INSTITUT PASTEUR

Archives or Palimpsests?
Bacterial Genomes
Unveil a Scenario for the Origin of Life

Institute for Systems Biology, Seattle

April 23, 2007

adanchin@pasteur.fr

AUTHORS

Génétique des Génomes Bactériens (in silico)

- Gang Fang
- Etienne Larsabal
- Yuko Makita
- Géraldine Pascal
- Eduardo Rocha
- Marie Touchon
- Tingzhang Wang

Génétique in silico

- Marc Bailly-Béchet
- Massimo Vergassola

Abdus Salam International Center in Theoretical Physics

- Mudassar Iqbal
- Matteo Marsili

INSTITUT PASTEUR

TAKE HOME MESSAGE

- Genomes are made of two classes of genes:
  - The first class constitutes the paleome, which makes a replicator and a constructor, as in a living Turing machine
  - The second class contitutes the cenome, which allows cells to live in and explore a particular niche
- Genes making the paleome constitute an archive of ubiquitous functions encoded by persistent genes
- Persistent genes group into a three layers network:
  - the first layer codes for functions creating basic building blocks
  - the second layer is organized around tRNA synthetases
  - the third layer is organized around information transfer

This organisation makes a scenario for the origin of life

adanchin@pasteur.fr

INSTITUT PASTEUR

- A computational view of what life is
- A map of the cell in the chromosome?
- Translation organises the bacterial genome
- The core genome: persistent genes
- The core genome is an archive of the origin of life

adanchin@pasteur.fr

WHAT LIFE IS … WITH A TWIST

Three processes are required to constitute life:

- **Information transfer => genomics unveils the organization of the program associated to the cell**

Forces coupling the genome structure to the structure of the cell (the cellular machine, not discussed today):

- **Metabolism**
- **Compartmentalization**

The cell is the atom of life, with two compartmentalization strategies: a simple envelope (prokaryotes), or the multiplication of membranes and skins (eukaryotes)

A first hint of a **link between genome organization and the architecture of the organism** is visible here: the general composition of genomes follow this distinction: prokaryotic genomes look random, eukaryotic genomes look repeated

INSTITUT PASTEUR

AN ALGORITHMIC VIEW OF BIOLOGICAL ACTIONS

Replication, transcription, translation: high parallelism

"Begin, Check Control Points, Repeat, End"

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

The processes of time control (check points) are rarely taken into account (except for the replication/division processes), but their role is essential to allow coordination of multiple actions in parallel

adanchin@pasteur.fr

INSTITUT PASTEUR

WHAT IS COMPUTING?

Two processes are needed for computing:

- A machine able to read and write
- A program on a physical support (typically, a punched or magnetic tape illustrates the sequential order of the symbols that make the program), split (in pratice, but not conceptually) into two entities :
  - Program (providing the goal)
  - Data (providing the context)

The machine is distinct from the program

adanchin@pasteur.fr

INSTITUT PASTEUR

CONTEXT: THE "GENETIC PROGRAM"

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

The « genetic program » metaphor has practical consequences: we know how to manipulate genes and gene products, can we push the metaphor to its ultimate consequences?

adanchin@pasteur.fr

INSTITUT PASTEUR

CELLS AND COMPUTERS

Genetics rests on the description of genomes as **texts written with a four letter alphabet**: but, **do cells behave as computers?**

Horizontal Gene Transfer
Virus
Genetic engineering => reconstruction of the hepatitis C virus
Animal cloning

all points to separation between

« Machine » (the cell factory)
and
Data + program

adanchin@pasteur.fr

INSTITUT PASTEUR

- **A computational view of what life is**
- **A map of the cell in the chromosome?**
- **Translation organises the bacterial genome**
- **The core genome: persistent genes**
- **The core genome is an archive of the origin of life**

INSTITUT PASTEUR

IS THERE A MAP OF THE CELL IN THE CHROMOSOME?

John von Neumann, trying to understand brain, suggested that if the computer was both to behave as a computer and to construct the machine itself, one should find an image of the machine somewhere in the machine.

The computer has to be split into a replicator and a constructor, which expresses the program for construction of both the replicator and the constructor.

The metaphor does not appear to apply to the brain, does it apply to the cell?

adanchin@pasteur.fr

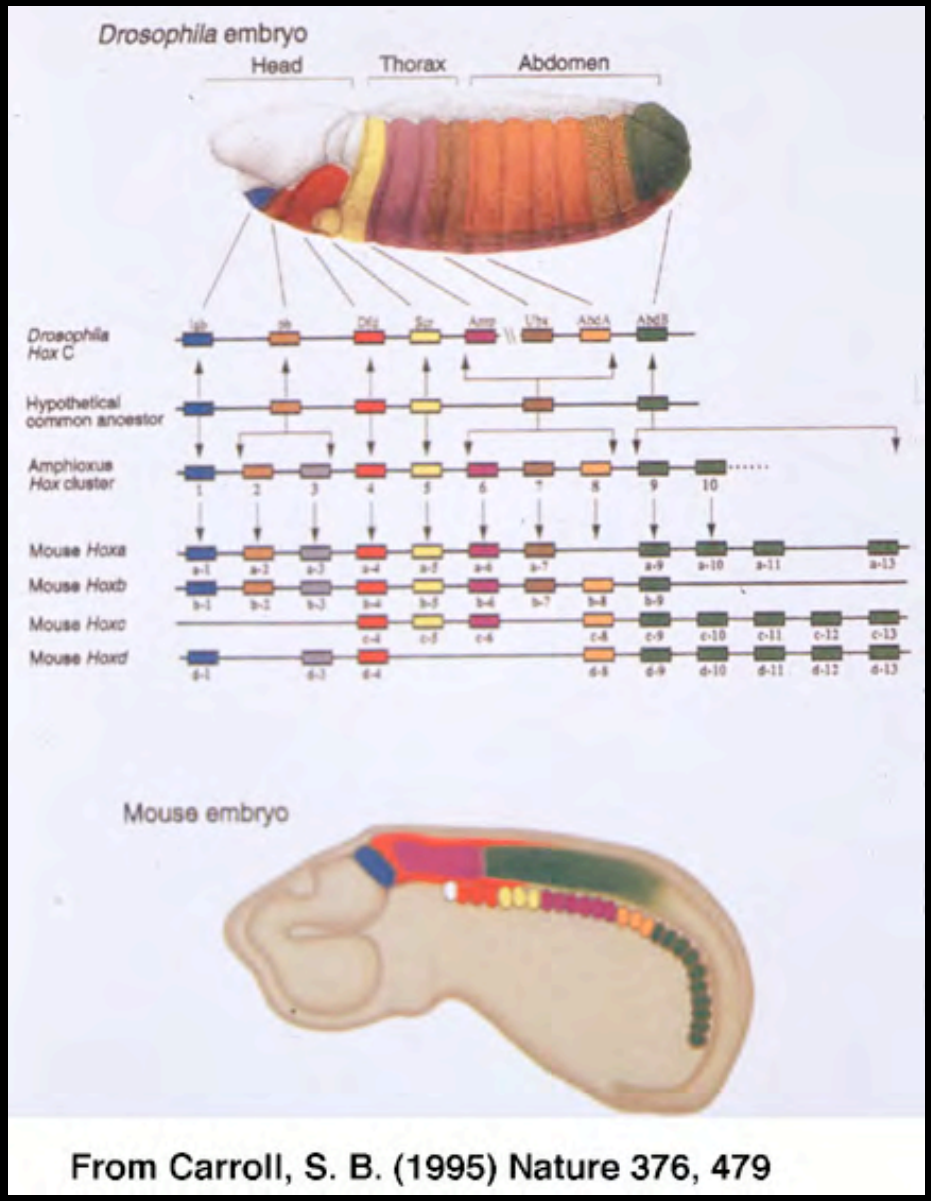
INSTITUT PASTEUR

The mystery of homeogenes' origin

Drosophiloculus,

Homunculus?

Celluloculus?



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

adanchin@pasteur.fr

INSTITUT PASTEUR

LOOKING FOR THE REPLICATOR AND THE CONSTRUCTOR

Are genes grouped randomly in the chromosomes?

Do we find different gene categories, in terms of the way they are organized?

