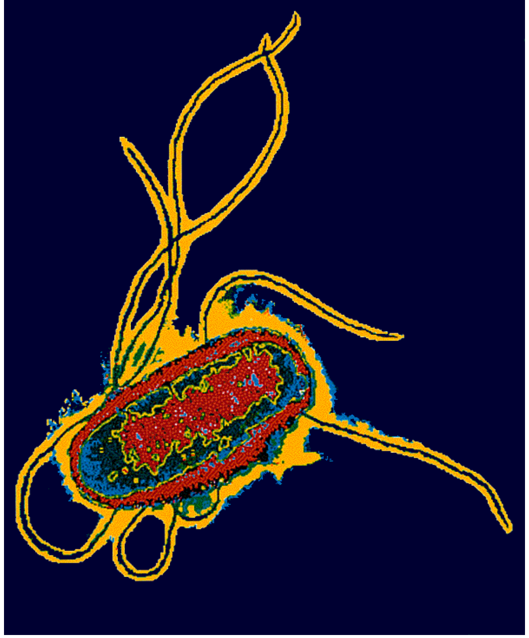
Codon bias and the space of micro-organisms

Alessandra Carbone
Université Pierre et Marie Curie, Paris



GGTACTTACCTTGGA

GAGATTCCATTACCG

CGCGTAGCGCTTAAT

TCCGCGAGATCGAT

CGATCGTGCATTCAA

TTCAGCGCATACGAT

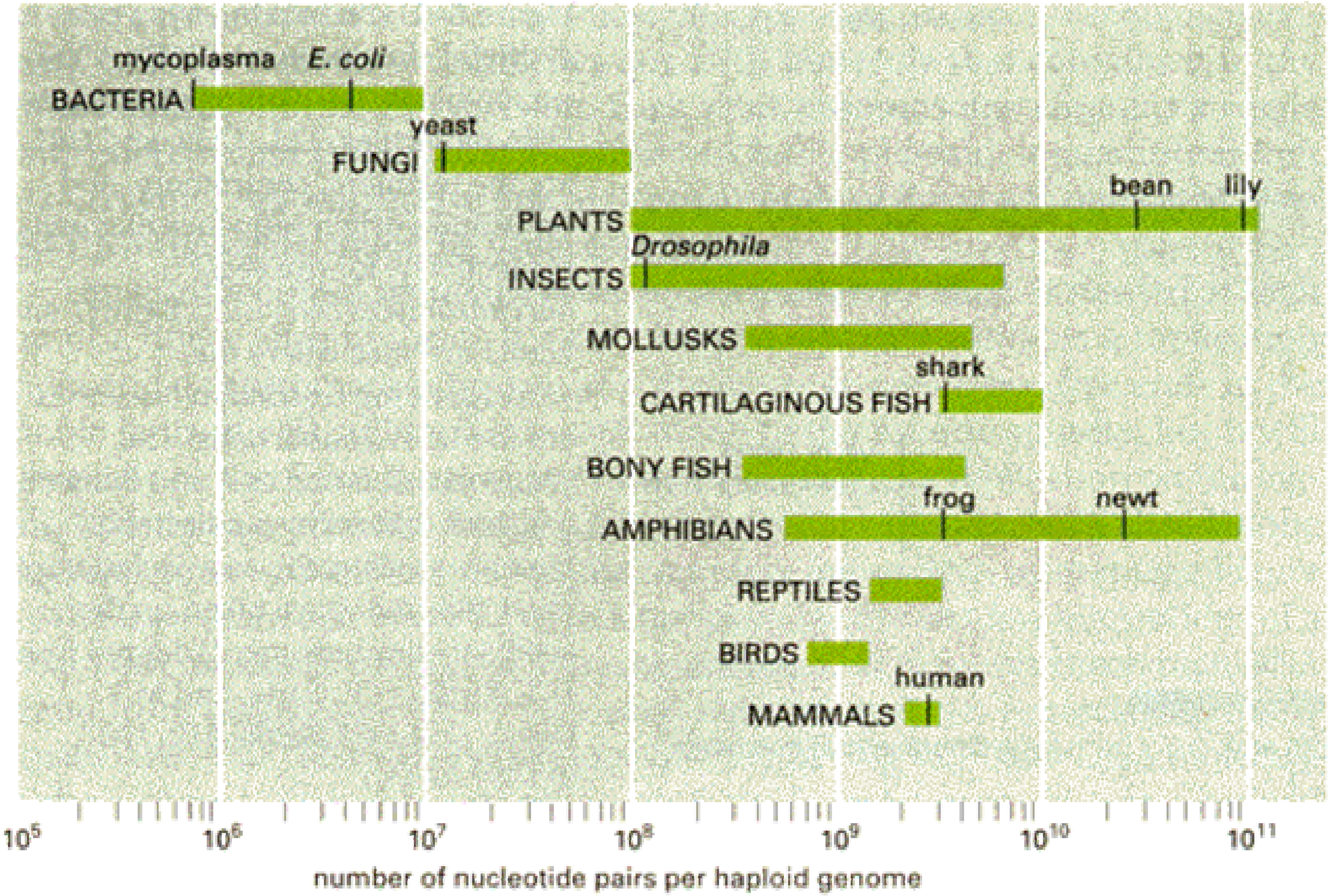
CGACTACTTCAGCG

What is coded in the genome?

Initial conditions are determined by the mother cell, but all the rest (architecture, consistency of initial conditions, and behavior) is coded in the genome.

Only a fraction of the genomic coding sequence is used at different moments along cell life. In particular, in multicellular organisms, differentiated cells use only a part of the genome.

A variety of genome sizes



…but strong genomic constraints

Symmetries in cellular organisation:

syntactical : repetitions, homologies, codons

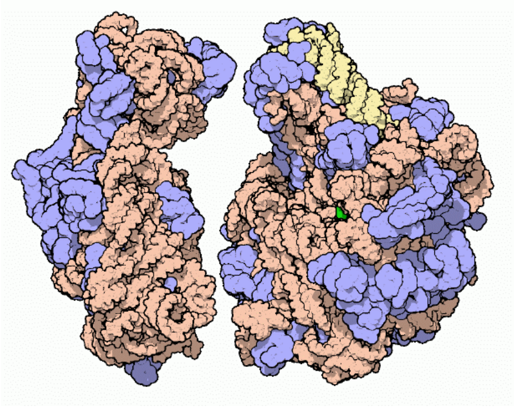
spatial : virus coats

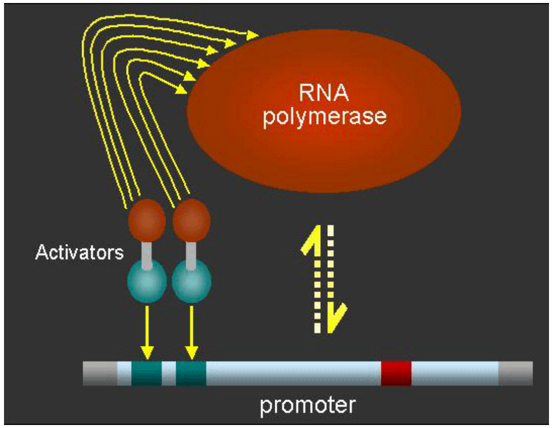
temporal : cyclic behavior of bio-chemical processes

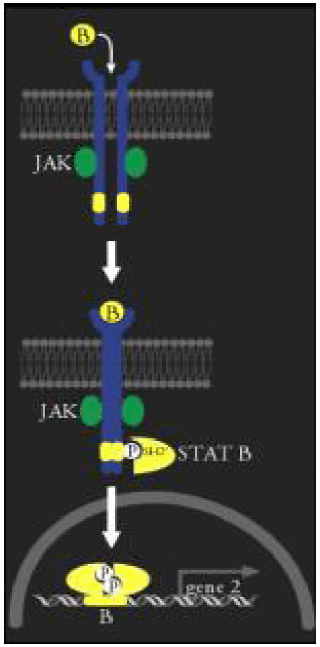
combinatorial : repetitions of sub-complexes

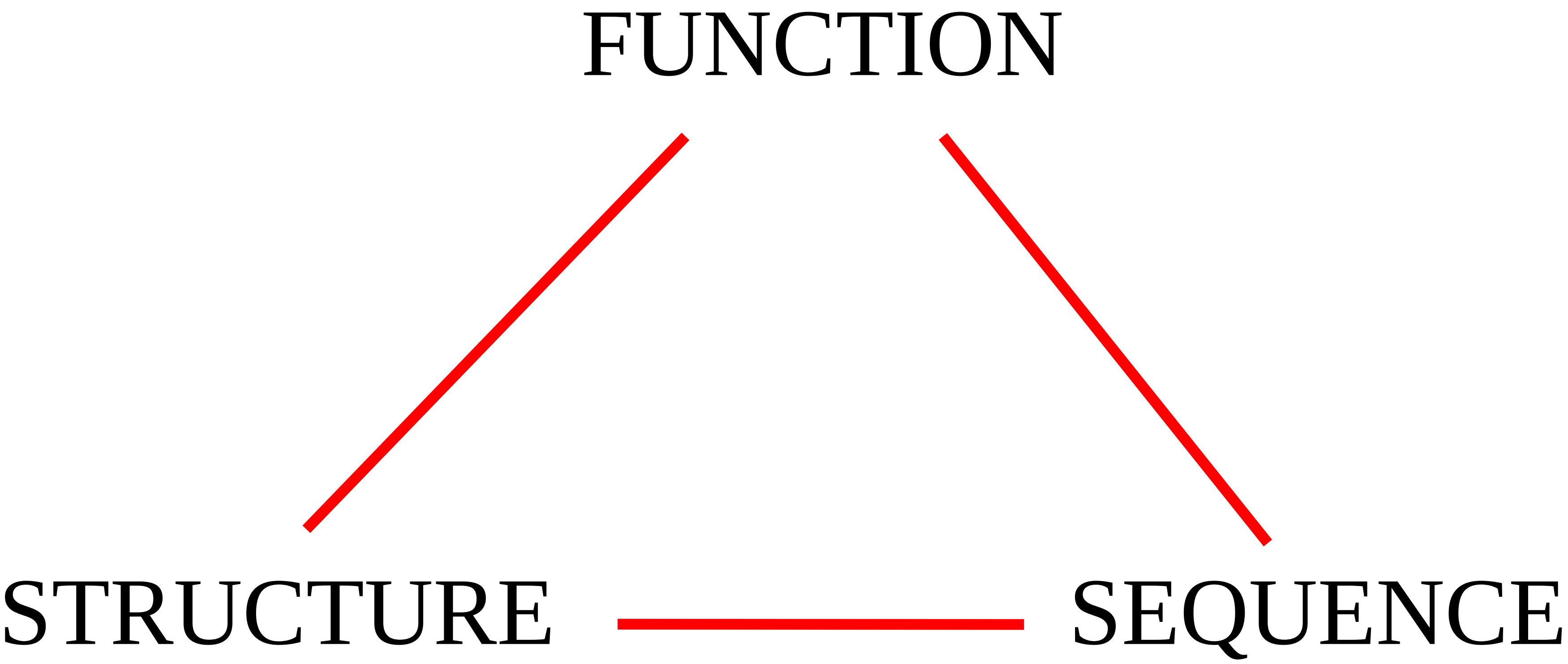
functional : function similarity of homologous complexes

