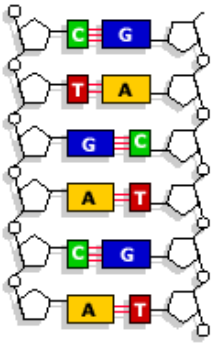
Universidade de Aveiro Departamento de Matemática
2010

Sara dos Santos Escudeiro Cruz

Análise Estatística de Dados de Biologia Molecular



Universidade de Aveiro Departamento de Matemática
2010

Sara dos Santos Escudeiro Cruz

Análise Estatística de Dados de Biologia Molecular

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática e Aplicações, realizada sob a orientação científica da Professora Doutora Vera Mónica Almeida Afreixo, Professora Auxiliar Convidada do Departamento de Matemática da Universidade de Aveiro, e da co-orientadora Professora Doutora Adelaide de Fátima Baptista Valente Freitas, Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro.

À minha família, em especial ao Luís, à Vitória e ao Henrique.

o júri

presidente

Professora Doutora Isabel Maria Simões Pereira
Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro

Professora Doutora Luzia Augusta Pires Gonçalves
Professora Auxiliar do Instituto de Higiene e Medicina Tropical da Universidade Nova de Lisboa

Professora Doutora Adelaide de Fátima Baptista Valente Freitas
Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro

Professora Doutora Vera Mónica Almeida Afreixo
Professora Auxiliar Convidada do Departamento de Matemática da Universidade de Aveiro

agradecimentos

À minha orientadora, Professora Doutora Vera Mónica Almeida Afreixo, pela disponibilidade, incentivo, conhecimentos transmitidos e orientação científica.

À Professora Doutora Adelaide de Fátima Baptista Valente Freitas, pelas suas óptimas sugestões.

Aos meus colegas, em especial ao Engenheiro Azevedo e ao Professor Durão, pelo incentivo que me deram, durante estes anos, à aprendizagem de novas matérias.

À minha colega de Mestrado, Joana, pela preocupação e apoio durante a parte escolar.

palavras-chave

Distâncias entre nucleótidos, classificação hierárquica, análise de componentes principais, mistura finita de distribuições paramétricas, algoritmo EM, ADN.

resumo

Nesta dissertação são analisados os genomas completos de 46 espécies de organismos, com o objectivo de investigar a existência, ou não, de características estatísticas discriminatórias da classe a que pertence cada uma das espécies em estudo, com base na distribuição empírica da *distância global entre nucleótidos iguais*. Esta distribuição resulta do mapeamento da estrutura primária do ADN proposto e avaliado por Afreixo *et al*. (2009).

São utilizadas metodologias estatísticas multivariadas de análise não-supervisionada e de redução da dimensionalidade, nomeadamente as classificações hierárquica e não-hierárquica e a análise de componentes principais. Verifica-se que o mapeamento da distância global entre nucleótidos iguais captura características essenciais do ADN das espécies analisadas, uma vez que a distribuição das primeiras distâncias determina uma possível assinatura genética capaz de permitir a diferenciação entre espécies. Esta diferenciação é conseguida não só a um nível geral, entre os dois grandes grupos de espécies eucariotas e procariotas, mas também a níveis mais especializados.
No que diz respeito ao ajustamento de modelos probabilísticos teóricos à distribuição empírica de cada espécie, são avaliados o modelo proposto em Afreixo *et al.* (2009) e também um modelo alternativo, ambos baseados em misturas finitas de distribuições geométricas. No caso deste último, é utilizado o algoritmo EM (*Expectation-Maximization*) para estimar os seus parâmetros. A qualidade do ajustamento dos modelos teóricos à distribuição empírica é investigada com o auxílio do teste de ajustamento do qui-quadrado e também com a utilização de medidas de similaridade. Os resultados obtidos permitem constatar que, na maioria das espécies em estudo, o modelo de mistura de quatro distribuições geométricas é aquele que melhor se ajusta à distribuição empírica da distância global entre nucleótidos iguais.

keywords

Inter-nucleotide distances, hierarchical classification, principal components analysis, finite mixture distributions, EM algorithm, DNA.

abstract

In this dissertation the complete genomes of 46 species of organisms are analysed, with the aim of investigating the possible existence of discriminatory statistical characteristics of the class to which each of the species under study belongs, based on the empirical distribution of the *global distance between equal nucleotides*. This distribution came about from the mapping scheme for the primary structure of DNA proposed and assessed by Afreixo *et al.* (2009).

Unsupervised multivariate statistical and dimensionality reduction methods are used in the present analysis, namely hierarchical classification, non hierarchical classification and principal component analysis. It is shown that the mapping of the global distance between equal nucleotides captures essential features of the DNA of the species studied, as it allows to infer that the distribution of the first distances represents a possible genetic signature capable of differentiating among species. This differentiation is achieved not only at a general level between the two major groups of species, eukaryotic and prokaryotic, but also at more specialized levels.
Furthermore, fittings of probabilistic models to the empirical distribution are investigated for each specie. More specifically, the model proposed by Afreixo *et al.* (2009) and an alternative model, both based on finite geometric mixture models, are analysed. In the latter case, the EM (*Expectation-Maximization*) algorithm is used to estimate its parameters. The goodness of fit of the theoretical models is assessed using a chi-square test and measures of similarity. For most species studied, the results show that four-component geometric mixture models are the ones that better fit to the empirical distribution of the global distance between equal nucleotides.

Conteúdo

Lista de Figuras

Lista de Tabelas

Simbologia

Básica

ADN	ácido desoxirribonucleico
$\mathcal{A}$ ={A, C, G, T }	alfabeto do ADN
A-T	Adenina-Timina
C-G	Citosina-Guanina
D^x, $x \in \mathcal{A}$	distância entre nucleótidos iguais
D	distância global entre nucleótidos iguais
$S = (S_1, S_2, \ldots, S_N)$	sequência simbólica de ADN
$S^x = (S_1^x, S_2^x, , \ldots, S_{N^x}^x)$	sequência cujos elementos são os índices das posições do nucleótido x na sequência S
$\left[S_1^{\mathrm{A}}\ S_1^{\mathrm{C}}\ S_1^{\mathrm{G}}\ S_1^{\mathrm{T}}\right]$	posição da primeira ocorrência do nucleótido x em S
Y	variável aleatória ou vector aleatório
y	observação de Y
v.a.	variável aleatória
i.i.d.	independente e identicamente distribuído
f.m.p.	função massa de probabilidade
Ψ	vector dos parâmetros
Θ	espaço paramétrico
$tr(X)$	traço da matriz X
$diag(X)$	diagonal principal da matriz X
Σ	matriz de covariâncias
ACP	análise de componentes principais
CPs	componentes principais
DVS	decomposição em valores singulares

Funções e estruturas de dados do R utilizadas

package - base

c()	*paste*()	*rep*()
colnames()	*rownames*()	*names*()
numeric()	*as.numeric*()	*vector*()
matrix()	*list*()	*NROW*()
max()	*min*()	*sum*()
sqrt()	*round*()	*abs*()
sort()	*dim*()	*length*()
apply()	*print*()	*svd*()
attr()	*is.infinite*()	*attributes()*

package - graphics

par()	*bxp*()	*plot*()
barplot()	*lines*()	*abline*()
text()	*points*()	

package - stats

sample()	*dgeom*()	*rgeom*()
na.omit()	*cor*()	*cov*()
dist()	*hclust()*	*kmeans()*
prcomp()	.$*rotation*	.$*x*
chisq.test()	*qchisq*()	*dendrapply()*

package - Hmisc

wtd.mean()	*wtd.var*()	*wtd.quantile*()

package - FactoMineR

PCA()	.$*eig*	.$*loadings*
.$*var*$*cor*	.$*var*$*cos2*	.$*var*$*contrib*

Capítulo 1

Introdução

Descoberta a existência do ácido desoxirribonucleico (ADN) no núcleo das células pelo bioquímico suíco Frederich Miescher em 1869 [9], apenas em 1944 foi sugerido que seria essa molécula, e não as proteínas como até então se pensava, que constituía o suporte da informação genética [4], isto é, da informação que define as características dos organismos vivos e que é transportada de geração em geração em consequência do processo de reprodução. A confirmação dessa possibilidade surgiu em 1952, em resultado do trabalho de Alfred Hershey e Martha Chase [9]. Desde então têm-se multiplicado os esforços de investigação multi-disciplinares sobre a molécula de ADN, passando, entre outros momentos importantes, pela descoberta da sua estrutura e mecanismos de replicação e, mais recentemente, pela sequenciação completa do ADN de um número crescente de organismos, incluindo o do ser humano, tendo esta última sido iniciada em 1990 e terminada em 2003 [39].
À luz do conhecimento actual constata-se que todos os organismos conhecidos utilizam a molécula de ADN como suporte para a informação de hereditariedade [3].

Apesar de terem vindo a ser feitos grandes avanços no conhecimento sobre o ADN e de haver um número crescente de aplicações práticas desse conhecimento com impacto directo nas nossas vidas, por exemplo na medicina e na análise forense, existe a certeza de que ainda resta muito para descobrir sobre o ADN.

Neste capítulo será feita uma pequena introdução a alguns conceitos básicos de biologia relacionados com o ADN. Serão também apresentados a motivação e os objectivos gerais desta dissertação, bem como a organização da mesma.

1.1 Conceitos biológicos

A estrutura actualmente aceite para a molécula de ADN foi descrita pela primeira vez por James Watson e Francis Crick, num artigo publicado em 1953 na revista Nature [43]. Nesse artigo foi proposta uma estrutura para a molécula de ADN radicalmente diferente de outras que haviam sido sugeridas até então, descrevendo-a como sendo constituída por duas cadeias helicoidais enroladas em torno do mesmo eixo, em que cada elo destas cadeias seria formado por uma pentose (desoxirribose), um grupo fostato e uma base azotada, e estaria interligado a um elemento idêntico na outra cadeia por ligações de hidrogénio entre as respectivas bases azotadas (ver Figura 1.1). A estes elos dá-se o nome de **nucleótidos** (ver Figura 1.2).

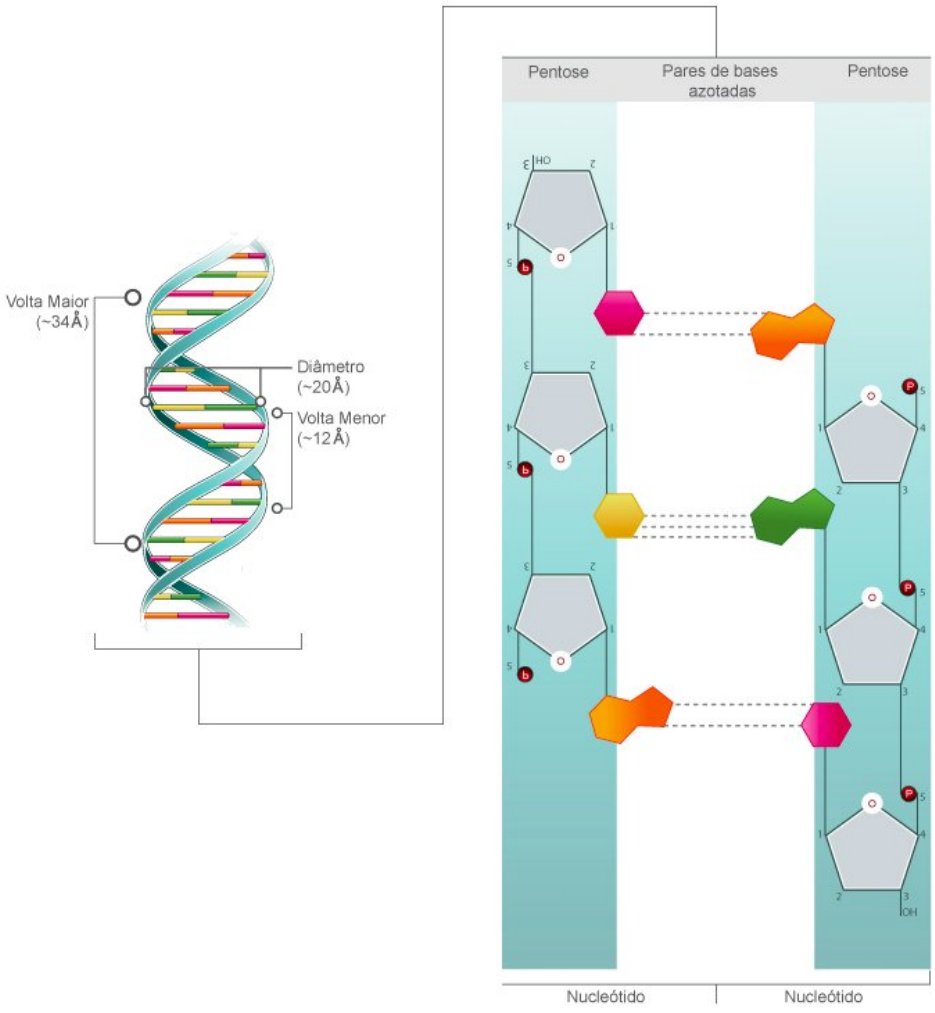


Figura 1.1: Representação simplificada da estrutura do ADN. Adaptação de uma figura publicada pelo Grupo de Ciências Biológicas do Instituto Superior Técnico [11].

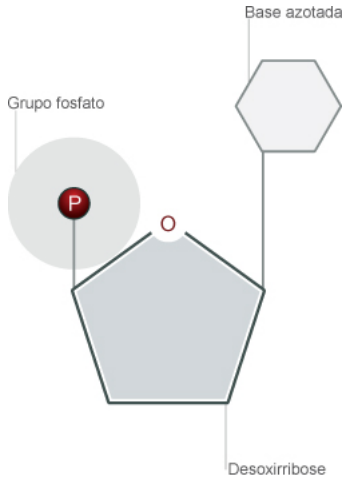


Figura 1.2: Estrutura básica de um nucleótido. Imagem publicada pelo Grupo de Ciências Biológicas do Instituto Superior Técnico [11].

No caso do ADN existem quatro tipos de bases azotadas, as quais são habitualmente designadas pela primeira letra do seu nome: A (Adenina), C (Citosina), G (Guanina) e T (Timina). As bases azotadas podem ser classificadas, de acordo com a sua estrutura, em purinas e pirimidinas. A Adenina e a Guanina são purinas, pois possuem uma estrutura com dois anéis; a Citosina e a Timina são pirimidinas, pois têm uma estrutura com apenas um anel. Estas bases apenas se emparelham entre si (por pontes de hidrogénio) sob as formas A-T e C-G, dizendo-se então que os elementos de cada par são complementares [3].

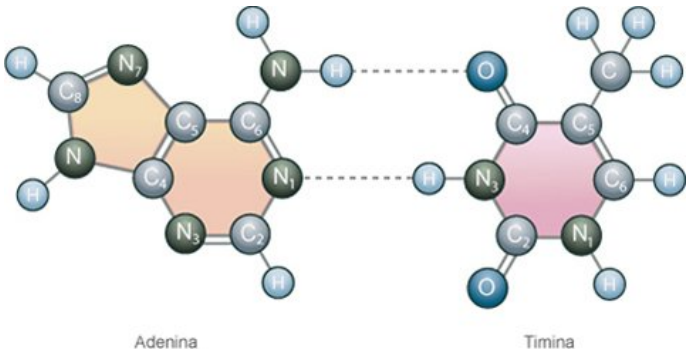


Figura 1.3: Duas pontes de hidrogéneo: ligação Adenina-Timina. Imagem publicada pelo Grupo de Ciências Biológicas do Instituto Superior Técnico [11].

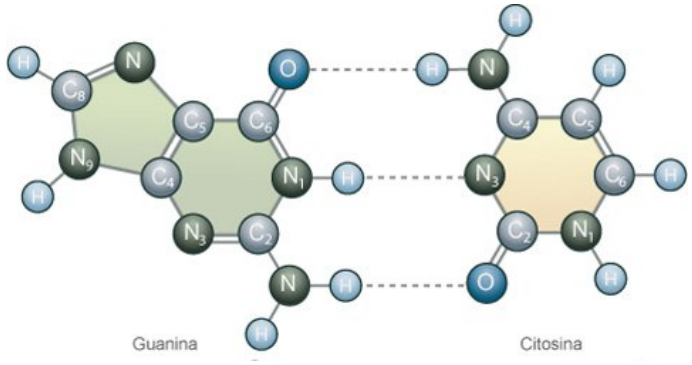


Figura 1.4: Três pontes de hidrogéneo: ligação Guanina-Citosina. Imagem publicada pelo Grupo de Ciências Biológicas do Instituto Superior Técnico [11].

Devido a esta complementaridade, a sequência de nucleótidos de uma cadeia determina a sequência de nucleótidos da outra cadeia, o que significa que se for conhecida uma das cadeias então facilmente se obtém a outra. É também com base nesta complementaridade que se desenrola o processo de reprodução das células. As duas cadeias separam-se e servem de modelo para a criação de duas novas moléculas de ADN idênticas à original, se excluirmos a possibilidade de ocorrerem erros genéticos durante este processo. A designação dos nucleótidos está relacionada com a base azotada que contêm. Existem, portanto, nucleótidos do tipo A, C, G e T. Assim sendo, é possível ler-se a cadeia de ADN como uma sequência de letras, por exemplo,

$$\text{AAGGTTATCCACTATGTTTTTCGATAAAAAGCTTAA} \cdots$$

A estrutura primária da molécula de ADN, isto é, a ordem específica da sequência dos nucleótidos que a compõem, determina a informação genética necessária para criar um organismo específico, com todas as suas particularidades. A sequência completa de ADN de cada célula chama-se **genoma**.

Dentro da sequência de ADN, os nucleótidos associam-se em grupos de três elementos, formando os chamados **codões**. Existem, portanto, 64 (4^3) codões distintos. Cada um dos codões contém as instruções necessárias para a produção de um aminoácido. Os **aminoácidos** são os componentes estruturais das proteínas. As **proteínas** são polímeros complexos de aminoácidos presentes em quase todos os aspectos da fisiologia e bioquímica dos organismos, funcionando, por exemplo, como componentes estruturais das células ou intervindo como catalisadores em reacções bioquímicas essenciais, na qualidade de enzimas. Aos grupos de codões que se encontram correctamente organizados no sentido de

serem capazes de produzir uma proteína específica, dá-se o nome de **genes**. O comprimento dos genes é variável. Uma vez que na construção das proteínas apenas são utilizados 20 aminoácidos diferentes e existem 64 codões, então alguns desses codões corresponderão ao mesmo aminoácido; adicionalmente, existem alguns codões com funções especiais que não a produção de proteínas [3]. O genoma humano contém mais de 30.000 genes. Ao longo do genoma, e intercalados com os genes, existem sequências de nucleótidos que não têm uma função codificante de proteínas. Faz-se assim a distinção entre regiões codificantes e regiões não codificantes no genoma, não sendo ainda clara a função destas últimas. Os genes organizam-se em **cromossomas**. Nos organismos **procariotas**, isto é, naqueles que não têm um núcleo celular bem definido, as células possuem normalmente um cromossoma circular. As bactérias são um exemplo de organismos procariotas. Nos organismos **eucariotas**, isto é, naqueles que possuem um núcleo celular bem definido, os seus cromossomas, os quais têm geralmente uma forma linear, localizam-se no núcleo celular e variam em número conforme a espécie. As plantas, os animais e os fungos são organismos eucariotas [16].

1.2 Motivação e objectivos gerais

As sequências de ADN são habitualmente representadas por sequências dos símbolos A, C, G e T. No entanto, esta forma de representação não é, normalmente, a mais conveniente do ponto de vista do tratamento matemático. Assim sendo, torna-se necessário recorrer a esquemas de mapeamento que permitam traduzir a representação simbólica do ADN para uma representação numérica capaz de ser analisada por aplicação de técnicas estatísticas e de sinal, entre outras [35]. Esta dissertação surge no seguimento do trabalho desenvolvido em [2], no qual é proposto e avaliado um novo mapeamento directamente relacionado com as características intrínsecas do ADN e que pode ser útil para a discriminação entre diferentes espécies. Esse mapeamento faz corresponder, a cada posição da sequência de ADN em análise, o valor da distância entre o nucleótido que se encontra nessa posição e o nucleótido igual que imediatamente lhe sucede na sequência. Caso não exista mais nenhum nucleótido desse tipo até ao final da sequência simbólica, volta-se ao início da sequência e prossegue-se a contagem até se encontrar o primeiro nucleótido desse tipo, isto é, considera-se a sequência de ADN como sendo cíclica. Este mapeamento é designado por *distância global entre nucleótidos iguais*. Os resultados em [2] permitiram concluir

que este mapeamento captura características essenciais do ADN das espécies analisadas, no sentido em que permite a construção de dendrogramas interpretáveis como árvores filogenéticas, por os mesmos estarem de acordo com as similaridades esperadas entre as espécies; por conseguinte, é aí sugerido que a distribuição das primeira distâncias representa uma possível assinatura genética capaz de permitir a diferenciação entre espécies. A distribuição modelo proposta em [2] para descrever as propriedades estatísticas do genoma baseou-se na lei da probabilidade total. A análise estatística da sequência de distâncias global entre nucleótidos iguais foi efectuada sobre o vector dos erros relativos entre a distribuição modelo e a distribuição empírica de cada espécie, tendo o estudo incidido sobre 28 espécies.

Um dos objectivos desta dissertação é a utilização de metodologias estatísticas adicionais, nomeadamente, a análise classificatória (hierárquica e não-hierárquica) e a análise de componentes principais (ACP), para expandir a análise efectuada em [2]; um segundo objectivo é tentar encontrar, para cada espécie, uma nova distribuição modelo que melhor se ajuste à sua distribuição empírica da sequência de distâncias global entre nucleótidos iguais. A representação gráfica da distribuição empírica de cada espécie, confrontada com a distribuição modelo proposta por [2], sugerem averiguar o ajustamento a uma nova distribuição definida por uma mistura finita de distribuições geométricas.

Os dados analisados nesta dissertação são constituídos pelo genoma completo de 46 espécies, incluindo as 28 usadas em [2]. Assume-se que, para caracterizar o genoma de uma espécie, foram sequenciados os genomas de k indivíduos escolhidos de forma aleatória de entre uma população constituída por todos os organismos dessa espécie. Esses k genomas foram então reduzidos a um só genoma, o qual se considera ser representativo da espécie em questão, sendo este o genoma que é disponibilizado pelas bases de dados públicas.

1.3 Organização da dissertação

Neste capítulo foi apresentada a motivação que esteve na base da elaboração desta dissertação e enunciados os objectivos gerais que se pretendem atingir com a mesma. Afloraram-se ainda alguns conceitos básicos de biologia relacionados com o ADN. A restante dissertação está organizada como se indica a seguir.

O **segundo capítulo** começa por descrever a origem dos genomas que irão ser objecto de análise e identificar as espécies a que correspondem. Em seguida, e após serem feitas algumas considerações genéricas sobre a necessidade da existência de esquemas de mapeamento que permitam converter a representação simbólica do ADN para uma representação numérica capaz de ser analisada recorrendo a técnicas (clássicas) de análises estatísticas e de sinal, são apresentados os dois esquemas de mapeamento que irão ser usados, designadamente o mapeamento da distância entre nucleótidos iguais, D^x, $x \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, e o mapeamento da distância global entre nucleótidos iguais, D.
Nesta altura, o trabalho prossegue com uma análise exploratória dos dados já mapeados, a que se segue a caracterização das distribuições dos mapeamentos D^x e de D propostas por [2]. Finalmente, e com o objectivo de melhor salientar a diferença entre a distribuição empírica de cada distância e a correspondente distribuição modelo proposta por [2], determina-se o vector dos erros relativos entre uma e outra, para cada espécie, apresentando-os sob a forma de uma matriz, sendo sobre esta matriz que irá incidir a análise efectuada no Capítulo 3.

No **terceiro capítulo** são aplicadas à matriz dos erros relativos obtida no Capítulo 2 algumas técnicas estatísticas multivariadas de análise não-supervisionada e de redução de dimensionalidade, designadamente a análise classificatória, hierárquica e não-hierárquica, e a análise de componentes principais, para diferenciar e caracterizar as espécies. Na classificação hierárquica das espécies foram gerados dois dendrogramas, tendo sido utilizada como medida de similaridade a distância euclidiana e como métodos de agregação o método de Ward e o método de ligação completa (*complete linkage*), este último já utilizado em [2]. Na classificação não-hierárquica o método utilizado foi o *K-means*.
Em relação à ACP, realizaram-se três análises distintas considerando-se, respectivamente, a matriz dos erros relativos padronizada, não padronizada e apenas centrada.

No **quarto capítulo** é apresentada a definição de misturas finitas de distribuições paramétricas (caso discreto), tendo como pano de fundo a tentativa de verificar a suposição de que uma nova distribuição modelo, resultante também de uma mistura finita de distribuições geométricas, poderá revelar um melhor ajustamento à distribuição empírica de cada espécie. Em seguida, é tratado o problema de identificabilidade dos modelos de misturas finitas de distribuições paramétricas.
Prossegue-se então com a apresentação da estrutura de dados incompletos para o caso de

misturas. Esta conceptualização do modelo de mistura em termos de dados incompletos é extremamente útil, na medida em que permite a estimação de máxima verosimilhança dos parâmetros da mistura através do algoritmo EM. Para além da descrição do algoritmo EM no contexto do modelo de misturas, e uma vez que os dados observados aos quais se irá aplicar este algoritmo se encontram categorizados, apresentar-se-á também uma adaptação desse algoritmo para este caso.
No final deste capítulo, são descritos o teste de ajustamento do qui-quadrado e algumas medidas de similaridade entre distribuições, com o objectivo de avaliar o ajustamento entre os modelos teóricos propostos e a distribuição empírica. São ainda apresentados, para cada espécie, as estimativas dos parâmetros da distribuição modelo originalmente proposta em [2], as estimativas dos parâmetros da mistura de quatro distribuições geométricas estimados pelo algoritmo EM e os resultados de duas medidas de similaridade.

O **quinto capítulo** apresenta as conclusões desta dissertação e menciona algumas ideias possíveis para trabalho futuro.

No apêndice A apresentam-se alguns resultados complementares não incluídos anteriormente.
Finalmente, o apêndice B inclui todo o código R desenvolvido durante esta dissertação, que inclui várias funções e *scripts*:

- construção das caixas de bigodes para as n espécies simultaneamente, com base em dados categorizados;
- representação da distribuição da sequência de distâncias entre nucleótidos iguais;
- função massa de probabilidade da mistura de distribuições geométricas;
- representação de uma mistura de duas distribuições geométricas e das suas componentes;
- representação dos dendrogramas;
- análise de componentes principais;
- decomposição em valores singulares;
- *K-means*;

- algoritmo EM;
- estimativas iniciais para aplicação do algoritmo EM (mistura de 4 geométricas) às 46 espécies;
- estimativas dos parâmetros da mistura de 4 geométricas obtidas via algoritmo EM às 46 espécies;
- funções usadas no cálculo das medidas de similaridade;
- cálculo das medidas de similaridade entre distribuições para as 46 espécies;
- teste de ajustamento do qui-quadrado;
- representação gráfica das distribuições: empírica, modelo e mistura de quatro geométricas (via algoritmo EM);
- matriz dos erros relativos.

Capítulo 2

Distâncias entre nucleótidos

Neste trabalho analisaram-se as sequências de ADN completas de 46 espécies. Destas, 43 foram obtidas do *National Center for Biotechnology Information (NCBI)*[14]. No caso das outras três espécies, *Populus_trichocarpa* (*California poplar*), *Xenopus_tropicalis* (*Western clawed frog*) e *Takifugu_rubripes*, o genoma da primeira foi obtido no *Joint Genome Institute* [21], o da segunda da *Xenbase* [44] e o da terceira no *Genome Project* [38]. No pré-processamento desta informação foram retiradas todas as ocorrências de símbolos que não sejam um dos quatros nucleótidos {A, C, G, T}. Na Tabela 2.1 é apresentada a lista de espécies que irão ser analisadas, bem como a versão dos ficheiros que contêm os respectivos genomas completos. Apenas com o objectivo de auxiliar a interpretação dos resultados, coloriram-se as espécies em estudo da seguinte forma: bactérias (vermelho), plantas (azul escuro), animais (preto), protozoários (azul claro) e fungos (verde).

Tabela 2.1: Lista das 46 espécies em estudo, com a designação abreviada de cada espécie.

