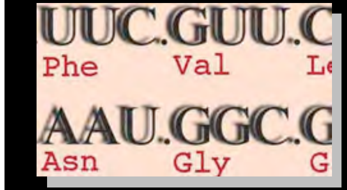
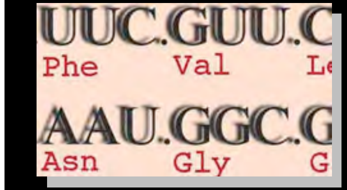


From Symplectic Biology to Synthetic Biology:

Universals in Bacterial Genomes

Σπετσες 4 september 2006

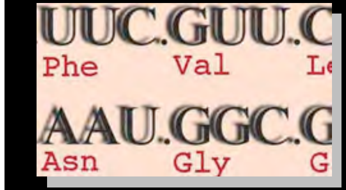
Génétique des Génomes Bactériens
http://www.pasteur.fr/recherche/unites/REG/



Ουδεν χρημα ματην γινεται αλλα
παντα εκ λογου τε και υπ’αναγκης
ΛΕΥΚΙΠΠΟΣ

No thing comes by itself [and without cause] but everything comes from a reason (logos) and is under the constraint of necessity

LEUCIPPUS



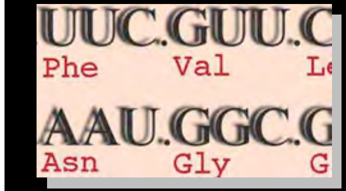
A twenty years old revolution: genome projects

2127 ongoing projects, 354 completed, mostly from microbes (228 with more than 1500 genes, more or less correctly annotated)

148,116,054,623 nucleotides at International Nucleotide Sequence Database Collaboration (INSDC)

Microbes make 50% of the Earth protoplasm

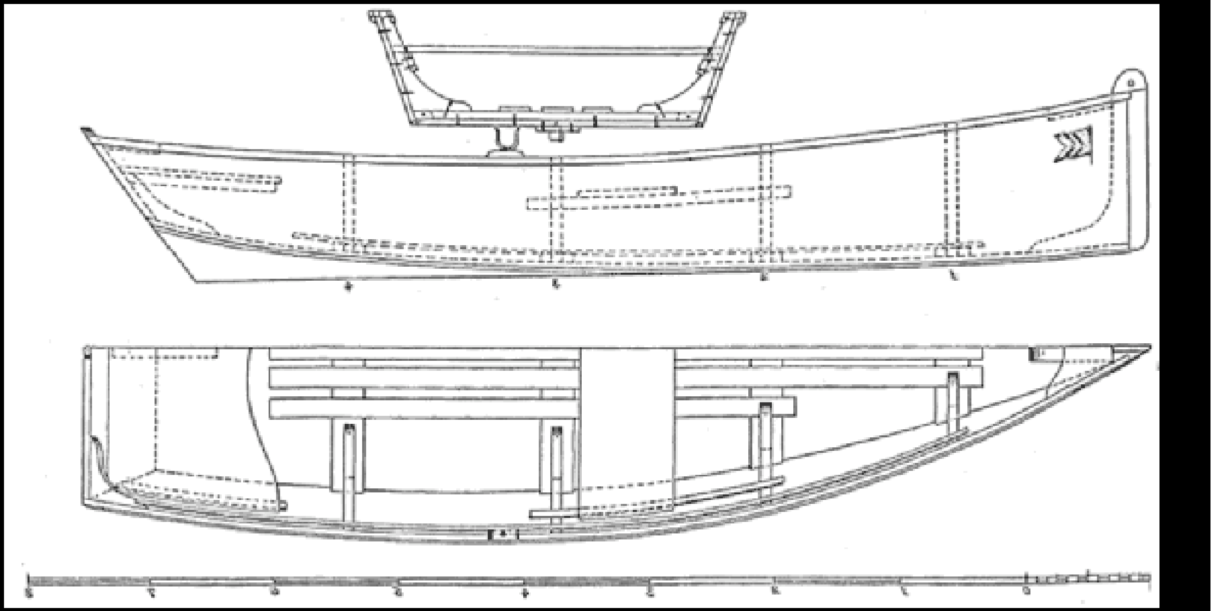
40-50% coding DNA sequences (CDSs) do not correspond to known functions; 10% correspond to the core genome (« persistent » genes)



A coming revolution: Synthetic Biology

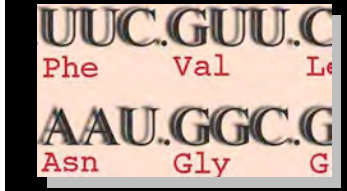
- However we need to remember that biology is a science of relationships between objects rather than of objects:

  it is **symplectic** (from συν together, πλεκτειν, to weave)
- As for constructing a boat, failing to understand their relationships will result in ultimate failure of synthetic biology
- If there is no intelligent design, how are relationships created?



The Delphic Boat: Harvard University

Press, février 2003

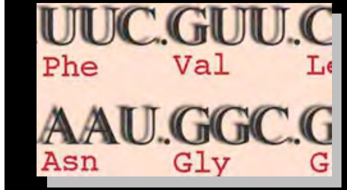


UUC.GUU.C
Phe Val Le
AAU.GGC.G
Asn Gly G

Context: the “genetic program”

- **Physics:** ***matter, energy, time***
- **Statistical physics:** **Physics +** ***information***
- **Biology:** **Physics +** ***information, coding, control...***
- **Arithmetics:** ***sequence of integers, recursivity, coding…***
- **Computation:** **Arithmetics +** ***program + machine...***

A metaphor with practical consequences, that of the genetic program: we know how to manipulate the genes and their products, can we push the metaphor to its ultimate consequences?



What life is

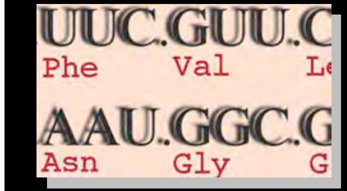
Three processes are needed for Life:

- **Information transfer (Living Computers?) => the goal of genomics is to decipher the program associated to the machine and its meaning**

Driving force for a coupling between the genome structure and the structure of the cell (not discussed today):

- **Metabolism**
- **Compartmentalisation**

The cell is the atom of life, with two compartmentalisation strategies: a single envelope (prokaryotes), or multiplication of membrane and skins (eukaryotes); **remarkably, this is correlated with the genome sequence: at first sight prokaryotic genomes look random** and **eukaryotic genomes look repeated**



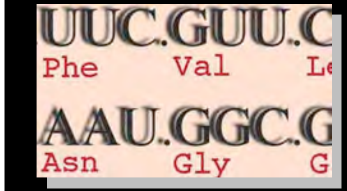
An algorithmic view of the biological action processes

Replication, transcription, translation: high parallelism

“Beginning, Repeated Routine and Check Points, End”

The action is always oriented, with a beginning and an end

The control process of Check Points is rarely taken into account in present research (except in replication/division), but its role is essential to permit coordination of multiple actions in parallel



What is computing?

Two processes are needed for computing:

- A read/write machine

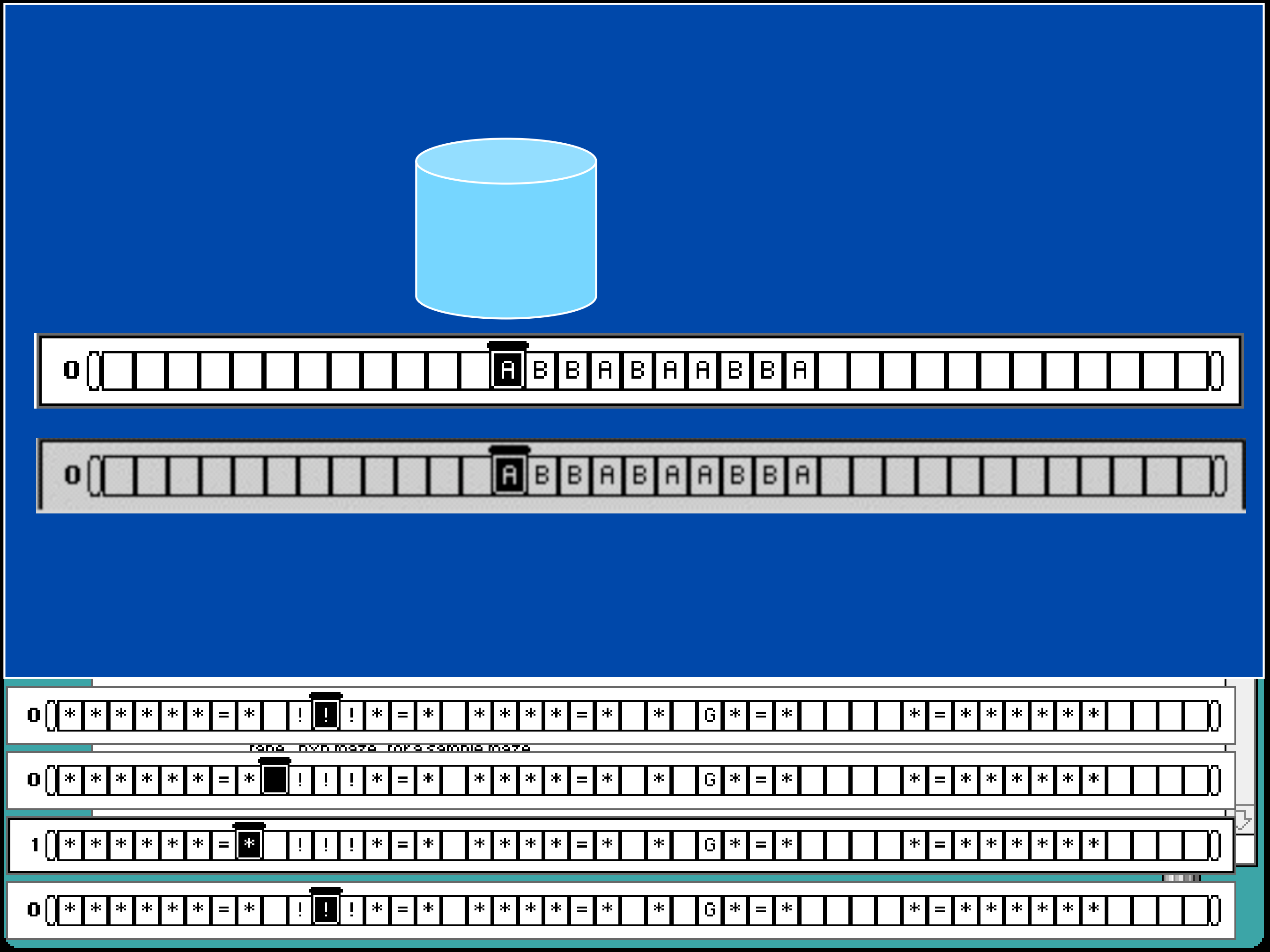
- A program on a physical support (typically, a tape illustrates the sequential string of symbols that makes up the program), split (in practice) into two entities:
  - Program (providing the goal)
  - Data (providing the context)

