Les petits ARN régulateurs

Apports de la génomique

Daniel Gautheret

V.2008.0 http://www.esil.univ-mrs.fr/~dgaut/Cours

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

- ARNt
- ARNr
- ARNsn (spliceosome)
- Autres ARNs
  - 4.5S, 10Sa, Spot42, DicF, MicF, OxyS, DsrA, 6S (procaryotes) produits des gènes XIST, H19, IPW, 7H4, His-1, NTT (mammifères), microARN (animaux et plantes), snoARN (partout)

ARNr 18S et 28S: promoteurs pol-I.

ARNt et ARNr 5S: promoteurs pol-III

Autres ARNnc: pol-I, II ou III

Les ARN régulateurs

- Procaryotes
- Eucaryotes

ARN régulateurs eucaryotes

L'interférence ARN

Medicine

The Nobel Prize in Physiology or Medicine 2006

"for their discovery of RNA interference - gene silencing by double-stranded RNA"

Photo: L. Cicero/Stanford

Andrew Z. Fire

1/2 of the prize

Photo: R. Carlin/UMMAS

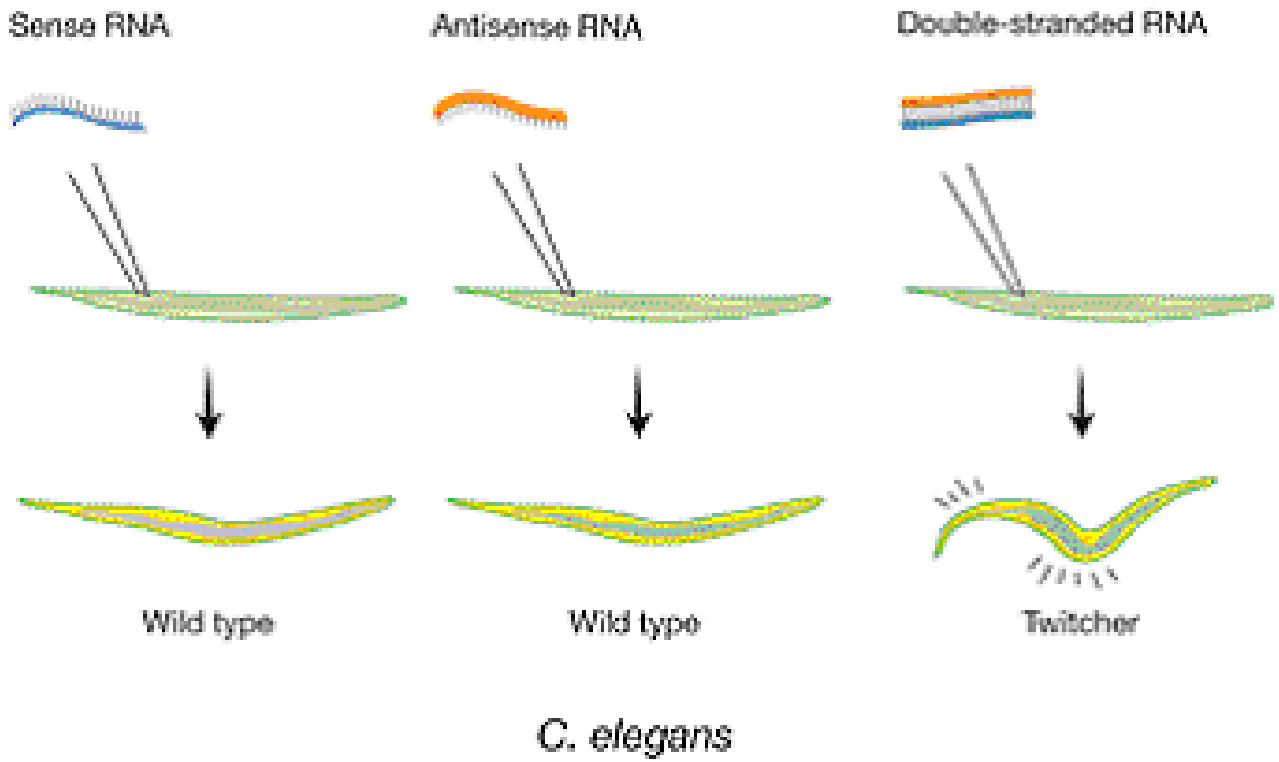
Craig C. Mello

1/2 of the prize

Comment ça marche?

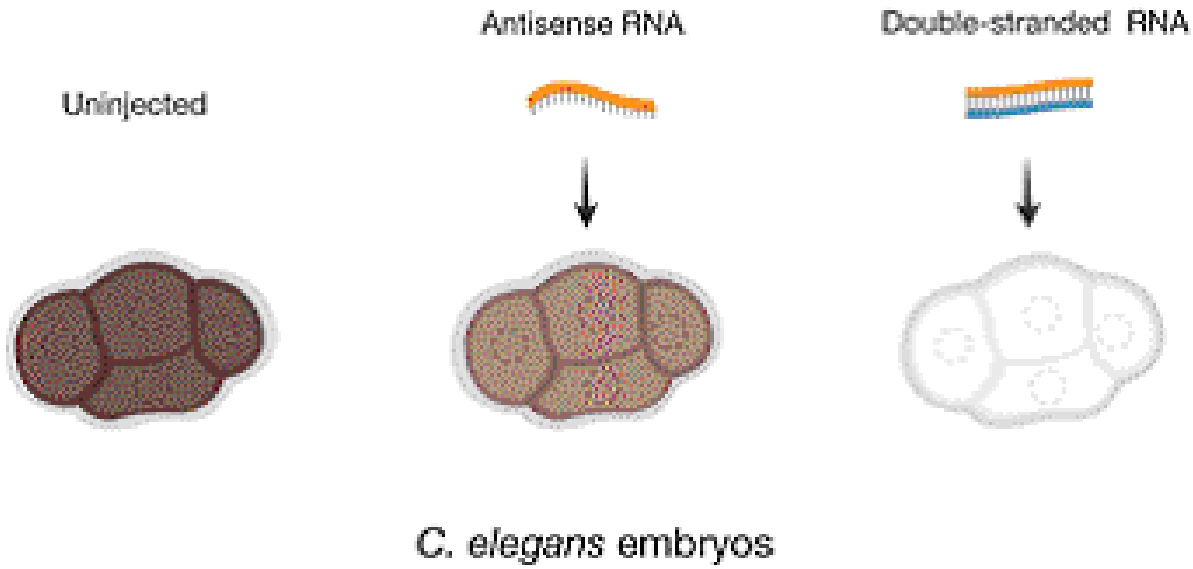
- Début années 90: ARN antisens
- Effets modestes et parfois incohérents

Article de Fire & Mello 1998



The *unc-22* gene encodes a myofilament protein. Decrease in *unc-22* activity is known to produce severe twitching movements (convulsions). Injected double-stranded RNA, but not single-stranded RNA, induced the twitching phenotype in the progeny.

Image: Bertil Daneholt. http://nobelprize.org/nobel_prizes/medicine/laureates/2006/adv.html



Injection of single-stranded or double-stranded *mex-3* RNA into the gonad of *C. elegans.*
The extent of brown colour reflects the amount of *mex-3* mRNA present. *mex-3* mRNA is abundant in the gonads and early embryos. The mRNA was lost after injection of double-stranded RNA, while injection of antisense RNA only reduced the content of mRNA to some extent.

Image: Bertil Daneholt. http://nobelprize.org/nobel_prizes/medicine/laureates/2006/adv.html

Observations importantes

- Le silencing ne fonctionne qu'avec une séquence de cet ARNm
- La séquence doit provenir d'ARN mature
  - Post-transcription / cytoplasmique
- L'ARNm ciblé semble être dégradé
- Quelques molécules d'ADNds suffisent
  - Soit amplification, soit catalyse
- L'effet peut se transmettre entre tissus, voire à la descendance
  - Transmission

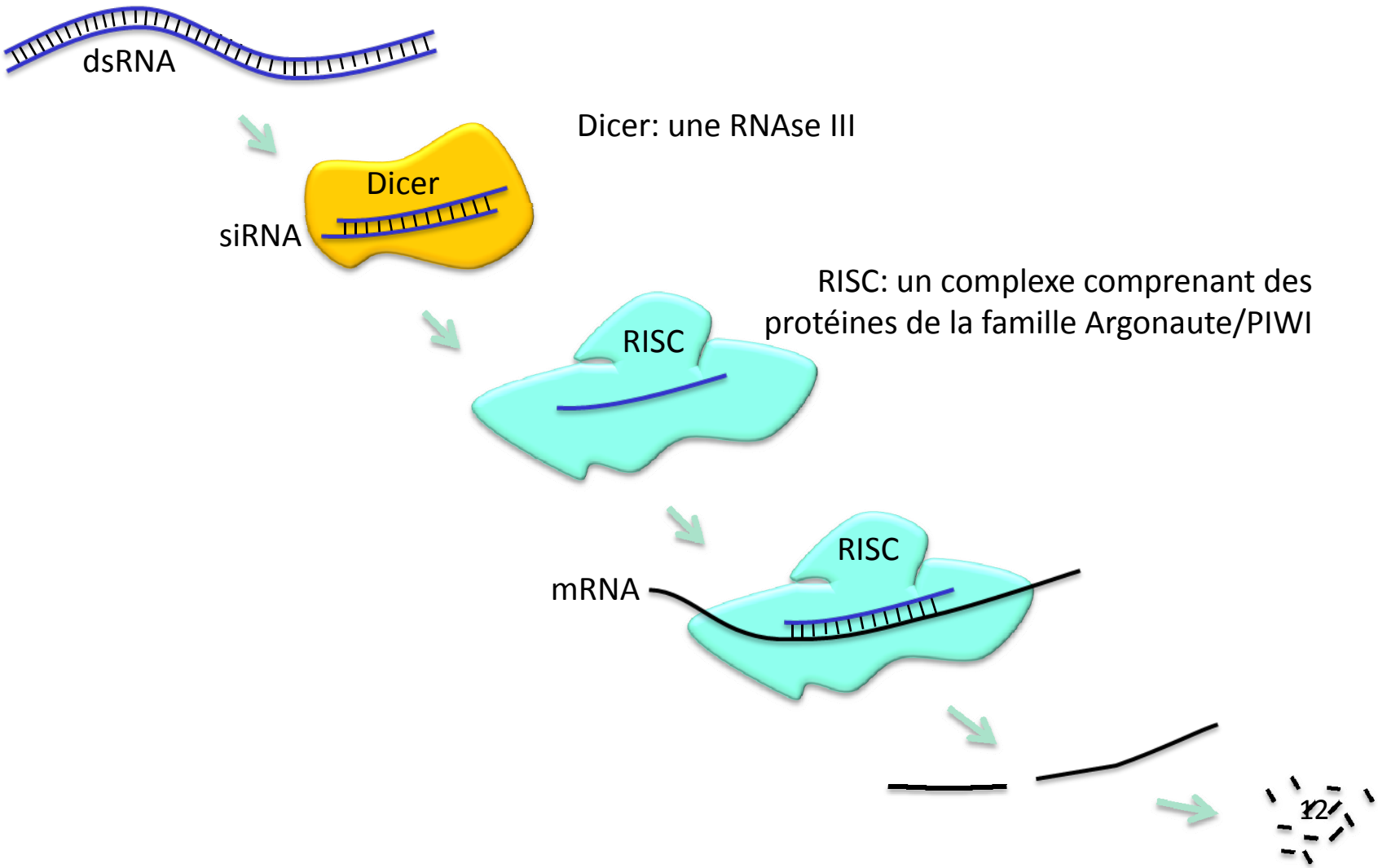
Généralisation

- Animaux (métazoaires)
  - Mammifères: seulement avec ARN de 21nt
- Plantes
- Tous les eucaryotes sauf les levures

1999: le mécanisme

- Présence de petits ARN de 20-25nt contenant les deux brins: siRNA (small interfering)
- Ces ARN sont produits par clivage de l'ARNds
- Ce sont les petits ARN qui interagissent avec l'ARNm

2000: découverte de la machinerie



Dicer: une RNAse III

RISC: un complexe comprenant des protéines de la famille Argonaute/PIWI

Importance de la découverte

- Protection contre les virus
- Silencing des éléments mobiles
- Maintient de l'état condensé de la chromatine
- Un nouvel outil pour réprimer spécifiquement les gènes
- Répression de la synthèse protéique et régulation du développement
  - Des siRNA naturels: les miRNAs

