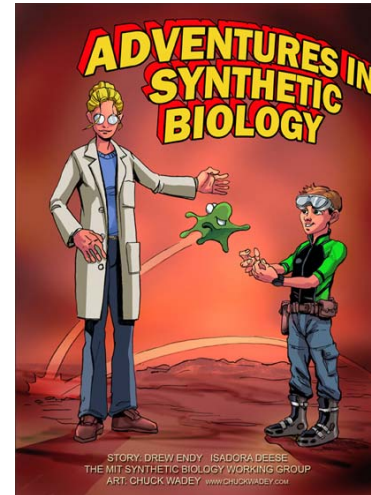


Genomic functional cores specific to different microbes and detection of essential metabolic pathways

Alessandra Carbone
Génomique Analytique
Université Pierre et Marie Curie, Paris

carbone@ihes.fr

1st step



Genome synthesis

2nd step



Genome programming

Craig Venter, November 2002

Synthesis of a bacterial genome

the chromosome will be inserted in a living cell (whose genetic material has been removed) to verify if it can direct normal functional activities of the organism.

Clyde Hutchison, 1999 (*Science* 286, 2165-2169):

Gene knock out (517) of *Mycoplasma genitalium* (580kb), and estimation of how many genes are necessary to life over 517: about 300 to survive.

Eckard Wimmer, 2002 (*Science* 297, 1016-1018):

Synthesis of a poliovirus that infects cells! (~7500b)

Search for a minimal genome

Why to do this :

Add genes to transform *Mycoplasma* in a "useful" bacteria

Remedy against environmental pollution, new industrial chemical substances production, insuline production...

To search for a minimal set is not easy...

Experiments : transposomal mutagenesis

<i>B.subtilis</i> 300 genes/~4000 (Itaya, 1995)	<i>M.genitalium</i> 265 genes / 482 (Hutchison et al., 1999) 382 genes / 482 (Hutchison et al., 2006)	<i>H.influenzae</i> 670 genes/ ~1272 (Akerley et al. 2002)	<i>E.coli</i> 620 genes / 3746 (Gerdes et al. 2003) 234 genes / 2994 (Hashimoto et al. 2005)
<i>S.cerevisiae</i> 1105 genes/ 5916 (Giaever et al. 2002)	<i>C.elegans</i> 1722 genes/ 19427 (Kamath et al. 2003)	<i>S.aureus</i> 150 genes (Yi et al. 2001)	<i>S.pneumoniae</i> 110 genes (Thanassi et al. 2002)

Comparative genomics

2 genomes 256 genes (Mushegian & Koonin 1996)	34 genomes 80 genes (Harris et al 2003)	100 genomes 60 genes (Koonin et al. 2003)	147 genomes 35 genes (Charlebois & Doolittle 2004)
---	---	---	--

Number of genes in the minimal set depends on

Experiments:

- life/environmental conditions of the organism during the experiment

Computational detection of sequence homology:

- parameters and tools to detect homologies

Genes relevant to environmental conditions are missing
Stress response genes are missing
Genes with uncharacterized functions are missing



$$g = [x_{1,g} \ x_{2,g} \ \dots \ x_{64,g}] \quad x_{i,g} \text{ relative frequency of codon } i \text{ in } g$$

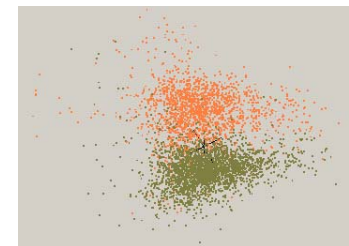
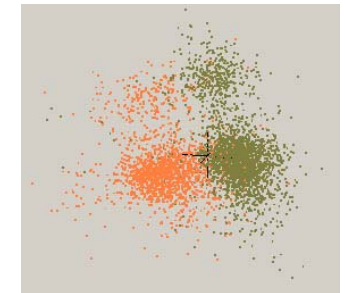
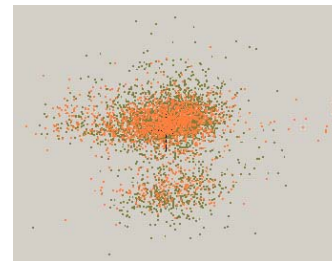
Normalisation:

$$(x_{i,g} - \underline{x}_i) / \sigma_i \quad \begin{array}{l} \underline{x}_i \text{ mean of frequencies } x_{i,g} \\ \sigma_i \text{ standard deviation of } x_{i,g} \end{array}$$

