

Genomes codon signatures and organisms lifestyle

Alessandra Carbone
Génomique Analytique
Université Pierre et Marie Curie, Paris

carbone@ihes.fr

We want to capture **global data sets** and integrate them together to get a coherent understanding of the biological system

It is important to define new spaces, new measures: space of genes, space of genomes, spaces of interactions...



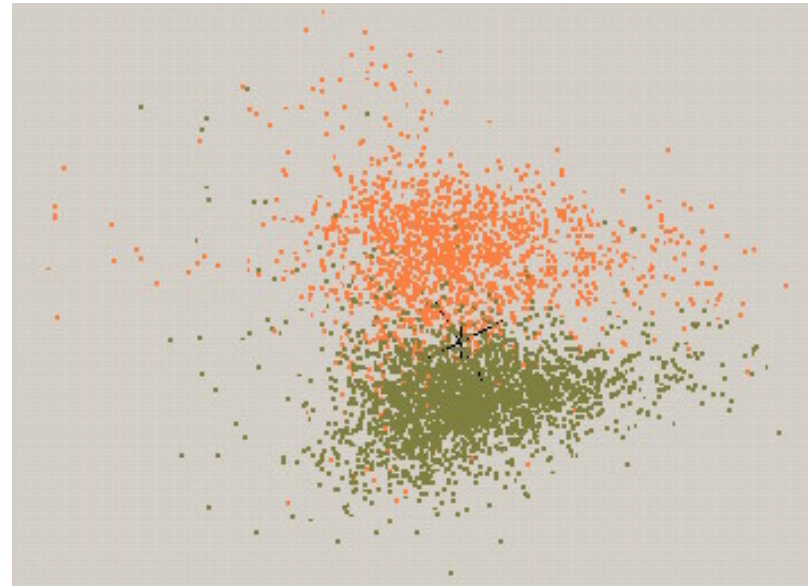
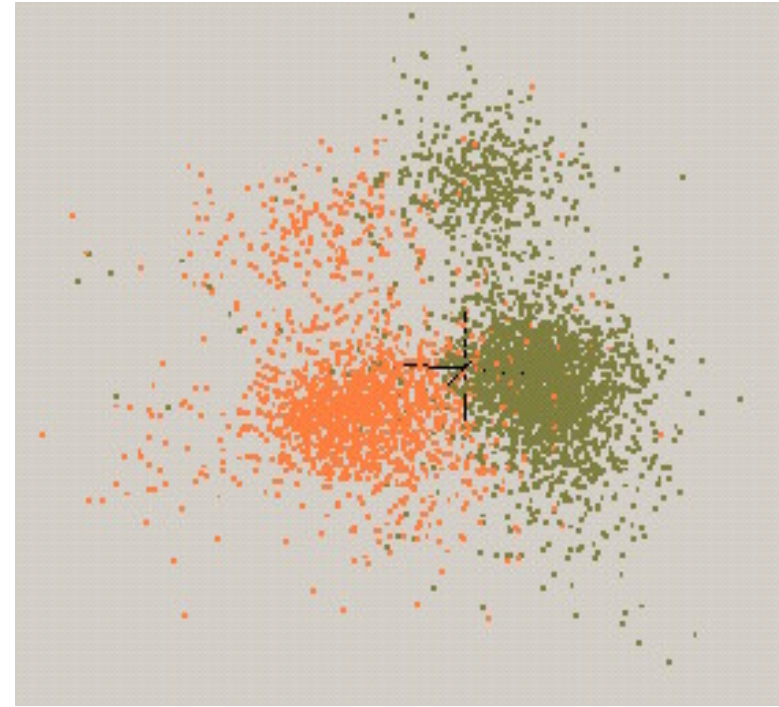
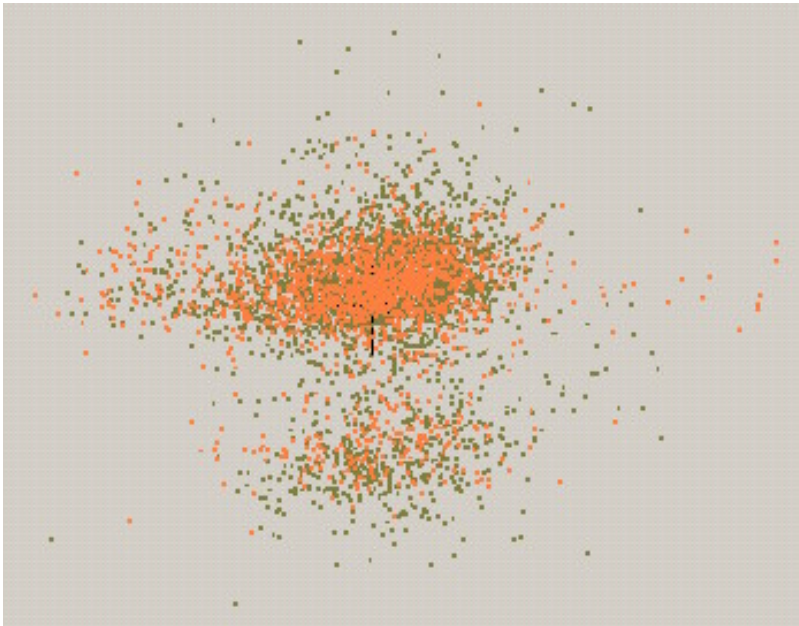
$g = [x_{1,g} \ x_{2,g} \ \dots \ x_{64,g}]$ $x_{i,g}$ relative frequency of codon i in g

Normalisation:

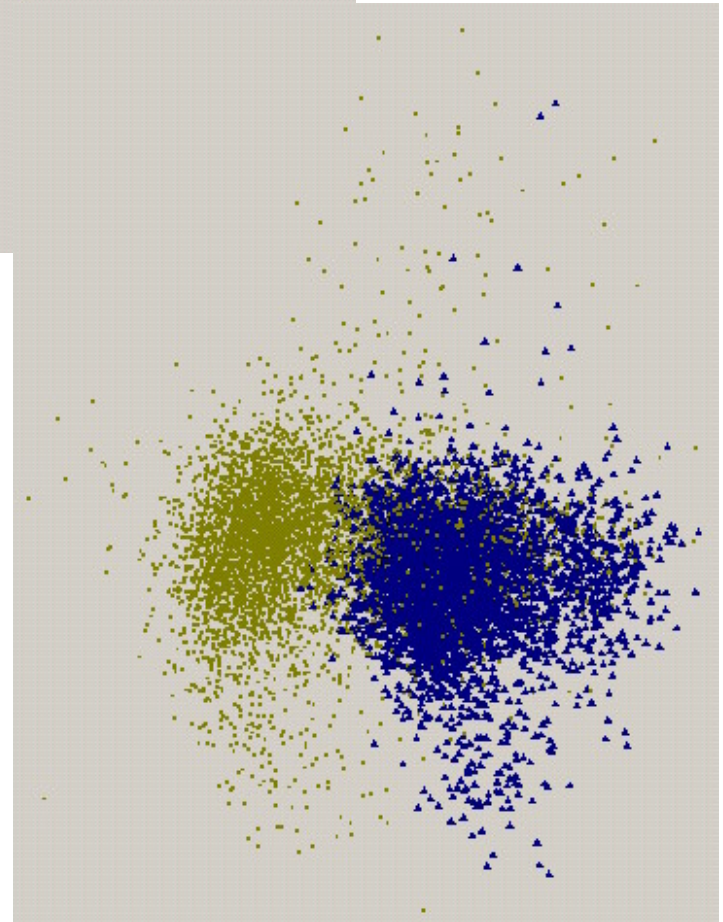
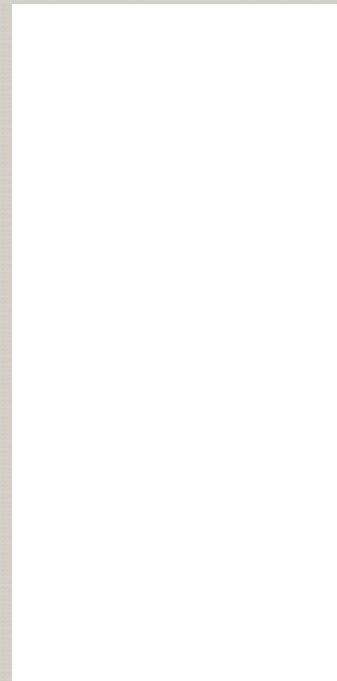
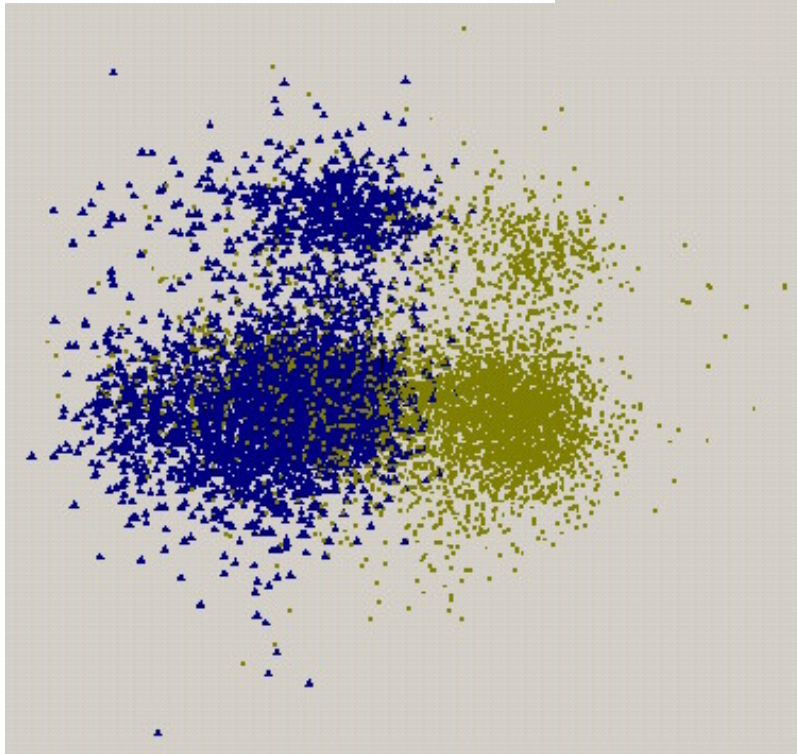
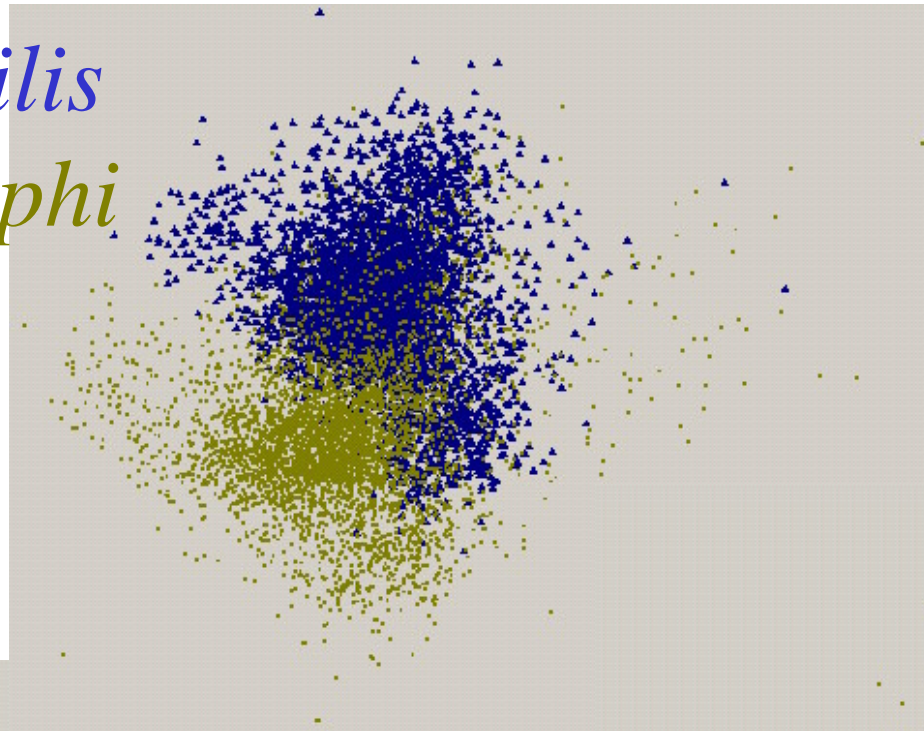
$(x_{i,g} - \underline{x}_i) / \sigma_i$ \underline{x}_i mean of frequencies $\underline{x}_{i,g}$
 σ_i standard deviation of $\underline{x}_{i,g}$

