CAN WE DESIGN DE NOVO A BACTERIAL GENOME?

Synthetic Biology 3.0 Conference
Zürich, 25 june 2007

Genetics of Bacterial Genomes
http://www.pasteur.fr/recherche/unites/REG/

AUTHORS

Génétique des Génomes Bactériens (in silico)

- Gang Fang
- Etienne Larsabal
- Eduardo Rocha

Génétique in silico

- Marc Bailly-Béchet
- Massimo Vergassola

Abdus Salam International Center in Theoretical Physics

- Mudassar Iqbal
- Matteo Marsili

INSTITUT PASTEUR

SYNOPSIS

- Physics will help us, not hinder our efforts to construct a synthetic cell, provided we take the right constraints into account [we should also remember chemical constraints]
- Biological constraints are essential, including:
  - The paleome, reminiscent of a scenario of the origin of life, which forms a replicator and a constructor, as in a « living computer »
  - The cenome, which allows cells to live in and explore a particular niche [cf « biocenose »]
- A synthetic genome needs to put together counterparts of the first class. The second class, which could be the subject of specific design will achieve the goal of the construct

- LIFE AND COMPUTATION
- SOME SIMPLE PHYSICAL CONSTRAINTS
- TRANSLATION ORGANIZES THE BACTERIAL GENOME
- THE PALEOME: CONSTRUCTOR AND REPLICATOR
- THE CENOME: THE “PURPOSE” OF THE MACHINE
- TOWARDS “SYNTHETIC BIOLOGY”

INSTITUT PASTEUR

WHAT LIFE IS

Three processes 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:

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

The cell is the atom of life

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

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

INSTITUT PASTEUR

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?

INSTITUT PASTEUR

WHAT COMPUTING IS

Two entities permit 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

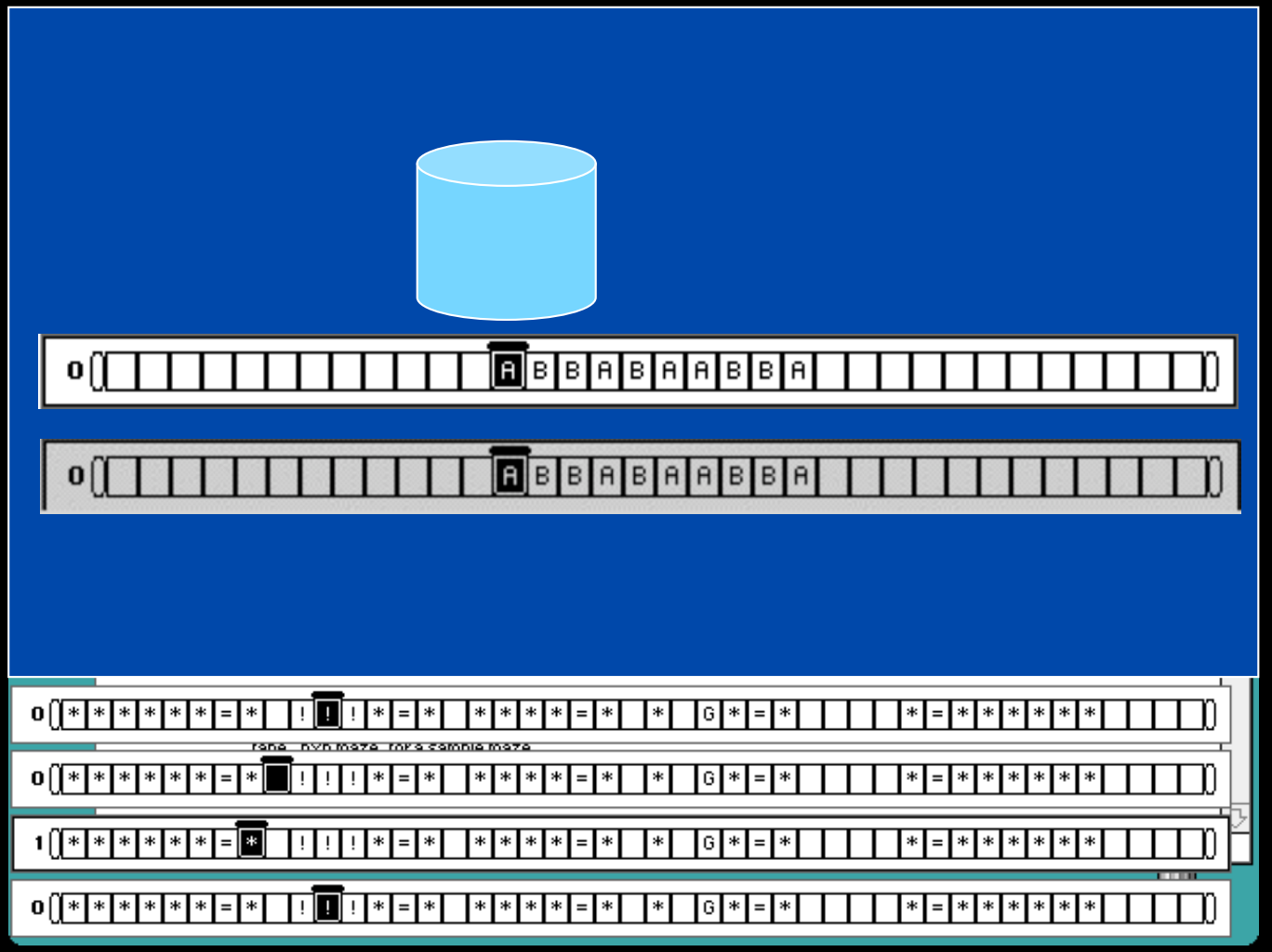
INSTITUT PASTEUR

THE TURING MACHINE

the machine
(read/write)

is physically distinct from

the program
(data)
as a linear
sequence
of symbols



INSTITUT PASTEUR

CELLS AND COMPUTERS

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

Horizontal Gene Transfer

Viruses

Genetic engineering

Animal cloning (and now direct transformation of a whole genome into a recipient cell)

all points to separation between

« Machine » (the cell factory)

and

Data + program

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**

INSTITUT PASTEUR

A RECURSIVE MACHINE

- **Replicator**: DNA specifies proteins that replicate DNA

- **Constructor**: DNA specifies proteins which form the machine that constructs the cell

INSTITUT PASTEUR

IS THERE A MAP OF THE CELL IN THE CHROMOSOME?

John von Neumann, trying to understand the brain, suggested that were the computer both to behave as a computer and to construct the machine itself, it should harbour an image of the machine somewhere.

That special computer had 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?