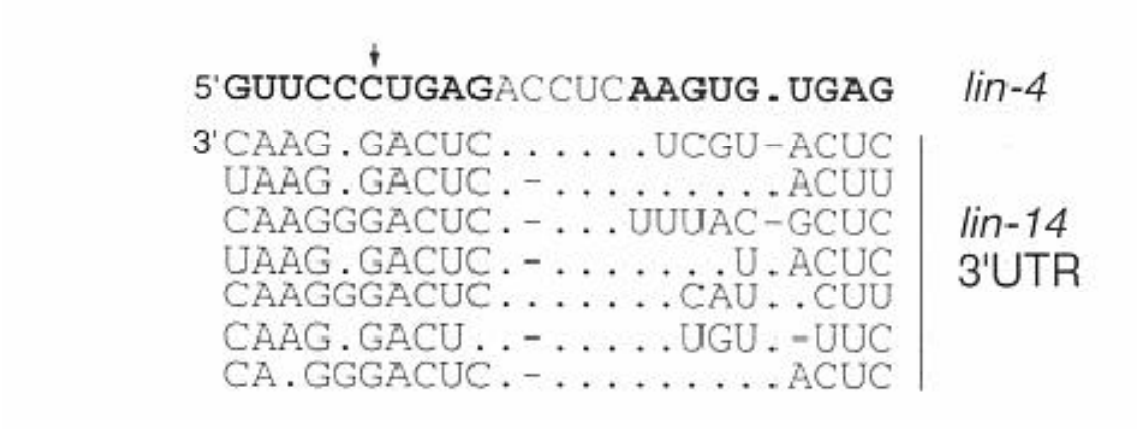
1993: découverte du premier microRNA

(Laboratoire de Victor Ambros)

- 4 stades larvaires: L1 à L4
- Ambros isole un mutant qui « réitère » le stade L1
- Ce mutant présente une délétion du gène lin-4
- Lorsqu'on insère dans un animal transgénique un fragment d'ADN contenant le gène lin-4, on retrouve le phénotype normal
- Curieusement lin-4 ne code pas pour une protéine mais un ARN de 22n

Répression par reconnaissance de l'UTR 3'

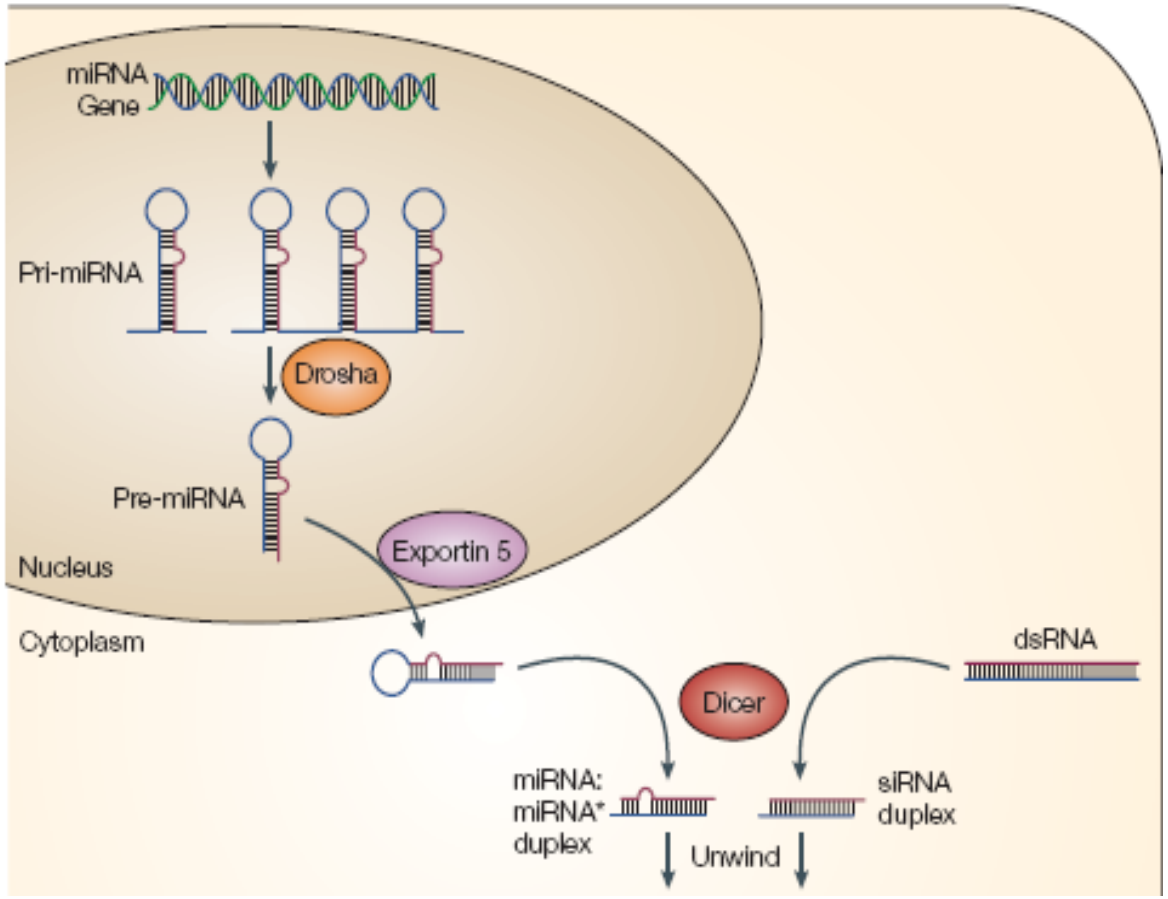
- Plusieurs fragments de l'UTR 3' de lin-14 sont complémentaires à lin-4



- En mutant lin-4 ou l'UTR, on est capable d'abolir ou restaurer la répression du gène lin-14

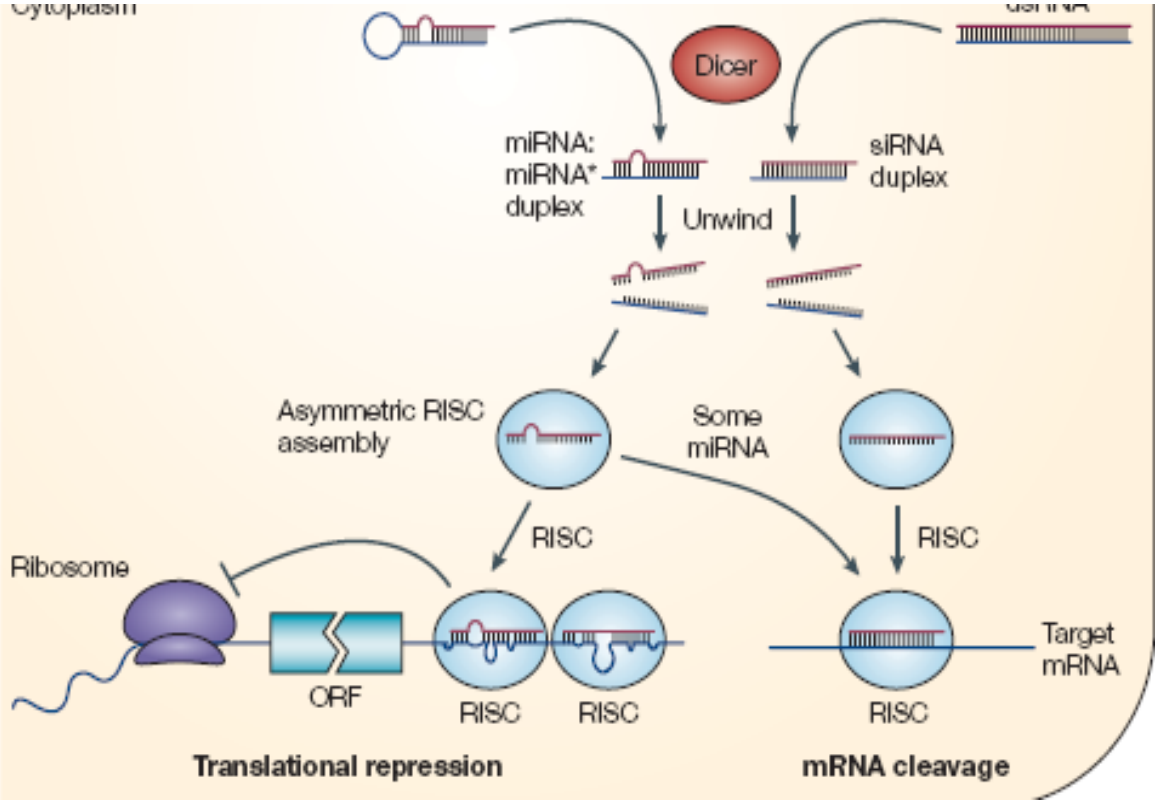
La biogénèse des miRNA: la jonction entre miRNA et siRNA

- Deux Rnase III nécessaires
  - Drosha
  - Dicer
- Les deux enzymes laissent un bout 3' libre de 2nt



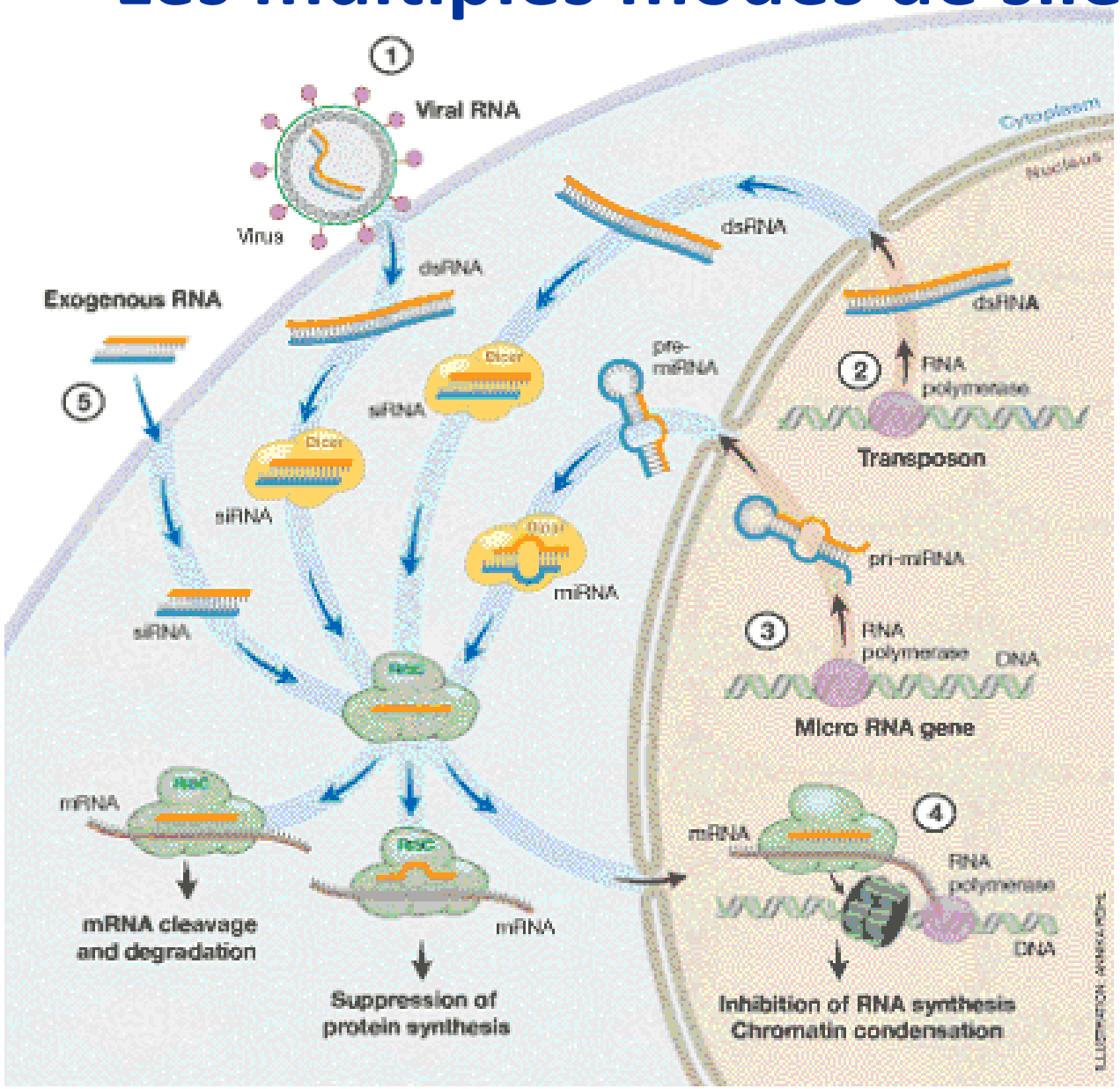
(He & Hannon, Nature reviews, 2004)