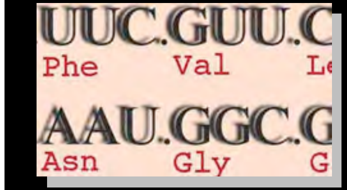
The machine is distinct from the program

A Turing machine

machine
(read/write)

programme
(data)


0 A B B A B A A B B A
0 A B B A B A A B B A
0 * * * * * * = * ! ! ! * = * * * * * = * * G * = * * = * * * * * *
0 * * * * * * = * ! ! ! * = * * * * * = * * G * = * * = * * * * * *
1 * * * * * * = * ! ! ! * = * * * * * = * * G * = * * = * * * * * *
0 * * * * * * = * ! ! ! * = * * * * * = * * G * = * * = * * * * * *



Cells as computers

Genetic studies rest on the description of genomes as texts written with a four letters alphabet: do cells behave as computers?

Horizontal Gene Transfer

Virus

Genetic engineering => reconstruction of the hepatitis C virus

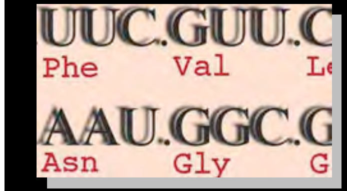
Animal cloning

all point to separation between

A « Machine » (the cell factory)

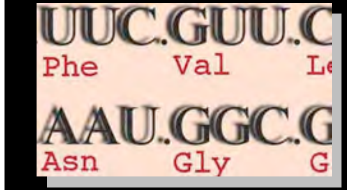
and

Data + Programme



A research program: is there a map of the cell in the chromosome?

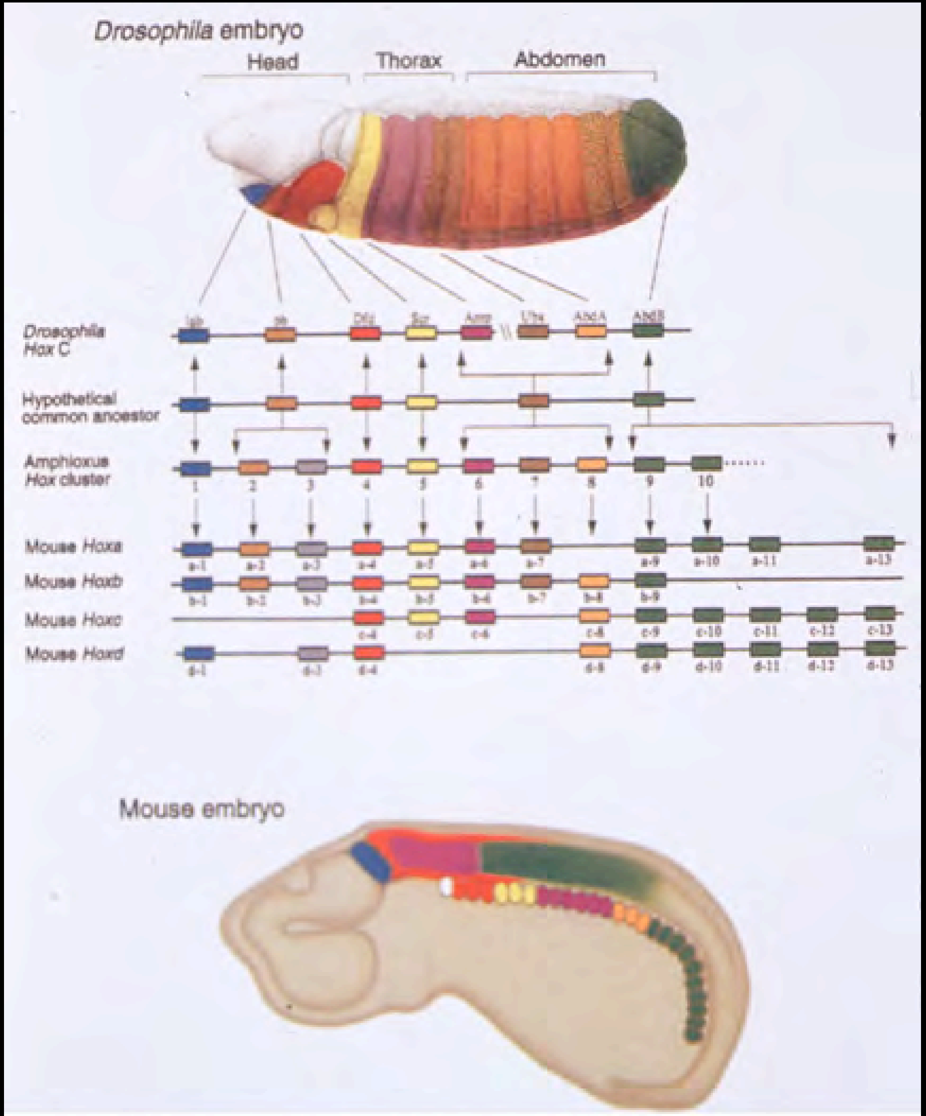
If the machine has not only to behave as a computer but has also to construct the machine itself, one must find an image of the machine somewhere in the machine (John von Neumann)



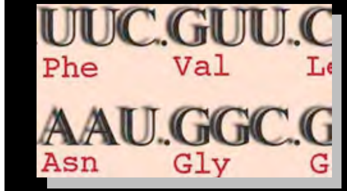
Drosophiloculus,

Homunculus?

Celluloculus?



From Carroll, S. B. (1995) Nature 376, 479



Genome organisation

Is the gene order random in the chromosomes?

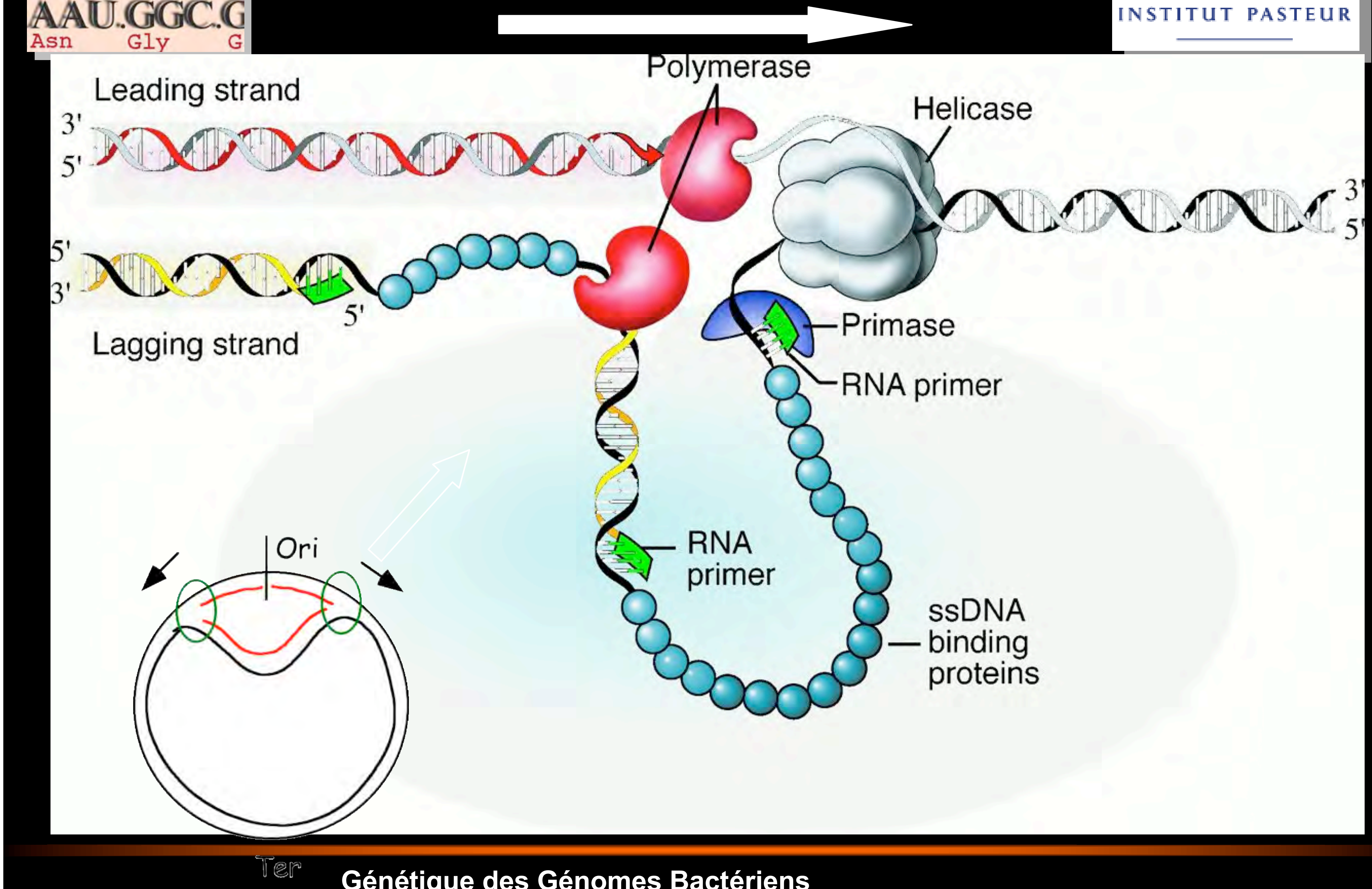
At first sight, consistent with different DNA management processes in different organisms not much is conserved, while genes transferred from other organisms are distributed throughout genomes