“Generic” machines:









FUNCTION

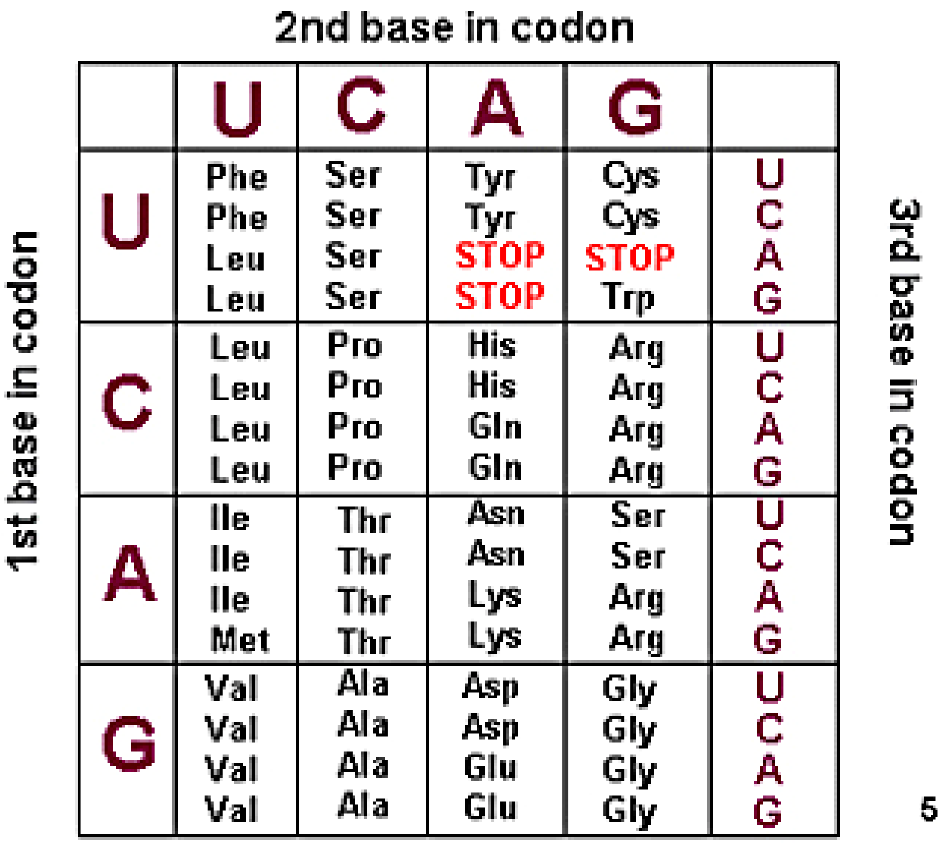
STRUCTURE

SEQUENCE

“Consistency principle”

Can **gene composition** tell us something about **gene expression** ?

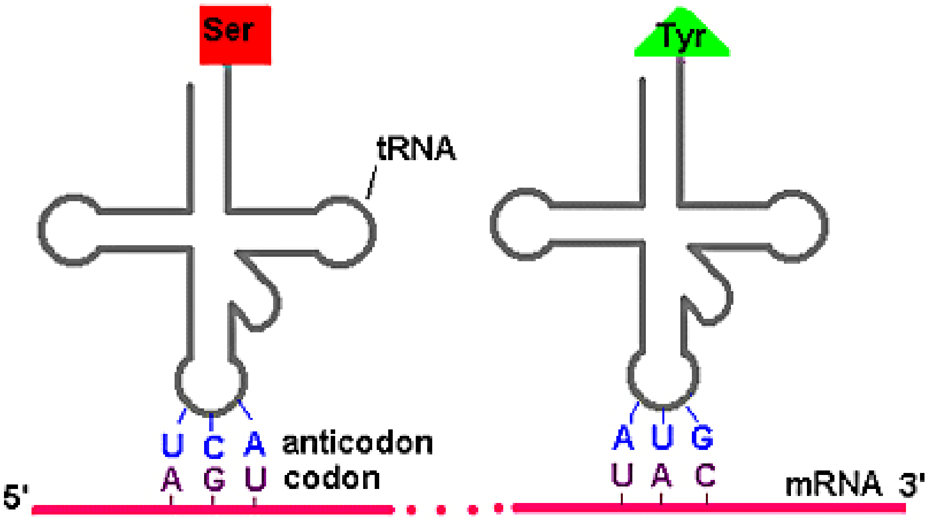
Redundancy of the genetic code



2nd base in codon

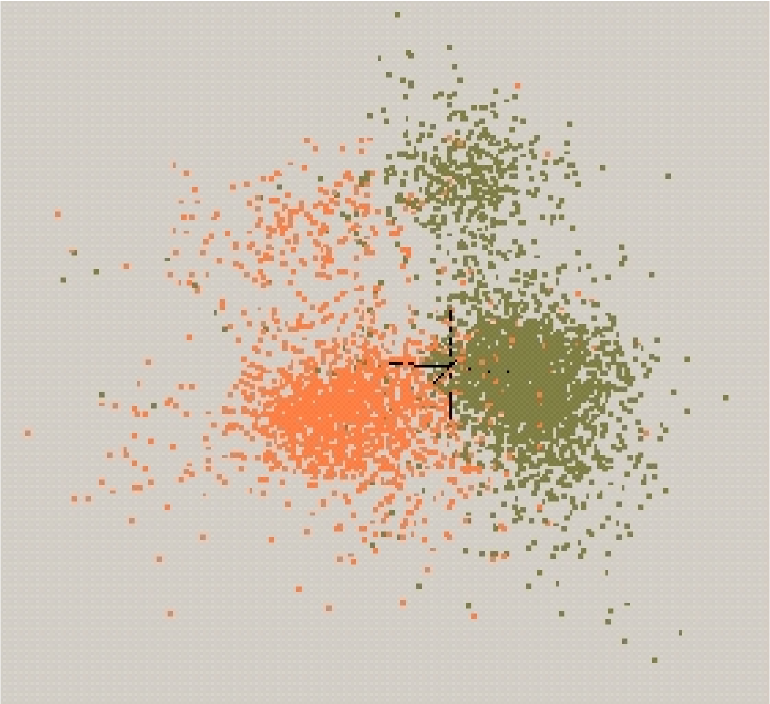
1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val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