Espécie	Abreviatura	Versão
Aeropyrum_pernix (bactéria)	*Ap*	NC000854
Halobacterium_salinarum_R1 (bactéria)	*Hr*	NC010364
		NC010366
		NC010369
Methanococcus_jannaschii (bactéria)	*Mj*	NC000909
		NC001732
		NC001732
Pyrococcus_furiosus (bactéria)	*Pf*	NC003413
Thermococcus_kodakarensis_KOD1 (bactéria)	*Tk*	AP006878
Bacillus_anthracis_Ames (bactéria)	*Ba*	NC003997
Bacillus_subtilis (bactéria)	*Bs*	NC000964

continua na página seguinte

Tabela 2.1 – continuação da página anterior

Espécie	Abreviatura	Versão
Chlamydia_trachomatis (bactéria)	*Ct*	NC000117
Clostridium_botulinum_A (bactéria)	*Cb*	NC009495
		NC009496
Desulfovibrio_vulgaris_DP4 (bactéria)	*Dv*	NC008741
		NC008751
E_coli (bactéria)	*Ec*	NC000913
Haemophilus_influenzae (bactéria)	*Hi*	NC000907
Helicobacter_pylori_26695 (bactéria)	*Hp*	NC000915
Mycoplasma_genitalium (bactéria)	*Mg*	NC000908
Pseudomonas_aeruginosa (bactéria)	*Pa*	NC002516
Staphylococcus_aureus_COL (bactéria)	*Sa*	NC002951
		NC006629
Streptococcus_mutans (bactéria)	*Sm*	NC004350
Streptococcus_pneumoniae_ATCC_700669 (bactéria)	*St*	NC011900
Arabidopsis_thaliana (planta)	*At*	AGI 7.2
Oryza_sativa (planta)	*Os*	NC008394
		NC008405
Populus_trichocarpa (planta)	*Po*	Build 1.0
Vitis_vinifera (planta)	*Vv*	Build 1.1
Bos_taurus (vaca)	*Bt*	Build 4.1
Cannis_familiaris (cão)	*Cf*	Build 2.1
Equus_caballus (cavalo)	*Eq*	Build 2.1
Gallus_gallus (galinha)	*Gg*	Build 2.1
Apis_mellifera (abelha)	*Am*	Build 4.1
Drosophila_melanogaster (mosca da fruta)	*Dm*	Build 4.1
M_musculus (rato)	*Mu*	Build 37.1
Caenorhabditis_elegans (minhoca)	*Ce*	NC003279
Rattus_norvegicus (rato)	*Rn*	Build 4.1
Xenopus_Tropicalis (sapo)	*Xt*	Build 4.1
H_sapiens (primata)	*Hs*	Build 36.3
Macaca_mulatta (primata)	*Mm*	Build 1.1
Pan_troglodytes (primata)	*Pt*	Build 2.1
D_rerio (peixe)	*Dr*	Build 3.1
Takifugu_rubripes (peixe)	*Fu*	fourth assembly
Ornithorhynchus_anatinus (ornitorinco)	*Oa*	Build 1.1
Dictyostelium_discoideum (protozoário)	*Dd*	Build 2.1
Leishmania_infantum (protozoário)	*Li*	NC009277
		NC009386
		NC009420
Plasmodium_falciparum (protozoário)	*Pl*	Build 2.1
Trypanosoma_brucei (protozoário)	*Tb*	NC005063
		NC007276

continua na página seguinte

Tabela 2.1 – continuação da página anterior

Espécie	Abreviatura	Versão
		NC007283
		NC007334
		NC008409
		NT165287:88
Candida_albicans (fungo)	*Ca*	NC007436
Neurospora_crassa (fungo)	*Nc*	NW001091935
		NW00102755
Saccharomyces_cerevisiae (fungo)	*Sc*	SGD 1
Schizosaccharomyces_pombe_OLD (fungo)	*Sp*	Build 1.1

2.1 Mapeamento do ADN em sequências de distâncias entre nucleótidos iguais

A representação de uma sequência de ADN por uma sequência de símbolos A, C, G e T não é, normalmente, a mais conveniente do ponto de vista da aplicação das técnicas clássicas de análise estatística. Existe, portanto, a necessidade de fazer a conversão da representação simbólica do ADN para um formato numérico cujas propriedades matemáticas reflictam, tanto quanto possível, as características biológicas relevantes da sequência simbólica original. A conversão entre as duas representações, feita por um esquema de mapeamento, é, portanto, um processo crítico que deverá minimizar a introdução de quaisquer perturbações capazes de provocar a alteração dos resultados da análise dos dados (ver Secção 1.3 de [13]).

A escolha do tipo de mapeamento poderá ser feita de forma a evidenciar certas características da sequência de ADN [35]. Podem ser encontradas em [1] e [36] referências para mapeamentos que vêm sendo utilizados por vários autores.

Os dados que irão ser analisados neste trabalho foram obtidos através dos mapeamentos propostos em [2], designadamente, o mapeamento da distância entre nucleótidos iguais e o mapeamento da distância global entre nucleótidos iguais.

Caracterização do mapeamento das distâncias

Seja $\mathcal{A} = \{\text{A,C,G,T}\}$, $S = (S_1, S_2, \ldots, S_N)$ uma sequência simbólica de ADN de comprimento N e $S^x = (S_1^x, S_2^x, , \ldots, S_{N^x}^x)$ uma nova sequência cujos elementos são os índices das posições do nucleótido x na sequência S. Denotar-se-ão por s e s^x as concretizações das sequências S e S^x, respectivamente. O comprimento da sequência S^x é igual ao número de nucleótidos do tipo x existentes na sequência simbólica de ADN, representando-se por N^x, $x \in \mathcal{A}$. Aplicando o mapeamento da distância entre nucleótidos iguais à sequência S^x, obtém-se a **sequência de distâncias entre nucleótidos iguais**, D^x, a qual é uma sequência numérica de comprimento N^x definida por

$$D^x = \left(D_1^x, D_2^x, \ldots, D_{N^x-1}^x, D_{N^x}^x\right),$$

onde

$$D_i^x = \begin{cases} S_{i+1}^x - S_i^x, & i = 1, 2, \ldots, N^x - 1 \\ N + S_1^x - S_i^x, & i = N^x \end{cases}$$

Denotar-se-á por d^x a concretização da sequência D^x.

Aplicando, como exemplo, o mapeamento da sequência de distâncias entre nucleótidos iguais ao fragmento de ADN dado pela sequência

$$\text{AAGGTTATCCACTAT}, \tag{2.1}$$

de comprimento $N = 15$, tem-se que

$$s = (\text{A, A, G, G, T, T, A, T, C, C, A, C, T, A, T})$$
$$s^{\text{A}} = (1, 2, 7, 11, 14) \quad s^{\text{C}} = (9, 10, 12) \quad s^{\text{G}} = (3, 4) \quad s^{\text{T}} = (5, 6, 8, 13, 15)$$

Consequentemente,

$$d_1^{\text{A}} = s_2^{\text{A}} - s_1^{\text{A}} = 2 - 1 = 1$$
$$d_2^{\text{A}} = s_3^{\text{A}} - s_2^{\text{A}} = 7 - 2 = 5$$
$$d_3^{\text{A}} = s_4^{\text{A}} - s_3^{\text{A}} = 11 - 7 = 4$$
$$d_4^{\text{A}} = s_5^{\text{A}} - s_4^{\text{A}} = 14 - 11 = 3$$
$$d_5^{\text{A}} = N + s_1^{\text{A}} - s_5^{\text{A}} = 15 + 1 - 14 = 2$$

Procedendo de igual forma para os restantes nucleótidos, obtêm-se as seguintes sequências de distâncias:

$$d^{\text{A}} = (1, 5, 4, 3, 2) \quad d^{\text{C}} = (1, 2, 12) \quad d^{\text{G}} = (1, 14) \quad d^{\text{T}} = (1, 2, 5, 2, 5).$$

Note-se que os comprimentos das sequências s^x e d^x são $N^{\mathrm{A}} = 5\,, N^{\mathrm{C}} = 3\,, N^{\mathrm{G}} = 2$ e $N^{\mathrm{T}} = 5$. Para além da sequência D^x, define-se também a **sequência de distâncias global entre nucleótidos iguais**, D. Esta sequência resulta da aplicação do mapeamento da distância global entre nucleótidos iguais à sequência S, o qual faz corresponder a cada posição dessa sequência o valor da distância entre o nucleótido que se encontra nessa posição e o nucleótido igual que imediatamente o sucede. Caso não exista mais nenhum nucleótido desse tipo até ao final da sequência S, volta-se ao início dessa sequência e prossegue-se a contagem até encontrar o primeiro nucleótido desse tipo, isto é, considera-se a sequência S como sendo cíclica. A sequência D pode ser definida analiticamente por

$$D_i = \begin{cases} \min\left\{n \in V : S_i = S_{i+n}\right\}, & \text{se } \exists\, n \in V : S_i = S_{i+n} \\ N - i + S_1^{S_i}, & \text{se } \nexists\, n \in V : S_i = S_{i+n} \end{cases}, \tag{2.2}$$

onde $i = 1, 2, \ldots, N,\ V = \{1, 2, \ldots, N - i\}$ e $S_1^{S_i}$ é a posição da primeira ocorrência do nucleótido S_i na sequência S. A sequência D tem o mesmo comprimento da sequência S e a sua concretização será denotada por d.

Aplicando (2.2) ao fragmento de ADN (2.1), sabendo que $s_1^{\mathrm{A}} = 1$, $s_1^{\mathrm{C}} = 9$, $s_1^{\mathrm{G}} = 3$ e $s_1^{\mathrm{T}} = 5$, os valores de d_1 e d_4 são:

$$d_1 = \min\{1, 6, 10, 13\} = 1$$
$$d_4 = 15 - 4 + 3 = 14$$

Repetindo o mesmo procedimento para as restantes posições, a sequência de distâncias global entre nucleótidos será então

$$d = (1, 5, 1, 14, 1, 2, 4, 5, 1, 2, 3, 12, 2, 2, 5)\,. \tag{2.3}$$

O comprimento da sequência D é igual à soma dos comprimentos das sequências D^x, ou seja,

$$N = \sum_{x \in \mathcal{A}} N^x.$$

Se for conhecida a posição da primeira ocorrência de cada nucleótido na sequência simbólica de ADN, $\left[S_1^{\mathrm{A}}\ S_1^{\mathrm{C}}\ S_1^{\mathrm{G}}\ S_1^{\mathrm{T}}\right]$, e a sequência de distâncias global D, é possível reconstruir a sequência simbólica $S = (S_1, S_2, \ldots, S_N)$ determinando iterativamente cada componente

da sequência através das fórmulas

$$S_i = \underset{x\in\mathcal{A}}{\arg\min}\ S_i^x \quad \text{e} \quad S_{i+1}^x = \begin{cases} D_i + S_i^x, & x = S_i \\ S_i^x, & x \neq S_i \end{cases},$$

com $i = 1, 2, \ldots, N$. Na Tabela 2.2 é ilustrado o procedimento iterativo para a sequência concreta (2.3).

Tabela 2.2: Obtenção da sequência de ADN a partir da distância d e da posição inicial de cada nucleótido x.

i	d_i	s_i^{A}	s_i^{C}	s_i^{G}	s_i^{T}	s_i
1	1	1	9	3	5	A
2	5	2	9	3	5	A
3	1	7	9	3	5	G
4	14	7	9	4	5	G
5	1	7	9	18	5	T
6	2	7	9	18	6	T
7	4	7	9	18	8	A
8	5	11	9	18	8	T
9	1	11	9	18	13	C
10	2	11	10	18	13	C
11	3	11	12	18	13	A
12	12	14	12	18	13	C
13	2	14	24	18	13	T
14	2	14	24	18	15	A
15	5	16	24	18	15	T

A sequência de distâncias D atrás definida é uma variação da sequência de distâncias in originalmente introduzida por Nair e Mahalakshmi [35], que aqui se definirá por

$$in(i) = \begin{cases} \min\{n \in V : S_i = S_{i+n}\}, & \text{se } \exists\, n \in V : S_i = S_{i+n} \\ N - i, & \text{se } \nexists\, n \in V : S_i = S_{i+n} \end{cases}, \tag{2.4}$$

onde $i = 1, 2, \ldots, N$ e $V = \{1, 2, \ldots, N - i\}$. Neste caso, a contagem da distância termina no final da sequência de ADN, ou seja, a contagem não é cíclica.

Aplicando (2.4) ao fragmento de ADN (2.1), os valores $in(1)$ e $in(6)$ são:

$$in(1) = \min\{1, 6, 10, 13\} = 1$$
$$in(4) = 15 - 4 = 11$$

Repetindo o mesmo procedimento para as restantes posições, a sequência de distâncias global entre nucleótidos será então

$$in = (1, 5, 1, 11, 1, 2, 4, 5, 1, 2, 3, 3, 2, 1, 0)\,.$$

Nesta definição de sequência de distâncias global, o conhecimento da posição da primeira ocorrência de cada nucleótido na sequência de ADN não é suficiente para reconstruir esta última, uma vez que não é possível determinar o tipo de nucleótido que ocupa a posição N, ou seja, o mapeamento que gera a sequência in não é reversível.

Análise exploratória

Para cada espécie foram registadas individualmente as sequências de distâncias entre nucleótidos iguais, d^x. Adicionalmente, foi também registada a sequência de distâncias global entre nucleótidos iguais, d. Os elementos de ambas as sequências foram calculados de acordo com os mapeamentos atrás definidos. Na Tabela 2.3 é apresentado um sumário de estatísticas referente à sequência de distâncias global observada para cada uma das espécies. É de destacar que, embora o número máximo de distâncias observadas difira de espécie para espécie, a média das distâncias para cada espécie é sempre quatro, uma vez que

$$\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i \Leftrightarrow \bar{d} = \frac{1}{N}\sum_{i=1}^{N}\left(d_i^{\mathrm{A}} + d_i^{\mathrm{C}} + d_i^{\mathrm{G}} + d_i^{\mathrm{T}}\right) \Leftrightarrow \bar{d} = \frac{4N}{N} \Leftrightarrow \bar{d} = 4\,.$$

No que diz respeito à variabilidade, nos organismos procariotas os maiores valores foram observados nas espécies *Mj* e *Cb*, enquanto que no caso dos organismos eucariotas os maiores valores foram observados nas espécies *Dd* e *Pl*.

Tabela 2.3: Sumário de estatísticas: bactérias, plantas, animais, protozoários e fungos.

Espécie	Min	1ºQ.	Med	3ºQ	Máx	Média	Desv.padrão
Ap	1	1	3	5	99	4	3.77
Hr	1	2	3	5	127	4	3.96
Mj	1	1	2	5	180	4	5.04
Pf	1	1	3	5	93	4	4.10
Tk	1	1	3	5	88	4	3.74
Ba	1	1	3	5	110	4	4.18
Bs	1	1	3	5	144	4	3.80
Ct	1	1	3	5	122	4	3.85
Cb	1	1	2	5	156	4	5.06
Dv	1	2	3	5	97	4	3.80
Ec	1	1	3	5	83	4	3.60
Hi	1	1	3	5	156	4	3.96
Hp	1	1	2	5	217	4	4.21
Mg	1	1	3	5	132	4	4.36
Pa	1	2	3	5	134	4	3.94
Sa	1	1	3	5	121	4	4.16
Sm	1	1	3	5	87	4	4.00
St	1	1	3	5	94	4	3.88
At	1	1	3	5	669	4	4.35
Os	1	1	3	5	1003	4	4.29
Po	1	1	2	5	1357	4	4.86
Vv	1	1	2	5	1866	4	4.81
Bt	1	1	3	5	1127	4	4.14
Cf	1	1	2	5	951	4	4.51
Eq	1	1	3	5	1125	4	4.21
Gg	1	1	3	5	1134	4	4.13
Am	1	1	2	5	902	4	5.26
Dm	1	1	3	5	1127	4	4.15
Mu	1	1	3	5	2691	4	4.58
Ce	1	1	2	5	888	4	4.38
Rn	1	1	3	5	2358	4	4.44
Xt	1	1	3	5	912	4	4.06
Hs	1	1	3	5	1819	4	4.34
Mm	1	1	3	5	2897	4	4.31
Pt	1	1	3	5	1541	4	4.31
Dr	1	1	3	5	2256	4	4.96
Fu	1	1	3	5	1010	4	4.21
Oa	1	1	3	5	1022	4	4.17
Dd	1	1	2	4	606	4	7.49

continua na página seguinte

Tabela 2.3 – continuação da página anterior

Espécie	Min	1ºQ.	Med	3ºQ	Máx	Média	Desv.padrão
Li	1	2	3	5	643	4	4.26
Pl	1	1	2	4	723	4	7.16
Tb	1	1	3	5	762	4	4.25
Ca	1	1	3	5	157	4	4.60
Nc	1	1	3	5	362	4	4.13
Sc	1	1	3	5	398	4	4.10
Sp	1	1	3	5	212	4	4.13

Na Figura 2.1 são apresentadas as caixas de bigodes[1] para as espécies em estudo. A distribuição empírica de cada uma das espécies é assimétrica positiva. Em relação à variabilidade dos dados, verifica-se uma maior concentração dos 50% dos valores mais centrais nas espécies *Hr*, *Dv*, *Pa* e *Li*. De salientar que 75% das distâncias são inferiores ou iguais a 5. Em todas as espécies verifica-se a presença de um número elevado de observações atípicas.

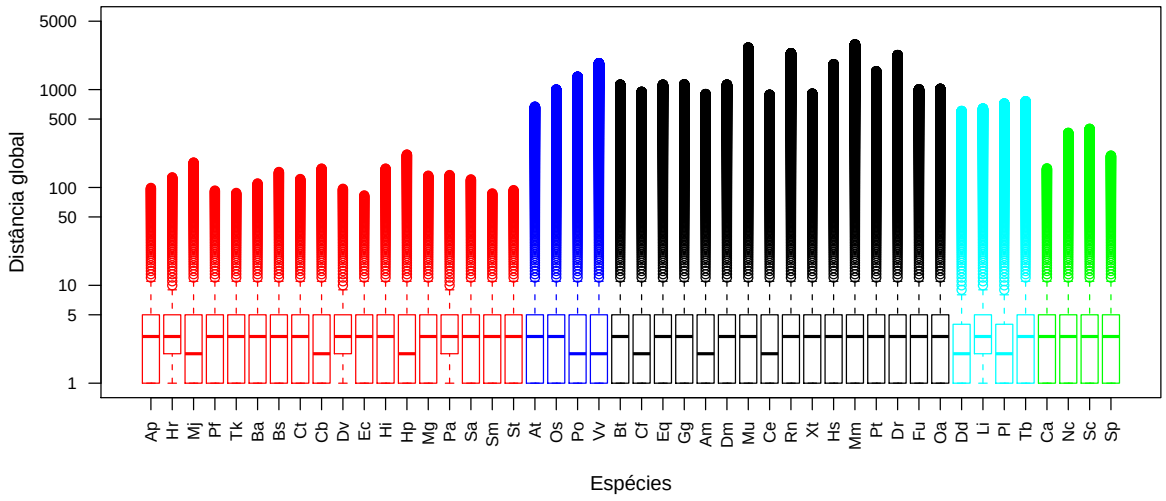


Figura 2.1: Caixas de bigodes das espécies da Tabela 2.1: bactérias, plantas, animais, protozoários e fungos.

[1] A caixa de bigodes permite a obtenção de informação sobre a localização central e variabilidade dos dados (altura da caixa reduzida e bigodes mais pequenos significam que existe uma maior concentração dos dados) e também sobre a assimetria da distribuição e a existência de observações atípicas.

2.2 Distribuição das distâncias

Admitindo que as sequências de nucleótidos foram geradas por um processo aleatório independente e identicamente distribuído (i.i.d.), uma distribuição teórica admissível para modelar as distâncias empíricas entre nucleótidos iguais seria dada pela distribuição geométrica. Esta possibilidade é reforçada pelo facto de a distribuição geométrica indicar a probabilidade de ocorrer um sucesso após um determinado número de provas, independentemente dos resultados das provas anteriores, propriedade que pode ser utilizada directamente para o cálculo da probabilidade da existência de uma distância k entre dois nucleótidos do mesmo tipo. Com base na lei da probabilidade total, uma possível distribuição teórica para a sequência de distâncias global, pode ser definida por uma mistura daquelas quatro geométricas, as quais correspondem às distribuições das distâncias entre os nucleótidos A, C, G e T.

Distribuição geométrica

Designam-se por provas de Bernoulli de parâmetro p $(0 < p < 1)$ uma sucessão de provas independentes realizadas nas mesmas condições, tendo cada prova apenas dois resultados possíveis, o *sucesso* e o *insucesso*. Em cada prova de Bernoulli a probabilidade de sucesso é constante e igual a p. Se a variável aleatória (v.a.) Y designar o número de provas de Bernoulli até à ocorrência do primeiro sucesso, tem-se que Y é uma v.a. discreta, tomando um número finito ou infinito numerável de valores com função massa de probabilidade

$$P(Y = y) = p\,(1-p)^{y-1},\ y = 1, 2, 3, \ldots. \tag{2.5}$$

Nestas condições, diz-se que Y tem uma **distribuição geométrica**[2] de parâmetro p, $Y \sim Geom(p)$. A sua função de distribuição é, por definição,

$$F(y) = P(Y \leq y) = \begin{cases} 0, & y < 1 \\ \sum\limits_{i=1}^{[y]} (1-p)^{i-1} p, & y \geq 1 \end{cases},$$

onde $[y]$ representa a parte inteira do número real y. Atendendo a que $F(y)$, para valores de $y \geq 1$, representa a sucessão das somas parciais de uma série geométrica de razão $1-p$,

[2] Esta distribuição é, por vezes, chamada de distribuição do tempo de espera por um sucesso.

com $|1-p|<1$, obtém-se[3],

$$F(y)=\begin{cases} 0\,, & y<1 \\ 1-(1-p)^y, & y\geq 1 \end{cases}. \tag{2.6}$$

O valor esperado e a variância de Y são, respectivamente,

$$E\,(Y)=\frac{1}{p} \quad \text{e} \quad Var\,(Y)=\frac{1-p}{p^2}\,. \tag{2.7}$$

De facto,

$$E\,(Y)=\sum_{i=1}^{+\infty} i(1-p)^{i-1}p=p\sum_{i=1}^{+\infty} i(1-p)^{i-1}=-p\sum_{i=1}^{+\infty}\frac{\mathrm{d}}{\mathrm{d}\,p}\left[(1-p)^i\right].$$

Sabendo que uma série de potências pode ser derivável termo a termo dentro do seu intervalo de convergência, obtém-se

$$E\,(Y)=-p\,\frac{\mathrm{d}}{\mathrm{d}\,p}\left(\sum_{i=1}^{+\infty}(1-p)^i\right)=-p\,\frac{\mathrm{d}}{\mathrm{d}\,p}\left(\frac{1-p}{1-(1-p)}\right)=-p\left(\frac{-1}{p^2}\right)=\frac{1}{p}.$$

De igual modo, prova-se que

$$E\left(Y^2\right)=\frac{2-p}{p^2}. \tag{2.8}$$

Consequentemente,

$$Var\,(Y)=E\left(Y^2\right)-\left(E\,(Y)\right)^2=\frac{1-p}{p^2}.$$

Distribuição das distâncias em sequências aleatórias com símbolos independentes

Como foi visto na Secção 2.1, a sequência de distâncias entre nucleótidos iguais

$$D^x=\left(D_1^x, D_2^x, \ldots, D_{N^x-1}^x, D_{N^x}^x\right),$$

[3] Seja $S_n=\sum_{i=0}^{n} u^i$ o termo geral da sucessão das somas parciais de uma série geométrica de razão $|u|<1$. Então $S_n=\frac{1-u^{n+1}}{1-u}$ e $\lim_{n\to\infty} Sn=\frac{1}{1-u}$.

é uma sequência numérica, tomando valores $1, 2, \ldots$, onde D_i^x representa a i-ésima distância entre nucleótidos do tipo x, com $x \in \mathcal{A} = \{\text{A,C,G,T}\}$. Designem-se por $p^{\text{A}}, p^{\text{C}}, p^{\text{G}}$ e p^T as probabilidades de ocorrência dos nucleótidos A, C, G e T, respectivamente, numa posição qualquer na sequência. Admitindo que as sequências de nucleótidos foram geradas por um processo aleatório i.i.d., então, para cada x,

$$D^x \sim Geom(p^x).$$

Atendendo a (2.5), a função massa de probabilidade, f^x, vem na forma

$$f^x(k) = P\left(D^x = k\right) = P\left(D = k|x\right) = p^x(1-p^x)^{k-1},\ k = 1, 2, \ldots. \tag{2.9}$$

Atendendo a (2.6), a função de distribuição, F^x, vem na forma

$$F^x(k) = P\left(D^x \leq k\right) = 1 - (1-p^x)^k,\ k \geq 1.$$

Atendendo a (2.7), o valor esperado e a variância são dados por

$$E\left(D^x\right) = \frac{1}{p^x} \qquad Var\left(D^x\right) = \frac{1-p^x}{\left(p^x\right)^2}. \tag{2.10}$$

Para estimar o parâmetro p^x da distribuição geométrica foi usada a frequência relativa

$$\hat{p^x} = \frac{N^x}{N}, \tag{2.11}$$

onde N^x é o comprimento da sequência D^x (coincidente com o número de nucleótidos do tipo x existentes na sequência original de ADN) e N o comprimento da sequência D.

A distância global

$$D \sim Modelo(p),\ p = \left(p^{\text{A}}, p^{\text{C}}, p^{\text{G}}, p^{\text{T}}\right), \tag{2.12}$$

cuja função massa de probabilidade, f, atendendo à lei da probabilidade total, vem na forma

$$f(k) = P\left(D = k\right) = \sum_{x \in \mathcal{A}} P\left(D = k|x\right) p^x = \sum_{x \in \mathcal{A}} p^x(1-p^x)^{k-1}p^x,\ k = 1, 2, \ldots. \tag{2.13}$$

O valor esperado de D, atendendo a (2.10), vem igual a

$$E\left(D\right) = \sum_{i=1}^{+\infty} i \sum_{x \in \mathcal{A}} p^x(1-p^x)^{i-1}p^x = \sum_{x \in \mathcal{A}} p^x E\left(D^x\right) = 4.$$

Para o cálculo da variância de D, atendendo a (2.8), tem-se que

$$E\left(D^2\right) = \sum_{x\in\mathcal{A}} p^x E\left[\left(D^x\right)^2\right] = \sum_{x\in\mathcal{A}} p^x \frac{2-p^x}{\left(p^x\right)^2} = \sum_{x\in\mathcal{A}} \left(\frac{2}{p^x}\right) - 4,$$

donde,

$$Var\left(D\right) = E\left(D^2\right) - \left(E\left(D\right)\right)^2 = \sum_{x\in\mathcal{A}} \left(\frac{2}{p^x}\right) - 20 = 2\sum_{x\in\mathcal{A}} E\left(D^x\right) - 20\,.$$

2.3 Distribuição empírica *vs* Distribuição modelo

O vector da sequência de distâncias global de cada uma das espécies foi reduzido a uma tabela de frequências

y	1	2	$\cdots$	L
f_y	f_1	f_2	$\cdots$	f_L

,

onde f_i representa a frequência absoluta da distância i e traduz a distribuição empírica das distâncias de cada uma das espécies.

Na Figura 2.2 são apresentadas graficamente[4] as distribuições geométricas (2.9) e as distribuições empíricas para as sequências de distâncias entre nucleótidos iguais da espécie *St.* Estes gráficos foram obtidos através da representação das distâncias observadas entre nucleótidos, d^x, e as curvas (linhas azuis) foram obtidas a partir da distribuição geométrica (2.9), com parâmetros constantes estimados através de (2.11), concretamente

$$D^A \sim Geom(0.3021) \qquad D^C \sim Geom(0.1982)$$
$$D^G \sim Geom(0.1967) \qquad D^T \sim Geom(0.3030)\,.$$

[4] Por uma questão de melhor visualização, apenas são apresentadas as vinte e cinco primeiras distâncias e é usada uma linha contínua para representar a distribuição geométrica, em vez de uma sequência de pontos.

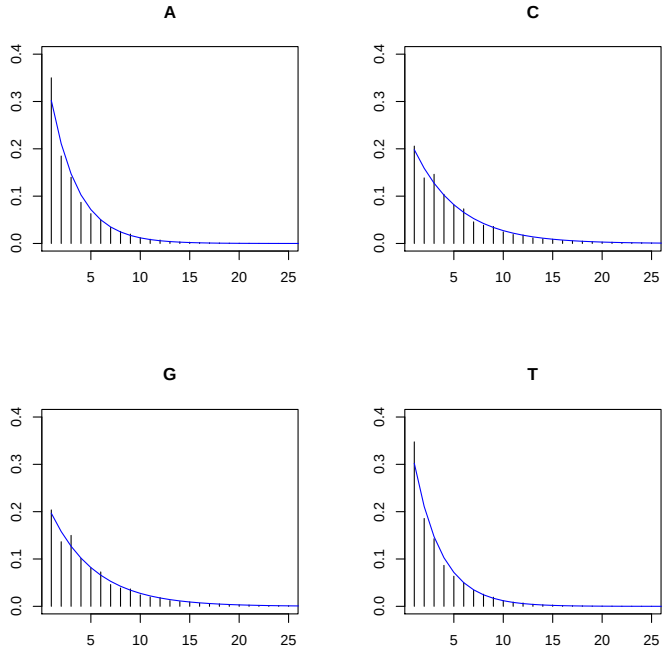


Figura 2.2: Distribuição empírica (gráfico de barras) e distribuição da sequência de distâncias entre nucleótidos iguais d^x, para a espécie *St* (conjunto de pontos contidos na curva representada a azul).

Na Figura 2.3 é apresentado o gráfico de barras para a mesma espécie, *St*, mas considerando agora a sequência de distâncias global d. A curva (linha azul) foi obtida a partir da distribuição modelo (2.13), com vector de parâmetros de componentes estimadas por (2.11), tendo resultado

$$\hat{p} = (0.3021, 0.1982, 0.1967, 0.3030)\,.$$

As representações gráficas da distribuição empírica de cada espécie e a forma da distribuição modelo proposta por [2] incentivam a procura por uma distribuição teórica que melhor se ajuste à distribuição empírica, sugerindo identificar a melhor mistura finita de distribuições geométricas ajustada. Abordar-se-á este assunto no Capítulo 4.

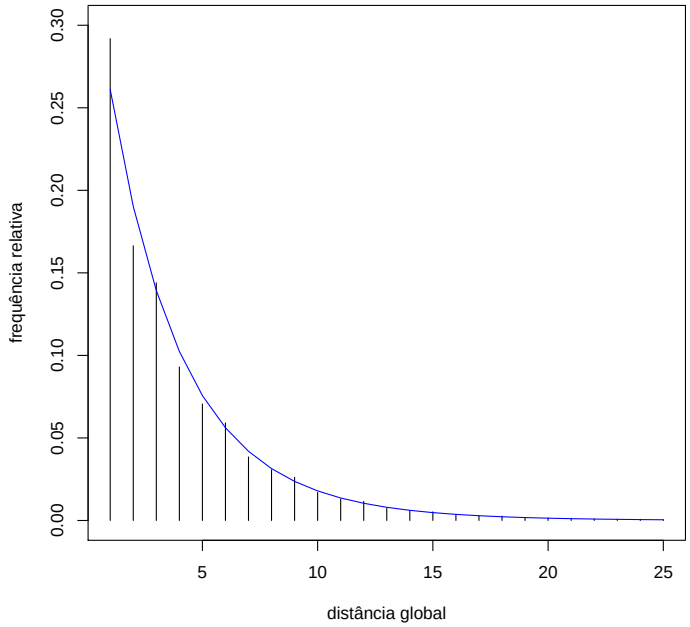


Figura 2.3: Distribuição empírica (gráfico de barras) e distribuição da sequência de distâncias global d, para a espécie St (conjunto de pontos contidos na curva representada a azul).

