Detecting non-Coding RNA in Genomic Sequences

I. Overview of ncRNAs
II. What's specific about RNA detection ?
III. Looking for known RNAs
IV. Looking for unknown RNAs

Daniel Gautheret
INSERM ERM 206 & Université de la Méditerranée

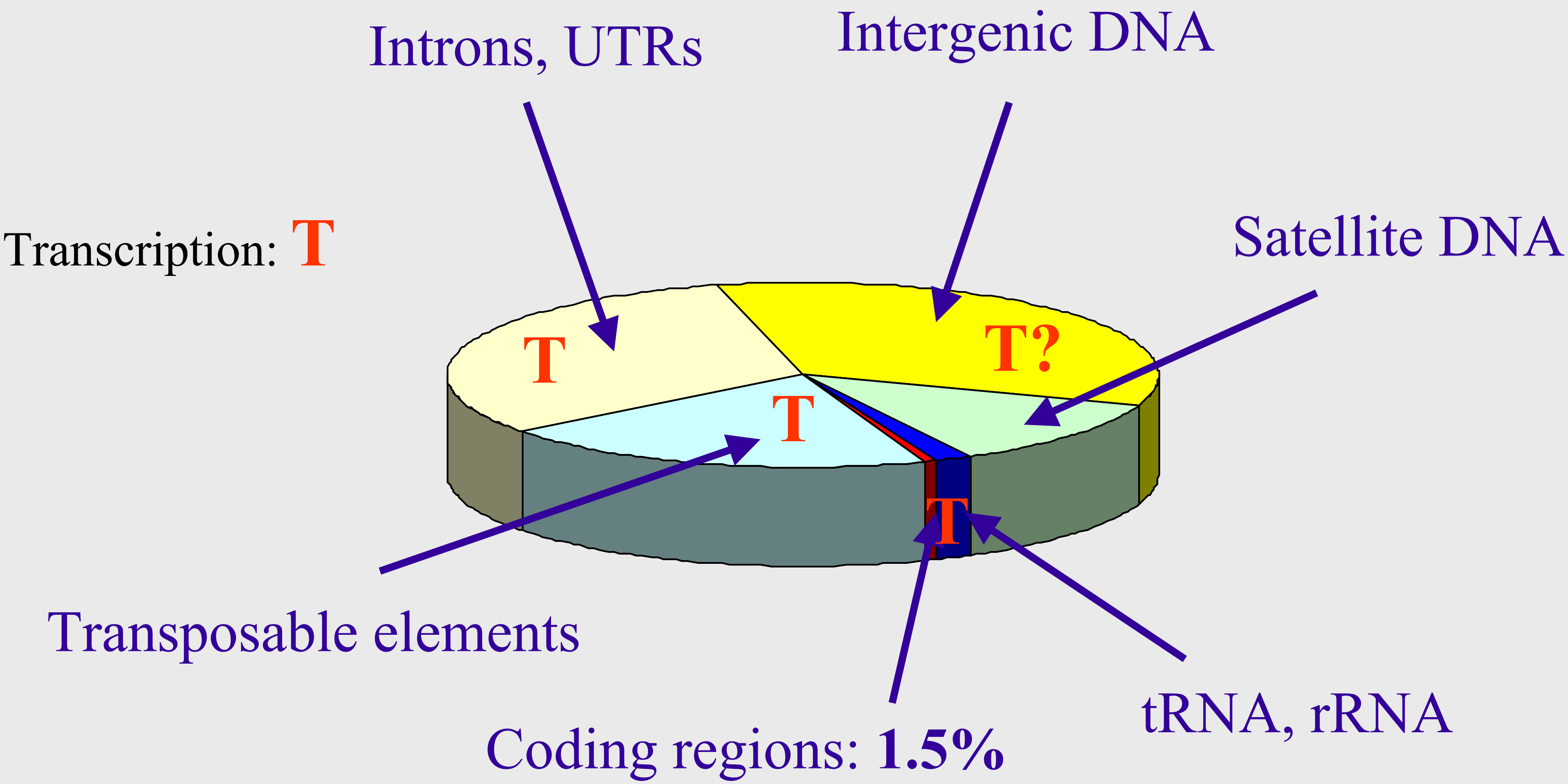
Non-coding or non-messenger RNA

- ★ All organisms
  - rRNA 5S/5.8S 15S/18S 23S/28S (5-300 copies)
  - RNAse P/MRP (1 copy)
  - tRNA (20 diff., 50-1000 copies)
- ★ Eukaryotes and Archaes
  - snoRNAs (H/ACA and C/D: 10-100 different)
- ★ Eukaryotes
  - miRNAs (100-200 diff.)
  - XIST, H19, IPW (vertebrates)
  - snRNAs (U1, U2, U4, U5, U6)
- ★ Procaryotes
  - rprA, csrB, oxyS,…
  - tmRNA

rRNA &
RNAse P:
catalytic!

Vertebrate Genomes: > 50% transcribed!

Vertebrate gene: 30kb (coding: 1,5kb)



Intrpremium, UTRs

How many other ncRNAs ?

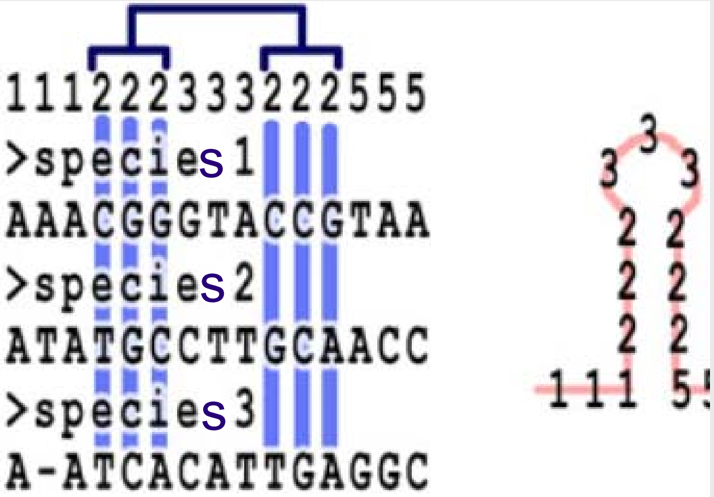
- ★ Rnomics*
  - Extract total RNA
  - Isolate small RNAs (> mRNA size)
  - Tag & Reverse transcribe
  - Clone & Sequence
- ★ Success
  - 80 ncRNAs identified in mouse
  - Lots of new snoRNAs
- ★ Limitations
  - Access to nuclear RNAs
  - Tissue/time specific expression may be frequent

* Huttenhofer et al., 2001

Bioinformatics

- Not sensitive to « rare » expression
- Highly succesful in identifying protein-coding genes
- Many complete genomes available
- Large computational toolbox
  - Statistics
  - Thermodynamics
  - Phylogeny

What’s Special About ncRNA Detection?



- ★ No ORF
- ★ No Markov model / sequence statistics
- ★ ncRNA is defined both by primary and secondary structure
- ★ « Substitution matrices » for nucleic acids are terrible compared to aminoacids counterparts

III. Looking for known ncRNAs

- ★ How can you detect ncRNA genes from known families?

Custom RNA search programs

Based on a variety of algorithms: seek regular expressions or base pairs, weight matrices, SCFGs, etc.

