

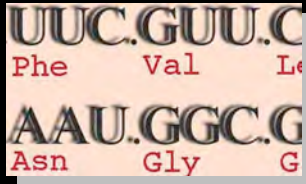


De la biologie symplectique à la biologie synthétique, quels gènes pour faire une cellule ?

Cours d'Analyse des Génomes

15 décembre 2006





- ➔ **Contexte**
- ➔ **Vie et calcul**
- ➔ **Du brin précoce au brin retardé**
- ➔ **La traduction organise le génome bactérien**
- ➔ **Une illustration**
- ➔ **Le coeur du génome: ce qui persiste**
- ➔ **Le cénome**
- ➔ **Le futur: vers la “biologie synthétique”**





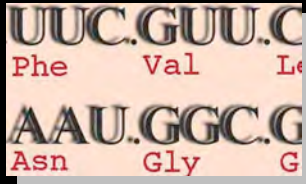
Une révolution vieille de vingt ans : les projets génomes



2210 projets en cours, 460 complets, principalement de microbes (295 de plus de 1500 gènes, plus ou moins bien annotés) font l'inventaire de tous les gènes d'un organisme
150 000 000 000 nucléotides à l'International Nucleotide Sequence Database Collaboration (INSDC) (<http://www.insdc.org>)

Les microbes forment 50% du protoplasme terrestre

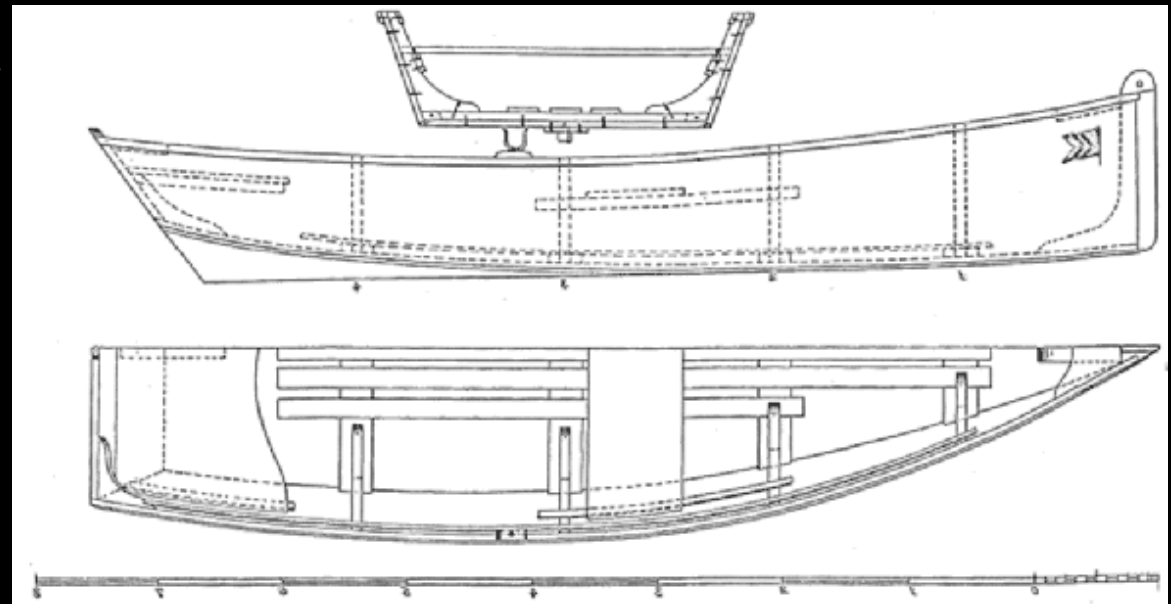
40-50% des régions codantes de l'ADN ne correspondent pas à des fonctions connues ; 10% correspondent au cœur du génome (gènes « persistants »)



La biologie est « symplectique »



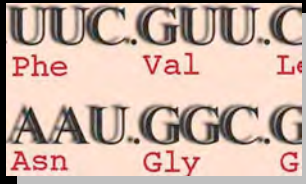
- La biologie est une science des relations entre objets : elle est **symplectique** (de συν ensemble, πλεκτειν, tisser)
- Comme dans la construction d'une barque, ne pas comprendre comment les objets interagissent conduira au naufrage
- Mais s'il n'y a pas d'instruction intelligente, comment les relations sont-elles donc créées ?



A. Danchin

La barque de Delphes, Odile Jacob, 1998
The Delphic Boat, Harvard University Press, 2003





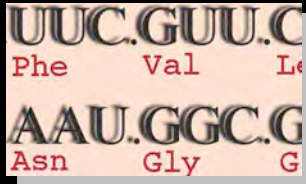
Contexte : le “programme génétique”



- **Physique:** *matière, énergie, temps*
- **Physique statistique:** *Physique + information*
- **Biologie:** *Physique + information, codage, contrôle...*
- **Arithmétique:** *suites d'entiers, récursivité, codage...*
- **Calcul:** *Arithmétique + programmes + machine...*

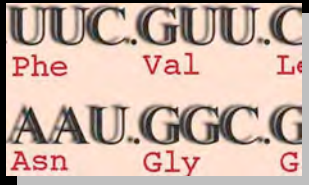
La métaphore du « programme génétique » a des conséquences pratiques : nous savons manipuler les gènes et leurs produits, **pouvons-nous pousser la métaphore jusqu'à ses conséquences ultimes ?**





- ➔ **Contexte**
- ➔ **Vie et calcul**
- ➔ **Du brin précoce au brin retardé**
- ➔ **La traduction organise le génome bactérien**
- ➔ **Une illustration**
- ➔ **Le cœur du génome: ce qui persiste**
- ➔ **Le cénome**
- ➔ **Le futur: vers la “biologie synthétique”**





Ce qu'est la vie



Deux entités et **trois processus** sont nécessaires à la vie :

→ **Un programme** (un “livre de recettes”)

→ 1. **Transfert d'information** => l'objet de la génomique est de déchiffrer le programme associé à la cellule et de comprendre sa signification

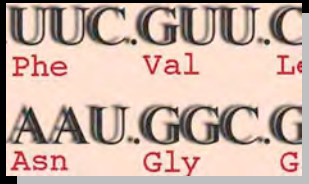
Forces couplant la structure du génome à celle de la cellule (l'**usine cellulaire**) :

→ **Une machine permettant au programme d'agir**

→ 2. **Métabolisme** (une dynamique)

→ 3. **Compartmentation** (la définition d'un intérieur et d'un extérieur)

La cellule est l'atome de vie, avec deux stratégies de compartimentation : une simple enveloppe (procaryotes), ou la multiplication des membranes et des peaux (eucaryotes) ; on peut remarquer que cela est corrélé avec la séquence des génomes : génomes **procaryotes semblent aléatoires** et les **génomes eucaryotes semblent répétés**



Transfert d'information



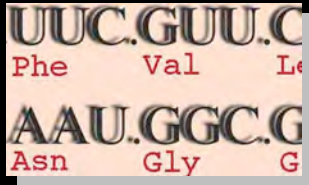
→ **Réplication** (loi: “complementarité”; concept: “effectif”)

→ **Transcription** (loi: “complementarité”; concept: “constructif”)

→ **Traduction** (loi: “chiffage”, le “code génétique”; concept: “prospectif”)

Myhill, J. (1952) Some philosophical implications of mathematical logic. Three classes of ideas. The Review of Metaphysics **6** : 165-198.





Vue algorithmique des actions biologiques

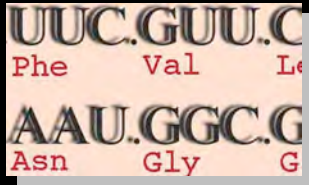


Réplication, transcription, traduction : parallélisme élevé

“Début, Routine répétitive et Points de contrôle, Fin”

L'action est toujours orientée, avec un début et une fin

Les processus de contrôle temporel (check points) sont rarement pris en compte (excepté pour les processus de réplication/division), **mais leur rôle est essentiel pour permettre la coordination de multiples actions en parallèle**



Qu'est-ce que calculer ?

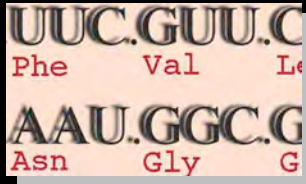


Deux processus sont nécessaires au calcul :

- **Une machine capable de lire et écrire**
- **Un programme sur un support physique** (typiquement, une bande perforée ou magnétique illustre la suite séquentielle des symboles qui forment le programme), séparé (en pratique, mais pas conceptuellement) en deux entités :
 - **Programme** (fournissant l'objectif)
 - **Données** (fournissant le contexte)

La machine est distincte du programme





La machine de Turing



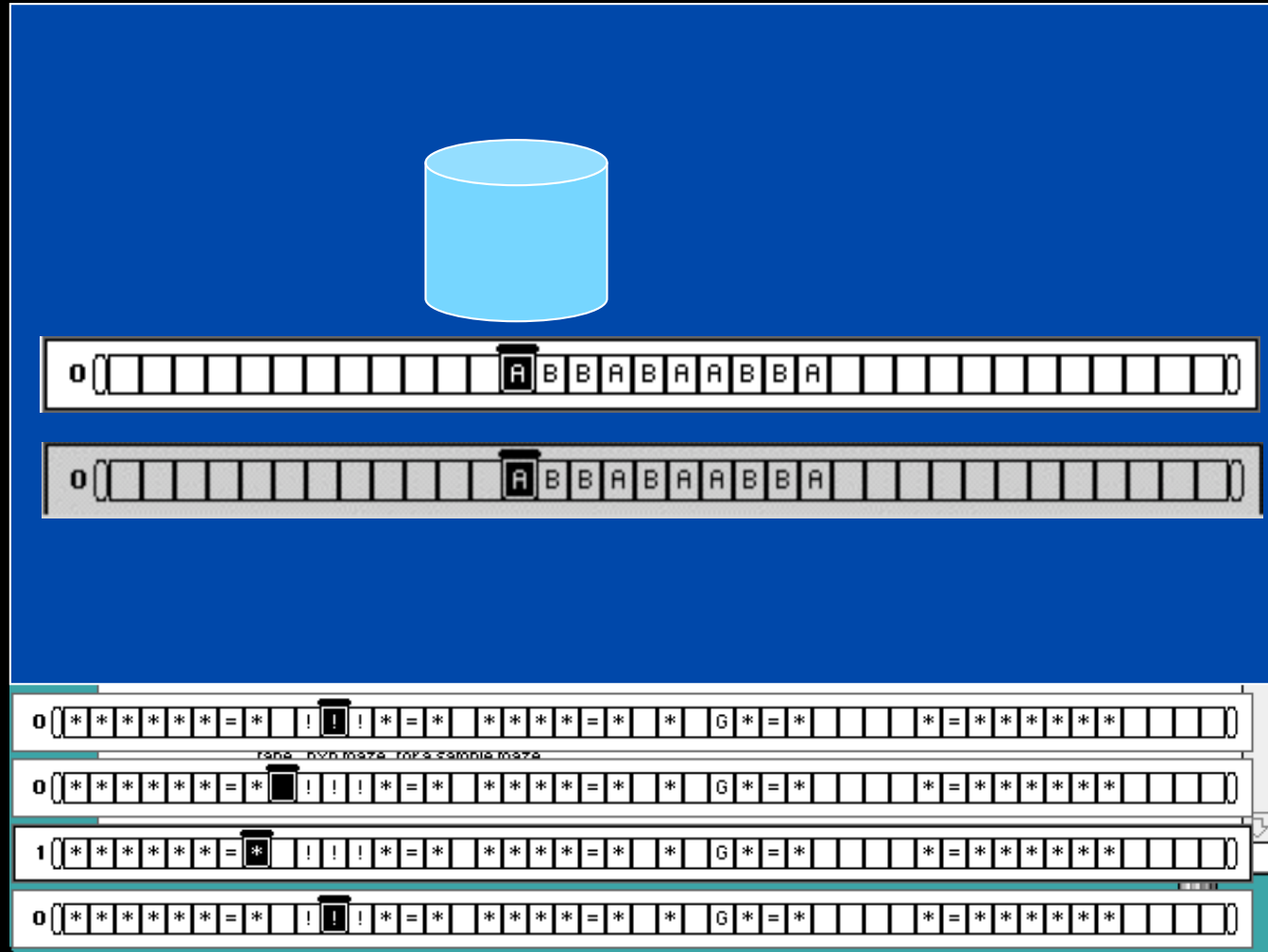
machine
(lire/écrire)

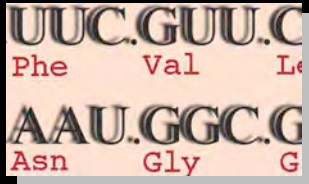
programme
(données)

sous la
forme d'une

suite

linéaire de
symboles





Cellules et ordinateurs



La génétique repose sur la description des génomes comme **textes écrits avec un alphabet de quatre lettres** : mais **les cellules se comportent-elles comme des ordinateurs ?**

Transfert Génétique Horizontal

Virus

Génie génétique => reconstruction du virus de l'hépatite C

Clonage animal

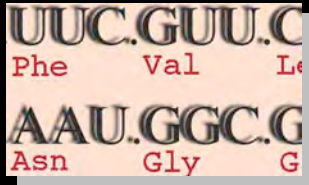
tout indique une séparation entre

« Machine » (l'usine cellulaire)

et

Données + programme





Y a-t-il une image de la cellule dans le chromosome?



Si la machine doit non seulement se comporter comme un ordinateur mais aussi construire la machine elle-même, on doit trouver une image de la machine quelque part dans la machine (John von Neumann, à propos du cerveau)

Cette réflexion, appliquée dans un contexte où elle ne peut l'être (le fonctionnement cérébral) est-elle applicable dans le cas de la cellule et du programme génétique ?

