SENCA: A codon substitution model to better estimate evolutionary processes

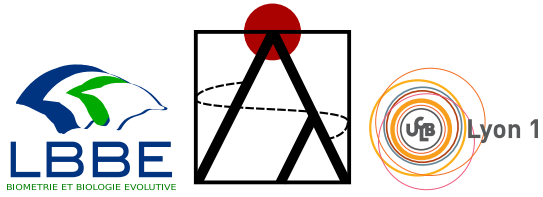
–

Fanny Pouyet

Marc Bailly-Bechet, Dominique Mouchiroud
and Laurent Guéguen

–

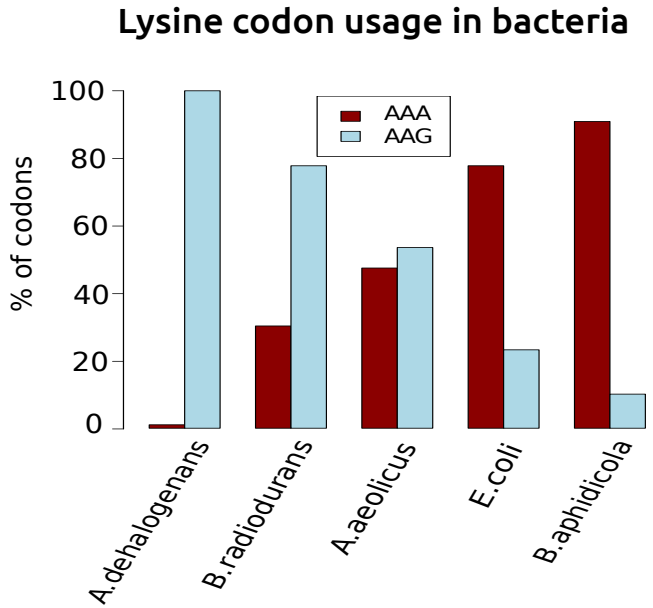
LBBE - UCB Lyon - UMR 5558
June 2015, Porquerolles, France.

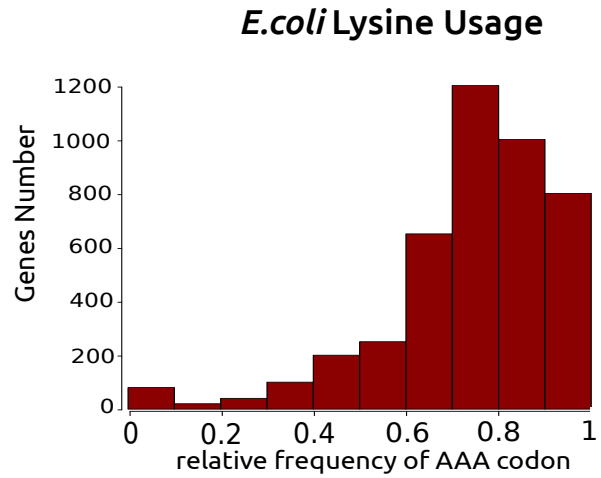


Codon Usage Bias - CUB

	U	C	A	G
U	UUU Phe UUC UUA Leu UUG	UCU UCC UCA Ser UCG	UAU Tyr UAC UAA *Stop* UAG	UGU Cys UGC UGA *Stop* UGG Trp
C	CUU CUC CUA Leu CUG	CCU CCC CCA Pro CCG	CAU His CAC CAA Gln CAG	CGU CGC CGA Arg CGG
A	AUU AUC AUA Ile AUG Met	ACU ACC ACA Thr ACG	AAU Asn AAC AAA Lys AAG	AGU Ser AGC AGA Arg AGG
G	GUU GUC GUA Val GUG	GCU GCC GCA Ala GCG	GAU Asp GAC GAA Glu GAG	GGU GGC GGA Gly GGG

61 sense codons, 20 amino acids

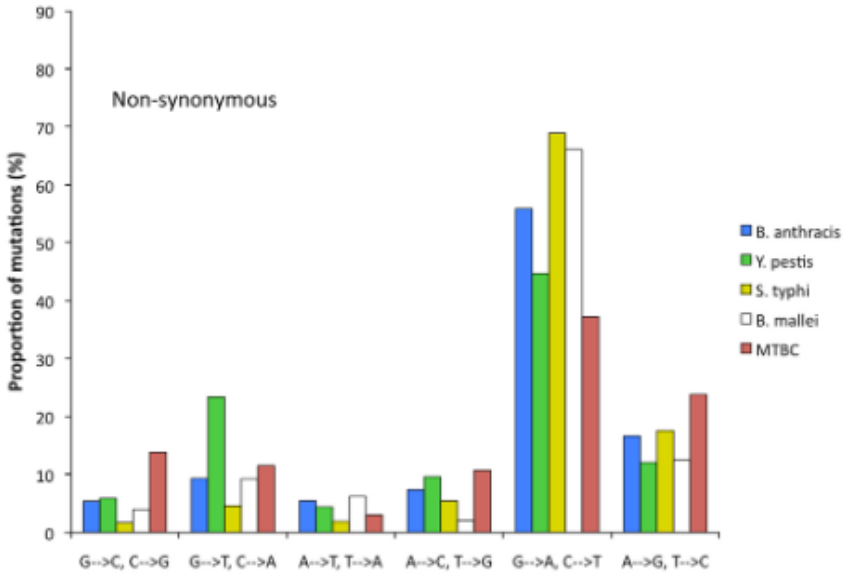


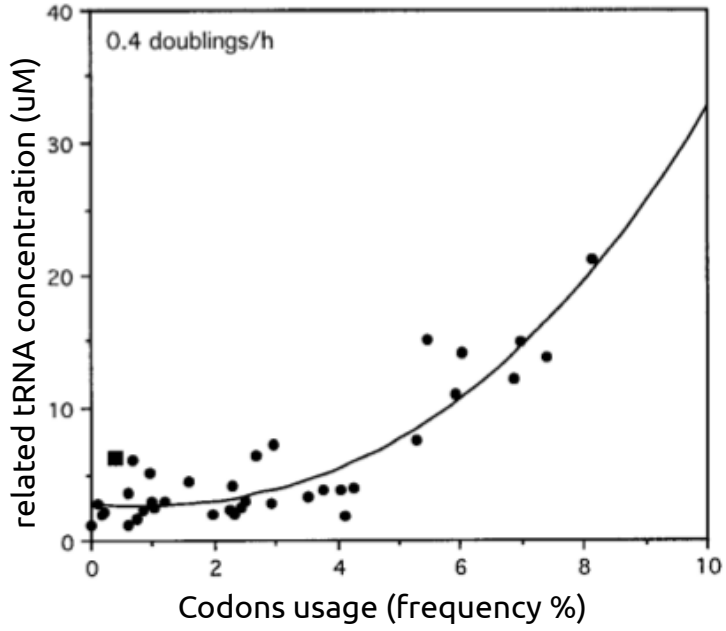


CUB

How to measure it?

Origins of the CUB: mutational *vs.* selective explanations.





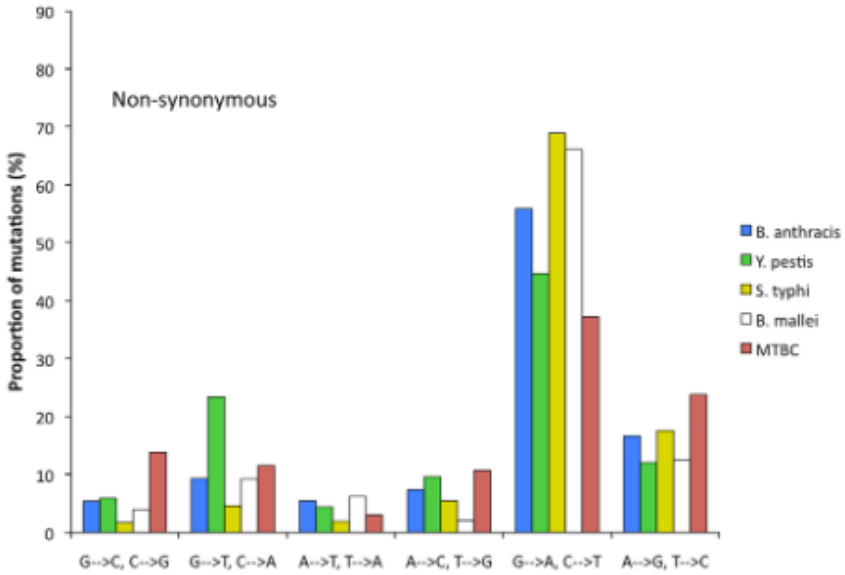
Objectives: Disentangling the two hypotheses