- tRNA
  - trnascan (Fichant & Burks 91)
  - trnascan-SE (Eddy 94)
- C/D Box snoRNAs (Lowe & Eddy 99)
  - One type of snoRNA
- miRNA
  - Not an automated procedure

Descriptor-based programs

- RnaMot / Rnamotif (Gautheret 91, Macke '02)
- Palingol (Viari 96)
- Patscan (Overbeek '00)
- PatSearch (Pesole '01)

```
h1 s1 h1 s2 h2 s3 h2

h1 5:5 1
h2 5:5 NNNNR:YNNNN
s1 7:7 NUNNNNN
s2 4:40
s3 7:7 UUCNNNN
```

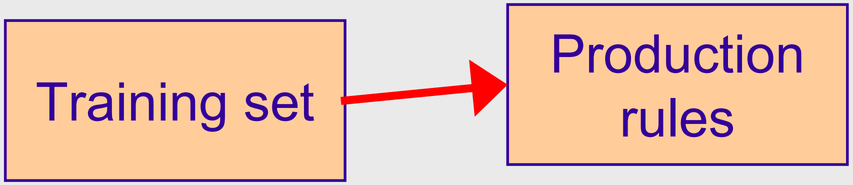
RnaMot descriptor for anticodon+TYC domain of tRNA

Descriptor-based programs

PROS	CONS
Draft descriptors can be **quickly sketched** and tested	Requires a **good prior knowledge** of secondary structure and sequence constraints
No alignment is required, although it is very helpful to have one	Requires **basic computer skills** to translate biological constraints into the descriptor language
Biologists decide what features are important or not (see also CONS!)	Biologists have the responsibility of correctly **weighting each important feature**

Probabilistic ncRNA search programs

Stochastic Context Free Grammars (first adaptation of CFG to RNA: Searls 94; SCFG: Eddy & Durbin 94)



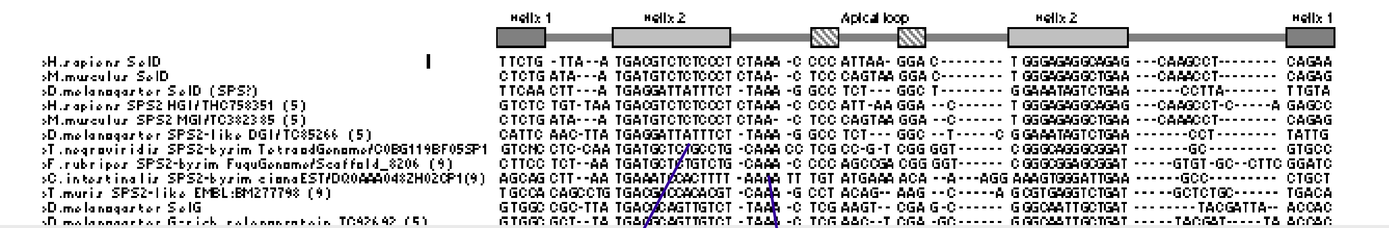
describe how to generate any structure in the training set

- Time cost = O(N^4) for sequence of length N
- Not « practical » for large alignments or genome-wide searches
- Pseudoknots not allowed

ERPIN: Profile-based search

Gautheret & Lambert, JMB, 2001, 313, p. 1005.

Training set



```
>H.sapiens SelD                                        TTCTG -TTA--A TGACGTCTCTCCCT CTAAA -C CCC ATTAA- GGA C--------- T GGGAGAGGCAGAG ---CAAGCCT--------- CAGAA
>M.musculus SelD                                       CTCTG ATA---A TGATGTCTCTCCCT CTAA- -C TCC CAGTAA GGA C--------- T GGGAGAGGCTGAA ---CAAACCT--------- CAGAG
>D.melanogaster SelD (SPS?)                            TTCAA CTT---A TGAGGATTATTTCT -TAAA -G GCC TCT--- GGC T--------- G GAAATAGTCTGAA ------CCTTA-------- TTGTA
>H.sapiens SPS2 HGI/THC758351 (5)                      GTCTC TGT-TAA TGACGTCTCTCCCT CTAAA -C CCC ATT-AA GGA --C------- T GGGAGAGGCAGAG ---CAAGCCT-C-----A GAGCC
>M.musculus SPS2 MGI/TC382385 (5)                      CTCTG ATA---A TGATGTCTCTCCCT CTAA- -C TCC CAGTAA GGA --C------- T GGGAGAGGCTGAA ---CAAACCT--------- CAGAG
>D.melanogaster SPS2-like DGI/TC85266 (5)              CATTC AAC-TTA TGAGGATTATTTCT -TAAA -G GCC TCT--- GGC --T-----C G GAAATAGTCTGAA -------CCT--------- TATTG
>T.negroviridis SPS2-bysim TetraodGenome/C0BG119BF05SP1 GTCNC CTC-CAA TGATGCTCTGCCTG -CAAA CC TCG CC-G-T CGG GGT------ C GGGCAGGGCGGAT --------GC---------- GTGCC
>F.rubripes SPS2-bysim FuguGenome/Scaffold_8206 (9)    CTTCC TCT--AA TGATGCTCTGTCTG -CAAA -C CCC AGCCGA CGG GGT------ C GGGCGGAGCGGAT -----GTGT-GC--CTTC GGATC
>C.intestinalis SPS2-bysim cionaEST/DQ0AAA048ZH02CP1(9) AGCAG CTT--AA TGAAATCCACTTTT -AAAA TT TGT ATGAAA ACA --A---AGG AAAGTGGGATTGAA ------GCC---------- CTGCT
>T.muris SPS2-like EMBL:BM277798 (9)                    TGCCA CAGCCTG TGACGACCACACGT -CAAA -G CCT ACAG-- AAG --C-----A G CGTGAGGTCTGAT -----GCTCTGC------ TGACA
>D.melanogaster SelG                                   GTGGC CGC-TTA TGACGCAGTTGTCT -TAAA -C TCG AAGT-- CGA G-C------ G GGCAATTGCTGAT ---------TACGATTA-- ACCAC
>D.melanogaster G-rich selenoprotein TC42692 (5)       GTGGC GCT--TA TGACGCAGTTGTCT -TAAA -C TCG AAC--T CGA -GC------ G GGCAATTGCTGAT -----TACGAT-----TA ACCAC
```

Helix profile (16xN)

A:A			
G:A			
C:A			
U:A			
A:G			
G:G			
C:G			
U:G			
…			
U:U			

A			
G			
C			
U			
-			

Single-strand profile (5xN)

Search algorithm combines **dynamic programming** for single strands and **profile search** for helices

Profile-based search

PROS	CONS
All constraints in the training set are **efficiently exploited**, resulting in highly specific detections	Alignment and secondary structure constraints must be accurate
After alignments and secondary structures are created, **no further programming is needed**	**Helices of variable length need to be reduced** to their shortest consensus
Scoring system is defined automatically	Program will not depart from initial alignment in terms of motif size
***E*-values** are provided for each hit	**Users still have to decide on search order** and masked elements