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

A. Danchin
The Delphic Boat; What genomes tell us. (2003) Harvard University Press

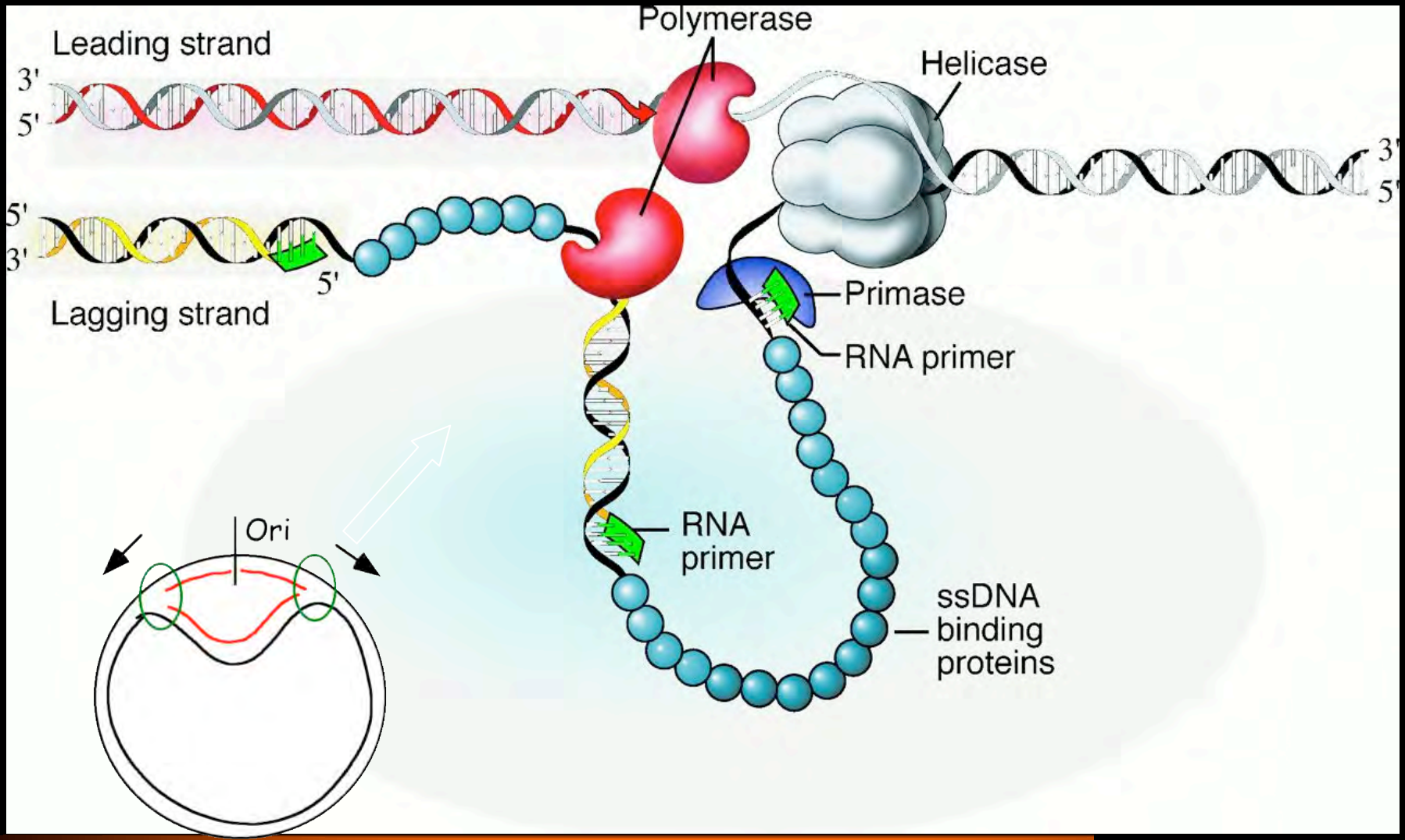
adanchin@pasteur.fr

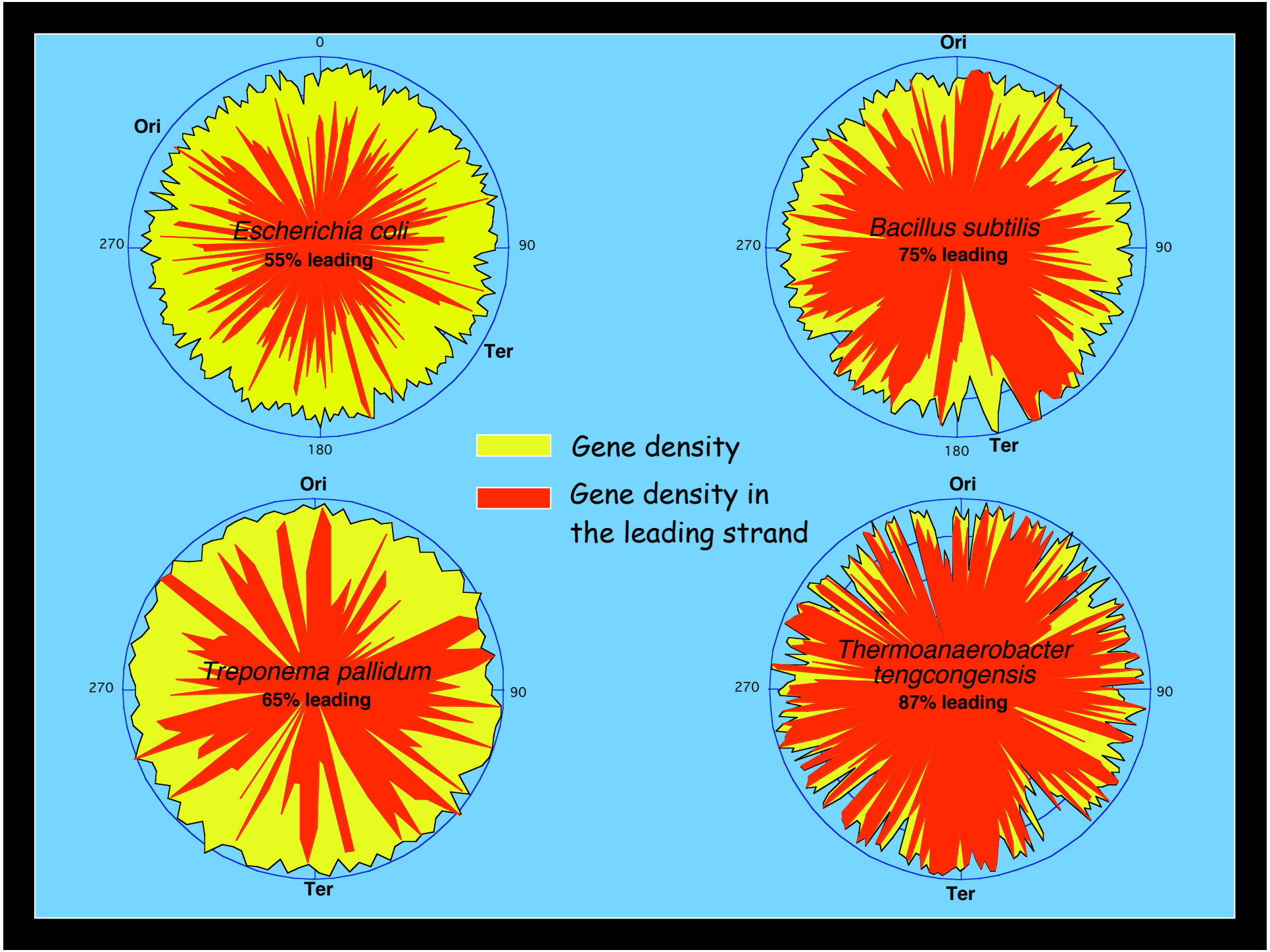
INSTITUT PASTEUR

A FIRST CONSTRAINT: TO LEAD OR TO LAG

In Bacteria, genes tend to prefer the replication leading strand. One observes however a very large variation, depending on the organism: Gram positive A+T-rich bacteria have a particularly large bias, parallel with the existence of two DNA polymerases III (DnaE and PolC)

adanchin@pasteur.fr

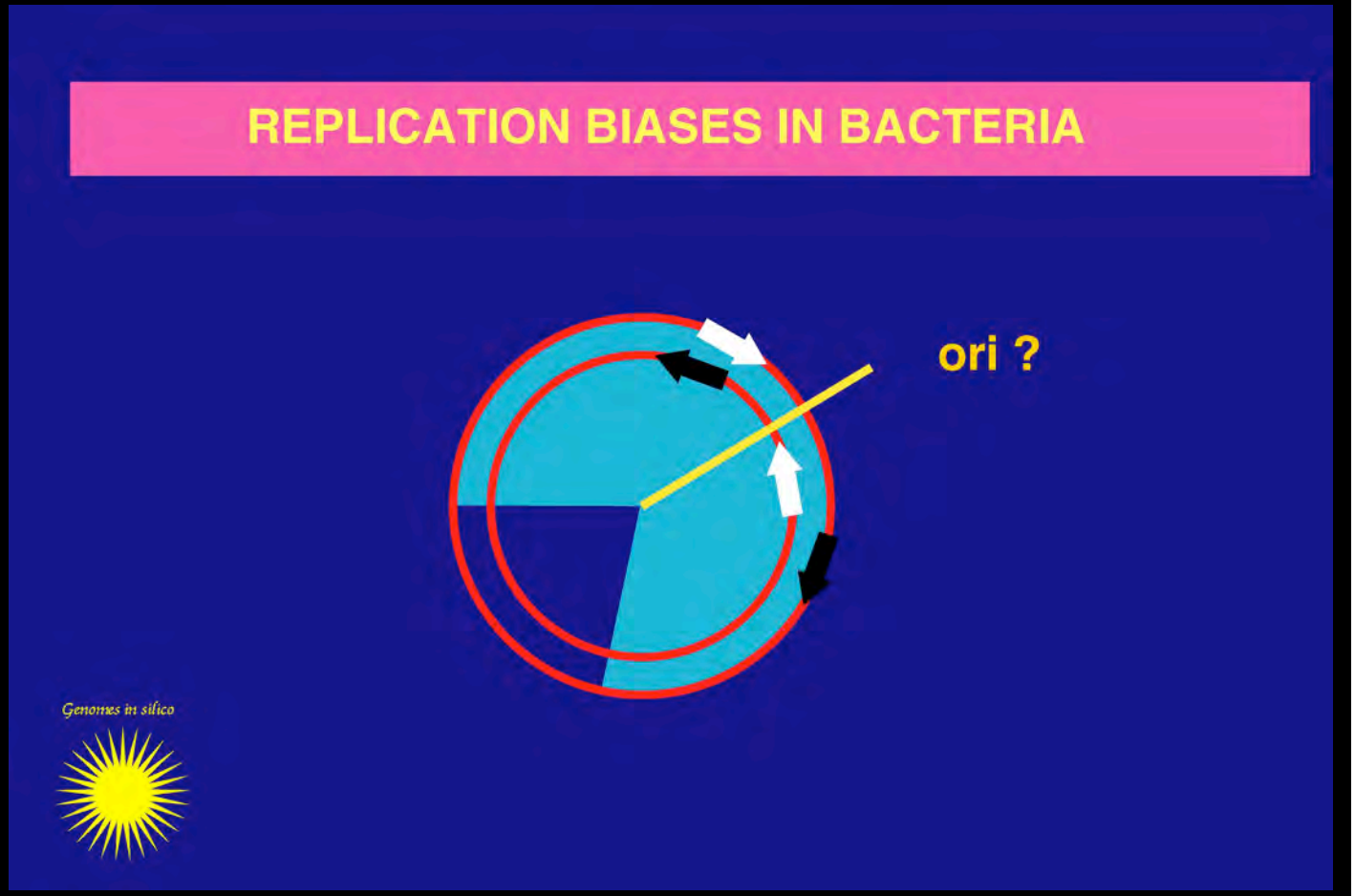

Polymerase
Leading strand
3'
5'
Helicase
3'
5'
5'
3'
5'
Lagging strand
Primase
RNA primer
RNA
primer
ssDNA
binding
proteins
Ori
Ter


0
Ori
270
90
Ter
180
Escherichia coli
55% leading
Ori
270
90
Ter
180
Bacillus subtilis
75% leading
Gene density
Gene density in
the leading strand
Ori
270
90
Ter
Treponema pallidum
65% leading
Ori
270
90
Ter
Thermoanaerobacter
tengcongensis
87% leading

INSTITUT PASTEUR

TO LEAD OR TO LAG...

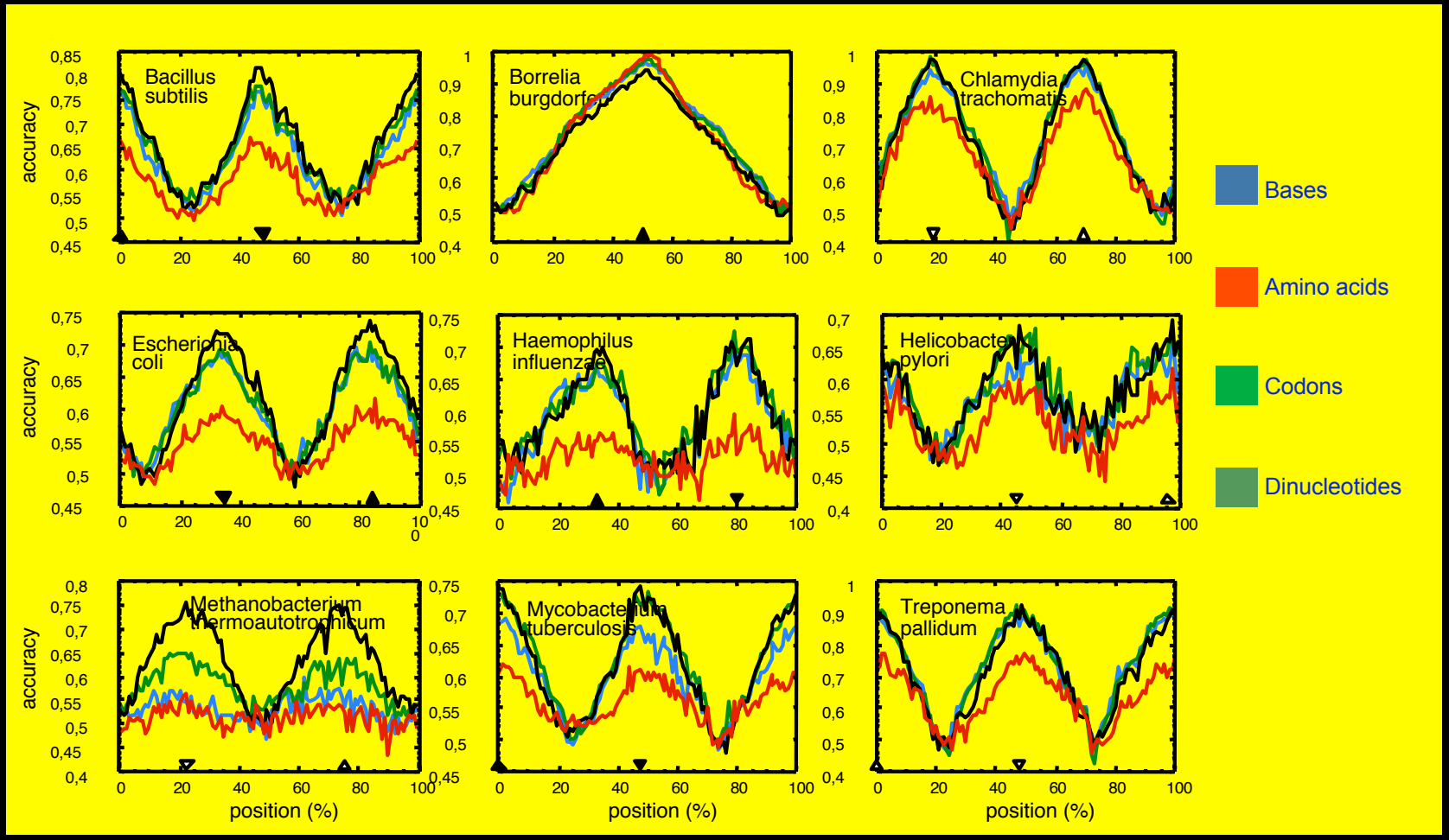
Chosing arbitrarily an origin of replication and a property of the strand (base composition, codon composition, codon usage, amino acid composition of the coded protein…) one can use discriminant analysis to see whether the hypothesis holds.



adanchin@pasteur.fr

INSTITUT PASTEUR

THAT IS THE QUESTION...

