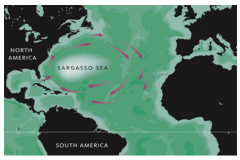
Genomes, metagenomes and environments: a perspective

Alessandra Carbone
Laboratoire de Genomique des Microorganismes
UMR7238 CNRS-Université Pierre et Marie Curie

The sargasso sea | **Santa Cruz whale carcass bone**

Different profiles of gene enrichment in environment specific-functions



The sargasso sea | **Santa Cruz whale carcass bone**

Nutrient poor environment, and its genes for ABC-type transporters dedicated to amino-acids transport and metabolism are translationally optimized.

Microbes live in an abundant food source. Translational optimization is shown in energy production and conversion genes.

This difference can reflect functional adaptation of microbes to different environmental conditions

The sargasso sea | **Santa Cruz whale carcass bone**

Nutrient poor environment, and its genes for ABC-type transporters dedicated to amino-acids transport and metabolism are **translationally optimized**.

Microbes live in an abundant food source. **Translational optimization** is shown in energy production and conversion genes.

This difference can reflect functional adaptation of microbes to different environmental conditions

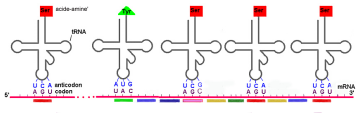
Three main results are shown for microbes:

1. Genome coding (codon bias) can be automatically studied and a set of optimized (most biased) genes can be identified for each genome
2. There exists a genome organization based on codon bias that reflects environmental living conditions
3. Codon biases reflect metabolic processes important for an organism

Are these statements true for microbial communities?

Codon usage and translational optimization

Transfert RNA



Preferred codon

Codon usage and translational optimization

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

high tRNA number
high expression

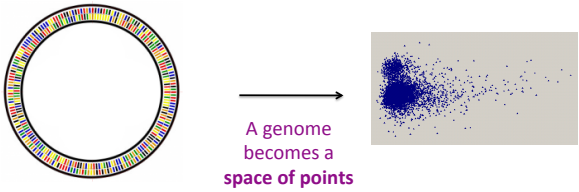
correlated to
(experimentally)

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.

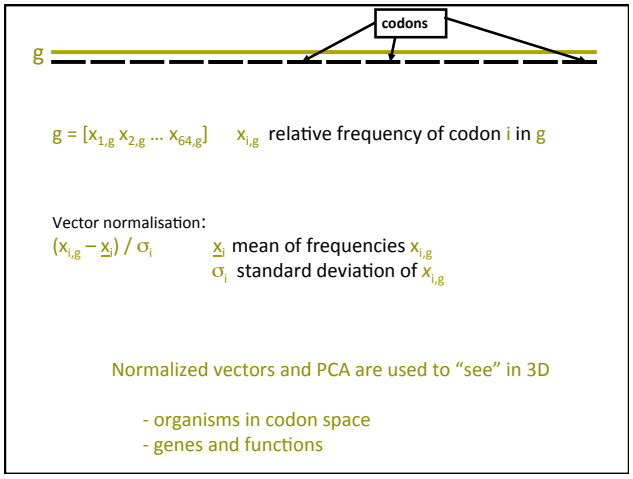
Towards the identification of a genome codon bias

A gene becomes
a **point**



A genome
becomes a
space of points

(Médigue et al., *J. Mol Biol* 1991)



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

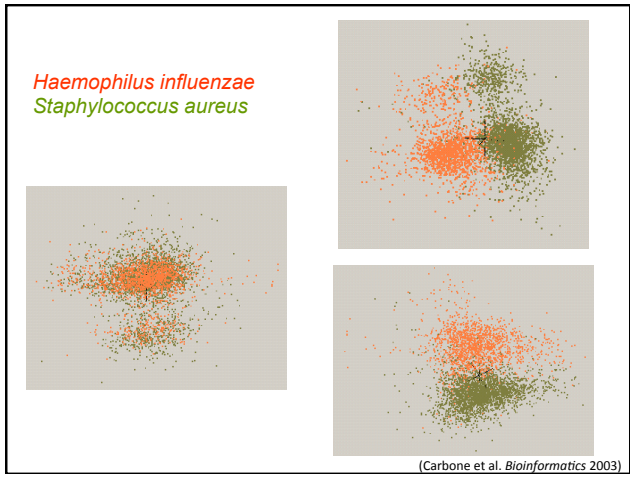
Vector normalisation:

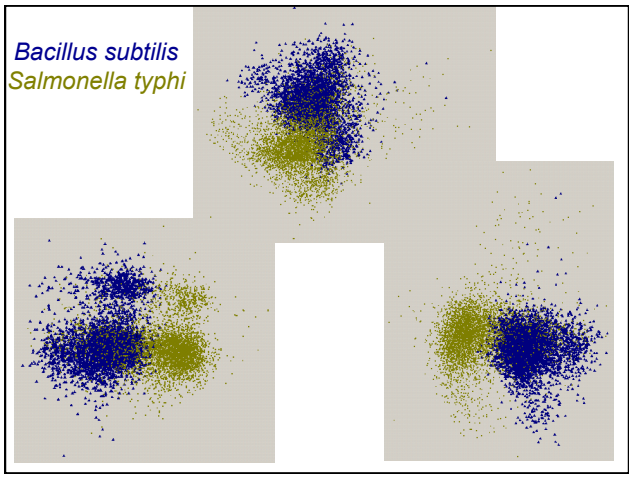
$(x_{i,g} - \underline{x}_i) / \sigma_i$ $\underline{x}_i$ mean of frequencies $x_{i,g}$