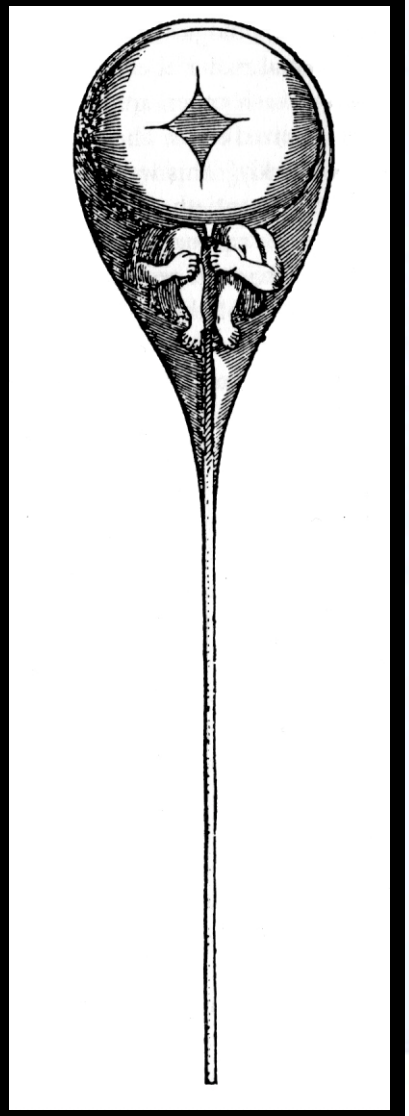
INSTITUT PASTEUR

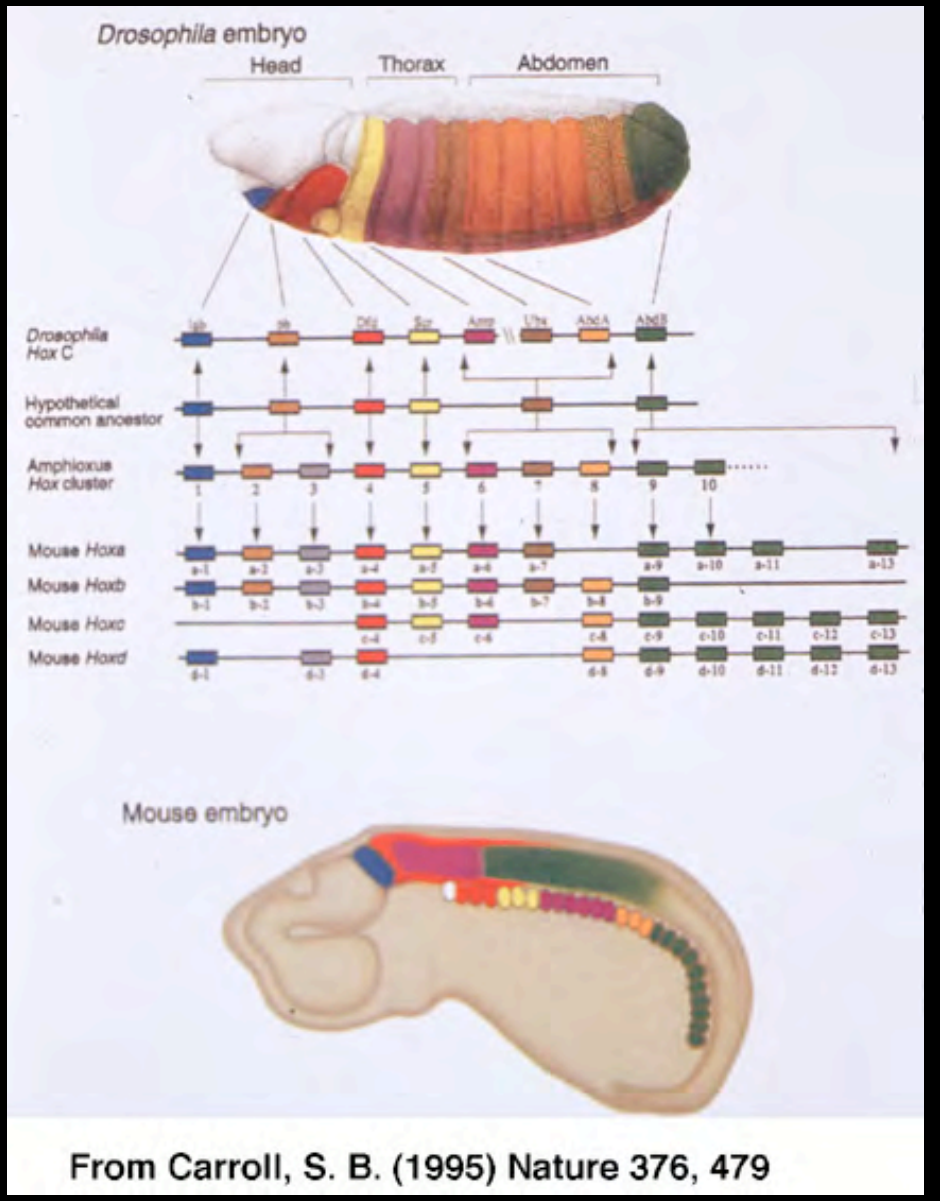
The mystery of homeogenes' origin

Drosophiloculus,

Homunculus?

Celluloculus?





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

- LIFE AND COMPUTATION
- SOME SIMPLE PHYSICAL CONSTRAINTS
- TRANSLATION ORGANIZES THE BACTERIAL GENOME
- THE PALEOME: CONSTRUCTOR AND REPLICATOR
- THE CENOME: THE "PURPOSE" OF THE MACHINE
- TOWARDS "SYNTHETIC BIOLOGY"

