

ABA :

ARNomique et bioinformatique de l'ARN

Master BIBS 2e année

Séance 2

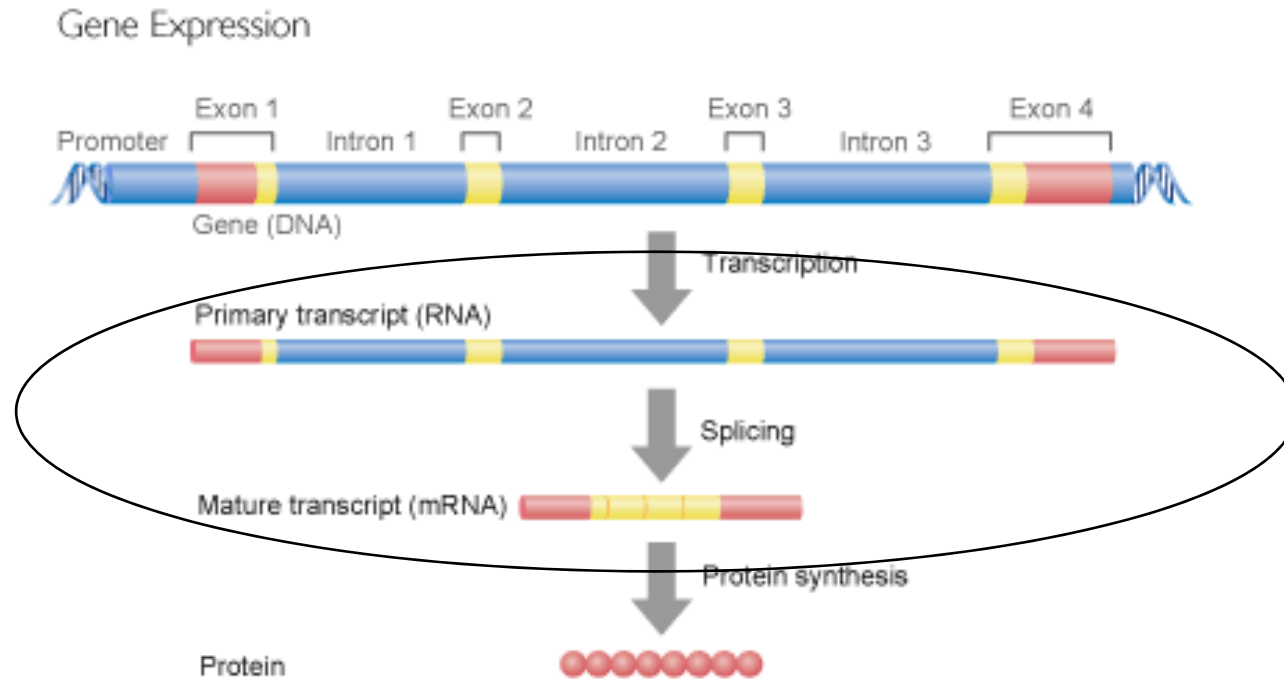


Daniel Gautheret

V.2009.1 rna.igmors.u-psud.fr/gautheret/cours/

La diversité des ARN

L'ARN messenger



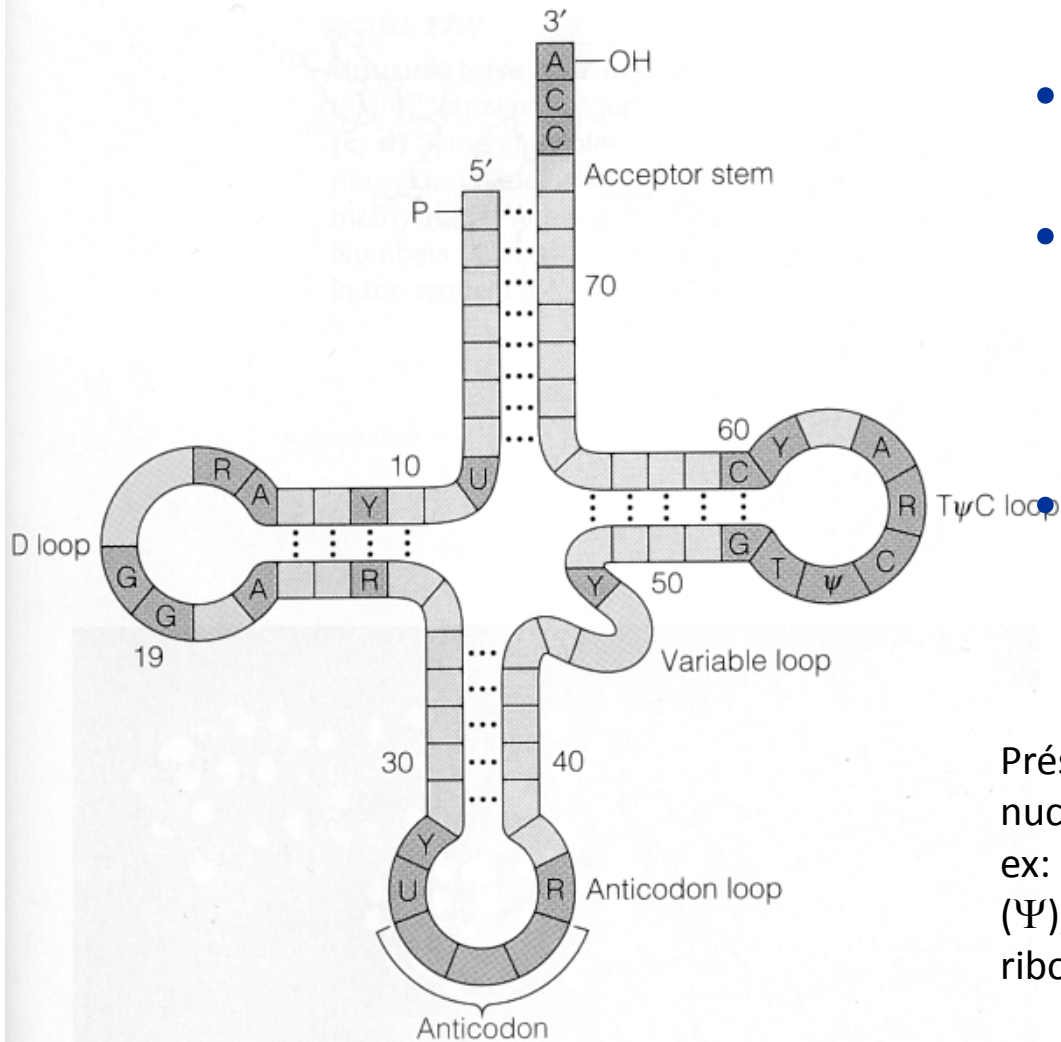
© Wellcome Trust

Les ARN non-messagers ou ARN non codants (ncRNA)

- ARNt
- ARNr
- ARNsn
- ARNsno
- micro-ARN
- Autres ARN 4.5S, 10Sa, Spot42, DicF, MicF, OxyS, DsrA, 6S (procaryotes) produits des gènes XIST, H19, IPW, 7H4, His-1, NTT (mammifères).

ARNr 18S et 28S: promoteurs pol-I. ARNt et ARNr 5S: promoteurs pol-III

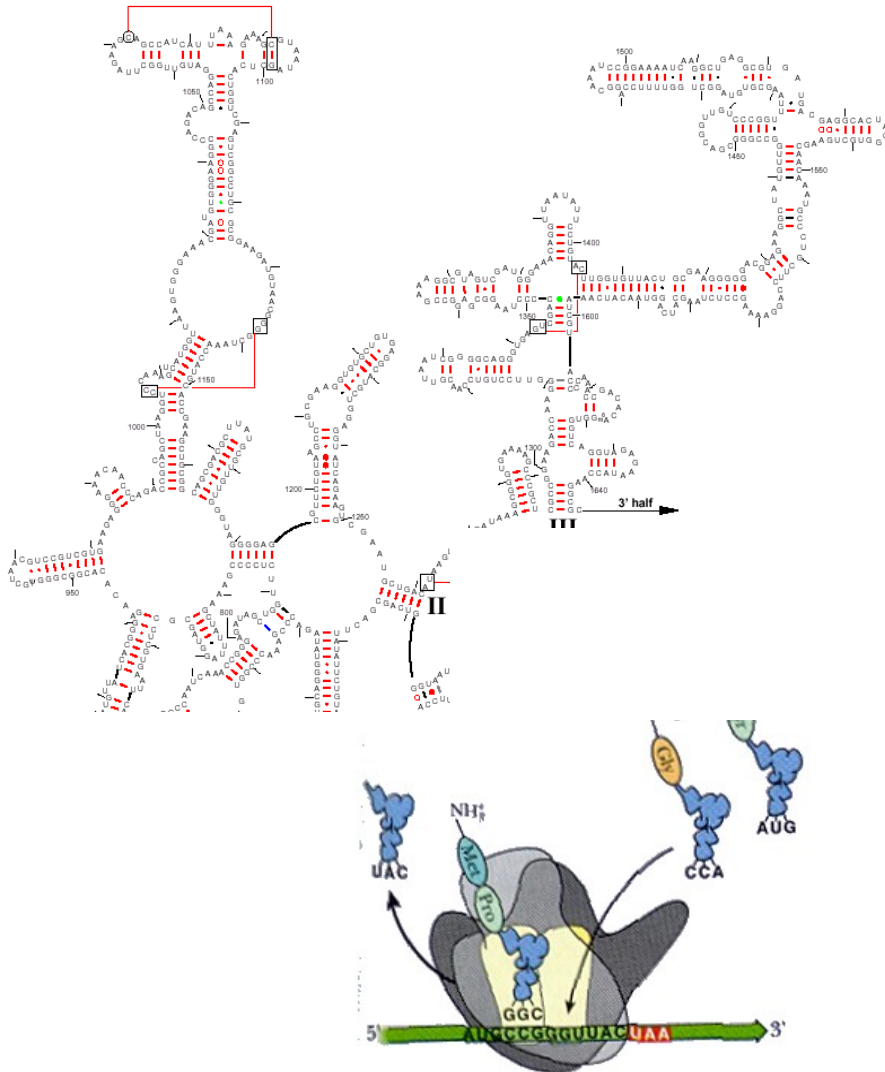
ARN de transfert (tRNA)



- Potentiellement 64 différents dans un génome
- En pratique une quarantaine dans les génomes microbiens. Qq centaines dans les génomes de mammifères.
- Le premier ARN observé en 3D (1974)

Présence de nombreux nucléosides inhabituels, par ex: inosine (I), pseudouridine (Ψ), dihydrouridine (D), ribothymidine (T), base Y, etc.