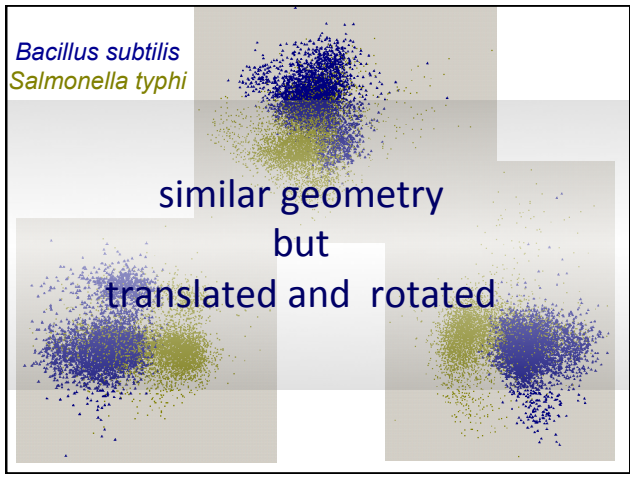
σ_i standard deviation of $x_{i,g}$

Normalized vectors and PCA are used to "see" in 3D

- organisms in codon space
- genes and functions

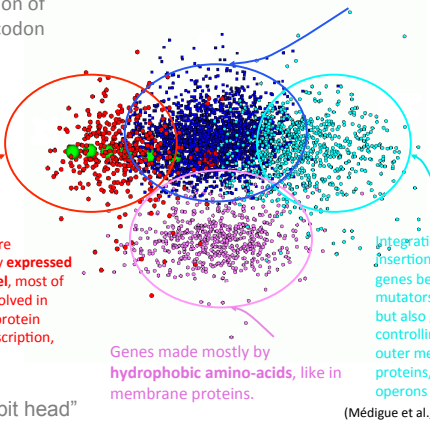






Functional organization of genes in codon space

Genes that maintain a **low or intermediary level of expression**, but that at times can be expressed at a very high level



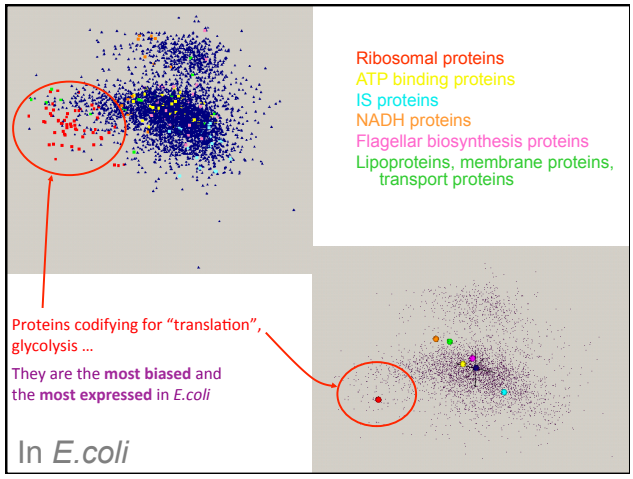
Genes that are constitutively **expressed at a high level**, most of them are involved in translation, protein folding, transcription, DNA binding

Integration host factors, insertion sequences, genes behaving as mutators when inactive, but also genes controlling cell division, outer membrane proteins, catabolic operons

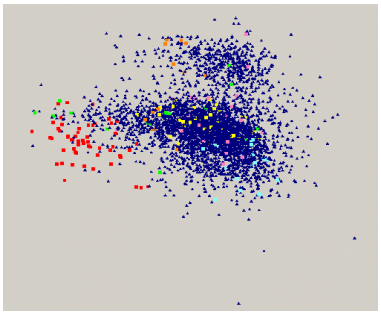
Genes made mostly by **hydrophobic amino-acids**, like in membrane proteins.

"The rabbit head"

(Médigue et al., *J. Mol Biol* 1991)

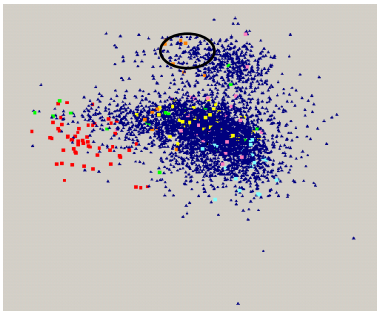


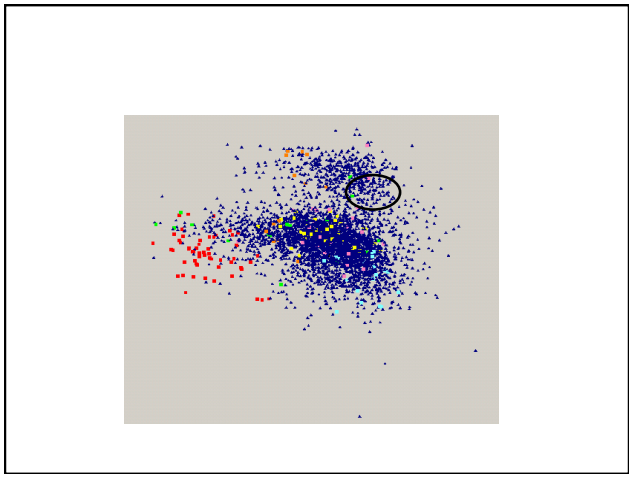
How to define "codon bias" and how to search for highly biased genes in an **automatic** manner?

