FRUSTRATION: PHYSICO-CHEMICAL PREREQUISITES FOR THE CONSTRUCTION OF A SYNTHETIC CELL

Antoine Danchin
唐善安東

Systems Chemistry, Beilstein Institut, Bolzano, 27 may 2008

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

ACKNOWLEDGEMENTS

The University of Hong Kong
Dpt of Mathematics and HKU-Pasteur Research Centre

- Stanislas Noria (collective name, working seminar in « Conceptual Biology »)

Génétique des Génomes Bactériens

- Gang Fang
- Etienne Larsabal
- Undine Mechold
- Géraldine Pascal
- Eduardo Rocha
- Agnieszka Sekowska
- Tingzhang Wang

Génoscope AGC

- Claudine Médigue

Marine Biological Laboratory, Woods Hole

- Monica Riley

INSTITUT PASTEUR

THREE REVOLUTIONS

- 1944 - 1985 MOLECULAR BIOLOGY
- 1985 - 2005 GENOMICS
- 2005 - … SYNTHETIC (SYMPLECTIC) BIOLOGY

(highly multidisciplinary !)

Mind media special interest groups!: « Symplectic » is in Greek (συν, together, πλεκτειν, to weave) the same word as « Complex » in Latin; used here to avoid the unwanted fuzzy connotations associated to « Complexity »; a connotation in Geometry will not interfere…

INSTITUT PASTEUR

GOALS OF SB

- A first aim of SB is to reconstruct life, in an endeavour to explore whether we understand what life is and learn missing entities from our failures
- A second aim is to keep the laws defining life, and to apply them using objects of a different physico-chemical nature
- A third aim is to see life from an engineering standpoint, trying to class and normalize « biobricks »

INSTITUT PASTEUR

WHAT LIFE IS

Life requires:

- A program (a “book of recipes”)
  - 1. Recursive information transfer => coding from one level to a second one as an essential element
- A machine allowing the program to be enacted
  - 2. Metabolism (a dynamic process)
  - 3. Compartmentalization (defining an inside and an outside)

The cell is the atom of life

INSTITUT PASTEUR

REPRODUCTION AND REPLICATION

- **The machine reproduces**
  - Reproduction can improve over time: it is always an aged organism that gives birth to a young one
- **The program replicates**
  - Replication keeps accumulating errors

INSTITUT PASTEUR

METABOLISM

- ~ 800 molecules are absolutely necessary ("anabolism and catabolism")
- Assimilation of C, H, N, O, S, P, and ions
- Management of energy (ATP and electron transfers)

"Coenzymes" are essential components of metabolism (generally forgotten in the common scenarios of the origin of life!)

INSTITUT PASTEUR

COMPARTMENTALIZATION

Cells define an inside and an outside

Two strategies separate the domains of life:

- A single envelope (more or less complex): prokaryotes
- Multiplication of membranes and skins: eukaryotes

INSTITUT PASTEUR

INFORMATION TRANSFER

As in the construction of a machine, one needs a « book of recipes » to construct the cell.

This asks to change the text of the recipe into something concrete.

This process transfers « information ».

In the cell, information transfer is organized by the genetic program.

INSTITUT PASTEUR

A NOVEL COMPONENT OF REALITY: INFORMATION

- Physics: *matter, energy, time*
- Statistical physics: Physics + *« information »*

Information reconciles classical physics and quantum physics

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

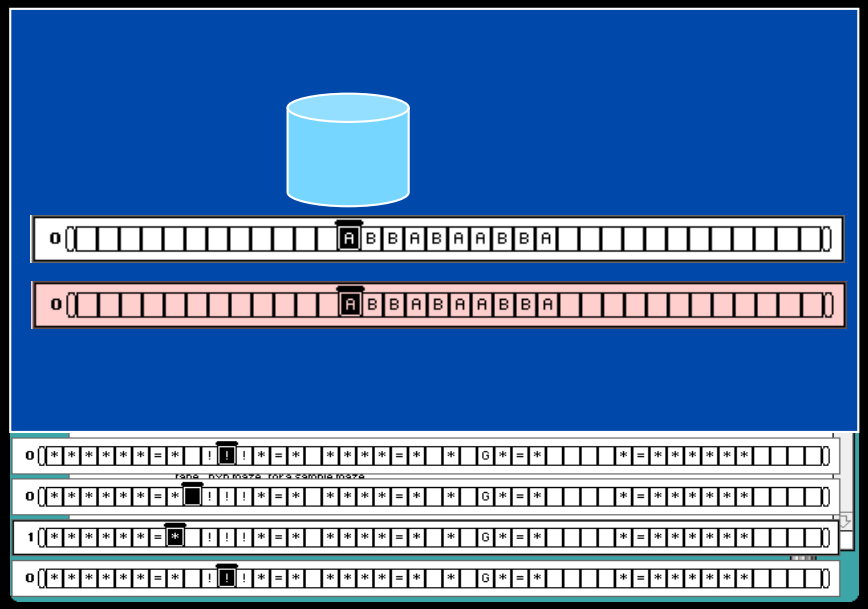
The « genetic program » metaphor has practical consequences: we know how to manipulate genes and gene products, do we have the conceptual tools to push the metaphor to its ultimate consequences?

INSTITUT PASTEUR

WHAT COMPUTING IS

Two entities permit computing:

- A machine able to read and write
- A program on a physical support, split by the human mind (not conceptually!) into two entities:
  - Program (providing the “goal”)
  - Data (providing the context)



The machine is distinct from the data/program

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

Direct transplantation of a naked genome into a recipient cell with subsequent change of the recipient machine into a new one (2007)

all points to separation between

«Machine» (the cell factory) and «Data/Program» (the genome)

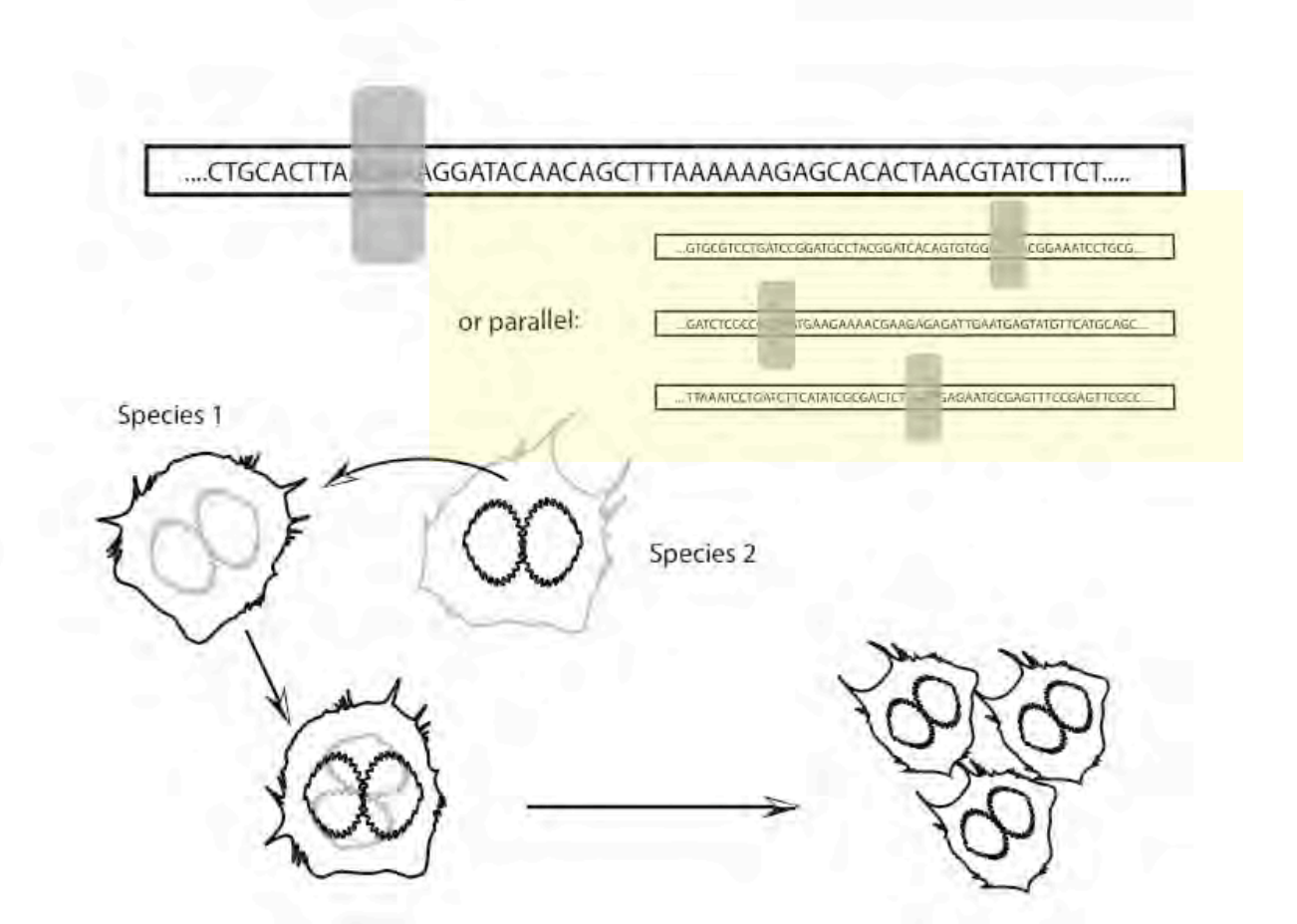
Need: conceptual analysis of biological information (algorithmic complexity, logical depth...)