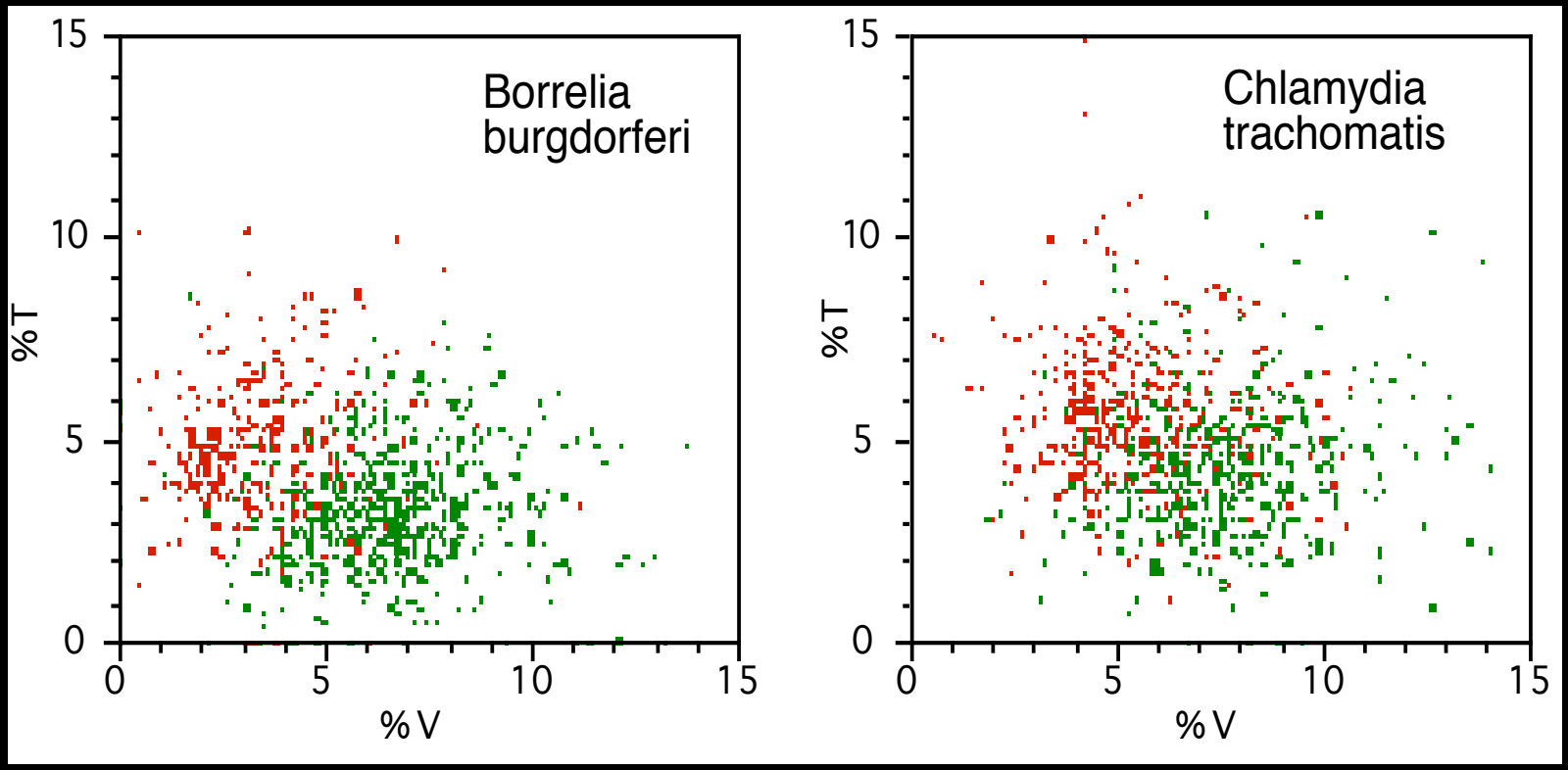


adanchin@pasteur.fr

INSTITUT PASTEUR

TO LEAD HAS A COST: BIAS VISIBLE IN PROTEINS…

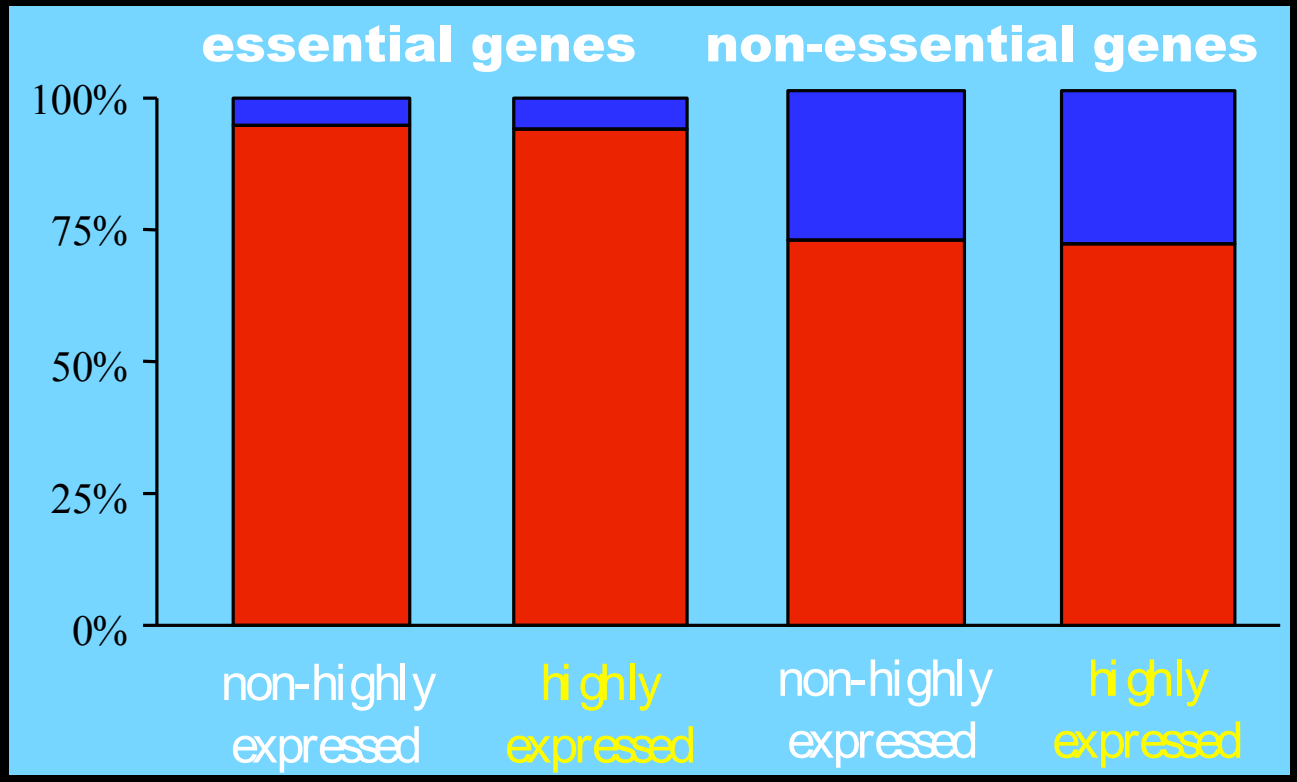
GT in the leading strand, CA in the lagging strand...



adanchin@pasteur.fr

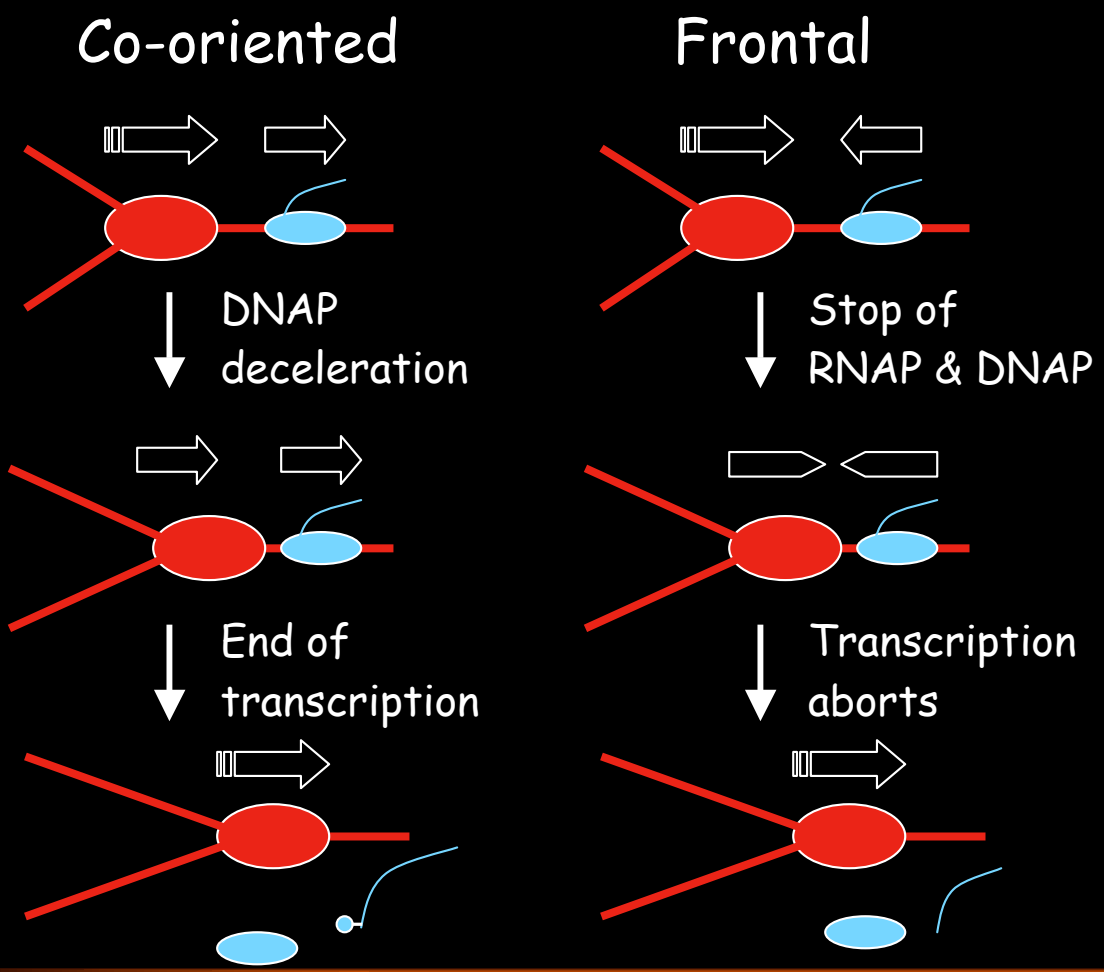
INSTITUT PASTEUR

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.