...we use normalized vectors and PCA

Haemophilus influenzae
Staphylococcus aureus

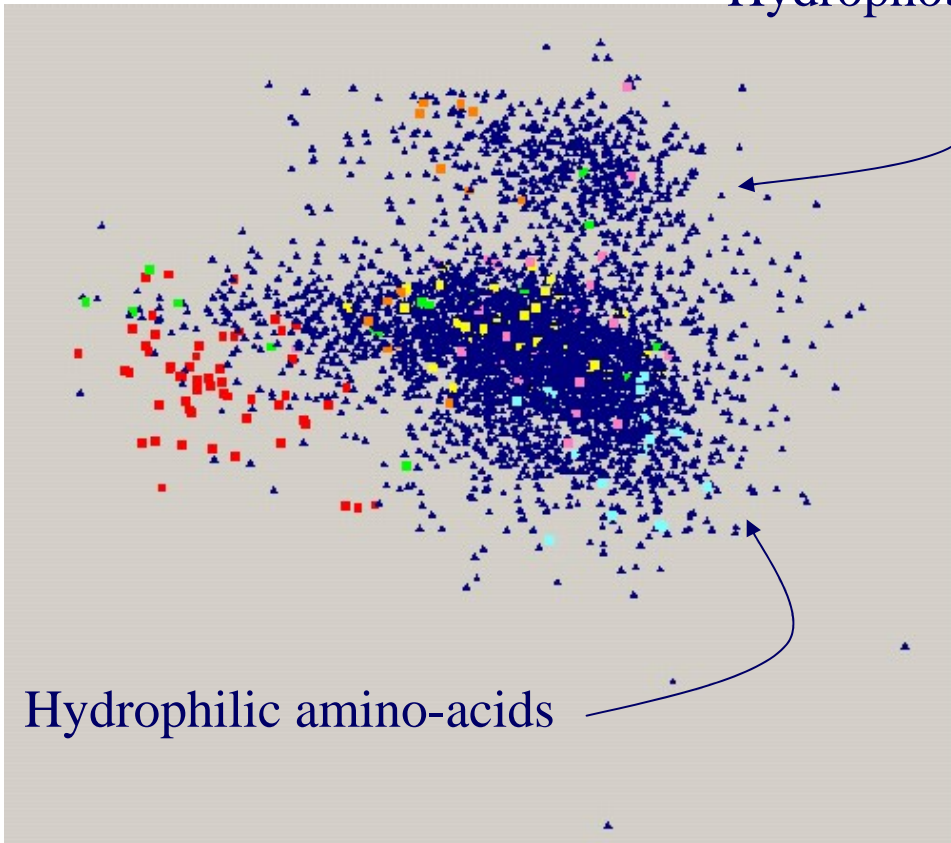


Bacillus subtilis
Salmonella typhi



E.coli

Hydrophobic amino-acids



Hydrophilic amino-acids

Do different genomes have the same centroids network ?

Ribosomal proteins

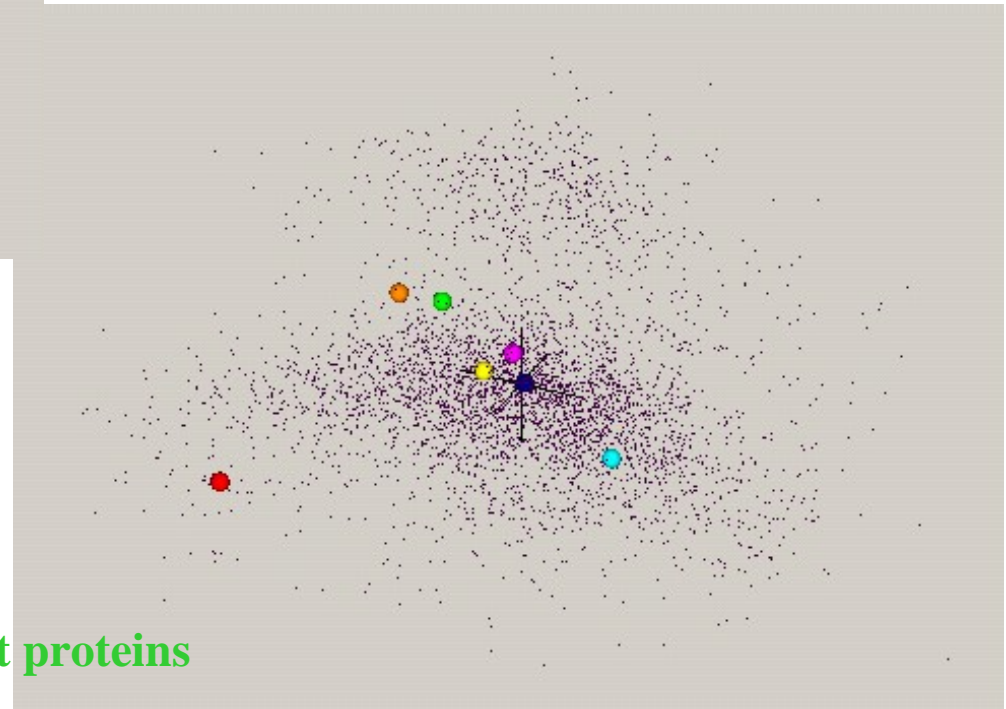
ATP binding proteins

IS proteins

NADH proteins

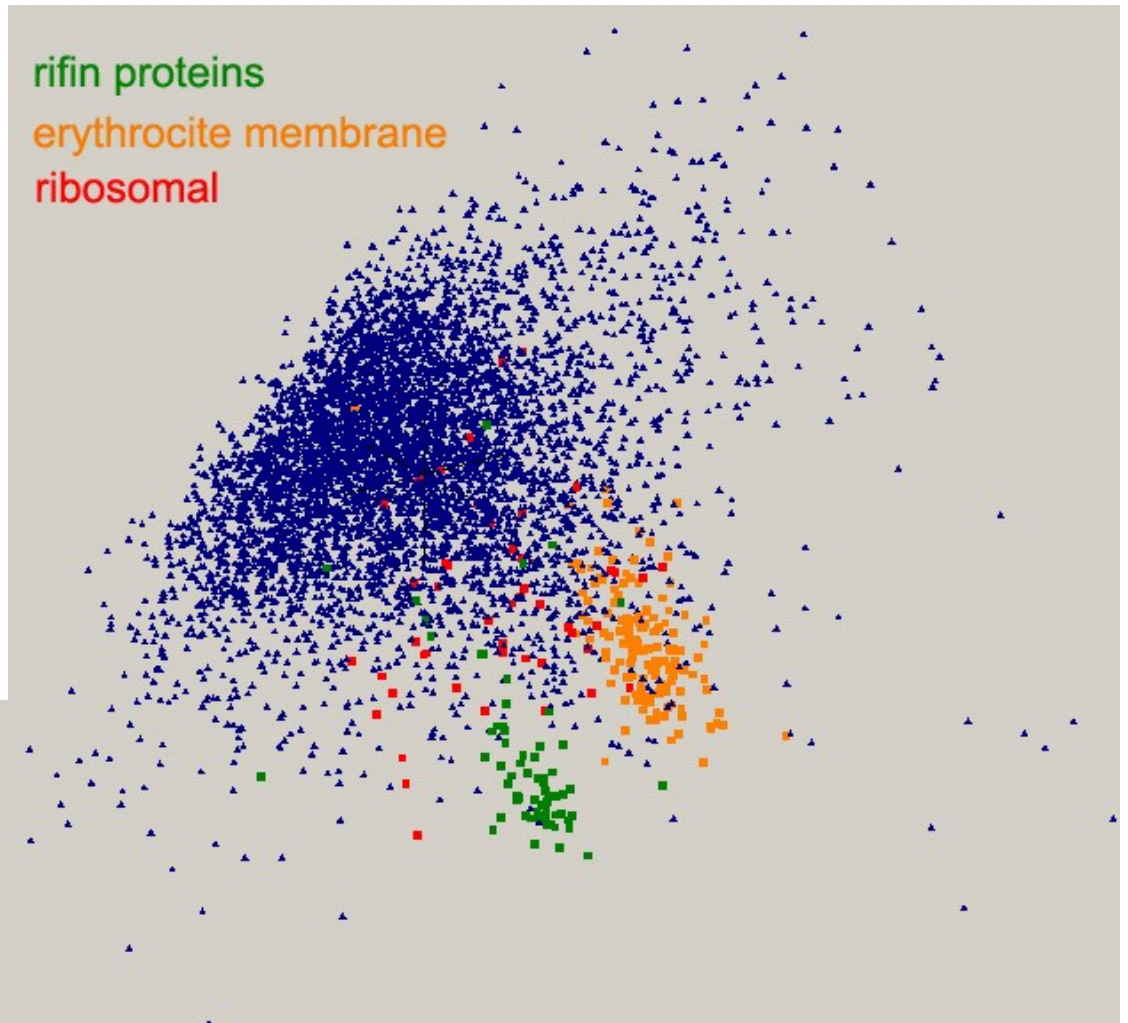
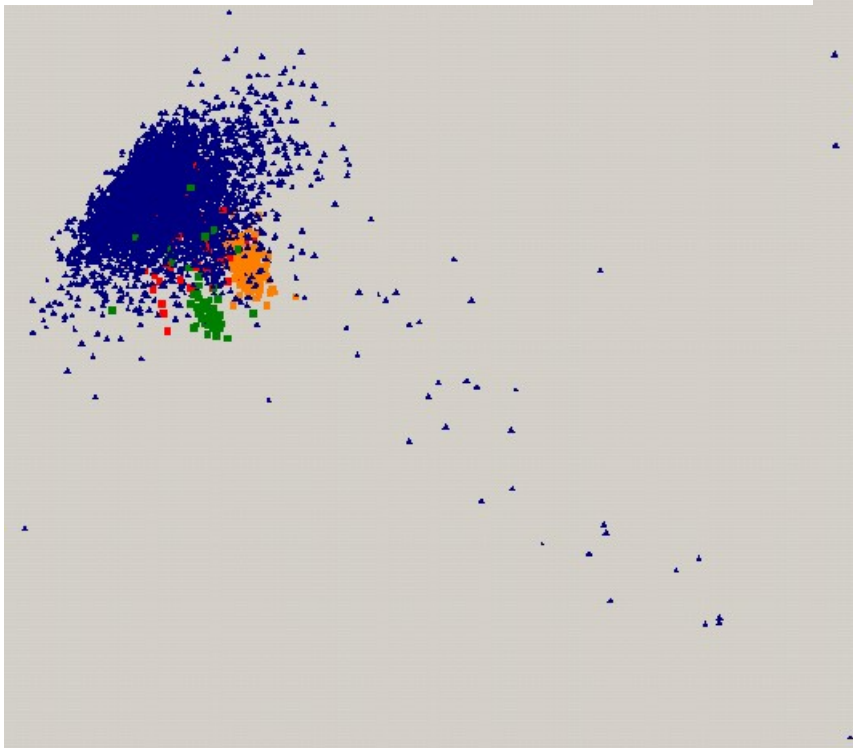
Flagellar biosynthesis proteins