VENTER'S DEMONSTRATION

The Turing machine

May exist in a parallel set up

Genome transplantation



INSTITUT PASTEUR

« FRUSTRATION »

- The organization just described is conceptual, but it needs to be implemented concretely
- Concrete objects have often properties that are not compatible with those of other objects. This implies « frustration » of possible mutually exclusive entities (because of constraints in space or energy states)
- The cell factory will require construction of « patches » to cope with these incompatibilities

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

Frustration 1 accounts for the origin of regulation: Check points introduce a requirement for regulation, with competition between different requirements, sometimes mutually exclusive.

INSTITUT PASTEUR

IS THERE A MAP OF THE CELL IN THE CHROMOSOME?

John von Neumann, trying to understand the brain, suggested that were a 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

PARADOX

If the machine has not only to behave as a Turing machine but also to make the machine, one must find a geometrical program somewhere in the machine (J. von Neumann)

Is there an image of the organism in the genome?

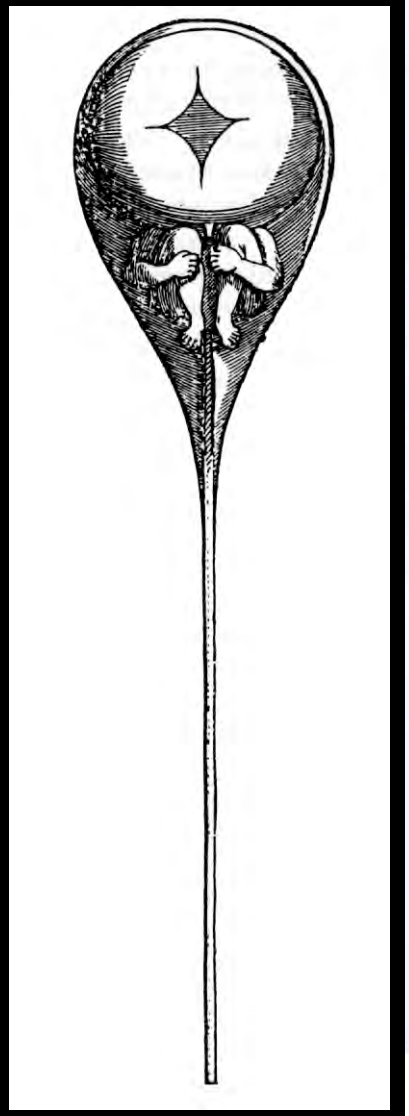
INSTITUT PASTEUR

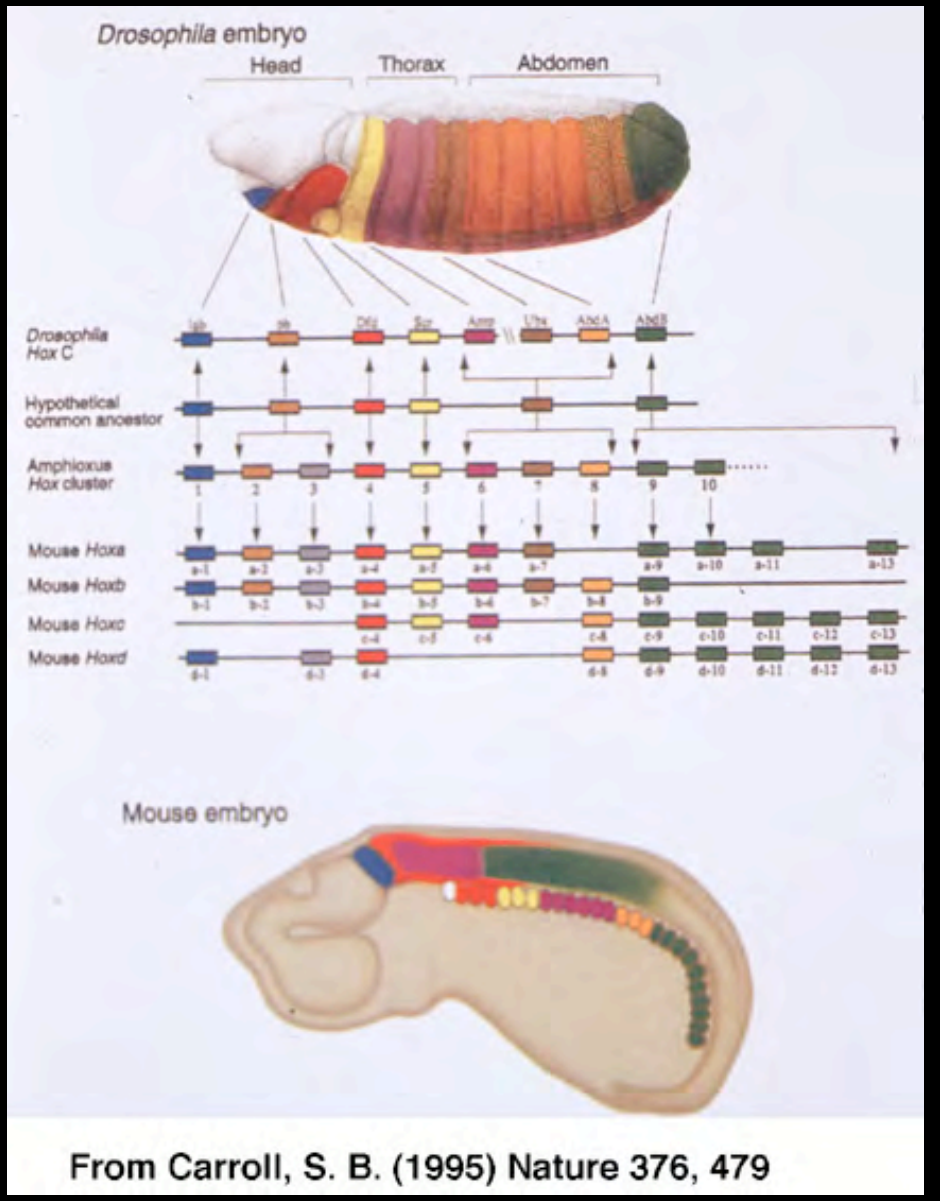
HOMEOGENES

Drosophiloculus,

Homunculus?

Celluloculus?



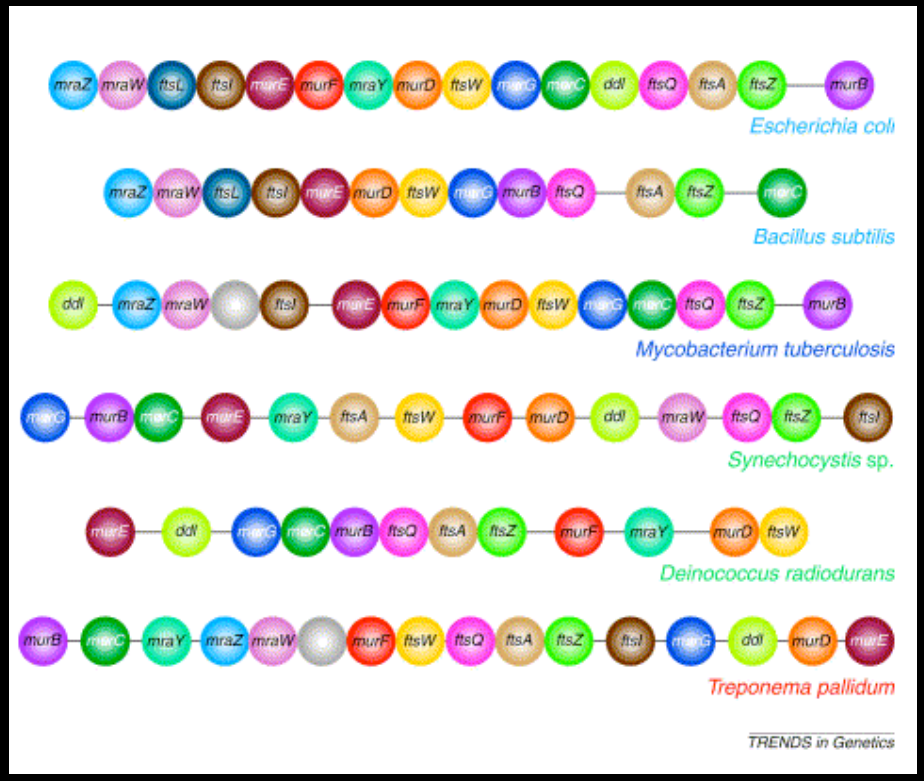


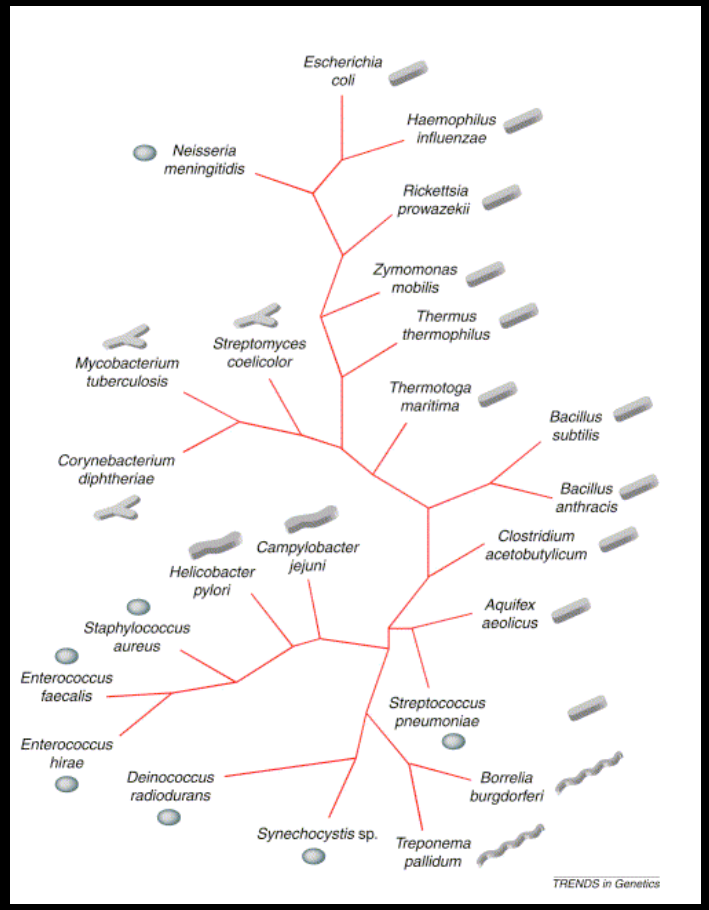
From Carroll, S. B. (1995) Nature 376, 479

