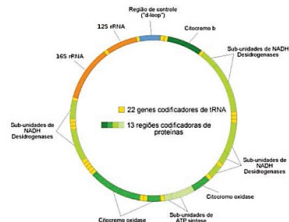


Análise de *codon usage bias* mitocondrial

Introdução

- Código genético - Degenerado
- Preferência por códons específicos
 - Viés de uso de codon (*codon usage bias*)
 - Comum no genoma mitocondrial
- Possíveis explicações
 - Maximizar a transcrição de seus genes (e.g. mais A-T)
 - Priorizar os códons do mitogenoma (22 tRNAs)
- Clado escolhido: Primates



Perguntas

Perguntas Biológicas

- 1 Quais aminoácidos apresentam *codon usage bias*?
- 2 Dentre os que apresentam:
 - a) Essa preferência está associada ao número de bases AT do códon?
 - b) Essa preferência está associada a quais codons o mitogenoma usa?

Perguntas experimentais

- 1 Há uma diferença entre o número de codons sinônimos usados para cada aa?
- 2
 - a) Há uma diferença entre o número de codons que possuem 1, 2 ou 3 Adeninas ou Timinas?
 - b) Há uma diferença entre o número de codons cujos tRNAs/anticodons são encontrados no mitogenoma e os que não são?

Obs: As perguntas são restritas à mitocôndrias de primatas.

Estudo observacional

```
##                               Species Translation Table Aminoacid Codon Anticodon
## 1 Allenopithecus_nigroviridis                2      Phe   TTT      <NA>
## 2 Allenopithecus_nigroviridis                2      Phe   TTC      GAA
##  codon_count /1000 Fraction
## 1          99 25.98      0.45
## 2         120 31.50      0.55
```

- Não usaremos todas as variáveis medidas
 - **Variáveis Dependentes:**
 - Codon count, Fraction
 - **Variáveis Independentes:**
 - Species, aminoacid, codon
- Unidade experimental: Espécie
 - Mitogenoma completo da espécie
 - Toda espécie consegue contribuir com um ponto por grupo - Critério de pareamento

Tamanho da amostra e poder estatístico

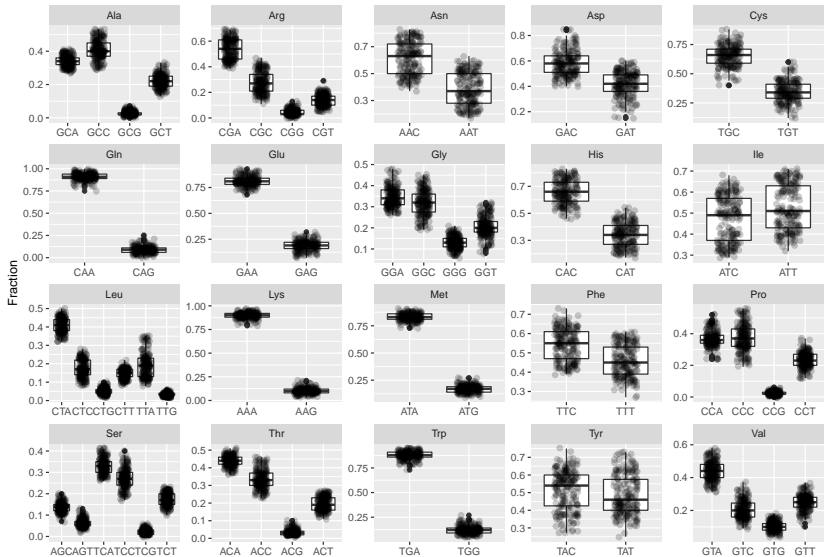
- Número de espécies

[1] 199

- Curva de poder estatístico

Resultados do experimento:

O que fazer aqui? Talvez plotar os gráficos?