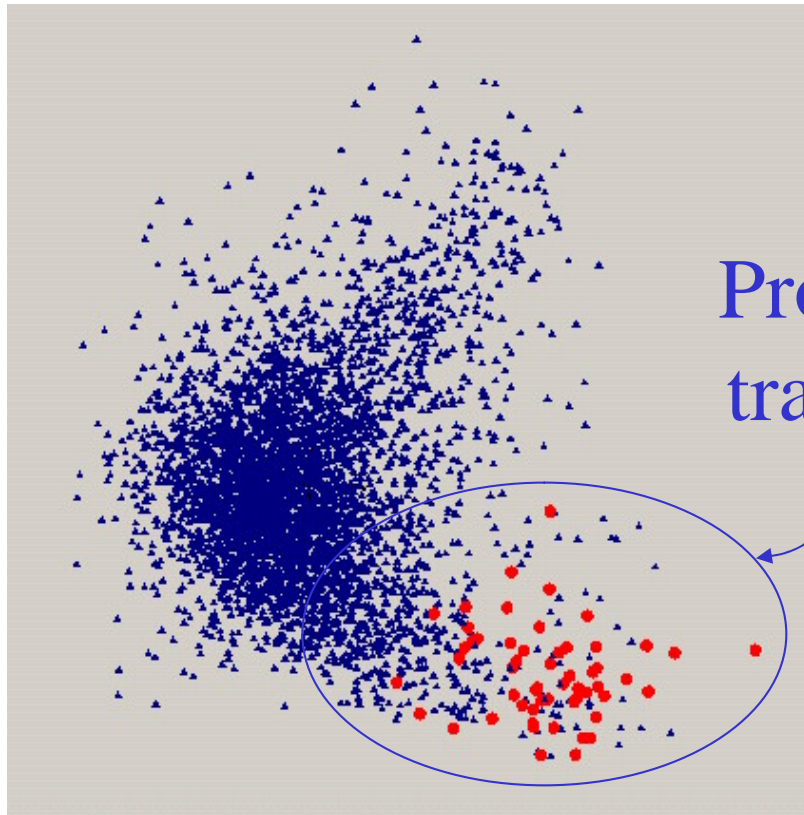
Lipoproteins, membrane proteins, transport proteins



*Plasmodium
falciparum*

rifin proteins
erythrocyte membrane
ribosomal





Proteins coding for the translational machinery

They are the most expressed in *E.coli*

Biological hypothesis: genes that are biased in the same manner are the most expressed.

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

Do they form the set of most biased genes?

An automatic detection of most biased genes

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

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

L number of codons in g

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

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

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

$$\text{CAI}(G/S) \leq \text{CAI}(S)$$

where G is the set of all genes

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

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

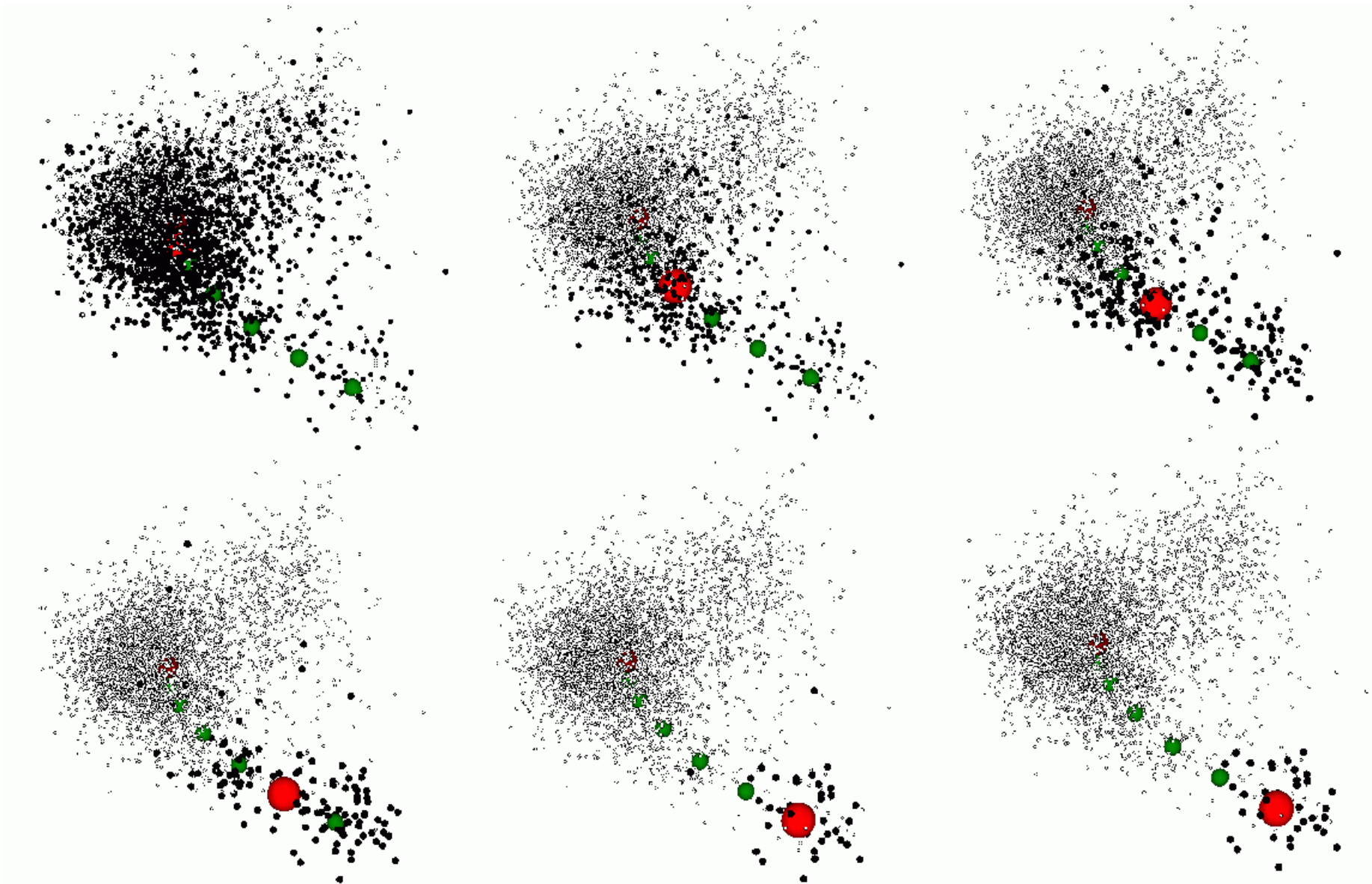
d_1, \dots, d_{20} preferred codons of G

we look for the set S such that

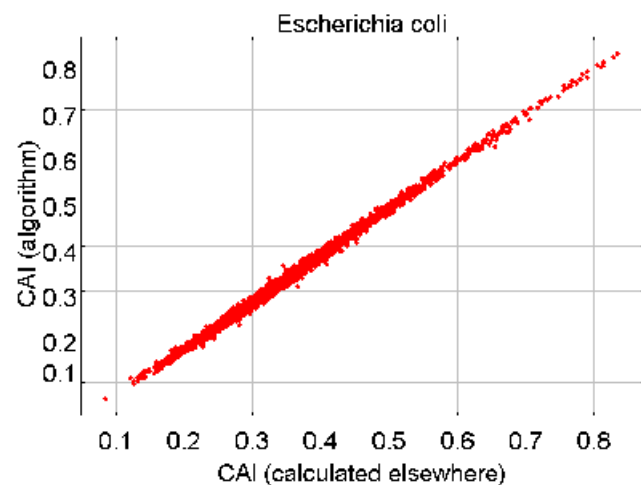
$$\sum_{i=1}^{20} \chi(c_i, d_i) \quad \text{is minimal}$$

- An exhaustive search is unfeasible
- Idea of the algorithm:
 - compute the weight of the codons over the whole genome and compute afterwards CAI values for all genes
 - Select the 50% of genes with the highest CAI 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 CAI until convergence is reached.

