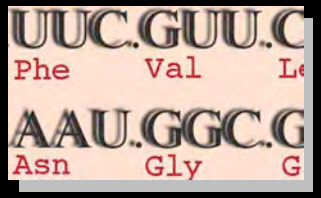
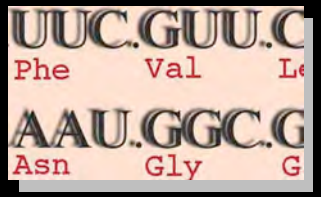


INSTITUT PASTEUR

Symplectic biology: the cell as a living computer

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

adanchin@pasteur.fr



Authors

in silico

- Gang Fang
- Etienne Larsabal
- Géraldine Pascal
- Eduardo Rocha
- Ivan Moszer
- Claudine Médigue

in vivo / in vitro

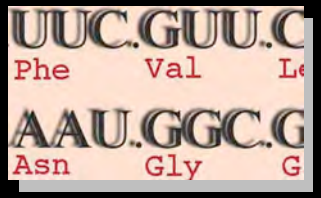
- Agnieszka Sekowska
- Anne Marie Gilles
- Octavian Barzu

Collective

- Stanislas Noria

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

adanchin@pasteur.fr

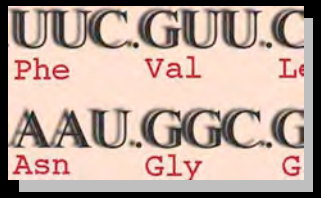


Background

INSTITUT PASTEUR

- **Physics:** *matter, energy, time*
- **Biology:** Physics + *information, coding, control...*
- **Arithmetics:** *strings of whole numbers, recursivity, coding…*
- **Computing:** Arithmetics + *program* + *machine...*

adanchin@pasteur.fr



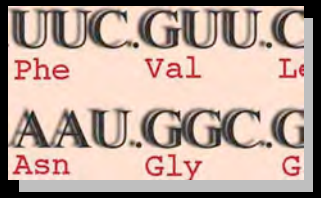
Information Transfer

INSTITUT PASTEUR

As is the case for building up a machine, one needs a book of recipe to build up a cell

This asks for changing the text of the recipe into something concrete: this transfers « information »

In a cell, information transfer is managed by the genetic program



What is Life?

INSTITUT PASTEUR

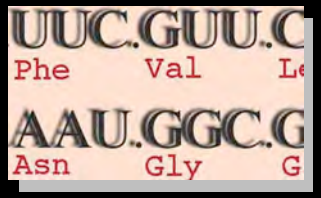
Three processes are needed for Life:

- **Information transfer** (Living Computers?) => **the goal of genomics is to decipher the blueprint of the "read-only" memory of the machine**

Driving force for a coupling between the genome structure and the structure of the cell:

- **Metabolism** (Internal organisation)
- **Compartmentalisation** (General structure)

adanchin@pasteur.fr



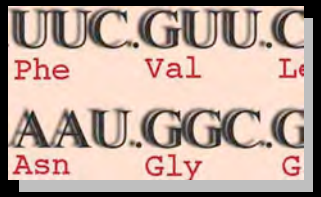
What is computing?

INSTITUT PASTEUR

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



Cells as computers

INSTITUT PASTEUR

Genomics rest on an alphabetic metaphor, that of a text written with a four-letter alphabet, acting as a program

Conjecture: do cells behave as computers?

Genetic engineering

Viruses

Horizontal gene transfer

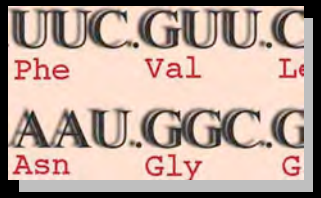
Cloning animal cells

all point to separation between

Machine

Data + Program

adanchin@pasteur.fr



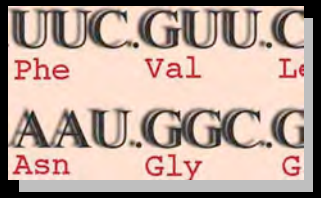
Is there a map of the cell in the chromosome?

INSTITUT PASTEUR

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 (J. von Neumann)

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

adanchin@pasteur.fr



Genome organisation

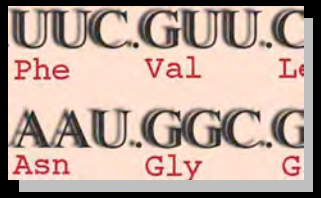
INSTITUT PASTEUR

Is the gene order random in the chromosomes?

At first sight, despite different DNA management processes not much is conserved, and genes transferred from other organisms are distributed throughout genomes

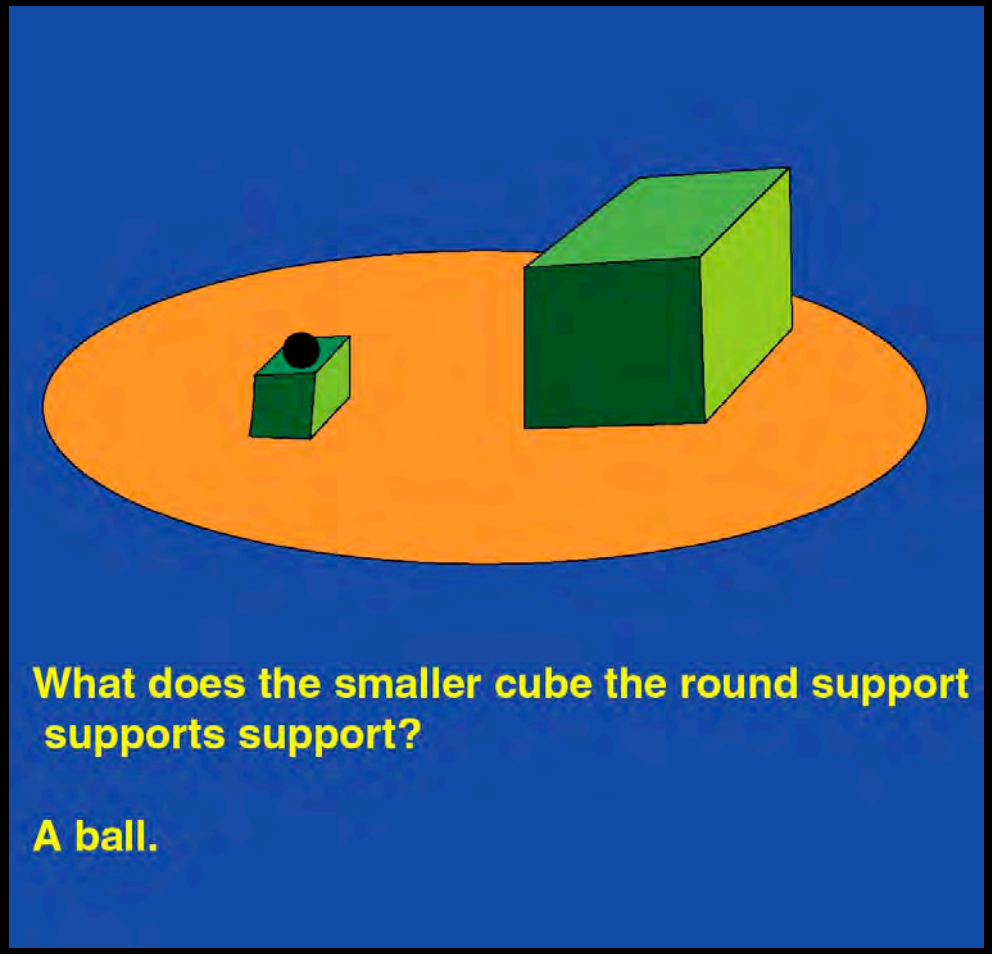
However, groups of genes such as operons or pathogenicity islands tend to cluster in specific places, and they code for proteins with common functions

First question: how are generated and where are located repeats in the genome sequence?



Caveat: Repeats are meaningful

INSTITUT PASTEUR

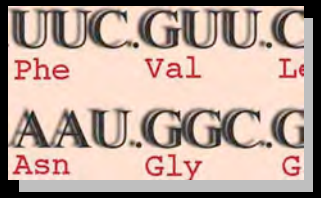


Remember also:

This clock has a minute minute hand

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

adanchin@pasteur.fr



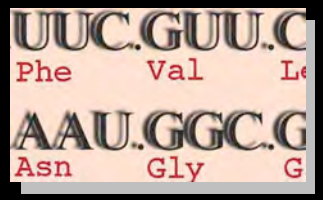
Repeats in bacteria

INSTITUT PASTEUR

- Abcissa: first occurrence of the repeat
- Ordinate: second position of the repeat

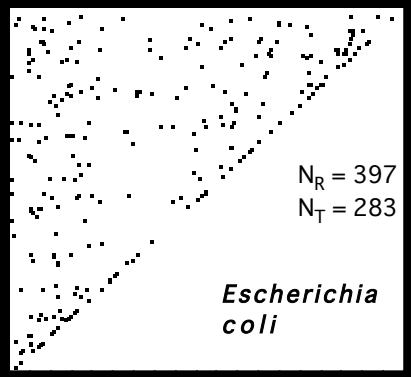
- Diagonal: repeats are located near to each other

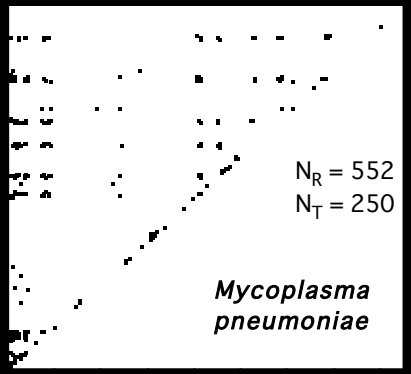
adanchin@pasteur.fr

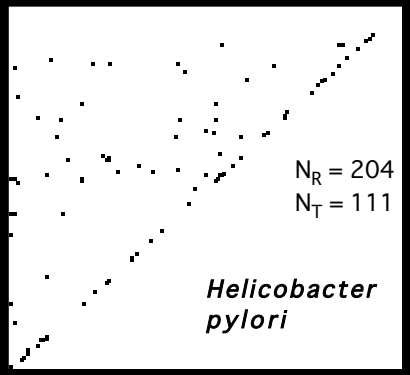


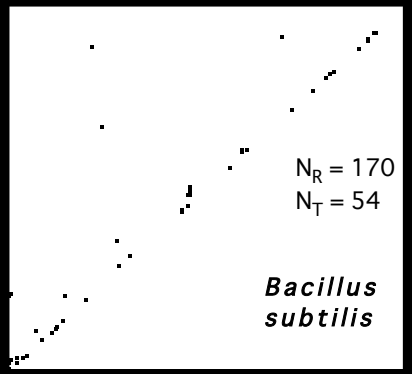
DNA management: Repeats in genomes

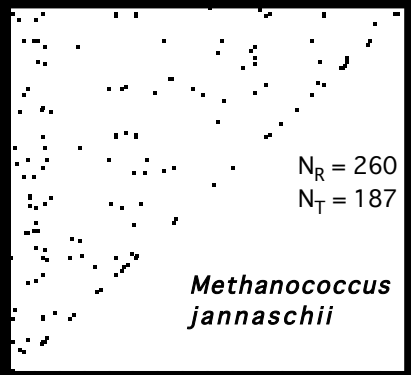
INSTITUT PASTEUR

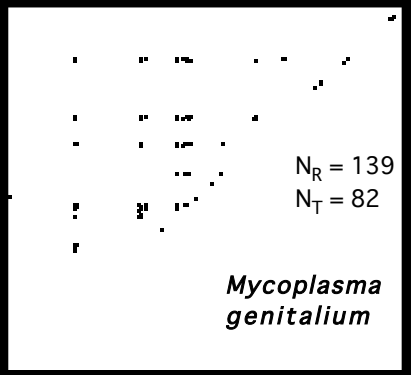


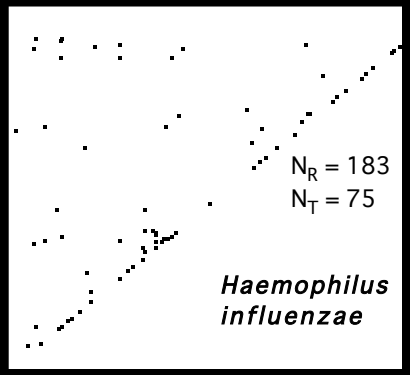


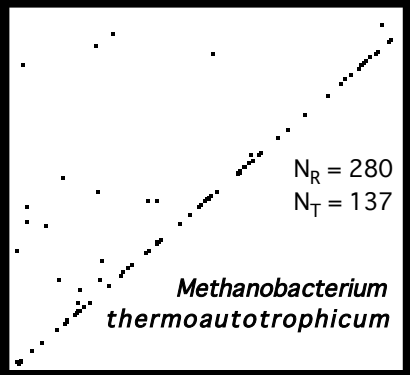








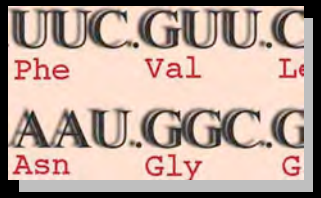




E. Rocha, A. Viari & A. Danchin Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in *Bacillus subtilis* and other competent prokaryotes. Mol. Biol. Evol. (1999) **16**: 1219-1230

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

adanchin@pasteur.fr

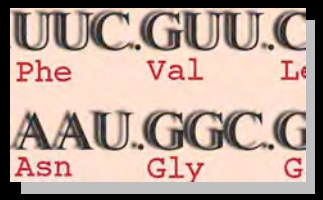


Genome organisation

INSTITUT PASTEUR

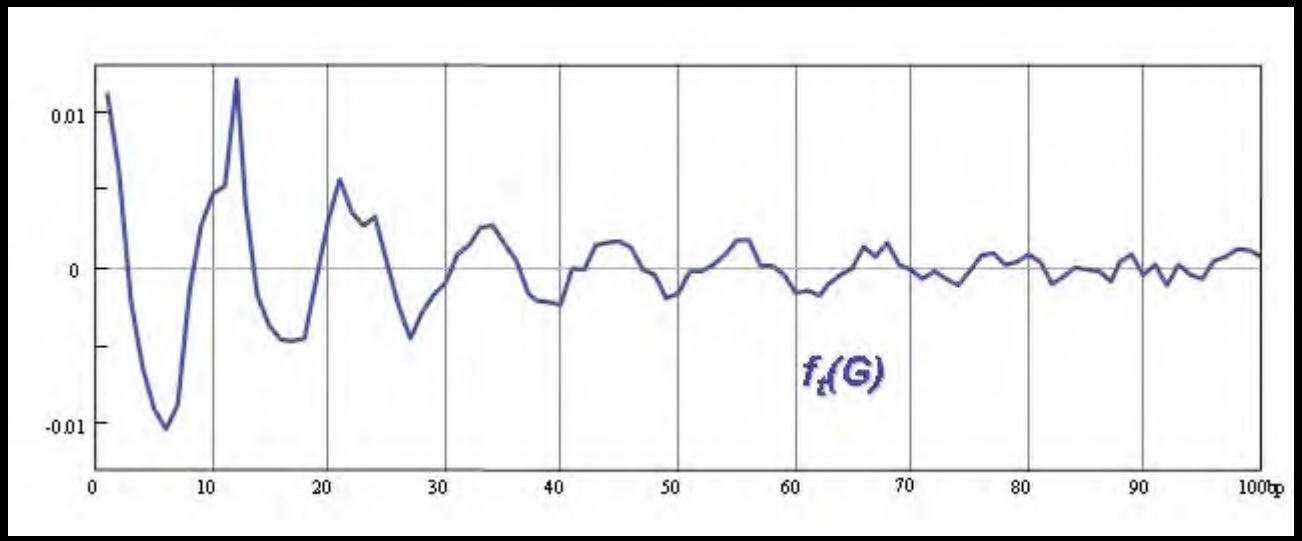
The genome organisation is so rigid that the overall result of selection pressure on DNA is visible in the genome text, which is full of « flexible patterns of class A »

adanchin@pasteur.fr



The period 10-11.5

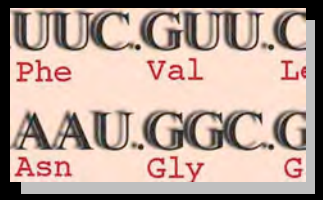
INSTITUT PASTEUR



The genome of *Helicobacter pylori* displays a period of 11 over regions spanning 60 nucleotides