INSTITUT PASTEUR

GENE ORDER AND CELL SHAPE

The *mur-fts* cluster

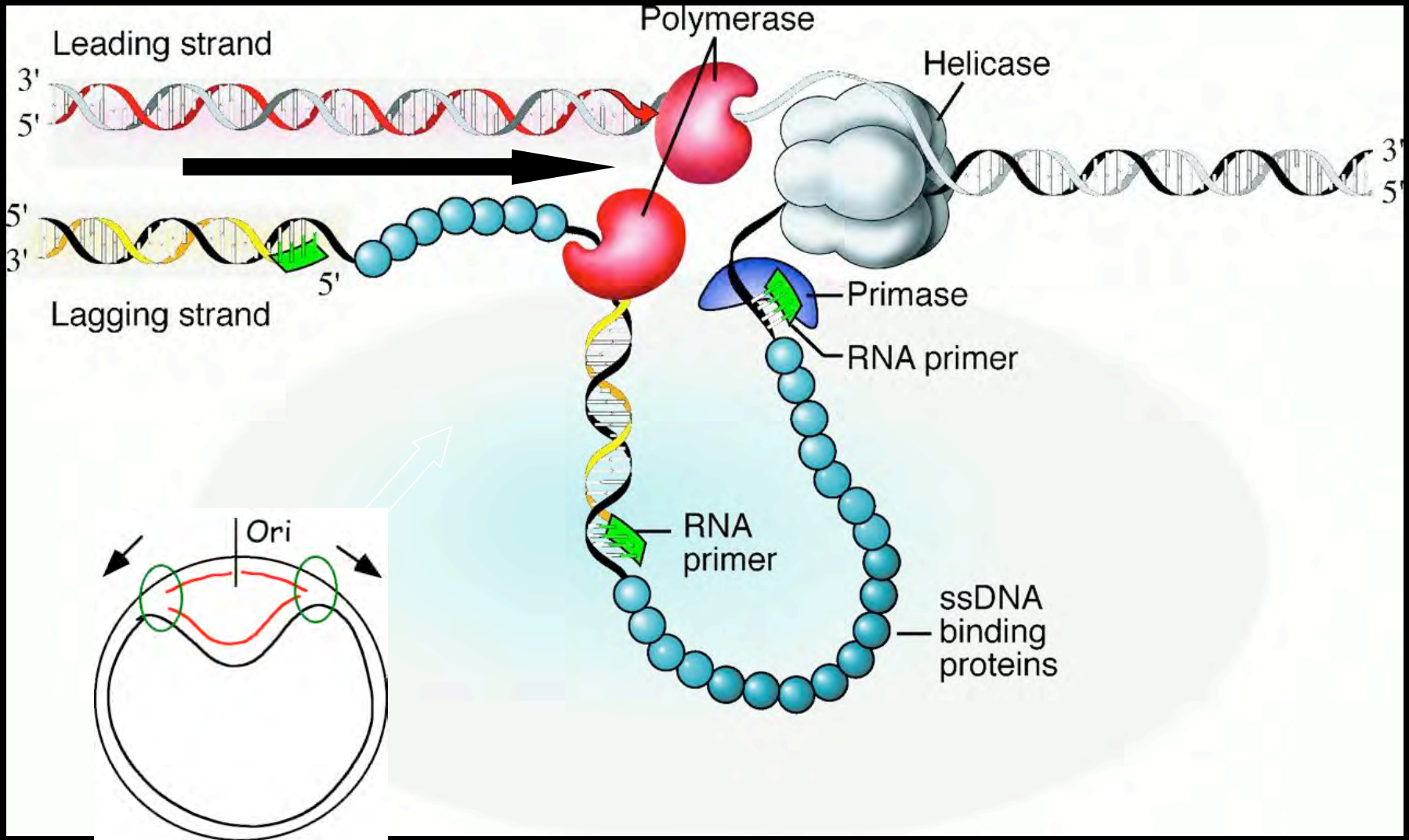




Tamames J, Gonzalez-Moreno M, Mingorance J, Valencia A, Vicente M
Bringing gene order into bacterial shape
Trends in Genetics (2001) **17**: 124-126

INSTITUT PASTEUR

DISSYMMETRY OF REPLICATION



INSTITUT PASTEUR

OPEN QUESTIONS

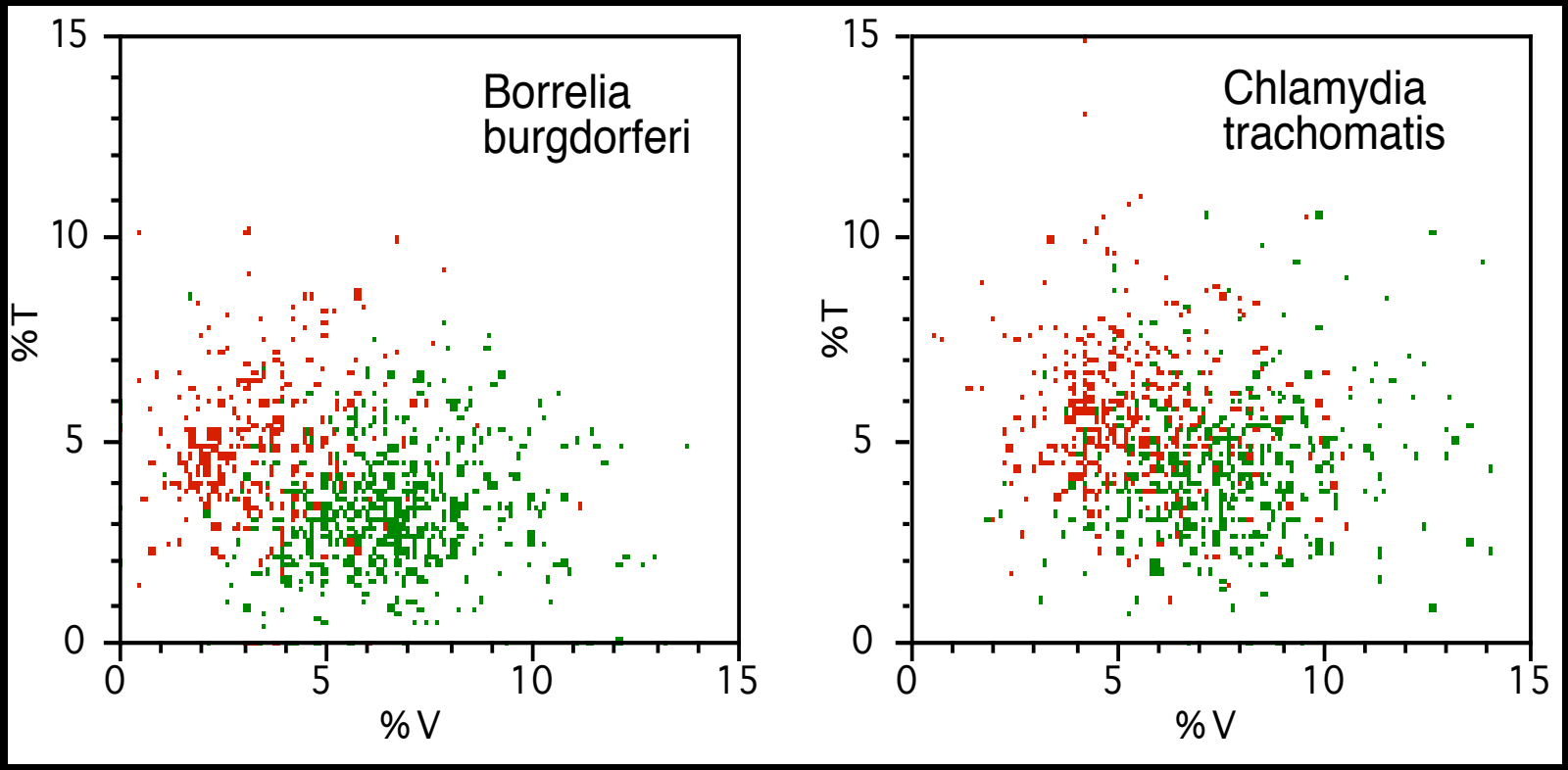
- The constraints resulting from the difference in leading and lagging replication are extremely important
- They may result in non random distribution of genes if some functions are associated to properties of proteins (amino acid frequency, formation of complexes, etc)

Frustration 2: conflict between protein structure and DNA structure, partially solved by the choice of the transcription strand.