Behavior of the algorithm for *E.coli*

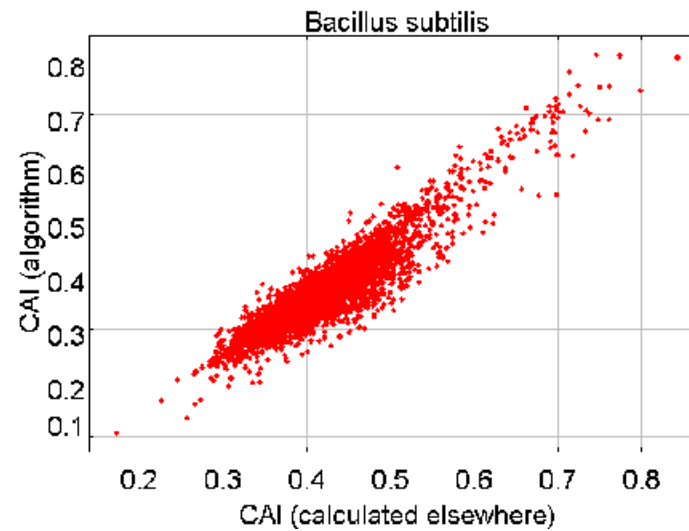
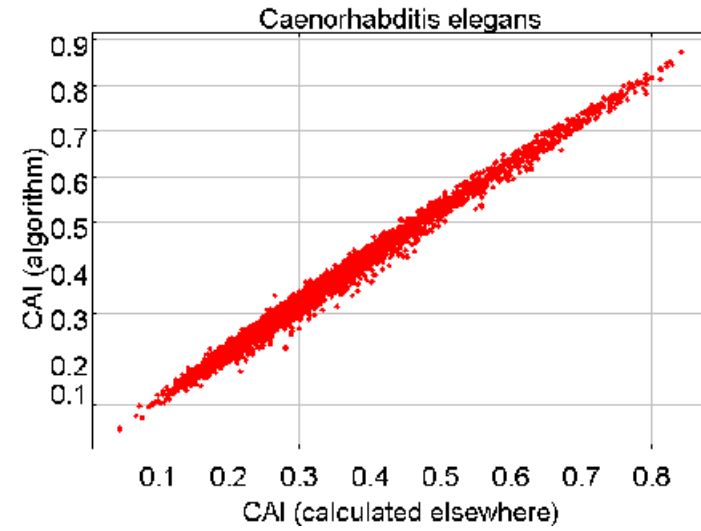
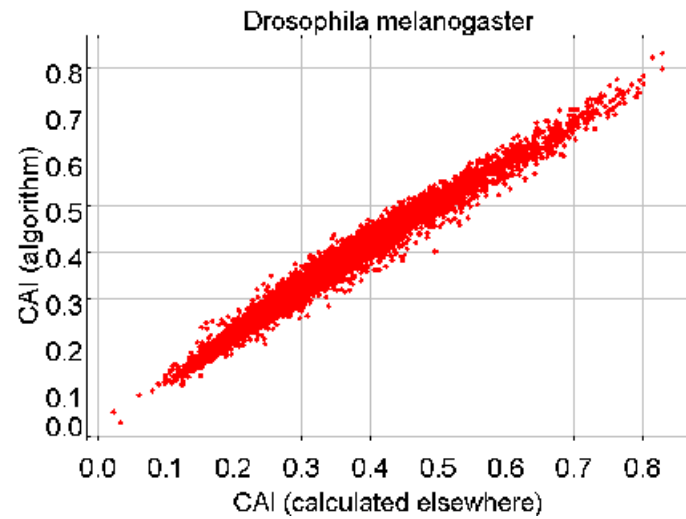


S found by the algorithm: *E.coli*

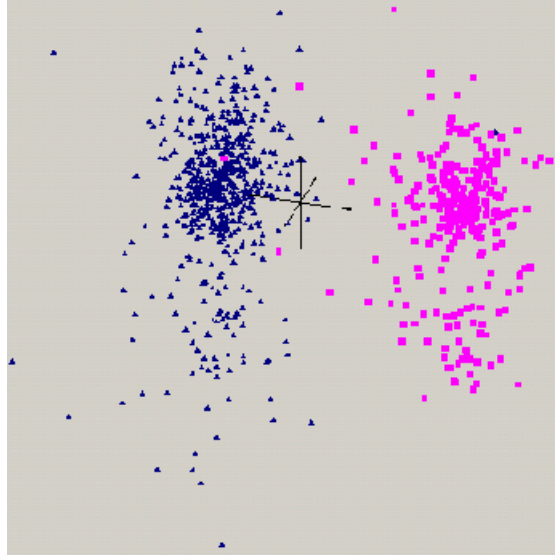


Gene	Annotation
tufA	protein chain elongation factor EF-Tu
tufB	protein chain elongation factor EF-Tu
tsf	protein chain elongation factor EF-Ts
fusA	GTP-binding protein chain elongation factor EF-G
mopA	chaperonin GroEL
dnaK	heat shock protein DnaK
cspA	cold shock protein 7.4
tig	trigger factor
ompA	outer membrane protein
ompX	outer membrane protein
ompC	outer membrane protein
lpp	murein lipoprotein
pal	peptidoglycan-associated lipoprotein
yaiU	putative flagellin structural protein
yfiD	putative formate acetyltransferase
eno	diadenosine tetraphosphatase
tpiA	triosephosphate isomerase
pgk	phosphoglycerate kinase
gapA	glyceraldehyde-3-phosphate dehydrogenase A
fba	fructose-bisphosphate aldolase class II
pykF	pyruvate kinase I
pflB	formate acetyltransferase 1
ahpC	alkyl hydroperoxide reductase C22 subunit
sodA	superoxide dismutase SodA
tktA	transketolase 1/2 isozyme
rpoC	RNA polymerase beta prime subunit
rpsI	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
rplI	50S ribosomal subunit protein L9
rplL	50S ribosomal subunit protein L7/L12
rplC	50S ribosomal subunit protein L3
rpmE	50S ribosomal subunit protein L31
rplB	50S ribosomal subunit protein L2
rplK	50S ribosomal subunit protein L11
rpmI	50S ribosomal subunit protein A
rpmA	50S ribosomal subunit protein L27
rplD	50S ribosomal subunit protein L4, regulates expression of S10 operon

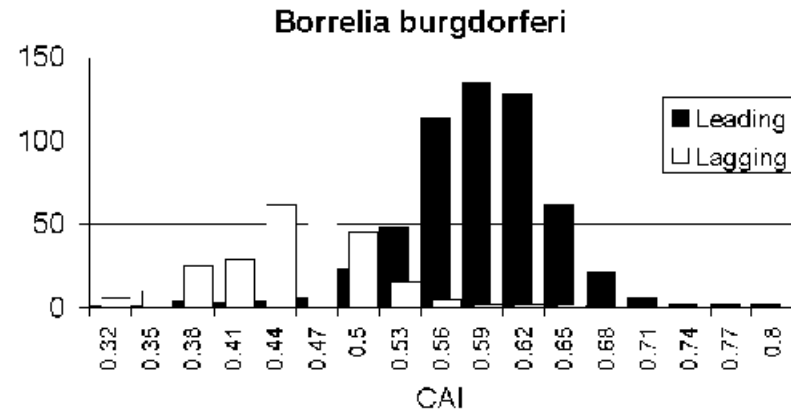
Validation on other fast growing organisms : translational bias