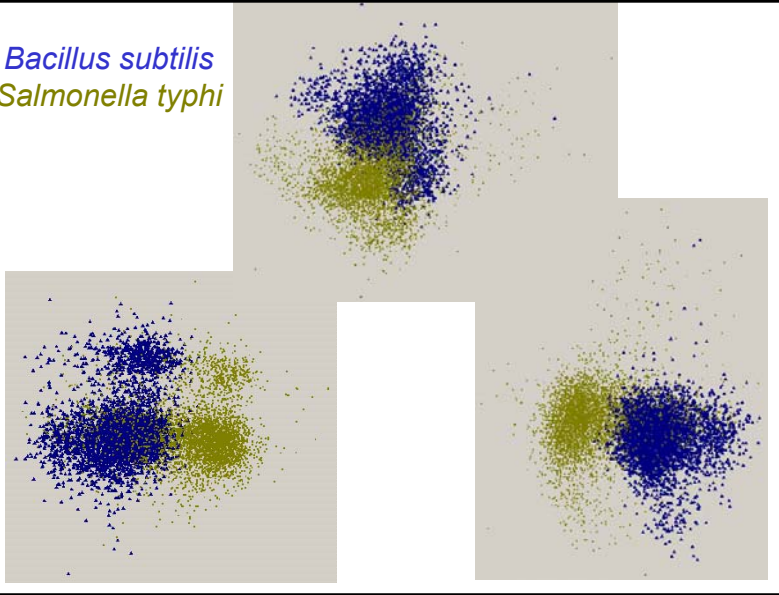
...we use normalized vectors and PCA to "see"

- organisms in codon space
- genes and functions

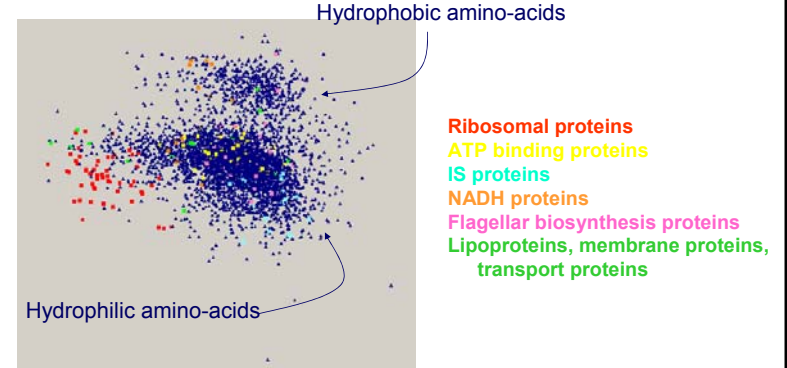
Haemophilus influenzae
Staphylococcus aureus



Bacillus subtilis
Salmonella typhi



E. coli



Some background:



		2nd base in codon				
		U	C	A	G	
1st base in codon	U	Phe Phe Leu	Ser Ser Ser	Tyr Tyr STOP	Cys Cys STOP	U C A G
	C	Leu Leu Leu	Pro Pro Pro	His Gln Gln	Arg Arg Arg	U C A G
	A	Ile Ile Met	Thr Thr Met	Aun Lys Lys	Ser Arg Arg	U C A G
	G	Val Val Val	Ala Ala Ala	Asp Asp Glu	Gly Gly Gly	U C A G

Preferential codons: codons that appear with higher frequency in most genes

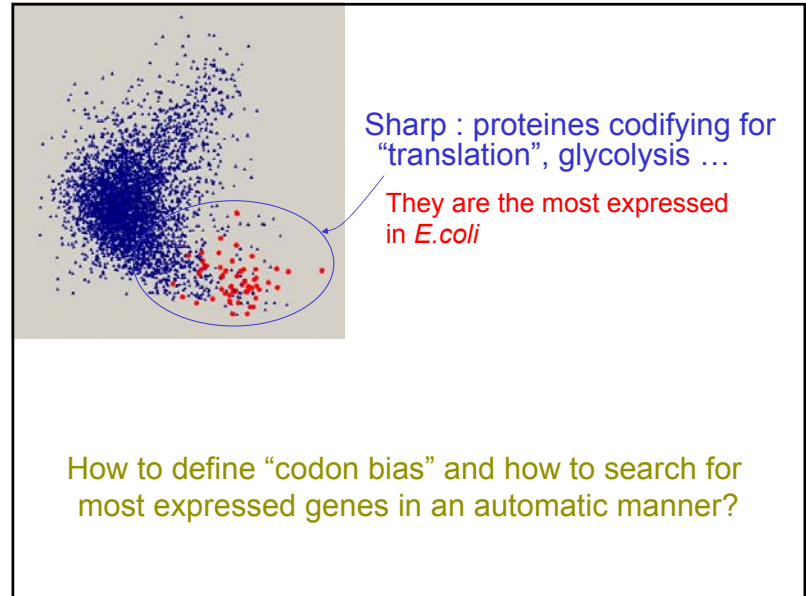
Correlation between high tRNA number and preferred codons

In 1987 P.Sharp states a new hypothesis on genomes

Organisms that reproduce rapidly have a set of genes :

- 1% of genes in the genome
- genes coded by preferential codons
- genes necessary to translation, genes that need to be translated fast and be present in the cell in large quantities at a given moment.

“Biased” set of genes



How to define “codon bias” and how to search for most expressed genes in an automatic manner?

