Genome synthesis and genomic functional cores

Alessandra Carbone
Département d'Informatique &
Equipe de Génomique Analytique, INSERM
Université Pierre et Marie Curie, Paris 6
Alessandra.Carbone@lip6.fr

Plan

1. algorithms + statistics to analyse microbial genomes without using biological information

   robust phenomena

2. space of genomes : from phylogeny to environmental classification

   new mathematical measures are needed

3. essential metabolic networks

4. essential genes

   looking at data from sequences
   &
   difficult comparison with experimental data

5. host-phage co-evolution

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 & RNA silencing

B.subtilis
300 genes/~4000
(Itaya, 1995)
248 genes/~4100
(Kobayashi, 2003)

M.genitalium
265 genes / 517
(Hutchison et al., 1999)
382 genes / 482
(Hutchison et al., 2006)

H.influenzae
670 genes/ ~1272
(Akerley et al. 2002)

E.coli
620 genes / 3746
(Gerdes et al. 2003)
234 genes / 2994
(Hashimoto et al. 2005)

S.cerevisiae
1105 genes/ 5916
(Giaever et al. 2002)

C.elegans
1722 genes/ 19427
(Kamath et al. 2003)

S.aureus
150 genes
(Yi et al. 2001)

S.pneumoniae
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

  ⟶ bacteria live in very good lab conditions

Computational detection of sequence homology:

- parameters and tools to detect homologies

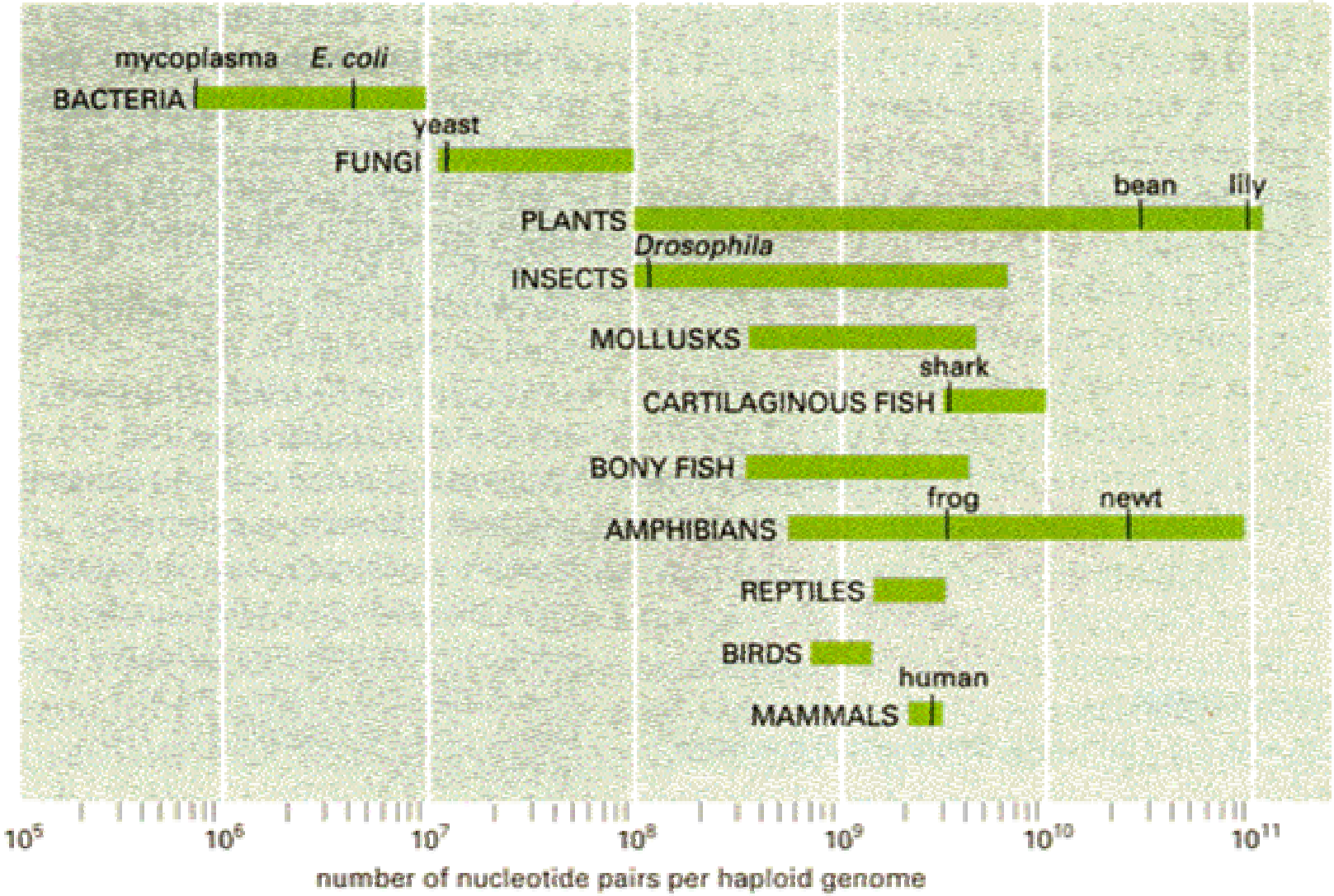
  ⟶ there are genomes with more than 60% of genes with unknown function

Genes relevant to **environmental conditions** are missing

Stress response genes are missing

Genes with **uncharacterized functions** are missing

Background : genomes and lengths



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

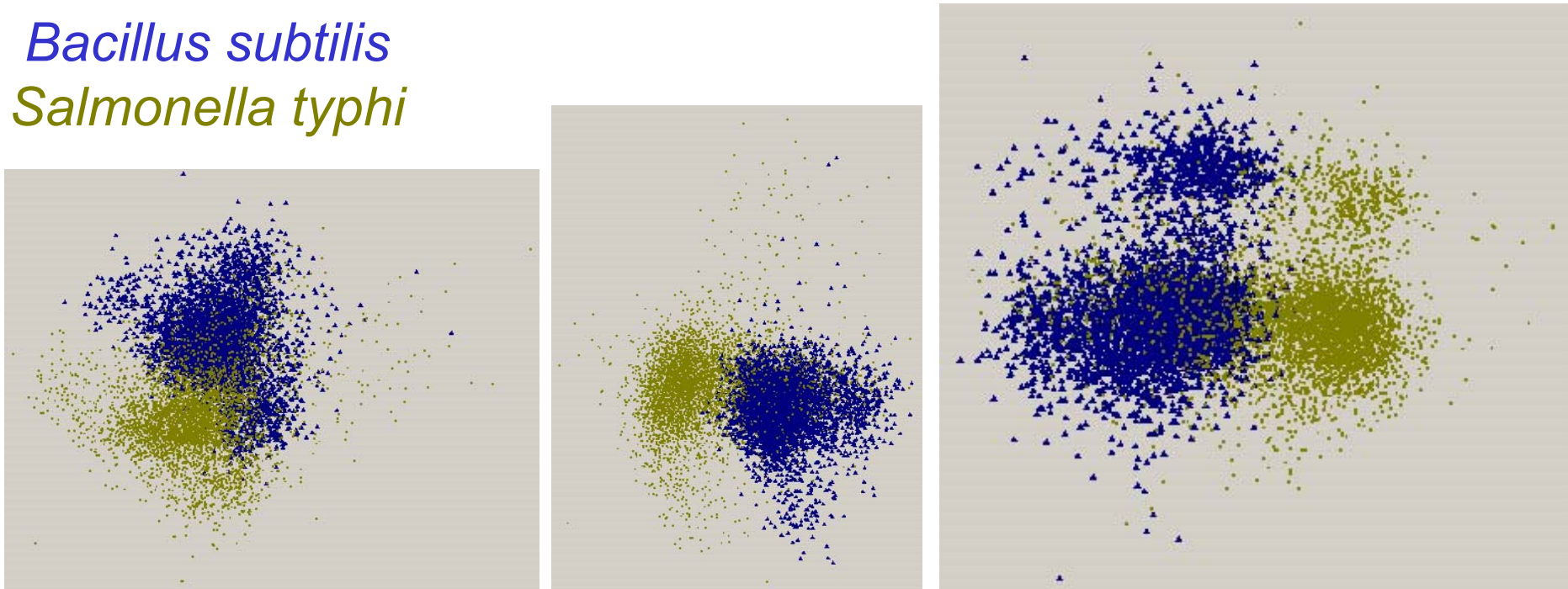
Vector normalisation:

$(x_{i,g} - \underline{x}_i) / \sigma_i$

$\underline{x}_i$ mean of frequencies $x_{i,g}$

σ_i standard deviation of $x_{i,g}$

Bacillus subtilis
Salmonella typhi



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

Vector normalisation:

$(x_{i,g} - \underline{x}_i) / \sigma_i$

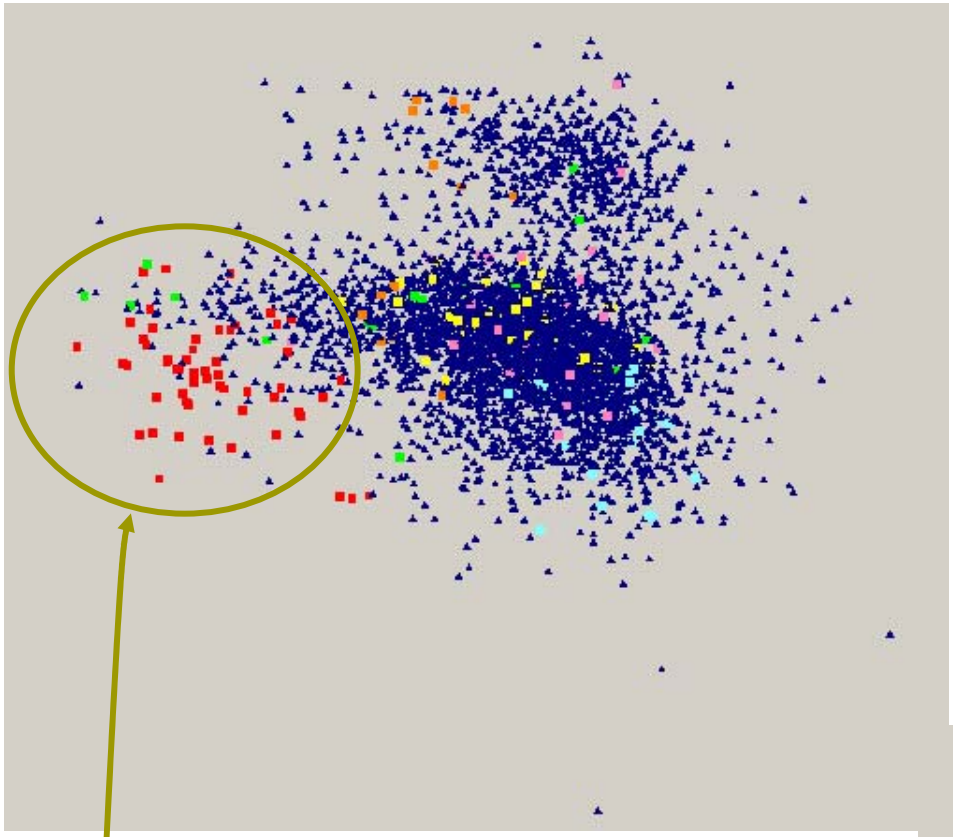
$\underline{x}_i$ mean of frequencies $x_{i,g}$

σ_i standard deviation of $x_{i,g}$

similar geometry

&

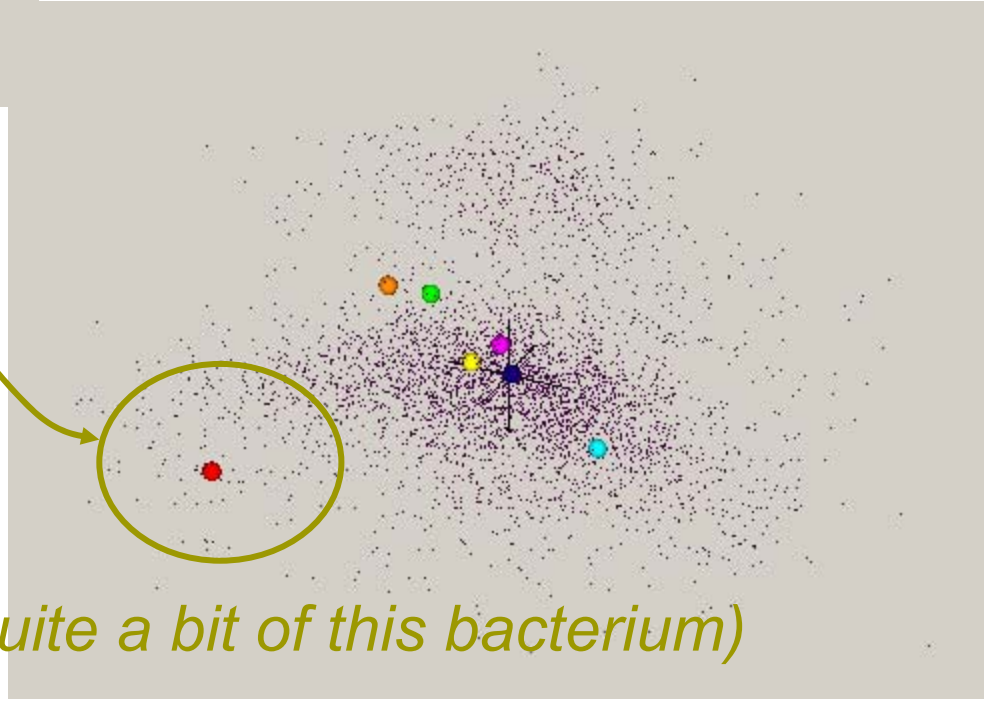
translation + rotation



Ribosomal proteins
ATP binding proteins
IS proteins
NADH proteins
Flagellar biosynthesis proteins
Lipoproteins, membrane proteins, transport proteins

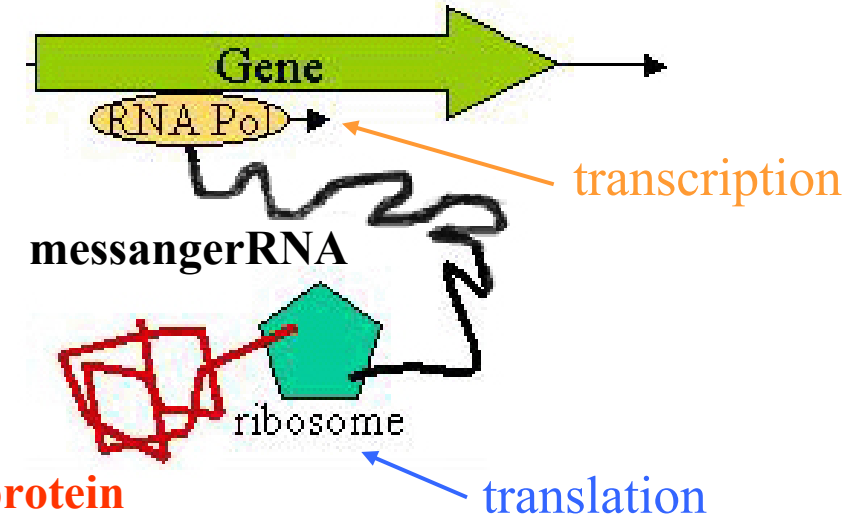
Proteins codifying for “translation”, glycolysis …

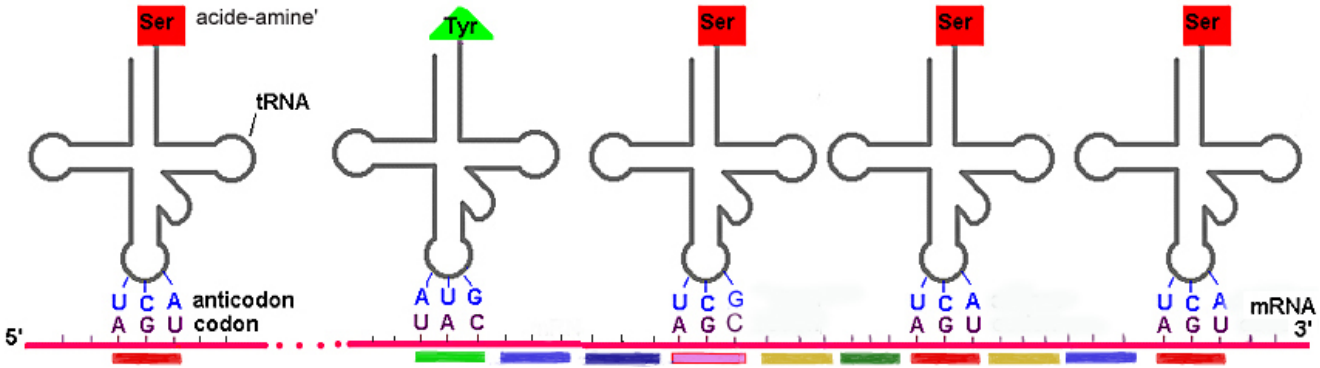
Roughly speaking, they are the most expressed in *E.coli*



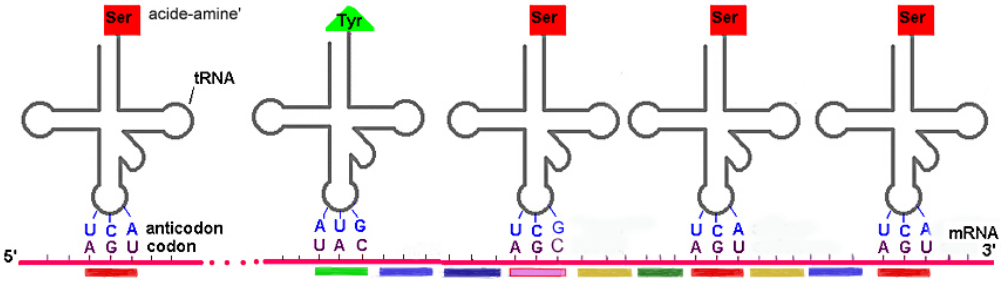
E.coli *(biologists know quite a bit of this bacterium)*

