Instituto de Ciências Matemáticas e de Computação
Ciência da Computação
SSC0800 e SSC0801

Análise da Distribuição de trincas por classificação taxonômica

Alunos: Pedro Fernandez Tonso,
(nomes protegidos)
Professor: (nome protegido)

Dezembro
2023

Conteúdo

1 Introdução

O código genético é a linguagem fundamental da vida, codificando e traduzindo informações genéticas em proteínas nos organismos. Essa codificação segue regras precisas, estabelecendo a correspondência entre sequências de nucleotídeos no código genético e os aminoácidos nas proteínas. A compreensão profunda do código genético é essencial para desvendar os processos biológicos fundamentais.

O estudo do código genético possui ampla abrangência e relevância, fornecendo informações valiosas sobre hereditariedade, evolução e o funcionamento de organismos vivos. Ao analisar trincas em códigos genéticos de diferentes grupos taxonômicos, surgem padrões, variações e similaridades entre espécies. Essa análise comparativa esclarece a relação entre a estrutura genética e a diversidade biológica, aprimorando a compreensão da evolução e adaptação das espécies ao longo do tempo.

Este projeto investiga a distribuição de trincas em bases genéticas de várias classificações taxonômicas. Buscamos identificar padrões nas proporções das trincas em diferentes segmentos da vida, utilizando métodos estatísticos para identificar assinaturas biológicas.

2 Desenvolvimento

O código genético é composto por moléculas, as quais os blocos fundamentais são os nucleotídeos. Cada nucleotídeo é formado por um par de base nitrogenada. Essas bases nitrogenadas são: adenina (A), timina (T), citosina (C) e guanina (G). As trincas, ou códons, referem-se a conjuntos de três dessas bases nitrogenadas, ordenadas, ao longo da cadeia genética, na qual cada trinca codifica um aminoácido específico.

Os aminoácidos são os blocos de construção das proteínas. Existem 20 tipos diferentes de aminoácidos que podem ser combinados em diversas sequências para formar uma grande variedade de proteínas. Durante o processo de síntese proteica, a informação genética é transcrita. Nesse processo os códons são lidos e "traduzidos"em aminoácidos. Esses aminoácidos são então ligados em uma sequência específica formando a proteína, conforme ditada pela sequência de códons. Visto que cada códon tem uma correspondência em aminoácidos, essa correspondência está descrita segundo a Figura 1.

Essa relação entre o código genético, as trincas de bases nitrogenadas e os aminoácidos é crucial para a expressão gênica e para a produção das proteínas essenciais ao funcionamento e estrutura celular. Pequenas variações ou erros nesse processo podem resultar em diferenças significativas na composição proteica e, consequentemente, nas funções celulares e no organismo como um todo.

Para analisar a distribuição de trincas em diferentes grupos de organismos, utilizamos um conjunto de dados contendo a frequência das trincas em diferentes espécies, oferecendo uma perspectiva do código genético exclusivo de cada espécie, acesso em: Codon Usage.

Destarte, esse analise visa investigar a relação entre classificações taxonômicas e a distribuição de trincas em códigos genéticos. Averiguando se há distinção entre as proporções genômicas e distintas taxonomias.

Com vistas a atingir nossos objetivos dividimos o trabalho em quatro etapas: limpeza e tratamento de dados, visualização e exploração, formulação de hipóteses e testes estatísticos.

2.1 Limpeza e tratamento

2.1.1 Composição da base

A base de dados consiste em 13.028 linhas e 68 colunas. 64 delas representam a frequência dos códons em cada espécie. As proporções das trincas, como: 'UUU', 'UUA', 'UUG', 'CUU', etc; são registradas como floats, apresentando decimais em 5 dígitos.

'Tipo de DNA' representa categoricamente a composição genômica: 0 - genômico, 1 - mitocondrial, 2 - cloroplasto, 3 - cianelo, 4 - plasto, 5 - nucleomorfo, 6 - endossimbionte secundário, 7 - cromoplasto, 8 - leucoplasto, 9 - NA, 10 - proplasto, 11 - apicoplasto, 12 - cinetoplasto.

'Ncodons' é a soma dos números associados aos diferentes códons em uma entrada do CUTG. As frequências dos códons são normalizadas pela contagem total de códons; assim, as frequências são obtidas dividindo o número de ocorrências por

Coluna	Nome	Tipo
Coluna 1	Reino	Inteiro
Coluna 2	Tipo de DNA	Inteiro
Coluna 3	ID da Espécie	Inteiro
Coluna 4	Ncodons	Inteiro
Coluna 5	Nome da Espécie	String
Coluna 6-69	Códon	Float

Tabela 1: Estruturação do dataset

'Ncodons' no arquivo de dados. O nome da espécie, representado por strings sem vírgulas, serve como rótulo descritivo para a interpretação dos dados.