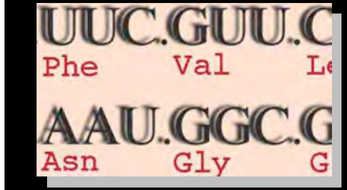
However, groups of genes such as operons or pathogenicity islands tend to cluster in specific places, and they code for proteins with common functions. « Persistent » genes are clustered together

Also, some « flexible » motifs in DNA generate a 10.5-12 bp autocorrelation period. They are ubiquitously present, suggesting general rules constraining genome organisation

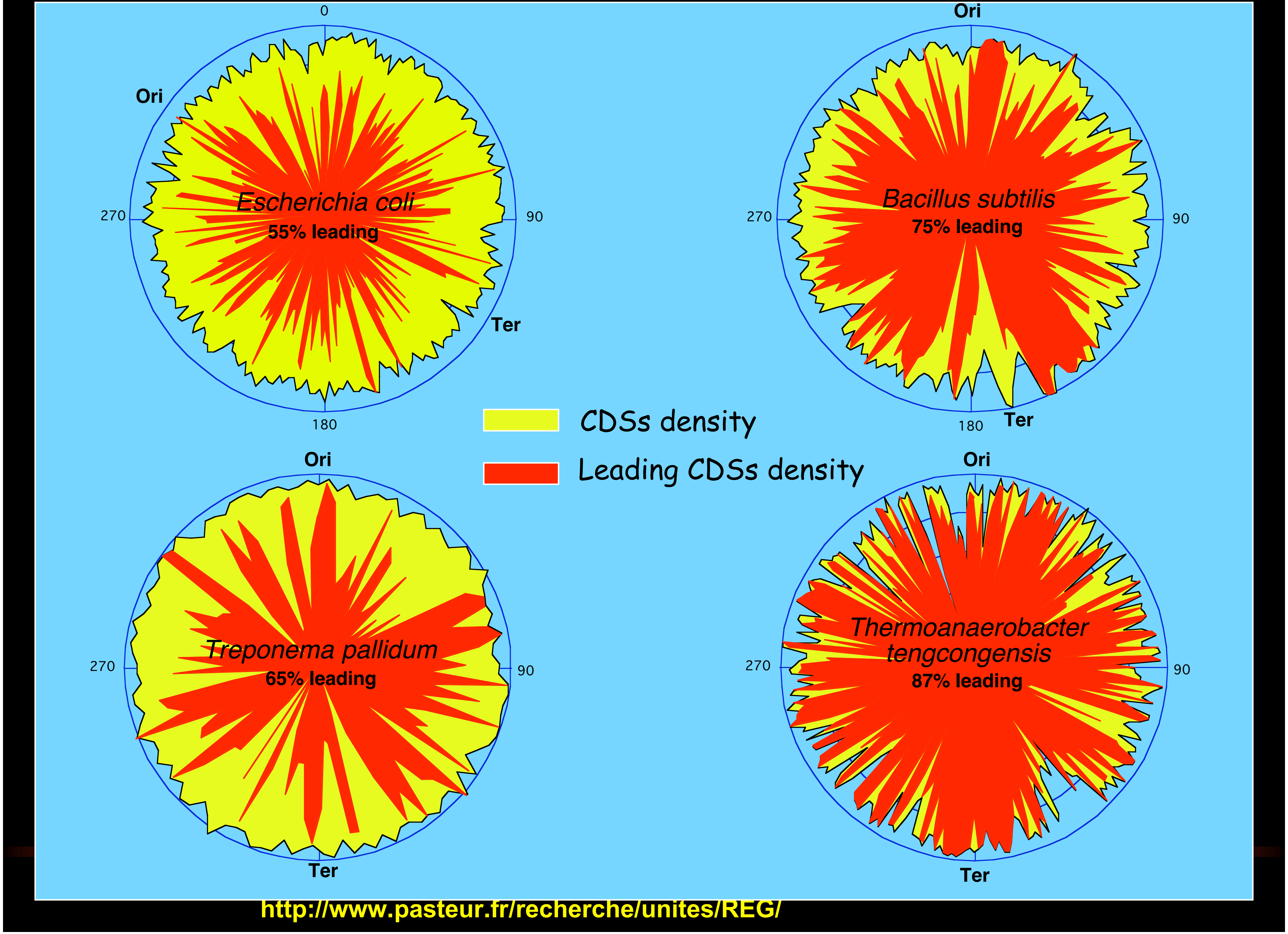
E Larsabal, A Danchin
Genomes are covered with ubiquitous 11bp periodic patterns, the "class A flexible patterns"
BMC Bioinformatics (2005) **6**: 206

From the leading strand to the lagging strand

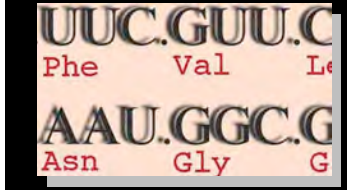

Polymerase
Leading strand
3'
5'
Helicase
3'
5'
5'
3'
5'
Lagging strand
Primase
RNA primer
RNA primer
ssDNA binding proteins
Ori
Ter



Genes are preferentially located in the leading replication strand in Bacteria. There is however much variation, depending on the organism, with a considerable bias in A+T-rich Gram-positive organisms



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

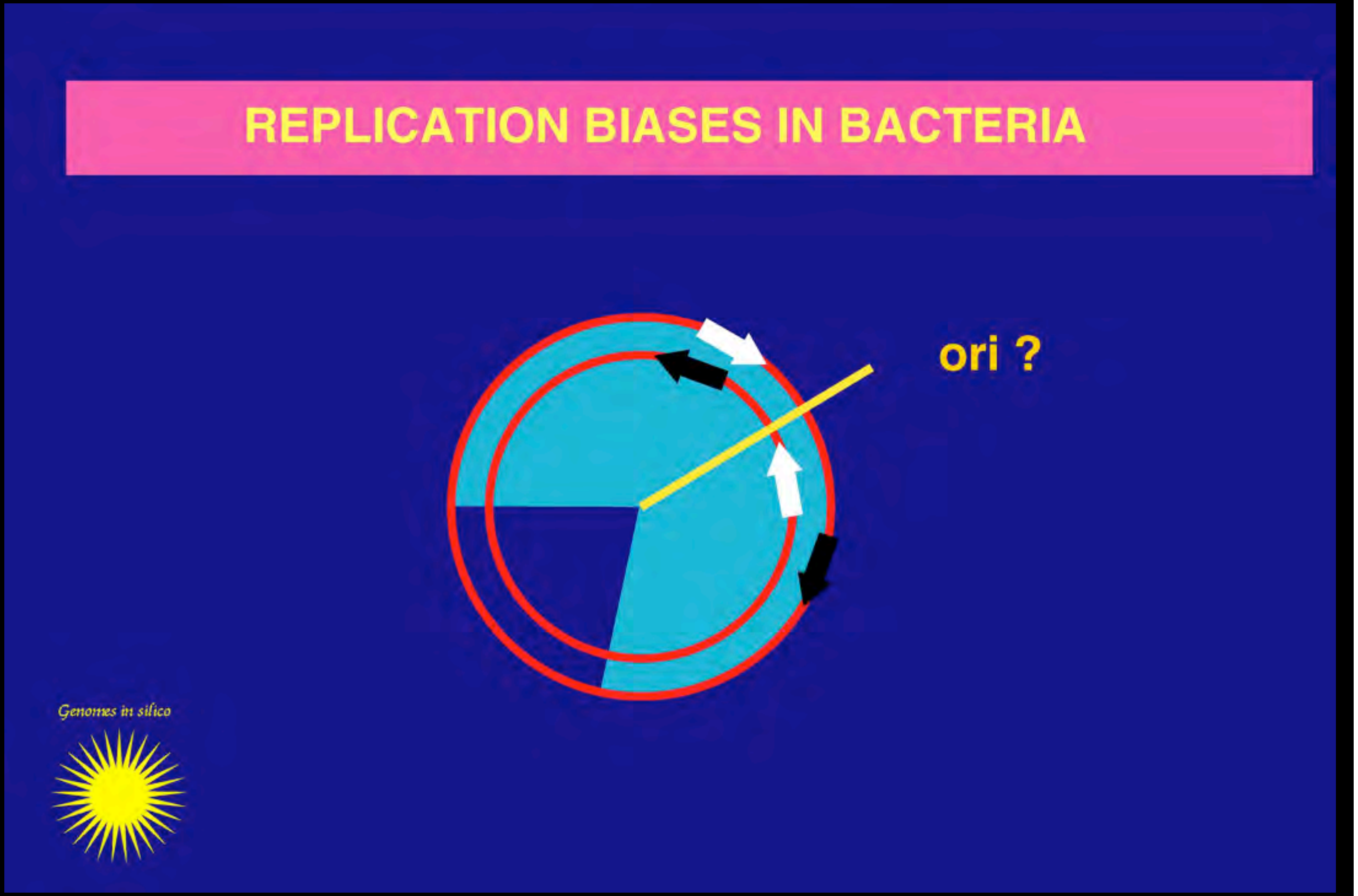


To lead or to lag...

Is it possible to see whether there is a difference in the nucleotide composition, between the leading and the lagging strand? Does that have a consequence on the codon biases? Does that have a consequence for the protein amino acid sequence?