An automatic detection of most biased genes

Let S be a set of genes and g be a gene

$$\text{CAI}(g) = (\prod_{k=1}^L w_k)^{1/L} \quad (\text{Sharp \& Li, 1987})$$

Codon Adaptation Index

$$\text{SCCI}(g) = (\prod_{k=1}^L w_k)^{1/L}$$

Self Consistent Codon Index

L number of codons in g

$w_k = \frac{|S_k|}{|S|}$ frequency of the k^{th} codon of g in S
frequency of the dominant synonymous codon in S

We look for S **automatically** in such a way that

1. S contains 1% of genes in the genome
2. SCCI values on genes in S are **maximal**

$\text{SCCI}(G/S) \leq \text{SCCI}(S)$
where G is the set of all genes

Self consistency
condition

3. S is **representative** of preferred codons

c_1, \dots, c_{20} preferred codons of S

d_1, \dots, d_{20} preferred codons of G

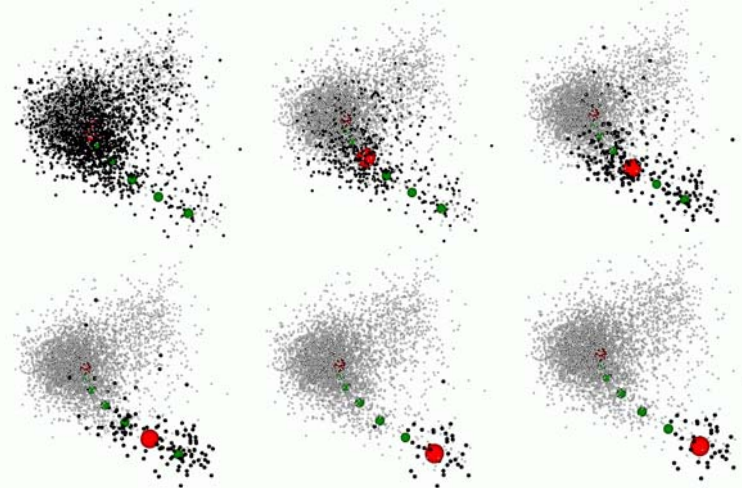
we look for the set S such that

$$\sum_{i=1}^{20} \chi(c_i, d_i) \text{ is minimal}$$

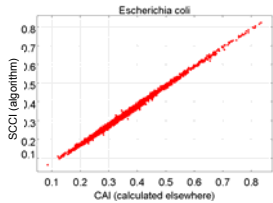
Idea of the algorithm:

- Compute the weight of the codons over the whole genome and compute afterwards SCCI values for all genes
- Select the 50% of genes with the highest SCCI value
- Repeat the iteration and select the 25% of the genes
- and so on... until we arrive to the 1% of genes in the original set.
- ... then repeat the iteration on the 1% of genes with highest SCCI until convergence is reached.

Behavior of the algorithm for *E.coli*

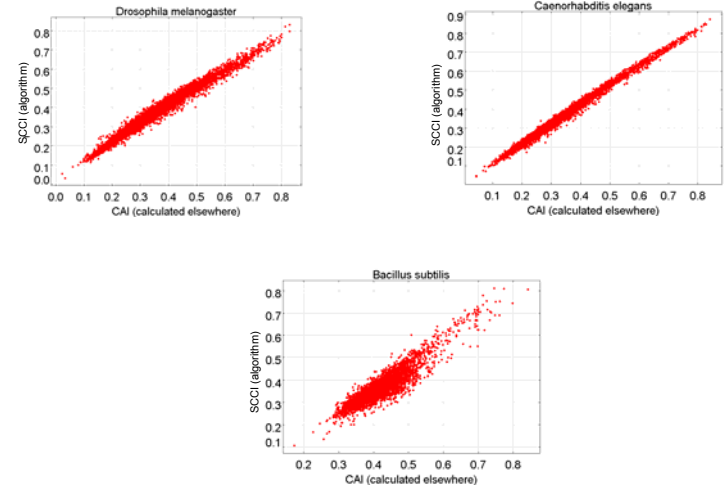


S found by the algorithm:
E.coli



Gene	Annotation
tufA	protein chain elongation factor EF-Tu
tufB	protein chain elongation factor EF-Tu
tsf	protein chain elongation factor EF-Ts
fusA	GTP-binding protein chain elongation factor EF-G
mopA	chaperonin GroEL
clxBK	heat shock protein DnaK
espA	cold shock protein 7.4
tig	trigger factor
ompA	outer membrane protein
ompX	outer membrane protein
ompC	outer membrane protein
lpp	murcin lipoprotein
pal	peptidoglycan-associated lipoprotein
yaiU	putative flagellin structural protein
yfD	putative formate acetyltransferase
eno	diadenosine tetraphosphatase
tpiA	triosephosphate isomerase
pgk	phosphoglycerate kinase
gapA	glyceraldehyde-3-phosphate dehydrogenase A
iba	fructose-bisphosphate aldolase class II
pykF	pyruvate kinase I
pfkB	formate acetyltransferase 1
ahpC	alkyl hydroperoxide reductase C22 subunit
sodA	superoxide dismutase SodA
tktA	transketolase 1/2 isozyme
rpoC	RNA polymerase beta prime subunit
rpsL	30S ribosomal subunit protein S9
rpsA	30S ribosomal subunit protein S1
rpsB	30S ribosomal subunit protein S2
rpsC	30S ribosomal subunit protein S3
rpsU	30S ribosomal subunit protein S21
rpLA	50S ribosomal subunit protein L1
rpLY	50S ribosomal subunit protein L25
rpH	50S ribosomal subunit protein L9
rpL	50S ribosomal subunit protein L7/L12
rpIC	50S ribosomal subunit protein L3
rpME	50S ribosomal subunit protein L31
rpLB	50S ribosomal subunit protein L2
rpIK	50S ribosomal subunit protein L11
rpmI	50S ribosomal subunit protein A
rpmA	50S ribosomal subunit protein L27
rpmD	50S ribosomal subunit protein L4, regulates expression of S10 operon