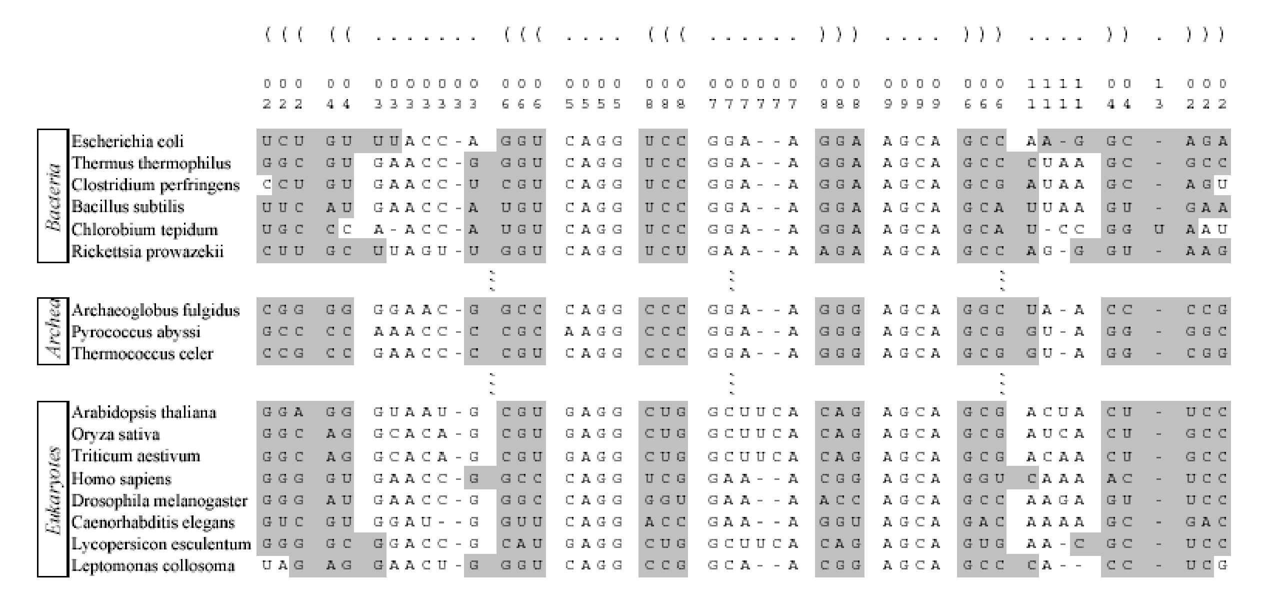
Running a successful ncRNA search

Example: the Signal Recognition Particle (SRP) RNA

- ★ 172 sequences available
- ★ All 3 kingdoms
- ★ Signature: 50-nt domain IV

Organize ncRNA information

- ★Alignment is a must
- ★Should be structure-based
- ★ClustalW OK only as a first attempt
- ★RNAalifold (Vienna package) can identify covarying basepairs



		(((	((	. . . .	(((	. . . .	(((	. . . . . .	)))	. . . .	)))	. . . .	))	.	)))
		0 0 0	0 0	0 0 0 0 0 0 0	0 0 0	0 0 0 0	0 0 0	0 0 0 0 0 0	0 0 0	0 0 0 0	0 0 0	1 1 1 1	0 0	1	0 0 0
		2 2 2	4 4	3 3 3 3 3 3 3	6 6 6	5 5 5 5	8 8 8	7 7 7 7 7 7	8 8 8	9 9 9 9	6 6 6	1 1 1 1	4 4	3	2 2 2
Bacteria	Escherichia coli	U C U	G U	U U A C C - A	G G U	C A G G	U C C	G G A - - A	G G A	A G C A	G C C	A A - G	G C	-	A G A
	Thermus thermophilus	G G C	G U	G A A C C - G	G G U	C A G G	U C C	G G A - - A	G G A	A G C A	G C C	C U A A	G C	-	G C C
	Clostridium perfringens	C C U	G U	G A A C C - U	C G U	C A G G	U C C	G G A - - A	G G A	A G C A	G C G	A U A A	G C	-	A G U
	Bacillus subtilis	U U C	A U	G A A C C - A	U G U	C A G G	U C C	G G A - - A	G G A	A G C A	G C A	U U A A	G U	-	G A A
	Chlorobium tepidum	U G C	C C	A - A C C - A	U G U	C A G G	U C C	G G A - - A	G G A	A G C A	G C A	U - C C	G G	U	A A U
	Rickettsia prowazekii	C U U	G C	U U A G U - U	G G U	C A G G	U C U	G A A - - A	A G A	A G C A	G C C	A G - G	G U	-	A A G
				⋮				⋮				⋮			
Archea	Archaeoglobus fulgidus	C G G	G G	G G A A C - G	G C C	C A G G	C C C	G G A - - A	G G G	A G C A	G G C	U A - A	C C	-	C C G
	Pyrococcus abyssi	G C C	C C	A A A C C - C	C G C	A A G G	C C C	G G A - - A	G G G	A G C A	G C G	G U - A	G G	-	G G C
	Thermococcus celer	C C G	C C	G A A C C - C	C G U	C A G G	C C C	G G A - - A	G G G	A G C A	G C G	G U - A	G G	-	C G G
				⋮				⋮				⋮			
Eukaryotes	Arabidopsis thaliana	G G A	G G	G U A A U - G	C G U	G A G G	C U G	G C U U C A	C A G	A G C A	G C G	A C U A	C U	-	U C C
	Oryza sativa	G G C	A G	G C A C A - G	C G U	G A G G	C U G	G C U U C A	C A G	A G C A	G C G	A U C A	C U	-	G C C
	Triticum aestivum	G G C	A G	G C A C A - G	C G U	G A G G	C U G	G C U U C A	C A G	A G C A	G C G	A C A A	C U	-	G C C
	Homo sapiens	G G G	G U	G A A C C - G	G C C	C A G G	U C G	G A A - - A	C G G	A G C A	G G U	C A A A	A C	-	U C C
	Drosophila melanogaster	G G G	A U	G A A C C - G	G G C	C A G G	G G U	G A A - - A	A C C	A G C A	G C C	A A G A	G U	-	U C C
	Caenorhabditis elegans	G U C	G U	G G A U - - G	G U U	C A G G	A C C	G A A - - A	G G U	A G C A	G A C	A A A A	G C	-	G A C
	Lycopersicon esculentum	G G G	G C	G G A C C - G	C A U	G A G G	C U G	G C U U C A	C A G	A G C A	G U G	A A - C	G C	-	U C C
	Leptomonas collosoma	U A G	A G	G A A C U - G	G G U	C A G G	C C G	G C A - - A	C G G	A G C A	G C C	C A - -	C C	-	U C G

Secondary Structure annotation

Will help identify sequence/ structure constraints: helix sizes, conserved bases, etc.

Want to publish your finds? Prepare Control Procedures

$$\text{Sensitivity: SN} = \frac{TP}{TP+FN}$$

Total « true » objects

$$\text{Specificity : SP} = \frac{TP}{TP+FP}$$

Total predictions

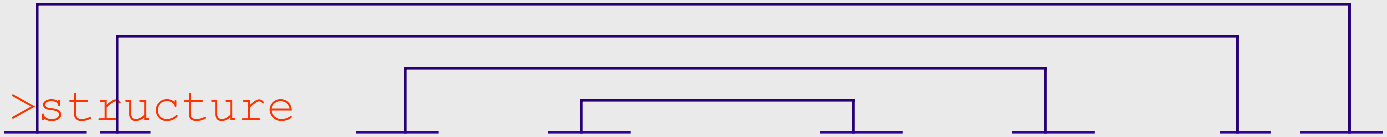
TP and **FN**: easy to obtain, using training set (*leave-one out*)

FP: harder! How do you know a hit is false?

