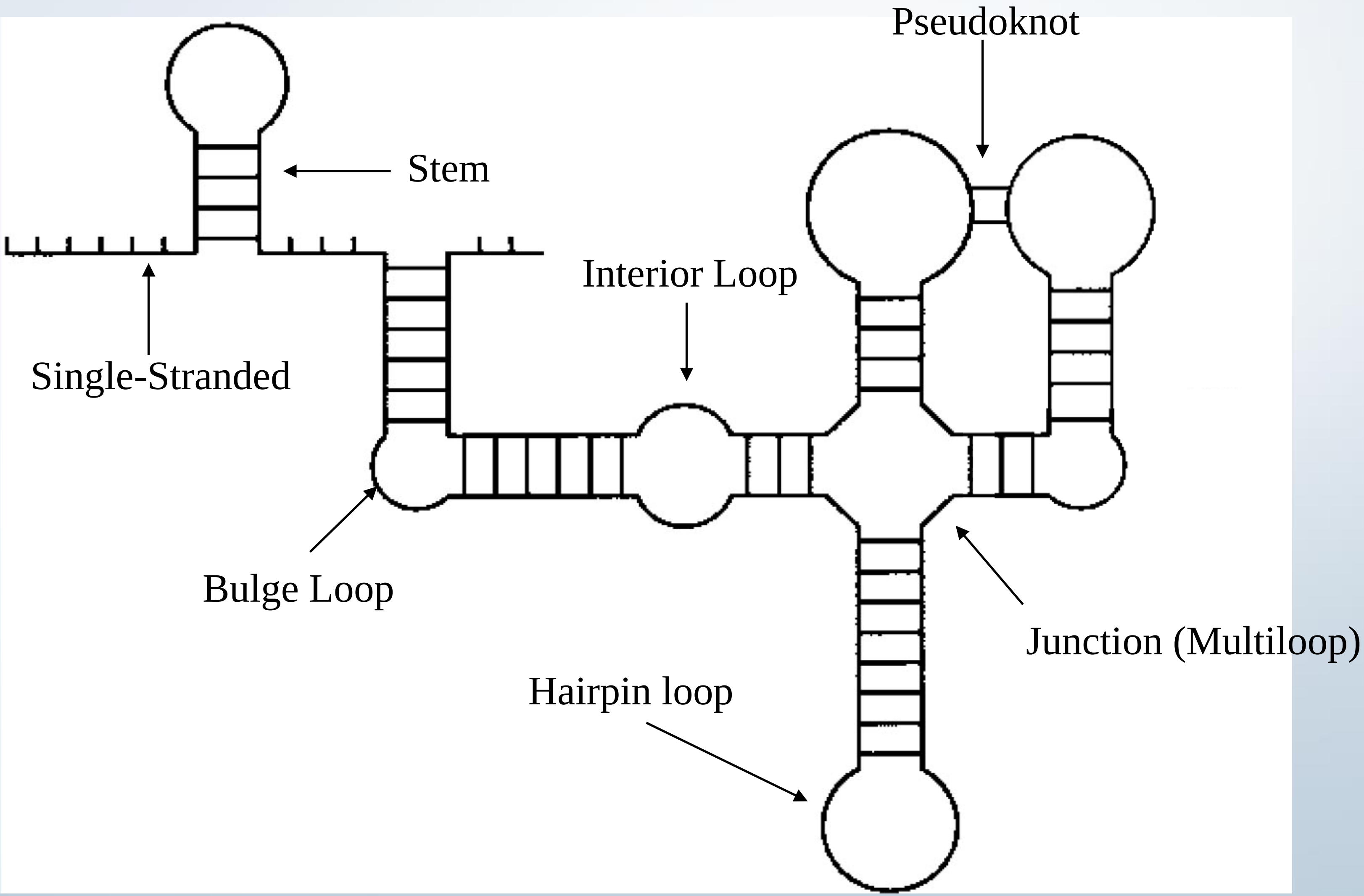
- Stacks: continuous nested basepairs (energetically favorable)

- Non-basepaired loops:
  - Hairpin loop
  - Bulge
  - Internal loop
  - Multiloop
  - Pseudo-knot



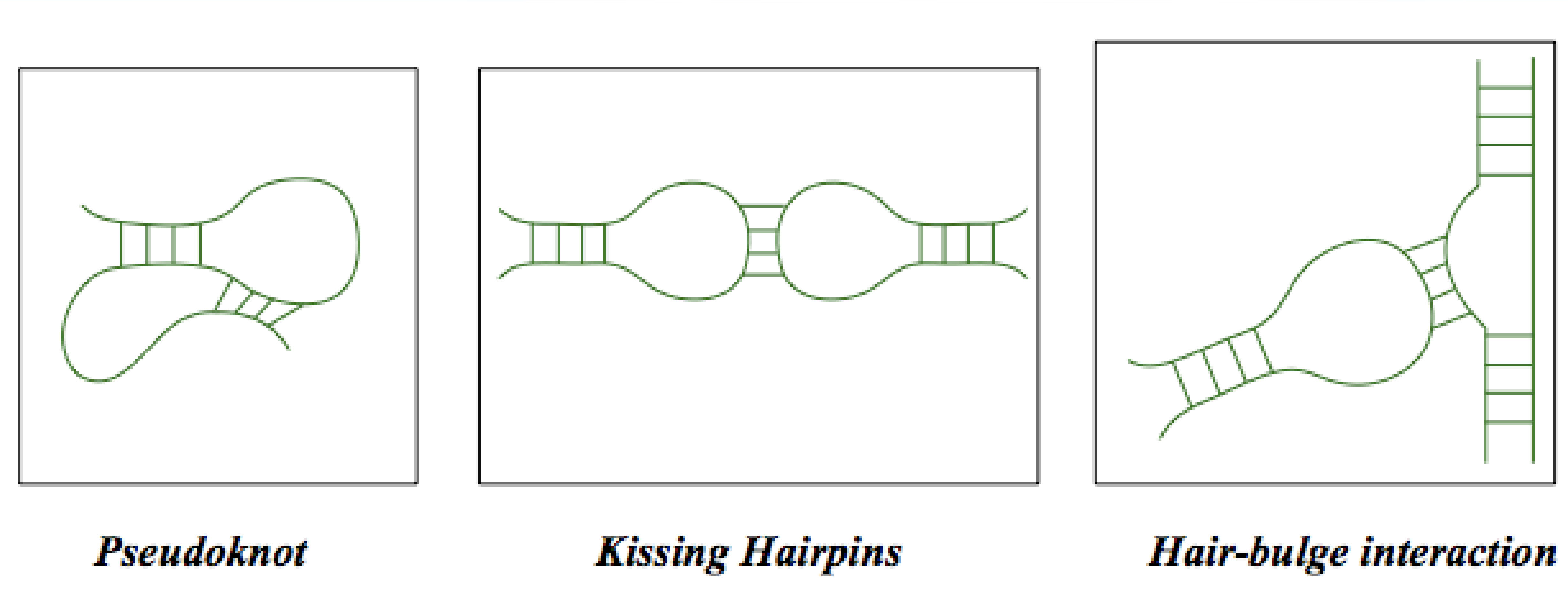
Bioinformatics: Life Sciences, Robert Lessick

RNA Secondary Structure



Bioinformatics: Life Sciences, Robert Lessick

Complex folds



Pseudoknot **Kissing Hairpins** **Hair-bulge interaction**

Bioinformatics: Life Sciences, Robert Lessick

Predicting RNA Secondary Structures

Main approaches to RNA secondary structure prediction

1. Energy minimization
   - dynamic programming approach
   - does not require prior sequence alignment
   - require estimation of energy terms contributing to secondary structure
2. Comparative sequence analysis
   - using sequence alignment to find conserved residues and covariant base pairs.
   - most trusted
3. Simultaneous folding and alignment (structural alignment)

1. Assumptions in energy minimization approaches

- Most likely structure similar to energetically most stable structure
- Energy associated with any position is only influenced by local sequence and structure
- Neglect pseudoknots

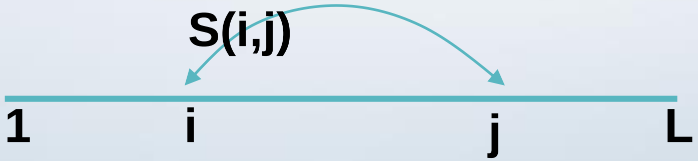
Base-pair maximization

- Find structure with the most base pairs
  - Only consider A-U and G-C and do not distinguish them

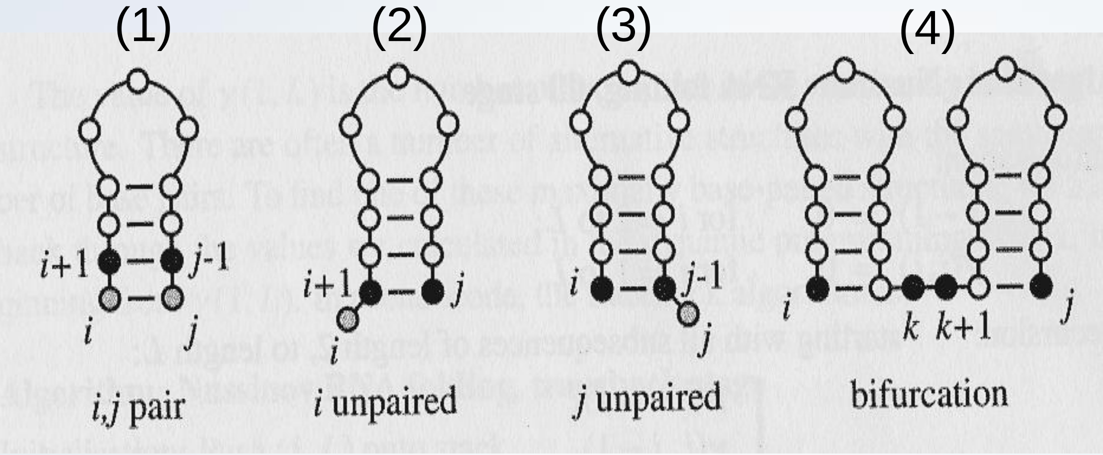
- Nussinov algorithm (1970s)
  - Too simple to be accurate, but stepping-stone for later algorithms

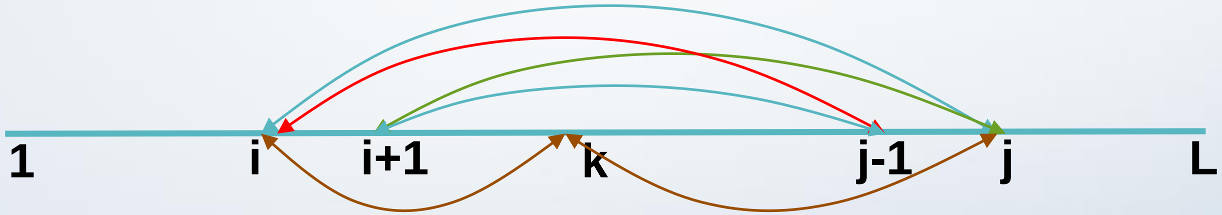
Nussinov algorithm

- Problem definition
  - Given sequence $X=x_1x_2\ldots x_{L,}$compute a structure that has maximum (weighted) number of base pairings
- How can we solve this problem?
  - Remember: RNA folds back to itself!
  - S(i,j) is the maximum score when $x_i..x_j$ folds optimally
  - S(1,L)?
  - S(i,i)?



“Grow” from substructures





$$S(i,j) = max \begin{cases} S(i+1, j-1) + w(i,j) & (1) \\ S(i+1, j) & (2) \\ S(i, j-1) & (3) \\ max_{i<k<j} S(i,k) + S(k+1, j) & (4) \end{cases}$$

$w(i,j) = 1$ if i, j are complementary (i.e., GC, CG, AU or UA); 0 otherwise

Bioinformatics: Life Sciences, Robert Lessick

Dynamic programming

- Compute S(i,j) recursively (dynamic programming)
  - Compares a sequence against itself in a dynamic programming matrix

- Three steps
  1. Initialization
  2. Recursion
  3. Traceback

Initialization

	G	G	G	A	A	A	U	C	C
G	0								
G	0	0							
G		0	0						
A			0	0					
A				0	0				
A					0	0			
U						0	0		
C							0	0	
C								0	0

Example:

GGGAAAUCC

$$S(i,i) = 0 \quad \forall \quad 1 \leq i \leq L$$

→ the main diagonal

$$S(i,i-1) = 0 \quad \forall \quad 2 \leq i \leq L$$

→ the diagonal below

L: the length of input sequence

Recursion

Fill up the table (DP matrix) -- diagonal by diagonal

→ *j*

↓ *i*

	G	G	G	A	A	A	U	C	C
G	0	0	0	0					
G	0	0	0	0	0				
G		0	0	0	0	0			
A			0	0	0	0	?		
A				0	0	0	1		
A					0	0	1	1	
U						0	0	0	0
C							0	0	0
C								0	0

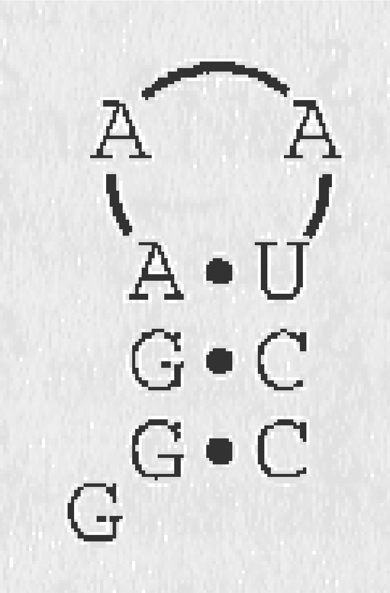
$$S(i,j) = max \begin{cases} S(i+1, j-1) + w(i,j) & (1) \\ S(i+1, j) & (2) \\ S(i, j-1) & (3) \\ max_{i<k<j} S(i,k) + S(k+1, j) & (4) \end{cases}$$

$$w(i,j) = \begin{cases} 1 & i,\, j \text{ are complementary} \\ 0 & otherwise \end{cases}$$

Traceback

	G	G	G	A	A	A	U	C	C
G	0	0	0	0	0	0	1	2	3
G	0	0	0	0	0	0	1	2	3
G		0	0	0	0	0	1	2	2
A			0	0	0	0	1	1	1
A				0	0	0	1	1	1
A					0	0	1	1	1
U						0	0	0	0
C							0	0	0
C								0	0

The structure is:



What are the other “optimal” structures?

An exercise

- Input: AUGACAU
- Fill up the table
- Trace back

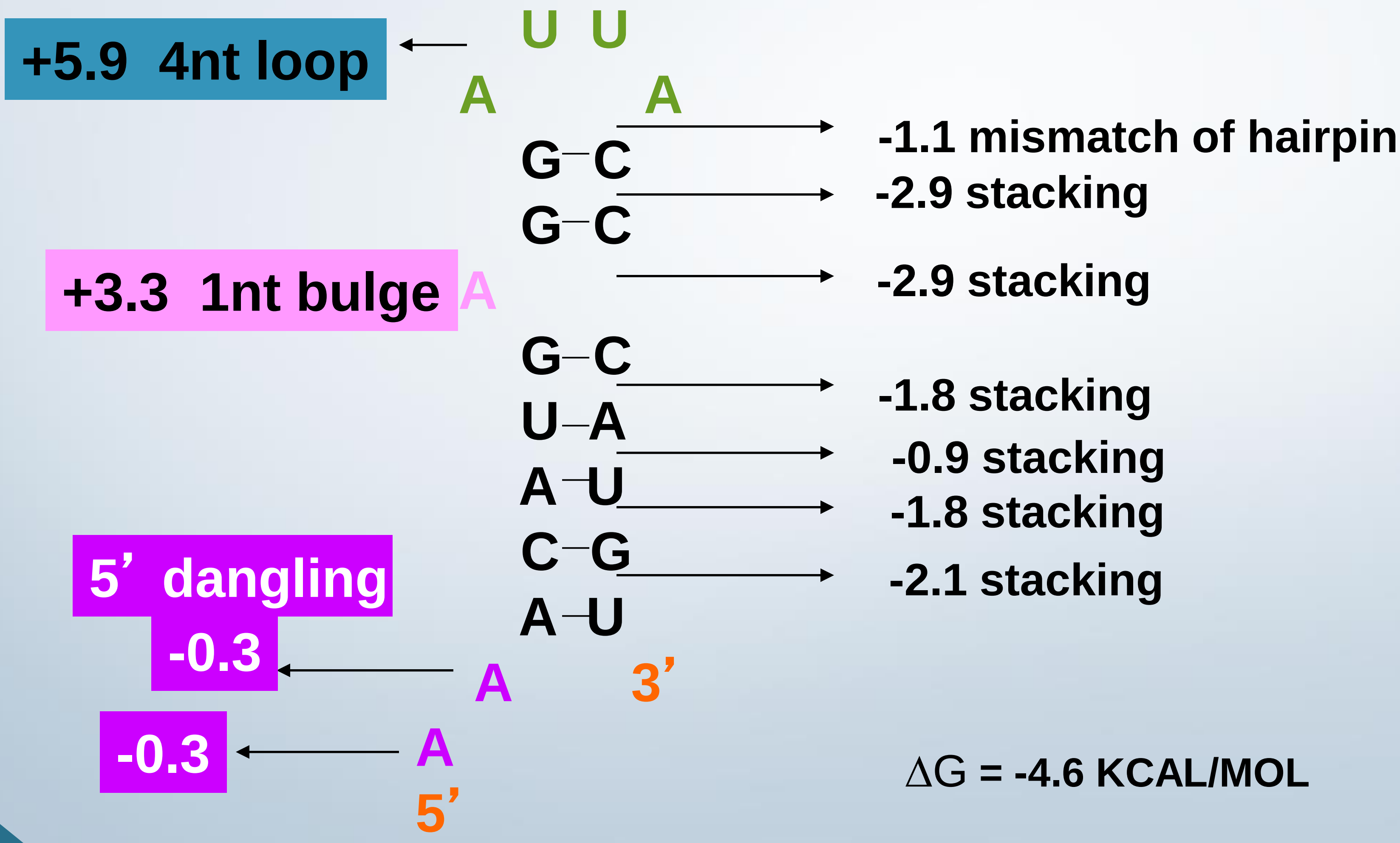
	A	U	G	A	C	A	U
A							
U							
G							
A							
C							
A							
U							