Validation on other fast growing organisms : [translational bias](#)

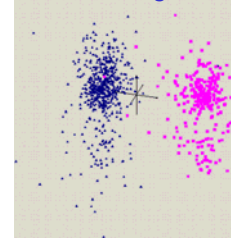


We demonstrated that the set of biased genes

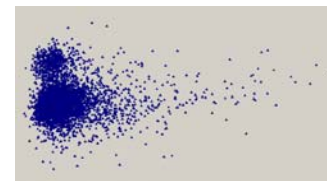
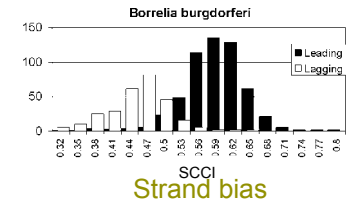
- is **unique** (for the organisms we checked, ~210)
- **exists** also for organisms that do not have an evolutionary tendency explained with translational pressure.

The “existence property” is universal and SCCI/CAI is a universal measure

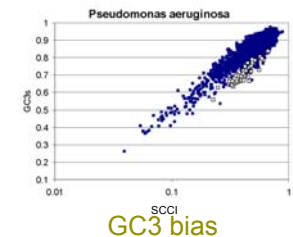
Borrelia burgdorferi



SCCI : a universal measure

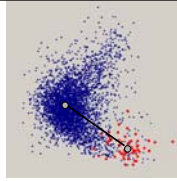


Pseudomonas aeruginosa

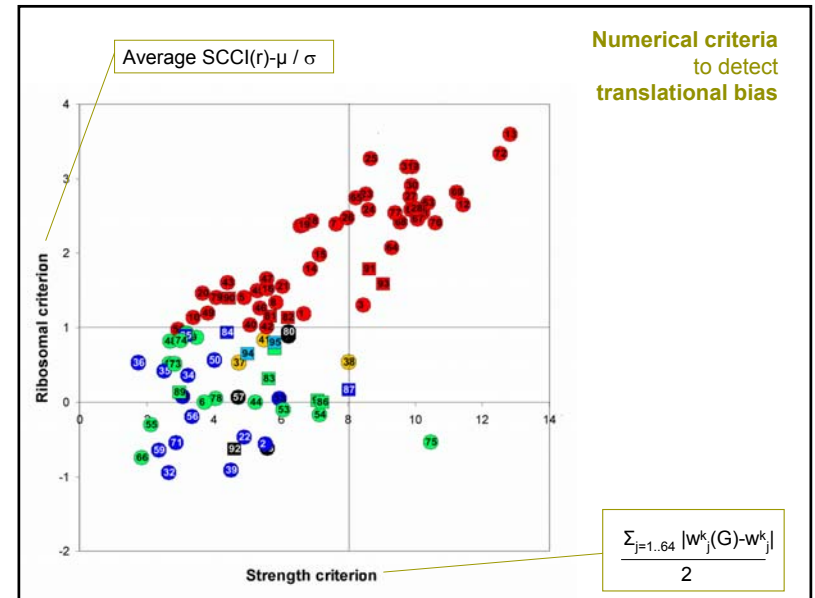


SCCI
GC3 bias

Strong and weak signals for organisms with a predisposition towards translational bias



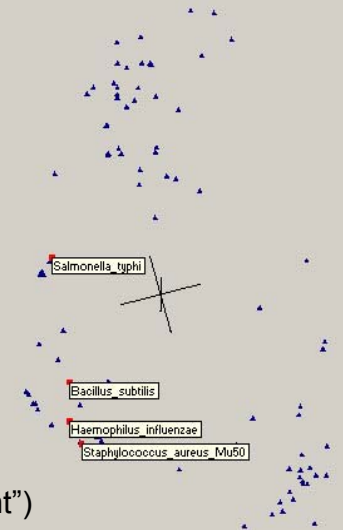
	Mean SCCI	σ	mean SCCI on ribosomal prot
<i>S.cerevisiae</i>	0.16	0.12	0.78
<i>E.coli</i>	0.30	0.10	0.60
<i>V.cholerae</i>	0.28	0.08	0.64
<i>B.subtilis</i>	0.37	0.07	0.64
<i>H.influenzae</i>	0.38	0.9	0.58
<i>M.acetivorans</i>	0.50	0.06	0.63