INSTITUT PASTEUR

TO LEAD HAS A COST: BIAS VISIBLE IN PROTEINS…

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



INSTITUT PASTEUR

CONSTRAINTS IN PROTEINS

Laplace-Gauss statistics

Principal Component Analysis uses the centered average and a simple distance (identity); it is the reference method but does not use properly information

Correspondence Analysis belongs to the same family; it uses the χ^2 measure as a distance (Benzécri, 1965) which is very close to a first level of information

Absence of normality (or log-normality)

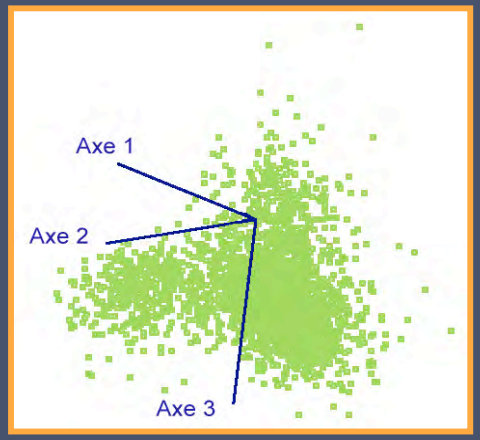
Independent Component Analysis uses the non gaussian character of the values associated to descriptors; it characterizes objects belonging to common independent clusters (the « cocktail party » theorem), (Hérault, 1984)

Further methods need to be developed

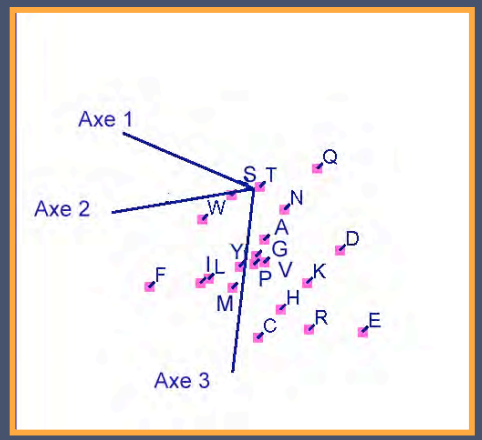
INSTITUT PASTEUR

Correspondence Analysis (CA)

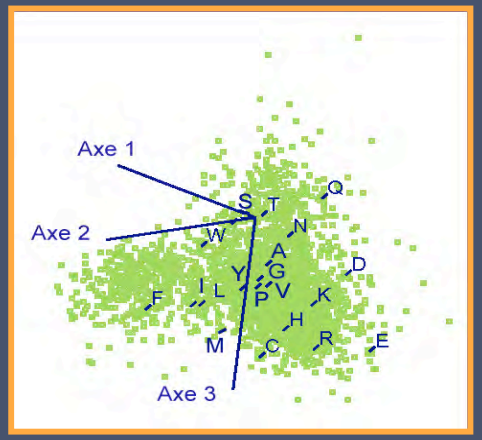
Factorial space of the proteins



Factorial space of the amino acids



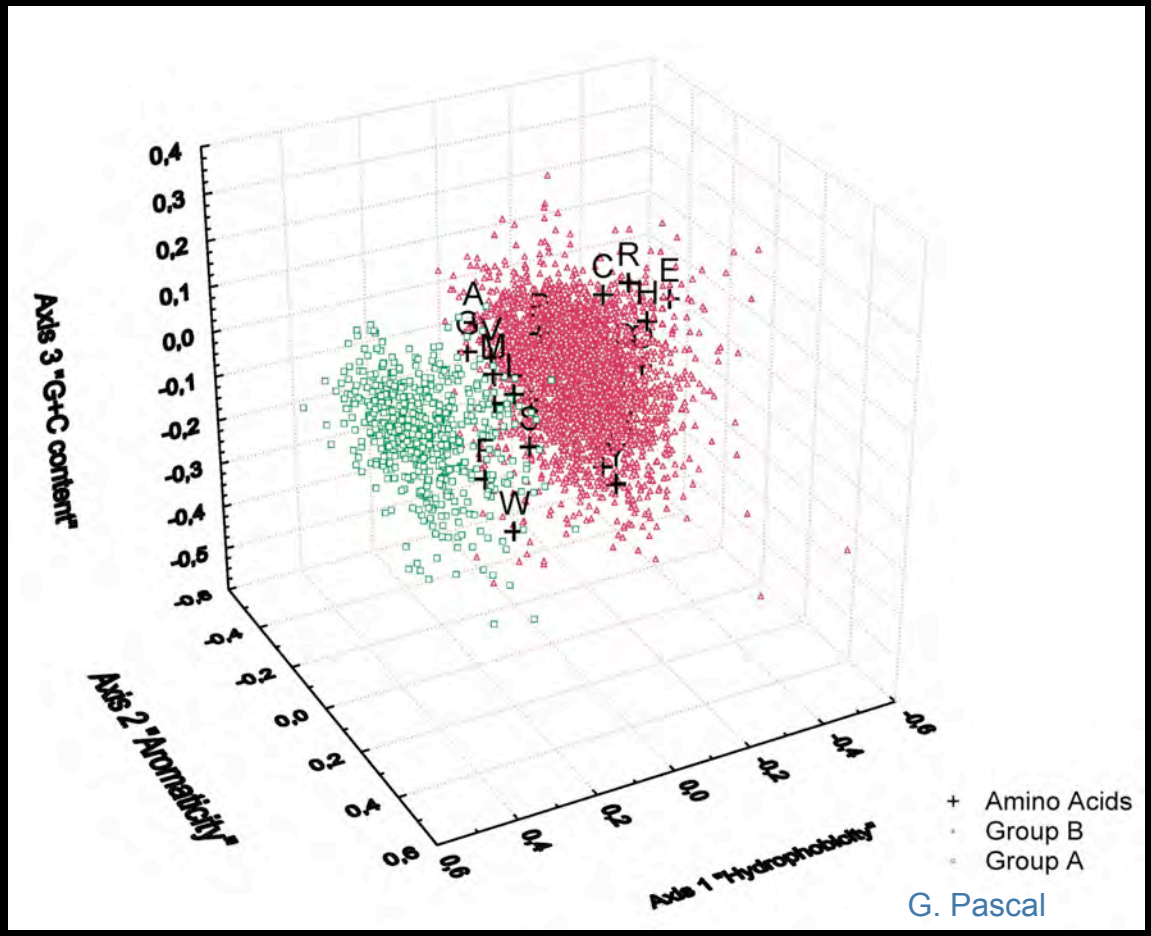
Superimposition of both spaces (clouds)



INSTITUT PASTEUR

BIAS IN AMINO ACID DISTRIBUTION

Neighborhood: distribution of aminoacids in the proteome



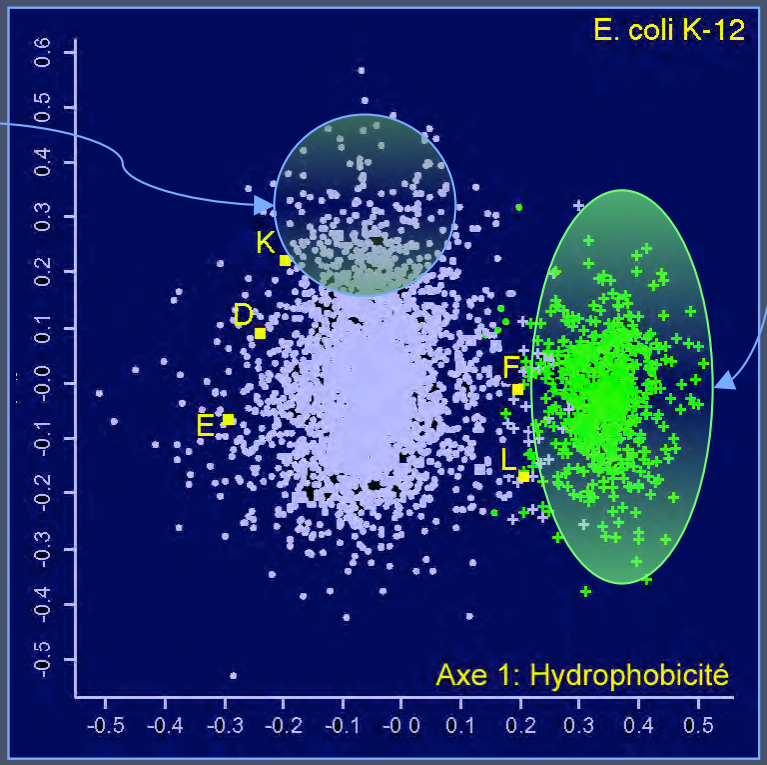
G. Pascal

INSTITUT PASTEUR

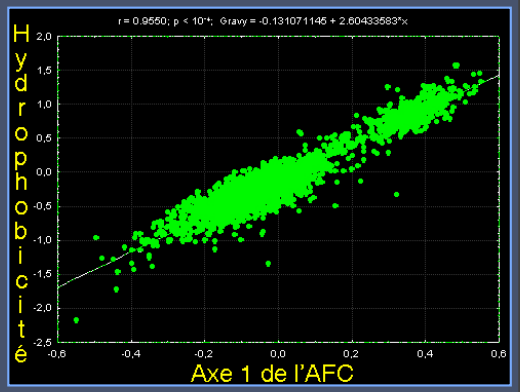
BIASES IN HYDROPHOBICITY OF PROTEINS

A strong bias opposing charged residues to hydrophobic residues

Proteins of the outer membrane: OmpT. OmpL…



Proteins of the inner membrane: LacY. SecE…



INSTITUT PASTEUR

UNIVERSAL BIASES IN PROTEIN AMINO ACID COMPOSITION

- **First axis:** separates Integral Inner Membrane Proteins (IIMP) from the rest; driven by opposition between charged and large hydrophobic residues