- Give the optimal structure
- What’s the size of the hairpin loop

Energy minimization methods

- Nussinov algorithm (base pair maximization) is too simple to be accurate
- Energy minimization algorithm predicts secondary structure by minimizing the free energy (ΔG)
- ΔG calculated as sum of individual contributions of:
  - loops
  - stacking

Free energy computation



+5.9 4nt loop

+3.3 1nt bulge

5' dangling -0.3

-0.3

-1.1 mismatch of hairpin

-2.9 stacking

-2.9 stacking

-1.8 stacking

-0.9 stacking

-1.8 stacking

-2.1 stacking

ΔG = -4.6 KCAL/MOL

Loop parameters (from Mfold)

DESTABILIZING ENERGIES BY SIZE OF LOOP

SIZE	INTERNAL	BULGE	HAIRPIN
1	.	3.8	.
2	.	2.8	.
3	.	3.2	5.4
4	1.1	3.6	5.6
5	2.1	4.0	5.7
6	1.9	4.4	5.4
.			
.			
12	2.6	5.1	6.7
13	2.7	5.2	6.8
14	2.8	5.3	6.9
15	2.8	5.4	6.9

Unit: Kcal/mol

Stacking energy
(from Vienna package)

```
# stack_energies
/*  CG    GC    GU    UG    AU    UA    @  */
  -2.0  -2.9  -1.9  -1.2  -1.7  -1.8    0
  -2.9  -3.4  -2.1  -1.4  -2.1  -2.3    0
  -1.9  -2.1   1.5   -.4  -1.0  -1.1    0
  -1.2  -1.4   -.4   -.2   -.5   -.8    0
  -1.7   -.2  -1.0   -.5   -.9   -.9    0
  -1.8  -2.3  -1.1   -.8   -.9  -1.1    0
     0     0     0     0     0     0    0
```

Mfold versus Vienna package

- Mfold
  - http://unafold.rna.albany.edu/?q=mfold
  - Suboptimal structures
    - The correct structure is not necessarily structure with optimal free energy
    - Within a certain threshold of the calculated minimum energy
- Vienna -- calculate the probability of base pairings
  - http://www.tbi.univie.ac.at/RNA/

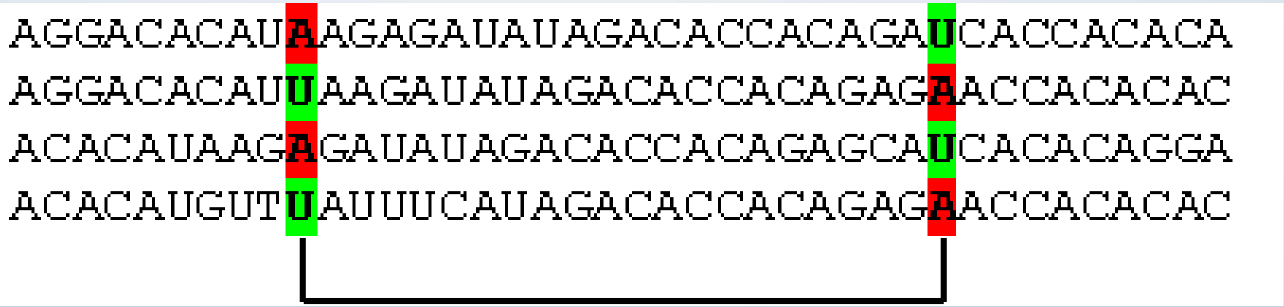
2. Inferring structure by comparative sequence analysis

- Need a multiple sequence alignment as input
- Requires sequences be similar enough (so that they can be initially aligned)
- Sequences should be dissimilar enough for covarying substitutions to be detected

"Given an accurate multiple alignment, a large number of sequences, and sufficient sequence diversity, comparative analysis alone is sufficient to produce accurate structure predictions" *(Gutell RR et al. Curr Opin Struct Biol 2002, 12:301-310)*

RNA Secondary Structure Predictions based on MSA

- The best predictions are based on multiple-sequence alignments
- They take advantage of covariation



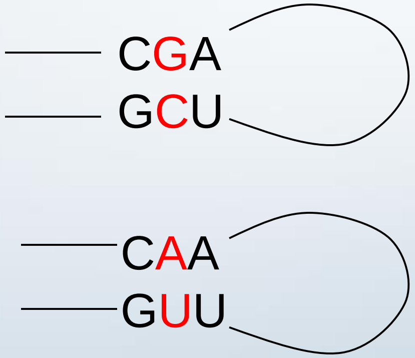
```
AGGACACAUAAGAGAUAUAGACACCACAGAUCACCACACA
AGGACACAUUAAGAUAUAGACACCACAGAGAACCACACAC
ACACAUAAGAGAUAUAGACACCACAGAGCAUCACACAGGA
ACACAUGUTUAUUUCAUAGACACCACAGAGAACCACACAC
```

Bioinformatics: Life Sciences, Robert Lessick

RNA variations

- Variations in RNA sequence maintain base-pairing patterns for secondary structures (conserved patterns of base-pairing)

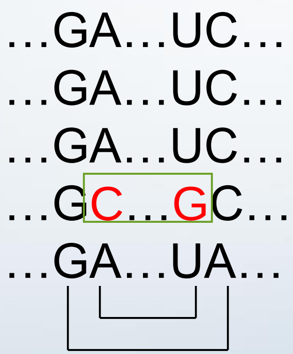
- When a nucleotide in one base changes, the base it pairs to must also change to maintain the same structure



- Such variation is referred to as ***covariation***.

If neglect covariation

- In usual alignment algorithms they are doubly penalized



…GA…UC…
…GA…UC…
…GA…UC…
…GC…GC…
…GA…UA…

Covariance measurements

- Mutual information (desirable for large datasets)
  - Most common measurement
  - Used in CM (Covariance Model) for structure prediction

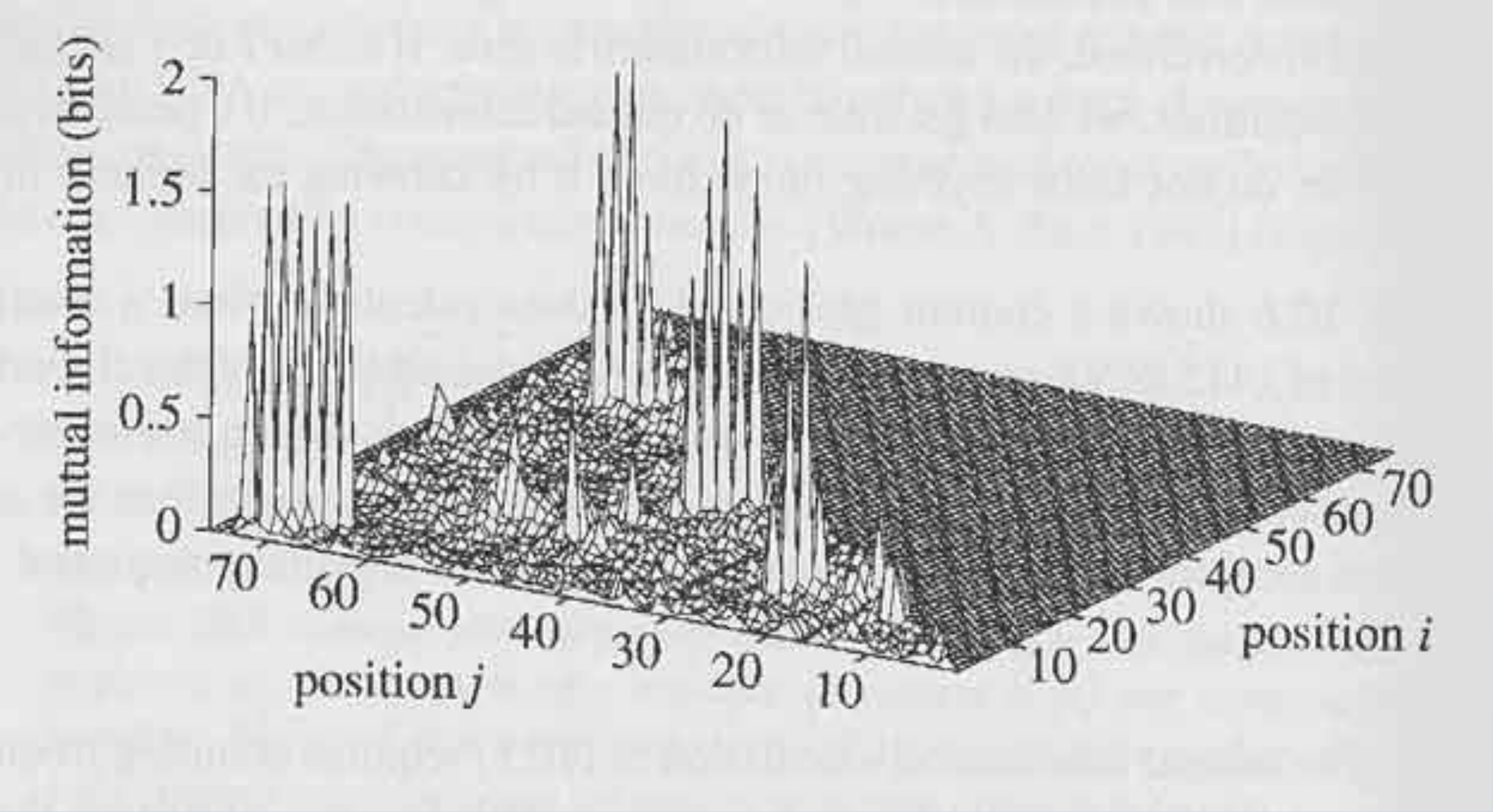
- Covariance score (better for small datasets)

Mutual information

$$MI_{ij} = \sum_{x_i y_j} f_{x_i y_j} log_2 \frac{f_{x_i y_j}}{f_{x_i} f_{x_j}}$$

- f_{x_i} : frequency of a base in column *i*
- $f_{x_i y_j}$: joint (pairwise) frequency of a base pair between columns *i* and *j*
- Mutual information should be reported in the range [0,H(A)], where zero corresponds to two totally uncorrelated images and H(A) corresponds to perfectly correlated images, case in which H(A)=H(B)
- If *i* and *j* are uncorrelated (independent), mutual information is 0

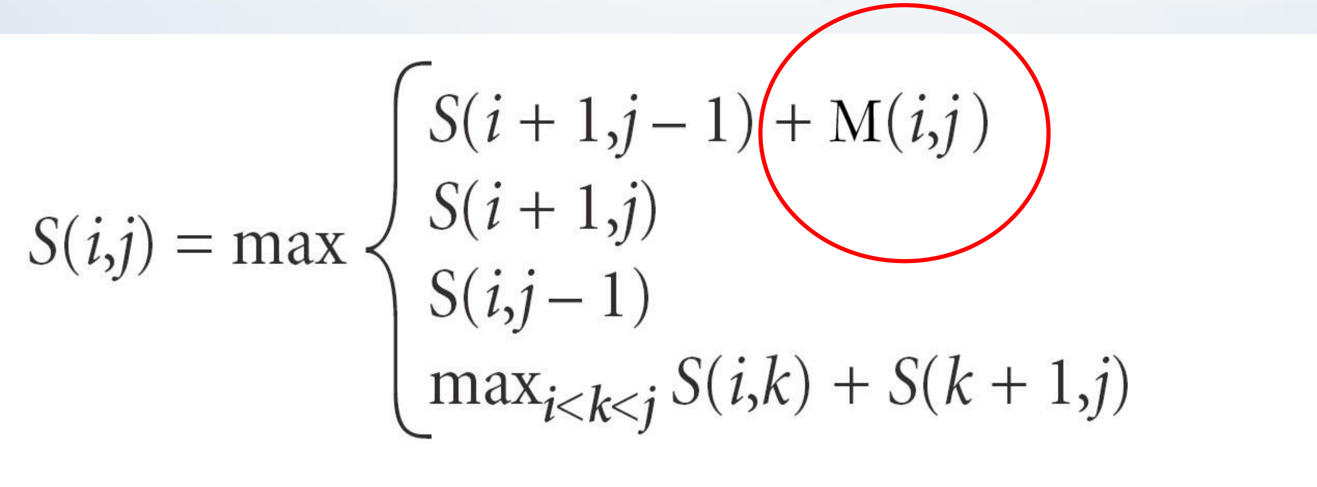
Mutual information plot



Bioinformatics: Life Sciences, Robert Lessick

Structure prediction using MI

- *S(i,j)* = Score at indices *i* and *j; M(i,j)* is the mutual information between i and j
- The goal is to maximize the total mutual information of input RNA
- The recursion is just like the one in Nussinov Algorithm, just to replace w(*i,j*) (1 or 0) with the mutual information *M(i,j)*



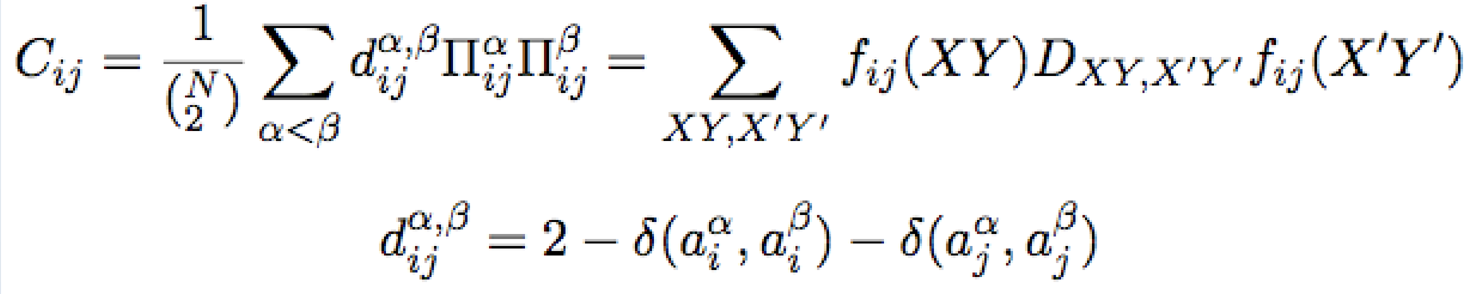
$$S(i,j) = \max \begin{cases} S(i+1, j-1) + M(i,j) \\ S(i+1, j) \\ S(i, j-1) \\ \max_{i<k<j} S(i,k) + S(k+1, j) \end{cases}$$

Covariance-like score

- RNAalifold
  - Hofacker et al. JMB 2002, 319:1059-1066
- Desirable for small datasets
- Combination of covariance score and thermodynamics energy

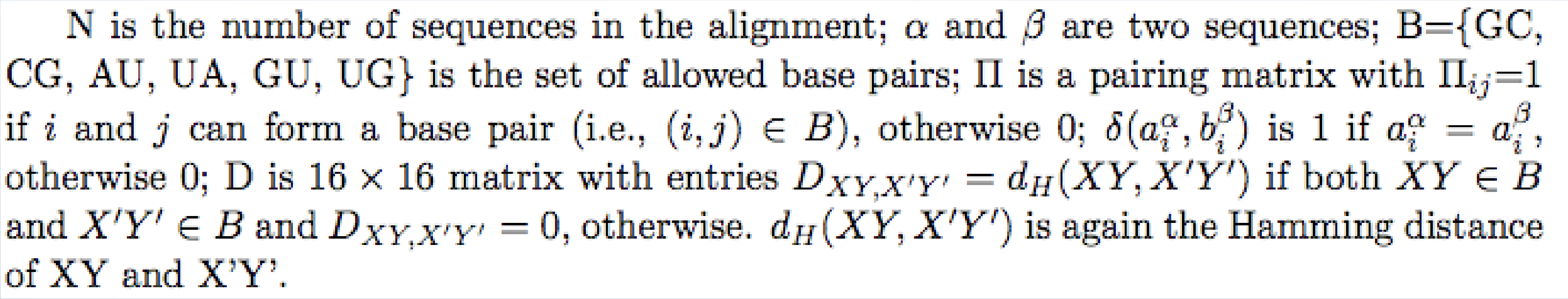
Covariance-like score calculation

The score between two columns *i* and *j* of an input multiple alignment is computed as following:



$$C_{ij} = \frac{1}{\binom{N}{2}} \sum_{\alpha<\beta} d_{ij}^{\alpha,\beta} \Pi_{ij}^{\alpha} \Pi_{ij}^{\beta} = \sum_{XY,X'Y'} f_{ij}(XY) D_{XY,X'Y'} f_{ij}(X'Y')$$

$$d_{ij}^{\alpha,\beta} = 2 - \delta(a_i^\alpha, a_i^\beta) - \delta(a_j^\alpha, a_j^\beta)$$