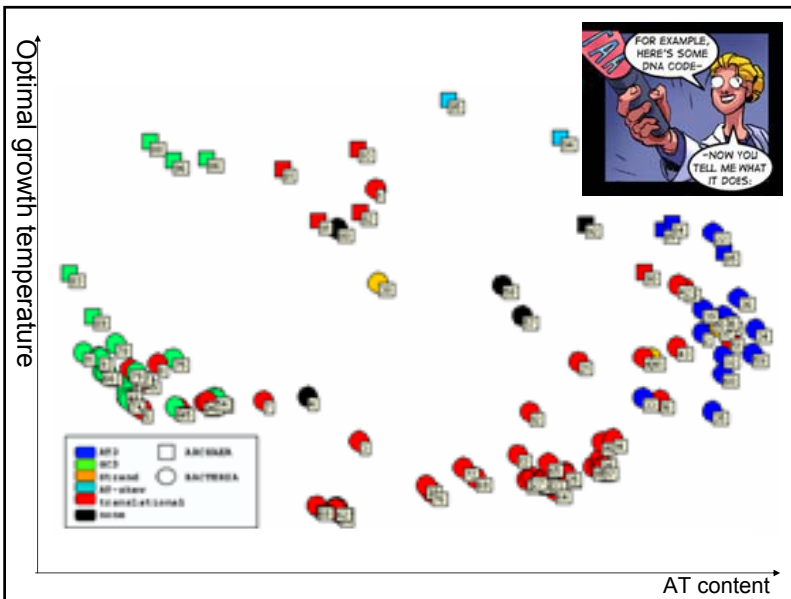
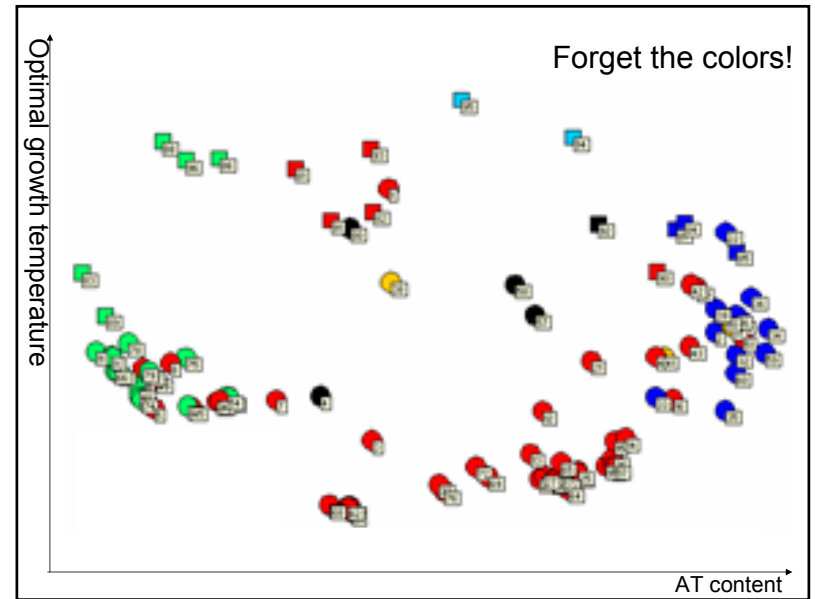
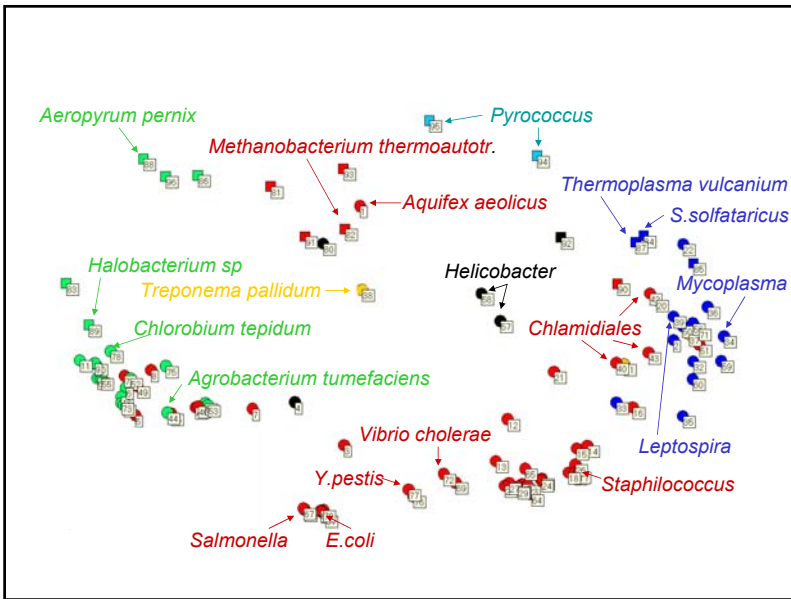
Randomised version