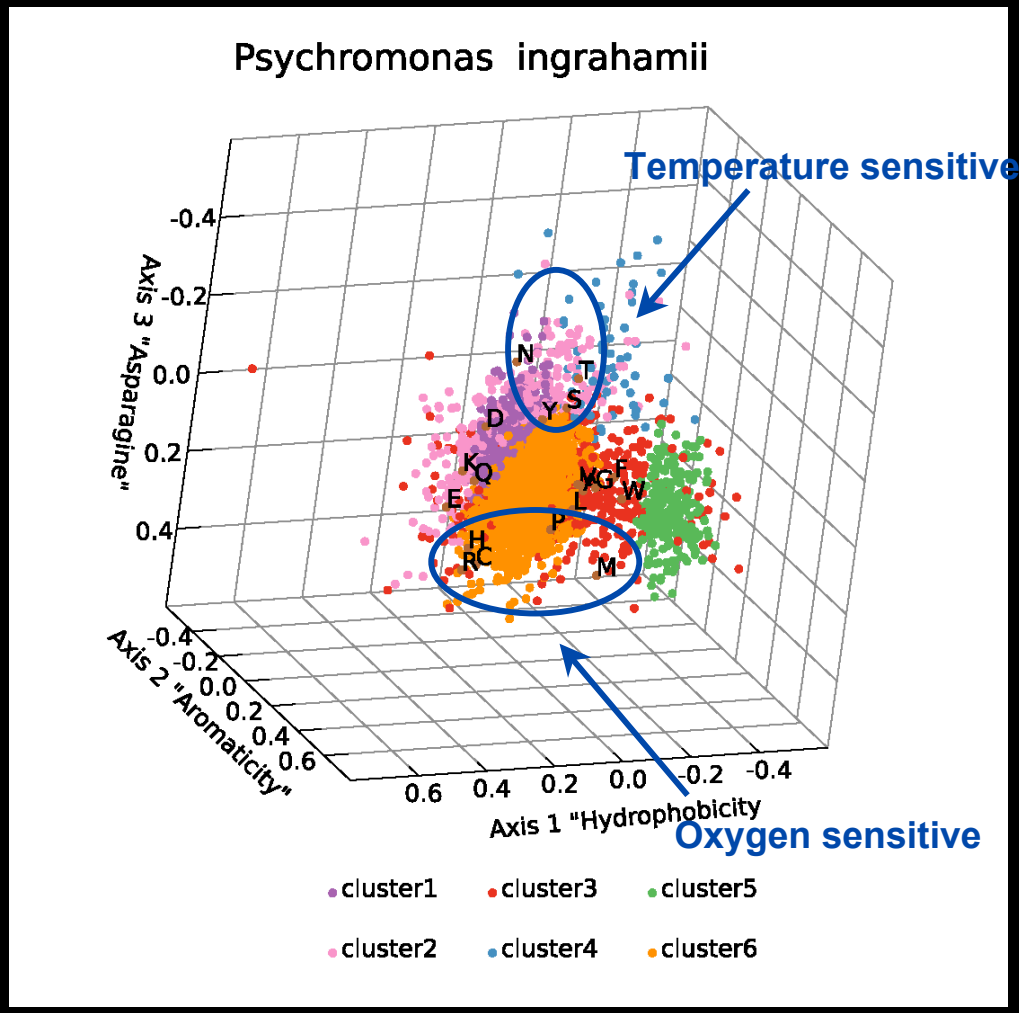
- **Second axis:** separates proteins by their content in aromatic amino acids; highly enriched in orphan proteins

- **Third axis:** separates proteins according to an opposition driven by the G+C content of the *first* codon base

INSTITUT PASTEUR

BIAS IN AMINO ACID DISTRIBUTION

Distribution of amino acids in the proteome of *Psychromonas ingrahamii*



Riley M, Staley JT, Danchin A, Wang T, Brettin TS, Hauser LJ, Land ML, Thompson LS. Genomics of an extreme psychrophile, *Psychromonas ingrahamii*. BMC Genomics. 2008 May 6;9(1):210

INSTITUT PASTEUR

TEMPERATURE-DEPENDENT BIASES IN PROTEIN AMINO ACID COMPOSITION

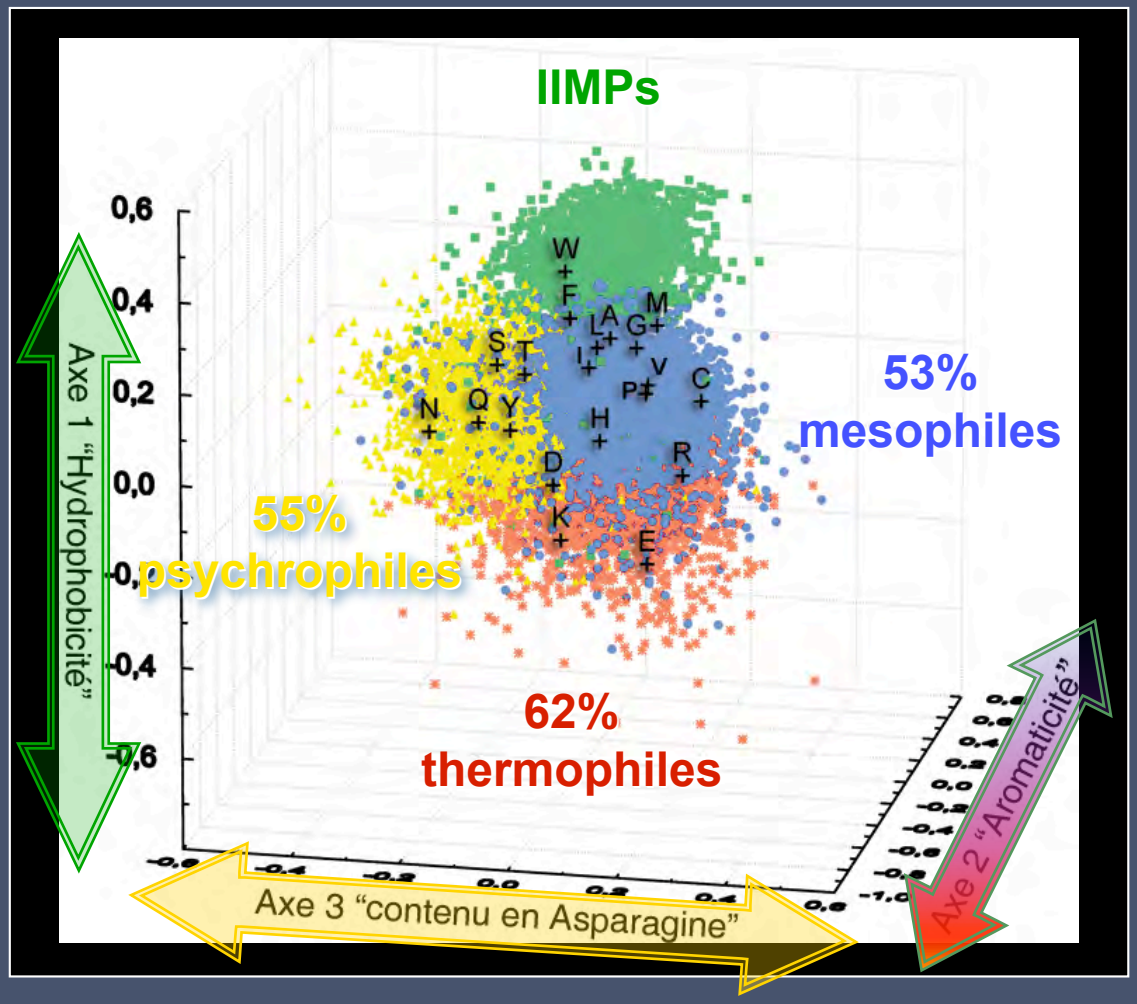
- The general trend of amino acid composition bias is to avoid some aminoacids at higher temperatures (associated to aging processes)
- Mesophilic bacteria belong to at least two different classes (in a 5-clusters analysis)
- Biases are always dominated by the IIMP clustering