Action des miRNA

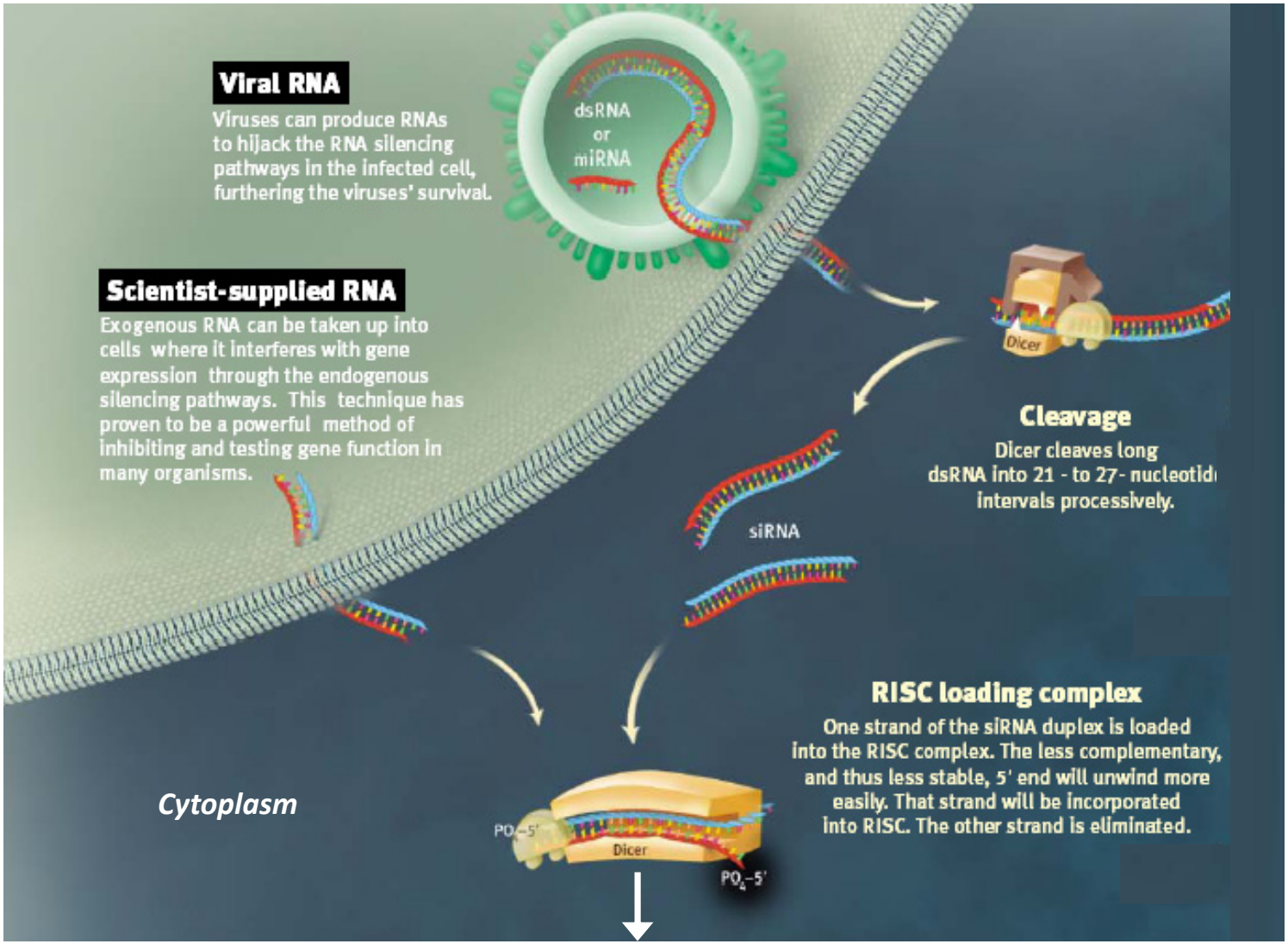


(He & Hannon, Nature reviews, 2004)

Les multiples modes de silencing



La voie d'extinction des ARN: I
(RNA silencing pathway)



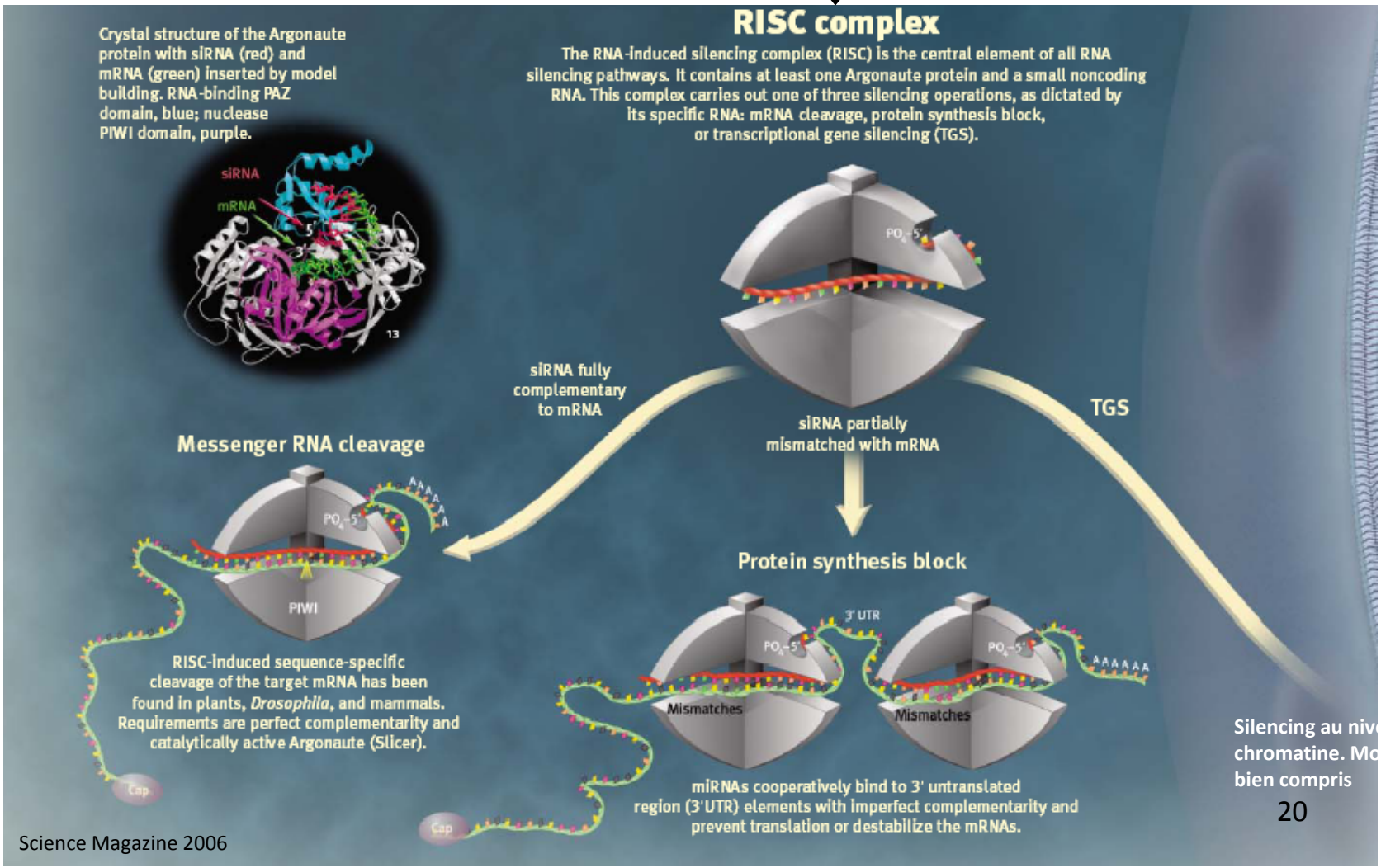
RISC complex

Science Magazine 2006

La voie d'extinction des ARN. II

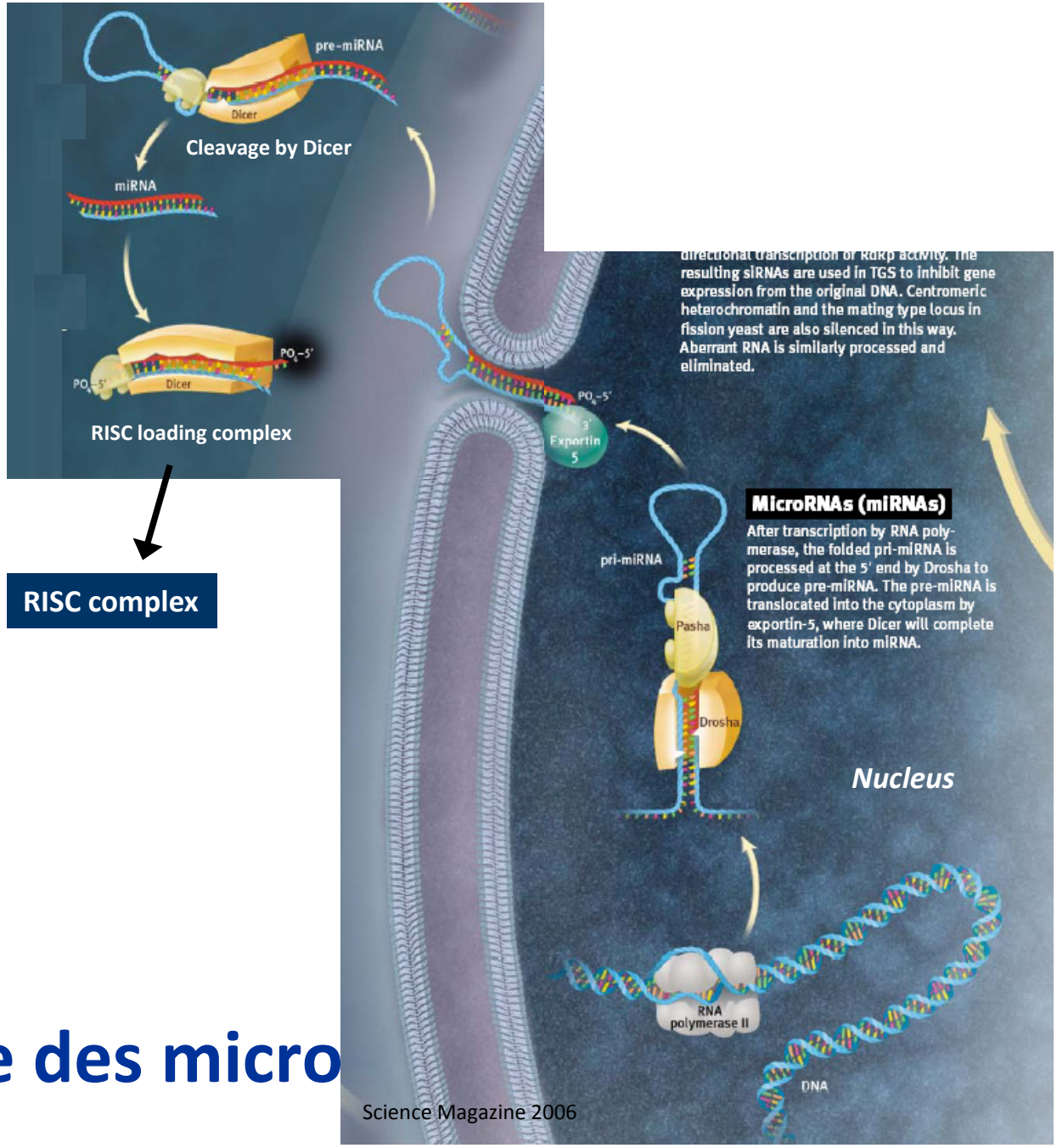
Le complexe RISC vise l'ARNm

RISC loading complex



Silencing au niv
chromatine. Mo
bien compris

Science Magazine 2006



La voie des micro

Science Magazine 2006

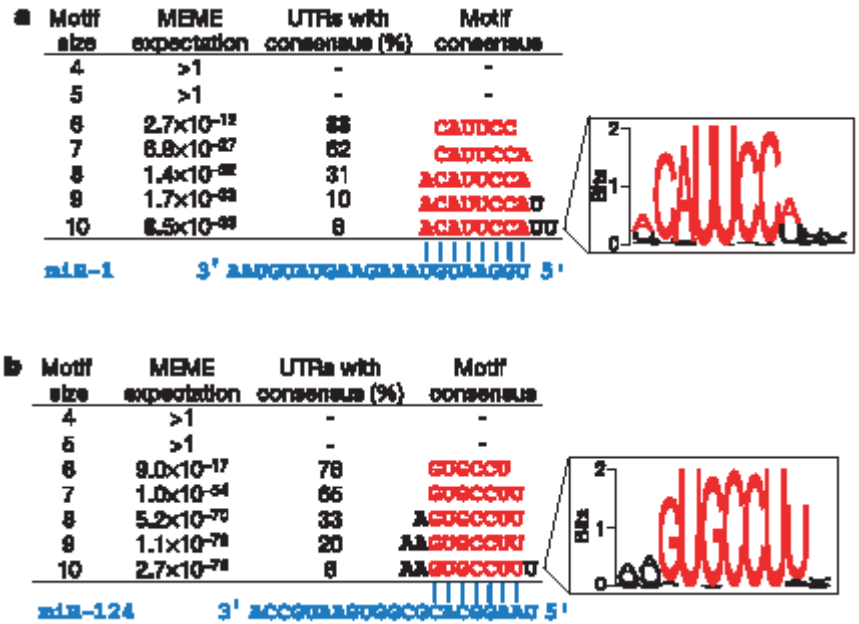
Un miRNA peut réprimer des centaines de transcrits

Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs

Lee P. Lim[1], Nelson C. Lau[2], Philip Garrett-Engele[1], Andrew Grimson[2], Janell M. Schelter[1], John Castle[1], David P. Bartel[2], Peter S. Linsley[1] & Jason M. Johnson[1]

- Injection de miR-1 (coeur + muscle) et miR-124 (cerveau) dans des cellules humaines
- Suivi de l'expression par puce ADN
- ~200 gènes réprimés
- Le miRNA de cerveau réprime des gènes qui <u>ne doivent pas s'exprimer</u> dans le cerveau en temps normal.
- IDEM pour miRNA de coeur.