2.4 A matriz dos erros relativos

É na diferença entre a distribuição empírica e a frequência esperada sob a hipótese de independência que se manifesta a selecção natural na evolução dos organismos [16]. Adicionalmente, uma forma de melhor salientar a diferença entre a distribuição empírica de cada distância e a correspondente distribuição modelo, é analisar o erro de uma relativamente à outra, por comparação das respectivas distribuições [2]. Na análise que se irá efectuar definiu-se o erro relativo da distância k para a espécie i como sendo

$$\delta_{ik} = \begin{cases} \frac{f_0^i(k) - f^i(k)}{f_0^i(k)}, & f_0^i(k) \neq 0 \\ 0\,, & f_0^i(k) = 0 \end{cases}, \tag{2.14}$$

onde $f_0^i(k)$ é a frequência relativa observada da distância k na i-ésima espécie[5] e $f^i(k)$ a função massa de probabilidade associada à distribuição modelo, isto é, (2.13), da i-ésima

[5] No caso das espécies com o genoma mais curto (ver Tabela 2.3), foi atribuído o valor zero ao erro relativo das distâncias para as quais não se registaram ocorrências.

espécie. Os erros relativos podem ser apresentados sob a forma de uma matriz

$$\Delta_{n\times p} = \begin{bmatrix} \delta_{11} & \delta_{12} & \dots & \delta_{1p} \\ \delta_{21} & \delta_{22} & \dots & \delta_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{n1} & \delta_{n2} & \dots & \delta_{np} \end{bmatrix}, \tag{2.15}$$

onde n representa o número de espécies (indivíduos) e p o número de valores possíveis tomados para a variável distância. Os valores dos erros relativos correspondentes à i-ésima espécie são representados por um vector linha $\delta_i^T = (\delta_{i1}, \delta_{i2}, \cdots, \delta_{ip})$. Por outro lado, o vector coluna $\delta_j = (\delta_{1j}, \delta_{2j}, \cdots, \delta_{nj})^T$ contém o erro relativo associado à distância j para cada espécie. O vector linha pode ser representado como um ponto no espaço $\mathbb{R}^p$ e o vector coluna como um ponto no espaço $\mathbb{R}^n$. Deste ponto de vista, a matriz $\Delta_{n\times p}$ é representável como uma nuvem de pontos no espaço $\mathbb{R}^p$ ou no espaço $\mathbb{R}^n$. Neste trabalho considerar-se-á a matriz Δ como uma nuvem de p pontos de coordenadas δ_j de $\mathbb{R}^n$. Um dos objectivos será projectar essa nuvem de pontos num espaço de menor dimensão, de forma a tentar deduzir propriedades estatísticas associadas às distâncias nas diferentes espécies.

Do mapeamento da distância entre nucleótidos da sequência de ADN das diferentes espécies analisadas resultou um elevado número de valores observados distintos para a variável distância. Os resultados obtidos em [2] demonstram que é possível obter informação sobre o genoma limitando a análise às cem primeiras distâncias, podendo estas serem interpretadas como uma assinatura genética. Assim, para a análise realizada na presente dissertação, considera-se a matriz dos erros relativos constituída pelas cem primeiras distâncias de todas as espécies da Tabela 2.1, isto é, $\Delta_{46\times 100}$.

Capítulo 3

Análise Multivariada - Comparação de Espécies

Classificação hierárquica, classificação não-hierárquica e análise de componentes principais são técnicas estatísticas de análise multivariada que é possível utilizar para melhor evidenciar a relação entre os indivíduos. Estas técnicas possuem fundamentos teóricos diferentes, podendo ser aplicadas independentemente. Com a classificação hierárquica é possível construir, utilizando todas as variáveis disponíveis, agrupamentos entre os indivíduos segundo um grau de similaridade que apresentam entre si e de acordo com um critério pré-definido. É possível também representar esses grupos no espaço bidimensional através de um dendrograma (gráfico em árvore). Com a classificação não-hierárquica, os indivíduos são distribuídos por k grupos especificados inicialmente, de acordo também com uma medida de similaridade. Com a análise de componentes principais pretende-se a redução da dimensionalidade original das variáveis sem perdas significativas de informação.
A aplicação destas técnicas incidirá sobre a matriz dos erros relativos $\Delta_{46\times100}$ da Secção 2.4.

3.1 Classificação hierárquica e não-hierárquica

O objectivo de uma classificação é reunir os objectos de uma amostra em grupos, satisfazendo a condição de que objectos que pertençam a um mesmo grupo sejam similares e objectos de grupos diferentes sejam dissimilares, face a um conjunto de variáveis. A ideia base é maximizar a homogeneidade de objectos dentro de grupos, ao mesmo tempo

que se maximiza a heterogeneidade entre grupos [17]. O critério em que assenta a decisão de similaridade ou dissimilaridade entre dois indivíduos baseia-se numa medida de proximidade. No caso do agrupamento de indivíduos, a medida de proximidade será uma medida de distância, pelo que a similaridade entre dois indivíduos será tanto maior quanto menor for a distância entre eles. Para um agrupamento de variáveis, a medida a usar será uma medida de associação, eventualmente baseada em coeficientes de correlação[1], pelo que quanto maior for o valor desta medida maior será a similaridade entre as duas variáveis [23].
A principal diferença entre a classificação hierárquica e a não-hierárquica reside no facto de que, na primeira, quando um indivíduo é atribuído a um grupo esse indivíduo não poderá transitar para outro grupo, enquanto que na segunda a atribuição de um indivíduo a um grupo pode ser alterada durante a execução do algoritmo.

Neste trabalho apenas será feito o agrupamento de indivíduos, ou seja, o agrupamento de espécies.

3.1.1 Medidas de proximidade

Como foi referido anteriormente, na análise de agrupamentos de indivíduos a similaridade ou dissimilaridade entre dois deles pode ser expressa como uma função de distância entre os dois pontos do espaço p-dimensional que os representam. Com base nesta distância, é então calculada a distância de cada ponto a todos os outros pontos, constituindo-se assim uma matriz de distâncias, d, designada por *matriz de proximidade*, a qual descreve a proximidade entre todos os indivíduos. A matriz d é uma matriz quadrada de ordem n, simétrica e com todos os elementos da diagonal principal nulos.

$$d = \begin{bmatrix} 0 & d_{(2,1)} & \dots & d_{(n,1)} \\ d_{(2,1)} & 0 & \dots & d_{(n,2)} \\ \vdots & \vdots & \ddots & \vdots \\ d_{(n,1)} & d_{(n,2)} & \dots & 0 \end{bmatrix}, \tag{3.1}$$

onde $d_{(i,j)}$ corresponde ao valor da distância entre os indivíduos de índices i e j. Esta medida de distância satisfaz as seguintes propriedades:

[1] Exemplos de coeficientes de correlação: Pearson, Spearman e Kendall.

- $d_{(i,j)} \geq 0\,, \forall\, i,j = 1,\ldots,n$;
- $d_{(i,i)} = 0\,, \forall\, i = 1,\ldots,n$;
- $d_{(i,j)} = d_{(j,i)}\,, \forall\, i,j = 1,\ldots,n$;
- $d_{(i,j)} \leq d_{(i,k)} + d_{(j,k)}\,, \forall\, i,j,k = 1,\ldots,n$.

As medidas de distâncias que se apresentam a seguir são as actualmente suportadas pela função *dist()* do R. Sejam $\delta_i^T = (\delta_{i1}, \delta_{i2}, \ldots \delta_{ip})$ e $\delta_j^T = (\delta_{j1}, \delta_{j2}, \ldots \delta_{jp})$ os vectores do i-ésimo e do j-ésimo indivíduo de uma matriz de dados de dimensão $(n \times p)$.

A **distância de *Minkowski*** é definida por

$$d_{(i,j)} = \sqrt[m]{\sum_{k=1}^{p} |\delta_{ik} - \delta_{jk}|^m}\,,\ m \in \mathbb{N}\,.$$

A **distância absoluta** (também conhecida por ***Manhattan*** ou ***city-block***) é um caso particular ($m = 1$) da distância de *Minkowski*, sendo definida por

$$d_{(i,j)} = \sum_{k=1}^{p} |\delta_{ik} - \delta_{jk}|\,.$$

A **distância euclidiana** é também um caso particular ($m = 2$) da distância de *Minkowski*, estando definida por

$$d_{(i,j)} = \sqrt{\sum_{k=1}^{p} (\delta_{ik} - \delta_{jk})^2}\,. \tag{3.2}$$

A **distância do máximo** ou l^∞ é mais um caso particular ($m \to \infty$) da distância de *Minkowski*, estando definida por

$$d_{(i,j)} = \max_k |\delta_{ik} - \delta_{jk}|\,.$$

A **distância de Canberra** para variáveis que apenas possam tomar valores não-negativos é definida por

$$d_{(i,j)} = \sum_{k=1}^{p} \frac{|\delta_{ik} - \delta_{jk}|}{|\delta_{ik} + \delta_{jk}|}.$$

A escolha de qual a função de distância entre indivíduos a utilizar depende do tipo de dados. Existe uma apetência natural pela utilização da distância euclidiana, uma vez que a mesma é habitualmente utilizada para medir a distância entre indivíduos no espaço bidimensional ou tridimensional.

3.1.2 Métodos hierárquicos

Os diferentes métodos de classificação hierárquica podem ser agrupados em duas categorias: os métodos ascendentes ou aglomerativos (*Bottom-Up*) e os métodos descendentes ou divisivos (*Top-Down*).

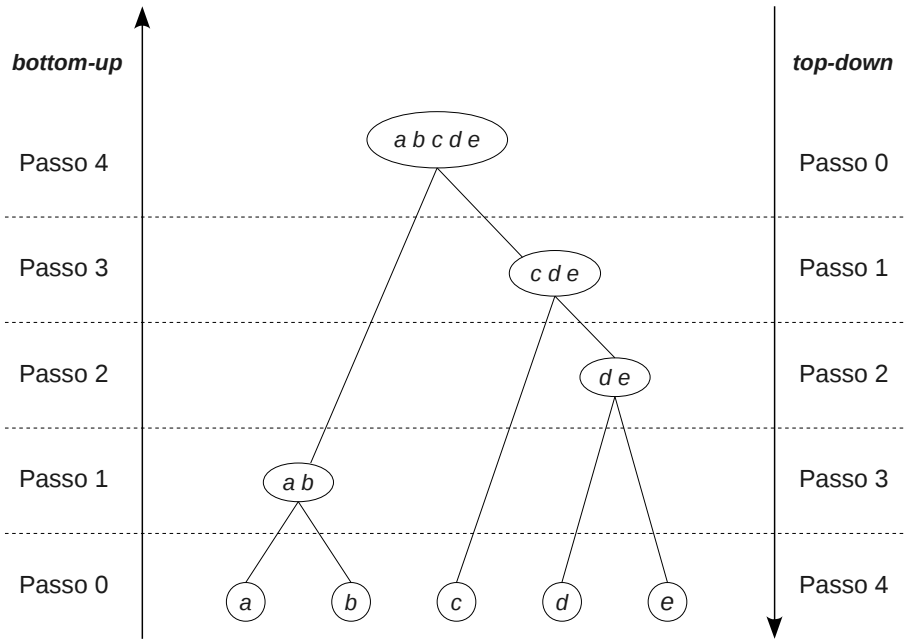


Figura 3.1: Agrupamento hierárquico aglomerativo (*bottom-up*) e divisivo (*top-down*).

Nos **métodos aglomerativos**, a cada indivíduo corresponde inicialmente um grupo do qual ele é o único elemento. Em cada passo do algoritmo são fundidos os dois grupos de indivíduos mais similares, constituindo-se assim um novo agrupamento. Geralmente não existe nenhum critério de paragem específico e esta operação é repetida até que todos os indivíduos estejam reunidos num único agrupamento. Note-se que, em cada passo, os indivíduos dos grupos que se agregam são cada vez mais dissimilares. O resultado deste procedimento designa-se por classificação hierárquica ascendente. Os **métodos divisivos** funcionam de maneira oposta à dos métodos aglomerativos, colocando inicialmente todos os indivíduos no mesmo grupo. Em cada passo do algoritmo, este grupo será dividido em dois outros grupos que contêm os objectos mais distintos, parando apenas quando a cada grupo corresponder apenas um indivíduo. O resultado deste procedimento designa-se por classificação hierárquica descendente. Nos métodos hierárquicos só se agregam ou dividem dois grupos de cada vez e, uma vez formado um grupo, este já não se divide. Os métodos

aglomerativos são os mais usados devido à sua eficiência computacional, existindo, por isso, menos implementações dos métodos divisivos [18].

Algoritmo aglomerativo

(a) Início: n grupos, cada um com apenas um indivíduo;

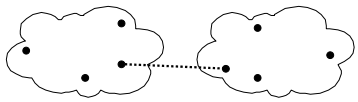
(b) Calcular a matriz de proximidade (3.1) de ordem n;

(c) Agrupar num só grupo os dois grupos cuja distância entre si é a menor;

(d) Criar uma nova matriz de proximidade de ordem $(n-1)$. A distância entre grupos com mais de um indivíduo será calculada de acordo com um critério de agregação, que pode ser, por exemplo, um dos seguintes:

1. **ligação única** (*single linkage*) - método do vizinho mais próximo: considerando-se todos os pares possíveis de membros dos dois grupos em que os elementos de cada par não pertencem ambos ao mesmo grupo, a distância entre dois grupos é a menor distância verificada entre os dois elementos de todos esses pares:

$$d_{\text{grupo1, grupo2}} = \min\left\{d_{(i,j)} : i \in \text{grupo1}, j \in \text{grupo2}\right\}$$

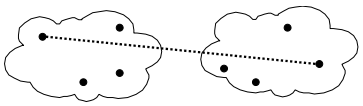


Este método tende a produzir grupos com efeito de ligação, com indivíduos que podem estar muito distantes entre si, mas pertencendo a um mesmo grupo[2].

2. **ligação completa** (*complete linkage*) - método do vizinho mais afastado: considerando-se todos os pares possíveis de membros dos dois grupos em que os elementos de cada par não pertencem ambos ao mesmo grupo, a distância entre dois grupos é a maior distância verificada entre os dois elementos de todos esses pares:

$$d_{\text{grupo1, grupo2}} = \max\left\{d_{(i,j)} : i \in \text{grupo1}, j \in \text{grupo2}\right\}$$

[2] Basta que exista um elemento de um grupo próximo de um único elemento do outro grupo para que estes sejam atraídos, independentemente de haver outros elementos dos grupos que estejam muito distantes entre si.

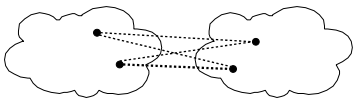


Este método tem uma forte tendência para produzir grupos compactos com diâmetros aproximadamente iguais dado que, em cada passo, tende a minimizar as distâncias intra-grupo.

3. **ligação média** (*average linkage*) - a distância entre dois grupos é obtida determinando-se a distância entre os elementos de cada par de elementos dos dois grupos, em que os elementos de cada par pertencem a grupos diferentes, e calculando-se, em seguida, o valor médio dessas distâncias.

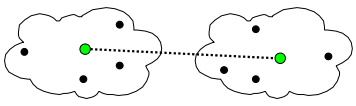
$$d_{\text{grupo1, grupo2}} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} d_{(i,j)} \, ,$$

onde n_1 e n_2 representam o número de elementos do grupo1 e grupo2, respectivamente.



Este método tende a juntar grupos com variâncias reduzidas e a produzir grupos com a mesma variância. Uma vez que considera todos os elementos do grupo, em vez de um único elemento, tende a ser menos influenciado por valores extremos quando comparado com outros métodos.

4. **método do centróide** (*centroid method*) - o centróide é o ponto médio de um grupo de pontos. É frequente o centróide não coincidir com um dos pontos do grupo. Neste método a distância entre dois grupos é definida como a distância entre os respectivos centróides.



Uma desvantagem deste método verifica-se no caso dos dois grupos possuírem dimensões muito diferentes. O centróide do novo agrupamento estará mais próximo do grupo de maior dimensão, pelo que as características do grupo de menor dimensão tenderão a perder-se.

5. **método de Ward** (*minimum variance method*) A distância entre dois grupos [22] é dada pelo quadrado da distância entre os vectores médios dos dois grupos dividido pela soma dos inversos aritméticos do número de elementos de cada grupo, n_1 e n_2, ou seja,

$$d_{grupo1,grupo2} = \frac{\left\|\bar{\delta}_{grupo1,j} - \bar{\delta}_{grupo2,j}\right\|^2}{\frac{1}{n_1} + \frac{1}{n_2}} = \frac{n_1 n_2}{n_1 + n_2} d^2_{(\bar{\delta}_{grupo1},\, \bar{\delta}_{grupo2})}\,.$$

   Este método tende a produzir grupos com um número aproximadamente igual de indivíduos.

(e) Repetir os passos (c) e (d) até todos os indivíduos estarem juntos num único grupo.

Dendrograma

Baseado numa matriz de proximidade e num critério de agregação, o processo de agrupamento hierárquico pode ser representado por um dendrograma (diagrama de árvore hierárquico). No eixo horizontal são colocados os indivíduos e no eixo vertical o índice de similaridade.

A interpretação de um dendrograma assenta no pressuposto básico de que, para cada ramo, quanto menor for a distância (vertical) entre dois pontos, maior será a semelhança entre os indivíduos correspondentes ou, por outras palavras, os valores das variáveis que modelam esses indivíduos serão mais semelhantes entre si. Isso significa que essas variáveis estarão mais próximas no espaço multidimensional. Assim sendo, os dendrogramas revelam especial utilidade na visualização de indivíduos representados por pontos em espaços de dimensão superior a três, para os quais a representação gráfica apresenta manifestas dificuldades. Os ramos da árvore fornecem a ordem das $(n-1)$ ligações, em que o primeiro nível representa a primeira ligação, o segundo nível a segunda ligação, e assim sucessivamente, até que todos os ramos se juntem, numa hierarquização baseada no grau de similaridade entre indivíduos, colocando em ramos adjacentes os grupos que possuem maior similaridade entre si. Uma outra característica útil dos dendrogramas é a possibilidade de, através da inspecção dos mesmos, ser possível sugerir o número de grupos a formar a partir da determinação (subjectiva ou analítica) do ponto de corte do dendrograma.

3.1.3 Métodos não-hierárquicos

Um dos métodos não-hierárquicos mais utilizado é o algoritmo *K-means*. Existem muitas variantes deste algoritmo. Nesta secção será abordada uma das variações mais frequentemente utilizadas, conhecida por algoritmo de Lloyd [24]. O processo de formação dos grupos é feito iterativamente estabelecendo-se, como parâmetros do algoritmo, o número de grupos pretendidos, que se denominará por k, e para cada um desses grupos um centróide inicial (uma semente). Os centróides iniciais podem ser definidos pelo utilizador ou determinados aleatoriamente. Em cada uma das iterações, e com base numa medida de proximidade, é associado um (novo) agrupamento de indivíduos a cada um dos centróides determinados na iteração anterior. De seguida, procede-se à actualização desses centróides de acordo com os indivíduos a ele associados na iteração corrente. O objectivo do algoritmo é formar os grupos de modo a que, para cada grupo, se atinja o menor erro interno entre os indivíduos que o compõem e os centróides respectivos. Define-se o erro interno como sendo

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \|c_i - x\|^2 = \sum_{i=1}^{k} \sum_{x \in C_i} d^2_{(c_i,x)} \,,$$

onde c_i representa o centróide do i-ésimo grupo C_i e k o número total de grupos.
O ciclo de iterações termina quando nenhum dos indivíduos muda de grupo, ou seja, quando deixam de ocorrer variações dos centróides. Esta situação corresponde a um mínimo local do erro E, mas não necessariamente a um mínimo global. Isto acontece porque o algoritmo não vai incidir sobre todos os k agrupamentos possíveis mas sim apenas sobre aqueles que correspondem aos centróides iniciais especificados [41].
Frequentemente, e uma vez que a maior parte da convergência ocorre nas primeiras iterações, utiliza-se como critério de paragem uma condição menos rígida como critério de convergência, tal como, por exemplo, a não ultrapassagem de uma percentagem máxima de mudança de indivíduos de um grupo para outro.

Algoritmo *K-means*

(1) Seleccionar k pontos como centróides iniciais (ou sementes);

(2) Formar k grupos, associando cada indivíduo ao centróide mais próximo;

(3) Actualizar o centróide de cada grupo com base nos indivíduos correntes desse grupo;

(4) Se o critério de paragem não for satisfeito, voltar ao passo 2 e repetir o processo, caso contrário, terminar.

3.1.4 Resultados experimentais

Métodos hierárquicos

A partir da matriz dos erros relativos (ver Secção 2.4) construiu-se a matriz de proximidade usando a distância euclidiana como medida de similaridade. Na Figura 3.2 e na Figura 3.3 encontram-se representados os dendrogramas correspondentes, usando como método de agregação das espécies a ligação completa e o método de Ward, respectivamente.

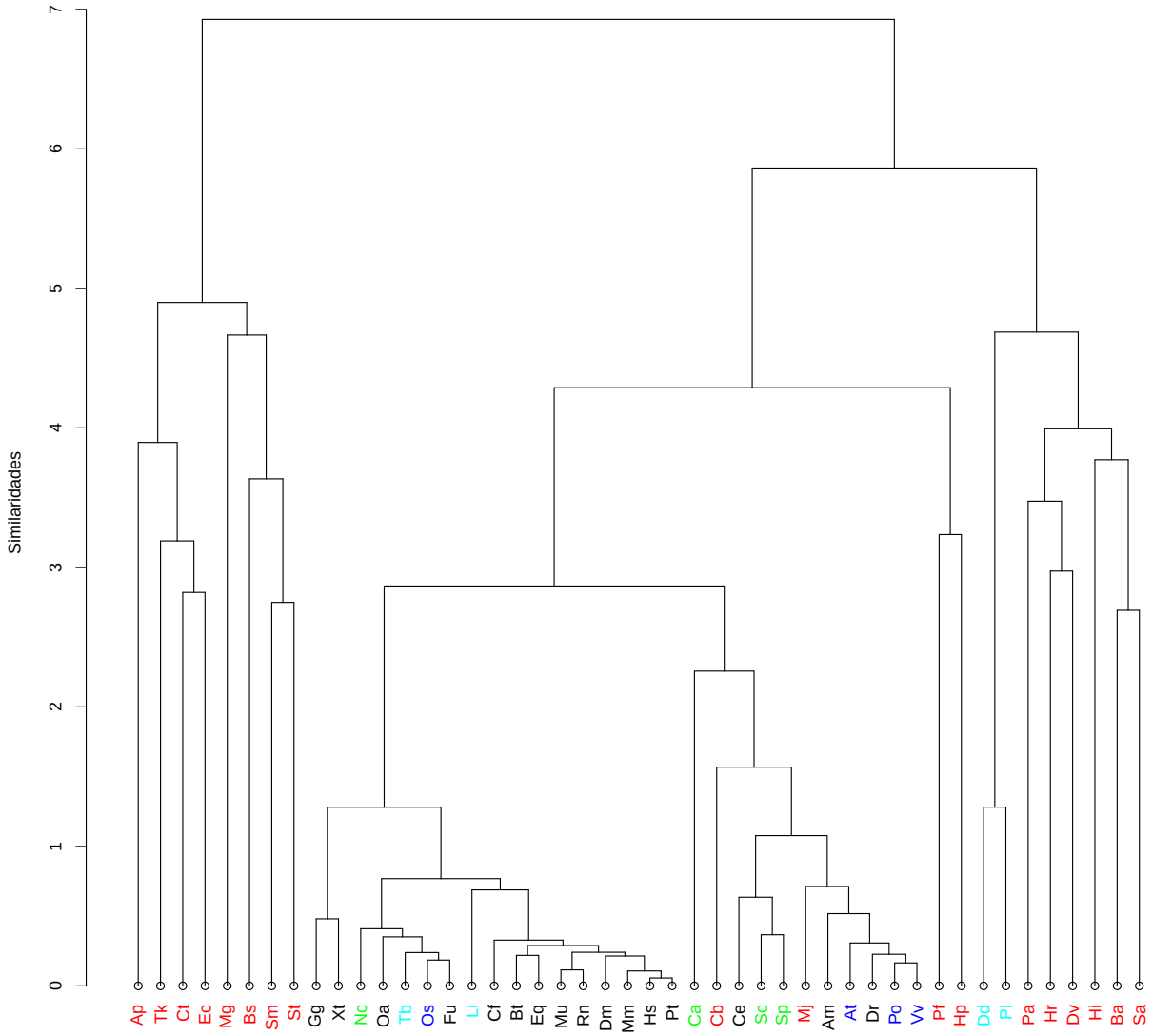


Figura 3.2: Dendrograma usando a distância euclidiana como medida de similaridade e a ligação completa como critério de agregação.

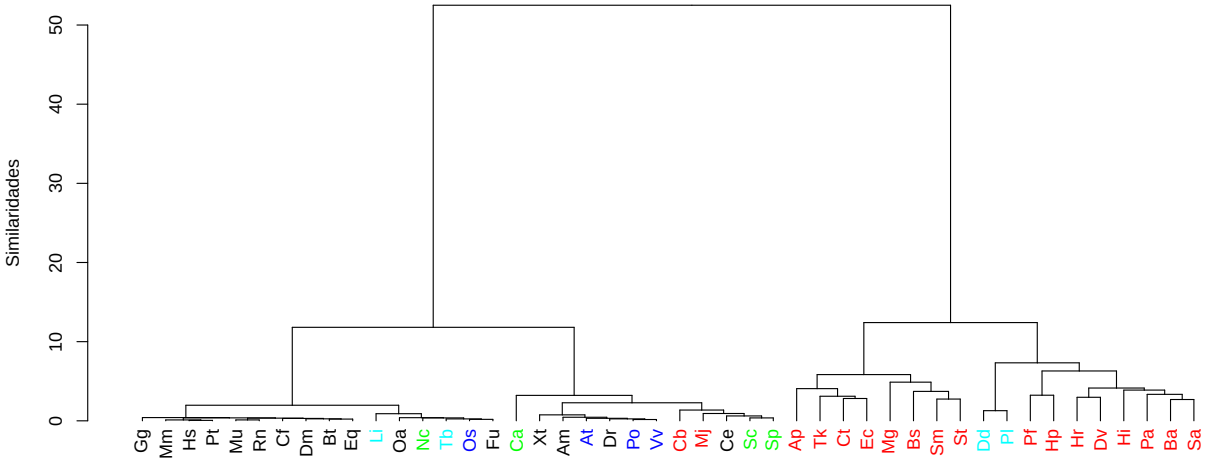


Figura 3.3: Dendrograma usando a distância euclidiana como medida de similaridade e o método de Ward como critério de agregação.

Das Figuras 3.2 e 3.3 é evidente que as ramificações dos dendrogramas dependem do critério de agregação utilizado. No dendrograma da Figura 3.3 é visível o agrupamento das espécies em procariotas (do lado direito) e eucariotas (do lado esquerdo). Existem, no entanto, duas espécies que, aparentemente, estão fora do agrupamento em que tradicionalmente seriam colocadas: os protozoários *Dd* e *Pl*, os quais deveriam estar no grupo dos eucariotas, e as bactérias *Mj* e *Cb*, que deveriam estar no grupo dos procariotas. Na Figura 3.2 a separação entre procariotas e eucariotas não é tão evidente, mas as quatro espécies atrás referidas como excepções continuam a sê-lo também neste dendrograma. Em relação às restantes espécies, e em ambos os dendrogramas, verifica-se que os primatas $\{Hs,\ Mm,\ Pt\}$, os ratos $\{Mu,\ Rn\}$, as leveduras $\{Sc,\ Sp\}$ e as bactérias $\{Sm,\ St\}$ estão bem agrupados. As espécies para as quais o grau de similaridade apresentado não é tão óbvio são, em ambas as figuras, o peixe *Dr*, o qual se encontra ligado ao ramo das plantas *Po* e *Vv* e, apenas na Figura 3.2, o sapo *Xt*, o qual se encontra ramificado com a galinha *Gg*.

Métodos não-hierárquicos

Em virtude do número elevado de variáveis tomadas na coluna da matriz dos erros ($p = 100$), os resultados que se apresentam a seguir dizem respeito à aplicação do algoritmo *K-means* apenas às dez primeiras variáveis ($\delta_1, \ldots, \delta_{10}$, segundo Secção 2.4). A

escolha do número inicial de agrupamentos para a aplicação do algoritmo foi feita com base nos resultados da classificação hierárquica, mais precisamente no dendrograma da Figura 3.3, tendo sido escolhidos dois grupos. A escolha dos centróides iniciais foi feita aleatoriamente. A variante do algoritmo *K-means* seleccionada para execução na função *kmeans()* do R foi o algoritmo de Lloyd[3], descrito na Secção 3.1.3.

Correndo várias vezes o algoritmo, apresenta-se a seguir, dos vários agrupamentos propostos, o agrupamento que apresentou menor erro interno entre os pontos que compõem cada grupo e o centróide desse grupo. O erro interno foi de 1.62 para o **grupo1** e de 0.69 para o **grupo2**. Na Tabela 3.1 encontram-se os centródes de cada grupo e na Tabela 3.12 a distribuição das espécies por grupo (23 espécies por grupo).

Tabela 3.1: Centróides do **grupo1** e do **grupo2** das primeiras dez variáveis δ_j, onde δ_j representa o erro relativo da frequência relativa da j-ésima distância.

centróides	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9	δ_{10}
grupo1	0.10	0.03	−0.05	−0.14	−0.15	−0.11	−0.16	−0.13	−0.09	−0.12
grupo2	0.09	−0.09	0.03	−0.15	−0.13	0.03	−0.12	−0.07	0.09	−0.11

Tabela 3.2: Distribuição das espécies por grupo.

grupo1	*Ap*	*Hr*	*Mj*	*Pf*	*Tk*	*Ba*	*Bs*	*Ct*	*Cb*	*Dv*	*Ec*	*Hi*
	Hp	*Mg*	*Pa*	*Sa*	*Sm*	*St*	*Ce*	*Dd*	*Ca*	*Sc*	*Sp*	
grupo2	*At*	*Os*	*Po*	*Vv*	*Bt*	*Cf*	*Eq*	*Gg*	*Am*	*Dm*	*Mu*	*Rn*
	Xt	*Hs*	*Mm*	*Pt*	*Dr*	*Fu*	*Oa*	*Li*	*Pl*	*Tb*	*Nc*	

A representação gráfica da Figura 3.4 não é mais do que uma colecção de projecções ortogonais da nuvem de $n = 46$ pontos (espécies) sobre os 45 planos coordenados definidos pelos possíveis pares dos 10 menores valores possíveis para a distância. Apesar de se terem considerado apenas as dez primeiras variáveis, a interpretação gráfica não é muito elucidativa. A fim de melhor ilustrar a distribuição das espécies em estudo por dois grupos far-se-á, na Secção 3.2.2, uma aplicação do algoritmo *K-means* considerando os resultados obtidos da análise de componentes principais.