INSTITUT PASTEUR

COMPARATIVE PROTEOMICS

A specific asparagine bias in psychrophiles

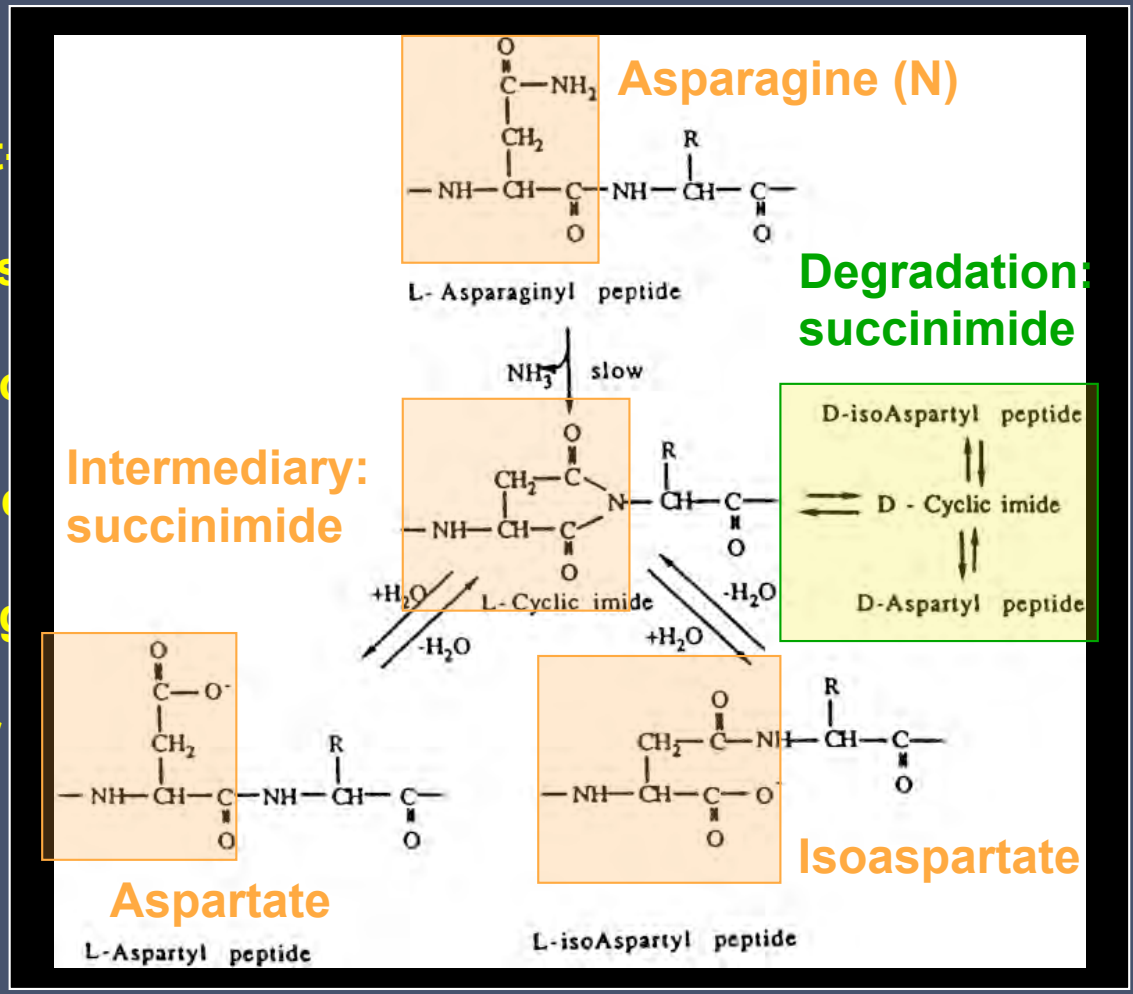
- Motility
- Cell wall, outer membrane
- Transport (TonB), secretion
- Adaptation to stress
- Metabolism of DNA and RNA



A CHEMICAL ANECDOTE

Asparagine deamidates: a major contribution to protein aging

- Main post
- Reaction s
- Spontaneo
- Affects th
- Role in reg
- Signal for



REVISITING THE INEVITABILITY OF AGING

- Proteins age, sometimes very fast (e.g. ribosomal protein S11 from *E. coli*, within minutes at 37°C)
- As a consequence, it is always an aged cell (or multicellular organism) that gives birth to a young one
- This implies that in the process of forming a progeny, there is creation of information
- We need to identify the genes acting in this process of accumulating information

INSTITUT PASTEUR

REVISITING INFORMATION

Improvement of metabolism can be conceptually tolerated as creation of information is reversible (Landauer, 1961; Bennett, 1982, 1988)

Open question: « making room » is needed to accomodate innovation; how is it obtained? Can we identify in genomes the genes coding for the functions required to put this process in action?

INSTITUT PASTEUR

FUNCTION OR STRUCTURE?

We need to look for ubiquitous functions, while we have only direct access to structures (or even worse, sequences)

INSTITUT PASTEUR

A WEALTH OF UNKNOWN GENES

3721 ongoing projects, 665 completed, mostly from microbes (546 Bacteria, more or less correctly annotated)

204,981,131,711 nucleotides at International Nucleotide Sequence Database Collaboration (INSDC)

Microbes make 50% of the Earth protoplasm

The first discovery of genomics (1991):

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

INSTITUT PASTEUR

ACQUISITIVE EVOLUTION MASKS FUNCTIONAL PERSISTENCE

Variation / Selection / Amplification

Stabilisation

Evolution

creates

Function

captures (recruits)

Structure

codes

Sequence

INSTITUT PASTEUR

FROM FUNCTIONAL UBIQUITY TO GENE PERSISTENCE

Functional ubiquity does not imply structural ubiquity
Efficient objects tends to persist through generations:

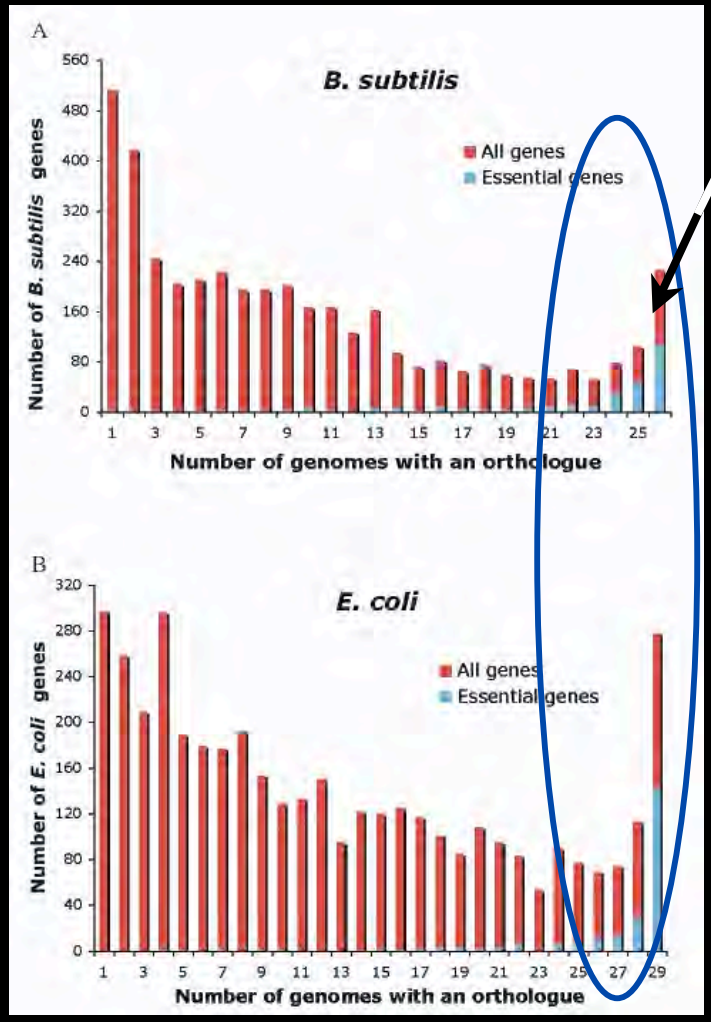
- Looking for « persistence » permits identification of (most) ubiquitous functions
- Is « ubiquitous » a synonym of « essential »?

« Laboratory essential » genes are located in the DNA leading strand

400-500 genes persist in 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: TWO GENE CATEGORIES



Persistent genes

Essential genes and ….

Stress, maintenance and repair

Fighting weathering (anti-aging)

PERSISTENT GENES ARE CLUSTERED TOGETHER

Persistent genes are functionally defined. They are located in the DNA replication leading strand

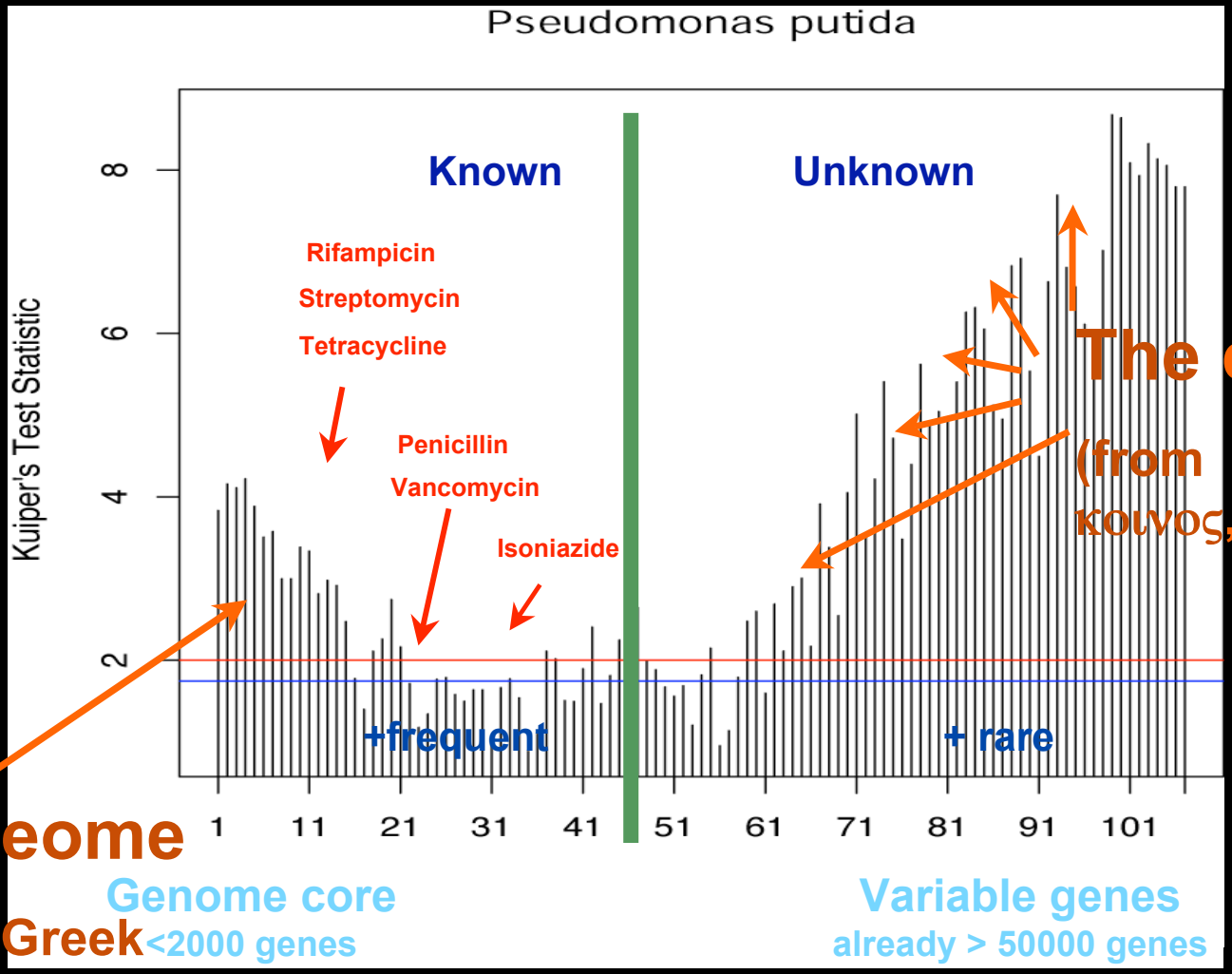
Depending of their tendency to remain clustered in genomes (in > 250 bacteria with genome length > 1,500) they form three families that reflect a scenario of the origin of life