Les gènes réprimés ont un motif commun dans leur 3’ UTR



Séquence de 7-8nt « seed » essentielle pour la reconnaissance de la cible

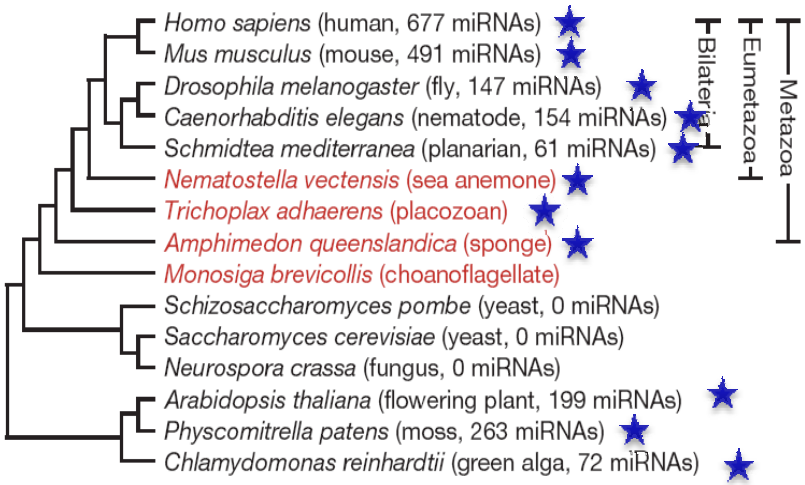
Les nouvelles familles d'ARN interférents naturels

- piRNA
  - PIWI-associated RNA
  - Protège le génome des transposons
  - Comprend: rasiRNA (repeat-associated RNAs)

- Utilisent différentes protéines Dicer et Argonaute (RISC).

Machinerie miRNA chez les eucaryotes

Grimson et al. Nature 455. 2008.



Dicer & argonaute proteins

Table 1 | The small-RNA machinery of representative eukaryotes

Species	Ago	Piwi	Dicer	Drosha	Pasha	Hen1
Homo sapiens	4	4	1	1	1	1
Drosophila melanogaster	2	3	2	1	1	1
*Caenorhabditis elegans**	5	3	1	1	1	1
Nematostella vectensis†	3	3	2	1	1	1
Trichoplax adhaerens†	1	0‡	5	1	0§	0‡
Amphimedon queenslandica†	2	3	4	1	1	2
Monosiga brevicollis	0‡	0‡	0‡	0	0	0‡
Saccharomyces cerevisiae	0‡	0‡	0‡	0	0	0‡
Schizosaccharomyces pombe‖	1	0‡	1	0	0	0‡
Arabidopsis thaliana	10	0‡	4	0	0	2
Physcomitrella patens	6	0‡	5	0	0	1
Chlamydomonas reinhardtii	2	0‡	3	0	0	1

ARN régulateurs bactériens

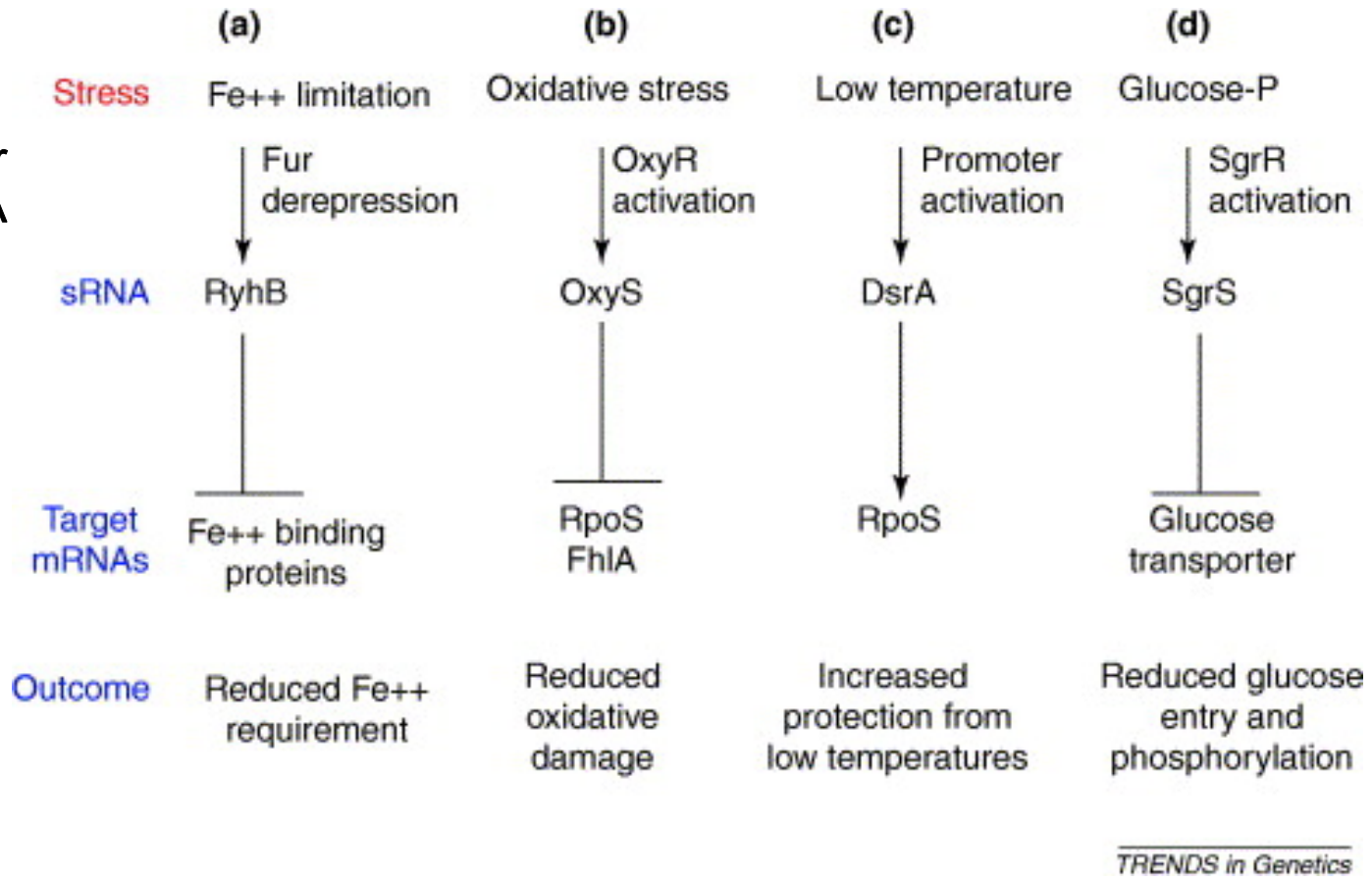
- Début des années 80:
  - Chez *Escherichia coli, de petits ARN* (~100nt) peuvent se lier à des séquences complémentaires et inhiber la traduction
  - Aujourd'hui, environ 25 cas connus d'ARN antisens régulateurs chez *E. coli*

Les sRNA: la principale famille d'ARN régulateurs chez les bactéries

- 80-100nt de long
- Pas de processing/clivage
- Interaction avec protéine chaperone Hfq
- Appariement aux ARNm
- Affectent capacité de traduction ou stabilité de l'ARNm

Les sRNA chez Coli sont souvent impliqués dans la réponse au Stress

Déclencheur synthèse sRNA



Les cibles sont souvent des régulateurs de transcription

La chaperone Hfq est nécessaire à l'action des sRNA

- Intéragit avec régions riches en AU sur le sRNA et sur la cible
  - Stabilise le sRNA
  - Stimule appariement entre sRNA et cible

Reconnaissance sRNA/mRNA

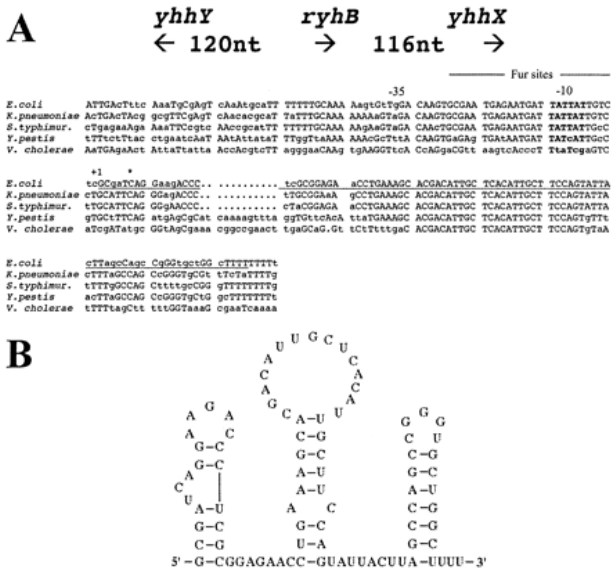
- Le plus souvent en 5’
  - Occlusion du site de fixation du ribosome, du codon Start, du départ de transcription
- Dégradation par RNAse E
  - Est-elle une conséquence de l’appariement?
  - Ou une conséquence de l’arrêt de traduction?
  - RNAse E dégrade aussi le sRNA.

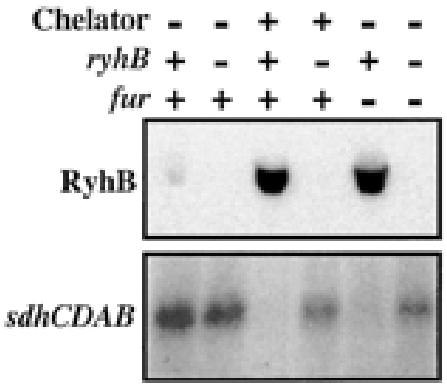
Ryhb et la synthèse de protéines liant le fer

- Répresseur: Fur (Ferric uptake regulator). Reconnaît le promoteur de Ryhb.
  - En présence de fer: FUR activé. Réprime Ryhb
  - Fer limitant: FUR inactif. Synthèse de Ryhb
- Ryhb se lie avec les mRNA de 5 opérons de protéines liant le fer
- Conduit à une rapide dégradation des mRNA
- -> Diminution des besoins cellulaires en fer -> redirection des ressources en fer vers protéines essentielles

Caractérisation de Ryhb (2002)

- Genome-wide search (bioinformatique)
- 26 sRNA potentiels
- Complementarités sRNA:mRNA
  - Plusieurs mRNA utilisant le fer
- Vérif: Niveau de sdh inversement proportionnel à Rhyb
- Caractérisation au moyen de souches Rhyb-; fur-; avec ou sans fer (iron chelator+)