Hint: express SP as: **FP / Mb in a random sequence**

Make it large enough and of same composition (mono & di-nt) as search database (e.g. with the *shuffle* program)

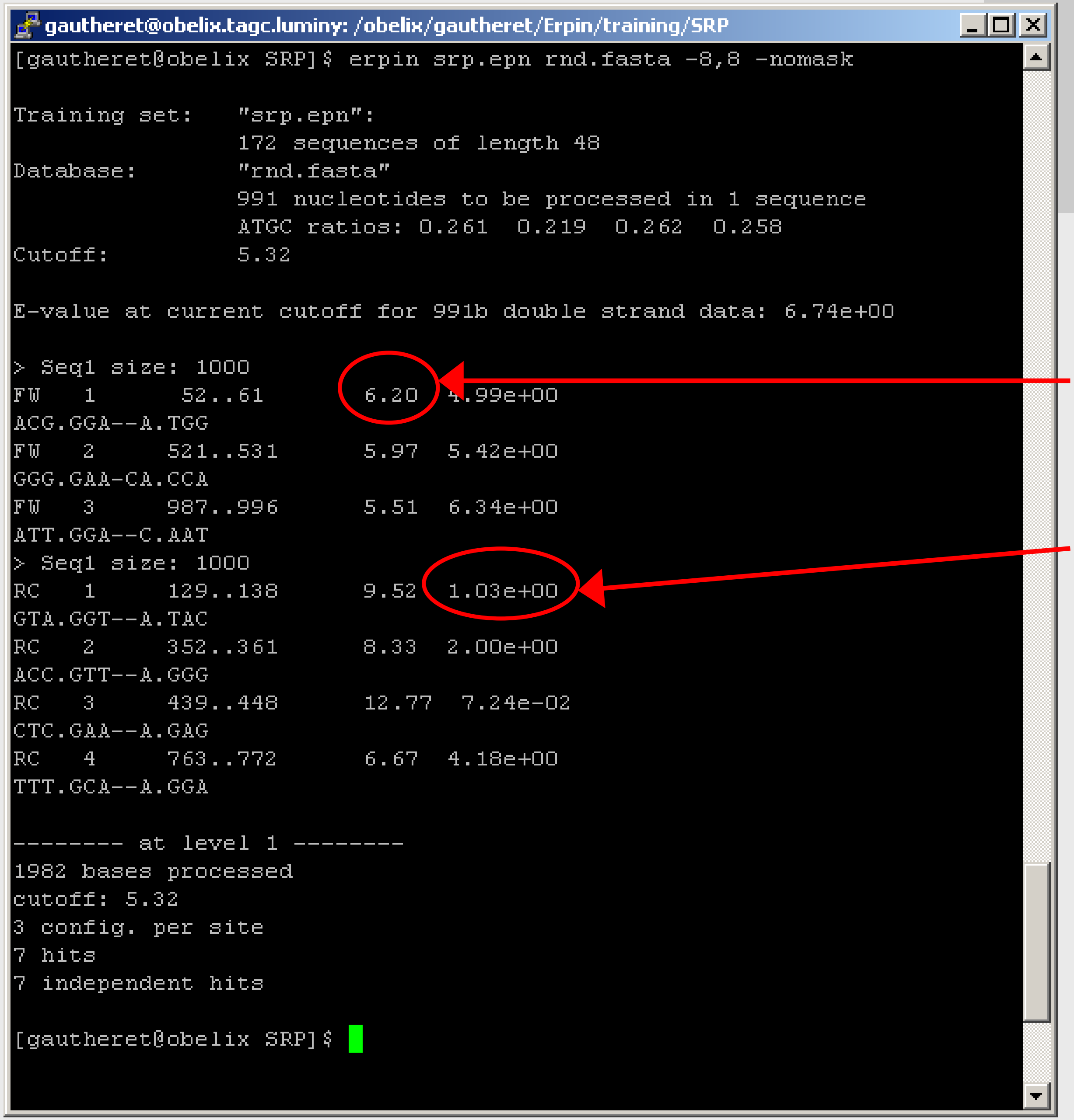
Using the ERPIN program



```
>structure
000000000000000000000000000000000000001111001000
222443333333666555588877777788899996661111443222
>AQU.AEO.
AGGGUGAACU-CCCCCAGGCCCGAA--AGGGAGCAAGGGUAAGC-CCG
>THE.THE.
GGCGUGAACC-GGGUCAGGUCCGGA--AGGAAGCAGCCCUAAGC-GCC
```

erpin srp.epn sequence.fasta -8,8 -nomask
erpin srp.epn sequence.fasta –2,2 -nomask
erpin srp.epn sequence.fasta -2,2 –umask 5 9 -nomask

ERPIN results


gautheret@obelix.tagc.luminy: /obelix/gautheret/Erpin/training/SRP

```
[gautheret@obelix SRP]$ erpin srp.epn rnd.fasta -8,8 -nomask

Training set:   "srp.epn":
                172 sequences of length 48
Database:       "rnd.fasta"
                991 nucleotides to be processed in 1 sequence
                ATGC ratios: 0.261  0.219  0.262  0.258
Cutoff:         5.32

E-value at current cutoff for 991b double strand data: 6.74e+00

> Seq1 size: 1000
FW   1      52..61       6.20  4.99e+00
ACG.GGA--A.TGG
FW   2     521..531      5.97  5.42e+00
GGG.GAA-CA.CCA
FW   3     987..996      5.51  6.34e+00
ATT.GGA--C.AAT
> Seq1 size: 1000
RC   1     129..138      9.52  1.03e+00
GTA.GGT--A.TAC
RC   2     352..361      8.33  2.00e+00
ACC.GTT--A.GGG
RC   3     439..448      12.77  7.24e-02
CTC.GAA--A.GAG
RC   4     763..772      6.67  4.18e+00
TTT.GCA--A.GGA

-------- at level 1 --------
1982 bases processed
cutoff: 5.32
3 config. per site
7 hits
7 independent hits

[gautheret@obelix SRP]$
```