Causality: avoiding collisions between RNA and DNA polymerases



~~Consequences:~~

~~1. Slowing down of replication~~

~~2. Loss of transcripts~~

Consequences:

1. Truncated transcripts
2. Truncated essential proteins

adanchin@pasteur.fr

INSTITUT PASTEUR

- A computational view of what life is
- A map of the cell in the chromosome?
- Translation organises the bacterial genome
- The paleome: persistent genes
- The paleome is an archive of the origin of life

adanchin@pasteur.fr

INSTITUT PASTEUR

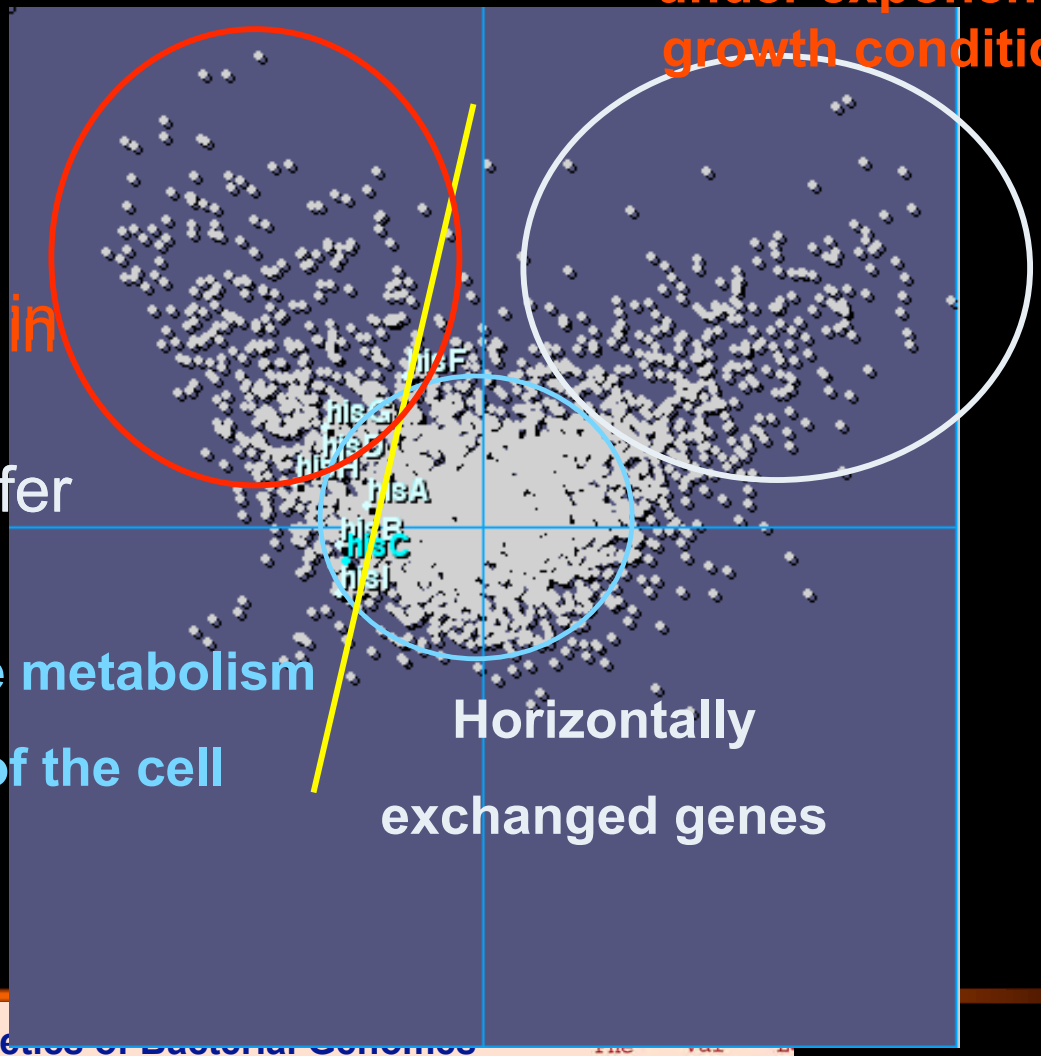
Biases in codon us
are functionally orde

Genes expressed at a high level under exponential growth conditions

Class I: core metabolism

Class II: high expression in exponential growth

Class III: horizontal transfer



adanchin@pasteur.fr

INSTITUT PASTEUR

LOCAL CODON USAGE BIASES

In Correspondence Analysis, genes with similar 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.

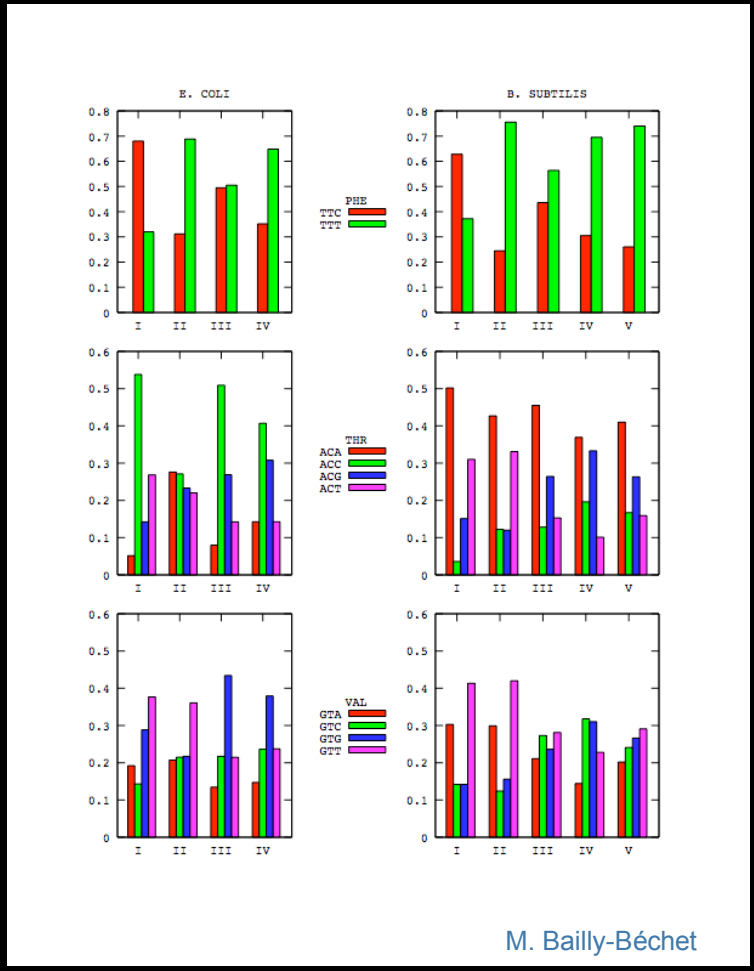
The method allows finding both the specific codon usage bias in a class and the most relevant number of classes (4 for *E. coli* and 5 for *B. subtilis*).

adanchin@pasteur.fr

GENOMIC ISLANDS

Genes with a similar bias are close together in the chromosome (groups of 30 genes), demonstrating a role of translation in structuring bacterial chromosomes.

This effect comes from the dynamic compartmentalization induced by the recycling of rare tRNAs, leading to gene expression rates dependent on their genomic and expression context.



M. Bailly-Béchet

INSTITUT PASTEUR

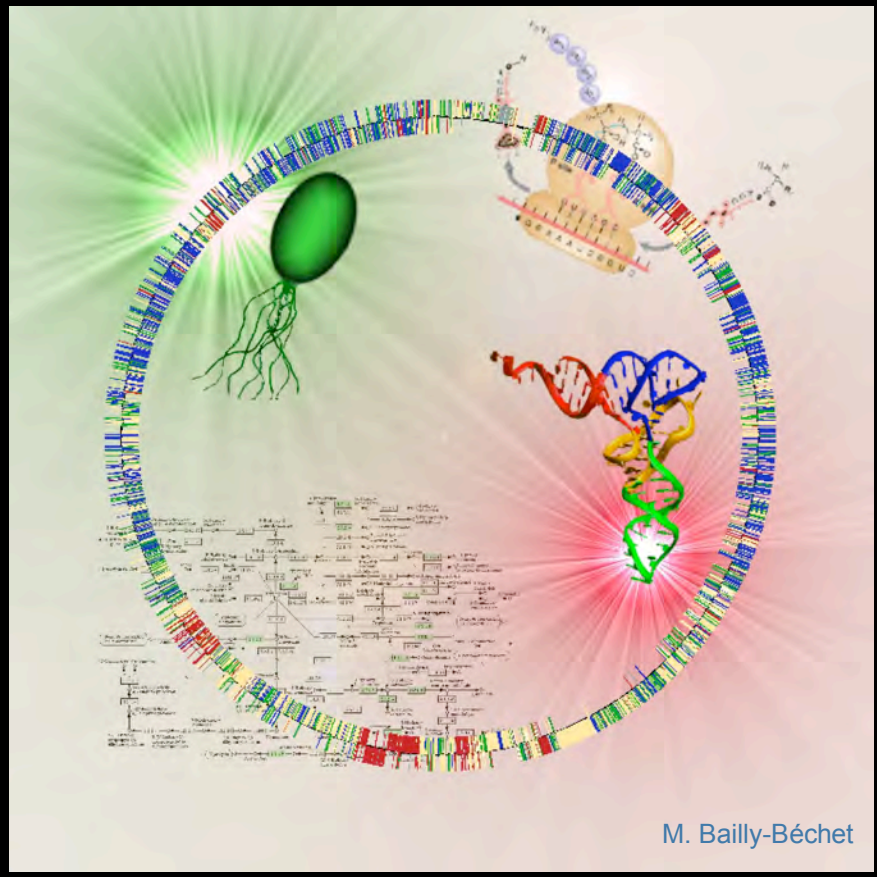
GENOMIC ISLANDS

High gene expression (blue)

Horizontally transferred genes (red)

Motility (yellow)

Intermediary metabolism (green)



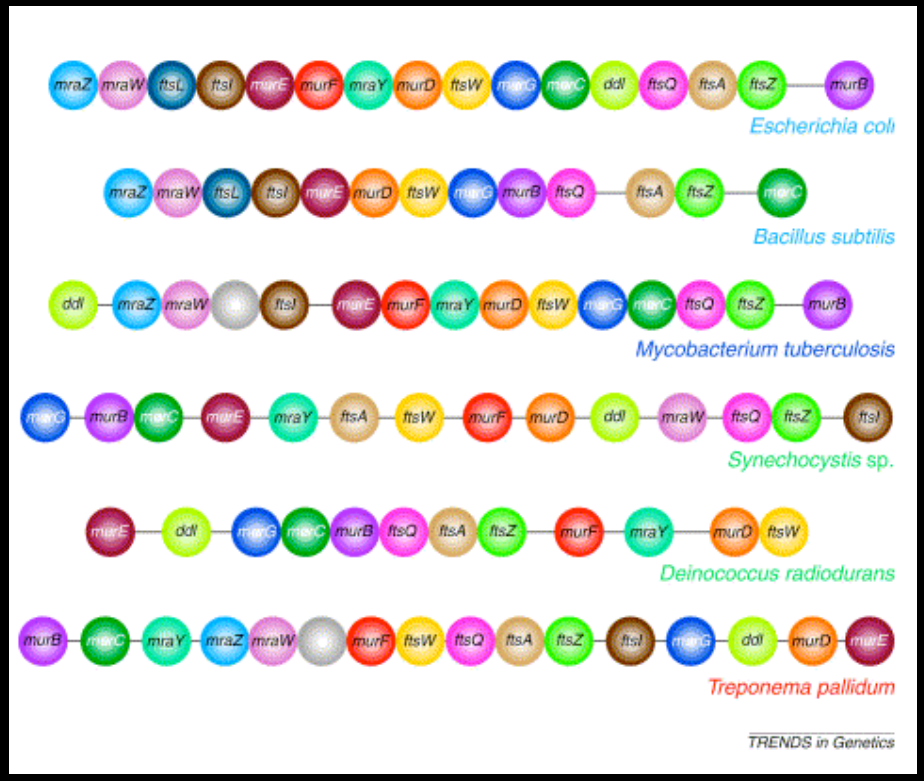
M Bailly-Bechet, A Danchin, M Iqbal, M Marsili, M Vergassola
Codon usage domains over bacterial chromosomes
PLoS Computational Biology (2006) **2**: e37