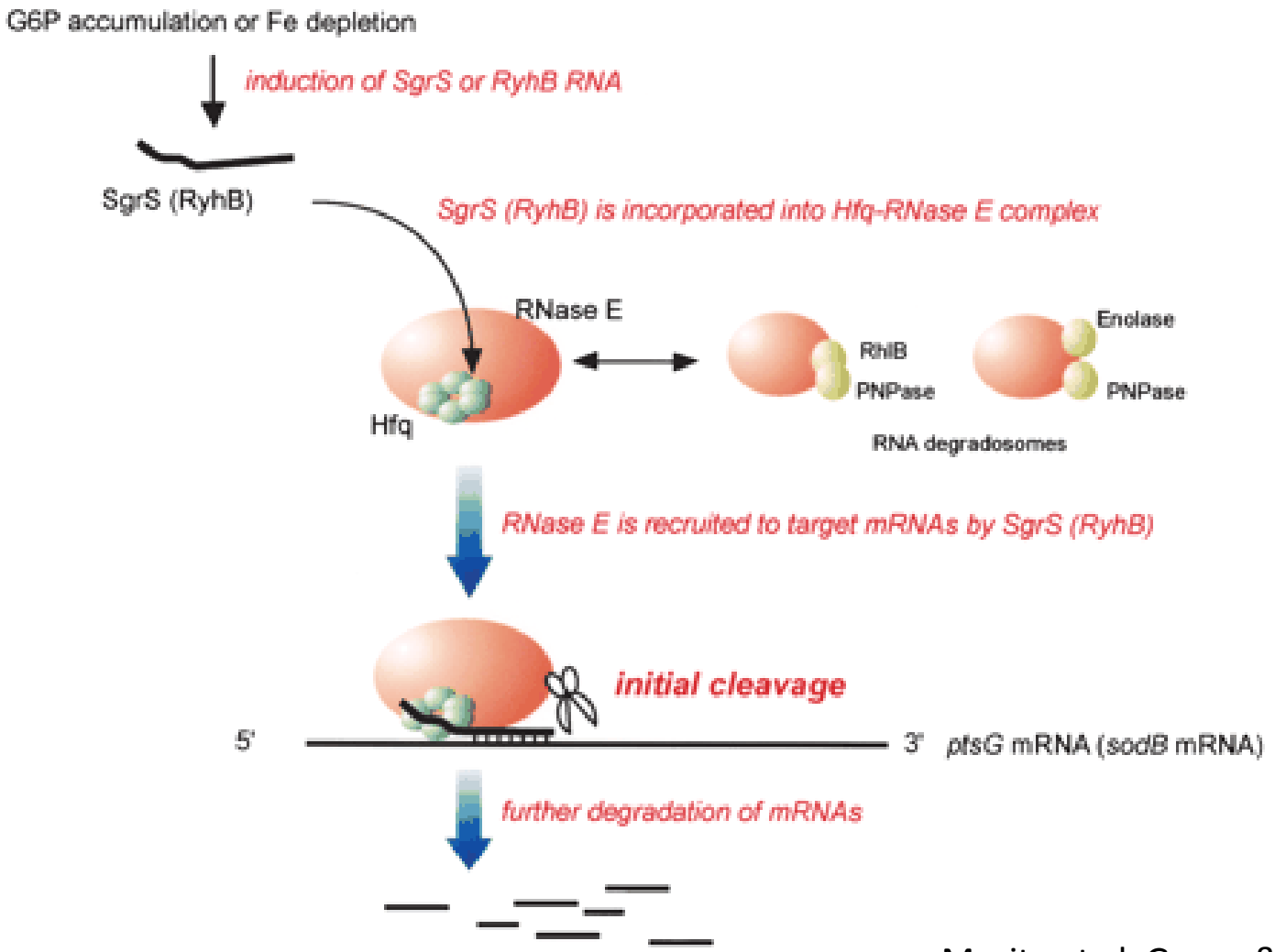




sdhCDAB: opéron succinate dehydrogenase C,D,A,B

Eric and Gottesman (2002) Proc. Natl. Acad. Sci. USA 99, 4620-4625

Action concertée de Hfq et Rnase E

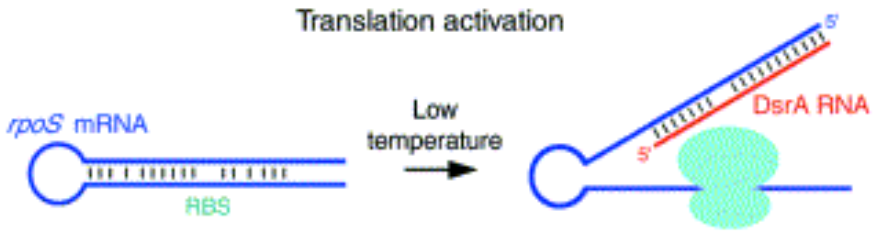


Morita et al. Genes & Dev. 2005

Régulation positive de RpoS par *DsrA/RprA*

- RpoS est un facteur Sigma alternatif très important pour l'expression de nombreux gènes en conditions de stress.
- Etat normal: éteint. La région 5' est repliée sur elle-même.
- Cible située ~70nt en amont du début de traduction
- DsrA se fixe sur la cible et libère le site d'entrée du ribosome (RBS)
- Le mode d'action de RprA est encore incertain

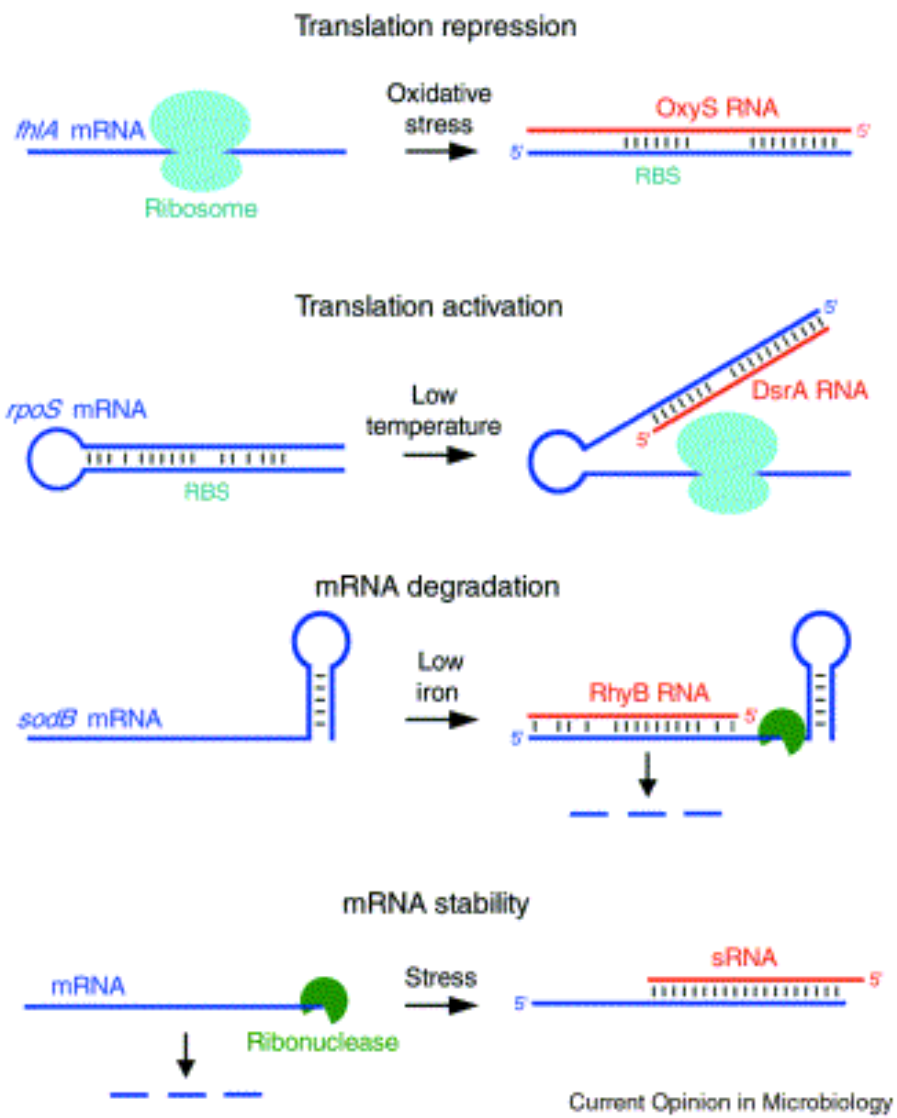
- Régulation positive par ARN: jamais vu chez les eucaryotes



Inhibition de RpoS par OxyS

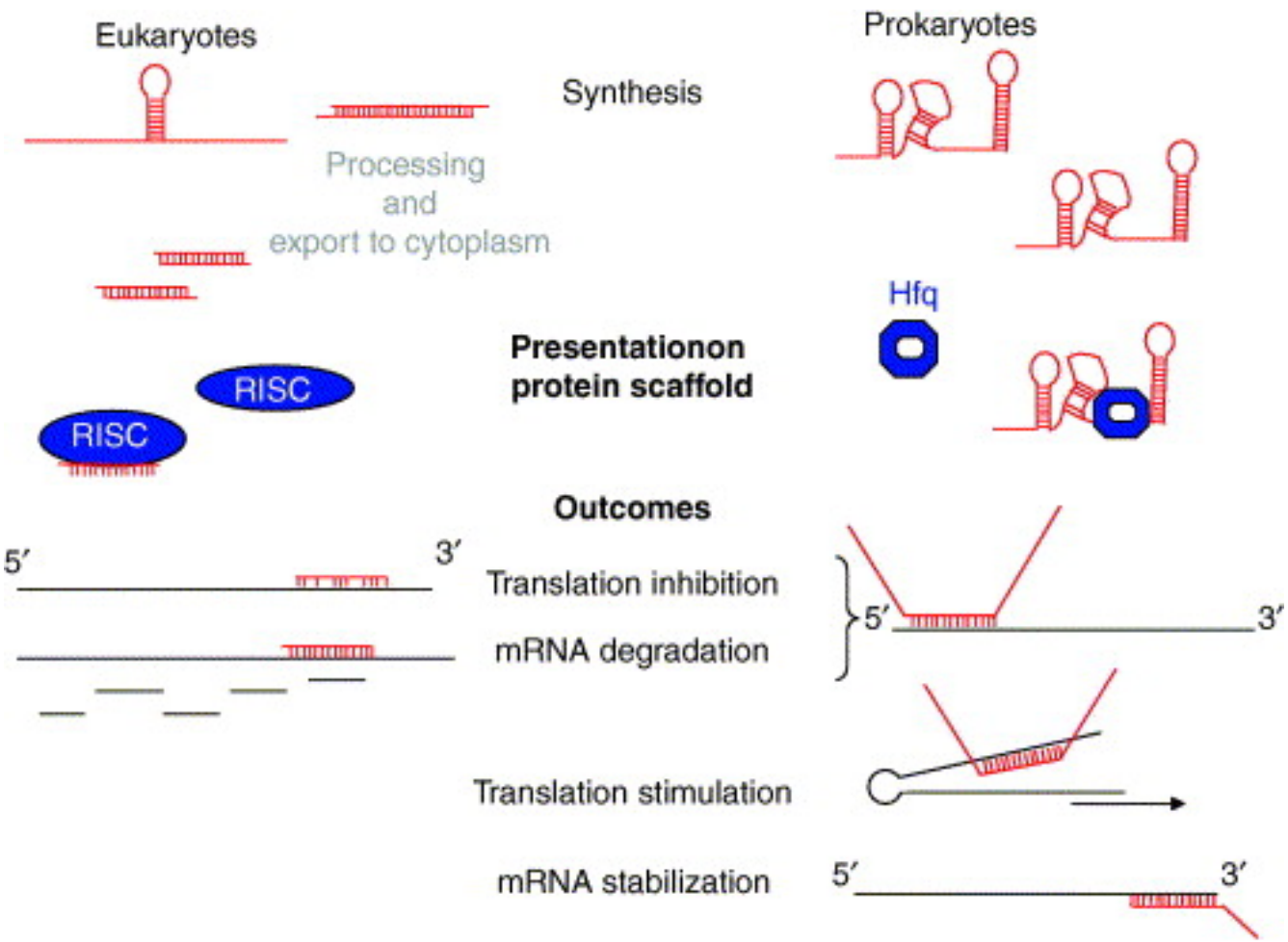
- En cas de stress oxydatif fort, la bactérie « éteint » RpoS
  - Induction du régulon OxyR comprenant un autre sRNA: OxyS
  - OxyS inhibe la traduction de RpoS, probablement par titration de l'Hfq disponible

Conclusion: une variété de mécanismes



(possible!)

