香港大學－巴斯德研究中心
HKU-Pasteur Research Centre

Exploration of neighbourhoods for inductive reasoning

Authors

in silico

- Eduardo Rocha
- Ivan Moszer
- Claudine Médigue

in vivo / in vitro

- Agnieszka Sekowska
- Anne Marie Gilles
- Octavian Barzu

Collective

- Stanislas Noria

A Chinese view for ....

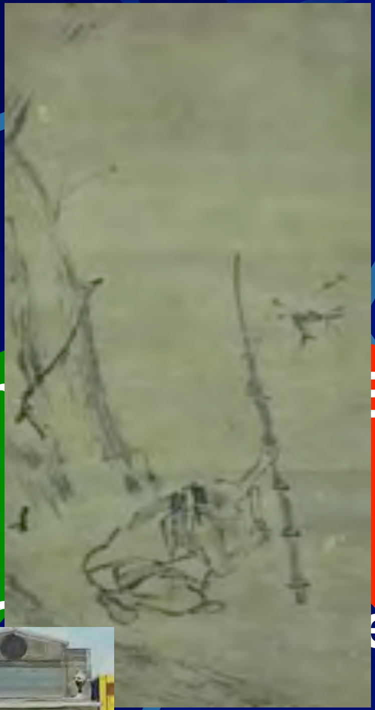
a virtuous circle

- Context
- Data
- Hypotheses => Today’s presentation

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

« Bombardment of th

Embassy in Belgrade »



en

H

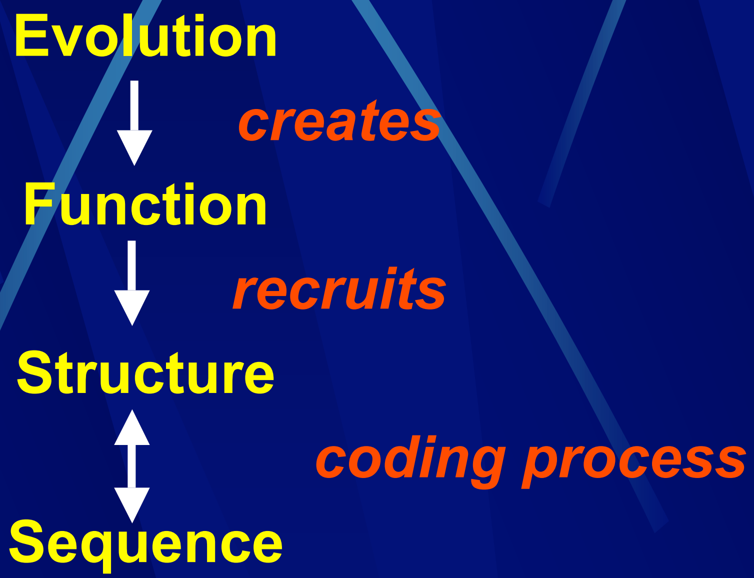
n

Data vs Hypotheses

What biological question are you asking?

Empedocle / Maupertuis / Malthus / Darwin

Variation / Selection / Amplification



Evolution

creates

Function

recruits

Structure

coding process

Sequence

What is Life?

- Physics: *matter, energy, time*

- Biology: Physics + *information, coding, control...*

What functions for Life? An extension of Cuvier’s view….

- Physical stability ([cyto]skeleton)
- Reproduction
- Respiration
- Locomotion
- Perception
- Transport (import / export)
- Circulation (internal fluxes)
- Digestion and recycling
- Assimilation
- Accommodation (regulation)
- Maintenance (repair)
- Etc…

What is Life?

Three processes are needed for Life:

- Information transfer (Living Turing Machines)

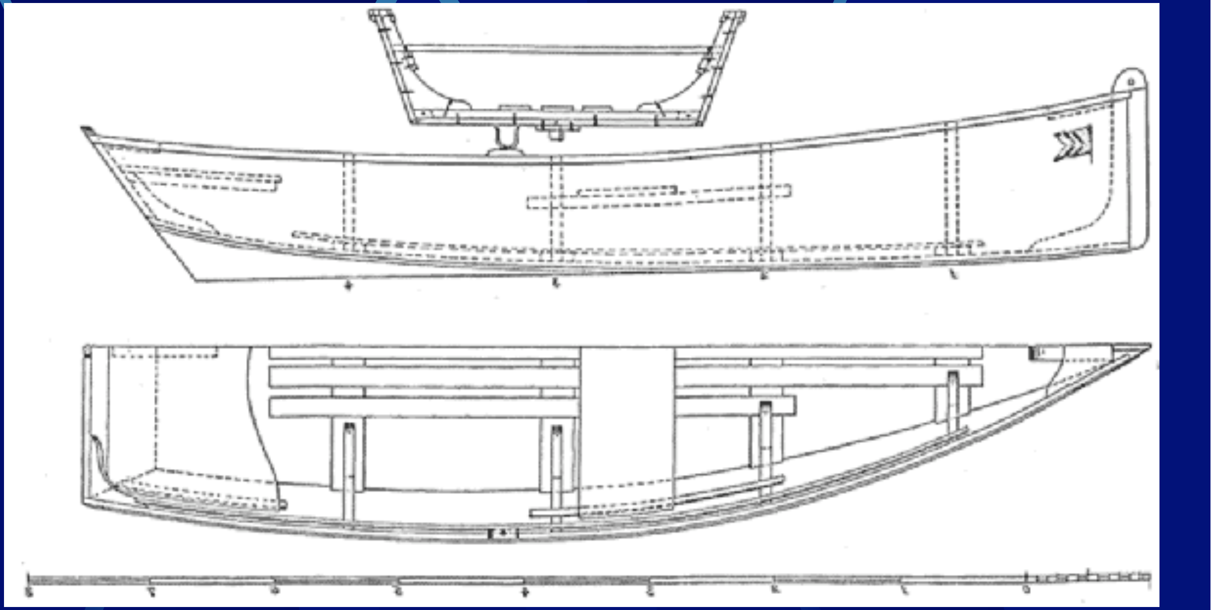
Driving force for a coupling between the genome structure and the structure of the cell:

- Metabolism (Internal organisation)
- Compartmentalization (General structure)

Because of these two processes, note that "concentration" usually does not have a meaning inside a cell

Inductive strategy: exploring “neighborhoods”

- 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

Induction: exploring neighborhoods

To make discoveries we explore the general « neighborhoods » of genes of interest: proximity in the chromosome, in evolution, in the literature, in biochemical complexes, in metabolism etc.

Comparative genomics is essential, hence the use of « subtractive » genomics (comparison of pairs or larger sets of similar genomes)

From sequence to function

Combining genome sequence data and *in silico* prediction (bioinformatics) we test our hypotheses using large scale genomics techniques (transcriptome and proteome analysis) as well as other types of neighborhoods, such as common electric charge or codon usage bias.

↓ Note that regulation evolves much faster than all other processes

Genome organisation

Is the gene order random in the chromosomes?

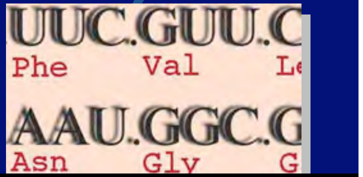
At first sight, despite different DNA management processes not much is conserved, and horizontally transferred genes are distributed throughout genomes

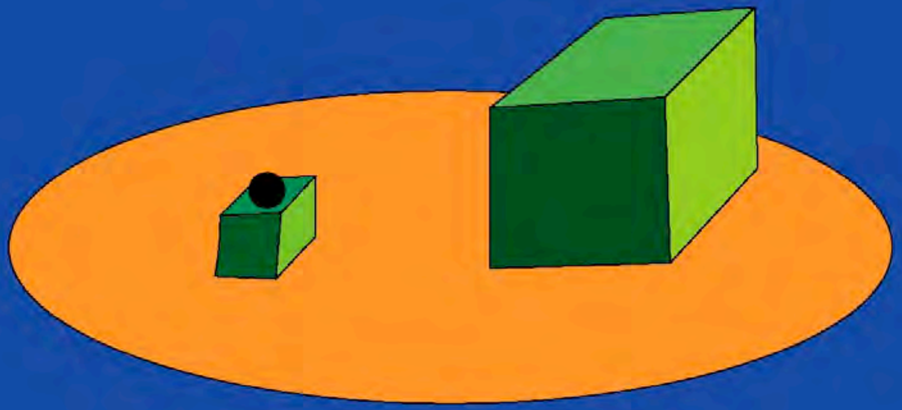
However, groups of genes, such as operons or pathogenicity islands tend to cluster in specific places, and they code for proteins with common functions

A question: where are located repeats?

香港大學－巴斯德研究中心
HKU-Pasteur Research Centre

Caveat:
Repeats are meaningful


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



What does the smaller cube the round support supports support?

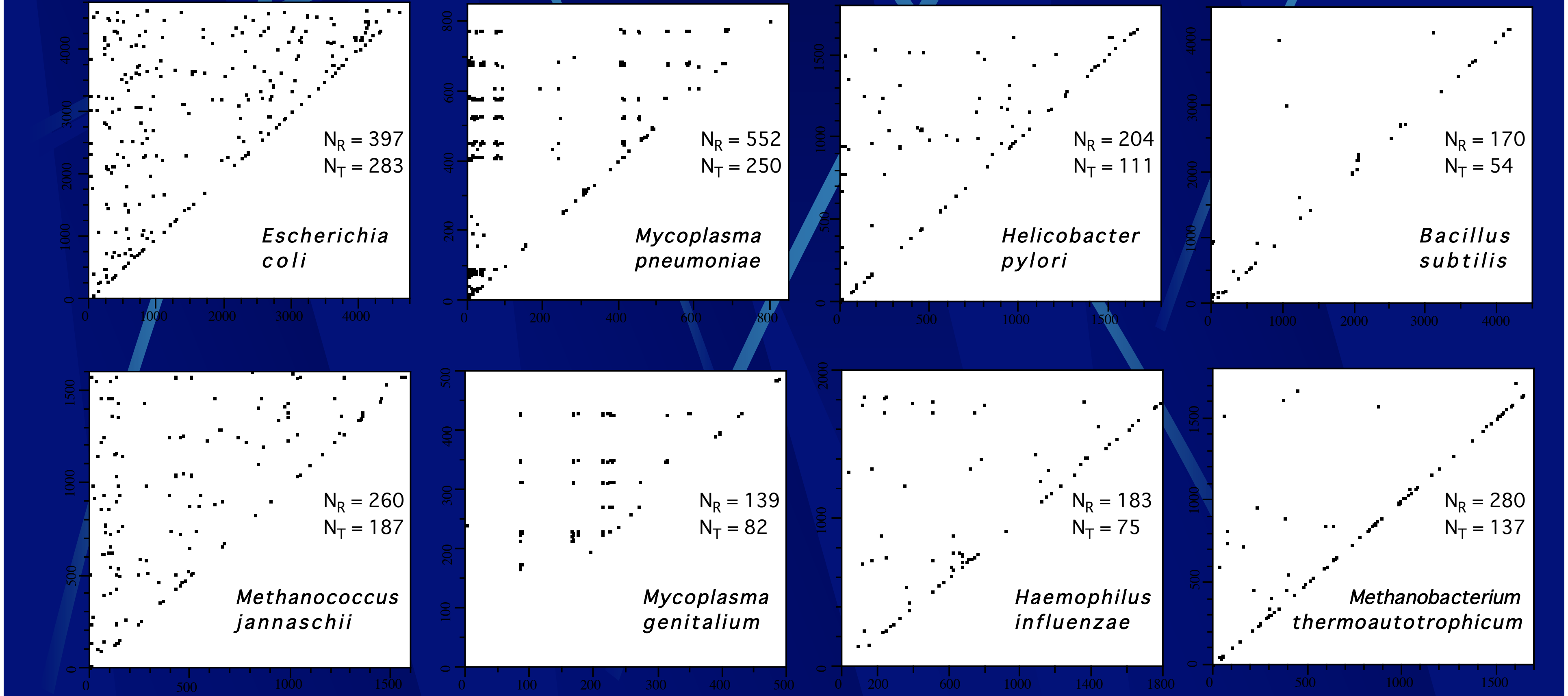
A ball.