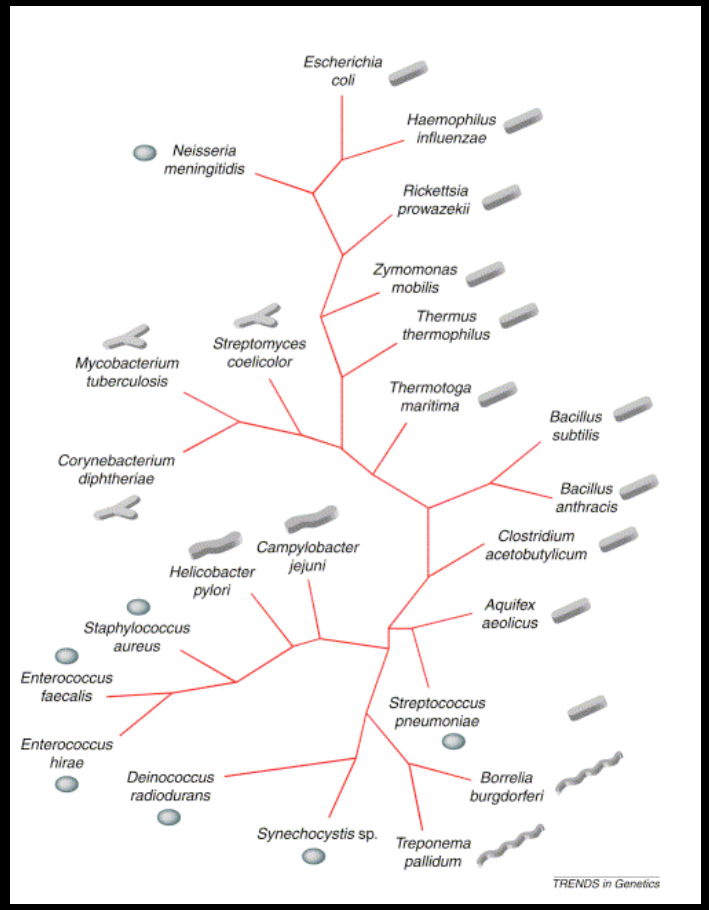
adanchin@pasteur.fr

INSTITUT PASTEUR

GENE ORDER AND CELL SHAPE

The mur-fts cluster





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

adanchin@pasteur.fr

INSTITUT PASTEUR

- A computational view of what life is
- A map of the cell in the chromosome?
- Translation organises the bacterial genome
- The paleome: persistent genes
- The paleome is an archive of the origin of life

adanchin@pasteur.fr

INSTITUT PASTEUR

THE FIRST DISCOVERY OF GENOMICS: GENES OF UNKNOWN FUNCTION

In 1991, at the first EU meeting on genome projects, the sequence of yeast's chromosome III and the first 100 kb of the *Bacillus subtilis* genome revealed that, contrary to expectation, **at least half of the genes newly uncovered were totally unknown, whether in the function or the structure of their product**

Several reasons account for this fact: on the one hand our knowledge of metabolism is extremely scant, on the other hand we do not know how new genes are created; finally evolution proceeds through selection of functions, by recruting pre-existing structures whose fitness will increase in parallel with fitness of the species in a given environment (acquisitive evolution)

If there are that many genes for a given function, is there nevertheless some common ones?

adanchin@pasteur.fr

ACQUISITIVE EVOLUTION MASKS FUNCTIONAL PERSISTENCE

Variation / Selection / Amplification

Stabilisation

Evolution

creates

Function

captures (recruits)

Structure

codes

Sequence

INSTITUT PASTEUR

PERSISTENT GENES

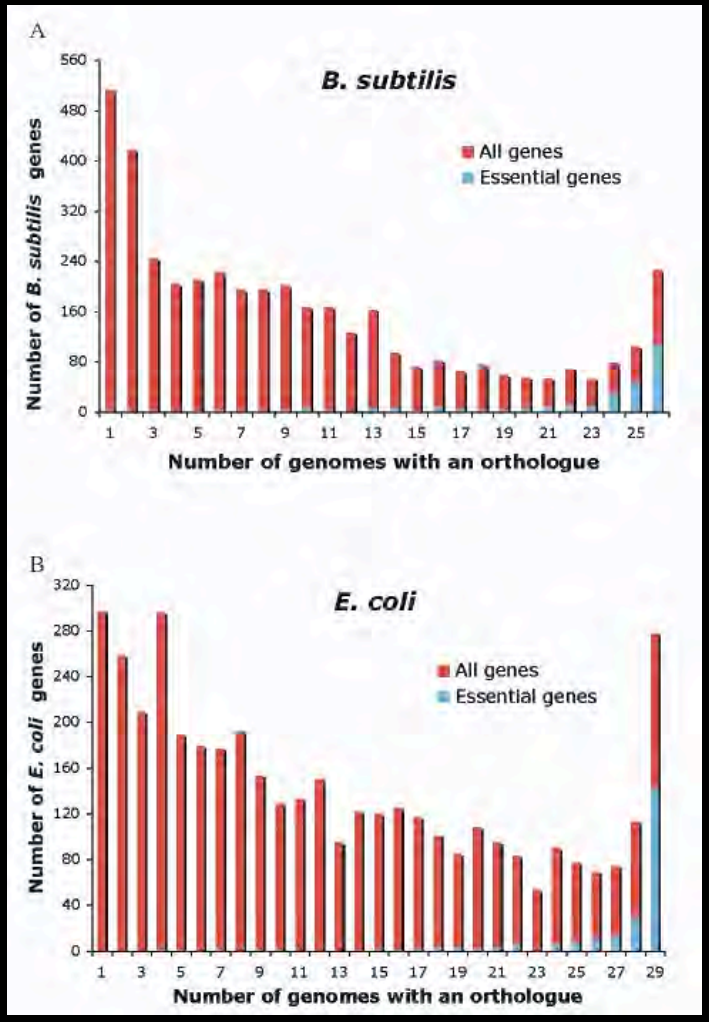
Laboratory essential genes are located in the leading strand. They are conserved in a majority of genomes. By contrast the genes that are conserved and located in the leading strand make a particular category, which doubles the number of « essential » genes.

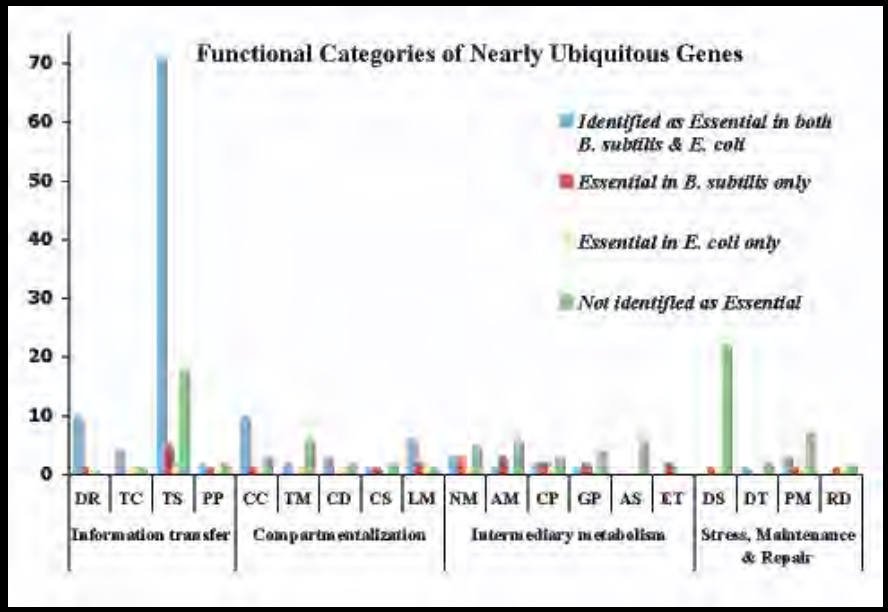
These genes make a universal category; 400-500 genes persist in a majority of bacterial genomes; they are not only involved in the three processes needed for life, but in maintenance and in adaptation to transient phenomena; a fraction manages the evolution of the organism.

adanchin@pasteur.fr

INSTITUT PASTEUR

AN EXTENSION OF ESSENTIALITY: GENE PERSISTENCE





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

adanchin@pasteur.fr

INSTITUT PASTEUR

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

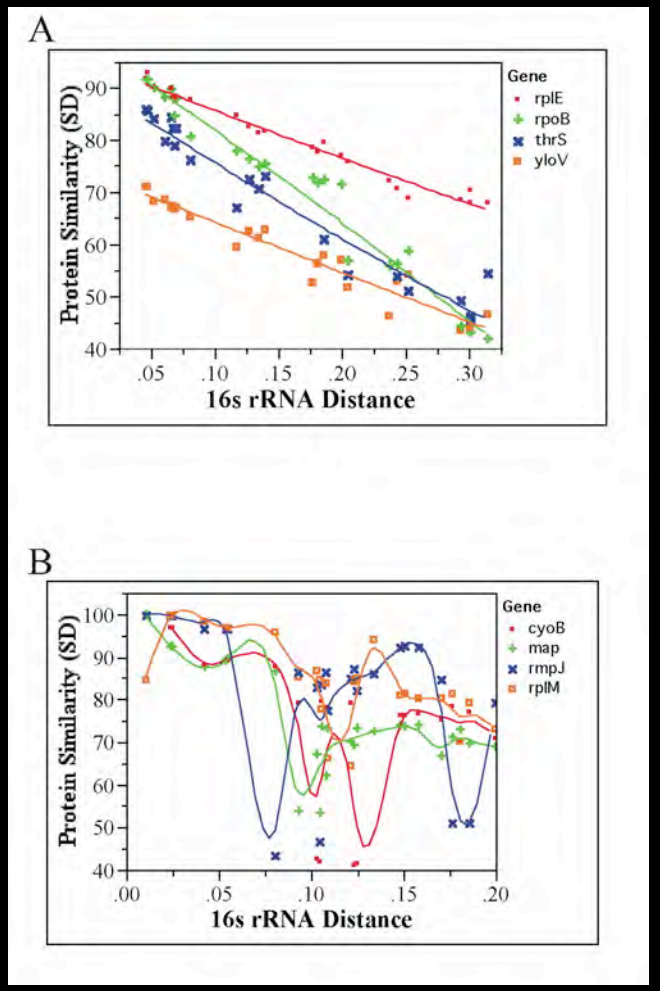
Two scenarios are observed, either a linear correlation with rRNA evolution (85%), or erratic evolution, implying horizontal gene transfer (15%)

adanchin@pasteur.fr

GENE PERSISTENCE

(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 sequence evolution.

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.

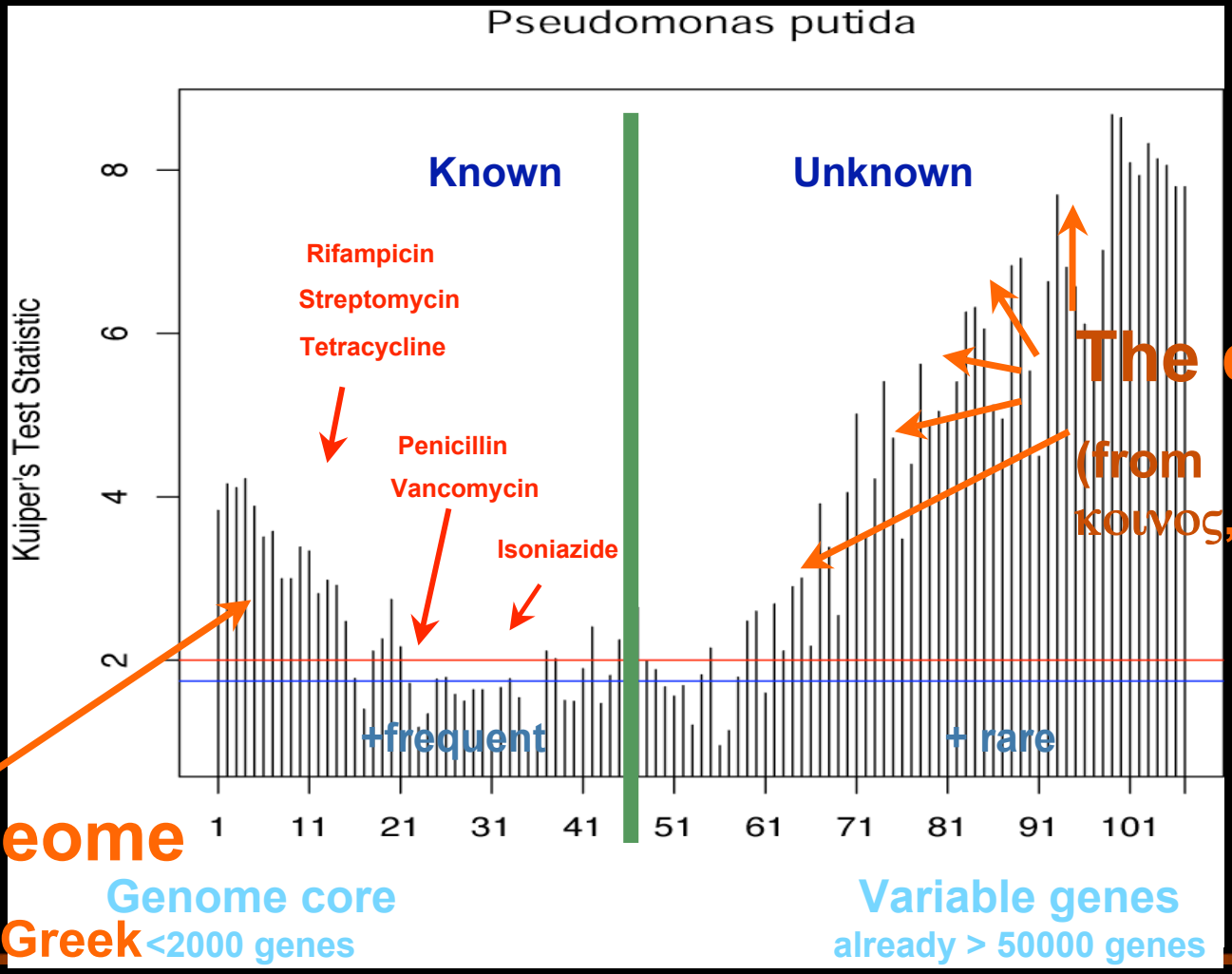


G Fang, EPC Rocha, A Danchin
How essential are non-essential genes?
Mol Biol Evol (2005) **22**: 2147-2156