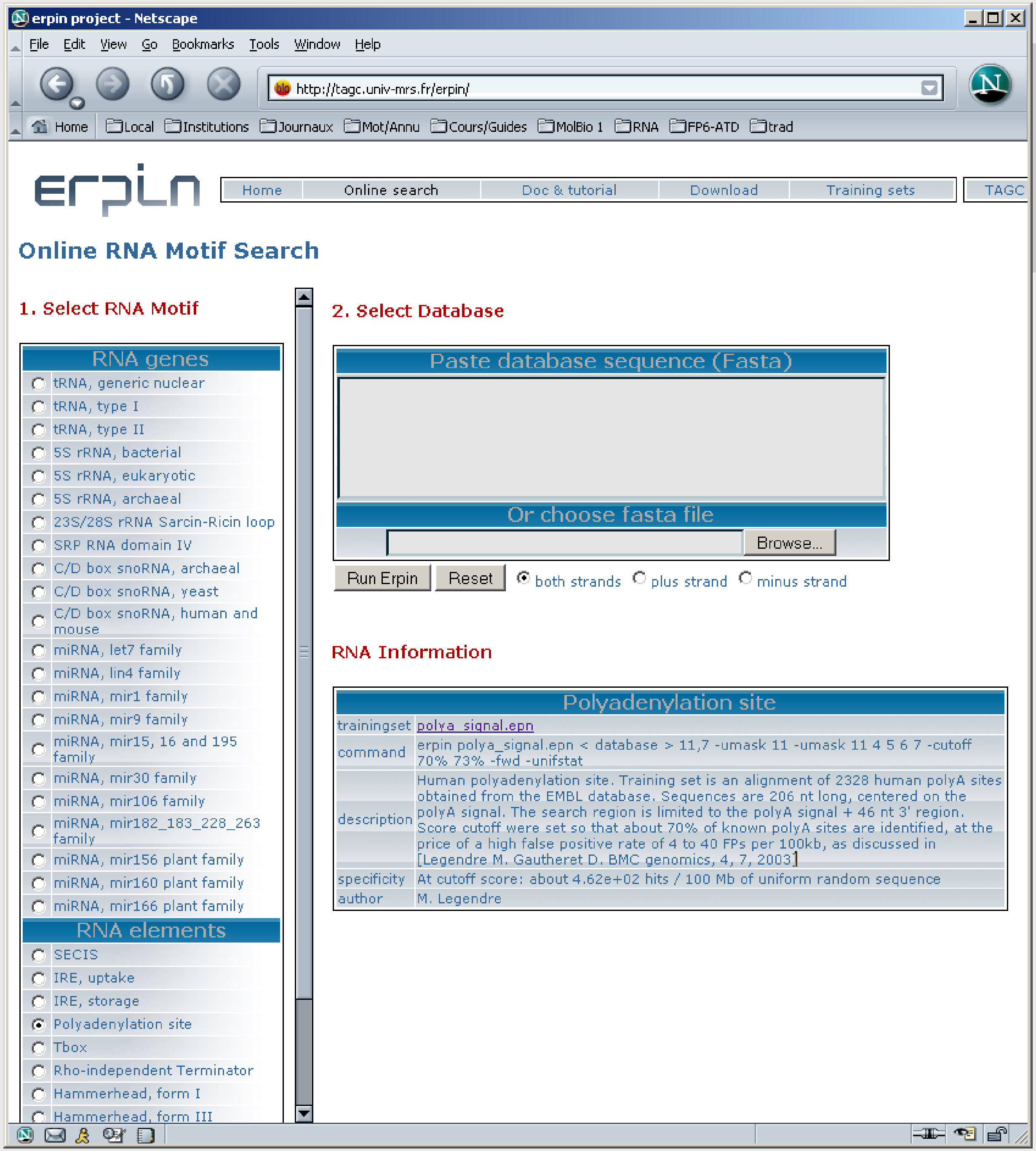
Score: based on profile values

E-value: How many hits expected at this score or higher?

No need for random sequence tests!

The ERPIN Server

http://tagc.univ-mrs.fr/erpin/



Online RNA Motif Search

1. Select RNA Motif

RNA genes

- ○ tRNA, generic nuclear
- ○ tRNA, type I
- ○ tRNA, type II
- ○ 5S rRNA, bacterial
- ○ 5S rRNA, eukaryotic
- ○ 5S rRNA, archaeal
- ○ 23S/28S rRNA Sarcin-Ricin loop
- ○ SRP RNA domain IV
- ○ C/D box snoRNA, archaeal
- ○ C/D box snoRNA, yeast
- ○ C/D box snoRNA, human and mouse
- ○ miRNA, let7 family
- ○ miRNA, lin4 family
- ○ miRNA, mir1 family
- ○ miRNA, mir9 family
- ○ miRNA, mir15, 16 and 195 family
- ○ miRNA, mir30 family
- ○ miRNA, mir106 family
- ○ miRNA, mir182_183_228_263 family
- ○ miRNA, mir156 plant family
- ○ miRNA, mir160 plant family
- ○ miRNA, mir166 plant family

RNA elements

- ○ SECIS
- ○ IRE, uptake
- ○ IRE, storage
- ● Polyadenylation site
- ○ Tbox
- ○ Rho-independent Terminator
- ○ Hammerhead, form I
- ○ Hammerhead, form III

2. Select Database

Paste database sequence (Fasta)

Or choose fasta file

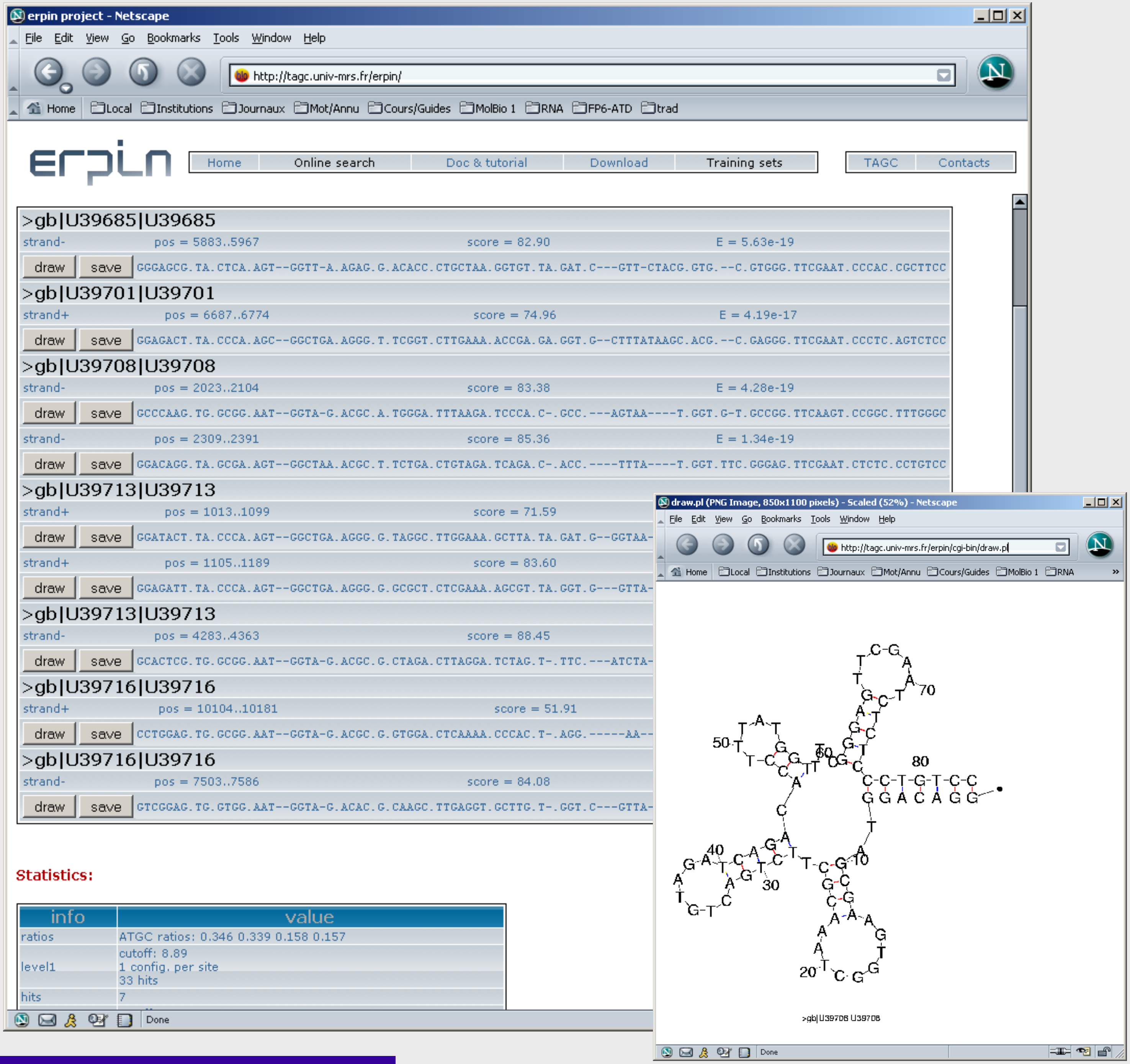
Browse...