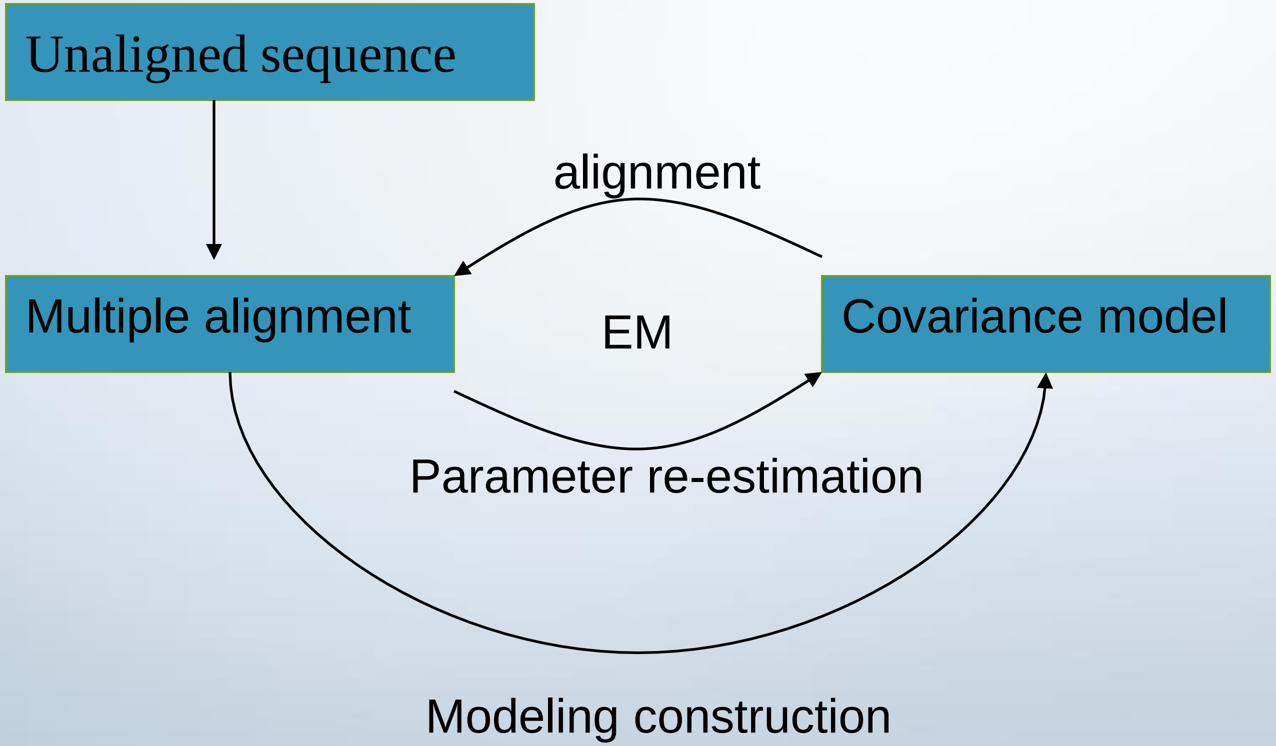


N is the number of sequences in the alignment; α and β are two sequences; B={GC, CG, AU, UA, GU, UG} is the set of allowed base pairs; Π is a pairing matrix with Π_{ij}=1 if i and j can form a base pair (i.e., $(i,j) \in B$), otherwise 0; $\delta(a_i^\alpha, b_i^\beta)$ is 1 if $a_i^\alpha = a_i^\beta$, otherwise 0; D is 16 × 16 matrix with entries $D_{XY,X'Y'} = d_H(XY, X'Y')$ if both $XY \in B$ and $X'Y' \in B$ and $D_{XY,X'Y'} = 0$, otherwise. $d_H(XY, X'Y')$ is again the Hamming distance of XY and X'Y'.

Covariance model (CM)

- A formal covariance model(CM), devised by Eddy and Durbin
  - A probabilistic model
  - ≈ A Stochastic Context-Free Grammer
  - Generalized HMM model
- A CM is like a sequence profile, but it scores a combination of sequence consensus and RNA secondary structure consensus
- Provides very accurate results
- Very slow and unsuitable for searching large genomes

CM - training algorithm



Unaligned sequence

alignment

Multiple alignment

EM

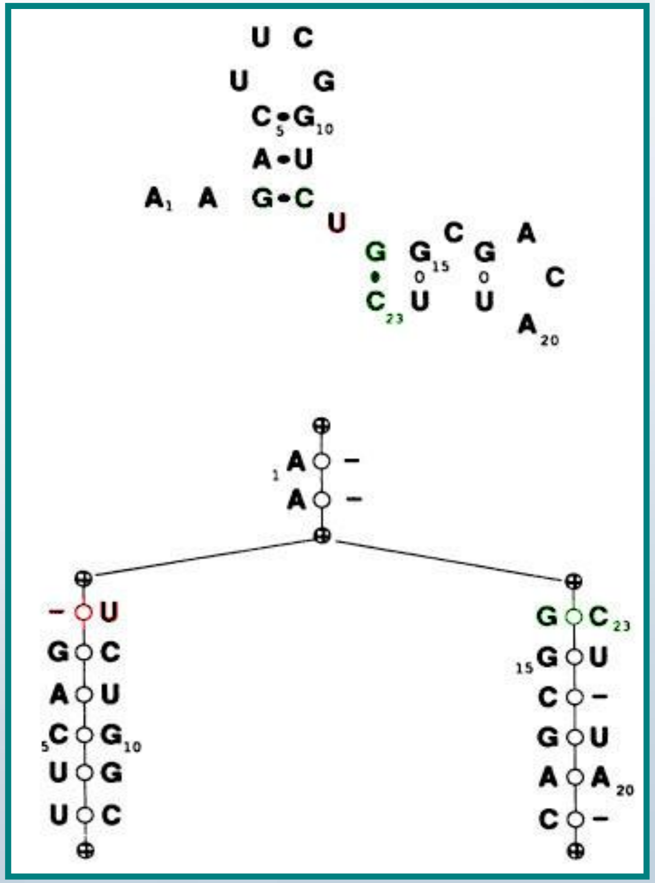
Covariance model

Parameter re-estimation

Modeling construction

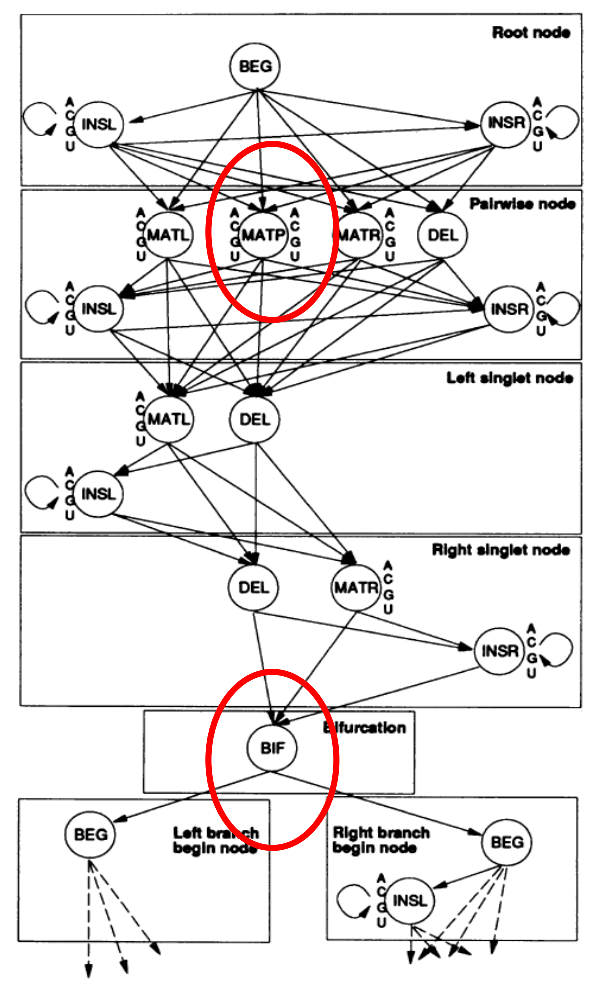
CM - Binary tree representation of RNA secondary structure

- Representation of RNA structure using Binary tree
- Nodes represent
  - Base pair if two bases are shown
  - Loop if base and “gap” (dash) are shown
- Pseudoknots still not represented
- Tree does not permit varying sequences
  - Mismatches
  - Insertions & Deletions



Bioinformatics: Life Sciences, Robert Lessick

CM - Overall architecture



Bioinformatics: Life Sciences, Robert Lessick

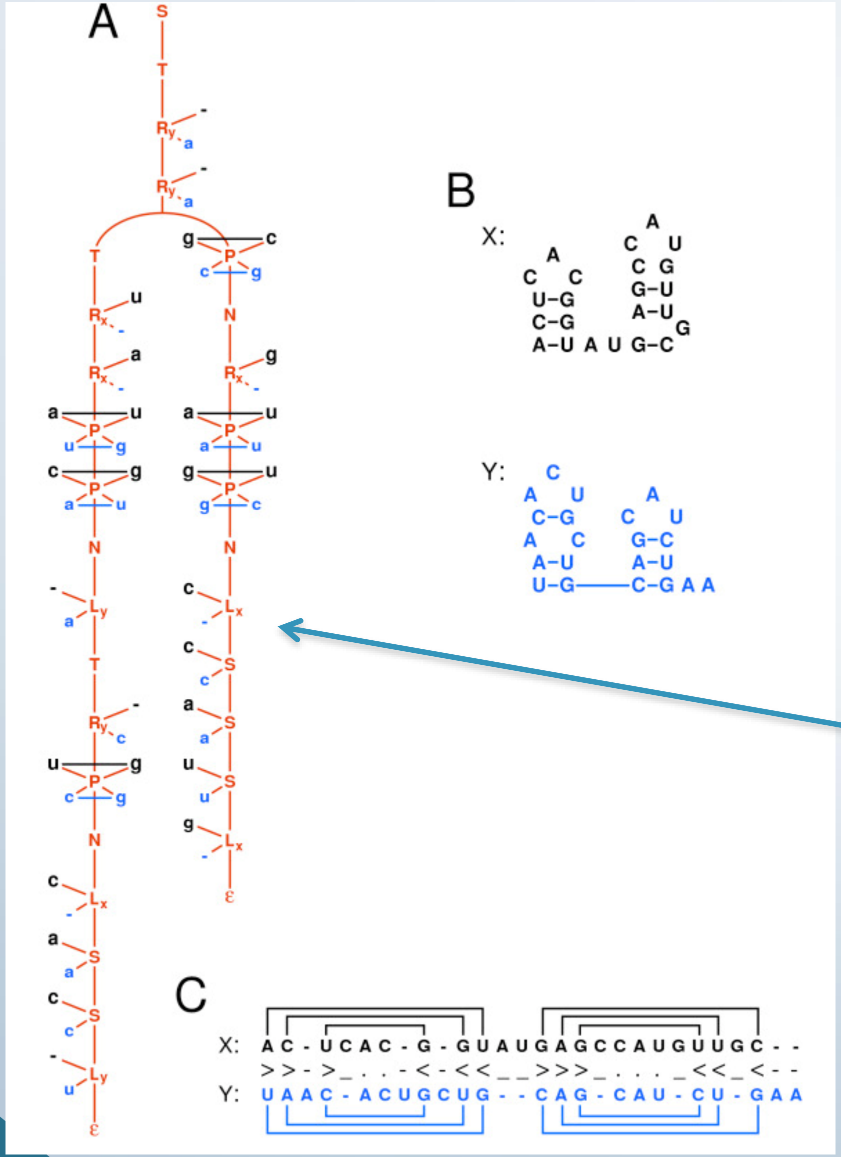
MATP emits pairs of bases: modeling of base pairing

BIF allows multiple helices (bifurcation)

CM - Drawbacks

- Needs to be well trained (large datasets)
- Not suitable for searches of large RNA
  - Structural complexity of large RNA cannot be modeled
  - Runtime
  - Memory requirements

3. Simultaneous structure prediction and alignment of ncRNAs



A
B
X:
Y:
C
X: AC-UCAC-G-GUAUGAGCCAUGUUGC--
>>->_..-<-<<__>>>_..._<<_<--
Y: UAAC-ACUGCUG--CAG-CAU-CU-GAA

The grammar emits two correlated sequences, x and y

http://www.biomedcentral.com/1471-2105/7/400

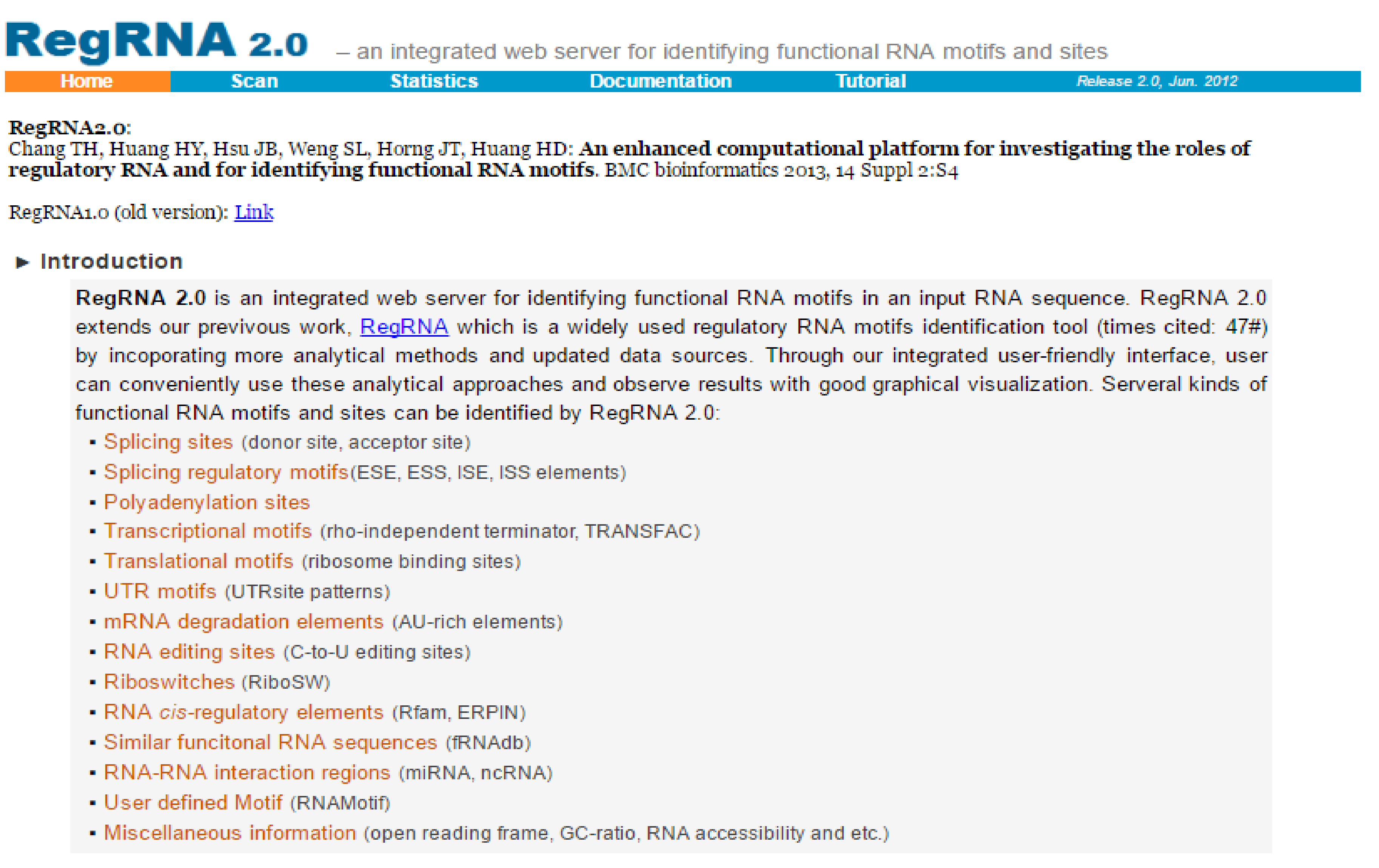
RNA motifs

Searching Databases for RNA motifs

- It is possible to search databases for RNA sequences that can fold according to a specified pattern

- Example:
  - RegRNA (http://regrna2.mbc.nctu.edu.tw/)

RegRNA



RegRNA 2.0
– an integrated web server for identifying functional RNA motifs and sites
Home
Scan
Statistics
Documentation
Tutorial
Release 2.0, Jun. 2012

RegRNA2.0:
Chang TH, Huang HY, Hsu JB, Weng SL, Horng JT, Huang HD: **An enhanced computational platform for investigating the roles of regulatory RNA and for identifying functional RNA motifs.** BMC bioinformatics 2013, 14 Suppl 2:S4