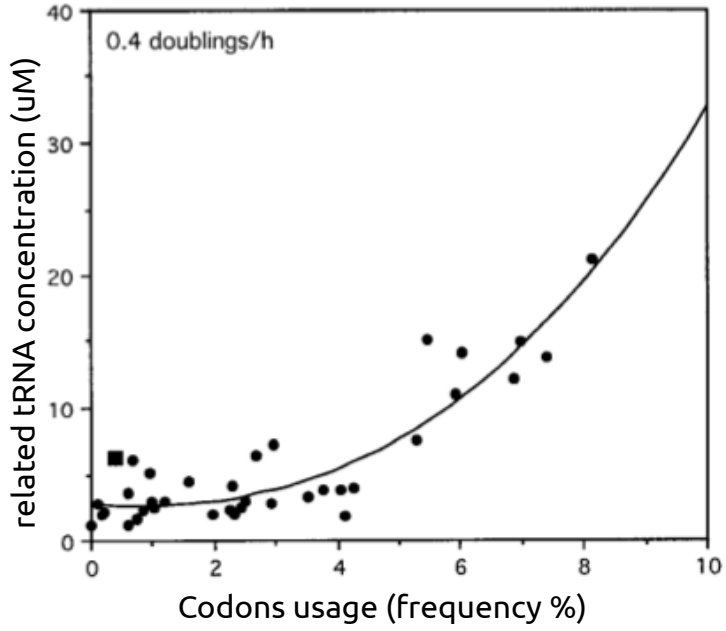
Modelling evolutive processes explaining the observed CUB: at the nucleotidic (N), the codons (C) and the AA layers (A).

CUB

How to measure it?

Origins of the CUB: mutational *vs.* selective explanations.



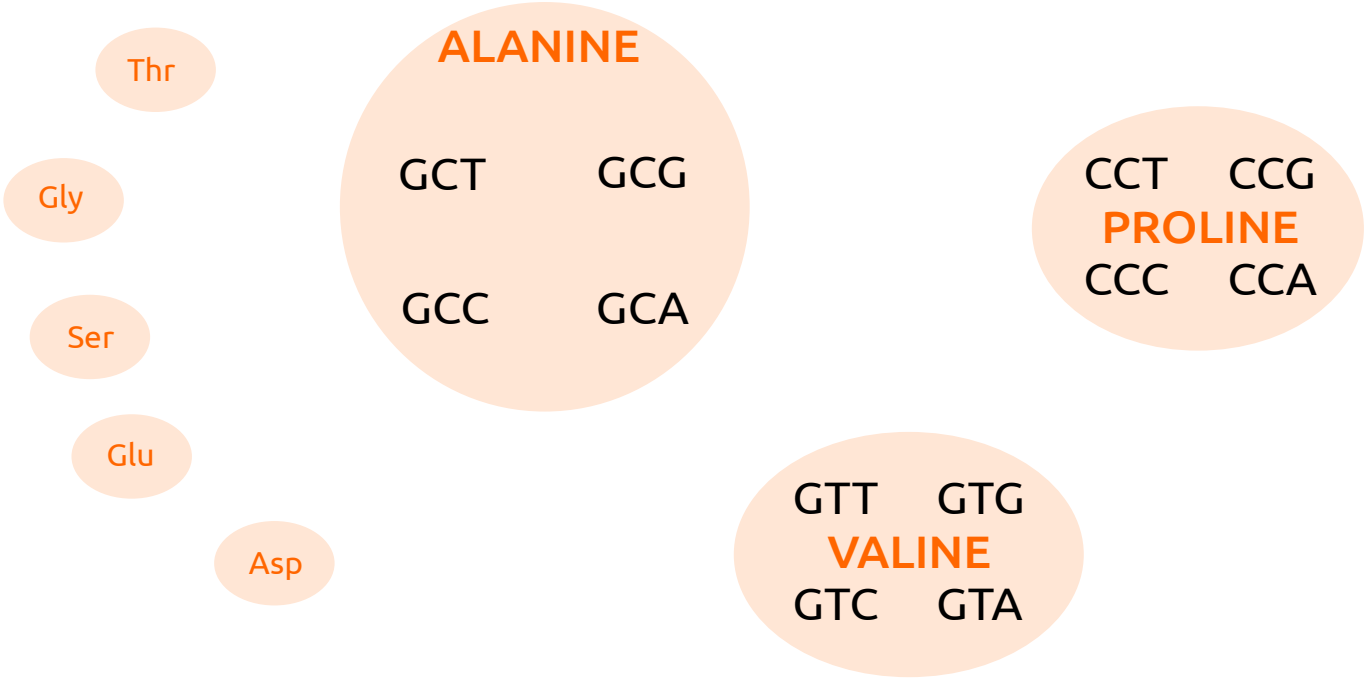


Objectives: Disentangling the two hypotheses

Modelling evolutive processes explaining the observed CUB: at the nucleotidic (N), the codons (C) and the AA layers (A).

SENCA: Sites Evolution of Nucleotides, Codons and Amino-acids

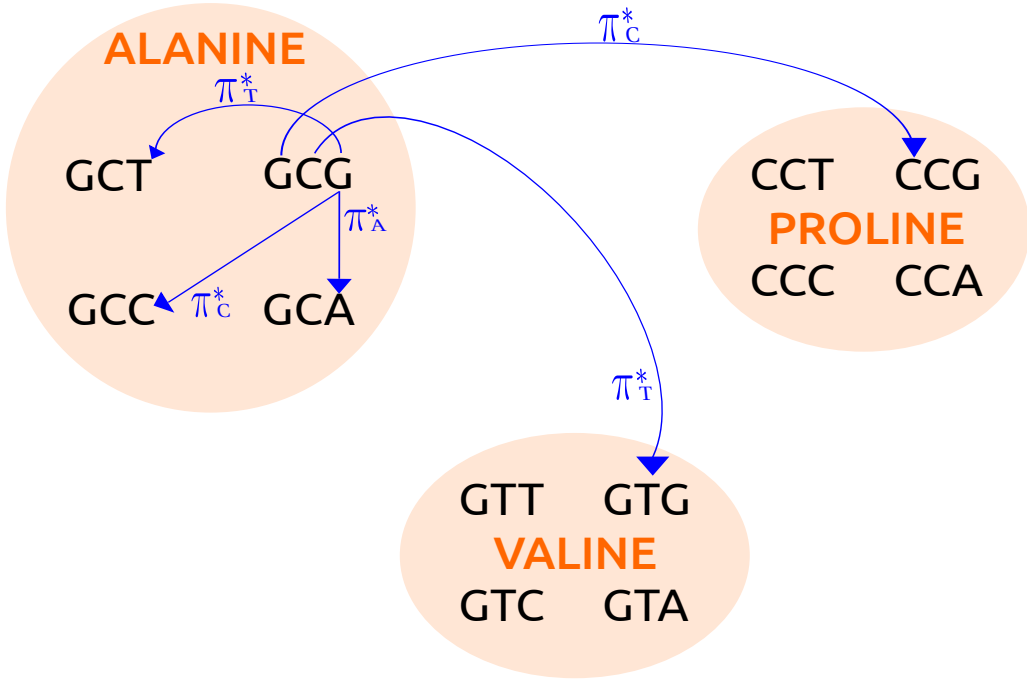
Let's have an example



Inspired by Yang and Nielsen 2008 (FMutSel).

SENCA: Sites Evolution of Nucleotides, Codons and Amino-acids

N layer

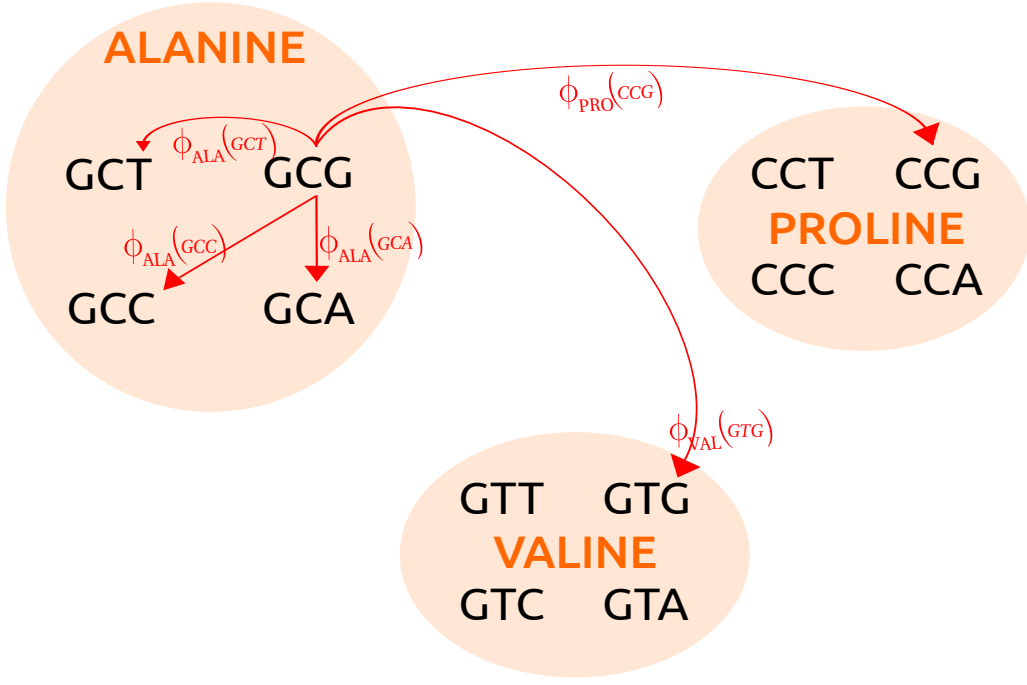


κ : transition/transversion

π^*_n :equilibrium frequency of nucleotide $n \in [A, C, G, T]$.