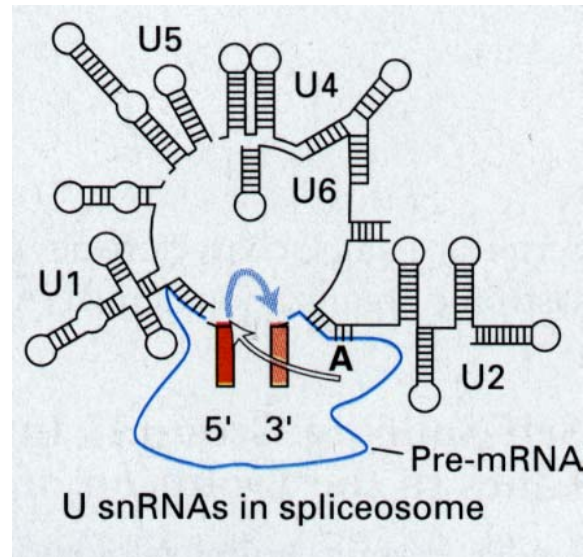
ARN ribosomique (rRNA)



- 5S, 16S, 23S pour les procaryotes (120, 1500 et 3000 bp)
- 5.5S, 18S et 28S pour les eucaryotes.
- Complexés avec protéines (~40)
- Observées en quelques exemplaires chez les procaryotes (7 opérons chez coli)
- Les génomes vertébrés peuvent accueillir plusieurs centaines de copies identiques. (clusters de gènes)
- Structure 3D en 1999

Small nuclear RNA (snRNA)

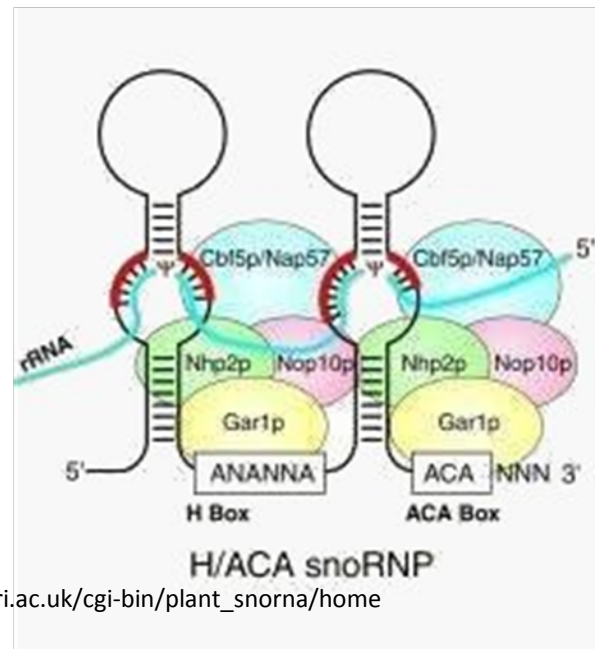
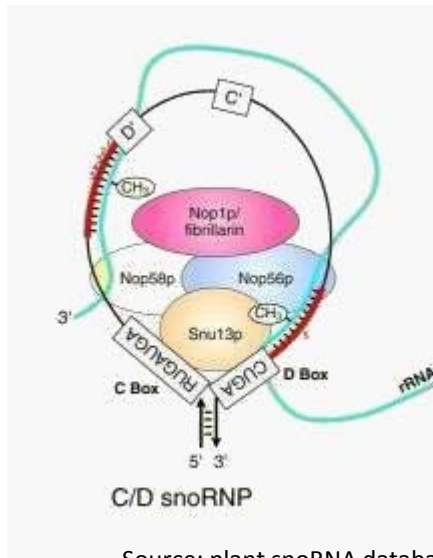
- Petits ARN du noyau impliqués dans l'épissage et la maintenance des télomères
- Complexés à des protéines
- Plusieurs sortes nommées U1, U2, U4, U6, U12...



(reference missing)

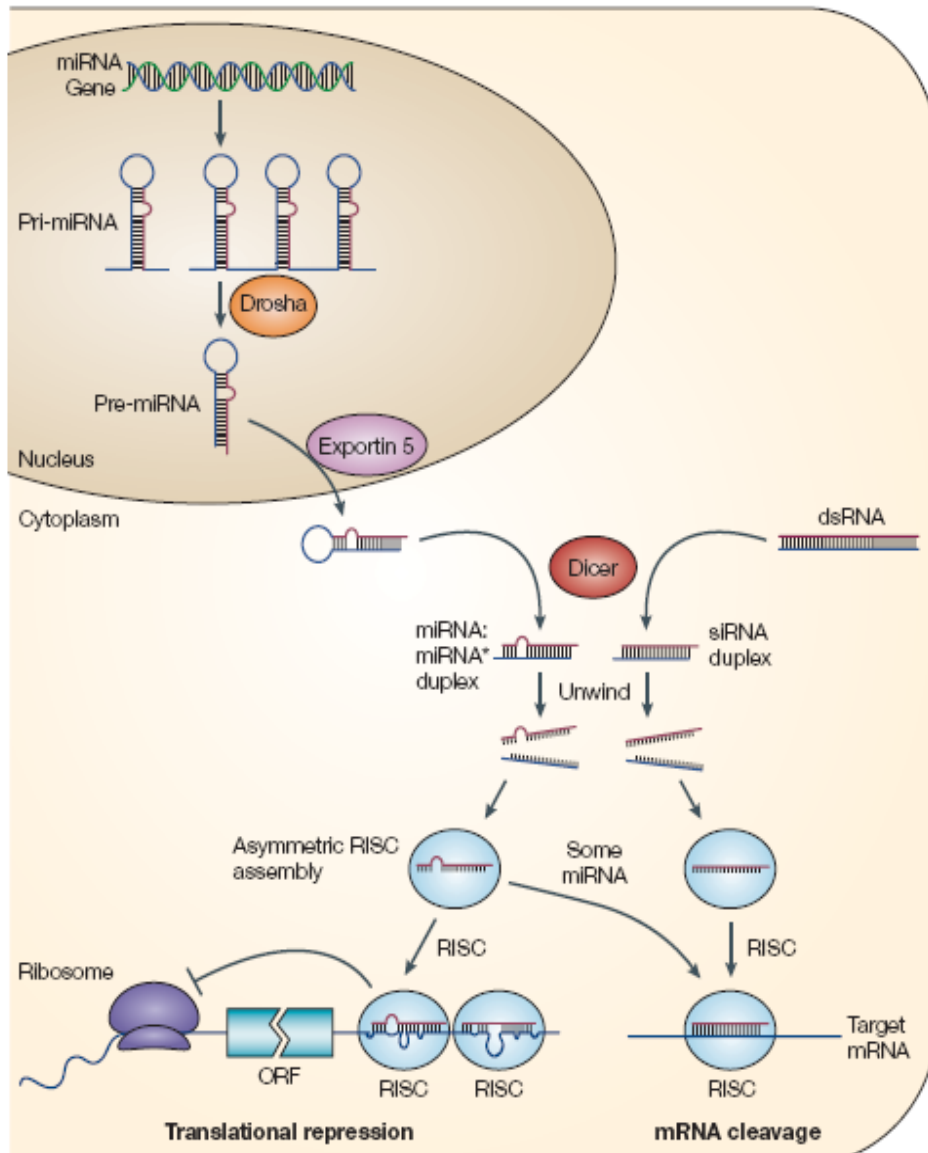
Small nucleolar RNAs (snoRNAs)

- Guide de méthylation ou de pseudo-uridylation des rRNA
- Eucaryotes et archae
- Deux familles: boîte C/D et boîte H/ACA
- Nomenclature: E2, E3, U3, U14, U23...



Source: plant snoRNA database http://bioinf.scri.sari.ac.uk/cgi-bin/plant_snorna/home

microRNA (miRNA)

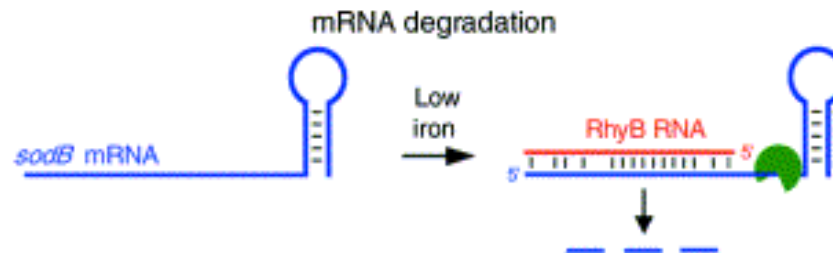


(He & Hannon, Nature reviews, 2004)

- Se trouvent chez les animaux et les plantes
- >700 chez l'homme
- Un seul miRNA peut cibler 100 gènes ou plus
- Recherche de miRNA
 - Conservation
 - Structure

Les « *small* » RNA bactériens

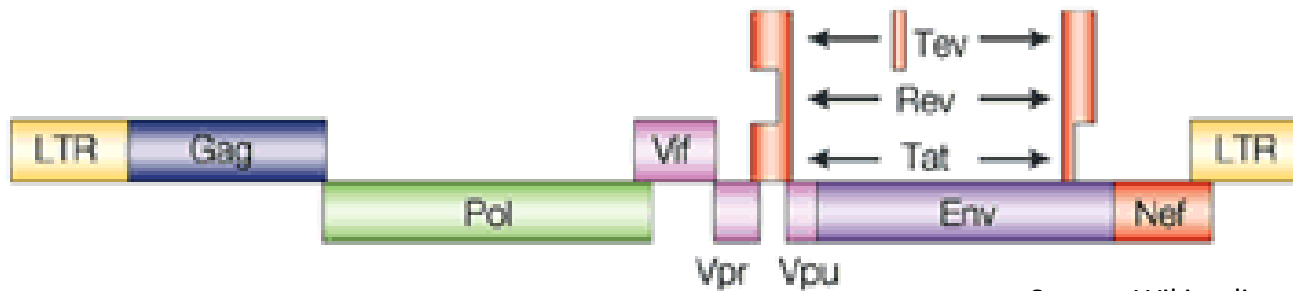
- Reconnaittent des séquences complémentaires dans les ARNm
- Bloquent la traduction ou déclenchent la dégradation de l'ARNm
- Un peu l'équivalent bactérien des miRNA, mais mécanisme très différent
- Plusieurs dizaines dans chaque génome



Exemple: le sRNA RyhB

Les génomes à ARN

- Virus à ARN
 - Double brin, simple brin
 - En partie codant, en partie non-codant



Source: Wikipedia

RFAM: The RNA database

Sam Griffiths-Jones, Simon Moxon, Mhairi Marshall, Ajay Khanna,
Sean R. Eddy and Alex Bateman.