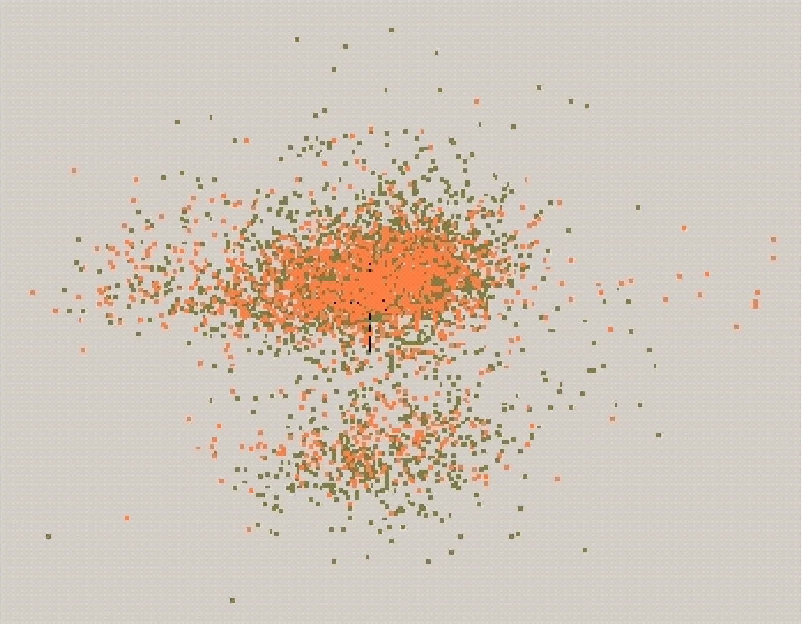
Transfer RNA

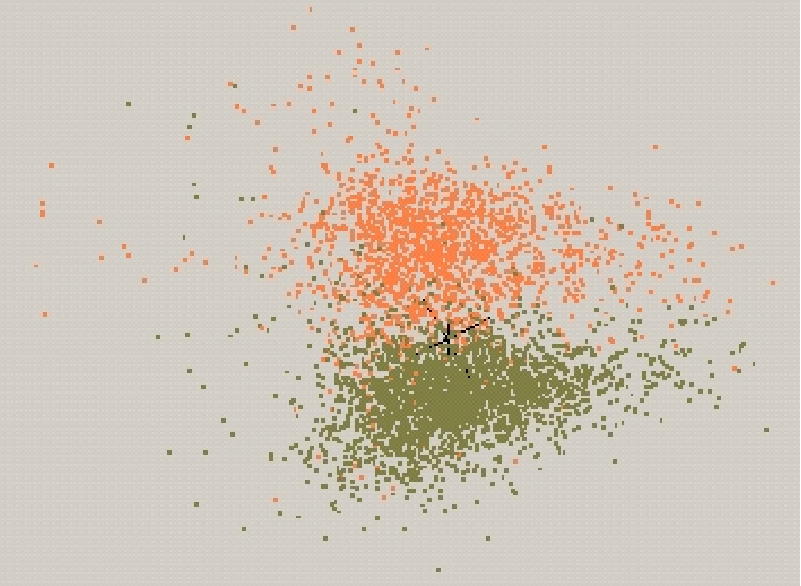


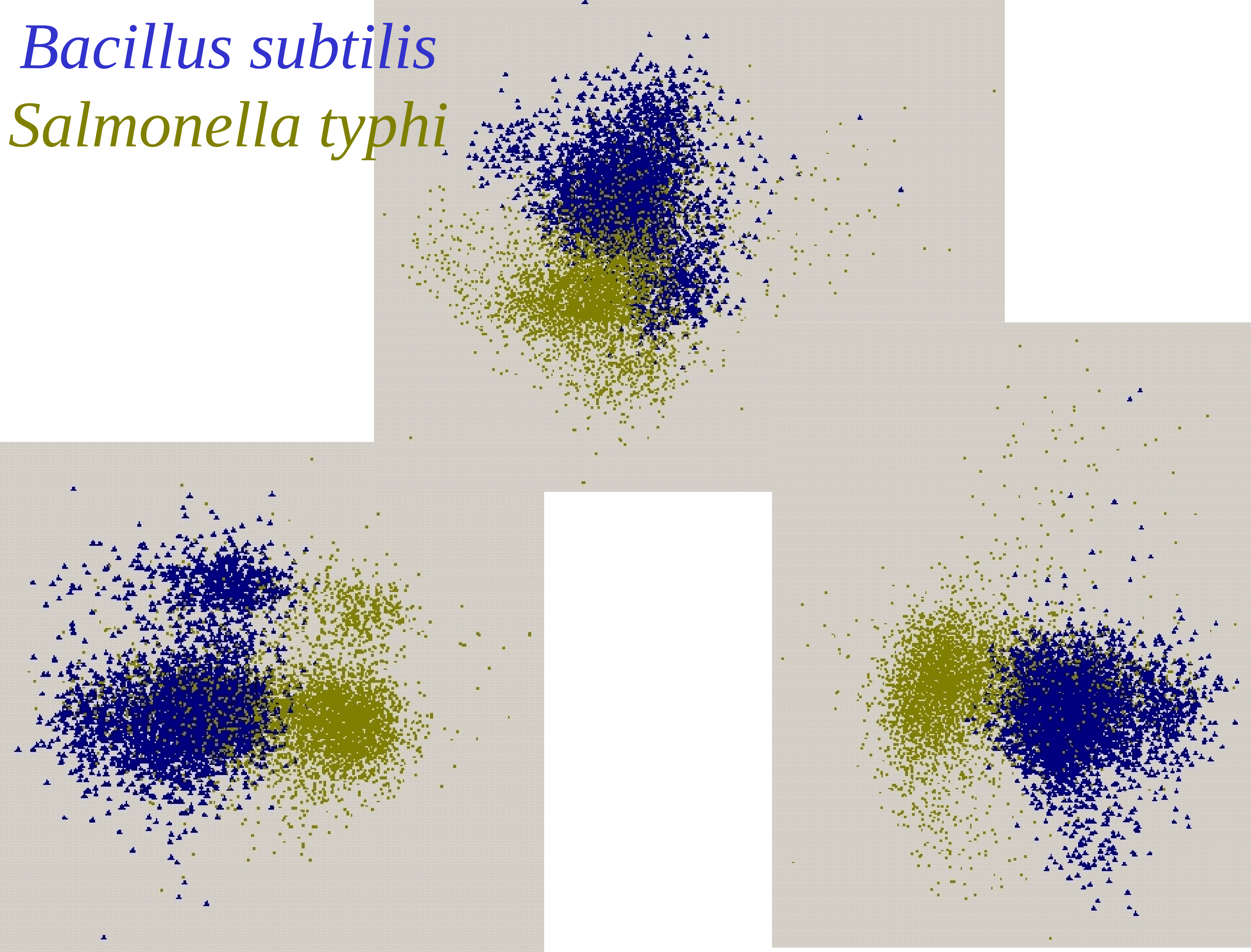
Haemophilus influenzae

Staphylococcus aureus



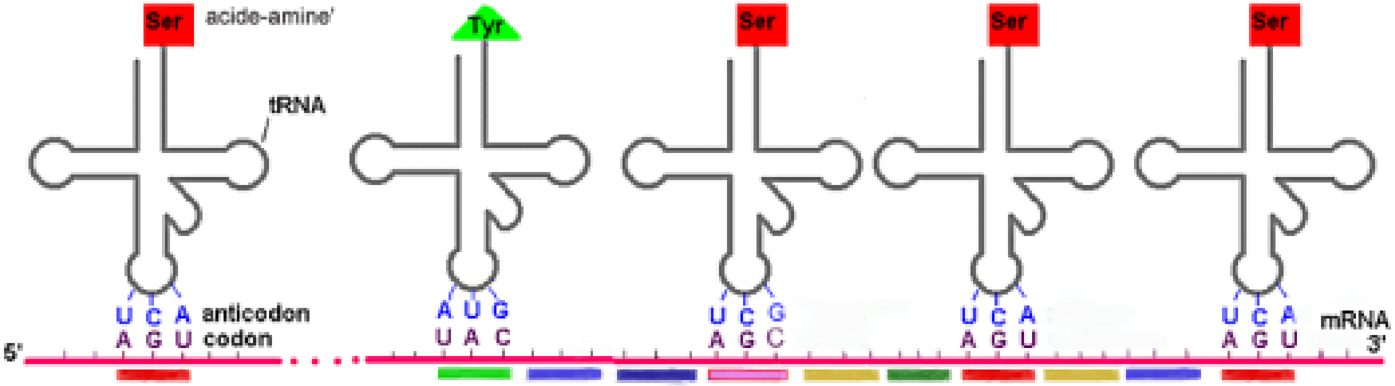





Bacillus subtilis
Salmonella typhi

Preferred codons

These are codons that appear with the higher frequency in most genes

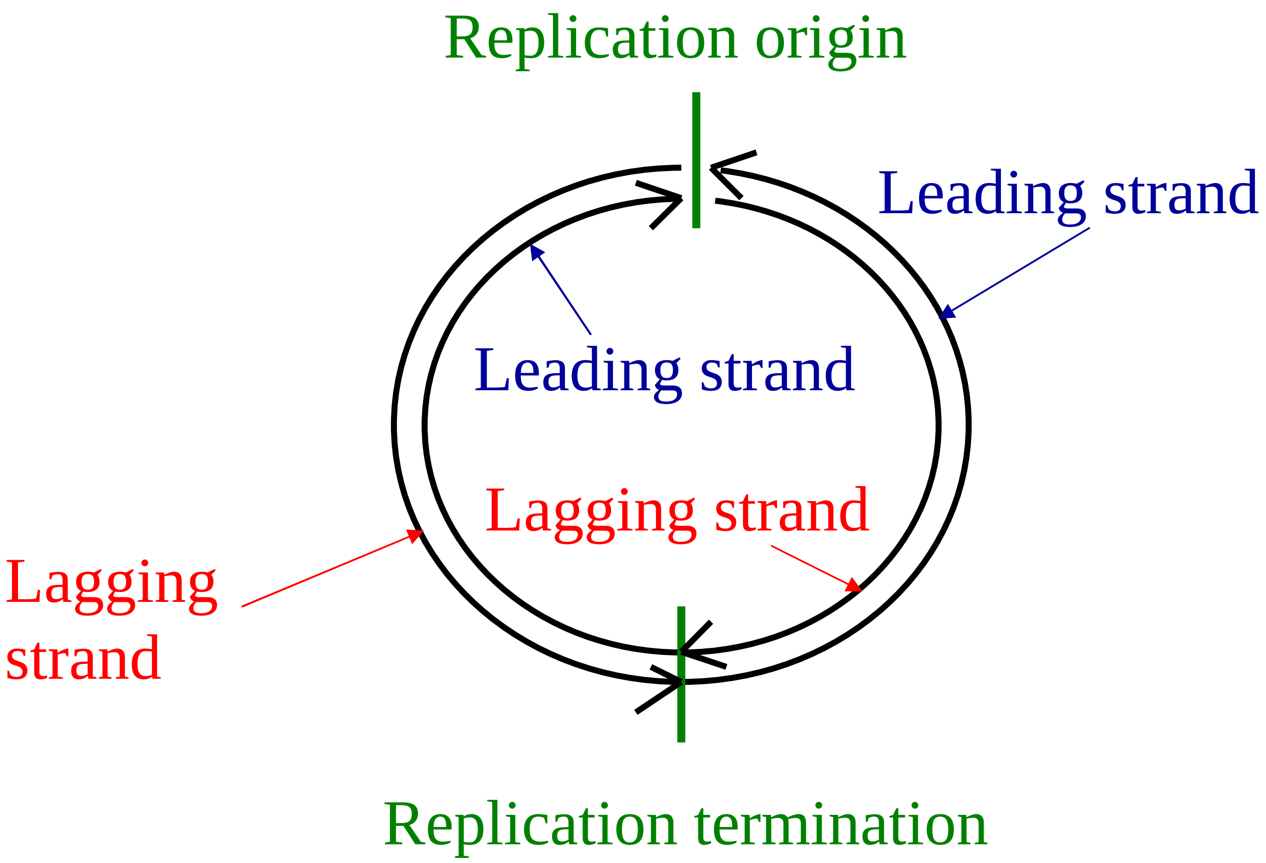


CODON BIAS

Codon bias in gene sequences

- G+C / A+T
- G+C in the third position, GC3 / AT3
- …
- Leading strand richer in G+T than lagging strand
- Horizontal gene transfert
- Translational bias (mRNA ⟹ protein)