SENCA: Sites Evolution of Nucleotides, Codons and Amino-acids

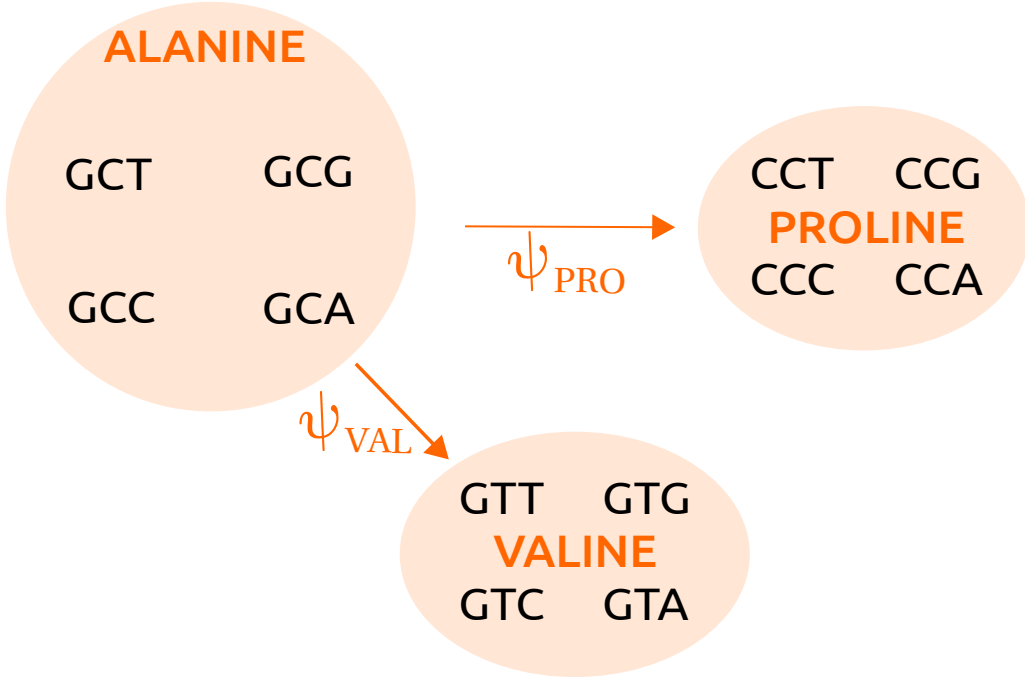
C layer



$\phi_{AA}(i)$: codon i preference – intra-AA.
$\sum_{I \in codons(AA)} \phi_{AA}(I) = 1$

SENCA: Sites Evolution of Nucleotides, Codons and Amino-acids

A layer

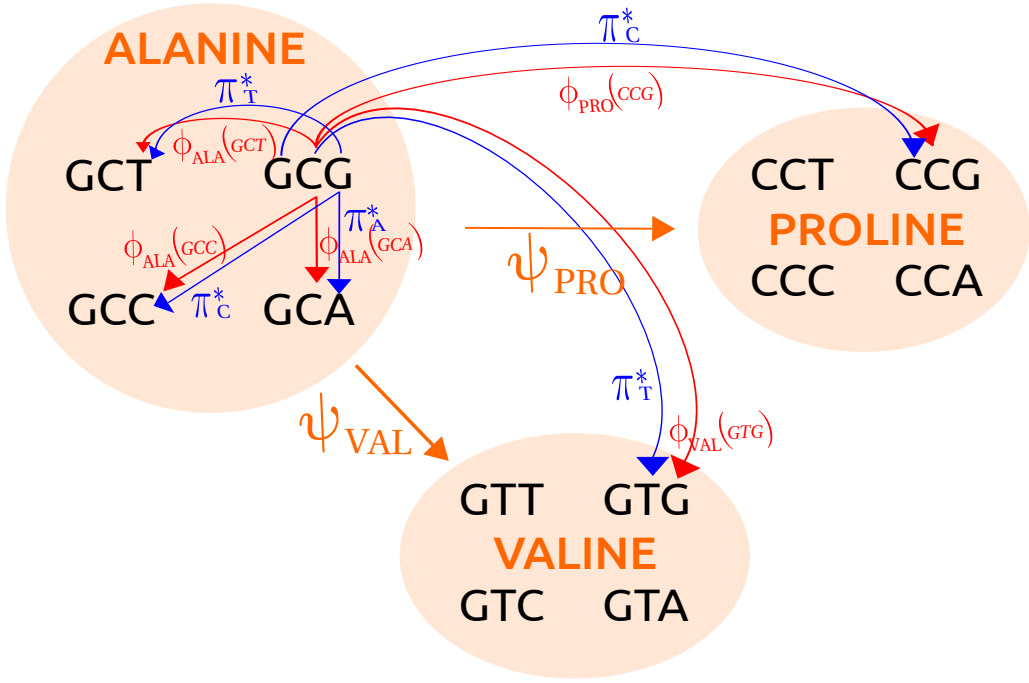


ω : non-synonymous/synonymous substitution rates

ψ_{AA}: preference AA. $\sum_{AA\in \text{ amino acids}} \psi(AA) = 1$

SENCA: Sites Evolution of Nucleotides, Codons and Amino-acids

Inspiration and structure



Implemented in Bio++

`http://biopp.univ-montp2.fr/.`

SENCA

Formulas

Instantaneous evolutionary rate from codon i to j :

$$q_{ij} = \begin{cases} 0 & \text{if 2 or 3 different positions,} \\ \pi^*_{j_k}\kappa \quad f(x_i, x_j) & \text{synonymous transition,} \\ \pi^*_{j_k}\kappa\omega f(x_i, x_j) & \text{non-synonymous transition,} \\ \pi^*_{j_k} \quad f(x_i, x_j) & \text{synonymous transversion,} \\ \pi^*_{j_k} \quad \omega f(x_i, x_j) & \text{non-synonymous transversion.} \end{cases}$$

with:

$$f(x_i, x_j) = \frac{-log(\frac{x_i}{x_j})}{1 - \frac{x_i}{x_j}}$$

and:

$$x_i = n_{aa_i}\phi_{aa}(i)\psi_{aa_i}$$

Data

21 pathogeneous bacteria and archea

From Lassalle et al., *PLoS Genet. 2015* $\Rightarrow$ concatenates of approx. 100 genes by increasing ENC (core genome, non-recombinant).

Dataset	Taxon Name	Nb. of strains	Nb of concatenates	Mean GC %
brucella	*Brucella* spp.	9	4	58.8
francis	*Francisella tularensis*	8	3	33.8
mycobacterium	*Mycobacterium tuberculosis* complex	7	1	66.1
burk_mal	*Burkholderia Pseudomallei* group	9	6	68.7
yersinia	*Yersinia pestis*	11	6	49.3
clamydia	*Clamydia trachomatis*	13	4	41.8
sulfo	*Sulfolobus* spp.	8	4	35.4
burk_ceno	*Burkholderia cenocepacia* complex	8	8	68.2
staph	*Staphylococcus aureus*	15	5	34.2
bifido	*Bifidobacterium longum*	6	3	61.9
acineto	*Acinetobacter* spp.	6	5	40.8
campylo	*Campylobacter jejunii*	6	4	31.6
clostridium	*Clostridium botulinum*	8	5	29.6
strep_pyo	*Streptococcus pyogenes*	12	3	39.6
strep_pneu	*Streptococcus pneumoniae*	13	3	42.0
salmonella	*Salmonella enterica*	14	6	54.6
listeria	*Listeria* spp.	8	3	38.8
B_anthracis	*Bacillus antharcis/aureus* group	17	3	37.0
nesseiria	*Nesseiria meningitidis*	8	2	55.3
escherichia	*Escherichia coli*	35	1	53.3
helicobacter	*Helicobacter pylori*	14	1	40.4

Implementation

- Non-stationnary and homogeneous run,
- Intra-species runs: amino-acids preferences stationnary,
- Optimisation by max. likelihood: 109 parameters.