The screenshot shows the Rfam website in a Mozilla Firefox browser window. The browser title is "Rfam: Rfam Home Page - Mozilla Firefox". The address bar shows the URL "http://www.sanger.ac.uk/Software/Rfam/". The website header features the Rfam logo and the text "RNA families database of alignments and CMs" along with the Wellcome Trust Sanger Institute logo. A navigation menu includes links for Home, Keyword Search, Sequence Search, Browse Rfam, Genomes, ftp, Help, and miRNA. The main content area contains an introductory paragraph about Rfam as a joint project, a list of features (view/download alignments, read annotations, examine species distribution, follow links), and information about using Rfam with the INFERNAL software suite. On the right side, there are two search boxes: "Enter your keyword(s) here" and "Enter an EMBL name or accession number", both with "Go" and "Example" buttons. At the bottom left, there are sections for "Rfam Mirror Servers Worldwide" (listing Sanger Institute in UK and St. Louis in USA) and "FTP access to Rfam". A "References" section at the bottom left includes a citation for a 2005 publication. The browser status bar at the bottom shows "Terminé".

Rfam: Rfam Home Page - Mozilla Firefox

Fichier Edition Affichage Aller à Marque-pages Outils ?

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

Local BioMol journaux Annuaire Trad

Rfam RNA families database of alignments and CMs

Wellcome Trust Sanger Institute

Home Keyword Search Sequence Search Browse Rfam Genomes ftp Help miRNA

Rfam Home Page

Rfam is a joint project involving researchers based at the [Wellcome Trust Sanger Institute](#), and [Washington University, St. Louis](#). Rfam is a large collection of multiple sequence alignments and covariance models covering many common non-coding RNA families. For each family in Rfam you can:

- View and download multiple sequence alignments
- Read family annotation
- Examine species distribution of family members
- Follow links to otherdatabases

In conjunction with the [INFERNAL](#) software suite, Rfam can be used to annotate sequences (including complete genomes) for homologues to known non-coding RNAs. Please read important information about [using Rfam for genome annotation](#). We provide pre-calculated lists of putative RNAs in over [100 complete genomes](#), and a [web search facility](#) for short sequences.

Rfam makes use of a large amount of available data, especially published multiple sequence alignments, and repackages these data in a single searchable and sustainable resource. We have made every effort to credit individual sources on family pages, and have compiled a list of links to these sources [here](#). If you find any of the data presented here useful, please also be sure to credit the primary source.

For more information on Rfam, and using this site, click [here](#).

Version 7.0
March 2005, 503 families

Enter your keyword(s) here
 Go Example

Enter an EMBL name or accession number
 Go Example

Rfam Mirror Servers Worldwide

- 🇬🇧 [Sanger Institute \(UK\)](#)
- 🇺🇸 [St. Louis \(USA\)](#)

FTP access to Rfam

You can also download the Rfam database and for instance search it locally using the [INFERNAL](#) covariance model software. [Hyperlink directly to the ftp site](#) or [View ftp site files](#)

References

If you find Rfam useful, please cite the following publication:
[Rfam: annotating non-coding RNAs in complete genomes.](#)
Sam Griffiths-Jones, Simon Moxon, Mhairi Marshall, Ajay Khanna, Sean R. Eddy and Alex Bateman.
Nucleic Acids Res. 2005 33:D121-D124.

Terminé

RFAM Stats (V.9.1)

- 1400 Familles d'ARN
 - une famille = un groupe d'ARN homologues et alignables.
 - Une « clique »: un groupe de familles de même fonction (miRNA, RNase P..)
 - >100 familles: miRNA, snoRNAs..
- Homme:
 - >100 familles
 - 3000 ARN différents annotés
- E. Coli: ~50 familles

Un alignement Rfam

AL606727.3/61372-61292	Dan.rer.	GCCUGGCUGUAGCAGCACGUA	AAUAUUGG	AGUCAAAAGCACUUGCGAAUC.	CUCCAGUAUUGACC	GUGCUGCU	GGAGUU	Next		
CR478286.10/109802-109722	Dan.rer.	UCCUCGCUUUAGCAGCACGUA	AAUAUUGG	UGUGUUUAUAGUCAAGGCCAA.	CCCAAUAUU	AUGUGUGCUGCU	UCAGUAA	Next		
AL669869.10/100495-100571	Mus.mus.	UCCUGGCUCUAGCAGCACAGAA	AAUAUUGG	CAUGGGGAAGUGAGUCUG.	CCCAAUAUU	GGCUGUGCUGCU	CCAGGCA	Next		
AC119847.12/19552-19632	Mus.mus.	GUUCCACUCUAGCAGCACGUA	AAUAUUGG	CGUAGUGAAAUAUUAAACA	CCAAUAUU	.AUUGUGCUGCU	UUAGUGU	Next		
AC154660.2/77731-77651	Mus.mus.	GCGGUGCUIUAGCAGCACGUA	AAUAUUGG	CGUUAAGAUCUGAAAUAAC.	CUCCAGUAUU	GACUGUGCUGCU	GAAAGUAA	Next		
AY865836.1/531-611	Mac.nem.	GUUCCACUCUAGCAGCACGUA	AAUAUUGG	CGUAGUGAAAUAUGUAUUAAACA	CCAAUAUU	.ACUGUGCUGCU	UCAGUGU	Next		
AY865955.1/411-487	Gor.gor.	CCUGGCUCUAGCAGCACAGAA	AAUAUUGG	CACAGGGGAAGCGAGUCUG.	CCCAAUAUU	GGCUGUGCUGCU	CCAGGCA	Next		
AF480551.1/1-81	Hom.sap.	GCAGUGCUIUAGCAGCACGUA	AAUAUUGG	CGUUAAGAUCUAAAAUUUAU.	CUCCAGUAUU	AACUGUGCUGCU	GAAAGUAA	Next		
AF480552.1/1-81	Hom.sap.	GUUCCACUCUAGCAGCACGUA	AAUAUUGG	CGUAGUGAAAUAUAUUAAACA	CCAAUAUU	.ACUGUGCUGCU	UUAGUGU	Next		
AY865833.1/553-633	Pan.tro.	GUUCCACUCUAGCAGCACGUA	AAUAUUGG	CGUAGUGAAAUAUAUGUUAAACA	CCAAUAUU	.ACUGUGCUGCU	UUAGUGU	Next		
AY866308.1/534-614	Lag.lag.	GCAGUGCUIUAGCAGCACGUA	AAUAUUGG	CGCUAAGAUCUAAAAUUUAU.	CUCCAGUAUU	AACUGUGCUGCU	GAAAGUAA	Next		
AY865835.1/569-649	Lag.lag.	GUUCCACUCUAGCAGCACGUA	AAUAUUGG	CGUAGUGAUUAUAUAUUAAACA	CCAAUAUU	.ACUGUGCUGCU	UUAGUGU	Next		
SS_cons		<<<<	<<<<<<<<	>>>>>>>>	>>>>	Next

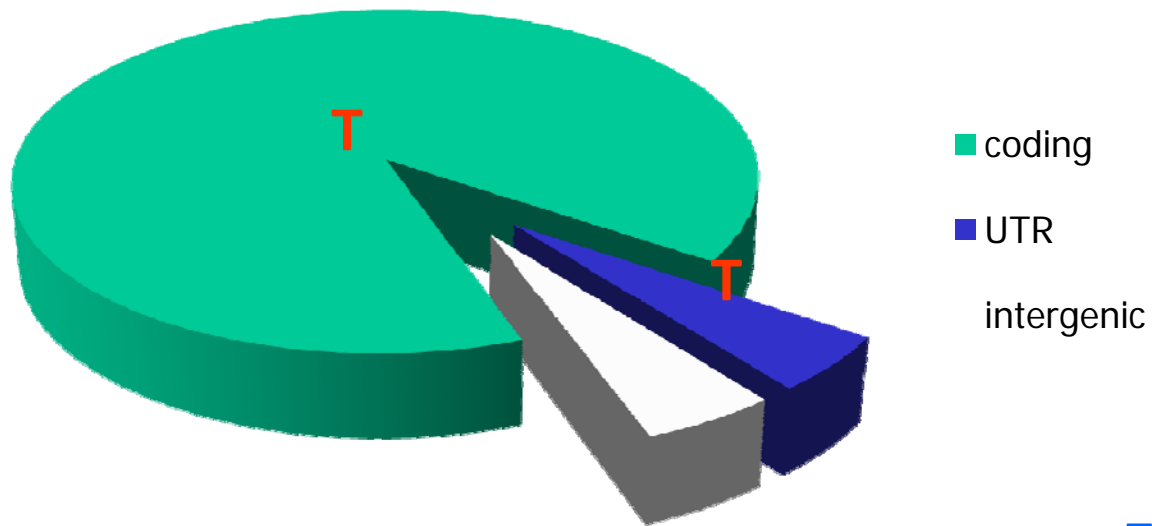
AL606727.3/61372-61292	Dan.rer.	GGC	Prev
CR478286.10/109802-109722	Dan.rer.	GGC	Prev
AL669869.10/100495-100571	Mus.mus.	GGG	Prev
AC119847.12/19552-19632	Mus.mus.	GAC	Prev
AC154660.2/77731-77651	Mus.mus.	GGU	Prev
AY865836.1/531-611	Mac.nem.	GAC	Prev
AY865955.1/411-487	Gor.gor.	GGG	Prev
AF480551.1/1-81	Hom.sap.	GGU	Prev
AF480552.1/1-81	Hom.sap.	GAC	Prev
AY865833.1/553-633	Pan.tro.	GAC	Prev
AY866308.1/534-614	Lag.lag.	GGU	Prev
AY865835.1/569-649	Lag.lag.	GAC	Prev
SS_cons		...	Prev

Micro RNA miR 16

Combien d'ARNnc à découvrir?