Background : translation





Background : redundant genetic code



translation

2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

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

Background : bias on codon usage & preferred codons

In *E.coli* and other organisms that reproduce rapidly

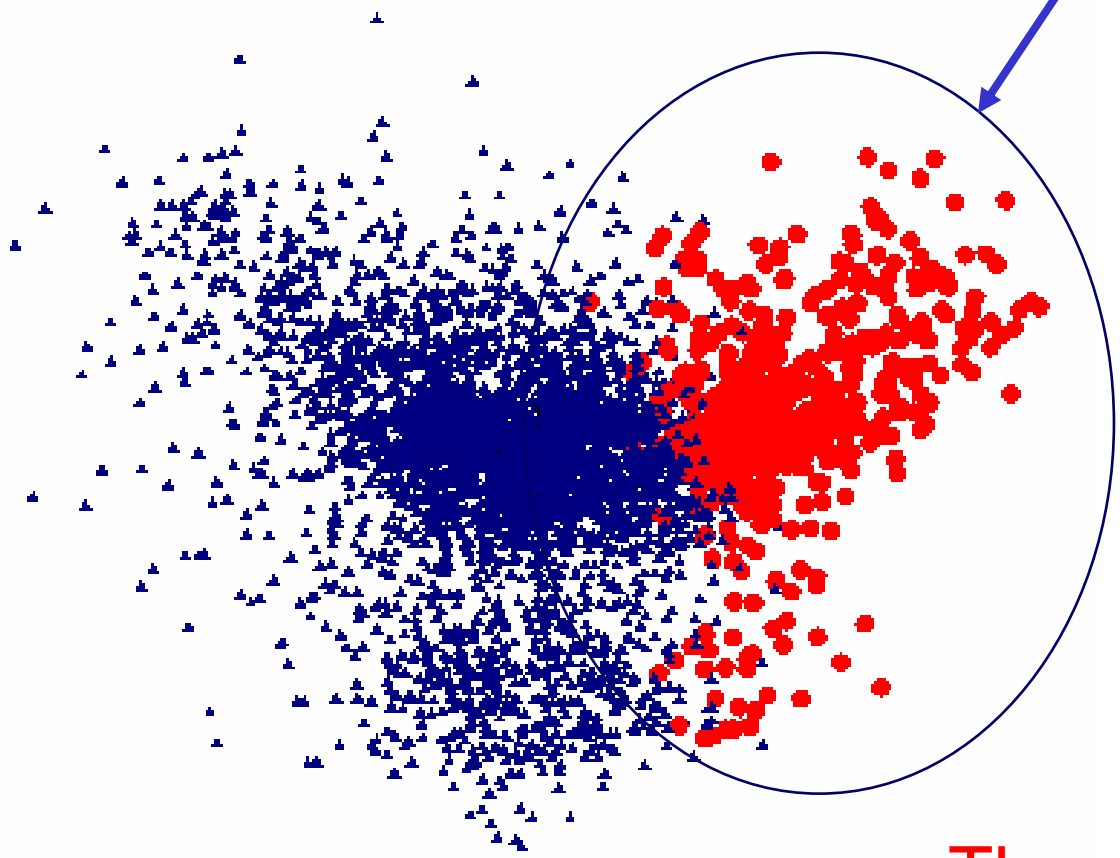
high tRNA number
high expression

correlated to
(experimentally)

codon preference

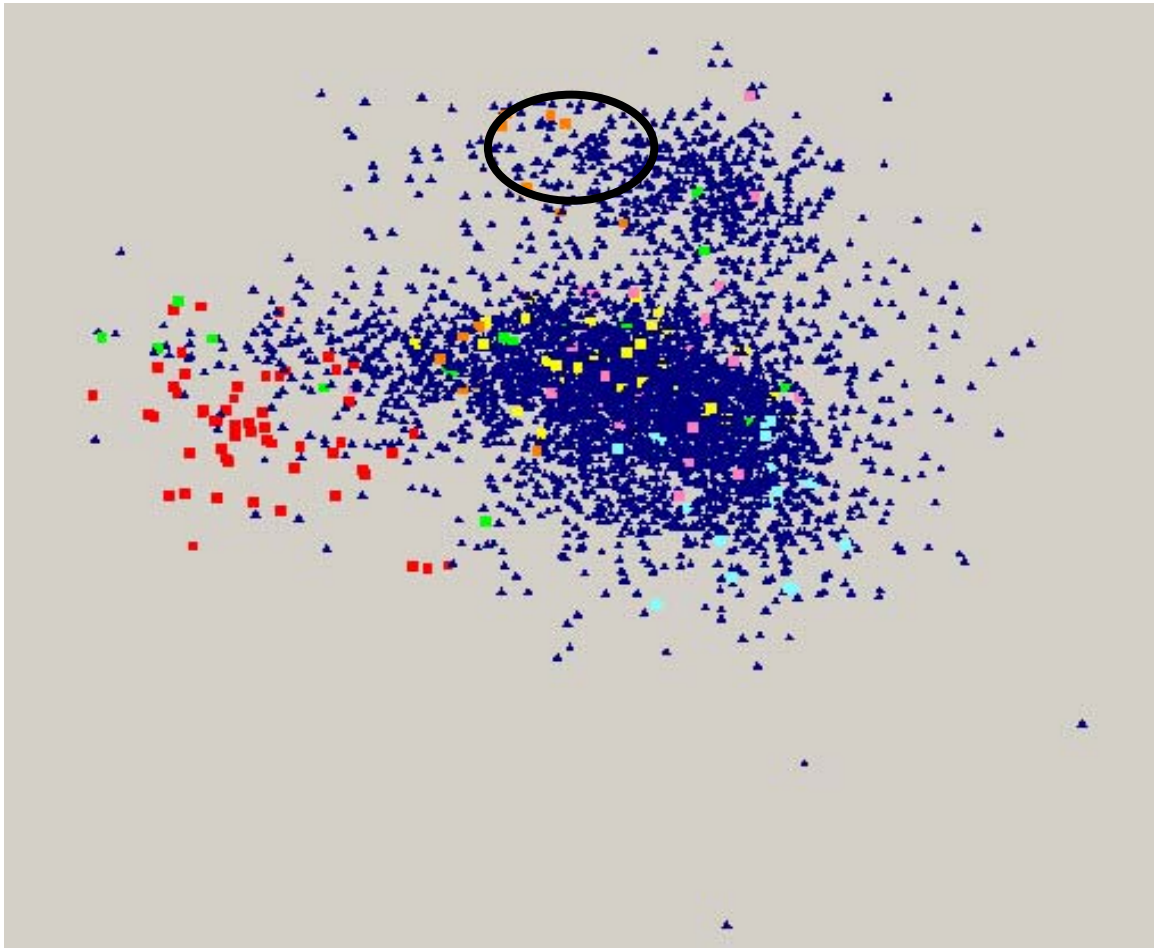
Codon preference and tRNA : Ikemura, 1985; Bennetzen and Hall, 1982; Bulmer, 1987; Gouy and Gautier, 1982.
tRNA and elongation rate : Varenne *et al*., 1984.
High expression and codon preference : Grantham *et al*., 1980; Wada *et al*., 1990; Sharp and Li, 1987; Sharp *et al*., 1986; Médigue *et al*., 1991; Shields and Sharp, 1987; Sharp *et al*., 1988; Stenico *et al*., 1994.

Are they among the **most essential** ?

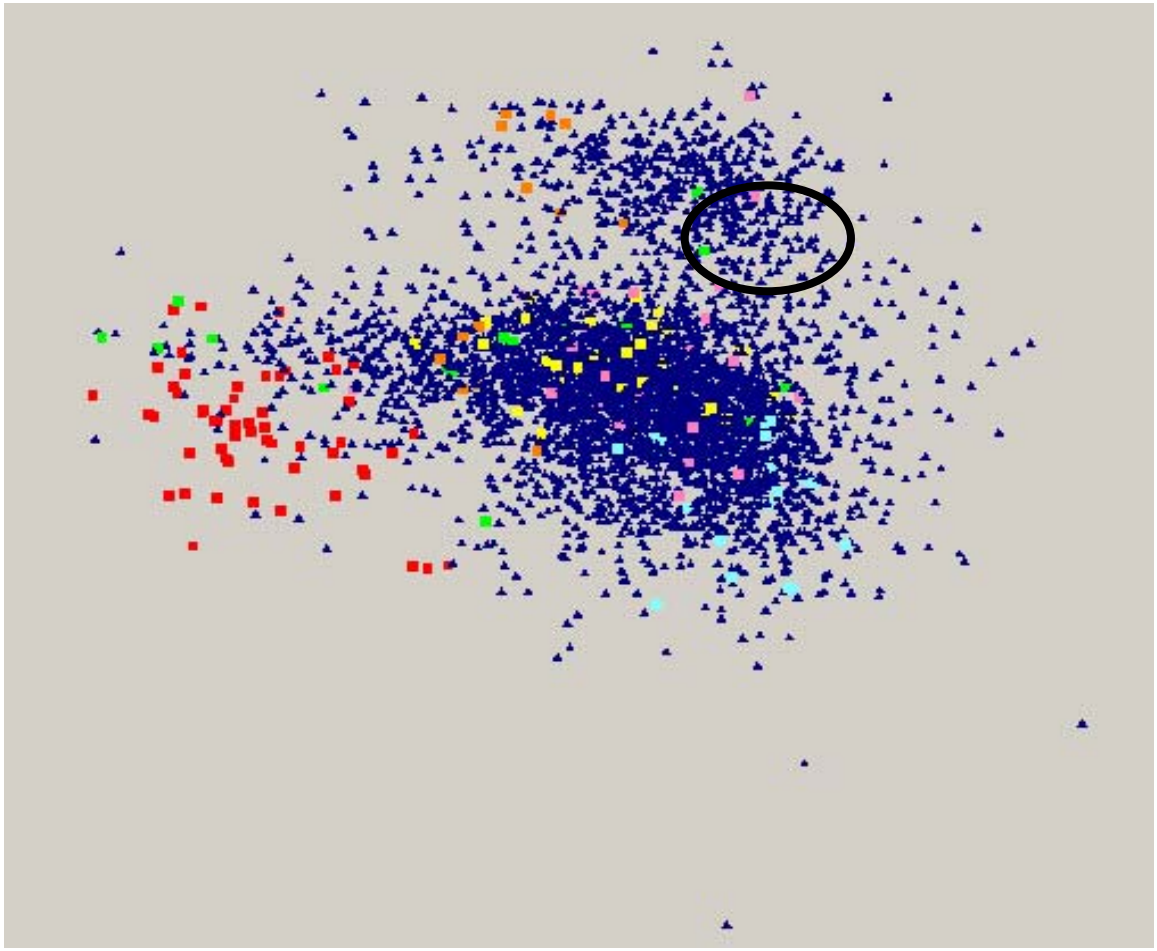


The most expressed

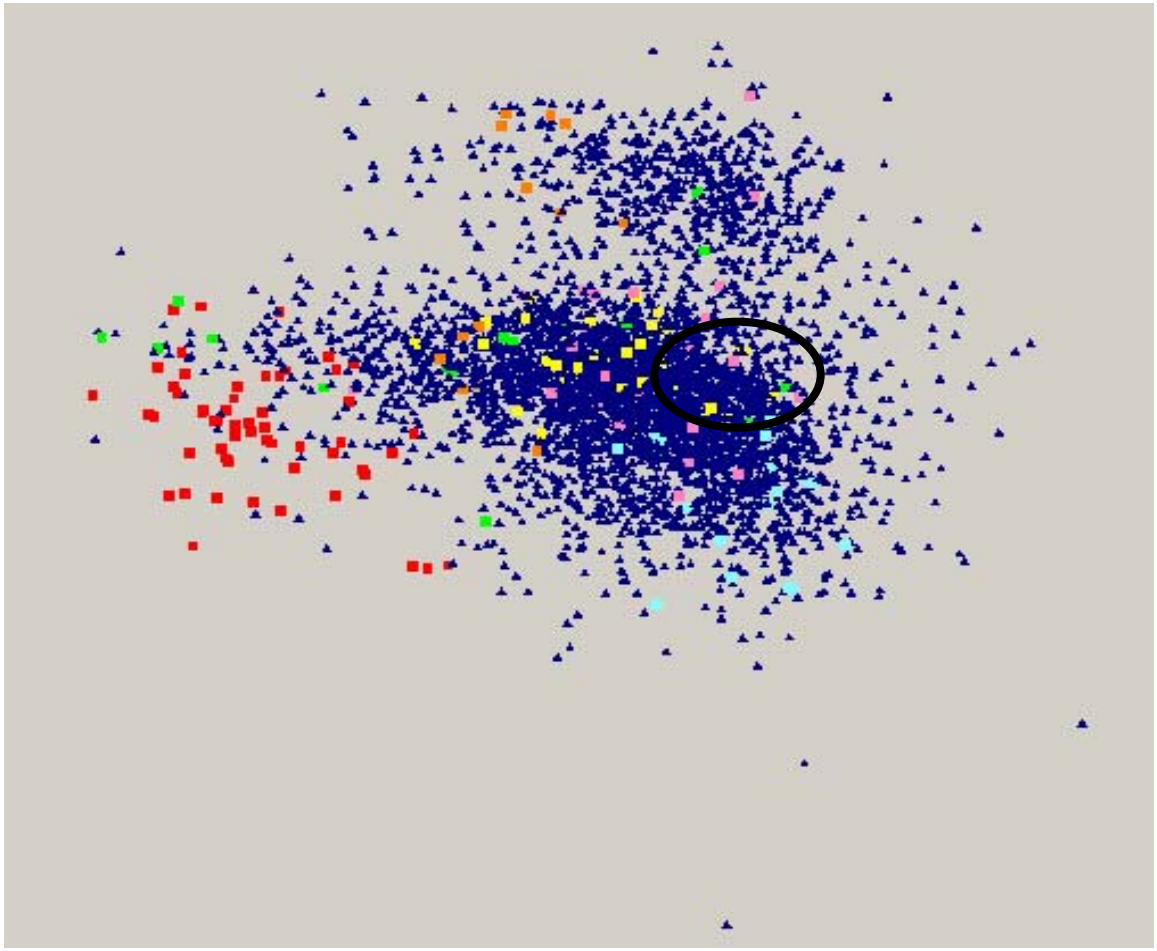
How to define “codon bias” and how to search for highly biased genes in an automatic manner?



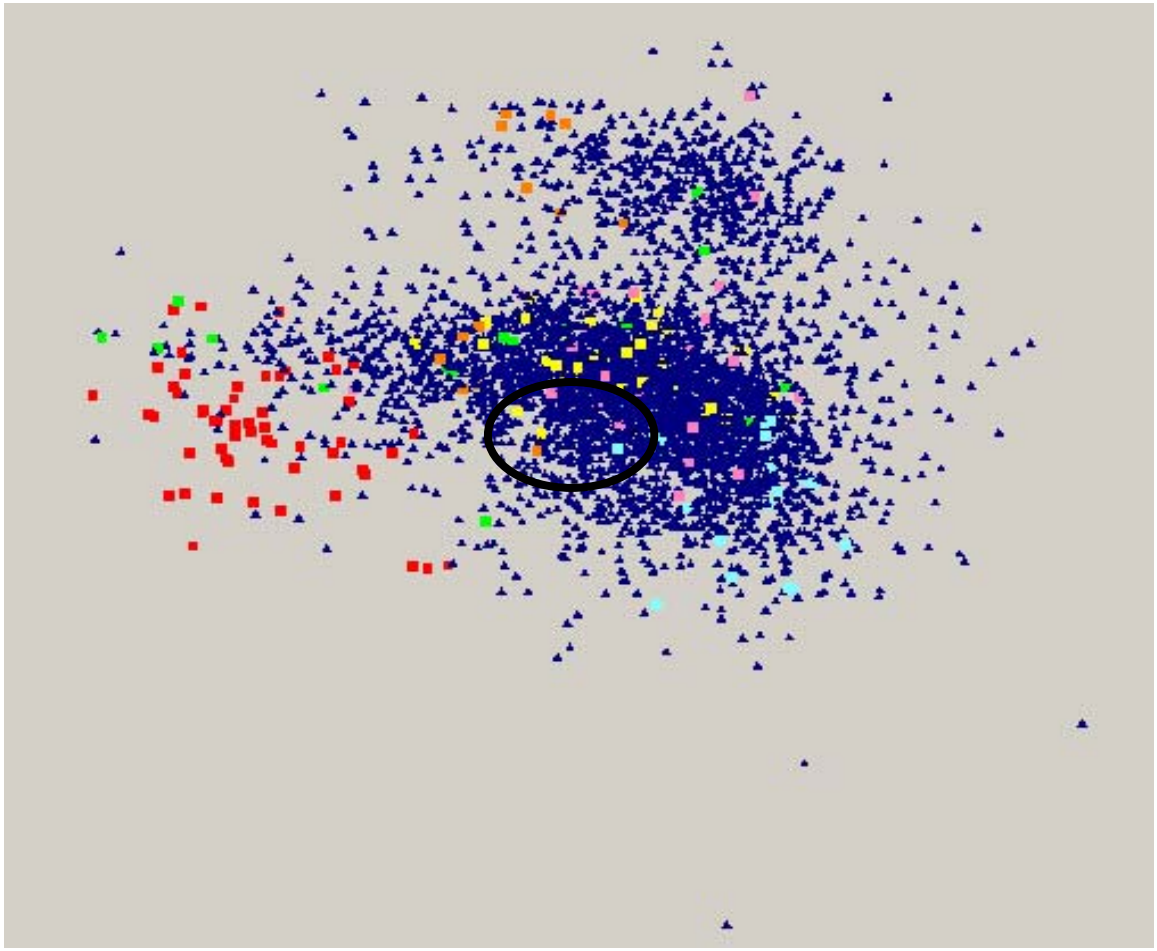
How to define “codon bias” and how to search for highly biased genes in an automatic manner?



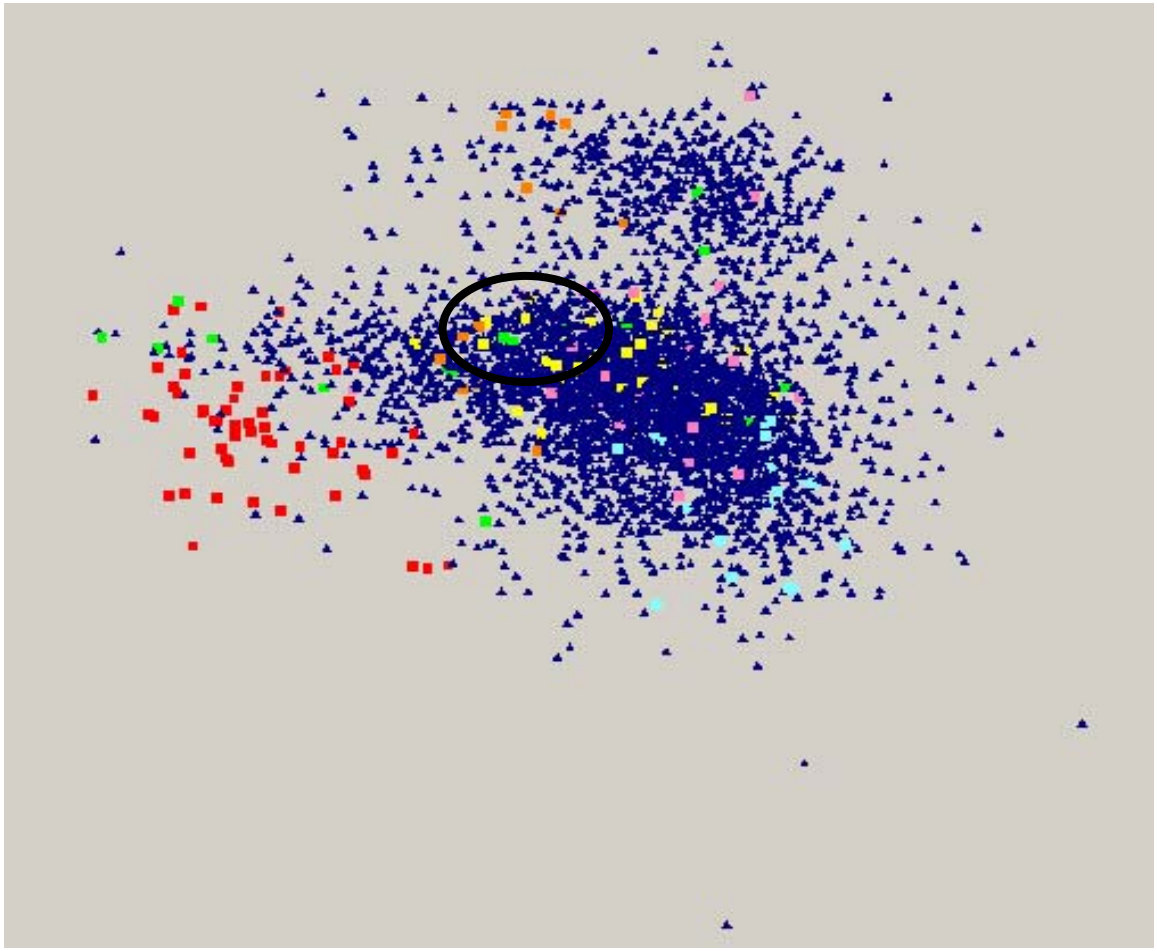
How to define “codon bias” and how to search for highly biased genes in an automatic manner?