[3] A função *kmeans()* do R, na sua versão actual, implementa também os métodos de Hartigan-Wong [19], Forgy [15] e MacQueen [28].

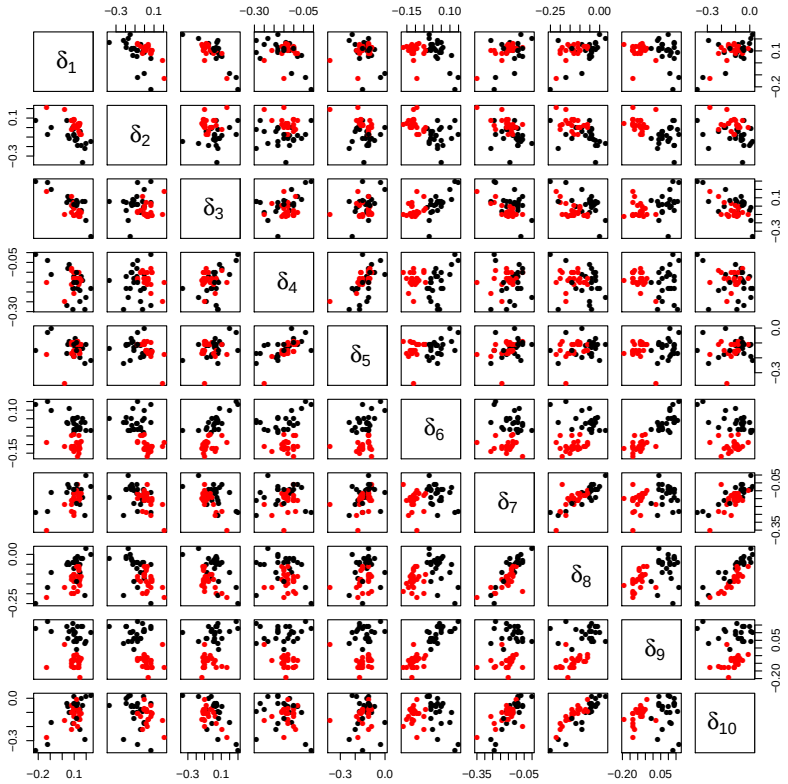


Figura 3.4: Representação das dez primeiras variáveis da matriz de erros relativos $\Delta_{46\times100}$, observadas em 23 espécies para cada um de dois grupos.

3.2 Análise de componentes principais

A análise de componentes principais (ACP) é um método de análise de dados multivariados que transforma um conjunto de p variáveis originais correlacionadas entre si, $X_1, X_2, \cdots, X_p$, num outro conjunto de novas variáveis não correlacionadas, CP_1, CP_2, $\cdots, CP_p$. Aos elementos deste segundo conjunto chamam-se componentes principais. Cada componente principal é uma combinação linear de todas as variáveis originais. As variáveis originais têm a mesma importância estatística, enquanto que as componentes principais são obtidas por ordem decrescente de máxima variância, ou seja, a componente principal CP_1 detém mais informação estatística que a componente principal CP_2, que por sua vez detém mais informação estatística que a componente principal CP_3, e assim por diante. Aproveitando este facto é possível conseguir-se uma redução da dimensionalidade original, pois consideram-se apenas as componentes principais que expliquem a maior parte da variação associada às variáveis iniciais. Assim, o tratamento dos dados é

facilitado visto que, sem perdas significativas de informação, a análise passará a incidir sobre um número reduzido de variáveis não correlacionadas.
A técnica da ACP foi originalmente descrita em 1901 por Karl Pearson, que na prática a usou para um máximo de três variáveis originais, e foi posteriormente consolidada por Hotelling em 1931.

3.2.1 Metodologia

Na ACP pretende-se transformar um vector p-dimensional $X = (X_1, X_2, \cdots, X_p)^T$ num vector s-dimensional $Y = (CP_1, CP_2, \cdots, CP_s)^T$, normalmente de dimensão menor, onde p representa o número de variáveis e s o número de componentes seleccionadas. A transformação da ACP é dada por uma matriz V de dimensão $s \times p$, tal que

$$Y = VX\,.$$

Existem vários métodos para estimar a matriz V. No método convencional, as colunas da matriz V correspondem aos vectores próprios da matriz de correlações ou covariâncias; a ordenação dos vectores próprios é feita em função dos valores próprios correspondentes, por ordem decrescente dos mesmos. A função *princomp()* do R utiliza este método[4]. Poderá haver ganhos de eficiência na determinação da matriz V se forem usados outros métodos de cálculo para a mesma. A escolha desses métodos alternativos depende de vários factores, incluindo o número de variáveis e/ou a dimensão das amostras. A Decomposição em Valores Singulares (DVS) é um desses métodos e é aquele que se irá usar dada a dimensão dos dados [40]. Existem no R pelo menos duas funções, *prcomp()* e *PCA()*, que usam o método DVS como parte do algoritmo que implementam para executar a ACP.

Decomposição em valores singulares

Seja X uma matriz real de dimensão $n \times p$ e característica r. Admite-se, sem perda de generalidade, que $n \geq p$ e, por conseguinte, $r \leq p$. A matriz $X^T X$, de ordem p, é uma matriz simétrica com p valores próprios reais não negativos $\lambda_1, \lambda_2, \cdots, \lambda_p$. Designam-se por **valores singulares** da matriz X as p raízes quadradas dos valores próprios da matriz

[4] Esta função não se aplica quando o número de indivíduos é inferior ao número de variáveis.

$X^T X$, isto é, $\sigma_i = \sqrt{\lambda_i}$, $i = 1, 2, \cdots, p$. A matriz X admite a decomposição

$$X = U\, S\, V^T, \tag{3.3}$$

chamada **decomposição em valores singulares**, onde U é uma matriz ortogonal $n \times p$, S uma matriz diagonal $p \times p$, e V^T uma matriz ortogonal $p \times p$. As colunas de U denominam-se *vectores singulares à esquerda*, $\{u_k\}$, e formam uma base ortonormada do espaço gerado pelas colunas de X. As linhas de V^T denominam-se *vectores singulares à direita*, $\{v_k\}$, e formam uma base ortonormada do espaço gerado pelas linhas de X. Pode mostrar-se que as colunas de U correspondem aos vectores próprios da matriz XX^T e as colunas de V correspondem aos vectores próprios da matriz $X^T X$. Os elementos da diagonal principal de S correspondem aos valores singulares da matriz X, ou seja, $S = diag(\sigma_1, \ldots, \sigma_p)$. No caso da matriz X possuir r valores singulares não nulos, tem-se que $\sigma_1 > \sigma_2 > \ldots > \sigma_r > 0$ e $\sigma_{r+1} = \sigma_{r+2} = \ldots = \sigma_p = 0$. Por convenção, a ordenação dos vectores singulares é feita em função dos valores singulares. No caso de X ser uma matriz quadrada simétrica, a DVS é equivalente à diagonalização[5] [42]. Para obter a DVS de uma matriz no R, pode usar-se a função *svd()*[6].

As componentes principais

Seja $X = (X_1, X_2, \ldots, X_p)^T$ o vector das variáveis originais. As **componentes principais** são combinações lineares das p variáveis originais correlacionadas entre si X_1, X_2, $X_3, \cdots, X_p$:

$$CP_j = e_{1j}X_1 + e_{2j}X_2 + \ldots + e_{pj}X_p = e_j^T X\,, \tag{3.4}$$

onde $j = 1, 2, \ldots, p$ e $e_j^T = (e_{1j}, e_{2j}, \ldots, e_{pj})$ são vectores de constantes. A variância da j-ésima componente principal é determinada por

$$Var\,(CP_j) = \sum_{i=1}^{n} \frac{\left(CP_{ij} - \overline{CP}_j\right)^2}{n-1}\,,$$

[5] Uma matriz A de ordem m é diagonalizável, $A = P\,\Lambda\,P^{-1}$, se e só se possui m vectores próprios linearmente independentes, sendo Λ uma matriz diagonal cujos elementos da diagonal principal são iguais aos valores próprios da matriz A, e P uma matriz que contém os vectores próprios associados aos valores próprios de A. Se A é uma matriz simétrica então é diagonalizável por uma matriz ortogonal Q obtida a partir dos vectores próprios de A, isto é, $A = Q\,\Lambda\,Q^T$.

[6] O número de valores singulares, vectores singulares à direita e vectores singulares à esquerda, é dado por $\min(n, p)$.

onde CP_{ij} corresponde ao valor da j-ésima componente principal para o i-ésimo indivíduo.

Os vectores dos coeficientes e_j^T são determinados de modo a satisfazerem as condições seguintes:

- $Var(CP_1) \geq Var(CP_2) \geq \ldots \geq Var(CP_p)$;
- $Corr\,(CP_i, CP_j) = 0$, $i, j = 1, 2, \ldots, p$, $i \neq j$, isto é, quaisquer duas componentes principais são não correlacionadas[7];
- $e_j^T e_j = 1$, $j = 1, 2, \ldots, p$, isto é, o vector e_j^T tem norma unitária.

Quando se usa a matriz de covariâncias Σ de X (ou a matriz de correlações de X) para a obtenção das componentes principais, prova-se que:

(a) os vectores dos coeficientes e_j^T, $j = 1, 2, \ldots, p$, correspondem aos p vectores próprios associados aos p valores próprios ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$) da matriz de covariâncias (ou de correlações);

(b) $Var(CP_j) = \lambda_j$, $j = 1, 2, \ldots, p$;

(c) $\sum_{j=1}^{p} Var(CP_j) = tr\,(\Sigma)$.

Quando se usa a decomposição em valores singulares da matriz $X = U\,S\,V^T$ para a obtenção das componentes principais, prova-se que:

(a) os vectores dos coeficientes e_j^T, $j = 1, 2, \ldots, p$, correspondem aos p vectores singulares à direita[8], $\{v_k\}$, associados aos valores próprios ($\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_p^2 \geq 0$) da matriz $X^T X$;

(b) $Var(CP_j) = \sigma_j^2$, $j = 1, 2, \ldots, p$;

(c) $\sum_{j=1}^{p} Var(CP_j) = tr\left(X^T X\right)$.

[7] No caso das variáveis originais seguirem uma distribuição normal $p-$variada, as componentes principais são independentes.

[8] Vectores próprios da matriz $X^T X$.

Para avaliar a contribuição das k primeiras componentes CP_k na explicação da variação total, calcula-se a **percentagem de variabilidade explicada** pelas primeiras k componentes principais através da fórmula

$$\frac{Var\,(CP_1) + \ldots + Var\,(CP_k)}{\sum_{k=1}^{p} Var\,(CP_k)} \times 100\,,\ 1 \leq k \leq p. \tag{3.5}$$

O **coeficiente de correlação** entre a j-ésima variável X_j e a k-ésima componente principal CP_k é definido por

$$\rho_{X_j CP_k} = \frac{e_{jk}\sqrt{\text{var}\,(CP_k)}}{\sqrt{\text{var}\,(X_j)}}\,. \tag{3.6}$$

A decisão sobre o número de componentes principais a considerar depende da percentagem de explicação das primeiras k componentes principais. Existem critérios práticos e empíricos para esse efeito (ver, por exemplo, [29]), tais como:

(1) Decidir com base na representação gráfica, por ordem decrescente, da percentagem de variação total explicada por cada componente;

(2) Incluir o número mínimo de componentes que expliquem pelo menos 70% da variação total;

(3) Reter somente aquelas componentes cujas variâncias são maiores do que um.

Em muitas situações, as variáveis originais $X_1, X_2, \ldots, X_p$ são medidas em escalas diferentes ou unidades diferentes, o que conduz a grandes discrepâncias das variâncias. Deste modo, surge a necessidade de se estabelecer uma certa uniformização dos dados, o que se consegue através da padronização das variáveis[9].

$$Z_j = \frac{X_j - \mu_j}{\sqrt{\sigma_{jj}}},\ j = 1, \cdots, p\,.$$

As variáveis Z_j, $j = 1, 2, \ldots p$ têm valor médio nulo e variância unitária. A matriz de covariâncias das variáveis Z_j é igual à matriz de correlações das variáveis X_j, isto é,

$$Cov(Z_i, Z_j) = Corr(X_i, X_j).$$

[9] A finalidade deste procedimento é uniformizar a importância estatística de todas as variáveis utilizadas. Aos valores observados de cada variável é subtraído o seu valor médio e divide-se pelo seu desvio padrão.

De facto, atendendo à definição de covariância e a que $E(Z_j) = 0$, vem

$$Cov\left(Z_i, Z_j\right) = E\left[\left(Z_i - E\left(Z_i\right)\right)\left(Z_j - E\left(Z_j\right)\right)\right] = E\left[\left(\frac{X_i - \mu_i}{\sqrt{\sigma_{ii}}}\right)\left(\frac{X_j - \mu_j}{\sqrt{\sigma_{jj}}}\right)\right]$$

Atendendo às propriedades de valor esperado e à definição de correlação,

$$Cov\left(Z_i, Z_j\right) = \frac{Cov\left(X_i, X_j\right)}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}} = Corr\left(X_i, X_j\right).$$

Em termos geométricos, na representação em $\mathbb{R}^p$, a centralização dos dados equivale a uma translação do centro de gravidade da nuvem (ponto constituído pelos valores médios das variáveis) para a origem do referencial, e cada eixo, de acordo com o valor do desvio padrão da variável correspondente, será estendido (se $\sqrt{\sigma_{jj}} < 1$) ou contraído (se $\sqrt{\sigma_{jj}} > 1$), com factores de alteração das escalas diferenciados para cada eixo. Com a redução dos dados elimina-se o problema da escala de medida das variáveis.

3.2.2 Resultados experimentais

Os vectores dos erros relativos das distâncias para as diferentes espécies apresentam, de modo geral, um erro relativo muito elevado a partir de uma determinada distância, que é menor para os seres menos complexos e maior para os seres mais complexos. O acto de centrar as componentes destes vectores em relação à sua média faz com que sejam tratadas, de forma desigual, os erros relativos associados às primeiras distâncias. Porém, de acordo com [2], são estas as que definem uma adequada caracterização e diferenciação das espécies. Por esse motivo, optou-se por aplicar a ACP a três situações diferentes, de forma a possibilitar a comparação de resultados: na primeira consideraram-se as variáveis padronizadas, na segunda as variáveis apenas centradas e na terceira as variáveis sem padronização. Os resultados relativos às variáveis padronizadas podem ser obtidos usando, por exemplo, a função *PCA()* com o parâmetro *scale.unit=TRUE*, ou a função *prcomp()* com os parâmetros *center=TRUE* e *scale=TRUE*. Ambas as funções fazem uso do método DVS, apesar de no seu *output* a primeira apresentar os valores próprios da matriz de correlações e a segunda a raiz quadrada dos valores próprios da matriz de correlações, em vez de apresentarem o quadrado dos valores singulares e os valores singulares da DVS, respectivamente. Os resultados relativos às variáveis centradas podem ser obtidos com as mesmas funções, mas considerando agora na função *PCA()* o parâmetro *scale.unit=FALSE* e na

função *prcomp()* o parâmetro *scale=FALSE*. As variâncias das componentes principais apresentadas no *output* da função *PCA()* são idênticas aos valores próprios da matriz de covariâncias dos dados, e os desvios padrão das componentes principais apresentadas no *output* da função *prcomp()* são iguais à raiz quadrada dos valores próprios da matriz de covariâncias dos dados. Os resultados relativos às variáveis sem padronização podem ser obtidos através da função *prcomp()*, com os parâmetros *center=FALSE* e *scale=FALSE*.

Nenhuma das componentes resultantes da ACP parece apresentar um significado óbvio como indicador de alguma característica importante associada à amostra em estudo.

Variáveis padronizadas

A seguir apresentam-se os resultados obtidos pela aplicação da função *PCA()* à matriz dos erros relativos padronizada. A Tabela 3.3 dá-nos uma primeira informação acerca da estrutura dos dados. São apresentadas para as 15 primeiras componentes principais a variância explicada (v.próprios), a percentagem de variância total (% var.total) e a percentagem de variância total acumulada (% var.total acum.). Verifica-se que para explicar mais de 80% da variância, é necessário considerar apenas as cinco primeiras componentes principais; este é um número bastante reduzido de componentes principais quando comparado com o número de variáveis originais. Em todo o caso, as percentagens de variância explicadas pelas componentes CP4 e CP5 são relativamente baixas, 3.37% e 2.79% respectivamente, quando comparadas com as percentagens de variância explicadas pelas componentes CP1 e CP2, que são de 47.77% e 20.02%, respectivamente. Tendo em conta os critérios de selecção do número de componentes a considerar (ver Secção 3.2.1), considerar-se-ão apenas as três primeiras componentes principais (ver Figura 3.5).

Tabela 3.3: Variação explicada pelas componentes principais.

c.p.	v.próprios	% var.total	% var.total acum.
CP1	47.7663	47.7663	47.7663
CP2	20.0212	20.0212	67.7875
CP3	6.1751	6.1751	73.9626
CP4	4.9042	4.9042	78.8668
CP5	3.3739	3.3739	82.2407
CP6	2.7920	2.7920	85.0328
CP7	2.1717	2.1717	87.2045

continua na página seguinte

Tabela 3.3 – continuação da página anterior

c.p.	v.próprios	% var.total expl.	% var.total acum.
CP8	1.8624	1.8624	89.0669
CP9	1.5606	1.5606	90.6275
CP10	1.2148	1.2148	91.8423
CP11	1.1527	1.1527	92.9950
CP12	1.0382	1.0382	94.0332
CP13	0.9080	0.9080	94.9412
CP14	0.8272	0.8272	95.7684
CP15	0.8198	0.8198	96.5882

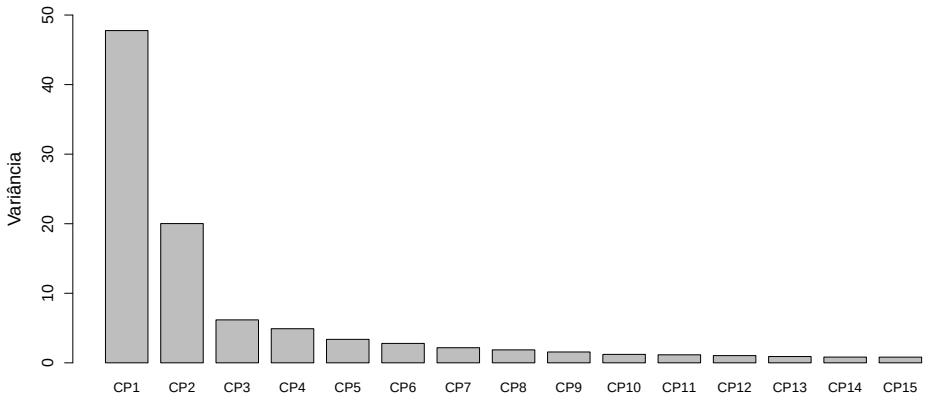


Figura 3.5: *Barplot* com os valores próprios da matriz de correlações.

Na Tabela 3.4 são apresentados os pesos das variáveis que contribuíram, em valor absoluto, com peso superior ou igual a 0.1 para a formação das três primeiras componentes principais[10]. As variáveis que mais contribuíram em valor absoluto para a formação da componente CP1 foram as variáveis δ_{52}, δ_{35}, δ_{37} e δ_{41}; no caso da componente CP2, as variáveis que mais contribuíram foram as variáveis δ_{77}, δ_{79} e δ_{75}; na formação da componentes CP3, a maior contribuição foi dada pelas variáveis δ_1 e δ_{10}.

[10] Os valores da Tabela 3.4 foram obtidos através da função *prcomp()*. O sinal dos valores dos vectores próprios são arbitrários, e portanto podem diferir entre implementações da ACP, e mesmo entre versões do R.

Tabela 3.4: Vectores próprios com peso absoluto igual ou superior a 0.1 nas três primeiras componentes principais.

δ	CP1	CP2	CP3	δ	CP1	CP2	CP3
δ_1			0.3004	δ_{51}	0.1290		
δ_2			−0.2116	δ_{52}	0.1387		
δ_3			−0.2638	δ_{53}	0.1267		
δ_4				δ_{54}	0.1258		−0.1010
δ_5				δ_{55}	0.1069		
δ_6	−0.1197			δ_{56}	0.1062		−0.1134
δ_7			0.2518	δ_{57}			
δ_8		0.1150	0.2571	δ_{58}			
δ_9	−0.1275			δ_{59}			
δ_{10}		0.1248	0.2973	δ_{60}	0.1134		−0.1158
δ_{11}		0.1481	0.2728	δ_{61}		−0.1241	
δ_{12}	−0.1164			δ_{62}		−0.1263	0.1063
δ_{13}		0.1211	0.2381	δ_{63}		−0.1331	
δ_{14}		0.1282	0.2089	δ_{64}			
δ_{15}		0.1367		δ_{65}		−0.1221	0.1049
δ_{16}		0.1147	0.1705	δ_{66}		−0.1369	0.1007
δ_{17}	0.1039	0.1286	0.1135	δ_{67}		−0.1158	
δ_{18}		0.1395		δ_{68}		−0.1140	
δ_{19}	0.1149	0.1165		δ_{69}		−0.1050	
δ_{20}	0.1153	0.1162		δ_{70}		−0.1141	
δ_{21}		0.1147	−0.1365	δ_{71}			
δ_{22}	0.1205	0.1091	0.0431	δ_{72}		−0.1213	
δ_{23}	0.1258	0.1002	0.0061	δ_{73}		−0.1205	
δ_{24}	0.1025	0.1000	−0.1338	δ_{74}		−0.1528	0.1134
δ_{25}	0.1268			δ_{75}		−0.1563	
δ_{26}	0.1318			δ_{76}		−0.1389	
δ_{27}	0.1171		−0.1141	δ_{77}		−0.1606	
δ_{28}	0.1299			δ_{78}		−0.1022	
δ_{29}	0.1298			δ_{79}		−0.1566	
δ_{30}	0.1218		−0.1122	δ_{80}		−0.1032	
δ_{31}	0.1309			δ_{81}		−0.1142	
δ_{32}	0.1321			δ_{82}		−0.1451	
δ_{33}	0.1257			δ_{83}		−0.1175	
δ_{34}	0.1321			δ_{84}		−0.1201	
δ_{35}	0.1331			δ_{85}		−0.1249	
δ_{36}	0.1327			δ_{86}		−0.1332	
δ_{37}	0.1329			δ_{87}		−0.1045	0.1237
δ_{38}	0.1311			δ_{88}		−0.1128	
δ_{39}	0.1310			δ_{89}	0.1027	−0.1109	

continua na página seguinte

Tabela 3.4 – continuação da página anterior

δ	CP1	CP2	CP3	δ	CP1	CP2	CP3
δ_{40}	0.1326			δ_{90}		−0.1425	
δ_{41}	0.1329			δ_{91}		−0.1237	
δ_{42}	0.1315			δ_{92}		−0.1270	
δ_{43}	0.1322			δ_{93}	0.1046		
δ_{44}	0.1304			δ_{94}	0.1043	−0.1037	
δ_{45}	0.1315			δ_{95}			0.1063
δ_{46}	0.1327			δ_{96}	0.1016		
δ_{47}	0.1290			δ_{97}		−0.1276	
δ_{48}	0.1285			δ_{98}	0.1079	−0.1130	
δ_{49}	0.1281			δ_{99}	0.1038		
δ_{50}	0.1032		−0.1192	δ_{100}		−0.1293	

Devido à padronização das variáveis, o comprimento dos vectores dos erros relativos das distâncias é inferior ou igual à unidade. Em termos geométricos, isto significa que os vectores se encontram dentro de uma hiperesfera de raio 1 cujo centro é a origem dos eixos. Na Figura 3.6 é apresentado o círculo de correlações em função das componentes CP1 e CP2. Nesse círculo, as variáveis são representadas graficamente por vectores. A projecção destes vectores sobre as componentes principais corresponde à correlação entre estas e as variáveis representadas por esses vectores (valores da Tabela A.1 em anexo). Constata-se que, das cem variáveis, oito delas, δ_3, δ_6, δ_7, δ_8, δ_9, δ_{11}, δ_{12} e δ_{15}, estão correlacionadas negativamente com a componente CP1. Destas, as variáveis δ_6, δ_9 e δ_{12} são as que apresentam uma maior correlação negativa (inferior a -0.80) com essa componente. As restantes variáveis apresentam uma correlação positiva com a componente CP1: δ_4, δ_5 e δ_{10} apresentam uma correlação muito fraca (inferior a 0.07) e as variáveis compreendidas entre δ_{23} e δ_{54}, com excepção de δ_{24} e δ_{50}, apresentam uma correlação relativamente forte (superior a 0.80)(ver Tabela A.1 em anexo). Esta constatação é também confirmada pela análise dos valores do cosseno quadrado apresentados na Tabela 3.5, pois a qualidade de representação de uma variável é medida pelo cosseno quadrado do ângulo entre o vector correspondente à variável e a projecção desse vector sobre a componente principal desejada[11]. Se o valor do cosseno quadrado estiver próximo de 1, isso significa que a variável está bem projectada sobre a componente principal em questão [27].

[11] Valores obtidos a partir da função *PCA()*.

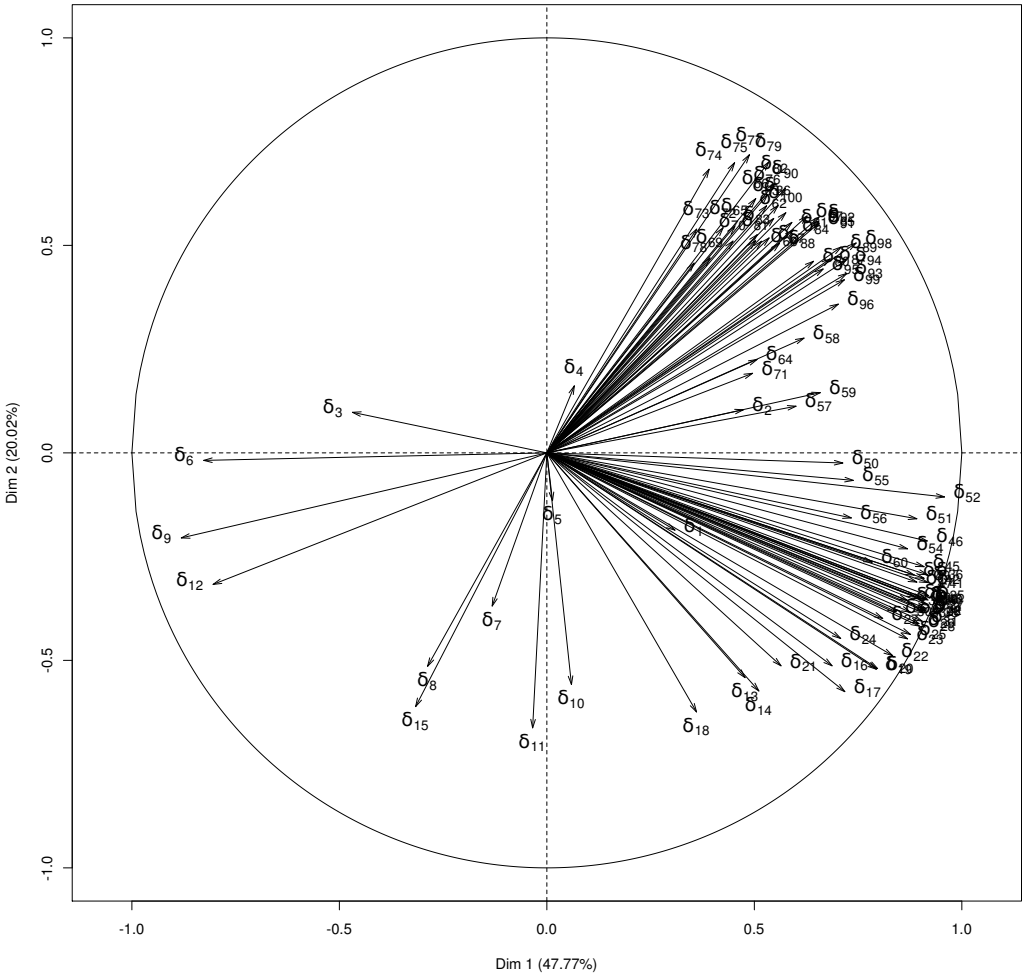


Figura 3.6: Círculo das correlações em função das componentes CP1 e CP2.

Mais de metade das primeiras variáveis, com excepção das variáveis δ_2, δ_3 e δ_4, apresentam uma correlação negativa com a componente CP2. A partir da variável δ_{57}, inclusive, e excluindo a variável δ_{60}, todas as variáveis apresentam uma correlação positiva com a componente CP2. Em relação à componente CP3, a maioria das variáveis apresenta uma correlação relativamente fraca (inferior, em valor absoluto, a 0.34) com esta componente, com excepção das variáveis, δ_2 e δ_3, as quais apresentam uma correlação inferior a -0.50, e das variáveis δ_1, δ_7, δ_8, δ_{10}, δ_{11}, δ_{13}, δ_{14} e δ_{16}, as quais apresentam uma correlação superior a 0.42.

Tabela 3.5: Valores do cosseno quadrado.