Bacterial genomes

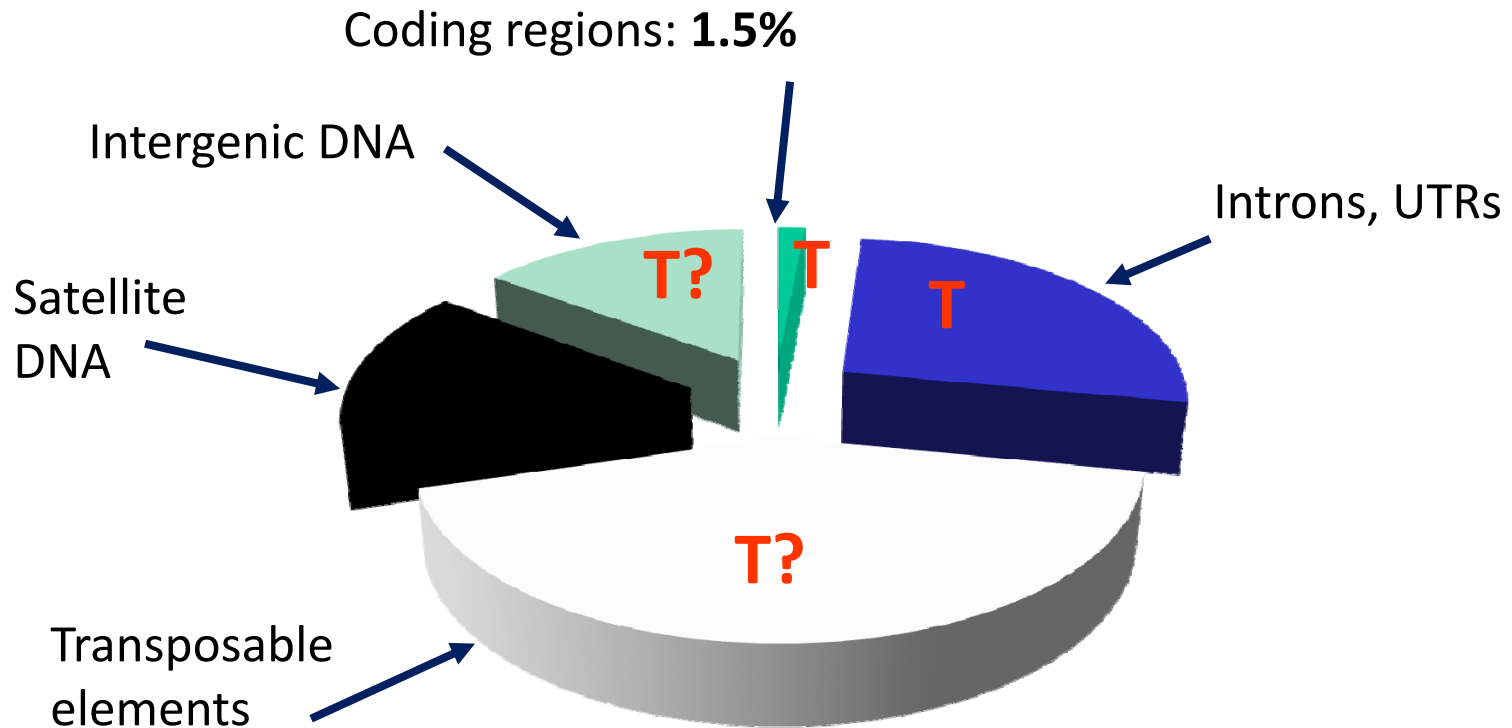
- 95% transcribed (~90% coding)



Transcription: **T**

Vertebrate Genomes

- >30% transcribed (93%?) / 1.5% coding



Transcription: **T**

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

**Les approches expérimentales:
l' « ARNominique »
(Rnomics)**

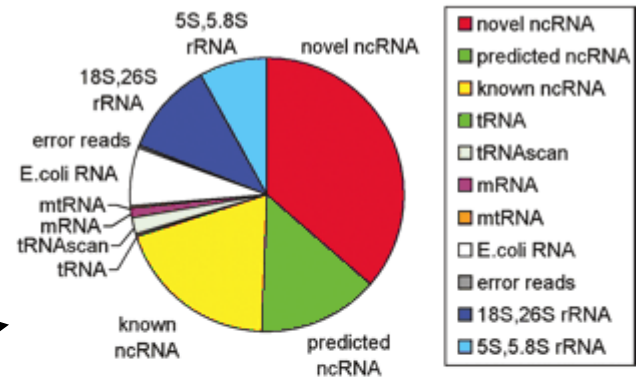
Avant 2007: cloning

– Rnomics*

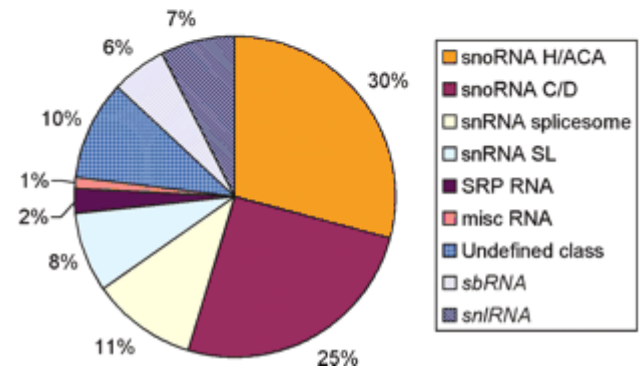
- Extraire ARN total
- Isoler petits RNAs
- Marquage et reverse transcription
- Clonage & Sequençage

- 200 ncRNA chez la souris
- 160 ncRNA chez *C. elegans*
- 100 ncRNA chez différentes bactéries

A Distribution of sequenced clones



B Functional class



* Huttenhofer et al., 2001

Limitations

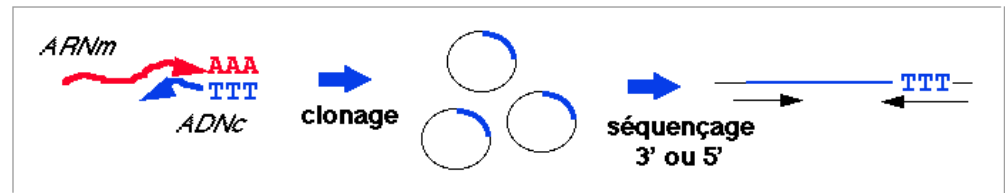
- Sensibilité: ncRNA rares, exprimés à des sites précis ou pendant une courte durée
- Non exhaustivité

Autres approches

– Full-length cDNA projects

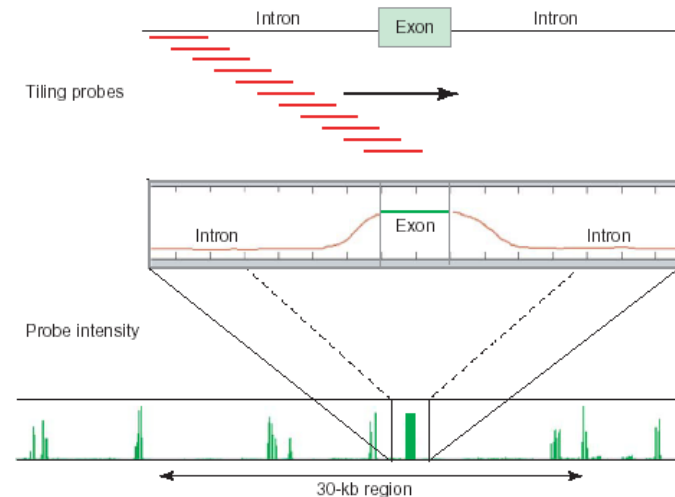
- FANTOM3:

- 100,000 mouse cDNAs
- 32,000 non-coding!



– Tiling arrays

- Half human transcriptome is polyA-, cytoplasmic and maps unannotated loci



Depuis 2007: RNA-seq

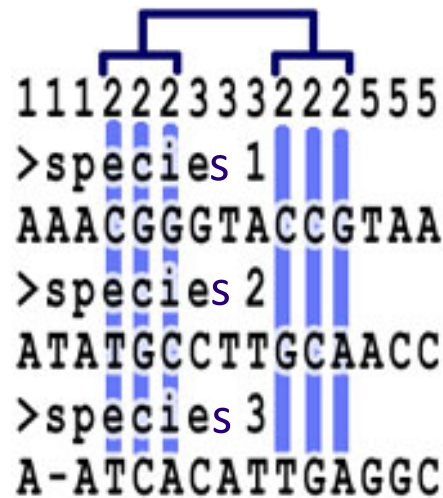
- Extraire ARN total
- Séparation petits ARN sur gel
- Amplification/ligation avec linkers
- Séquençage massif sur séquenceur haut débit
Illumina / Solid /454 (1M à 100M de reads 30-100nt)
 - Nombreux travaux sur microRNA, piRNA etc.

Limitations des techniques haut-débit

- Nombreux transcrits non fonctionnels (« fuites » de la transcription)
- Pas de preuve de fonction en tant qu'ARN

Recherche Bioinformatique de gènes ARN

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

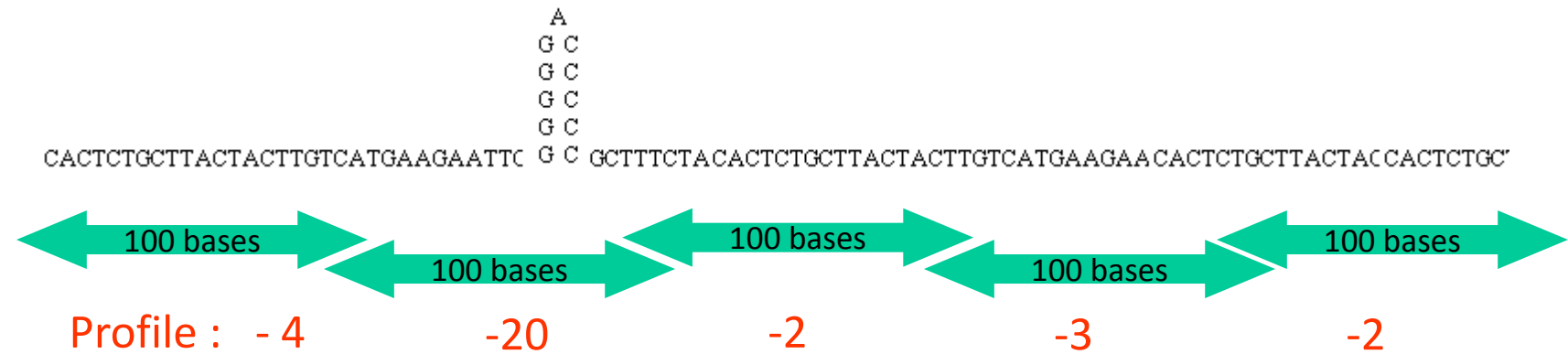
Deux classes de programmes

- « *de novo* »: recherche de gènes de familles inconnues
- Par homologie: recherche de gènes de familles connues