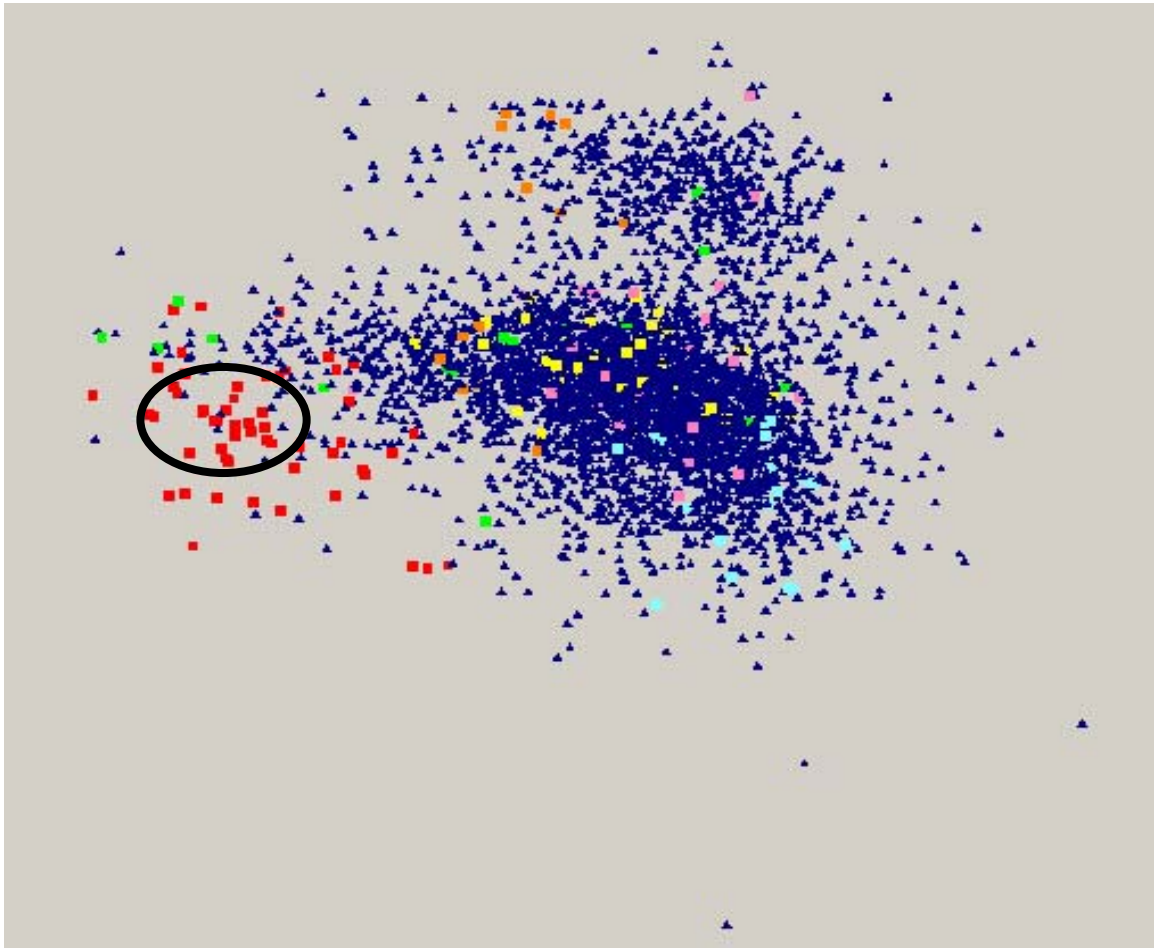
How to define “codon bias” and how to search for highly biased genes in an automatic manner?

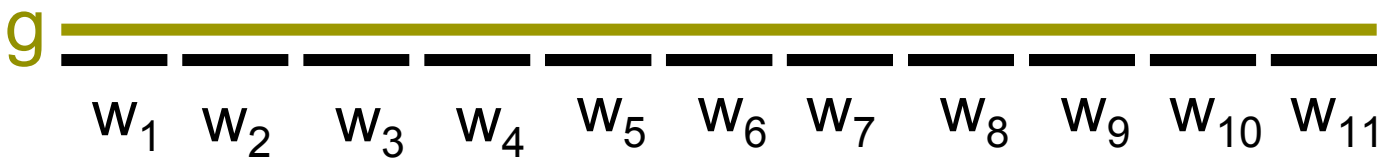


How to define “codon bias” and how to search for highly biased genes in an automatic manner?



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





$$(\Pi_{k=1\ldots11}\, w_k)^{1/11}$$

$$CAI(g) = (\Pi_{k=1\ldots L}\, w_k)^{1/L}$$

Codon Adaptation Index

(Sharp & Li, 1987)

L number of codons in g

w_k $\dfrac{\text{frequency of the } k^{th} \text{ codon of g in S}}{\text{frequency of the dominant synonymous codon in S}}$

proteines codifying for “translation”, glycolysis …

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

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

Codon Adaptation Index

L number of codons in g

w_k $\dfrac{\text{frequency of the } k^{th} \text{ codon of g in S}}{\text{frequency of the dominant synonymous codon in S}}$

proteines codifying for “translation”, glycolysis …

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

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

Codon Adaptation Index

L number of codons in g

w_k $\dfrac{\text{frequency of the } k^{th} \text{ codon of g in S}}{\text{frequency of the dominant synonymous codon in S}}$

we compute S

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

$$SCCI(g) = (\Pi_{k=1\ldots L}\, w_k)^{1/L}$$

Self Consistent Codon Index

L number of codons in g

w_k $\dfrac{\text{frequency of the } k^{th} \text{ codon of g in S}}{\text{frequency of the dominant synonymous codon in S}}$

we compute S

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

$$SCCI(g) = (\Pi_{k=1...L} w_k)^{1/L}$$

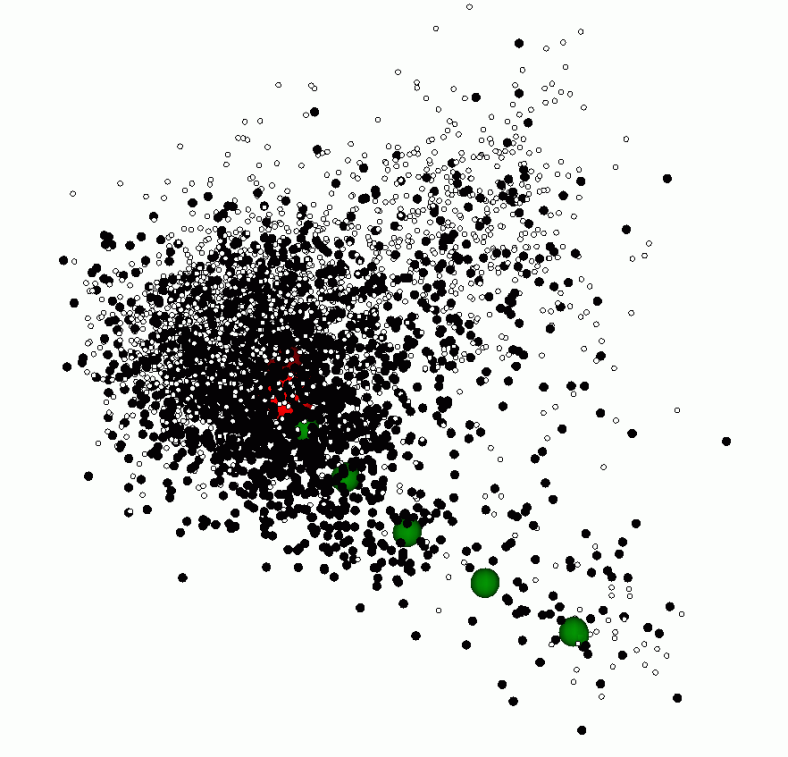
Self Consistent Codon Index

Self consistency condition

SCCI values on genes in S are **maximal** :
$SCCI(G/S) \leq SCCI(S)$, G is the set of all genes

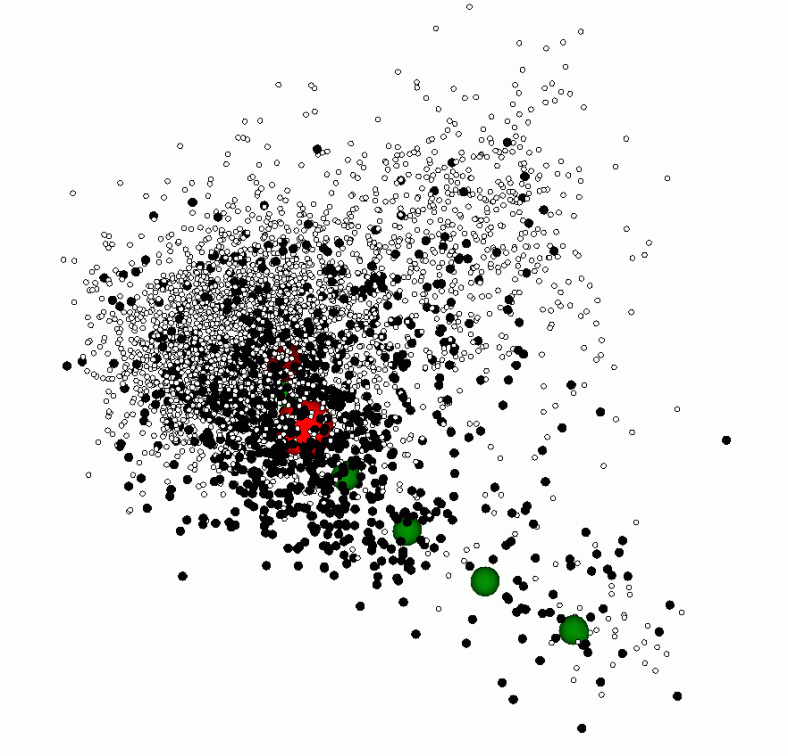
Idea of the algorithm:

- 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.



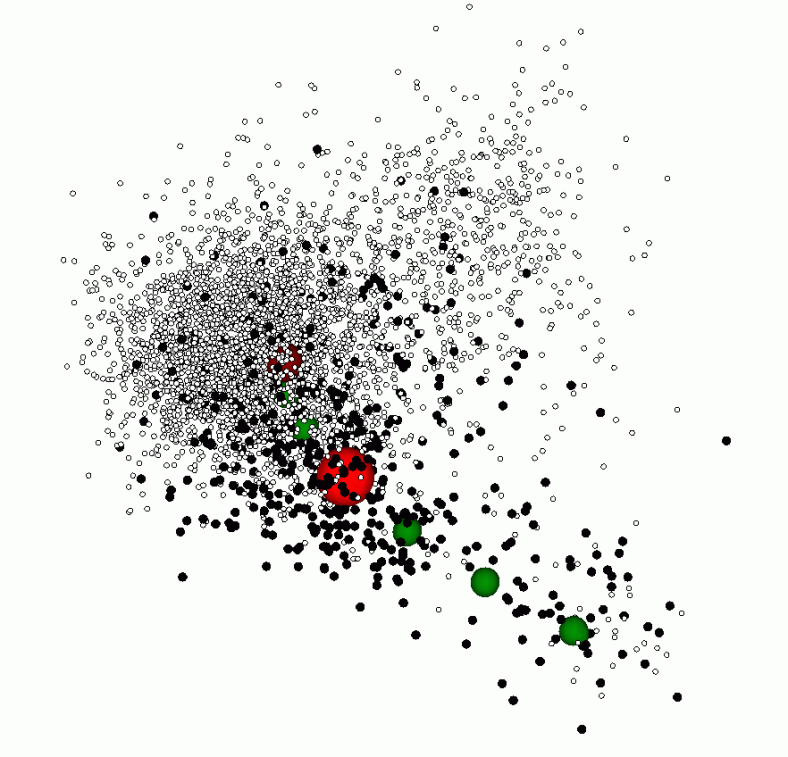
Idea of the algorithm:

- 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.



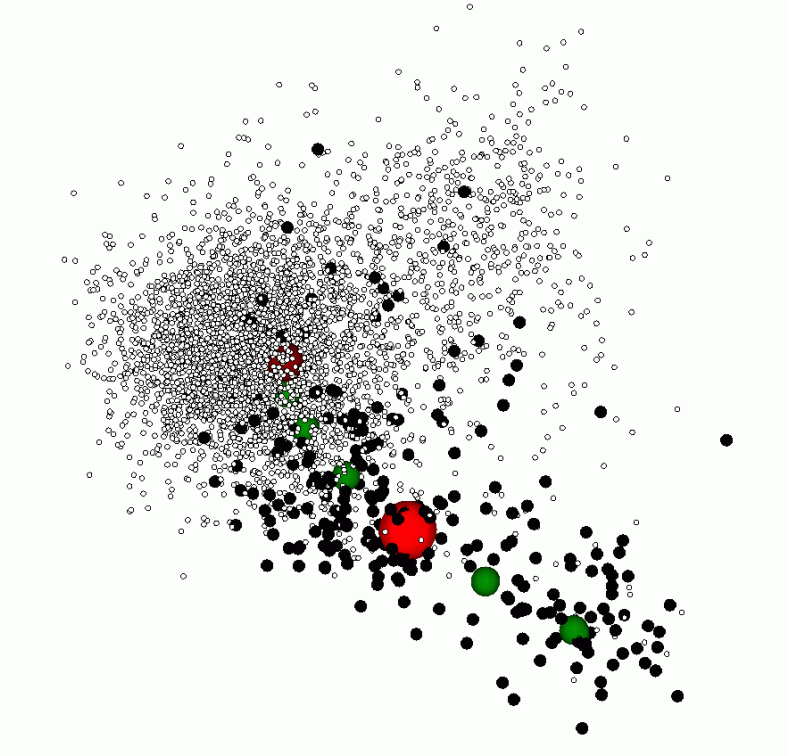
Idea of the algorithm:

- 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.



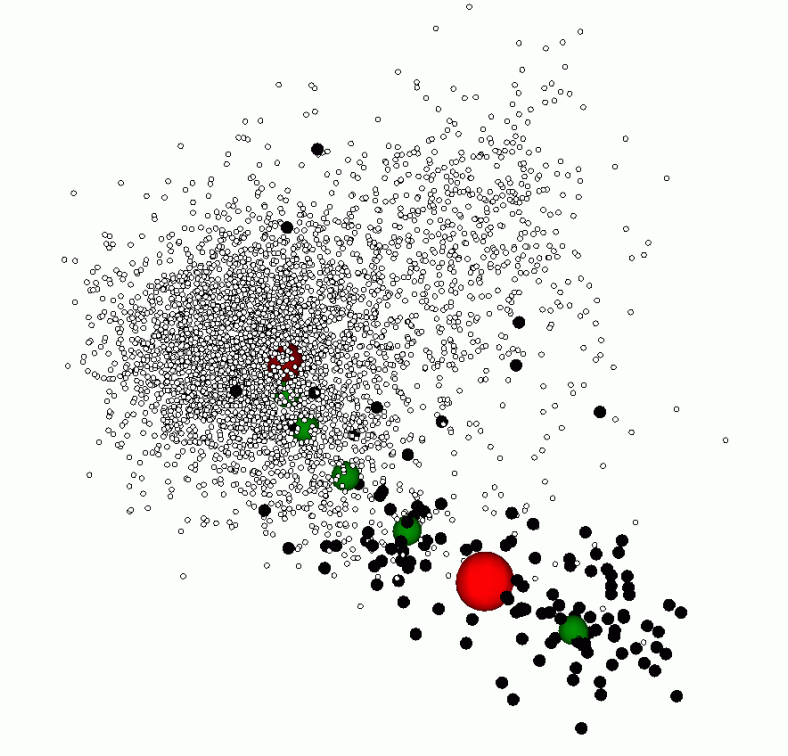
Idea of the algorithm:

- 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.



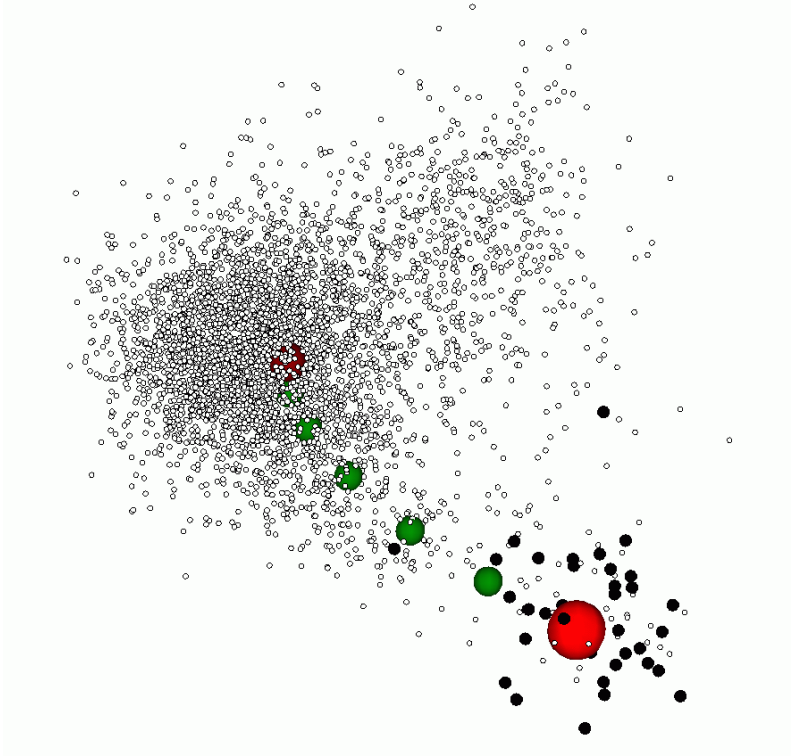
Idea of the algorithm:

- 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.