Run Erpin Reset ● both strands ○ plus strand ○ minus strand

RNA Information

Polyadenylation site	
trainingset	polya_signal.epn
command	erpin polya_signal.epn < database > 11,7 -umask 11 -umask 11 4 5 6 7 -cutoff 70% 73% -fwd -unifstat
description	Human polyadenylation site. Training set is an alignment of 2328 human polyA sites obtained from the EMBL database. Sequences are 206 nt long, centered on the polyA signal. The search region is limited to the polyA signal + 46 nt 3' region. Score cutoff were set so that about 70% of known polyA sites are identified, at the price of a high false positive rate of 4 to 40 FPs per 100kb, as discussed in [Legendre M. Gautheret D. BMC genomics, 4, 7, 2003]
specificity	At cutoff score: about 4.62e+02 hits / 100 Mb of uniform random sequence
author	M. Legendre

All searches parameterized to scan a bacterial genome in less than 5 minutes


erpin project - Netscape
File Edit View Go Bookmarks Tools Window Help
http://tagc.univ-mrs.fr/erpin/
Home Local Institutions Journaux Mot/Annu Cours/Guides MolBio 1 RNA FP6-ATD trad
erpin
Home Online search Doc & tutorial Download Training sets
TAGC Contacts

```
>gb|U39685|U39685
strand-            pos = 5883..5967              score = 82.90           E = 5.63e-19
draw  save  GGGAGCG.TA.CTCA.AGT--GGTT-A.AGAG.G.ACACC.CTGCTAA.GGTGT.TA.GAT.C---GTT-CTACG.GTG.--C.GTGGG.TTCGAAT.CCCAC.CGCTTCC
>gb|U39701|U39701
strand+             pos = 6687..6774              score = 74.96           E = 4.19e-17
draw  save  GGAGACT.TA.CCCA.AGC--GGCTGA.AGGG.T.TCGGT.CTTGAAA.ACCGA.GA.GGT.G--CTTTATAAGC.ACG.--C.GAGGG.TTCGAAT.CCCTC.AGTCTCC
>gb|U39708|U39708
strand-            pos = 2023..2104              score = 83.38           E = 4.28e-19
draw  save  GCCCAAG.TG.GCGG.AAT--GGTA-G.ACGC.A.TGGGA.TTTAAGA.TCCCA.C-.GCC.---AGTAA----T.GGT.G-T.GCCGG.TTCAAGT.CCGGC.TTTGGGC
strand-            pos = 2309..2391              score = 85.36           E = 1.34e-19
draw  save  GGACAGG.TA.GCGA.AGT--GGCTAA.ACGC.T.TCTGA.CTGTAGA.TCAGA.C-.ACC.----TTTA----T.GGT.TTC.GGGAG.TTCGAAT.CTCTC.CCTGTCC
>gb|U39713|U39713
strand+             pos = 1013..1099              score = 71.59
draw  save  GGATACT.TA.CCCA.AGT--GGCTGA.AGGG.G.TAGGC.TTGGAAA.GCTTA.TA.GAT.G--GGTAA-
strand+             pos = 1105..1189              score = 83.60
draw  save  GGAGATT.TA.CCCA.AGT--GGCTGA.AGGG.G.GCGCT.CTCGAAA.AGCGT.TA.GGT.G---GTTA-
>gb|U39713|U39713
strand-            pos = 4283..4363              score = 88.45
draw  save  GCACTCG.TG.GCGG.AAT--GGTA-G.ACGC.G.CTAGA.CTTAGGA.TCTAG.T-.TTC.---ATCTA-
>gb|U39716|U39716
strand+           pos = 10104..10181                score = 51.91
draw  save  CCTGGAG.TG.GCGG.AAT--GGTA-G.ACGC.G.GTGGA.CTCAAAA.CCCAC.T-.AGG.-----AA--
>gb|U39716|U39716
strand-            pos = 7503..7586              score = 84.08
draw  save  GTCGGAG.TG.GTGG.AAT--GGTA-G.ACAC.G.CAAGC.TTGAGGT.GCTTG.T-.GGT.C---GTTA-
```

Statistics:

info	value
ratios	ATGC ratios: 0.346 0.339 0.158 0.157
level1	cutoff: 8.89 1 config. per site 33 hits
hits	7

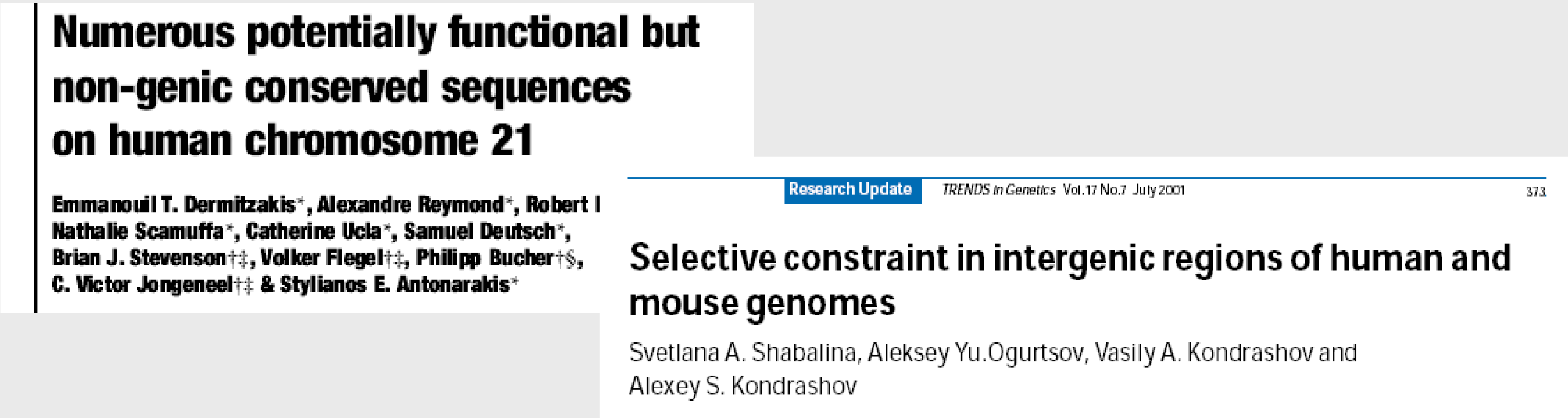
Done
draw.pl (PNG Image, 850x1100 pixels) - Scaled (52%) - Netscape
File Edit View Go Bookmarks Tools Window Help
http://tagc.univ-mrs.fr/erpin/cgi-bin/draw.pl
Home Local Institutions Journaux Mot/Annu Cours/Guides MolBio 1 RNA
70
80
50
60
40
30
10
20
>gb|U39708 U39708
Done

ncRNA detection

IV. *De novo* ncRNA finding

★ How can we detect ncRNA genes when no prior sequence/structure data is available?

Exciting times for comparative genomics



Numerous potentially functional but non-genic conserved sequences on human chromosome 21

Emmanouil T. Dermitzakis[*], Alexandre Reymond[*], Robert I
Nathalie Scamuffa[*], Catherine Ucla[*], Samuel Deutsch[*],
Brian J. Stevenson[†‡], Volker Flegel[†‡], Philipp Bucher[†§],
C. Victor Jongeneel[†‡] & Stylianos E. Antonarakis[*]

Research Update TRENDS in Genetics Vol.17 No.7 July 2001 373

Selective constraint in intergenic regions of human and mouse genomes

Svetlana A. Shabalina, Aleksey Yu.Ogurtsov, Vasily A. Kondrashov and
Alexey S. Kondrashov

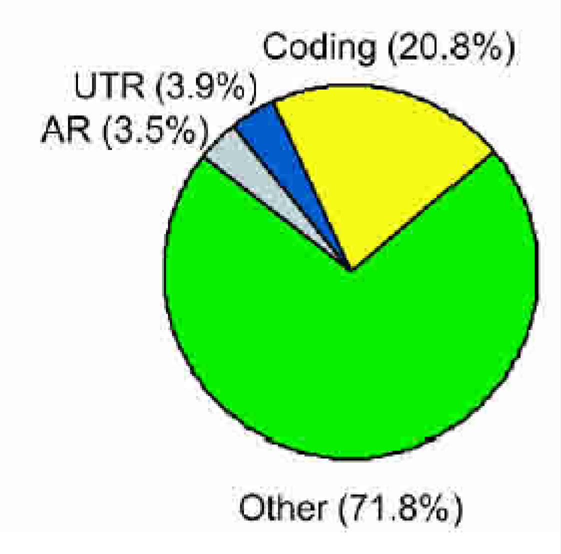
- ★ 5-6% of mamalian genome under selection vs 1.5% coding
- ★ 3 times as much as in nematodes!
- ★ *« Intergenic regions might hold the key to the complexity of mammals »*

Functional assignment of conserved regions

- ★ Coding exons
- ★ Regulatory non coding exons and introns
- ★ Promoters
- ★ ncRNA
- ★ Ancestral repeats
- ★ Others (matrix attachment, etc.)