2.1.2 Numérico e categórico

O dataset, inicialmente composto por 65 colunas numéricas, foi automaticamente classificado pelo pandas no que diz respeito ao tipo. Contudo, identificamos que duas colunas de códons ("UUU", "UUC") estavam incorretamente categorizadas como categóricas. Tentando converter os dados para flutuantes, observamos que apenas alguns valores eram necessários para converter ambas as colunas em formatos totalmente compatíveis com float. Optamos por impor esses valores utilizando o método "Entrada por amostra aleatória".

O método consiste em adicionar dados faltantes a uma coluna por amostragem aleatória dos dados existentes na coluna. Para adotar essa técnica, assume-se que a coluna é não MNAR, o que significa que o motivo para os dados estarem faltando é independente do valor que os dados assumiria naquela posição.

Por serem poucos, os dados substituidos, as hipóteses futuras difícilmente seriam afetada. Desenvolvemos, portanto, uma função que identifica e armazena os valores de uma série convertíveis em float. Em seguida, substituímos os valores ausentes na série original por amostras aleatórias desses valores float. Esse método preserva a distribuição dos dados, ao contrário de abordagens como a Imputação Média, que pode enviesar a distribuição em direção ao centro.

2.1.3 Conversão para aminoácido

Os dados referentes à frequência dos códons, originalmente organizados em 64 colunas, apresentavam uma complexidade excessiva para análise devido à alta dimensionalidade e à impossibilidade de analisá-los separadamente. Com o objetivo de simplificar as visualizações, reduzimos a dimensionalidade agrupando os dados por aminoácidos, resultando em uma redução de 64 para 21 colunas.

Devido à sua correspondência com os aminoácidos na síntese proteica, o agrupamento das trincas em seus respectivos aminoácidos foi, bioquimicamente, conciso e representou uma redução significativa na dimensionalidade para nossas análises.

Essa conversão foi realizada pela soma das frequências de cada códon que referencia o mesmo aminoácido ou indica o término na leitura do gene, como vemos na Figura 1. A exemplo, no caso da Fenilalanina (Phe), obtemos sua frequencia por meio da soma das frequências das trincas UUU e UUC.

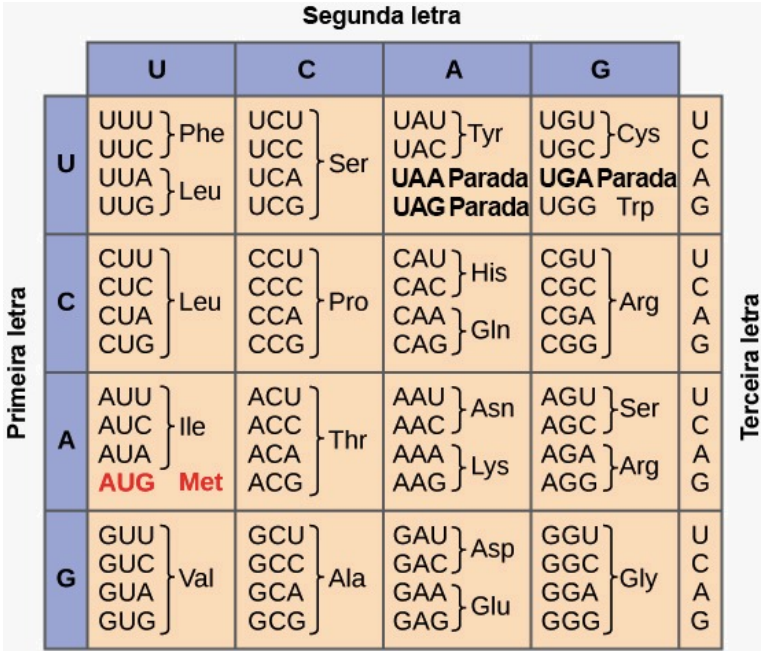

Figura 1: O conjunto completo de relações entre códons e aminoácidos (ou sinais de parada), acesso em: Khan Academy 15/12/2023

i	k=3	k=4	k=6
1	0-33%	0-25%	0-17%
2	33-66%	25-50%	17-33%
3	66-99%	50-75%	50-67%
4		75-100%	67-83%
5		75-100%	83-100%

Tabela 2: Frequências de corte para agrupamento i, dado um valor de k

2.1.4 Agrupando por frequência

Outra forma de agrupar os dados foi referente a frequência total de aparições de cada códon nos dados. Optou-se por examinar a diversidade de distribuição dos códons e categorizá-los de maneira mais eficaz. Empregou-se a função quantil para criar um classificador com k categorias, definidas pela frequência de ocorrência dos códons. Assim, obteve-se k conjuntos de códons, organizados com base em sua frequência. Consulte a Figura 2 para obter as frequências referentes a cada agrupamento.