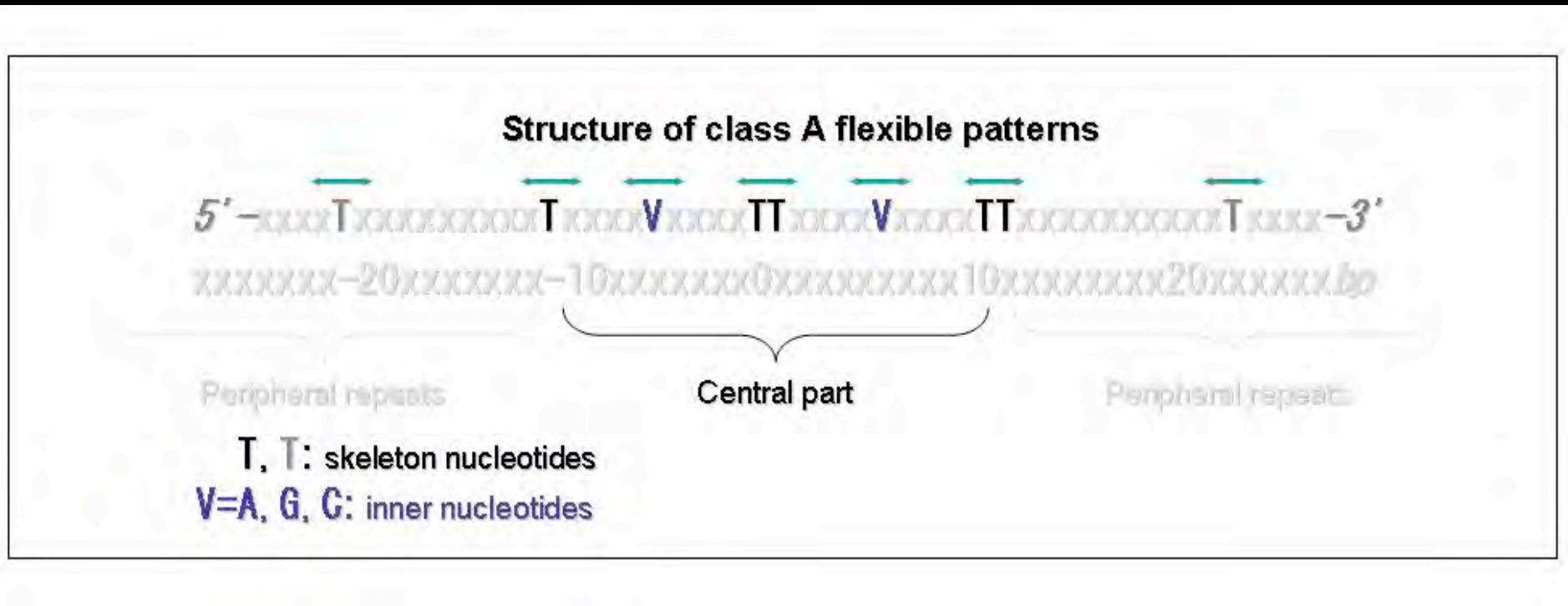
http://www.pasteur.fr/recherche/unites/REG

adanchin@pasteur.fr

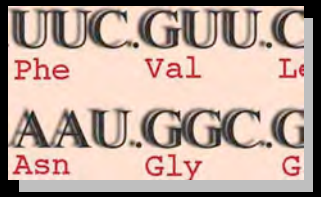


Class A flexible patterns

INSTITUT PASTEUR



The period 10-11.5 is explained by the presence of omnipresent patterns the class A flexible patterns



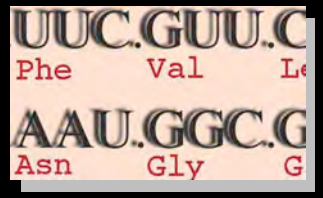
Class A flexible patterns are ubiquitous

INSTITUT PASTEUR

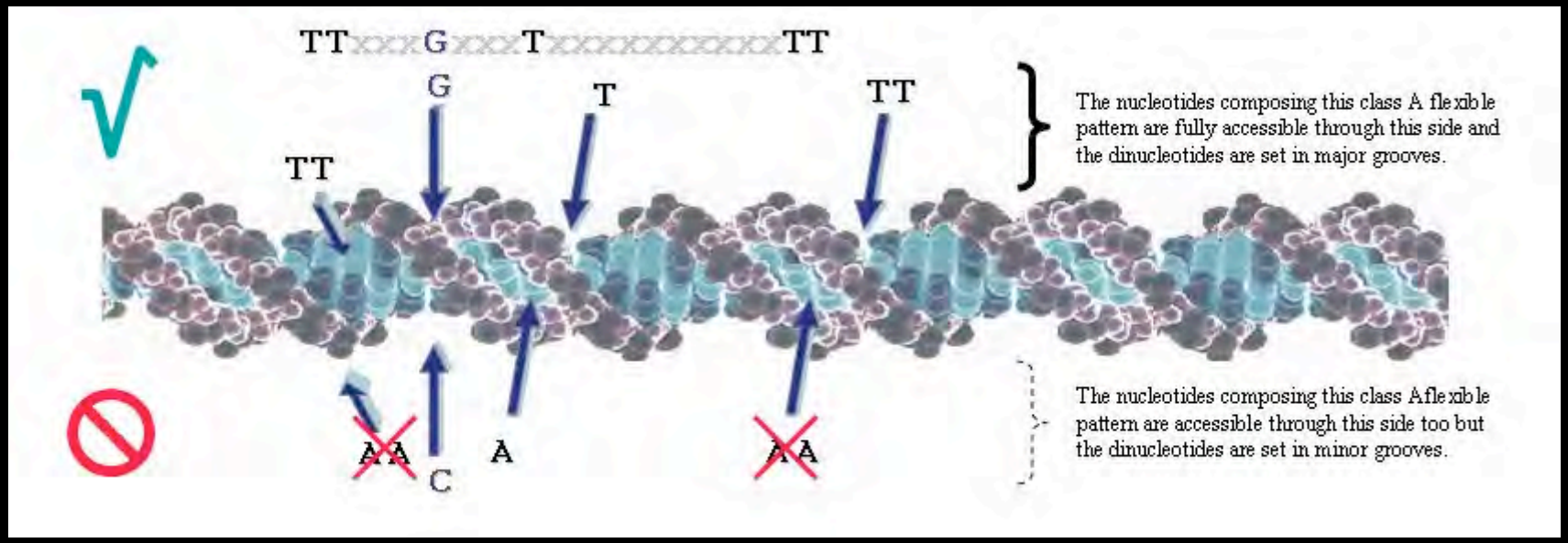
1-	A	T	A	TT	A	T	A	All kindoms
2-			G	TT	C	T		Proteobacteria
4-		T	AG	TT		T		Archaea
5'-								-3'

The period 10-11.5 is explained by the presence of omnipresent patterns the class A flexible patterns

adanchin@pasteur.fr

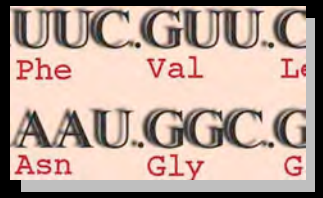


Class A flexible patterns



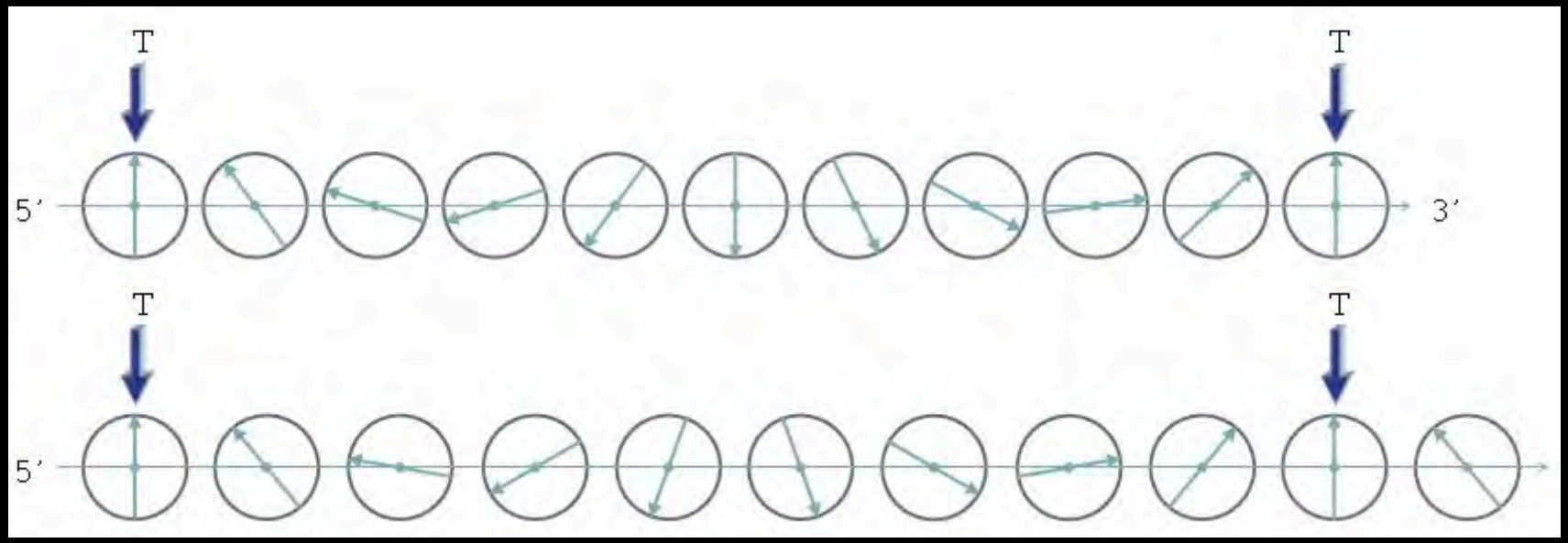
The period 10-11.5 is explained by the presence of omnipresent patterns the class A flexible patterns

adanchin@pasteur.fr



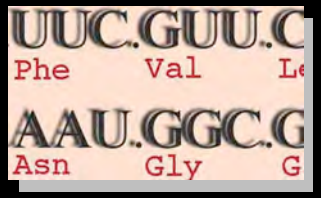
A universal rule: class A flexible patterns

INSTITUT PASTEUR



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

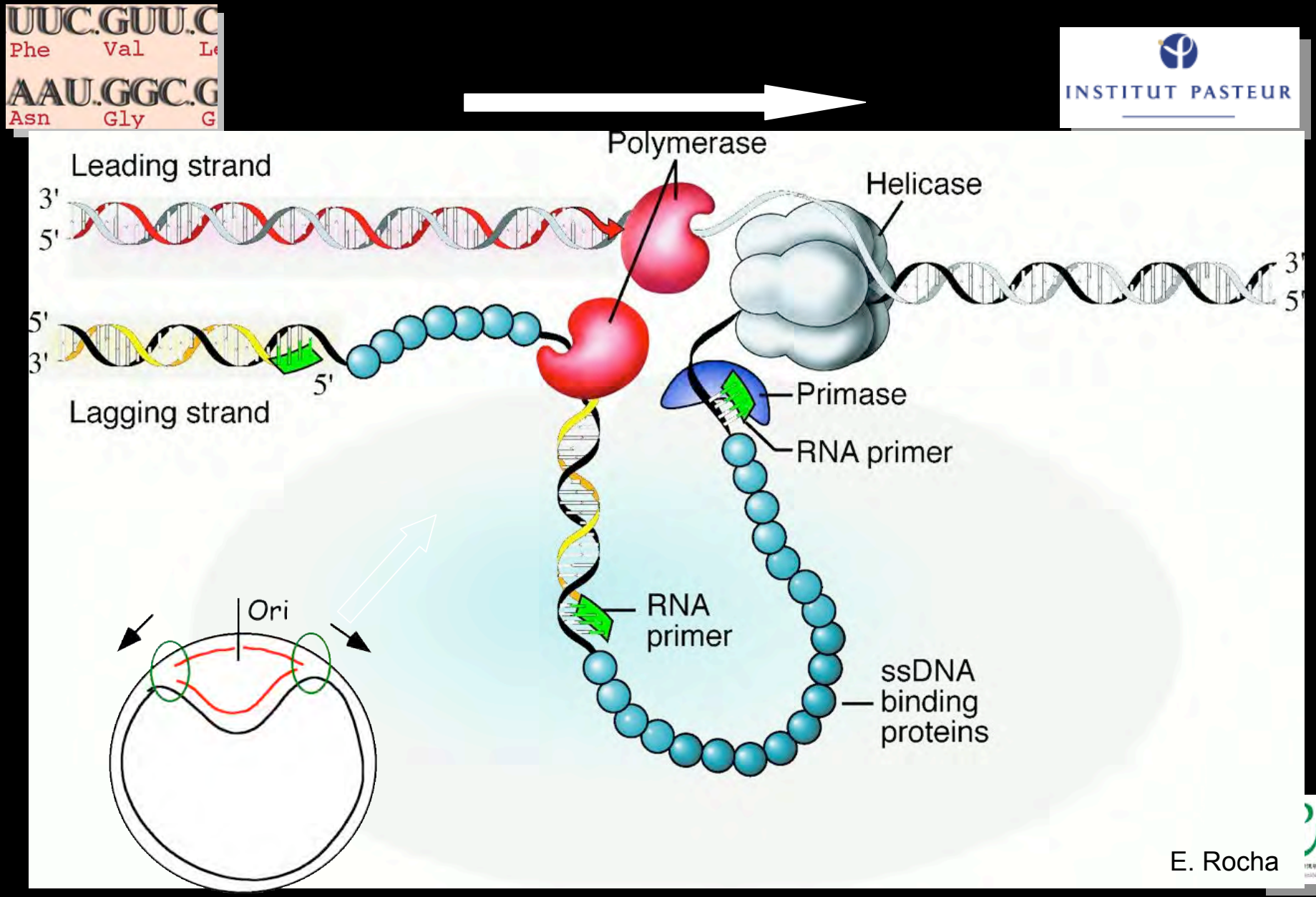
http://www.pasteur.fr/recherche/unites/REG
adanchin@pasteur.fr