INSTITUT PASTEUR

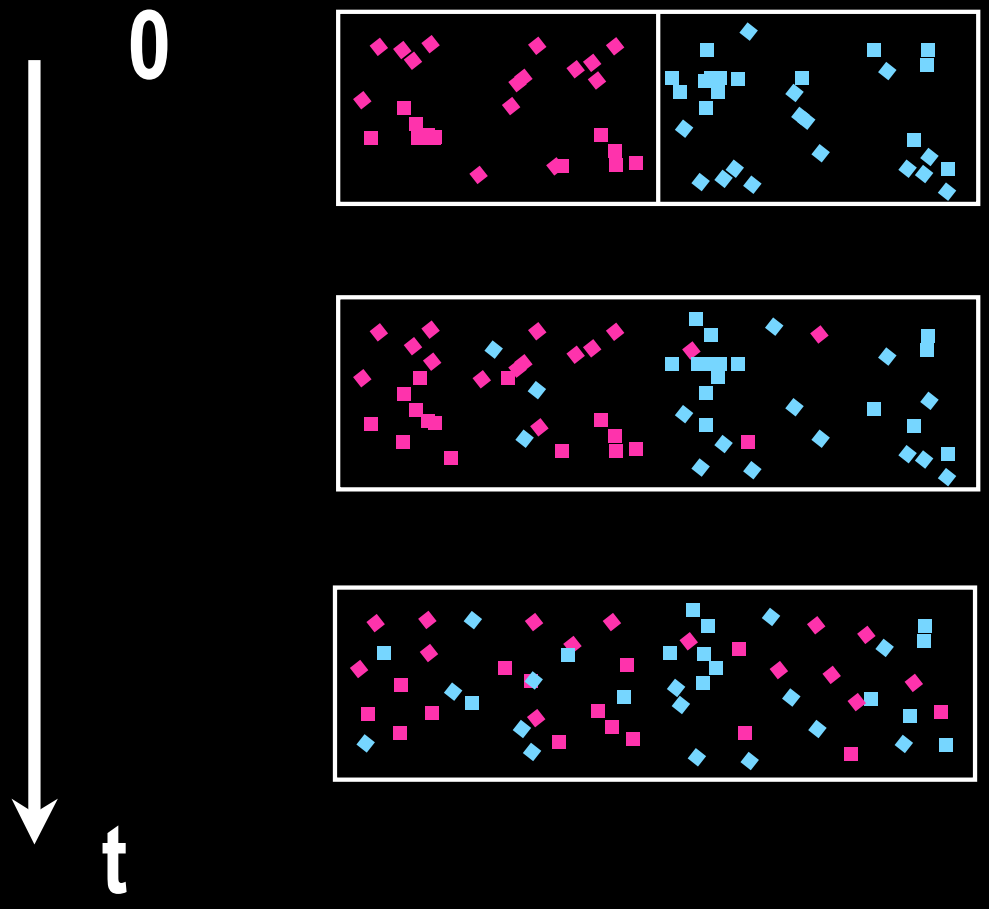
THE PHYSICS OF REPLICATION

- DNA forms a long folded thread: how do the daughter molecules separate?
- Are physical constraints reflected in the sequence
- [Replication is oriented: the physics of a strand cannot be that of its complement]

INSTITUT PASTEUR

A widely spread idea links entropy with disorder

$$S = k \log \Omega$$



Benjamin Crowell, licensed under the Creative Commons Attribution-ShareAlike license

INSTITUT PASTEUR

However an increase in entropy is enough to separate chromosomes

Jun S, Mulder B. Entropy-driven spatial organization of highly confined polymers: lessons for the bacterial chromosome .Proc Natl Acad Sci U S A. 2006 103:12388-93.

A NEED FOR SYNTHETIC BIOLOGY

Evolution optimises this physical phenomenon, while DNA needs also to support gene sequences

This is witnessed by:

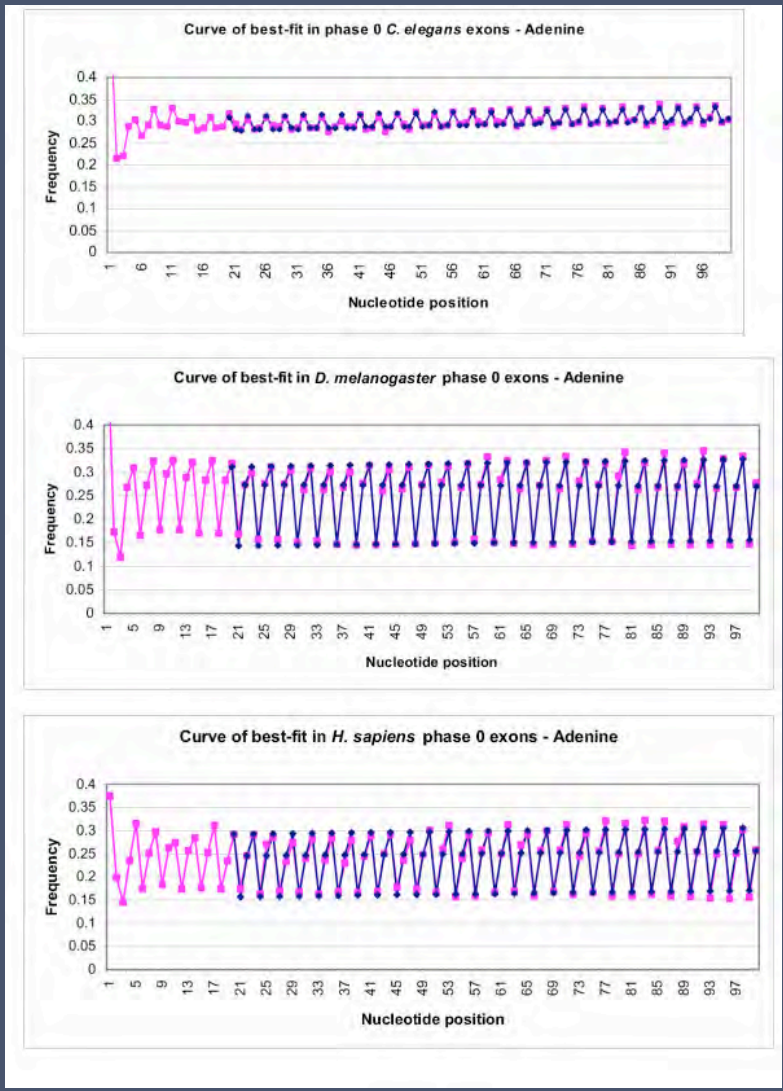
- A period 3, signature of the codon succession in genes (constrained by the genetic code rule)

- A period 10-11.5 of yet unknown function...