2.2 Visualização e exploração

2.2.1 Estatísticas decritivas

Observam-se algumas estatísticas descritivas sobre os dados na Tabela 34. o total das estatisticas descritivas estão anexos ao relatório.

Tabela 3: Dataset com 13028 linhas e 72 colunas

	SpeciesID	Ncodons	AAA	AUU	GAA
count	13028	13028	13028	13028	13028
mean	130451	79605.8	0.028504	0.0283515	0.0282903
std	124787	719701	0.0178897	0.0175071	0.0143424
min	7	1000	0	0	0
25%	28850.8	1602	0.017315	0.01636	0.01736
50%	81971.5	2927.5	0.025315	0.025475	0.026085
75%	222891	9120	0.03726	0.0381125	0.0368
max	465364	4.06626e+07	0.14601	0.15406	0.14489

Tabela 4: Estatísticas descritivas dos códons, agrupados por frequência

2.2.2 visualizações

A fim de identificar as relações entre as diferentes trincas, nos utilizamos analises visuais, como boxplot, gráfico de barras e heatmap. esses gráficos nos auxiliaram a compreender as disposições dos dados.

Analisamos a distribuição de frequências dos códons, Figura 2, multiplicando a proporção das trincas pelos total, "Ncodons", e agregando valores correspondentes aos mesmos códons. Os dados foram ordenados para destacar os códons mais frequentes.

Com a distribuição de frequências dos códons no primeiro gráfico, pudemos observar a distribuição de códons pelos nossos dados. Observamos que trata-se de uma distribuição em escala linear, permitindo o agrupamento em quantil sem grandes correções.

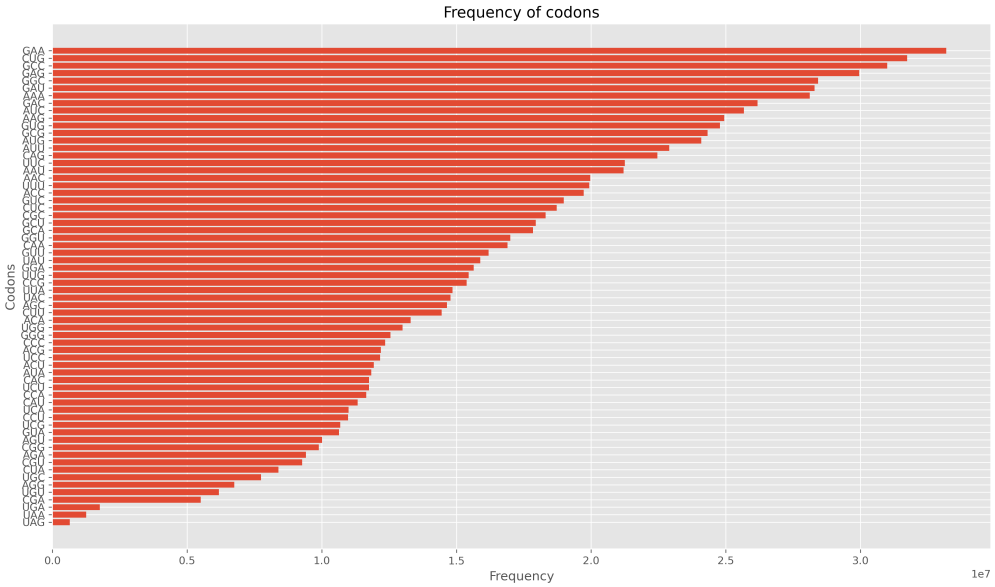

Figura 2: Distribuição sortida dos códons por frequência total

Para tentar compreender melhor os padrões das distribuições das trincas, também calculamos a matriz de correlação de Pearson, como mostrado na Figura 3. Infelizmente, constatamos que os padrões existem, porém não podemos afirmar como eles ocorrem, pois nem todos os aminoácidos apresentam correlação, apenas alguns. É importante ressaltar que o coeficiente de correlação de Pearson é sensível a valores aberrantes, portanto, apenas seu valor não é suficiente para afirmar categoricamente uma correlação.

