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

61 sense codons, 20 amino acids

Lysine codon usage in bacteria

E.coli Lysine Usage
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

\[ \kappa : \text{transition/transversion} \]
\[ \pi_n^* : \text{equilibrium frequency of nucleotide } n \in [A, C, G, T]. \]
SENCA: Sites Evolution of Nucleotides, Codons and Amino-acids

C layer

\[ \phi_{AA}(i) : \text{codon } i \text{ preference – intra-}AA. \]
\[ \sum_{i \in \text{codons}(AA)} \phi_{AA}(i) = 1 \]
SENCA: Sites Evolution of Nucleotides, Codons and Amino-acids

A layer

\[ \omega : \text{non-synonymous/synonymous substitution rates} \]
\[ \psi_{AA}: \text{preference } AA. \sum_{AA \in \text{amino acids}} \psi(AA) = 1 \]
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 $j$:

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

with:

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

and:

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

21 pathogenaceous bacteria and archaea

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

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

\[ \text{ENC} = 61 \quad \text{ENC} = 32 \quad \text{ENC} = 20 \]
At equilibrium, frequency of codon $i$ is proportional to:

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

$$\rightarrow ENC_{SENCA}^* \approx ENC_{OBS}$$

Species ordered by incr. $GC_{obs}$
**ENC index**

Compute $\text{ENC}_N^*$:

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

$\rightarrow$ $\text{ENC}_N^* \approx \text{uniform CUB}$,

Species ordered by incr. $GC_{obs}$
ENC index

![Graph showing ENC index for various species]

Compute $ENC^*_C$:

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

$\rightarrow$ $ENC^*_C \approx ENC^*_SENCA$

Species ordered by incr. $GC_{obs}$
ENC index

- $\text{ENC}_{\text{SENCA}} \approx \text{ENC}_{\text{OBS}} < \text{ENC}_{\text{YN98}}$
- Is $\text{ENC}_{\text{SENCA}}^*$ mostly driven by $\text{ENC}_C^*$? Quantifying.

Species ordered by incr. $G_{\text{obs}}$
Quantification

\[ d\text{ENC}_{\text{SENCA}}^* \approx d\text{ENC}_C^* + d\text{ENC}_N^* \]

- CUB measured as distance to uniform usage: \( d\text{ENC}^* = 61 - \text{ENC}^* \),
- High correlation (\( R = 0.95 \), p-value<10e-16, Pearson correlation test),
- Relative importance of \( C \) and \( N \): \( \frac{d\text{ENC}_C^*}{d\text{ENC}_{\text{SENCA}}^*} > \frac{d\text{ENC}_N^*}{d\text{ENC}_{\text{SENCA}}^*} \).
What about genome composition?

**GC**\(^*\) content

- \(GC_{SENCA} < GC_{obs}\)
- SENCA combined effects of the 3 layers \(\rightarrow\) AT enrichment at equilibrium.
What about genome composition?

**GC3** content

- \( GC3^*_{SENCA} \) more biased than \( GC3^*_{YN98} \): in agreement with \( GC3_{obs} \),

\( \Rightarrow \) Which layer has the most important impact on GC bias?
Quantification

\[ dGC^* \approx dGC_N^* + dGC_C^* + dGC_A^* \]

\[ GC^* \text{ bias measured as } dGC^* = 0.5 - GC^* \text{ (distance to uniform composition).} \]

High correlation (R=0.99, p-value<10e-16, Pearson correlation test).

Slope=1.04, intersect fixed to 0.
Quantification

\[ dGC^* \approx dGC_N^* + dGC_C^* + dGC_A^* \]

\( N, C \) and \( A \) impact on \( GC^* \).
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 $\omega$ 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

![Scatter plot showing Omega comparisons between YN98 and SENCA. The plot displays a diagonal line indicating perfect correspondence, with some points deviating from the line.]
SENCAN Site Evolution of Nucleotides, Codons and Amino-acids
Implementation and parameters

SENCAN with 65 parameters:

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

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