CH You, HY Lu, A Sekowska, G Fang, YP Wang, AM Gilles, A Danchin
The two authentic methionine aminopeptidase genes are differentially expressed in *Bacillus subtilis*BMC *Microbiology* (2005) **5**: 57

adanchin@pasteur.fr

CONSERVATION OF GENE CLUSTERING



The paleome

(from the Greek παλαιος, ancient)

The cenome

(from the Greek κουνος, common)

Frequency in genomes

INSTITUT PASTEUR

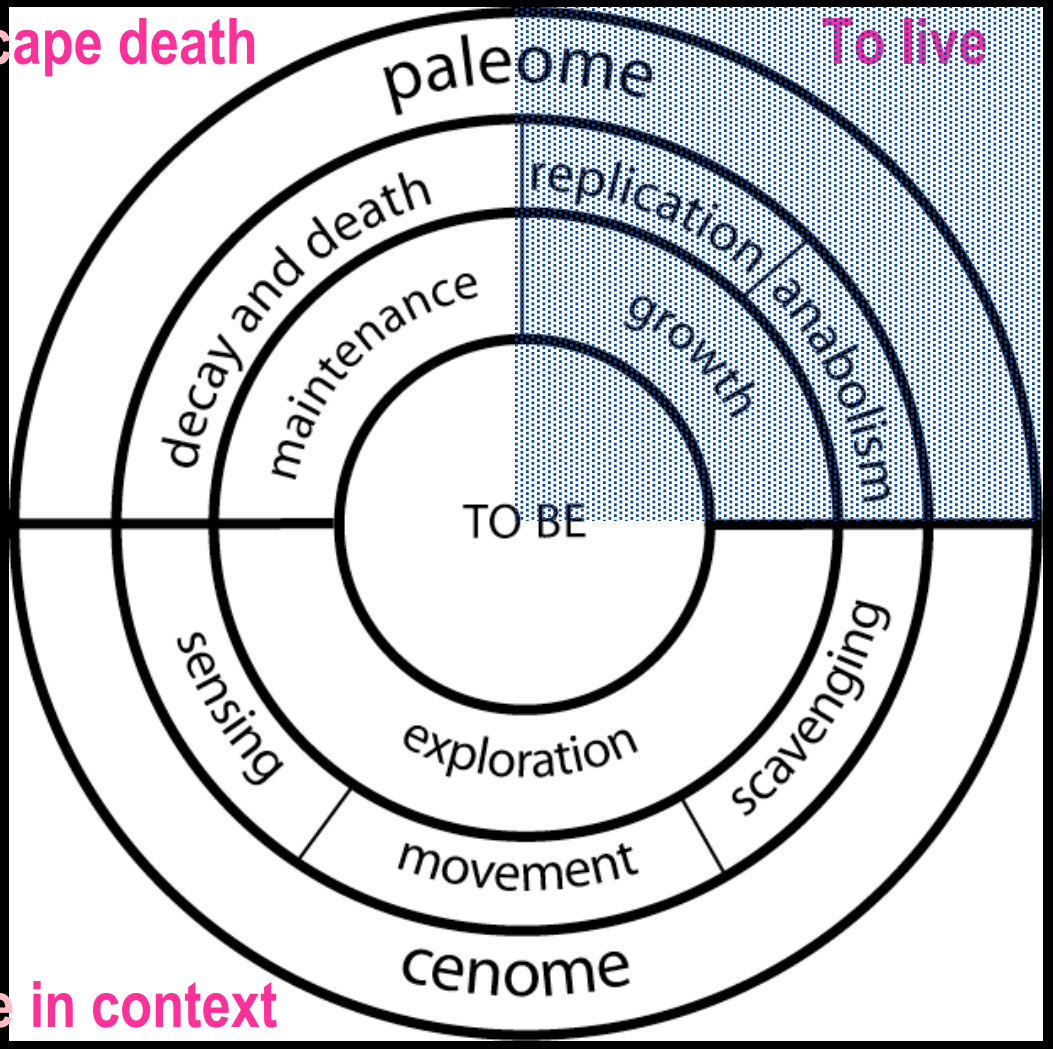
THE COMPOSITE GENOME

- Expecting **two genome components**, we need to separate between between the **replicator/constructor** and ancillary functions
- Extant genomes should comprize ubiquitous functions (not genes!) which would correspond to the former (here named the **paleome** for reasons which will become apparent soon) and functions specific to the environment of the organism (named the **cenome** — as in "biocenose" — to express the fact that these genes correspond to a specific niche)

adanchin@pasteur.fr

FROM THE PALEOME TO THE CENOME

Life manifests first by growth and repair of weathering: the corresponding genome exists since the origin, it is the paleome. Exploration of the environment is an inevitable consequence of existence, it results from continuous creation and exchange of the genes which form the cenome.



adanchin@pasteur.fr

INSTITUT PASTEUR

PERSISTENT GENES ARE GROUPED TOGETHER

Analysis of 228 genomes each coding for more than 1500 genes, shows that persistent genes that tend to stay clustered.

This « mutual attraction » constitutes a network made of three concentric circles that group together the functions identified in the following scenario of the origin of life

adanchin@pasteur.fr

INSTITUT PASTEUR

- A computational view of what life is
- A map of the cell in the chromosome?
- Translation organises the bacterial genome
- The paleome: persistent genes
- The paleome is an archive of the origin of life

adanchin@pasteur.fr

INSTITUT PASTEUR

CAVEAT

“ The experimental method whereby it is proposed to find the evolutionary precursors of protoplasm is to examine present-day biochemical reactions in protoplasm and seek to relate them to reactions that may have occurred and may still occur in the minerals around us. ”

Samuel Granick, 1957

“Arguments that attempt to extrapolate from modern biochemistry back to the origin of life are futile”

Steven Benner, 1987

adanchin@pasteur.fr

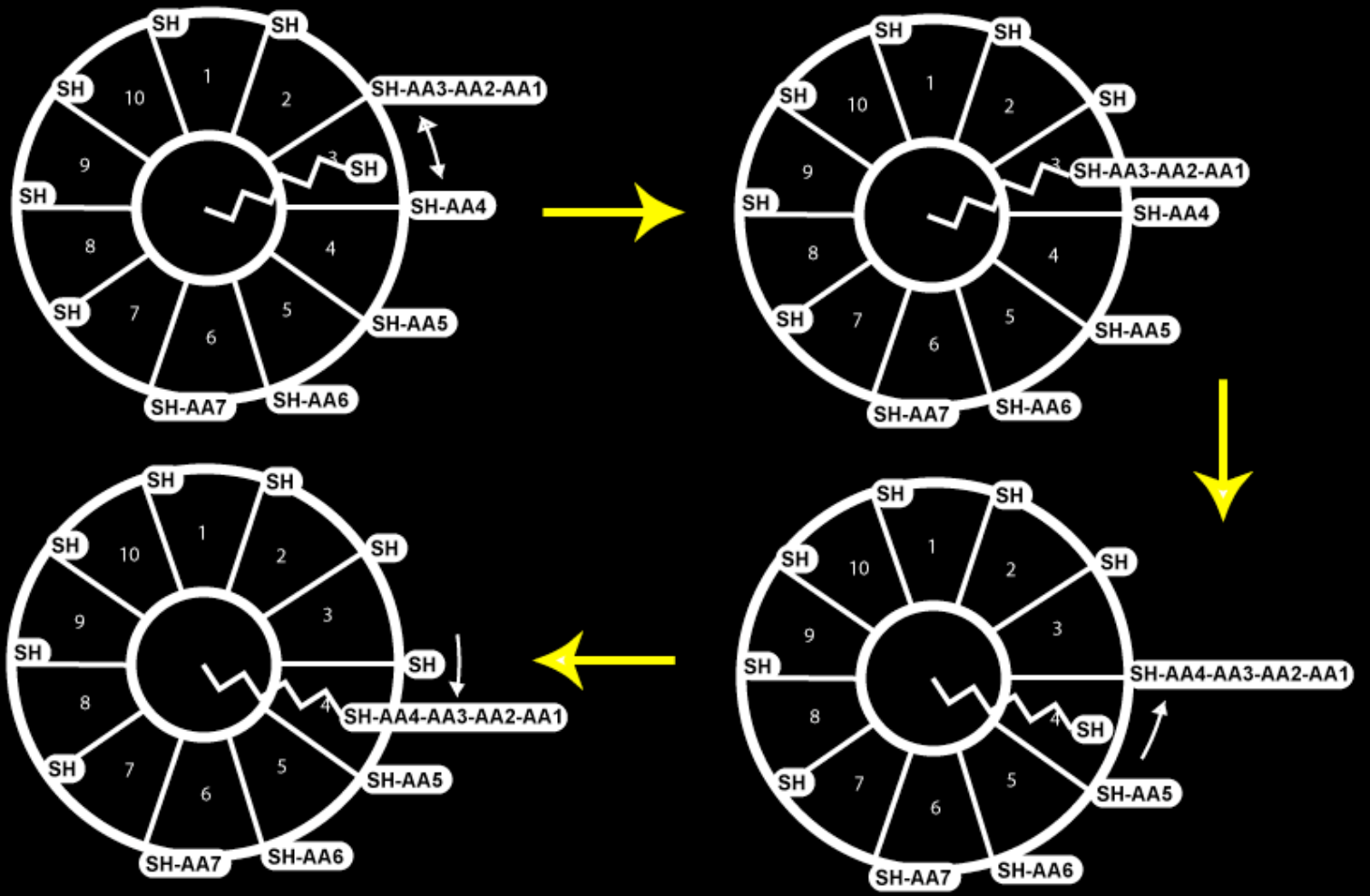
FIRST STEP: SURFACE METABOLISM

IN THIS SCENARIO, CHARGED SOLID SURFACES SELECT MOLECULES FROM THE ENVIRONMENT AND FAVOUR POLYMERISATION

BASIC BUILDING BLOCKS (AMINO ACIDS, SHORT (ISO)PEPTIDES, NUCLEOTIDES, LIPIDS, COENZYMES) ARE CONSTRUCTED IN A PROCESS OF « VARIATION UPON A THEME » HOMEOTOPIC TRANSFORMATION

A Danchin
Homeotopic transformation and the origin of translation
Prog Biophys Mol Biol (1989) **54**:81-86.

NON RIBOSOMAL PEPTIDE (AND LIPID) SYNTHESIS

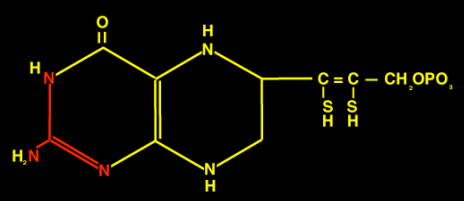


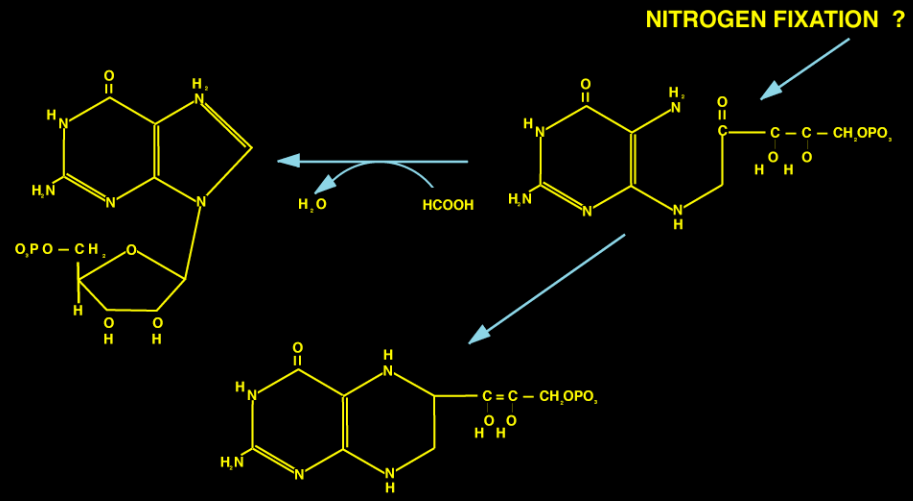
Origin

adanchin@pasteur.fr

NUCLEOTIDES ARE VERY UNSTABLE: THEY MUST BE PRODUCED CONTINUOUSLY
A SULFUR COENZYME, MOLYBDOPTERIN, MIGHT GIVE AN IDEA OF A SCENARIO PRODUCING NUCLEOTIDES AS A « LEAK » FROM THE PROCESS OF NITROGEN FIXATION.