Validation of the model:

Comparisons with AIC criterium to standard codons model YN98xF61.

SENCA is better in 68 over 78 concats. Exceptions are:

- 2 concats. of *Brucella* spp. and *Burkholderia peusomallei*,
- 1 of *Salmonella enterica* and *Yersinia pestis*,
- all of *Sulfolobus* spp.

Implementation

- Non-stationnary and homogeneous run,
- Intra-species runs: amino-acids preferences stationnary,
- Optimisation by max. likelihood: 109 parameters.

Validation of the model:

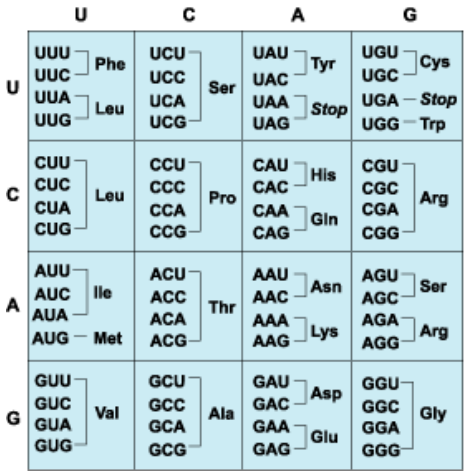
Comparisons with AIC criterium to standard codons model YN98xF61.

SENCA is better in 68 over 78 concats. Exceptions are:

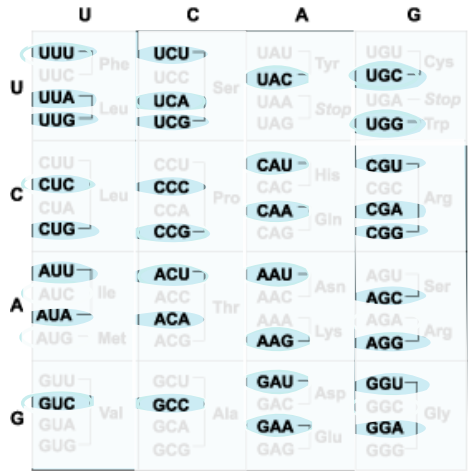
- 2 concats. of *Brucella* spp. and *Burkholderia peusomallei*,
- 1 of *Salmonella enterica* and *Yersinia pestis*,
- all of *Sulfolobus* spp.

What about the CUB?

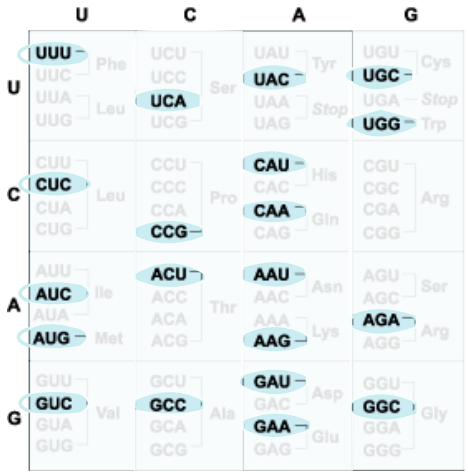
- To disentangle between the 2 hypotheses: compute the CUB by comparison with expected if uniform –> ENC index.
- Effective Number of Codons – ENC – varies between 61 and 20.