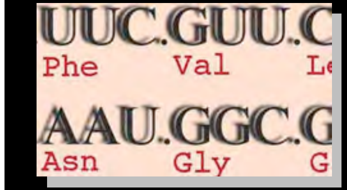
To lag or to lead...

Chosing arbitrarily an origin of replication and a property of the strand (base composition, codon composition, codon usage, amino acid composition of the coded protein…) one can use discriminant analysis to see whether the hypothesis holds.

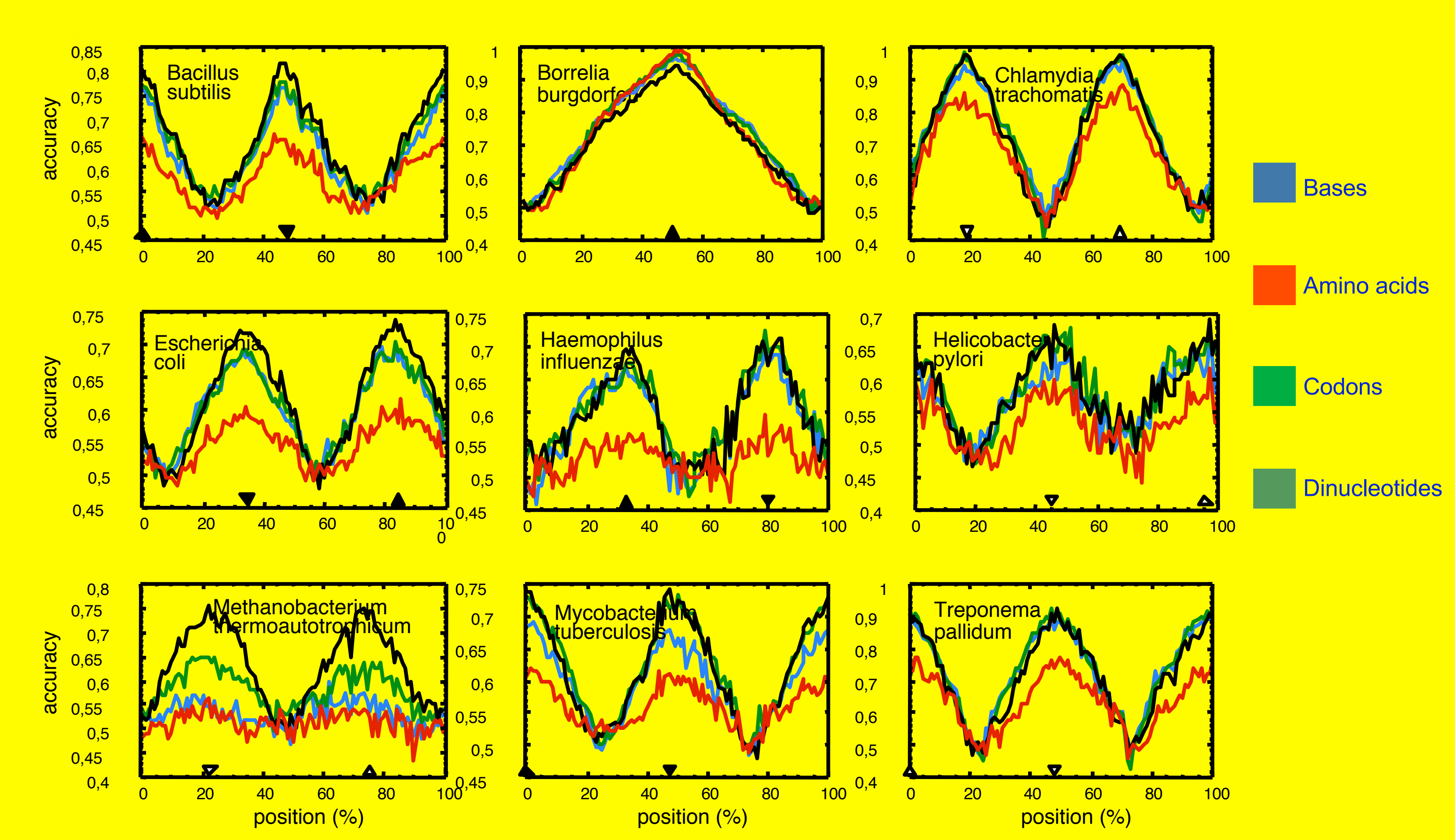


REPLICATION BIASES IN BACTERIA

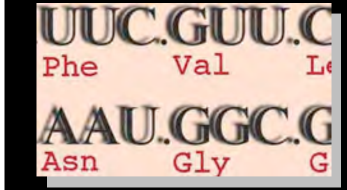
E. Rocha, A. Danchin & A. Viari Universal replication biases in bacteria. Mol. Microbiol. (1999) 32: 11-16



That is the question…



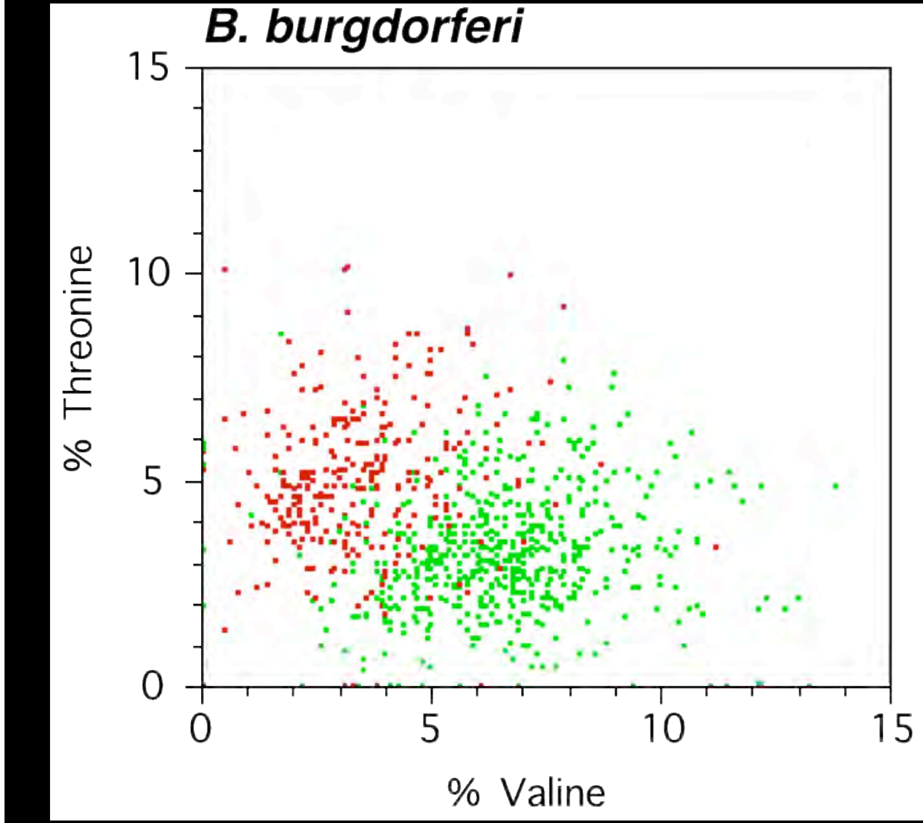
Génétique des Génomes Bactériens
http://www.pasteur.fr/recherche/unites/REG/