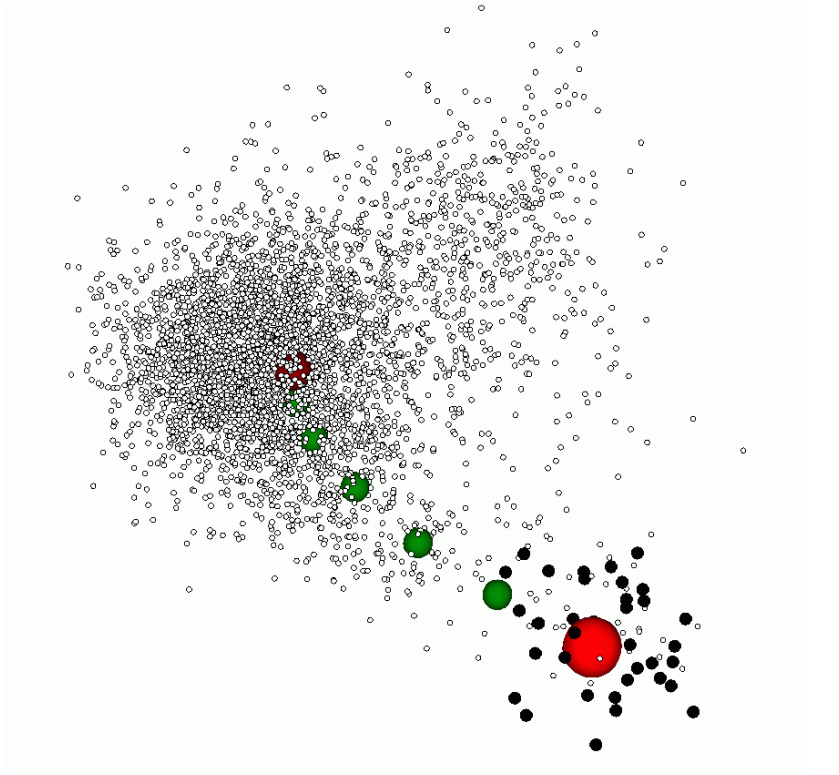
Idea of the algorithm:

- 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.



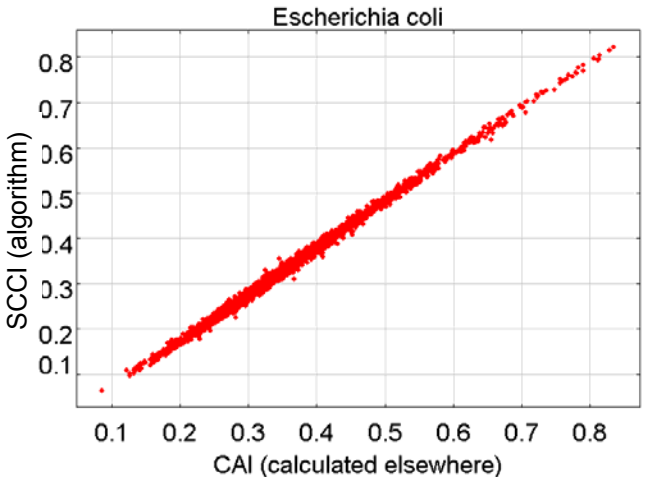
Idea of the algorithm:

- 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.



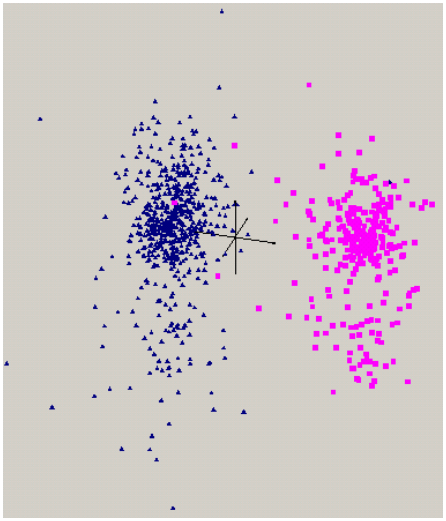
S found by the algorithm: *E.coli*

(*E.coli* reproduce rapidly)



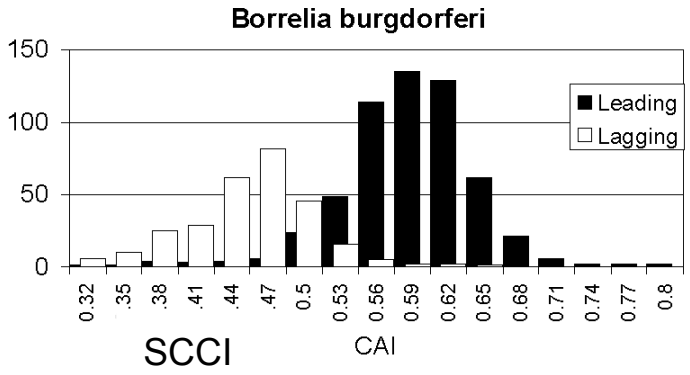
Gene	Annotation
tufA	protein chain elongation factor EF-Tu
tufB	protein chain elongation factor EF-Tu
tsf	protein chain elongation factor EF-Ts
fusA	GTP-binding protein chain elongation factor EF-G
mopA	chaperonin GroEL
dnaK	heat shock protein DnaK
cspA	cold shock protein 7.4
tig	trigger factor
ompA	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

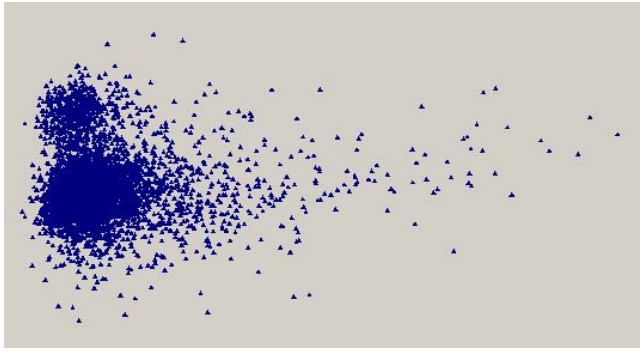
SCCI : a universal measure



Borrelia burgdorferi

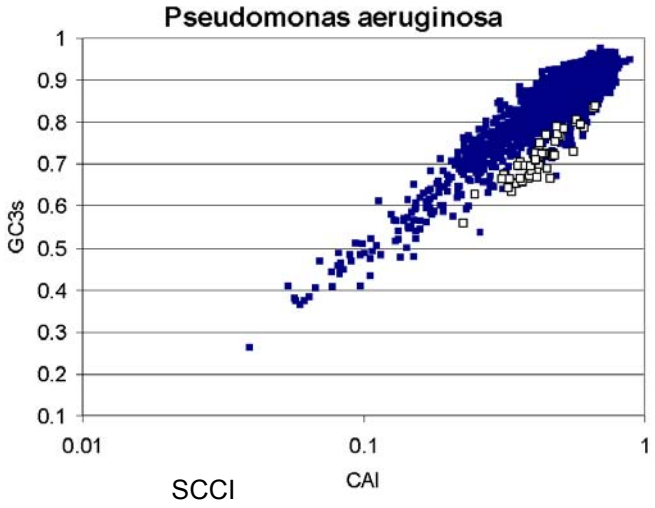
Strand bias





Pseudomonas aeruginosa

GC3 bias



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.

For **any** bacteria we can compute:

+ dominant bias: strand bias, GC3, AT, ...
+ numerical criteria to determine the strength of translational bias

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

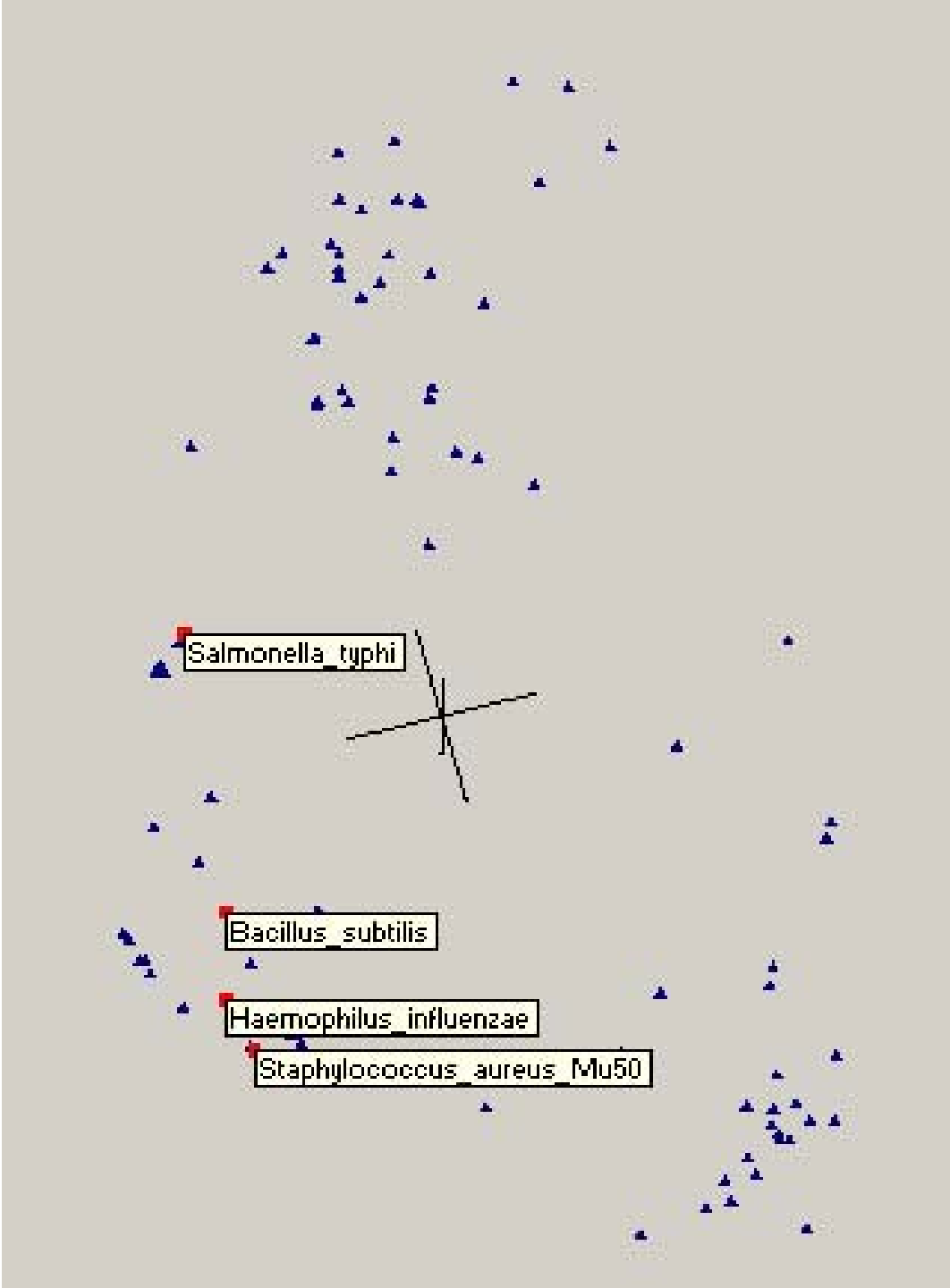
Bacteria and Archaea in SCCI codon space