All together they make
“codon bias signatures”



Replication origin

Leading strand

Leading strand

Lagging strand

Lagging
strand

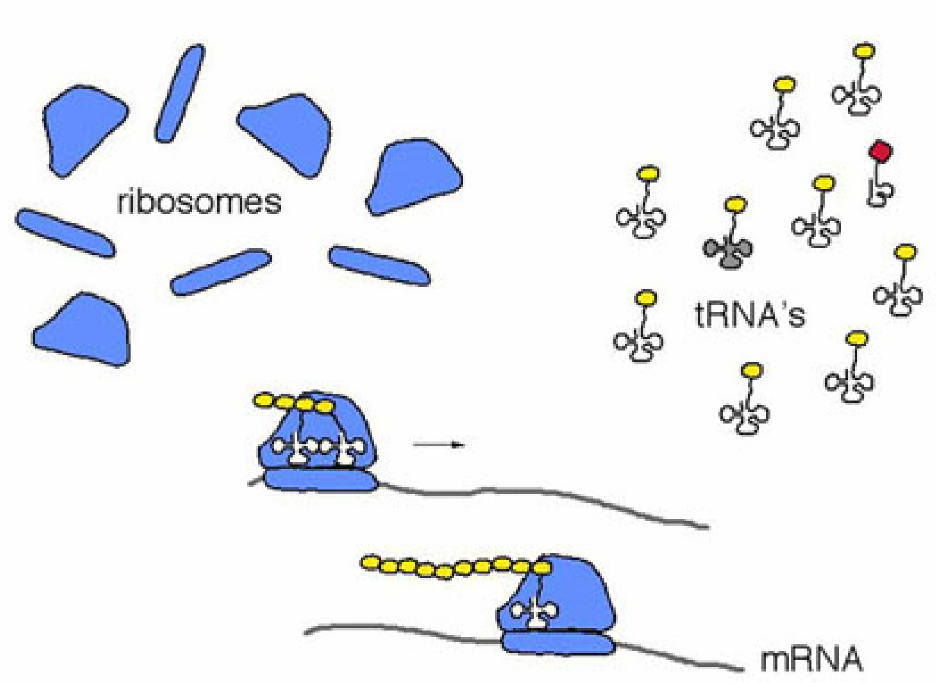
Replication termination

Translational bias: three facts

- **Highly expressed genes** use a limited number of codons (→ preferred codons).

- Preferred codons and associated transfer RNA in the cell exhibit a strong positive correlation

- Use of these codons may make translation faster or more efficient and may decrease misincorporation.



Fix a set S of **highly expressed** genes and, for all genes g in the genome, compute

$$\mathrm{CAI}(g) = \left(\prod_{k=1\ldots L} w_k\right)^{1/L}$$

(Codon Adaptation Index -

formula introduced by Sharp & Li, 1987)

L number of codons in g

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

Manual choice of S

- Ribosomal proteins
- Elongation factors
- Glycolitic proteins
- …

They need to be expressed fast and/or in large quantities

From S, biologists **ranked** all genes

We propose an algorithm to detect **dominant codon bias** in a genome.

The algorithm is based on a simple and precise mathematical formulation of the problem, that lead us to use the **Codon Adaptation Index** as a *universal* statistical measure of codon bias.

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

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

L number of codons in g

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

We look for S automatically in such a way that

1. S contains the 1% of genes of the genome

2. CAI values on S are **maximal**, i.e.

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

   where G is the set of all genes

3. S is **representative** of preferred codons, i.e.

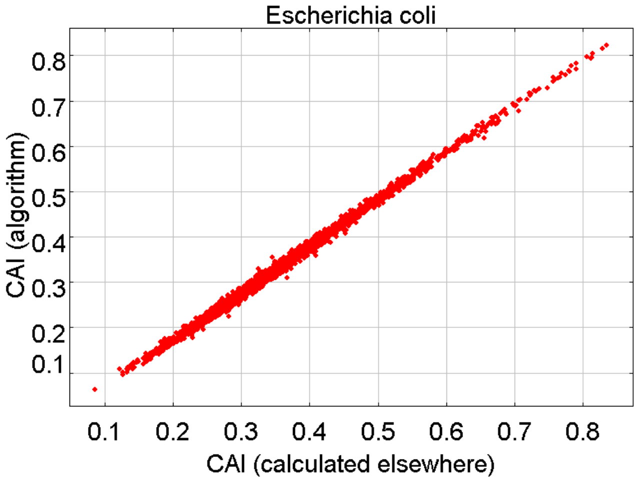
   $c_1,\ldots,c_{20}$ preferred codons for S
   $d_1,\ldots,d_{20}$ preferred codons for G
   we look for the set S for which

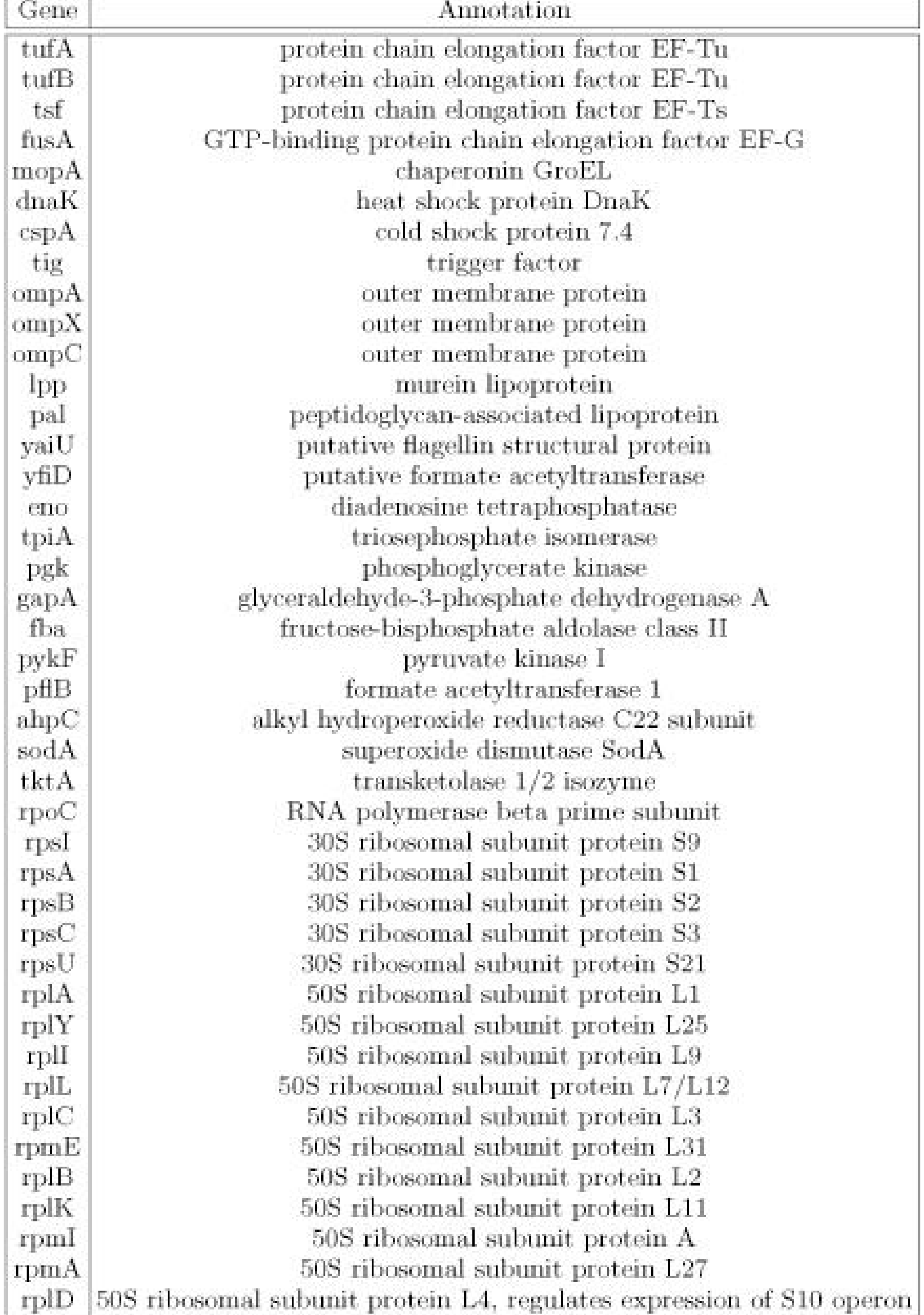
$$\Sigma_{i=1}^{20}\,\chi(c_i,d_i) \quad \text{is minimal}$$

- 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.