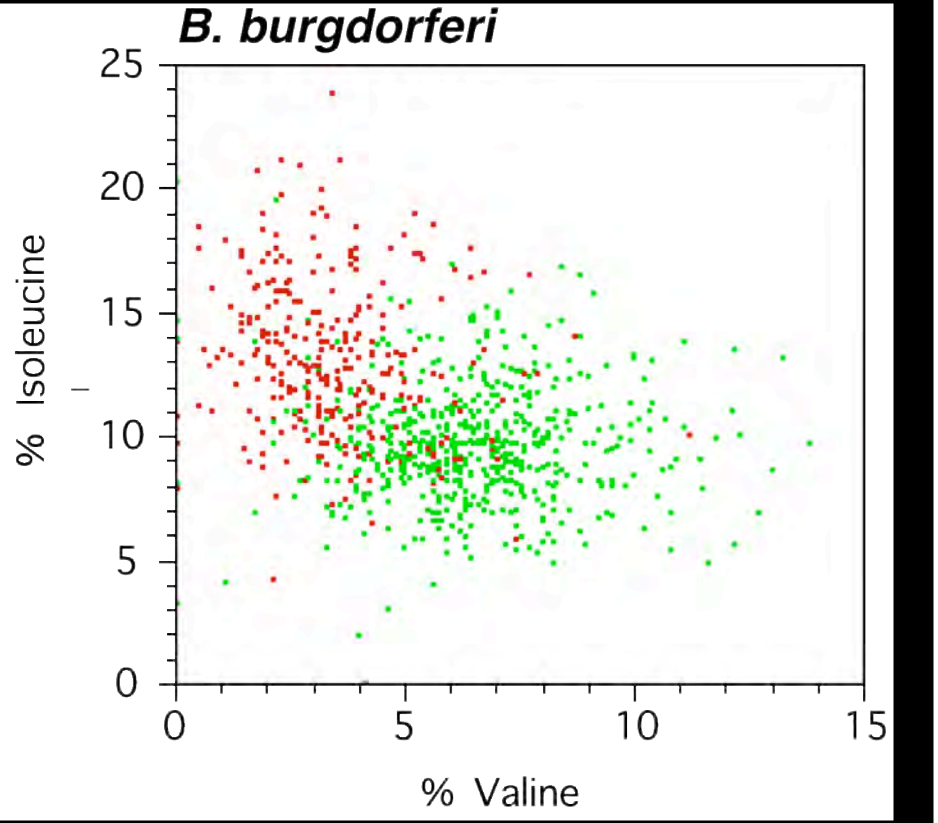


INSTITUT PASTEUR

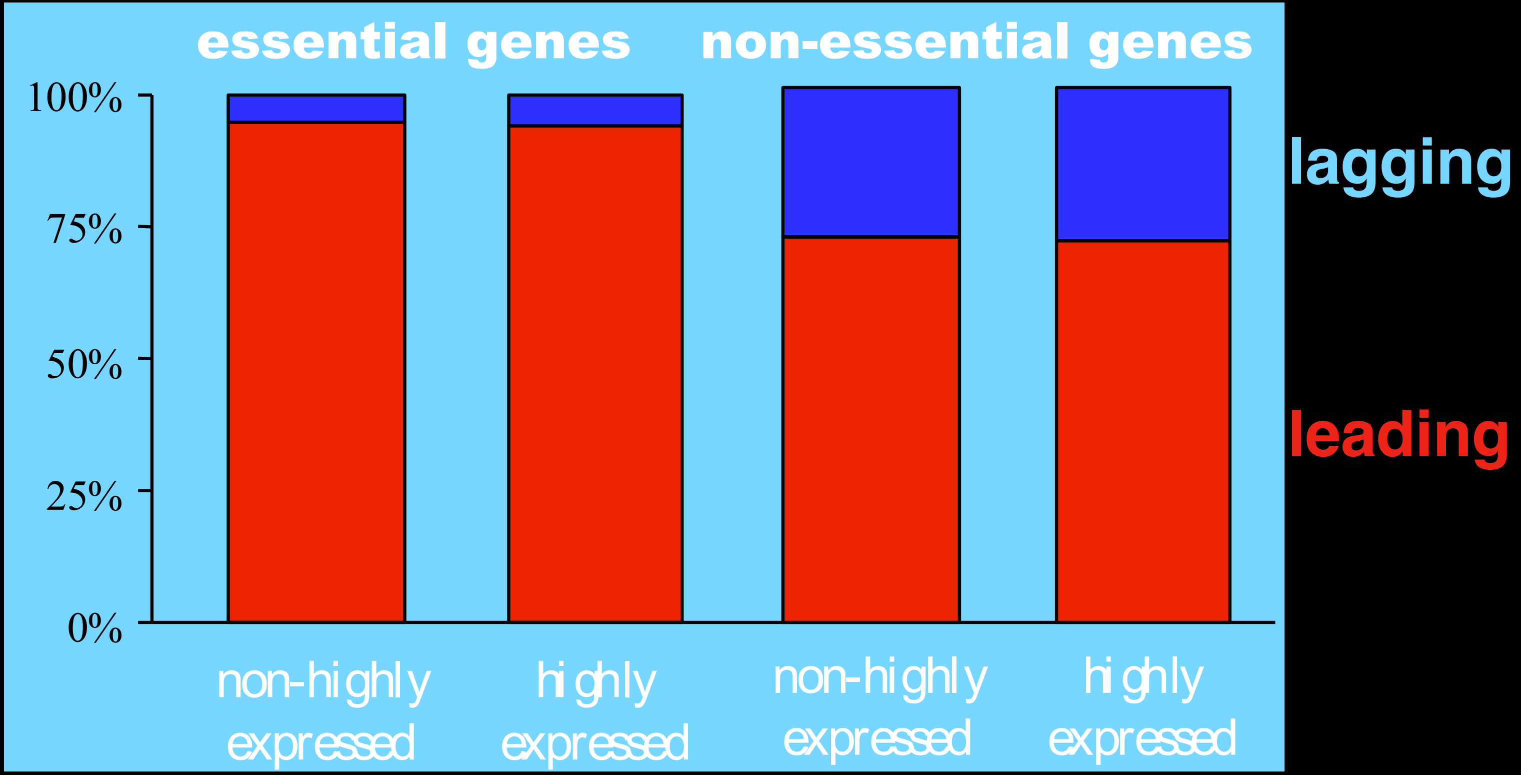
Visible in proteins…

GT on the leading strand, CA on the lagging strand...





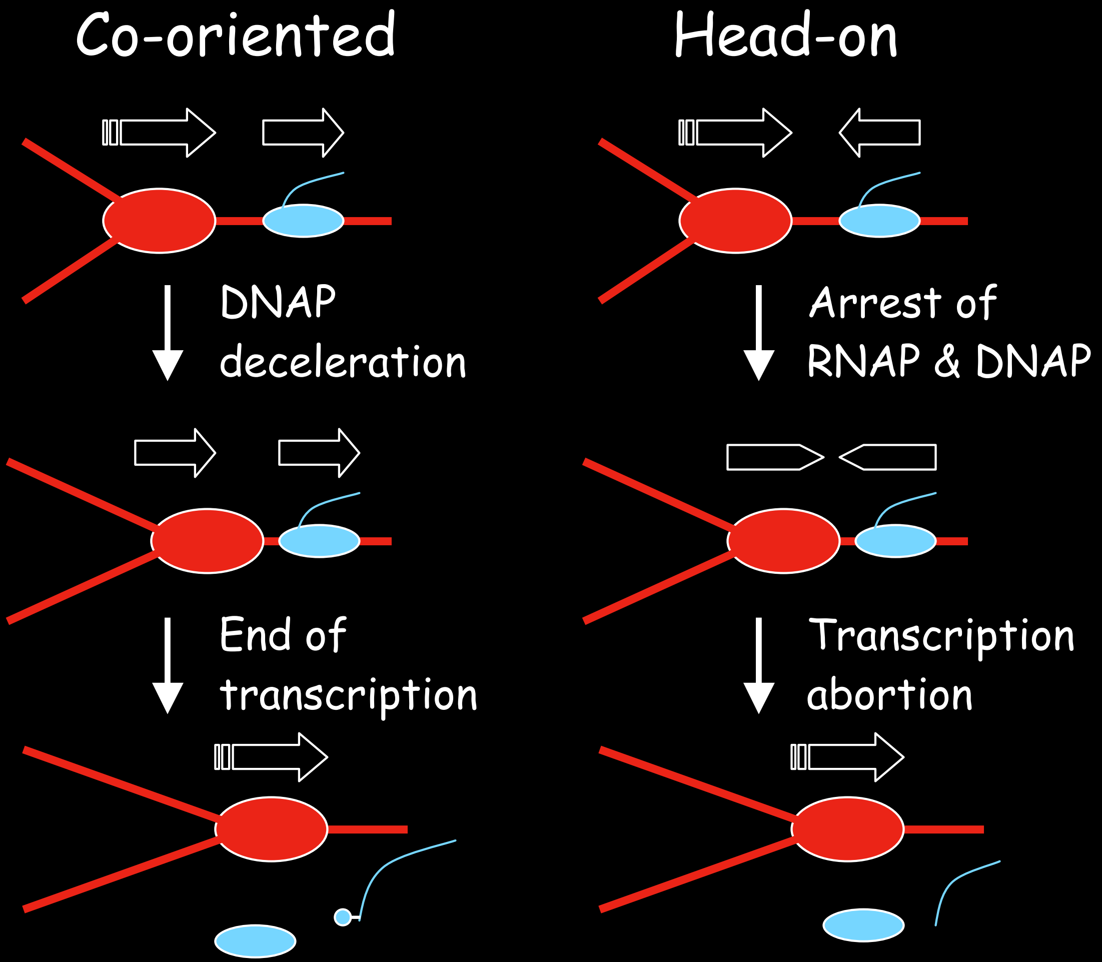
Essential genes locate in the leading strand



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

When polymerases collide

INSTITUT PASTEUR



Co-oriented

DNAP deceleration

End of transcription

Head-on

Arrest of RNAP & DNAP

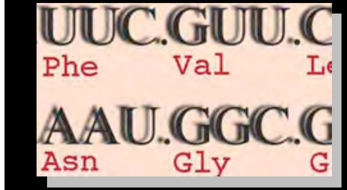
Transcription abortion

~~Consequences:~~

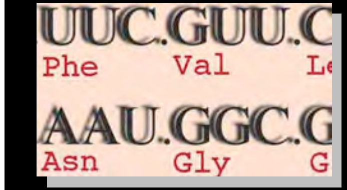
1. ~~Replication slow-down~~
2. ~~Loss of transcripts~~

Consequences:

1. Aborted transcripts
2. Truncated essential proteins



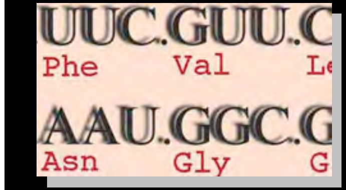
From function to structure



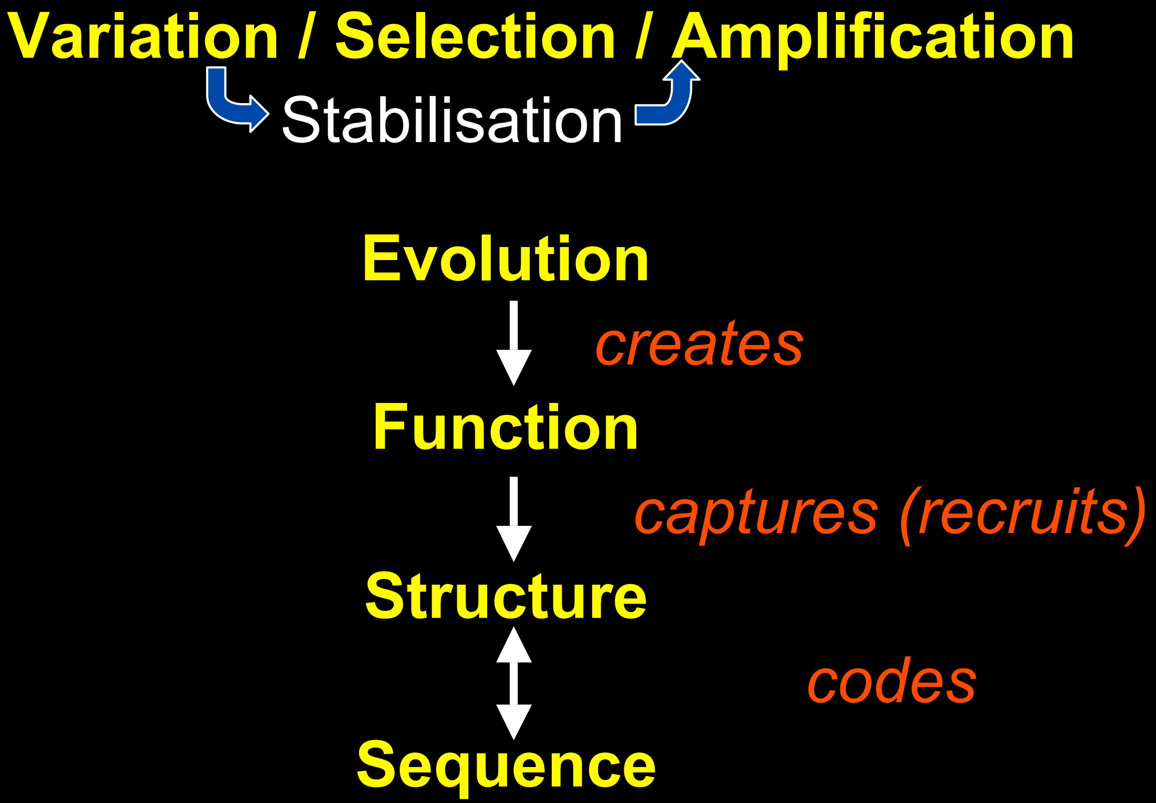
The first discovery of genomics

In 1991, at the EU meeting on genome programs in Elounda, Greece, the presentation of the yeast chromosome III and the first 100 kb of the *Bacillus subtilis* genome revealed that, contrary to expectation (the only cases where this had been observed were phages, because they evolve so fast), **at least half of the genes uncovered were totally unknown, whether in structure or in function**

Among reasons for that is our present lack of deep knowledge of metabolism, as well as our lack of knowledge of the way new genes are created, selecting function first, then recruiting a structure that will be improved as it is submitted to natural selection for increased fitness of its host (acquisitive evolution)



The darwinian trio



Variation / Selection / Amplification

Stabilisation

Evolution

↓ *creates*

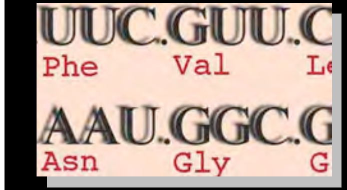
Function

↓ *captures (recruits)*

Structure

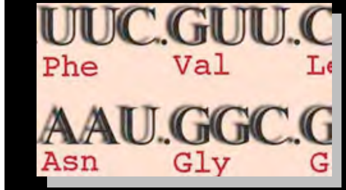
↕ *codes*

Sequence



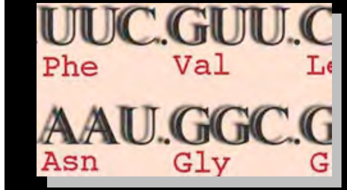
What functions for life?
Extending Cuvier's vision

- We need to separate between **root function** and **helper functions** [the root function of a printer is "printing", "feeding paper", "supplying energy" are helper functions]
- **To be** — to persist in time — can be proposed as the root function of living organisms
  - **Self-consistence** implies correlation of forms
  - **Fighting weathering** implies chemical turnover (metabolism) and protection (compartmentalisation)
  - **Exploration**, associated to sensing and memorizing is the discovery that made life as we know it



What functions for life?

<table>
<tr><td colspan="5">exploration</td></tr>
<tr><td colspan="5">sensing</td></tr>
<tr><td colspan="5">being</td></tr>
<tr><td colspan="3">individualisation</td><td>prevention and correction of erosion</td><td>representation (memory)</td></tr>
<tr><td colspan="3">compartmentalisation</td><td>metabolism</td><td>information transfer</td></tr>
<tr><td colspan="4">energisation</td><td>replication</td></tr>
<tr><td colspan="3">shaping</td><td rowspan="2">construction of biomass precursors</td><td rowspan="2">transcription</td></tr>
<tr><td>making an envelope</td><td>making a skeleton</td><td>making appendages</td></tr>
<tr><td colspan="4">phospholipid and envelope biosynthesis</td><td>translation</td></tr>
<tr><td colspan="3">transport</td><td>degradation</td><td>editing</td></tr>
<tr><td colspan="3">circulation (chanelling)</td><td>salvage</td><td>folding/scaffolding</td></tr>
<tr><td colspan="3">protection</td><td>cleaning</td><td>control</td></tr>
<tr><td colspan="3">partitioning</td><td>labelling</td><td></td></tr>
<tr><td colspan="3">storage</td><td>inactivation</td><td></td></tr>
<tr><td colspan="3"></td><td colspan="2">maintenance (repair, degradation)</td></tr>
<tr><td colspan="3"></td><td colspan="2">modification (labelling, maturation, addressing, stabilisation, protection, control)</td></tr>
</table>



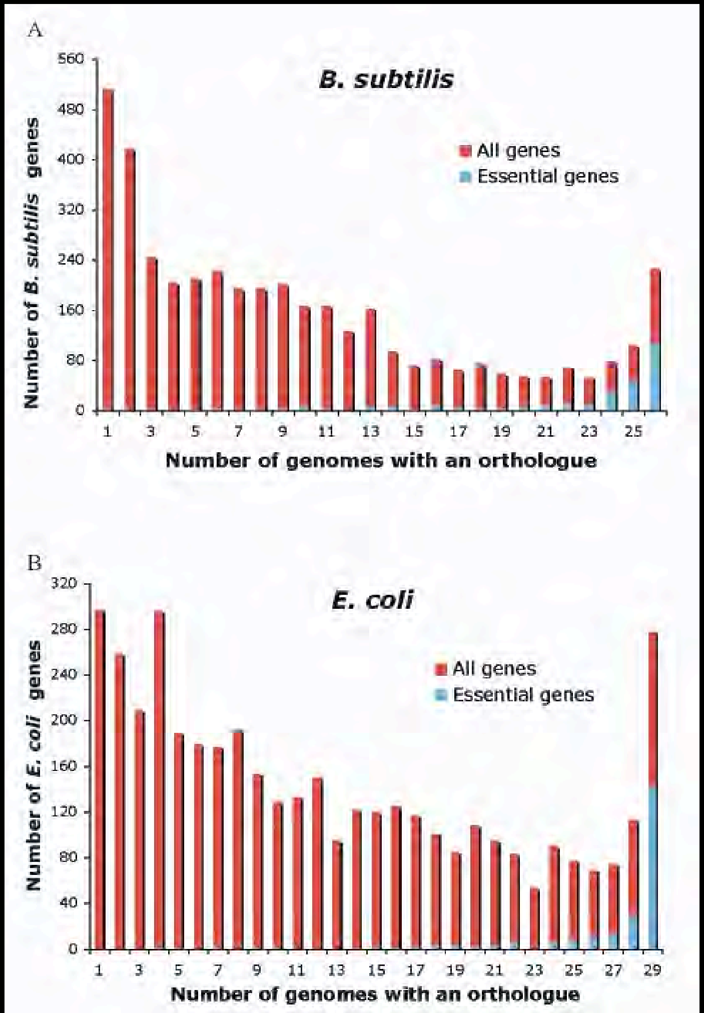
The core genome: looking for persistence

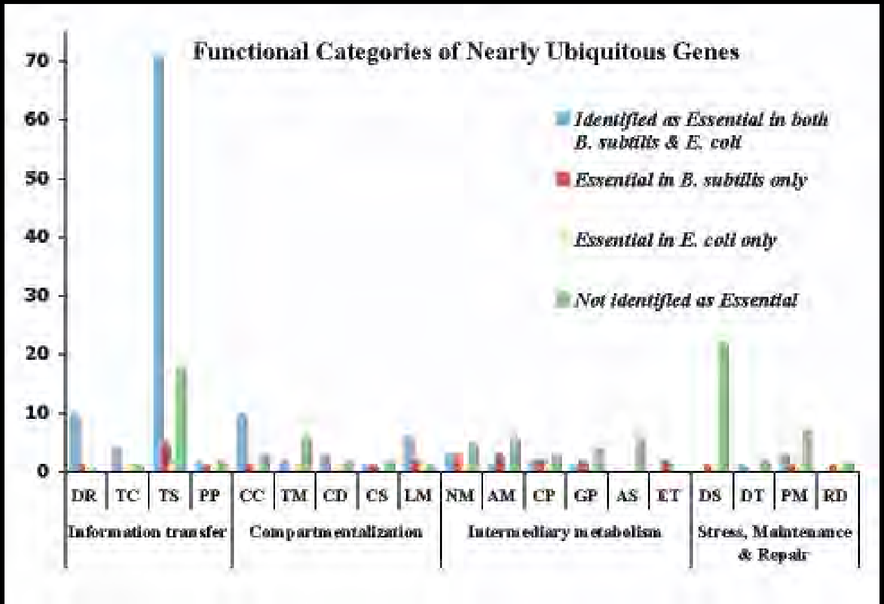
Persistent genes

« Laboratory essential » genes are located in the leading strand, they are also conserved in a majority of genomes. Could we reverse the procedure, and identify genes which are present in a majority of genomes and located in the leading strand?