An organism is a 64-dim vector where

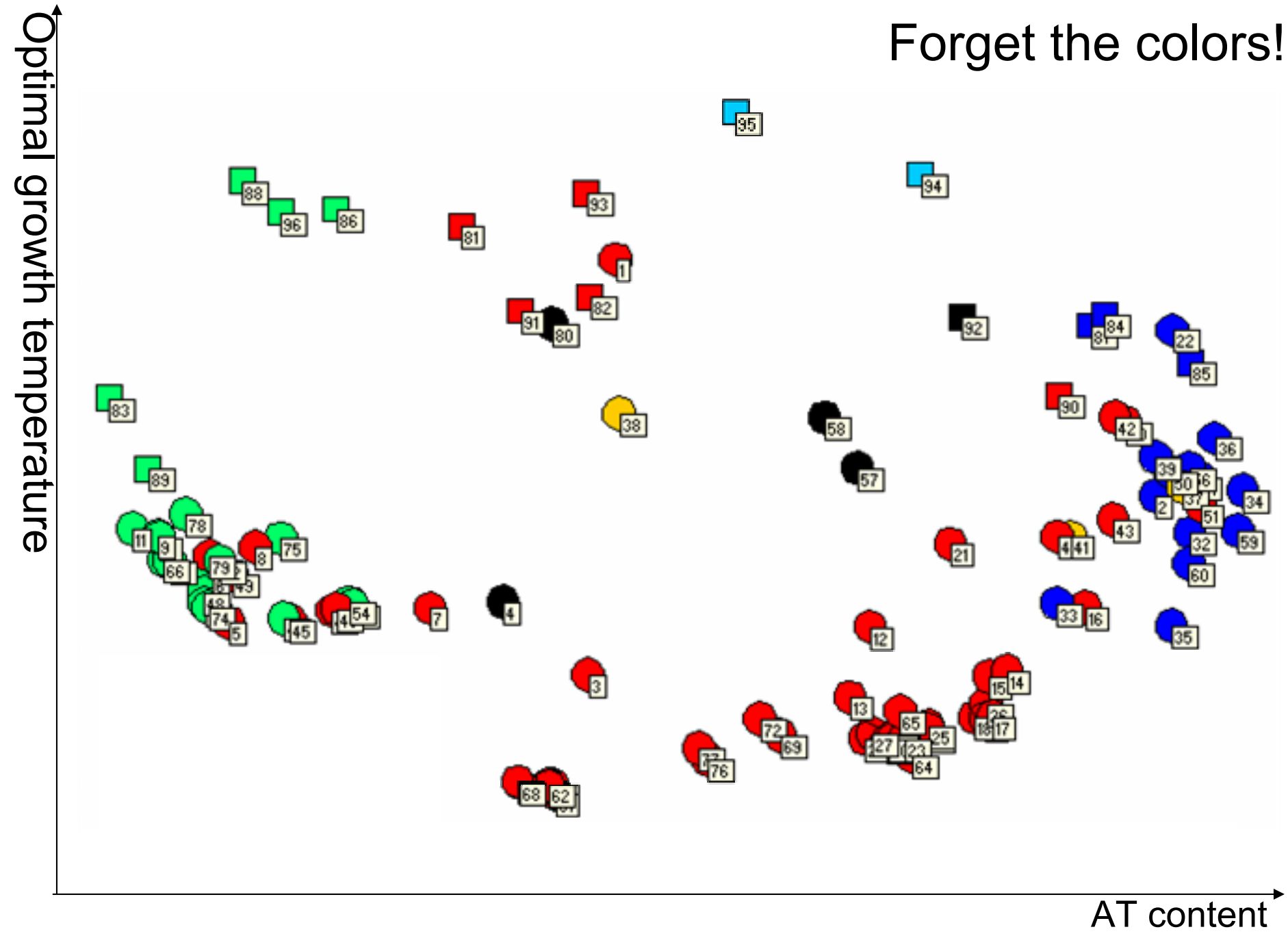
coordinate

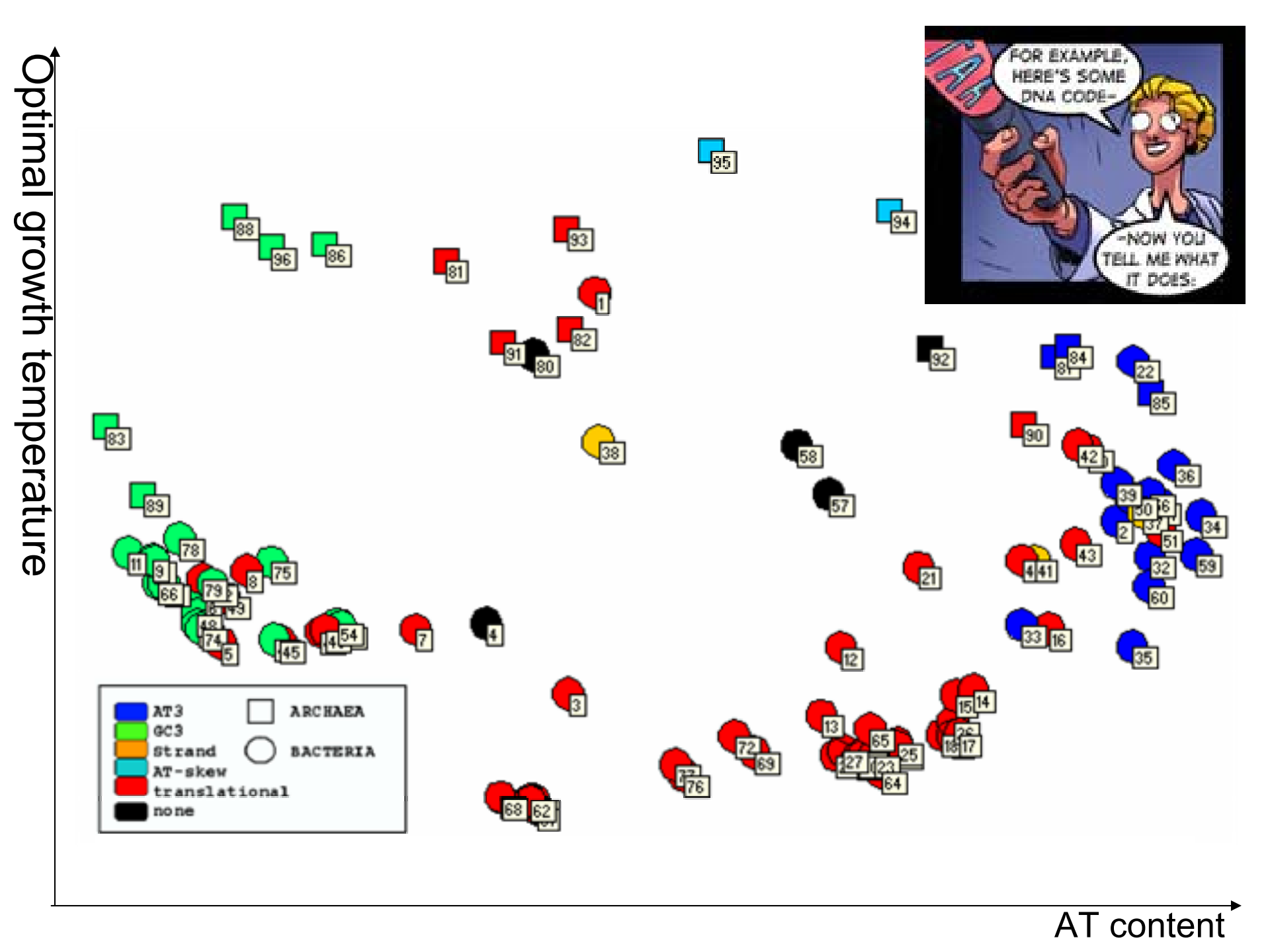
=

SCCI codon weight



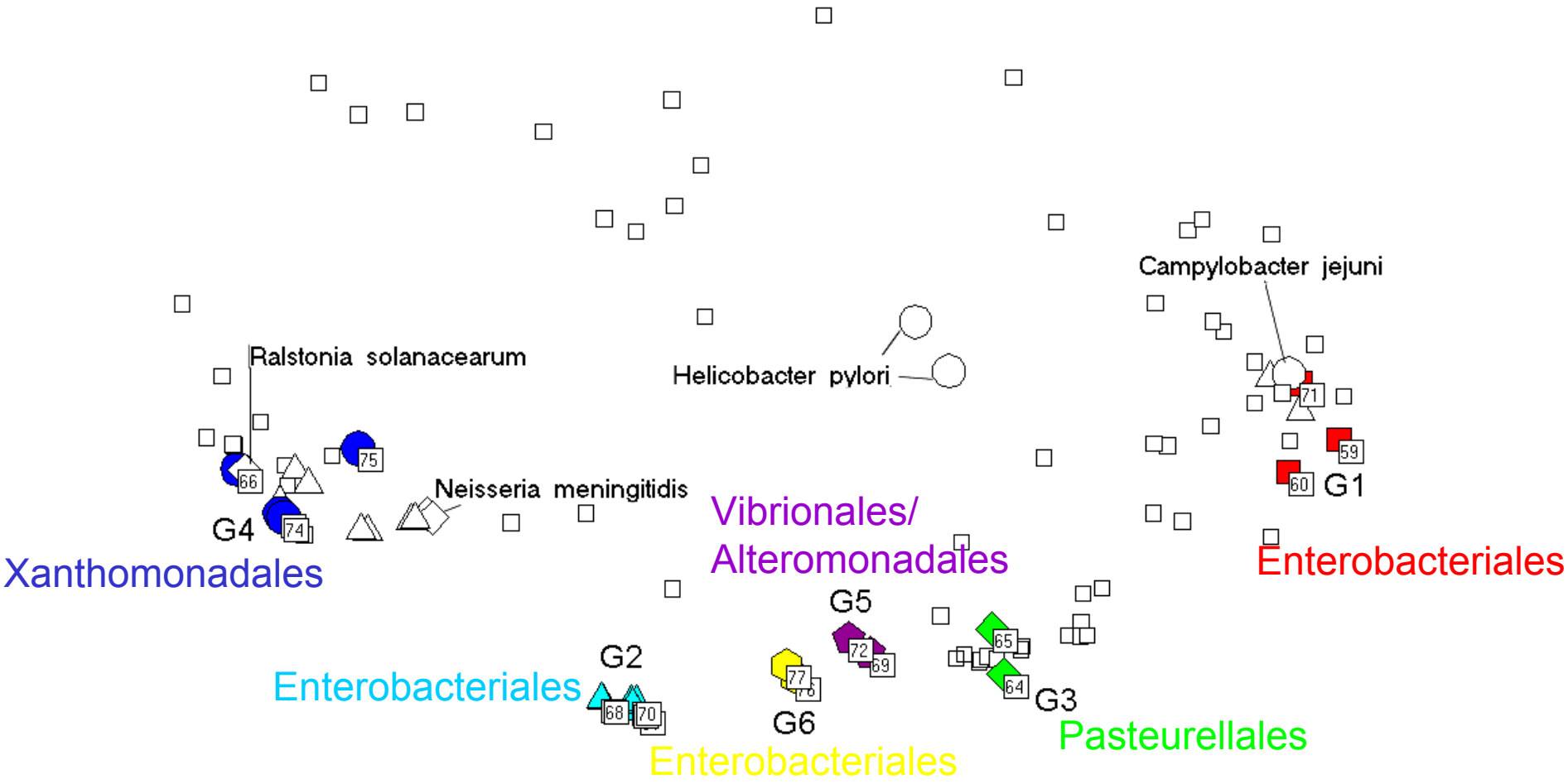
Forget the colors!




Optimal growth temperature
AT content
AT3
GC3
Strand
AT-skew
translational
none
ARCHAEA
BACTERIA
FOR EXAMPLE, HERE'S SOME DNA CODE-
-NOW YOU TELL ME WHAT IT DOES:

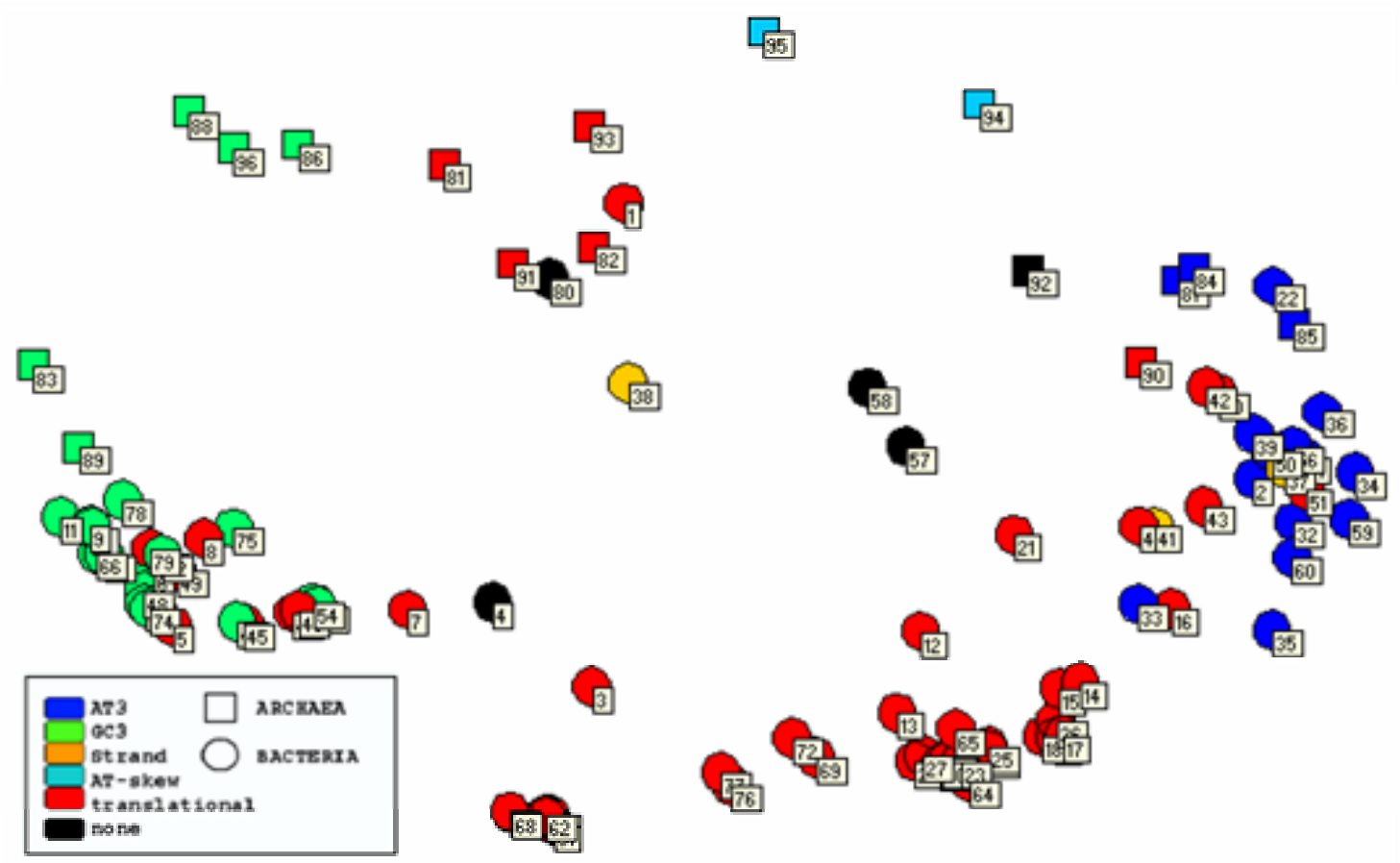
Can we exploit the geometry of the space to derive functional characteristics of groups of organisms?

Phylogenetically related families : γ-proteobacteria



Similar physiology and habitat

Organisms at small distance: similar physiology and habitat



Environmental clusters :

soil bacteria
enterics
symbions

spore formers
small intercellular pathogens
small extracellular pathogens

Coherence in the organisms space based on SCCI

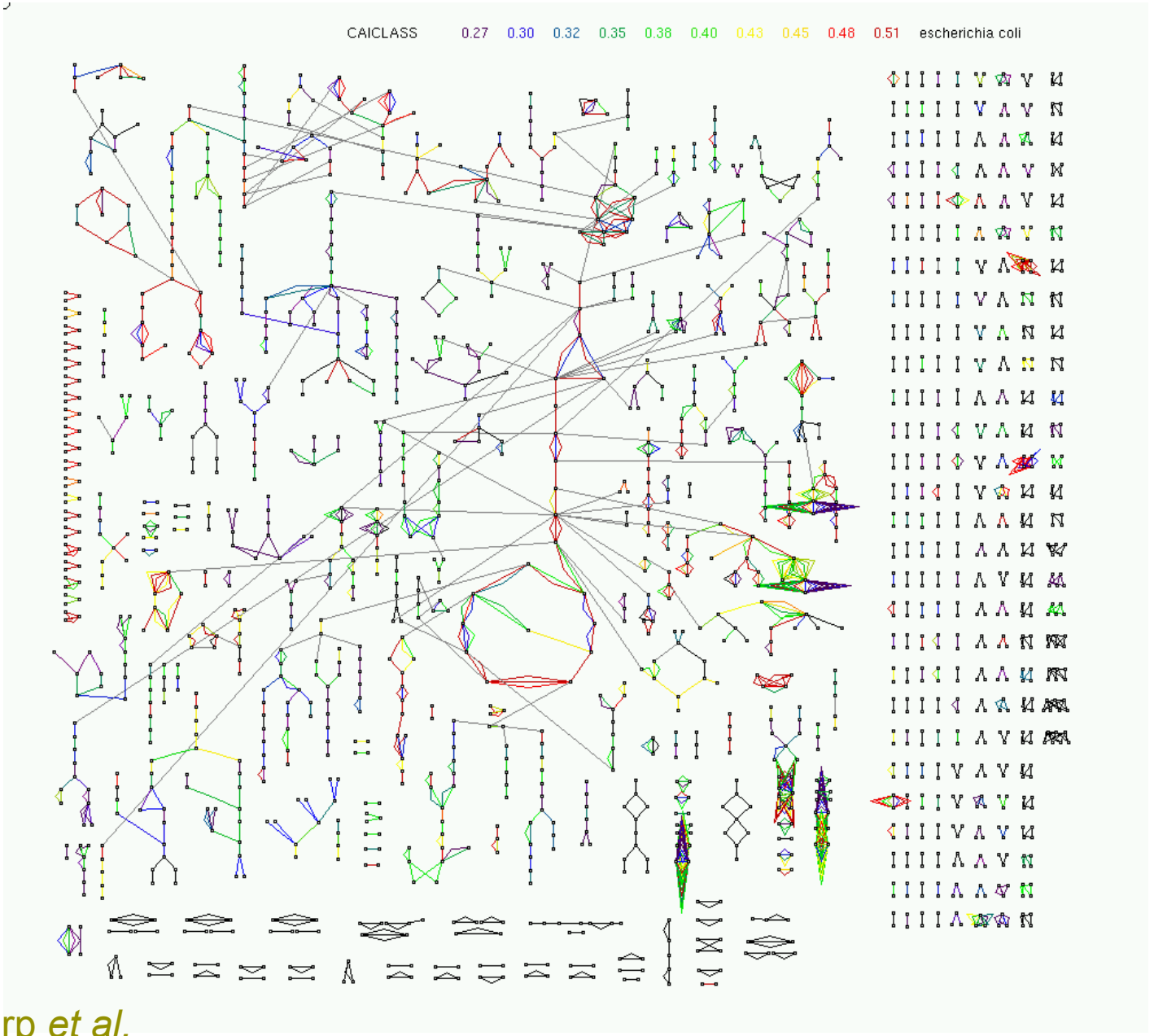
Can we use this signal to deduce some more biological information ?