Remember also:

This clock has a minute minute hand

香港大學－巴斯德研究中心
HKU-Pasteur Research Centre

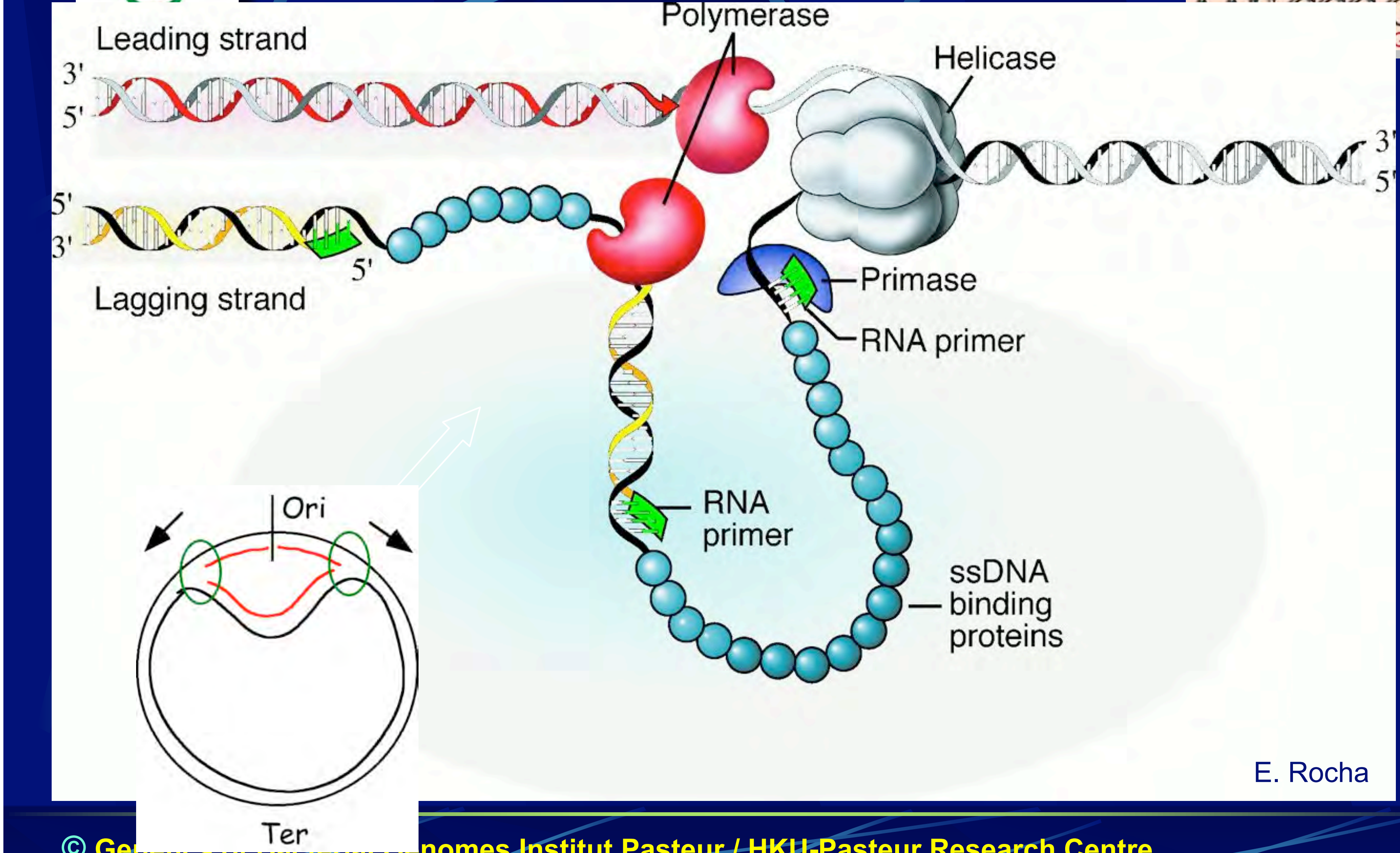
DNA management:
Repeats in genomes



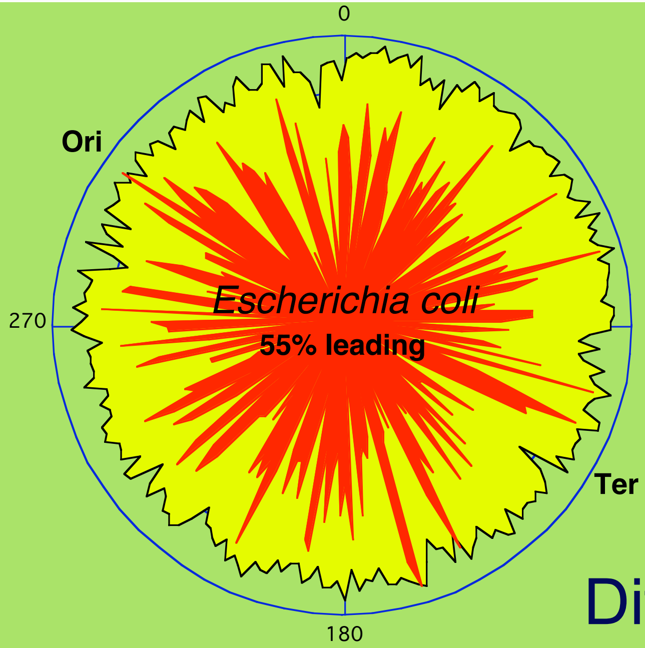
E. Rocha, A. Viari & A. Danchin Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in *Bacillus subtilis* and other competent prokaryotes. Mol. Biol. Evol. (1999) **16**: 1219-1230

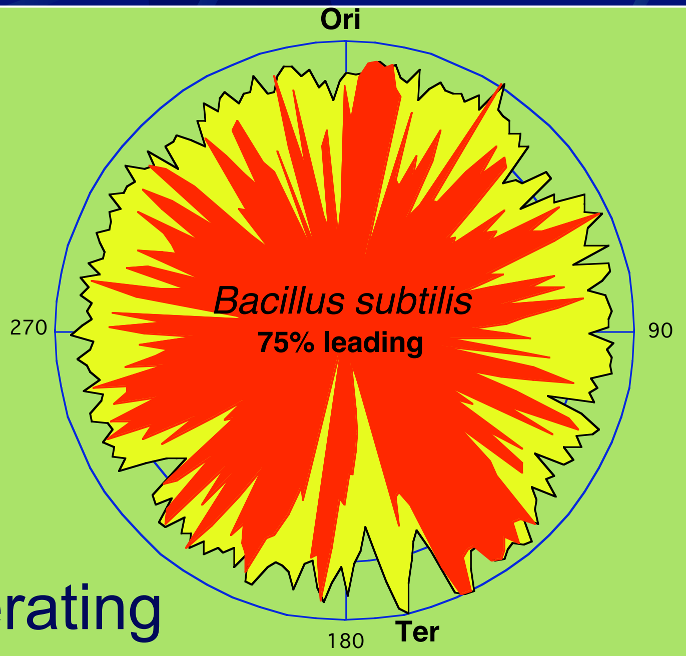
Genome organisation

The genome organisation is so rigid that the overall result of selection pressure on DNA is visible in the genome text, which differentiates the leading strand from the lagging strand

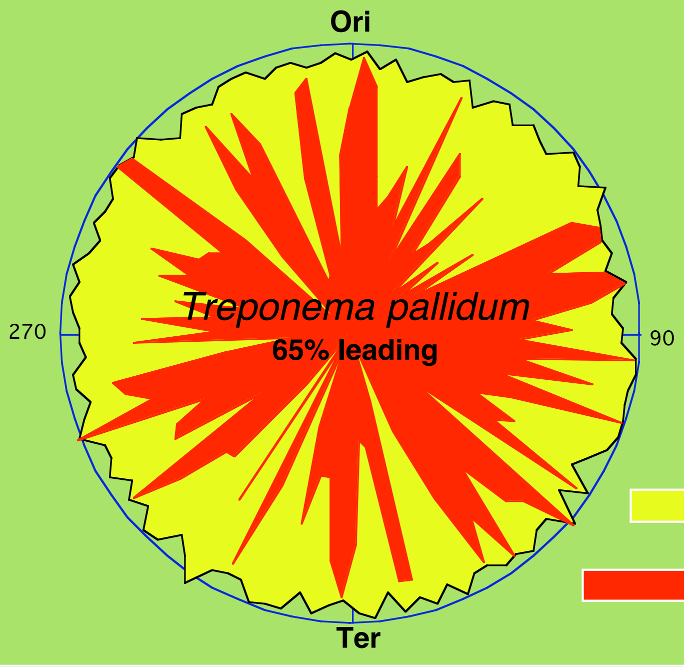

Polymerase
Leading strand
3'
5'
Helicase
3'
5'
5'
3'
5'
Lagging strand
Primase
RNA primer
RNA primer
ssDNA binding proteins
Ori
Ter

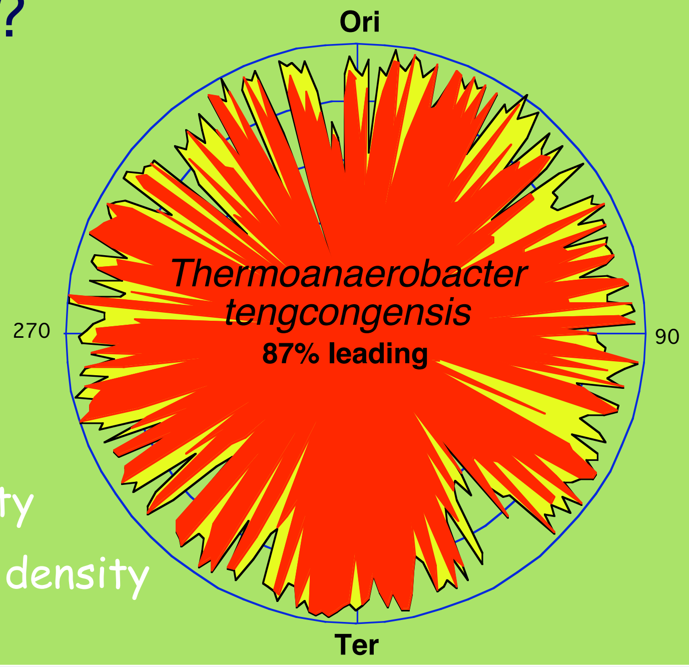
E. Rocha





Different “Operating Systems”?