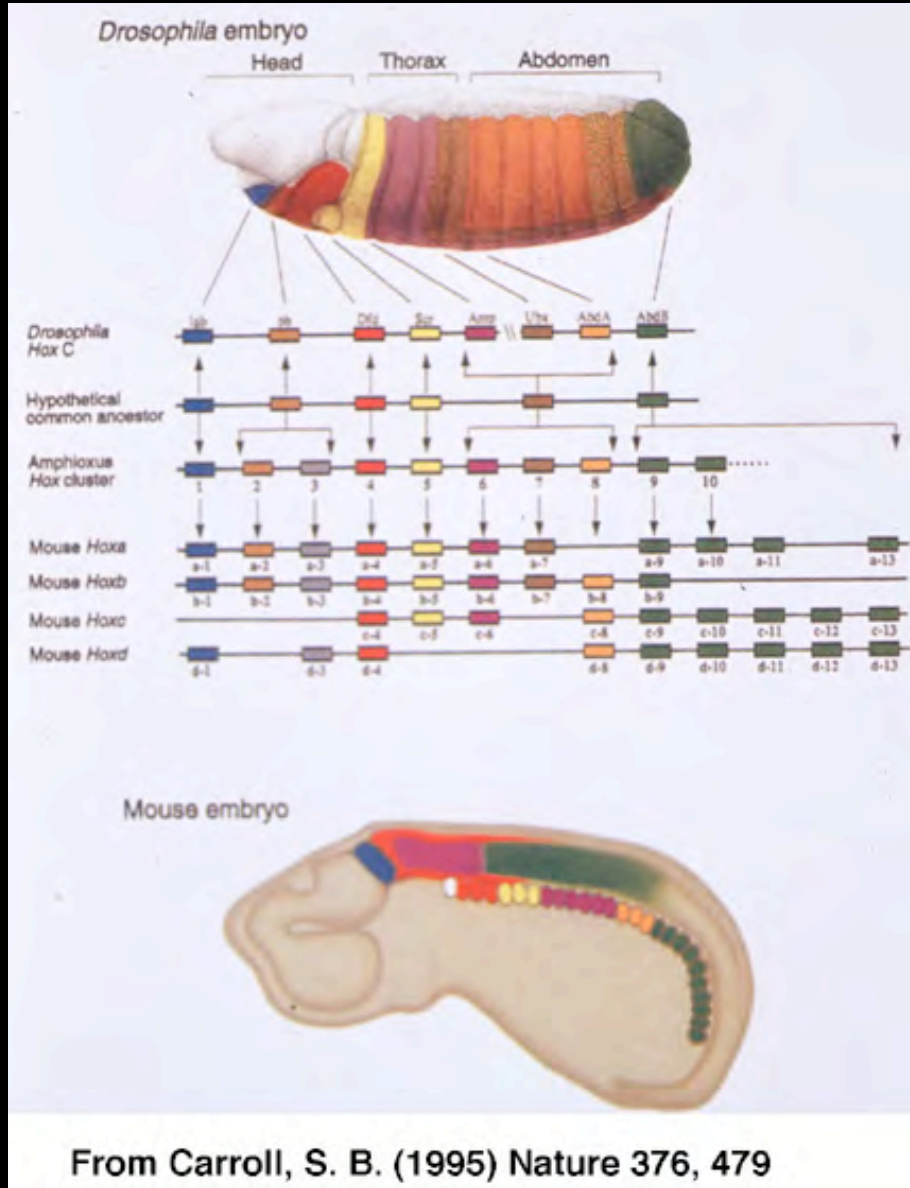


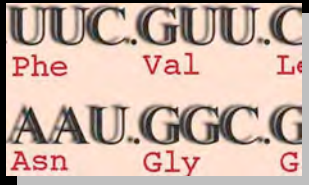
UUC.GUU.C
Phe Val Leu
AAU.GGC.G
Asn Gly G

Drosophiloculus,

Homunculus?

Celluloculus?





Organisation des génomes

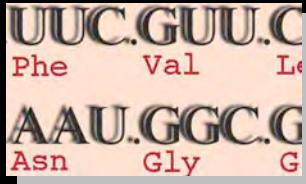


L'ordre des gènes dans les chromosomes est-il aléatoire ?

A première vue, en accord avec la grande variété de mécanismes de manipulation de l'ADN chez les différentes espèces, on observe peu de conservation de l'organisation des gènes, d'autant plus que le transfert génétique « horizontal » a un rôle très important et redistribue les gènes

Pourtant, des groupes de gènes comme les **opérons** ou les **îlots de pathogénicité** démontrent un regroupement systématique, associé à la présence de fonctions communes. Les **gènes « persistants »** sont regroupés

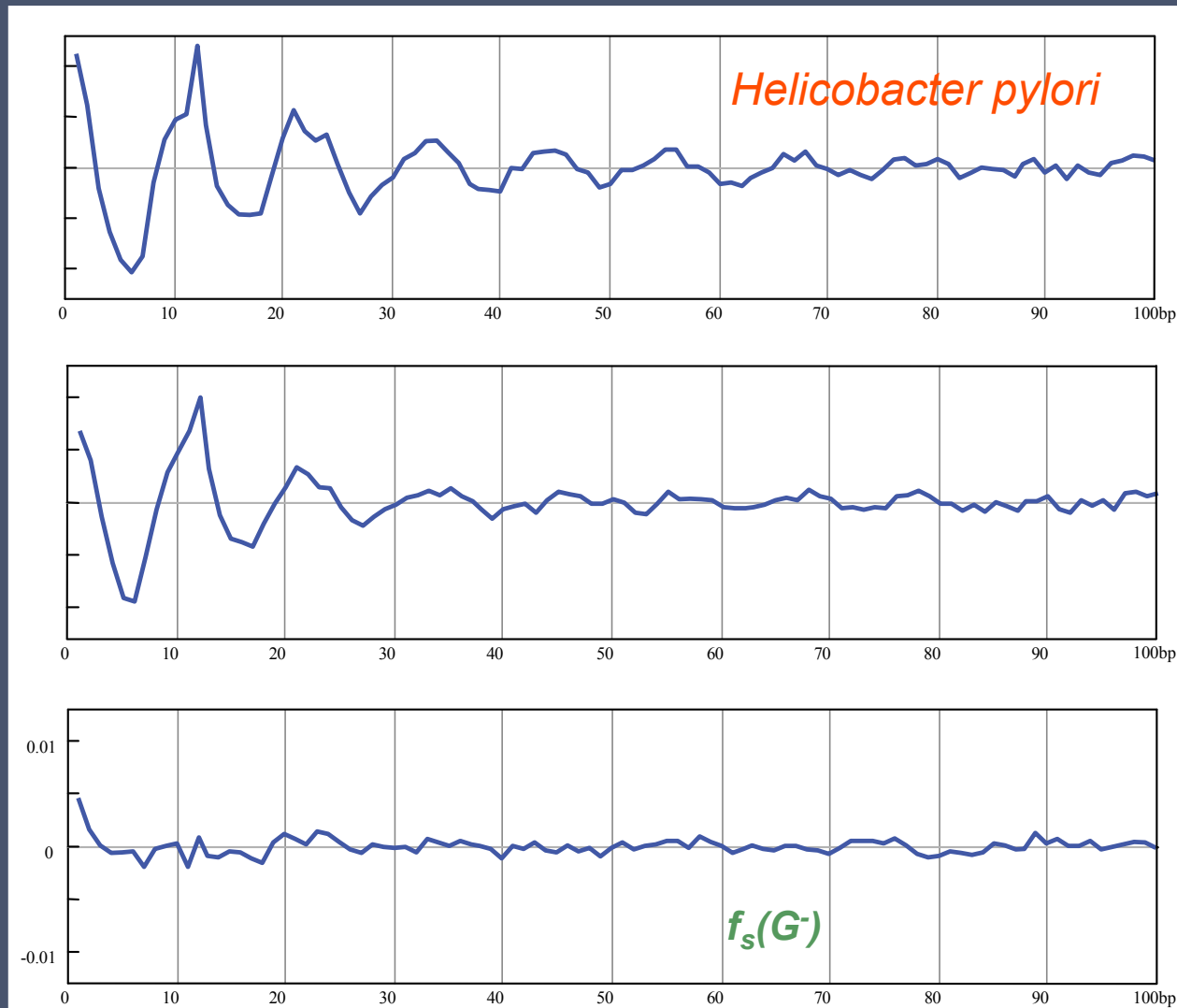




- **Contexte**
- **Vie et calcul**
- **Du brin précoce au brin retardé**
- **La traduction organise le génome bactérien**
- **Une illustration**
- **Le cœur du génome: ce qui persiste**
- **Le cénome**
- **Le futur: vers la “biologie synthétique”**



Un caractère universel du texte du programme : la période 10-11,5



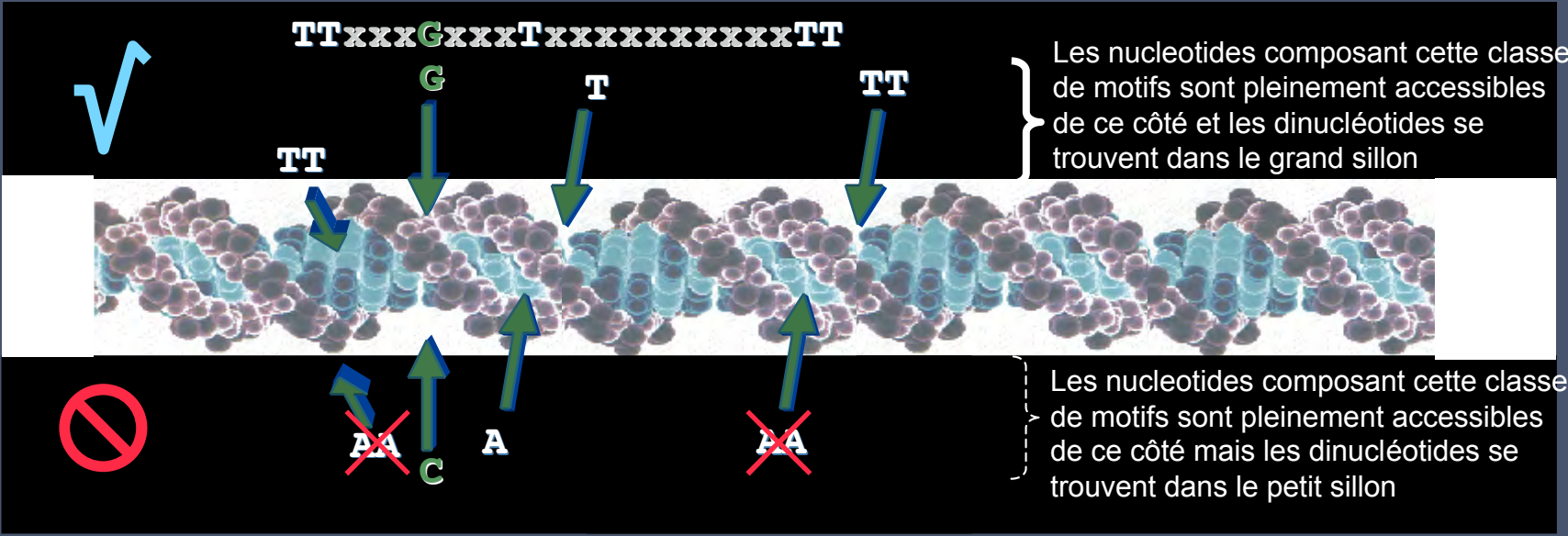
réel

modèle

différence

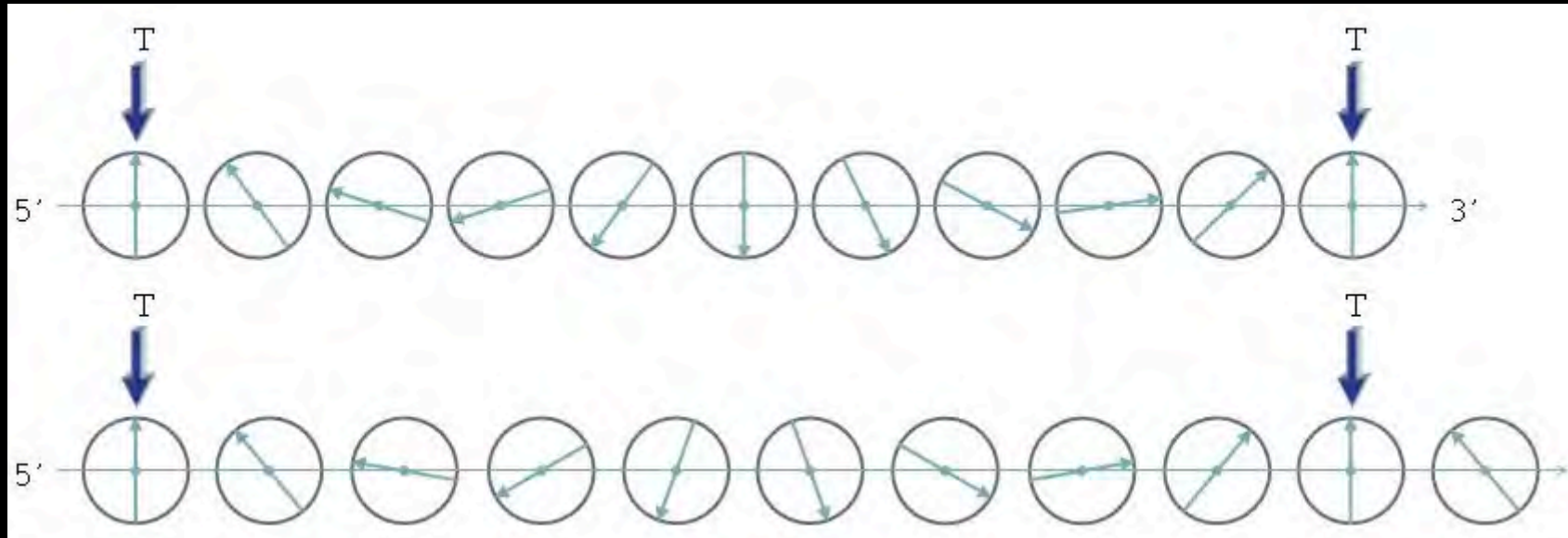
Motifs flexibles de type A

\longleftrightarrow \longleftrightarrow \longleftrightarrow \longleftrightarrow \longleftrightarrow \longleftrightarrow \longleftrightarrow
 1-xAxxxxTxxxxAxxxxTTxxxxxAxxxxTxxxxAxxx: Tous règnes
 2-xxxxxxxxxxxxGxxxxTTxxxGxxxxTxxxxxxxx: Proteobacteria
 4-xxxxxTxxxxAGxxxTTxxxxxxxxTxxxxxxxx: Archaea
 5'-xxx-10xxxxxxxx0xxxxxxxx10xxxxxbp-3'