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

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

Metabolic networks

E.coli



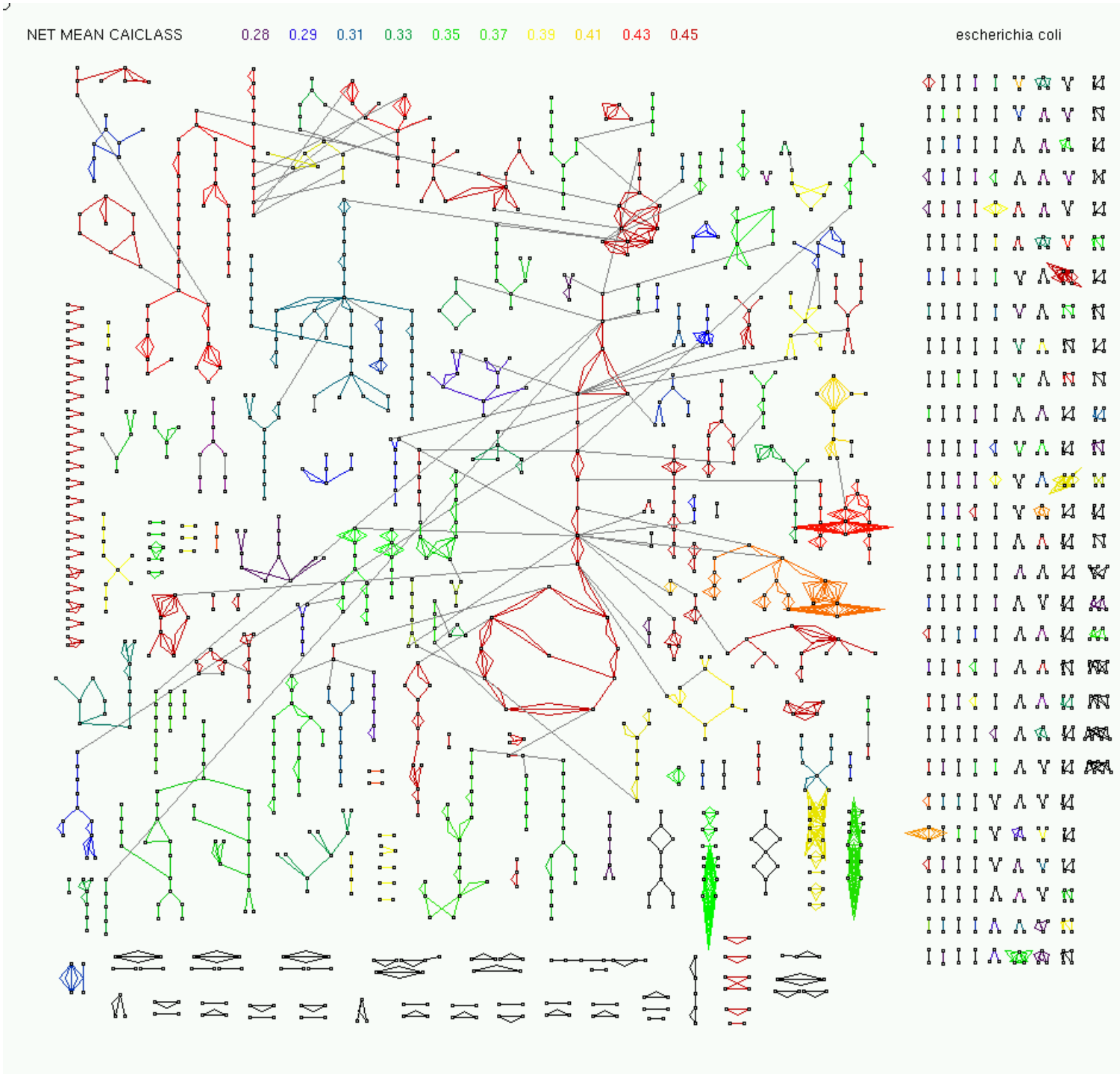
EcoCyc network, P.Karp *et al.*

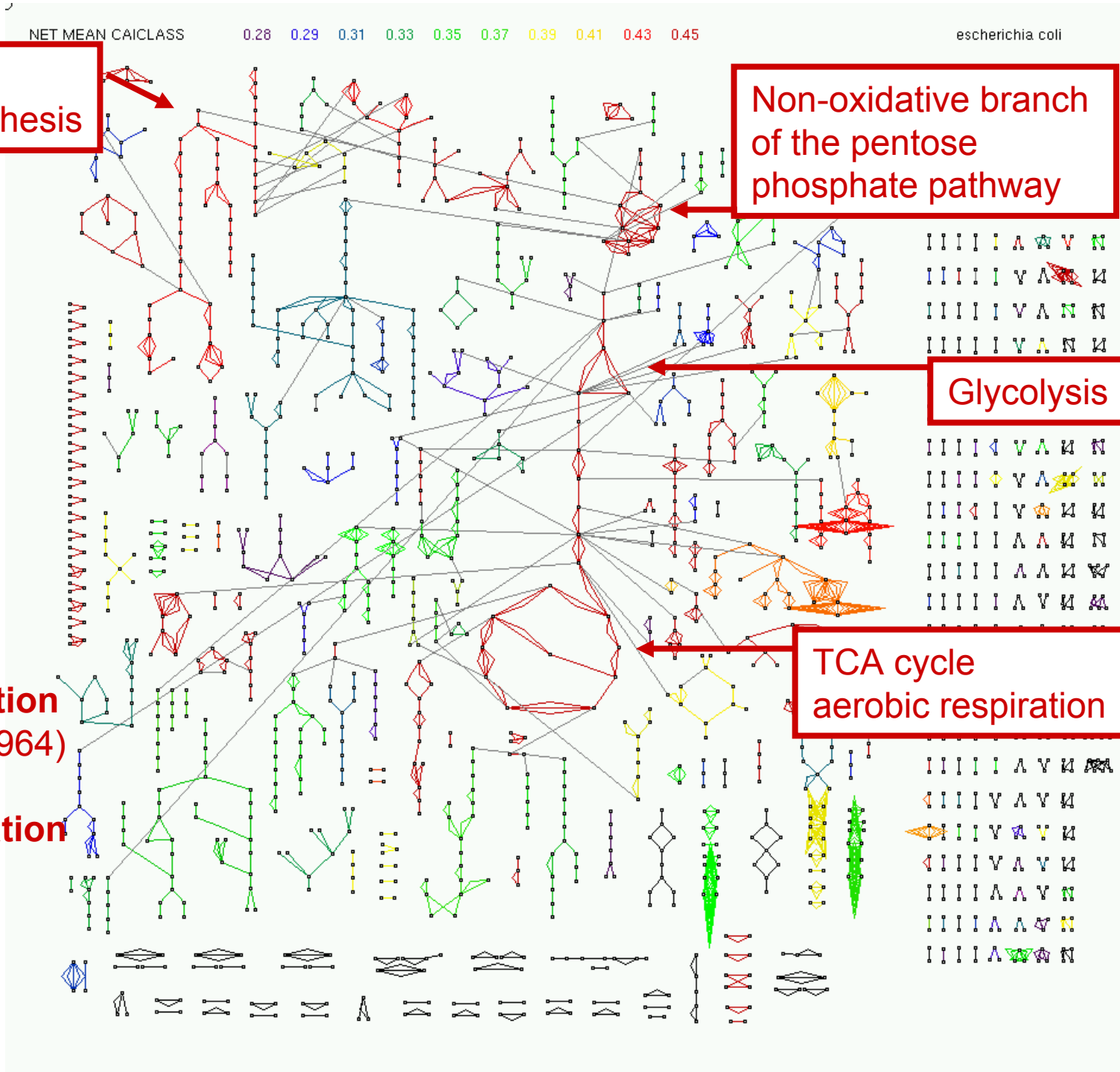
Pathway Index

$$PI(P) = \text{mean } SCCI(g)_{g \in P}$$

Relative Pathway Index

$$RPI(P) = (PI(P)-\mu_M)/\sigma_M$$





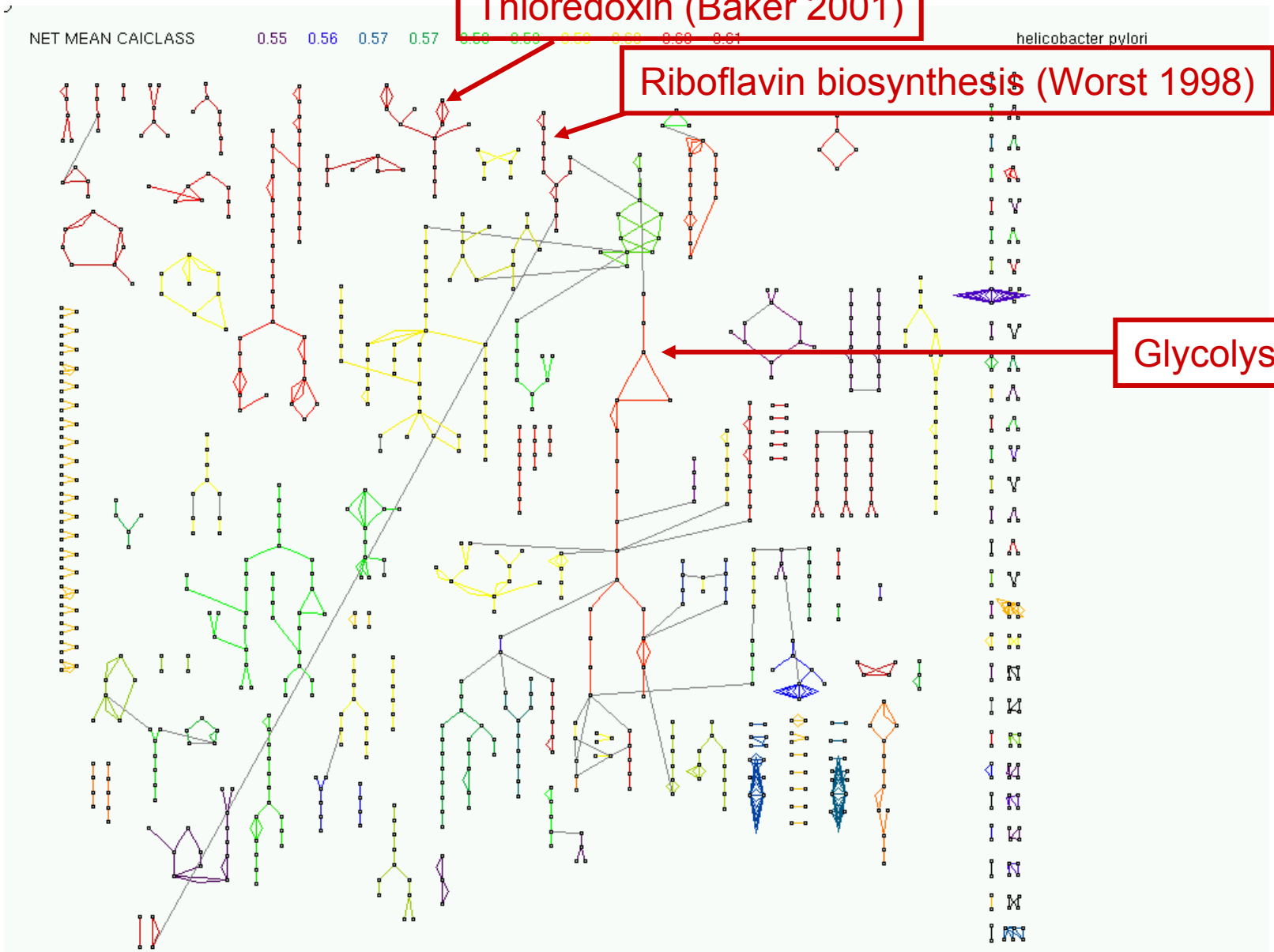
...and also :

L-serine degradation
(Pizer&Potochny 1964)

Ammonia assimilation Pathway
(Reitzer 1986, Helling 1994)

Helicobacter Pylori

BioCyc network, P.Karp *et al.*



Even genomes that do not grow rapidly might have signals of translational bias

Metabolic pathways essential to *Mycobacterium tuberculosis*

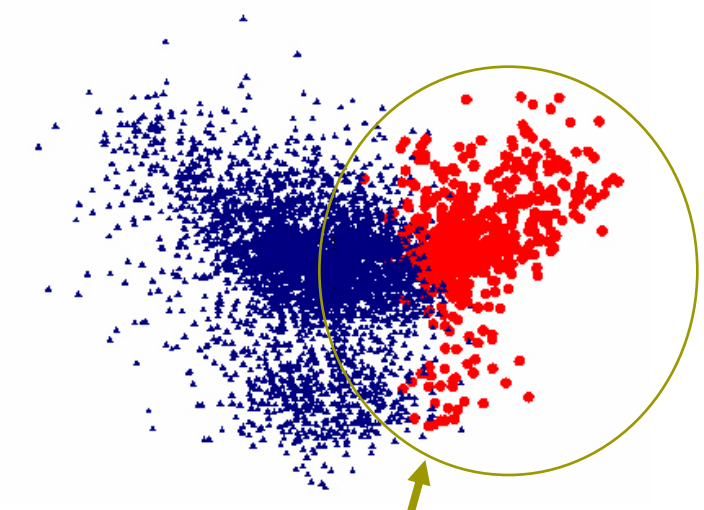
Essential to *M.tuberculosis* **but not** to other bacteria

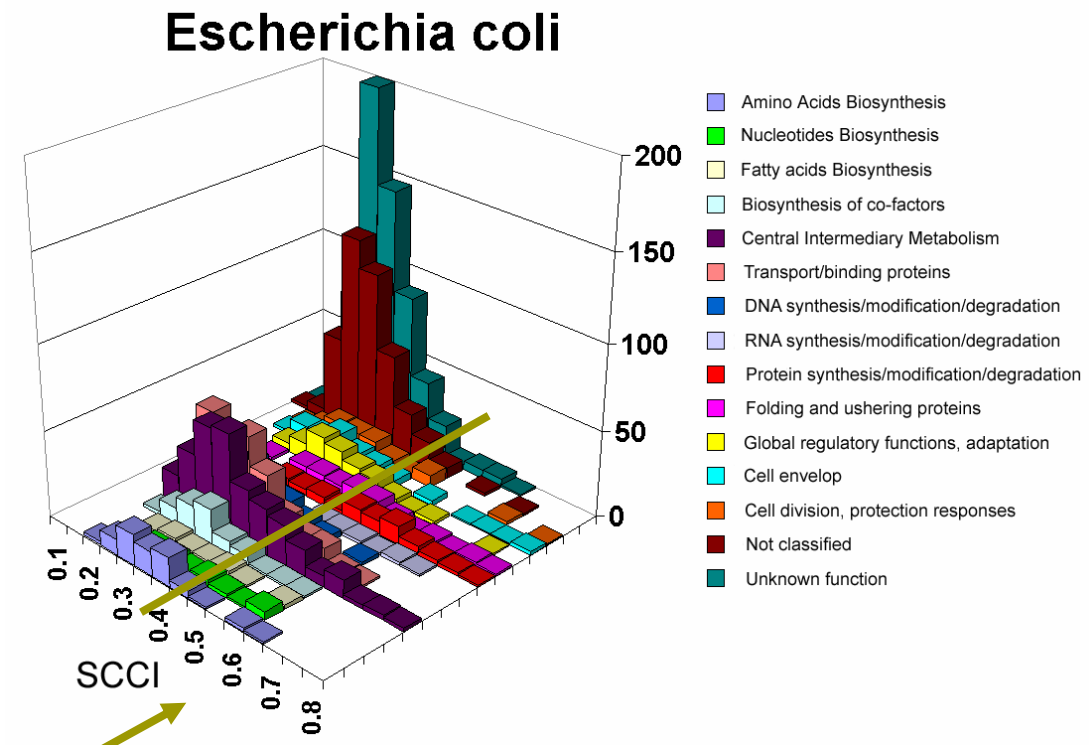
Biotin synthesis	(Norman et al. 1994)
Chorismate biosynthesis	(Parish and Stoker 2002)
Aspargine degradation	(Sassetti et al. 2003)
Pyridoxal 5'phosphate biosynthesis	(Sassetti et al. 2003)
Valine degradation	(Sassetti et al. 2003)
Leucine biosynthesis	(Sassetti et al. 2003)
ppGpp	(Primm et al. 2000)

Genes in 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





Core genes

We look at the tail

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

Map of core genes of 27 organisms (based on 200 most biased genes)

	Aci	Bha	Bsu	Bth	Bba	Cdi	Efa	Eca	Eco	Hin	Lpl	Lla	Mac	Pmu	Plu	Pab	Sty	Sat	Son	Sfl	Sag	Smu	Spn	Spy	Syn	Vch	Ype
INFORMATION STORAGE AND PROCESSING																											
J Translation and associated functions																											
ribosomal proteins (including subunits)	49	65	48	34	11	49	49	41	45	46	50	53	39	51	47	49	47	46	48	44	52	46	51	52	22	53	51
elongation factors	5	4	4	4	1	4	4	5	5	5	4	4	2	5	5	3	5	5	6	5	5	4	6	5	3	7	3
initiation factors	2	2	1	1	3	1		1	1	1	1	1	2	2	2	3	1	1	1	1	2	2		2		2	2
aminoacyl-transfer-RNA-synthetases	1		2	13	5	5	5	7	9	6	6	8	6	6	9	5	7	7	7	11	6	7	11	9		10	2
polyribonucleotide nucleotidyltransferase	1	1			1		1	1	1	1		1	2	1	1		1	1	1	1			1			1	1
ribosome recycling/releasing/binding factors	1		1	1			1		2	1	1			1	2		1	1	1	1	2	1	1	1		1	1
K Transcription																											
cold shock proteins	2	1	3	5		1	1	2	3		3	2		2	3		3	3	2	3	1			1		3	7
RNA polymerase	3	1	4	1	1	3	3	4	5	3	5	5	2	3	4	4	3	3	3	5	6	4	4	5	5	4	4
transcription antiterminator					1	1	1		1	1	1	1		1	1		1	1		1		1					
trascription terminator					1		1					1		1			2	2	1	1			1	1		2	1
L DNA replication, recombination and repair																											
Bacterial nucleoid DNA-binding protein	1	1	1					1	2	1		1		1	1		1	1		2			1	1		2	2
RNA helicase					1	1	1	1	1		1	1	1		1		1	1		1	1			1	1	1	1
single-strand binding protein					1				1	1					1		1	1		1		1		1			1
Recombination protein					1				1			1			1		1			1	1			1			1
CELLULAR PROCESSES AND SIGNALING																											
D Cell division and chromosome partitioning																											
cell division proteins							1	1				1	1		1	1	1	1			2	2	1				1
O Posttranslational modification, protein turnover, chaperons																											
chaperone proteins	3	3	3	2	3	4	3	3	2	3	3	3	2	3	5		3	3	3	3	1	2	3	2	4	5	3
peptidyl-prolyl cis-trans isomerase	2	1	1	1	2	1		3	3	3		1	2	2	3		3	3	3	3				1	1	2	3
thioredoxin	1	3	1	1			1			1	2	2		1	1				1		1	1	2	1	2	1	
alkyl hydroperoxide reductase protein			1	2			1	1	1						1		1	2	1	1	1	1		1		1	1
trigger factor	1	1	1	1			1	1	1	1	1	1		1	1		1	1	1	1	1	1	1	1		1	1
Clp protease			1	1		2	1	1	1		1	1										2	1		1		
ribose-phosphate pyrophosphokinase		1	1	1		1			1					1			1	1	1							1	
cell division							1				1	1			1		1		1			1		1	2		1
M Cell envelop biogenesis, outer membrane																											
channel forming, conductance	1			1		1			1					1			1	1		1	1	1		1			
lipoproteins					1			2	1	1				1	2		3	2	2	2						2	2
outer membrane proteins								4	5	4				3	3		8	7	9	5						6	7
N Cell mobility and secretion																											
secretory proteins	1				2	1	1	2	3	2				3	3		3	4	3	4	2				1		2
flagellin proteins		1	2	1	1			1							1	3			1							1	3
membrane GTP-binding proteins	1			1	1	1		1			1	1			1		1	1		1							
P Inorganic ion transport and metabolism																											
superoxide dismutase	1	1	1	1		1	1	1	2	1		1		1	1		1	1	1	2	1	1	1	1	1	1	1
phosphate binding proteins		2			1			3	2		1	1					3	3		2		1		1	2	1	1
metal-ion binding proteins		4		2						2			1	4	1						2		2	1			

	Aci	Bha	Bsu	Bth	Bba	Cdi	Efa	Eca	Eco	Hin	Lpl	Lla	Mac	Pmu	Plu	Pab	Sty	Sat	Son	Sfl	Sag	Smu	Spn	Spy	Syn	Vch	Ype
METABOLISM																											
C Energy production and conversion																											
hydratase	2	1	1			1		1	1	1				1	1		1	1	1	1							
dehydrogenases	7	8	5	10	7	7	6	5	6	5	5	6	14	6	5	6	6	4	11	5	2	3	5	5	3	6	6
membrane-bound ATP-synthase	4	3	2	3	2	5	2	4	3	4	3	3	3	5	4	3	3	3	4	3	3	2	2	3	3	4	3
succinyl-CoA synthetase	2	2		1				2	3						2	1	2	2	4	3						2	2
ferredoxin	2	2	1	1									5		1	1						1			2		1
other reductases								2	3	2			2	1				1	1	3	1					2	2
cytochrome oxidases	2		1			1									1		1	2	2	1						1	
cytochrome		1	1			2			1	3				3		1		1	6	1					5		
inorganic pyrophosphatase	1	1			1	1	1	1	1	1				1	1	1	1	1	1	1	1	1	1	1		1	1
phosphate acetyltransferase			1	1			1	1	1	1			1	1	1		1	1		1			1	1		1	1
formate acetyltransferase								1	1	1		1		1	1		1	1	1	1	1	1	1	1		1	1
acetate kinase				1		1	1	1	1	1			1	1	1		1	1		1	1			1		1	1
flavodoxin									1	1		1		1					1		1	1	1	1		1	1
G Carbohydrate transport and metabolism																											
enolase	1	1	1	1	1	1	1	1	1	1	1	1		1	1		1	1	1	1	1	1	1	1	1	1	1
1,6-bisphosphate aldolase	1	1	1	1		1	1	1	1	1	1	1		1	2		1	1	1	1	1	1	2	2	2	2	2
glyceraldehyde-3-phosphate dehydrogenase	1	1	1	2	1	1	1	1	1	1	1	2		1	1	1	1	1	1	1	2	2	1	2	1	1	1
inositol-1-monophosphatase	1								1	1				1	1		1	1	1	1							
transketolase	1		1	1				2	1	1				2	1		1	1	1	1	1	1	1	1	1	2	1
transporter (ABC and others)		3						1		3	2	4	1	3						2	1		7	3		3	2
PTS system		2	2	2			2	2	5	1	8	6					5	6		4	9	13	12	9		3	2
triosephosphate isomerase		1	1	1			1	1	1	1	1	1		1	1		1	1		1	1		1	1		1	1
phosphoglycerate kinase/mutase		1	2	1	1	1	2	2	2	2	2	2		2	1		3	3		3	3	3	2	3	1	2	3
pyruvate kinase			1	1	1		1	1	1	1	1	1	1	1	1	1	1	1		1	1		1	1		1	1
6-phosphofructokinase				1			1		1	1	1	1					1	1		1	1		1	1		1	
glucose-6-phosphate isomerase				1			1		1		2	1					1	1		1	1	1	1	1			1
transaldolase				1					1	1				1	1		1	1		1					1	1	1
6-phosphogluconate dehydrogenase (gnd)							1	1		1	1	1		1	1								1			1	1
isomerases									1		1	1				1			1	1	1	1	2	2	1		
E Amino acids transport and metabolism																											
transporters	1	2			3			1	1	2	1	2	2	1		2	1	2		1	5	4	2			3	1
glutamine synthetase	1	1	1	1	1	1		1	1		1	1	2		1		1	1	1	1	1	1	1		1		1
serine hydroxymethyltransferase	1		1	1		1		1	1	1				1	1	1	1	1	1	1		1					1
aminotransferases	2	1		2	1		1	1		2	1	2	2	1	1	4	1	2		1	1		1	1		1	2
ketol-acid reductoisomerase	1	1	1					1	1	1		1	1		1		1	1				1					1
dehydrogenases				3		2		1		2			3		2	3			1			3	2				
synthases	1	1						2				1	4		1	1						2	1		2	1	1
lyases								1	1	2				1	1		1	1		1			1			1	
F Nucleotide transport and metabolism																											
nucleoside diphosphate kinase	1	1				1		1	1				1		1				1	1					1		1
other kinases						1		1	1	1	1	1	1	1	1		2	2	2	1	1	1	1	1	1	1	1
nucleoside di(tri)phosphate reductase							1	2		1	1		1	2	2						2	1	1			1	
GMP reductase			1	1		1	1	1	1	1		1					1	1	1	1	1	1	1	1		1	1
GMP synthase			1	1		1	1	1	1	1		1				2	1	1		1	1		1	1			
CTP synthetase	1		1	1			1					1	1						1		1	1				1	
adenylosuccinate synthetase	1			1		1	1	1	1	1				1	1		1	1		1	1						1
transferases			1	2				1	1		1	2		1		1				1	2		3	1			
purine-nucleoside phosphorylase				2		1		1	1	1		1		1	1		1	1	1	1	1			2		1	1
deoxyribose-phosphate aldolase				1		1			1			1	1		1		1	1		1	1	1		1			
H Coenzyme metabolism																											
S-adenosylmethionine synthetase	1		1	1	1			1	1	1	1				1		1	1	1	1					1	1	1
I Lipid metabolism																											
acyl carrier protein	1	1	1				1	2	1	1		1		1	1		1	1	2	1	1		1	1	1	1	1
acetylCoA carboxylase	3	2	1			1		2	1	1				2	2		2	2		2						1	2
beta-ketoacyl-acyl carrier protein synthase	1								1	1		1		1			1	1		1		1					1

Genes with **specific metabolic functions** are in the tail

Photosynthesis metabolism : *Synechocystis*

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

Methan metabolism : *Methanosarcina acetivorans*

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

Ferrodoxin 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

Stress response genes are in the tail

Comparison with data from **comparative genomics**

Most represented functional classes of genes issued by comparing *M.genitalium* and *H.influenzae* (Mushegian and Koonin, 1996) **correspond to** most represented functional classes in functional genomic cores

Core genes expected to be essential but **missed** in (Mushegian&Koonin) :