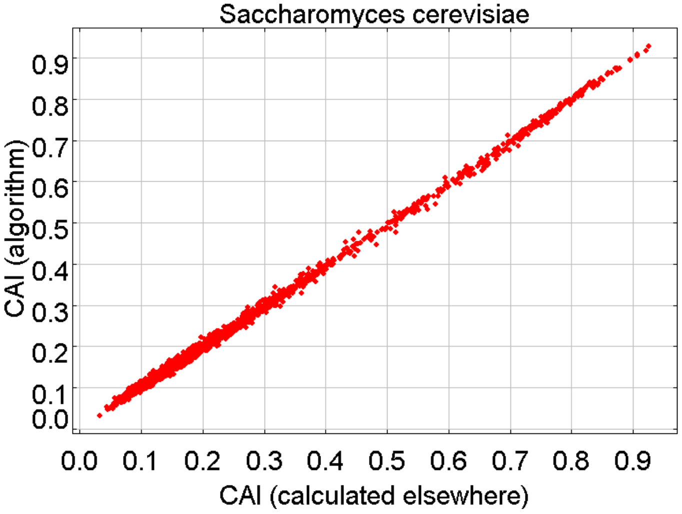
S chosen by the algorithm: *E.coli*

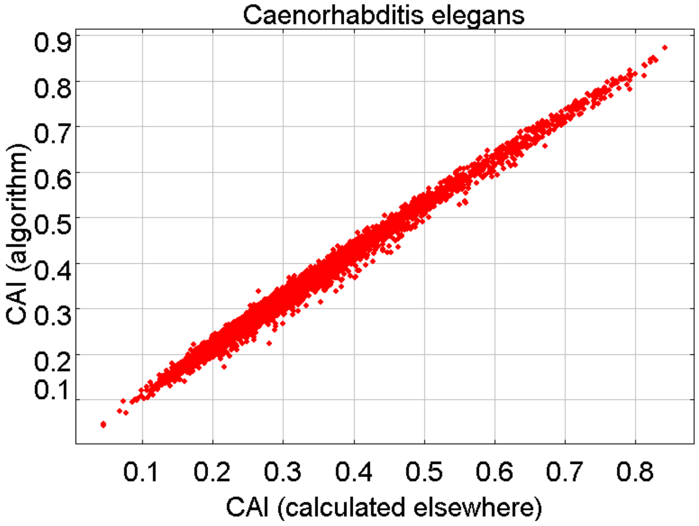


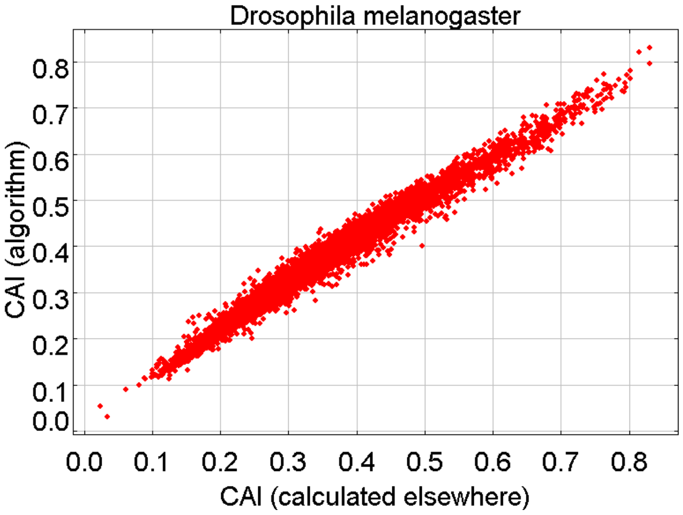


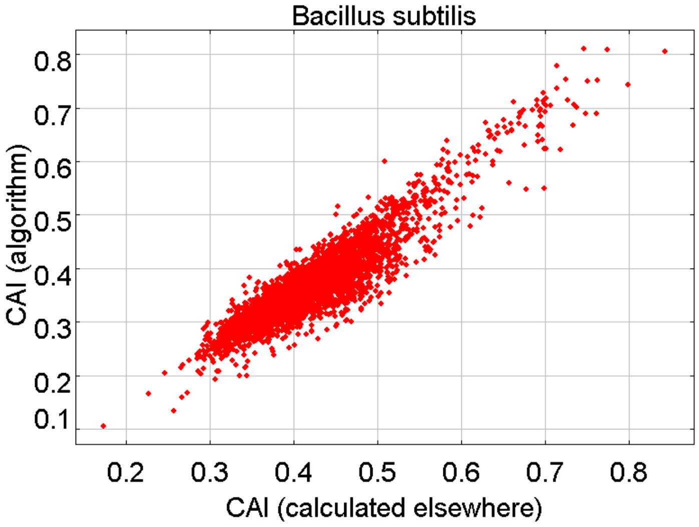
Gene	Annotation
tufA	protein chain elongation factor EF-Tu
tufB	protein chain elongation factor EF-Tu
tsf	protein chain elongation factor EF-Ts
fusA	GTP-binding protein chain elongation factor EF-G
mopA	chaperonin GroEL
dnaK	heat shock protein DnaK
cspA	cold shock protein 7.4
tig	trigger factor
ompA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

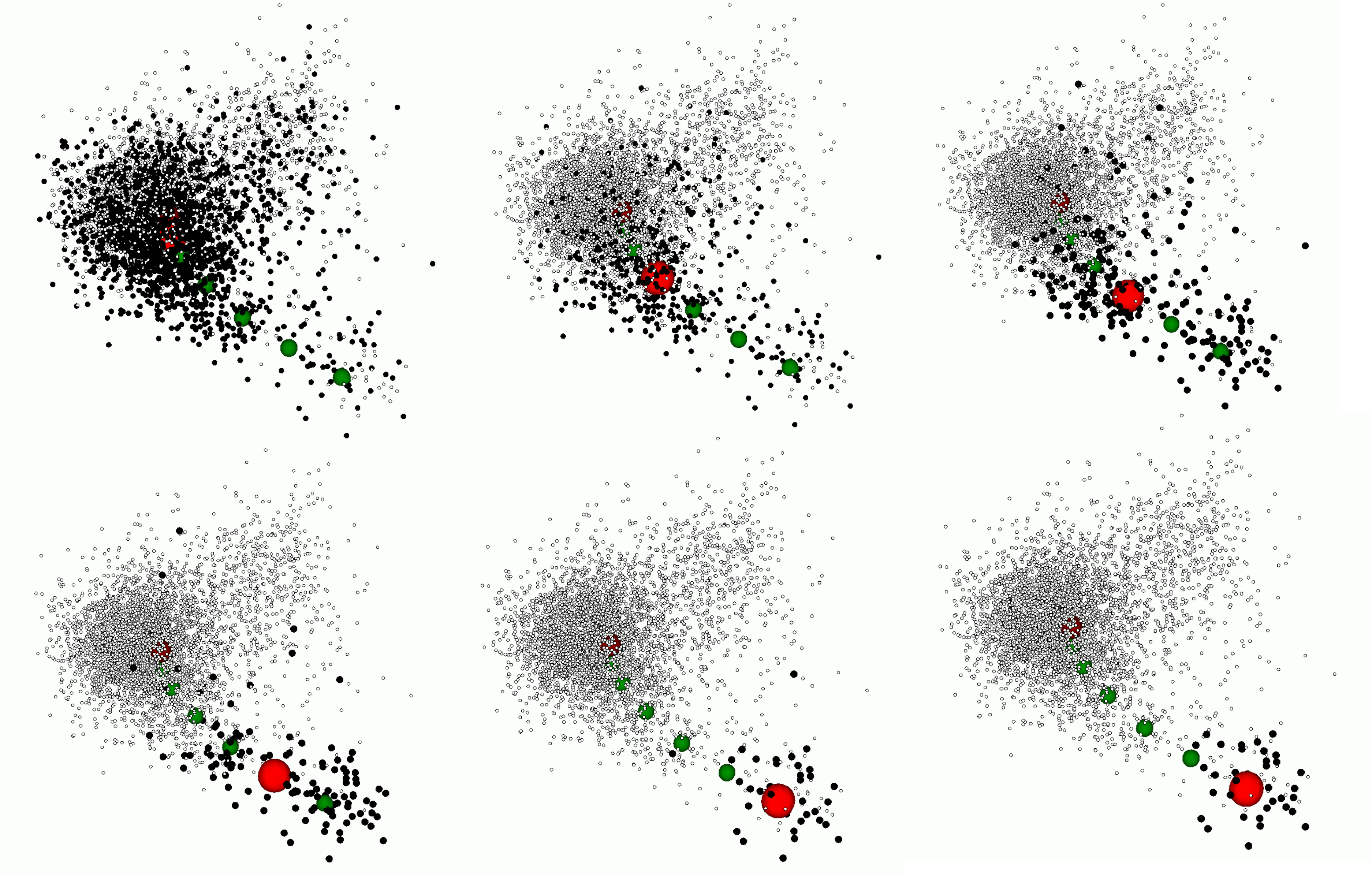




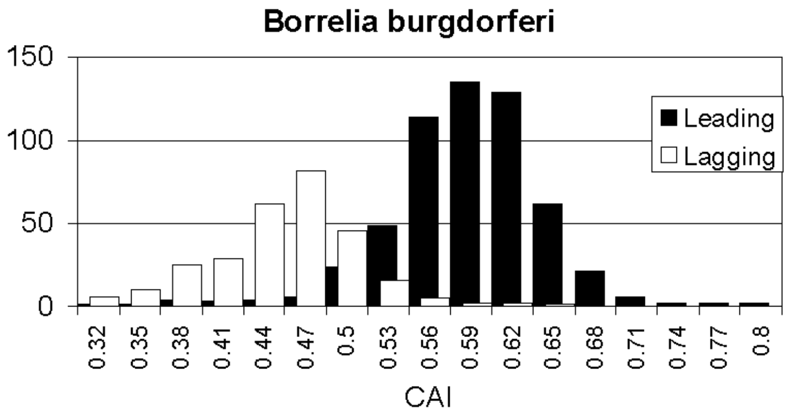




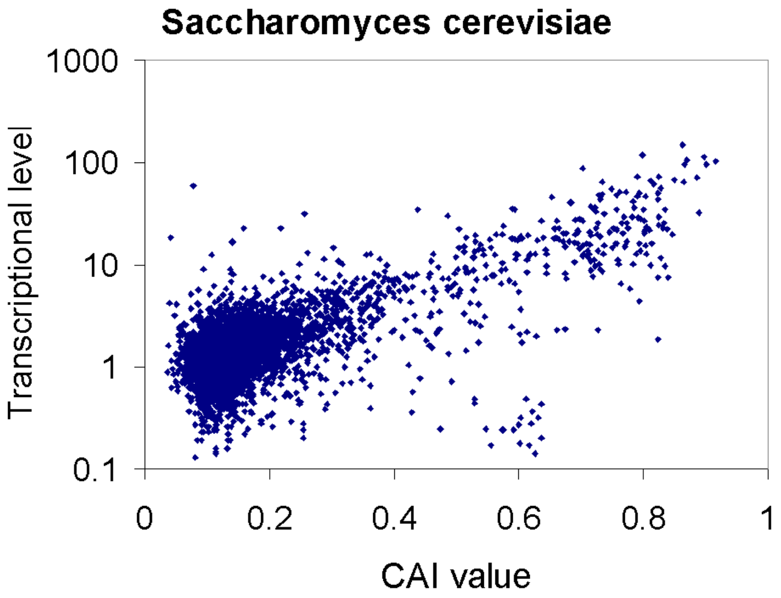
Algorithmic behavior on *Bacillus subtilis*