PERIODS IN GENOMES

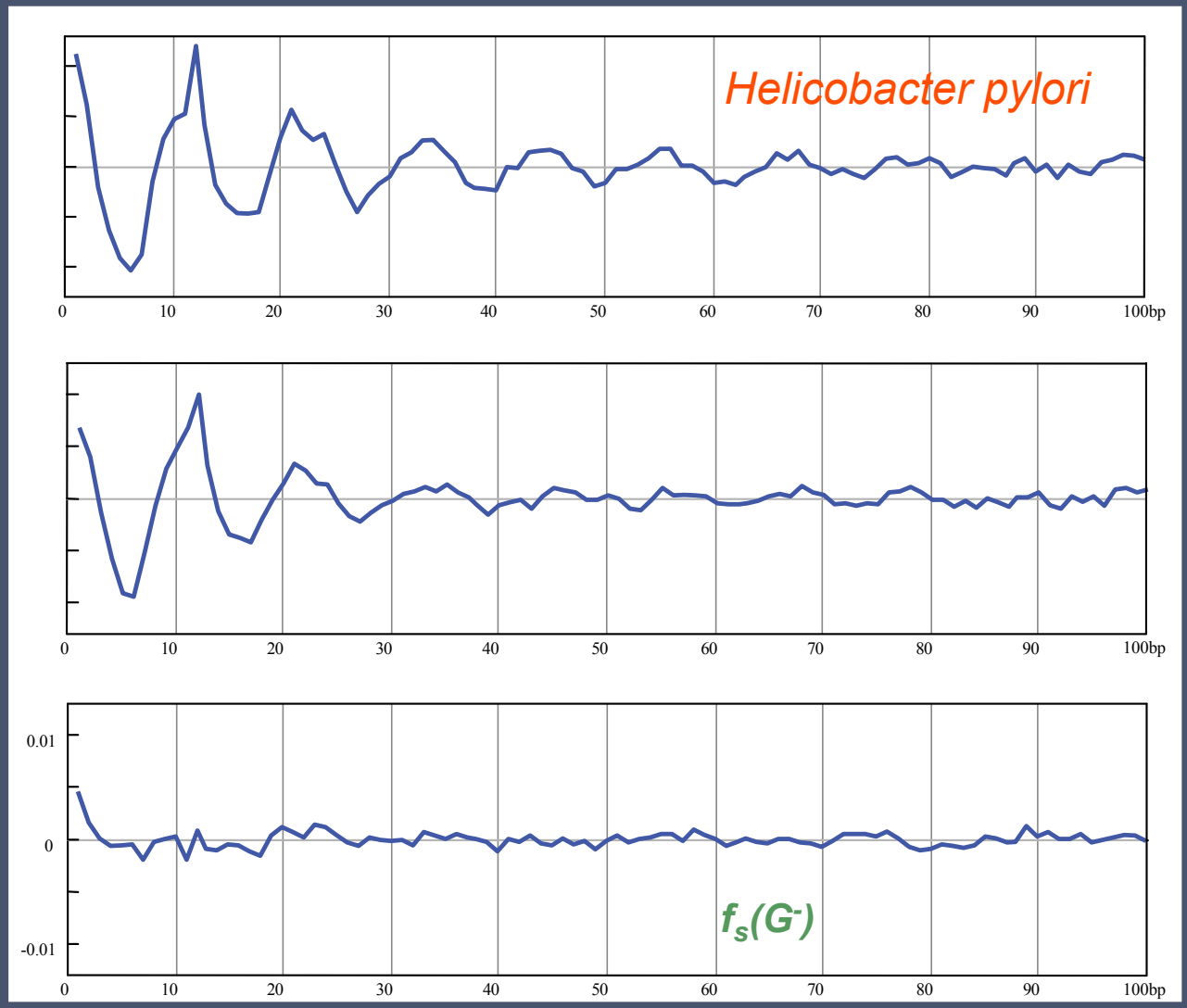
One observes a correlation between base pairs with period three.

After deconvolution of this period there remains a somewhat fuzzy period of 10 to 11.5 base pairs