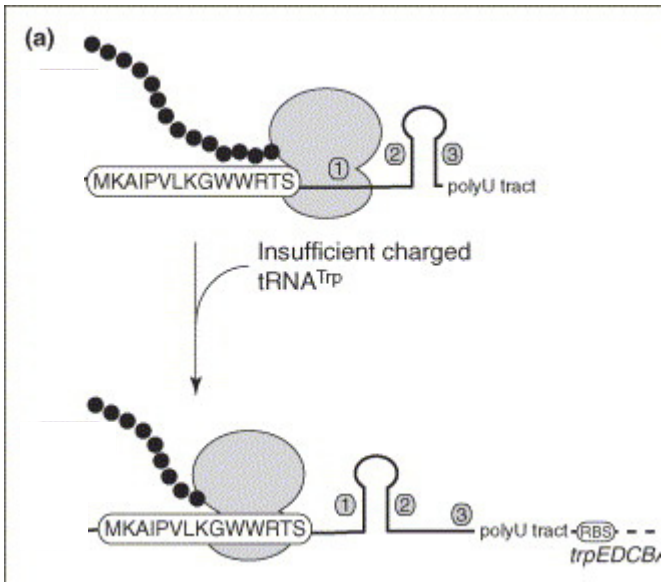
Eucaryotes et procaryotes: quelques similitudes



S. Gottesman TRENDS in Genetics

Les régulations par les ARN <u>en *cis*</u>

- ARN présent sur l'ARNm lui même
- Des « capteurs » (sensors) de métabolites
- Un cas classique, l'atténuation par le tryptophane:

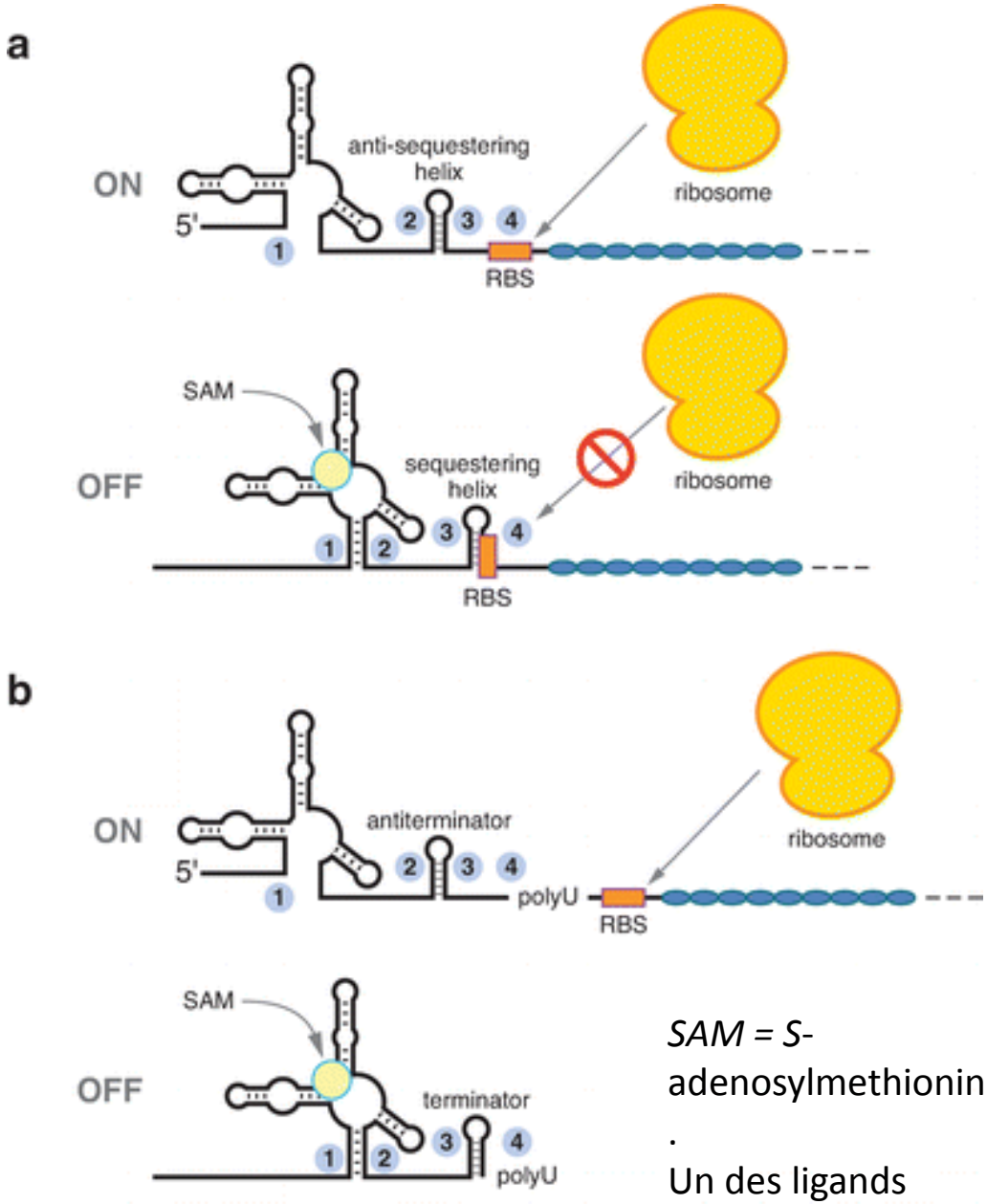


Les riboswitches

- Dans la région 5’ des gènes
- Deux conformations
- Se lient directement aux métabolites ou par l’intermédiaire de protéines ou d’ARN
- Blocage transcription ou traduction

Blocage traduction

Blocage transcription



Winkler WC, Breaker RR. 2005.
Annu. Rev. Microbiol. 59:487–517

*SAM = S-*adenosylmethionine

.

Un des ligands reconnus par les riboswitches

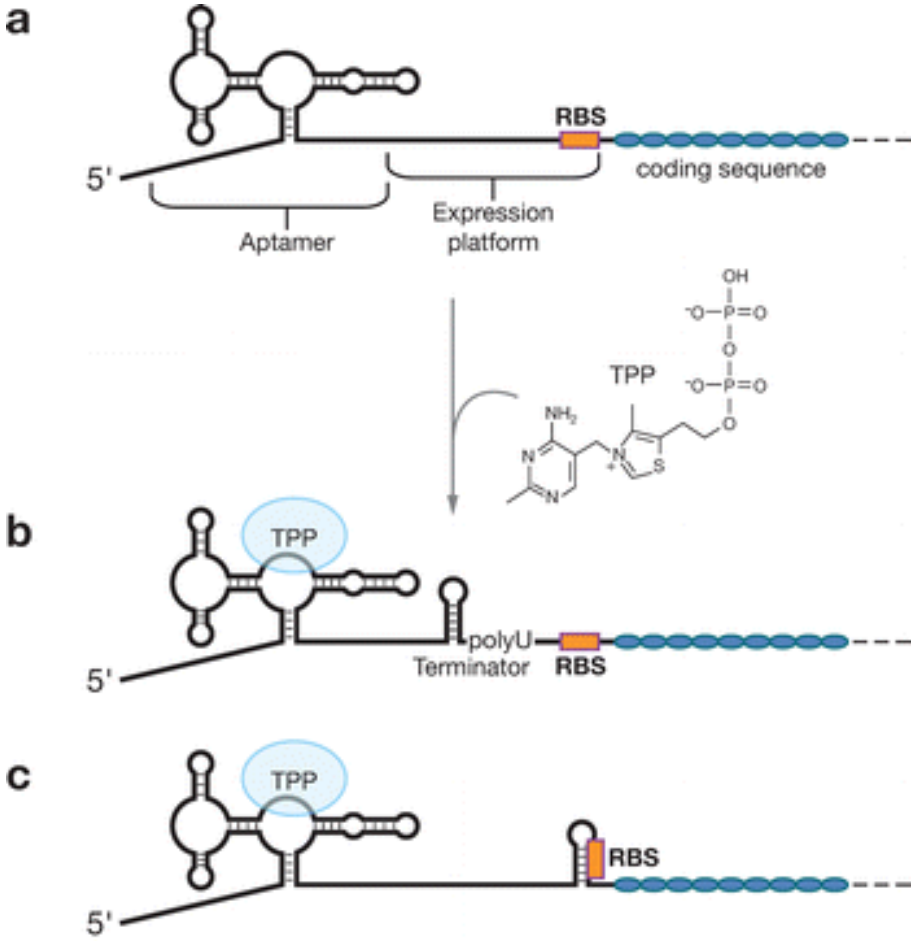
Reconnaissance directe du ligand par le riboswitch

- Coenzyme B_{12}
- TPP (dérivé de Thiamine)
- FMN (flavin mononucleotide)
- SAM (*S*-adenosylmethionine)
- Lysine
- Guanine
- Adenine
- GlcN6P (glucosamine-6-phosphate)
- Glycine

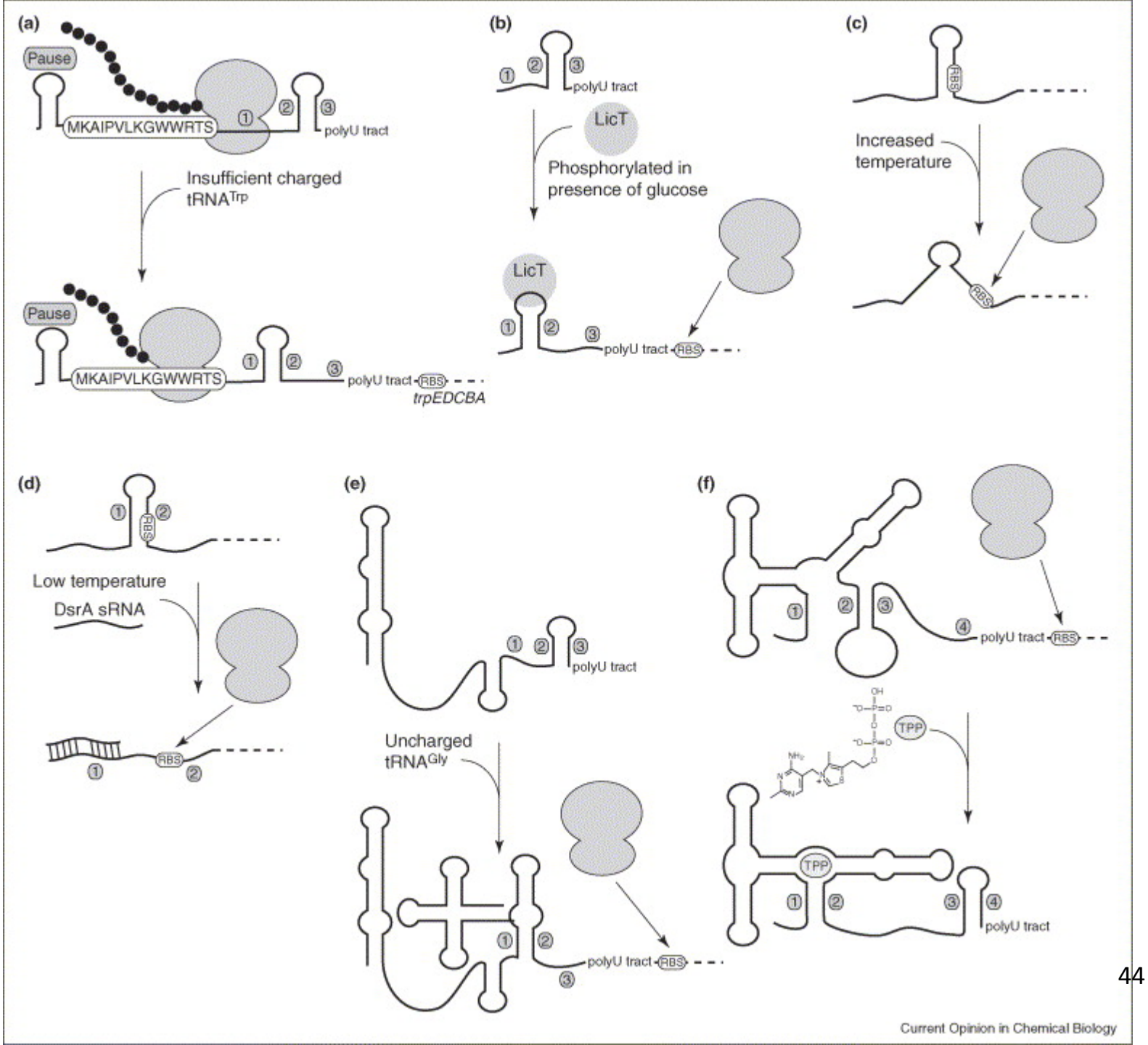
Exemple des riboswitches capteurs de TPP

- Aptamère: structure d'ARN capable de reconnaître un ligand
- Après fixation du ligand: changement de conformation

- (Ici deux mécanismes différents selon espèces)