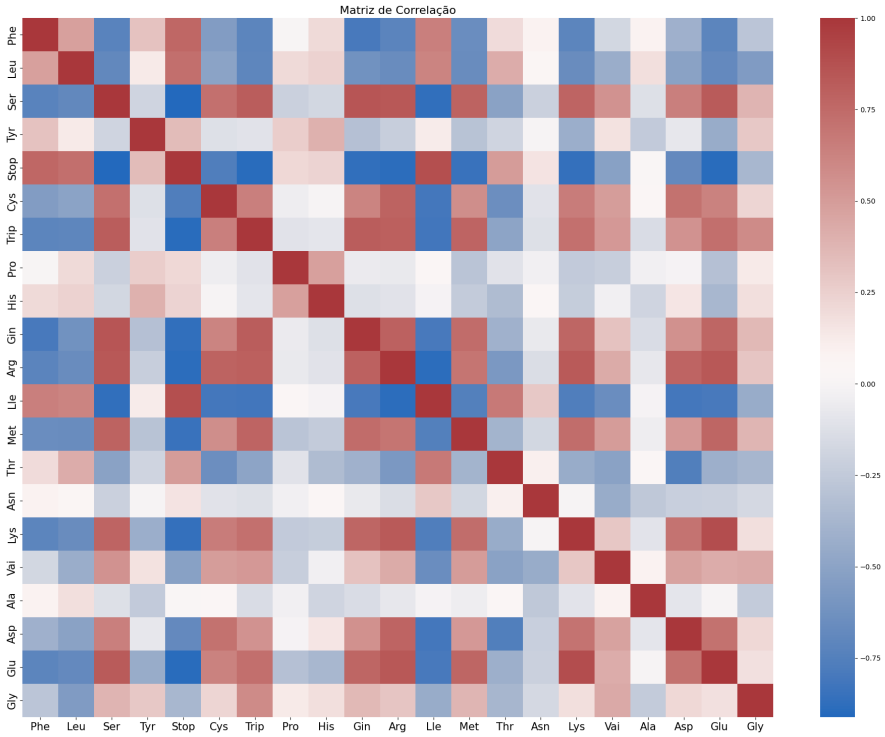

Figura 3: Heat map, matriz de correlção entre aminoácidos nos mamíferos.

Nós também analisamos as distribuições dos aminoácidos e das trincas em cada classificação taxonômica. As marcantes diferenças nas frequências dos aminoácidos levantam a hipótese de que as frequências dos códons seguem um padrão; no entanto, não podemos descrevê-las apenas por meio dos boxplots, conforme mostrado na Figura 4.

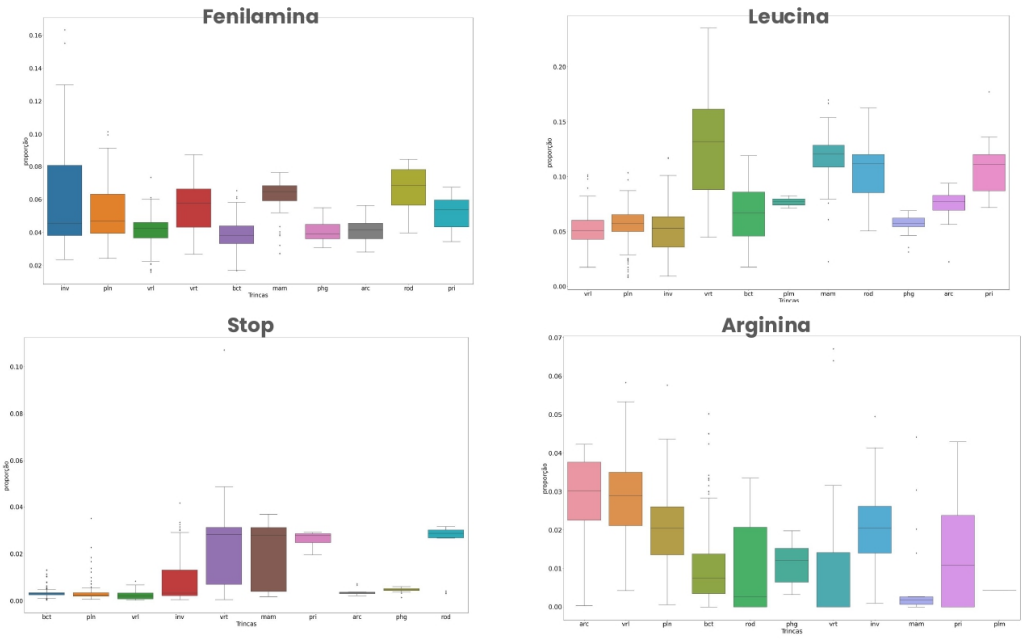

Figura 4: Distribuição dos quatro aminoácidos mais comuns agrupado por classificação taonômica
Em ordem, as classificações: Invertebrados, plantas, vírus, vertebrados, bactéria, mamíferos, bacteriófago, archaea, roedores, primatas