RegRNA1.0 (old version): Link

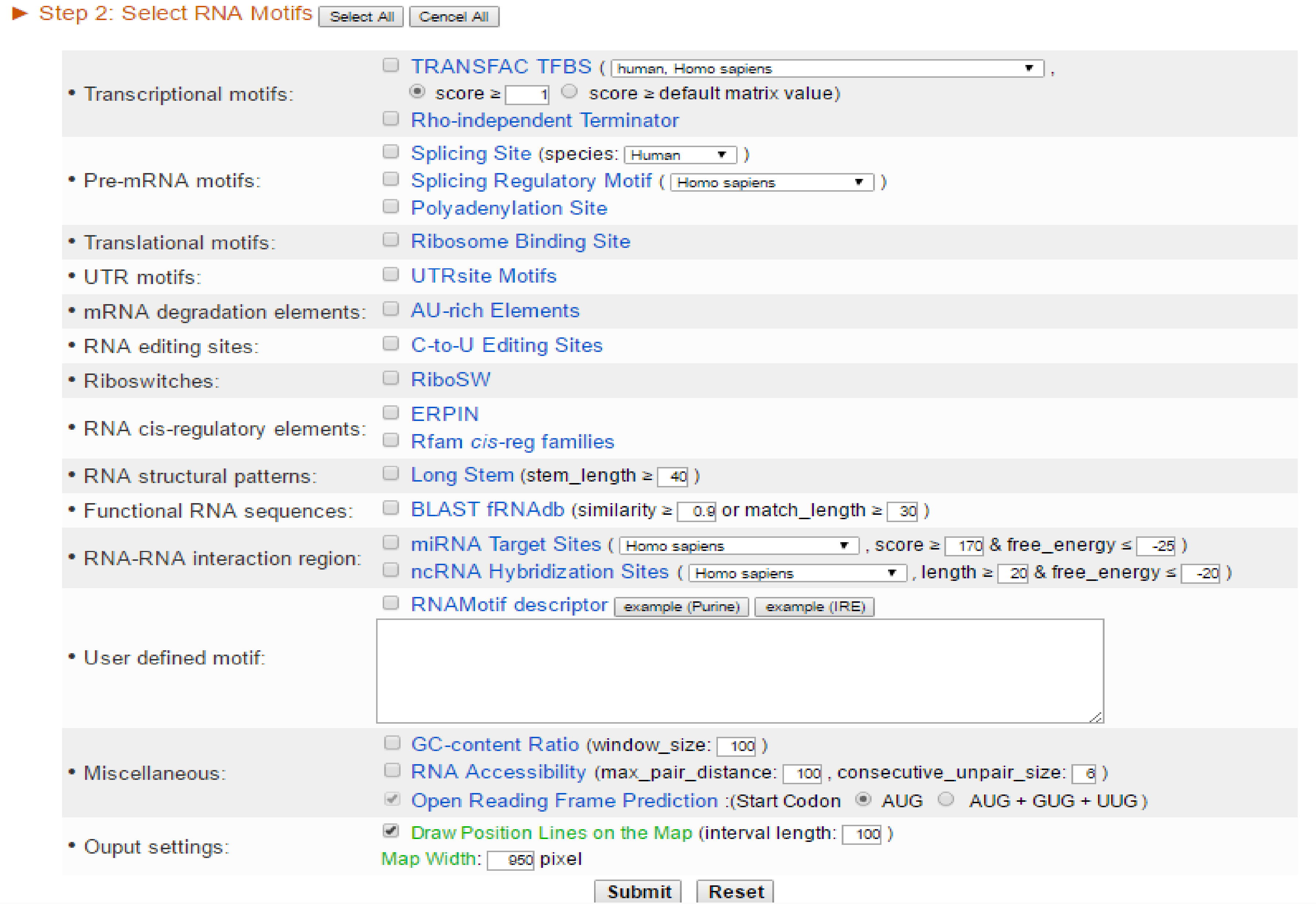
▸ **Introduction**

RegRNA 2.0 is an integrated web server for identifying functional RNA motifs in an input RNA sequence. RegRNA 2.0 extends our previvous work, RegRNA which is a widely used regulatory RNA motifs identification tool (times cited: 47#) by incoporating more analytical methods and updated data sources. Through our integrated user-friendly interface, user can conveniently use these analytical approaches and observe results with good graphical visualization. Serveral kinds of functional RNA motifs and sites can be identified by RegRNA 2.0:

- Splicing sites (donor site, acceptor site)
- Splicing regulatory motifs(ESE, ESS, ISE, ISS elements)
- Polyadenylation sites
- Transcriptional motifs (rho-independent terminator, TRANSFAC)
- Translational motifs (ribosome binding sites)
- UTR motifs (UTRsite patterns)
- mRNA degradation elements (AU-rich elements)
- RNA editing sites (C-to-U editing sites)
- Riboswitches (RiboSW)
- RNA *cis*-regulatory elements (Rfam, ERPIN)
- Similar funcitonal RNA sequences (fRNAdb)
- RNA-RNA interaction regions (miRNA, ncRNA)
- User defined Motif (RNAMotif)
- Miscellaneous information (open reading frame, GC-ratio, RNA accessibility and etc.)

http://regrna2.mbc.nctu.edu.tw

RegRNA

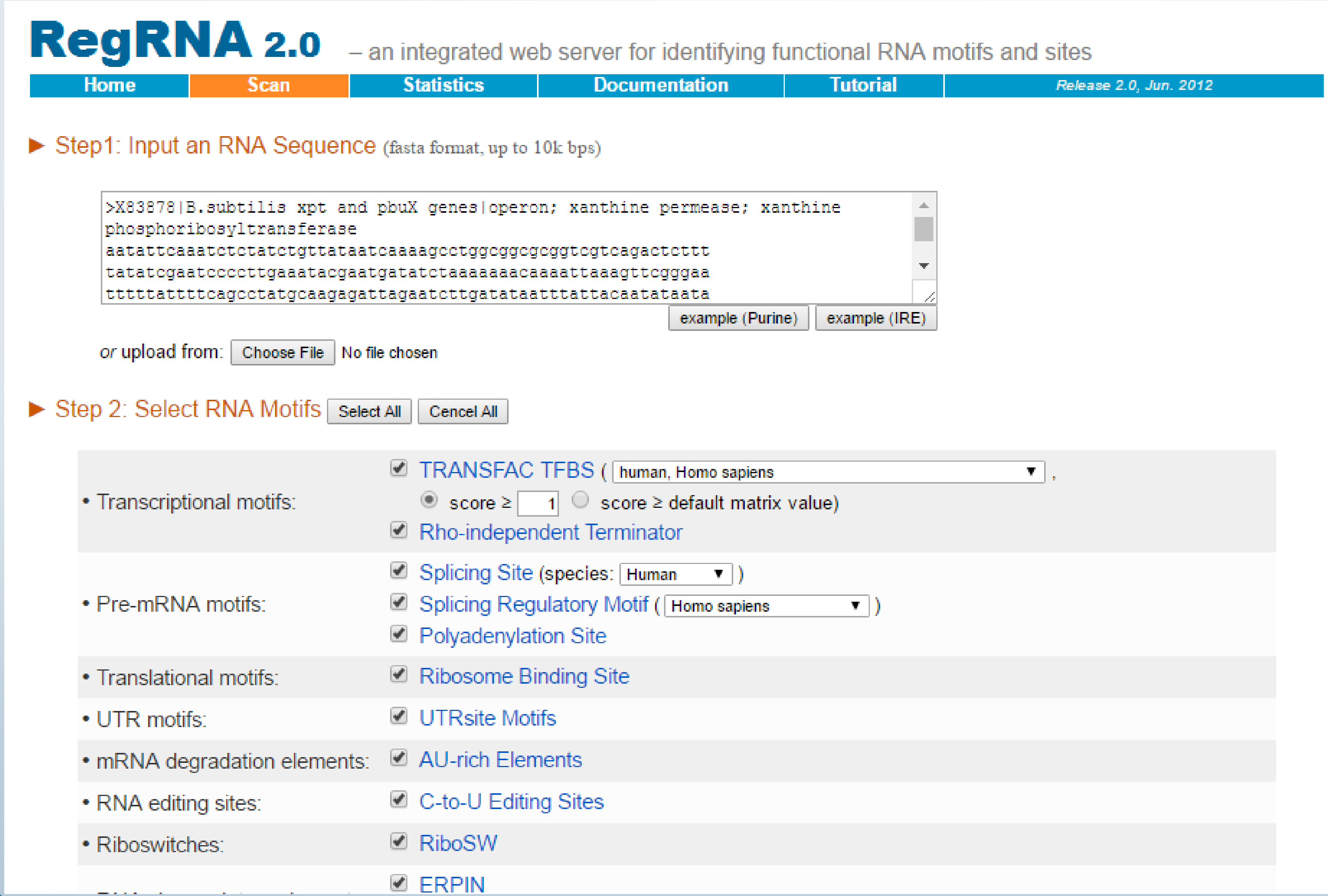


▶ Step 2: Select RNA Motifs [Select All] [Cencel All]

• Transcriptional motifs:	□ TRANSFAC TFBS (human, Homo sapiens ▾), ● score ≥ 1 ○ score ≥ default matrix value) □ Rho-independent Terminator
• Pre-mRNA motifs:	□ Splicing Site (species: Human ▾) □ Splicing Regulatory Motif (Homo sapiens ▾) □ Polyadenylation Site
• Translational motifs:	□ Ribosome Binding Site
• UTR motifs:	□ UTRsite Motifs
• mRNA degradation elements:	□ AU-rich Elements
• RNA editing sites:	□ C-to-U Editing Sites
• Riboswitches:	□ RiboSW
• RNA cis-regulatory elements:	□ ERPIN □ Rfam *cis*-reg families
• RNA structural patterns:	□ Long Stem (stem_length ≥ 40)
• Functional RNA sequences:	□ BLAST fRNAdb (similarity ≥ 0.9 or match_length ≥ 30)
• RNA-RNA interaction region:	□ miRNA Target Sites (Homo sapiens ▾ , score ≥ 170 & free_energy ≤ -25) □ ncRNA Hybridization Sites (Homo sapiens ▾ , length ≥ 20 & free_energy ≤ -20)
• User defined motif:	□ RNAMotif descriptor [example (Purine)] [example (IRE)]
• Miscellaneous:	□ GC-content Ratio (window_size: 100) □ RNA Accessibility (max_pair_distance: 100 , consecutive_unpair_size: 6) ☑ Open Reading Frame Prediction :(Start Codon ● AUG ○ AUG + GUG + UUG)
• Ouput settings:	☑ Draw Position Lines on the Map (interval length: 100) Map Width: 950 pixel

[Submit] [Reset]

RegRNA - Example


RegRNA 2.0 – an integrated web server for identifying functional RNA motifs and sites
Home
Scan
Statistics
Documentation
Tutorial
Release 2.0, Jun. 2012

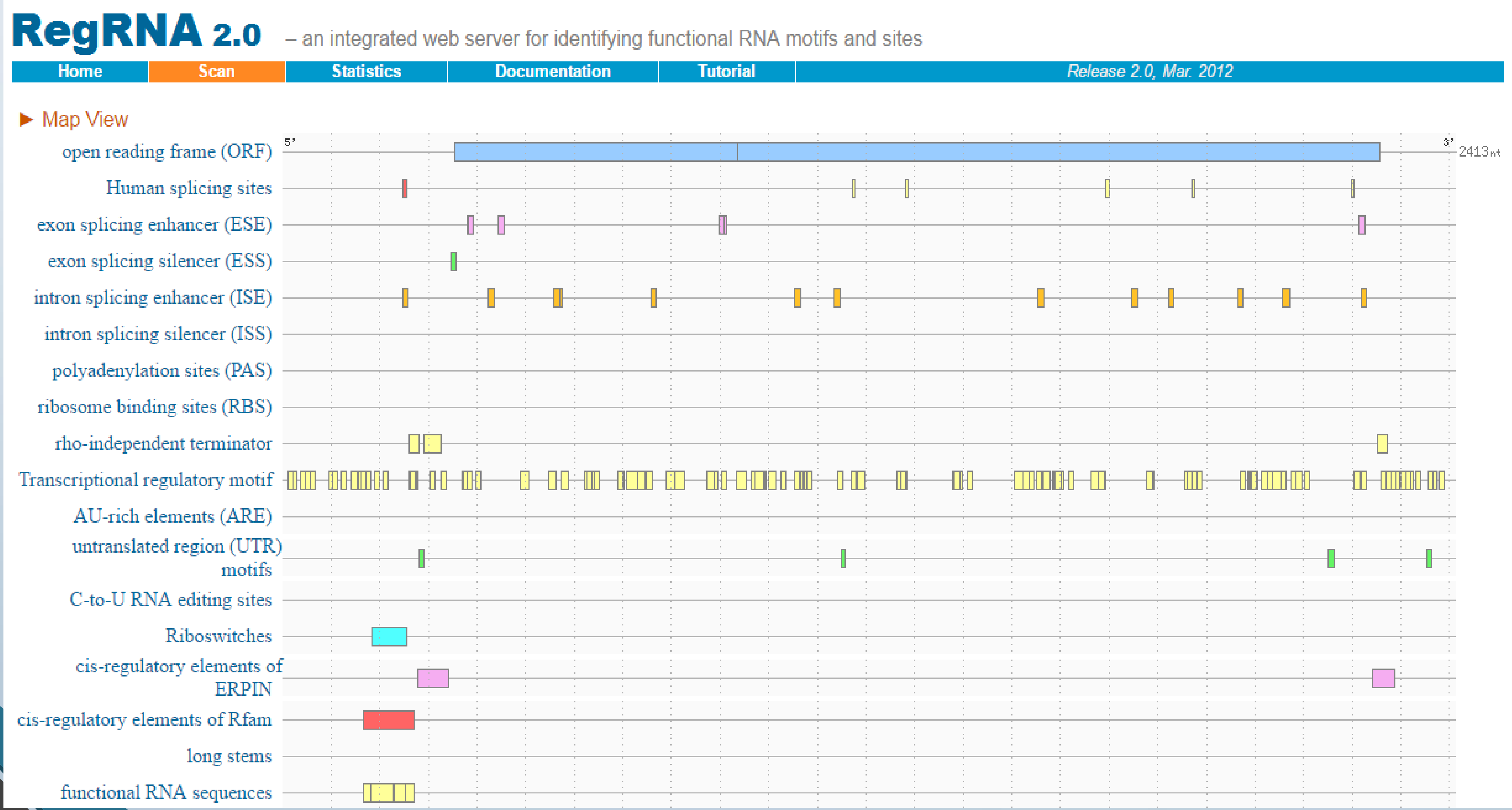
▶ Step1: Input an RNA Sequence (fasta format, up to 10k bps)

```
>X83878|B.subtilis xpt and pbuX genes|operon; xanthine permease; xanthine
phosphoribosyltransferase
aatattcaaatctctatctgttataatcaaaagcctggcggcgcggtcgtcagactcttt
tatatcgaatccccttgaaatacgaatgatatctaaaaaaacaaaattaaagttcgggaa
tttttattttcagcctatgcaagagattagaatcttgatataatttattacaatataata
```

▶ Step 2: Select RNA Motifs

- Transcriptional motifs: ☑ TRANSFAC TFBS (human, Homo sapiens ▾ , ● score ≥ 1 ○ score ≥ default matrix value) ☑ Rho-independent Terminator
- Pre-mRNA motifs: ☑ Splicing Site (species: Human ▾) ☑ Splicing Regulatory Motif (Homo sapiens ▾) ☑ Polyadenylation Site
- Translational motifs: ☑ Ribosome Binding Site
- UTR motifs: ☑ UTRsite Motifs
- mRNA degradation elements: ☑ AU-rich Elements
- RNA editing sites: ☑ C-to-U Editing Sites
- Riboswitches: ☑ RiboSW
- ☑ ERPIN