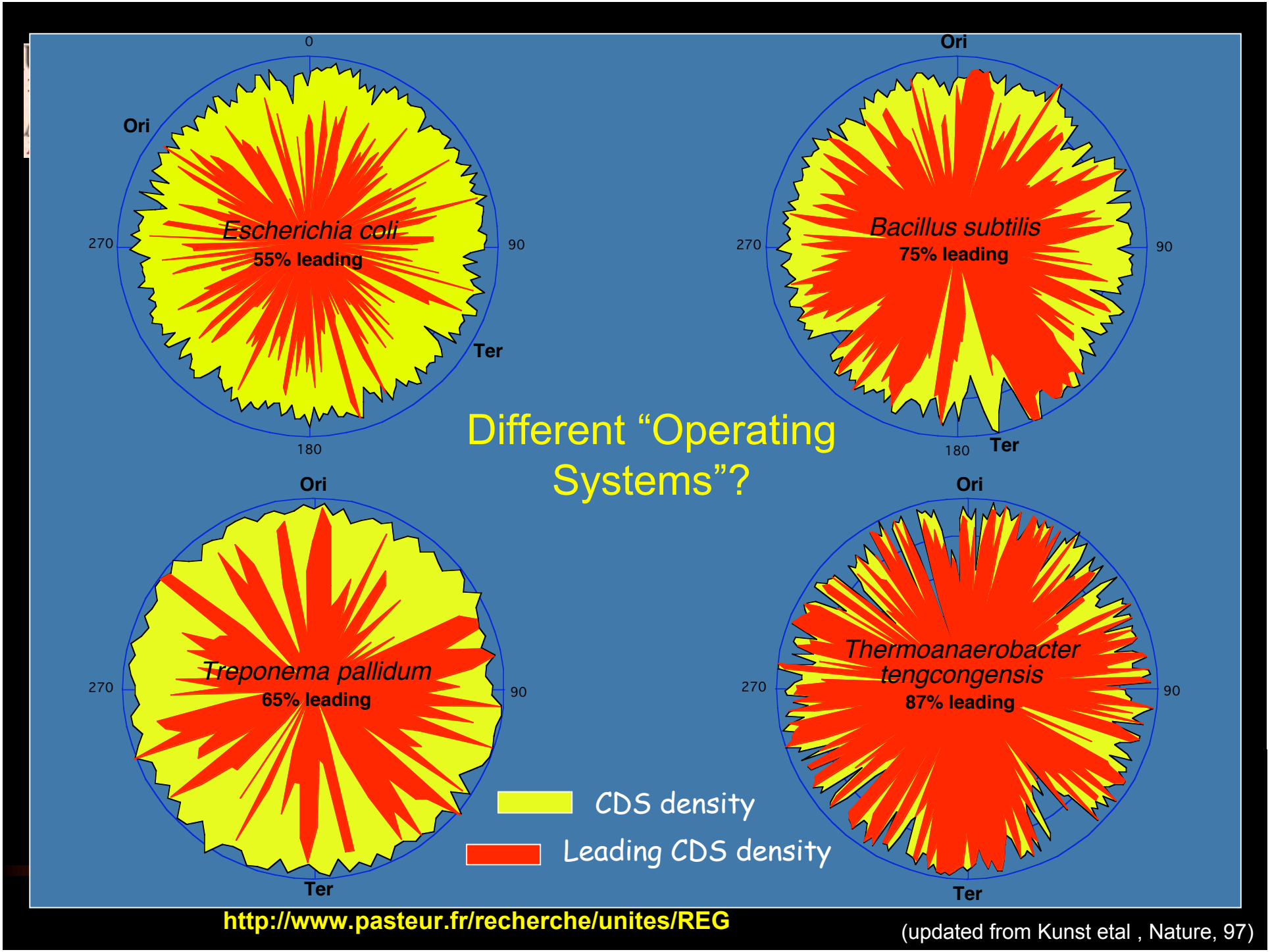


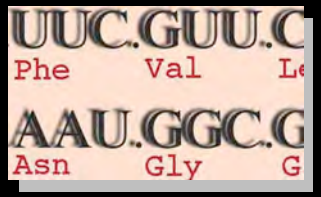
Genome organisation

INSTITUT PASTEUR

The genome organisation is so rigid that the overall result of selection pressure on DNA is visible in the genome text, where the constraints of replication are visible in the leading and the lagging strand


UUC.GUU.C
Phe Val Le
AAU.GGC.G
Asn Gly G
INSTITUT PASTEUR
Polymerase
Leading strand
3'
5'
Helicase
3'
5'
5'
3'
5'
Lagging strand
Primase
RNA primer
RNA
primer
ssDNA
binding
proteins
Ori
E. Rocha

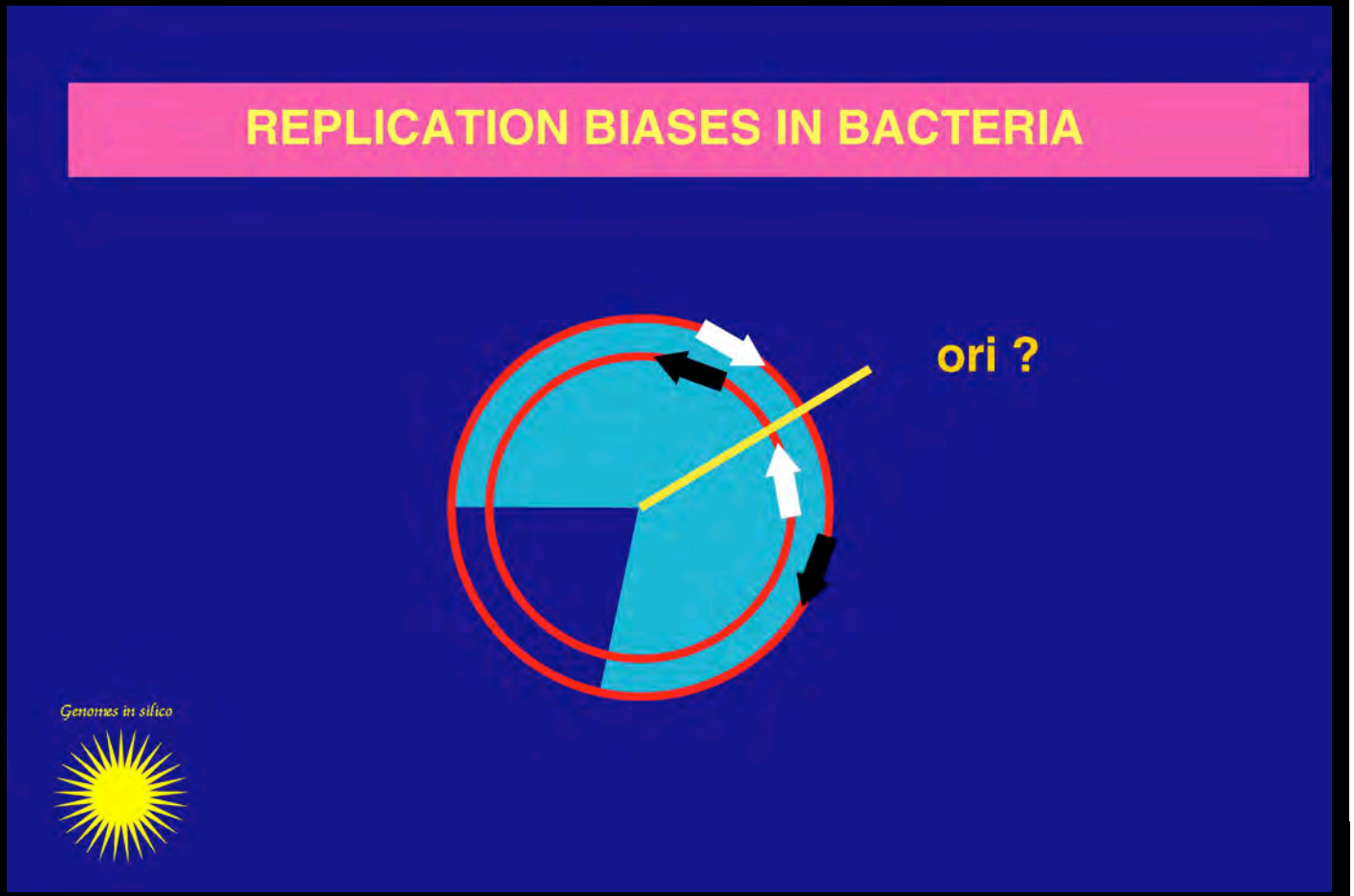

0
Ori
270
90
180
Ter
Escherichia coli
55% leading
Ori
270
90
180
Ter
Bacillus subtilis
75% leading
Different "Operating Systems"?
Ori
270
90
Ter
Treponema pallidum
65% leading
Ori
270
90
Ter
Thermoanaerobacter tengcongensis
87% leading
CDS density
Leading CDS density
http://www.pasteur.fr/recherche/unites/REG
(updated from Kunst etal , Nature, 97)



To lag or to lead...

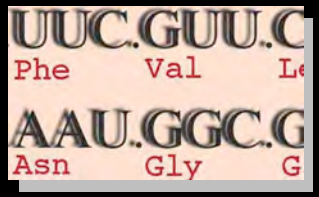
INSTITUT PASTEUR

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.



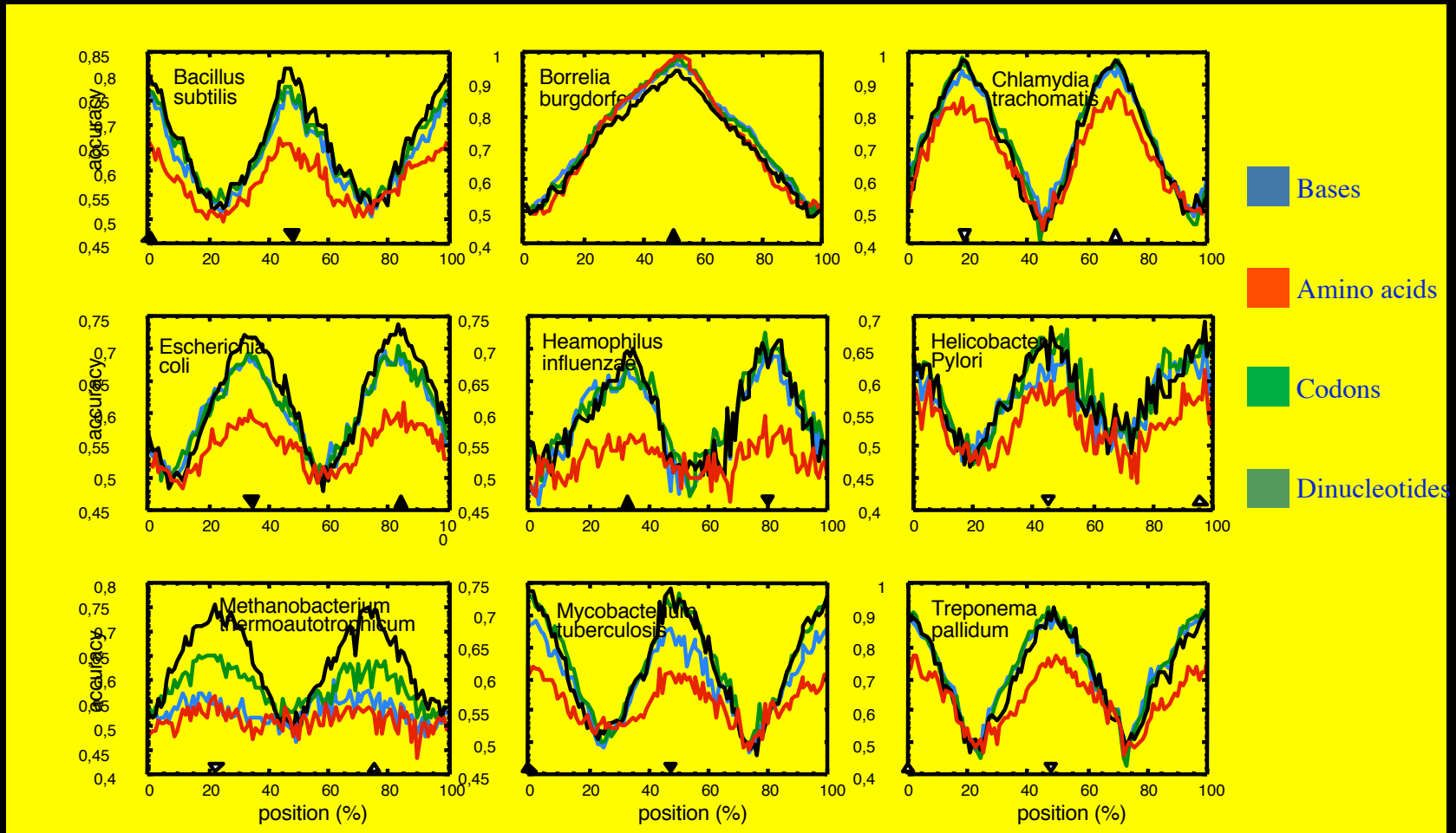
E. Rocha, A. Danchin & A. Viari Universal replication biases in bacteria. Mol. Microbiol. (1999) 32: 11-16

adanchin@pasteur.fr



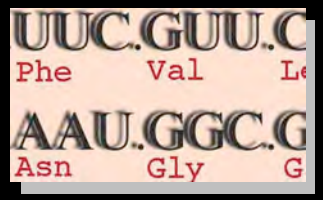
To lag or to lead, that is the question

INSTITUT PASTEUR



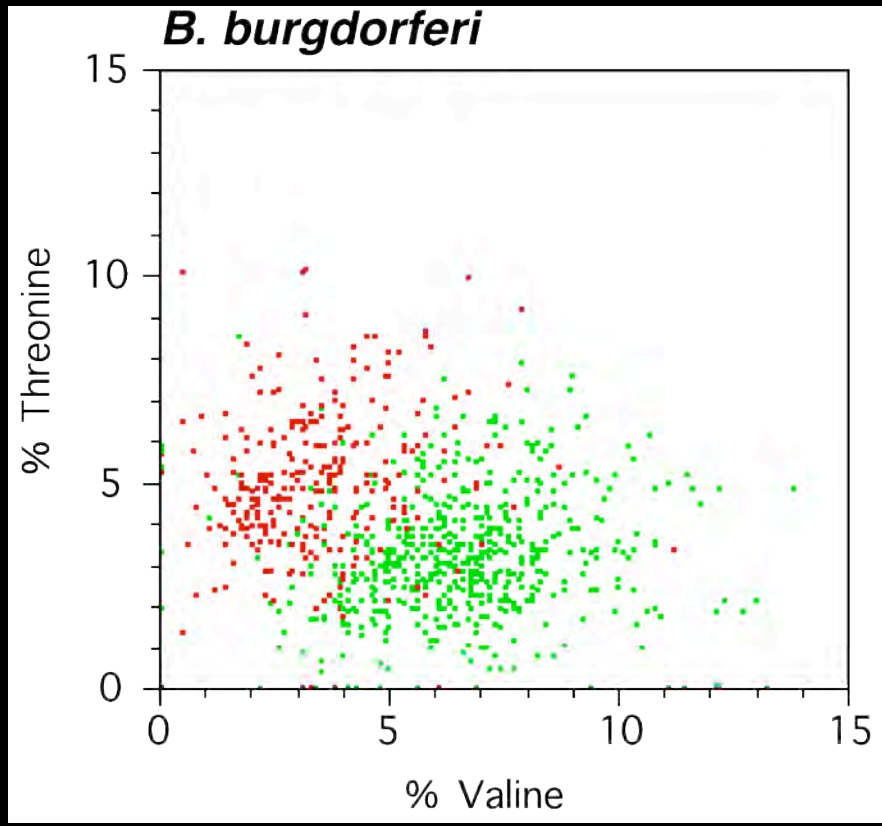
http://www.pasteur.fr/recherche/unites/REG