2.2.3 Redução de dimensionalidade

Outra forma de observar os dados foi aplicando a redução dimensional. Com ela pode-se obter informações valiosas sobre a estrutura e padrões de frequências dos códons.

Aliamos a redução dimensional uma técnica de agrupamento, para que os gráficos apresentem, com ainda mais clareza, quais são os grupamentos observados.

Aplicamos o algoritmo de aprendizado não supervisionado KMeans para agrupar os dados em torno das médias dos agrupamentos. O KMeans atribui cada ponto a um dos agrupamentos com base na proximidade aos valores esperados. Observando os gráficos de redução e variando o valor de k, definiu-se empiricamente $k = 3$ como melhor solução.

O agrupamento permite a criação de uma coluna categórica para os dados, que poderá ser futuramente explorada. Além disso, oferece uma cor para os gráficos de redução, contribuindo para o entendimento dos agrupamentos.

Investigamos a distribuição espacial dos códons em um subespaço 64-dimensional, questionando se os aminoácidos formam grupos homogêneos, ou se apresentam grupos facilmente identificáveis. Para isso, utilizamos técnicas de ciência de dados como o Principal Component Analysis (PCA) e o t-istributed Stochastic Neighbor Embedding (t-SNE), para reduzir a dimensionalidade dos dados e obter perspectivas visuais distintas do espaço de códons.

2.3 Teste estatístico

A análise exploratória inicial revelou a complexidade nos padrões de distribuição, evidenciando a insuficiência das visualizações isoladas para uma compreensão abrangente. Diante disso, torna-se essencial uma análise mais aprofundada, utilizando métodos mais sofisticados.

Essa abordagem refinada exige a diferenciação da disposição das trincas em diferentes classificações taxonômicas, já que sem essa distinção, os dados não forneceriam informações relevantes.

Com o objetivo de identificar informações estatisticamente significativas, consideramos a aplicação do teste ANOVA para investigar se a disposição estatística dos códons varia entre as distintas classificações taxonômicas.

No entanto, antes de empregar o teste ANOVA, realizamos o teste de Shapiro-Wilk para verificar se os dados de frequência dos aminoácidos seguem uma distribuição normal. Observamos que, considerando um nível de significância de 0.01, apenas alguns conjuntos de dados assumem uma distribuição normal. Por exemplo, nos arqueas, os aminoácidos Phe, Leu, Pro, entre outros, exibem comportamento normal, enquanto em outras classificações, como nos plasmídios, essa distribuição é observada para todos os aminoácidos.

Diante dessa constatação, optamos por um teste não paramétrico, o teste de Kruskal-Wallis, para evitar suposições de normalidade nos dados. Esse teste revelou diferenças estatisticamente significativas em todas as proporções de aminoácidos entre diferentes classificações taxonômicas. Por exemplo, observamos que os aminoácidos His e Asn mostraram-se estatisticamente distintos em suas frequências entre as classificações taxonômicas, com um nível de significância de $4.580784083491248e-289$ e $2.425446571076577e-261$, respectivamente.

Os resultados completos dos testes de Shapiro-Wilk e Kruskal-Wallis, bem como as médias populacionais e níveis de significância para as médias serem iguais, encontram-se em anexo para referência detalhada.