- Randomly choose the 1% of genes in S
- Compute the weights and the SCCI values
- Select the 1% of genes with highest SCCI value
- Repeat the iteration until the algorithm converges

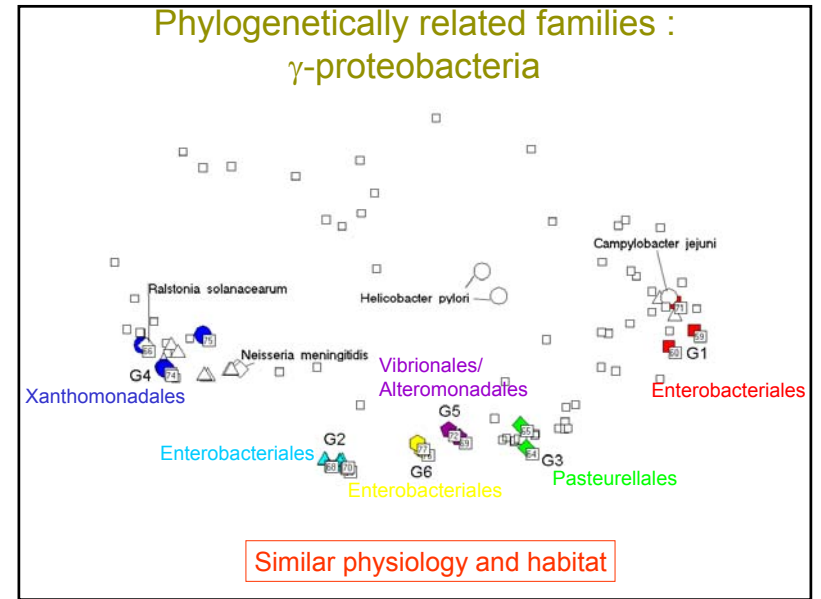
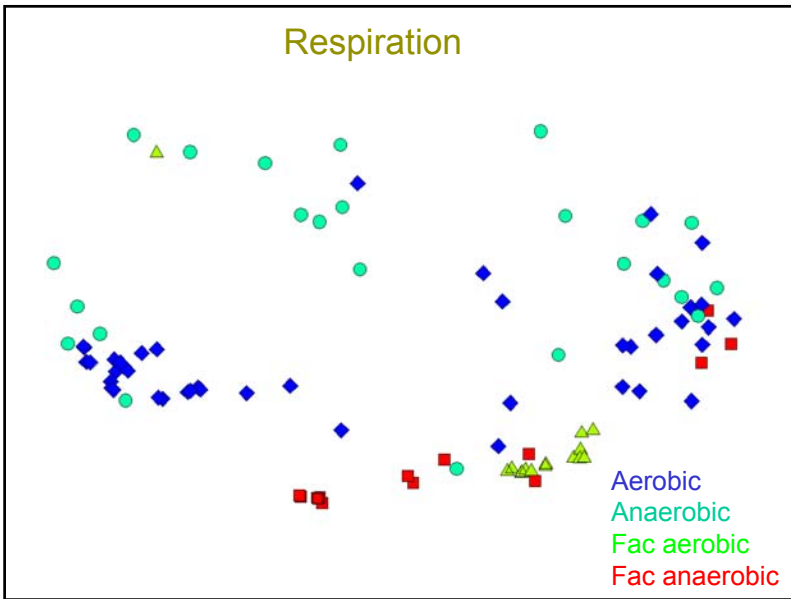
Bacteria and Archaea in codon space



An organism is a vector on 64 coordinates (codon “weight”)



Can we exploit the geometry of the space to derive functional characteristics of groups of organisms?