🟨 CDS density

🟥 Leading CDS density

(updated from Kunst etal , Nature, 97)

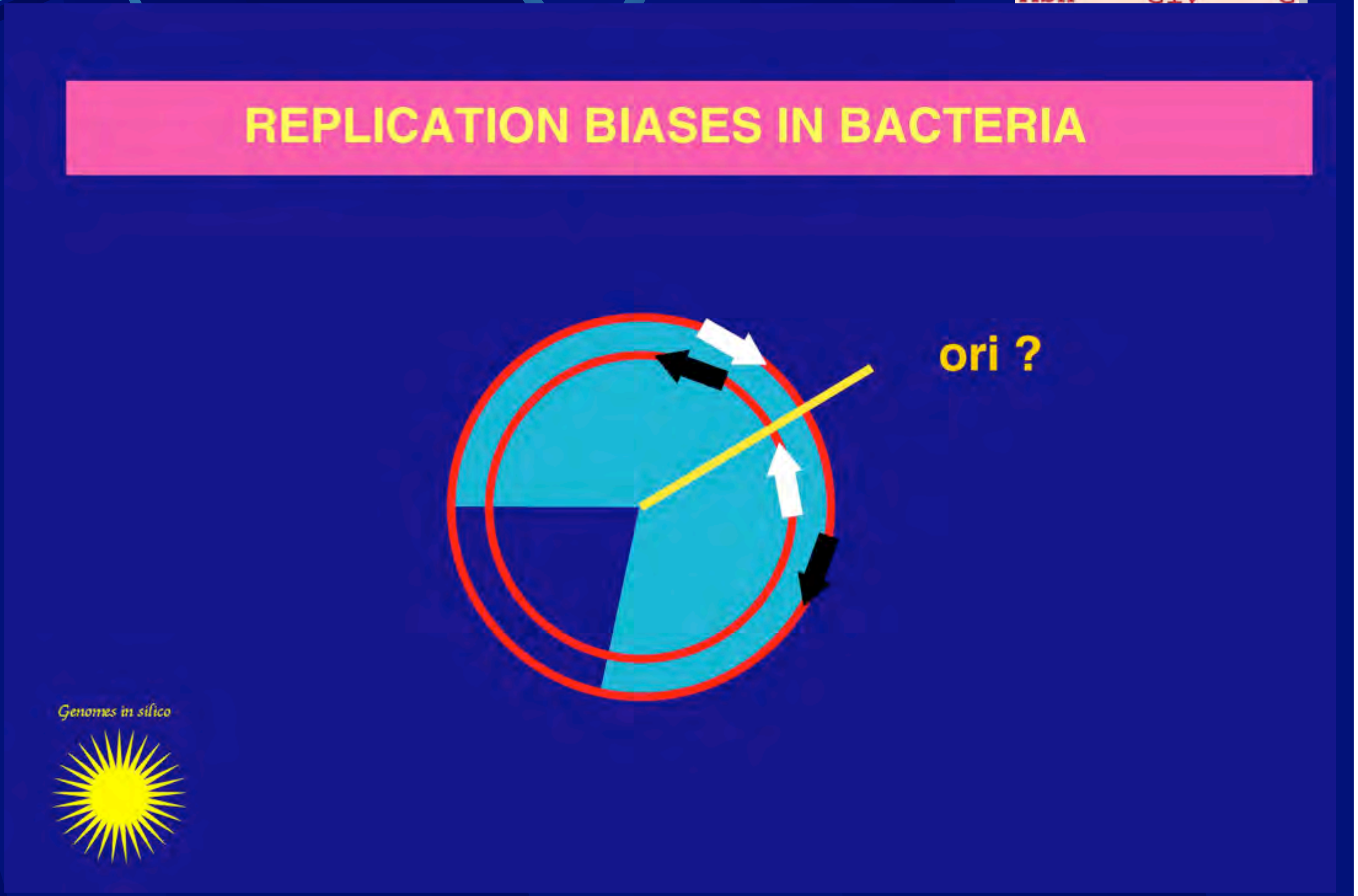
To lead or to lag...

Is it possible to see whether the position of genes in the chromosome is randomly distributed on the leading and lagging strand?

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.

To lag or to lead...

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

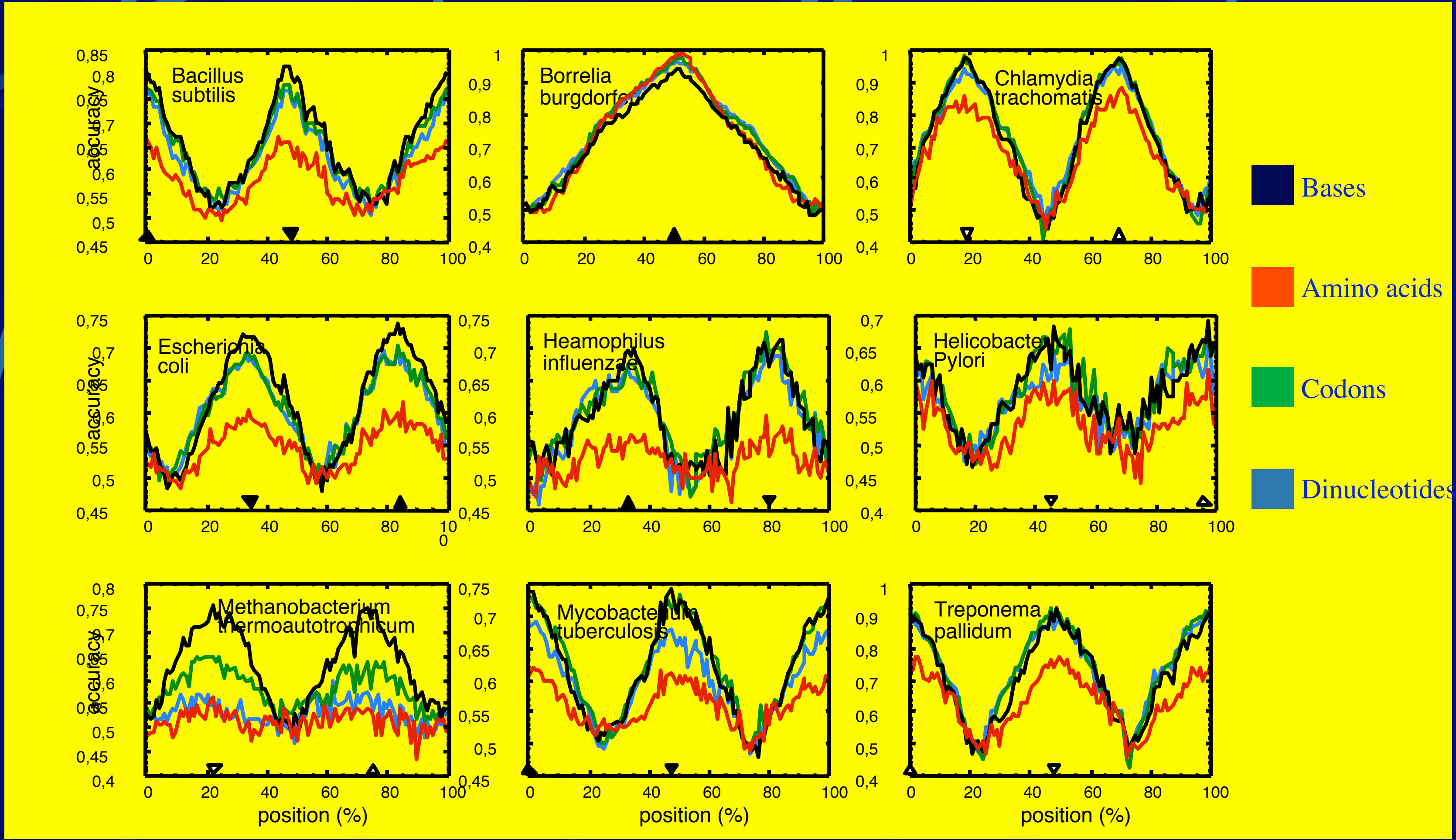


REPLICATION BIASES IN BACTERIA

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

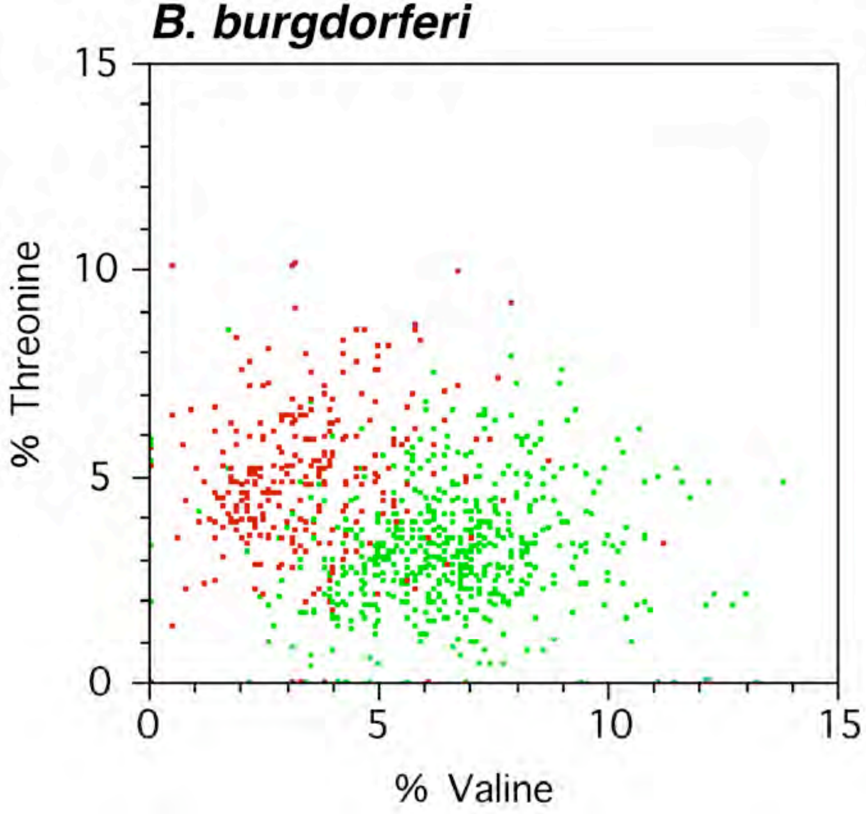
To lag or to lead, that is the question

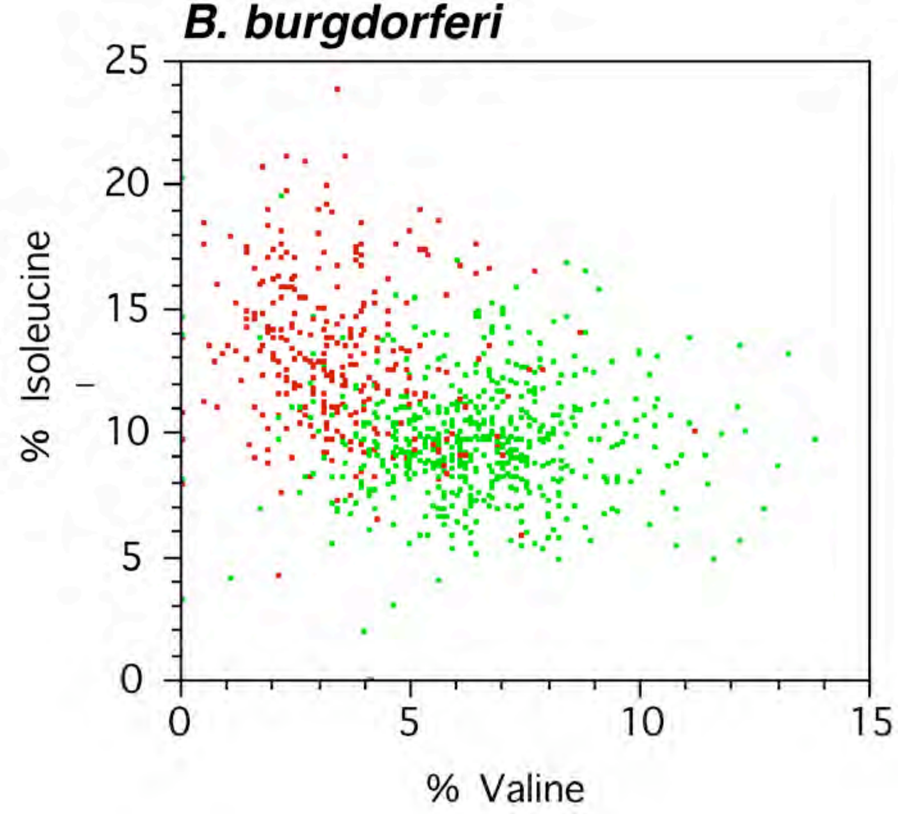
UUC.GUU.C
Phe Val Le
AAU.GGC.G
Asn Gly G