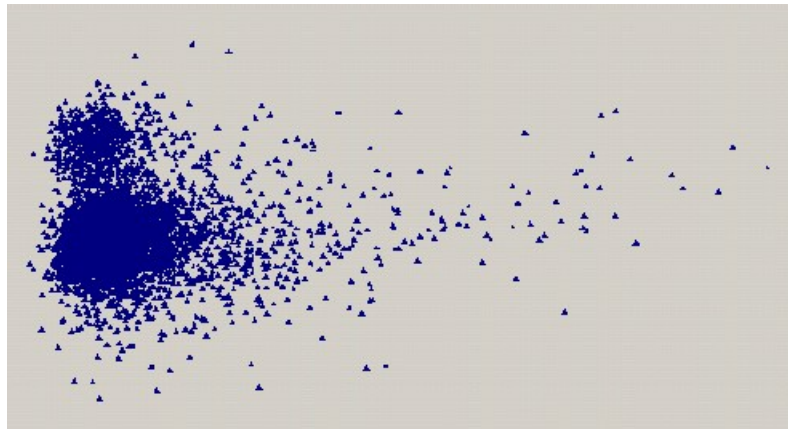
Borrelia burgdorferi



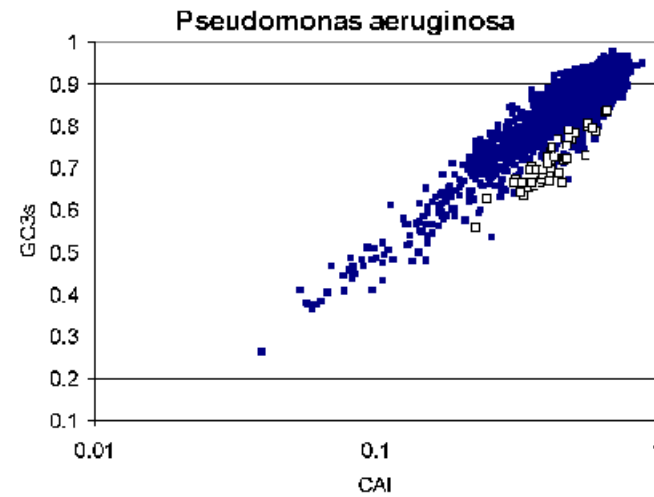
CAI : a universal measure



Strand bias

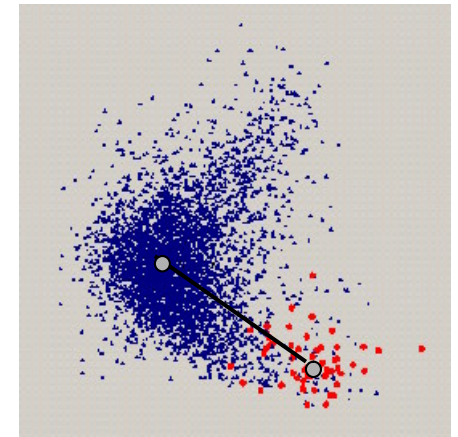


Pseudomonas aeruginosa



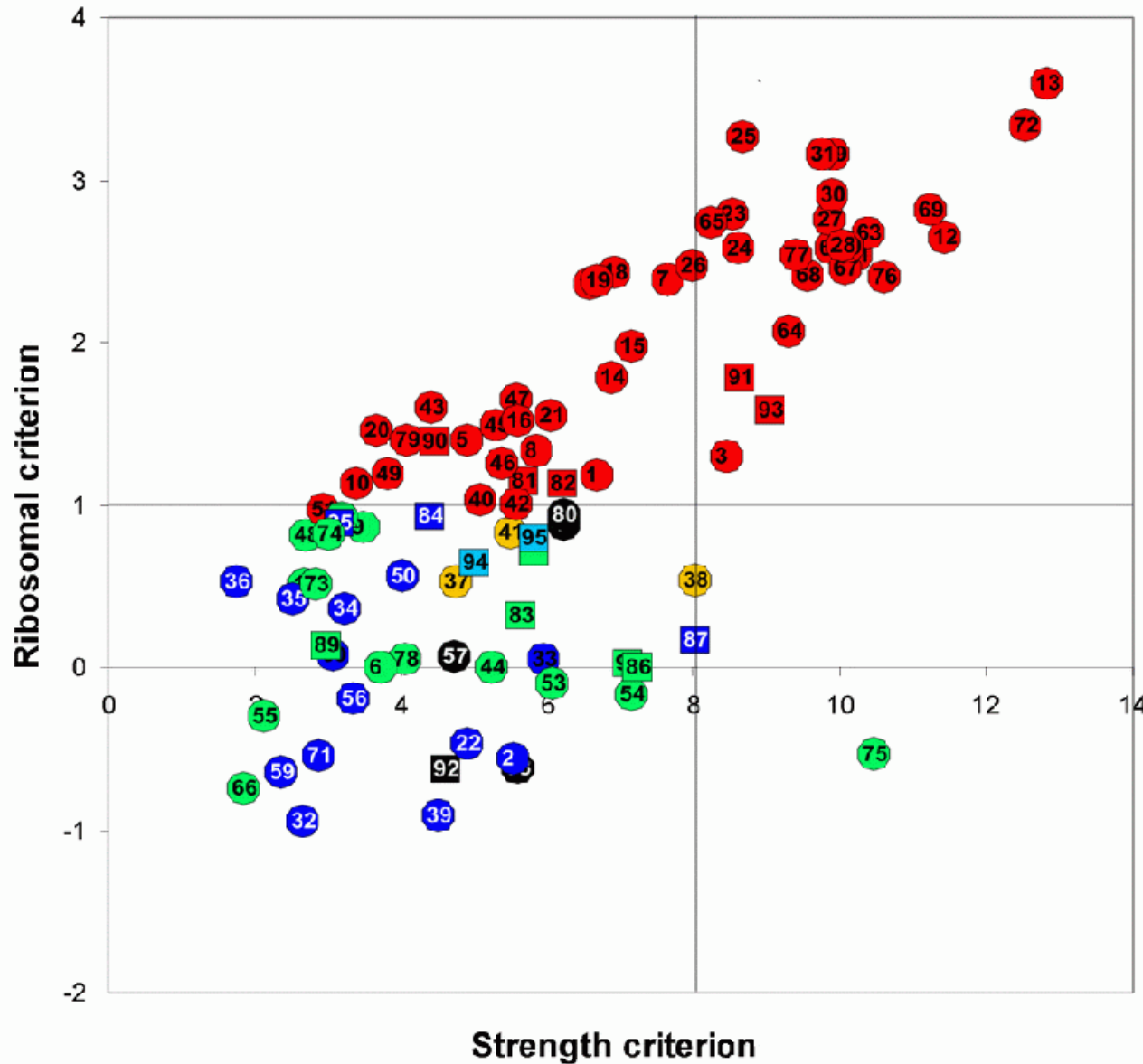
GC3 bias

Strong and weak signals for translationally biased organisms



	Mean CAI	σ	mean CAI 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>Methanosarcina acetivorans</i>	0.50	0.06	0.63

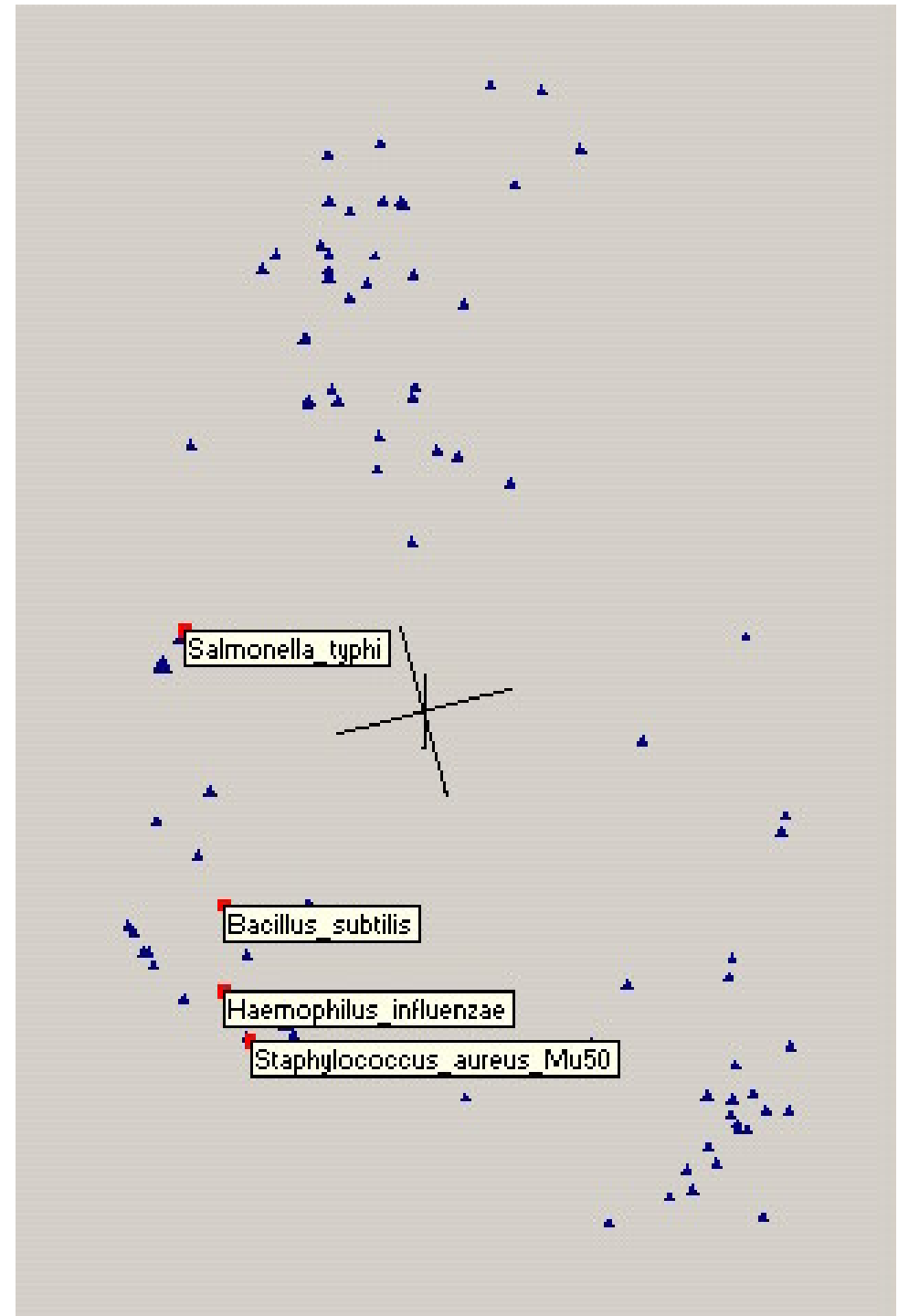
Numerical criteria to speak about organism signatures : strength of the biases



Genomes show
a projection of
biases

Bacteria and archaea in codon space

An organism is a vector
of 64 codon weights



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

Aeropyrum pernix

Pyrococcus

Methanobacterium thermoautotr.

Aquifex aeolicus

Thermoplasma vulcanium

S. solfataricus

Halobacterium sp

Treponema pallidum

Helicobacter

Mycoplasma

Chlorobium tepidum

Chlamidiales

Agrobacterium tumefaciens

Vibrio cholerae

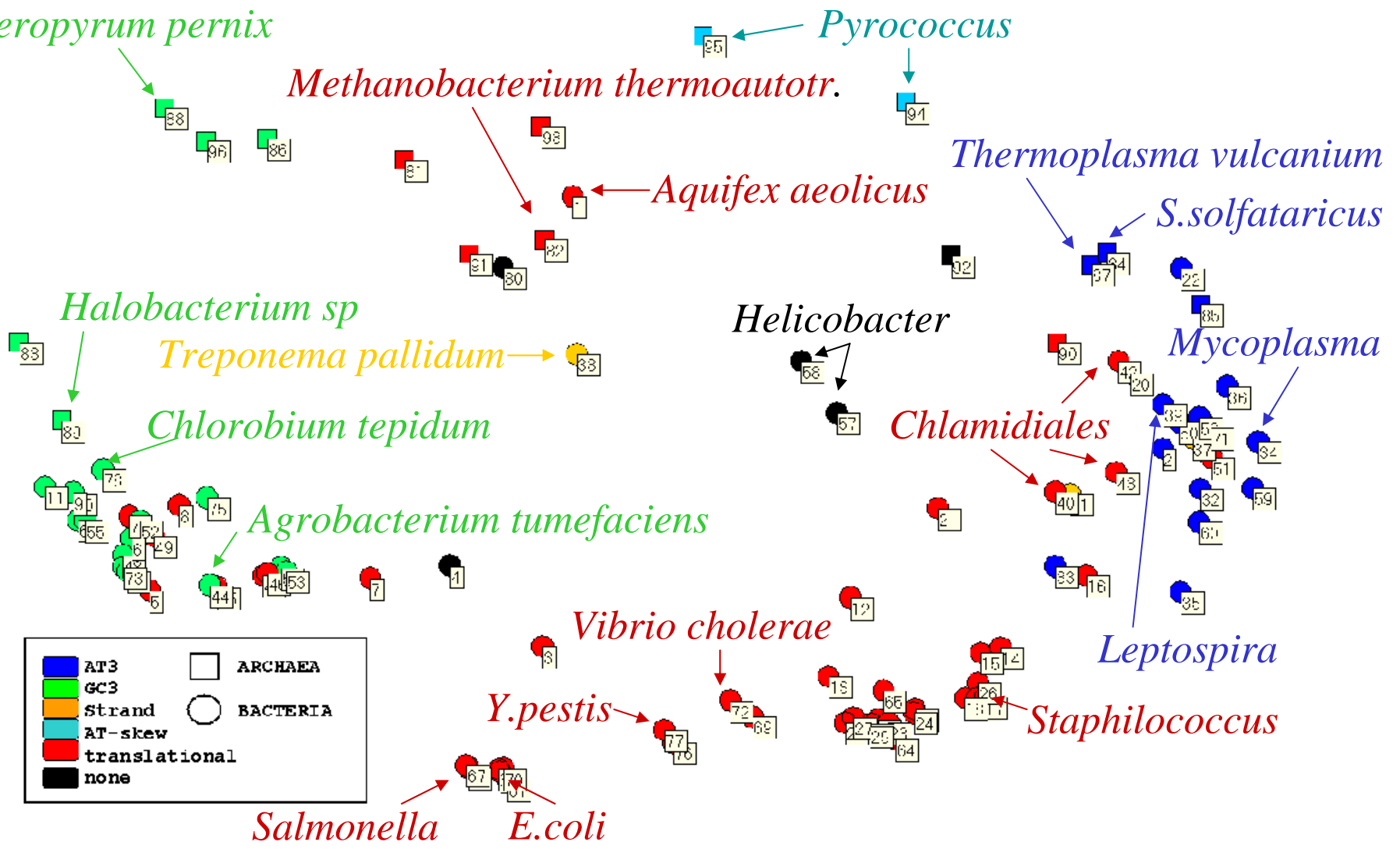
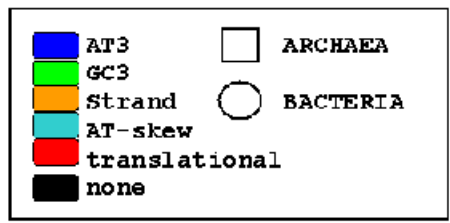
Leptospira

