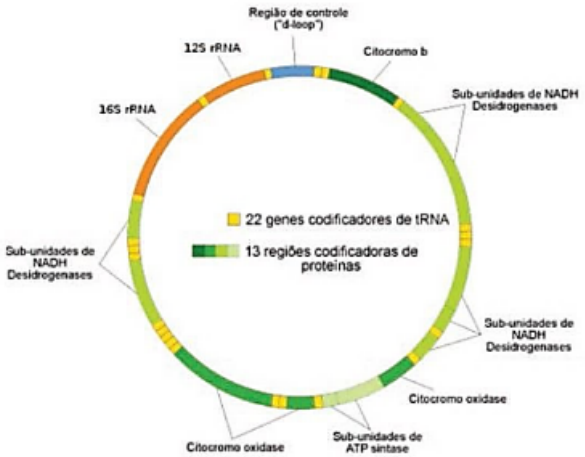
Análise de *codon usage bias* mitocondrial

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

Pergunta Biológica

Quais aminoácidos apresentam *codon usage bias* (se é que algum apresenta)?

Pergunta quantitativa

O quão provável é os valores observados na contagem de códons sinônimos mitocondriais para um dado aminoácido serem tão diferentes dos valores esperados caso não haja preferência por nenhum códon?

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
```

- Amostragem: 199 espécies
- Não usaremos todas as variáveis medidas
  - **Variável Dependente**:
    - Codon count
  - **Variável Independente**:
    - Codon
- Unidade experimental: Espécie
  - Mitogenoma completo da espécie
- Logo, minha pergunta e possíveis extrapolações estão **restritos à mitocôndrias de primatas**.

Teste estatístico:

- Teste para cada aminoácido
  - Uma variável qualitativa independente (codons) em cada teste
  - Essa variável qualitativa única pode ter duas ou mais categorias
  - A variável dependente é a contagem dos códons
  - Queremos saber quão provável é obter a contagem de códons observados aleatóriamente, dado que, se não houver bias, esperamos que todos os códons sejam encontrados em igual quantidade/proporção.
- Em outras palavras. . .
  - Queremos testar o quão provável é uma variável pertencer a uma determinada distribuição teórica.
  - **Qui-quadrado de aderência**

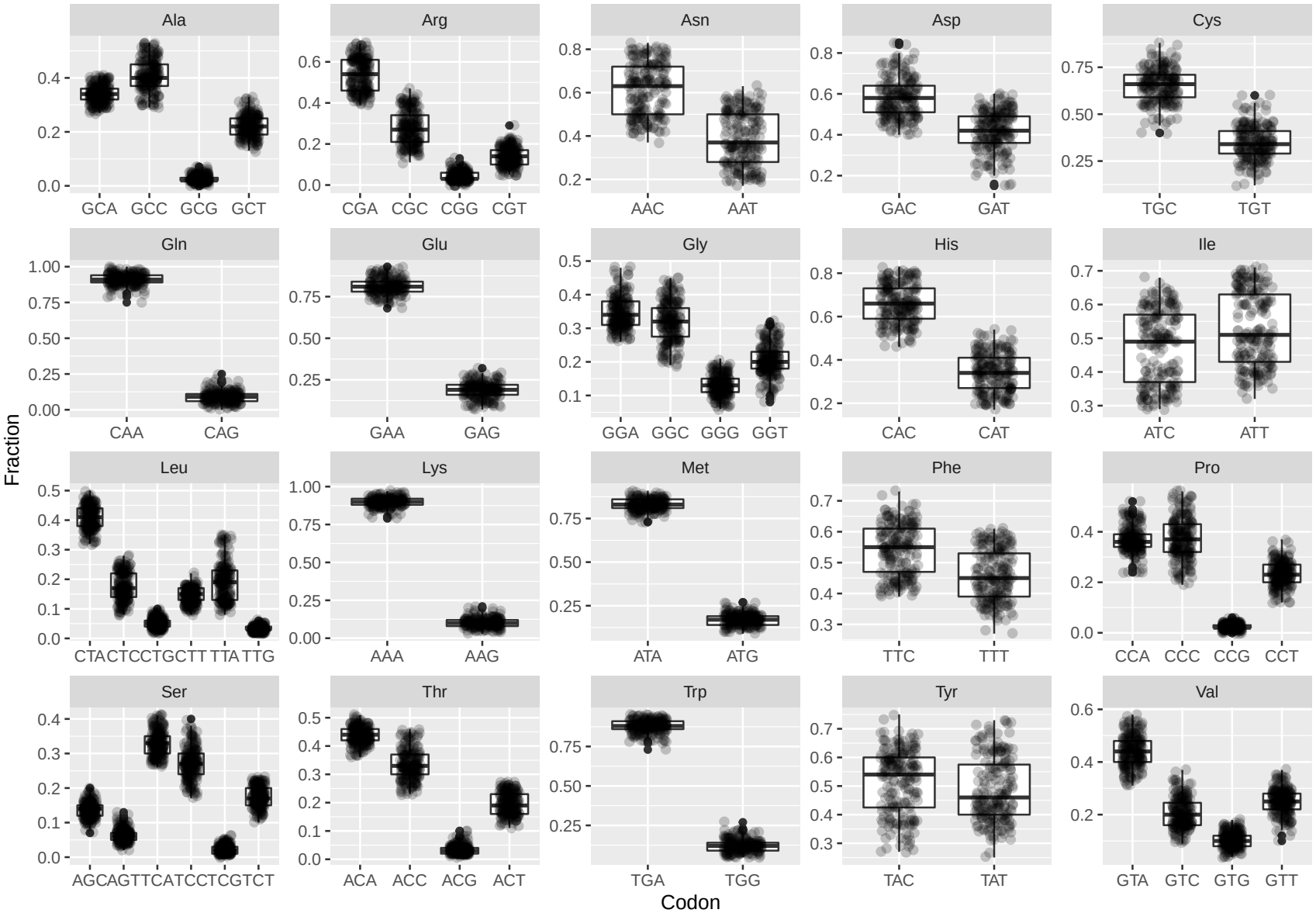
Resultados dos testes:

```
## [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-73"
## [1] "Valor de p para aminoacido Cys : 5.19476270305044e-103"
## [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-56"
## [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-60"
## [1] "Valor de p para aminoacido Pro : 0"
## [1] "Valor de p para aminoacido Ser : 0"
## [1] "Valor de p para aminoacido Thr : 0"
## [1] "Valor de p para aminoacido Trp : 0"
## [1] "Valor de p para aminoacido Tyr : 2.2724420099098e-05"
## [1] "Valor de p para aminoacido Val : 0"
```

O que os resultados significam mesmo?

- Hipótese nula:
  - A diferença entre os valores observados e esperados das contagens de codons é devida ao acaso.
- Dado que a hipótese nula é verdadeira, a probabilidade de obter um resultado onde a **diferença entre valores observados e esperados** é **igual ou maior** aos que eu encontrei é muito baixa (até demais).
- Logo, seria razoável rejeitar a hipótese nula e aceitar uma hipótese alternativa
  - No caso, a hipótese de que há um viés no uso de códons.

Será que os resultados fazem sentido?



Será que os resultados fazem sentido?

	Aminoacid	Codon	Observed	Expected	Residuals	Stdres	p-value
1	Ala	GCA	16459	12078.25	39.86	46.03	0
2	Ala	GCC	19850	12078.25	70.72	81.66	0
3	Ala	GCG	1309	12078.25	-97.99	-113.15	0
4	Ala	GCT	10695	12078.25	-12.59	-14.53	0
5	Arg	CGA	6878	3187	65.38	75.5	0
6	Arg	CGC	3552	3187	6.47	7.47	0
7	Arg	CGG	553	3187	-46.66	-53.88	0
8	Arg	CGT	1765	3187	-25.19	-29.09	0
9	Asn	AAC	20191	16433.5	29.31	41.45	0
10	Asn	AAT	12676	16433.5	-29.31	-41.45	0
11	Asp	GAC	7535	6504	12.78	18.08	4.63309121213369e-73
12	Asp	GAT	5473	6504	-12.78	-18.08	4.63309121213369e-73
13	Cys	TGC	3176	2425.5	15.24	21.55	5.19476270305044e-103
14	Cys	TGT	1675	2425.5	-15.24	-21.55	5.19476270305044e-103
15	End	AGA	147	643	-19.56	-22.59	0
16	End	AGG	114	643	-20.86	-24.09	0
17	End	TAA	1964	643	52.1	60.15	0
18	End	TAG	347	643	-11.67	-13.48	0
19	Gln	CAA	16832	9239	79	111.72	0
20	Gln	CAG	1646	9239	-79	-111.72	0
21	Glu	GAA	14658	9017.5	59.4	84	0
22	Glu	GAG	3377	9017.5	-59.4	-84	0
23	Gly	GGA	14461	10450	39.24	45.31	0
24	Gly	GGC	13303	10450	27.91	32.23	0
25	Gly	GGG	5438	10450	-49.03	-56.61	0

Outras perguntas

Pergunta Biológica

A preferência está associada ao número de bases AT presente no códon?

Pergunta experimental

Os valores observados na contagem de códons mitocondriais com 0, 1, 2 ou 3 Adeninas/Timinas são diferentes dos valores esperados caso não haja preferência por nenhuma dessas categorias?

- Abordagem:
  - Unidade experimental: Espécie
    - Os codons pertencentes à mesma categoria dentro de uma espécie serão usados como réplicas
  - Qui-quadrado de aderência
    - 4 categorias: Onde está a diferença?
    - Comparar o quarto grupo (3 A/T) com todos os outros com correção do valor de p por Bonferroni para 3 comparações.

Pitch - Biologia Computacional e Reprodutibilidade

WILDLIFE SOCIETY BULLETIN

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 nesses artefatos de pesquisa?

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.

Não é uma idéia totalmente original. . .

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