UUC.GUU.C
Phe Val Le
AAU.GGC.G
Asn Gly G

Une règle universelle : les motifs flexibles de type A



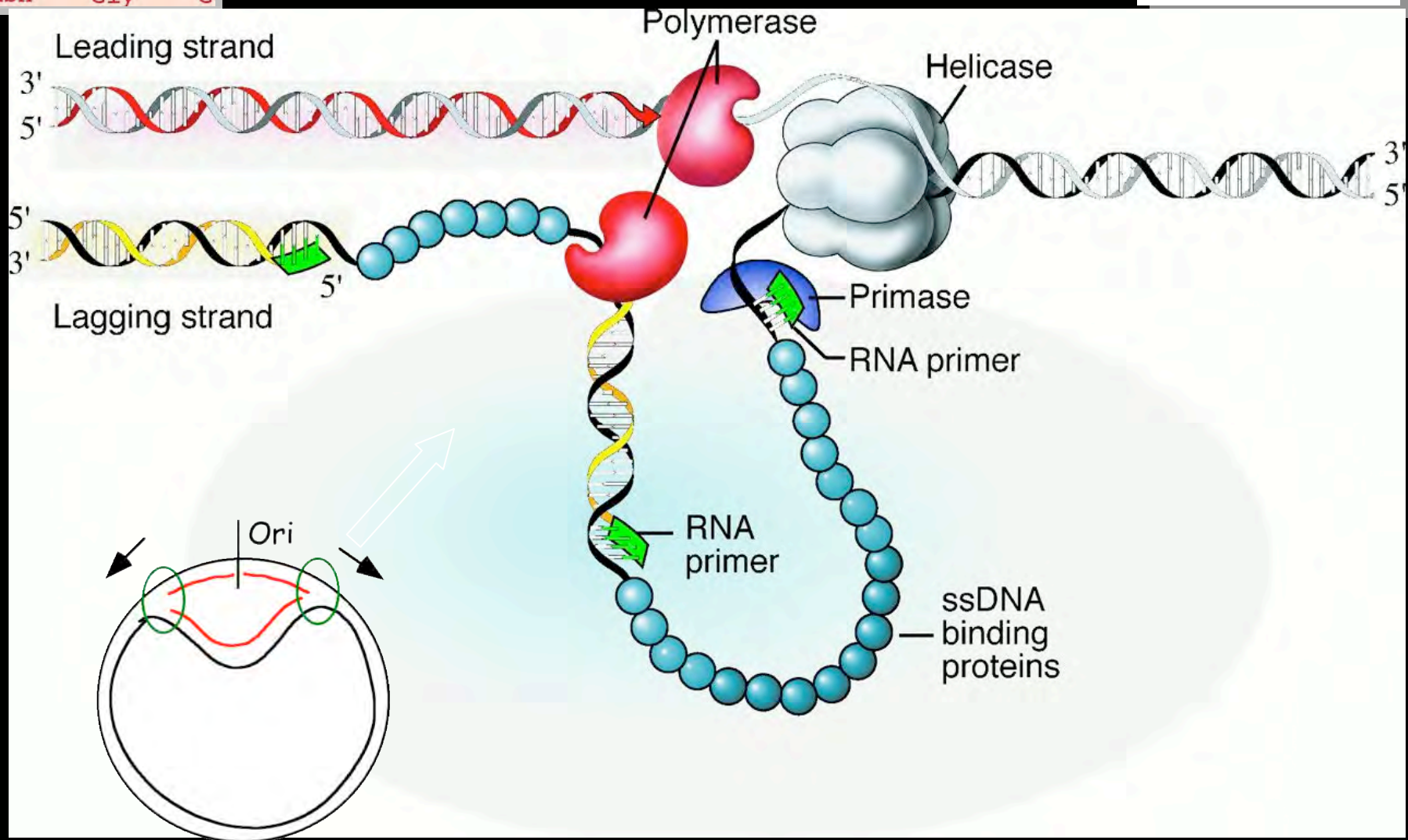
La flexibilité de ces motifs permet à l'ADN de prendre en compte les supertours et les courbures locales

Larsabal E, Danchin A.

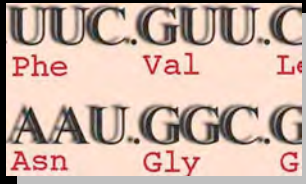
Genomes are covered with ubiquitous 11 bp periodic patterns, the "class A flexible patterns »
BMC Bioinformatics. 2005 6:206



UUC.GUU.C
Phe Val Leu
AAU.GGC.G
Asn Gly G

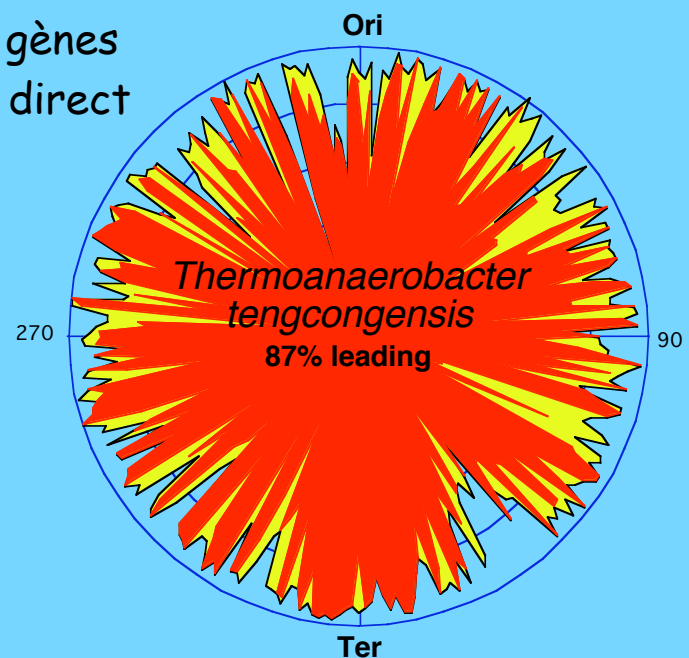
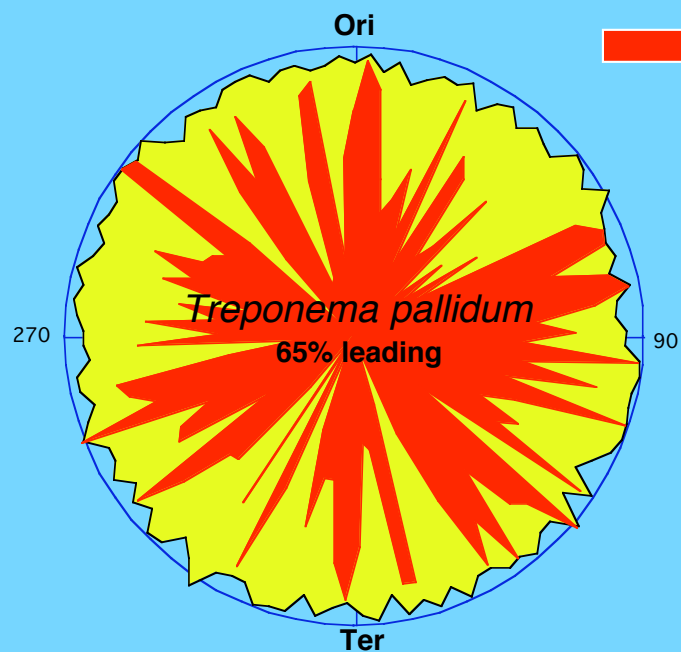
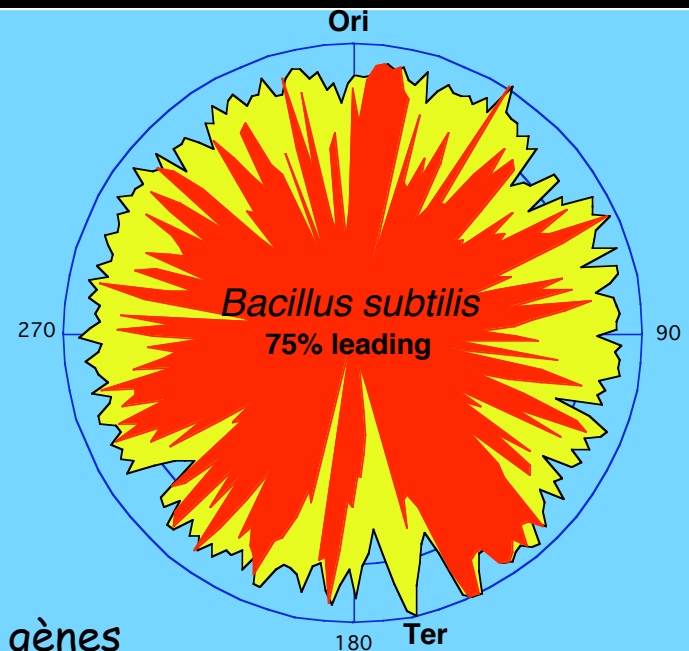
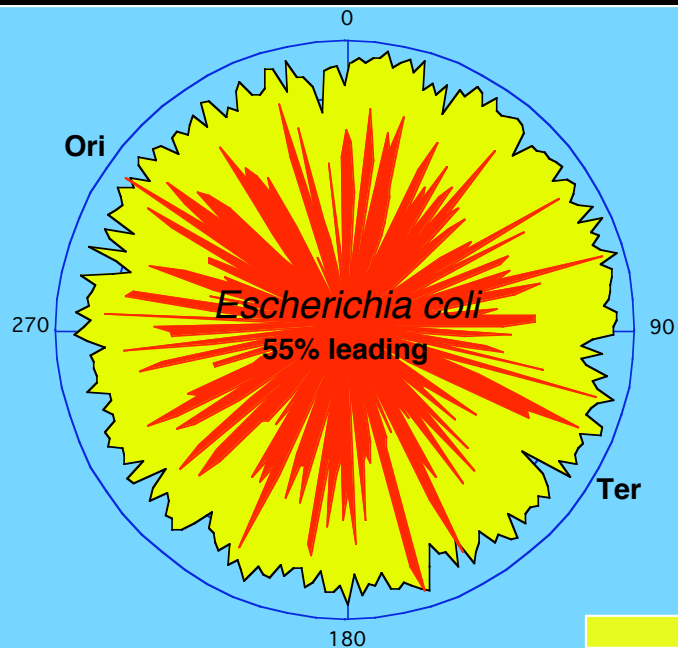




Ter

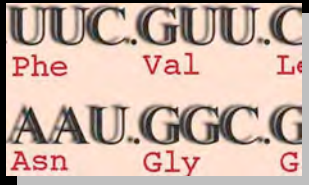


Les gènes ont tendance à préférer le brin précoce de la réplication chez les Bactéries. On observe cependant une très grande variation, qui dépend de l'organisme : les bactéries à coloration de Gram positive et dont l'ADN est riche en A+T ont un biais particulièrement important