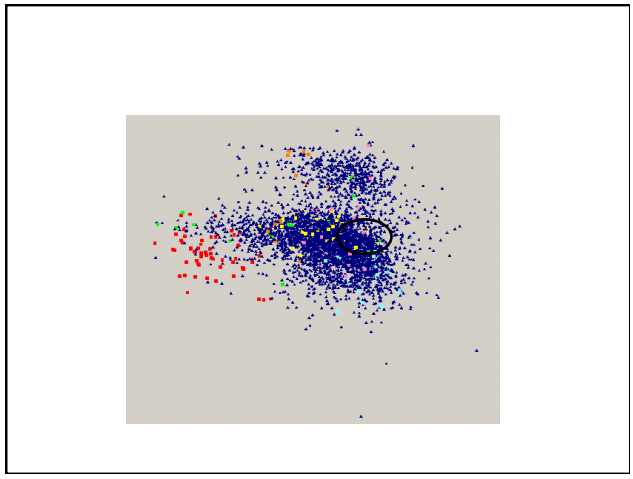


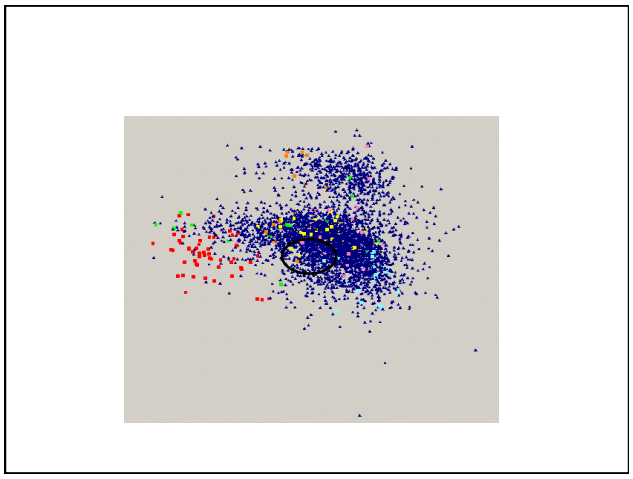
We look for a **dominant** bias

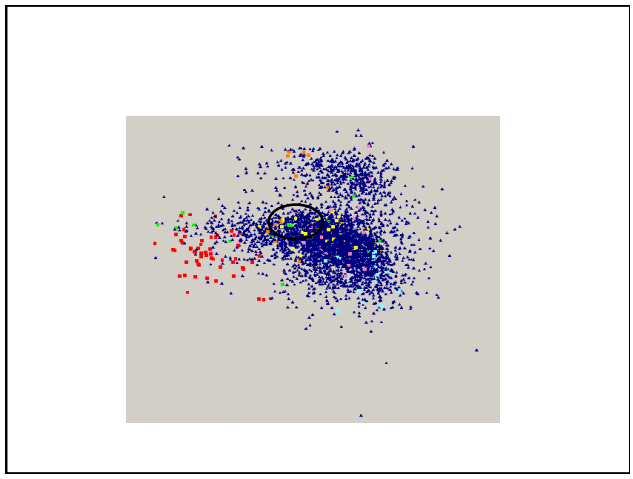
(Carbone et al. *Bioinformatics* 2003)

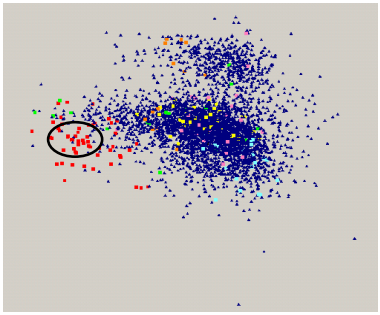












g

w_1 w_2 w_3 w_4 w_5 w_6 w_7 w_8 w_9 w_{10} w_{11}

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

L number of codons in g

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

How to compute S for indexing the genes?

- Take S by choosing randomly the 1% of genes in G
- Compute weights and index values
- Select, as new S, the 1% of genes with the highest index
- Repeat the iteration until convergence

$(\Pi_{k=1\ldots11}\, w_k)^{1/L}$ Self-Consistent Codon Index

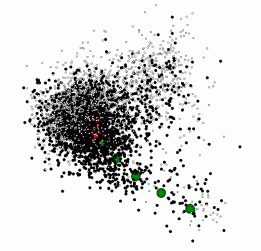
L number of codons in g

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

- Take S by choosing randomly the 1% of genes in G
- Compute weights and index values
- Select, as new S, the 1% of genes with the highest index
- Repeat the iteration until convergence

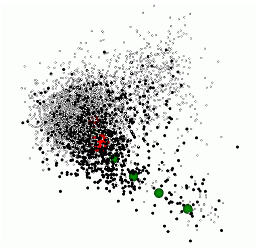
Idea of the "full genome search" algorithm:

- 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.



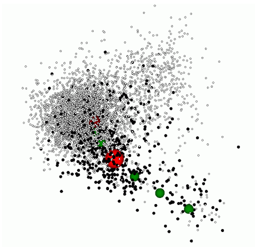
Idea of the algorithm:

- 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.



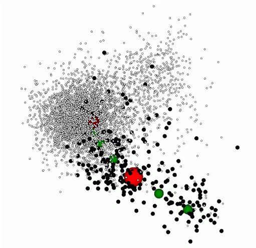
Idea of the algorithm:

- 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.



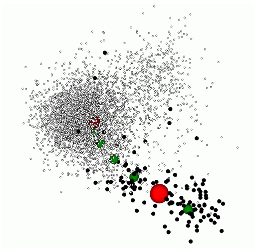
Idea of the algorithm:

- 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.