Molybdopterin





Origin

adanchin@pasteur.fr

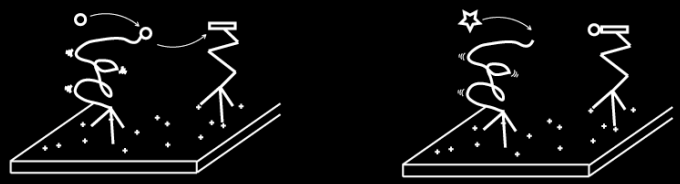
SECOND STEP: THE RNA WORLD

IN THIS SCENARIO TRANSFER RNA PLAYS A CENTRAL ROLE.

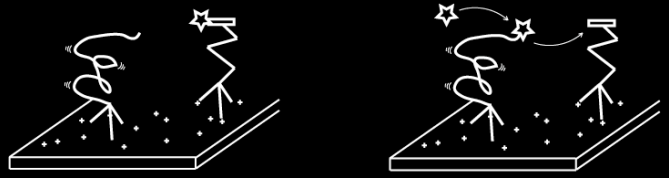
tRNA SYNTHETASES WOULD MAKE THE LINK BETWEEN THE RNA WORLD AND THE WORLD OF ENZYMES. INDEED IT HAS BEEN OBSERVED THAT SOME SYNTHESES (e.g. AMINOLEVULINATE SYNTHESIS) REQUIRE tRNA IN NON TRANSLATION-RELATED PROCESSES.

RNA CAN DISPLAY AN AUTONOMOUS CATALYTIC ACTIVITY, BIND SMALL MOLECULES (APTAMERS), AND ACT AS TEMPLATES FOR COMPLEMENTARY SEQUENCES.

adanchin@pasteur.fr



HOMEOTOPIC TRANSFORMATION



Origin

tRNA OUT OF TRANSLATION

FORMATION OF A PEPTIDIC LINK
- Gly~ARNt^{Gly} murein
- Phe~ARNt^{Phe} N-terminal
- Leu~ARNt^{Leu} N-terminal
- Arg~ARNt^{Arg} Ubiquitine

HOMEOTOPY
- Met~ARNt^{FMet} –> FMet~ARNt^{FMet}
- Glu~ARNt^{Gln} –> GlnARNt^{Gln}
- Ser~ARNt^{SeCys} –> SeCys~ARNt^{SeCys}

DIVERSE
- Lys~ARNt^{Lys} Lipids
- Glu~ARNt^{Glu} Aminolevulinate

OTHER MODIFICATIONS
- Base moodifications

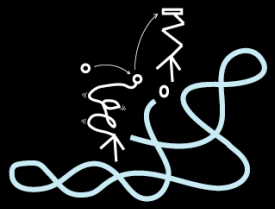
Origine

PATHWAYS USING PHOSPHATE RESIDUES WITH NO "REASON"

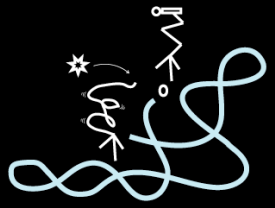
- SERINE (SERINE~PHOSPHATE)
- PYRIDOXAL PHOSPHATE
- ISOPRENYL ~ PYROPHOSPHATE
- 4-PHOSPHOPANTETHEINE
- THIAMINE ~ PYROPHOSPHATE

CONTAINING NUCLEOTIDES :

- CYTIDYLATE (LIPIDES)
- ADENYLATE, GUANYLATE, URIDYLATE (SUGARS)
- COENZYMES (NAD, CoA…)



TRANSFER RNA SUBSTITUTES
TO SURFACES



Origine

Origin

adanchin@pasteur.fr

THIRD STEP: INFORMATION TRANSFER

IN THE RNA WORLD, DISCOVERY OF THE COMPLEMENTARITY LAW DRIVES RNA TO BE USED BOTH AS ENZYME AND AS TEMPLATE.

RNA-MEDIATED ENHANCEMENT OF THE SPECIFICITY OF SYNTHESIZING PROTEINS DISCOVERS THE GENETIC CODE.

IN A WORLD WHERE CELLS FUSE AND SPLIT INSIDE LARGE POPULATIONS, A GENETIC BUILD UP IS DISCOVERED WITH INVENTION OF DEOXYRIBONUCLEOTIDES AND DNA.

A SCENARIO FOR THE ORIGIN OF LIFE

- THE SURFACE OF CHARGED SOLIDS SUCH AS IRON PYRITE (FE-S) ALLOWS THE SELECTION AND PRIMITIVE COMPARTIMENTALIZATION OF CHARGED MOLECULES; POLYMERISATION WITH ELIMINATION OF WATER MOLECULES IS FAVOURED BY AN ENTROPIC EFFECT; THIS FIRST STEP FORMS SOME AMINOACIDS, THE MAIN COENZYMES, FATTY ACIDS AND RIBONUCLEOTIDES

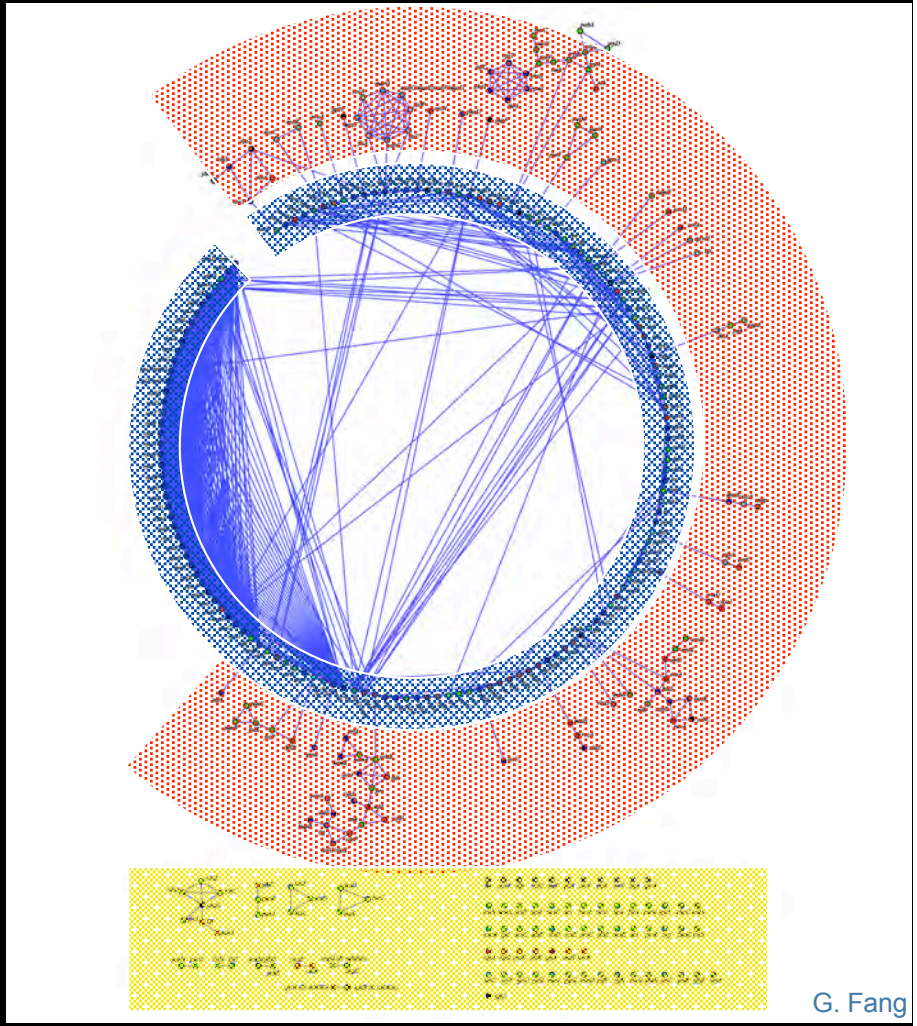
- ONCE COMPARTMENTALISED, METABOLISM CREATES SUBSTITUTES FOR SURFACES (THE RNA WORLD) VIA POLYMERISATION OF RIBONUCLEOTIDES IN THE PRESENCE OF PEPTIDES; RNAs DISCOVER THE COMPLEMENTARITY LAW, AND THE GENETIC CODE IS INVENTED

- NUCLEIC ACIDS ARE STABILISED BY THE INVENTION OF DEOXYRIBONUCLEOTIDES, A THE TIME WHEN THE RULES CONTROLLING INFORMATION TRANSFER ARE DISCOVERED, FIRST WITHIN THE RNA WORLD WHERE VESICLES CARRYING THE ANCESTORS OF GENES SPLIT AND FUSE RANDOMLY, BEFORE FORMATION OF THE FIRST GENOMES

INSTITUT PASTEUR

PERSISTENT GENES NETWORKING RECAPITULATES THE ORIGIN OF LIFE

The external network, made of genes of intermediary metabolism (nucleotides and coenzymes, lipids), is highly fragmented; the middle network is built around class I tRNA synthetases, and the inner network, almost continuous, organised around the ribosome, transcription and replication manages information transfers



G. Fang

A Danchin, G Fang, S Noria
The extant core bacterial proteome is an archive of the origin of life
Proteomics. (2007) **7**:875-889