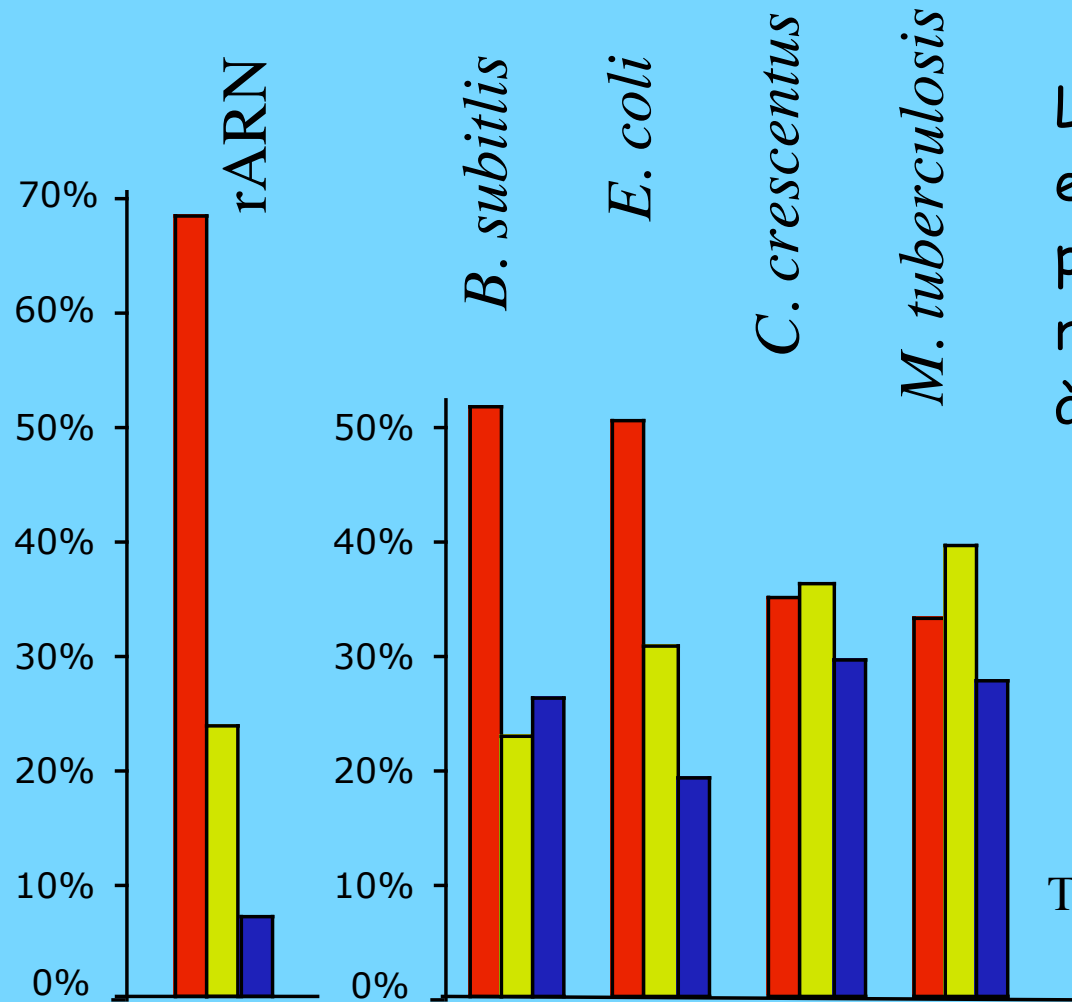




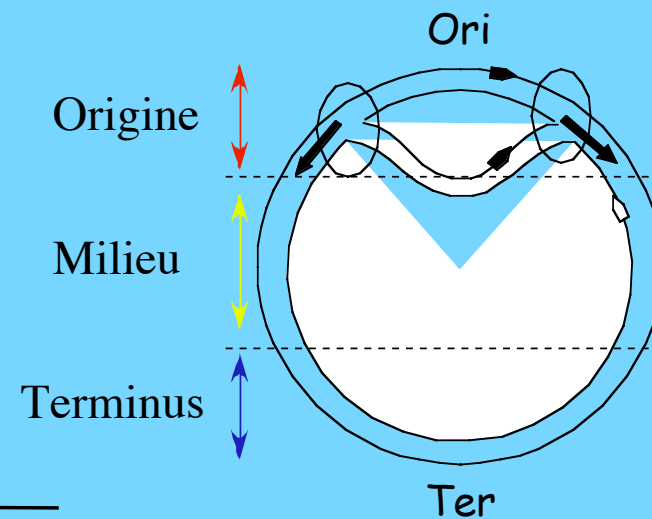
 densité des gènes
 densité des gènes dans le brin direct (précoce)

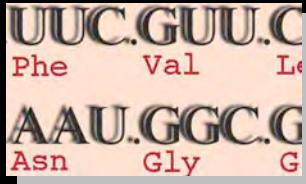


Répartition des gènes fortement exprimés



Les gènes fortement exprimés sont rassemblés près de l'origine de réplication chez les bactéries à croissance rapide





Tôt ou tard...



Est-il possible de voir s'il y a une différence dans la composition en nucléotides, entre le brin précoce et le brin retardé ? Est-ce que cela influence le biais d'usage du code ? Est-ce que cela a une influence sur la composition en acides aminés des protéines ?



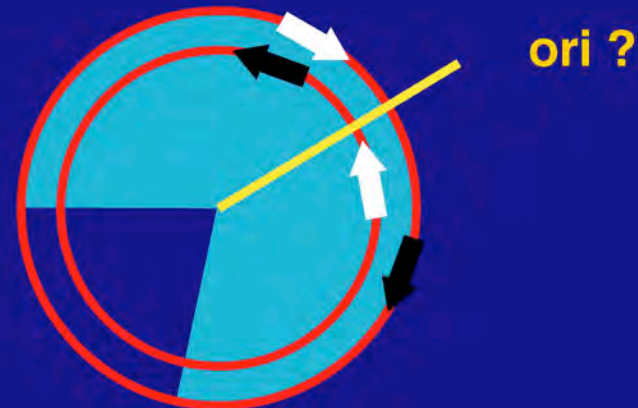
UUC.GUU.C
Phe Val Le
AAU.GGC.G
Asn Gly G

Tôt ou tard ...



Prenant une origine arbitraire pour la réplication et une propriété du brin (composition en bases, en codons, composition en acides aminés de la protéine codée...) l'analyse discriminante linéaire permet de découvrir s'il y a une origine, et s'il y a un biais entre les brins

REPLICATION BIASES IN BACTERIA



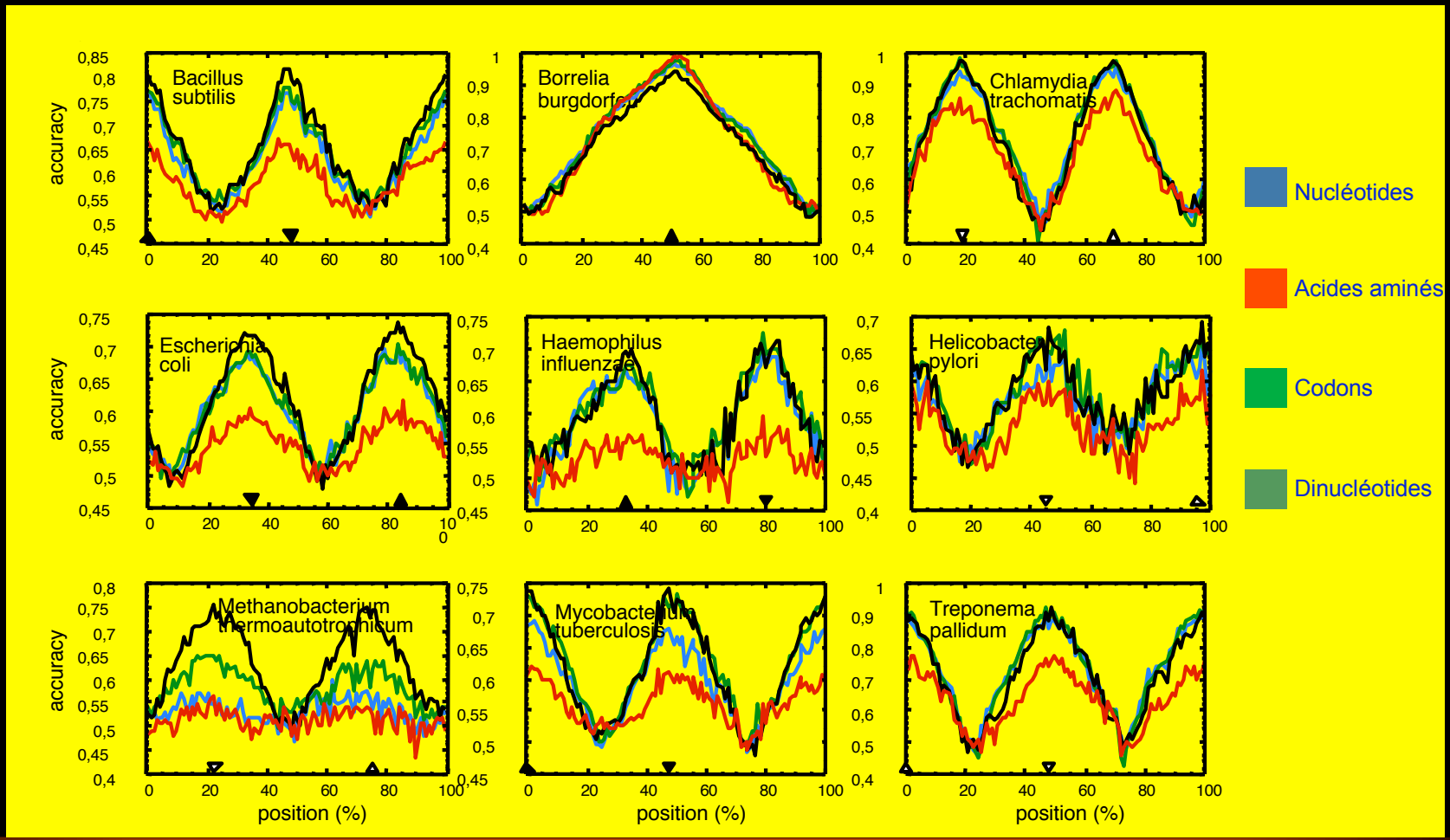
Genomes in silico



E. Rocha, A. Danchin & A. Viari Universal replication biases in bacteria. Mol. Microbiol. (1999) 32: 11-16

UUC.GUU.C
Phe Val Leu
AAU.GGC.G
Asn Gly G

C'est la question...



Génétique des Génomes Bactériens

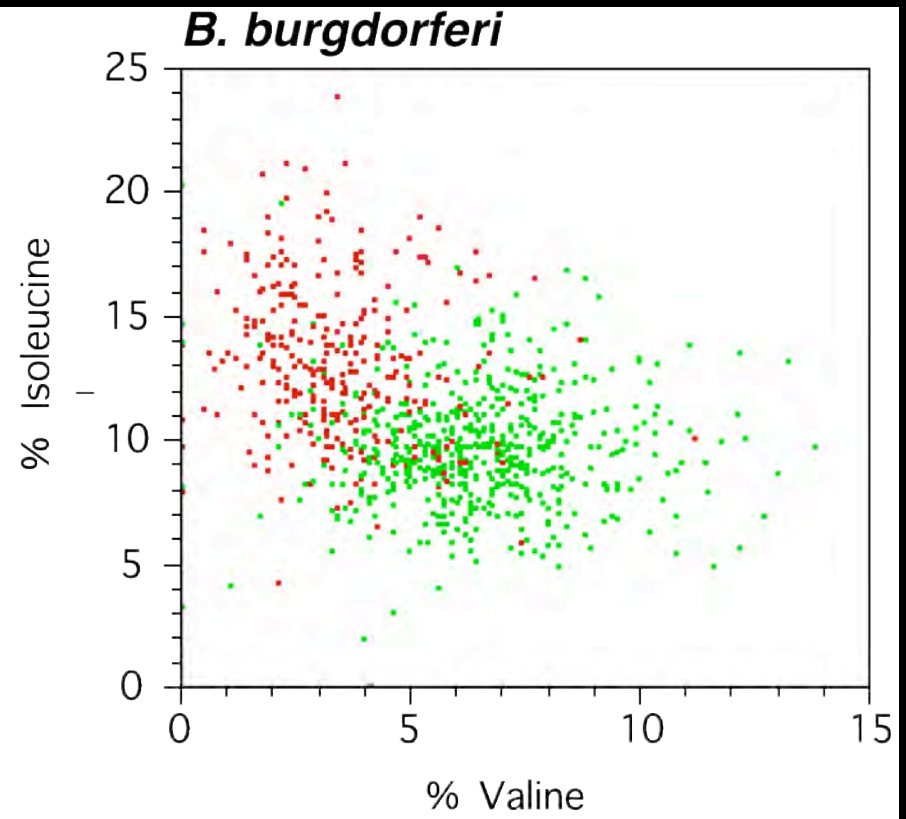
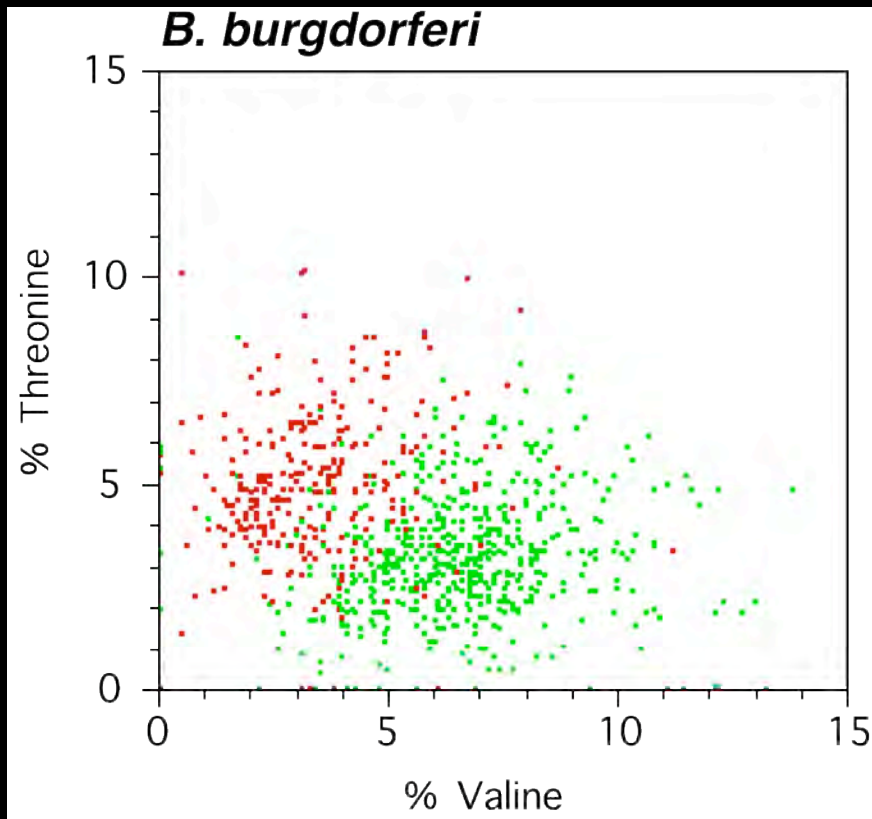
<http://www.pasteur.fr/recherche/unites/REG/>

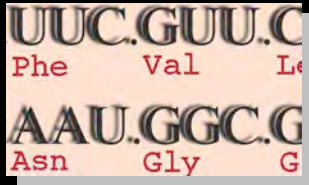


Visible dans les protéines...