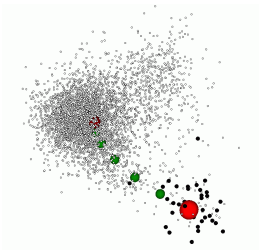
Idea of the algorithm:

- 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.



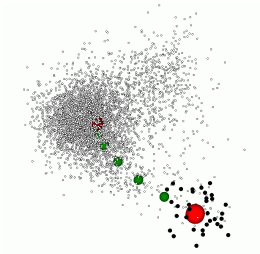
Idea of the algorithm:

- 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:

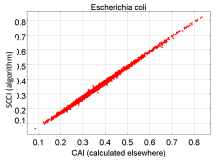
- Compute the weight of the codons over the whole genome and compute afterwards SCCI values for all genes
- Select the 50% of genes with the highest SCCI value
- Repeat the iteration and select the 25% of the genes
- and so on… until we arrive to the 1% of genes in the original set.
- … then repeat the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



The algorithm associates to each genome a **vector of weights**, one for each codon, representing the occurrence of the codon within the most biased set of genes of the genome.

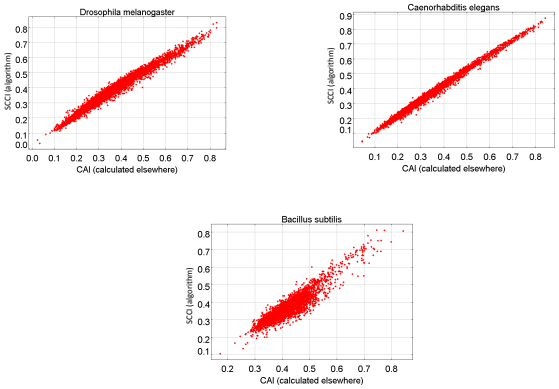
Most biased genes in *E.coli*

(*E.coli* reproduce rapidly)

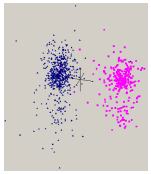


Gene	Annotation
tufA	protein chain elongation factor EF-Tu
tufB	protein chain elongation factor EF-Tu
tsf	protein chain elongation factor EF-Ts
fusA	GTP-binding protein chain elongation factor EF-G
mopA	chaperonin GroEL
dnaK	heat shock protein DnaK
cspA	cold shock protein 7.4
tig	trigger factor
ompA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

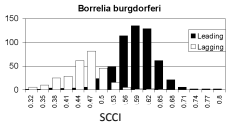


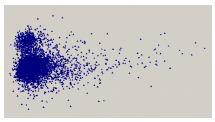
SCCI : a universal measure for dominant bias



Borrelia burgdorferi

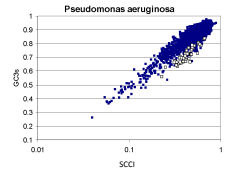
Strand bias





Pseudomonas aeruginosa

GC3 bias

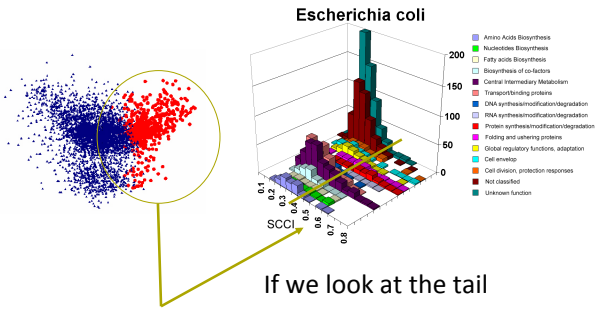


The set of biased genes S

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

What are these genes doing?

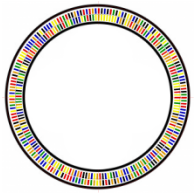
- Highly expressed genes (belonging to most species)
- Genes with uncharacterised function
- Genes dependent on specific environmental conditions
- Stress response genes
- Non-orthologous genes