δ	CP1	δ	CP1	δ	CP1	δ	CP1
δ_1	0.0954	δ_{26}	0.8296	δ_{51}	0.7951	δ_{76}	0.2825
δ_2	0.2244	δ_{27}	0.6551	δ_{52}	0.9185	δ_{77}	0.2384
δ_3	0.2196	δ_{28}	0.8061	δ_{53}	0.7671	δ_{78}	0.1260
δ_4	0.0044	δ_{29}	0.8045	δ_{54}	0.7554	δ_{79}	0.2859
δ_5	0.0002	δ_{30}	0.7088	δ_{55}	0.5455	δ_{80}	0.4132
δ_6	0.6839	δ_{31}	0.8181	δ_{56}	0.5385	δ_{81}	0.2533
δ_7	0.0174	δ_{32}	0.8330	δ_{57}	0.3607	δ_{82}	0.3016
δ_8	0.0826	δ_{33}	0.7550	δ_{58}	0.3842	δ_{83}	0.2572
δ_9	0.7761	δ_{34}	0.8329	δ_{59}	0.4344	δ_{84}	0.3528
δ_{10}	0.0035	δ_{35}	0.8468	δ_{60}	0.6146	δ_{85}	0.4301
δ_{11}	0.0012	δ_{36}	0.8408	δ_{61}	0.3487	δ_{86}	0.3104
δ_{12}	0.6470	δ_{37}	0.8442	δ_{62}	0.2989	δ_{87}	0.4664
δ_{13}	0.2282	δ_{38}	0.8206	δ_{63}	0.2804	δ_{88}	0.3141
δ_{14}	0.2611	δ_{39}	0.8195	δ_{64}	0.2568	δ_{89}	0.5034
δ_{15}	0.1001	δ_{40}	0.8395	δ_{65}	0.2051	δ_{90}	0.3309
δ_{16}	0.4725	δ_{41}	0.8434	δ_{66}	0.2534	δ_{91}	0.4303
δ_{17}	0.5157	δ_{42}	0.8265	δ_{67}	0.2869	δ_{92}	0.4304
δ_{18}	0.1301	δ_{43}	0.8353	δ_{68}	0.2679	δ_{93}	0.5225
δ_{19}	0.6306	δ_{44}	0.8128	δ_{69}	0.1546	δ_{94}	0.5201
δ_{20}	0.6354	δ_{45}	0.8266	δ_{70}	0.2013	δ_{95}	0.4433
δ_{21}	0.3185	δ_{46}	0.8405	δ_{71}	0.2459	δ_{96}	0.4935
δ_{22}	0.6937	δ_{47}	0.7950	δ_{72}	0.1805	δ_{97}	0.3933
δ_{23}	0.7554	δ_{48}	0.7891	δ_{73}	0.1303	δ_{98}	0.5557
δ_{24}	0.5014	δ_{49}	0.7840	δ_{74}	0.1530	δ_{99}	0.5142
δ_{25}	0.7677	δ_{50}	0.5087	δ_{75}	0.2048	δ_{100}	0.3317

Na Figura 3.7 é apresentada a distribuição das espécies em função das componentes CP1 e CP2. O eixo das abcissas representa os *scores*[12] para a componente CP1 e o eixo das ordenadas representa os *scores* para a componente CP2. Em relação à componente CP1, verifica-se que todas as bactérias apresentam *scores* negativos, com excepção das bactérias *Mj* e *Pf*. Por outro lado, as espécies eucariotas apresentam *scores* positivos, com excepção do fungo *Sp* (*score* quase nulo) e dos protozoários *Dd* e *Pl*. A disposição das espécies na Figura 3.7 torna aparente uma divisão relativamente clara entre, pelo menos, espécies eucariotas e procariotas.

[12] Coordenadas das observações no novo sistema de eixos formado pelas componentes principais.

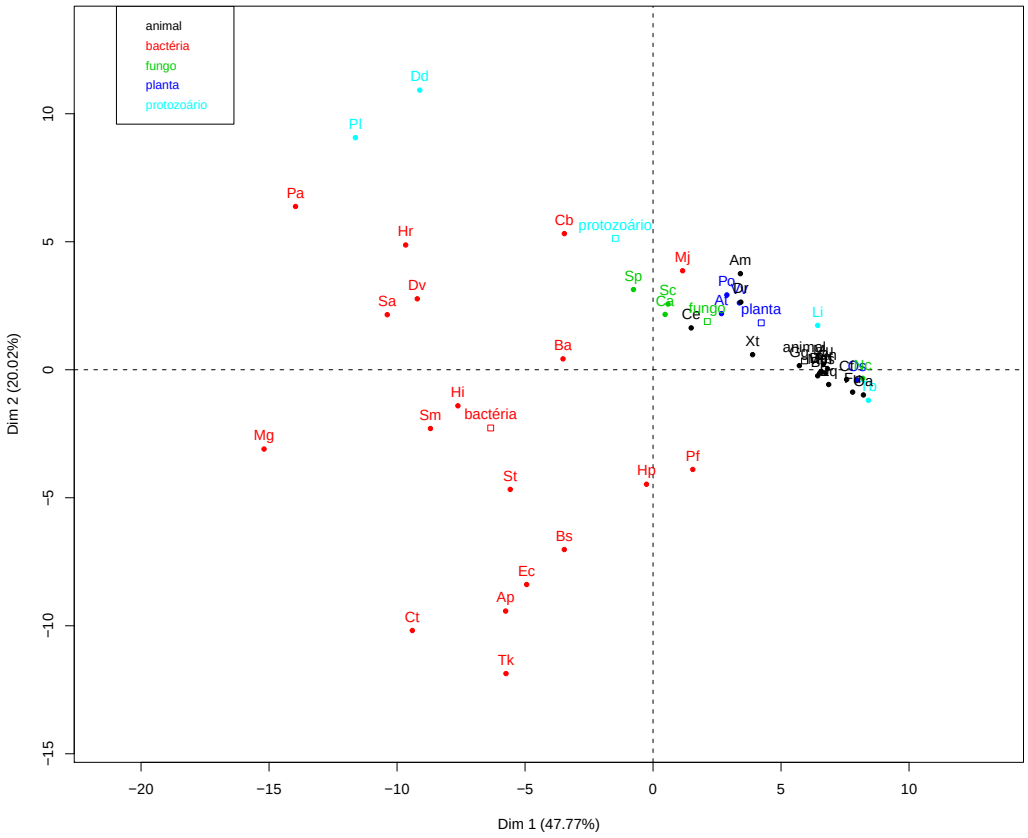


Figura 3.7: Representação das espécies entre CP1 e CP2 (variáveis originais padronizadas).

As representações do círculo de correlações e da distribuição das espécies, em função das componentes CP1 e CP3, encontram-se em anexo na Figura A.1 e na Figura A.2, respectivamente; a Figura A.3 e a Figura A.4 apresentam o círculo de correlações e a distribuição das espécies, em função das componentes CP2 e CP3, respectivamente.

Variáveis apenas centradas

A seguir mostram-se os resultados obtidos pela aplicação da função *PCA()* à matriz dos erros relativos centrada. Na Tabela 3.6 são apresentadas, para as 15 primeiras componentes principais, a variância explicada, a percentagem de variância total e a percentagem de variância total acumulada. Verifica-se que para explicar mais de 80% da variância, é necessário considerar apenas as quatro primeiras componentes principais. As percentagens de variância explicadas pelas componentes CP1, CP2 e CP3 são de 54.90%, 71.06% e 76.61%, respectivamente. Tendo em conta os critérios de selecção do número de componentes a considerar, considerar-se-ão apenas as três primeiras componentes principais (ver Figura 3.8).

Tabela 3.6: Variação explicada pelas componentes principais.

c.p.	v.próprios	% var.total	% var.total acum.
CP1	4.3239	54.8983	54.8983
CP2	1.2726	16.1571	71.0554
CP3	0.4375	5.5541	76.6094
CP4	0.2768	3.5143	80.1238
CP5	0.2467	3.1322	83.2560
CP6	0.2112	2.6817	85.9377
CP7	0.1853	2.3521	88.2898
CP8	0.1721	2.1855	90.4754
CP9	0.1280	1.6254	92.1008
CP10	0.1064	1.3503	93.4510
CP11	0.0896	1.1375	94.5886
CP12	0.0780	0.9901	95.5787
CP13	0.0695	0.8825	96.4612
CP14	0.0617	0.7837	97.2449
CP15	0.0589	0.7483	97.9932

Na Tabela 3.7 são apresentados os pesos das variáveis que contribuíram, em valor absoluto, com peso superior ou igual a 0.1 para a formação das três primeiras componentes principais[13]. As variáveis que mais contribuíram em valor absoluto para a formação da componente CP1 foram as últimas, a partir da variável δ_{88}; no caso da componente CP2, as variáveis que mais contribuíram foram as variáveis δ_{31}, δ_{28}, δ_{37}, δ_{34} e δ_{25}; na formação da componente CP3, a maior contribuição deveu-se às variáveis δ_{99}, δ_{68}, δ_{62} e δ_{61}.

Atendendo à Tabela A.2, em anexo, e relativamente à CP1, constata-se que a partir da variável δ_{89} todas as variáveis apresentam uma correlação positiva relativamente forte (superior a 0.83) com essa componente principal. O mesmo acontece também para as variáveis δ_{85} e δ_{87}. As variáveis δ_9 e δ_{12} apresentam uma correlação negativa relativamente forte (inferior a -0.83) e a correlação da variável δ_{18} é praticamente nula. Em relação à componente CP2, as variáveis δ_{13}, δ_{14} e as variáveis compreendidas entre δ_{16} e δ_{54} (excluindo δ_{50}), são aquelas que apresentam maior correlação positiva com esta componente; as variáveis δ_3, δ_4, δ_6, δ_9, δ_{12} e todas as variáveis a partir de δ_{61}, inclusive, apresentam uma correlação negativa relativamente fraca (superior a -0.39) ou mesmo praticamente nula; as variáveis δ_1, δ_2, δ_5, δ_7, δ_8, δ_{10}, δ_{11}, δ_{15}, δ_{55}, δ_{57}, δ_{58} e δ_{59} apresentam

[13] Os valores da Tabela 3.4 foram obtidos através da função *prcomp()*.

uma correlação positiva relativamente fraca (inferior a 0.33). Finalmente, e em relação à componente CP3, praticamente todas as variáveis apresentam uma correlação relativamente fraca (inferior, em valor absoluto, a 0.39) com essa componente, à excepção de δ_{61}, δ_{62}, δ_{63} e δ_{68}, δ_{74}, δ_{65} e δ_{66}, as quais apresentam correlação positiva.

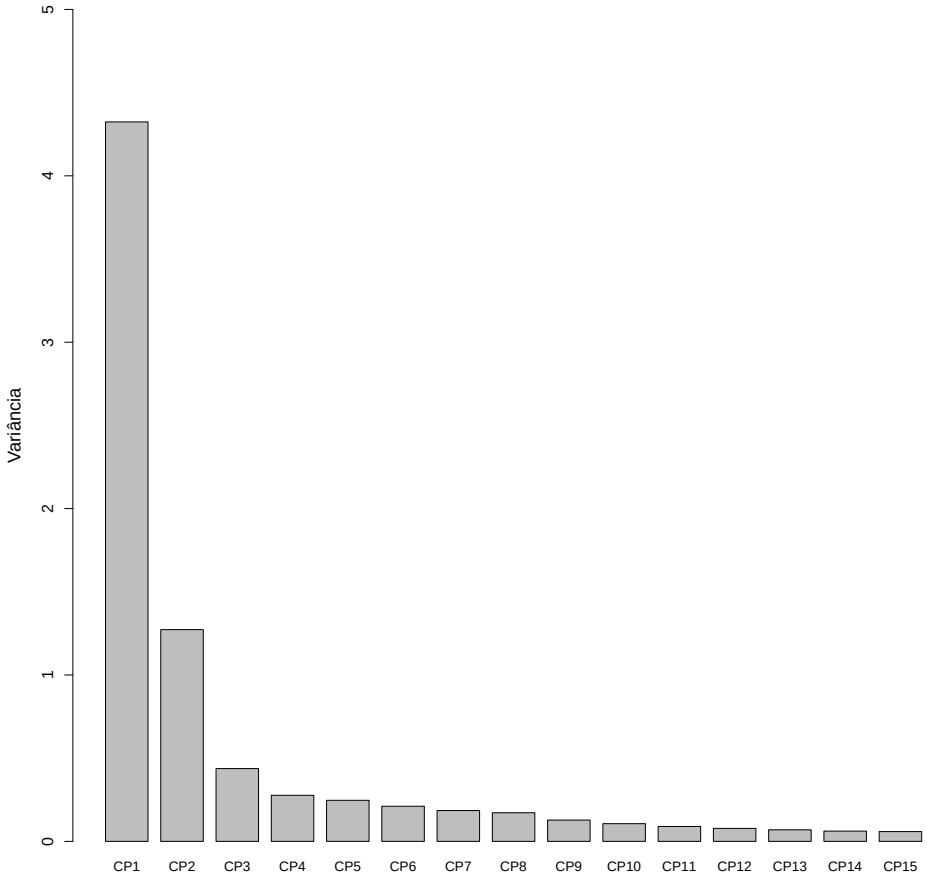


Figura 3.8: *Barplot* com os valores próprios da matriz de covariâncias.

Tabela 3.7: Vectores próprios com peso absoluto igual ou superior a 0.1 nas três primeiras componentes principais.

δ	CP1	CP2	CP3	δ	CP1	CP2	CP3
δ_1				δ_{51}			
δ_2				δ_{52}			
δ_3				δ_{53}			
δ_4				δ_{54}			
δ_5				δ_{55}			
δ_6				δ_{56}			

continua na página seguinte

Tabela 3.7 – continuação da página anterior

δ	CP1	CP2	CP3	δ	CP1	CP2	CP3
δ_7				δ_{57}			
δ_8				δ_{58}			0.1078
δ_9				δ_{59}			
δ_{10}				δ_{60}			
δ_{11}				δ_{61}			0.2222
δ_{12}				δ_{62}			0.2278
δ_{13}				δ_{63}			0.2166
δ_{14}				δ_{64}			
δ_{15}				δ_{65}			0.1871
δ_{16}				δ_{66}	0.1049		0.1905
δ_{17}		0.1045		δ_{67}	0.1016		0.1742
δ_{18}				δ_{68}			0.2344
δ_{19}		0.1400		δ_{69}			0.1057
δ_{20}		0.1306		δ_{70}			0.1287
δ_{21}				δ_{71}			
δ_{22}		0.1645		δ_{72}	0.1118		
δ_{23}		0.1594		δ_{73}			0.1633
δ_{24}				δ_{74}		−0.1145	0.2173
δ_{25}		0.1836		δ_{75}	0.1184	−0.1223	0.2071
δ_{26}		0.1723		δ_{76}	0.1259		0.1875
δ_{27}		0.1203		δ_{77}	0.1354	−0.1332	0.1706
δ_{28}		0.1966		δ_{78}			
δ_{29}		0.1831		δ_{79}	0.1439	−0.1227	0.1463
δ_{30}		0.1303		δ_{80}	0.1603		
δ_{31}		0.1972		δ_{81}	0.1387		
δ_{32}		0.1834		δ_{82}	0.1574	−0.1161	0.1031
δ_{33}		0.1323		δ_{83}	0.1415		−0.1910
δ_{34}		0.1845		δ_{84}	0.1561		
δ_{35}		0.1685		δ_{85}	0.1755		
δ_{36}		0.1276		δ_{86}	0.1603		
δ_{37}		0.1849		δ_{87}	0.1705		0.1015
δ_{38}		0.1718		δ_{88}	0.1650		−0.1843
δ_{39}		0.1330		δ_{89}	0.2016		−0.1489
δ_{40}		0.1680		δ_{90}	0.1803	−0.1143	−0.1306
δ_{41}		0.1574		δ_{91}	0.1811		
δ_{42}		0.1138		δ_{92}	0.1874		
δ_{43}		0.1426		δ_{93}	0.1792		
δ_{44}		0.1417		δ_{94}	0.1981		−0.1558
δ_{45}		0.1090		δ_{95}	0.1876		−0.1814
δ_{46}		0.1275		δ_{96}	0.1842		−0.2116
δ_{47}		0.1260		δ_{97}	0.1896		−0.1417
δ_{48}		0.1018		δ_{98}	0.2137		−0.1902

continua na página seguinte

Tabela 3.7 – continuação da página anterior

δ	CP1	CP2	CP3	δ	CP1	CP2	CP3
δ_{49}		0.1193		δ_{99}	0.1977		−0.2821
δ_{50}		0.1228		δ_{100}	0.1732		

Na Figura 3.9 é apresentada a distribuição das espécies em função das componentes CP1 e CP2. Tal como no caso padronizado, também aqui se verifica uma divisão geral entre as espécies procariotas e eucariotas. Em relação à CP1, verifica-se que todas as bactérias apresentam *scores* negativos, com excepção das bactérias *Mj* e *Cb*. Por outro lado, as espécies eucariotas apresentam *scores* positivos, com excepção do protozoário *Pl*. Neste caso, o protozoário *Dd* apresenta um *score* quase nulo.

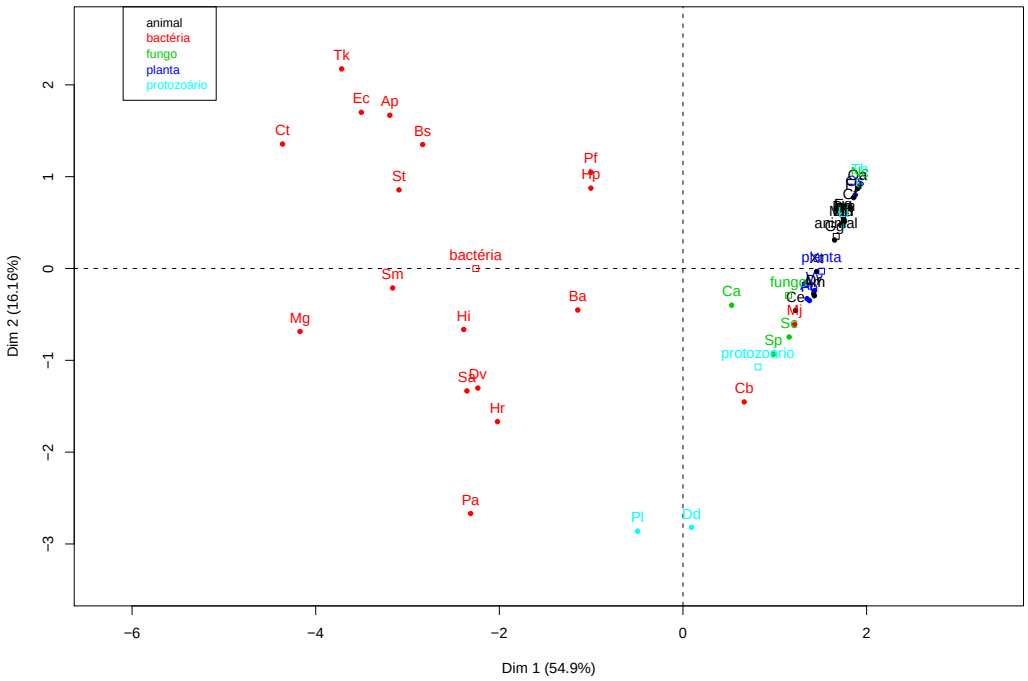


Figura 3.9: Representação das espécies entre CP1 e CP2 (variáveis originais apenas centradas).

A representação das espécies em função das componentes CP1 e CP3 encontra-se em anexo na Figura A.5 e a representação das espécies em função das componentes CP2 e CP3 encontra-se em anexo na Figura A.6.

Variáveis não padronizadas

A seguir mostram-se os resultados obtidos pela aplicação da função *prcomp()* à matriz dos erros relativos. Na Tabela 3.8 são apresentadas, para as 5 primeiras componentes principais, a variância explicada, a percentagem de variância total e a percentagem de variância total acumulada. Verifica-se que as duas primeiras componentes CP1 e CP2 explicam mais de 94% da variância: a CP1 contribui com 91.18% e a CP2 contribui com 2.93%.

Tabela 3.8: Variação explicada pelas componentes principais.

c.p.	v.próprios	% var.total	% var.total acum.
CP1	47.6292	91.18	91.18
CP2	1.5280	2.93	94.10
CP3	0.9868	1.89	95.99
CP4	0.3318	0.64	96.63
CP5	0.2525	0.48	97.11

A percentagem de variância total acumulada é calculada através da fórmula (3.5), onde a variância explicada por cada componente, $Var(CP_j)$, é igual ao quadrado dos valores singulares da DVS da matriz dos erros relativos. A seguir são apresentados os valores singulares[14], por ordem decrescente.

Tabela 3.9: Valores Singulares da matriz dos erros relativos (variáveis não padronizadas).

46.2959	8.2921	6.6639	3.8641	3.3706	3.2919	2.9209	2.8179	2.7418
2.4245	2.2044	2.0112	1.7908	1.7839	1.6514	1.5429	1.4651	1.3159
1.0397	0.8603	0.6073	0.5377	0.4105	0.3013	0.2680	0.2099	0.1695
0.1418	0.1343	0.1238	0.1206	0.1008	0.0907	0.0843	0.0648	0.0636
0.0560	0.0556	0.0507	0.0466	0.0432	0.0365	0.0349	0.0281	0.0192
0.0177								

Atendendo a que o traço da matriz dos erros relativos é igual a 2350.758, as percentagens de variância total acumulada para as duas primeiras componentes são as seguintes:

$$\%\,\text{var.total acum.(CP1)} = \frac{(46.2959)^2}{2350.758} \times 100 = 91.18$$

[14] Os valores singulares foram obtidos usando a função *svd()*.

$$\%\,\text{var.total acum.(CP2)} = \frac{(46.2959)^2 + (8.2921)^2}{2350.758} \times 100 = 94.10$$

Na Tabela 3.10 são apresentados os pesos das variáveis que contribuíram, em valor absoluto, com peso superior ou igual a 0.1 para a formação das duas primeiras componentes principais[15]. As variáveis compreendidas entre δ_{51} e δ_{87}, inclusive, com excepção da δ_{86}, foram as variáveis que mais contribuíram para a formação da componente CP1, com pesos relativamente próximos. No caso da componente CP2, as variáveis que mais contribuíram foram as últimas variáveis, a partir da variável δ_{88}, com excepção das variáveis δ_{93} e δ_{96}.

Tabela 3.10: Vectores próprios com peso absoluto igual ou superior a 0.1 nas duas primeiras componentes principais.

δ	CP1	CP2	δ	CP1	CP2	δ	CP1	CP2
δ_1			δ_{35}		0.1074	δ_{69}	−0.1307	
δ_2			δ_{36}		0.1032	δ_{70}	−0.1325	
δ_3			δ_{37}		0.1158	δ_{71}	−0.1335	
δ_4			δ_{38}		0.1220	δ_{72}	−0.1255	
δ_5			δ_{39}	−0.1026	0.1188	δ_{73}	−0.1308	
δ_6			δ_{40}	−0.1007	0.1219	δ_{74}	−0.1304	
δ_7			δ_{41}	−0.1029	0.1131	δ_{75}	−0.1270	
δ_8			δ_{42}	−0.1091	0.1133	δ_{76}	−0.1278	
δ_9			δ_{43}	−0.1087	0.1251	δ_{77}	−0.1249	−0.1198
δ_{10}			δ_{44}	−0.1100	0.1221	δ_{78}	−0.1276	
δ_{11}			δ_{45}	−0.1147	0.1178	δ_{79}	−0.1264	−0.1244
δ_{12}			δ_{46}	−0.1135	0.1127	δ_{80}	−0.1218	−0.1054
δ_{13}			δ_{47}	−0.1157	0.1215	δ_{81}	−0.1201	−0.1089
δ_{14}			δ_{48}	−0.1197	0.1258	δ_{82}	−0.1220	−0.1444
δ_{15}			δ_{49}	−0.1196	0.1272	δ_{83}	−0.1206	−0.1129
δ_{16}			δ_{50}	−0.1177		δ_{84}	−0.1222	−0.1151
δ_{17}			δ_{51}	−0.1242	0.1140	δ_{85}	−0.1204	−0.1380
δ_{18}			δ_{52}	−0.1226		δ_{86}	−0.1195	−0.1447
δ_{19}			δ_{53}	−0.1252	0.1320	δ_{87}	−0.1206	−0.1038
δ_{20}			δ_{54}	−0.1279	0.1211	δ_{88}	−0.1136	−0.1447
δ_{21}			δ_{55}	−0.1271	0.1075	δ_{89}	−0.1105	−0.1750
δ_{22}			δ_{56}	−0.1283	0.1227	δ_{90}	−0.1148	−0.1913
δ_{23}			δ_{57}	−0.1294		δ_{91}	−0.1183	−0.1483
δ_{24}			δ_{58}	−0.1283		δ_{92}	−0.1157	−0.1635
δ_{25}			δ_{59}	−0.1289		δ_{93}	−0.1186	−0.1044

continua na página seguinte

[15] Os valores da Tabela 3.10 foram obtidos através da função *prcomp()*.

Tabela 3.10 – continuação da página anterior

δ	CP1	CP2	δ	CP1	CP2	δ	CP1	CP2
δ_{26}			δ_{60}	-0.1327	0.1269	δ_{94}	-0.1105	-0.1515
δ_{27}			δ_{61}	-0.1265		δ_{95}	-0.1124	-0.1572
δ_{28}			δ_{62}	-0.1280		δ_{96}	-0.1125	-0.1219
δ_{29}		0.1015	δ_{63}	-0.1283		δ_{97}	-0.1097	-0.1795
δ_{30}		0.1036	δ_{64}	-0.1303		δ_{98}	-0.1086	-0.1860
δ_{31}		0.1076	δ_{65}	-0.1289		δ_{99}	-0.1103	-0.1549
δ_{32}		0.1069	δ_{66}	-0.1277		δ_{100}	-0.1175	-0.1701
δ_{33}		0.1074	δ_{67}	-0.1285				
δ_{34}		0.1094	δ_{68}	-0.1317				

Atendendo à Tabela A.3, em anexo, constata-se que a maioria das variáveis estão correlacionadas negativamente com a componente CP1, com excepção das variáveis δ_3, δ_5, δ_6, δ_7, δ_8, δ_9, δ_{10}, δ_{11}, δ_{12} e δ_{15}. Destas, as variáveis δ_9 e δ_{12} apresentam uma correlação positiva relativamente forte (superior a 0.84) com a componente CP1. Em relação à componente CP2, as variáveis δ_2, δ_4, δ_{50}, δ_{57}, δ_{58}, δ_{59} e todas as restantes a partir de δ_{61} apresentam correlação negativa, registando-se os maiores valores de correlação a partir da variável δ_{74}. As restantes variáveis apresentam uma correlação positiva relativamente fraca, com excepção das variáveis δ_{11}, δ_{15} e δ_{18}.

Atendendo à Tabela 3.8, tem-se que mais de 94% da variabilidade total dos dados é preservada pela projecção da nuvem de pontos ($n = 46$ espécies) sobre o sub-espaço bidimensional de $\mathbb{R}^{100}$ que é gerado pelas duas primeiras componentes CP1 e CP2. Este resultado significa que a distribuição das espécies na Figura 3.10 é uma representação bastante fidedigna da nuvem de pontos original.
À semelhança dos outros dois casos, variáveis padronizadas e variáveis apenas centradas, também aqui é possível identificar as espécies procariotas e eucariotas, as quais aparecem em grupos bem separados na projecção dos dados sobre a componente CP1, com excepção das bactérias *Mj* e *Cb* e dos protozoários *Dd* e *Pl*.

Pode assim concluir-se que os resultados de todas as análises efectuadas são concordantes.

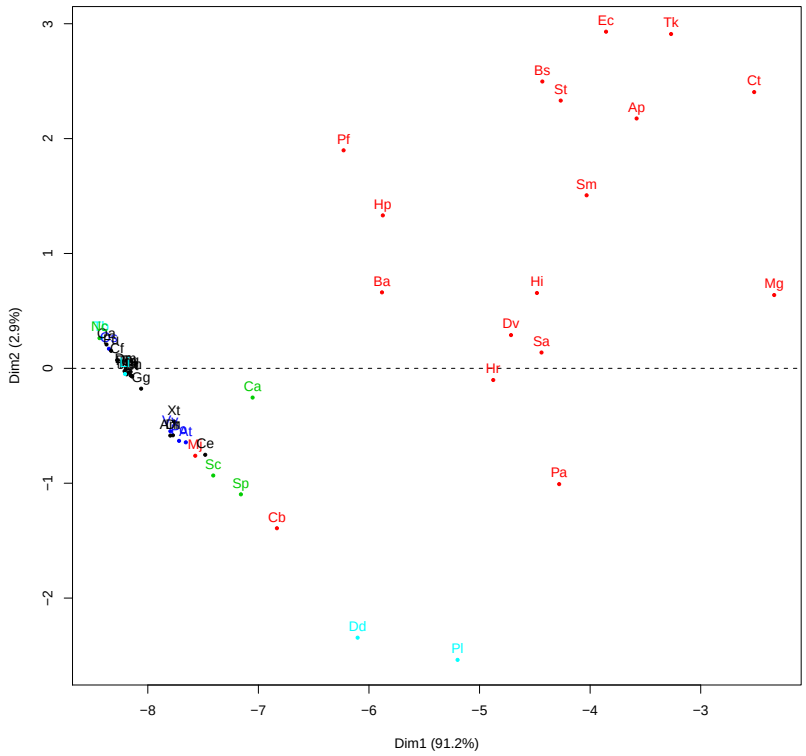


Figura 3.10: Representação das espécies entre CP1 e CP2 (variáveis originais não padronizadas).

Aplicação do algoritmo *K-means*

Dos resultados da análise de componentes principais efectuada às variáveis originais não padronizadas, concluiu-se que mais de 94% da variabilidade total dos dados é explicada pelas componentes CP1 e CP2. Considerando os *scores* destas componentes, apresenta-se a seguir uma aplicação da classificação não-hierárquica baseada no algoritmo *K-means*, descrito na Secção 3.1.3. Como foi referido na Secção 3.1.4, a escolha do número de agrupamentos foi baseada nos resultados obtidos pela classificação hierárquica. Para $k = 2$ grupos é possível observar na Figura 3.11 uma divisão entre as espécies eucariotas e procariotas, com excepção das bactérias *Mj* e *Cb*, que continuam a aparecer no grupo das espécies eucariotas. Recorde-se que, de entre os organismo procariotas, são estas duas espécies aquelas que apresentam o maior desvio padrão. Em relação aos organismos eucariotas, os protozoários *Dd* e *Pl* são os que possuem maior desvio padrão (ver Tabela 2.3).

O agrupamento que apresentou menor erro interno entre os pontos que compõem cada grupo e o centróide desse grupo é aquele que se apresenta na Figura 3.11. O erro interno

foi de 39.22 para o **grupo1** e de 29.01 para o **grupo2**. Na Tabela 3.11 encontram-se os centróides de cada grupo e na Tabela 3.12 a distribuição das espécies por grupo (16 no grupo1 e 30 no grupo2).