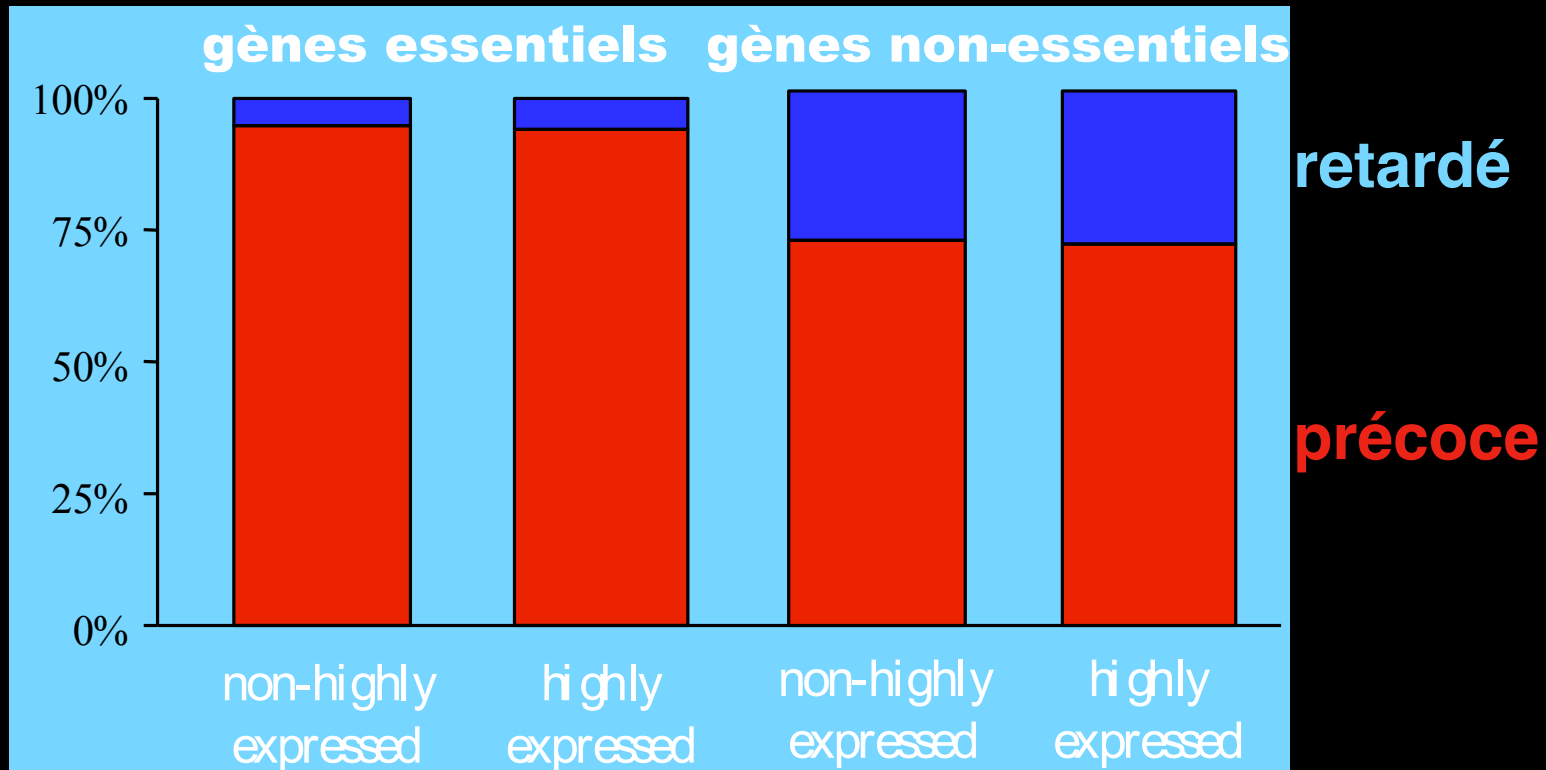


GT dans le brin précoce, CA dans le brin retardé...





Les gènes essentiels sont dans le brin précoce



Rocha EP, Danchin A.
Essentiality, not expressiveness, drives gene-strand bias in bacteria
Nature Genetique. 2003 34:377-378.

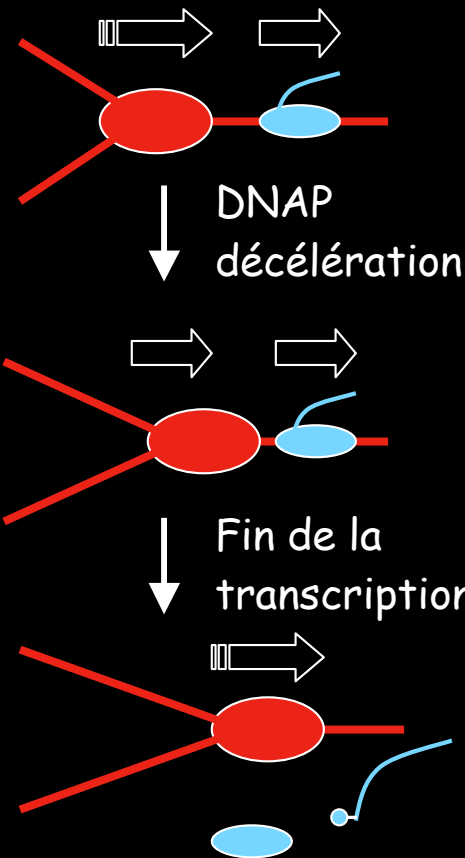


UUC.GUU.C
Phe Val Le
AAU.GGC.G
Asn Gly G

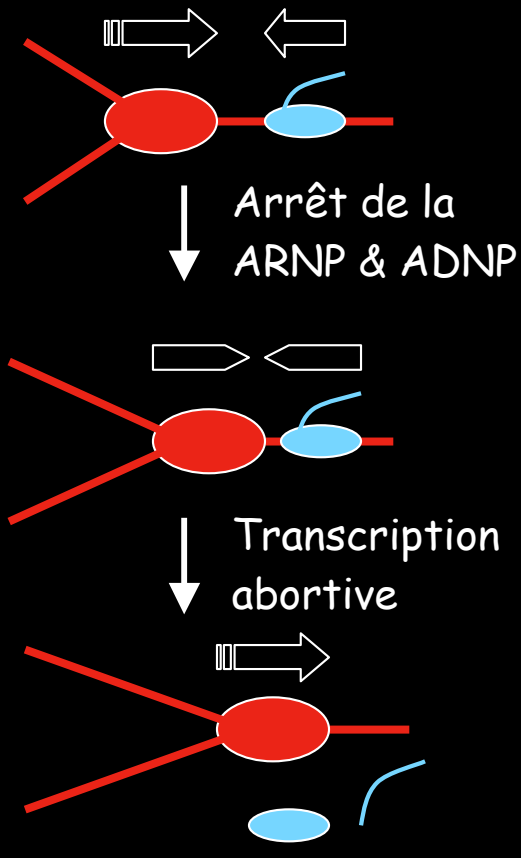
Collision des polymérase



Co-orientées



Frontale



Conséquences:

1. Ralentissement de la réplication
2. Perte des transcrits

Conséquences:

1. Transcrits interrompus
2. Protéines essentielles tronquées



- ➔ **Contexte**
- ➔ **Vie et calcul**
- ➔ **Du brin précoce au brin retardé**
- ➔ **La traduction organise le génome bactérien**
- ➔ **Une illustration**
- ➔ **Le cœur du génome: ce qui persiste**
- ➔ **Le cénome**
- ➔ **Le futur: vers la “biologie synthétique”**





Analyses multivariées



A l'opposé de la génétique habituelle, la génomique analyse de grandes collections de gènes et de produits de gènes

Les analyses multivariées cherchent à extraire l'information en réduisant le plus possible les descripteurs des objets étudiés

L'Analyse en Composantes Principales utilise la valeur centrée réduite des caractères étudiés est une méthode de base

L'Analyse Factorielle des Correspondances est de la même famille, mais utilise la mesure du χ^2 comme distance. Cela permet non seulement d'étudier des objets dont les descripteurs sont hétérogènes, mais aussi de travailler sur l'espace des descripteurs et sur l'espace des objets simultanément

L'Analyse en Composantes Indépendantes utilise le caractère non gaussien de la répartition des valeurs, etc



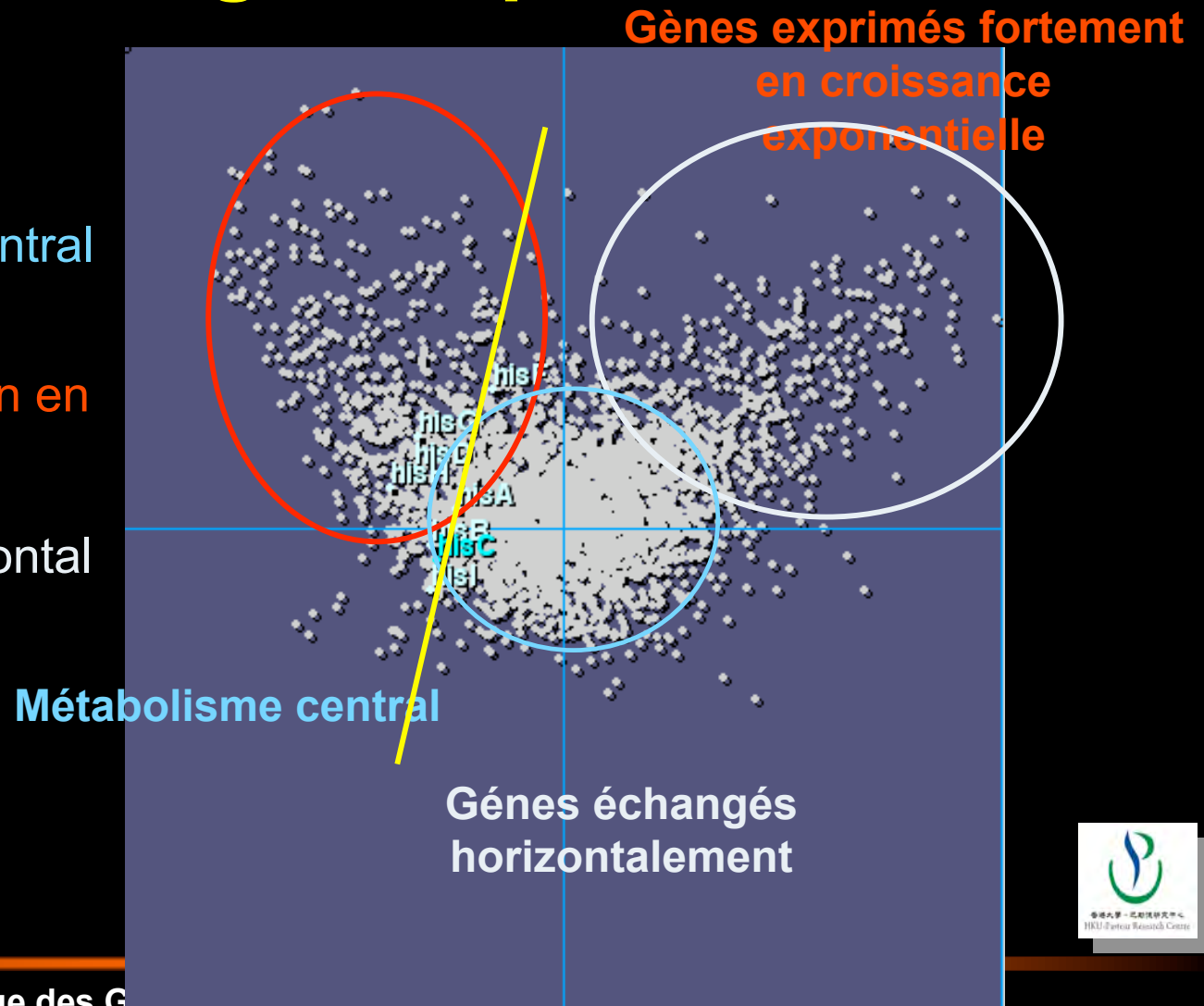


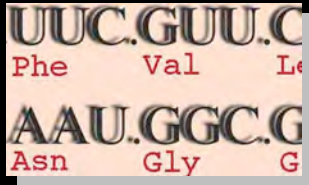
Biais dans l'usage du code génétique

Classe I: métabolisme central

Classe II: forte expression en croissance exponentielle

Classe III: transfert horizontal





Biais locaux de l'usage du code



L'Analyse Factorielle des Correspondances montre que les gènes qui ont un biais similaire dans l'usage des codons sont apparentés fonctionnellement. Comment est-ce que cela se répartit dans le chromosome ?