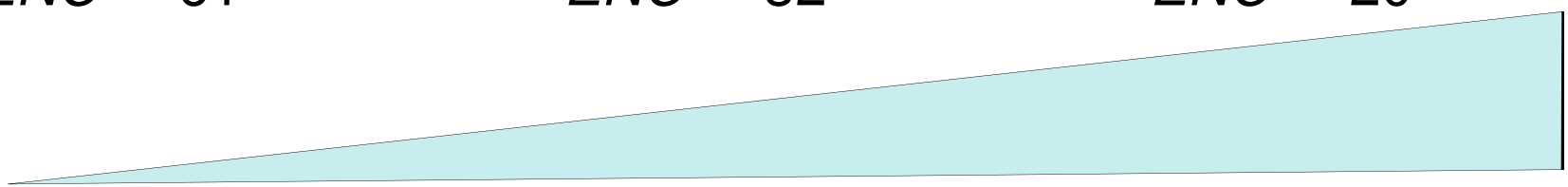
$ENC = 61$



$ENC = 32$

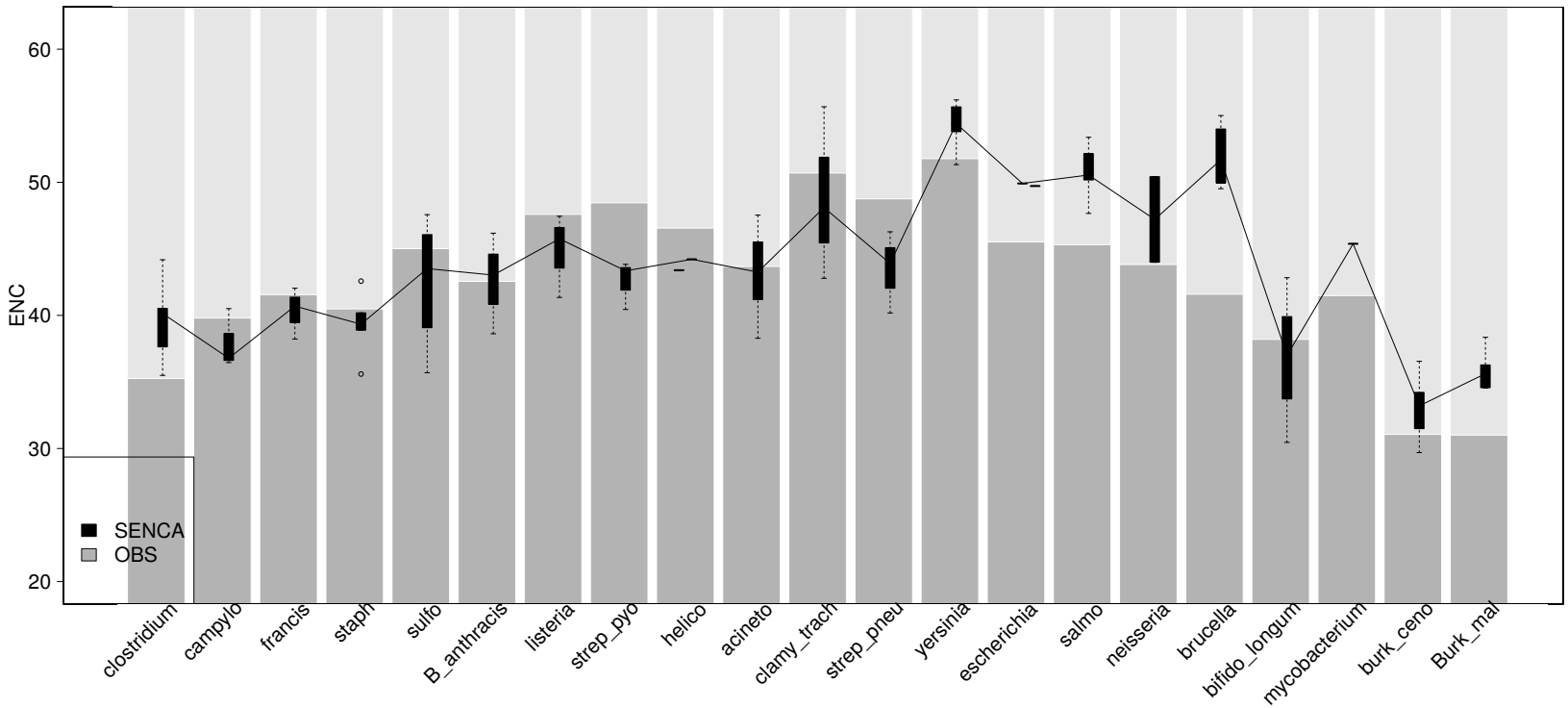


$ENC = 20$



Strength of CUB

ENC index



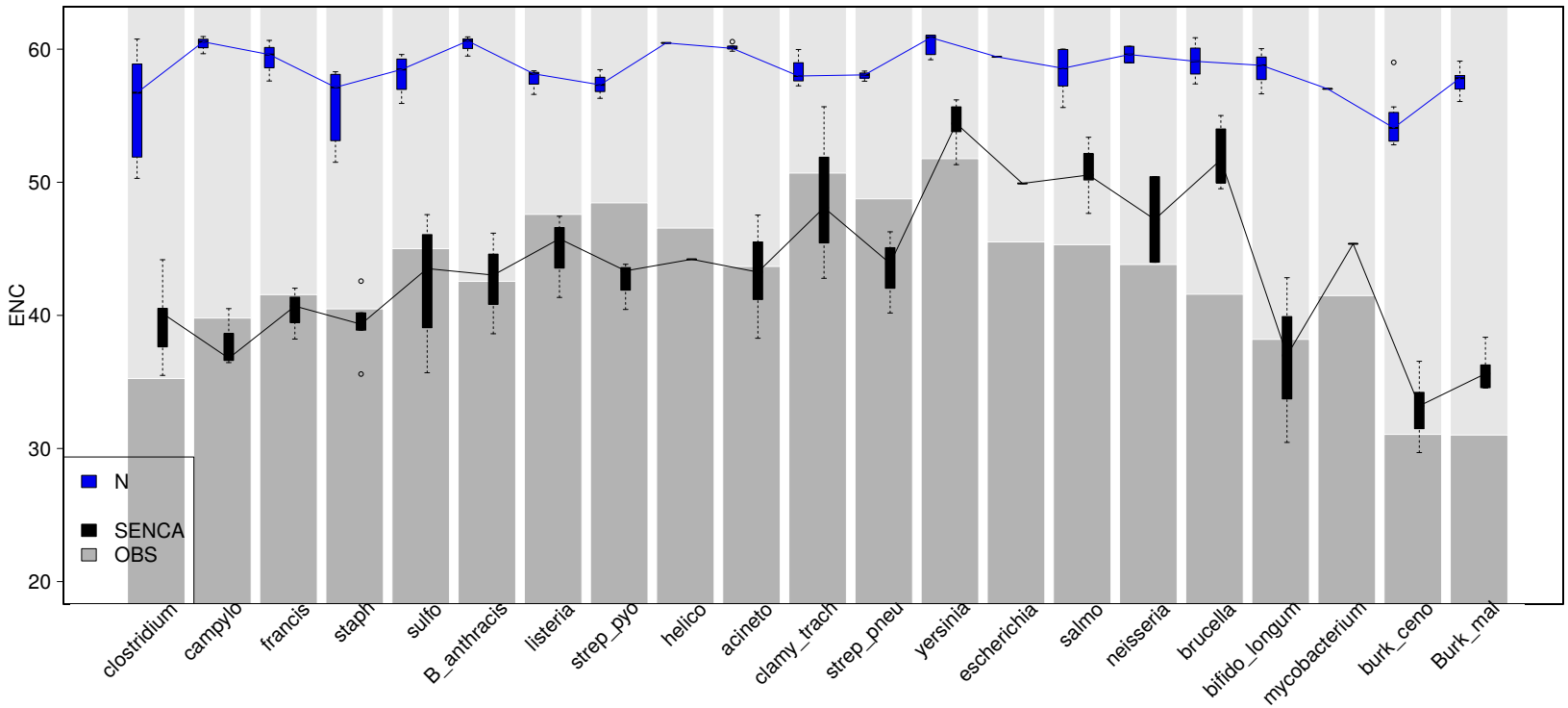
At equilibrium, frequency of codon *i* is proportional to:

$$f^*(i) \propto \prod_{k=1}^{3} \pi^*_{i_k} \times \phi_{aa_i}(i) n_{aa_i} \times \psi_{aa_i}$$

$\rightarrow ENC^{\star}_{SENCA} \approx ENC_{OBS}$

Species ordered by incr. GC_{obs}

ENC index



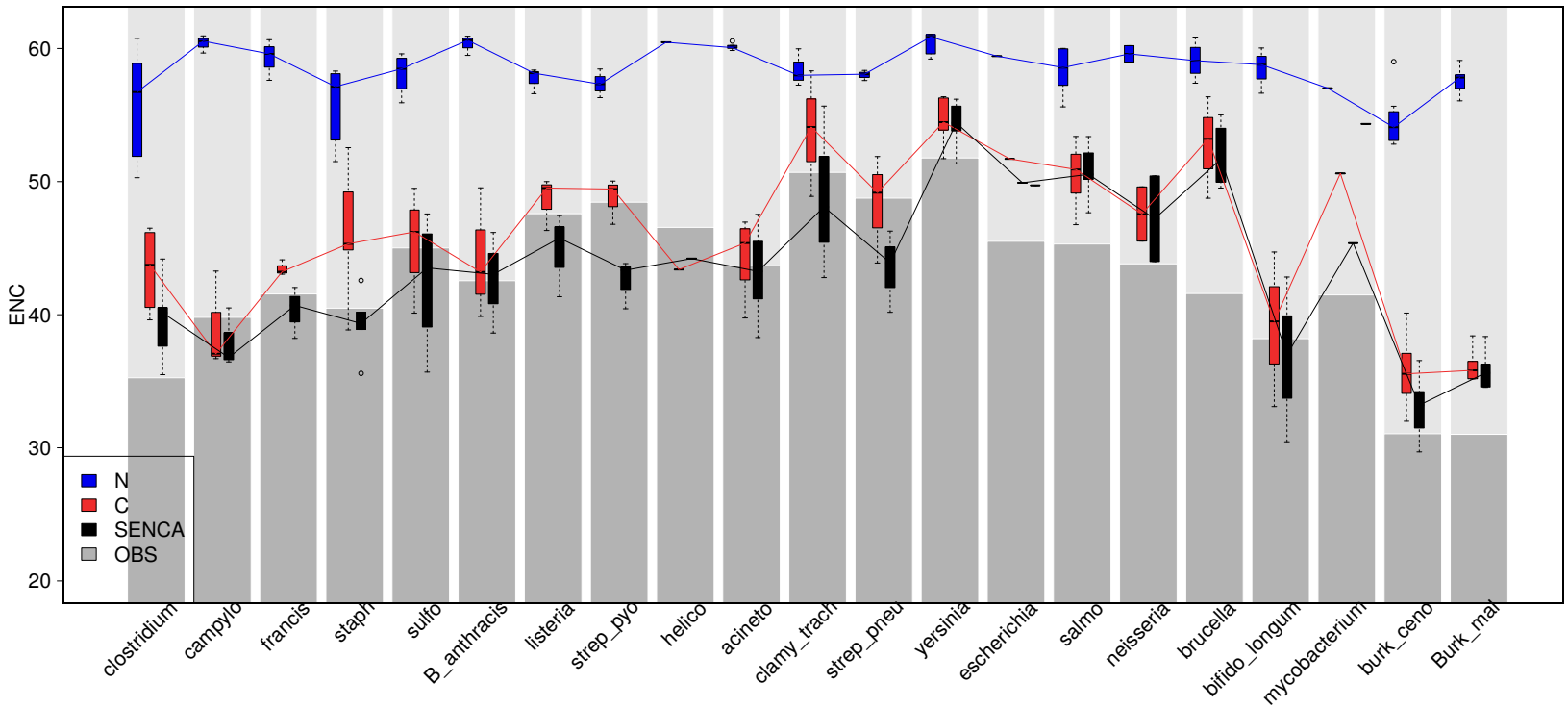
Compute $ENC^{\star}_{N}$:

$$f^{*}(i) \propto \prod_{k=1}^{3} \pi^{*}_{i_k} \times 1 \times 1$$

–> $ENC^{\star}_{N} \approx$ uniform CUB,

Species ordered by incr. GC_{obs}

ENC index



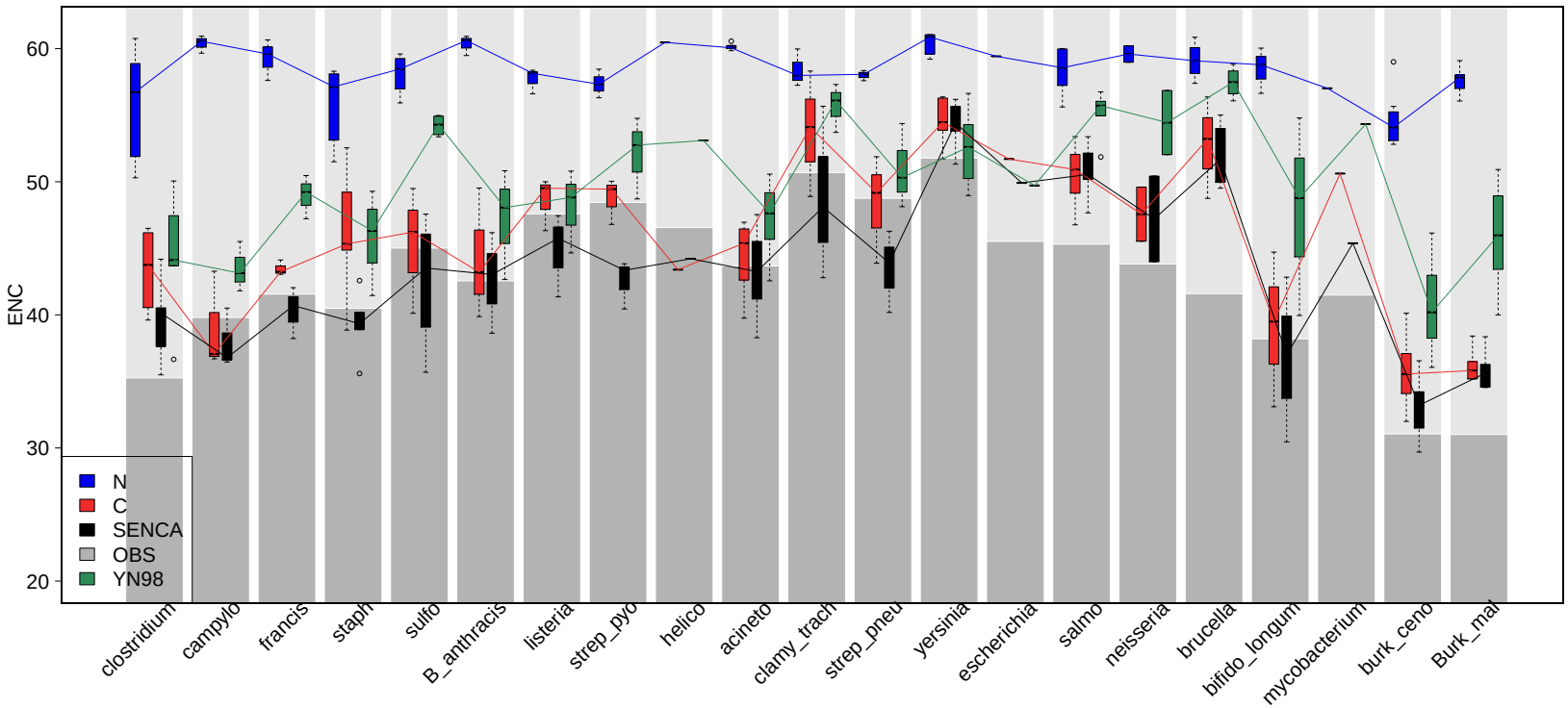
Compute $ENC^{\star}_{C}$:

$$f^{*}(i) \propto 1 \times \phi_{aa_i}(i) n_{aa_i} \times 1$$

$\rightarrow$ $ENC^{\star}_{C} \approx ENC^{\star}_{SENCA}$

Species ordered by incr. GC_{obs}

ENC index



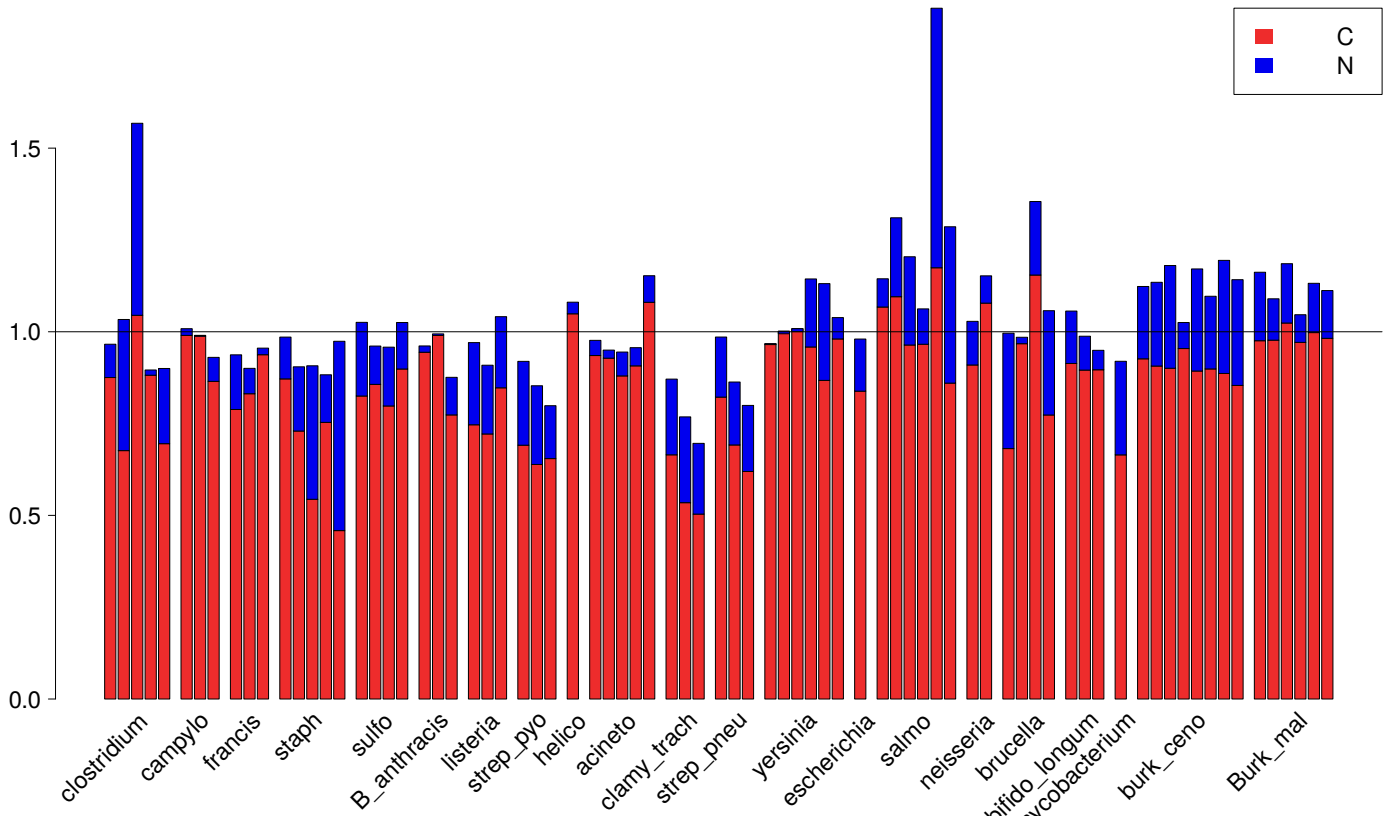
- $ENC^{\star}_{SENCA} \approx ENC_{OBS} < ENC^{\star}_{YN98}$,
- Is $ENC^{\star}_{SENCA}$ mostly driven by $ENC^{\star}_{C}$? Quantifying.