Transcription : Sigma factors (rpo), termination factors (rho), chaperons (hsp90)
Energy metabolism : PTS proteins
Translation : no tRNA nucleotidyltransferase is found (consistently with comparative genomics)

Comparison with **experimental data**

difficult to make since:

1. there are no a priori false positives nor false negatives

2. *E.coli*:

620 essential genes (Gerdes *et al*., 2003)
234 essential genes (Hashimoto *et al*., 2005)

E.coli (Gerdes, 2003) :

620 essential genes over 3746 analyzed ones
520 core genes: 62.5% are essential

Enolase (*eno*) is a core gene and it does not belong to the 620 genes claimed to be essential for *E.coli*

E.coli (Hashimoto et al., 2005) :

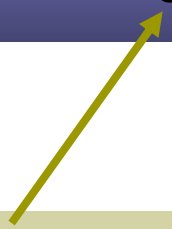
234 essential genes, 1890 non-essential, 900 unknown behavior
over 2994 analyzed ones (after genome minimization)
520 core genes : 129 essential, 278 non-essential,
53 unknown behavior,
63 deleted after minimization

Most are stress response genes

B.subtilis (Kobayashi et al., 2003) :

248 essential genes

519 core genes : 126 essential



Most genes involved in Embden-Meyerhof-Parnas pathway are core genes in agreement with their unexpected essentiality for (Kobayashi et al. 2003)

Synthetic biology

Bacteria and environment

Phages ⟷ Bacteria **are** the phage environment

A significant fraction of the procaryotic community is infected by phages

The total number of viruses, which is much larger than the total prokaryotic abundance, varies strongly in different environments and it is correlated with bacterial abundance and activity

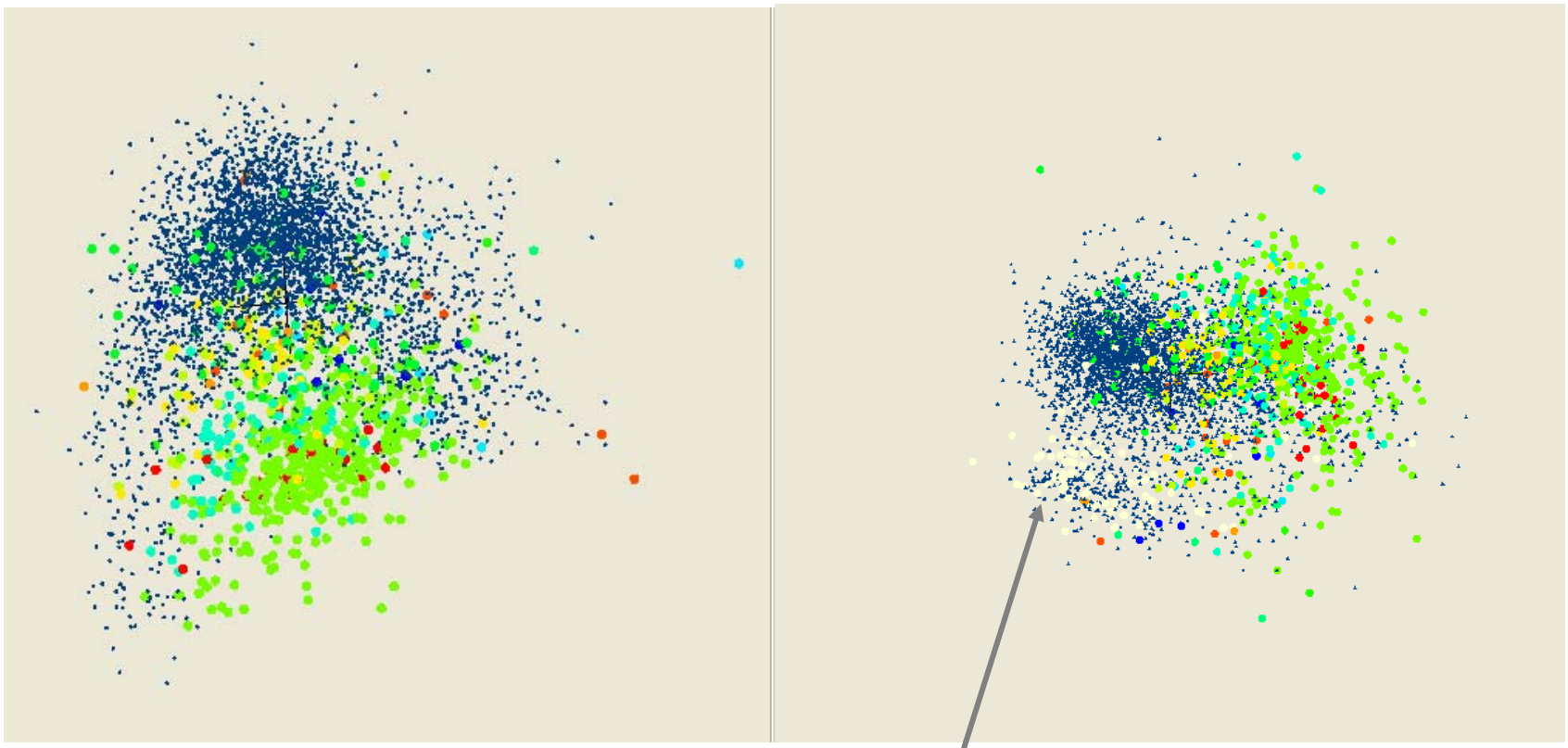
Codon bias is a major factor explaining phage evolution in transl biased hosts

116 phages of 11 translationally biased host bacteria :

actinobacteria	*Mycobacterium smegmatis*
	Mycobacterium tuberculosis
proteobacteria gamma	*Escherichia coli*
	Vibrio cholerae
	Salmonella thyphimurium
firmicutes bacillales	*Bacillus subtilis*
	Listeria monocytogenes
	Staphylococcus aureus
firmicutes lactobacillales	*Lactococcus lactis*
	Streptococcus pyogenes
chlamydiales	*Chlamydophila caviae*

102 of these phages display at least one gene with high bias

Vibrio cholerae + 10 phages



Membrane proteins and
permease proteins («hydrophobic» aa)

Functional classification of highly biased phage genes

Phage	cap	mor	tai	lys	hyp	oth	hig
Enterobacteria phage RB43	5		4		35	7	51
Enterobacteria phage RB49	3	2	2		26	6	39
Enterobacteria phage RB69	4	2	1	1	8	4	20
Enterobacteria phage T4	2	2			3	3	10
Enterobacteria phage T7	2	1			2	3	8
Enterobacteria phage K1E	1			1	3		5
Enterobacteria phage K1F	1	1			1	1	4
Enterobacteria phage Mu	1				2		3
Enterobacteria phage P1	1		1			1	3
Enterobacteria phage N15	2				1		3
Enterobacteria phage $\alpha 3$	1		1				2
Enterobacteria phage λ	1				1		2
Enterobacteria phage P2	1	1					2
Enterobacteria phage 186	1						1
Enterobacteria phage G4	1						1
Enterobacteria phage HK022	1						1
Enterobacteria phage I2-2	1						1
Enterobacteria phage Ike	1						1
Enterobacteria phage ϕK	1						1
Enterobacteria phage $\phi X174$						1	1
Enterobacteria phage 933W						1	1
Enterobacteria phage HK97	1						1
Enterobacteria phage $\epsilon 15$					6	1	7
Salmonella phage SETP3	1				2		3
Enterobacteria phage P22	1				1		2
Enterobacteria phage S13						1	1
Bacillus phage SPBc2					50	9	59
Bacillus phage GA-1 virus	1	1		2	8	3	15
Bacillus phage PZA	1	1		1	4		7
Bacillus phage B103	1		2	1	1	1	6
Mycobacterium phage D29	1		5	1	38	3	48
Mycobacterium phage Che12	2	1	4	3	33	9	52

Phage	cap	mor	tai	lys	hyp	oth	hig
Lactococcus phage P335	2			2	1		5
Lactococcus phage bIL170	2				3		5
Lactococcus phage r1t	1				4		5
Lactococcus phage BK5-T	2			1		1	4
Lactococcus phage jj50	2			1	2		4
Lactococcus phage $\phi LC3$	1	1			2		4
L.lactis phage ul36	1	1		1			3
Lactococcus phage P008	2				1		3
Lactococcus phage 712	2						2
Lactococcus phage c2	1				1		2
Lactococcus phage Tuc2009	1				1		2
Listeria phage A118	1			1	4		6
Listeria phage P100					5	1	6
Streptococcus phage 315-5				1	6		7
Streptococcus phage 315-2					3	1	4
Streptococcus phage 315-1	1		1		1		3
Staphylococcus phage K	1		1	2	18	3	25
S.aureus phage $\phi P68$	3				3	1	7
Staphylococcus phage 44AHJD	3				3	1	7
S.aureus phage $\phi NM3$ provirus					4	1	5
S.aureus phage PVL provirus					33		3
S.aureus phage phi11 provirus	2					1	3
Staphylococcus phage ϕETA	1			1	1		3
Staphylococcus phage ϕNM	1			1	1		3
Vibrio phage KVP40	2	1	4		178	57	242
Vibriophage VP4	4	1	2		4	15	25
Vibrio phage VP2					12	1	13
Vibrio phage K139	1	1	1				3
Vibrio phage $VGJ\phi$	1				1	1	3
Chlamydia phage chp1	4				3	2	9
Chlamydia phage chp2	1						1
Chlamydia phage $\phi CPG1$	1						1

Cap = capsid proteins
Mor = morphogenesis
tai = tail proteins
lys = lysis proteins
hyp = hypothetical proteins
oth = other functions

Functional classification of biased coliphage genes

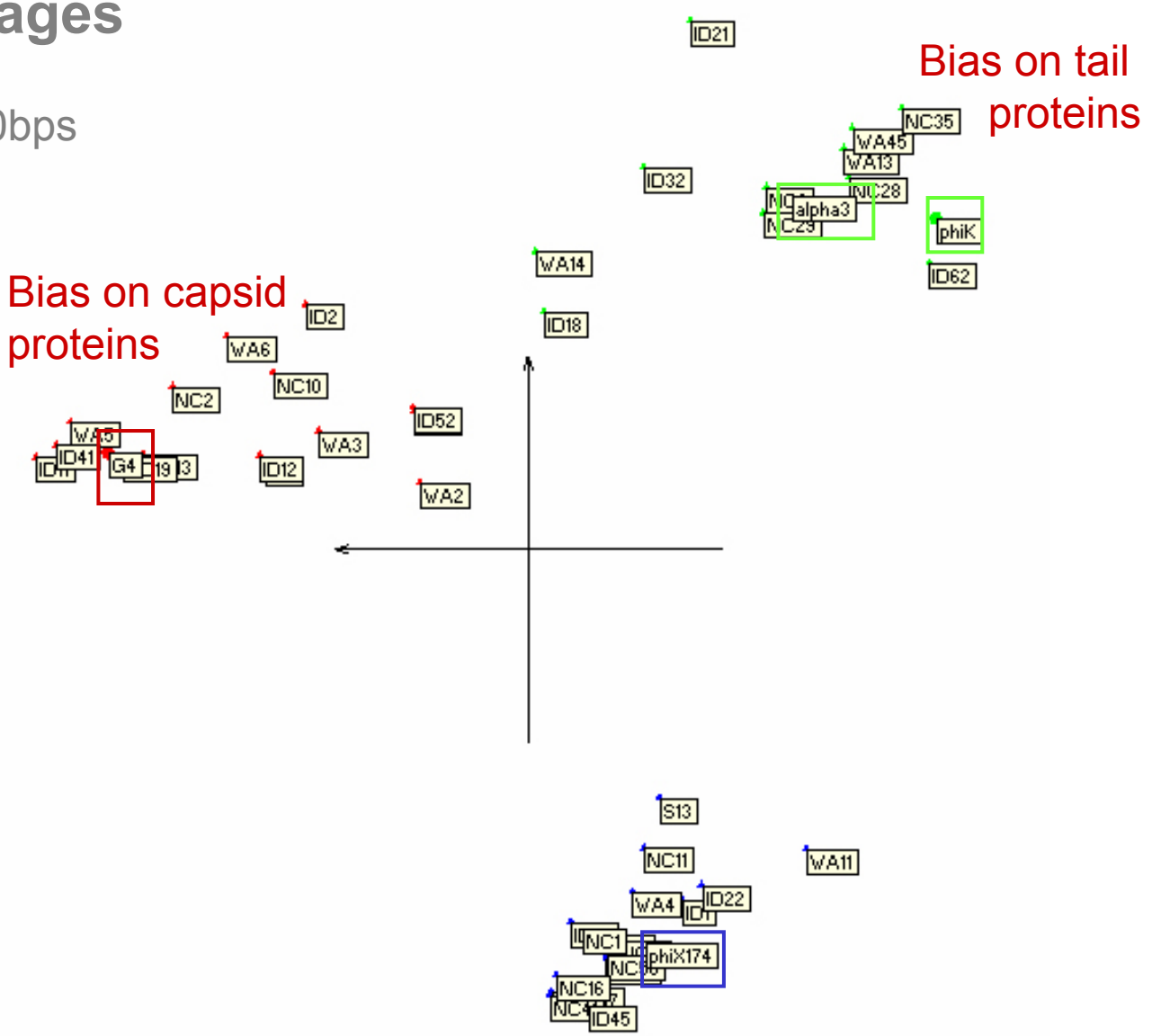
	RB43	T4	RB69	RB49	P1	VT2-sa	933W	T1	λ	K1E	HK97	T7	N15	HK620	HK022	Mu	K1F	P2	186	P4	ϕX174	G4	M13	α3	If1	Ike	ϕK	I2-2
number of genes	292	288	275	274	110	83	80	78	71	62	61	60	60	58	57	55	43	43	42	14	11	10	10	10	10	10	10	9
number of biased genes	168	57	87	113	16	11	11	22	20	19	10	20	19	10	5	17	18	11	16	2	4	4	4	4	1	3	4	4
percentage of biased genes	57	19	31	41	14	13	13	28	28	30	16	33	31	17	9	31	42	38	38	14	36	40	40	40	10	30	40	44
capsid/head/coat proteins	5	5	5	4	2			1	2	1	1	2	1	1	1	1	1	1	2		1	1	2	1	1	2	2	2
portal proteins	1			1	1											1		1	1									
terminase subunits	1																	2	2									
capsid protease	1															1												
head-tail joining proteins	1			1						1																		
tape measure proteins													1															
virion morphogenesis	3	3	2	5	4				1	1		1	4	2	1	1	3	2	2		2	1		1			1	1
lysis	1	4	4	1	1	2	2		1	1	2	1				1	4		2									
tails/tail fibers proteins	12	2	2	3	3			1	5		2		3		1	1	1	2	3			1		1			1	
Other functions																												
host interaction and inhibitors of host functions	4	3	2	2	2	1	1		1	1		2		1			1		2									
inhibitors of phage functions	1		2	1																			1					
transcription regulation	2	3	3	2	1				1					1		1		1										
translational regulator		1	2																									
DNA condensation and stabilization									2													1		1				
DNA destabilization (ssDNA binding)	1	1	1									1					1				1		1			1		1
dsDNA binding protein	1	1	1	1																								
exonuclease ∞	1					1	1	1	1			1					1											
endonuclease	1	1	1	1						1							1											
RNA replicase																												
dUTPase, dCTPase	1	1	1	1																								
RNA polymerase and RNA pol. binding proteins	1	1	1	2						1		1					1											
DNA polymerase accessory proteins	3	2	3	2								1																
reductase	3		1	4																								
phosphatase	1	1	1	1																								
recombination protein	1		2	2		1	1	1	1																			
DNA topoisomerase	1	2	1																									
DNA replication protein							1		1				1															
ligase (DNA or RNA)	2											1																
RNAse ribonuclease			1	1																								
dCMP protein		1		1																								
replicase																												
endoprotease																												
DNA primase																				1								
transposase																1												
synthetase	1																											
helicase (DNA or RNA)	3																											
thioredoxin				1																								
unknown function	115	25	51	76	2	6	5	18	4	12	4	9	8	5	2	9	4	2	2	1	0	0	0	0	0	0	0	0

If phage genomes were to contain a pool of essential genes, these functional classes could suggest appropriate candidate genes

Adaptation explains phage classification

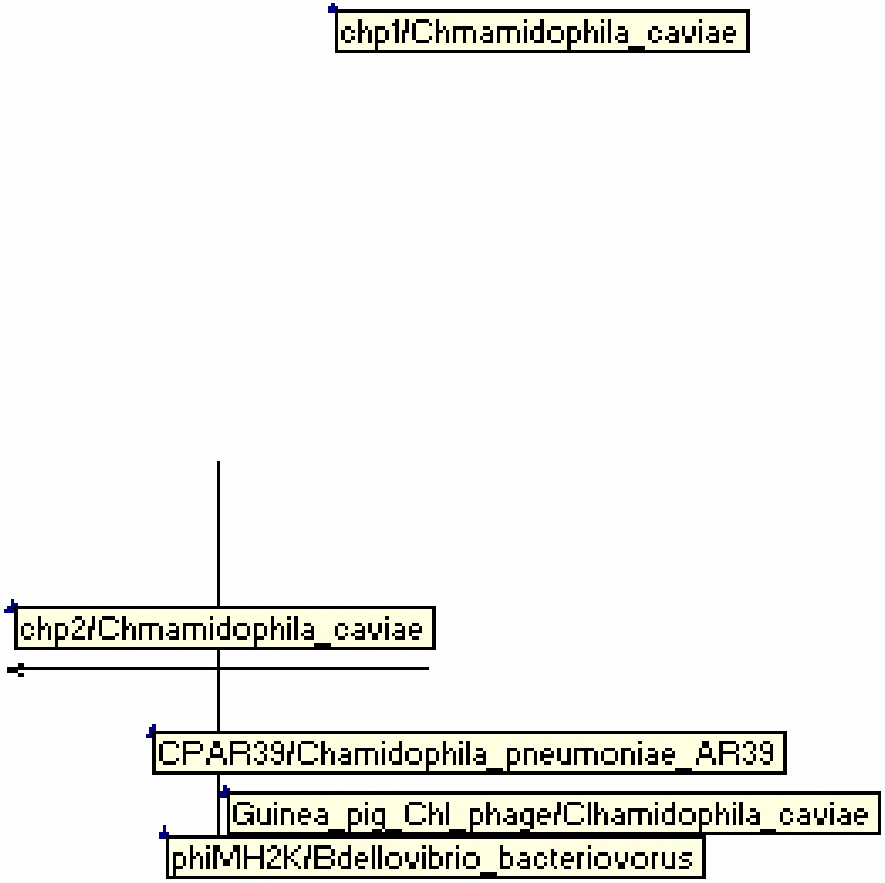
Microviridae phages

Host: *E.coli*
Genome length: ~6000bps
Genes: 10 in common



Coordinates
=
10-dim vector
of SCCI

Patterns of adaptation go beyond host species



Phage classification does not reflect host phylogeny

Collaborations and references

Algorithm and microbial SCCI 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 & IGI (USA)

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.