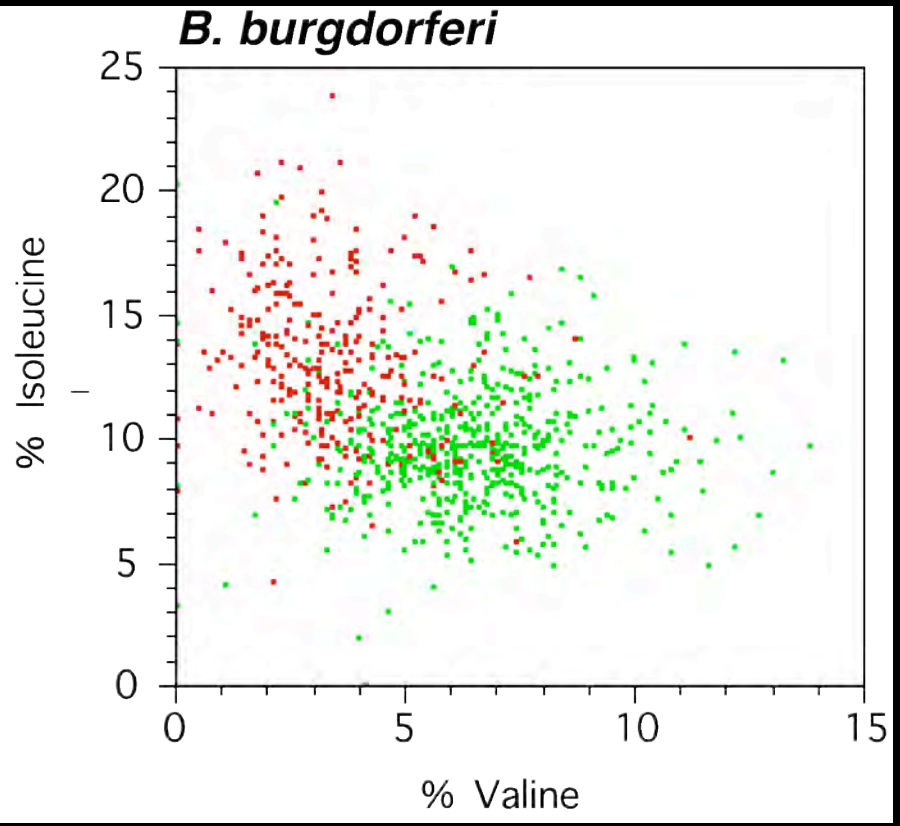
adanchin@pasteur.fr



Visible even in proteins…

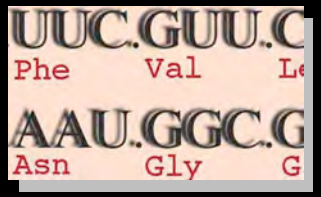
INSTITUT PASTEUR





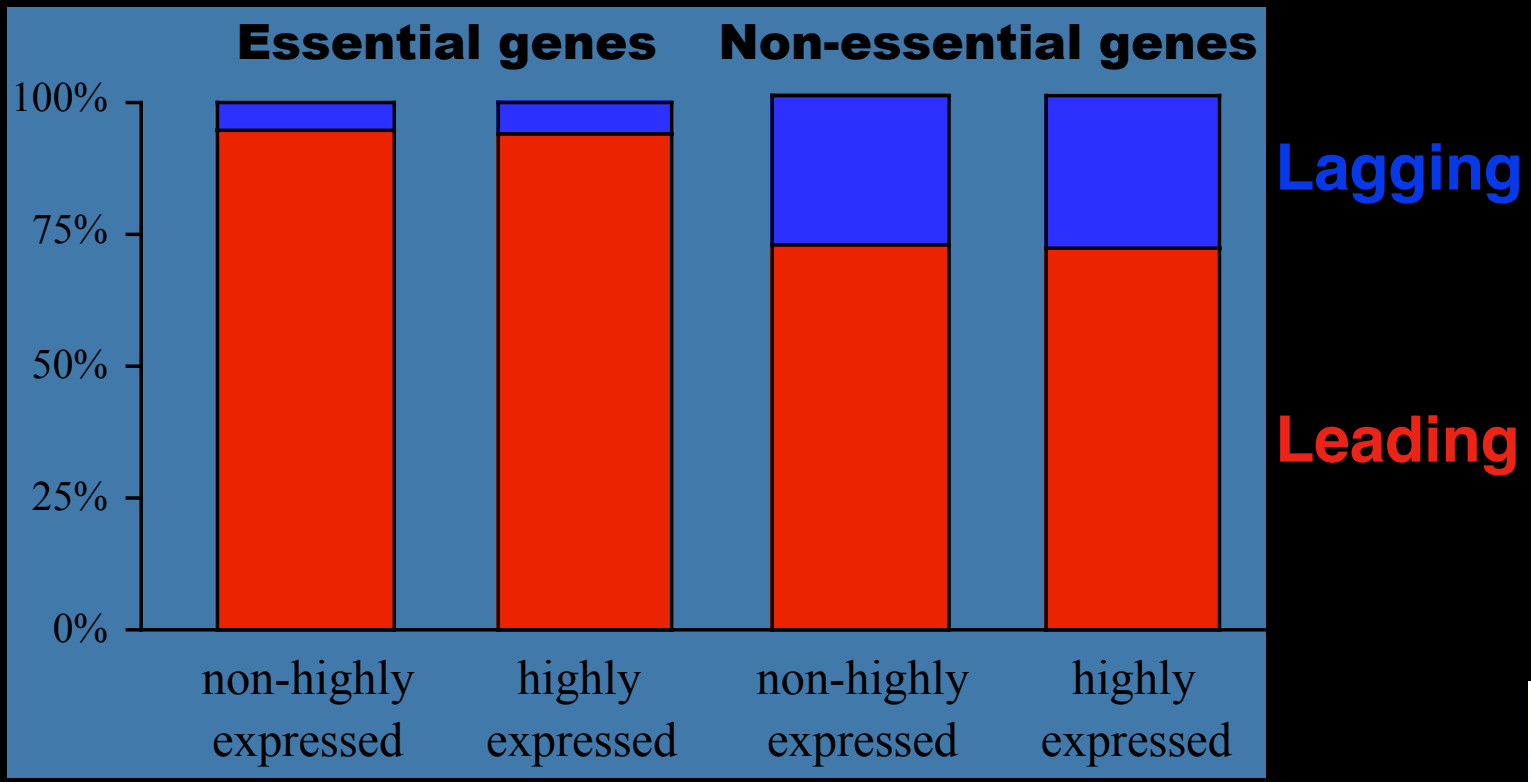
http://www.pasteur.fr/recherche/unites/REG

adanchin@pasteur.fr



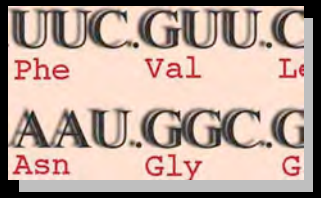
Essentiality in *B. subtilis*

INSTITUT PASTEUR



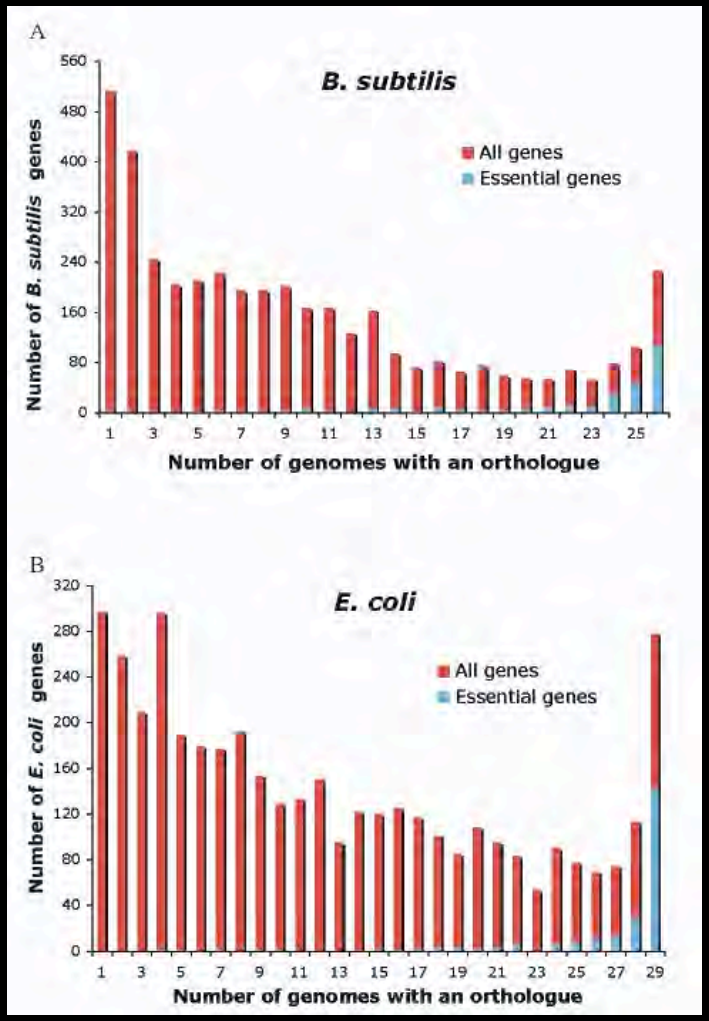
http://www.pasteur.fr/recherche/unites/REG

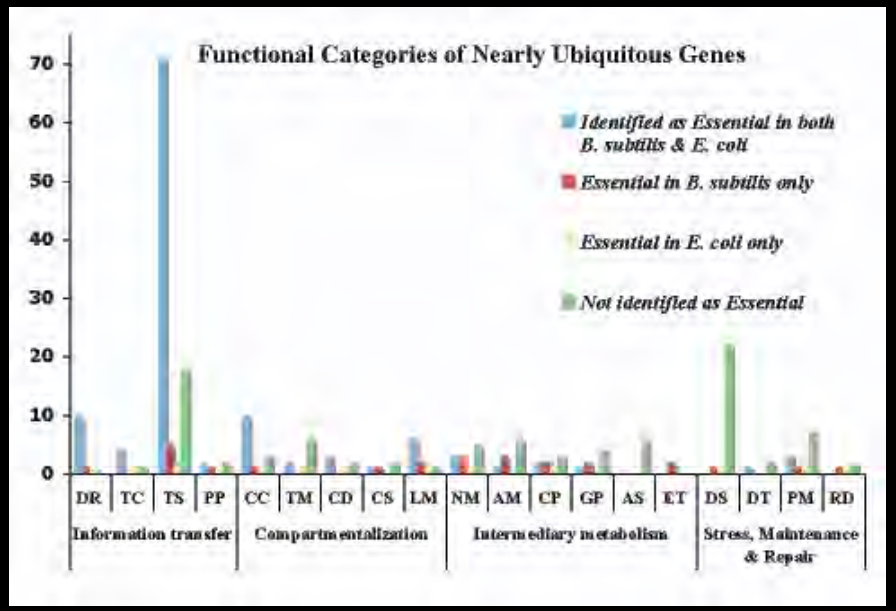
adanchin@pasteur.fr



Gene persistence

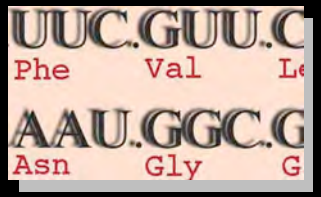
INSTITUT PASTEUR





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

adanchin@pasteur.fr

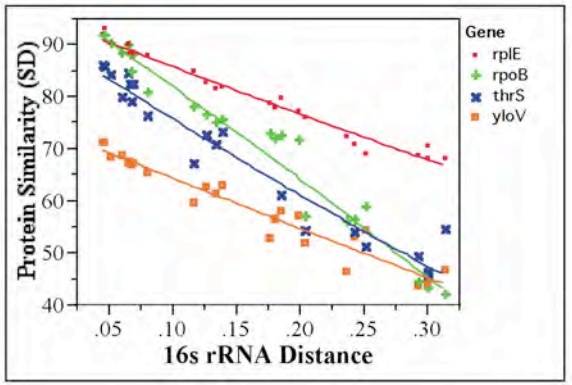


Gene persistence

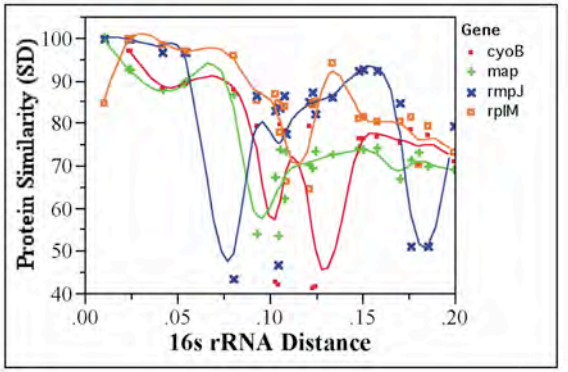
INSTITUT PASTEUR

Some of the 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. As expected, the correlations were high. For example (Figure A), 38% (resp. 48%) of *B. subtilis* (resp. *E. coli*) persistent genes showed a correlation coefficient >0.9 between the sequence similarity of the pair of orthologs and the 16S. In contrast, some genes (Figure 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. However, the genes presenting such an erratic pattern are rare in the persistent set.