If we look at the tail

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

Towards a space of genomes

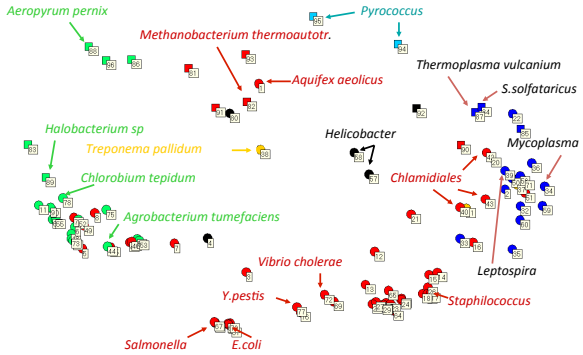


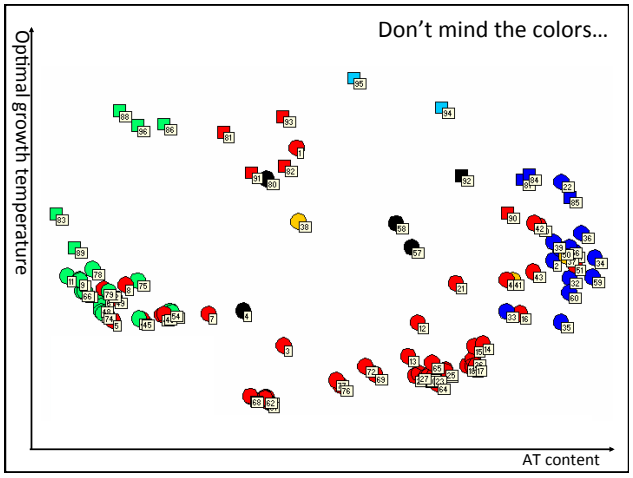
A genome
becomes a point

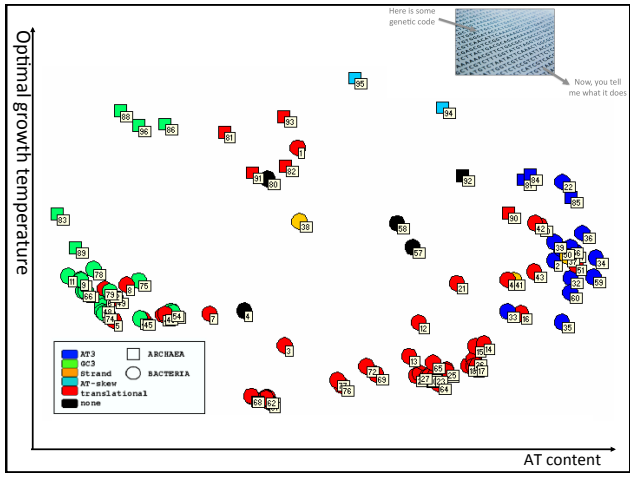
The vector of weights is taken as a signature for the genome

(Carbone et al. *Mol Biol Evol* 2004)

Bacteria and archaea in codon space

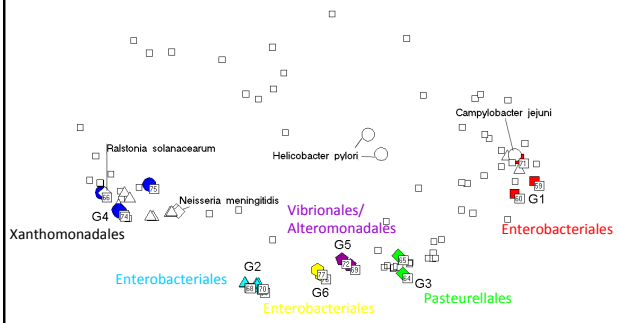






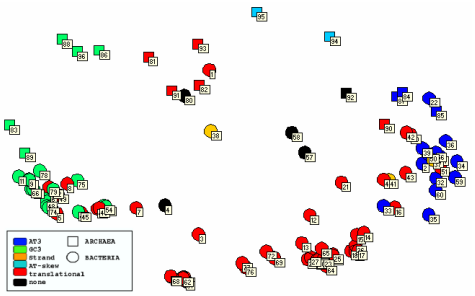
Can we exploit the geometry of the space to derive functional characteristics of groups of organisms?

Phylogenetically related families : γ-proteobacteria



Similar physiology and habitat

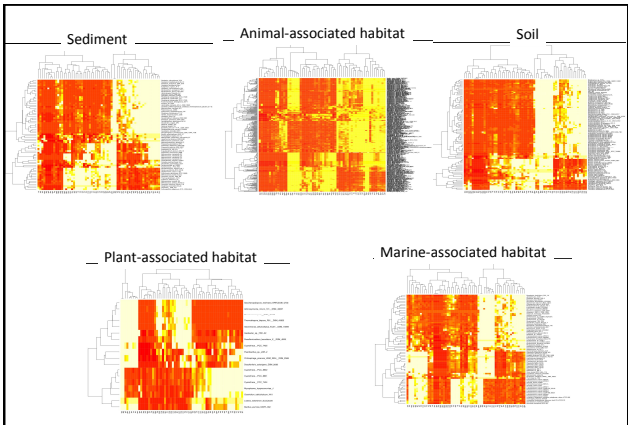
Organisms at small distance: similar physiology and habitat



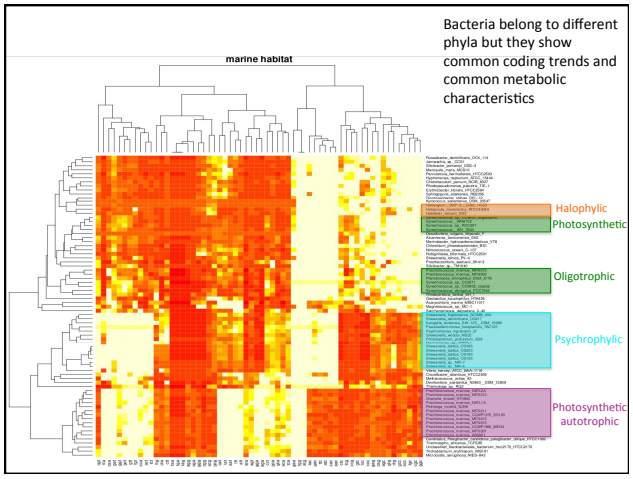
Environmental clusters : on 323 bacteria

soil bacteria
enterics
symbions
spore formers
small intercellular pathogens
small extracellular pathogens

(Willenbrock et al. *Gen Biol* 2006)



A study of 855 bacterial genomes

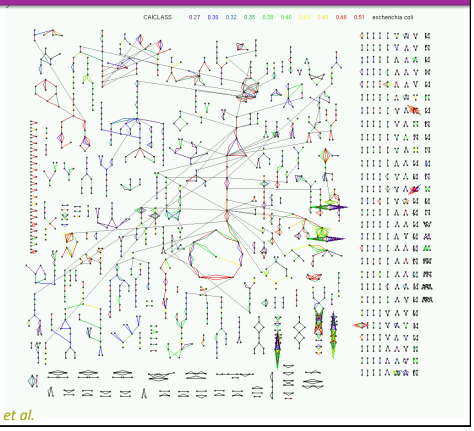


Can we use this signal to identify the most important **metabolic networks** in an organism?