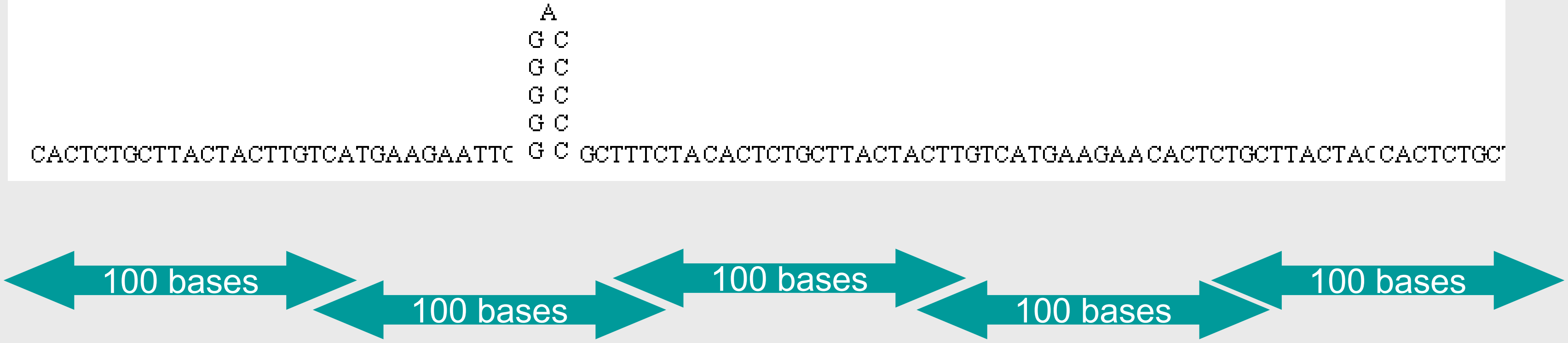
Detect this!

Fraction of conserved sequences in.. (AR=ancestral repeats)



Margulies et al, 2003

Thermodynamic Profiling (Le *et al*. 88)


A
G C
G C
G C
G C
CACTCTGCTTACTACTTGTCATGAAGAATTC G C GCTTTCTACACTCTGCTTACTACTTGTCATGAAGAA CACTCTGCTTACTAC CACTCTGCT
100 bases
100 bases
100 bases
100 bases
100 bases

Profile : - 4 -20 -2 -3 -2

$$\text{Z-score} = \frac{\text{window free energy} - \text{mean (energy of rnd seq.)}}{\sqrt{\text{Var(energy of rnd seq.)}}}$$

→ New software by Hofacker *et al*: (RNALfold)

The problem with thermodynamics

- OK for strong local structures (some success in viral genomes)
- However: true ncRNA (tRNA, rRNA) do not display higher folding energy than random sequences of same composition (di-nt: Rivas & Eddy 2000)

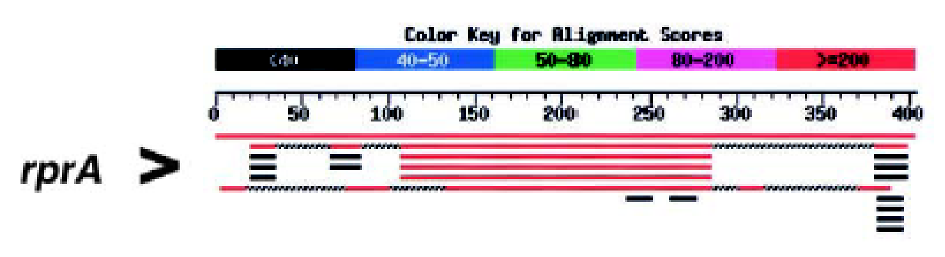
G+C content

- G+C content alone is a better ncRNA predictor than free energy
- In high A+T background (thermophilic archaebacteria), ncRNA stand out clearly.
- Combining (G+C)% and CpG% provides the best discriminant (Schattner ’02).
- Does not work in genomes with « normal » G+C contents, except as a complement to other methods (thermodynamics, etc.)

→ See software RNAGenie (Carter, Dubchak & Holbrook, 2001).
- Combines energy and G+C contents

Comparative Genomics + experiments

Bacteria: microarray + Northern in different growth conditions



From Wassarman et al. ’01
Seq: Escherichia, Salmonella, Klebsiella

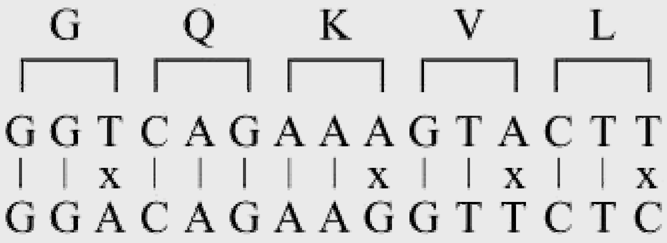
Wassarman *et. al.* ‘01: 60 ncRNA predicted, 23 confirmed

★ Argaman et al. ’01: 24 predicted, 14 confirmed

Q-RNA (Rivas & Eddy 2001)

★ Analysis of Blast alignment (SCFG based)

- Model for protein coding gene

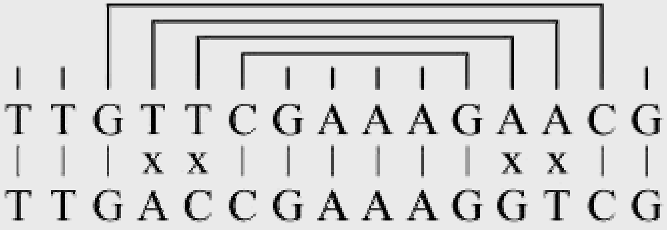

G Q K V L
G G T C A G A A A G T A C T T
| | x | | | | | x | | x | | x
G G A C A G A A G G T T C T C

P(GGT-GGA)*P(CAG-CAG)*...

Synonymous mutations

- Model for ncRNA

(also include loop probabilities obtained from training set of real ncRNA)


T T G T T C G A A A G A A C G
| | | x x | | | | | | x x | |
T T G A C C G A A A G G T C G

P(T-T)*P(T-T)*P(GC-GC)*P(TA-AT)*...

Compensatory mutations

Q-RNA results

★ Limited range for similarity (65%-85%)

- Too dissimilar: incorrect Blast alignments
- Too similar: no covariation

→ Problem: Human/mouse/rat ncRNAs not in this range!