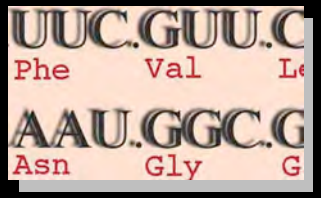
A



B



adanchin@pasteur.fr



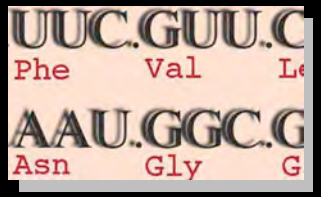
Replication transcription conflicts

INSTITUT PASTEUR

- Transcription may proceed opposite to the movement of the replication fork movement
- This will abort transcription, leading to truncated mRNA
- If translated truncated mRNA may lead to truncated proteins, this will become negative dominant if in complexes…

E.P.C. Rocha & A. Danchin Essentiality, not expressiveness, drives gene-strand bias in bacteria. Nature Genetics (2003) 34 : 377-378

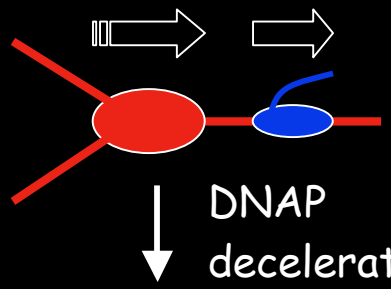
adanchin@pasteur.fr

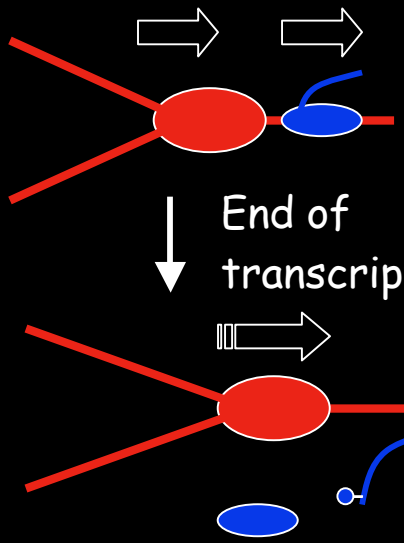


When polymerases collide

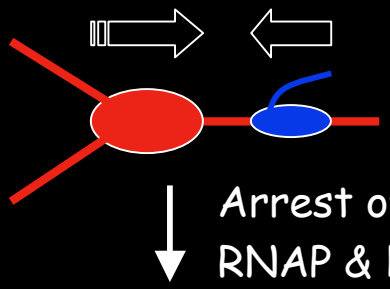
INSTITUT PASTEUR

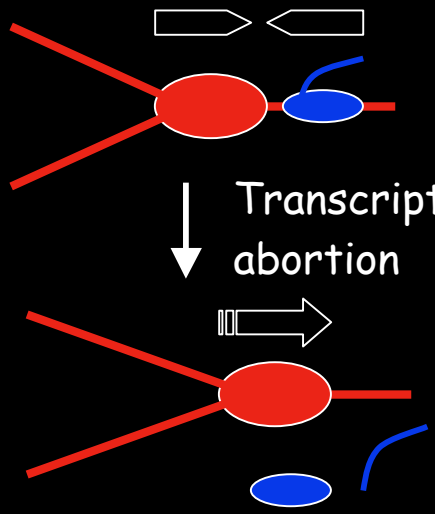
Co-oriented





Head-on





~~Consequences:~~

~~1. Replication slow-down~~

~~2. Loss of transcripts~~

Consequences:

1. Aborted transcripts
2. Truncated essential proteins

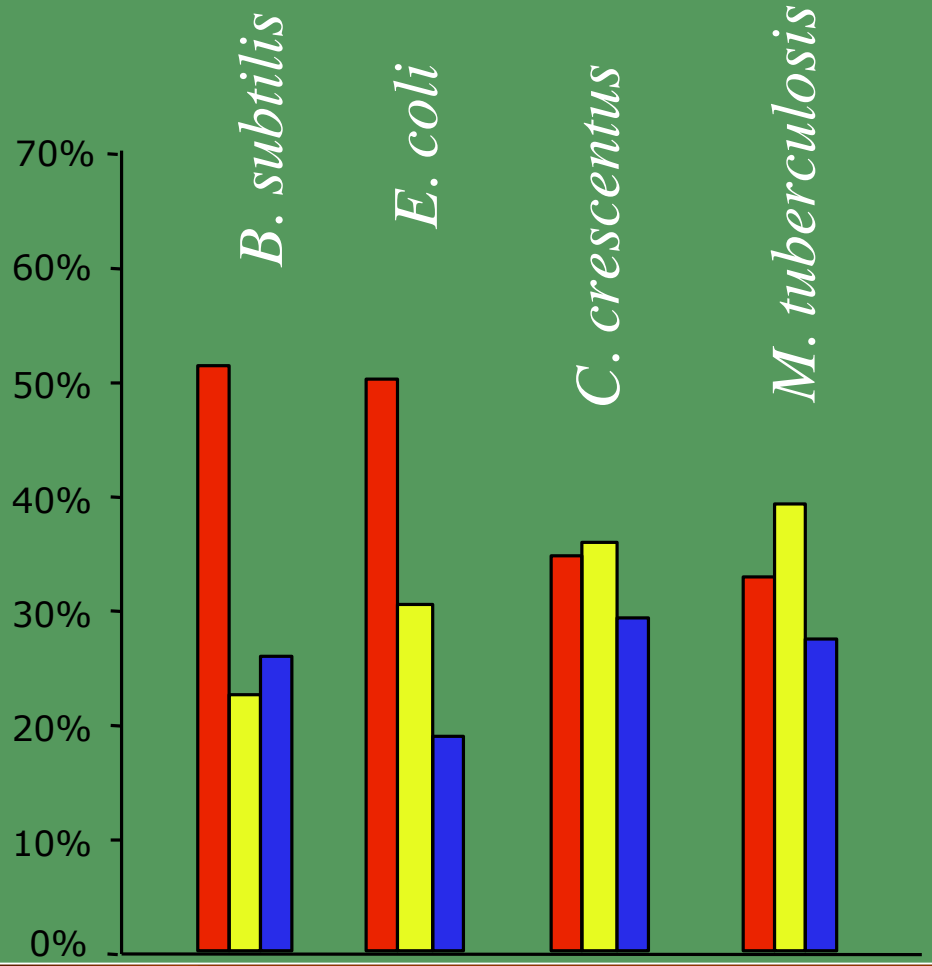
E. Rocha

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

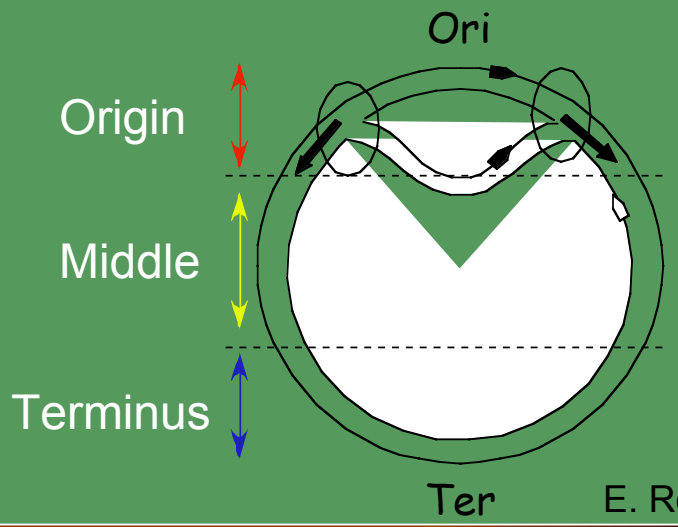
adanchin@pasteur.fr

Distribution of highly expressed genes

Fast growers | Slow growers

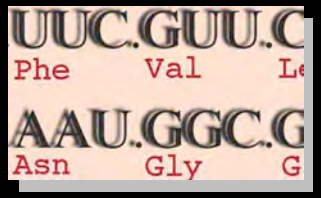


Highly expressed genes cluster near the origin in fast-growing bacteria



E. Rocha

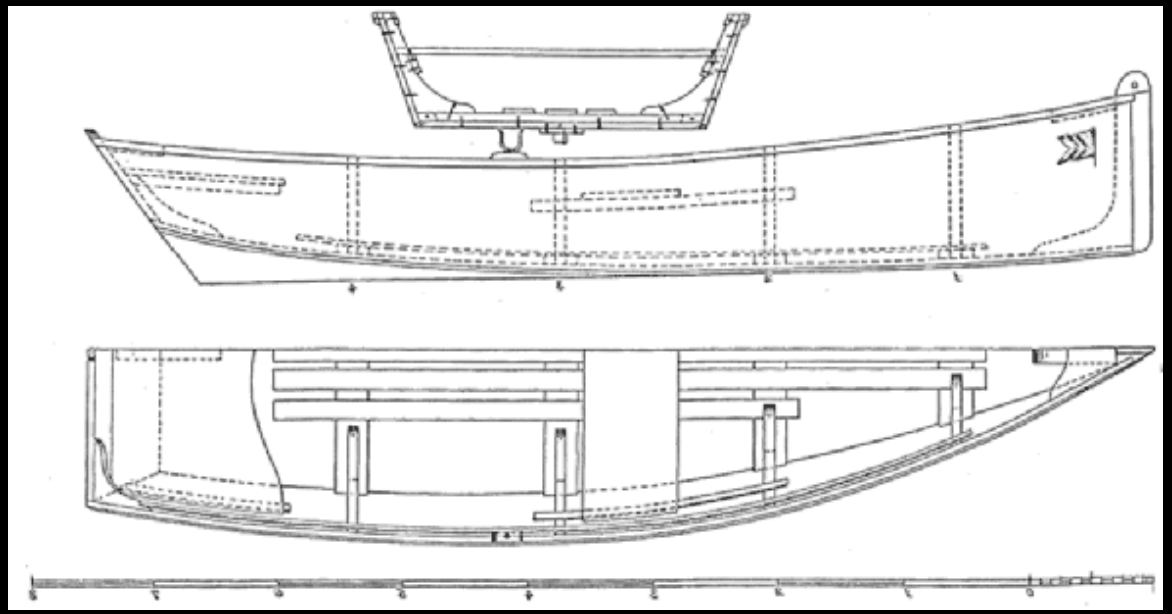
adanchin@pasteur.fr



Symplectic biology: The Delphic Boat

INSTITUT PASTEUR

- Genes do not operate in isolation
- Proteins are part of complexes, as are parts in an engine
- It is important to understand their relationships, as those in the planks which make a boat

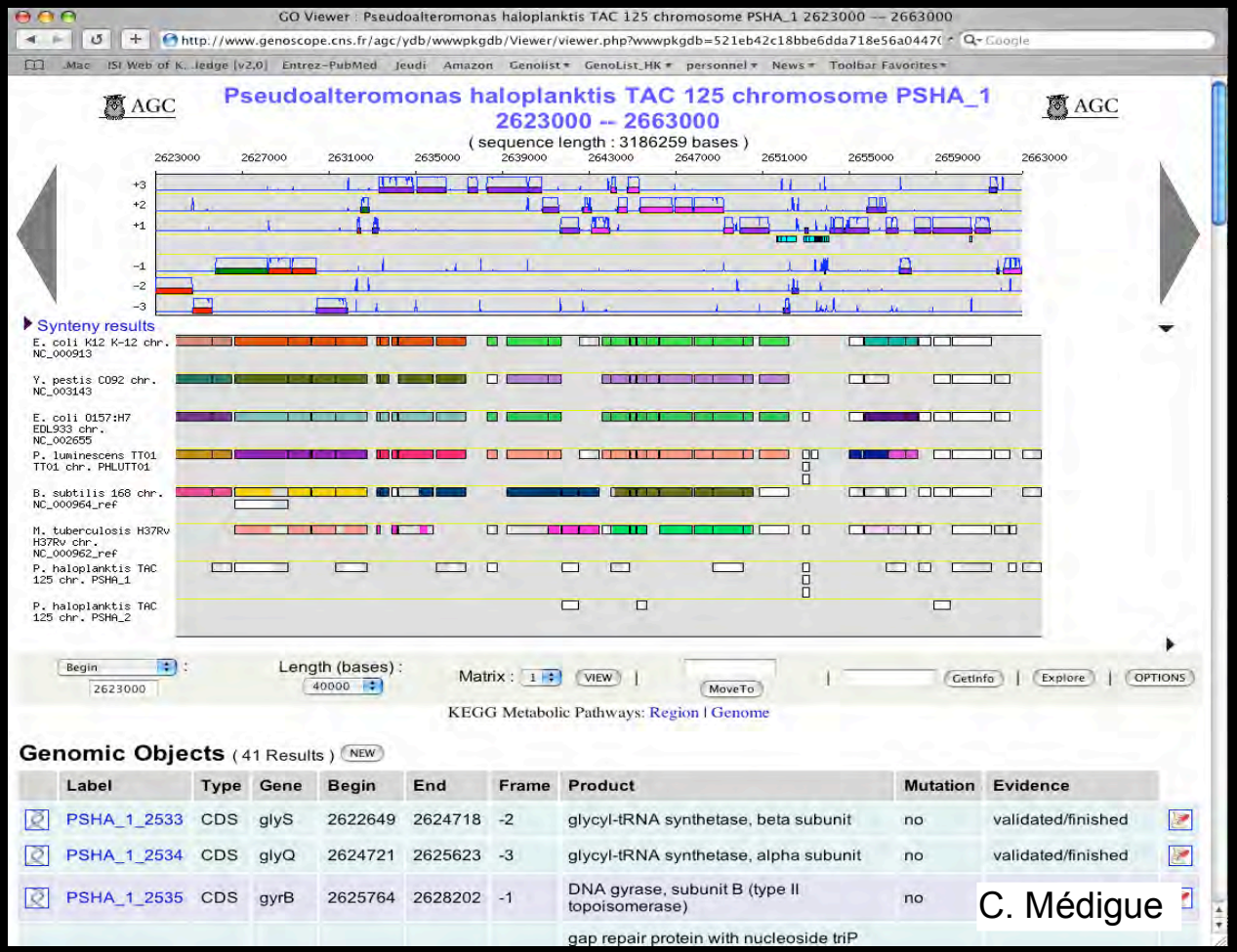


The Delphic Boat: Harvard University Press, february 2003

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

adanchin@pasteur.fr

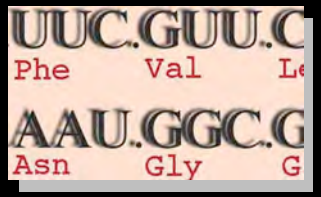
Gene vicinity: synteny



Label	Type	Gene	Begin	End	Frame	Product	Mutation	Evidence
PSHA_1_2533	CDS	glyS	2622649	2624718	-2	glycyl-tRNA synthetase, beta subunit	no	validated/finished
PSHA_1_2534	CDS	glyQ	2624721	2625623	-3	glycyl-tRNA synthetase, alpha subunit	no	validated/finished
PSHA_1_2535	CDS	gyrB	2625764	2628202	-1	DNA gyrase, subunit B (type II topoisomerase)	no	
						gap repair protein with nucleoside triP		

C. Médigue

adanchin@pasteur.fr



Multivariate Analyses

In contrast to standard genetics, genomics analyses large collections of genes and gene products.