RegRNA - Example



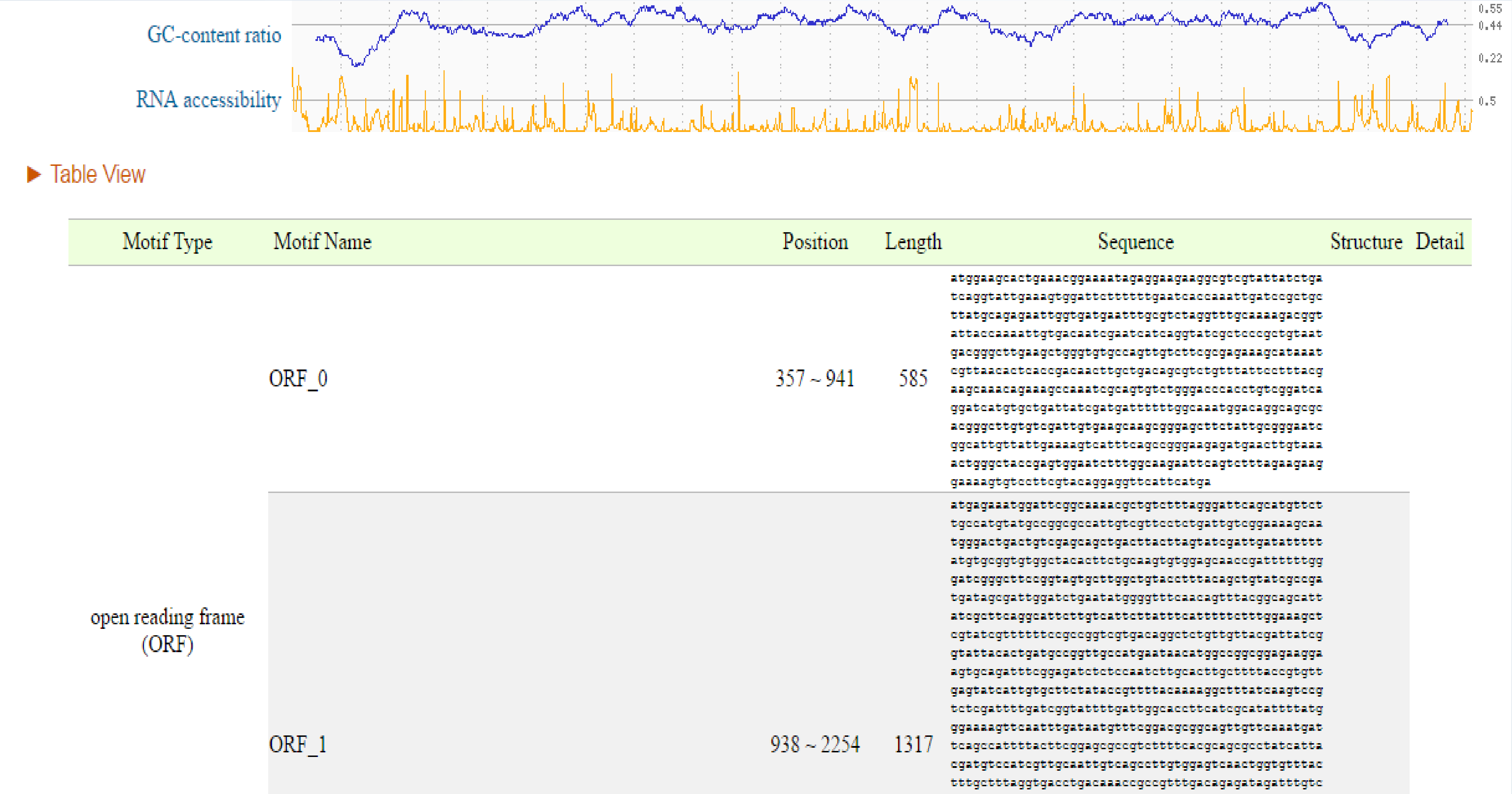
RegRNA 2.0 – an integrated web server for identifying functional RNA motifs and sites

Home
Scan
Statistics
Documentation
Tutorial
Release 2.0, Mar. 2012

▶ Map View

5'
3'
2413nt
open reading frame (ORF)
Human splicing sites
exon splicing enhancer (ESE)
exon splicing silencer (ESS)
intron splicing enhancer (ISE)
intron splicing silencer (ISS)
polyadenylation sites (PAS)
ribosome binding sites (RBS)
rho-independent terminator
Transcriptional regulatory motif
AU-rich elements (ARE)
untranslated region (UTR) motifs
C-to-U RNA editing sites
Riboswitches
cis-regulatory elements of ERPIN
cis-regulatory elements of Rfam
long stems
functional RNA sequences

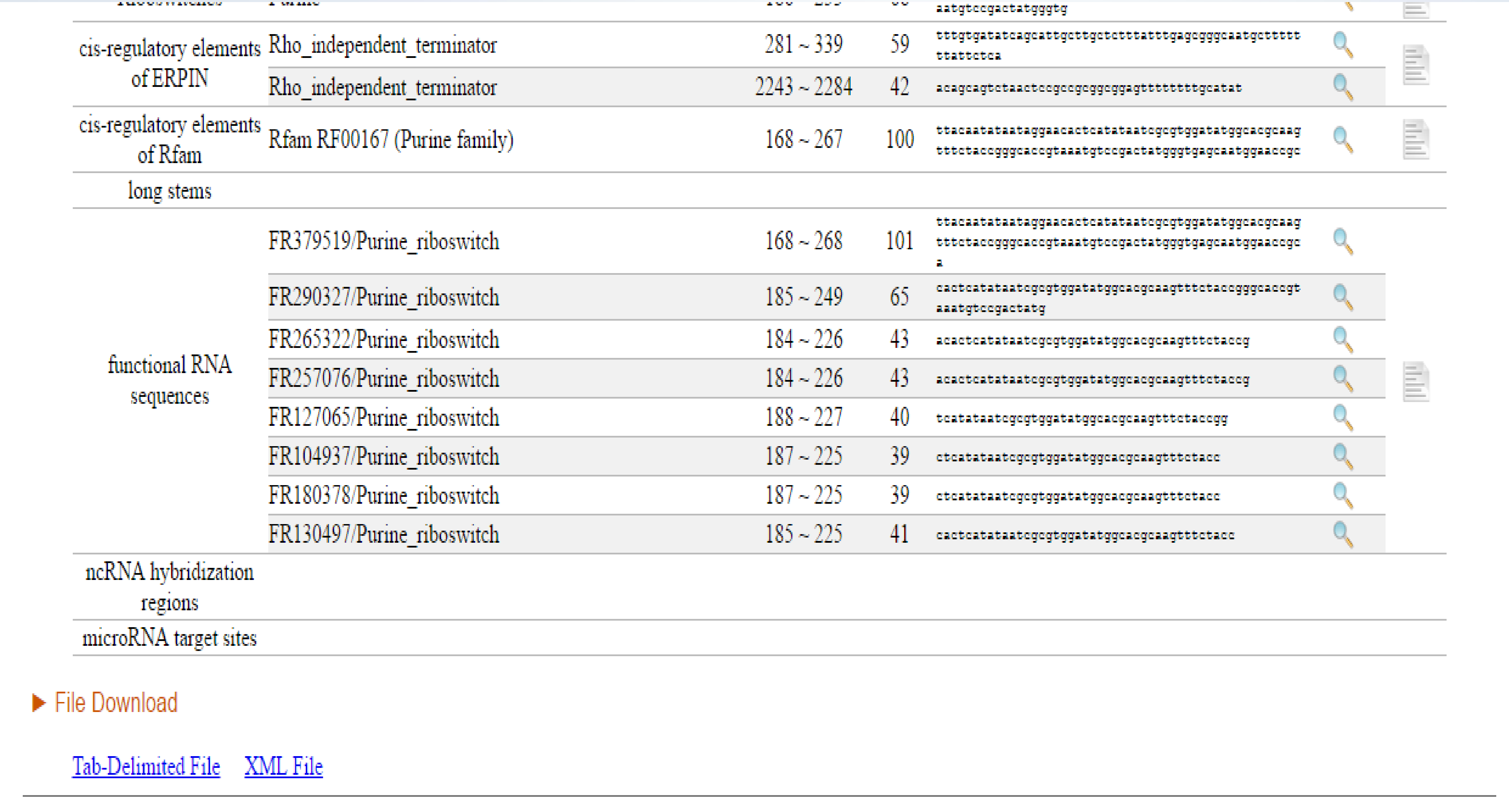
RegRNA - Example



▶ Table View

Motif Type	Motif Name	Position	Length	Sequence	Structure	Detail
open reading frame (ORF)	ORF_0	357 ~ 941	585	atggaagcactgaaacggaaaatagaggaagaaggcgtcgtattatctga tcaggtattgaaagtggattctttttttgaatcaccaaattgatccgctgc ttatgcagagaattggtgatgaatttgcgtctaggtttgcaaaagacggt attaccaaaattgtgacaatcgaatcatcaggtatcgctcccgctgtaat gacgggcttgaagctgggtgtgccagttgtcttcgcgagaaagcataaat cgttaacactcaccgacaacttgctgacagcgtctgtttattcctttacg aagcaaacagaaagccaaatcgcagtgtctgggacccacctgtcggatca ggatcatgtgctgattatcgatgatttttttggcaaatggacaggcagcgc acgggcttgtgtcgattgtgaagcaagcgggagcttctattgcgggaatc ggcattgttattgaaaagtcatttcagccgggaagagatgaacttgtaaa actgggctaccgagtggaatctttggcaagaattcagtctttagaagaag gaaaagtgtccttcgtacaggaggttcattcatga		
	ORF_1	938 ~ 2254	1317	atgagaaatggattcggcaaaacgctgtctttagggattcagcatgttct tgccatgtatgccggcgccattgtcgttcctctgattgtcggaaaagcaa tgggactgactgtcgagcagctgacttacttagtatcgattgatattttt atgtgcggtgtggctacacttctgcaagtgtggagcaaccgatttttttgg gatcgggcttccggtagtgcttggctgtacctttacagctgtatcgccga tgatagcgattggatctgaatatggggtttcaacagtttacggcagcatt atcgcttcaggcattcttgtcattcttatttcatttttctttggaaagct cgtatcgtttttttccgccggtcgtgacaggctctgttgttacgattatcg gtattacactgatgccggttgccatgaataacatggccggcggagaagga agtgcagatttcggagatctctccaatcttgcacttgcttttaccgtgtt gagtatcattgtgcttctataccgttttacaaaaggctttatcaagtccg tctcgattttgatcggtattttgattggcaccttcatcgcatattttatg ggaaaagttcaatttgataatgtttcggacgcggcagttgttcaaatgat tcagccattttacttcggagcgccgtcttttcacgcagcgcctatcatta cgatgtccatcgttgcaattgtcagccttgtggagtcaactggtgtttac tttgctttaggtgacctgacaaaccgccgtttgacagagatagatttgtc		

RegRNA - Example



Riboswitches	Purine	[illegible]	[illegible]	aatgtccgactatgggtg		
cis-regulatory elements of ERPIN	Rho_independent_terminator	281 ~ 339	59	tttgtgatatcagcattgcttgctctttatttgagcgggcaatgcttttt ttattctca	🔍	📄
	Rho_independent_terminator	2243 ~ 2284	42	acagcagtctaactccgccgcggcggagtttttttttgcatat	🔍	
cis-regulatory elements of Rfam	Rfam RF00167 (Purine family)	168 ~ 267	100	ttacaatataataggaacactcatataatcgcgtggatatggcacgcaag tttctaccgggcaccgtaaatgtccgactatgggtgagcaatggaaccgc	🔍	📄
long stems						
functional RNA sequences	FR379519/Purine_riboswitch	168 ~ 268	101	ttacaatataataggaacactcatataatcgcgtggatatggcacgcaag tttctaccgggcaccgtaaatgtccgactatgggtgagcaatggaaccgc a	🔍	📄
	FR290327/Purine_riboswitch	185 ~ 249	65	cactcatataatcgcgtggatatggcacgcaagtttctaccgggcaccgt aaatgtccgactatg	🔍	
	FR265322/Purine_riboswitch	184 ~ 226	43	acactcatataatcgcgtggatatggcacgcaagtttctaccg	🔍	
	FR257076/Purine_riboswitch	184 ~ 226	43	acactcatataatcgcgtggatatggcacgcaagtttctaccg	🔍	
	FR127065/Purine_riboswitch	188 ~ 227	40	tcatataatcgcgtggatatggcacgcaagtttctaccgg	🔍	
	FR104937/Purine_riboswitch	187 ~ 225	39	ctcatataatcgcgtggatatggcacgcaagtttctacc	🔍	
	FR180378/Purine_riboswitch	187 ~ 225	39	ctcatataatcgcgtggatatggcacgcaagtttctacc	🔍	
	FR130497/Purine_riboswitch	185 ~ 225	41	cactcatataatcgcgtggatatggcacgcaagtttctacc	🔍	
ncRNA hybridization regions						
microRNA target sites						

▶ File Download

Tab-Delimited File   XML File

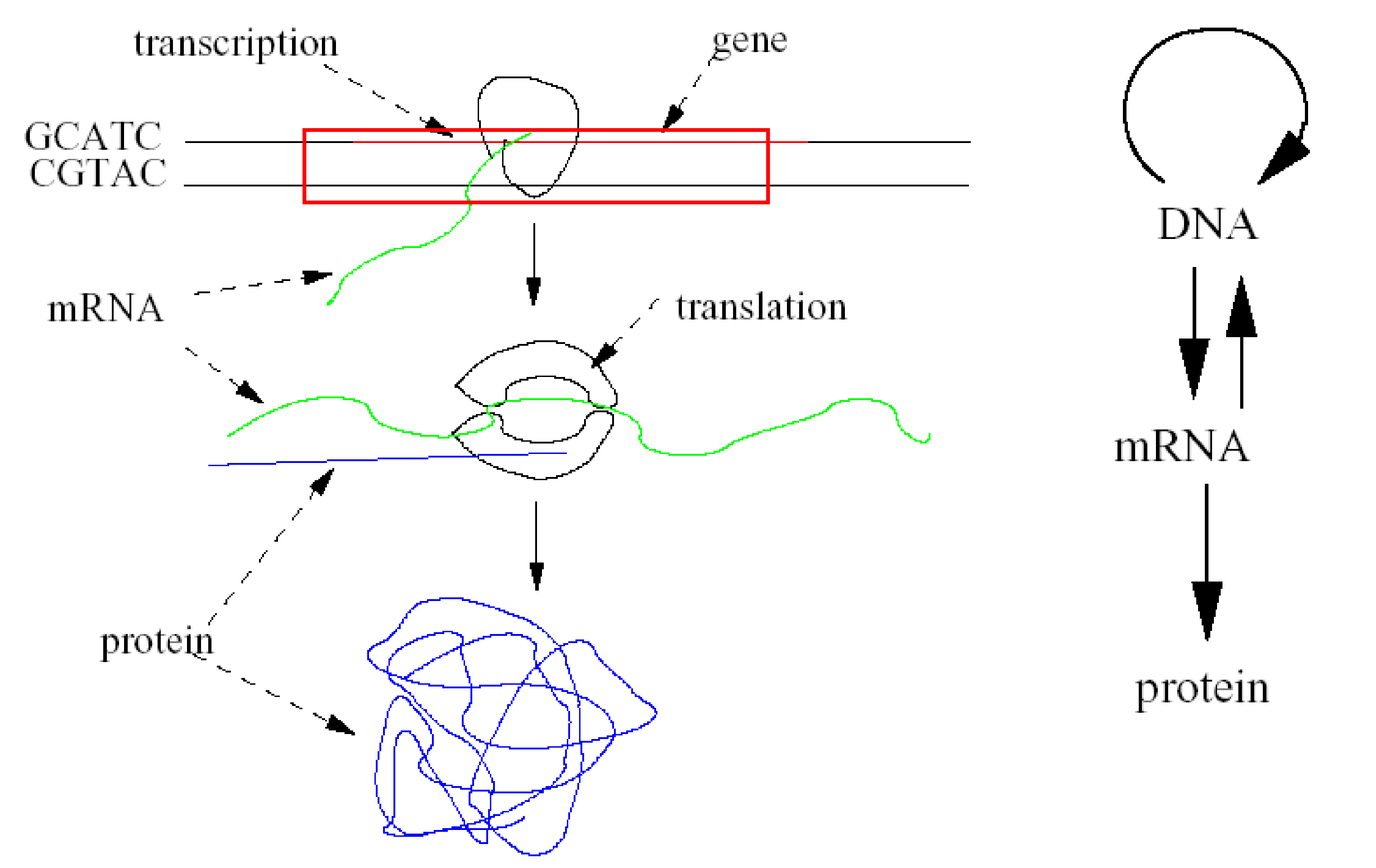
Finding RNA Genes

Finding RNA genes

- Most gene prediction methods only work well for protein coding genes

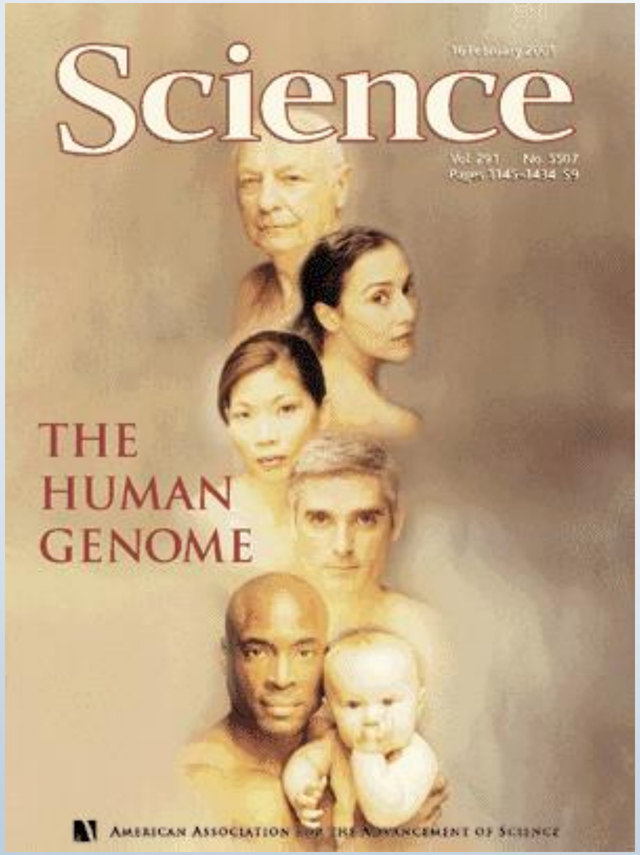
- In these case, one could use the same strategy as DNA gene finding approaches

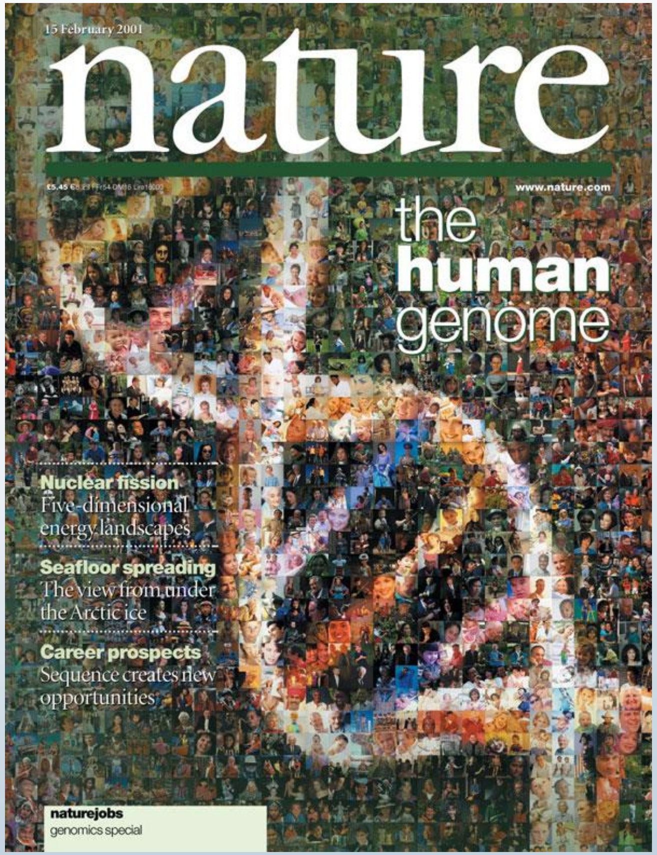
Central dogma



Bioinformatics: Life Sciences, Robert Lessick

Human genome





How many genes do we have?