★ *E.coli* vs *Salmonella typhy*: analysis of ~5000 Blast hits

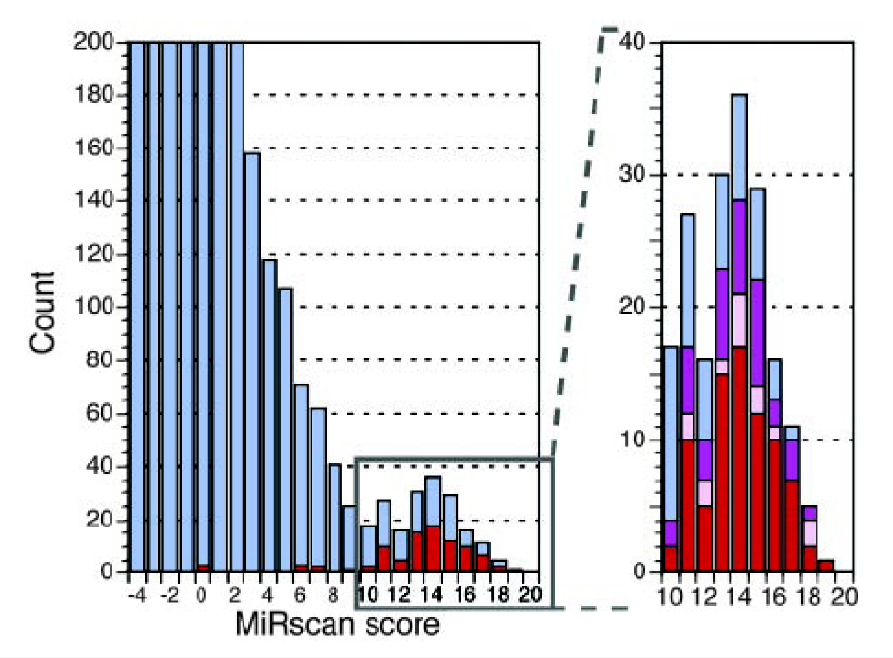
- 115 true ncRNAs
- 33 with Blastn alignments in the 65%-85% range
- 33 detected as ncRNA
- 440 other candidates (half of them known elements: terminators, palindromic repetitive elements, etc.)

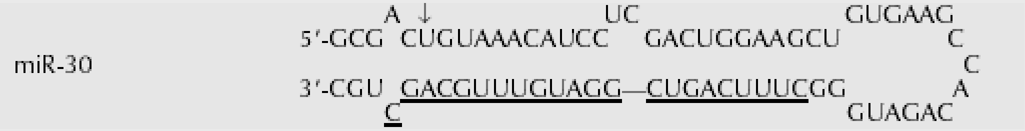
Comparative Genomics & miRNA

Lim et al. Science 2003

★ Criteria:

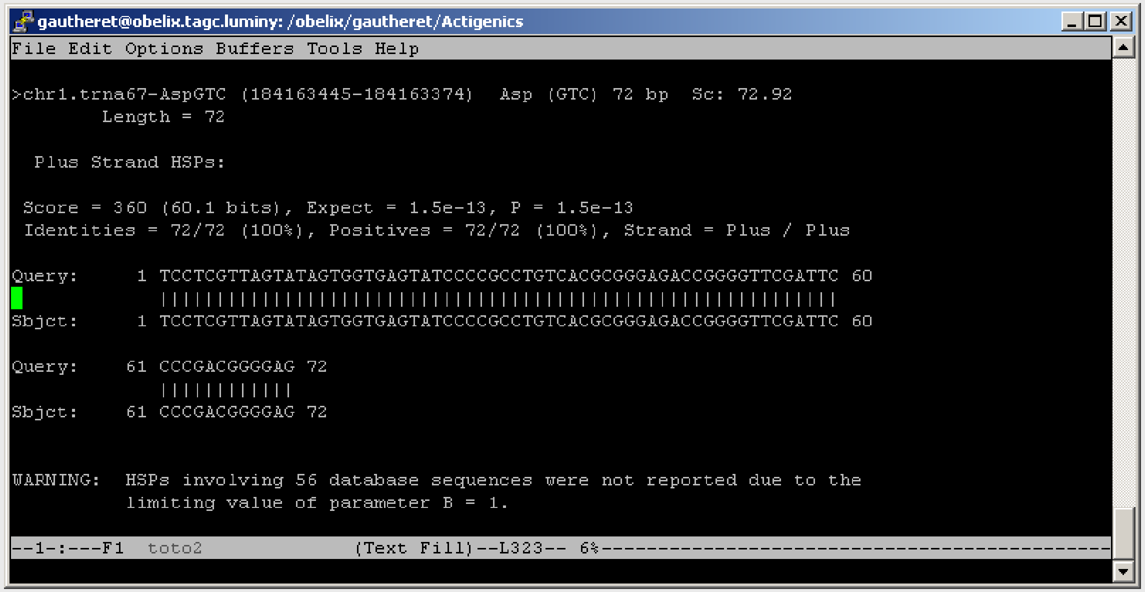
- Loose conservation human/mouse/fugu
- Fall outside of protein coding gene
- Predicted to form stem-loop
  - ➢ 15000 hits
- Score based on resemblance to 21mer miRNA
- 107 potential new miRNAs




miR-30
A ↓
UC
GUGAAG
5′-GCG CUGUAAACAUCC GACUGGAAGCU
C
C
3′-CGU GACGUUUGUAGG—CUGACUUUCGG
A
C
GUAGAC

The right species for ncRNA detection?

- Human/mouse ncRNA: ~98-100% id
- 18S fugu/xenopus/human: 95% id! Still too close
- Obvious interest for older animals



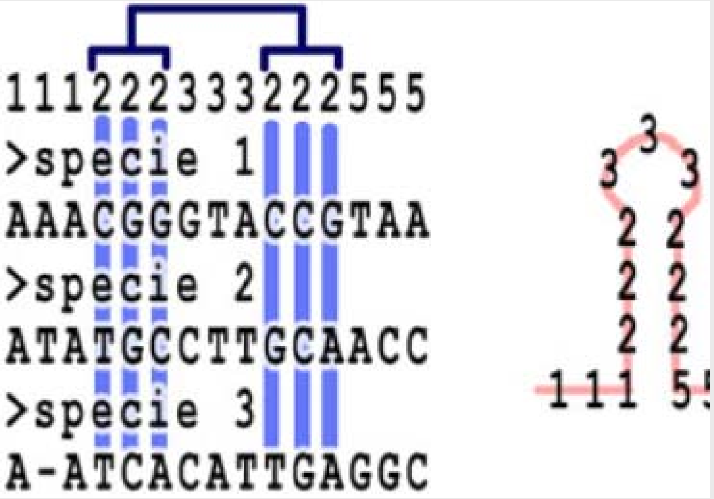
gautheret@obelix.tagc.luminy: /obelix/gautheret/Actigenics
File Edit Options Buffers Tools Help

```
>chr1.trna67-AspGTC (184163445-184163374)  Asp (GTC) 72 bp  Sc: 72.92
        Length = 72

  Plus Strand HSPs:

 Score = 360 (60.1 bits), Expect = 1.5e-13, P = 1.5e-13
 Identities = 72/72 (100%), Positives = 72/72 (100%), Strand = Plus / Plus

Query:     1 TCCTCGTTAGTATAGTGGTGAGTATCCCCGCCTGTCACGCGGGAGACCGGGGTTCGATTC 60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:     1 TCCTCGTTAGTATAGTGGTGAGTATCCCCGCCTGTCACGCGGGAGACCGGGGTTCGATTC 60

Query:    61 CCCGACGGGGAG 72
             ||||||||||||
Sbjct:    61 CCCGACGGGGAG 72


WARNING:  HSPs involving 56 database sequences were not reported due to the
          limiting value of parameter B = 1.

--1-:---F1  toto2           (Text Fill)--L323-- 6%-----------------------------------------
```

Human/mouse Asp tRNAs

Multiple species is the key

- Multiple alignments will enable covariation detection
- Covariation + GC-content + energy will provide enough evidence for ncRNA status


111222333222555
>specie 1
AAACGGGTACCGTAA
>specie 2
ATATGCCTTGCAACC
>specie 3
A-ATCACATTGAGGC
3
3
3
2 2
2 2
2 2
111 55