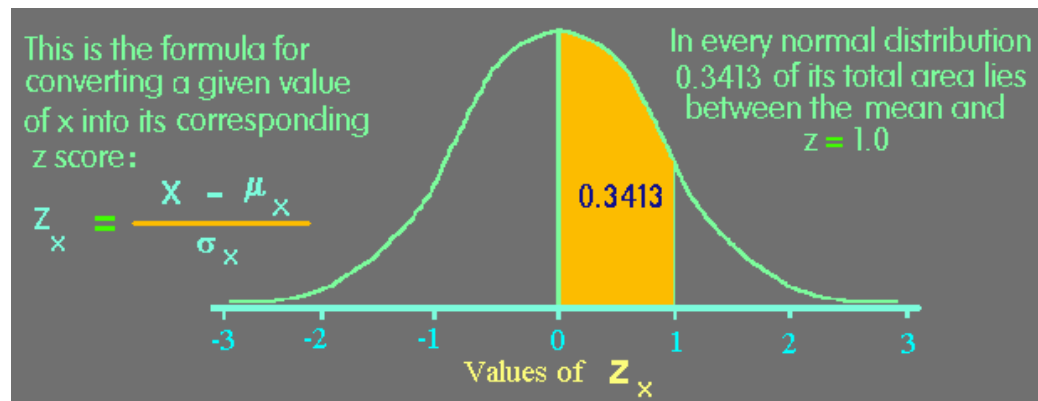
De novo ncRNA finding

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

Thermodynamic Profiling (Le et al. 88)



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



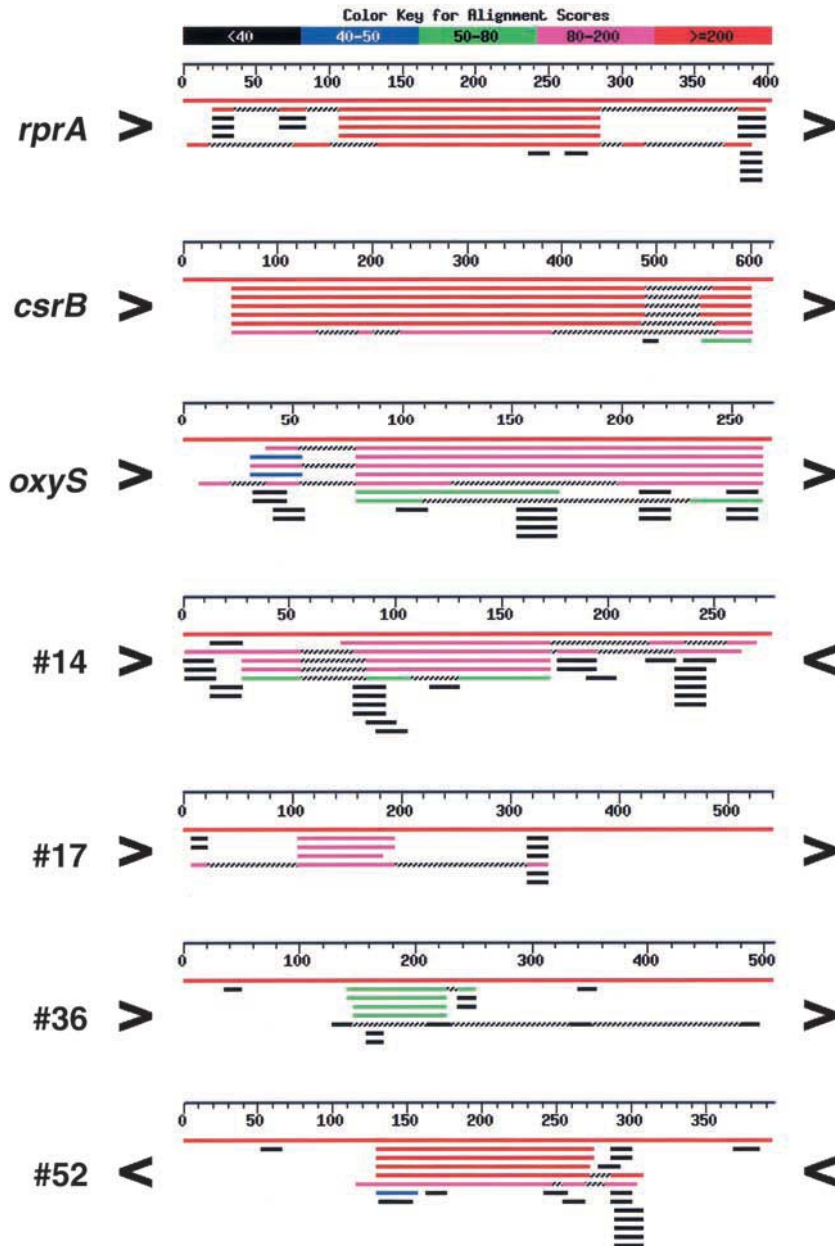
Limite de l'approche thermodynamique

- ★ OK pour structures locales fortes (qq réussites avec génomes viraux)
- ★ Mais: les vrais ARNnc (tRNA, rRNA) n'ont pas une meilleure énergie que des séquences aléatoires de même composition (en di-nt: Rivas & Eddy 2000)
- ★ La méthode rate de nombreux ARNnc
- ★ Le contenu en G+C à lui seul est un meilleur prédicteur de ncRNA que l'énergie libre

Contenu en G+C

- ★ Dans un génome riche en A+T (ex. bactéries thermophiles), les ARNnc se distinguent nettement.
- ★ Une combinaison de (G+C)% et CpG% fournit la meilleure discrimination (Schattner '02).
- ★ Qq dizaines d'ARNnc ont été ainsi prédits et confirmés expérimentalement dans *M. jannaschii* et *P. furiosus*.
- ★ Ne fonctionne pas dans genomes à contenu en G+C « normal », sauf en complément avec d'autres méthodes (thermodynamique, etc.)

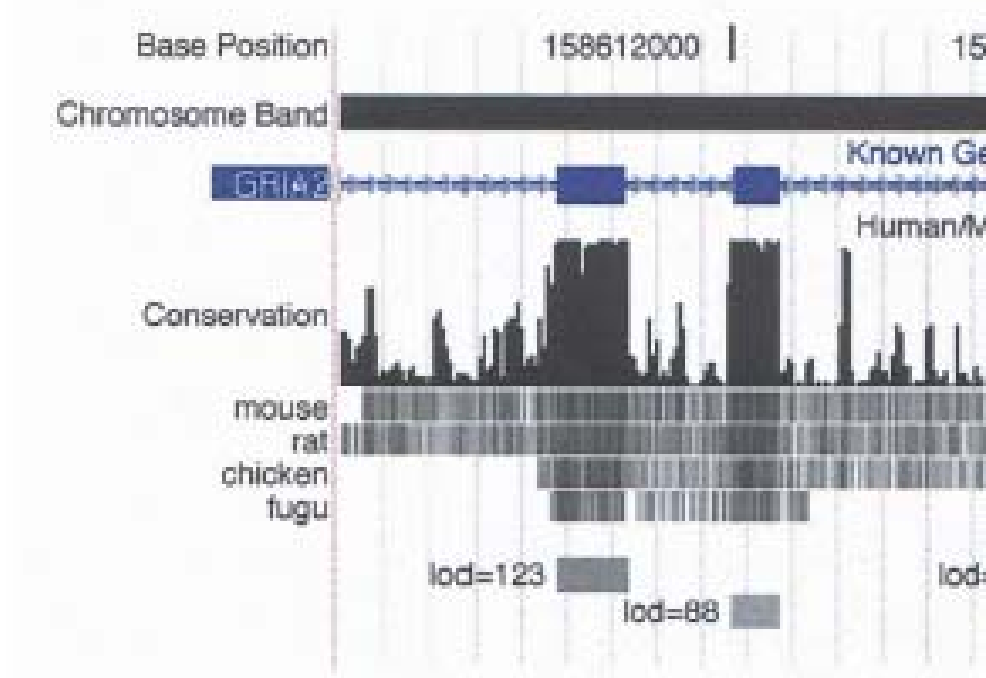
La génomique comparative



- Wassarman *et. al.* '01: comparaison Escherichia, Salmonella, Klebsiella : 60 ncRNA predicted, 23 confirmed
- Many ongoing projects in bacteria, Xenopus, Ciona, human

Wassarman et al. Genes & Dev. 2001 BLAST alignments of representative Intergenic regions.

Comparative genomics in human



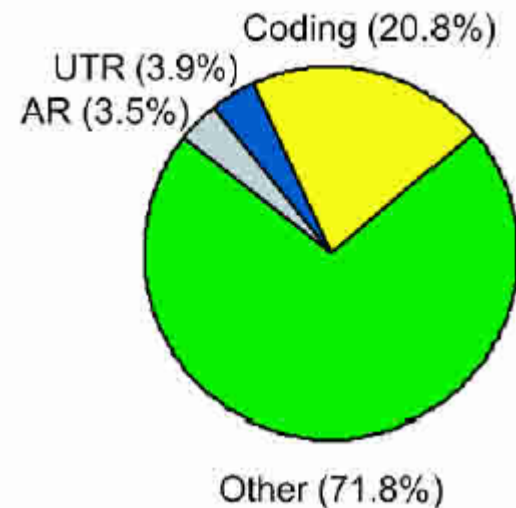
- 5-6% of mammalian genome under selection vs 1.5% coding (3 times as much as in nematodes)

Functional assignment of conserved regions

- Coding exons
- Regulatory sequences in exons and introns
- Promoters
- ncRNA
- Ancestral repeats
- Others (matrix attachment, etc.)

Detect this!