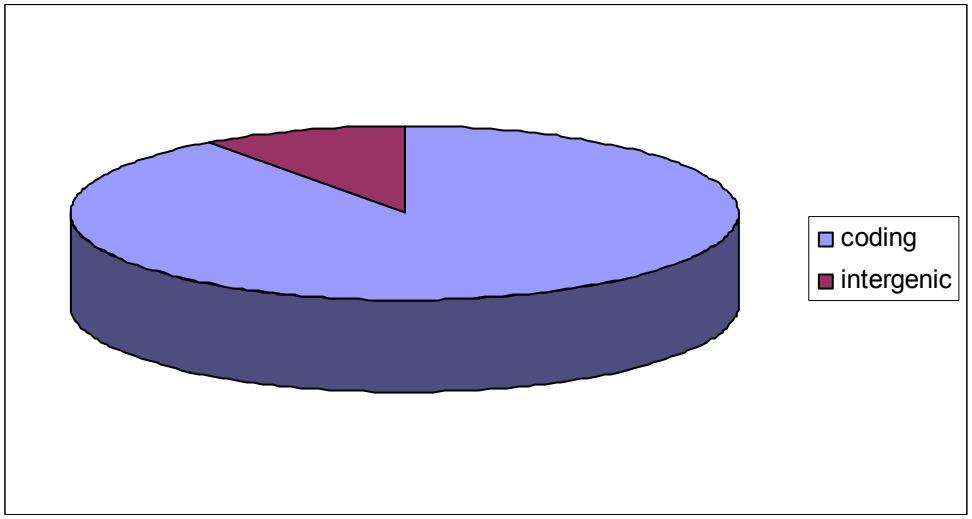
Winkler & Breaker. Ann. Rev. Microb. 2005


(a)
Pause
MKAIPVLKGWWRTS
polyU tract
Insufficient charged
tRNA^Trp
Pause
MKAIPVLKGWWRTS
polyU tract
RBS
trpEDCBA
(b)
polyU tract
LicT
Phosphorylated in
presence of glucose
LicT
polyU tract
RBS
(c)
RBS
Increased
temperature
RBS
(d)
RBS
Low temperature
DsrA sRNA
RBS
(e)
polyU tract
Uncharged
tRNA^Gly
polyU tract
RBS
(f)
polyU tract
RBS
TPP
TPP
polyU tract
Current Opinion in Chemical Biology

Recherche de nouveaux ARNnc

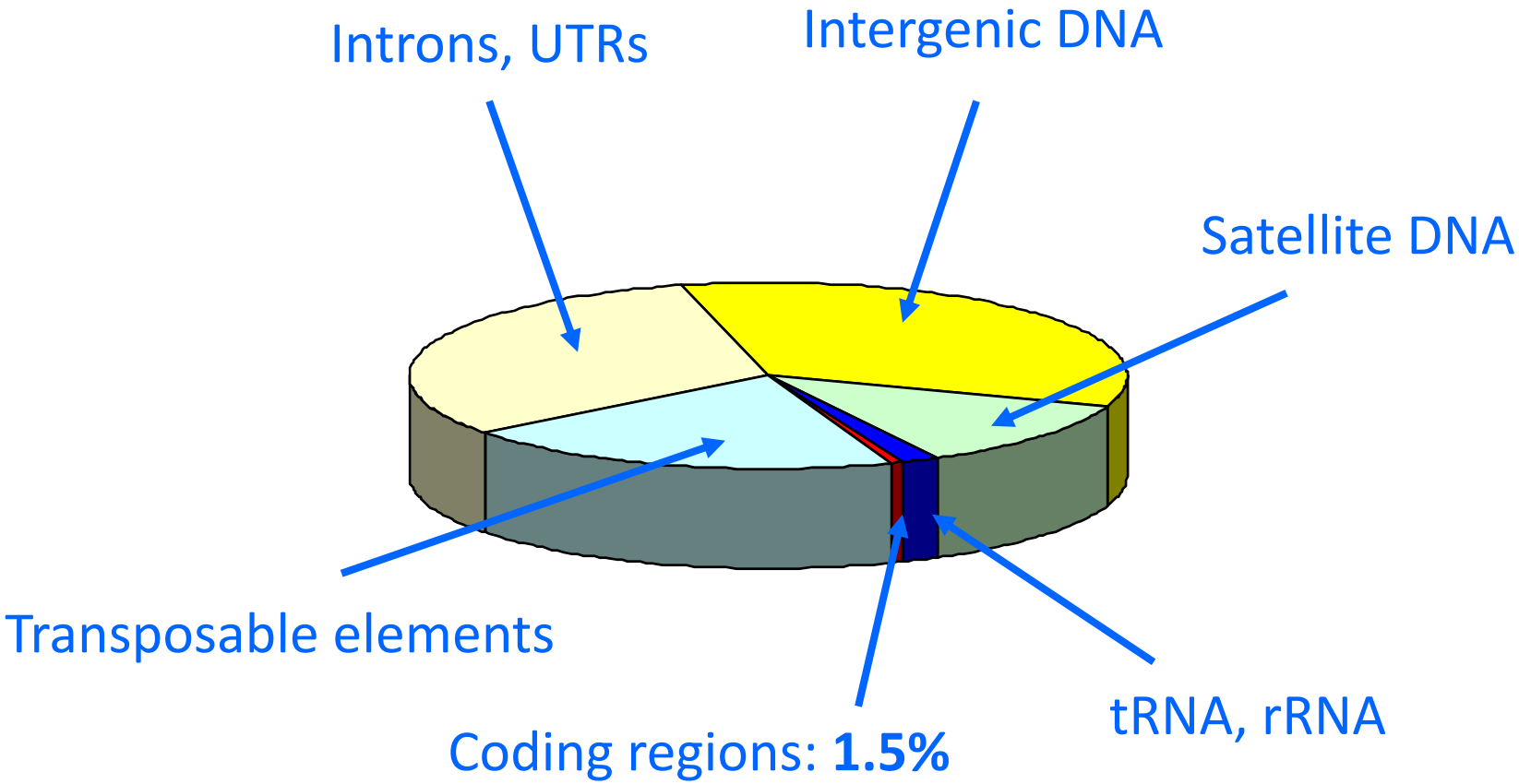
Les transcrits dans les génomes bactérien

- 95% transcrit (~90% codant)



95% du génome se retrouve à un moment donné sous forme d'ARN

Les Génomes vertébrés: 90% transcrit aussi!



90% du génome se trouverait à un moment donné sous forme d'ARN (*cf* Affymetrix results *in* Encode project)

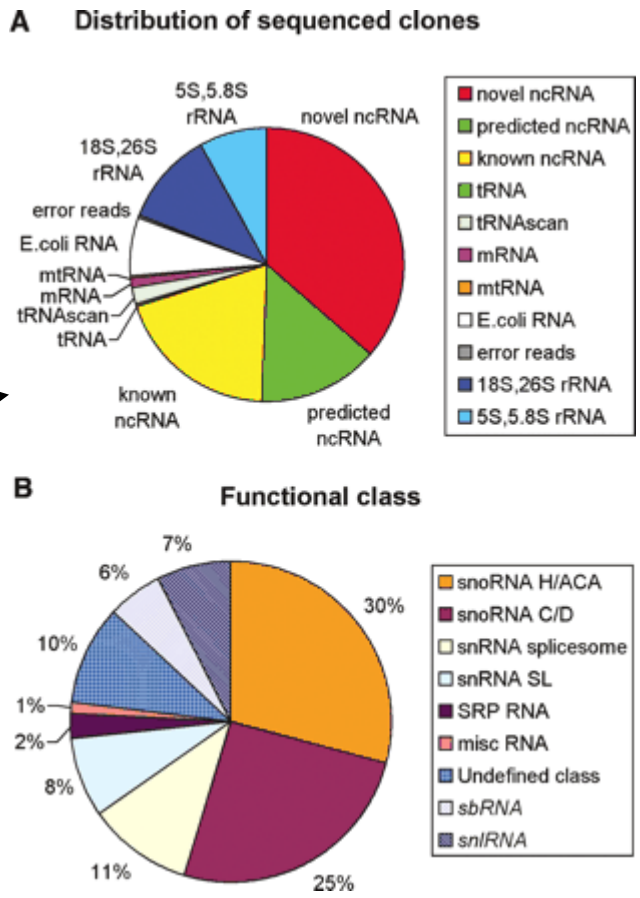
- Quels sont les transcrits non codants fonctionnels?
- Une aiguille dans une botte de foin!

2.1 Approches expérimentales

Cloning

- Rnomics*
  - ExtraireARN 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



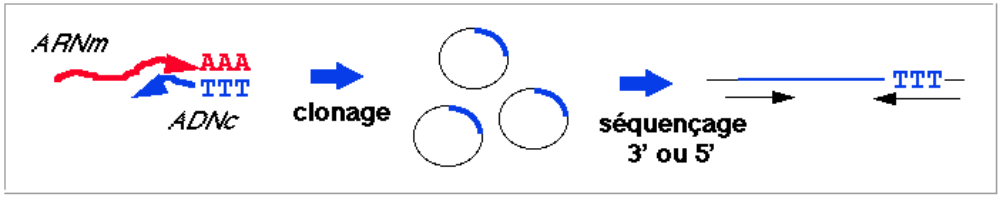
* Huttenhofer et al., 2001

Limitations

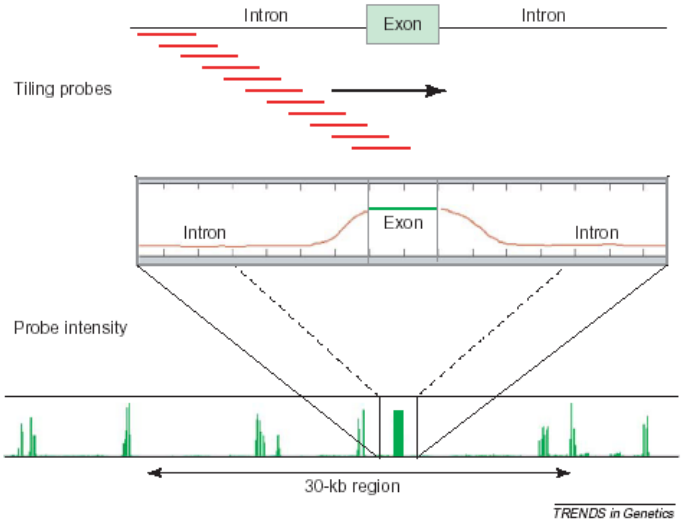
- Sensibilité: ncRNA rares, exprimés à des sites précis ou pendant une courte durée
- Non exhaustivité

Nouvelles approches

- Full-length cDNA projects
  - FANTOM3:
    - 100,000 mouse cDNAs
    - 32,000 non-coding!



- Tiling arrays
  - Half human transcriptome is polyA-, cytoplasmic and maps unnanotated loci



Sequençage massif

- Réalisé chez les eucaryotes et bactéries
- Landgraf et al. Cell 2007:
  - 336 bibliothèques de petits ARN homme,rat,souris
  - 330.000 séquences (75% miRNA)

Limitations

- Nombreux transcrits non fonctionnels (« fuites » de la transcription)
- Pas de preuve de fonction en tant qu'ARN
- Reséquençage des mêmes ARN
  - Crible bactérien récent: 90% 5S!

Recherche Bioinformatique de gènes ARN

Pourquoi la détection des ARN est-elle particulière?

- Pas d'ORF
- Pas de statistiques sur les sequences
- Pas de bonnes Matrices de Substitution pour l' ARN/ADN
- Les ARNnc sont définis à la fois par la structure primaire et secondaire

Recherche d'ARNnc connus

- Comment détecter de nouveaux membres d'une famille connue?