Y. pestis

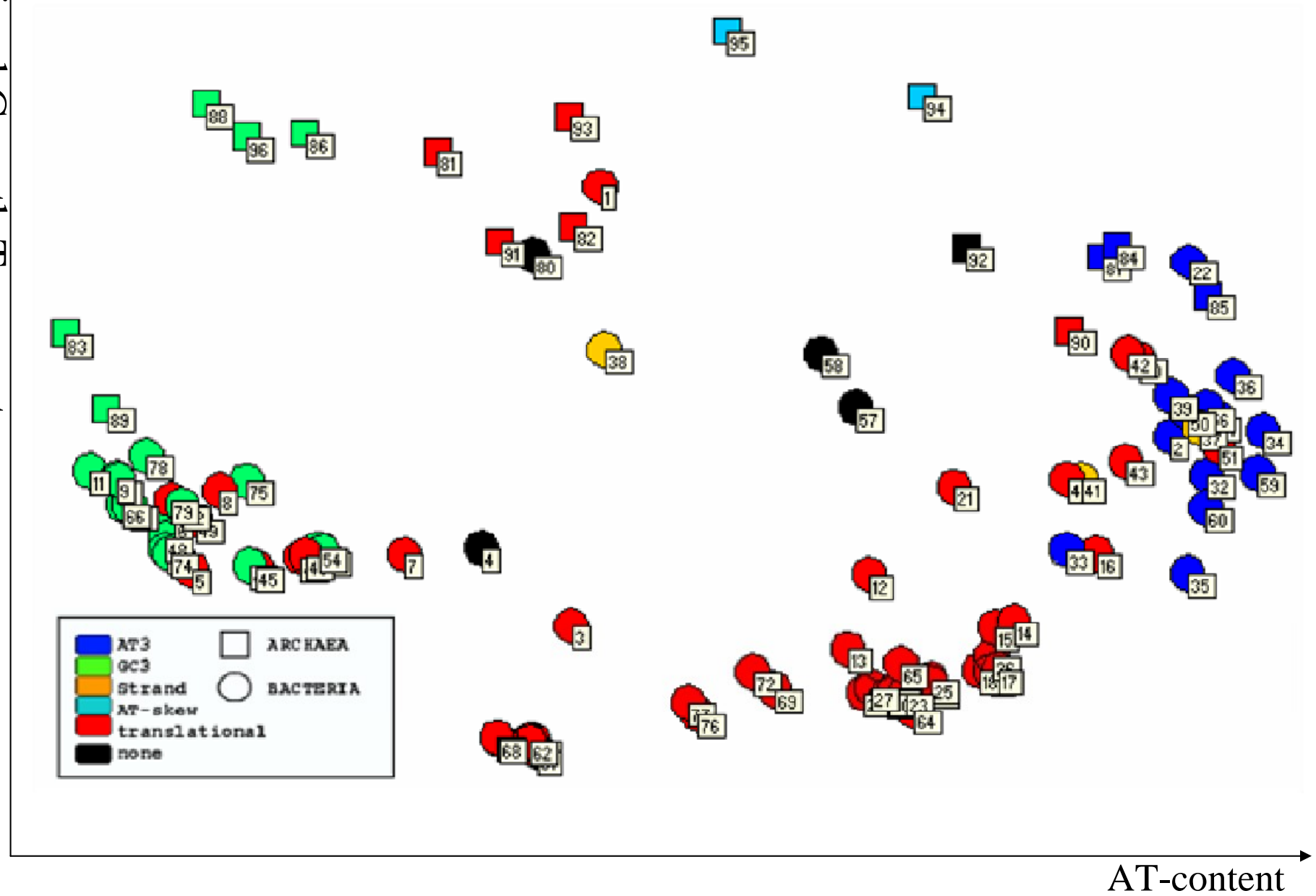
Staphilococcus

Salmonella

E. coli

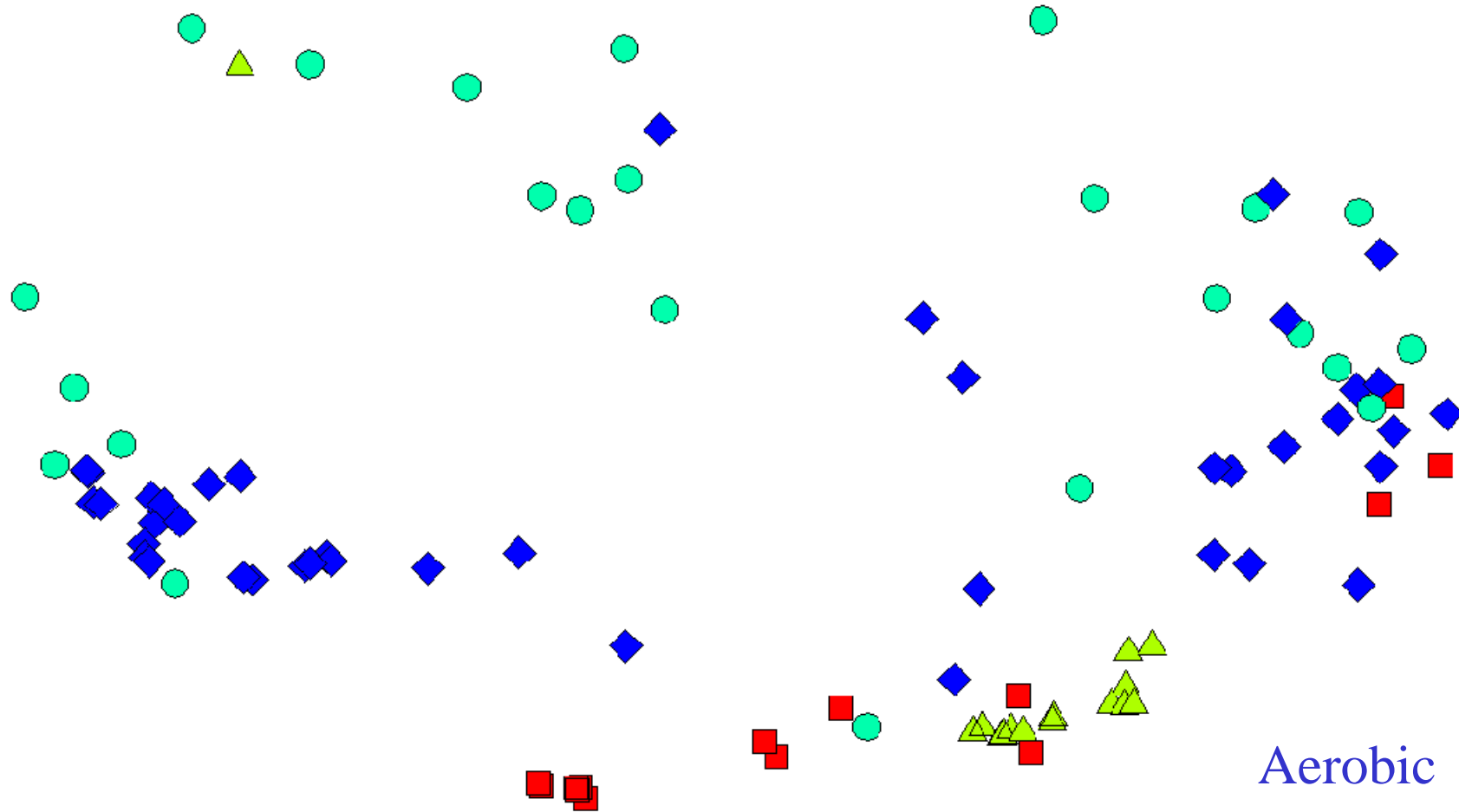


Optimal Growth Temperature



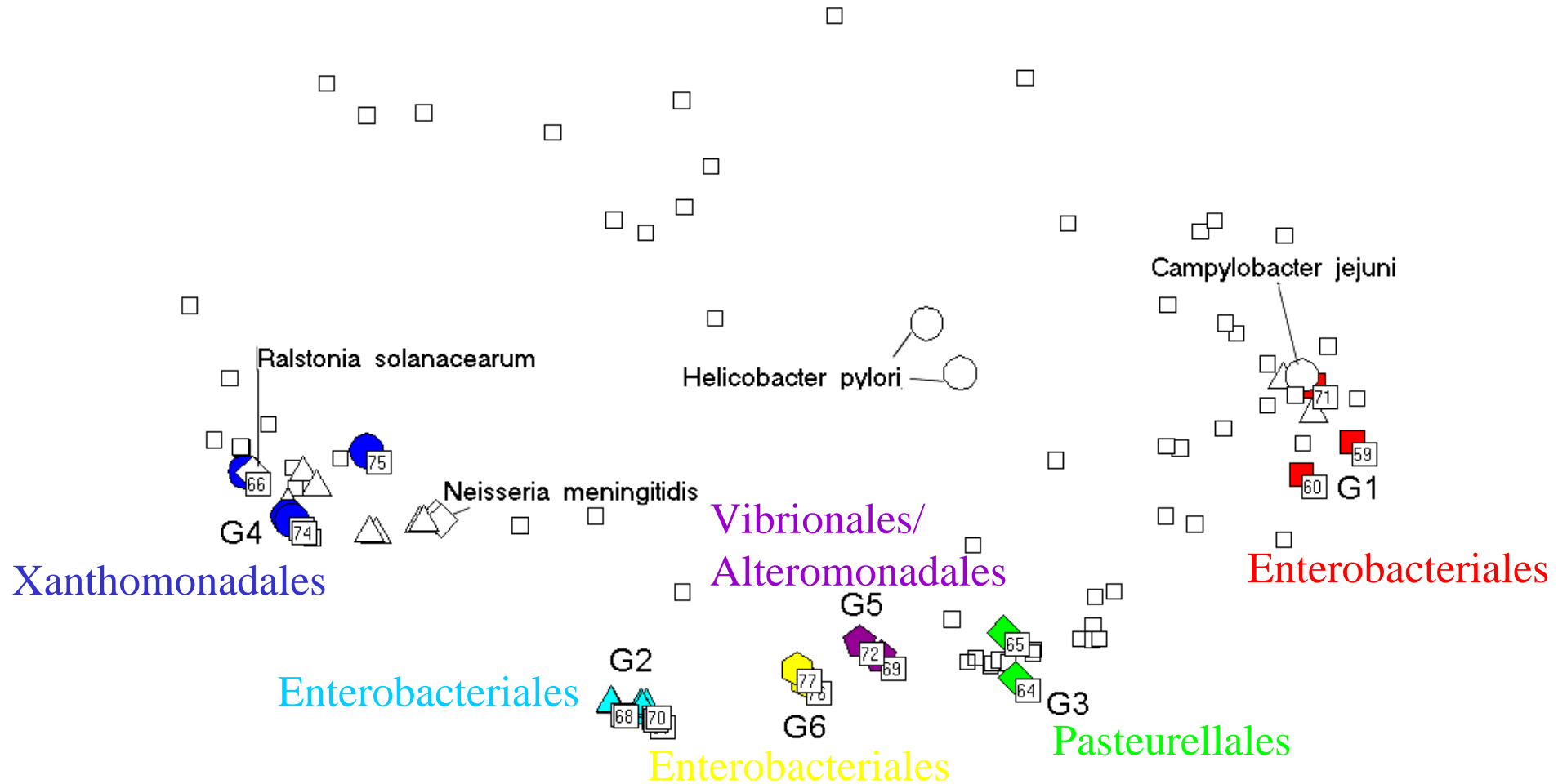
AT-content

Respiration

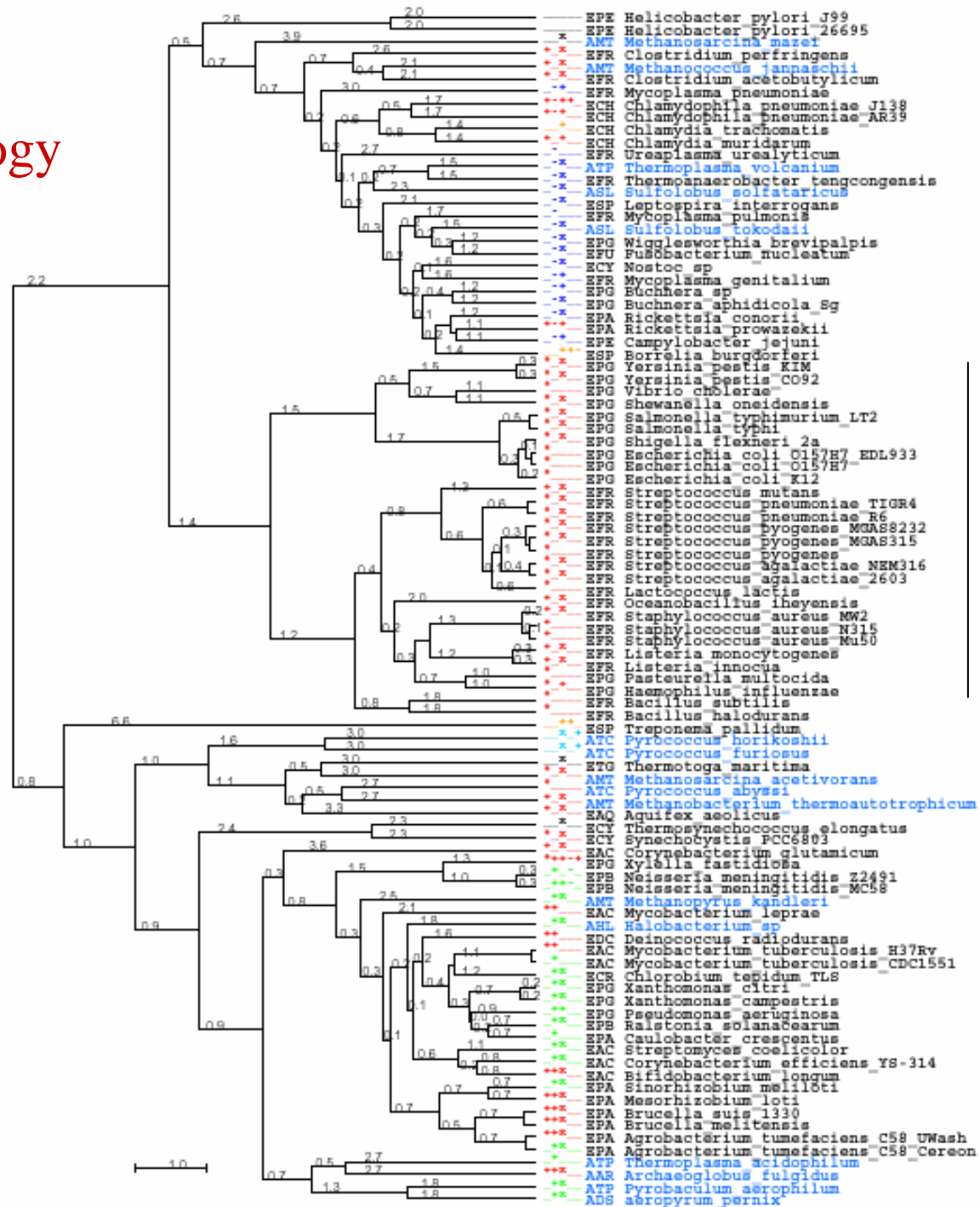


Aerobic
Anaerobic
Fac aerobic
Fac anaerobic

Phylogenetically related families : γ -proteobacteria



Similar physiology
and habitat :



AT-rich

Transl. bias

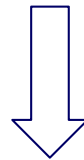
GC-rich

Signals of translational selection



Codon bias values for genes

CAI is a universal biological signal,
not only linked to translational selection



Coherence of the space of organisms

Can we use further this biological signal ?