3 Resultados

3.1 Visuais

3.1.1 Aminoácidos produzíveis por espécie

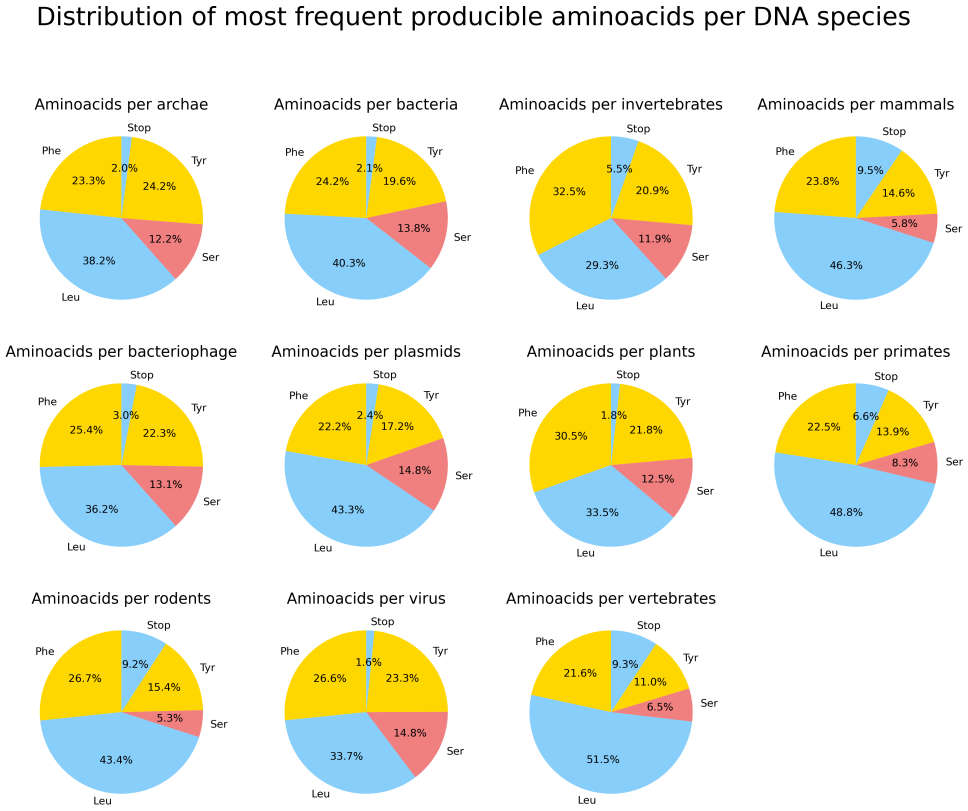

Figura 5: Distribuição dos aminoácidos mais frequentes por espécies
Para cada uma das 11 espécies, a distribuição dos aminoácidos mais frequentes

No contexto da limpeza dos dados, optamos por consolidar os códons conforme os aminoácidos que eles geram, resultando em uma redução para 20 categorias distintas. Posteriormente, procedemos à identificação dos cinco aminoácidos mais frequentes em todas as espécies, analisando sua distribuição individual em cada uma delas. Essa abordagem aprofunda a compreensão da funcionalidade específica de cada aminoácido nos organismos vivos.

Destacamos, de modo especial, a análise da distribuição dos aminoácidos de parada. Observamos que, em alguns reinos, sua ocorrência é significativamente mais frequente em comparação com outros aminoácidos. Este é o caso dos vertebrados (9.3%), mamíferos (9.5%) e os roedores (9.2%). Essa constatação pode sugerir padrões distintos na complexidade ou tamanho das proteínas que compõe estes reinos.