Programmes avec descripteurs

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

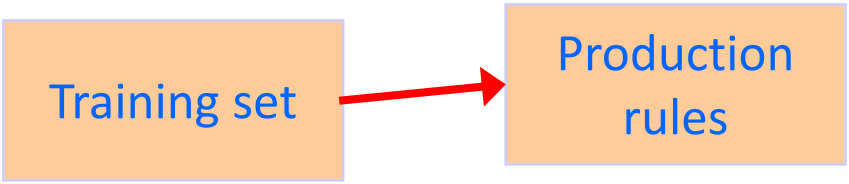
```
h1 s1 h1 s2 h2 s3 h2

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

Programmes Probabilistes

- Grammaires Stochastiques Hors Contexte (first adaptation of CFG to RNA: Searls 94; SCFG: Eddy & Durbin 94)



describe how to generate any structure in the training set

  - Temps de calcul = $O(N^4)$ pour séquence de longueur N
  - Pas réaliste pour de grands alignements ou des recherches dans les génomes

Matrices poids-position

Alignement de miRNA de la famille Mir-133:

```
(( - ((((((( ------ ((( - (((( ---------------------- )))) ))) ------ ))))))) - ))
TC t GGCTGGT caaac- GGA a CCAA gtccgtcttcctgagaggt--- 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 atcaacccagaaaaaggat--- TTGG TCC CCTTCA ACCAGCC g CA
TA t GGCTGGT caaac- GGA a CCAA gtccgtcttccttagaggt--- TTGG TCC CCTTCA ACCAGCT a TT
AG t TGCTGGT aaaac- GGA a CCAA gtcgggtgtttgcgagaggt-- TTGG TCC CTTTCA ACCAGCT a CT
TG t GGCTGGT caaat- GGA a CCAA gtcaggtgtttctgcgaggt-- TTGG TCC CCTTCA ACCAGCT a CT
```

A	
G	
C	
U	
-	

$Score_b = \log(Obs_b / Exp_b)$

Découverte d'ARNnc *de novo*

- Comment détecter des ARNnc pour lesquels aucune information de structure ou de séquence n'est disponible?

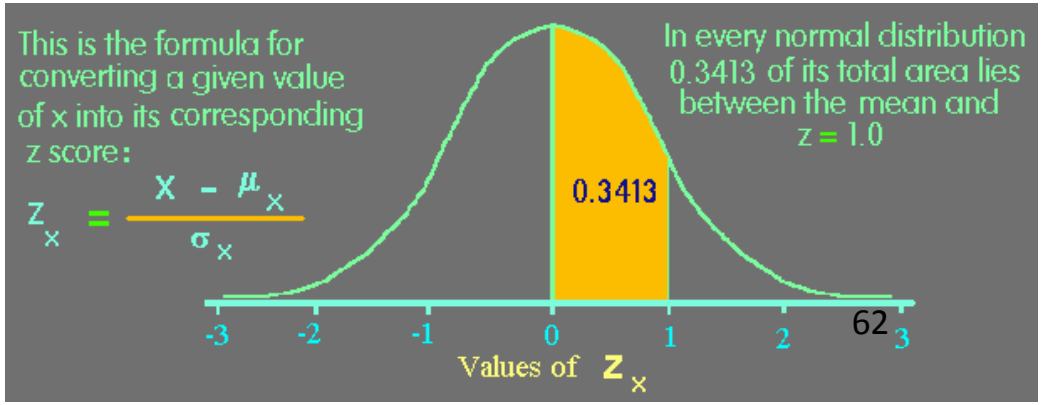
Profils Thermodynamiques (Le *et al*. 88)

```
                                A
                               G C
                               G C
                               G C
                               G C
CACTCTGCTTACTACTTGTCATGAAGAATTC G C GCTTTCTACACTCTGCTTACTACTTGTCATGAAGAACACTCTGCTTACTACCACTCTGC
```

100 bases | 100 bases | 100 bases | 100 bases | 100 bases

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

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



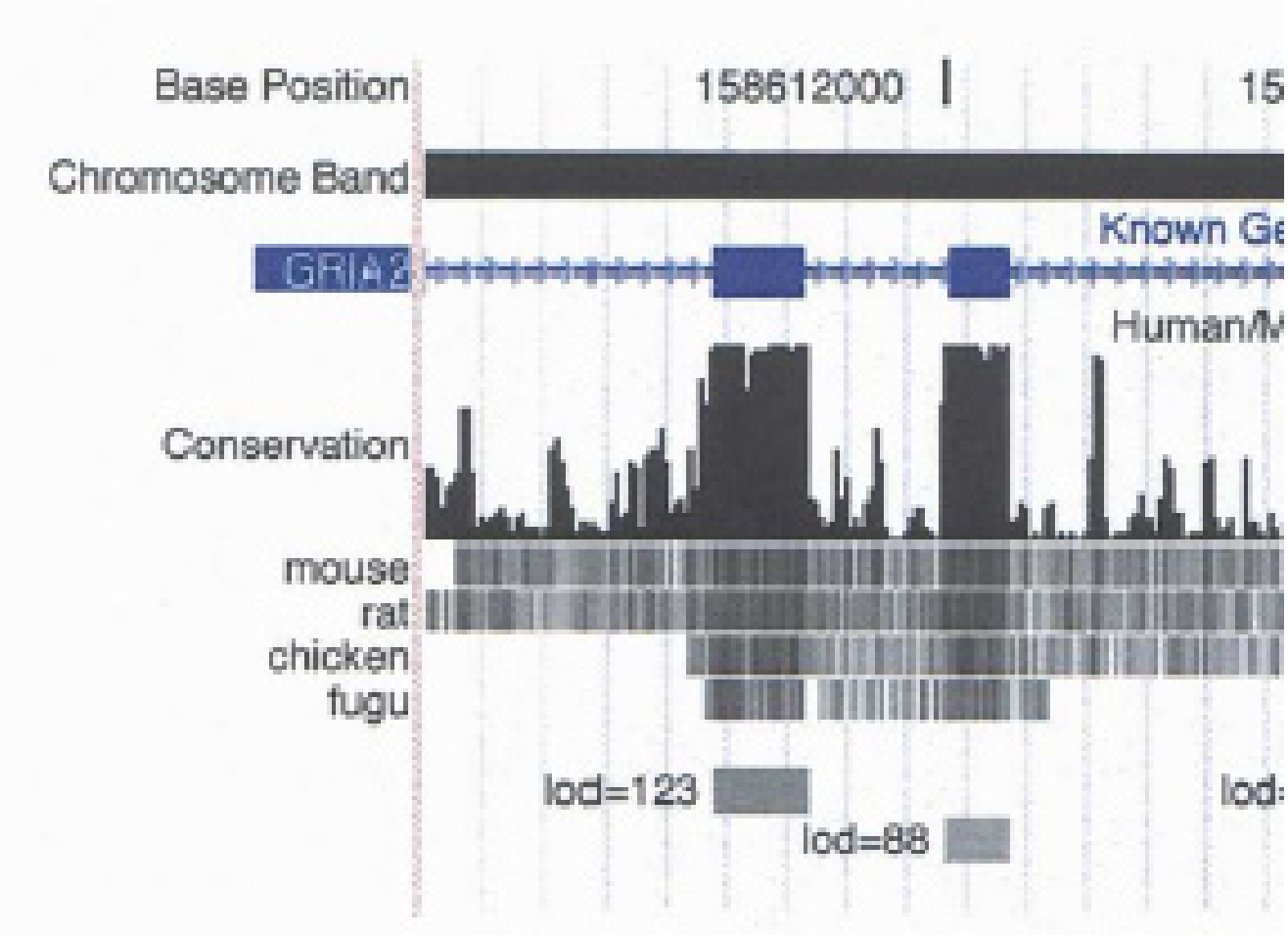
Le problème avec la recherche de structures stables

- ★ OK pour les structures locales fortes (certains génome viraux, le précurseurs de mIRNA)
- ★ Mais: les véritables ncRNA (tRNA, rRNA) n'ont pas un repliement plus stable qu'une séquence aléatoire de même composition (Rivas & Eddy 2000)
- ★ De nombreux ncRNA ne sont pas détectés.
- ★ La composition en G+C est à elle seule un meilleur prédicteur d'ARNnc que l'energie de repliement.

Le contenu en G+C

- ★ Dans les génomes riches en A+T (thermophilic archaebacteria), les ncRNA se distinguent clairement.
- ★ La combinaison (G+C)% et CpG% donne les meilleurs résultats (Schattner '02).
- ★ 10-20 ncRNA prédits de cette façon sont conformés chez *M. jannaschii* et *P. furiosus*.
- ★ Insuffisant dans les genomes à contenu en GC « normal »

La génomique comparative



Comparative genomics: a major source of ncRNA in eukaryotes

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

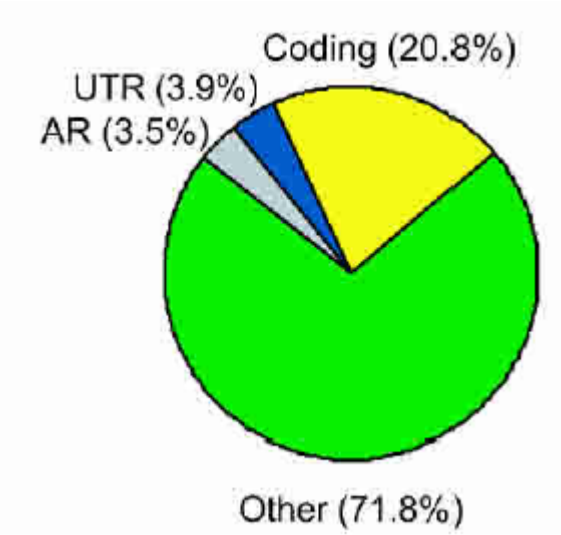
– 5-6% of mamalian genome under selection vs 1.5% coding (3 times as much as in nematodes)

Problem: Functional assignment of conserved regions

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

Detect this!

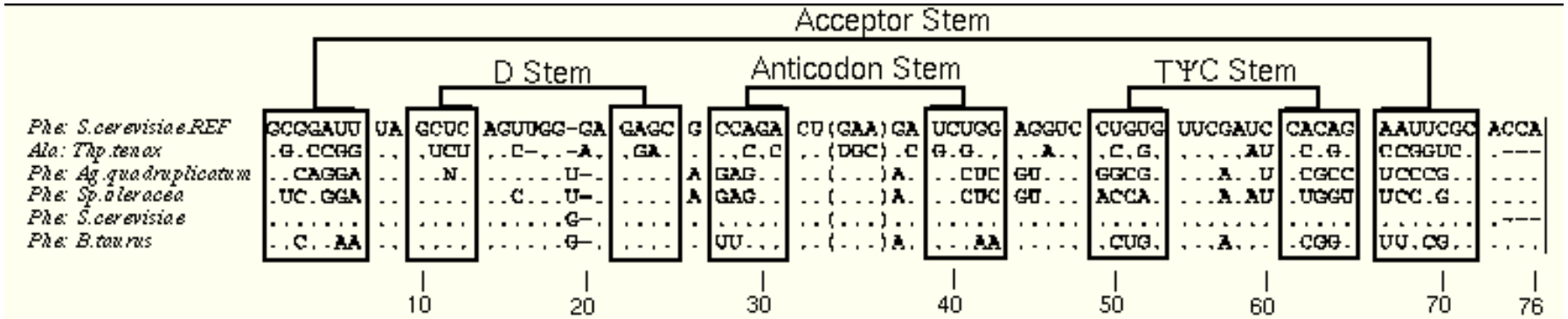
Fraction of conserved sequences in.. (AR=ancestral repeats)



Margulies et al, 2003

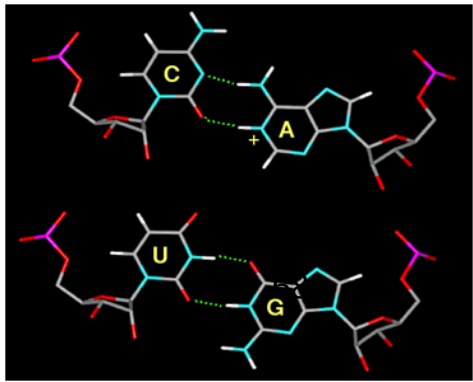
→ Need classification method!

Détecter un ARN par analyse comparative



- L'analyse comparative est la façon la plus fiable de déterminer les structures secondaires
- Elle repose sur la détection de covariations
- Fonctionne même pour les paires non Watson-Crick!

- Mais requiert un excellent alignement



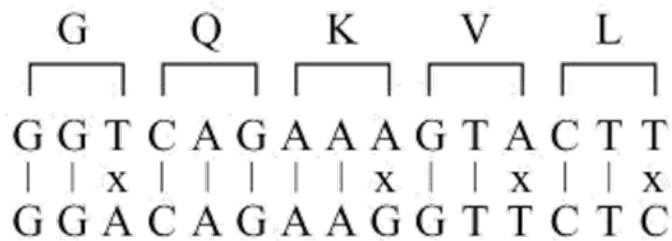
Premiers travaux chez les bactéries

- Wassarman *et al.* (2001)
  - Comparaison *Escherichia, Salmonella, Klebsiella*
  - Analyse visuelle des alignements
  - 60 ncRNA predits, 23 confirmés

Q-RNA (Rivas & Eddy 2001)

– Analysis of Blast alignment (SCFG based)

- Model for protein coding gene

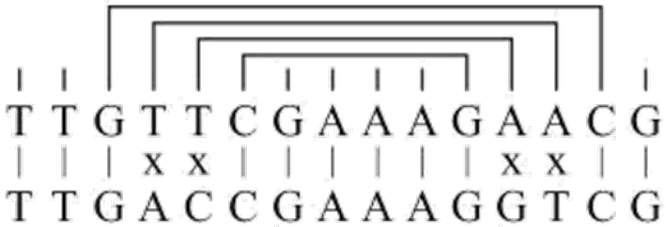


P(GGT-GGA)*P(CAG-CAG)*...

Synonymous mutations

- Model for ncRNA

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



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

Compensatory mutations

RNAz (Washietl et al. 2004)

- Utilise un alignment multiple en entrée
- Deux composantes:
  - (1) Mesure de la conservation de la structure secondaire (exploitant les covariations)
  - (2) Mesure de la stabilité thermodynamique

.

Résultats RNAz

- C. elegans (comparaison avec C. briggsae)
  - 3500 ncRNAs
  - Sensibilité et spécificité: 50%.
  - Estimation: 3000 à 4000 ncRNA conservés dans le génome
  - Beaucoup de familles inconnues
- Homme+mammifères
  - Environ 10000 ARN non codants prédits

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