香港大學－巴斯德研究中心
HKU-Pasteur Research Centre

Visible even in proteins…



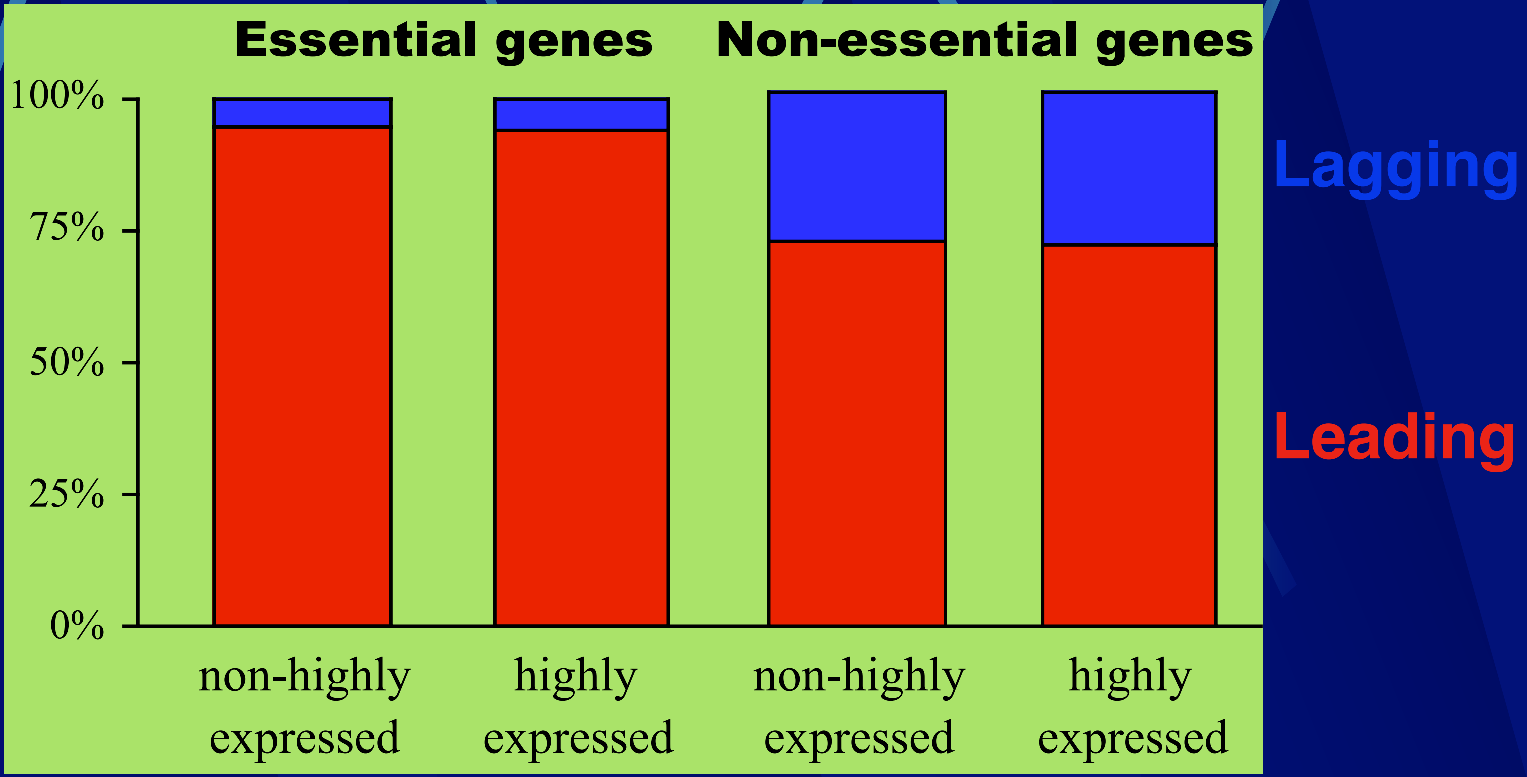


Replication transcription conflicts

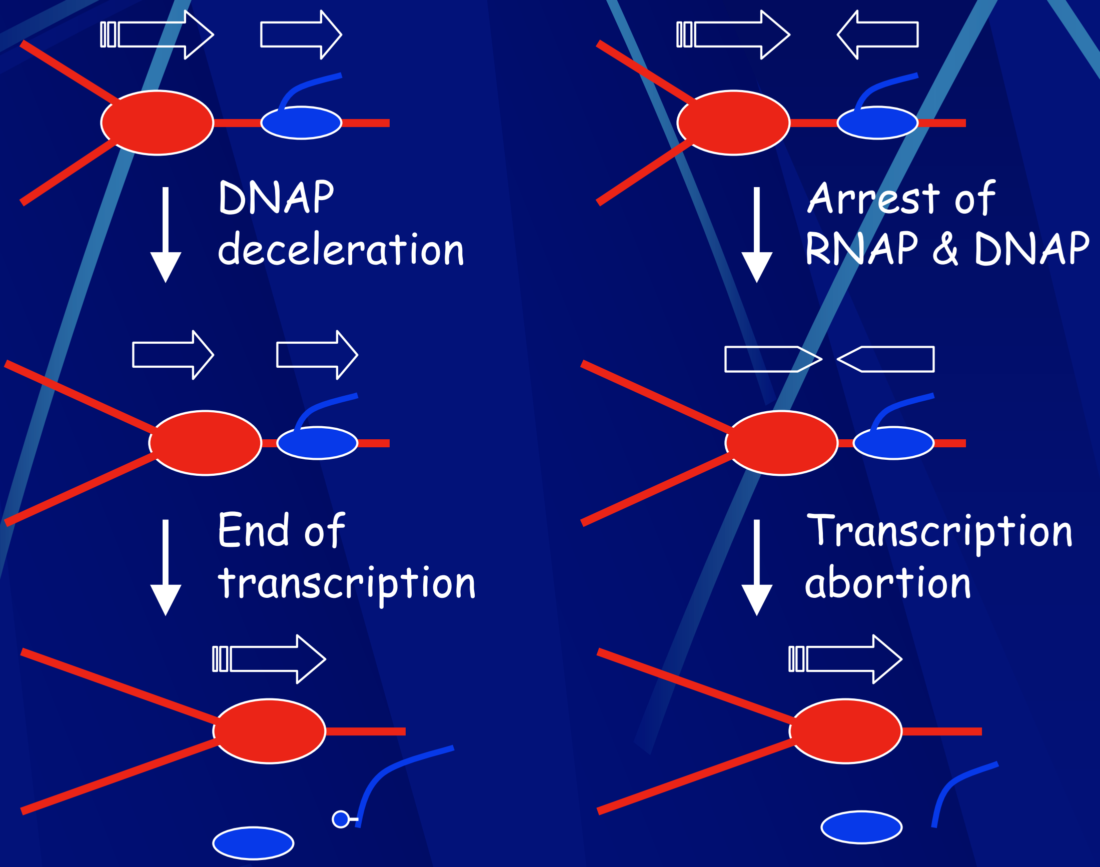
- Transcription may proceed opposite to the movement of the replication fork movement
- This will abort transcription, leading to truncated mRNA
- If translated truncated mRNA may lead to truncated proteins, this will become negative dominant if in complexes…

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

Essentiality in *B. subtilis*



When polymerases collide (II)



Co-oriented

DNAP deceleration

End of transcription

Head-on

Arrest of RNAP & DNAP

Transcription abortion

~~Consequences:~~

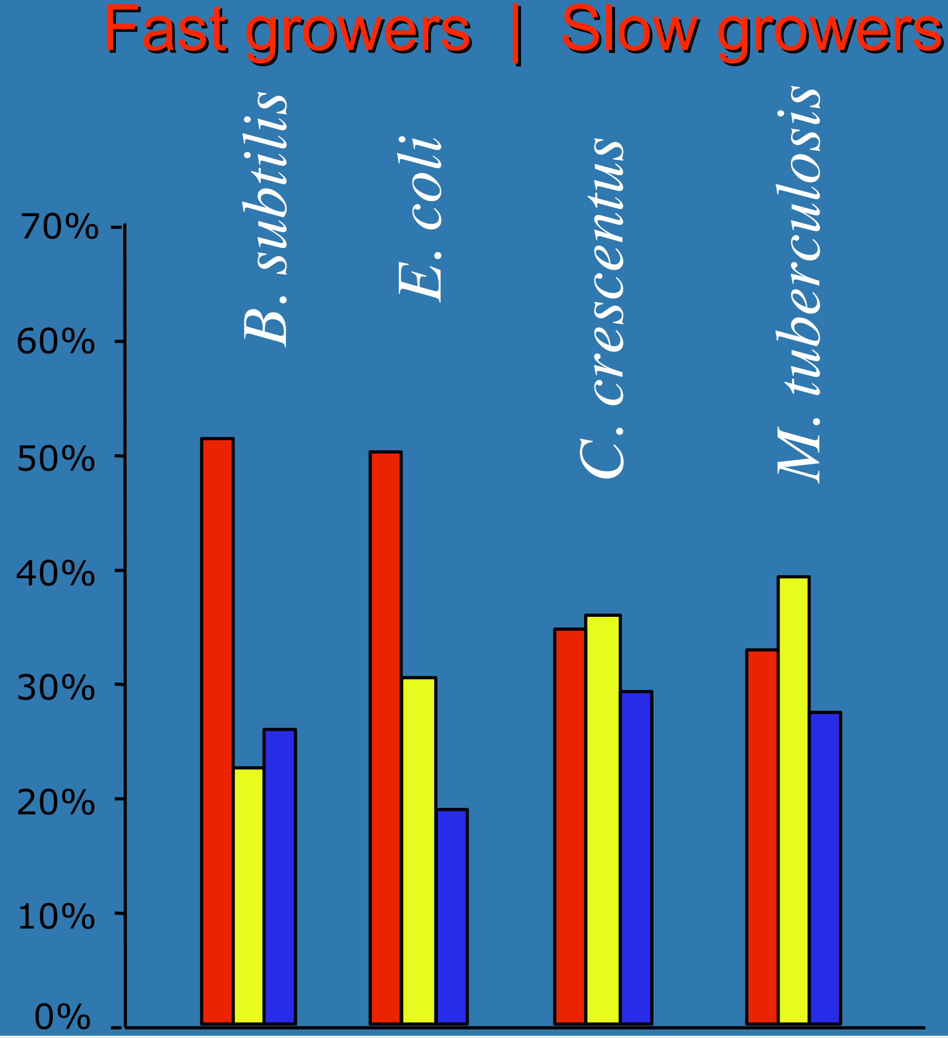
1. ~~Replication slow-down~~
2. ~~Loss of transcripts~~