Une méthode de groupage fondée sur l'analyse des biais d'**usage des codons** au moyen d'une théorie de l'information regroupe les gènes en familles homogènes, qui ne se répartissent pas au hasard dans le chromosome. La méthode permet à la fois d'identifier des biais cohérents et de trouver le nombre pertinent de classes à considérer (4 pour *E. coli* et 5 pour *B. subtilis*)

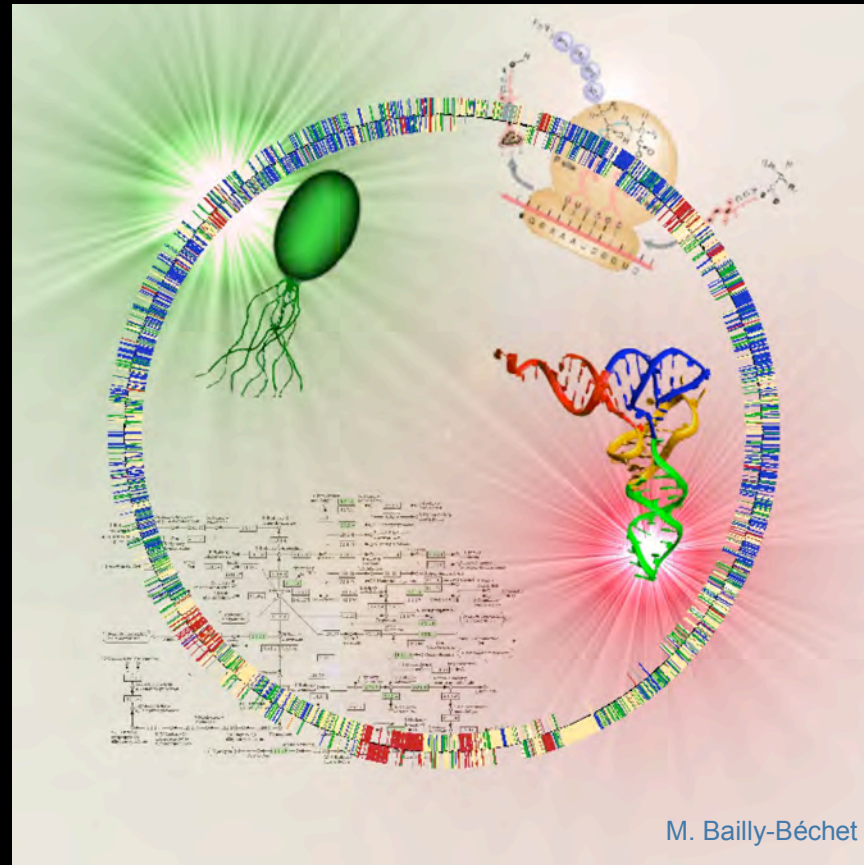


UUC.GUU.C
Phe Val Le
AAU.GGC.G
Asn Gly G

Ilots génomiques

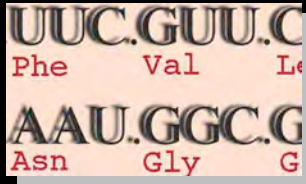


Un groupe correspond au niveau d'expression (bleu). Les autres groupes sont fonctionnellement cohérents : gènes du transfert horizontal (rouge), motilité (jaune) et métabolisme intermédiaire (vert).



M Bailly-Béchet, A Danchin, M Iqbal, M Marsili, M Vergassola
Codon usage domains over bacterial chromosomes
PLoS Computational Biology (2006) 2: april 20th



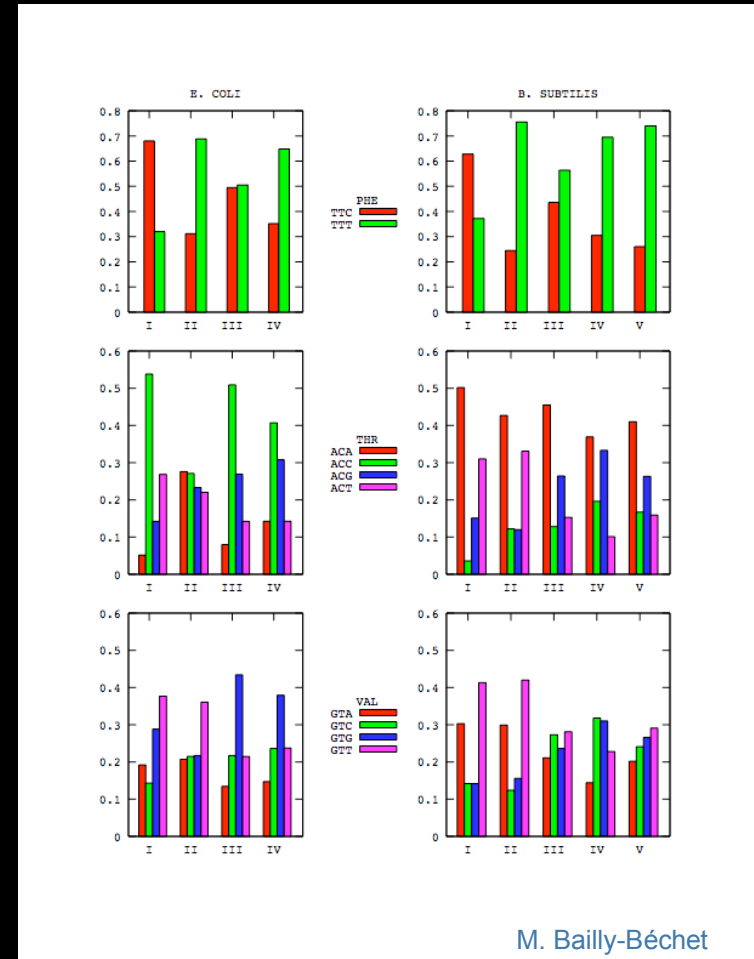


Ilots génomiques



Les gènes qui ont un biais similaire sont organisés en régions plus étendues que les opérons, ce qui démontre un rôle de la traduction dans la structuration du chromosome bactérien.

Une part importante de la contribution à cette effet vient du recyclage des ARN de transfert rares.

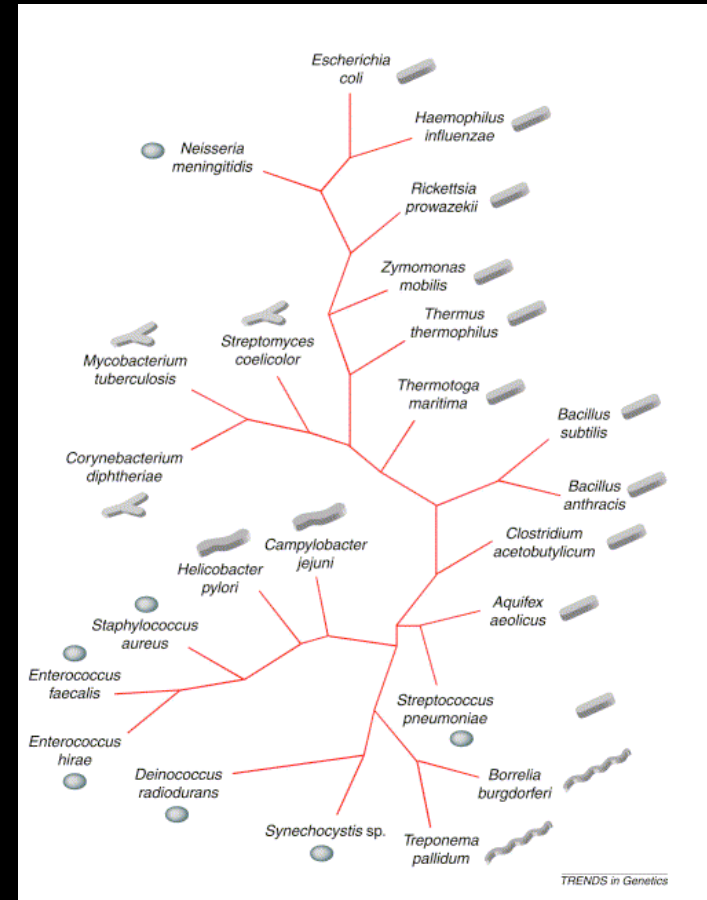
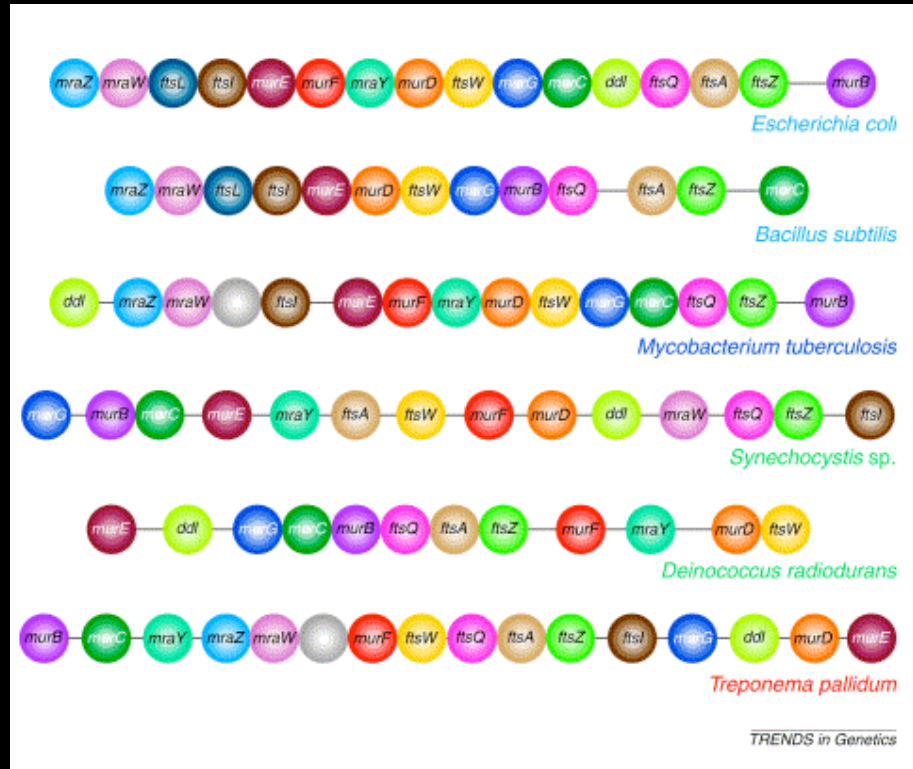


UUC.GUU.C
Phe Val Le
AAU.GGC.G
Asn Gly G

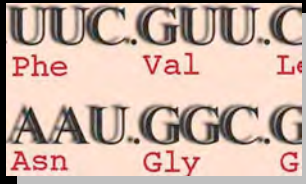
Ordre des gènes et forme des bactéries



L'îlot *mur-fts*



Tamames J, Gonzalez-Moreno M, Mingorance J, Valencia A, Vicente M
Bringing gene order into bacterial shape
Trends in Genetics (2001) 17: 124-126



- **Contexte**
- **Vie et calcul**
- **Du brin précoce au brin retardé**
- **La traduction organise le génome bactérien**
- **Une illustration**
- **Le cœur du génome : ce qui persiste**
- **Le cénome**
- **Le futur: vers la “biologie synthétique”**

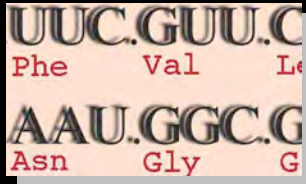


UUC.GUU.C
Phe Val Le
AAU.GGC.G
Asn Gly G



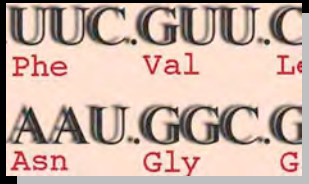
[Non présenté]





- ➔ **Contexte**
- ➔ **Vie et calcul**
- ➔ **Du brin précoce au brin retardé**
- ➔ **La traduction organise le génome bactérien**
- ➔ **Une illustration**
- ➔ **Le cœur du génome: ce qui persiste**
- ➔ **Le cénome**
- ➔ **Le futur: vers la “biologie synthétique”**





La première découverte de la génomique

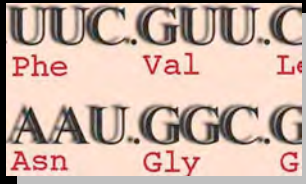


En 1991, à la réunion de l'UE sur les programmes génomes à Elounda, en Grèce, la présentation du chromosome III de la levure et des premiers 100 kb du génome de *Bacillus subtilis* a révélé que, au contraire de ce qu'on pensait jusqu'alors, **au moins la moitié des gènes découverts étaient totalement inconnus, que ce soit par leur fonction ou la structure de leur produit**

Plusieurs raisons rendent compte de ce fait, toujours vrai aujourd'hui ; d'une part notre connaissance du métabolisme est très imparfaite, d'autre part nous ne savons pas comment se créent de nouveaux gènes ; enfin, l'évolution procède par la **sélection de fonctions**, en **recrutant des structures** dont l'adaptation va s'accroître en parallèle avec l'adéquation de l'individu à son environnement (**évolution acquise**)

S'il y a tant de nouveaux gènes, existe-t-il cependant un socle commun ?