Tabela 3.11: Centróides do grupo1 e do grupo2 da CP1 e CP2.

centróides	CP1	CP2
grupo1	−4.32	1.33
grupo2	−7.77	−0.43

Tabela 3.12: Distribuição das espécies por grupo.

grupo1	*Ap*	*Hr*	*Pf*	*Tk*	*Ba*	*Bs*	*Ct*	*Dv*	*Ec*	*Hi*	*Hp*	*Mg*
	Pa	*Sa*	*Sm*	*St*								
grupo2	*Mj*	*Cb*	*At*	*Os*	*Po*	*Vv*	*Bt*	*Cf*	*Eq*	*Gg*	*Am*	*Dm*
	Mu	*Ce*	*Rn*	*Xt*	*Hs*	*Mm*	*Pt*	*Dr*	*Fu*	*Oa*	*Dd*	*Li*
	Pl	*Tb*	*Ca*	*Nc*	*Sc*	*Sp*						

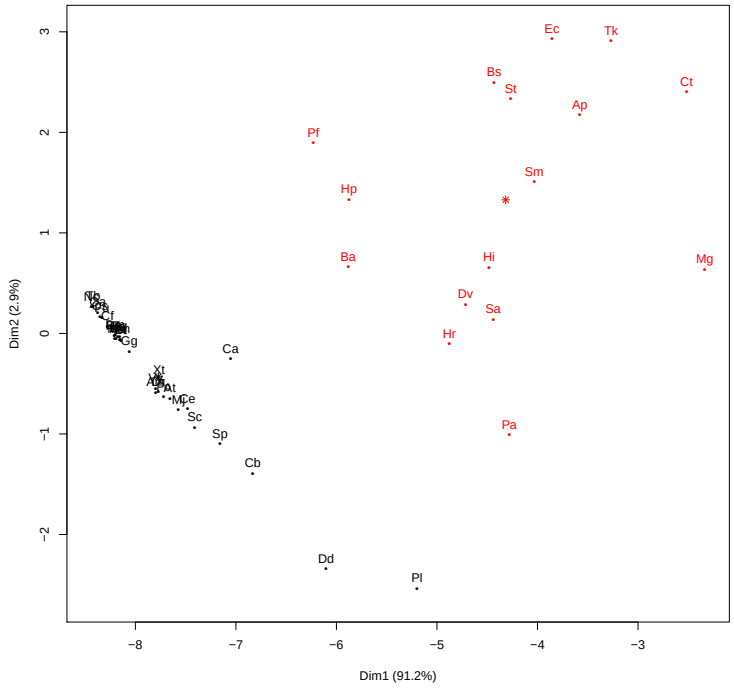


Figura 3.11: Algoritmo *K-means* aplicado aos *scores* das componentes CP1 e CP2 (variáveis originais não padronizadas).

Capítulo 4

Modelação da distribuição das distâncias

A representação gráfica da distribuição empírica das várias espécies e a distribuição modelo (2.13) proposta por [2], sugeriram averiguar a existência de outro modelo teórico alternativo, também definido por mistura de geométricas, que determine um melhor ajustamento da distribuição empírica. Neste capítulo tentar-se-á concretizar este modelo teórico alternativo. A estimação dos parâmetros do modelo será feita pelo método da máxima verosimilhança, através do algoritmo iterativo EM (*Expectation - Maximization*). A aplicação do algoritmo EM tem a vantagem de permitir considerar os efeitos do agrupamento dos dados e de simplificar o processo de obtenção das estimativas de máxima verosimilhança para a mistura de distribuições [31]. A fim de avaliar a qualidade do ajustamento dos vários modelos probabilísticos teóricos à distribuição empírica, utilizar-se-á o teste de ajustamento do qui-quadrado e medidas de similaridade, designadamente uma medida baseada numa distância e a medida de Kullback-Liebler.

4.1 Mistura finita de distribuições

Distribuições baseadas em misturas de outras distribuições ocorrem quando a população é constituída por subgrupos heterogéneos, cada qual representado por uma distribuição de probabilidade diferente [25]. Neste trabalho apenas será tratado o caso de uma mistura finita de distribuições paramétricas (caso discreto).

Seja $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ uma amostra aleatória de dimensão n, onde $\mathbf{Y}_j$ é um vector aleatório p-dimensional com função massa de probabilidade $f(\mathbf{y}_j) \in \mathbb{R}^p$. Seja $\mathbf{Y} = \left(\mathbf{Y}_1^T, \mathbf{Y}_2^T, \ldots, \mathbf{Y}_n^T\right)^T$ o vector que representa a amostra total[1]. Uma realização do vector aleatório $\mathbf{Y}$ será denotada por $\mathbf{y} = \left(\mathbf{y}_1^T, \mathbf{y}_2^T, \ldots, \mathbf{y}_n^T\right)^T$, onde $\mathbf{y}_j$ é o valor observado do vector aleatório $\mathbf{Y}_j$.

De acordo com McLachlan *et al.* [30], diz-se que $\mathbf{Y}_j$ tem distribuição pertencente a uma **mistura finita de distribuições paramétricas** com g componentes se $f(\mathbf{y}_j)$ puder ser escrita na forma

$$f\left(\mathbf{y}_j \,|\Psi\right) = \sum_{m=1}^{g} \pi_m \, f_m\left(\mathbf{y}_j \,|\theta_m\right) \,, \tag{4.1}$$

onde

- $\mathbf{y}_j \in \mathbb{R}^p$;
- $f_m\left(\mathbf{y}_j \,|\theta_m\right)$ são funções massa de probabilidade conhecidas a menos de parâmetros (desconhecidos);
- as quantidades $\pi_1, \pi_2, \ldots, \pi_g$ são escalares não negativos tais que $\sum_{m=1}^{g} \pi_m = 1$;
- Ψ é um vector que contém todos os parâmetros desconhecidos do modelo da mistura e pertence ao espaço paramétrico

$$\Omega = \left\{ \left(\pi_1, \pi_2, \ldots, \pi_{g-1}, \, \xi^T\right)^T : \sum_{m=1}^{g} \pi_m = 1 \text{ e } \pi_{\mathrm{m}} \geq 0, \, \theta_m \in \Theta_m \,, 1 \leq m \leq g \right\}, \tag{4.2}$$

  em que ξ é o vector que contém todos os parâmetros $\theta_1, \theta_2, \ldots, \theta_g$, inicialmente distintos, e Θ_m representa o espaço paramétrico para θ_m.

As funções $f_m\left(\mathbf{y}_j \,|\theta_m\right)$ são designadas por componentes da mistura e $\pi_1, \pi_2, \ldots, \pi_g$ por pesos ou proporções da mistura. A proporção de mistura π_g é determinada por

$$\pi_g = 1 - \sum_{m=1}^{g-1} \pi_m \,.$$

Atendendo a que as funções $f_m\left(\mathbf{y}_j \,|\theta_m\right), m = 1, \ldots, g$, são funções massa de probabilidade, a expressão (4.1) define uma função massa de probabilidade. De facto,

$$\sum_{\mathbf{y}_j} f\left(\mathbf{y}_j \,|\Psi\right) = \sum_{\mathbf{y}_j} \left(\sum_{m=1}^{g} \pi_m f_m\left(\mathbf{y}_j \,|\theta_m\right) \right) = \sum_{m=1}^{g} \pi_m \sum_{\mathbf{y}_j} f_m\left(\mathbf{y}_j \,|\theta_m\right) = 1 \,.$$

[1] O vector $\mathbf{Y}$ é um n-uplo de pontos em $\mathbb{R}^p$.

Na maioria das aplicações é frequente as componentes da mistura pertencerem a uma mesma família paramétrica, pelo que a mistura finita de distribuições (4.1) vem na forma

$$f\left(\mathbf{y}_j\,|\Psi\right) = \sum_{m=1}^{g} \pi_m\, f\left(\mathbf{y}_j\,|\theta_m\right)\,, \tag{4.3}$$

onde $f\left(\cdot\,|\theta\right)$ representa um elemento genérico da família paramétrica $\left\{f\left(\mathbf{y}_j\,|\theta\right) : \theta \in \Theta\right\}$.

Na formulação do modelo de mistura (4.1) considera-se o número de componentes g como sendo fixo, mas em muitas aplicações o valor de g é desconhecido e tem de ser inferido a partir dos dados disponíveis, juntamente com as proporções da mistura e os parâmetros das componentes da mistura. McLachlan *et al.* [30], afirmam que o teste para o número de componentes g numa mistura é um problema importante mas muito difícil, o qual ainda não foi completamente resolvido.

Concretização da mistura finita ao caso de geométricas

Considere-se uma mistura de distribuições geométricas (parâmetros diferentes) com g componentes. Atendendo à definição de função massa de probabilidade (2.5), as componentes da mistura na expressão (4.3) são dadas por

$$f(k\,|p_m) = p_m(1-p_m)^{k-1},\; k = 1, 2, \ldots,\; m = 1, 2, \ldots g,\; (0 \leq p_m \leq 1)\,. \tag{4.4}$$

Deste modo,

$$f\left(k\,|\Psi\right) = \sum_{m=1}^{g} \pi_m\, p_m(1-p_m)^{k-1},\; k = 1, 2, \ldots,\; (0 \leq p_m \leq 1)\,. \tag{4.5}$$

As proporções da mistura π_m são não negativas, a sua soma é igual a 1 e o vector Ψ dos parâmetros desconhecidos é constituído por

$$\Psi = (\pi_1, \pi_2, \ldots, \pi_{g-1}, p_1, p_2, \ldots, p_g)^T. \tag{4.6}$$

No caso de duas componentes,

$$f\left(k\,|\Psi\right) = \pi_1\, p_1\,(1-p_1)^{k-1} + (1-\pi_1)\, p_2\,(1-p_2)^{k-1},\; k = 1, 2, \ldots, \tag{4.7}$$

$$\Psi = (\pi_1, p_1, p_2)^T,\; 0 \leqslant p_i \leqslant 1, i = 1, 2\,.$$

Na Figura 4.1 encontram-se representadas duas distribuições geométricas de parâmetros $p_1 = 0.3$ e $p_2 = 0.5$, bem como a curva (cor azul) correspondente à representação da mistura destas duas distribuições com pesos $\pi_1 = 0.4$ e $\pi_2 = 0.6$.

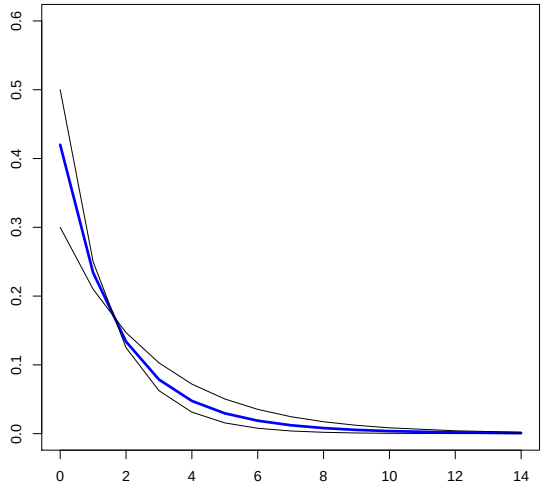


Figura 4.1: Mistura de duas distribuições geométricas com $\Psi = (0.4, 0.3, 0.5)$

4.1.1 Identificabilidade de misturas de distribuições

A estimação do vector dos parâmetros Ψ no modelo (4.1), com base nas observações $\mathbf{y}_j$, tem significado apenas se Ψ for identificável [30]. Em geral, uma família paramétrica de função massa de probabilidade $f\left(\mathbf{y}_j \,|\Psi\right)$ diz-se identificável se valores distintos de Ψ determinam membros distintos da família $\{f\left(\mathbf{y}_j \,|\Psi\right) : \Psi \in \Omega\}$, isto é,

$$f\left(\mathbf{y}_j \,|\Psi\right) = f\left(\mathbf{y}_j \,|\Psi^*\right) , \tag{4.8}$$

se e só se

$$\Psi = \Psi^* .$$

No caso de misturas finitas de distribuições, a definição de identificabilidade é ligeiramente diferente. Suponha-se que a função massa de probabilidade $f\left(\mathbf{y}_j \,|\Psi\right)$ em (4.1) tem duas componentes de mistura $f_i\left(\mathbf{y}_j \,|\theta_i\right)$ e $f_h\left(\mathbf{y}_j \,|\theta_h\right)$ pertencentes ambas à mesma família paramétrica. No caso de se permutarem os índices i e h em $\Psi = (\pi_i, \pi_h; \theta_i, \theta_h)$, a função massa $f\left(\mathbf{y}_j \,|\Psi\right)$ terá o mesmo valor para cada $\mathbf{y}_j$, isto é, a igualdade (4.8) é verificada. Embora esta classe de misturas possa ser identificável, o vector Ψ não o é. De facto, se todas as g componentes da mistura (4.1) pertencerem à mesma família paramétrica, então a função massa da mistura $f\left(\mathbf{y}_j \,|\Psi\right)$ será invariante para as $g!$ permutações dos índices das componentes de Ψ.

Sejam

$$f\left(\mathbf{y}_j\,|\Psi\right)=\sum_{m=1}^{g}\pi_m\,f_m\left(\mathbf{y}_j\,|\theta_m\right)\quad\text{e}\quad f\left(\mathbf{y}_j\,|\Psi^*\right)=\sum_{m=1}^{g^*}\pi_m^*\,f_m\left(\mathbf{y}_j\,|\theta_m^*\right)\tag{4.9}$$

duas quaisquer funções massa de probabilidade pertencentes a uma classe de misturas finitas de distribuições paramétricas. Esta classe de misturas finitas diz-se **identificável** para $\Psi\in\Omega$ se

$$f\left(\mathbf{y}_j\,|\Psi\right)=f\left(\mathbf{y}_j\,|\Psi^*\right),\tag{4.10}$$

se e só se, $g=g^*$ e ainda for possível permutar os índices das componentes de modo a que

$$\pi_m=\pi_m^*\quad\text{e}\quad f_m\left(\mathbf{y}_j\,|\theta_m\right)=f_m\left(\mathbf{y}_j\,|\theta_m^*\right)\,,m=1,2,\ldots,g\,.\tag{4.11}$$

A modelação incorrecta de uma mistura de $g-1$ componentes por uma mistura de g componentes pode ser tratada de duas maneiras:

(1) Um dos pesos na mistura de g-componentes pode ser igualado a zero;

(2) Duas componentes na mistura de g-componentes podem ser encaradas como sendo a mesma.

4.1.2 Estimação de máxima verosimilhança

Partindo do pressuposto que $\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_n$ são realizações independentes do vector $\mathbf{Y}$, a função de verosimilhança para o vector Ψ dos parâmetros da mistura (4.1) é dada por

$$L\left(\Psi\right)=\prod_{j=1}^{n}f\left(\mathbf{y}_j\,|\Psi\right)$$

e a função log-verosimilhança por

$$\log L\left(\Psi\right)=\sum_{j=1}^{n}\log\left(\sum_{m=1}^{g}\pi_m f_m(\mathbf{y}_j\,|\theta_m)\right).\tag{4.12}$$

O método da máxima verosimilhança consiste na maximização da verosimilhança $L\left(\Psi\right)$ como uma função de Ψ, sobre o espaço paramétrico Ω definido em (4.2), ou seja,

$$\frac{\partial L\left(\Psi\right)}{\partial\Psi}=\mathbf{0}.$$

Equivalentemente[2],

$$\frac{\partial \log L(\Psi)}{\partial \Psi} = \mathbf{0}. \tag{4.13}$$

O objectivo da estimação de máxima verosimilhança é determinar uma estimativa $\hat{\Psi}$ para cada n, de modo a que se defina uma sequência de raízes de (4.13) que seja consistente e assimptoticamente eficiente[3]. Sabe-se que tal sequência existe sob condições de regularidade apropriadas. Com probabilidade tendendo para 1, esta sequência de raízes corresponde ao máximo local no interior do espaço paramétrico. Para modelos de estimação em geral, a verosimilhança tem habitualmente um máximo global no interior do espaço paramétrico. Então, uma sequência de raízes da equação de verosimilhança com as propriedades assimptoticamente desejadas é obtida considerando-se $\hat{\Psi}$ para cada n como sendo a raiz que maximiza globalmente a função de verosimilhança $L(\Psi)$, isto é, $\hat{\Psi}$ é o maximizador global da verosimilhança. Nestas condições, diz-se que $\hat{\Psi}$ é o estimador de máxima verosimilhança [30].

Derivando parcialmente a equação (4.12) em relação aos parâmetros π_m e θ_m, as equações de verosimilhança vêm na forma

$$\frac{\partial \log L(\Psi)}{\partial \pi_m} = \sum_{j=1}^{n} \left\{ \frac{f_m(\mathbf{y}_j|\theta_m)}{f(\mathbf{y}_j|\hat{\Psi})} - \frac{f_g(\mathbf{y}_j|\theta_g)}{f(\mathbf{y}_j|\hat{\Psi})} \right\},\ m = 1, 2, \ldots, g-1, \tag{4.14}$$

$$\frac{\partial \log L(\Psi)}{\partial \theta_m} = \sum_{j=1}^{n} \frac{\pi_m}{f(\mathbf{y}_j|\hat{\Psi})} \frac{\partial f_m}{\partial\, \theta_m} \left(\mathbf{y}_j\,|\theta_m\right),\ m = 1, 2, \ldots, g. \tag{4.15}$$

Igualando as equações (4.14) e (4.15) a zero, não é imediata a obtenção da solução explícita para a estimativa de máxima verosimilhança

$$\hat{\Psi} = \left(\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_{g-1}, \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_g\right)^T.$$

Ao longo dos anos, uma grande variedade de métodos têm sido usados para estimar os parâmetros de misturas de distribuições, tais como, por exemplo, métodos gráficos, o

[2] Dado que $L(\Psi) > 0$ e a função logaritmo é monótona crescente, a maximização de $L(\Psi)$ equivale à maximização da função $\log L(\Psi)$.

[3] A demonstração de que a sequência de raízes da equação de verosimilhança é consistente e assimptoticamente eficiente pode ser encontrada em [26].

método dos momentos, o método por distância mínima[4], o método de Newton-Raphson, o método da máxima verosimilhança e abordagens Bayesianas [30]. Neste trabalho, usar-se-á o método da máxima verosimilhança, via algoritmo EM.

4.2 Algoritmo EM em modelos de misturas

O algoritmo EM assumiu ao longo do tempo um papel de crescente importância no conjunto de ferramentas disponíveis em Estatística Computacional, sendo amplamente utilizado em quase todos os campos onde se recorrem a técnicas estatísticas [31]. É um algoritmo muito usado no cálculo iterativo de estimativas de máxima verosimilhança nos modelos de misturas finitas. A formulação geral deste algoritmo e das suas propriedades básicas foi realizada por Dempster, Laird e Rubin, no seu trabalho de 1977 [12], apesar de antes da sua publicação já terem sido desenvolvidos e aplicados algoritmos semelhantes em várias situações [25].

A metodologia do algoritmo EM consiste em reformular o problema de dados incompletos num problema de dados completos, estabelecendo uma relação entre as funções de verosimilhança destes dois problemas. Embora inicialmente um problema possa não aparentar ser um problema de dados incompletos, poderá ser vantajoso formulá-lo artificialmente como tal, a fim de facilitar a estimação de máxima verosimilhança. Isto deve-se ao facto de o algoritmo EM explorar a redução na complexidade da estimação de máxima verosimilhança quando aplicado aos dados completos [31].

4.2.1 Estrutura de dados incompletos

Antes de se definir a estrutura de dados incompletos para o problema de misturas, apresenta-se a seguir a metodologia para a geração de vectores pseudo-aleatórios de uma mistura de funções massa de probabilidade.

Uma forma de se gerar um vector aleatório $\mathbf{Y}_j$ da função massa de probabilidade (4.1) consiste em considerar uma variável aleatória categorizada Z_j que assuma os valores $1, 2, \ldots, g$, com probabilidades $\pi_1, \pi_2, \ldots, \pi_g$, respectivamente, e supor que a função massa de probabilidade condicional de $\mathbf{Y}_j$, dado $Z_j = m$, é $f_m(\mathbf{y}_j\,|\theta_m)$, $m = 1, 2, \ldots, g$. Então,

[4] Uma maneira de estimar o vector Ψ num modelo de mistura é usando o valor de Ψ que minimiza a distância entre a distribuição da mistura F_Ψ e a distribuição empírica $\hat{F}_n$, $\delta(\hat{F}_n, F_\Psi)$.

a função massa de probabilidade marginal de $\mathbf{Y}_j$ será dada por (4.1). De facto,

$$\begin{aligned} P\left(\mathbf{Y}_j = \mathbf{y}_j\right) &= \sum_{m=1}^{g} P\left(\mathbf{Y}_j = \mathbf{y}_j, Z_j = m\right) \\ &= \sum_{m=1}^{g} P\left(Z_j = m\right) P\left(\mathbf{Y}_j = \mathbf{y}_j \,|Z_j = m\right) \\ &= \sum_{m=1}^{g} \pi_m f_m(\mathbf{y}_j \,|\theta_m) \\ &= f\left(\mathbf{y}_j \,|\Psi\right). \end{aligned}$$

Neste contexto, a variável Z_j pode ser interpretada como uma variável latente do vector $\mathbf{Y}_j$, indicando a componente da qual o vector $\mathbf{Y}_j$ é proveniente. Em vez de se considerar a variável aleatória categorizada Z_j, é conveniente trabalhar com um vector g-dimensional $\mathbf{Z}_j = (Z_{1j}, Z_{2j}, \ldots, Z_{gj})^T$ onde o m-ésimo elemento de $\mathbf{Z}_j$ é definido por

$$Z_{mj} = \begin{cases} 1, & \text{se } \mathbf{Y}_j \text{ pertence à componente } f_m \\ 0, & \text{caso contrário} \end{cases} \tag{4.16}$$

Uma vez que cada vector $\mathbf{Y}_j$ provem exactamente de uma componente, tem-se que

$$\sum_{m=1}^{g} Z_{mj} = 1\,.$$

Atendendo a (4.16), $\mathbf{Z}_j$ segue uma distribuição multinomial

$$\mathbf{Z}_j \sim Mult_g\left(1, \pi\right),\ \pi = (\pi_1, \pi_2, \ldots, \pi_g)^T,$$

onde

$$P(\mathbf{Z}_j = \mathbf{z}_j) = \pi_1^{z_{1j}} \pi_2^{z_{2j}} \ldots \pi_g^{z_{gj}}. \tag{4.17}$$

Uma situação onde o modelo da mistura de distribuições (4.1) é directamente aplicável sucede quando o vector $\mathbf{Y}_j$ é extraído de uma população G constituída por g grupos $G_1, G_2, \ldots, G_g$, nas proporções $\pi_1, \pi_2, \ldots, \pi_g$, respectivamente.

O vector $\mathbf{Z}_j$ é um vector de dados que não é observável e tem apenas como finalidade associar à j-ésima observação da amostra uma das g componentes da mistura. O conceito da existência deste vector como uma variável latente do vector $\mathbf{Y}_j$ é muito útil, apesar de, em termos físicos, nem sempre ser apropriado ver o modelo de mistura neste sentido.

Será visto que esta conceptualização do modelo de mistura em termos de $\mathbf{Y}_j$ e $\mathbf{Z}_j$ é extremamente útil, na medida que permite a estimação de máxima verosimilhança através do algoritmo EM [30].

Considerem-se $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ como sendo as n realizações dos vectores aleatórios i.i.d. $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ com função massa de probabilidade comum $f(\mathbf{y}_j)$ dada por (4.1). Então,

$$\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n \overset{i.i.d.}{\sim} F\,,$$

onde $F(\mathbf{y}_j)$ representa a função de distribuição correspondente à função massa de probabilidade $f(\mathbf{y}_j)$. No âmbito da infraestrutura do algoritmo EM, as realizações $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ são vistas como sendo incompletas pois os vectores $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$, indicadores de componentes, são não observáveis. Deste modo, o vector de dados completo é definido por

$$\mathbf{x}_c = \left(\mathbf{y}^T, \mathbf{z}^T\right)^T, \tag{4.18}$$

onde $\mathbf{y} = \left(\mathbf{y}_1^T, \mathbf{y}_2^T, \ldots, \mathbf{y}_n^T\right)^T$ é o vector dos dados observados (ou dados incompletos) e $\mathbf{z} = \left(\mathbf{z}_1^T, \mathbf{z}_2^T, \ldots, \mathbf{z}_n^T\right)^T$ é o vector não observável das variáveis indicadoras de componentes. Os vectores $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ são realizações dos vectores aleatórios $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n$, para os quais, sob a hipótese de independência, é apropriado assumir [30] que seguem uma distribuição multinomial

$$\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n \overset{i.i.d.}{\sim} Mult_g\left(1, \pi\right),\ \pi = \left(\pi_1, \pi_2, \ldots, \pi_g\right)^T.$$

A m-ésima proporção da mistura, π_m, pode ser interpretada como a probabilidade *a priori* de que a observação pertença à m-ésima componente da mistura ($m = 1, 2, \ldots, g$), enquanto que a probabilidade *a posteriori* de que a observação pertença à m-ésima componente, sabendo que $\mathbf{y}_j$ já foi observado, é dada por

$$\begin{aligned}
\tau_m(\mathbf{y}_j|\Psi) &= P\left(\text{observação} \in m\text{−ésima componente} \,\middle|\, \mathbf{y}_j\right) \\
&= P\left(Z_{mj} = 1 \,\middle|\, \mathbf{y}_j\right) \\
&= \frac{P\left(Z_{mj} = 1, \mathbf{Y}_j = \mathbf{y}_j\right)}{P\left(\mathbf{Y}_j = \mathbf{y}_j\right)} \\
&= \frac{\pi_m f_m(\mathbf{y}_j|\theta_m)}{f\left(\mathbf{y}_j|\Psi\right)},\ m = 1, 2, \ldots g;\ j = 1, 2, \ldots, n.
\end{aligned} \tag{4.19}$$

Se z_{mj} fosse observável, então a estimativa de máxima verosimilhança de π_m (considerando

os dados completos) seria dada por

$$\hat{\pi}_m = \sum_{j=1}^{n} \frac{z_{mj}}{n},\ m = 1, 2, \ldots, g \tag{4.20}$$

e a estimativa de θ_m poderia ser obtido a partir das observações pertencentes à m-ésima componente.

4.2.2 Formulação do algoritmo

No seguimento da secção anterior, sendo $\mathbf{y} = \left(\mathbf{y}_1^T, \mathbf{y}_2^T, \ldots, \mathbf{y}_n^T\right)^T$ o vector dos dados observados ou incompletos, $\mathbf{x}_c = \left(\mathbf{y}^T, \mathbf{z}^T\right)^T$ o vector de dados completos definido em (4.18) e $\Psi = \left(\pi^T, \xi^T\right)^T$ o vector dos parâmetros desconhecidos do modelo de mistura (4.1), a função de verosimilhança dos dados completos é definida por

$$\begin{aligned} L_c(\Psi) &= \prod_{j=1}^{n} P\left(\mathbf{Y}_j = \mathbf{y}_j, \mathbf{Z}_j = \mathbf{z}_j\right) \\ &= \prod_{j=1}^{n} P\left(\mathbf{Z}_j = \mathbf{z}_j\right) P\left(\mathbf{Y}_j = \mathbf{y}_j \,|\mathbf{Z}_j = \mathbf{z}_j\right) \\ &= \prod_{j=1}^{n} \pi_1^{z_{1j}} \pi_2^{z_{2j}} \ldots \pi_g^{z_{gj}} f_1(\mathbf{y}_j\,|\theta_1)^{z_{1j}} f_2(\mathbf{y}_j\,|\theta_2)^{z_{2j}} \ldots f_g(\mathbf{y}_j\,|\theta_g)^{z_{gj}} \\ &= \prod_{j=1}^{n} \prod_{m=1}^{g} \left(\pi_m f_m(\mathbf{y}_j\,|\theta_m)\right)^{z_{mj}}. \end{aligned} \tag{4.21}$$

A função log-verosimilhança completa, atendendo às propriedades dos logaritmos, vem na forma

$$\log L_c(\Psi) = \sum_{j=1}^{n} \sum_{m=1}^{g} z_{mj} \left[\log(\pi_m) + \log\left(f_m(\mathbf{y}_j\,|\theta_m)\right)\right]. \tag{4.22}$$

O algoritmo EM lida indirectamente com o problema de resolver a equação log-verosimilhança dos dados incompletos (4.13), procedendo iterativamente em termos da função log-verosimilhança dos dados completos $\log L_c(\Psi)$ [30]. A seguir apresenta-se a variante do algoritmo EM que irá ser implementada.

Seja $\Psi^{(0)} = \left(\pi_1^0, \pi_2^0, \ldots, \pi_g^0, \theta_1^0, \theta_2^0, \ldots, \theta_g^0\right)^T$ o valor inicial da estimativa de Ψ e $\Psi^{(k)}$ o valor aproximado da estimativa de Ψ obtido na k-ésima iteração do algoritmo.

- 1ª iteração do algoritmo EM:

  Passo-E: calcular a esperança matemática condicional da função log-verosimilhança completa (4.22), dado o vector dos dados observados ou incompletos $\mathbf{y}$, ou seja,

  $$Q\left(\Psi;\Psi^{(0)}\right) = E_{\Psi^{(0)}}\left[\log L_c\left(\Psi\right)|\,\mathbf{y}\right].$$

  Passo-M: escolher $\Psi^{(1)}$ como sendo um valor de $\Psi \in \Omega$ que maximiza $Q\left(\Psi;\Psi^{(0)}\right)$, ou seja, escolher $\Psi^{(1)}$ tal que

  $$Q\left(\Psi^{(1)};\Psi^{(0)}\right) \geq Q\left(\Psi;\Psi^{(0)}\right), \forall\,\Psi \in \Omega.$$

  Na iteração seguinte, o valor de $\Psi^{(0)}$ é substituído por $\Psi^{(1)}$.