Consequences:

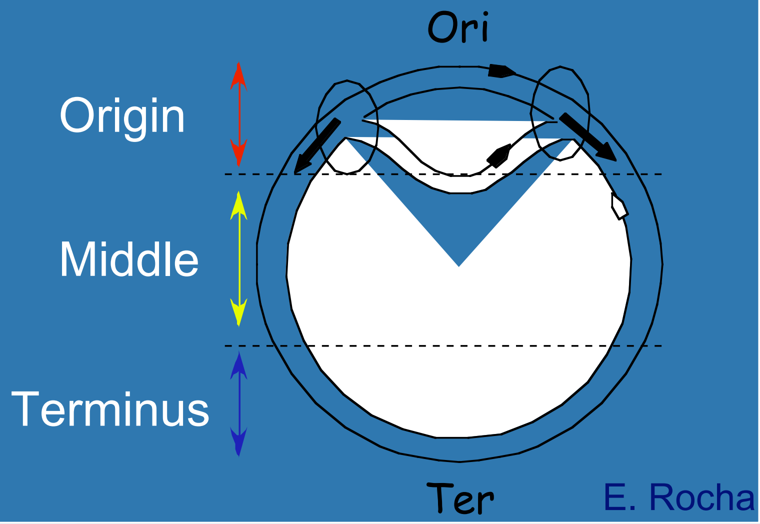
1. Aborted transcripts
2. Truncated essential proteins

E. Rocha

Distribution of highly expressed genes

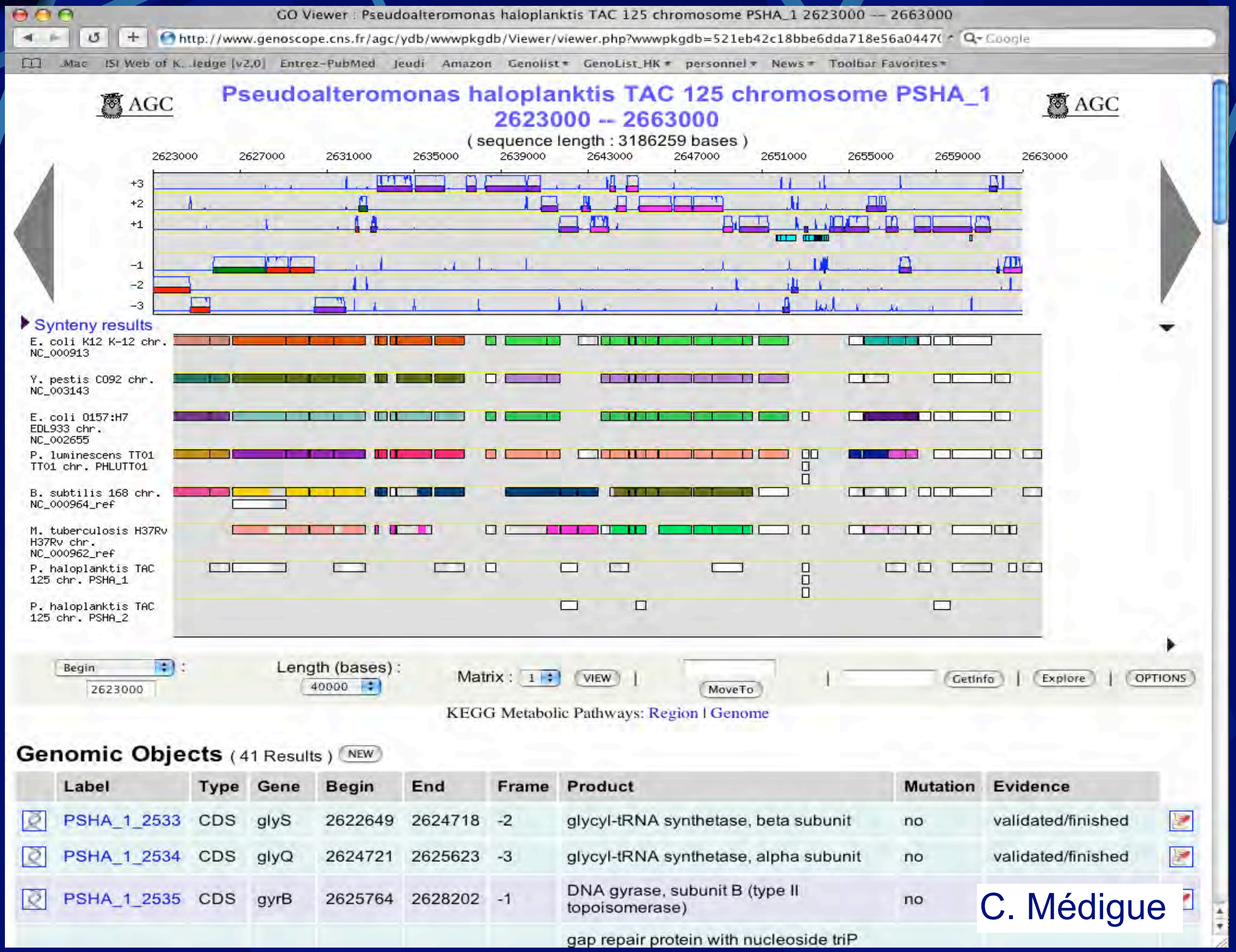


Highly expressed genes cluster near the origin in fast-growing bacteria



香港大學－巴斯德研究中心
HKU-Pasteur Research Centre

Gene vicinity: synteny


GO Viewer : Pseudoalteromonas haloplanktis TAC 125 chromosome PSHA_1 2623000 -- 2663000
http://www.genoscope.cns.fr/agc/ydb/wwwpkgdb/Viewer/viewer.php?wwwpkgdb=521eb42c18bbe6dda718e56a0447(
Google
.Mac ISI Web of K...ledge [v2,0] Entrez-PubMed Jeudi Amazon Genolist GenoList_HK personnel News Toolbar Favorites
AGC
AGC
2623000 2627000 2631000 2635000 2639000 2643000 2647000 2651000 2655000 2659000 2663000
+3
+2
+1
-1
-2
-3

Pseudoalteromonas haloplanktis TAC 125 chromosome PSHA_1
2623000 -- 2663000
(sequence length : 3186259 bases)

Synteny results

E. coli K12 K-12 chr. NC_000913

Y. pestis CO92 chr. NC_003143

E. coli O157:H7 EDL933 chr. NC_002655

P. luminescens TT01 TT01 chr. PHLUTT01

B. subtilis 168 chr. NC_000964_ref

M. tuberculosis H37Rv H37Rv chr. NC_000962_ref

P. haloplanktis TAC 125 chr. PSHA_1

P. haloplanktis TAC 125 chr. PSHA_2

Begin : 2623000

Length (bases) : 40000

Matrix : 1

KEGG Metabolic Pathways: Region | Genome

Genomic Objects (41 Results)

	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

香港大學－巴斯德研究中心
HKU-Pasteur Research Centre

Multivariate Analyses

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

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.

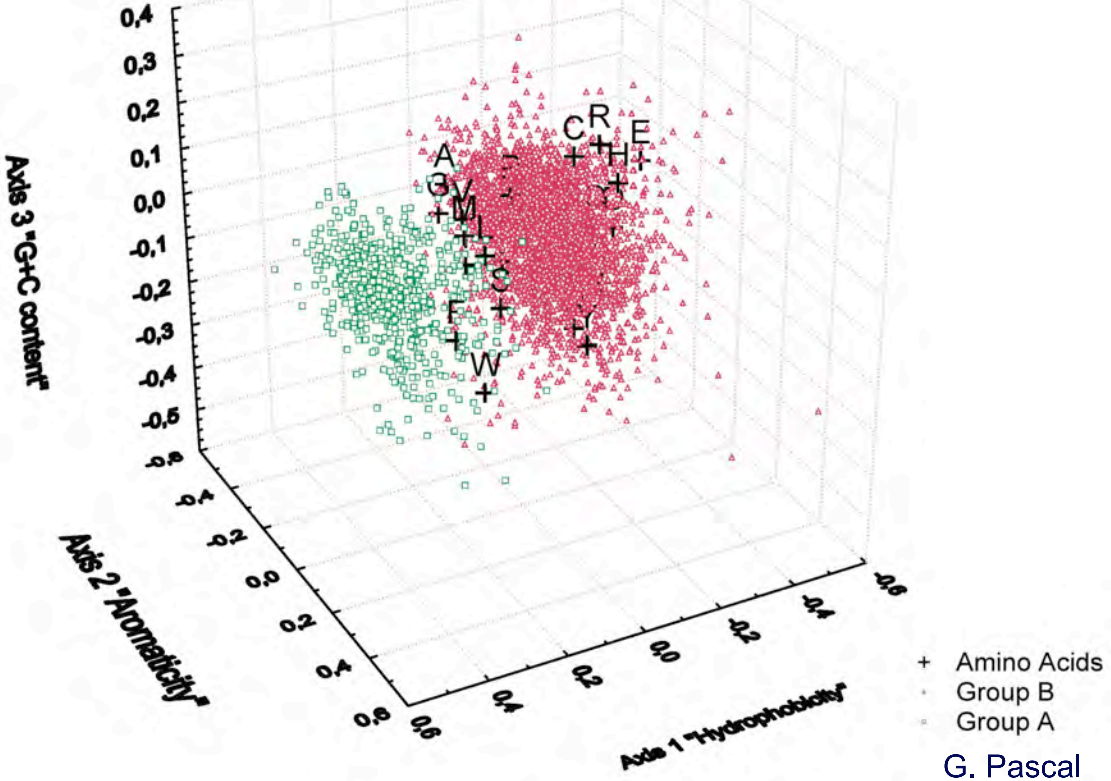
Principal Component Analysis uses the centered average and a simple distance (identity); it is the reference method.