(Carbone & Madden, *J Mol Evol* 2005)

Metabolic networks and coding

E.coli



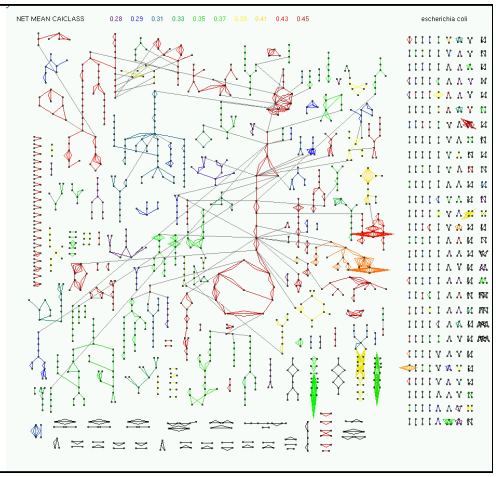
EcoCyc network, P.Karp *et al.*

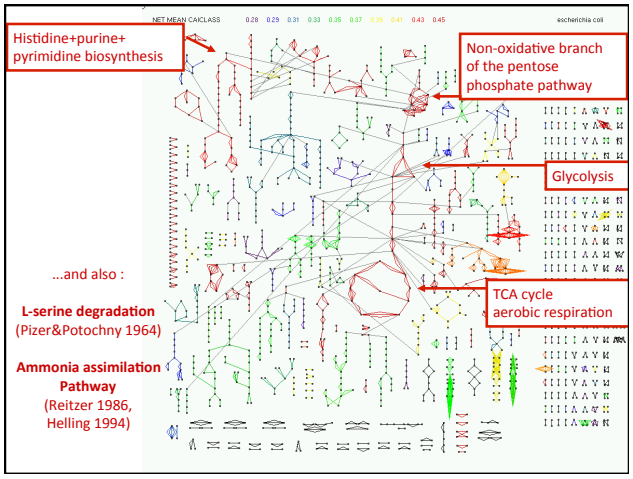
Pathway Index

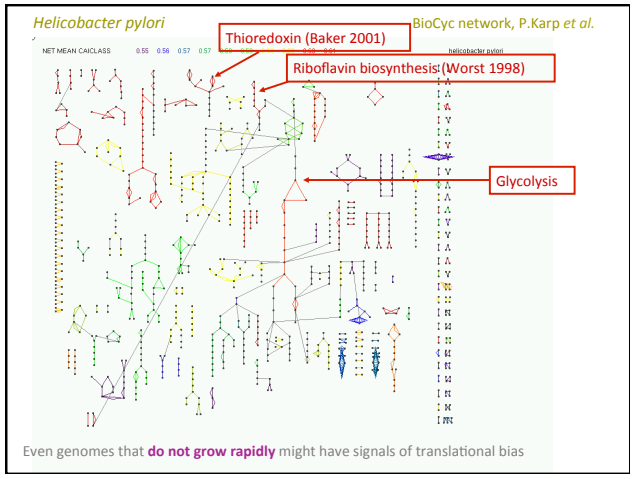
PI(P)
=
mean SCCI(g)
g∈P

Relative Pathway Index

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





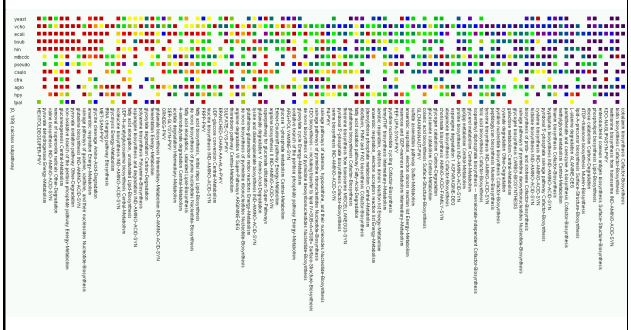


Metabolic pathways essential to *Mycobacterium tuberculosis*

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)

Towards a large scale mapping of metabolic preferences

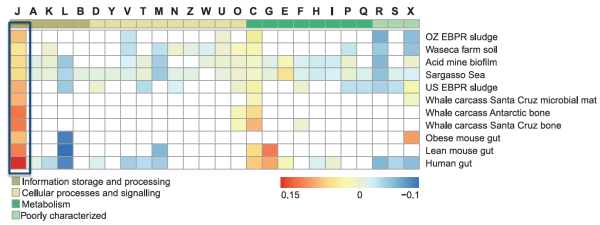


(Carbone & Madden, *J Mol Evol* 2005)

Is there a signature of a **microbial community**, similarly to the genomic signature seen for single bacterial genomes?

(Alaeitabar & Carbone, unpublished, 2011)

Enrichment of functions within highly expressed genes in **metagenomes**



Metagenomes show different functional enrichment patterns that are consistent with environmental requirements (e.g. metabolite transport functions [E] in the Sargasso Sea or energy conversion [C] in the whale carcass metagenome). Non-significant enrichments are shown in white. Letters at the top represent COG supercategories: [J] Translation, ribosomal structure and biogenesis; [A] RNA processing and modification; [K] Transcription; [L] Replication, recombination and repair; [B] Chromatin structure and dynamics; [D] Cell cycle control, cell division, chromosome partitioning; [Y] Nuclear structure; [V] Defence mechanisms; [T] Signal transduction mechanisms; [M] Cell wall/membrane/envelope biogenesis; [N] Cell motility; [Z] Cytoskeleton; [W] Extracellular structures; [U] Intracellular trafficking, secretion and vesicular transport; [O] Posttranslational modification, protein turnover, chaperones; [C] Energy production and conversion; [G] Carbohydrate transport and metabolism; [E]Amino acid transport and metabolism; [F] Nucleotide transport and metabolism; [H] Coenzyme transport and metabolism; [I] Lipid transport and metabolism; [P] Inorganic ion transport and metabolism; [Q] Secondary metabolites biosynthesis, transport and catabolism; [R] General function prediction only; [S] Function unknown; [X] Uncharacterized.