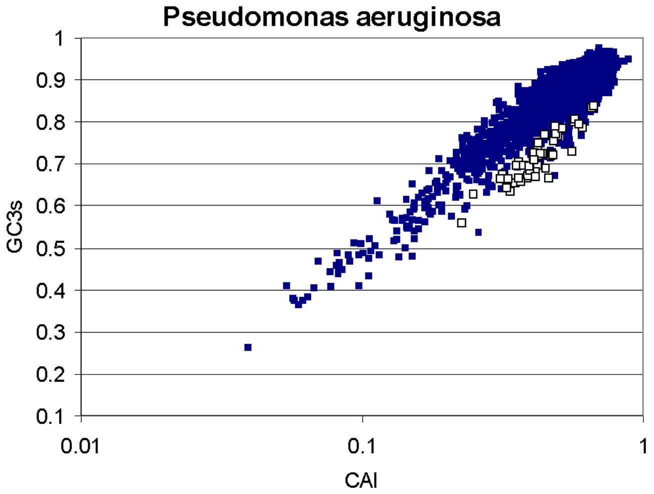


Other dominant biases



Strand bias





Transcriptional bias

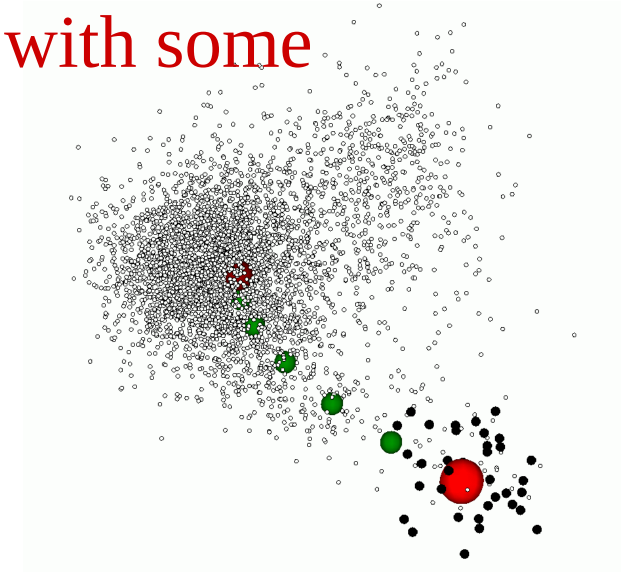
GC3 bias

Randomised version

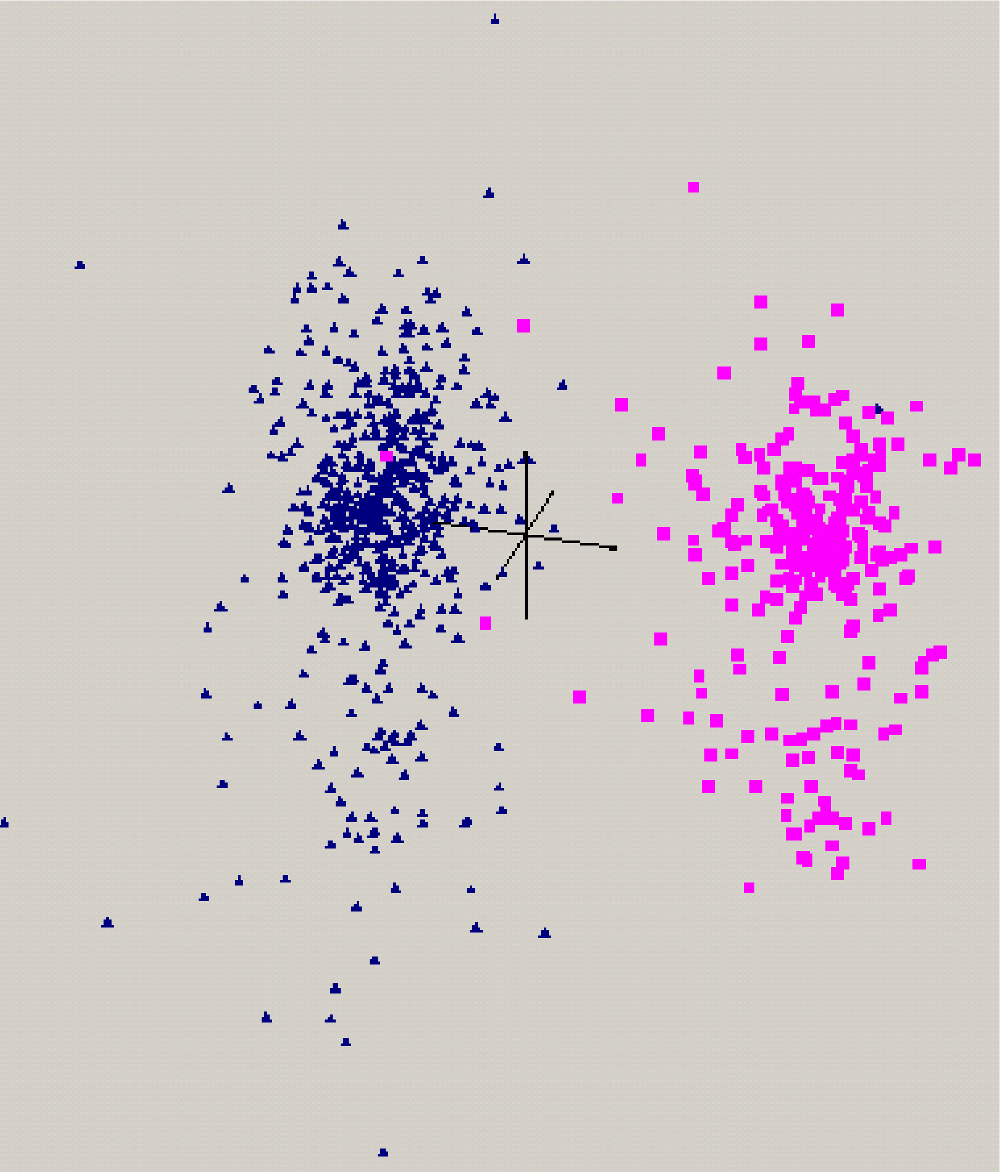
- Randomly choose the 1% of genes in S
- Compute the weights and the CAI values
- Select the 1% of genes with highest CAI value
- Repeat the iteration until the algorithm converges

We usually converge to the same set, with some exception :

1. the set contains **horizontally transferred genes**



2. in the presence of a leading-lagging strand bias in *B.burgdorferi*



Leading strand (565)

has a G+T rich bias

Lagging strand (286)

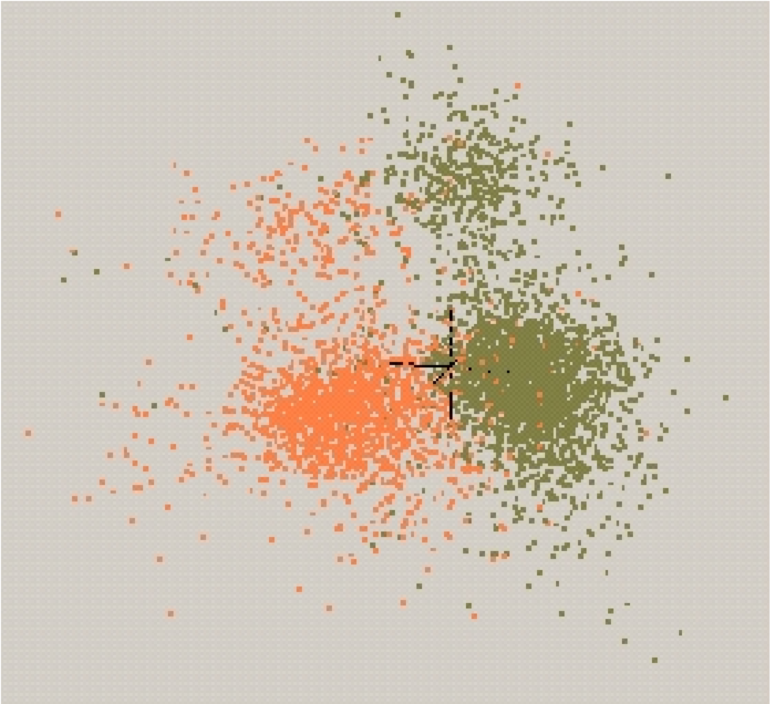
(no plasmids included)