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



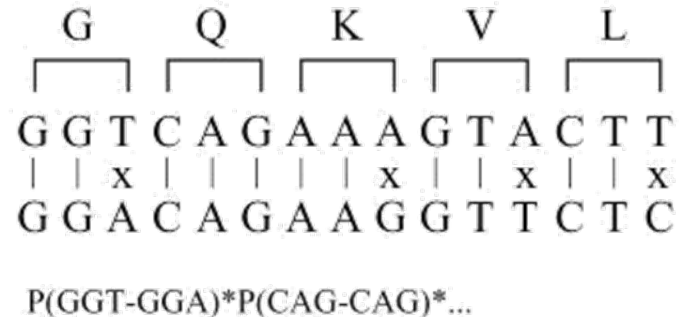
Margulies et al, 2003

- Need classification software
- QRNA (Rivas & Eddy, 2001)
 - RNAz (Hofacker & Stadler, 2005)

Q-RNA (Rivas & Eddy 2001)

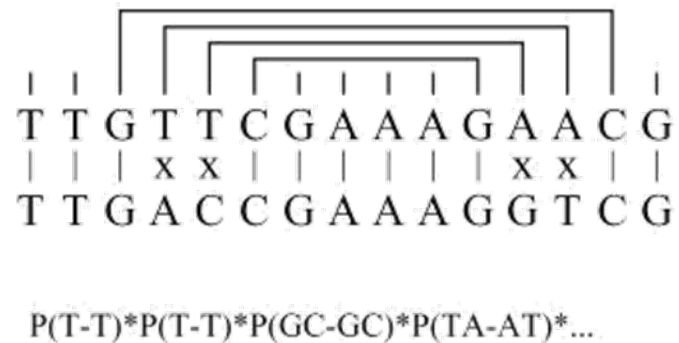
– Analysis of Blast alignment (SCFG based)

- Model for protein coding gene



Synonymous mutations

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



Compensatory mutations

RNA-Alifold (Hofacker, Stadler)

- Originally: secondary structure prediction
- Predicts best common structure for a set of aligned RNAs
- Dynamic programming, averaging:
 - Energies of aligned bases
 - Covariation term

RNAz (Washietl et al. 2004)

- Uses multiple alignments.
- two basic components:
 - (1) Measure RNA secondary structure conservation based on computing an RNAalifold consensus secondary structure (Structure Conservation Index)
 - (2) Measure thermodynamic stability, based on a Z-score normalized for sequence length and base composition and can be calculated without sampling from shuffled sequences (SVM)
 - An SVM again to combine 1 and 2

Q-RNA & RNAz

– QRNA

- Limited range for similarity (65%-85%): too dissimilar= incorrect Blast alignments, too similar=no covariation

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

- Pairwise: limited covariation

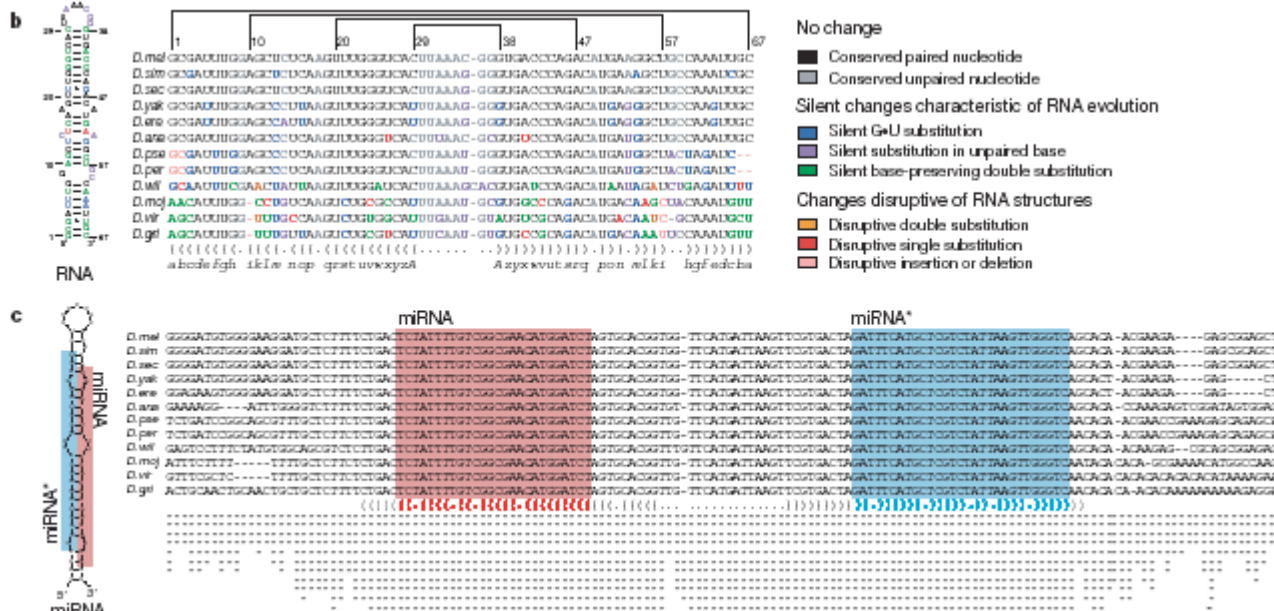
– RNAz

- Multiple alignment: better use of sequence variation
- Already applied to mammals (5-species alignment) & Ciona
 - Human: 10,000 ncRNAs predicted

De multiples études comparatives en cours

- ENCODE: région de 30Mb, 14 mammifères
 - Tous les programmes découvrent des ARN différents!
- 12 Drosophiles
 - Nature 2008

Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures



Looking for known ncRNAs

- How can we detect ncRNA genes from known families?

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
```

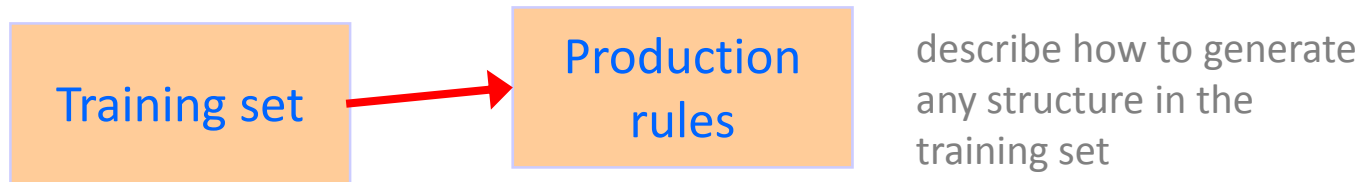
RnaMot descriptor for
anticodon+TYC domain of
tRNA

Descriptor-based programs

PROS	CONS
Draft descriptors can be quickly sketched and tested	Requires a good prior knowledge of secondary structure and sequence constraints
Alignment is not compulsory , although it is very helpful to have one	Requires basic computer skills to translate biological constraints into computer script
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)



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

ERPIN: Secondary Structure Profiles

Alignment

1	2	3	4	5	6	7	8	9
A	C	C	C	G	A	G	G	U
A	C	C	C	A	A	G	G	U
G	C	G	C	G	-	U	G	C
A	C	C	-	G	-	G	G	U
A	C	G	-	G	A	C	G	U

Weight matrices
Implemented in
PSI-Blast or
Prosite

- 16-row matrix captures base correlations and base-pair freqs.

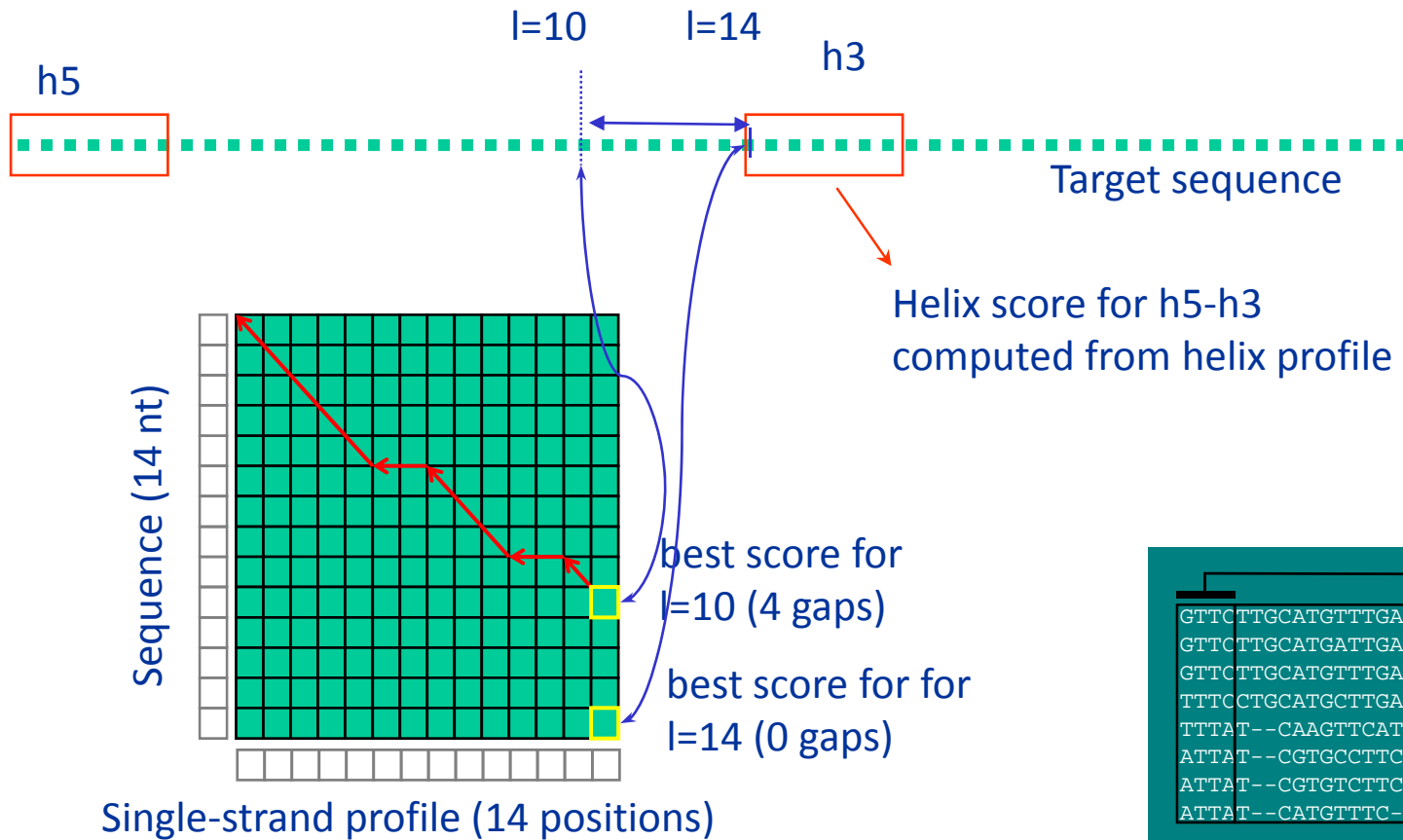
$$S_{b1,b2} = \log(F_{b1b2} / F_{b1}F_{b2})$$

A:A			
G:A			
C:A			
U:A			
A:G			
G:G			
C:G			
U:G			
...			

A			
G			
C			
U			
-			

Usual 5-row matrix for single strands

Profile Search Strategy



GTTC	TTGCATG	TTTGACG	GAAC
GTTC	TTGCATG	ATTGACG	GAAC
GTTC	TTGCATG	TTTGACG	GAAC
TTTC	CTGCATG	CTTGACG	GAAC
TTTAT	--CAAG	TTCAT-	ATAAA
ATTAT	--CGT	GCCTTC-	ATAAT
ATTAT	--CGT	GTCTTC-	ATAAT
ATTAT	--CATG	TTTC--	ATAAT

Training set

Un problème très fréquent dans toutes les approches par profils:

Un training set pauvre. Par ex Mir-133:

```
(( - (((((( - ----- ((( - ((( - ----- )))) )) - ----- ))))))) - ))
TC t GGCTGGT caaac- GGA a CCAA gtcggtcttcctgagaggt--- TTGG TCC CCTTCA ACCAGCT a CA
TG t GGCTGGT caaac- GGA a CCAA gtcaggtgtttctgtgaggt-- TTGG TCC CCTTCA ACCAGAC t AT
TG t GGCTGGT aaaac- GGA a CCAA gtcaggtgtttttgtgaggt-- TTGG TCC CCTTCA ACCAGCT a TG
TG c GGCTGGT gaaaa- GGA a CCAC atcaaccagaaaaaggat--- TTGG TCC CCTTCA ACCAGCC g CA
TA t GGCTGGT caaac- GGA a CCAA gtcggtcttccttagaggt--- TTGG TCC CCTTCA ACCAGCT a TT
AG t TGCTGGT aaaac- GGA a CCAA gtcgggtgtttgagagaggt-- TTGG TCC CTTTCA ACCAGCT a CT
TG t GGCTGGT caaat- GGA a CCAA gtcaggtgtttctgagaggt-- TTGG TCC CCTTCA ACCAGCT a CT
```

100% C:G

Other scores = $\log(\text{obs}/\text{expected})$ = arbitrary low value!

What about G:C or A:U in this column?

Is it as bad as C:C or A:G?

Pseudocounts

- Principle: fill columns with expected counts, based on a reasonable model
- Example: column c contains 7 C:Gs, we know C:G often substitutes for G:C, let's allow for *some* G:Cs.
- We need substitution matrices!

Pseudocomptes de Henikoff & Henikoff

(pour un alignement de protéines)

$$b_{ca} = B_c * \sum_{i=1}^{20} \text{Probability}(i|\text{column } c) * \text{Probability}(a|i)$$

Nb total de pseudocomptes en colonne c

a remplacé par i

Nb de a ajoutés en colonne c

Avec exemple précédent:

Colonne c 100% C:G

Probabilité(C:G)=1, autres = 0

Nb de A:Ts = $B_c * 1 * \text{Probabilité}(C:G | A:T)$

Nb de A:As = $B_c * 1 * \text{Probabilité}(C:G | A:A)$, etc.

RNA substitution matrices

Obtained from euk+archae+bac 16S/18S rRNA alignment

	AA	AT	AG	AC	TA	TT	TG	TC	GA	GT	GG	GC	CA	CT	CG	CC
AA	6.54e-04	5.20e-06	3.88e-05	4.22e-05	2.13e-05	5.51e-06	1.21e-05	3.84e-05	8.52e-05	1.28e-05	1.76e-04	2.89e-06	1.47e-05	6.47e-06	3.19e-06	4.69e-06
AT	7.96e-05	9.00e-04	5.19e-05	1.78e-04	1.69e-04	1.43e-04	8.85e-05	1.86e-04	4.15e-05	1.69e-04	1.22e-04	1.99e-04	8.73e-05	2.44e-04	1.25e-04	3.30e-04
AG	1.00e-04	8.72e-06	1.35e-03	1.27e-04	1.72e-05	5.09e-06	3.10e-05	1.38e-04	5.74e-05	1.59e-05	1.01e-04	8.22e-06	9.99e-06	1.62e-05	1.33e-05	2.56e-05
AC	4.11e-05	1.13e-05	4.81e-05	9.79e-04	2.79e-06	7.02e-06	2.79e-06	4.47e-05	4.93e-06	1.97e-05	3.05e-05	8.06e-06	5.40e-06	1.55e-05	2.47e-06	7.30e-05
TA	4.23e-04	2.19e-04	1.33e-04	5.69e-05	1.16e-03	2.21e-04	2.35e-04	2.78e-04	9.59e-05	1.18e-04	1.79e-04	1.08e-04	3.54e-04	2.04e-04	2.24e-04	9.28e-05
TT	1.05e-05	1.80e-05	3.80e-06	1.38e-05	2.14e-05	9.30e-04	2.57e-05	7.75e-05	5.79e-06	2.33e-05	4.87e-05	1.18e-05	1.57e-05	8.72e-05	1.83e-05	5.25e-04
TG	1.05e-04	5.03e-05	1.04e-04	2.49e-05	1.03e-04	1.16e-04	1.14e-03	1.80e-04	4.69e-05	4.56e-05	1.25e-04	4.26e-05	1.70e-04	2.15e-04	7.52e-05	3.23e-05
TC	1.45e-05	4.59e-06	2.03e-05	1.73e-05	5.30e-06	1.52e-05	7.82e-06	1.60e-04	4.55e-06	8.99e-06	4.77e-06	3.66e-06	9.00e-06	6.17e-05	2.95e-06	1.61e-05
GA	2.57e-04	8.19e-06	6.74e-05	1.53e-05	1.46e-05	9.11e-06	1.63e-05	3.64e-05	1.47e-03	2.50e-05	8.70e-05	2.12e-05	3.02e-05	2.83e-05	4.40e-06	8.02e-06
GT	1.24e-04	1.07e-04	6.02e-05	1.96e-04	5.81e-05	1.18e-04	5.10e-05	2.31e-04	8.04e-05	1.28e-03	9.39e-05	8.77e-05	2.53e-05	9.12e-05	3.55e-05	4.58e-05
GG	1.82e-04	8.24e-06	4.08e-05	3.24e-05	9.35e-06	2.61e-05	1.49e-05	1.30e-05	2.97e-05	9.98e-06	5.62e-04	6.96e-06	6.83e-06	8.80e-06	1.32e-05	1.06e-05
GC	1.14e-04	5.14e-04	1.26e-04	3.27e-04	2.16e-04	2.44e-04	1.94e-04	3.84e-04	2.78e-04	3.57e-04	2.67e-04	1.49e-03	1.07e-04	5.26e-04	2.57e-04	2.87e-04
CA	1.30e-05	5.04e-06	3.43e-06	4.90e-06	1.58e-05	7.22e-06	1.73e-05	2.10e-05	8.85e-06	2.30e-06	5.85e-06	2.40e-06	5.30e-04	5.30e-05	1.58e-05	7.68e-06