Correspondence Analysis belongs to the same family, but it uses the χ 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

Bias in amino acid distribution

Neighborhoo: distribution of aminoacids in the proteome



G. Pascal

香港大學－巴斯德研究中心
HKU-Pasteur Research Centre

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 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

Temperature-dependent biases in protein amino acid composition

- 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

Codon usage biases

- 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

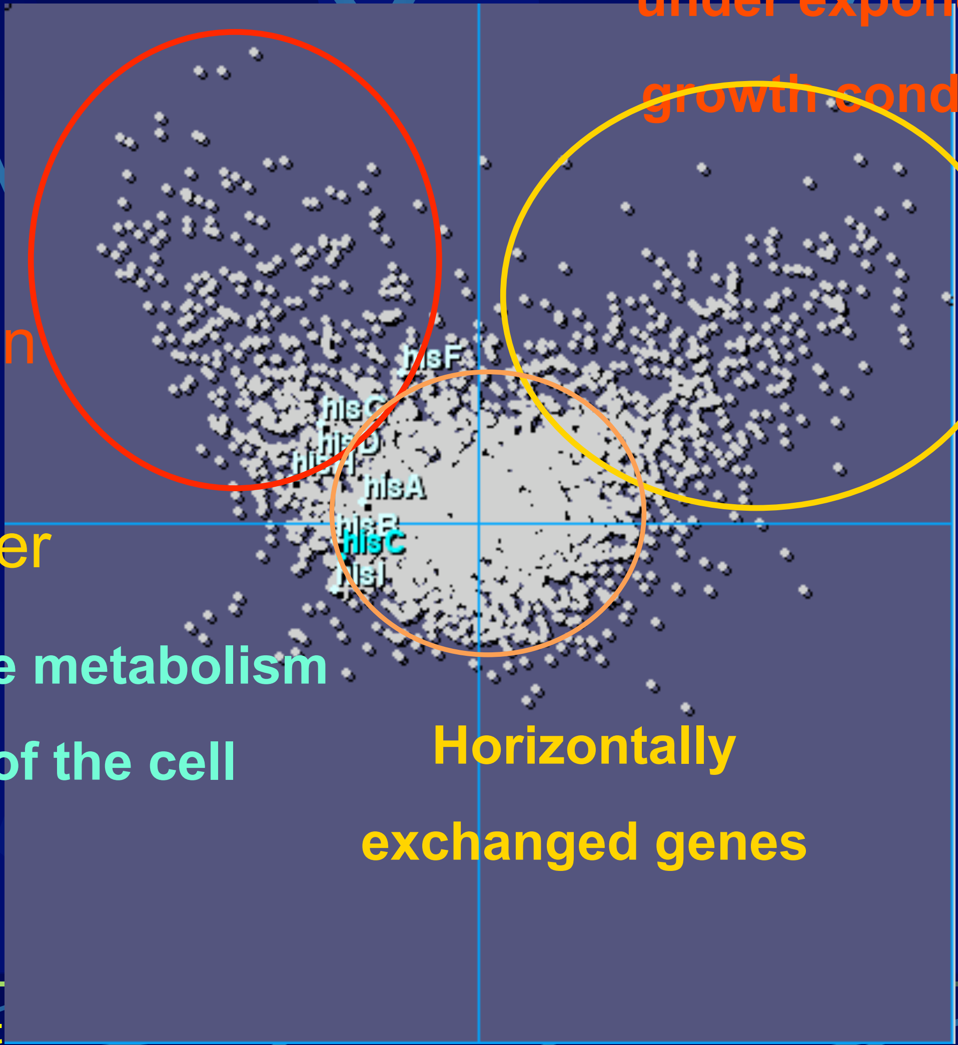
香港大學－巴斯德研究中心
HKU-Pasteur Research Centre

Gene exchange

Class I: core metabolism

Class II: high expression in exponential growth

Class III: horizontal transfer



Genes expressed at a high level under exponential growth conditions

Core metabolism of the cell

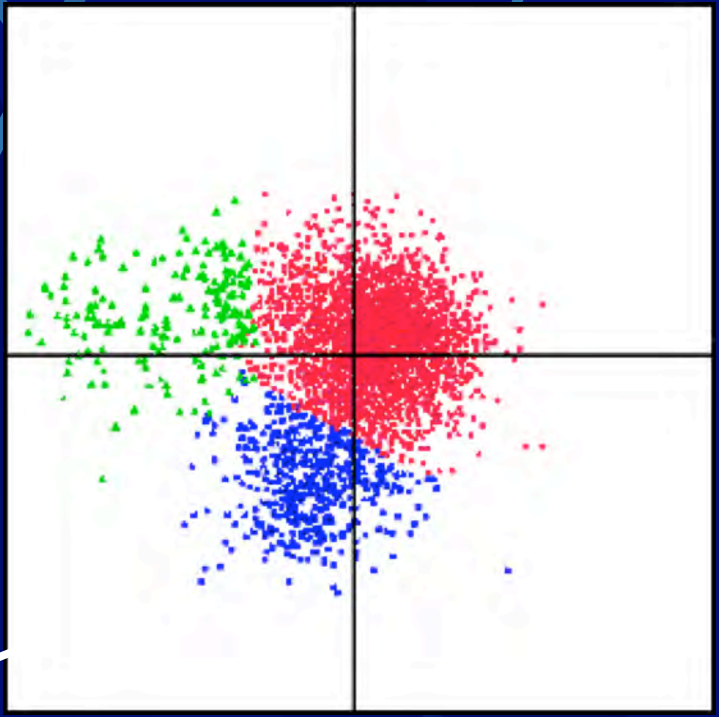
Horizontally exchanged genes

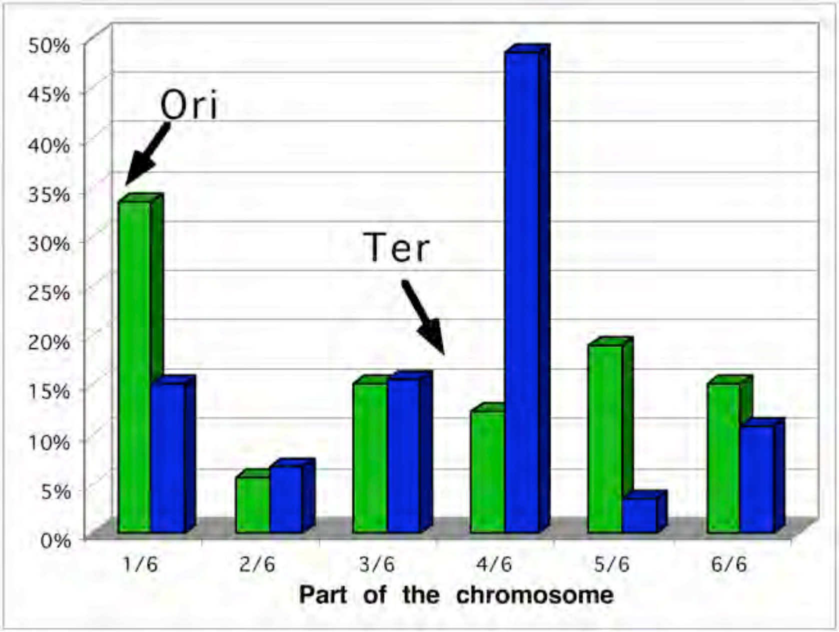
Codon usage, organisation and evolution of the *B. subtilis* genome

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

	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

香港大學－巴斯德研究中心
HKU-Pasteur Research Centre

The cell organizers

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….

Selection pressure for organisation: Oxido-reduction

- 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

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

- 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

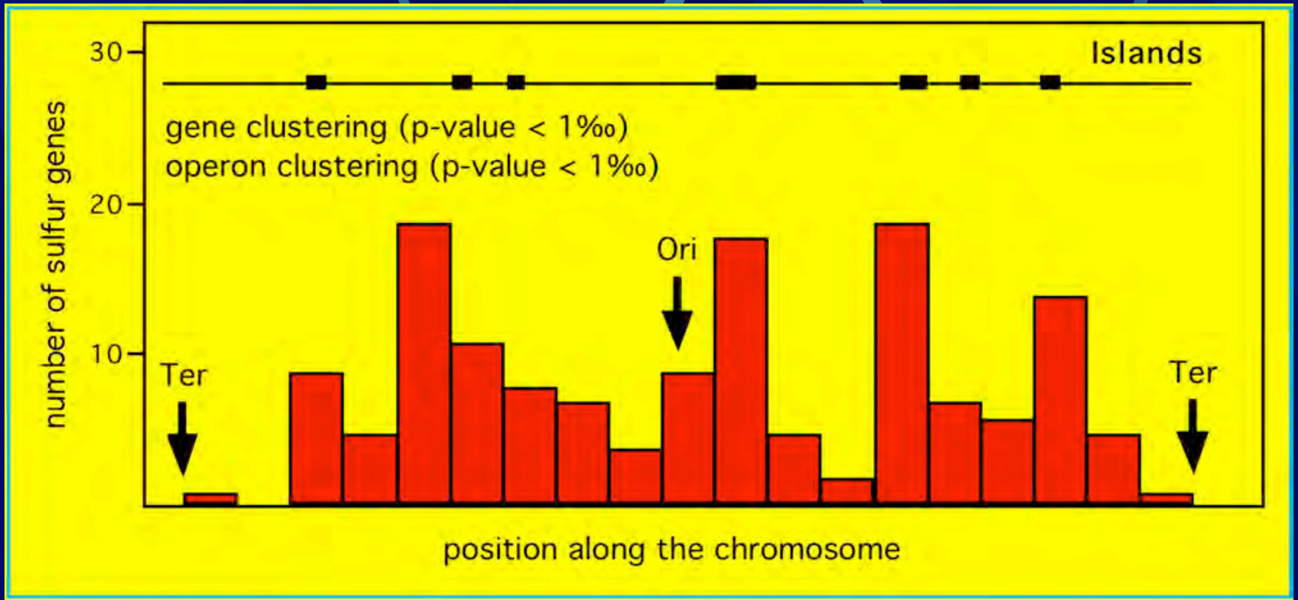
- Their genes are very poor in codons ATA, AGA and TCA

- 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

香港大學－巴斯德研究中心
HKU-Pasteur Research Centre

Proximity in the chromosome
Sulphur islands



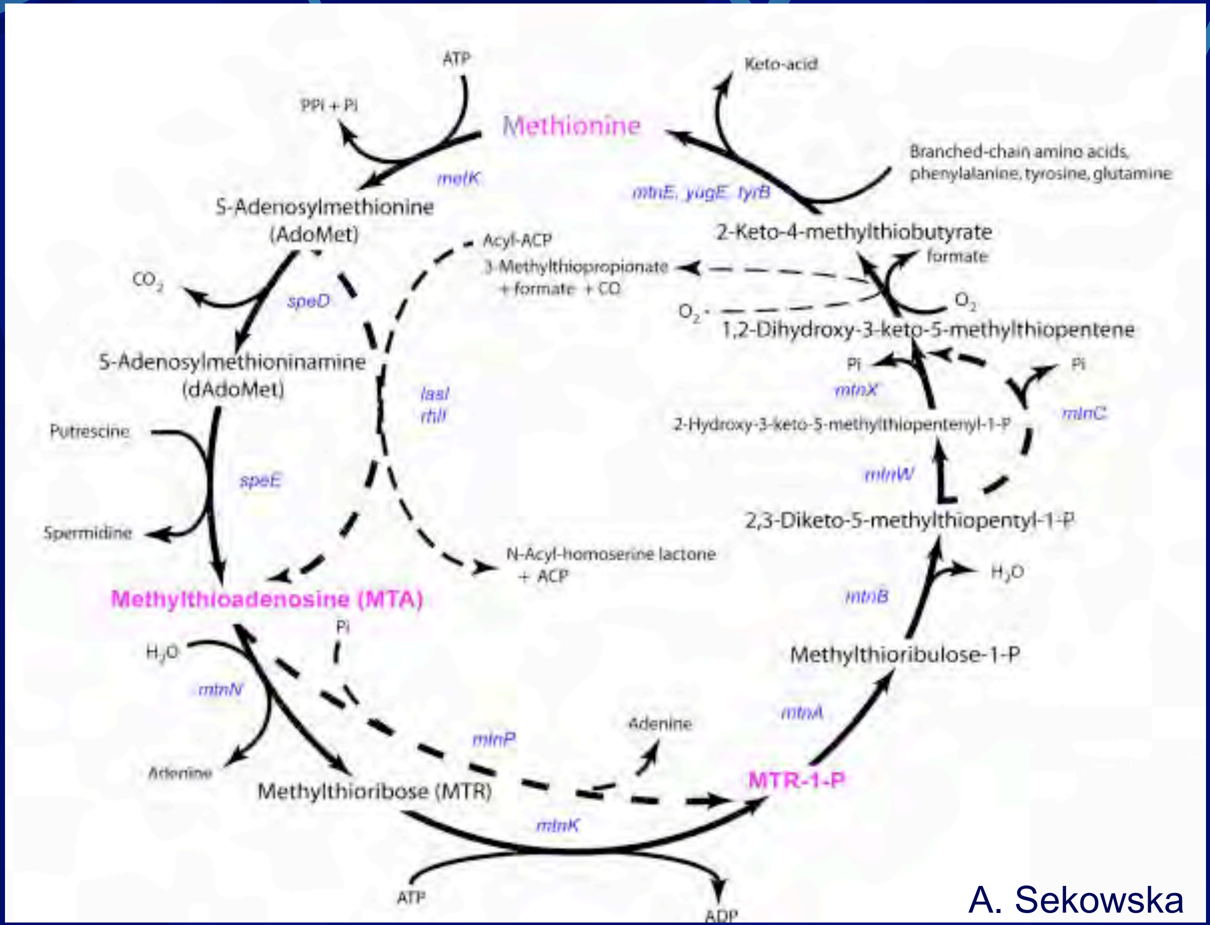
E.P.C. Rocha, A. Sekowska & A. Danchin Sulfur islands in the *Escherichia coli* genome: markers of the cell's architecture? FEBS Lett. (2000) 476: 8-11

The error catastrophe

- 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* annotated as « translation factor » is a component of sulfur metabolism!