Eskesen et coll. BMC Molecular Biology Volume 5, 12, 2004

A UNIVERSAL FEATURE OF THE GENOME TEXT: 10-11.5



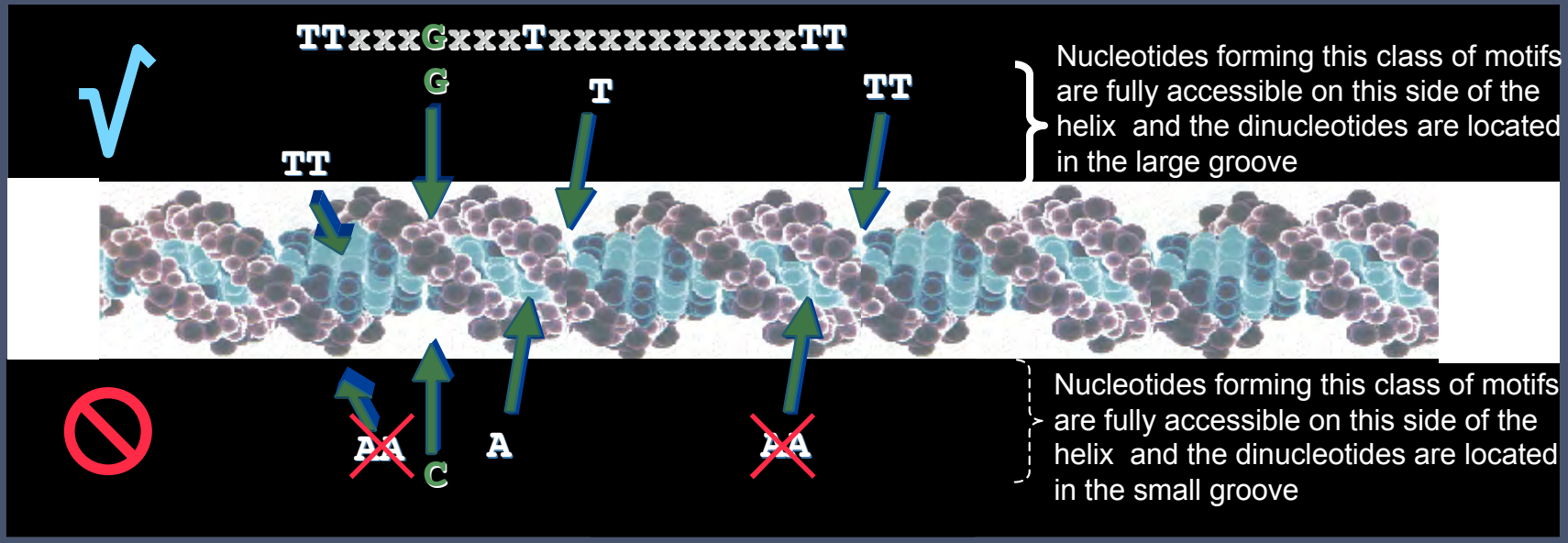
TYPE A FLEXIBLE MOTIFS

1-xAxxxxTxxxxAxxxxTTxxxxxAxxxxTxxxxAxxx: All domains

2-xxxxxxxxxxxGxxxxTTxxxCxxxxxTxxxxxxxxx: Proteobacteria

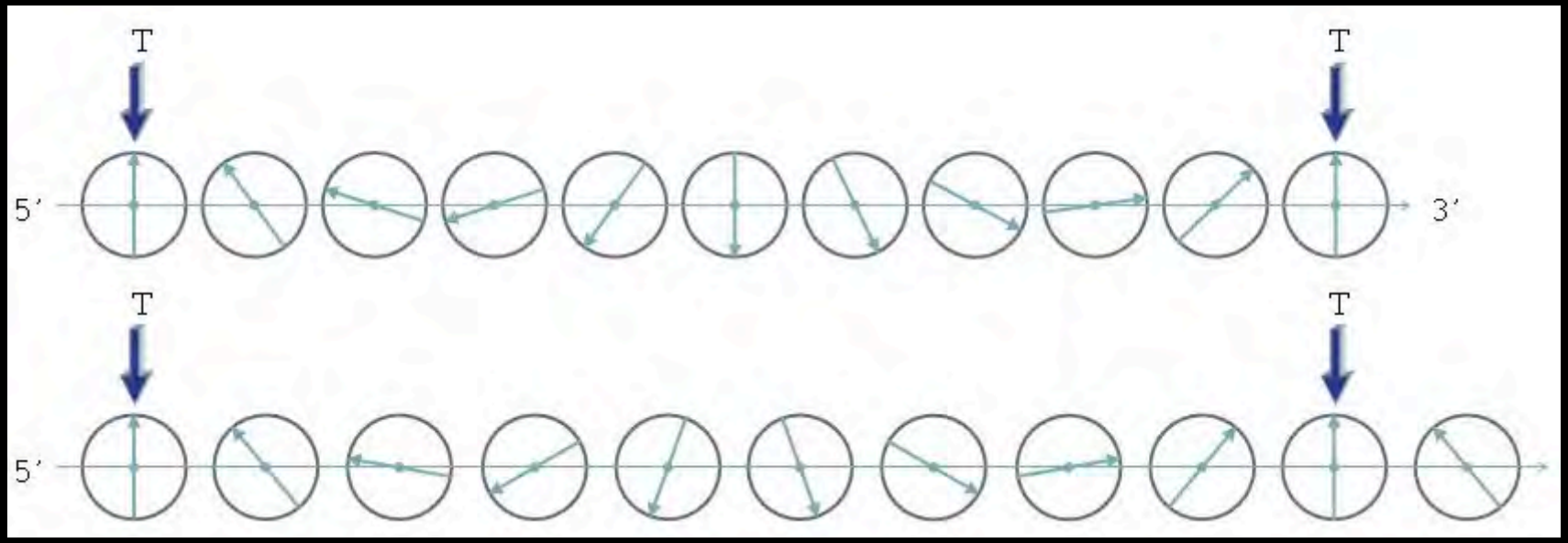
4-xxxxxxTxxxxAGxxxTTxxxxxxxxTxxxxxxxxxx: Archaea

5'-xxx-10xxxxxxxxx0xxxxxxxx10xxxxxxbp-3'



INSTITUT PASTEUR

FLEXIBLE MOTIFS ACCOMMODATE LOCAL VARIATIONS OF THE DNA STRUCTURE



The flexiblility of these motifs allow DNA to take into account superturns and bends

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

- LIFE AND COMPUTATION
- SOME SIMPLE PHYSICAL CONSTRAINTS
- TRANSLATION ORGANIZES THE BACTERIAL GENOME
- THE PALEOME: CONSTRUCTOR AND REPLICATOR
- THE CENOME: THE “PURPOSE” OF THE MACHINE
- TOWARDS “SYNTHETIC BIOLOGY”

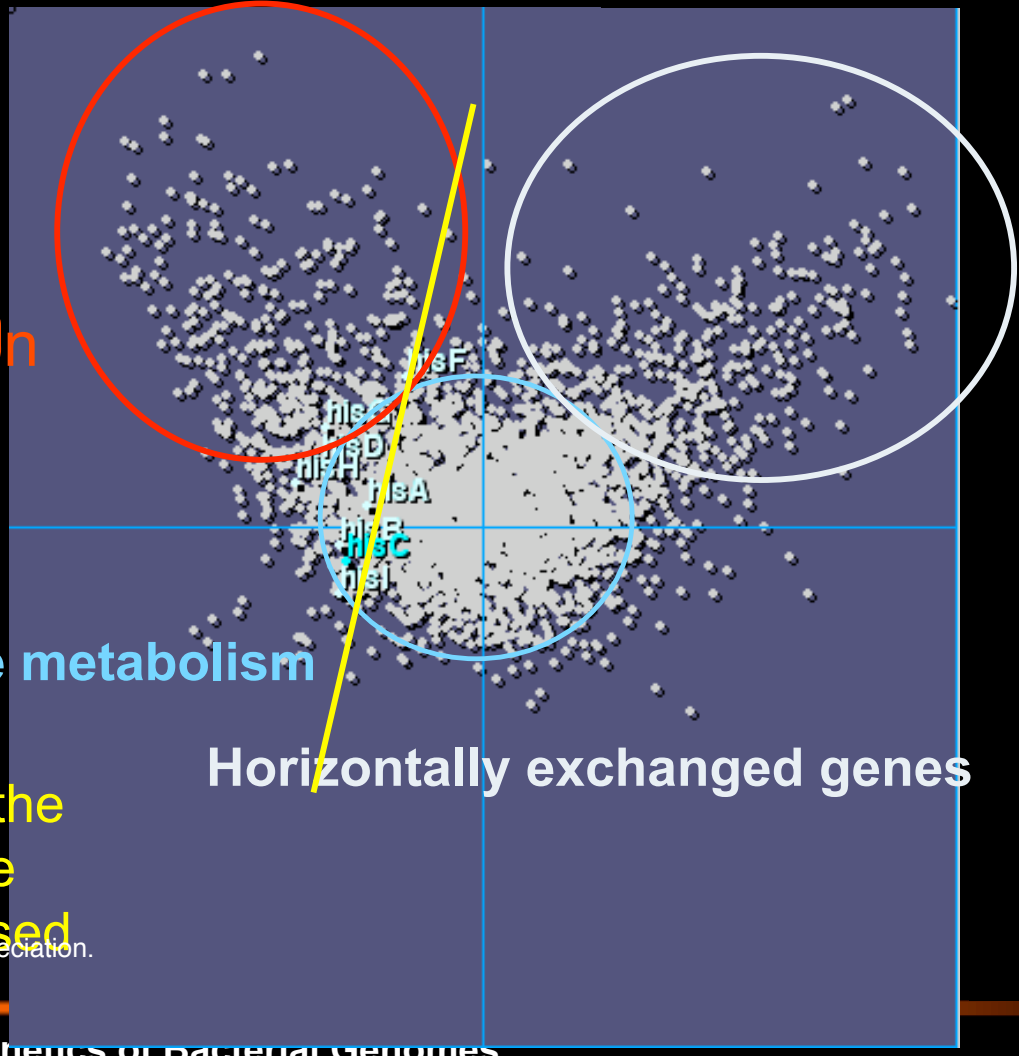
Codon usage biases are associated to func

Genes highly expressed in exponential growth

Class I: central metabolism

Class II : high expression in exponential growth

Class III : horizontal transfer



e.g. the codon usage of the gene involved in histidine metabolism is highly biased

Medigue C, Rouxel T, Vigier P, Henaut A, Danchin A.
Evidence for horizontal gene transfer in *Escherichia coli* speciation.
J Mol Biol. 1991 222:851-856.