Can we exploit gene bias to derive **lifestyle** information of an organism?

Can “important” metabolic networks be detected ?

YES

For translationally biased organisms

Photosynthetic metabolism : *Synechocystis*

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

Methane metabolism : *Methanosarcina acetivorans*

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

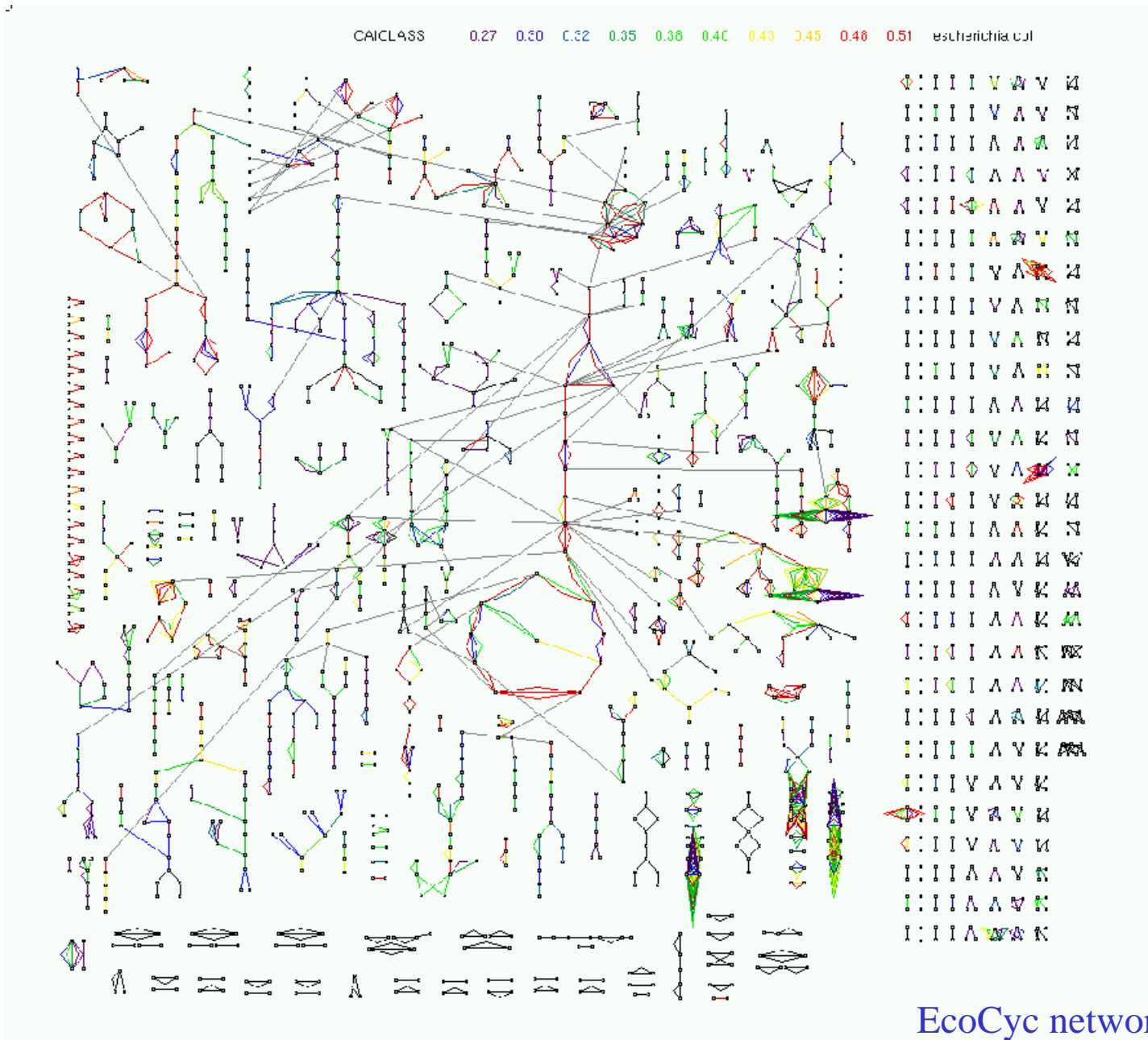
Ferredoxin metabolism : *Pyrococcus abissi*

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

Metabolism of carbohydrates : *Streptococcus mutans*

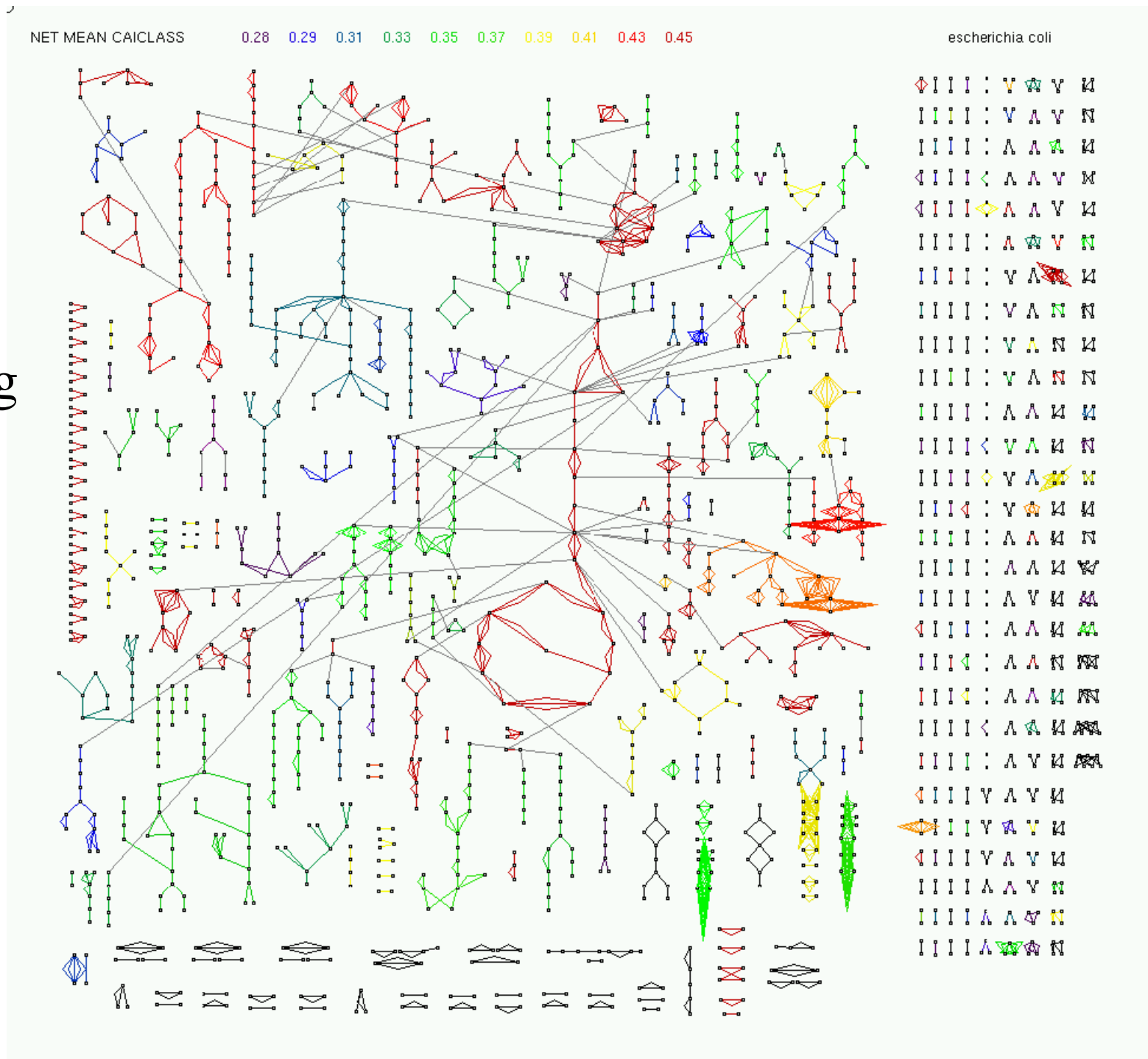
Transport and metabolism of cellobiose, sucrose, beta-glucoside
Metabolism of mannitol
Genes for metabolism of glucose, fructose, mannose, maltose/maltodextrin