Microbial genes are of infinite diversity but there exists universals; about 10% of their genes are of persistent and recognized function: they are present in most genomes but approximately half only are essential under laboratory growth conditions

Gene Persistence





- Information transfer
- Compartmentalisation
- Intermediary metabolism
- Stress, Maintenance and Repair

Phylogeny of persistence

Some of the essential genes missing from the list of persistent genes have diverged considerably

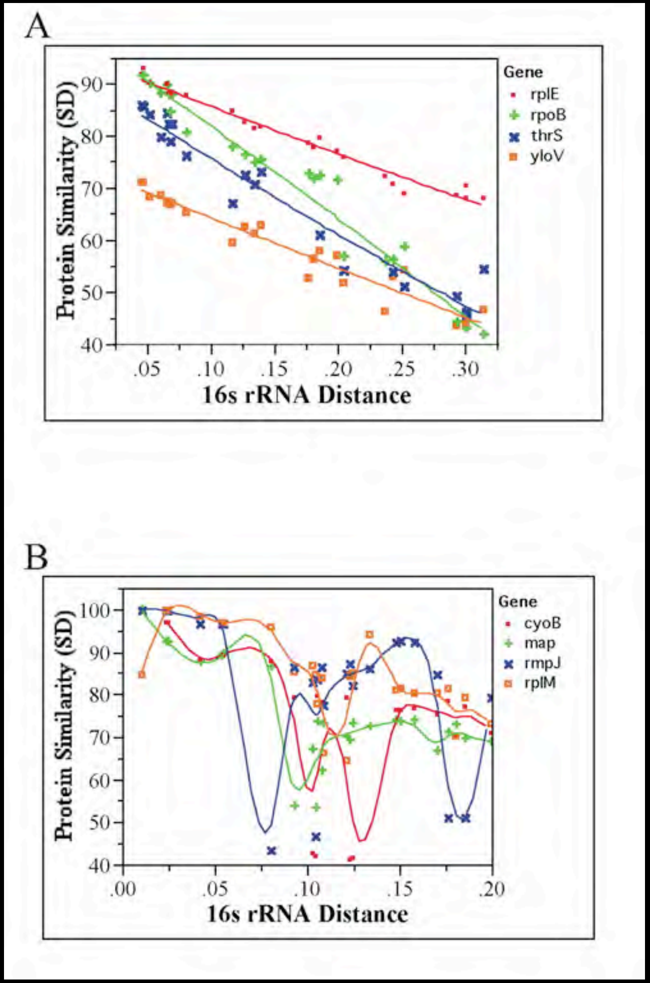
To assess the contribution of this effect we measured for each pair of genomes the correlation between the similarity of orthologous pairs and that of the 16S rRNA

Two scenarios are observed, either a linear correlation with rRNA evolution (85%), or erratic evolution, implying horizontal gene transfer (15%)

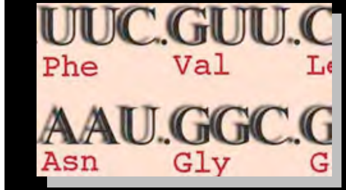
Gene persistence

For example (A), 38% (resp. 48%) of *B. subtilis* (resp. *E. coli*) persistent genes show a correlation coefficient >0.9 between the sequence similarity of the pair of orthologs and the 16S RNA.

In contrast, some genes (B) evolve in an erratic way. This may be due to horizontal gene transfer, local adaptations leading to faster or slower evolutionary pace, or simply wrong assignments of orthology. The latter can be a significant problem, especially in large protein families. The genes presenting such an erratic pattern are seldom found in the persistent set.



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

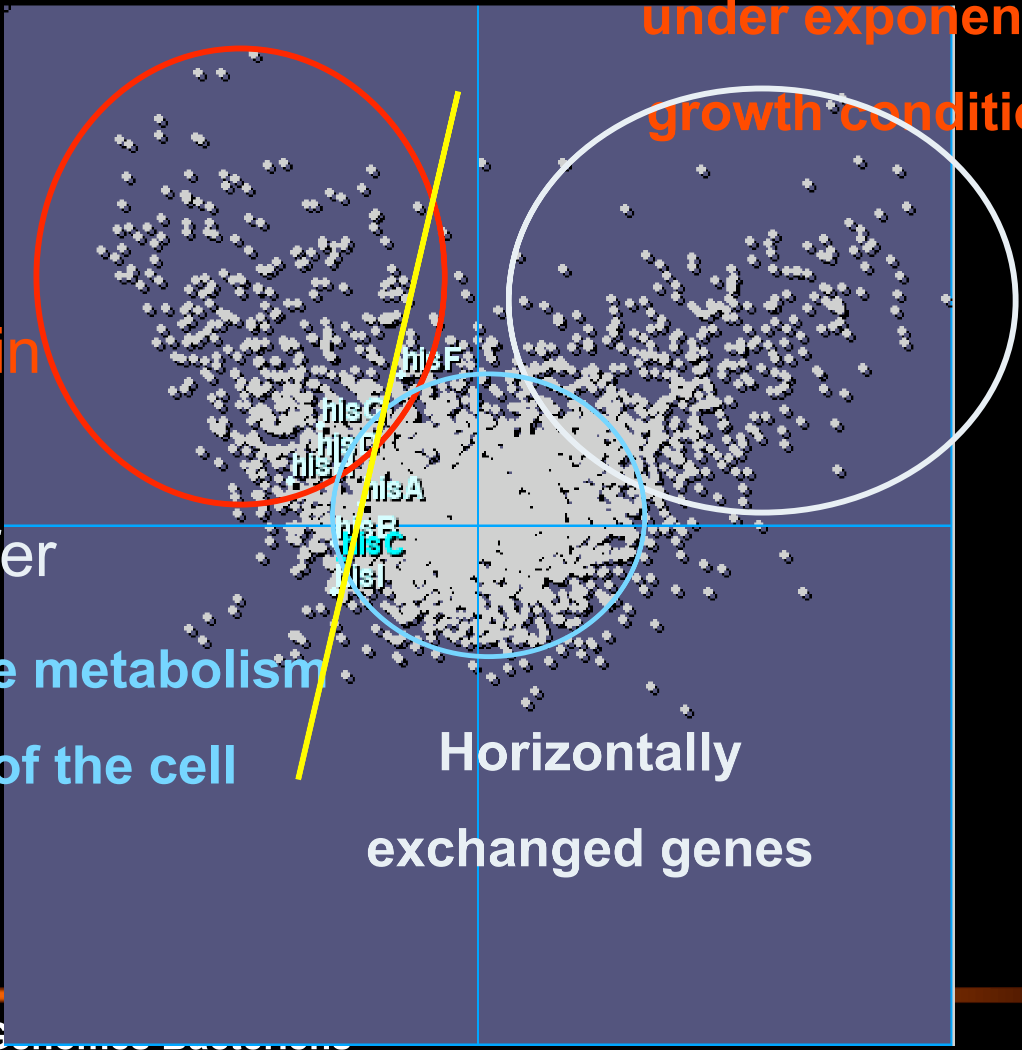


Biases in the codon usage

Class I: core metabolism

Class II: high expression in exponential growth

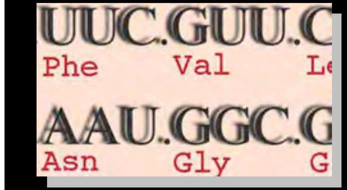
Class III: horizontal transfer



Genes expressed at a high level under exponential growth conditions

Core metabolism of the cell

Horizontally exchanged genes



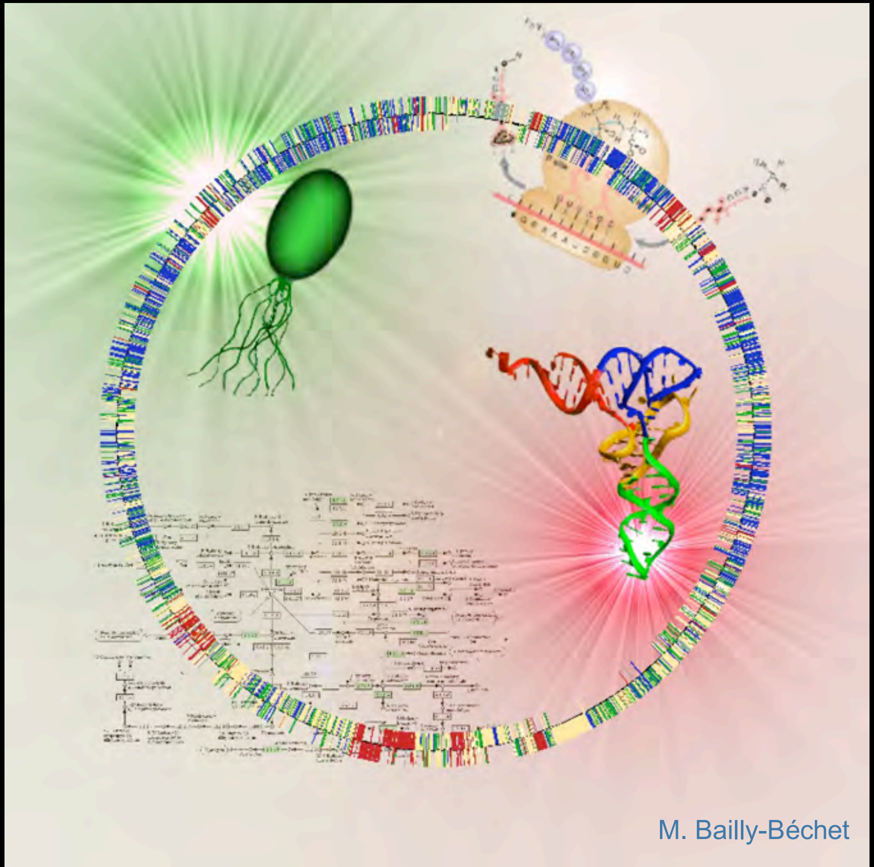
Local codon usage biases

Correspondence Analysis shows that genes with neighbouring codon usage biases are functionally related. How does this extrapolate in the distribution of genes in the chromosome?

A clustering method based on the analysis of codon usage biases using an information theory groups the genes into homogeneous clusters, which are not distributed randomly in the chromosome. The method allows finding both the specific codon usage bias in a class and the most relevant number of classes (4 for *E. coli* and 5 for *B. subtilis*).

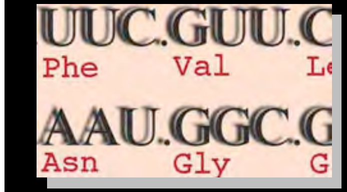
Genomic islands

One cluster is related to gene expression (blue). Other groups feature an over-representation of genes belonging to different functional groups: horizontally transferred genes (red), motility (yellow) and intermediary metabolism (green).