- Only about 30,000 to 40,000 protein-coding genes in the human genome

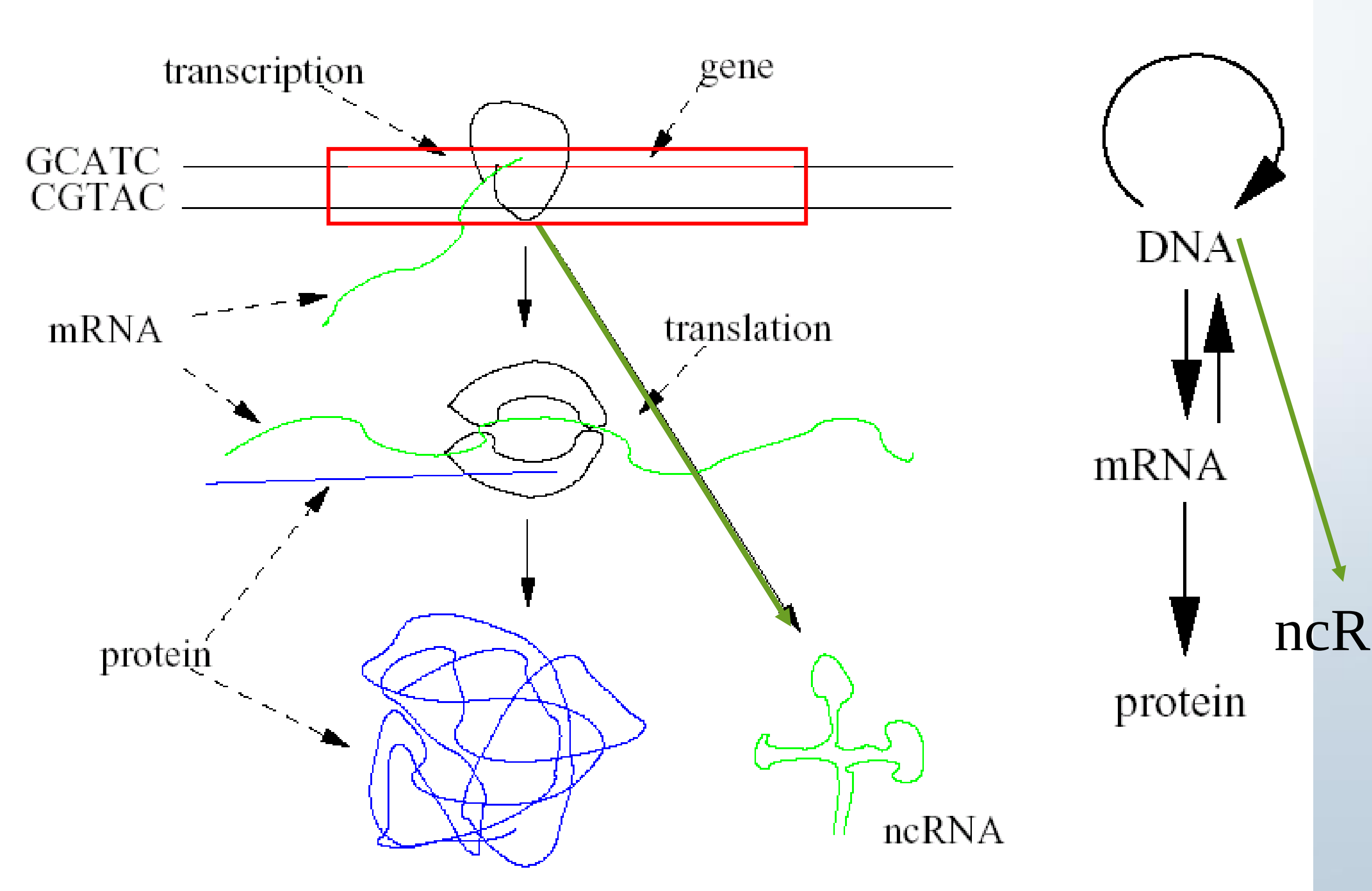
Lander et al. Nature (2001), Venter et al. Science (2001)

- Total protein coding gene length is only about 1.5 percent of the human genome. (3×10^9 bases)

What did we miss out?

- Non-coding RNA genes are undetected because they do not encode proteins
- Modern RNA world hypothesis:
  - There are many unknown but functional ncRNAs *[Eddy Nature Reviews (2001)]*
  - Many ncRNAs may play important role in the unexplained phenomenon *[Storz Science (2002)]*

Central dogma



ncRNA

Bioinformatics: Life Sciences, Robert Lessick

Non-coding RNA

- Non-coding RNA (ncRNA)
  - RNA acting as functional molecule
  - Not translated into protein

- Non-coding RNA gene
  - The region of DNA coding ncRNA

Non-coding RNA

Question:

If there are many ncRNAs, what are they doing?

Question:

Biologically, why do we need functional ncRNAs in addition to protein?

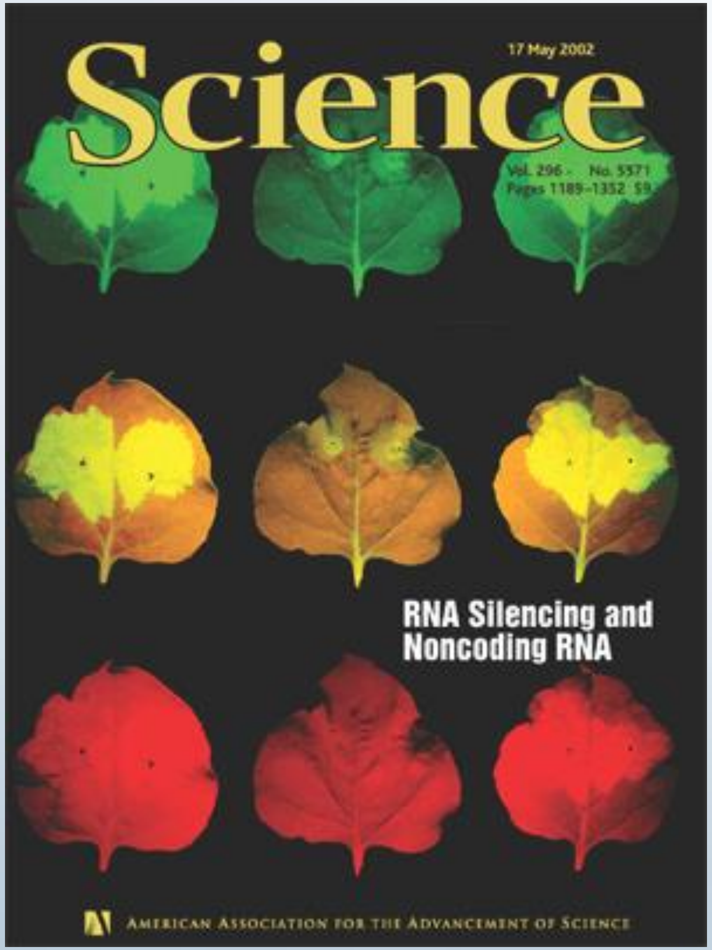
Why do we need ncRNAs?

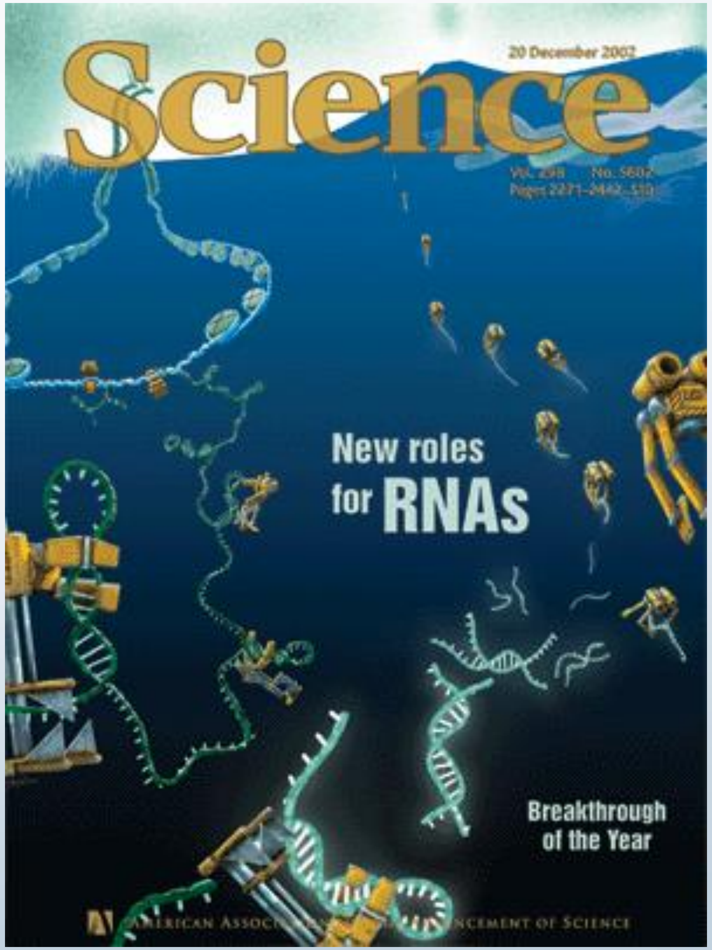
- ncRNAs involve sequence specific recognition of other nucleic acids (e.g. mRNAs, DNAs)
- ncRNA is an ideal material for this role
  - DNA is big and packaged and can do this job
- *Base complementary* allows ncRNA to be sequence specific
- For example:
  - small interfering RNAs (siRNA) is used to protect our genome
  - It recognizes invading foreign RNAs/DNAs based on the sequence specificity
  - And helps to degrade the foreign RNAs
- Nobel Prize 2006 : Fire and Melo for discovering **RNA interference**
- Some scientists believe that miRNAs are the future of medicine; find them at these sites:
  - sirna.cgb.ki.se
  - microrna.sanger.ac.uk

What do they do?

- RNA-protein machine:
  - Transfer RNA (tRNA)
  - Ribosomal RNA (rRNA)
  - RNAs (snRNAs) in spliceosome
- Catalytic RNAs (ribozymes): catalyzing some functions
- Micro RNAs (miRNAs): regulatory roles
- Small interfering RNAs (siRNAs): RNA silencing
  - The genome’s immune system. *[Plasterk, Science (2002)]*
  - The breakthrough of the year by Science magazine in 2002
- Riboswitch RNAs: a genetic control element, to control gene expression
  - found in prokaryotes and plants. eukaryotes?
- Small nucleolar RNAs (snoRNAs): help the modification of rRNAs
- tmRNA (tRNA like mRNA): direct abnormal protein degradation

Non-coding RNA – new era


17 May 2002
Science
Vol. 296 · No. 5571
Pages 1189–1352 $9
RNA Silencing and Noncoding RNA
American Association for the Advancement of Science



How can we find such ncRNA genes in the genome?

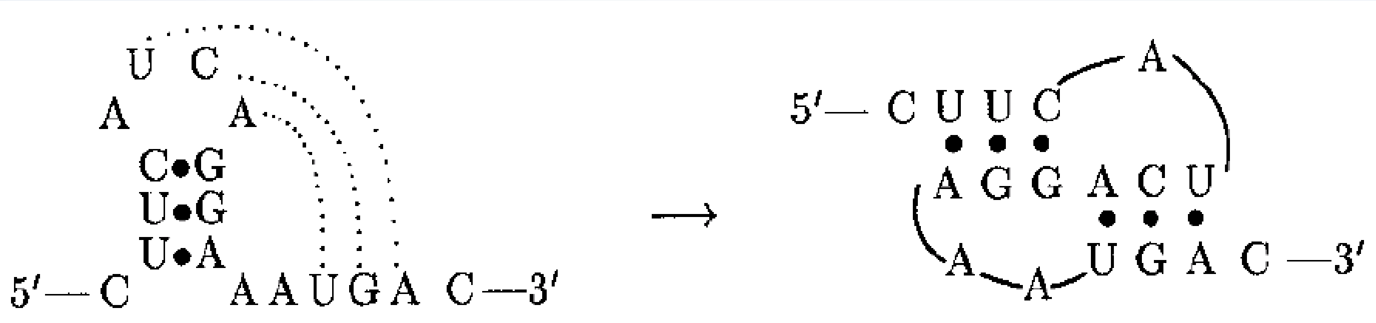
RNA secondary structure

- ncRNA is not a random sequence.
- Most RNAs fold into particular base-paired secondary structure

- Canonical basepairs:
  - Watson-Crick basepairs:
    - G - C
    - A - U
  - Wobble basepair:
    - G – U

Pseudoknots

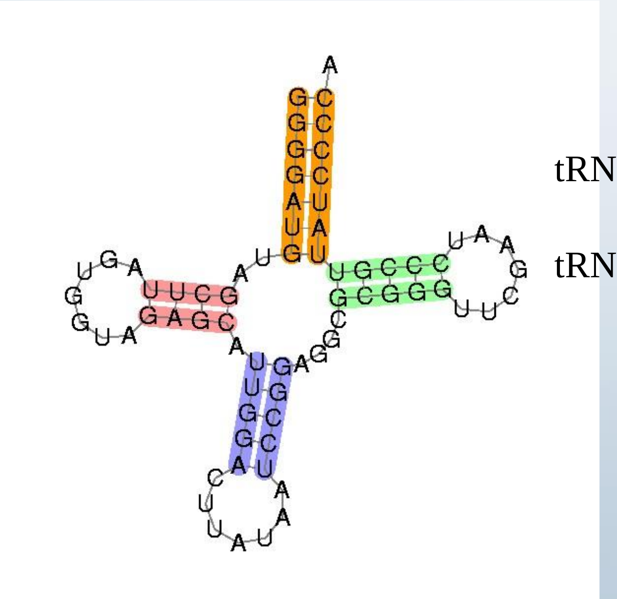
- Pseudoknots are important for certain ncRNAs
- Violate the non-crossing assumption.
- Pseudoknots make most problems harder
- We assume there are no pseudoknots otherwise noted



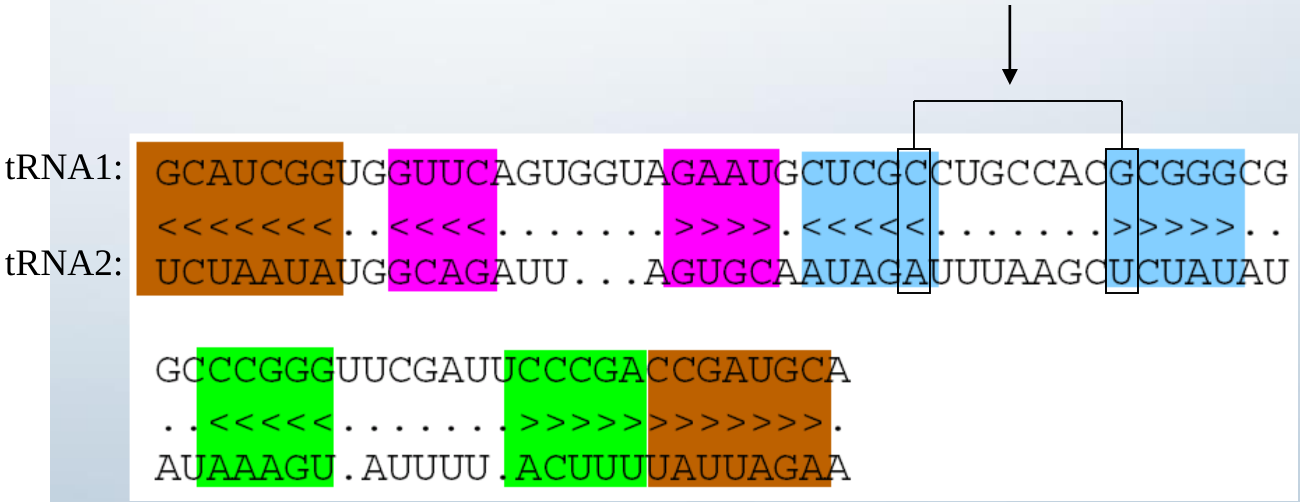
Bioinformatics: Life Sciences, Robert Lessick

ncRNA evolution
is constrained by it secondary structure

- Drastic sequence changes can be tolerated
- *Compensatory mutations* are very common
  - One basepair mutates into another basepair
  - Doesn’t change its secondary structure