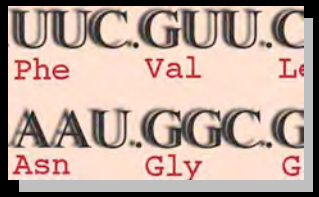
Multivariate analyses try to extract information by simplifying the number of relevant descriptors in the objects of interest.

Principal Component Analysis uses the centered average and a simple distance (identity); it is the reference method.

Correspondence Analysis belongs to the same family, but it uses the $\chi 2$ measure as a distance. This allows the user not only to work with highly heterogeneous objects but also to work simultaneously on the space of objects and on the space of descriptors.

Independent Component Analysis uses the non gaussian character of the values associated to descriptors

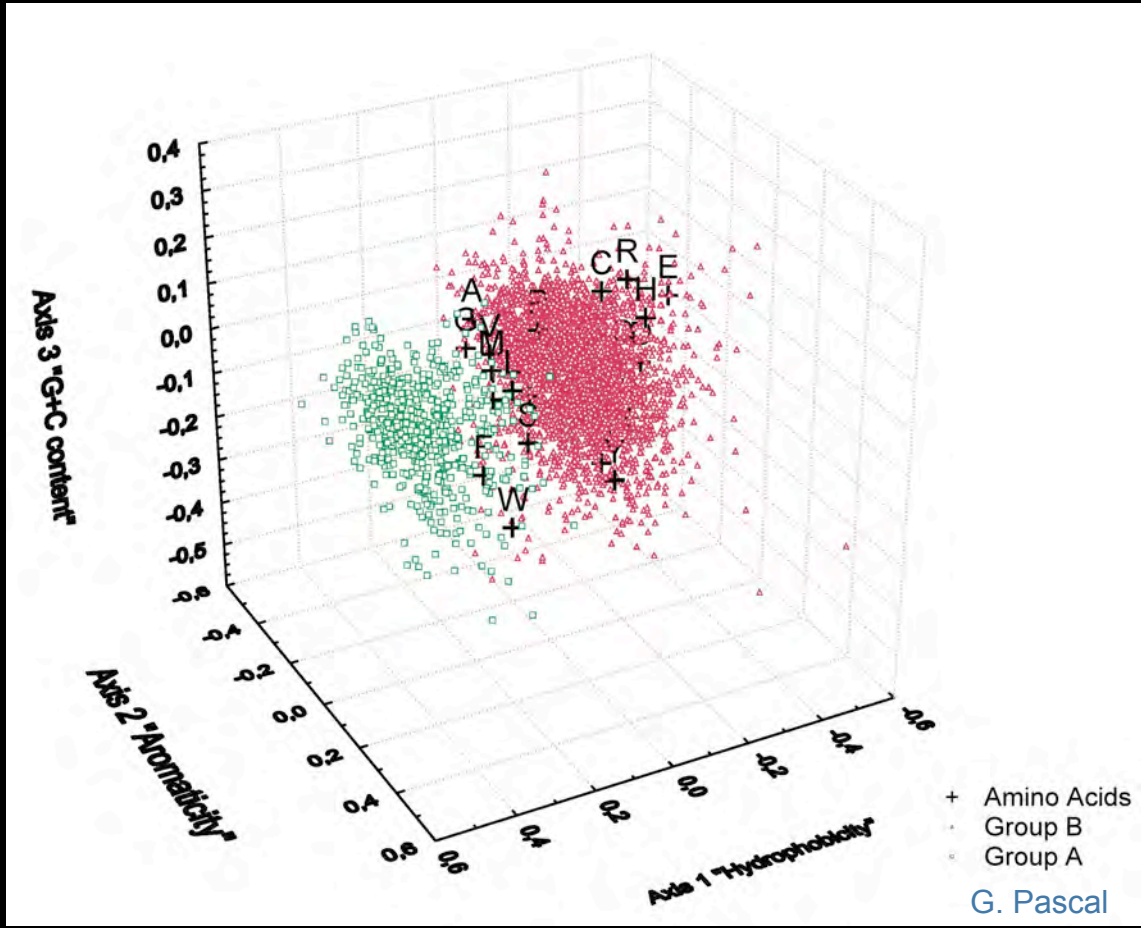
adanchin@pasteur.fr



Bias in amino acid distribution

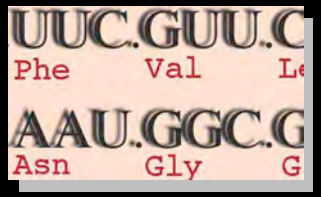
INSTITUT PASTEUR

Neighbourhood: distribution of aminoacids in the proteome



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

adanchin@pasteur.fr



Universal biases in protein amino acid composition

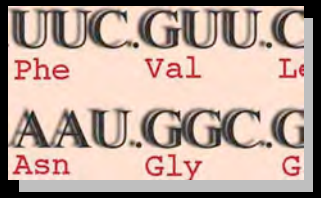
INSTITUT PASTEUR

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

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

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

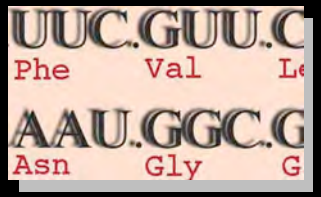
adanchin@pasteur.fr



The “gluons”

- There is an aromatic residues-oriented bias in all genomes
- With proteins of the same size this opposes ribosomal proteins to orphan proteins
- Hypothesis: orphans are “self”-specific proteins that stabilise complexes, they act as “gluons”

adanchin@pasteur.fr

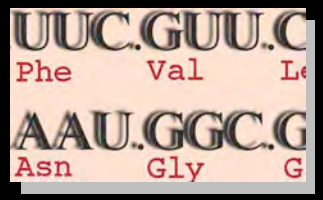


Temperature-dependent biases in protein amino acid composition

INSTITUT PASTEUR

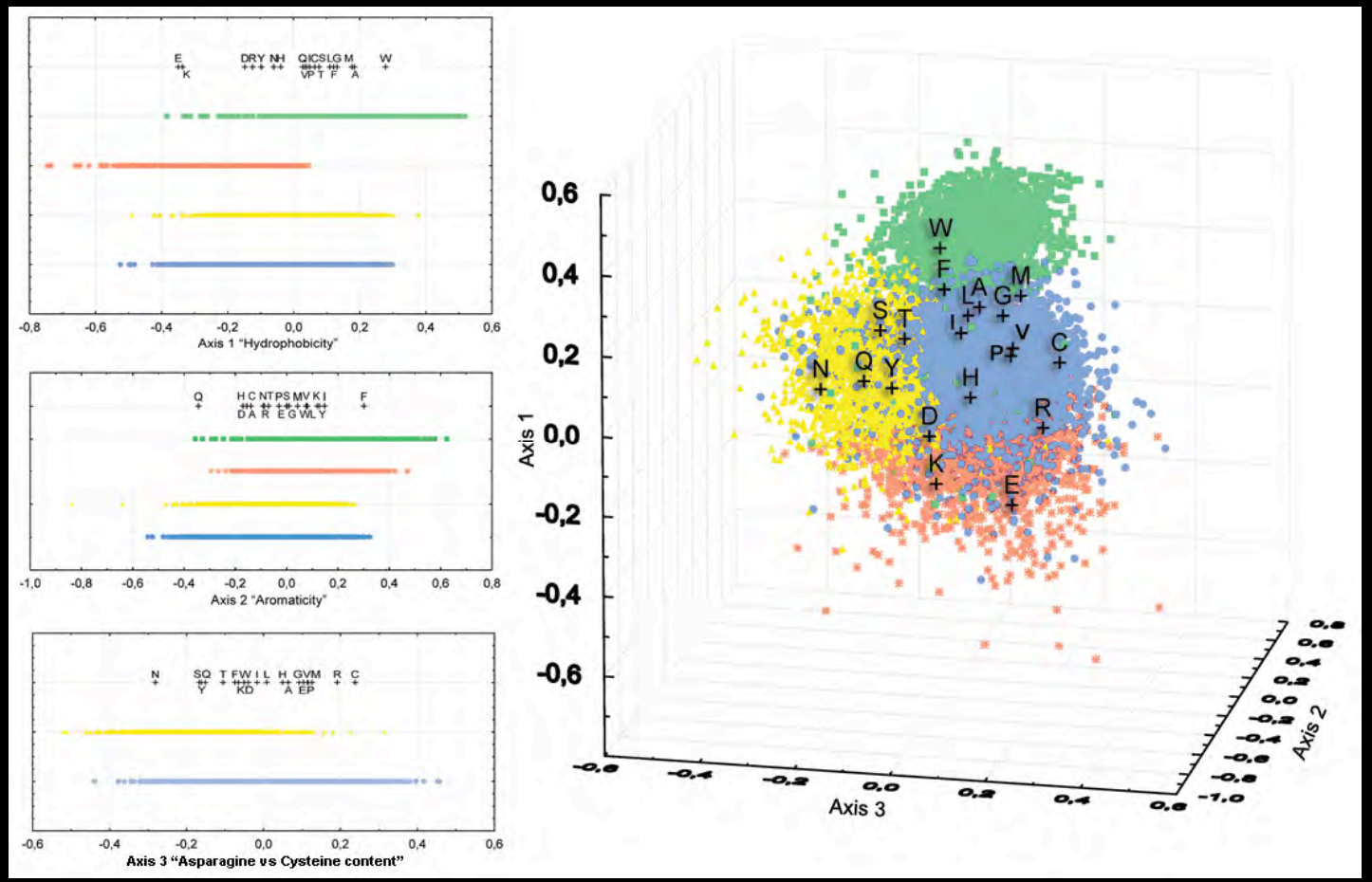
- The amino acid composition of proteins depends heavily on the phylogeny => need to compare organisms related to each other
- The general trend of amino acid composition bias is to avoid some aminoacids at higher temperatures
- Mesophilic bacteria belong to at least two different classes (in a 5-clusters analysis)
- Biases are always dominated by the IIMP clustering

adanchin@pasteur.fr



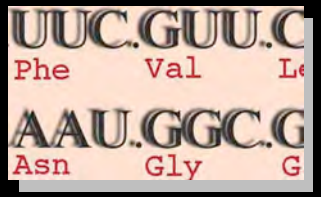
Temperature-dependent amino acid biases

INSTITUT PASTEUR



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

adanchin@pasteur.fr



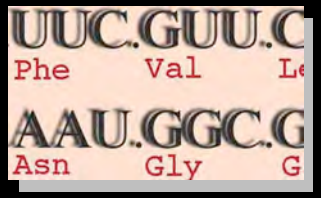
Codon usage biases

INSTITUT PASTEUR

- 20 amino acids 61 codons
- Study of the genes in the codon space, using Correspondence Analysis (χ^2 measure)
- At least three classes of genes, including one corresponding to horizontal transfer

C. Médigue, T. Rouxel, P. Vigier, A. Hénaut & A. Danchin. Evidence for horizontal gene transfer in *Escherichia coli* speciation J. Mol. Biol. (1991) 222 pp. 851-856

adanchin@pasteur.fr



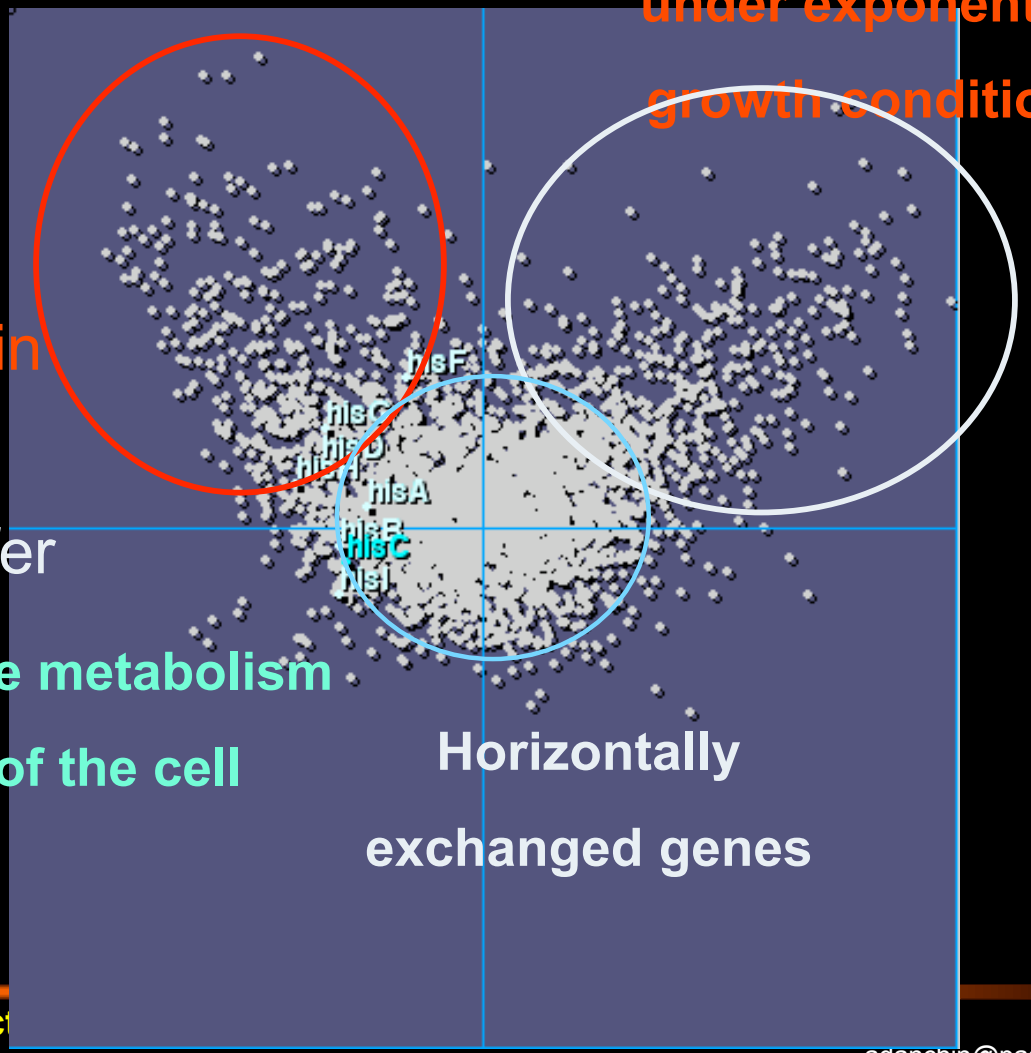
Gene exchange

Genes expressed at a high level under exponential growth conditions

Class I: core metabolism

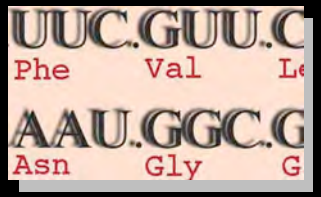
Class II: high expression in exponential growth