Une contrainte universelle, l'évolution



Variation / Sélection / Amplification



Evolution



crée

Fonction



capture (recrute)

Structure



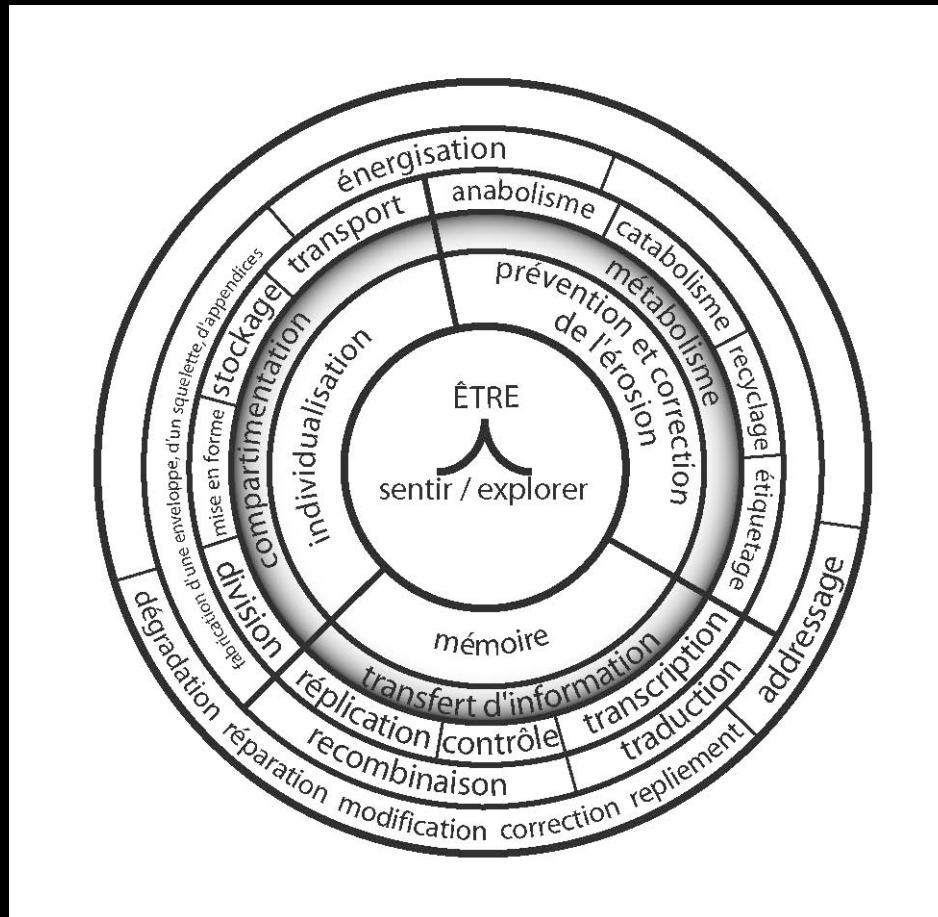
code

Séquence



UUC.GUU.C
Phe Val Le
AAU.GGC.G
Asn Gly G

Fonctions pour la vie





Gènes persistants



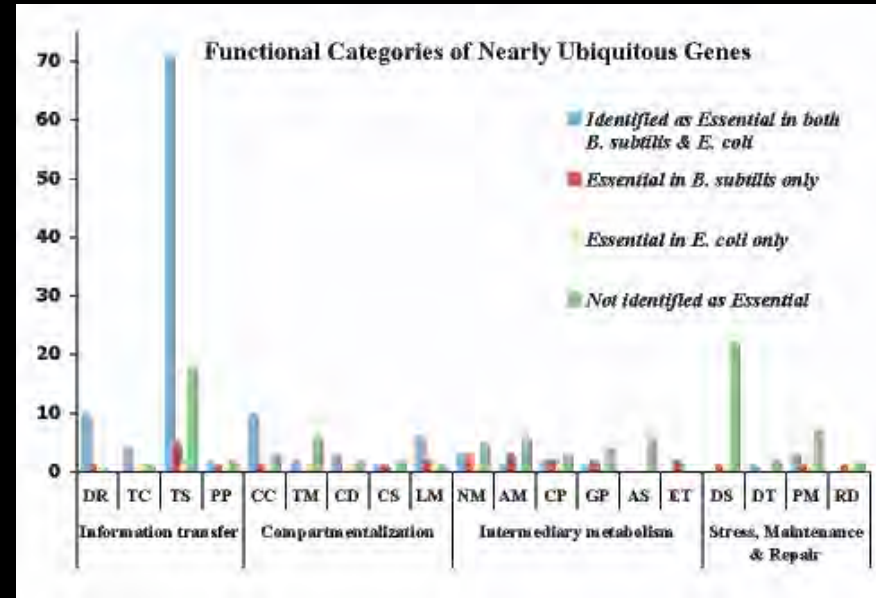
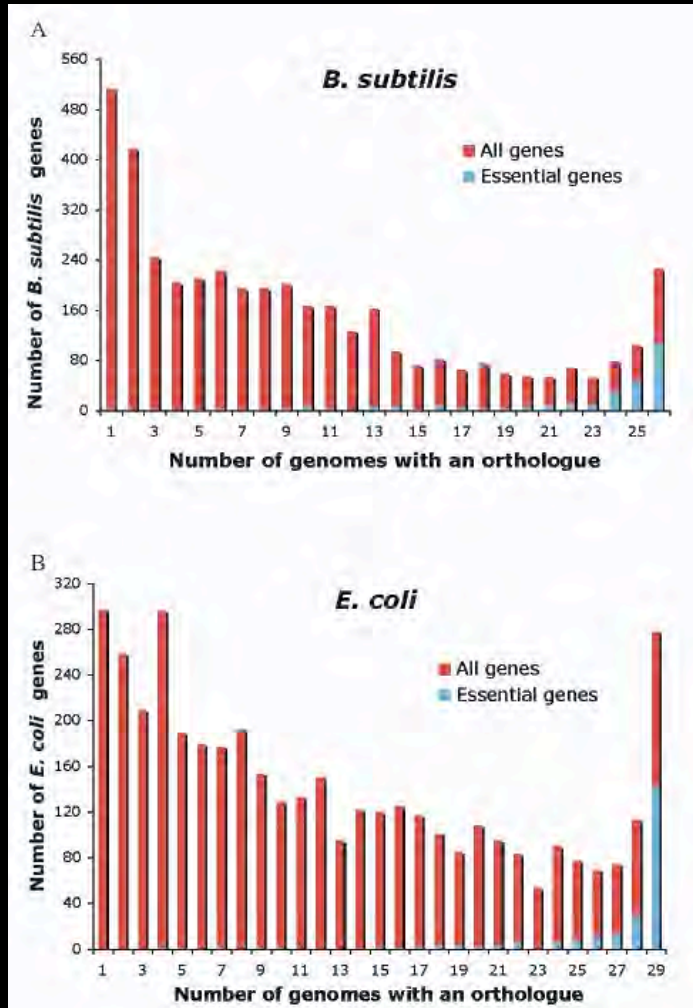
Les gènes essentiels en laboratoire sont situés dans le brin précoce. Ils sont aussi conservés dans une majorité de génomes. A l'inverse les gènes qui sont conservés et se trouvent dans le brin précoce forment une catégorie supplémentaire, qui double pratiquement celle des gènes essentiels

Ces gènes forment des **universaux** ; 400-500 gènes persistent dans un grand nombre de génomes ; ils sont impliqués non seulement dans les trois processus nécessaires à la vie, mais dans la **maintenance** et dans l'**adaptation aux phénomènes transitoires** ; une fraction gère l'**évolution** de l'organisme

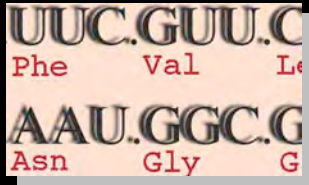




Persistance des gènes



- Transfert de l'Information
- Compartimentation
- Métabolisme intermédiaire
- Stress, Maintenance et Réparation



Phylogénie de la persistance

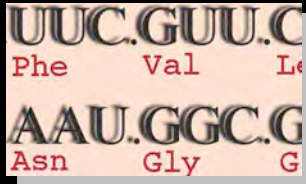


Quelques uns des gènes essentiels absents de la liste ont évolué particulièrement rapidement

Pour mesurer la contribution de cet effet, on la met en parallèle avec l'évolution de l'ARN ribosomique 16S

Deux scénarios distincts apparaissent : dans 85% des cas une évolution linéaire, et dans 15% des cas une évolution erratique





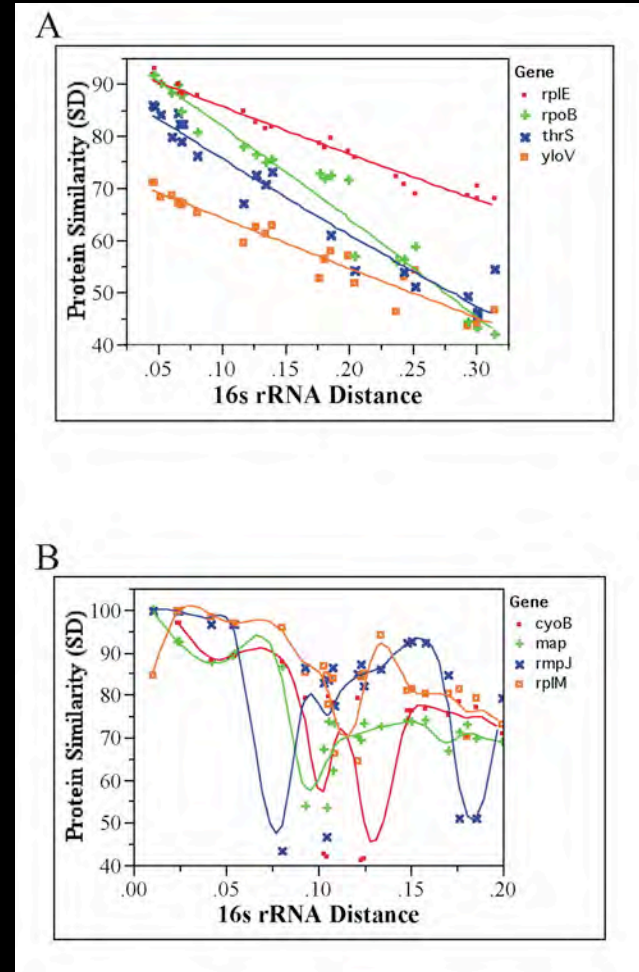
Persistance des gènes

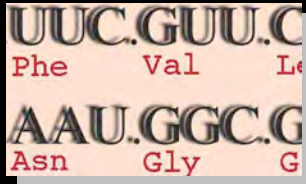


Par exemple (A), 38% (resp. 48%) of *B. subtilis* (resp. *E. coli*) persistent genes show a correlation coefficient >0.9 la similitude de séquence de paires d'orthologues et l'ARN 16S.

Au contraire, certains gènes (B) évoluent de façon erratique. Ce peut être dû au transfert génétique horizontal, à des adaptations particulières, ou tout simplement à des erreurs dans l'attribution de l'orthologie. Ce dernier problème est important dans les grandes familles de gènes apparentés. Cependant les gènes ayant cette propriété sont une minorité (un sixième) des gènes persistants.

G Fang, EPC Rocha, A Danchin
How essential are non-essential genes?
Mol Biol Evol (2005) 22: 2147-2156





Quelles fonctions pour la vie ? Un scénario de l'origine de la vie



- La surface de solides chargés comme la pyrite de fer (Fe-S) permet la sélection et une compartimentation primitive de molécules chargées ; la polymérisation avec élimination d'eau est favorisée par effet entropique
- Une fois compartimenté, le métabolisme crée des substituts des surfaces (le monde ARN)
- L'exploration, associée à la perception et à la mémoire (transfert d'information transfer) est la découverte qui a fait la vie telle que nous la connaissons

A Danchin. Une Aurore de Pierres. Aux origines de la vie, Le Seuil 1990



Génétique des Génomes Bactériens

<http://www.pasteur.fr/recherche/unites/REG/causeries/causeries.html>

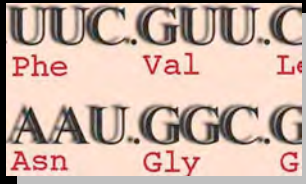


Les gènes persistants sont regroupés



Avec l'analyse de 228 génomes codant chacun plus de 1500 gènes et des annotations « correctes », on peut identifier les gènes persistants qui tendent à rester à proximité les uns des autres ; cette « attraction mutuelle » constitue un réseau remarquable fait de trois cercles concentriques qui regroupent les fonctions identifiées dans ce scénario de l'origine de la vie : un métabolisme compartimenté (à la surface des pierres) suivi de l'invention d'un substitut des surfaces, l'ARN, puis de l'usage des acides nucléiques pour gérer les interactions avec l'environnement au travers de l'information correspondante



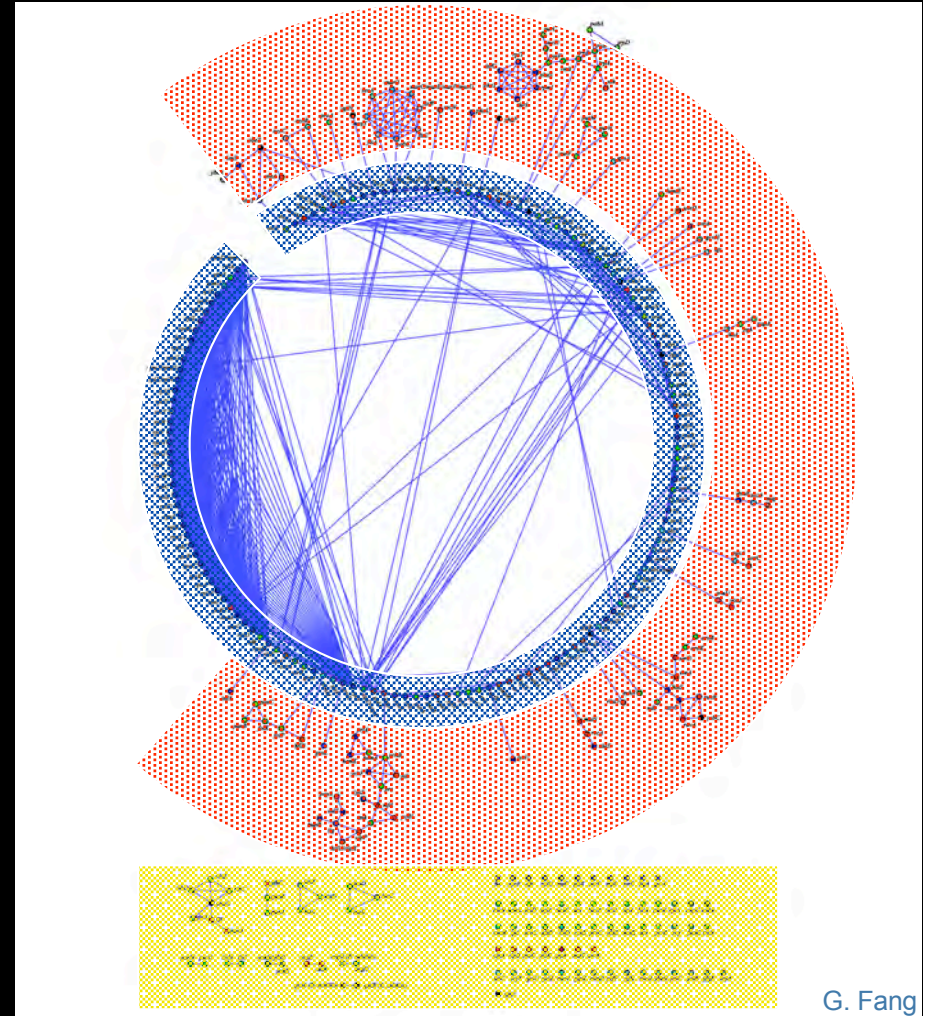


Les gènes persistants récapitulent l'origine de la vie

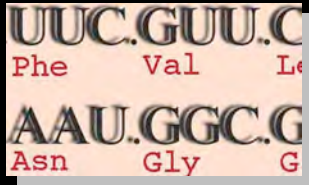


Le **réseau externe**, fait de gènes du métabolisme intermédiaire (nucléotides et coenzymes, lipides), est très fragmenté ; le **réseau médian** est formé autour des ARNt synthétases de classe I, et le **réseau intérieur**, presque continu, organisé autour du ribosome, de la transcription et de la réplication gère les transferts d'information