3.1.2 Redução de dimensionalidade e Agrupamento

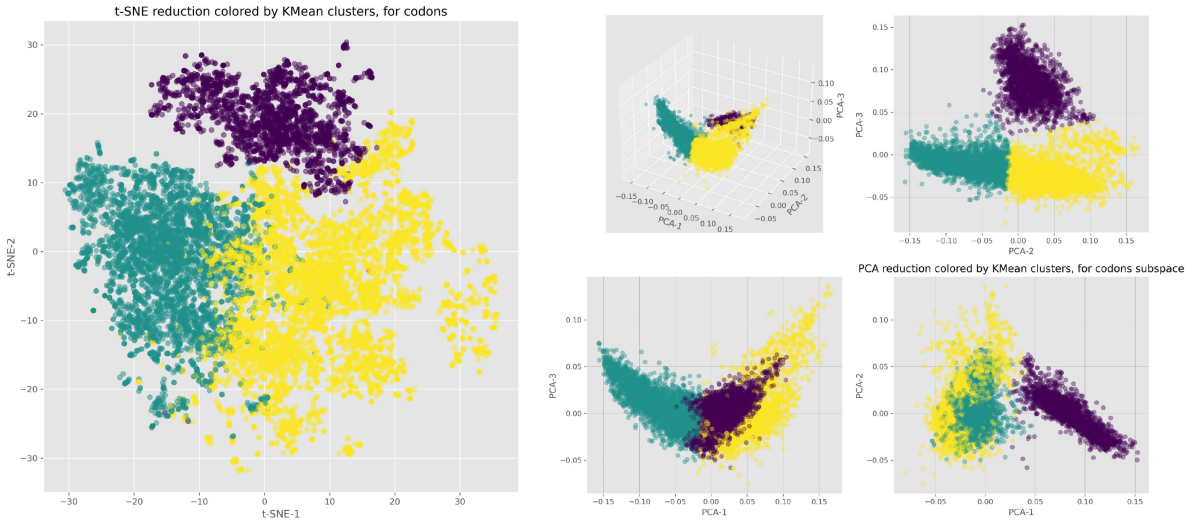

Figura 6: t-SNE a esquerda e PCA a direita pintados por agrupamento kMeans
Duas técnicas de redução de dimensionalidade, coloridos pelos 3 agrupamentos do KMeans. Á esquerda: Redução para um subespaço de 2 dimensões, minimizando a t-Students das distâncias. À direita: Algoritmo linear de redução de dimensionalidade para minimizar norma de projeções euclideanas.

Seja observando o PCA de três dimensões, ou o t-SNE, torna-se claro o porquê da escolha de $k = 3$ grupos. Existem de fato agrupamentos neste objeto multidimensional de trincas.

Note especial atenção neste agrupamento roxo, ele se destaca como um verdadeiro aglomerado a parte. Isso significa que devem existir uma porção de códons substancialmente diferentes dos outros. A partir desse ponto, resolveu-se investigar esse cluster roxo. Resolvemos identificar de quais espécies eles ocorrem.

3.1.3 Reinos por kMeans

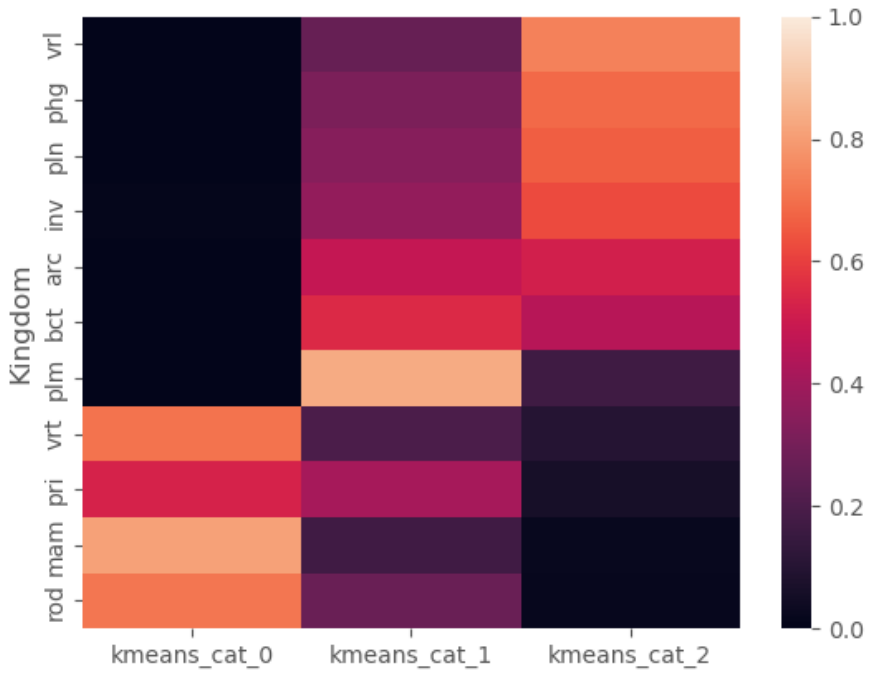

Figura 7: Reinos por agrupamentos do kmeans
Probabilidade do reino ser de cada um dos $k = 3$ agrupamentos

Para tanto, optamos por um Heatmap representando os reinos e as categorias, apresentado na Figura 7. Note que o valor de 0.0 a 1.0 representa a probabilidade de cada reino pertencer a cada uma das categorias do KMeans.

Em especial, 'kmeans-cat-0' representa o agrupamento roxo encontrado na redução. Note que essa categoria é máxima para os reinos dos roedores, mamíferos, primatas e vertebrados.

Agora relembre que os reinos, os roedores, os mamíferos e os vertebrados foram os seres vivos com mais códons de parada em seu DNA, proporcionalmente.

Portanto, é possível identificar que o grupo de roedores, mamíferos e vertebrados apresentam alguma característica fundamentalmente distinta em seu DNA, pela alta frequência de códons de paradas e similaridades nas probabilidades de códons.