Class III: horizontal transfer



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

adanchin@pasteur.fr



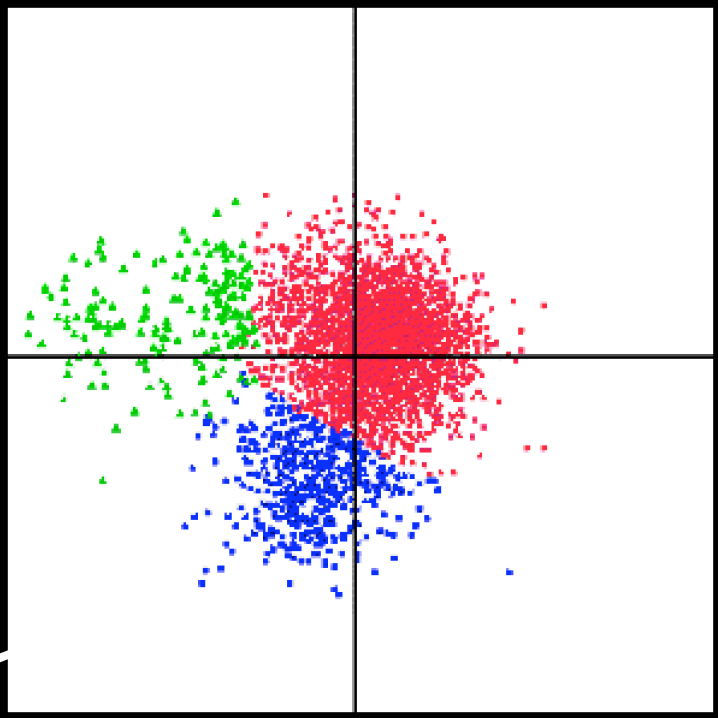
Codon usage, organisation and evolution of the *B. subtilis* genome

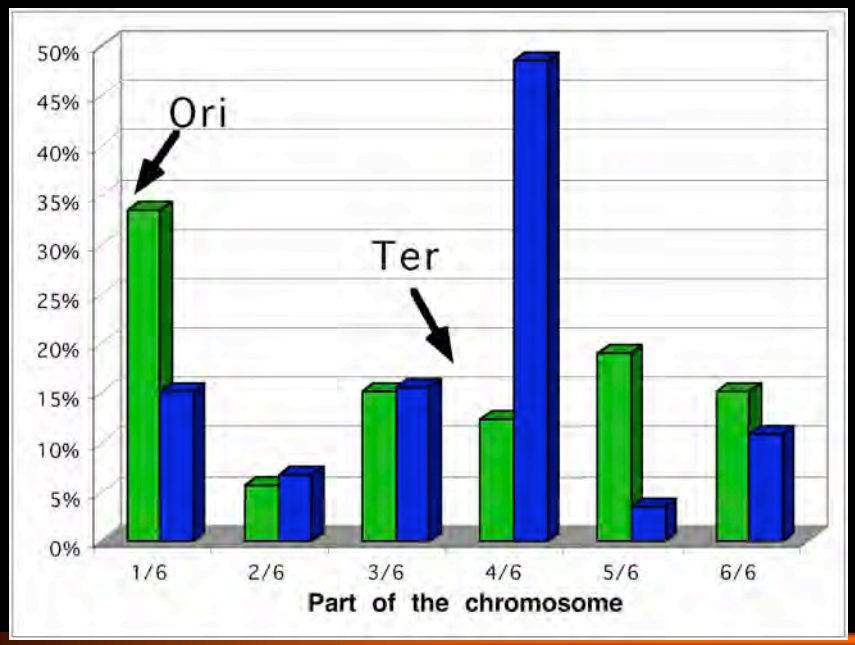
INSTITUT PASTEUR

	AAA	AAC	AAG ...
gene1	n_{11}	n_{21}	n_{31}
gene2	n_{12}	n_{22}	n_{32}
gene3	n_{13}	n_{23}	n_{33}
....	....	....	....
geneN	n_{1n}	n_{2n}	n_{3n}

Correspondence analysis

Classification

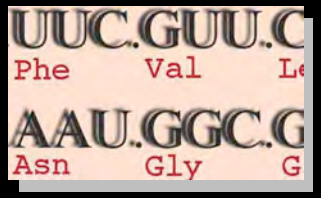




Highly expressed
Atypical / HGT
Others

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

adanchin@pasteur.fr

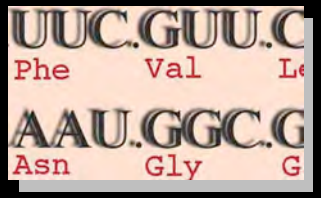


The cell organizers

INSTITUT PASTEUR

It is too early to understand the selection pressures that organize the cell architecture. However, at least in bacteria, the role of gasses and chemical highly reactive radicals play probably a major role. Most of the corresponding genes are still unknown….

adanchin@pasteur.fr



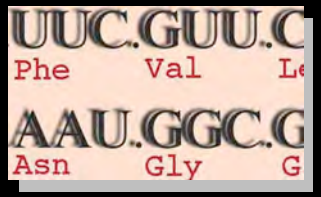
Selection pressure for organisation: oxido-reduction

INSTITUT PASTEUR

- Sulfur undergoes oxido-reduction reactions from -2 to +6
- Incorporation of sulfur into metabolism usually requires reduction to the gaseous form H_2S
- H_2S is highly reactive, in particular towards dioxygen
- => These two gasses, despite their diffusion properties, must be kept separate as much as possible
- Sulfur scavenging is energy-costly
- => Sulfur containing molecules have to be recycled

A. Sekowska, H-F. Kung & A. Danchin Sulfur metabolism in *Escherichia coli* and related bacteria, facts and fiction.
J. Mol. Microbiol. Biotechnol. (2000) 2: 145-177

adanchin@pasteur.fr



Sulfur metabolism: an unexpected organiser of the cell 's architecture

INSTITUT PASTEUR

- Sulfur metabolism-related proteins are more acidic (average pI 6.5) than bulk proteins (richer in asp and glu), they are poor in serine residues

- They are significantly poor in sulfur-containing amino-acids

- Their genes are very poor in codons ATA, AGA and TCA

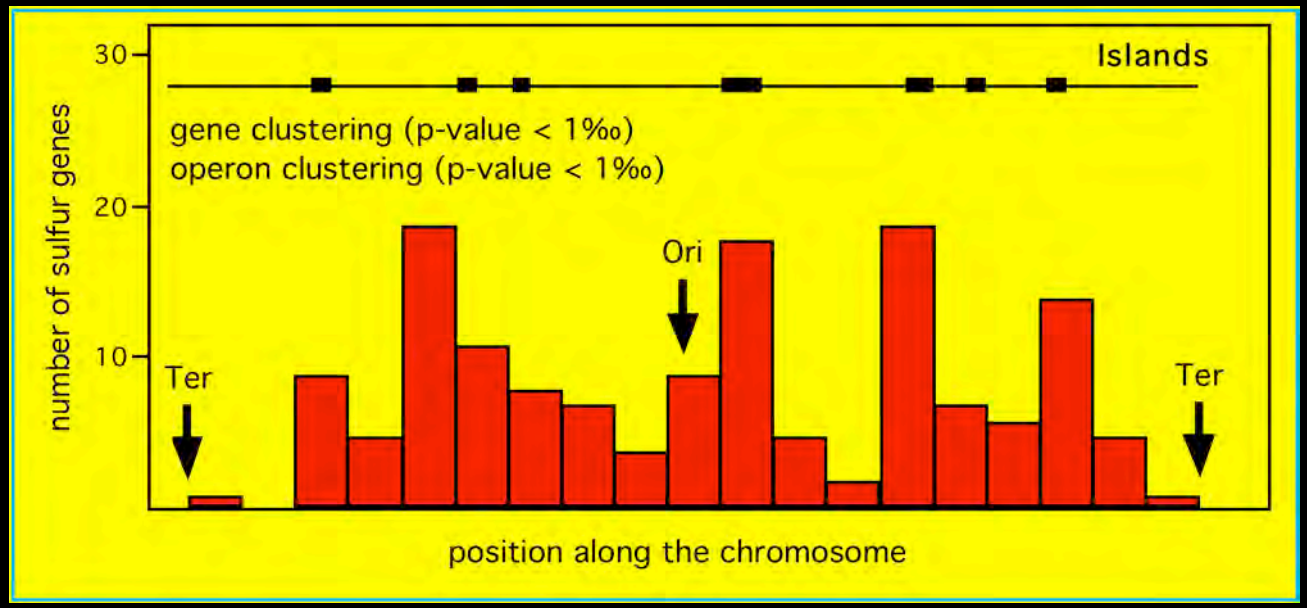
- There are no class III (horizontal transfer) genes in the class (only 2 in 150 genes)

- => sulfur-metabolism genes are ancestral and may for a core structure for the *E. coli* genome

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

adanchin@pasteur.fr

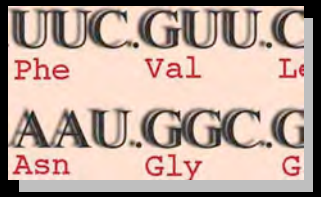
Proximity in the chromosome
Sulphur islands



E.P.C. Rocha, A. Sekowska & A. Danchin Sulfur islands in the *Escherichia coli* genome: markers of the cell's architecture? FEBS Lett. (2000) 476: 8-11

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

adanchin@pasteur.fr



The error catastrophe

INSTITUT PASTEUR

- Similarity in sequence leads to functional inference
- Because of recruitment of pre-existing structures, there is often no obvious link between a structure and a function (the book-paperweight)
- Hence a propagation of annotation errors
- *ykrS* (*mtnA*) annotated as « translation factor » is a component of sulfur metabolism!

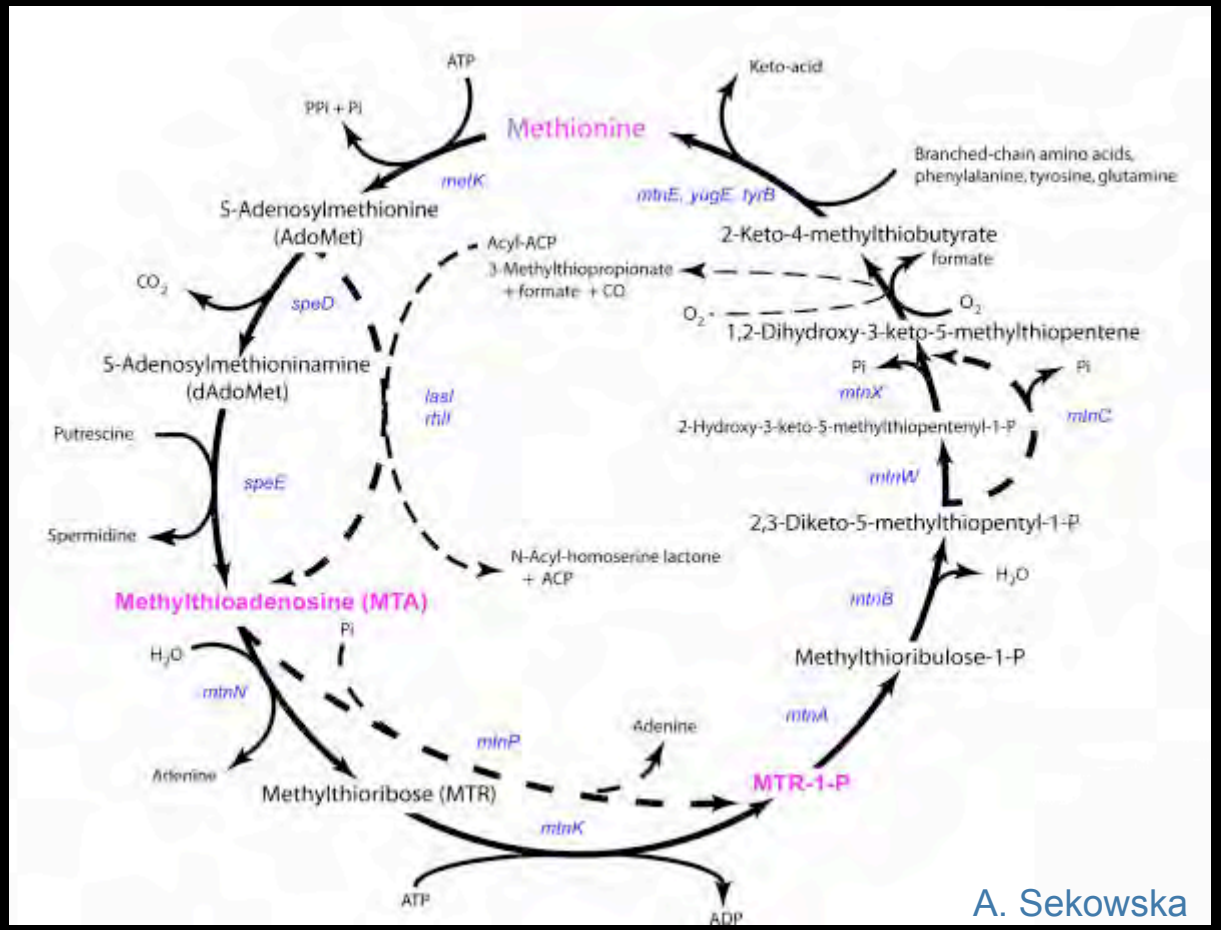
A Sekowska, V Dénervaud, H Ashida, K Michoud, D Haas, A Yokota, A Danchin Bacterial variations on the methionine salvage pathway *BMC Microbiol* (2004) **4:** 9

adanchin@pasteur.fr

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

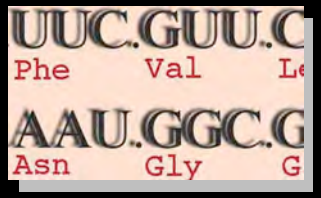
A new metabolic pathway

INSTITUT PASTEUR



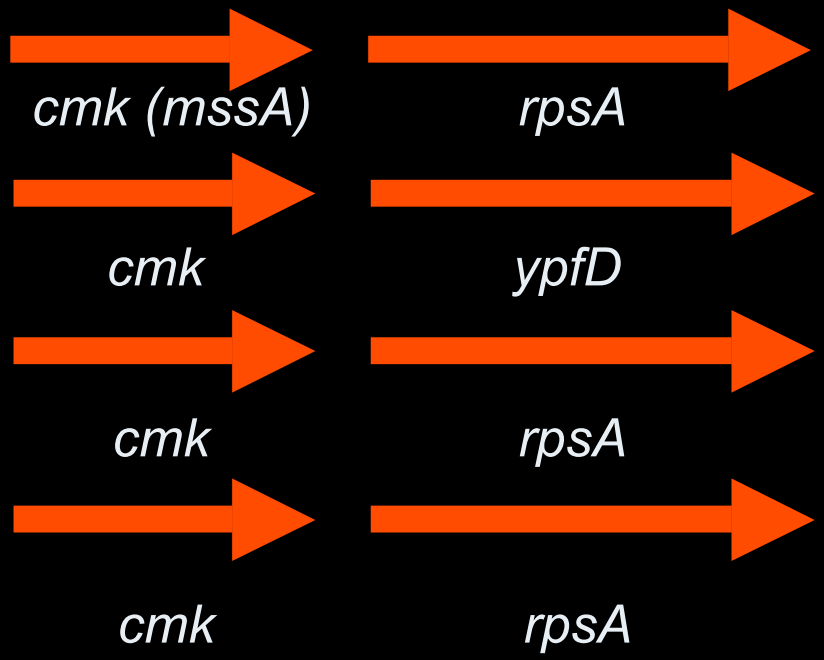
A. Sekowska

adanchin@pasteur.fr



Just so story: proximity in the genome

INSTITUT PASTEUR



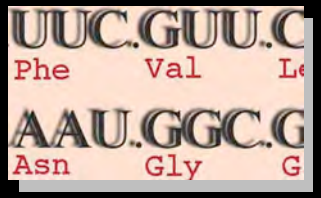
Escherichia coli

Bacillus subtilis no rpsA !!!