M. Bailly-Béchet

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

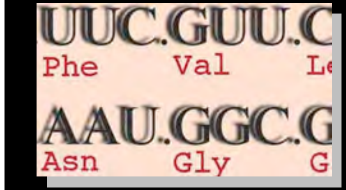


What functions for life?
Scenario for the origin of life

To be — to persist in time — can be proposed as the root function of living organisms

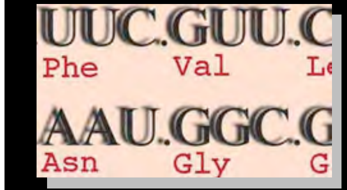
- Fighting weathering implies chemical turnover (metabolism) on solid surfaces and immobility requires protection (compartmentalisation)
- Compartementalised metabolism creates surface substitutes (RNA)
- Exploration, associated to sensing and memorizing (information transfer) is the discovery that made life as we know it

A Danchin_Homeotopic transformation and the origin of translation *Progress in Biophysics and Molecular Biology* (1989) **54:** 81-86



Persistent genes connectivity

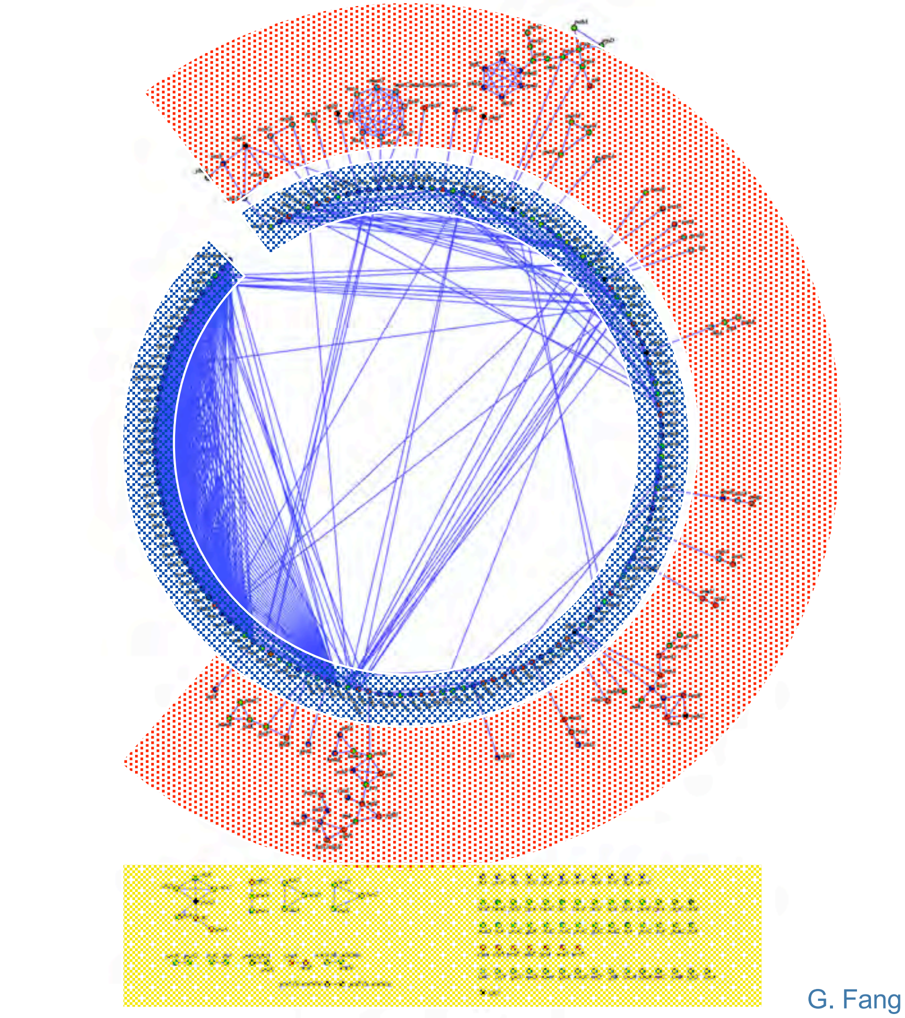
Using 228 genomes with more than 1500 genes and « correct » annotations, we have identified genes that tend to remain close to one another; this « mutual attraction » constructs a remarkable network made of three layers



Persistent genes organisation recapitulates the origin of life

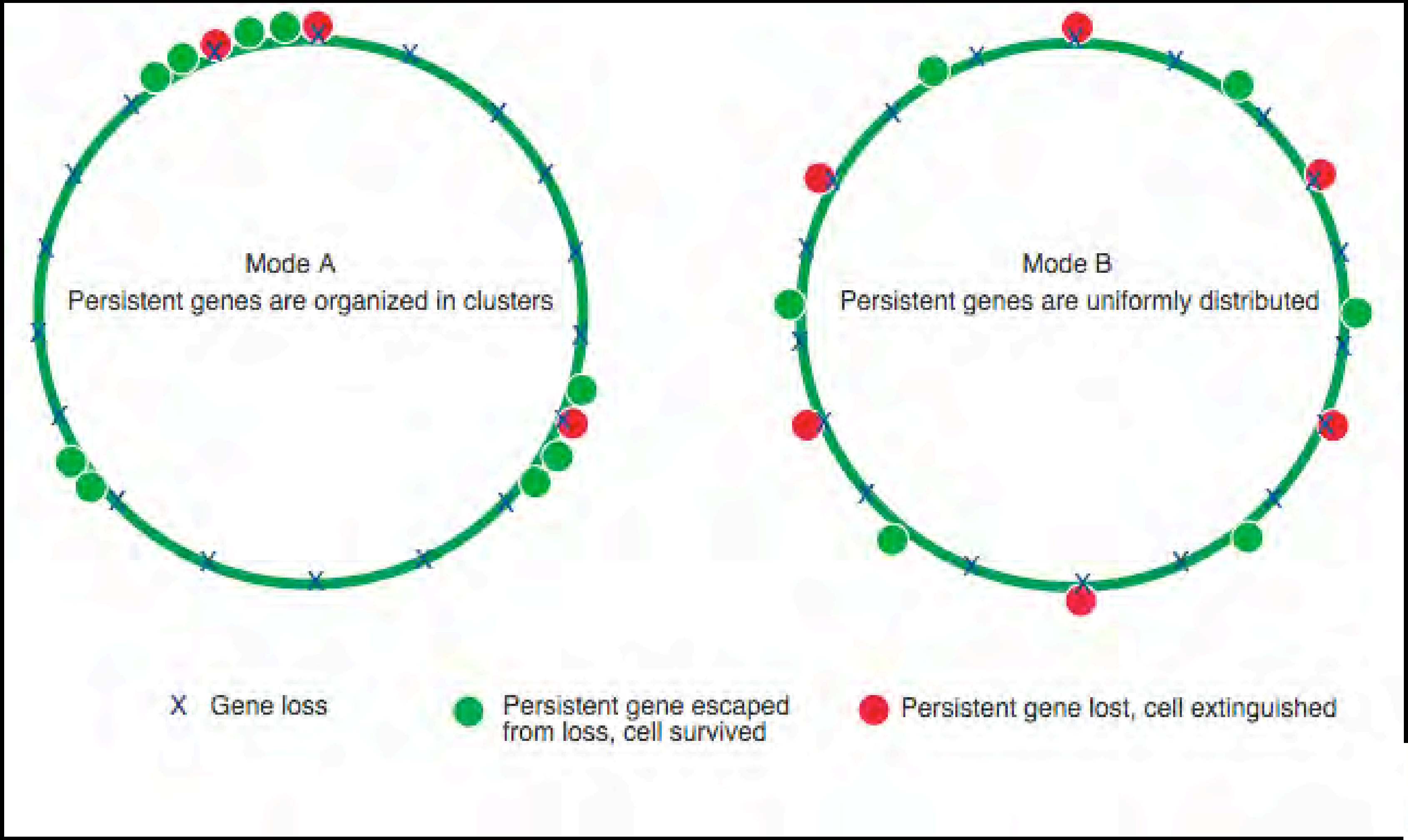
The external network, made from genes of intermediary metabolism (nucleotides and coenzymes, lipids), is highly fragmented; the middle network has class I tARN synthetases at its core, and the internal network, almost continuous, makes the core of information transfer around the ribosome, transcription and replication

This is consistent with a scenario where coenzymes and basic blocks led to a tRNA world organising metabolism, followed by a template-driven RNA world



G. Fang

Existence implies clustering



Mode A
Persistent genes are organized in clusters

Mode B
Persistent genes are uniformly distributed

X Gene loss

🟢 Persistent gene escaped from loss, cell survived

🔴 Persistent gene lost, cell extinguished

Existence implies clustered persistence

Why are persistent genes clustered? A simple model shows that if, in addition to horizontal gene transfer, there is a process deleting genes in bundles in genomes, then any gene contributing to fitness frequently enough over generations will tend to cluster with other genes with similar properties. This accounts for clustering of essential genes, but most probably also for clustering of antibiotic resistance genes in bacteria found in hospitals....

As a consequence gene clustering will precede not derive from co-transcription or protein-protein interaction!

Authors

Génétique des Génomes Bactériens

- Gang Fang
- Etienne Larsabal
- Géraldine Pascal
- Eduardo Rocha

Genoscope

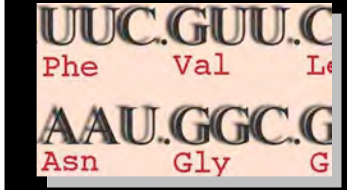
- Géraldine Pascal
- Claudine Médigue

Génétique in silico

- Marc Bailly-Béchet
- Massimo Vergassola

Abdus Salam International Center in Theoretical Physics

- Mudassar Iqbal
- Matteo Marsili



Σας ευχαριστω

Thank you