- $(k+1)$-ésima iteração do algoritmo EM:

  Passo-E: calcular $Q\left(\Psi;\Psi^{(k)}\right)$, onde

  $$Q\left(\Psi;\Psi^{(k)}\right) = E_{\Psi^{(k)}}\left[\log L_c\left(\Psi\right)|\mathbf{y}\right]. \tag{4.23}$$

  No contexto do modelo de misturas, e como a função $\log L_c\left(\Psi\right)$ é linear na variável não observável z_{mj}, neste passo do algoritmo apenas será necessário calcular a esperança condicional de Z_{mj} dado o vector de dados observados $\mathbf{y}$, sendo Z_{mj} a variável aleatória correspondente a z_{mj}. Deste modo,

  $$\begin{aligned} E_{\Psi^{(k)}}\left[\log L_c\left(\Psi\right)|\mathbf{y}\right] &= E_{\Psi^{(k)}}\left[Z_{mj}\,|\,\mathbf{y}\right] \\ &= P_{\Psi^{(k)}}\left(Z_{mj}=1\,|\,\mathbf{y}\right) \\ &= z_{mj}^{(k)}, \end{aligned}$$

  onde $z_{mj}^{(k)}$ é a probabilidade *a posteriori* da j-ésima observação pertencer à m-ésima componente da mistura, f_m. Atendendo à fórmula (4.19), tem-se

  $$z_{mj}^{(k)} = \tau_m(\mathbf{y}_j\,|\Psi^{(k)}) = \frac{\pi_m^{(k)} f_m(\mathbf{y}_j\,|\theta_m^{(k)})}{f\left(\mathbf{y}_j\,|\Psi^{(k)}\right)},\ m=1,2,\ldots g,\ \ j=1,2,\ldots,n. \tag{4.24}$$

  Deste modo, no passo-E estabelece-se que

  $$Q\left(\Psi;\Psi^{(k)}\right) = \sum_{j=1}^{n}\sum_{m=1}^{g} z_{mj}^{(k)}\left[\log\left(\pi_m\right) + \log\left(f_m(\mathbf{y}_j\,|\theta_m)\right)\right]. \tag{4.25}$$

Passo-M: escolher $\Psi^{(k+1)}$ como sendo um valor de $\Psi \in \Omega$ que maximiza $Q\left(\Psi; \Psi^{(k)}\right)$, ou seja, escolher $\Psi^{(k+1)}$ tal que

$$Q\left(\Psi^{(k+1)}; \Psi^{(k)}\right) \geq Q\left(\Psi; \Psi^{(k)}\right), \forall\, \Psi \in \Omega.$$

No contexto do modelo de misturas, as estimativas $\pi_m^{(k+1)}$ das proporções da mistura π_m são calculadas independentemente das estimativas $\xi^{(k+1)}$ do vector ξ dos parâmetros das componentes da mistura. Como no passo-E apenas se substitui z_{mj} pela sua correspondente esperança condicional $z_{mj}^{(k)}$ na função log-verosimilhança completa, então a estimativa actualizada de π_m é dada pela substituição de cada z_{mj} por $z_{mj}^{(k)}$ na expressão (4.20), obtendo-se

$$\pi_m^{(k+1)} = \sum_{j=1}^{n} \frac{z_{mj}^{(k)}}{n},\ m = 1, 2, \ldots, g. \tag{4.26}$$

Assim, no cálculo da estimativa de π_m na $(k+1)$-ésima iteração do algoritmo haverá, de cada observação $\mathbf{y}_j$, uma contribuição igual à sua probabilidade *a posteriori* de pertencer à m-ésima componente do modelo de mistura.

No que diz respeito à actualização da estimativa do vector ξ na $(k+1)$-ésima iteração, decorre de (4.25) que $\xi_m^{(k+1)}$ será uma raiz de

$$\sum_{j=1}^{n} \sum_{m=1}^{g} z_{mj}^{(k)} \frac{\partial \log\left(f_m(\mathbf{y}_j\,|\theta_m)\right)}{\partial \xi} = \mathbf{0}. \tag{4.27}$$

Os passos -E e -M são alternados repetidamente, até que se verifique convergência, de acordo com o critério

$$L\left(\Psi^{(k+1)}\right) - L\left(\Psi^{(k)}\right) < \varepsilon,$$

onde ε é um valor arbitrariamente pequeno previamente fixado. Dempster *et al.* [12] mostraram que a função de verosimilhança (dados incompletos) é não decrescente em cada iteração do algoritmo EM [31], ou seja,

$$L\left(\Psi^{(k+1)}\right) \geq L\left(\Psi^{(k)}\right), \tag{4.28}$$

para $k = 0, 1, 2, \ldots$. A função de verosimilhança será crescente se a desigualdade (4.28) se verificar no sentido estrito. Assim, para uma sequência limitada superiormente de valores da verosimilhança $\left\{L\left(\Psi^{(k)}\right)\right\}_{k\in\mathbb{N}_0}$, $L\left(\Psi^{(k)}\right)$ converge monotonamente para algum L^*. Em quase todas as aplicações, L^* é um valor estacionário,

ou seja, $L^* = L(\Psi^*)$, para algum ponto Ψ^* em que

$$\frac{\partial \log L(\Psi)}{\partial \Psi} = 0.$$

Algumas propriedades do algoritmo

O algoritmo EM possui algumas propriedades que o tornam atraente [31], entre as quais as seguintes:

- é numericamente estável, aumentando a verosimilhança em cada iteração;
- a sua convergência é fiável sob condições bastante gerais, pois a convergência ocorre quase sempre para um máximo local, mesmo quando a escolha de $\Psi^{(0)}$ não é a melhor;
- a sua implementação é relativamente fácil, quer analiticamente quer computacionalmente. Em particular, é geralmente fácil de programar e requer pouco espaço de armazenamento. Observando o crescimento monótono da verosimilhança durante o processo iterativo, é fácil monitorizar a convergência e os erros de programação;
- o custo por iteração é geralmente baixo, o que pode minorar a importância do facto de o algoritmo EM necessitar geralmente de um número de iterações mais elevado que outros algoritmos;
- pode ser usado para estimar parâmetros de dados incompletos.

O algoritmo EM também possui algumas fraquezas, entre as quais as seguintes:

- por vezes é muito lento a convergir, mesmo quando aplicado a problemas aparentemente simples e a problemas onde exista muita informação em falta;
- em alguns casos, pode não ser possível tratar analiticamente o passo-E.

Aplicação do algoritmo a dados categorizados

Com o objectivo de ajustar um modelo de mistura de distribuições geométricas (4.5) à distribuição empírica da distância global entre nucleótidos, será descrito nesta secção o algoritmo EM para a estimação de máxima verosimilhança do vector dos parâmetros desse modelo. O desenvolvimento teórico apresentado a seguir será feito com base na Secção 4.2.2. Tomar-se-á em consideração o facto de os dados observados aos quais se irá aplicar o algoritmo EM já se encontrarem categorizados[5]:

$$\begin{array}{c|cccc} y & 1 & 2 & \cdots & L \\ \hline f_y & f_1 & f_2 & \cdots & f_L \end{array},$$

onde f_y representa a frequência absoluta da distância y e

$$\sum_{y=1}^{L} f_y = N\,. \tag{4.29}$$

Substituindo em (4.21) as componentes da mistura (4.4), a função verosimilhança completa, para dados agrupados em categorias, é dada por

$$L_c\left(\Psi\right) = \prod_{y=1}^{L}\prod_{m=1}^{g}\left[\left(\pi_m p_m\left(1-p_m\right)^{y-1}\right)^{z_{my}}\right]^{f_y} \tag{4.30}$$

e a função log-verosimilhança completa vem na forma

$$\begin{aligned}\log L_c\left(\Psi\right) &= \sum_{y=1}^{L}\sum_{m=1}^{g} z_{my} f_y\left[\log\left(\pi_m\right) + \log\left(p_m\right) + \left(y-1\right)\log\left(1-p_m\right)\right] \\ &= \sum_{m=1}^{g}\left[n_m\left(\log\left(\pi_m\right) + \log\left(p_m\right)\right) + \left(\sum_{y=1}^{L} z_{my}\, y f_y - n_m\right)\log\left(1-p_m\right)\right],\end{aligned}$$

onde

$$n_m = \sum_{y=1}^{L} z_{my}\, f_y.$$

Para a mistura de g distribuições geométricas, o passo-E na $(k+1)$-ésima iteração do algoritmo EM (4.25) pode ser escrito como

$$Q\left(\Psi;\Psi^{(k)}\right) = \sum_{m=1}^{g}\left[n_m^{(k)}\left(\log\left(\pi_m\right) + \log\left(p_m\right)\right) + \left(\sum_{y=1}^{L} z_{my}^{(k)}\, y f_y - n_m^{(k)}\right)\log\left(1-p_m\right)\right], \tag{4.31}$$

[5] Os dados observados dizem respeito à sequência de distâncias global de cada espécie em estudo.

onde

$$n_{m}^{(k)} = \sum_{y=1}^{L} z_{my}^{(k)} f_y \tag{4.32}$$

e, atendendo à expressão (4.24), tem-se que

$$z_{my}^{(k)} = \frac{\pi_m^{(k)} p_m^{(k)} \left(1 - p_m^{(k)}\right)^{y-1}}{f\left(y \,|\Psi^{(k)}\right)} = \frac{\pi_m^{(k)} p_m^{(k)} \left(1 - p_m^{(k)}\right)^{y-1}}{\sum_{h=1}^{g} \pi_h^{(k)} p_h^{(k)} \left(1 - p_h^{(k)}\right)^{y-1}},\ m = 1, 2, \ldots, g. \tag{4.33}$$

No passo-M, as estimativas actualizadas das proporções da mistura, atendendo a (4.26), (4.29) e a (4.32), são dadas por

$$\pi_m^{(k+1)} = \frac{n_{m}^{(k)}}{N} = \frac{\sum_{y=1}^{L} z_{my}^{(k)} f_y}{\sum_{y=1}^{L} f_y}, \tag{4.34}$$

e as estimativas actualizadas dos parâmetros das componentes da mistura, derivando a equação (4.31) em ordem a p_m e igualando a zero, vêm dadas por

$$p_{m}^{(k+1)} = \frac{n_{m}^{(k)}}{\sum_{y=1}^{L} z_{my}^{(k)}\, y f_y}\,,\ m = 1, 2, \ldots, g\,. \tag{4.35}$$

4.2.3 Resultados experimentais

Nesta secção são apresentados os resultados da aplicação do algoritmo EM à determinação das estimativas dos parâmetros da mistura de duas, três e quatro distribuições geométricas, para a espécie *St.* Para as restantes espécies, os resultados da estimação dos parâmetros da mistura de quatro geométricas pelo algoritmo EM encontram-se na Secção 4.4.

A escolha das estimativas iniciais, no caso da mistura de duas e quatro distribuições, foi feita com base na frequência relativa $\hat{p}^x$ da ocorrência do nucleótido do tipo $x \in \mathcal{A}$, enquanto que no caso da mistura de três distribuições a escolha foi arbitrária. O critério de paragem do algoritmo EM baseou-se na diferença entre valores consecutivos da função log-verosimilhança dos dados observados, ou seja,

$$\left|\log L\left(\Psi^{(k+1)}\right) - \log L\left(\Psi^{(k)}\right)\right| < \varepsilon\,, \tag{4.36}$$

onde ε é um valor fixo pré-definido. Para além das estimativas dos parâmetros das misturas, são também apresentados o gráfico da sequência de valores da função log-verosimilhança,

$$\left\{\log L\left(\Psi^{(k)}\right)\right\}_{k\in\mathbb{N}}$$

e o gráfico das distribuições resultantes das misturas.

Mistura de duas distribuições geométricas - espécie *St*

Na Tabela 4.1 são apresentadas as estimativas de máxima verosimilhança para os parâmetros de uma mistura de duas distribuições geométricas. O cálculo dos valores iniciais para a aplicação do algoritmo EM foi baseado na fórmula (2.11), tendo-se atendido ao facto de os nucleótidos {A,T} e {C,G} estarem presentes no genoma na mesma proporção. Assim sendo, consideraram-se para as componentes da mistura os valores $p_1^{(0)} = \hat{p}^{\mathrm{A}} = 0.3021$ e $p_2^{(0)} = \hat{p}^{\mathrm{C}} = 0.1982$. Aos pesos da mistura, uma vez que a probabilidade de ocorrência de qualquer nucleótido na sequência de ADN é a mesma, foram atribuídos os valores $\pi_1^{(0)} = 0.5$ e $\pi_2^{(0)} = 0.5$. O critério (4.36) foi aplicado com $\varepsilon = 10^{-5}$.

Tabela 4.1: Resultados do algoritmo EM para uma mistura de duas distribuições geométricas. Os dados observados dizem respeito à espécie *St*.

Iter.(k)	$\pi_1^{(k)}$	$\pi_2^{(k)}$	$p_1^{(k)}$	$p_2^{(k)}$	Critério
0	0.5	0.5	0.3021	0.1982	...
1	0.5062	0.4938	0.3175	0.2052	3439.93
...	...	...	...	...	...
10	0.5113	0.4887	0.3487	0.1929	31.6274
...	...	...	...	...	...
50	0.5050	0.4950	0.3555	0.1919	0.3616
...	...	...	...	...	...
1000	0.3719	0.6281	0.4019	0.2043	0.1229
...	...	...	...	...	
4000	0.2240	0.7760	0.5075	0.2181	0.00697
...	...	...	...	...	...
8631	0.1889	0.8111	0.5549	0.2216	9.9e-05
...	...	...	...	...	...
11033	0.1858	0.8142	0.5600	0.222	1.2e-05
...	...	...	...	...	...
11277	0.1857	0.8143	0.5602	0.2220	9.99e-06

Como se pode ver na Tabela 4.1 o algoritmo foi relativamente lento a satisfazer o critério estabelecido, o que apenas sucedeu na iteração 11277. De acordo com a Figura 4.2, que diz respeito à sequência de valores da função log-verosimilhança dos dados observados, verifica-se que a função log-veromilhança é monótona crescente em cada iteração[6] e que a partir de certa altura ocorre uma estabilização na evolução das estimativas.

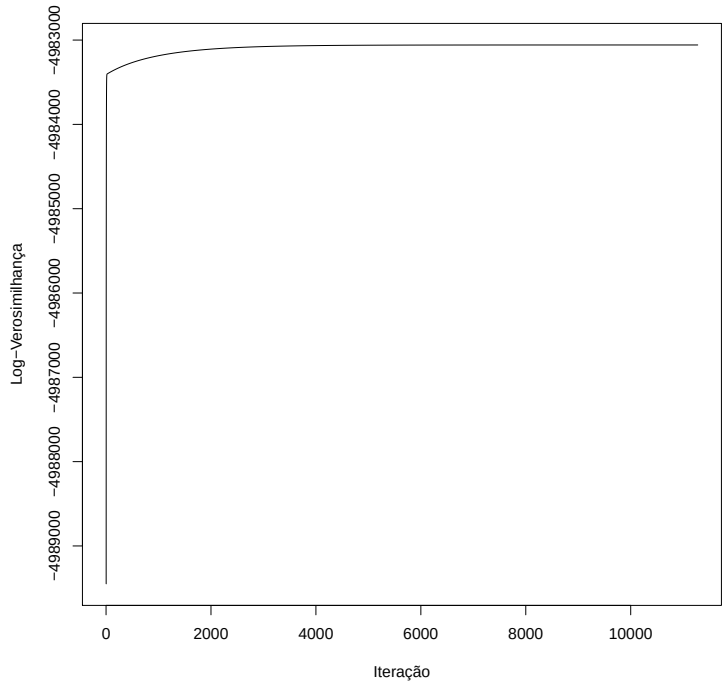


Figura 4.2: Sequência de valores da função log-verosimilhança $\{\log L\left(\Psi^{(k)}\right)\}_{k\in\mathbb{N}}$ para a mistura de duas geométricas, referente à espécie *St*.

A Figura 4.3 mostra uma representação da distribuição empírica[7], juntamente com a curva (linha azul) da distribuição modelo (2.13) e as curvas que resultaram da mistura de duas distribuições geométricas com parâmetros diferentes (linha vermelha e linha verde).

[6] Como referido na Secção 4.2.2, sucede quando existe convergência do algoritmo EM.

[7] Apenas são apresentadas as primeiras 25 distâncias, por uma questão de melhor visualização.

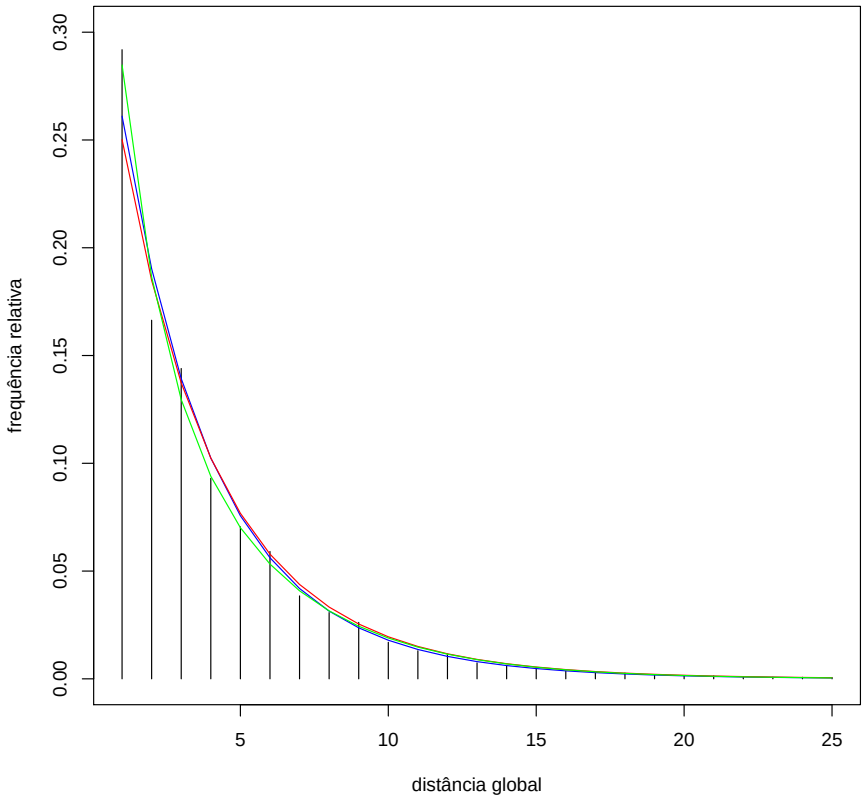


Figura 4.3: Mistura de duas distribuições geométricas. A curva a azul corresponde à distribuição modelo (2.13). A curva a vermelho corresponde ao vector dos parâmetros $\hat{\Psi} = (\pi^{(0)}, p^{(0)})$, onde $\pi^{(0)} = (0.5, 0.5)$ e $p^{(0)} = (0.3021, 0.1982)$. A curva a verde corresponde ao vector dos parâmetros $\hat{\Psi} = (\pi^{(11277)}, p^{(11277)})$, obtido a partir do algoritmo EM, onde $\pi^{(11277)} = (0.1857, 0.8143)$ e $p^{(11277)} = (0.5602, 0.2220)$.

A curva a vermelho foi obtida considerando os valores iniciais do algoritmo EM como sendo estimativas dos parâmetros da mistura. A curva a verde resultou das estimativas dos parâmetros obtidas na iteração 11277 do algoritmo EM. Pela observação da Figura 4.3, constata-se que há uma ligeira diferença entre a distribuição modelo (2.13) e as misturas de distribuições de quatro geométricas. A distribuição que aparenta ajustar-se melhor à distribuição empírica é aquela que resultou da aplicação do algoritmo EM. Apesar das diferenças, visualmente poderá concluir-se que estas distribuições são ambas uma aproximação relativamente razoáveis da distribuição empírica.

Mistura de três distribuições geométricas - espécie *St*

Na Tabela 4.2 são apresentadas as estimativas de máxima verosimilhança para os parâmetros de uma mistura de três distribuições geométricas. A determinação dos valores iniciais para a aplicação do algoritmo EM resultou de uma escolha arbitrária. Os valores

escolhidos foram os seguintes:

$$\pi_1^{(0)} = 0.1 \quad \pi_2^{(0)} = 0.2 \quad \pi_3^{(0)} = 0.7 \qquad p_1^{(0)} = 0.1 \quad p_2^{(0)} = 0.5 \quad p_3^{(0)} = 0.7 \tag{4.37}$$

e o critério (4.36) foi aplicado com $\varepsilon = 10^{-5}$.

Constata-se que há uma grande diferença entre a distribuição modelo proposta por [2] e a mistura de distribuições que usa como parâmetros os valores das estimativas iniciais (4.37). Apesar disso, o algoritmo EM acabou sempre por satisfazer o critério estabelecido. Verificou-se que a função log-veromilhança é também monótona crescente em cada iteração (ver Figura 4.4) e que, a partir de certa altura, ocorre uma estabilização na evolução das estimativas. A aplicação do algoritmo EM a este caso é uma ilustração da propriedade que o algoritmo EM possui (ver Secção 4.2.2) de convergir quase sempre para um máximo local da função log-verosimilhança, mesmo quando as estimativas iniciais não são as melhores.

Tabela 4.2: Resultados do algoritmo EM para uma mistura de três distribuições geométricas. Os dados observados dizem respeito à espécie *St*.

Iter.(k)	$\pi_1^{(k)}$	$\pi_2^{(k)}$	$\pi_3^{(k)}$	$p_1^{(k)}$	$p_2^{(k)}$	$p_3^{(k)}$	Critério
0	0.1	0.2	0.7	0.1	0.5	0.7	...
1	0.2657	0.2314	0.5029	0.1204	0.2977	0.4947	766945.1
2	0.2842	0.2445	0.4713	0.1425	0.2632	0.4381	51499.83
...	...	...	...	...	...	...	...
10	0.2982	0.2607	0.4411	0.1760	0.2348	0.3691	50.9377
...	...	...	...	...	...	...	...
3500	0.1	0.8487	0.0583	0.1398	0.2609	1	0.0066
...	...	...	...	...	...	...	...
5008	0.0926	0.8487	0.0589	0.1373	0.2598	1	1.32e-05
...	...	...	...	...	...	...	...
5074	0.0926	0.8487	0.0587	0.1373	0.2598	1	9.98e-06

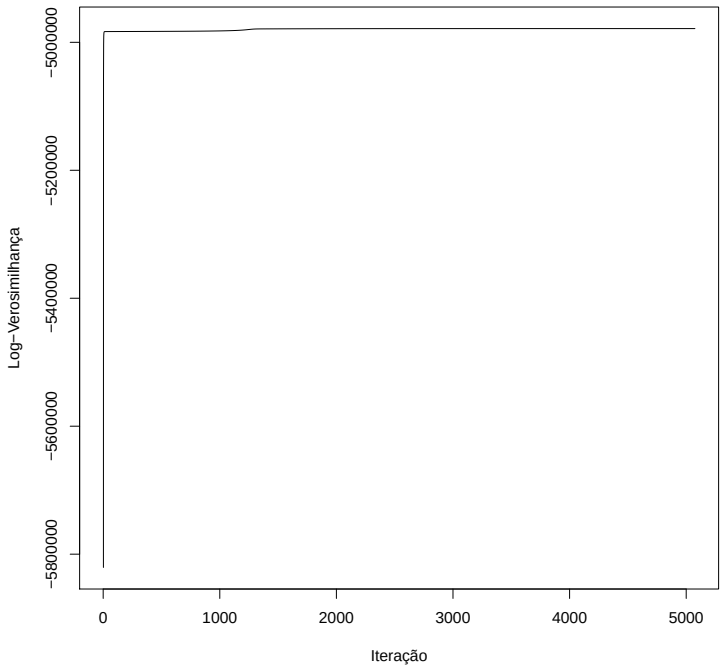


Figura 4.4: Sequência de valores da função log-verosimilhança $\{\log L\left(\Psi^{(k)}\right)\}_{k\in\mathbb{N}}$ para a mistura de três geométricas, referente à espécie *St*.

A Figura 4.5 mostra uma representação da distribuição empírica[8], juntamente com a curva (linha azul) da distribuição modelo (2.13) e as curvas que resultaram da mistura de três distribuições geométricas com parâmetros diferentes (linha vermelha e linha verde). A curva a vermelho foi obtida considerando-se os valores iniciais do algoritmo EM como sendo estimativas dos parâmetros da mistura e a curva a verde resultou das estimativas dos parâmetros obtidas na iteração 5074 do algoritmo EM. A distribuição que graficamente aparenta ajustar-se melhor à distribuição empírica é a mistura de três geométricas que resultou da aplicação do algoritmo EM.

[8] Apenas são apresentadas as primeiras 25 distâncias, por uma questão de melhor visualização.

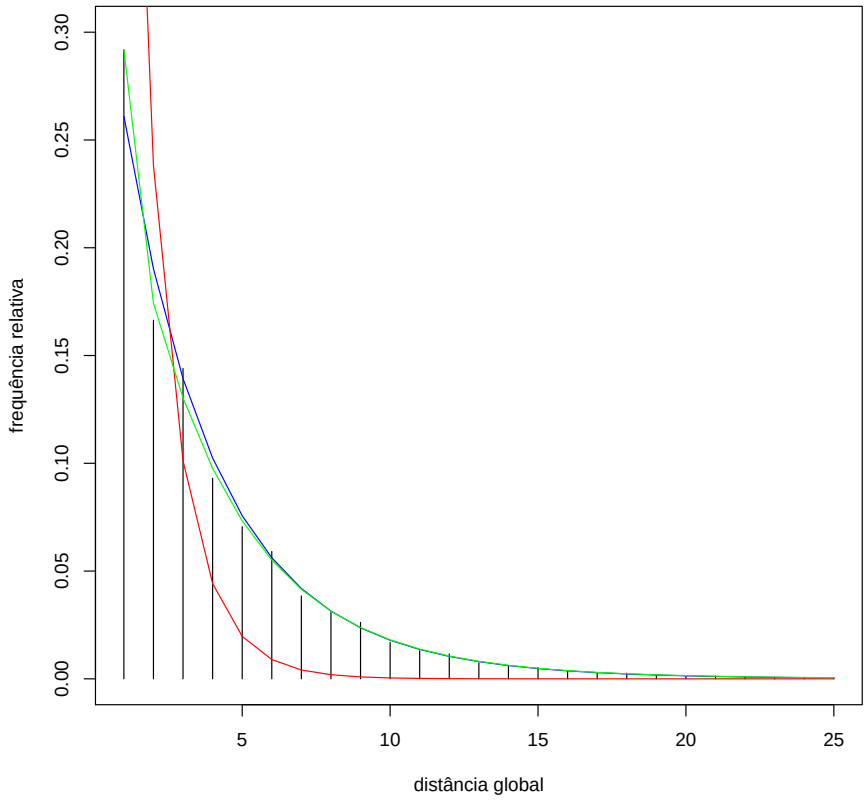


Figura 4.5: Mistura de três distribuições geométricas. A curva a azul corresponde à distribuição modelo (2.13). A curva a vermelho corresponde ao vector dos parâmetros $\hat{\Psi} = (\pi^{(0)}, p^{(0)})$, onde $\pi^{(0)} = (0.1, 0.2, 0.7)$ e $p^{(0)} = (0.1, 0.5, 0.7)$. A curva a verde corresponde ao vector dos parâmetros $\hat{\Psi} = (\pi^{(5074)}, p^{(5074)})$, onde $\pi^{(5074)} = (0.0926, 0.8487, 0.0587)$ e $p^{(5074)} = (0.1373, 0.2598, 1)$, obtido a partir do algoritmo EM.

Mistura de quatro distribuições geométricas - espécie *St*

Na Tabela 4.3 são apresentadas as estimativas de máxima verosimilhança para os parâmetros de uma mistura de quatro distribuições geométricas. O cálculo dos valores iniciais para as componentes da mistura foi baseado na fórmula (2.11). No caso dos pesos da mistura atendeu-se ao facto de a probabilidade de ocorrência dos nucleótidos {A,C,G,T} na sequência de ADN ser a mesma. Assim sendo, consideraram-se para pesos da mistura $\pi_m^{(0)} = 0.25$, $m = 1, 2, 3, 4$ e para as componentes da mistura os valores

$$p_1^{(0)} = \hat{p}^{\mathrm{A}} = 0.3021 \quad p_2^{(0)} = \hat{p}^{\mathrm{C}} = 0.1982 \quad p_3^{(0)} = \hat{p}^{\mathrm{G}} = 0.1967 \quad p_4^{(0)} = \hat{p}^{\mathrm{T}} = 0.3030\,.$$

O critério (4.36) foi aplicado com $\varepsilon = 10^{-5}$, tendo sido satisfeito na iteração 3997. Também neste caso, verifica-se que a função log-veromilhança é monótona crescente em cada

iteração (ver Figura 4.6) e que a partir de certa altura ocorre uma estabilização na evolução das estimativas.

Tabela 4.3: Resultados do algoritmo EM para uma mistura de quatro distribuições geométricas. Os dados observados dizem respeito à espécie *St*.

Iter.(k)	$\pi_1^{(k)}$	$\pi_2^{(k)}$	$\pi_3^{(k)}$	$\pi_4^{(k)}$	$p_1^{(k)}$	$p_2^{(k)}$	$p_3^{(k)}$	$p_4^{(k)}$	Critério
0	0.25	0.25	0.25	0.25	0.3021	0.1982	0.1967	0.3030	...
1	0.2532	0.2468	0.2467	0.2533	0.3178	0.2058	0.2040	0.3187	3614.95
2	0.2537	0.2463	0.2463	0.2537	0.3237	0.2033	0.2013	0.3247	756.19
...	...	...	...	...	...	...	...	...	...
127	0.2451	0.2548	0.2547	0.2454	0.3513	0.2012	0.1864	0.3666	0.9846
128	0.2450	0.2549	0.2548	0.2453	0.3512	0.2012	0.1864	0.3668	1.0033
...	...	...	...	...	...	...	...	...	...
244	0.2365	0.2671	0.2633	0.2331	0.2971	0.2209	0.1809	0.4323	9.6063
...	...	...	...	...	...	...	...	...	...
1000	0.3527	0.4006	0.1911	0.0556	0.2726	0.2726	0.1618	1	0.3058
...	...	...	...	...	...	...	...	...	...
3650	0.3973	0.4512	0.0929	0.0586	0.2599	0.2599	0.1374	1	4.34e-05
...	...	...	...	...	...	...	...	...	...
3997	0.3974	0.4513	0.0926	0.0587	0.2598	0.2598	0.1373	1	9.96e-06

Adicionalmente, verifica-se que a partir da iteração 1000 os valores das estimativas dos parâmetros p_1 e p_2 são iguais. Somando os valores das estimativas dos pesos π_1 e π_2, obtém-se $\pi^{(3997)} = (0.8487, 0.0926, 0.0587)$ e $p^{(3997)} = (0.2598, 0.1373, 1)$. Permutando o primeiro índice com o segundo nestes vectores, conclui-se que esta mistura de quatro geométricas pode ser modelada pela mistura de três geométricas anteriormente obtida (ver Secção 4.1.1). Este facto também pode ser constatado na Figura 4.7, que mostra uma representação da distribuição empírica, juntamente com a curva (linha azul) da distribuição modelo (2.13) e as curvas que resultaram da mistura de quatro distribuições geométricas com parâmetros diferentes (linha vermelha e linha verde). A curva a vermelho foi obtida considerando-se os valores iniciais do algoritmo EM como sendo estimativas dos parâmetros da mistura e a curva a verde resultou das estimativas dos parâmetros obtidas na iteração 3997 do algoritmo EM. Na Figura 4.7 e na Figura 4.5 a curva a verde corresponde à mesma distribuição.