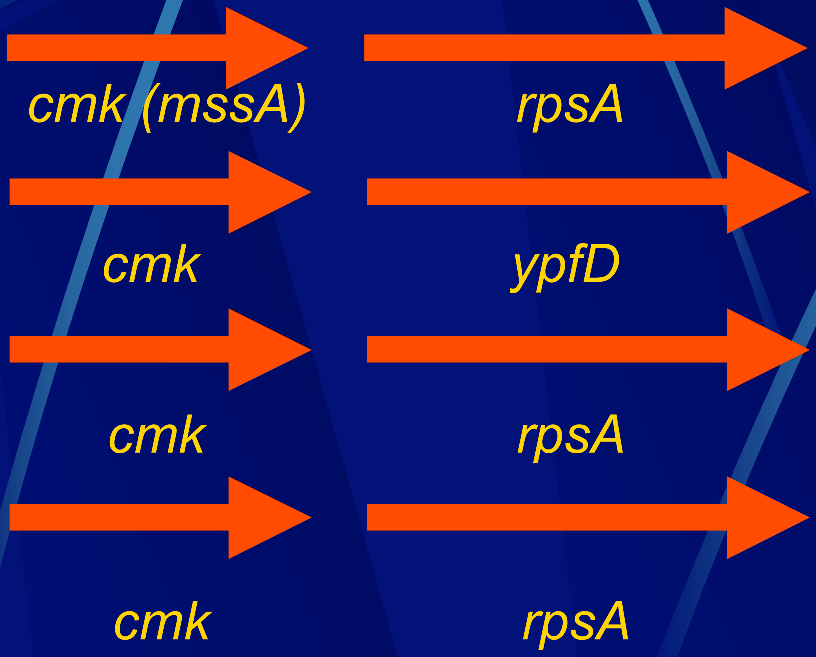
A Sekowska, V Dénervaud, H Ashida, K Michoud, D Haas, A Yokota, A Danchin Bacterial variations on the methionine salvage pathway *BMC Microbiol* (2004) **4:** 9

A new metabolic pathway



A. Sekowska

Just so story: proximity in the genome



Escherichia coli

***Bacillus subtilis* no *rpsA* !!!**

Haemophilus influenzae

Sinorhizobium meliloti

The pyrimidine diphosphate paradox

In order to make deoxyribonucleotides the cell uses ribonucleosides **di**phosphates, not **tri**phosphates

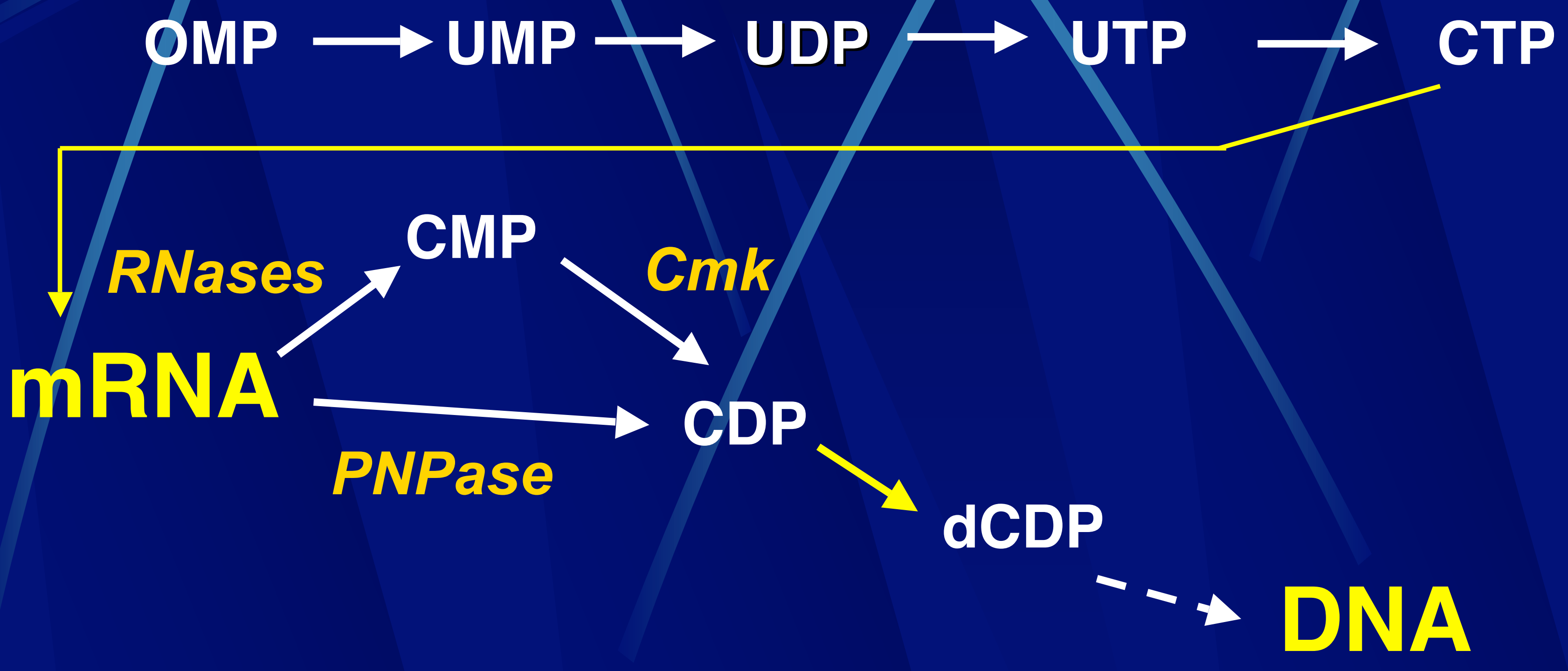
$$\mathrm{NDP} \xrightarrow[\textit{NDR}]{} \mathrm{dNDP} \xrightarrow[\textit{NDK}]{} \mathrm{dNTP}$$

And here is the paradox:

$$\mathrm{OMP} \longrightarrow \mathrm{UMP} \longrightarrow \mathrm{UDP} \longrightarrow \mathrm{UTP} \longrightarrow \mathrm{CTP}$$

no CDP !!!

How is the paradox resolved?



OMP → UMP → UDP → UTP → CTP

mRNA

RNases

CMP

Cmk

PNPase

CDP

dCDP

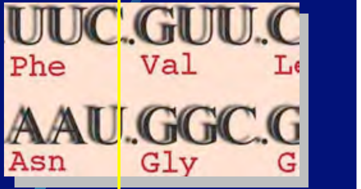
DNA

Phylogenetic neighbors: the S1 box

- *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…

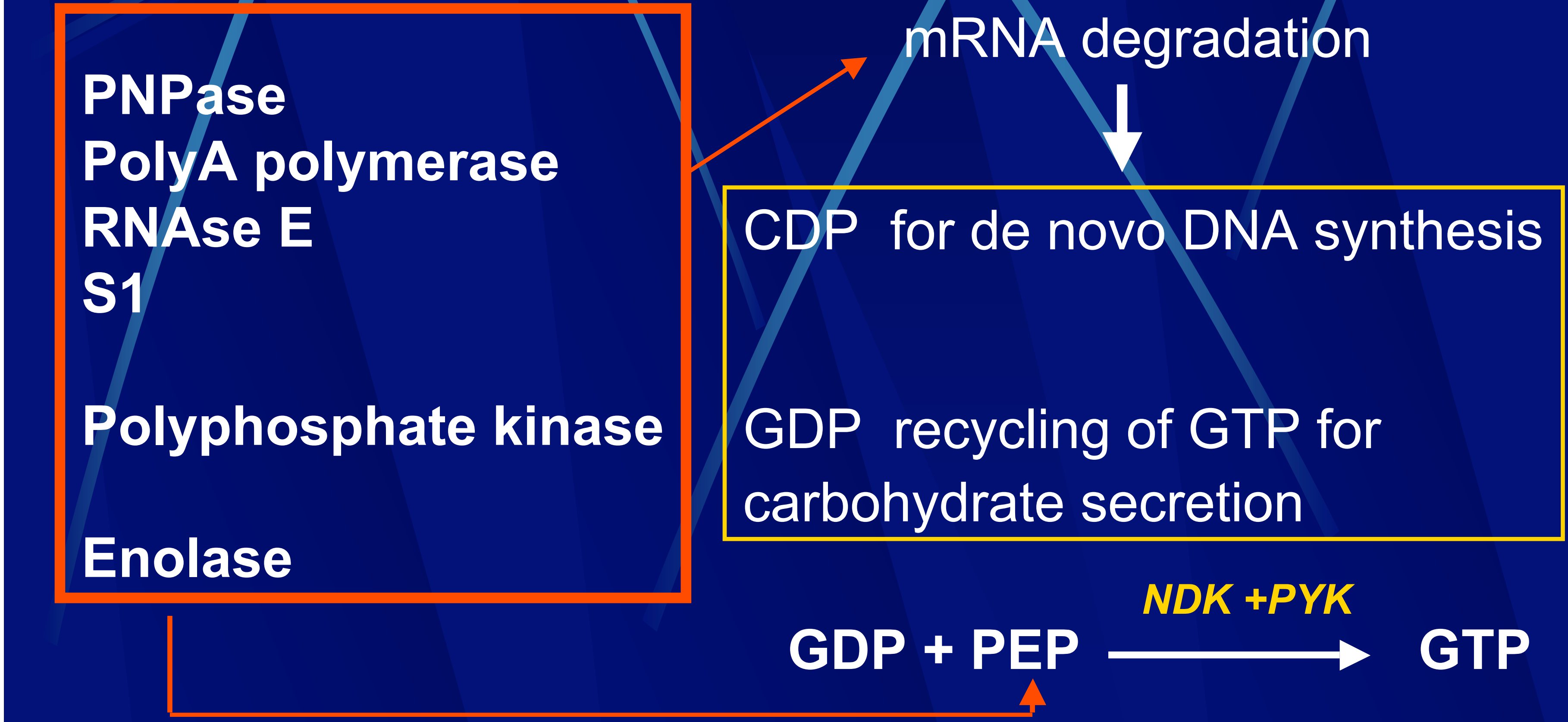
=> All this points to a function for S1 in RNA metabolism

Codon usage bias neighbors



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

Protein complexes: the Degradosome



PNPase
PolyA polymerase
RNAse E
S1

Polyphosphate kinase

Enolase

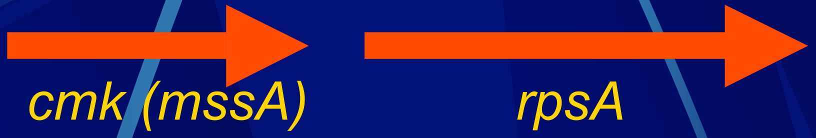
mRNA degradation

CDP for de novo DNA synthesis

GDP recycling of GTP for carbohydrate secretion

$$\text{GDP + PEP} \xrightarrow{\textit{NDK +PYK}} \text{GTP}$$

Just so story: the *cmk rpsA* operon



cmk (mssA)

rpsA

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.