Species ordered by incr. GC_{obs}

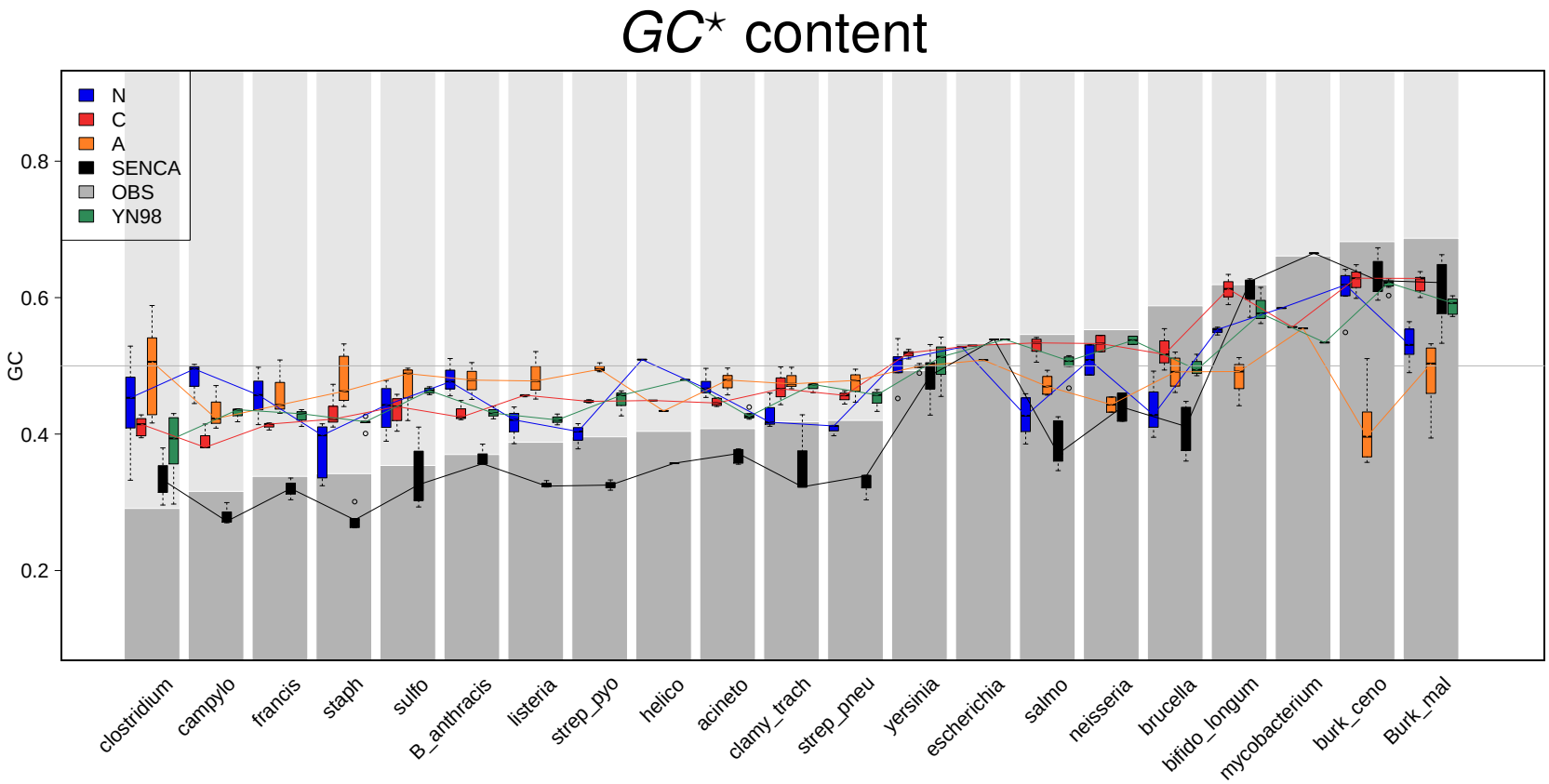
Quantification

$$dENC^{\star}_{SENCA} \approx dENC^{\star}_{C} + dENC^{\star}_{N}$$

- CUB measured as distance to uniform usage: $dENC^{\star} = 61 - ENC^{\star}$,
- High correlation ($R = 0.95$, p-value<10e-16, Pearson correlation test),
- Relative importance of C and N: $\frac{dENC^{\star}_{C}}{dENC^{\star}_{SENCA}} > \frac{dENC^{\star}_{N}}{dENC^{\star}_{SENCA}}$.

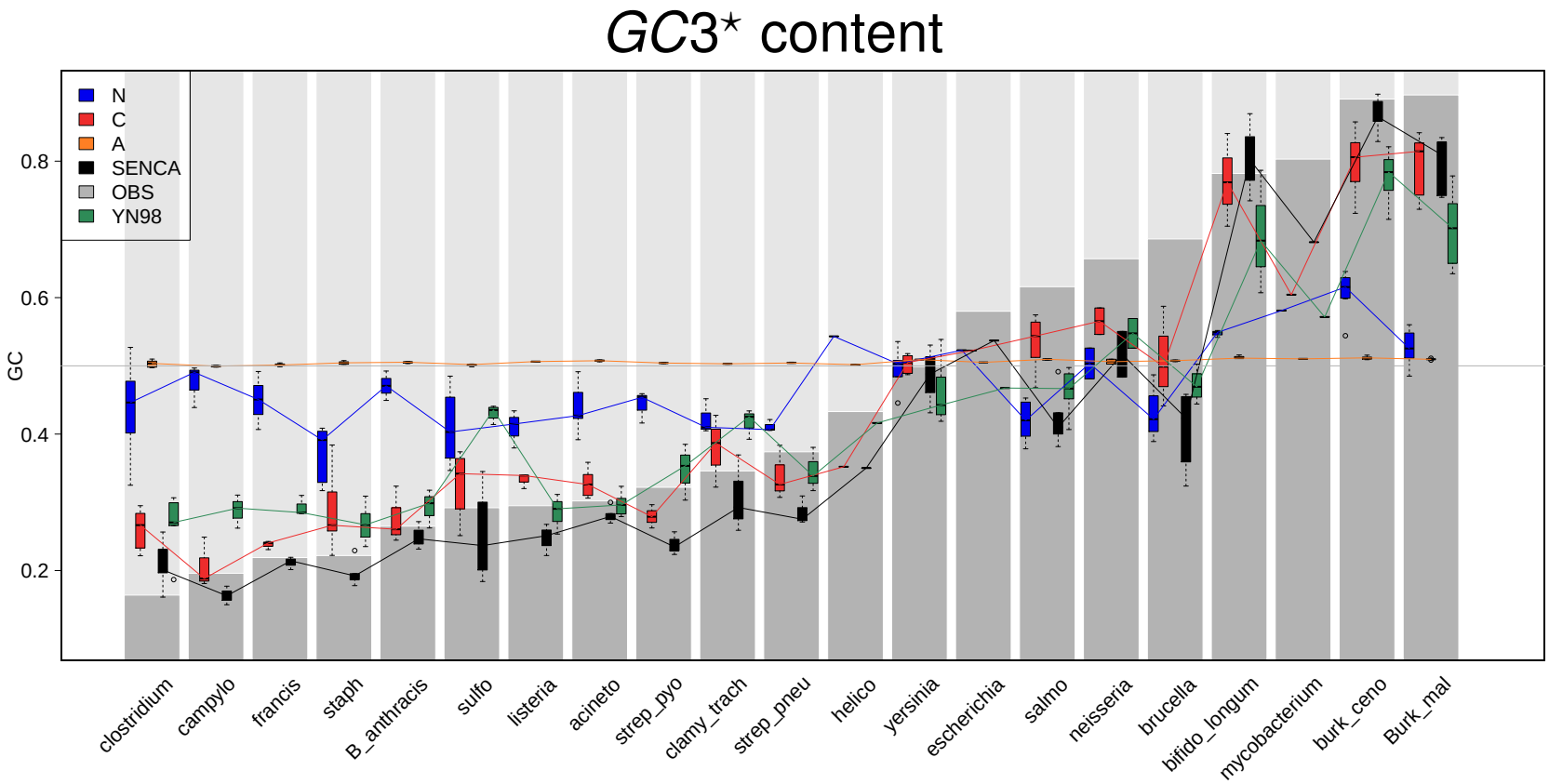


What about genome composition?



- $GC^{\star}_{SENCA} < GC_{obs}$
- SENCA combined effects of the 3 layers –> AT enrichment at equilibrium.

What about genome composition?

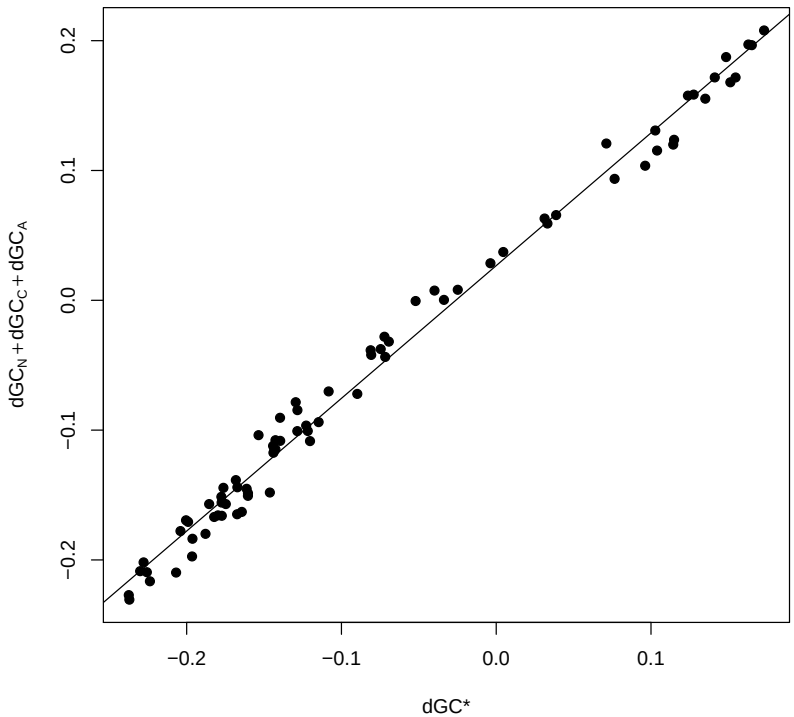


- $GC3^{\star}_{SENCA}$ more biased than $GC3^{\star}_{YN98}$: in agreement with $GC3_{obs}$,

$\Rightarrow$ Which layer has the most important impact on *GC* bias?