Compensatory mutation



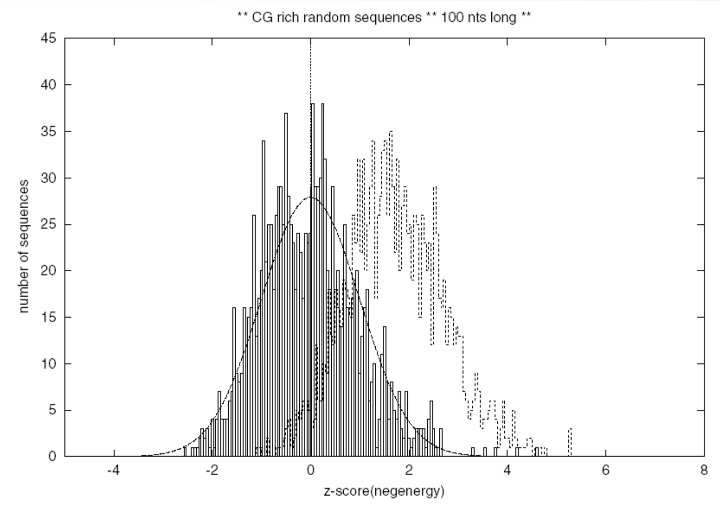
```
tRNA1: GCAUCGGUGGUUCAGUGGUAGAAUGCUCGCCUGCCACGCGGGCG
       <<<<<<<..<<<<.......>>>>.<<<<<.......>>>>>..
tRNA2: UCUAAUAUGGCAGAUU...AGUGCAAUAGAUUUAAGCUCUAUAU

       GCCCGGGUUCGAUUCCCGACCGAUGCA
       ..<<<<<.......>>>>>>>>>>>>.
       AUAAAGU.AUUUU.ACUUUUAUUAGAA
```

Non-coding RNA gene finding

- *De novo* ncRNA gene finding
  - Folding energy
  - Number of sub-optimal RNA structures
- Homology ncRNA gene searching
  - Sequence-based
  - Structure-based
  - Sequence and structure-based

Non-coding RNA gene finding Problems

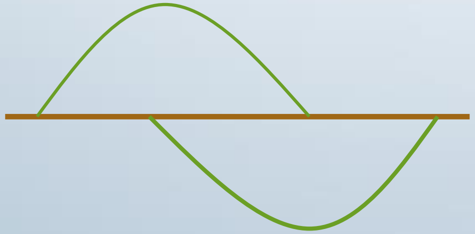
- *de novo* prediction:
  - Find stable secondary structure from genome *[Shapiro et al. (1990)]*
    - **Problem:**
      - The stability of ncRNA secondary structure is not sufficiently different from the predicted stability of a random sequence [Rivas and Eddy (2000)]
  - Look transcript signals *[Wassarman et al. (2001), Argaman et al. (2000)]*
    - **Problem :**
      - ncRNA transcript signals are not strong
      - protein coding gene signals (open reading frame, promoter)

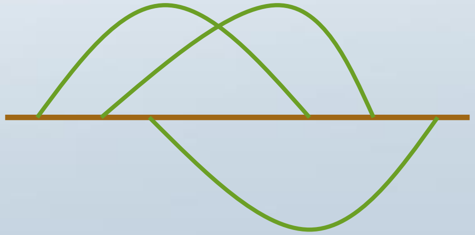


Bioinformatics: Life Sciences, Robert Lessick

Non-coding RNA gene finding Problems (prediction with pseudoknots)

- Base pair maximization allowing crossing pairs can be solved in polynomial time.
- Ieong et al. (2003) proved that base pairing maximization problem allowing crossing pairs in a *planar* secondary structure is NP-hard.



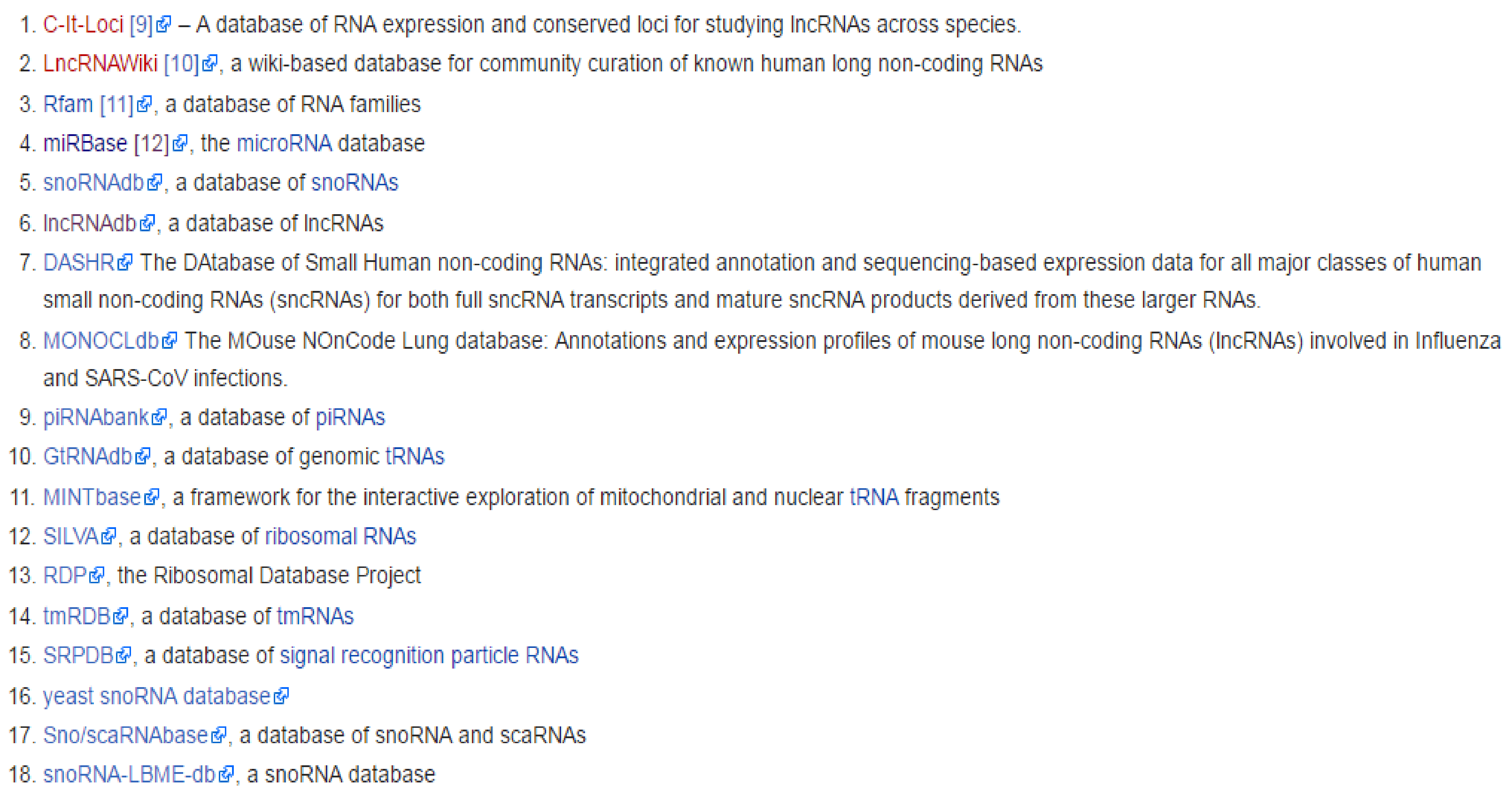


Prediction with pseudoknots – some references

- Prediction allowing generalized pseudoknots with energy functions depending on adjacent basepairs is NP-hard.
  - Akutsu (2000) (longest common subsequence for multiple sequences (LCS)).
  - Lyngsø and Pedersen (2000) (3SAT).
  - similar to Zuker-Sankoff minimum energy model.
- Pseudoknots in structure-known RNAs.
  - Biologists are not interested in the approximation solutions.
  - Most pseudoknots are planar.
  - Not too many variations.
- Rivas and Eddy (1999) presented a $O(n^6)$ solution allowing most types of pseudoknots in known ncRNAs.

Some RNA Databases

RNA Databases



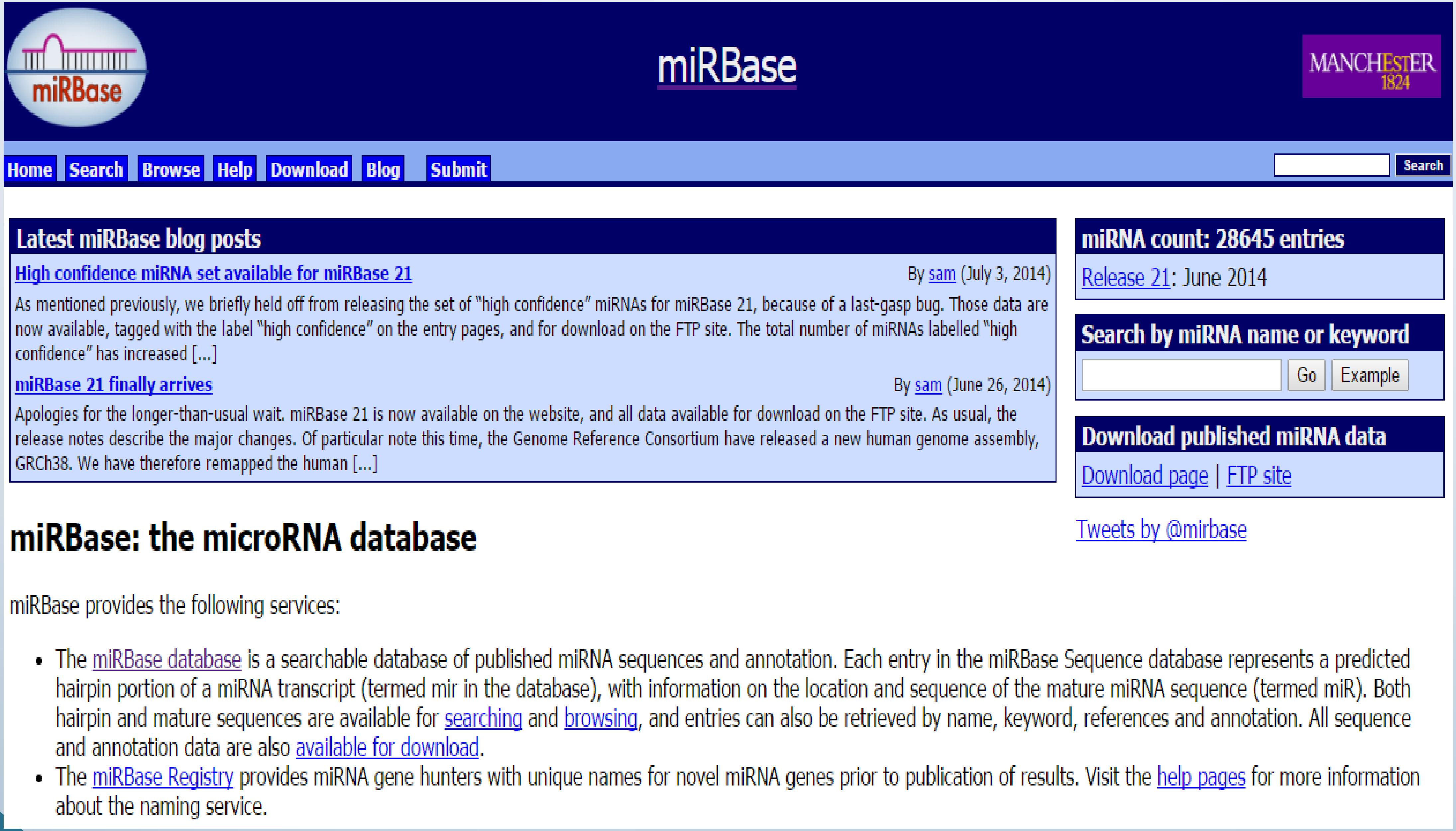
1. C-It-Loci [9] – A database of RNA expression and conserved loci for studying lncRNAs across species.
2. LncRNAWiki [10], a wiki-based database for community curation of known human long non-coding RNAs
3. Rfam [11], a database of RNA families
4. miRBase [12], the microRNA database
5. snoRNAdb, a database of snoRNAs
6. lncRNAdb, a database of lncRNAs
7. DASHR The DAtabase of Small Human non-coding RNAs: integrated annotation and sequencing-based expression data for all major classes of human small non-coding RNAs (sncRNAs) for both full sncRNA transcripts and mature sncRNA products derived from these larger RNAs.
8. MONOCLdb The MOuse NOnCode Lung database: Annotations and expression profiles of mouse long non-coding RNAs (lncRNAs) involved in Influenza and SARS-CoV infections.
9. piRNAbank, a database of piRNAs
10. GtRNAdb, a database of genomic tRNAs
11. MINTbase, a framework for the interactive exploration of mitochondrial and nuclear tRNA fragments
12. SILVA, a database of ribosomal RNAs
13. RDP, the Ribosomal Database Project
14. tmRDB, a database of tmRNAs
15. SRPDB, a database of signal recognition particle RNAs
16. yeast snoRNA database
17. Sno/scaRNAbase, a database of snoRNA and scaRNAs
18. snoRNA-LBME-db, a snoRNA database

https://www.wikipedia.org/

Rfam & Infernal

- Rfam 9.1 contains 1379 families (December 2008)
- **Rfam 10.0 contains 1446 families (January 2010)**
- Rfam is a collection of **multiple sequence alignments and covariance models** covering many common non-coding RNA families
- Infernal searches Rfam covariance models (CMs) in genomes or other DNA sequence databases for homologs to known structural RNA families

http://rfam.xfam.org/



Latest miRBase blog posts

High confidence miRNA set available for miRBase 21 By sam (July 3, 2014)

As mentioned previously, we briefly held off from releasing the set of "high confidence" miRNAs for miRBase 21, because of a last-gasp bug. Those data are now available, tagged with the label "high confidence" on the entry pages, and for download on the FTP site. The total number of miRNAs labelled "high confidence" has increased […]

miRBase 21 finally arrives By sam (June 26, 2014)