Haemophilus influenzae

Sinorhizobium meliloti

adanchin@pasteur.fr



The pyrimidine diphosphate paradox

INSTITUT PASTEUR

In order to make deoxyribonucleotides the cell uses ribonucleosides diphosphates, not triphosphates

NDP ⟶ dNDP ⟶ dNTP

NDR *NDK*

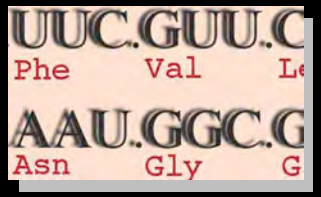
And here is the paradox:

OMP ⟶ UMP ⟶ UDP ⟶ UTP ⟶ CTP

no CDP !!!

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

adanchin@pasteur.fr



How is the paradox resolved?

INSTITUT PASTEUR

OMP → UMP → UDP → UTP → CTP

RNases

CMP

Cmk

mRNA

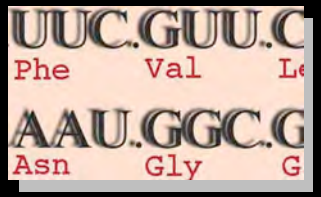
CDP

PNPase

dCDP

DNA

adanchin@pasteur.fr



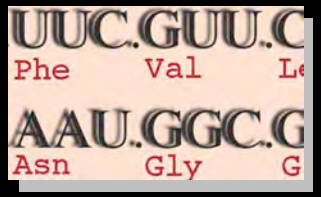
Phylogenetic neighbours: the S1 box

INSTITUT PASTEUR

- *rpsA* codes for ribosomal protein S1. It contains the S1 box (PROSITE PS50126). Many other proteins contain a similar box: polynucleotide phosphorylase, RNases E, G and R, RNA helicases etc.
- protein RegB of bacteriophage T4, associated to S1, cuts mRNA at GAGG motifs.
- S1 is a subunit of bacteriophage Qβ replicase…

=> All this points to a function for S1 in RNA metabolism

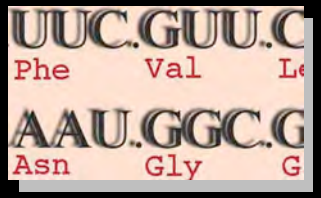
adanchin@pasteur.fr



Codon usage bias neighbours

Gene	Comment
bla *cat* *dicB* *lpp* *ompA*	long mRNA turnover
pyrF	pyrimidine metabolism
hflB *ftsH* *mrsACF* *lpp*	cell architecture
nusA *pcnB* *metY* *pnp* *rna* *rnb* *rnc* *rne/ams* *rng* *rph*	RNA maturation and turnover
trxA	oxido-reduction, subunit of T7 replicase, needed for synthesis of deoxyribonucleotides

adanchin@pasteur.fr



Protein complexes: the Degradosome

INSTITUT PASTEUR

PNPase
PolyA polymerase
RNAse E
S1

Polyphosphate kinase

Enolase

mRNA degradation

CDP for de novo DNA synthesis

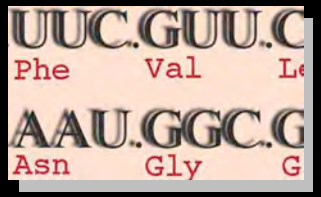
GDP recycling of GTP for carbohydrate secretion

NDK +PYK

GDP + PEP → **GTP**

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

adanchin@pasteur.fr



Just so story: the *cmk rpsA* operon

Escherichia coli

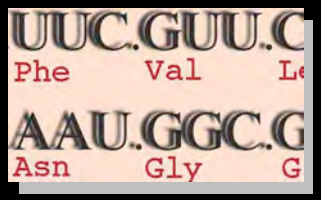
mssA **was discovered as a suppressor of** ***smbA (pyrH)*****, itself a suppressorof MukB, a myosin-like protein involved in chromosome segregation**

=> DNA synthesis is involved in the function.

Conclusion:

The function of the *cmk rpsA* operon is to make CDP for DNA synthesis

adanchin@pasteur.fr



Selection pressure for compartmentalisation: a dangerous intermediate

INSTITUT PASTEUR

Uridylate kinase (UMK)
pyrH (smbA)

No CDP: no DNA…

OMP → UMP → UDP → UTP → CTP

UDP → dUDP → dUTP → dUMP + PPi

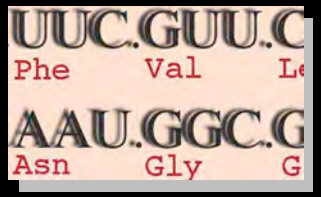
dUMP + PPi → dTMP → dTDP → dTTP → DNA

dUTP → ~~DNA~~

S. Noria & A. Danchin Just so genome stories : what does my neighbor tell me? International Congress Series 1246 Elsevier Science (2002) 3-13

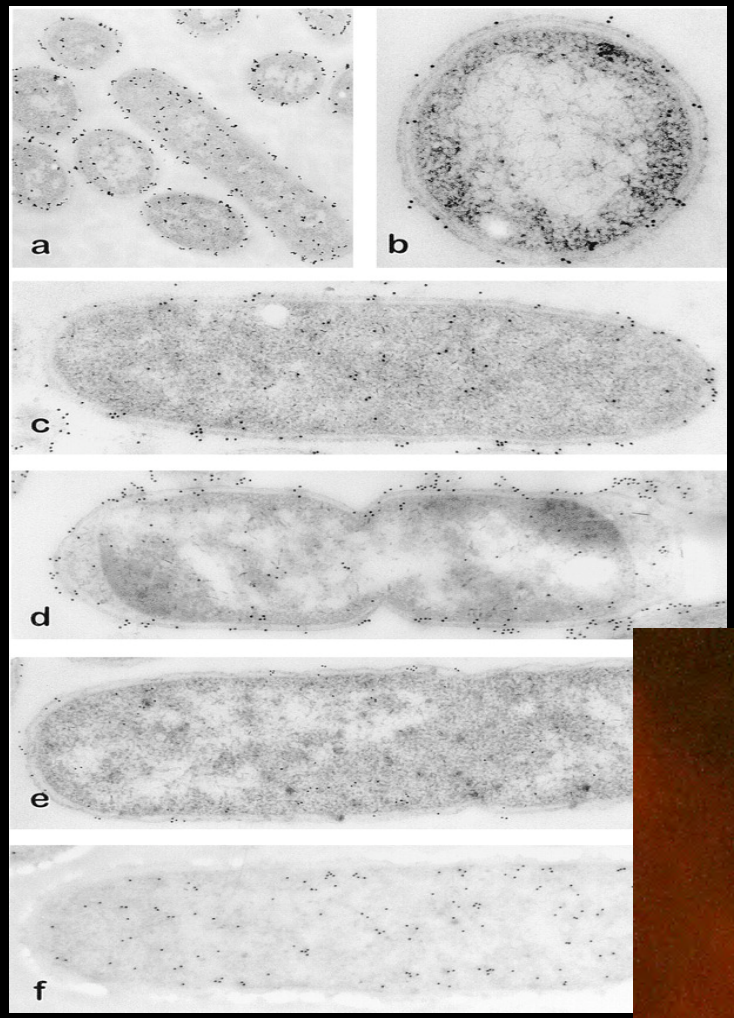
http://www.pasteur.fr/recherche/unites/REG

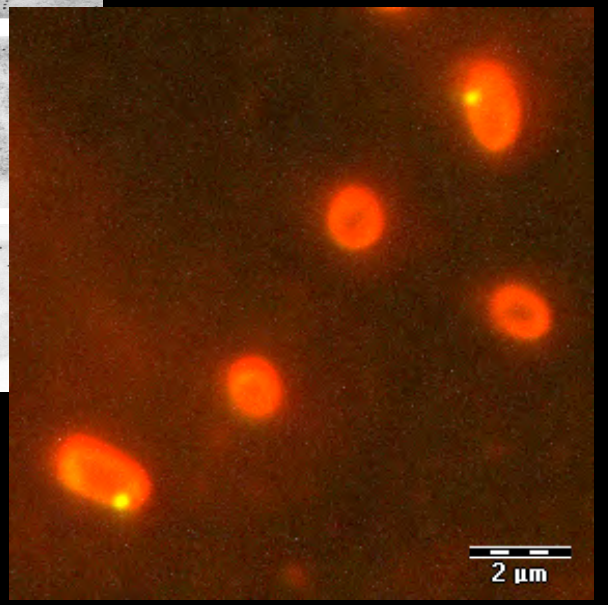
adanchin@pasteur.fr



In conclusion:

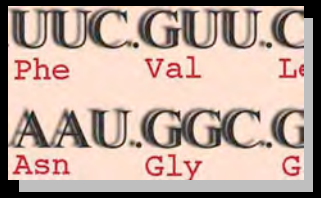
UMK must be compartmentalised





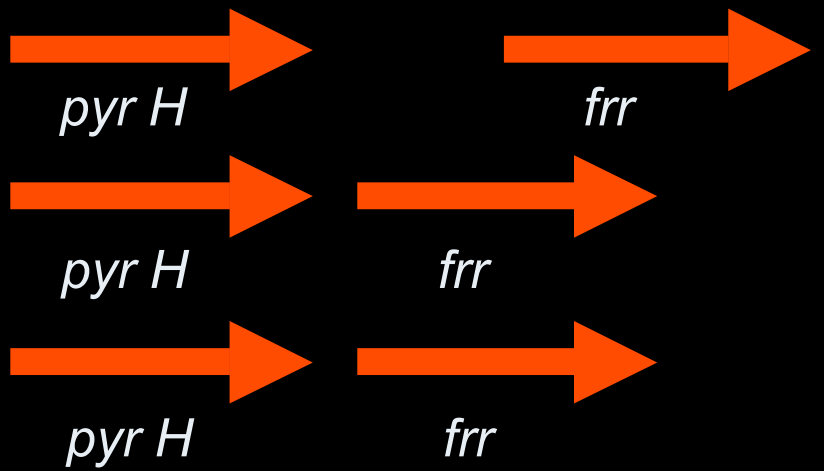
S. Landais, P. Gounon, C. Laurent-Winter, J.C. Mazié, A. Danchin, O. Barzu & H. Sakamoto Immunochemical analysis of UMP kinase from *Escherichia coli.* J. Bacteriol. (1999) **181:** 833-840

adanchin@pasteur.fr



A prediction: ribosome recycling and UTP

INSTITUT PASTEUR



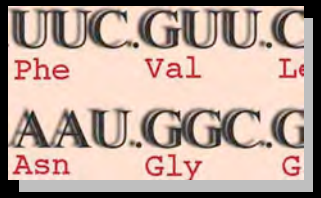
Escherichia coli

Bacillus subtilis

Photorhabdus luminescens

This organisation is conserved in most Gram+ and Gram- bacteria. **Why ?**

adanchin@pasteur.fr



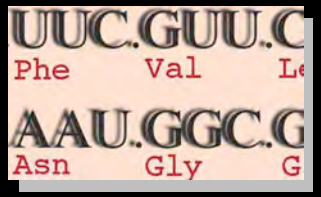
Ribosome recycling and UTP

INSTITUT PASTEUR

frr codes for the ribosome recycling factor, that allows 70S ribosomes to split into 30S and 50S subunits. In polycistronic operons, the 70S ribosome can go on from one gene to the next one without recycling (this requires formylation of the first methionine). At the end of the message, the ribosomes must recycle. This happens in a context where transcripts make stem and loops, ending with a polyU sequence.

Conjecture: is UTP controlling the activity of Frr? Remember that one cannot speak of « concentrations » of molecules in a cell. 1 micromolar would mean 600 molecules. There are 20,000 ribosomes, therefore 1 mM means only **30 individual molecules** in the immediate vicinity of each ribosome...

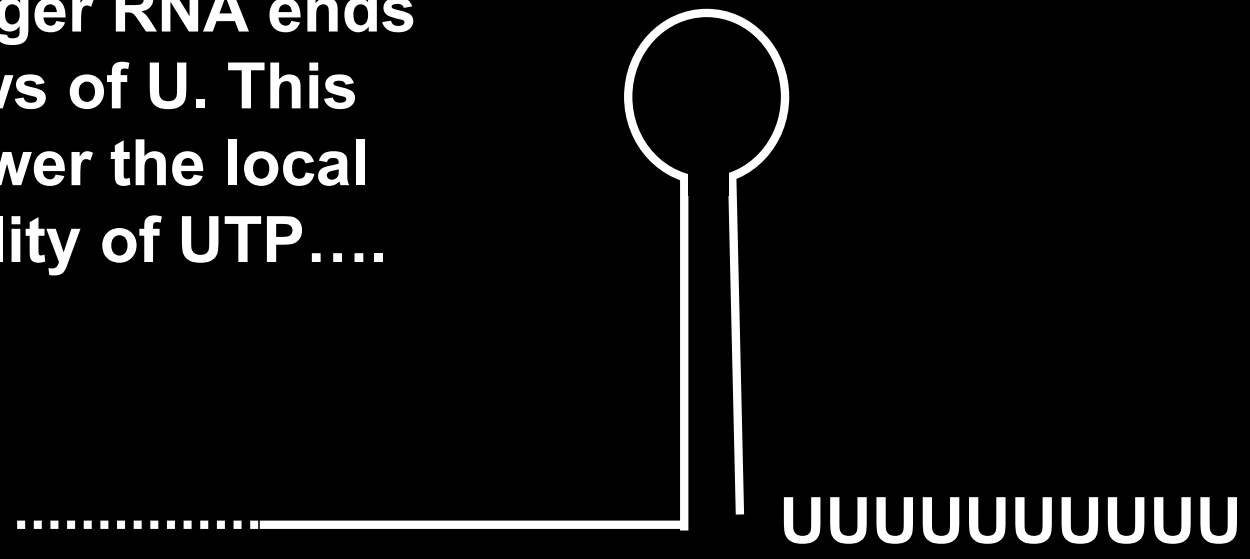
adanchin@pasteur.fr



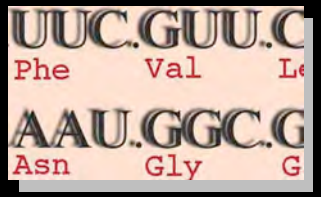
Transcription termination

INSTITUT PASTEUR

At Rho-independent sites for termination of transcription the messenger RNA ends with rows of U. This must lower the local availability of UTP….



This suggests Frr as a drug target, with analogs of UTP as leads...



A preconceived ideology

INSTITUT PASTEUR