Metabolic maps



“Averaging out”

Relative Pathway Index
 $(RPI(p) = (PI(P) - \mu_M) / \sigma_M)$



NET MEAN CAICLASS

0.28 0.29 0.31 0.33 0.35 0.37 0.39 0.41 0.43 0.45

escherichia coli

Histidine+purine+
pyrimidine biosynthesis

Non-oxidative branch
of the pentose
phosphate pathway

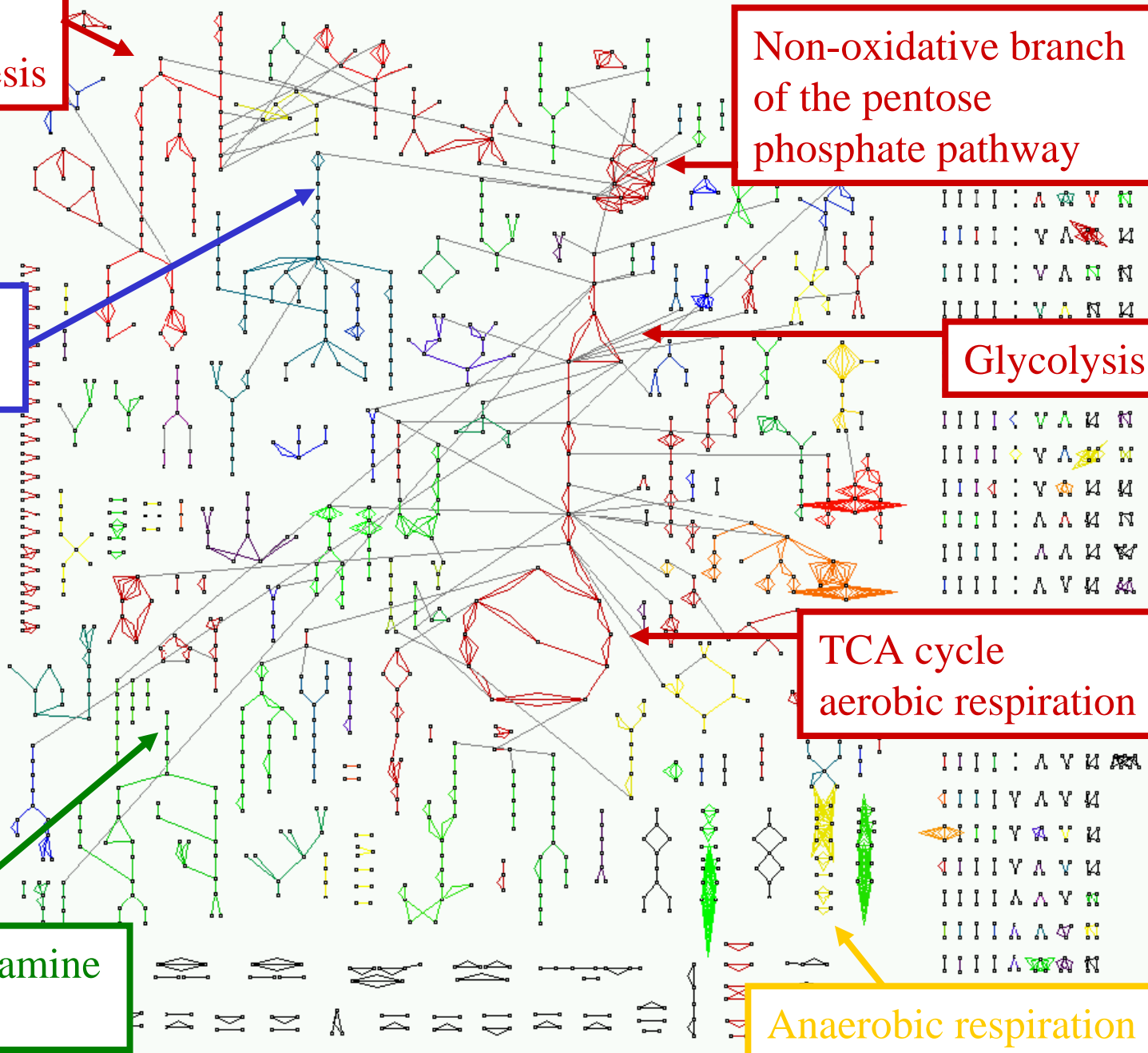
Chorismate and folic
acids biosynthesis

Glycolysis

TCA cycle
aerobic respiration

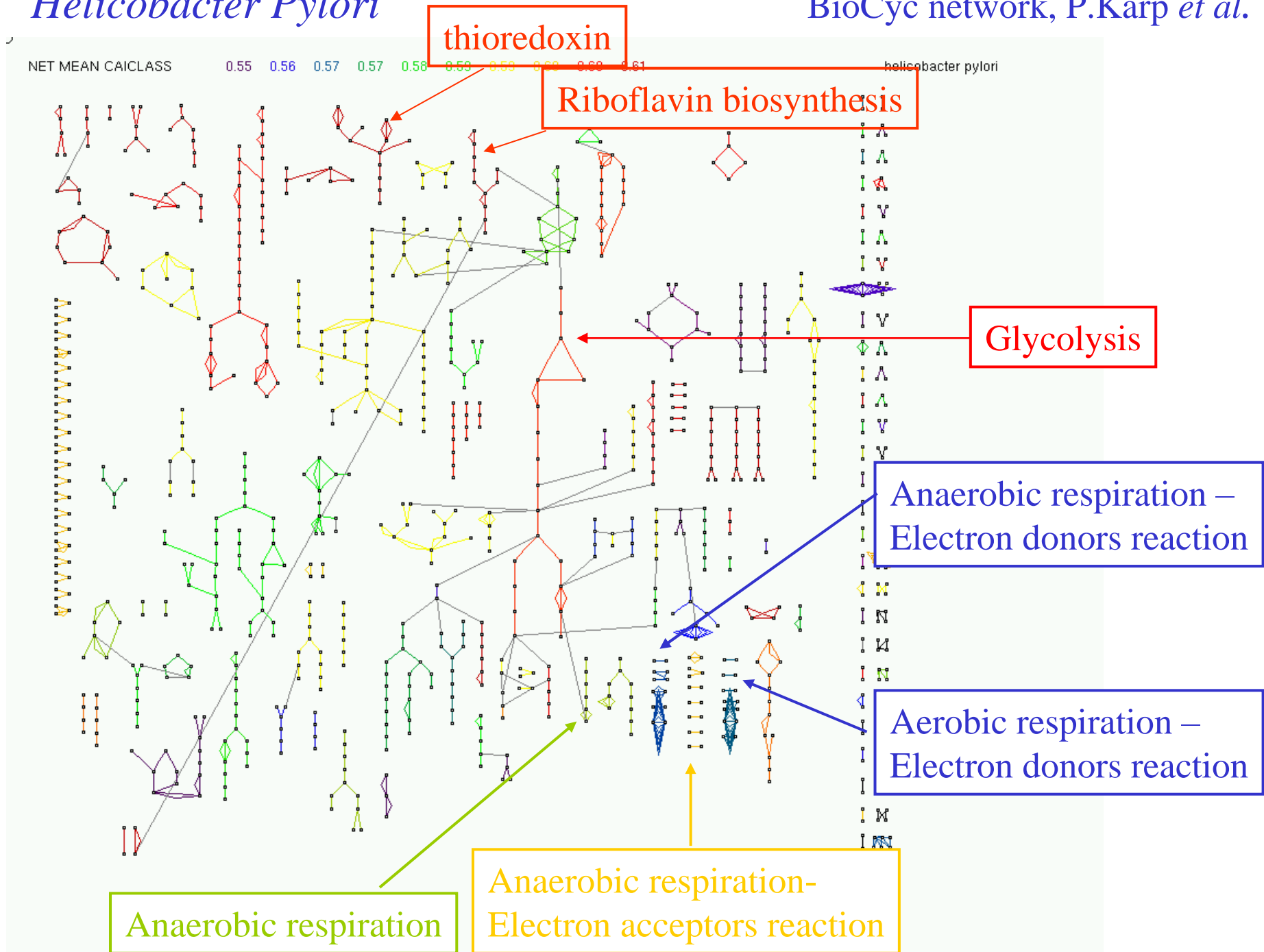
UDP-N-acetylglucosamine
biosynthesis

Anaerobic respiration



Helicobacter Pylori

BioCyc network, P.Karp et al.



Thioredoxin pathway:

it has a thioredoxin-dependent peroxiredoxin system playing a critical role in the defense against oxygen toxicity that is essential for survival and growth, even in microaerophilic environments (Baker et al. 2001).

Riboflavin biosynthesis:

crucial role in ferric-iron reduction and iron acquisition (Worst et al. 1998). As other pathogenic bacteria, *H.pylori* encounters an iron-limiting environment when it attempts to colonize or invade a mammalian host.

Essential pathways for *Mycobacterium tuberculosis*

highly ranked for *M.tuberculosis* and not highly ranked
for 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)

Analysis of *Plasmodium falciparum*

Collaborations

- F.Képès, CNRS and génopole Evry
- D.Madden, IHÉS and Sherbrooke University, Canada
- A.Zinovyev, IHÉS and Institut Curie (Paris)