INSTITUT PASTEUR

LOCAL BIASES OF CODON USAGE

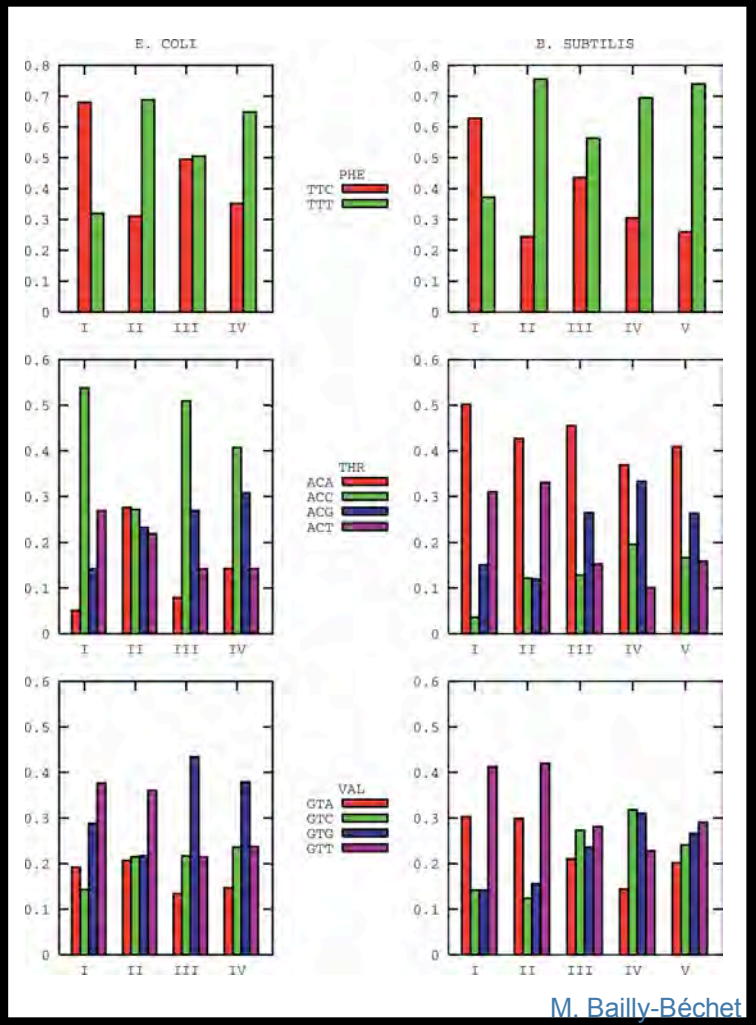
Correspondence Analysis shows that genes with similar biases are functionnally related. How is this reflected in the chromosome?

A clustering method based on information theory groups the genes into homogeneous families, which appear not to be randomly spread in the chromosome. The method identifies 4 classes in *E. coli* and 5 in *B. subtilis*)

Genomic translation islands

Genes with similar bias are organized into groups longer than operons, showing some translation-driven organization of the chromosome

A major part of this effect comes from the recycling or rare transfer RNA molecules



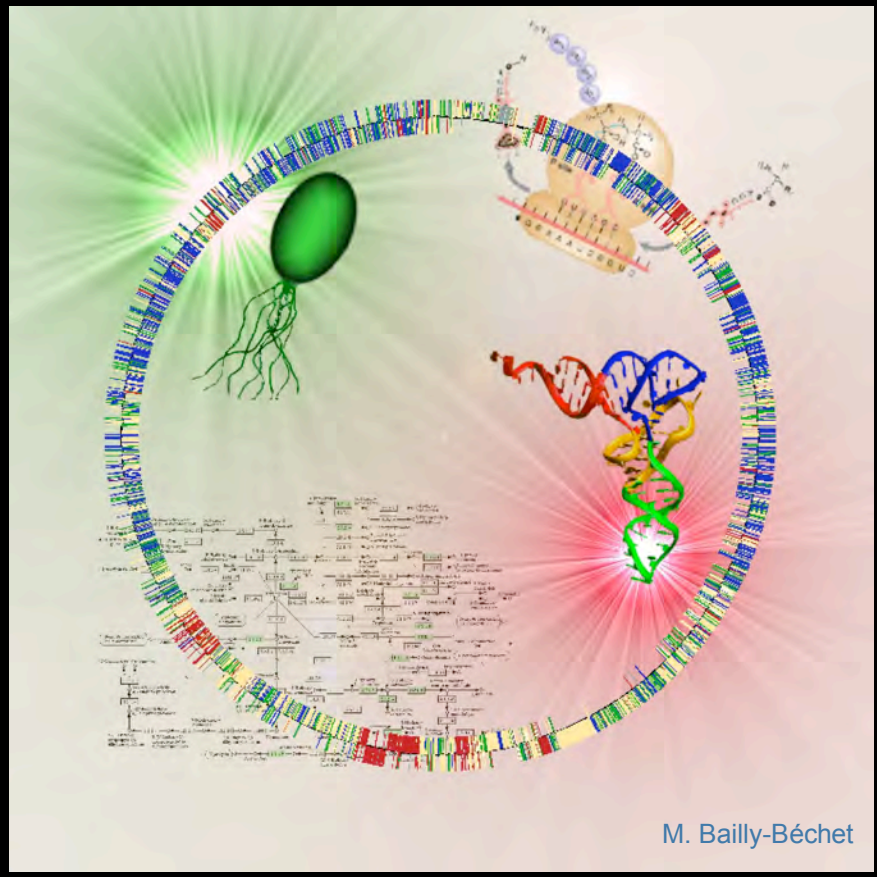
M. Bailly-Béchet

INSTITUT PASTEUR

TRANSLATION ISLANDS

One groups is associated to high expression (blue).

The other groups are also functionnally consistent: horizontally transferred genes (red), motility (yellow) and 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**: april 20th

- LIFE AND COMPUTATION
- SOME SIMPLE PHYSICAL CONSTRAINTS
- TRANSLATION ORGANIZES THE BACTERIAL GENOME
- THE PALEOME: CONSTRUCTOR AND REPLICATOR
- THE CENOME: THE “PURPOSE” OF THE MACHINE
- TOWARDS “SYNTHETIC BIOLOGY”