CT	3.86e-06	9.54e-06	3.78e-06	9.51e-06	6.16e-06	2.71e-05	1.48e-05	9.78e-05	5.60e-06	5.61e-06	5.10e-06	7.94e-06	3.58e-05	2.95e-04	5.22e-06	3.52e-05
CG	1.04e-04	2.68e-04	1.70e-04	8.32e-05	3.71e-04	3.12e-04	2.83e-04	2.56e-04	4.77e-05	1.19e-04	4.21e-04	2.12e-04	5.86e-04	2.86e-04	1.35e-03	2.50e-04
CC	2.13e-06	9.81e-06	4.54e-06	3.41e-05	2.12e-06	1.24e-04	1.69e-06	1.94e-05	1.21e-06	2.14e-06	4.70e-06	3.31e-06	3.95e-06	2.68e-05	3.48e-06	5.45e-04

	A	T	G	C
A	9.13e-04	8.22e-05	1.05e-04	9.35e-05
T	5.57e-05	6.70e-04	7.98e-05	1.41e-04
G	6.94e-05	7.78e-05	7.32e-04	5.03e-05
C	4.09e-05	9.15e-05	3.33e-05	6.03e-04

Une E-value pour les motifs ARN?

- ★ E-value:

- ★ Expectation value

- # Nb de hits attendus de score $> S$, par hasard

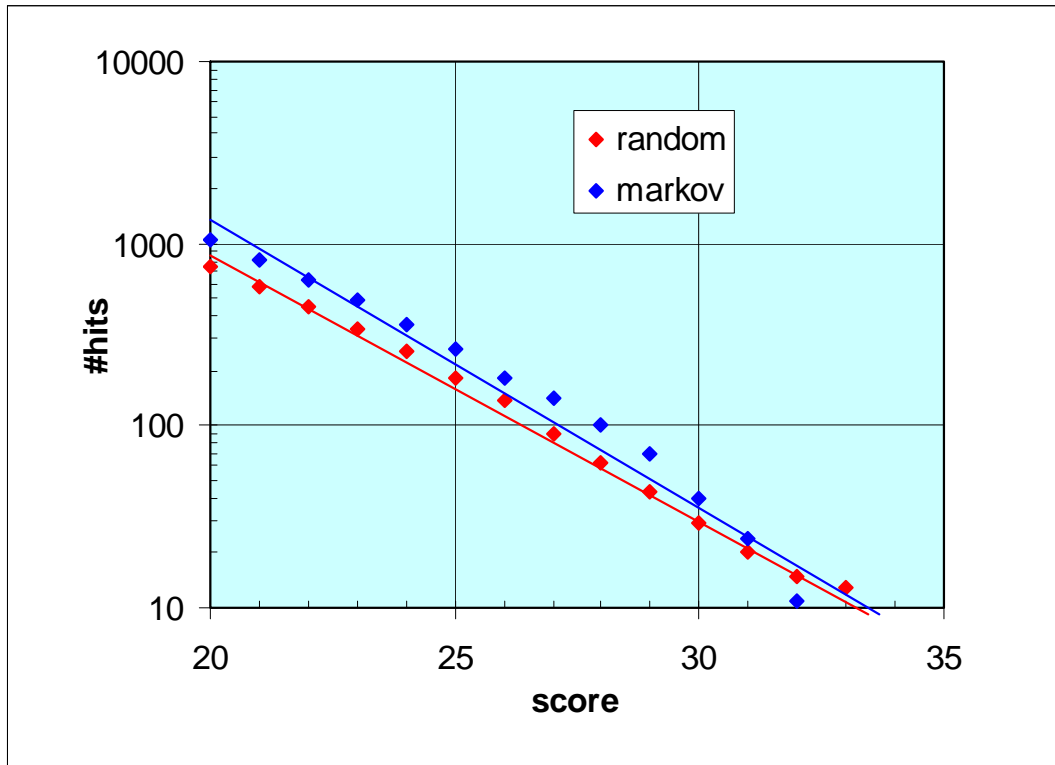
- Peut-on l'estimer?

- ★ Idée 1:

- ★ Chercher motif dans base de données aléatoire et calculer la distribution des scores

- ★ Extrapoler pour tout score

Les hauts scores se comportent bien



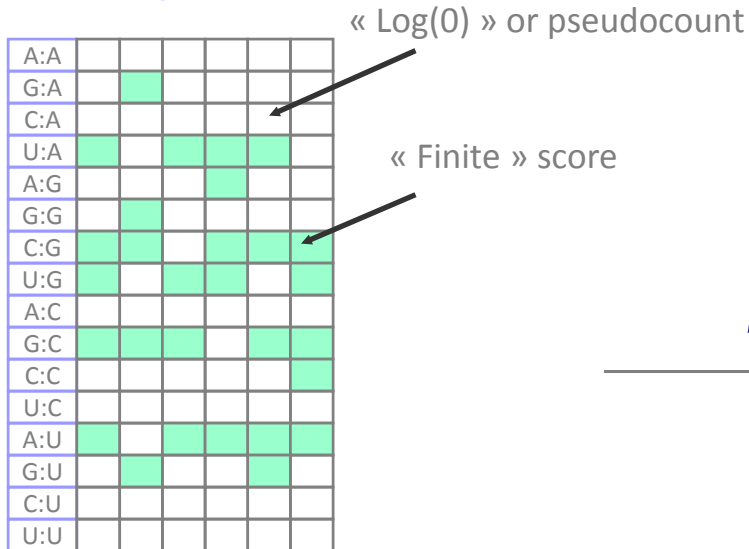
SECIS hits in
700 mb randomized sequences

- Score 30:
• 3.8 hits/100mb
- Score 40:
• 0.13 hits/100mb
- $\#hits = 6.6 \cdot 10^5 e^{-0.3 s}$
- $E = Kmn e^{-\lambda s}$
- (extreme value distrib.)

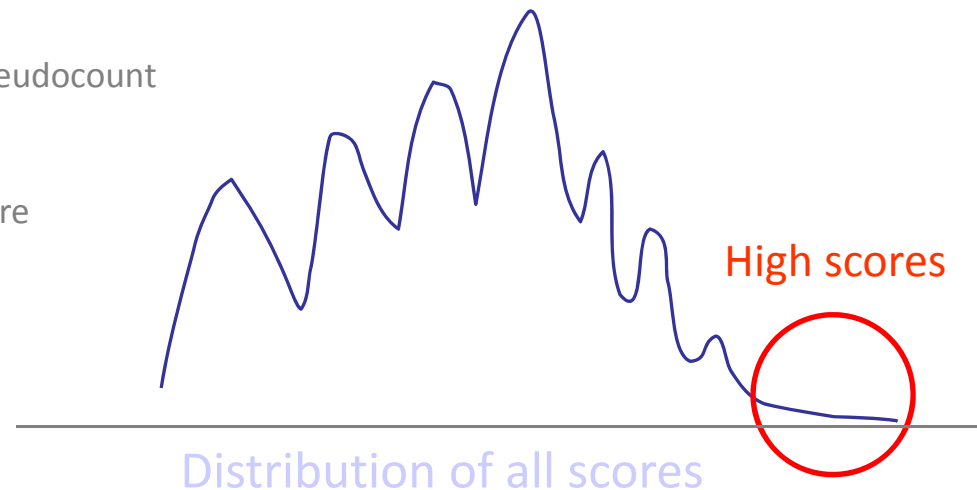
- An E-value is possible
- 1 day to run simulation!

La distribution globale n'est pas si régulière...

Helix profile



(same for single-strand profile)



- Not Gaussian
- How can we model it?

Produit de convolutions discrètes

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



computation
simulation

Fig 3a: polyA strands 11 4 5 ; 10Mb data

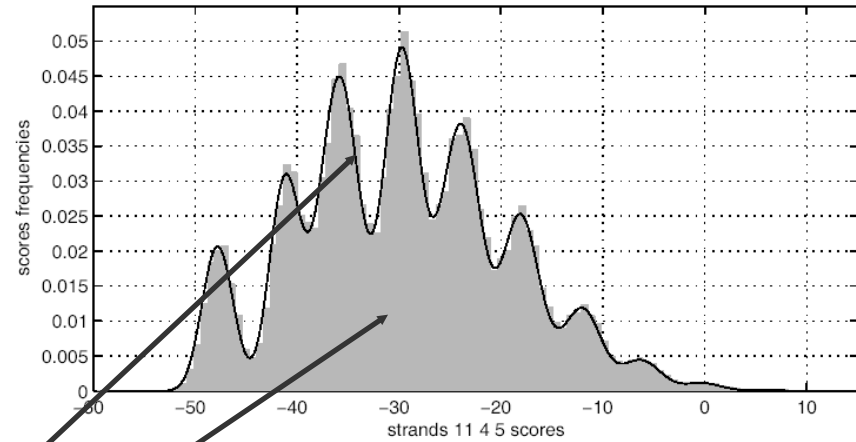
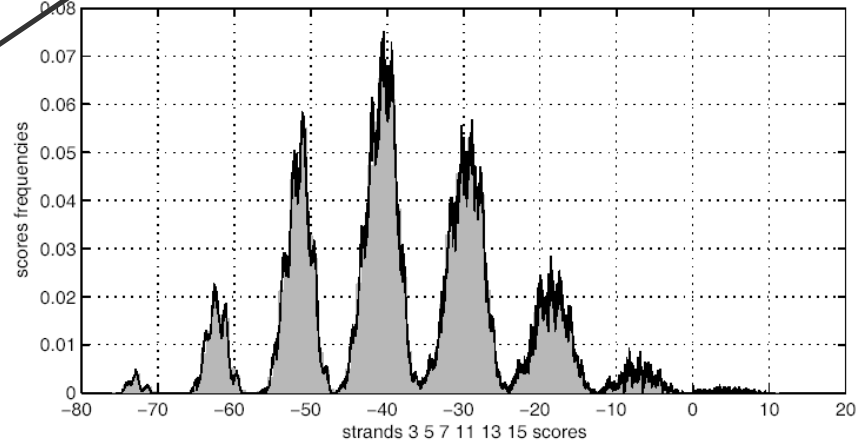
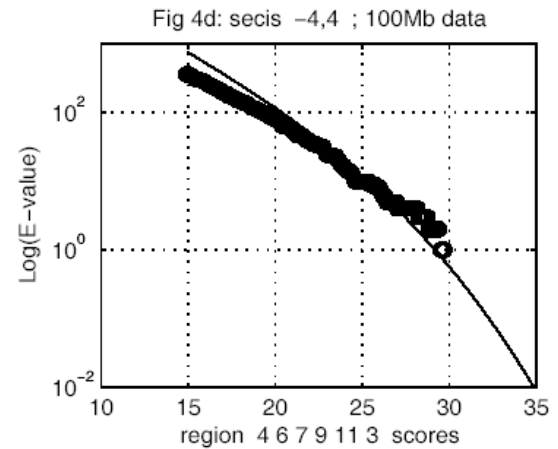
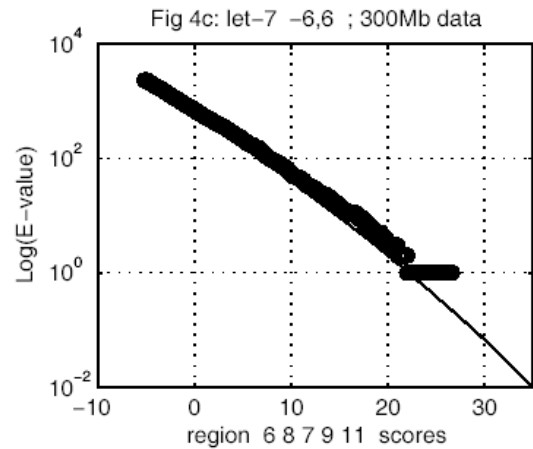
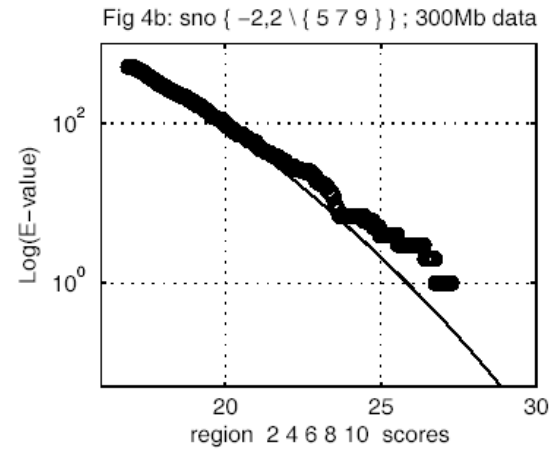
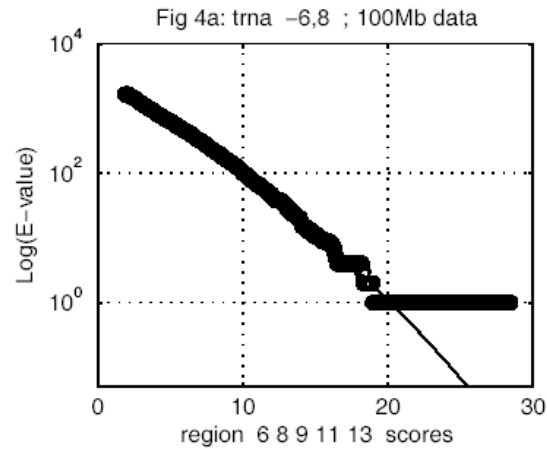


Fig 3b: let-7 strands 3 5 7 11 13 15 ; 10Mb data



E-values de motifs complets

○ simulated
— computed



Recherche par profils

PROS	CONS
All constraints in the training set are efficiently exploited , resulting in highly specific detections	Alignment and secondary structure constraints must be accurate
No 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	

Réussir une recherche d'ARNnc

★ Rnamotif, Pastcan, Palingol, InfeRNAI, Erpin...

Organiser l'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 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 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
				⋮			⋮				⋮			
Archaea	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 eeler	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 U G	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.

Evaluer les motifs trouvés: soit E-value, soit procédure de contrôle

Contrôle de spécificité:

TP: true positive

FP: false positive

$$\text{Spécificité : SP} = \frac{\text{TP}}{\text{TP+FP}} \quad \text{Nb total prédictions}$$

TP et **FP** difficiles à obtenir! Comment savoir qu'un hit est vrai/faux?

Truc: exprimer SP en **FP / Mb dans une séquence aléatoire**

Prendre grande séquence / même composition (mono & di-nt) que la base de recherche (p. ex. avec programme *shuffle*)

Sensibilité

TP: true positive
FN: false negative

Sensibilité: SN = $\frac{TP}{TP+FN}$ — Nb total objets « vrais »

Ici: TP & FN: facile à obtenir sur training set (*retirer la moitié du tr-set et lancer la recherche sur l'autre moitié*)