Ce réseau est compatible avec un scénario où les coenzymes et les maillons élémentaires des acides nucléiques et des protéines ont conduit à un métabolisme dont les support était l'ancêtre des ARN de transfert, suivi par un monde ARN où apparaît la loi de complémentarité et le concept de matrice, puis le code génétique



G. Fang



« Etre » implique le regroupement

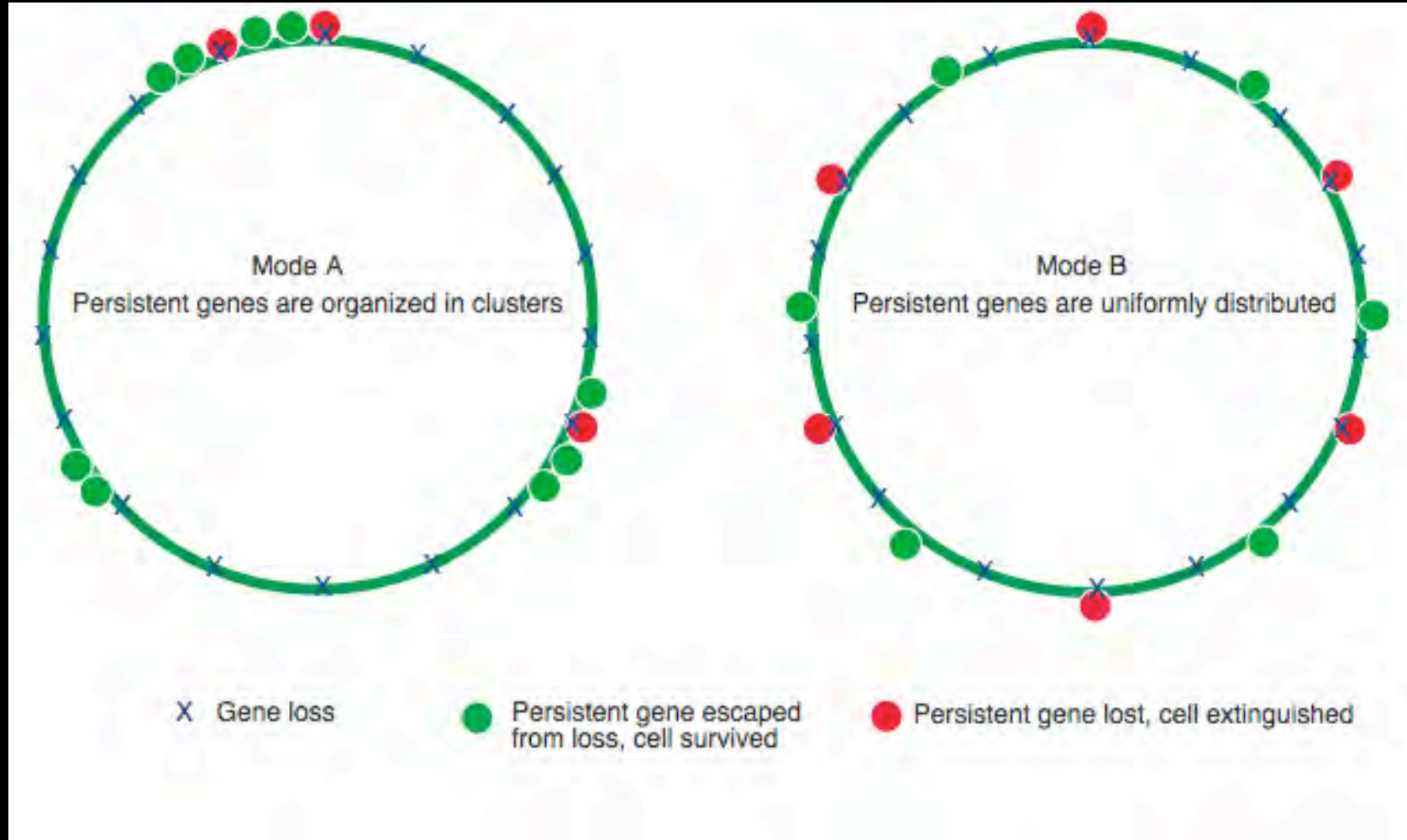


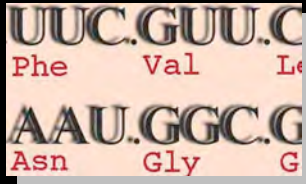
Pourquoi les gènes persistants sont-ils regroupés ? Un modèle très simple montre que si, pour compenser le transfert horizontal des gènes, les gènes tendent à disparaître en groupes, alors tout gène contribuant à l'adéquation de l'organisme à son environnement assez fréquemment tendra à se regrouper avec les gènes ayant la même propriété. Cela explique le regroupement des gènes essentiels, mais peut-être aussi celui des gènes de résistance aux antibiotiques....

En conséquence le regroupement **précède et ne résulte pas de la** co-transcription ou des interactions protéine-protéine !



L'existence implique le regroupement





- **Contexte**
- **Vie et calcul**
- **Du brin précoce au brin retardé**
- **La traduction organise le génome bactérien**
- **Le cœur du génome: ce qui persiste**
- **Une illustration**
- **Le cénome**
- **Le futur: vers la “biologie synthétique”**

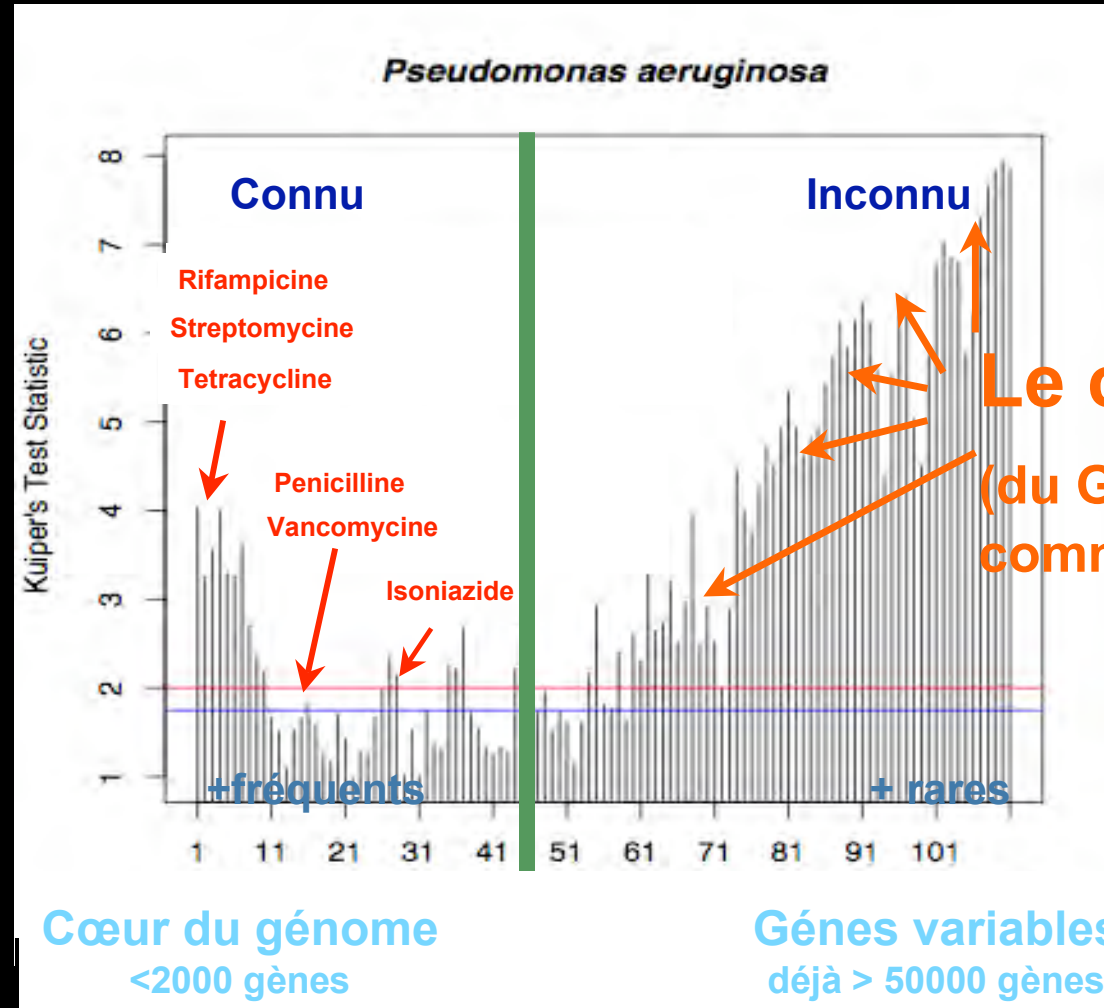


UUC.GUU.C
Phe Val Le
AAU.GGC.G
Asn Gly G

Le cénome



Fréquence de liaison



Le cénome
(du Grec κοινος, communauté)

Fréquence dans les génomes

Antibiotiques

Virulence

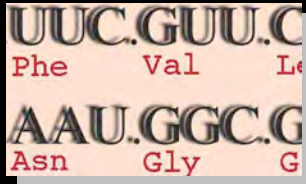


UUC.GUU.C
Phe Val Le
AAU.GGC.G
Asn Gly G



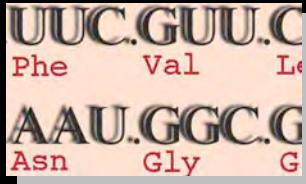
[Non présenté]





- **Contexte**
- **Vie et calcul**
- **Du brin précoce au brin retardé**
- **La traduction organise le génome bactérien**
- **Une illustration**
- **Le cœur du génome: ce qui persiste**
- **Le cénome**
- **Le futur: vers la “biologie synthétique”**





La biologie synthétique



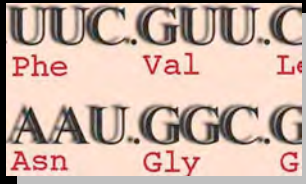
→ Comprendre et conserver les principes et les composants de base

- Encore réduire le plus petit génome : vers le génome minimal (*Mycoplasma genitalium*, Smith, Venter)
- Reconstruire ce qu'on connaît (repressilator, Leibler)
- Réécrire en simplifiant (génie logiciel) (T7, Endy ; promoteurs, Hwa)
- Reprogrammer : l'usine cellulaire (HCV, Liang ; carbone, Marlière)
- Reconstruire une cellule ancestrale « Jurassic Park moléculaire » (protéines ancestrales, Benner)
- ...

→ Conserver les principes et introduire une nouvelle chimie

- Changer les acides aminés (Cohen, Wong...)
- Changer le code génétique (Schultz, Marlière)
- Changer la chimie du support de l'hérédité (S-2L, PNA, TNA, GNA, etc)
- ...

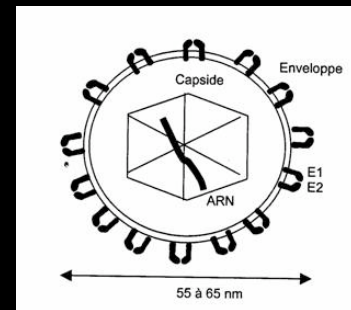




Reprogrammer

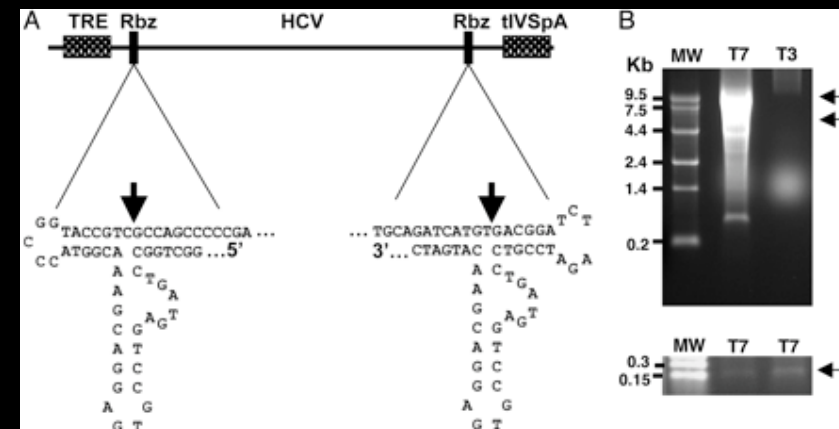


→ Le virus de l'hépatite C comporte un **ARN** à brin unique de 9600 nucléotides, codant directement une protéine multifonctionnelle de 3010 acides aminés qui est secondairement scindée en au moins 10 protéines tardives de maturation virale. **Fragile, il se réplique mal**



Pr. Amine SLIM Laboratoire de Microbiologie - CHU Charles Nicolle Tunis

→ Par génie génétique Liang et collègues ont construit un **ADN** autorépliatif codant l'ARN du virus, entouré de deux régions capables sous forme d'ARN de se cliver spontanément en libérant l'ARN viral. **Cette protection/déprotection, entièrement artificielle, donne lieu à la production continue de virus normalement infectieux**



Heller T et al. An in vitro model of hepatitis C virion production. Proc Natl Acad Sci U S A. 2005 102:2579-2583.