Teste estatístico:

- Variáveis qualitativas (aa e codons)
- Lidando com porcentagens observadas e esperadas
- Qui-quadrado de aderência
 - Feito para cada aminoácido

Resultados dos testes:

```
## `summarise()` has grouped output by 'Aminoacid'. You can
## [1] "Valor de p para aminoacido Ala : 0"
## [1] "Valor de p para aminoacido Arg : 0"
## [1] "Valor de p para aminoacido Asn : 0"
## [1] "Valor de p para aminoacido Asp : 4.63309121213369e-
## [1] "Valor de p para aminoacido Cys : 5.19476270305044e-
## [1] "Valor de p para aminoacido End : 0"
## [1] "Valor de p para aminoacido Gln : 0"
## [1] "Valor de p para aminoacido Glu : 0"
## [1] "Valor de p para aminoacido Gly : 0"
## [1] "Valor de p para aminoacido His : 0"
## [1] "Valor de p para aminoacido Ile : 1.15392845961021e-
## [1] "Valor de p para aminoacido Leu : 0"
## [1] "Valor de p para aminoacido Lys : 0"
## [1] "Valor de p para aminoacido Met : 0"
## [1] "Valor de p para aminoacido Phe : 1.0411507177594e-6
## [1] "Valor de p para aminoacido Pro : 0"
```


O que os resultados significam mesmo?

Será que os resultados fazem sentido?



In My Opinion |  Full Access |

Wildlife biology, big data, and reproducible research

Keith P. Lewis , Eric Vander Wal, David A. Fifield

First published: 14 January 2018 | <https://doi.org/10.1002/wsb.847> | Citations: 12

ABSTRACT

Changes in technology have made it possible to gather vast amounts of data, often of high quality, that in turn can improve the quality of wildlife biology. However, with this growth in data, practices such as data management, exploratory data analysis, data-sharing, and reproducibility of an analysis have become increasingly complex. These practices often depend heavily on computer scripting languages, and are often hidden from the peer-review process despite their influence on the final results. Although these

Pitch - Biologia Computacional e Reprodutibilidade

- “We used a custom python/perl script to...”
- “An In-house script was used to...”
- “The 199 mitogenomes were downloaded from NCBI. The complete coding sequence was extracted and codon occurrences were counted for each species.”
- Daí você procura o código e ele não está em local algum...
- Você pode contactar o autor do paper e pedir pelo script/jupyter notebook/rmarkdown...
 - Processo demorado: Mais rápido vc mesmo escrever seu programa
 - E isso é péssimo em termos de reprodutibilidade...
- E mesmo que o código esteja disponível, isso ainda não garante que o trabalho seja replicável...

Pitch - Biologia Computacional e Reprodutibilidade

- E se focássemos nisso?
- ① Selecionar uma amostra de papers que manipulem dados usando linguagens de programação
- Palavras-chave: **“custom script/program”**, **“jupyter notebook/.ipynb”**, **“Rmarkdown/.Rmd”**, etc...
- ② Se por acaso o código não estiver disponível em lugar nenhum, contactar os autores e requisitar o código...
- ③ Tentar reproduzir as análises computacionais.

An empirical analysis of journal policy effectiveness for computational reproducibility

 Victoria Stodden, Jennifer Seiler, and Zhaokun Ma

PNAS March 13, 2018 115 (11) 2584-2589; first published March 12, 2018; <https://doi.org/10.1073/pnas.1708290115>

A key component of scientific communication is sufficient information for other researchers in the field to reproduce published findings. For computational and data-enabled research, this has often been interpreted to mean making available the raw data from which results were generated, the computer code that generated the findings, and any additional information needed such as workflows and input parameters. Many journals are revising author guidelines to include data and code availability. This work evaluates the effectiveness of journal policy that requires the data and code necessary for reproducibility be made available postpublication by the authors upon request. We assess the effectiveness of such a policy by (i) requesting data and code from authors and (ii) attempting replication of the published findings. We chose a random sample of 204 scientific papers published in the journal *Science* after the implementation of their policy in February 2011. We found that we were able to obtain artifacts from 44% of our sample and were able to reproduce the findings for 26%. We find this policy—author remission of data and code postpublication upon request—an improvement over no policy, but currently