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

INSTITUT PASTEUR

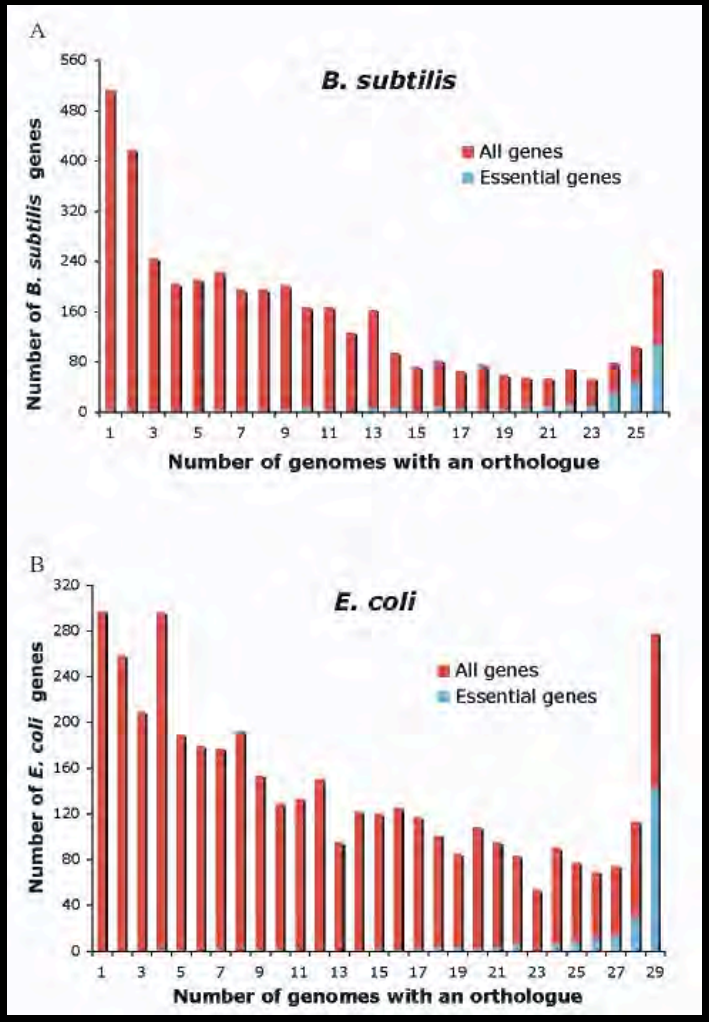
PERSISTENT GENES

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.

INSTITUT PASTEUR

GENE PERSISTENCE



Which functional category?

- Information transfer
- Compartmentalization
- Intermediary metabolism
- Stress, maintenance and repair

Highly non random!

INSTITUT PASTEUR

PERSISTENT GENES ARE CLUSTERED

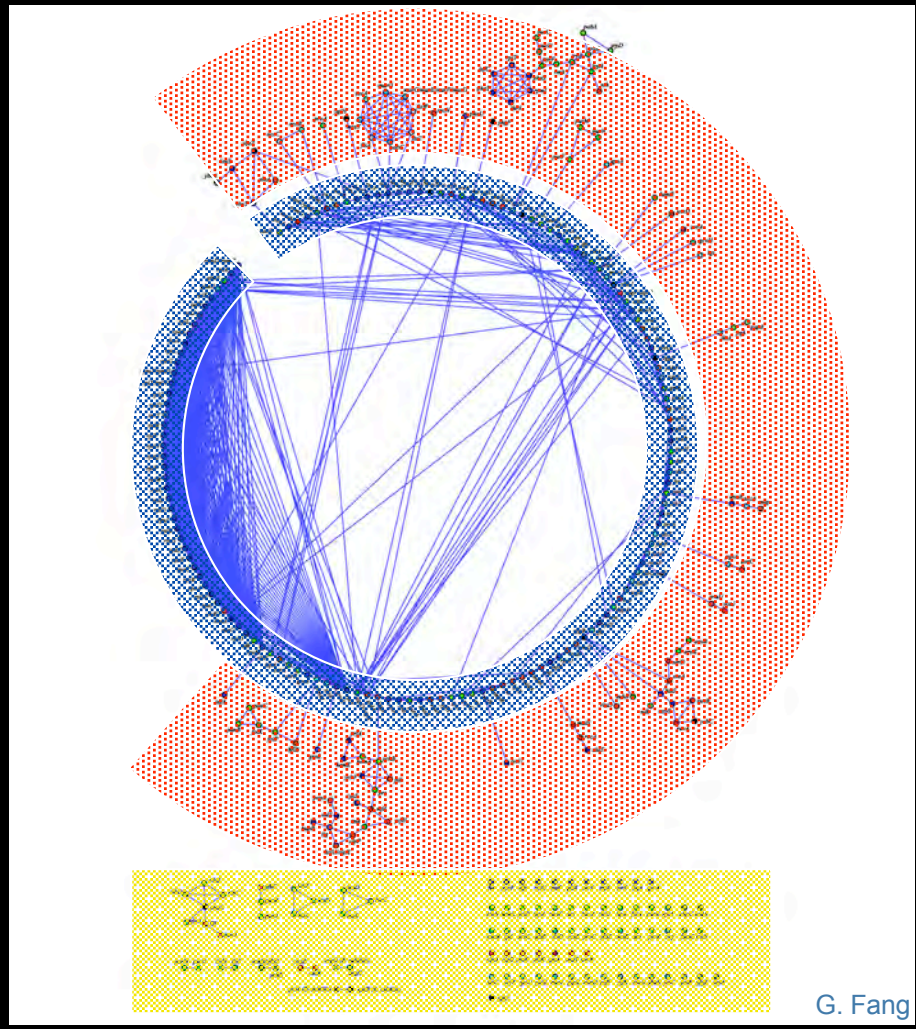
Persistent genes are functionally defined

The way they group along chromosomes in more than 250 bacteria displays three clusters that reflect a scenario of the origin of life. This is why it is proposed to name paleome (from παλαιος, ancient) this group of core genes

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

Persistent genes recapitulate 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, organized 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

- LIFE AND COMPUTATION
- SOME SIMPLE PHYSICAL CONSTRAINTS
- TRANSLATION ORGANIZES THE BACTERIAL GENOME
- THE PALEOME: CONSTRUCTOR AND REPLICATOR
- THE CENOME: THE "PURPOSE" OF THE MACHINE
- TOWARDS "SYNTHETIC BIOLOGY"

INSTITUT PASTEUR

THE COMPOSITE GENOME

- Expecting **two genome components**, coding for the machine and for the "purpose" of the machine, we need to separate between the **replicator/constructor** and secondary functions