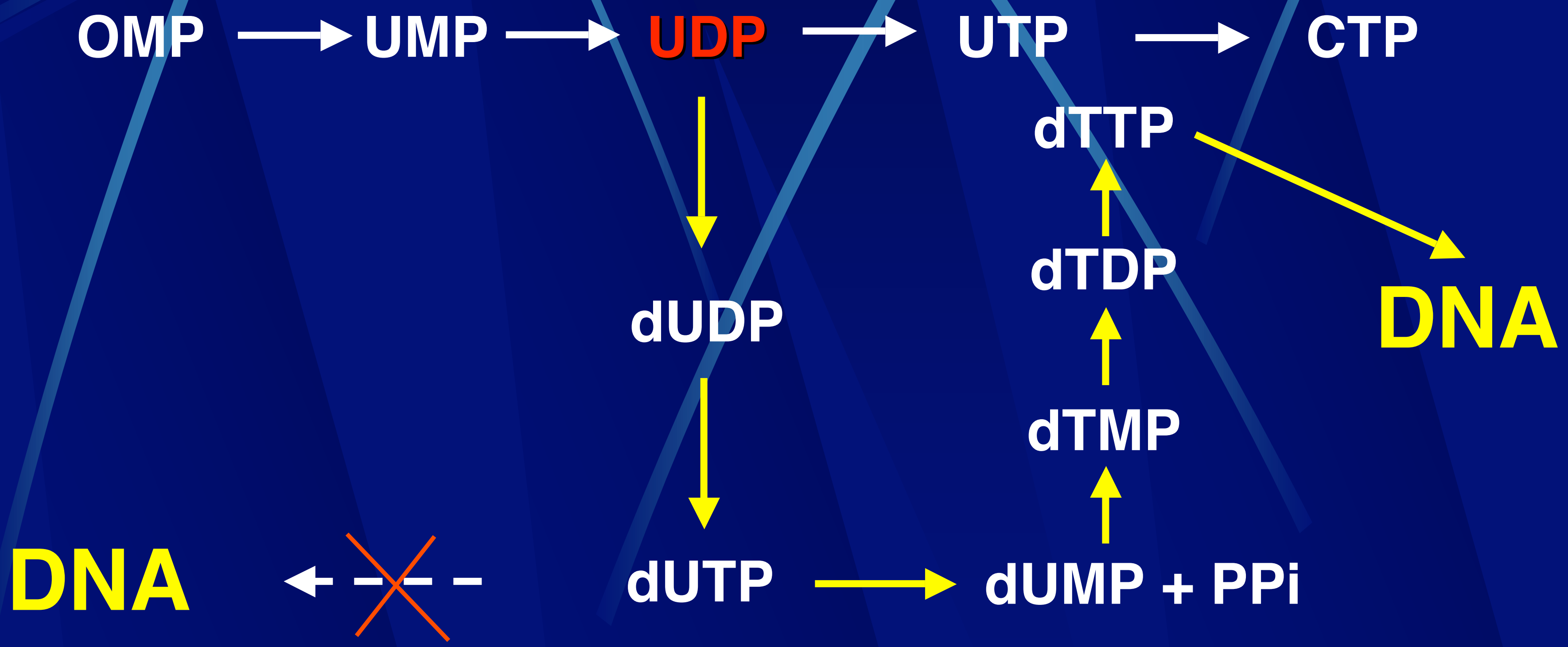
Conclusion:

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

Selection pressure for compartmentalization: a dangerous intermediate

Uridylate kinase (UMK)
pyrH (smbA)

No CDP: no DNA…

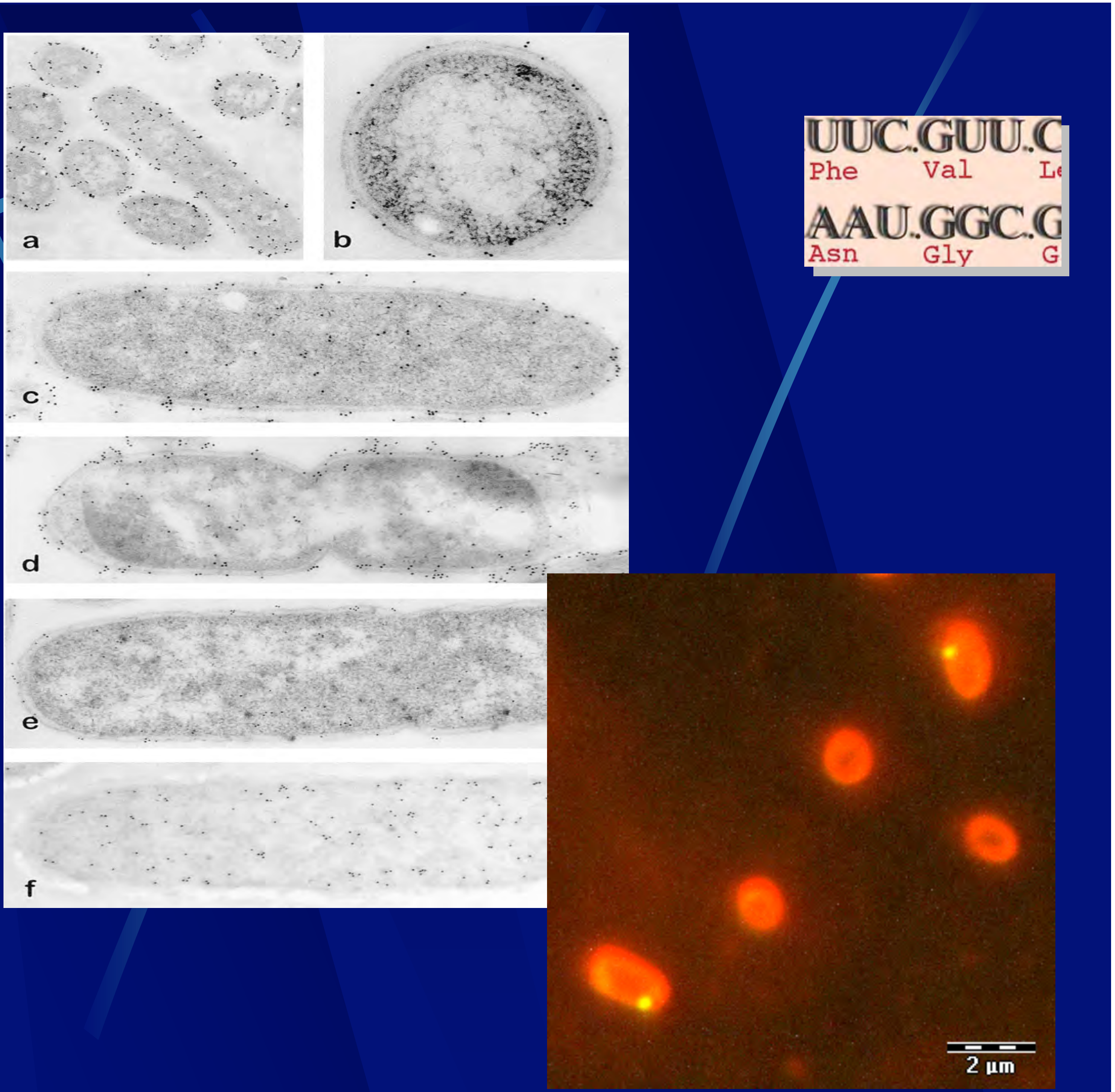

OMP → UMP → UDP → UTP → CTP

UDP → dUDP → dUTP → 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

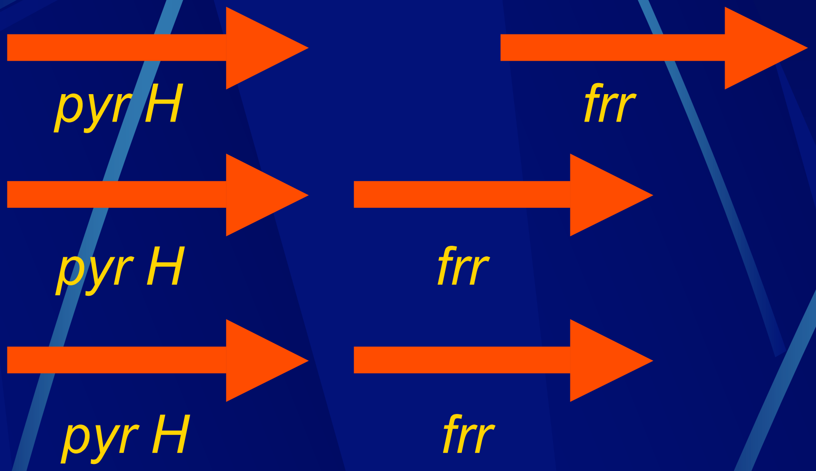
In conclusion:



UMK must be compartmentalized

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

A prediction: ribosome recycling and UTP



Escherichia coli

Bacillus subtilis

Photorhabdus luminescens

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

Ribosome recycling and UTP

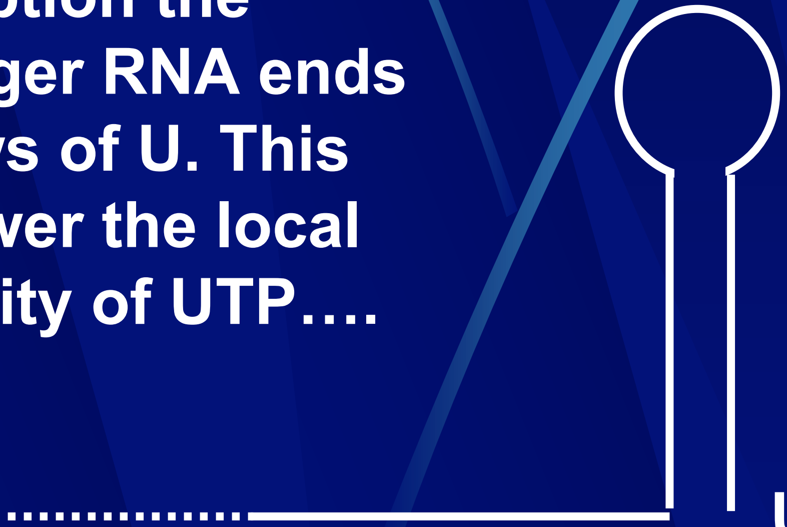
UUC.GUU.C
Phe Val Le
AAU.GGC.G
Asn Gly G

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...

Transcription termination

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



UUUUUUUUUU

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