Quantification

$$dGC^\star \approx dGC^\star_N + dGC^\star_C + dGC^\star_A$$

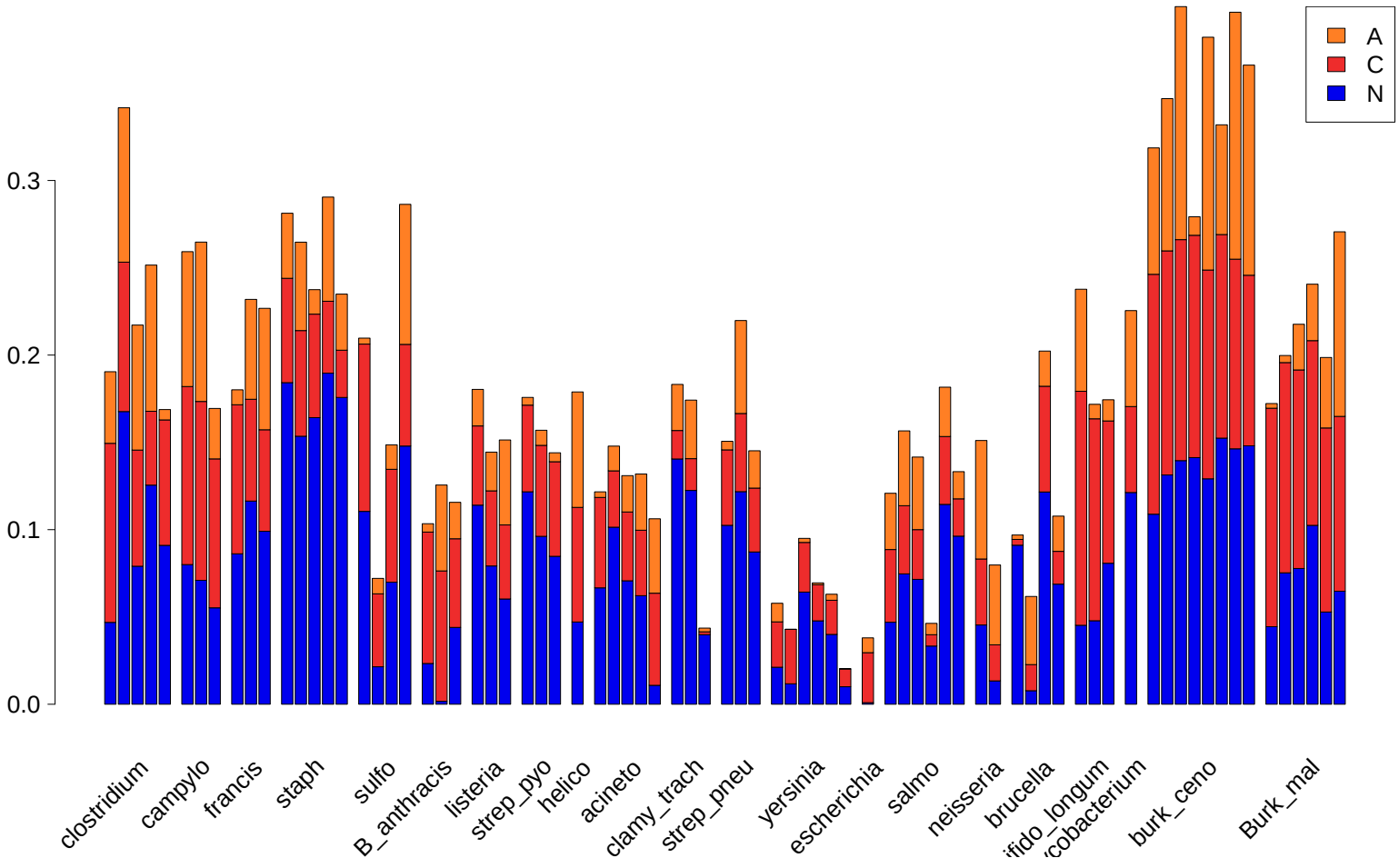


$GC^\star$ bias measured as $dGC^\star = 0.5 - GC^\star$ (distance to uniform composition).
High correlation (R=0.99, p-value<10e-16, Pearson correlation test).

Slope=1.04, intersect fixed to 0.

Quantification

$$dGC^\star \approx dGC^\star_N + dGC^\star_C + dGC^\star_A$$



N, *C* and *A* impact on $GC^\star$.

Conclusion

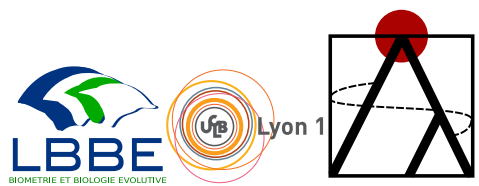
- Methodologically:
  - Non-stationary model
  - Multilayered model
  - New statistical tools: GC and ENC of layers

- Biologically:
  - Bias towards AT at equilibrium
  - Importance of N, C and A for genomic bias.
  - CUB mostly due to C.

Perspectives

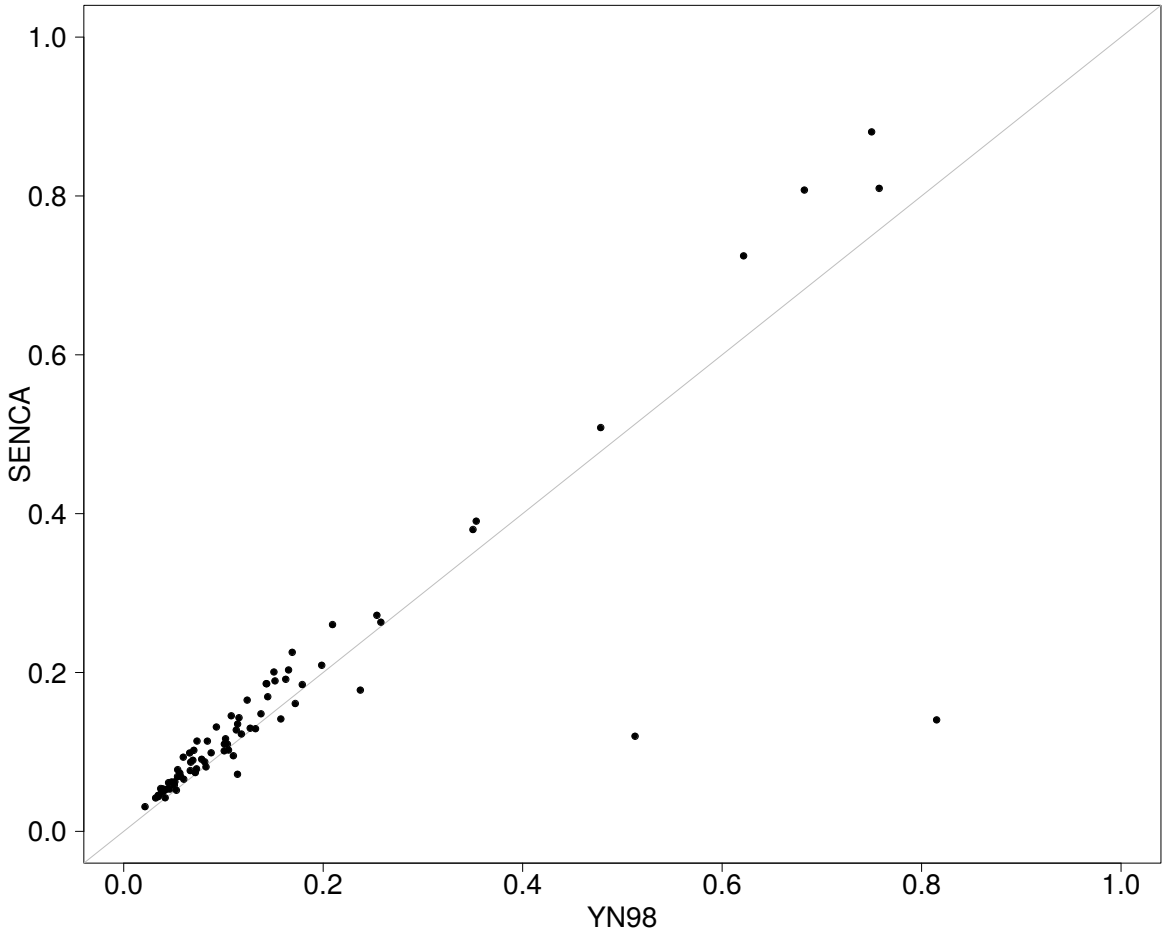
- Extension of the model:
  - Gene expression factor : measure of CUB intensity within a genome
- Biological questions:
  - Influence on ω estimation,
  - HIV – Human interaction and consequences on HIV CUB,
  - Can we detect recombination patterns: BGC?

Thank you for your attention!



- Acknowledgments:
  Laurent Guéguen, Marc Bailly-Bechet, Dominique Mouchiroud, Florent Lassalle, Vincent Daubin

Omega comparisons



SENCA: Site Evolution of Nucleotides, Codons and Amino-acids

Implementation and parameters

SENCA with 65 parameters:

$\pi^*_{j_k}$ equil. frequency of mutational process of j_k
κ transition/transversion
$\phi_{aa}(i)$ codon preference. For each AA, we have: $\sum_i \phi_{aa}(i) = 1$
n_{aa_i} degenerescence of the AA
$\omega = \frac{dN}{dS}$ mutation rate of non-synynymous dN and synonymous dS
ψ_{aa} AA preference. We have: $\sum_{aa} \psi_{aa} = 1$

In Bio++ (`http://biopp.univ-montp2.fr/`). Optimisation by max. likelihood. Homogeneous, non-stationary analysis (ψ_{aa} stat. because intra-species analyses)