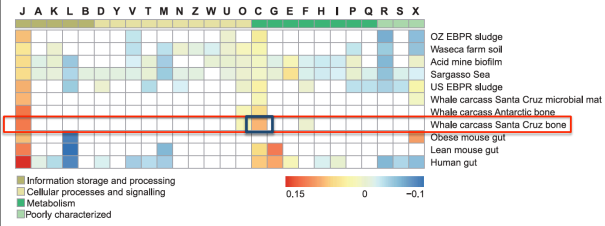
Roller M et al. Nucl. Acids Res. 2013;nar.gkt673

Enrichment of functions within highly expressed genes in metagenomes

Metagenomes show different functional enrichment patterns that are consistent with environmental requirements (e.g. metabolite transport functions [E] in the Sargasso Sea or energy conversion [C] in the whale carcass metagenome). Non-significant enrichments are shown in white. Letters at the top represent COG supercategories: [J] Translation, ribosomal structure and biogenesis; [A] RNA processing and modification; [K] Transcription; [L] Replication, recombination and repair; [B] Chromatin structure and dynamics; [D] Cell cycle control, cell division, chromosome partitioning; [Y] Nuclear structure; [V] Defence mechanisms; [T] Signal transduction mechanisms; [M] Cell wall/membrane/envelope biogenesis; [N] Cell motility; [Z] Cytoskeleton; [W] Extracellular structures; [U] Intracellular trafficking, secretion and vesicular transport; [O] Posttranslational modification, protein turnover, chaperones; [C] Energy production and conversion; [G] Carbohydrate transport and metabolism; [E]Amino acid transport and metabolism; [F] Nucleotide transport and metabolism; [H] Coenzyme transport and metabolism; [I] Lipid transport and metabolism; [P] Inorganic ion transport and metabolism; [Q] Secondary metabolites biosynthesis, transport and catabolism; [R] General function prediction only; [S] Function unknown; [X] Uncharacterized.

Roller M et al. Nucl. Acids Res. 2013;nar.gkt673

Enrichment of functions within highly expressed genes in metagenomes



Metagenomes show different functional enrichment patterns that are consistent with environmental requirements (e.g. metabolite transport functions [E] in the Sargasso Sea or energy conversion [C] in the whale carcass metagenome). Non-significant enrichments are shown in white. Letters at the top represent COG supercategories: [J] Translation, ribosomal structure and biogenesis; [A] RNA processing and modification; [K] Transcription; [L] Replication, recombination and repair; [B] Chromatin structure and dynamics; [D] Cell cycle control, cell division, chromosome partitioning; [Y] Nuclear structure; [V] Defence mechanisms; [T] Signal transduction mechanisms; [M] Cell wall/membrane/envelope biogenesis; [N] Cell motility; [Z] Cytoskeleton; [W] Extracellular structures; [U] Intracellular trafficking, secretion and vesicular transport; [O] Posttranslational modification, protein turnover, chaperones; [C] Energy production and conversion; [G] Carbohydrate transport and metabolism; [E]Amino acid transport and metabolism; [F] Nucleotide transport and metabolism; [H] Coenzyme transport and metabolism; [I] Lipid transport and metabolism; [P] Inorganic ion transport and metabolism; [Q] Secondary metabolites biosynthesis, transport and catabolism; [R] General function prediction only; [S] Function unknown; [X] Uncharacterized.

Roller M et al. Nucl. Acids Res. 2013;nar.gkt673

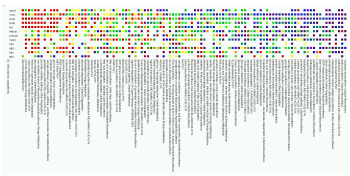
Conclusions

- These findings suggest that microbial communities are representable by **genomic signatures**, specific to different communities, as single genomes are.
- This might be true for bacterial communities but also for eukaryotic ones. For these latter, **assembly** is much harder and our approach does not ask for large contigs for the analysis.
- The community-wide "optimization effect" is an important metagenomic feature with **predictive power**: genes with unknown function that are potentially important for the community can be identified

Conclusions

- Likely, we will be able to rank metabolic functions and orthologous groups of genes at the **system level.** Such effort is likely to be important to understand the adaptation of the entire metagenome to its particular environment.

Can we draw a metabolic map for communities?



Conclusions

Analysis at the systemic level should go parallel to an improvement of **gene annotation**.

We should work at the **annotation level**:

On a test realized on 51 bacterial genomes containing 159 930 CDS and 189 726 annotated domains (by Pfam) we found 28 107 new domains.