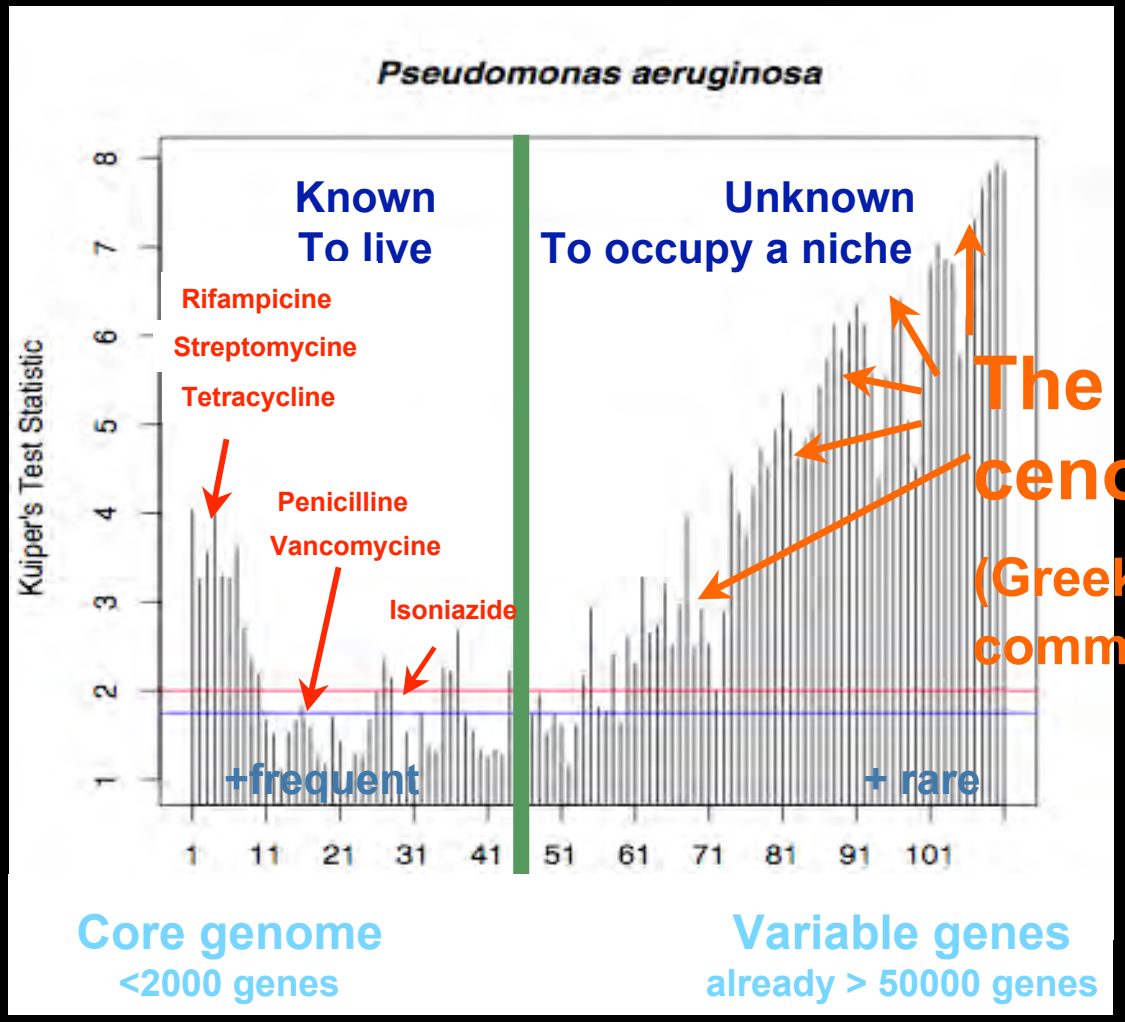
- Extant genomes should comprize ubiquitous functions (not genes!) which would correspond to the former (here named the **paleome**) and functions specific to the environment of the organism (named the **cenome** — as in "biocenose" — to express the fact that these genes correspond to a specific niche)

INSTITUT PASTEUR

THE CENOME

Linkage frequency

vertical bars group 50 genes with similar behaviour together; colored horizontal lines indicate the level of significance of linkage

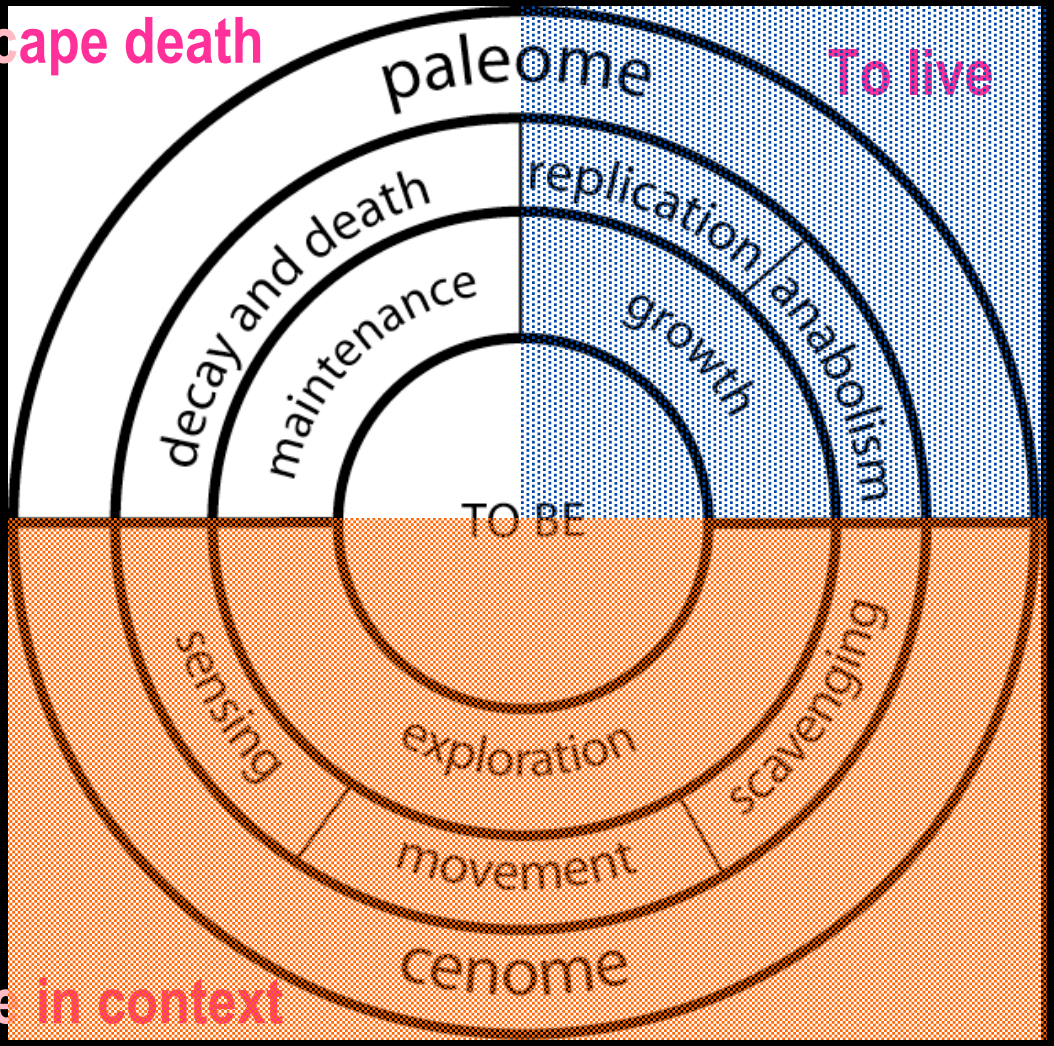


Frequency in genomes

- LIFE AND COMPUTATION
- SOME SIMPLE PHYSICAL CONSTRAINTS
- TRANSLATION ORGANIZES THE BACTERIAL GENOME
- THE PALEOME: CONSTRUCTOR AND REPLICATOR
- THE CENOME: THE “PURPOSE” OF THE MACHINE
- TOWARDS “SYNTHETIC BIOLOGY”

SYNTHESIS: A TALE OF TWO GENOMES

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



A. Danchin. Archives or Palimpsests? Bacterial Genomes Unveil a Scenario for the Origin of Life Biological Theory (MIT Press) (2007) 2: 52-61.

THANK YOU