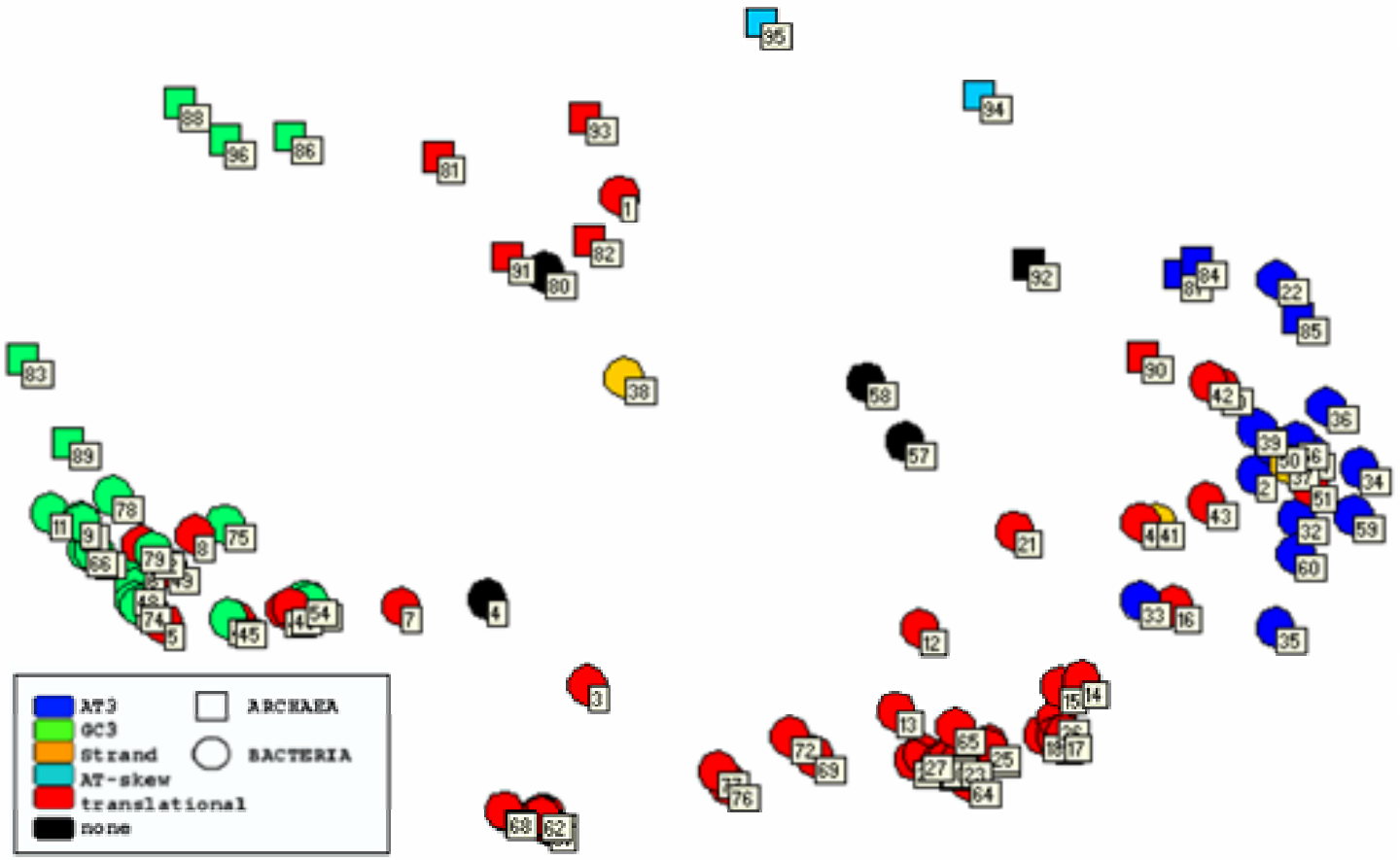
Minimal gene sets :

A.Carbone, Computational prediction of genomic functional cores specific to different microbes, *Journal of Molecular Evolution*, **63,** 733-746, 2006.

Host-phage co-evolution:

A.Carbone, Codon bias is a major factor explaining phage evolution in translationally biased hosts, *Journal of Molecular Evolution*, 2007. In press.

Bootstrapping information from translationally biased organisms



translationally biased organisms are everywhere

Small genomes : *M.genitalium* and *B.aphidicola*

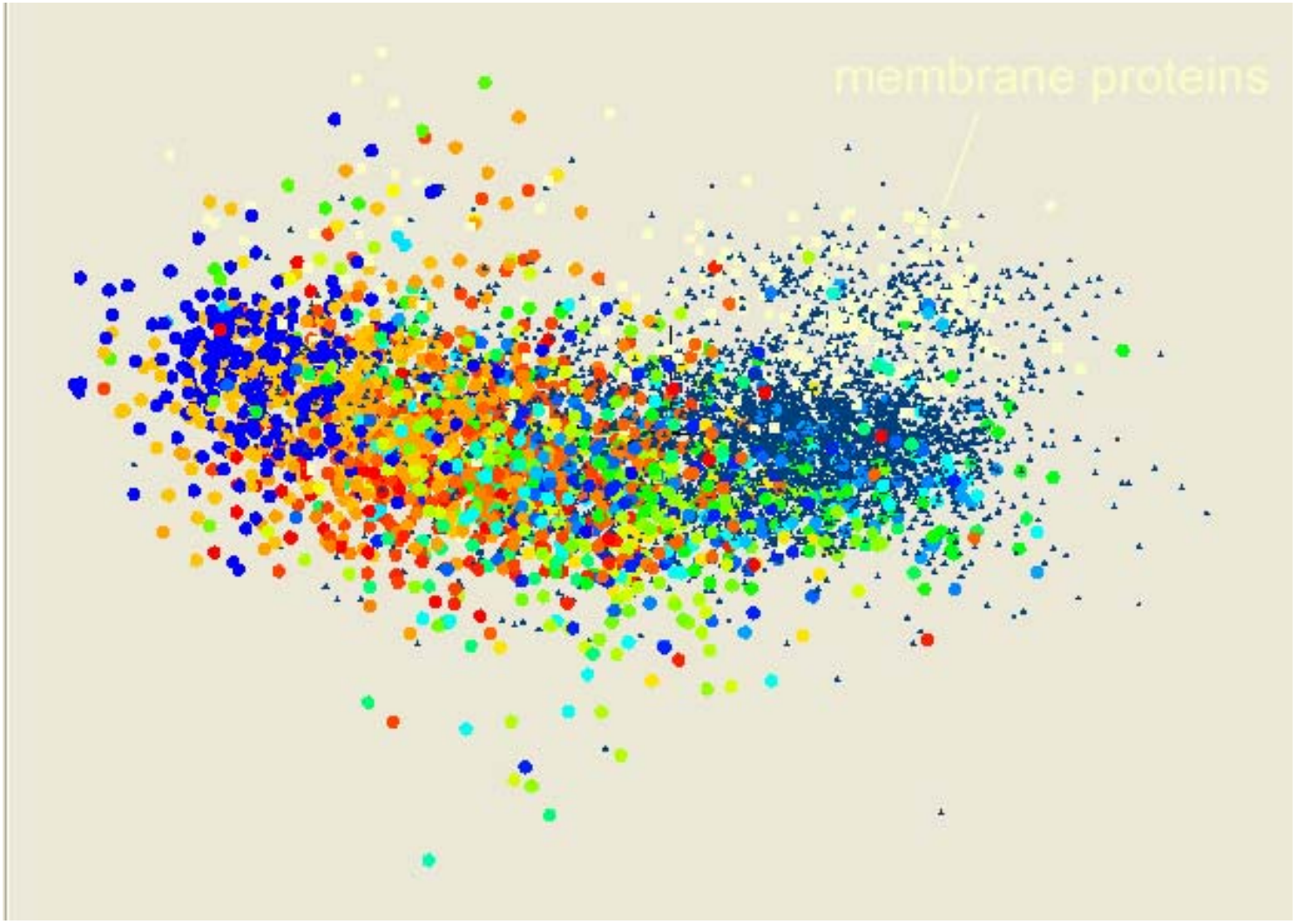
Buchnera aphidicola str Bp	504 coding genes 498 genes homologous to *E.coli* genes

Mycoplasma genitalium	484 coding genes 266 genes homologous to *E.coli* genes

189 genes are shared by *B.aphidicola* and *M.genitalium*

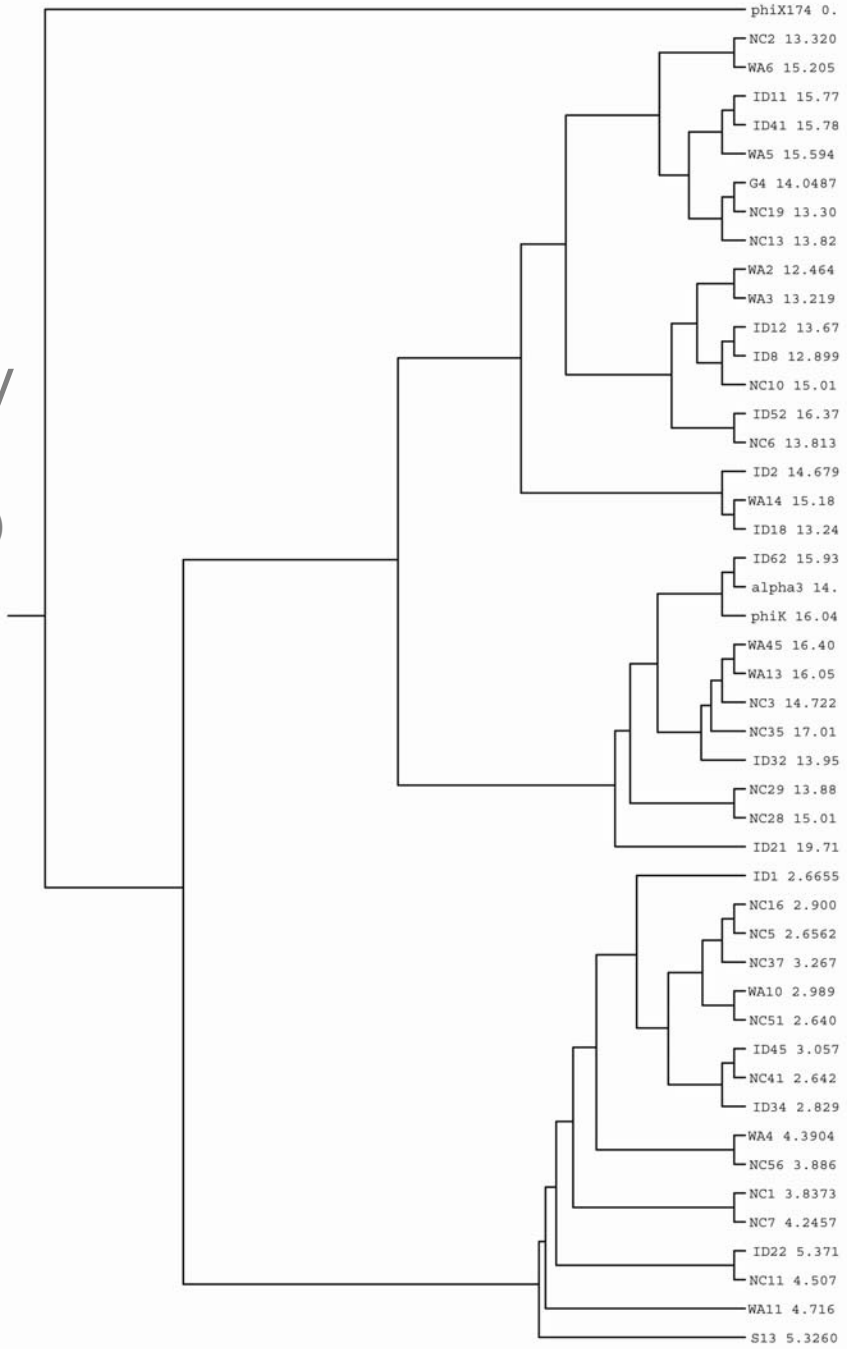
129 of these genes have high SCCI in *E.coli*

E. coli + 28 phages



Equivalent to the tree obtained from full genome comparison proposed as microviridae phylogeny

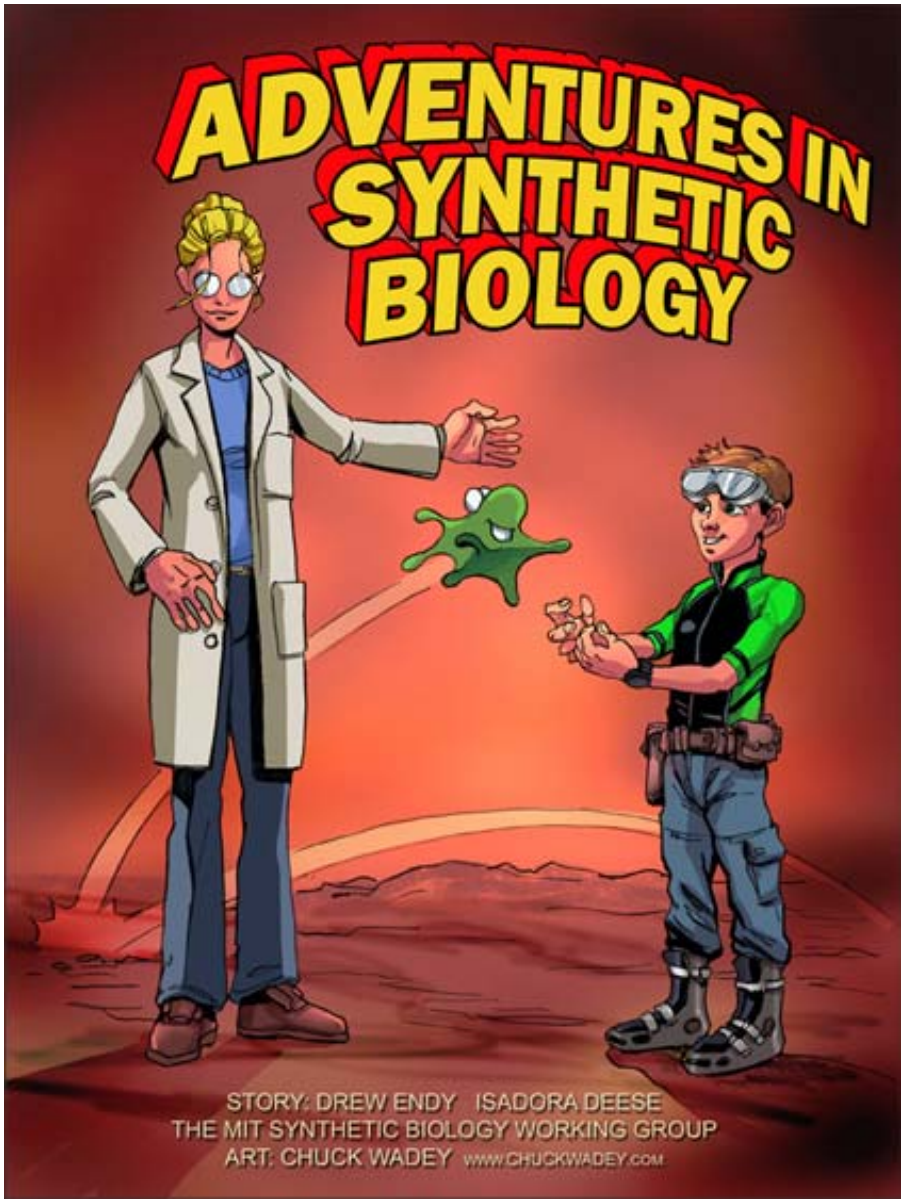
(Rokyta, Wichman et al. 2006)



Bias on capsid proteins

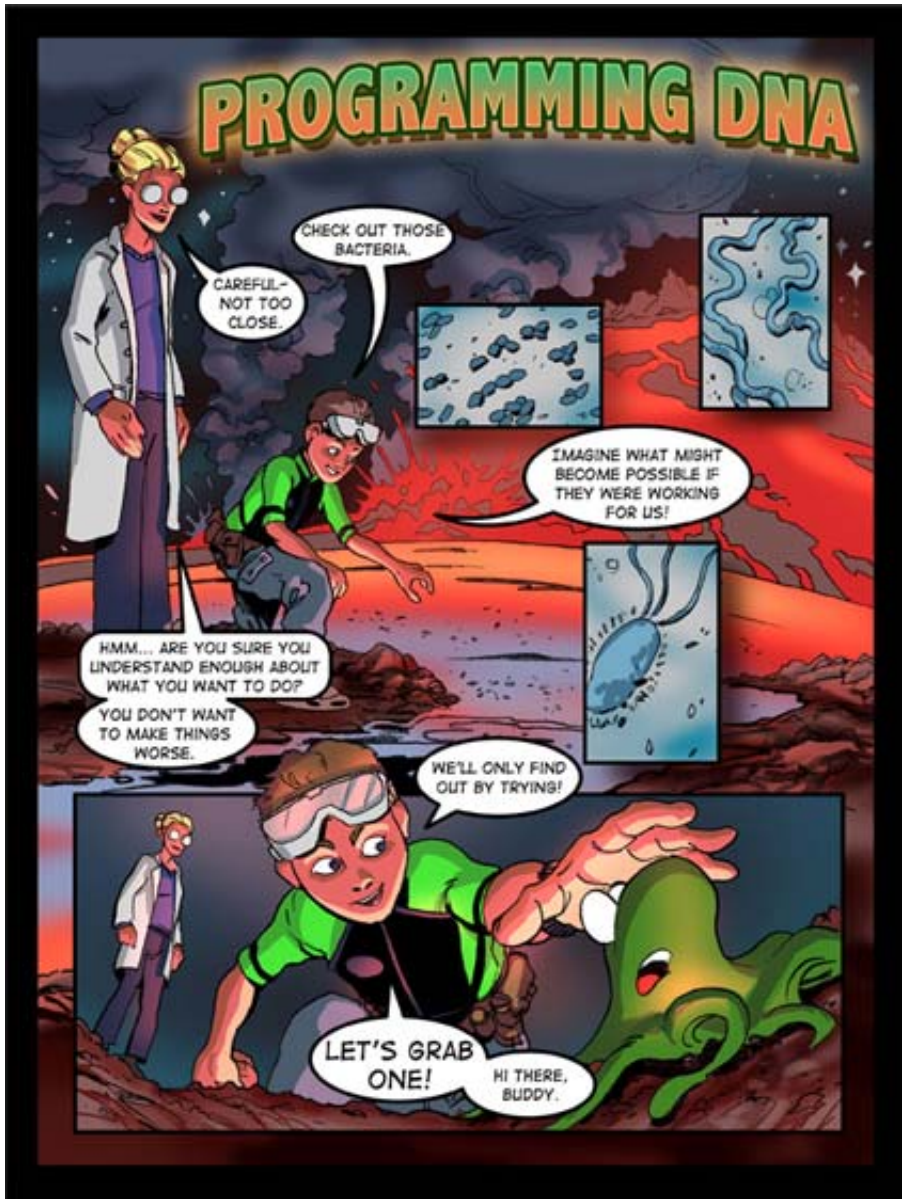
Bias on tail proteins

1st step



Genome synthesis

2nd step



Genome programming