(Bernardes et al. submitted, 2013)
(Ugarte et al, in preparation, 2013)

Conclusions

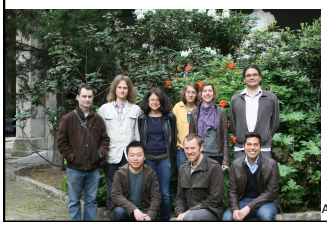
We should work at the **annotation level**:

Table 1: **Percentage of proteins with unknown function in different genomes**

Kingdom	Clade	Species	number of proteins	% of unknown proteins
Eukaryota	Metazoa	*Brugia malayi*	11472	20
		Caenorhabditis elegans	26047	24
		Drosophila melanogaster	27752	19
		Anopheles Gambiae	14576	22
		Ciona intestinalis	4122	25
	Fungi	*Saccharomyces cerevisiae*	6607	23
	Amoebozoa	*Entamoeba histolytica*	8201	44
		Plasmodium falciparum	5491	30
		Dictyostelium discoideum	12646	25
	Diplomonadida	*Giardia intestinalis*	5012	49
	Cryptophyta	*Giardia theta*	24822	51
		Bigelowiella natans	21000	47
	Stramenopiles	*Phaeodactylum tricornutum*	10408	20
	Parabasalia	*Trichomonas vaginalis*	59681	43
Bacteria	Proteobacteria	*Salmonella enterica*	4720	33
		Helicobacter pylori	1593	32
	Elusimicrobia	*Elusimicrobium minutum*	1548	35
	Firmicutes	*Staphylococcus aureus*	2620	62
		Enterococcus faecalis	3278	27
		Streptococcus pneumoniae	1990	23
	Actinobacteria	*Mycobacterium tuberculosis T46*	4134	42
	Cyanobacteria	*Microcystis aeruginosa*	5356	50
Archaea	Euryarchaeota	*Methanobrevibacter smithii*	1710	25
		Methanococcus maripaludis	1807	31
		Halobacterium salinarum	2749	47
	Korarchaeota	*Candidatus Korarchaeum cryptofilum*	1612	25
	Thaumarchaeota	*Cenarchaeum symbiosum A*	2017	48
	Crenarchaeota	*Pyrobaculum aguniense TE7*	2869	38

Juliana Bernardes Tina Alaeitabar

Angela Falciatore



Ari Ugarte

Juliana Bernardes Tina Alaeitabar

Ari Ugarte

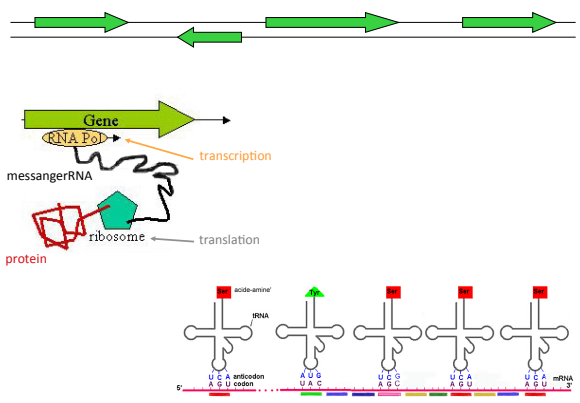
Angela Falciatore

Intra-species variations

Percentage of proteins with unknown function in different genomes

Kingdom	Clade	Species	number of proteins	% of unknown proteins
Eukaryota	Metazoa	*Brugia malayi*	11472	20
		Caenorhabditis elegans	26047	24
		Drosophila melanogaster	27752	19
		Anopheles Gambiae	14576	22
		Ciona intestinalis	4122	25
	Fungi	*Saccharomyces cerevisiae*	6607	23
	Amoebozoa	*Entamoeba histolytica*	8201	44
		Plasmodium falciparum	5491	30
		Dictyostelium discoideum	12646	25
	Diplomonadida	*Giardia intestinalis*	5012	49
	Cryptophyta	*Giardia theta*	24822	51
		Bigelowiella natans	21000	47
	Stramenopiles	*Phaeodactylum tricornutum*	10408	20
	Parabasalia	*Trichomonas vaginalis*	59681	43
Bacteria	Proteobacteria	*Salmonella enterica*	4720	33
		Helicobacter pylori	1593	32
	Elusimicrobia	*Elusimicrobium minutum*	1548	35
	Firmicutes	*Staphylococcus aureus*	2620	62
		Enterococcus faecalis	3278	27
		Streptococcus pneumoniae	1990	23
	Actinobacteria	*Mycobacterium tuberculosis T46*	4134	42
	Cyanobacteria	*Microcystis aeruginosa*	5356	50
Archaea	Euryarchaeota	*Methanobrevibacter smithii*	1710	25
		Methanococcus maripaludis	1807	31
		Halobacterium salinarum	2749	47
	Korarchaeota	*Candidatus Korarchaeum cryptofilum*	1612	25
	Thaumarchaeota	*Cenarchaeum symbiosum A*	2017	48
	Crenarchaeota	*Pyrobaculum oguniense TE7*	2869	38

Genes and redundancy of the genetic code



$$CAI(g) = (\Pi_{k=1...L} w_k)^{1/L}$$
Codon Adaptation Index

(Sharp & Li, 1987)

L number of codons in g

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

proteines codifying for "translation", glycolysis …

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

$$CAI(g) = (\Pi_{k=1...L} w_k)^{1/L}$$
Codon Adaptation Index

(Sharp & Li, 1987)

L number of codons in g

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

proteines codifying for "translation", glycolysis …

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

$$CAI(g) = (\Pi_{k=1...L} w_k)^{1/L}$$
Codon Adaptation Index

(Sharp & Li, 1987)

L number of codons in g

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

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

L number of codons in g

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

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

Can we classify microbial organisms starting from codon optimization?

Here is some genetic code

Now, you tell me what it does

Bacteria and Archaea in SCCI codon space