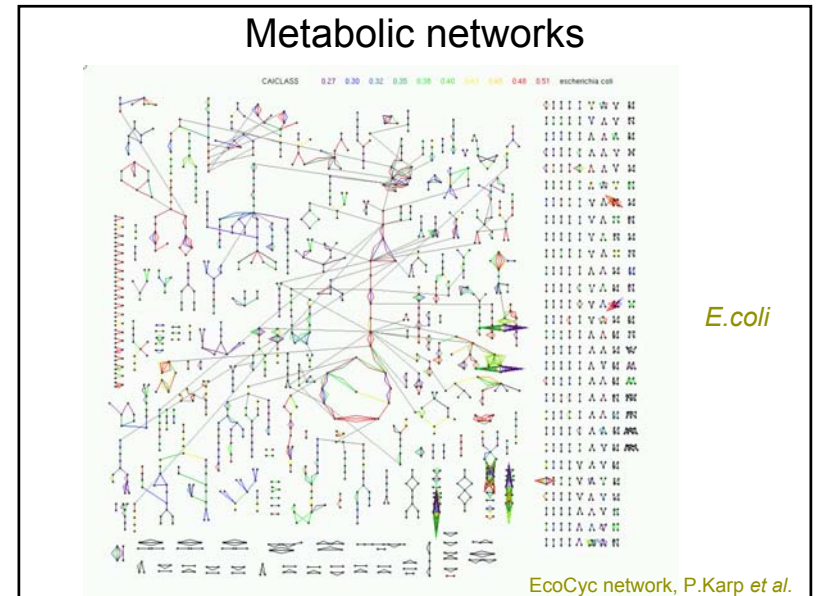


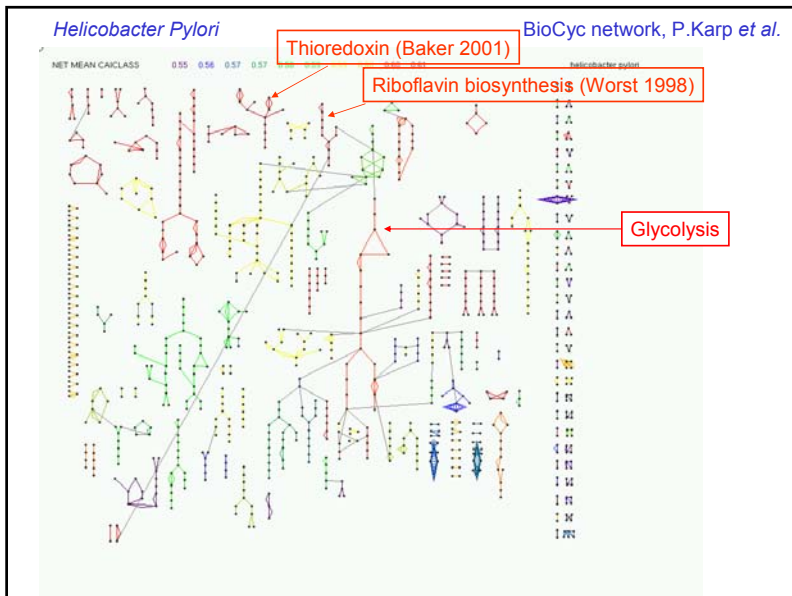
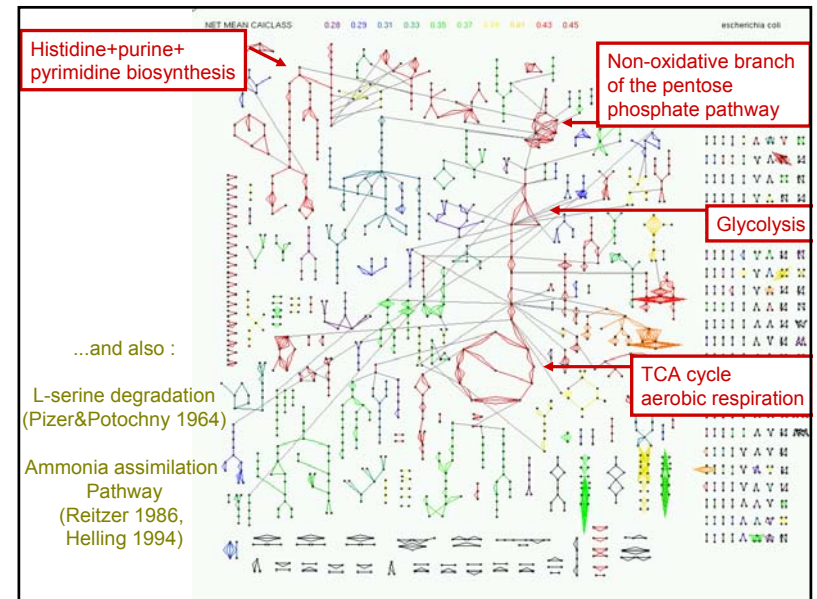
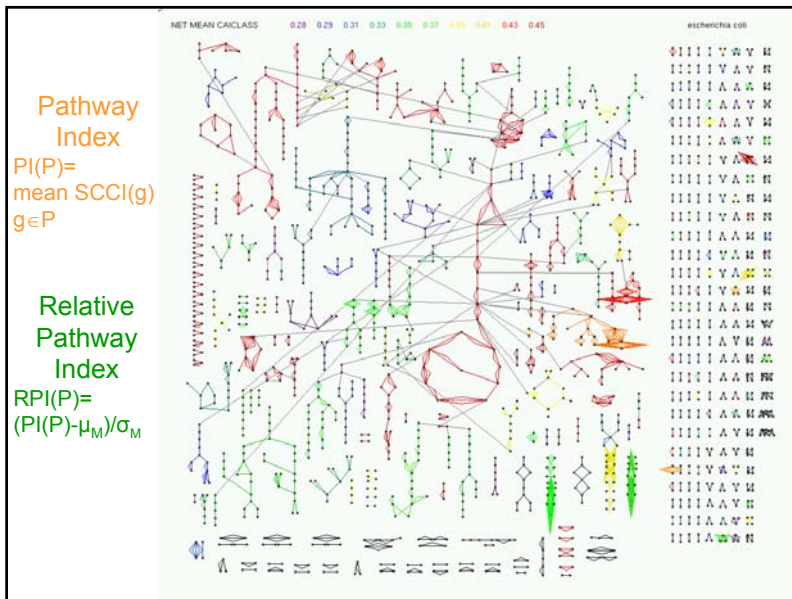
Coherence in the organisms space
based on SCCI

Can we use this signal to deduce some
more biological information ?