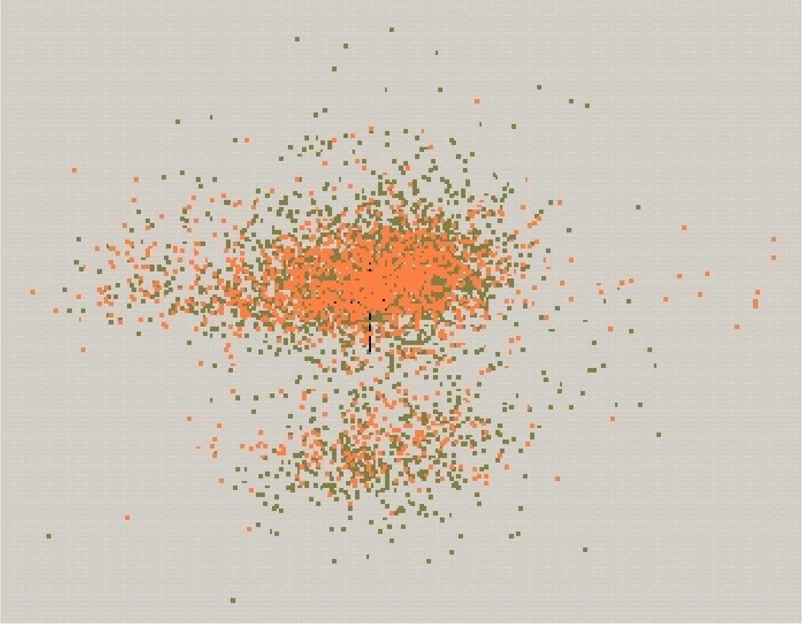
- Analysis done on **96 *bacterial* genomes, *S.cerevisiae, C.elegans, D.melanogaster***;
  we found all main results already known in the area, and **NEW** ones

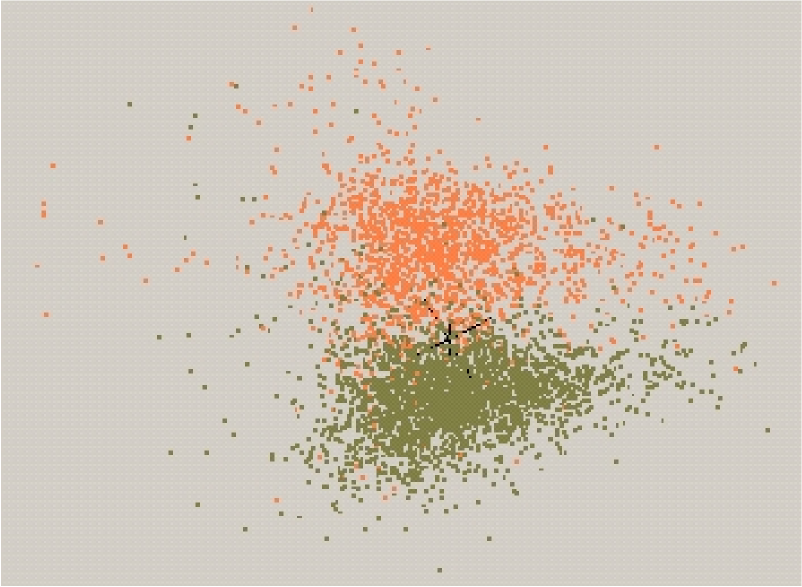
- Interesting perspectives open in the context of **new sequenced genomes** :

1. genome comparison
2. identification of binding sites in promoter regions
3. metabolic networks and codon bias