This group of core genes form the paleome (from παλαιος, ancient)

INSTITUT PASTEUR

CONSERVATION OF GENE CLUSTERING

Clustering frequency



The paleome

(from the Greek παλαιος, ancient)

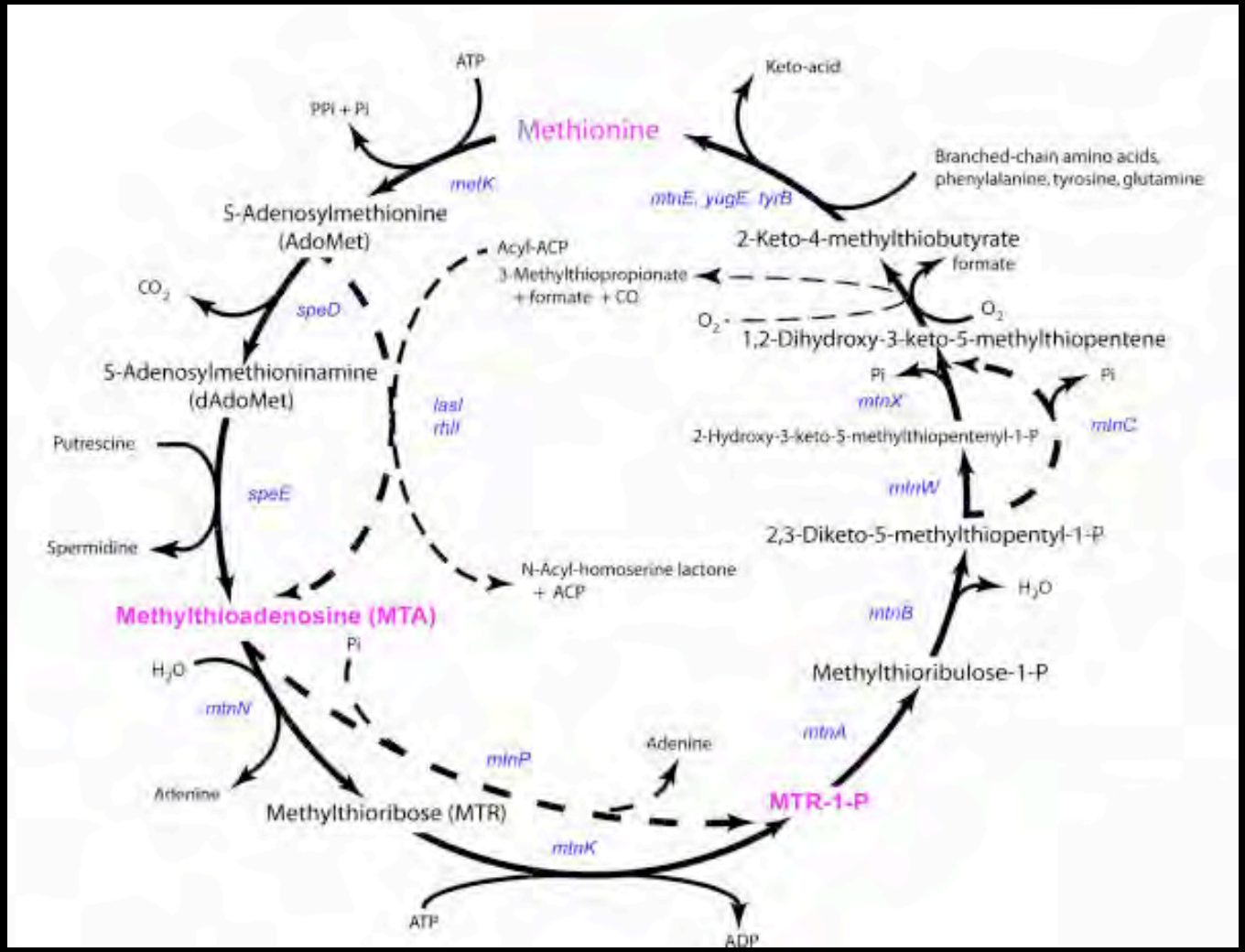
The cenome

(from the Greek κουνος, common)

Frequency in genomes

INSTITUT PASTEUR

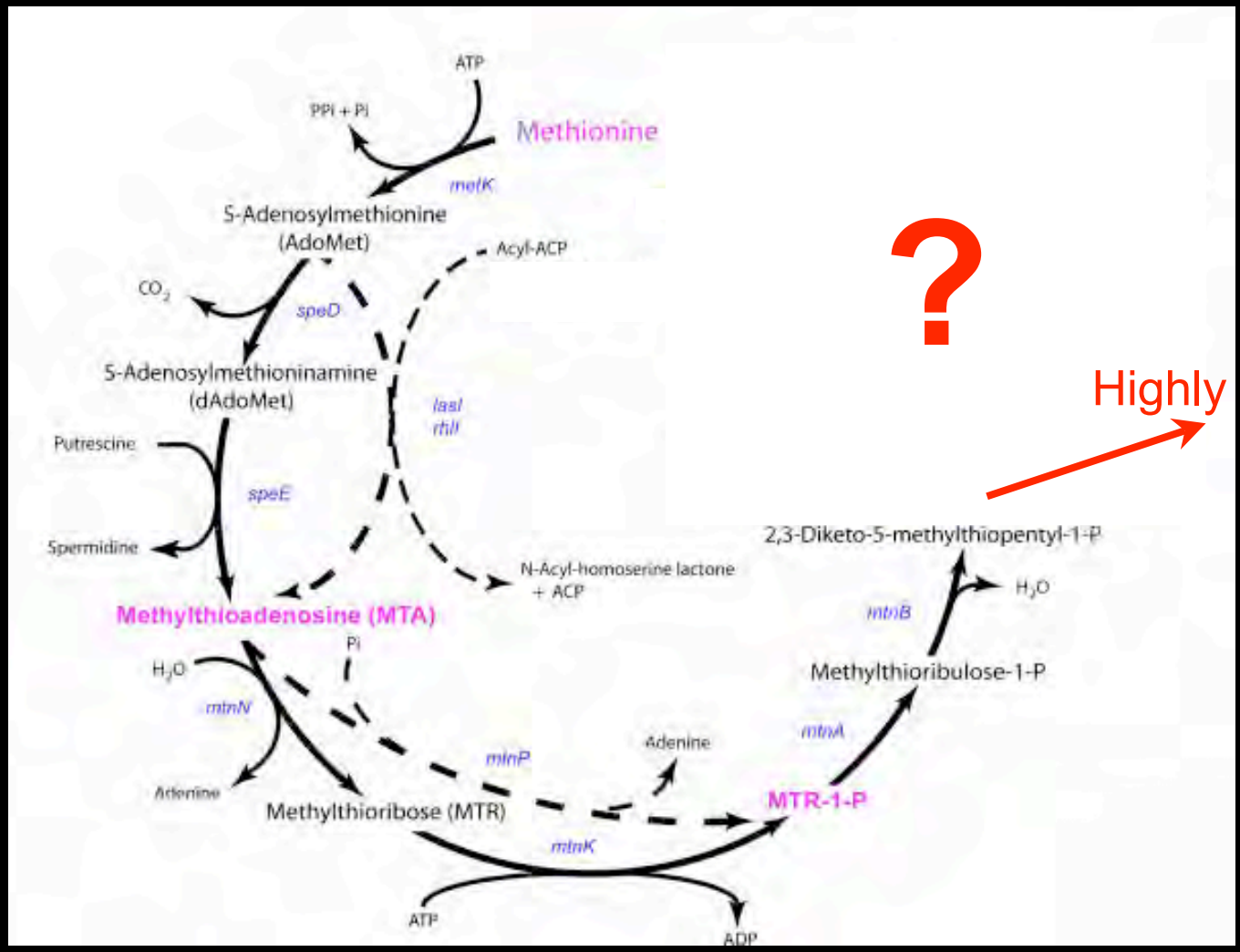
A PATHWAY IN THE TWILIGHT ZONE



Sekowska A, Dénervaud V, Ashida H, Michoud K, Haas D, Yokota A, Danchin A.