3.1.4 Soma dos quadrados das correlações

Usando técnicas computacionais exaustivas, observamos os gráficos de aminoácidos de cada espécie. Em especial observamos diversas matrizes de correlação dos aminoácidos, uma para cada espécie. Notamos que algumas espécies apresentavam gráficos com números mais extremos, seja para cima ou para baixo, e outras apresentavam gráficos mais próximos do zero.

Queríamos representar esse padrão por de maneira completa, sem apresentar dezenas de matrizes. Assim, montamos a soma dos quadrados das correlações mútuas, na Figura 8. Com este último gráfico, conseguimos identificar a 'correlação esperada' entre os aminoácidos.

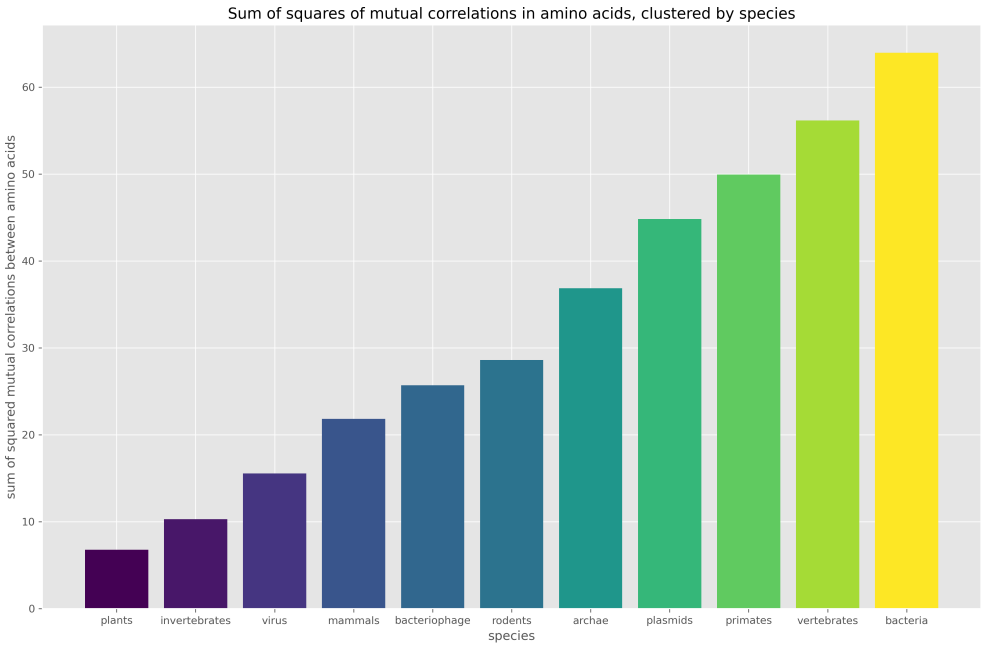

Figura 8: Soma dos quadrados das correlações mútuas entre aminoácidos, agrupado por espécies

Para cada reino, o valor S_k é a soma das correlações de todos os i, j aminoácidos: $V_k = \sum_{(i \neq j)} [Cor(i,j)]^2$

Nota-se que as bactérias lideram a correlação, revelando que as correlações entre as frequências de diferentes aminoácidos é elevada, em números absolutos. Isso pode indicar estruturações mais simples de ordenamento entre aminoácidos, de modo que eles apareçam sempre um em sequência do outro.

Por outro lado, o DNA dos reinos que apresentam as frequências de aminoácidos pouco correlacionada pode indicar um padrão informacional mais aleatória ou padrões mais complexos de ordenação mútua. Assim, o gráfico pode ser entendido como uma medida informacional do padrão de ordenamento dos aminoácidos.

4 Conclusão

Diante disso, procedemos à análise da probabilidade de ocorrência dos códons, empregando diversas técnicas de tratamento de dados, descrições estatísticas e visualizações. Destaca-se que realizamos uma demonstração estatística, evidenciando que as distribuições dos aminoácidos apresentam diferenças significativas entre si.

Adicionalmente, utilizamos algoritmos de agrupamento e redução dimensional para discernir um padrão específico nas distribuições dos códons. Aprofundamos a investigação deste agrupamento e identificamos quais as classificações taxonômicas em que eles, provavelmente, pertenciam.

Finalmente, procedemos à análise das correlações internas das frequências de aminoácidos, buscando desvelar padrões subjacentes na ordenação dos mesmos. Destacamos que, embora vertebrados e primatas tenham apresentado uma correlação significativa, tal relação não se verificou de maneira equivalente para mamíferos e roedores, de modo que não podemos afirmar a relação entre alta correlação de aminoácidos e alta incidência de códons de parada.


ICMC
USP
SÃO CARLOS