Apologies for the longer-than-usual wait. miRBase 21 is now available on the website, and all data available for download on the FTP site. As usual, the release notes describe the major changes. Of particular note this time, the Genome Reference Consortium have released a new human genome assembly, GRCh38. We have therefore remapped the human […]

miRNA count: 28645 entries

Release 21: June 2014

Search by miRNA name or keyword

Download published miRNA data

Download page | FTP site

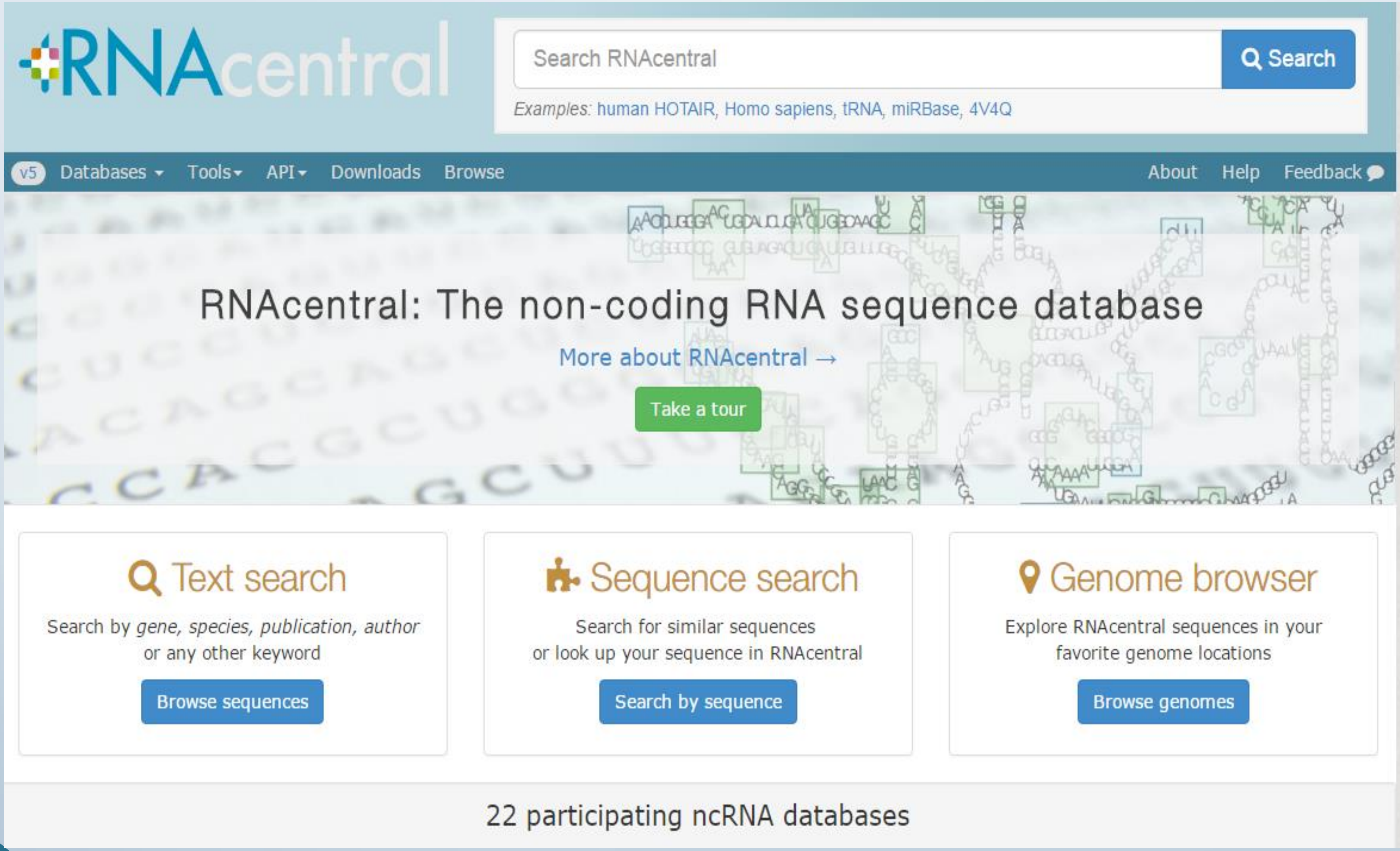
Tweets by @mirbase

miRBase: the microRNA database

miRBase provides the following services:

- The miRBase database is a searchable database of published miRNA sequences and annotation. Each entry in the miRBase Sequence database represents a predicted hairpin portion of a miRNA transcript (termed mir in the database), with information on the location and sequence of the mature miRNA sequence (termed miR). Both hairpin and mature sequences are available for searching and browsing, and entries can also be retrieved by name, keyword, references and annotation. All sequence and annotation data are also available for download.
- The miRBase Registry provides miRNA gene hunters with unique names for novel miRNA genes prior to publication of results. Visit the help pages for more information about the naming service.

http://www.mirbase.org/



RNAcentral: The non-coding RNA sequence database

Text search

Search by *gene, species, publication, author* or any other keyword

Sequence search

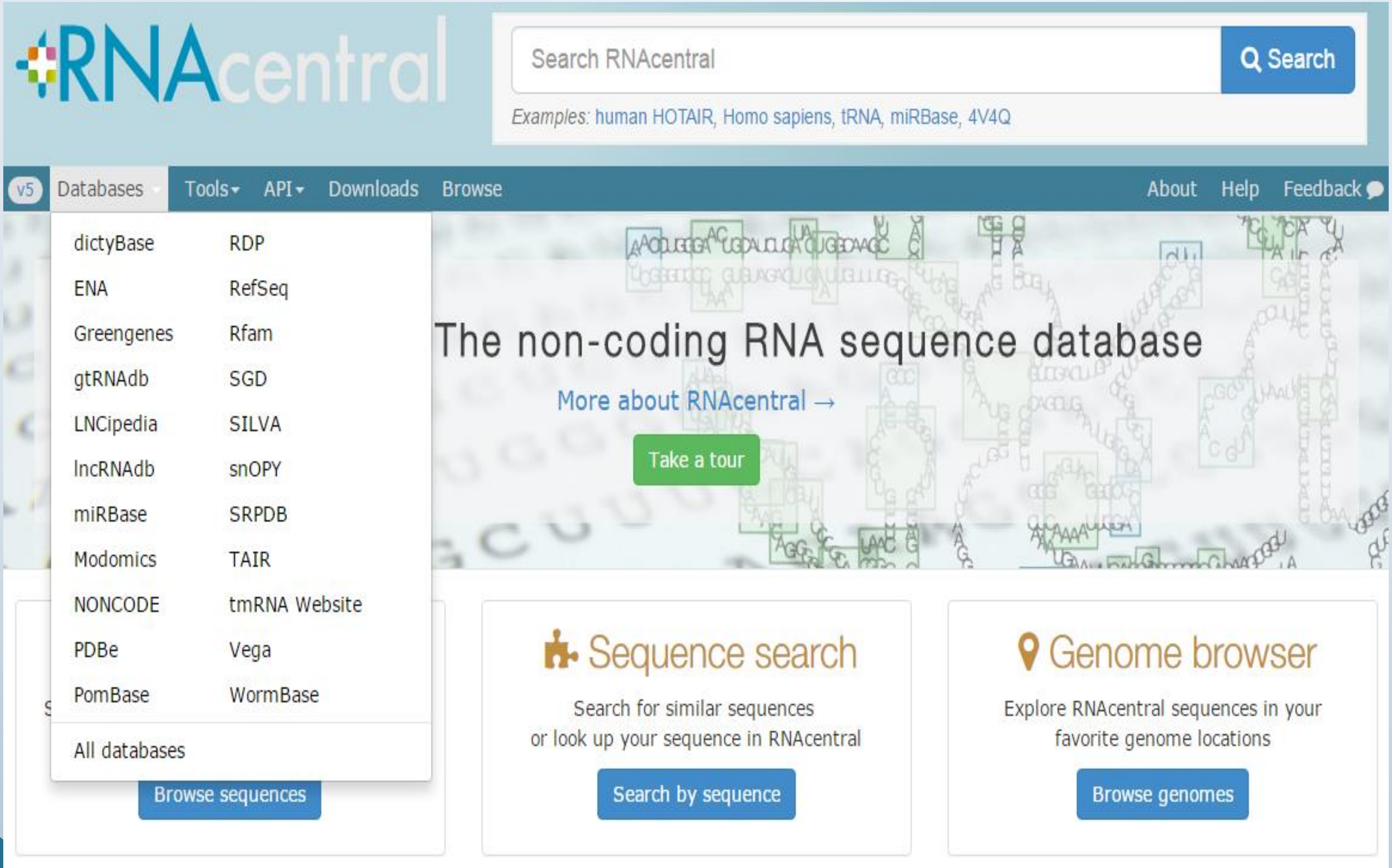
Search for similar sequences or look up your sequence in RNAcentral

Genome browser

Explore RNAcentral sequences in your favorite genome locations

22 participating ncRNA databases

http://rnacentral.org/



The non-coding RNA sequence database

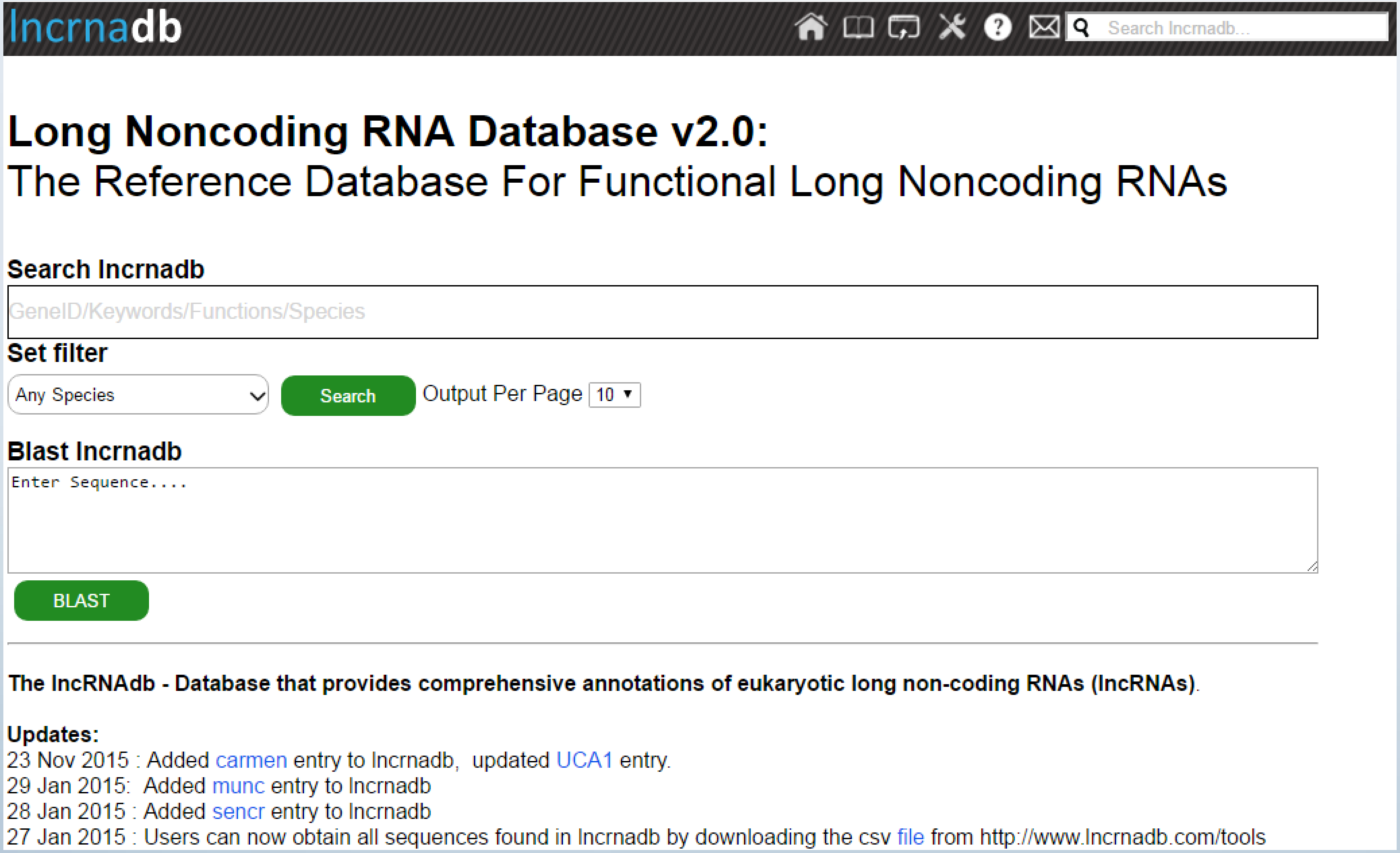
Sequence search

Search for similar sequences or look up your sequence in RNAcentral

Genome browser

Explore RNAcentral sequences in your favorite genome locations

http://rnacentral.org/



Long Noncoding RNA Database v2.0:
The Reference Database For Functional Long Noncoding RNAs

Search lncrnadb

GeneID/Keywords/Functions/Species

Set filter

Any Species | Search | Output Per Page 10

Blast lncrnadb

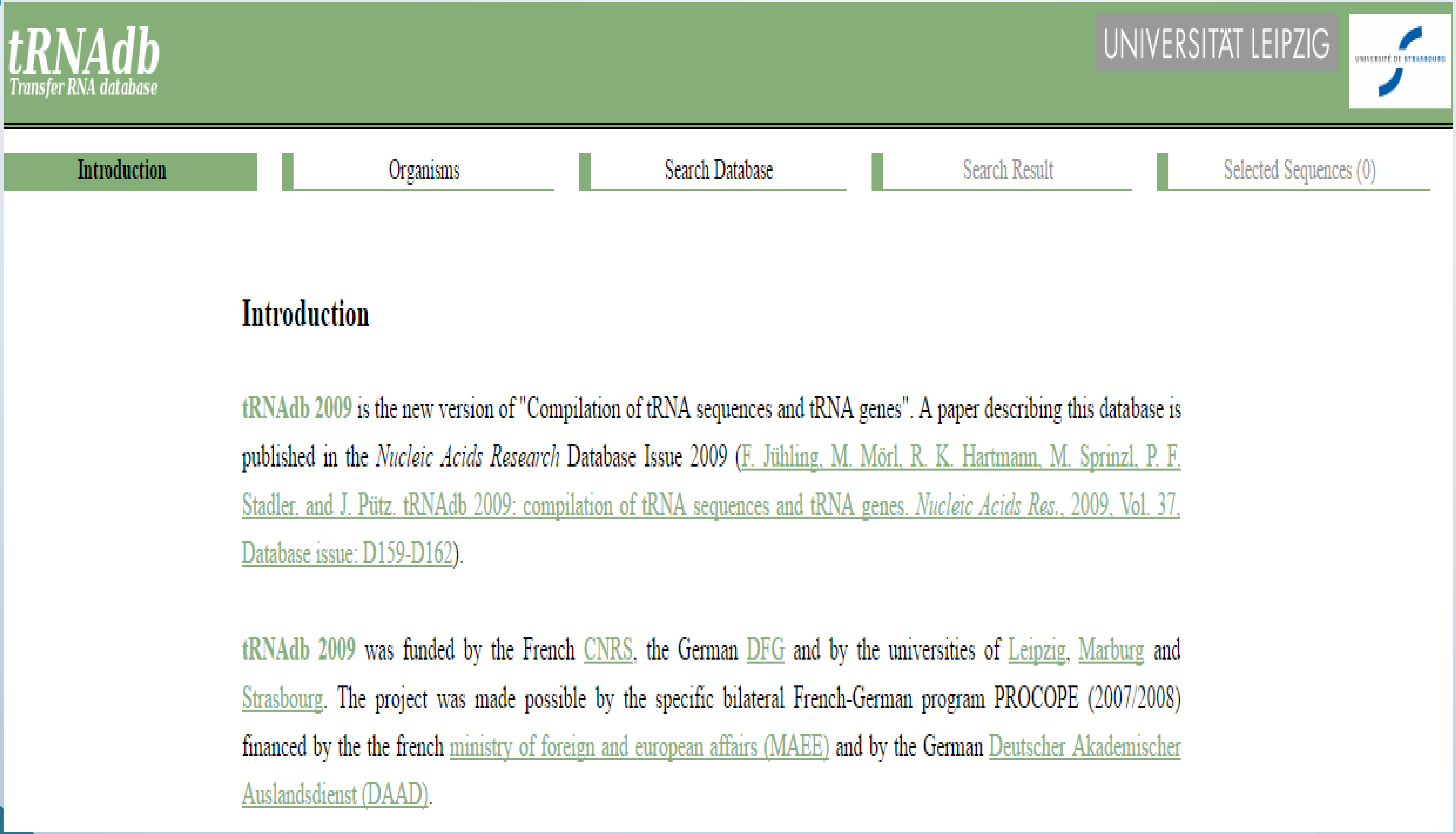
```
Enter Sequence....
```

BLAST

The lncRNAdb - Database that provides comprehensive annotations of eukaryotic long non-coding RNAs (lncRNAs).

Updates:
23 Nov 2015 : Added carmen entry to lncrnadb, updated UCA1 entry.
29 Jan 2015: Added munc entry to lncrnadb
28 Jan 2015 : Added sencr entry to lncrnadb
27 Jan 2015 : Users can now obtain all sequences found in lncrnadb by downloading the csv file from http://www.lncrnadb.com/tools

http://www.lncrnadb.org/



Introduction

tRNAdb 2009 is the new version of "Compilation of tRNA sequences and tRNA genes". A paper describing this database is published in the *Nucleic Acids Research* Database Issue 2009 (F. Jühling, M. Mörl, R. K. Hartmann, M. Sprinzl, P. F. Stadler, and J. Pütz. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res.*, 2009, Vol. 37, Database issue: D159-D162).

tRNAdb 2009 was funded by the French CNRS, the German DFG and by the universities of Leipzig, Marburg and Strasbourg. The project was made possible by the specific bilateral French-German program PROCOPE (2007/2008) financed by the the french ministry of foreign and european affairs (MAEE) and by the German Deutscher Akademischer Auslandsdienst (DAAD).

http://trna.bioinf.uni-leipzig.de/DataOutput/Welcome

References

- How Do RNA Folding Algorithms Work? Eddy. Nature Biotechnology, 22:1457-1458, 2004 (a short nice review)
- Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids. Durbin, Eddy, Krogh and Mitchison. 1998 Chapter 10, pages 260-297
- Secondary Structure Prediction for Aligned RNA Sequences. Hofacker et al. JMB, 319:1059-1066, 2002 (RNAalifold; covariance-like score calculation)
- Optimal Computer Folding of Large RNA Sequences Using Thermodynamics and Auxiliary Information. Zuker and Stiegler. NAR, 9(1):133-148, 1981 (Mfold)
- A computational pipeline for high throughput discovery of cis-regulatory noncoding RNAs in Bacteria, PLoS CB 3(7):e126
- Riboswitches in Eubacteria Sense the Second Messenger Cyclic Di-GMP, Science, 321:411 – 413, 2008
- Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline, Nucl. Acids Res. (2007) 35 (14): 4809-4819.
- CMfinder—a covariance model based RNA motif finding algorithm. Bioinformatics 2006;22:445-452