Can we determine the most important **metabolic networks**
in a (translationally biased) organism ?

Can we determine genes belonging to **minimal gene sets** ?



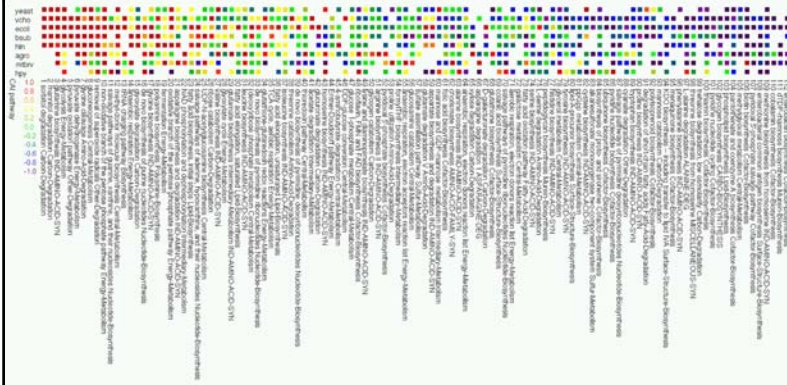


Metabolic pathways essential to *Mycobacterium tuberculosis*

Essential to *M.tuberculosis* but not to other bacteria

Biotin synthesis	(Norman et al. 1994)
Chorismate biosynthesis	(Parish and Stoker 2002)
Asparagine degradation	(Sasseti et al. 2003)
Pyridoxal 5'phosphate biosynthesis	(Sasseti et al. 2003)
Valine degradation	(Sasseti et al. 2003)
Leucine biosynthesis	(Sasseti et al. 2003)
ppGpp	(Primm et al. 2000)

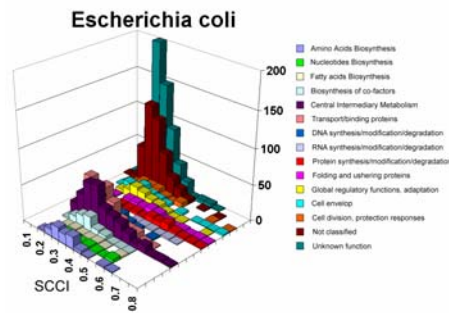
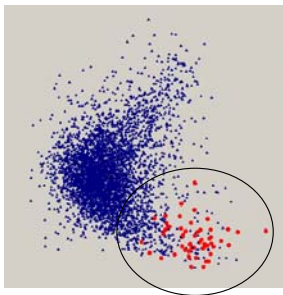
Metabolic networks map



Can we determine genes belonging to minimal gene sets ?

$$SCCI(g) > \mu + \sigma$$

- Genes with uncharacterised function
- Genes dependent on specific environmental conditions
- Stress response genes
- Highly expressed genes (belonging to most species)
- Non-orthologous genes



$$SCCI(g) > \mu + \sigma = 0.42$$

Genes with specific metabolic functions

Photosynthesis metabolism : *Synechocystis*

Phycobilisome proteins
Photosystem I and II
Fructose-1,6-bisphosphate-aldolase

Methan metabolism : *Methanosarcina acetivorans*

Methanol-5 hydroxybenzimidazolycobamideco methyltransferase
Methyl coenzyme M reductase
Methylcobamide methyltransferase isozyme M
Corrinoid proteins
Ack, Pta, cdhA

Ferredoxin metabolism : *Pyrococcus abyssi*

Ferredoxin
Ferredoxin oxidoreductase
Keto-valine-ferredoxin oxidoreductase γ -chain

Carbohydrates metabolism : *Streptococcus mutans*

Transport and metabolism of cellobiose, sucrose, beta-glucoside
Metabolism of mannitol
Genes for metabolism of glucose, fructose, mannose, maltose/maltodextrin

Collaborations and references

Algorithm and microbial codon space :

- F.Képès, CNRS & génopole Evry
- A.Zinovyev, IHÉS & Institut Curie (Paris)

A. Carbone, A. Zinovyev, F. Képès, Codon adaptation index as a measure of dominating codon bias, *Bioinformatics*, **19**, 2005–2015, 2003.

A. Carbone, F. Képès, A. Zinovyev, Codon Bias Signatures, Organization of Microorganisms in Codon Space, and Lifestyle, *Molecular Biology and Evolution*, **22**, 547–561, 2004.

Metabolic networks comparison :

- D.Madden, IHÉS & Sherbrooke University (Canada)

A. Carbone, R. Madden, Insights on the Evolution of Metabolic Networks of Unicellular Translationally Biased Organisms from Transcriptomic Data and Sequence Analysis, *Journal of Molecular Evolution*, **59**, 1–25, 2005.

Minimal gene sets :

A. Carbone, Computational prediction of genomic functional cores specific to different microbes, *Journal of Molecular Evolution*, 2006, under revision.