Haemophilus influenzae
Staphylococcus aureus



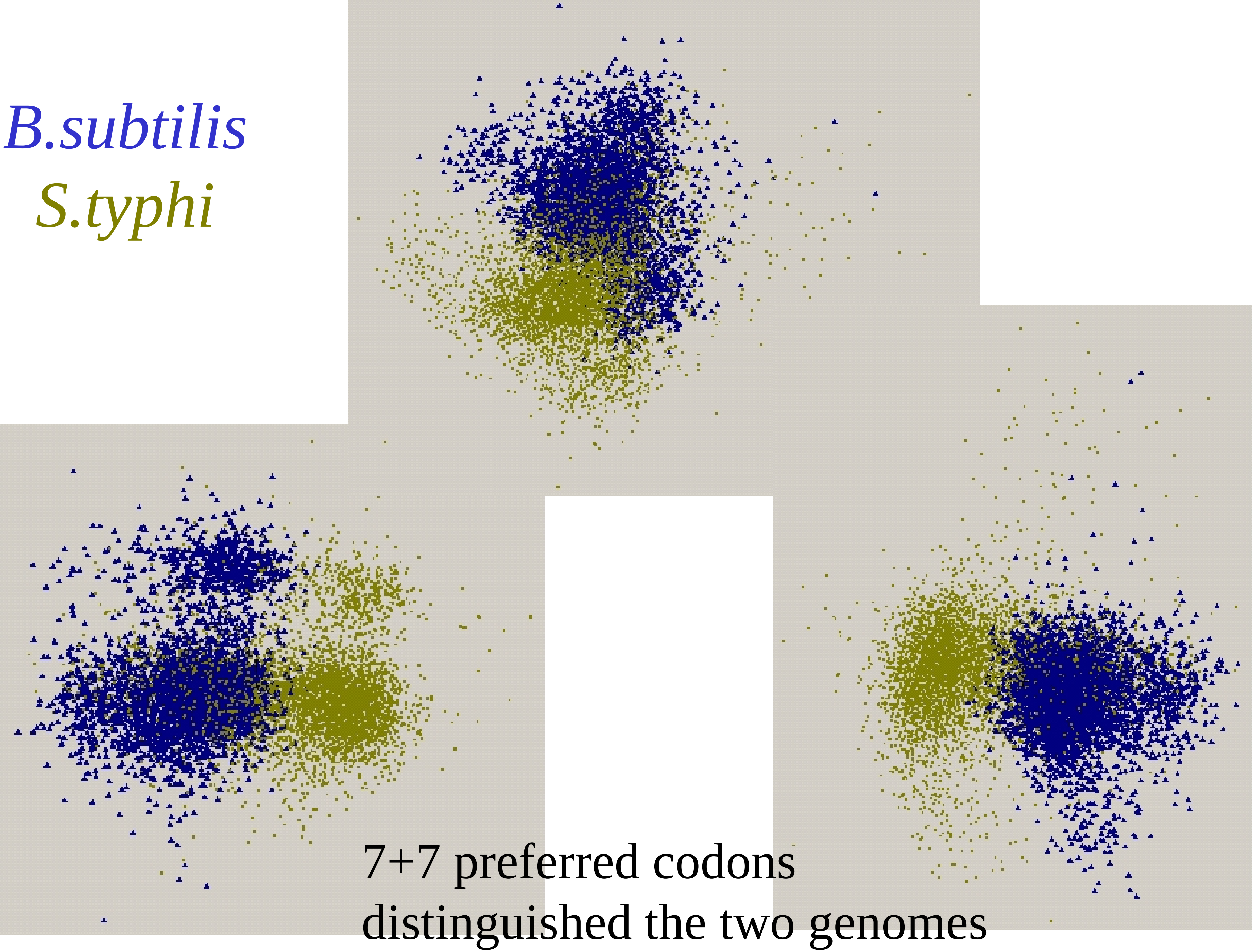




Only 2+2 preferred codons distinguished the two genomes

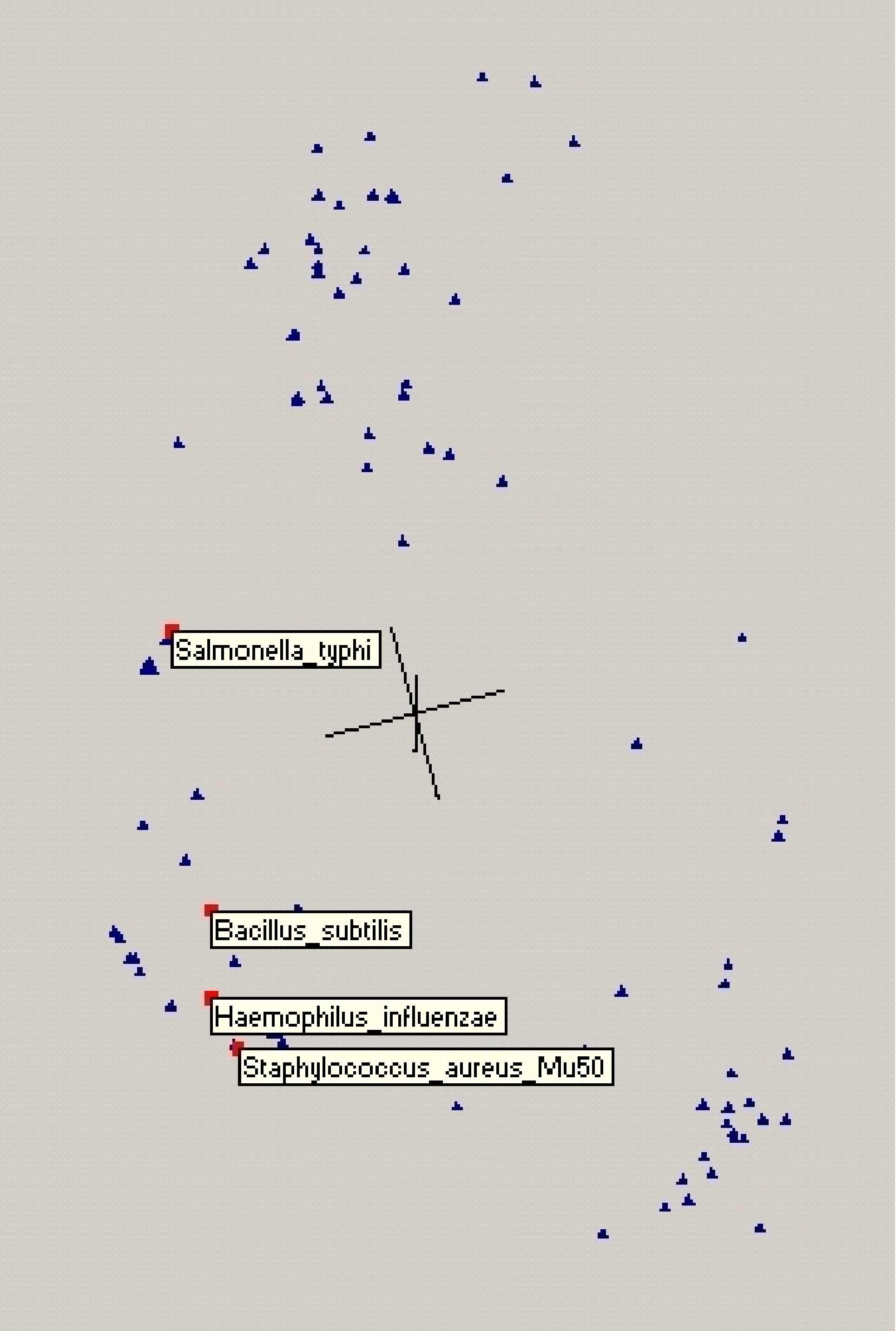
B.subtilis

S.typhi



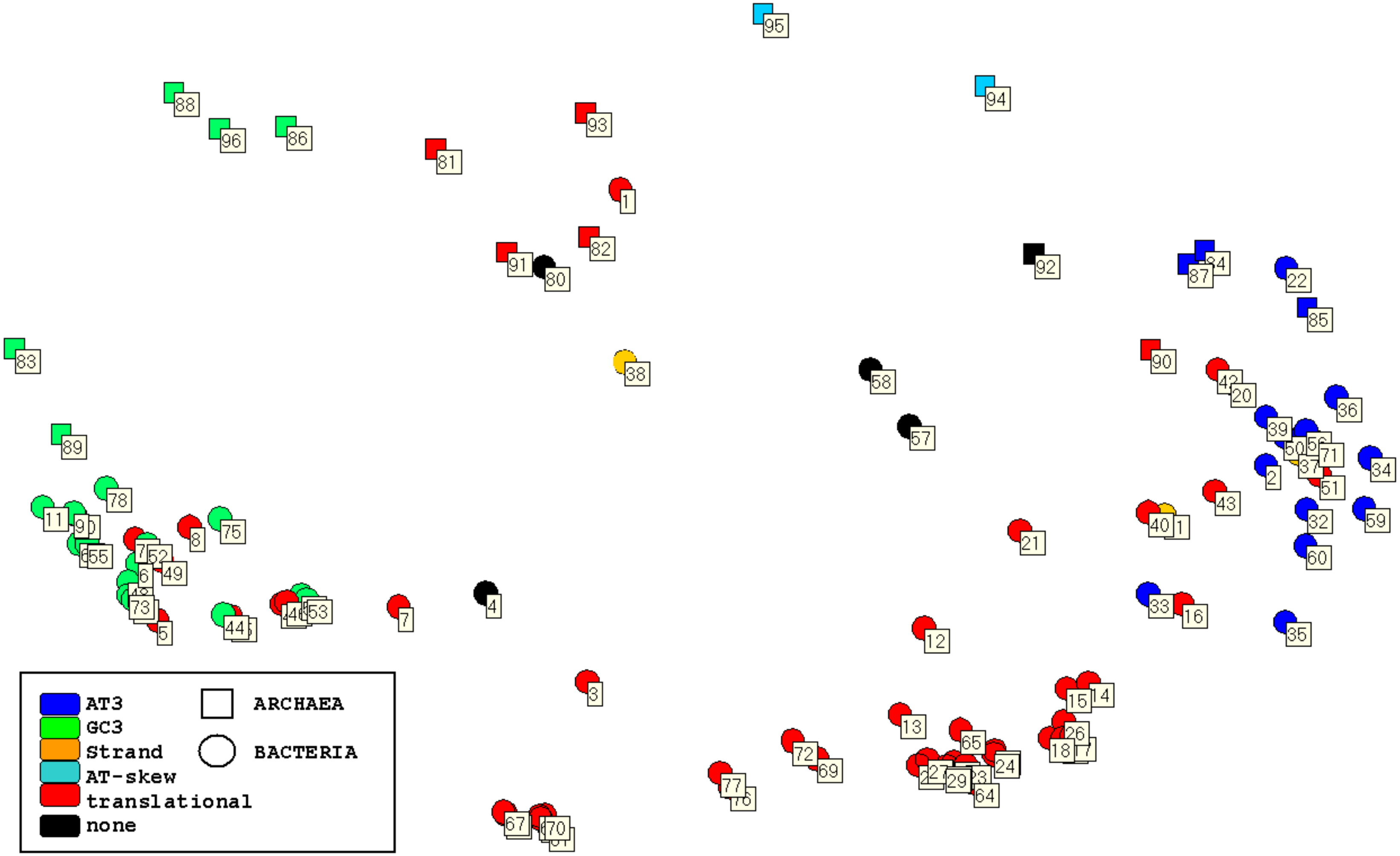
7+7 preferred codons
distinguished the two genomes

The four prokaryotic organisms in codon space



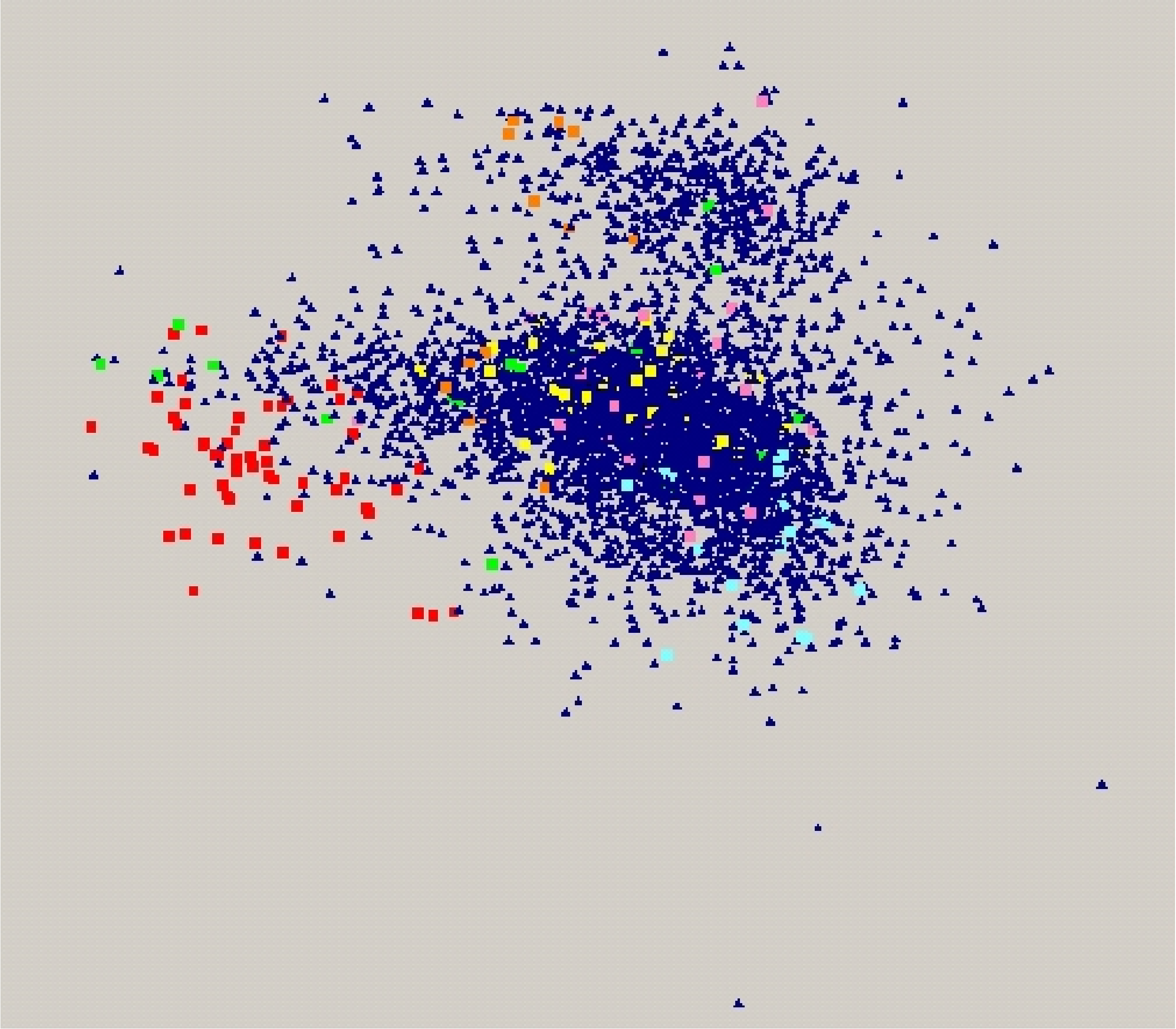
An organism is the vector of its codon weights

Prokaryotic space



Interactive interface at www.ihes.fr/~materials

E.coli



Ribosomal proteins

ATP binding proteins

IS proteins

NADH proteins

Flagellar biosynthesis proteins

Lipoproteins, membrane proteins, transport proteins

A.Carbone, A.Zinovyev, F.Képès
“Codon Adaptation Index as a measure of dominating codon bias”,
Bioinformatics 2003, to appear.
Preprint and data at http://www.ihes.fr/ and
http://www.ihes.fr/~materials

Collaborations

Organism organisation : F.Képès (CNRS, génopole Evry)
A.Zinovyev (IHÉS)

Analysis of promoter regions : Jacques van Helden (ULB, Bruxelles)

Metabolic pathways : R.Madden (IHÉS)