Bacterial variations on the methionine salvage pathway. BMC Microbiol. 2004 4:9.

METHYLTHIOADENOSINE DEGRADATION



INSTITUT PASTEUR

PERSISTENT GENES CONNECTIVITY

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

To understand its meaning, let us explore a scenario of the origin of life.

A MINERAL SCENARIO FOR THE ORIGIN OF LIFE

- The surface of charged solids (e.g. pyrite (Fe-S)) selects and compartimentalizes charged molecules; this first step forms some aminoacids, the main coenzymes, fatty acids and ribonucleotides; polymerisation with elimination of water molecules increases entropy

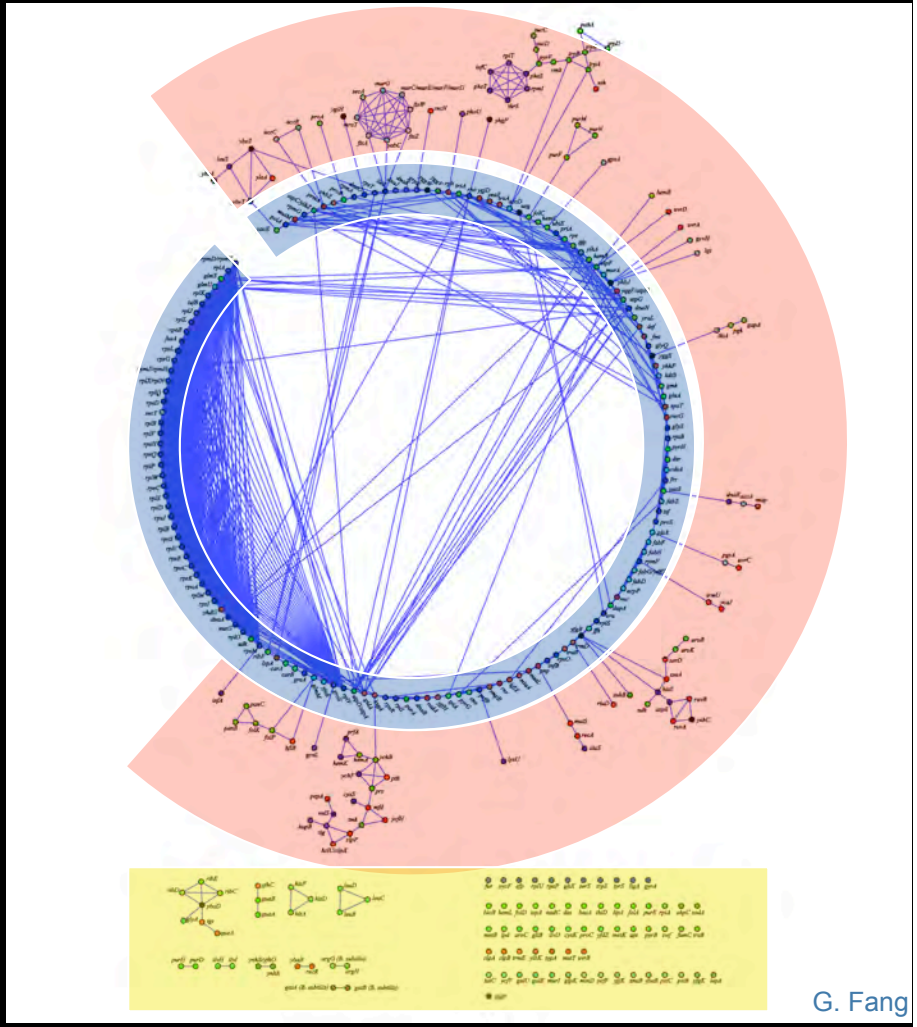
- Compartmentalised metabolism creates surface substitutes via polymerisation of ribonucleotides in the presence of peptides (the RNA world); RNAs discover the complementarity law, and the genetic code is invented

- 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 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

INSTITUT PASTEUR

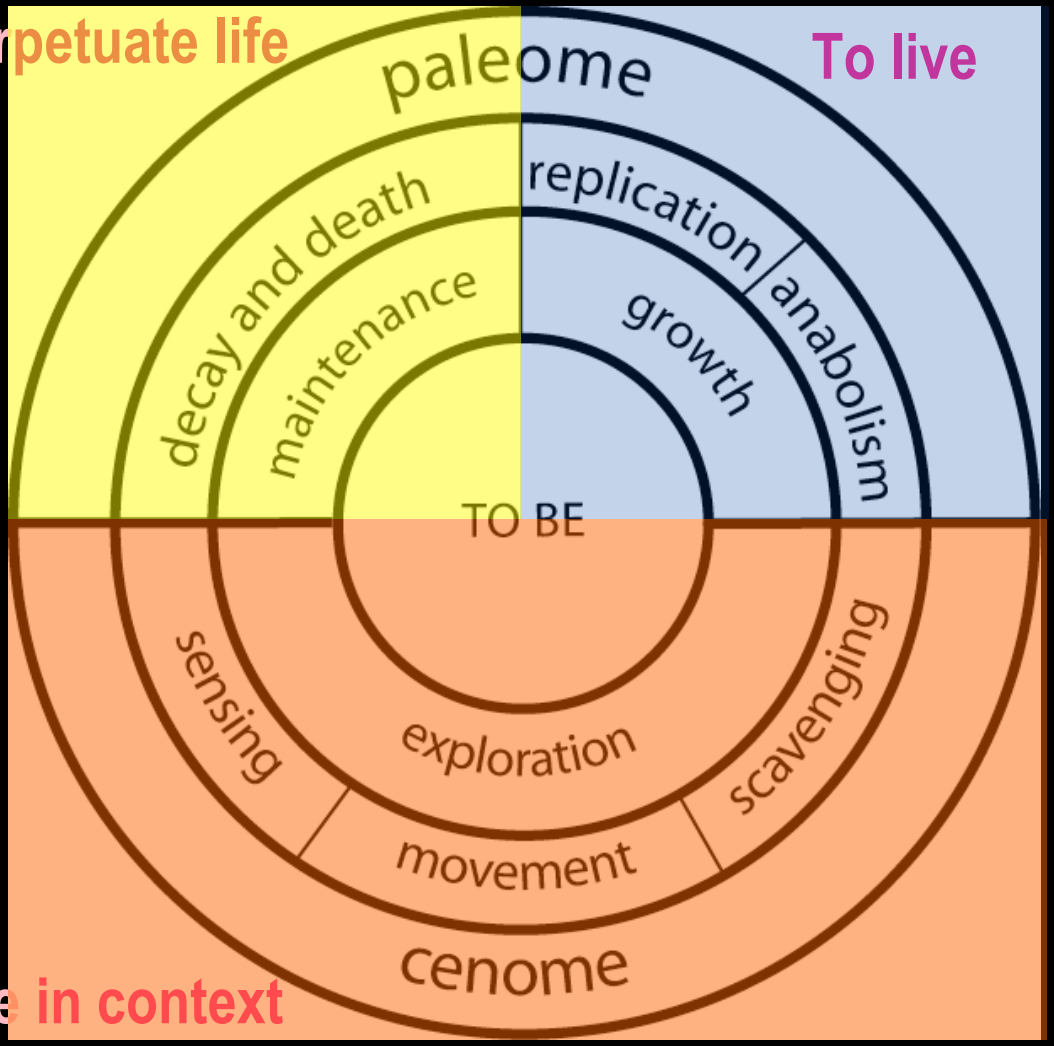
METABOLISM AND REPLICATION

This scenario emphasizes the separation between metabolism and replication, the latter being a secondary invention of prebiotic systems:

Building blocks => nucleotides => tRNA => ribosome => DNA

THE PALEOME AND 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



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

INSTITUT PASTEUR

A SPLIT PALEOME

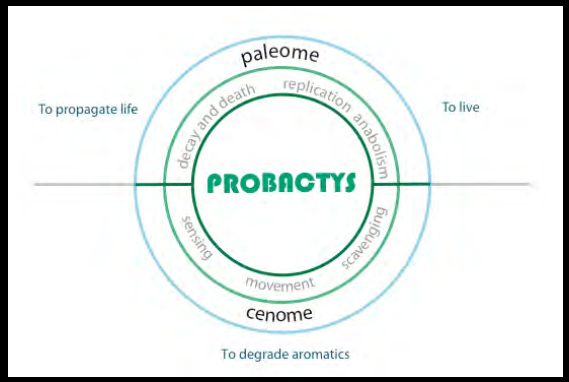
- **Paleome 1 (essential genes)**
  - **Constructor**: DNA specifies proteins which form the machine that constructs the cell (reproduction)
  - **Replicator**: DNA specifies proteins that replicate DNA (replication)
- **Paleome 2 (persistent non essential genes)**

  Perennisation of life, requires identification of functional from non functional objects

INSTITUT PASTEUR

THANK YOU

谢谢



FONDATION SCIENTIFIQUE FOURMENTIN-GUILBERT

POUR LE RAYONNEMENT DE LA BIOLOGIE

BIOSAPIENS NETWORK

BlastSets

EPG

Genetics of Bacterial Genomes

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