adanchin@pasteur.fr

INSTITUT PASTEUR

EXISTENCE IMPLIES CLUSTERED PERSISTENCE

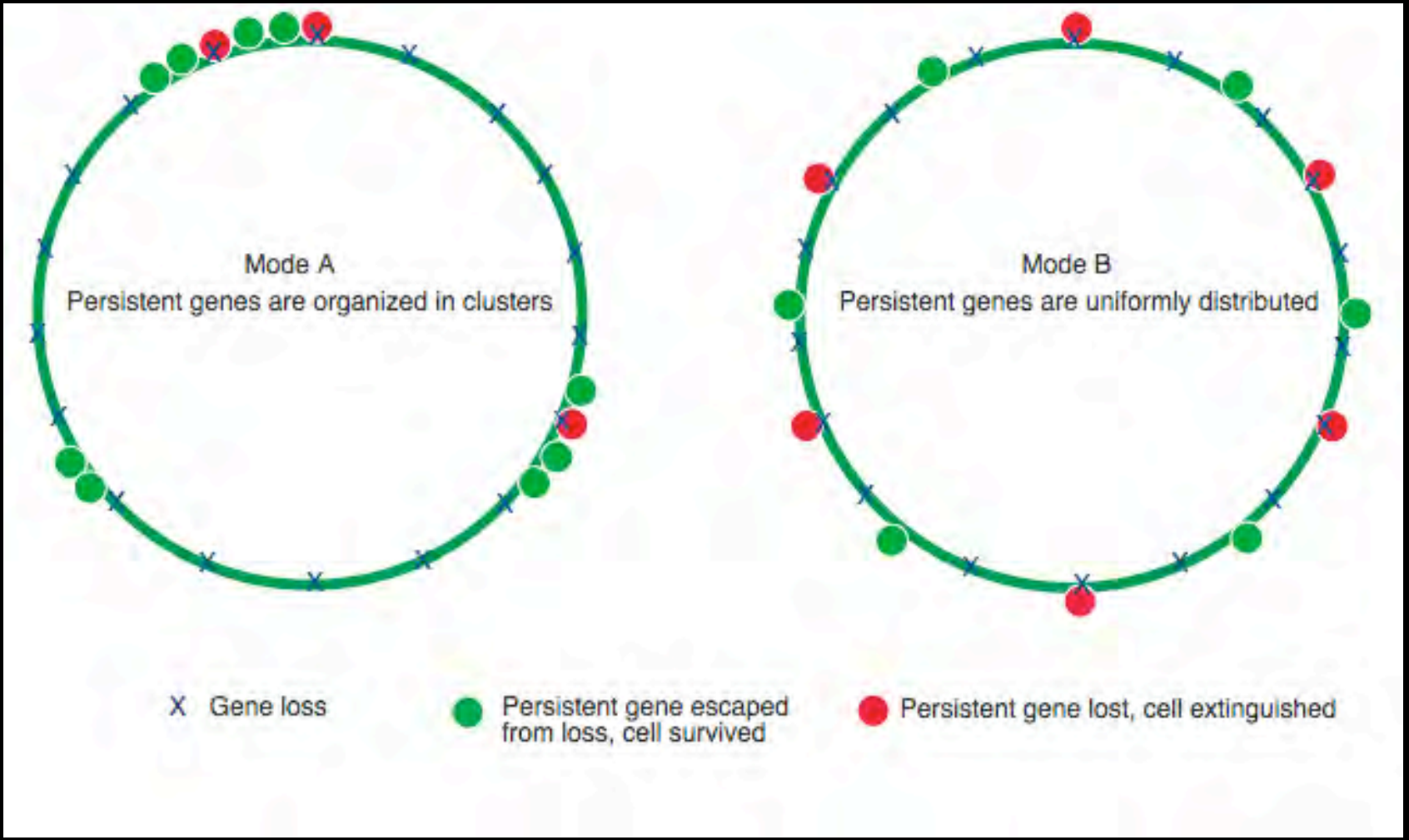
Why are persistent genes clustered? A simple model shows that if, in addition to horizontal gene transfer, there is a process deleting genes in bundles, then any gene contributing to fitness frequently enough over generations will tend to cluster with other genes with similar properties. This accounts 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!

adanchin@pasteur.fr

INSTITUT PASTEUR

EXISTENCE IMPLIES CLUSTERING: NO INTELLIGENT DESIGN!



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

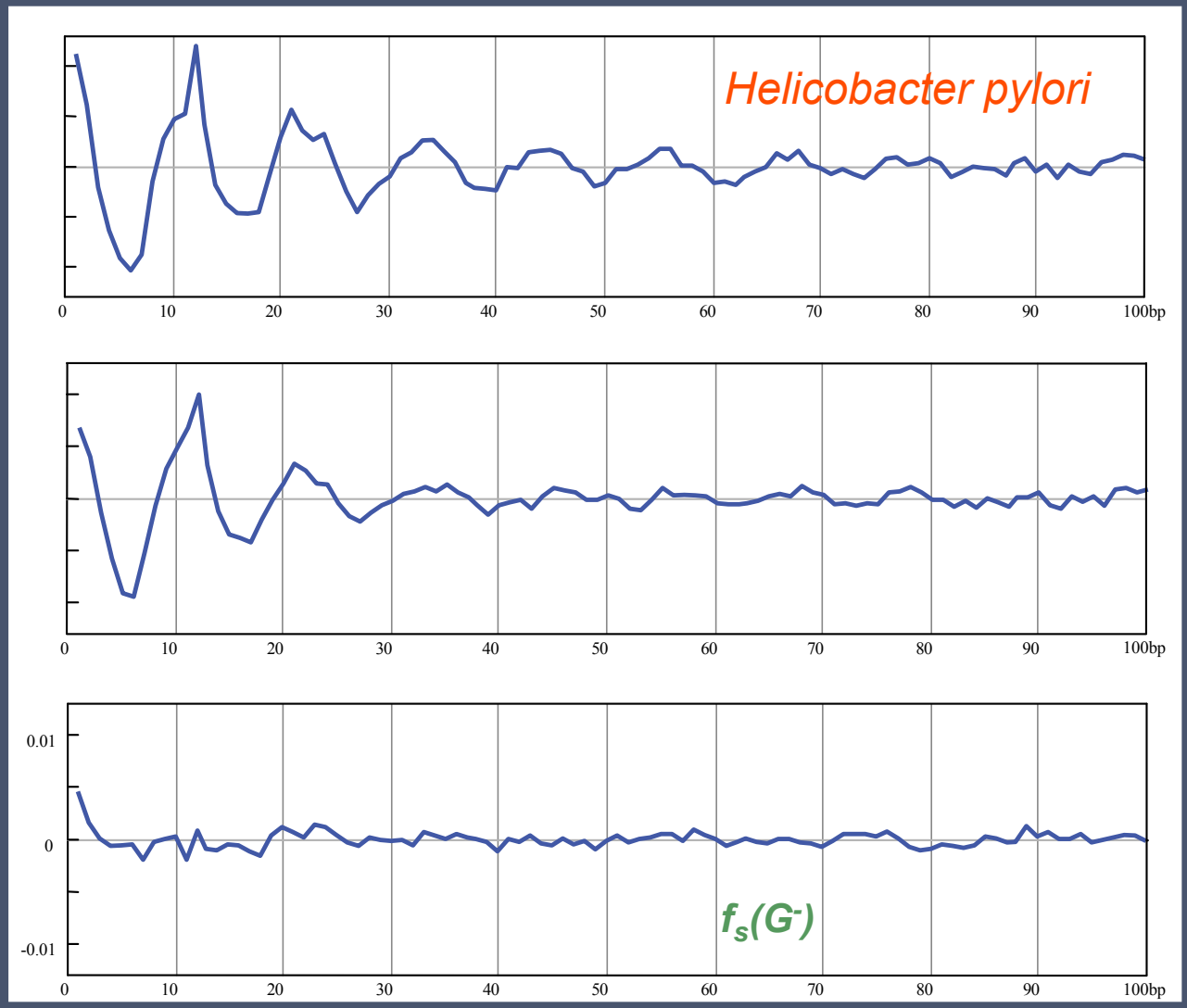
adanchin@pasteur.fr

INSTITUT PASTEUR

THANK YOU

adanchin@pasteur.fr

A universal feature of the program: the period of 10-11.5



adanchin@pasteur.fr

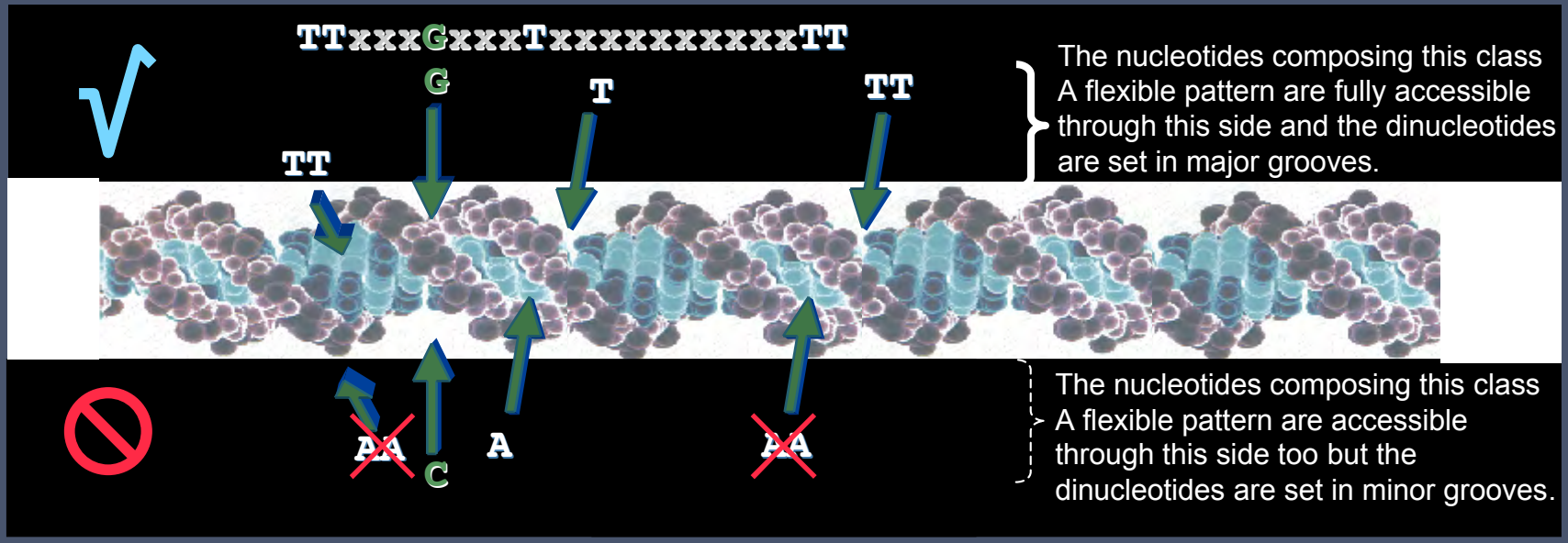
Flexible motifs of type A

1-xAxxxxTxxxxAxxxxTTxxxxxAxxxxTxxxxAxxx: All kindoms

2-xxxxxxxxxxxGxxxxTTxxxCxxxxxTxxxxxxxxx: Proteobacteria

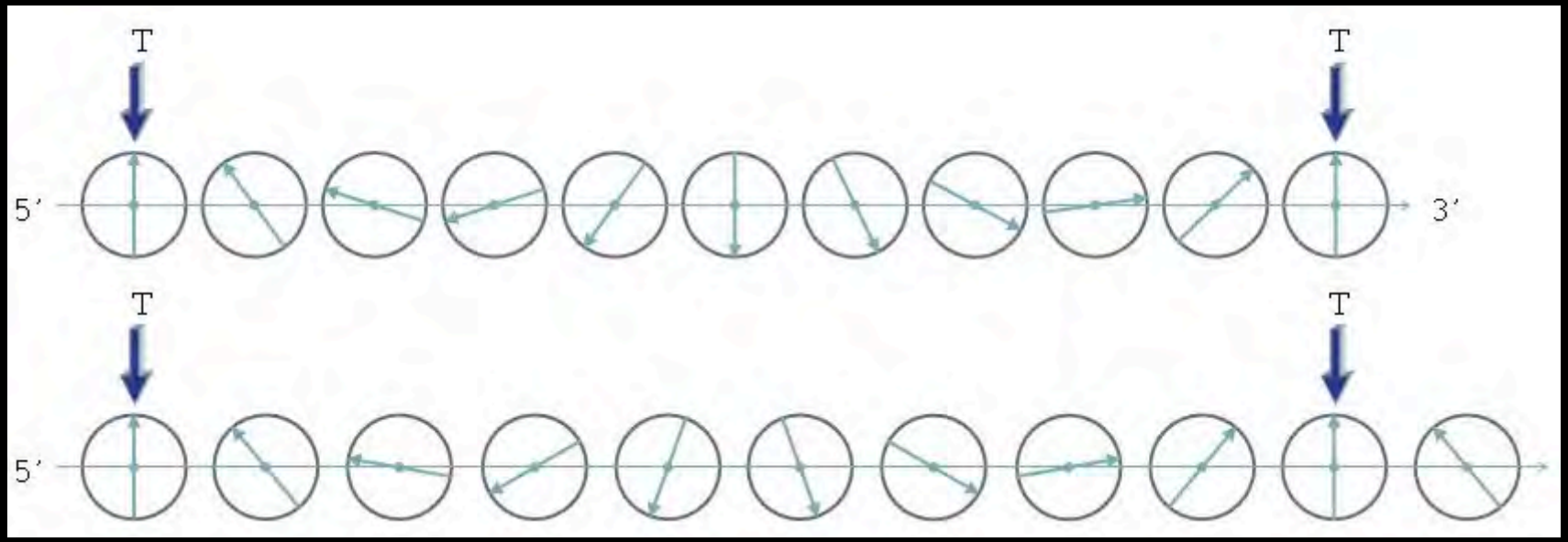
4-xxxxxxTxxxxAGxxxTTxxxxxxxxTxxxxxxxxxx: Archaea

5'-xxx-10xxxxxxxxx0xxxxxxxx10xxxxxxbp-3'



adanchin@pasteur.fr

A universal rule: class A flexible patterns



The flexible nature of the patterns permits DNA to accomodate superturns or local bending

adanchin@pasteur.fr