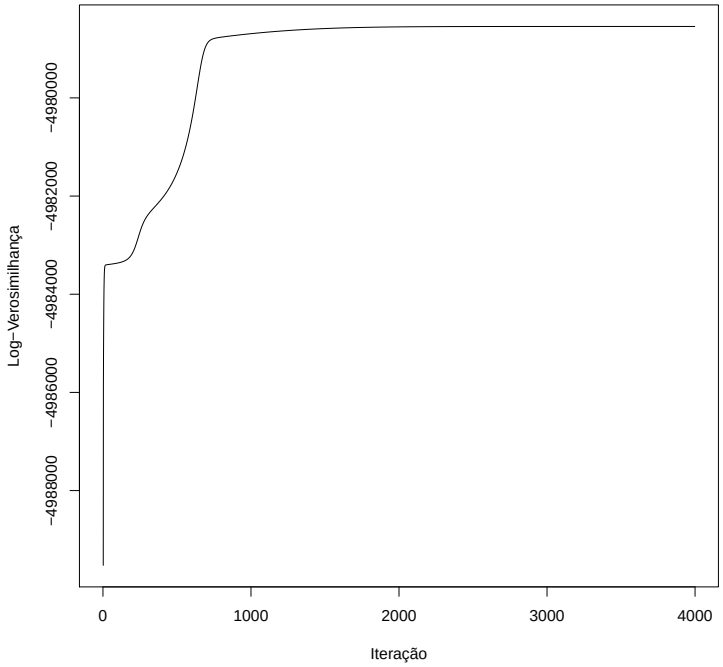


Figura 4.6: Sequência de valores da função log-verosimilhança $\{\log L\left(\Psi^{(k)}\right)\}_{k\in\mathbb{N}}$ para a mistura de quatro geométricas, referente à espécie *St*.

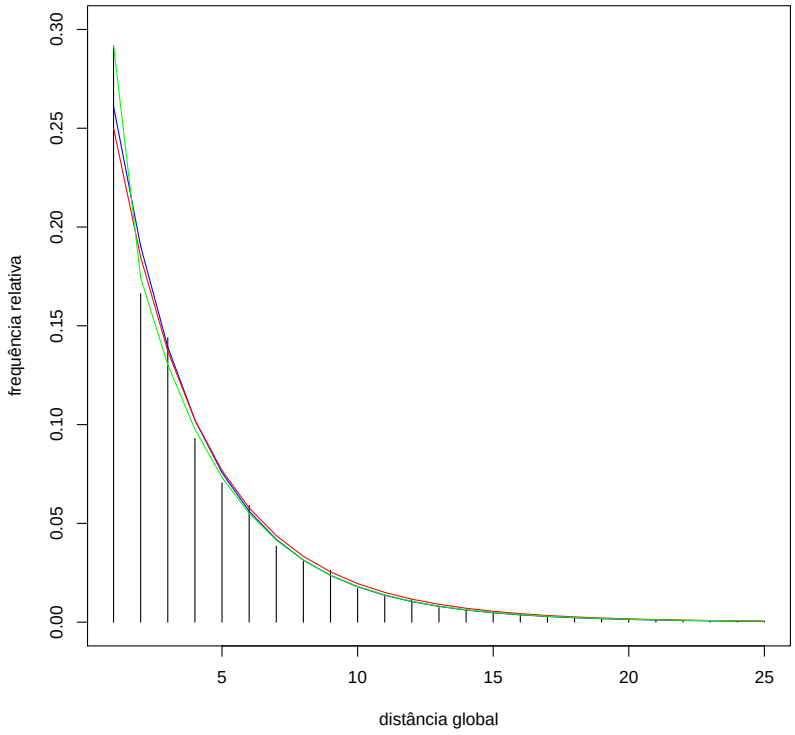


Figura 4.7: Mistura de quatro distribuições geométricas. A curva a azul corresponde à distribuição modelo (2.13). A curva a vermelho corresponde ao vector dos parâmetros $\hat{\Psi} = (\pi^{(0)}, p^{(0)})$, onde $\pi^{(0)} = (0.25, 0.25, 0.25, 0.25)$ e $p^{(0)} = (0.3021, 0.1982, 0.1967, 0.3030)$. A curva a verde corresponde ao vector dos parâmetros $\hat{\Psi} = (\pi^{(3997)}, p^{(3997)})$, onde $\pi^{(3997)} = (0.3974, 0.4513, 0.0926, 0.0587)$ e $p^{(3997)} = (0.2598, 0.2598, 0.1373, 1)$.

4.3 Teste de ajustamento e medidas de similaridade

Em face dos resultados obtidos na Secção 4.2.3, não se exclui a possibilidade de uma das distribuições teóricas $Modelo(p)$, $Mgeom(\Psi^{(0)})$ e $Mgeom(\Psi^{(EM)})$ se ajustar à distribuição empírica da sequência de distâncias global D, onde a primeira dessas distribuições teóricas corresponde à distribuição modelo (2.13) com parâmetro $p = \left(p^{\mathrm{A}}, p^{\mathrm{C}}, p^{\mathrm{G}}, p^{\mathrm{T}}\right)$, a segunda a uma mistura finita de distribuições geométricas com $\Psi^{(0)}$ representando o vector dos parâmetros iniciais do algoritmo EM, e a terceira a uma mistura finita de distribuições geométricas com $\Psi^{(EM)}$ representando o vector dos parâmetros obtido pelo algoritmo EM.

Teste do qui-quadrado

Para testar a qualidade do ajustamento entre a distribuição empírica e, por exemplo, a distribuição $Mgeom(\Psi^{(EM)})$, realizar-se-á o teste de ajustamento do qui-quadrado. As hipóteses a serem testadas são:

$$H_0 : D \sim Mgeom(\Psi^{(EM)}) \;\; vs \;\; H_1 : D \nsim Mgeom(\Psi^{(EM)})$$

Denote-se por:

- L o número de categorias (distâncias) e N o comprimento da sequência D;
- $n_{obs,k}$ o número de observações, ou frequência absoluta observada, da categoria k;
- p_k a probabilidade de se obter uma observação na categoria k, assumindo que a observação foi extraída de uma mistura finita de distribuições geométricas.

A frequência esperada da categoria k, quando a hipótese H_0 é verdadeira, é dada por

$$n_{esp,k} = n\, p_k \,.$$

A estatística de qui-quadrado de *Pearson* é dada por:

$$\chi^2_P = \sum_{k=1}^{L} \frac{\left(n_{obs,k} - n_{esp,k}\right)^2}{n_{esp,k}} \,. \tag{4.38}$$

Sendo verdadeira a hipótese H_0, esta estatística tem distribuição assimptótica de um qui-quadrado com $(L-r-1)$ graus de liberdade, onde r é o número de parâmetros desconhecidos da distribuição especificada em H_0 estimados a partir da amostra. Se a hipótese H_0 for verdadeira, a diferença entre o valor observado e o respectivo valor esperado, $n_{obs,k} - n_{esp,k}$,

não deve ser muito grande e, consequentemente, o valor observado da estatística de teste $\chi^2_{P_{obs}}$ será pequeno. Deste modo, é-se levado a concluir que as frequências observadas são provenientes de uma mistura finita de distribuições geométricas.

Em relação à espécie *St*, tem-se $L = 74$ (número de distâncias com frequência absoluta superior ou igual a 1) e $N = 2221315$. Os valores das probabilidades p_k foram obtidos da mistura

$$p_k = \sum_{m=1}^{3} \pi_m \, p_m (1 - p_m)^{k-1}, \; k = 1, 2, \ldots, L \, . \tag{4.39}$$

Como estimativas dos parâmetros da mistura, consideraram-se os valores obtidos na iteração 5074 do algoritmo EM, ou seja,

$$\hat{\pi}_m = (0.0926, 0.8487, 0.0587) \;\; \text{e} \;\; \hat{p}_m = (0.1373, 0.2598, 1) \, .$$

De salientar que apenas 20% das frequências esperadas têm valor inferior a 5. O valor observado da estatística de teste é $\chi^2_{P_{obs}} = 7392.62$ e o valor p igual a 2×10^{-16}. Perante estes resultados, pode concluir-se que, a um nível de significância de 1%, rejeita-se a hipótese de uma mistura de três distribuições geométricas se ajustar à distribuição empírica. Contudo, este resultado não é surpreendente dado o elevado número de observações que constitui a sequência de distâncias global. Nestes casos, os testes de ajustamento tendem a ser não conservativos, pelo que se rejeita sempre a hipótese H_0.

Medidas de similaridades

Para avaliar a similaridade entre a distribuição empírica e o modelo teórico poder-se-á utilizar a medida de distância

$$S^1 = 1 - \frac{\sum_{k=1}^{L} |f_0(k) - f(k)|}{\sum_{k=1}^{L} (|f_0(k)| + |f(k)|)} \, , \tag{4.40}$$

onde $f_0(k)$ representa a frequência relativa observada da distância k e $f(k)$ a f.m.p. associada à distribuição teórica [6]. O valor da medida S^1 está compreendido entre 0 e 1. Quanto mais próximo o seu valor estiver de 1, maior será a similaridade entre as duas distribuições.

Além da medida S^1 existem outras medidas, tais como o coeficiente de correlação linear de Pearson e a entropia relativa (ou divergência de Kullback-Liebler). O coeficiente de correlação linear de Pearson produz resultados que são qualitativamente similares aos obtidos pela medida S^1. No entanto, na utilização do coeficiente de correlação linear de Pearson existem diferenças importantes que devem ser consideradas na sua utilização, tais como, por exemplo, o facto de ser sensível a observações atípicas [6]. A entropia relativa ou divergência de Kullback-Liebler é definida [5] por

$$D_{KL}(f_0, f) = \sum_{k=1}^{L} f_0(k) \log\left(\frac{f_0(k)}{f(k)}\right).$$

Tem-se que $D_{KL} \geq 0$ e, em geral, $D_{KL}(f_0, f) \neq D_{KL}(f, f_0)$. Deste modo, D_{KL} não é uma distância, embora muitas vezes seja denominada distância de Kullback-Liebler. No uso desta medida utilizam-se as seguintes convenções:

- $0 \log(0) = 0$
- $f_0(k) \log\left(\frac{f_0(k)}{0}\right) = \infty$, se $f_0(k) > 0$
- $0 \log\left(\frac{0}{0}\right) = 0$

4.4 Resultados experimentais

São apresentados a seguir, para cada uma das espécies em estudo, os resultados das estimativas dos parâmetros da distribuição $Modelo(p)$, das estimativas dos parâmetros da mistura de quatro distribuições geométricas $Mgeom(\Psi^{(EM)})$, do teste de ajustamento do qui-quadrado e dos valores das medidas de similaridade S^1 e Kullback-Liebler entre a distribuição empírica e os modelos teóricos $Modelo(p)$, $Mgeom(\Psi^{(0)})$ e $Mgeom(\Psi^{(EM)})$. Na análise efectuada foram consideradas todas as distâncias.

Na Tabela 4.4 encontram-se os valores obtidos através da fórmula (2.11) para as estimativas do parâmetro $\hat{p}^x$, $x \in \mathcal{A}$, da distribuição da sequência de distâncias entre nucleótidos D^x, para cada uma das espécies em estudo.

Tabela 4.4: Estimativa do parâmetro $\hat{p}^x$, $x \in \mathcal{A}$, da distribuição da sequência de distâncias entre nucleótidos D^x, para cada uma das espécies em estudo.

Esp.	$\hat{p}^{\mathrm{A}}$	$\hat{p}^{\mathrm{C}}$	$\hat{p}^{\mathrm{G}}$	$\hat{p}^{\mathrm{T}}$	Esp.	$\hat{p}^{\mathrm{A}}$	$\hat{p}^{\mathrm{C}}$	$\hat{p}^{\mathrm{G}}$	$\hat{p}^{\mathrm{T}}$
Ap	0.2156	0.2835	0.2796	0.2213	*Cf*	0.2937	0.2064	0.2064	0.2936
Hr	0.1706	0.3286	0.3286	0.1723	*Eq*	0.2924	0.2074	0.2076	0.2926
Mj	0.3442	0.1555	0.1574	0.3429	*Gg*	0.2921	0.2077	0.2078	0.2924
Pf	0.2962	0.2037	0.2040	0.2961	*Am*	0.3366	0.1636	0.1632	0.3367
Tk	0.2410	0.2604	0.2596	0.2390	*Dm*	0.2888	0.2112	0.2111	0.2889
Ba	0.3224	0.1779	0.1759	0.3238	*Mu*	0.2918	0.2080	0.2081	0.2921
Bs	0.2818	0.2181	0.2171	0.2830	*Ce*	0.3226	0.1775	0.1773	0.3226
Ct	0.2942	0.2065	0.2066	0.2927	*Rn*	0.2905	0.2096	0.2096	0.2903
Cb	0.3549	0.1429	0.1395	0.3627	*Xt*	0.2987	0.2013	0.2013	0.2986
Dv	0.1839	0.3162	0.3153	0.1845	*Hs*	0.2952	0.2045	0.2046	0.2957
Ec	0.2462	0.2542	0.2537	0.2459	*Mm*	0.2956	0.2043	0.2044	0.2957
Hi	0.3102	0.1916	0.1899	0.3083	*Pt*	0.2964	0.2034	0.2035	0.2968
Hp	0.3030	0.1961	0.1926	0.3082	*Dr*	0.3171	0.1830	0.1829	0.3169
Mg	0.3457	0.1578	0.1591	0.3374	*Fu*	0.2726	0.2273	0.2273	0.2728
Pa	0.1686	0.3357	0.3299	0.1658	*Oa*	0.2725	0.2276	0.2273	0.2726
Sa	0.3359	0.1630	0.1652	0.3360	*Dd*	0.3881	0.1123	0.1118	0.3877
Sm	0.3146	0.1854	0.1829	0.3171	*Li*	0.2021	0.2983	0.2970	0.2025
St	0.3021	0.1982	0.1967	0.3030	*Pl*	0.4031	0.0969	0.0970	0.4030
At	0.3200	0.1802	0.1801	0.3197	*Tb*	0.2667	0.2322	0.2317	0.2694
Os	0.2823	0.2177	0.2178	0.2822	*Ca*	0.3313	0.1670	0.1681	0.3336
Po	0.3316	0.1687	0.1685	0.3311	*Nc*	0.2507	0.2492	0.2496	0.2506
Vv	0.3275	0.1728	0.1727	0.3271	*Sc*	0.3098	0.1909	0.1906	0.3087
Bt	0.2910	0.2087	0.2088	0.2915	*Sp*	0.3193	0.1804	0.1803	0.3199

Na aplicação do algoritmo EM consideraram-se como estimativas iniciais os valores do vector $\Psi^{(0)}$ da distribuição $Mgeom(\Psi^{(0)})$, em que os pesos da mistura são iguais para todas as espécies, $\pi^{(0)} = (0.25, 0.25, 0.25, 0.25)$, e como componentes da mistura os valores da Tabela 4.4.

Na Tabela 4.5 encontra-se, para cada espécie, o número da iteração (coluna Iter.(k)) em que o critério $\varepsilon = 10^{-5}$ do algoritmo EM foi atingido, bem como as estimativas do vector dos parâmetros da distribuição $Mgeom(\Psi^{(EM)})$ nessa iteração. Dos resultados aí apresentados conclui-se que a modelação, no caso das bactérias *Hr*, *Dv* e *Pa* e do protozoário *Li*, poderá ser feita considerando uma mistura de apenas duas geométricas. No caso das bactérias *Ap*, *Pf*, *Ct*, *Hp* e *St*, do protozoário *Pl* e do fungo *Nc*, a modelação poderá

ser feita considerando uma mistura de apenas três geométricas. Na classe dos animais e das plantas não se identificou nenhum caso em que fosse possível reduzir para menos de quatro o número de componentes da mistura.

Tabela 4.5: Resultados das estimativas do vector dos parâmetros $\Psi^{(EM)} = (\hat{\pi}, \hat{p})$, obtidas pelo algoritmo EM com $\varepsilon = 10^{-5}$, onde $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3, \hat{\pi}_4)$ e $\hat{p} = (\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4)$, para cada uma das espécies em estudo.

Esp.	Iter.(k)	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$
Ap	20592	0.01310	0.24743	0.34700	0.39247	0.11476	0.42621	0.22372	0.22372
Hr	585	0.09074	0.32966	0.32966	0.24993	0.11618	0.28247	0.28247	0.28247
Mj	115071	0.19404	0.01696	0.11194	0.67706	0.67741	0.06623	0.10645	0.28141
Pf	3726	0.14713	0.09101	0.49204	0.26982	0.69915	0.12488	0.24891	0.24891
Tk	30103	0.45483	0.14992	0.38009	0.01515	0.23169	0.54062	0.23169	0.12730
Ba	25088	0.50960	0.40840	0.04032	0.04168	0.34429	0.19501	0.10504	1.00000
Bs	87227	0.43343	0.50867	0.00654	0.05136	0.30411	0.20668	0.10501	1.00000
Ct	19617	0.21604	0.00662	0.45966	0.31768	0.53253	0.10065	0.22030	0.22030
Cb	73282	0.47628	0.09740	0.02748	0.39884	0.24709	0.11165	0.07026	0.49303
Dv	1473	0.09740	0.32292	0.32289	0.25680	0.13387	0.27582	0.27582	0.27582
Ec	32524	0.45912	0.01680	0.45872	0.06537	0.25649	1.00000	0.25649	0.16150
Hi	55778	0.09326	0.21977	0.00038	0.68658	1.00000	0.16700	0.04779	0.26583
Hp	3904	0.29086	0.46642	0.03907	0.20365	0.22301	0.22301	0.10789	0.84099
Mg	1165	0.12597	0.21460	0.33092	0.32851	1.00000	0.13818	0.28410	0.28414
Pa	824	0.24578	0.32063	0.32055	0.11304	0.28516	0.28516	0.28516	0.12706
Sa	3775	0.36821	0.19404	0.39603	0.04171	0.29794	0.13934	0.29779	1.00000
Sm	4700	0.33228	0.41001	0.17768	0.08003	0.27015	0.26937	0.15213	1.00000
St	3997	0.39739	0.45135	0.09261	0.05866	0.25982	0.25982	0.13727	1.00000
At	44601	0.19058	0.17827	0.00394	0.62721	0.53736	0.13845	0.05879	0.27381
Os	6910	0.34099	0.00775	0.31031	0.34094	0.34063	0.06283	0.16554	0.34061
Po	8808	0.06110	0.29176	0.02044	0.62670	1.00000	0.16297	0.06123	0.34534
Vv	16869	0.09156	0.21138	0.01262	0.68444	0.89279	0.14480	0.05486	0.31005
Bt	149024	0.62318	0.01356	0.27732	0.08594	0.29908	0.08831	0.16916	0.69643
Cf	29256	0.20210	0.00105	0.11575	0.68110	0.56839	0.03230	0.11871	0.25830
Eq	66883	0.62547	0.00077	0.19516	0.17860	0.27054	0.05218	0.14342	0.57118
Gg	38733	0.63042	0.00005	0.06374	0.30578	0.22857	0.01338	0.11842	0.43705
Am	23968	0.57132	0.08542	0.00424	0.33902	0.24903	0.09074	0.03826	0.51868
Dm	52733	0.74100	0.20943	0.00412	0.04545	0.29497	0.15149	0.06868	1.00000
Mu	21013	0.57483	0.00042	0.02545	0.39930	0.21128	0.01607	0.08445	0.41949
Ce	18634	0.05104	0.74506	0.08992	0.11398	1.00000	0.24815	0.11149	0.81433
Rn	32105	0.41398	0.02219	0.00039	0.56344	0.41420	0.08173	0.01968	0.20799
Xt	111125	0.14980	0.16840	0.00015	0.68164	0.62797	0.14810	0.03933	0.26013
Hs	16947	0.70534	0.00019	0.12346	0.17101	0.25989	0.01991	0.12454	0.59992
Mm	22959	0.71364	0.00014	0.10167	0.18456	0.25202	0.01810	0.11987	0.59075
Pt	21466	0.70285	0.00019	0.12529	0.17167	0.25994	0.02209	0.12512	0.60030

continua na página seguinte

Tabela 4.5 – continuação da página anterior

Esp.	Iter.(k)	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$
Dr	19073	0.46100	0.03924	0.00066	0.49910	0.40556	0.09193	0.01548	0.20849
Fu	56355	0.45332	0.00921	0.00014	0.53733	0.19056	0.07419	0.01556	0.36107
Oa	30612	0.75982	0.14642	0.00039	0.09337	0.27580	0.13437	0.03410	0.64871
Dd	1387	0.12160	0.15919	0.02416	0.69505	1.00000	0.12324	0.03299	0.37481
Li	487	0.11393	0.32848	0.32818	0.22941	0.11228	0.29681	0.29681	0.29681
Pl	599	0.39015	0.03813	0.18158	0.39014	0.45253	0.04271	0.13130	0.45253
Tb	7934	0.23714	0.44061	0.01422	0.30802	0.36779	0.18972	0.07282	0.36780
Ca	7866	0.68267	0.01508	0.25251	0.04974	0.33214	0.06781	0.15098	1.00000
Nc	7385	0.42038	0.00343	0.15670	0.41950	0.29588	0.05153	0.14312	0.29588
Sc	11051	0.03937	0.42100	0.00188	0.53776	1.00000	0.18370	0.04647	0.33022
Sp	34415	0.70287	0.22921	0.00353	0.06439	0.29140	0.15562	0.06961	1.00000

Aplicou-se também o algoritmo EM apenas às cem primeiras distâncias e, à semelhança de resultados obtidos por [2] com outras metodologias, também aqui os resultados são muito próximos daqueles que se obtiveram considerando todas as distâncias no algoritmo EM.

À semelhança do que aconteceu para a espécie *St*, também para as restantes espécies a aplicação do teste de ajustamento do qui-quadrado levou à rejeição da hipótese de uma mistura de quatro distribuições geométricas se ajustar à distribuição empírica. Este resultado já era esperado, dado o elevado número de observações que constituem o conjunto das distâncias entre nucleótidos por genoma sequenciado.

Na Tabela 4.6 encontram-se os valores da medida de similaridade S^1, definida em (4.40), calculados entre a distribuição empírica e cada um dos modelos teóricos *Modelo(p)*, $Mgeom(\Psi^{(0)})$ e $Mgeom(\Psi^{(EM)})$. Das três distribuições, aquela que melhor se ajusta à distribuição empírica, para cada uma das espécies, foi a que resultou da aplicação do algoritmo EM, isto é, a distribuição $Mgeom(\Psi^{(EM)})$. A qualidade do ajustamento é melhor nas espécies para as quais a modelação corresponde a uma mistura de quatro geométricas. Os piores resultados do ajustamento verificam-se nas espécies para as quais a modelação da mistura corresponde a uma mistura de duas geométricas.
Os resultados referentes à medida de Kullback-Liebler apresentam-se em anexo na Tabela A.4, mas não fornecem informação relevante para além daquela já obtida através da distância S^1.

Tabela 4.6: Resultados da aplicação da medida de similaridade S^1 entre a distribuição empírica e cada uma das distribuições teóricas: *Modelo*(p), mistura de quatro distribuições geométricas com os parâmetros iniciais do algoritmo EM, $Mgeom(\Psi^{(0)})$, e mistura de quatro distribuições geométricas com parâmetro obtidos pelo algoritmo EM, $Mgeom(\Psi^{(EM)})$, para cada uma das espécies em estudo.

Esp.	Modelo (p)	$Mgeom(\Psi^{(0)})$	$Mgeom(\Psi^{(EM)})$	Esp.	Modelo (p)	$Mgeom(\Psi^{(0)})$	$Mgeom(\Psi^{(EM)})$
Ap	0.9583	0.9544	**0.9642**	*Cf*	0.9360	0.9276	**0.9908**
Hr	0.9112	0.9062	**0.9144**	*Eq*	0.9436	0.9361	**0.9946**
Mj	0.9405	0.9015	**0.9828**	*Gg*	0.9629	0.9548	**0.9925**
Pf	0.9453	0.9418	**0.9858**	*Am*	0.9395	0.8976	**0.9928**
Tk	0.9508	0.9506	**0.9669**	*Dm*	0.9521	0.9482	**0.9943**
Ba	0.9581	0.9408	**0.9794**	*Mu*	0.9451	0.9373	**0.9860**
Bs	0.9535	0.9515	**0.9804**	*Ce*	0.9242	0.9097	**0.9912**
Ct	0.9616	0.9572	**0.9847**	*Rn*	0.9444	0.9374	**0.9852**
Cb	0.9591	0.8904	**0.9791**	*Xt*	0.9564	0.9465	**0.9955**
Dv	0.9528	0.9489	**0.9550**	*Hs*	0.9444	0.9350	**0.9924**
Ec	0.9641	0.9640	**0.9735**	*Mm*	0.9446	0.9348	**0.9899**
Hi	0.9402	0.9352	**0.9821**	*Pt*	0.9449	0.9350	**0.9924**
Hp	0.9050	0.9018	**0.9853**	*Dr*	0.9574	0.9332	**0.9872**
Mg	0.9197	0.8925	**0.9641**	*Fu*	0.9521	0.9495	**0.9835**
Pa	0.9266	0.9334	**0.9309**	*Oa*	0.9474	0.9453	**0.9894**
Sa	0.9589	0.9335	**0.9717**	*Dd*	0.9071	0.8183	**0.9645**
Sm	0.9417	0.9309	**0.9709**	*Li*	0.9129	0.9089	**0.9245**
St	0.9550	0.9486	**0.9778**	*Pl*	0.9369	0.7924	**0.9479**
At	0.9642	0.9388	**0.9923**	*Tb*	0.9428	0.9413	**0.9797**
Os	0.9499	0.9444	**0.9878**	*Ca*	0.9444	0.9163	**0.9734**
Po	0.9438	0.9199	**0.9946**	*Nc*	0.9543	0.9543	**0.9771**
Vv	0.9422	0.9217	**0.9950**	*Sc*	0.9615	0.9500	**0.9840**
Bt	0.9530	0.9458	**0.9956**	*Sp*	0.9575	0.9391	**0.9844**

A titulo de exemplo, apresentam-se na Figura 4.8 as representações gráficas das distribuições empíricas e das distribuições *Modelo(p)* (linha azul) e $Mgeom(\Psi^{(EM)})$ (linha verde) referentes às espécies *Mj*, *Pf*, *Hp* e *Dv*. Para a espécie *Dv* a diferença entre os dois modelos teóricos é mínima. Em anexo, na Figura A.7, apresentam-se as representações das distribuições para a classe das plantas.

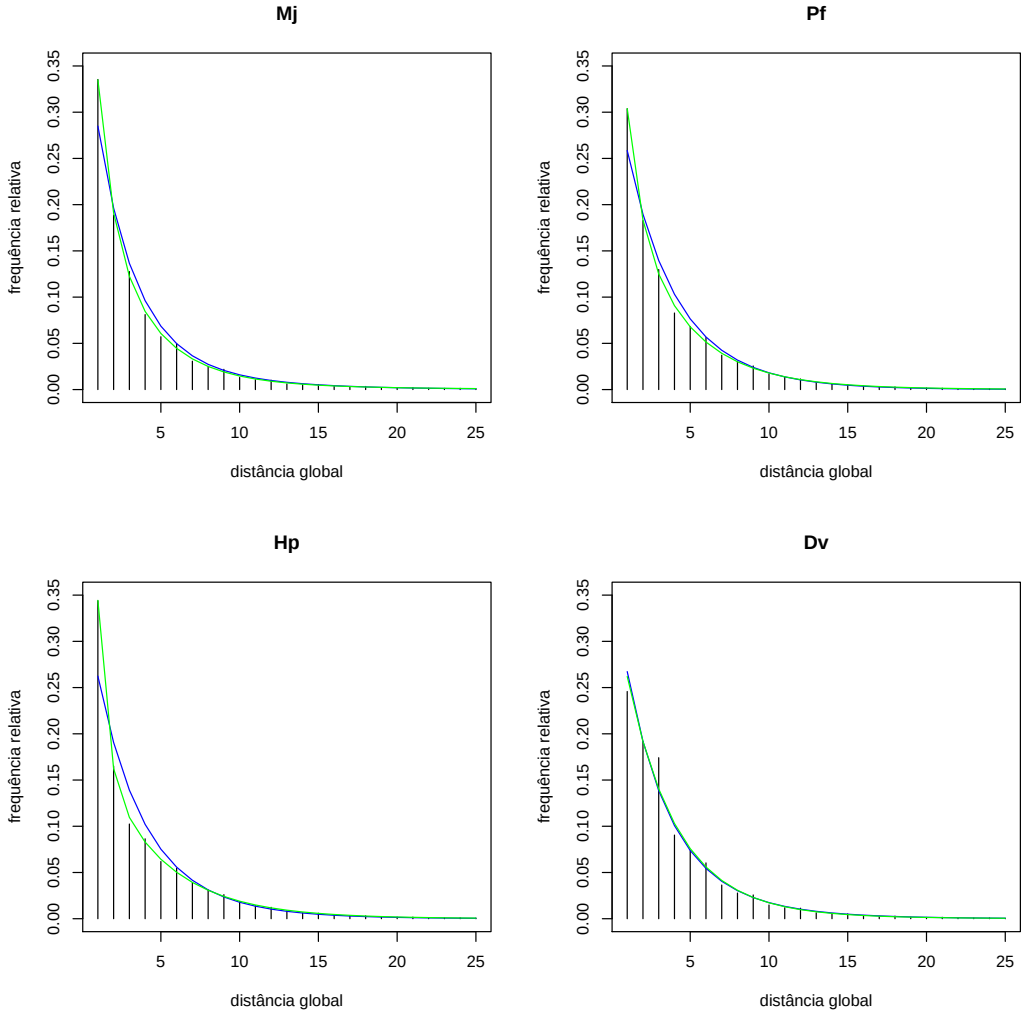


Figura 4.8: Distribuição empírica (gráfico de barras) *vs* Distribuições teóricas *Modelo(p)* (linha azul) e $Mgeom(\Psi^{(EM)})$ (linha verde), das espécies *Mj*, *Pf*, *Hp* e *Dv*.

Capítulo 5

Conclusões e trabalho futuro

Nesta dissertação foram aplicadas algumas ferramentas de estatística multivariada aos genomas completos de 46 espécies de organismos, com o objectivo de explorar e identificar a existência de possíveis regras na estrutura primária desses genomas. Em particular, investigaram-se as propriedades das distribuições empíricas das distâncias entre nucleótidos resultantes do mapeamento das sequências de ADN proposto por [2]. Este mapeamento é designado por distância global entre nucleótidos iguais.

As metodologias estatísticas multivariadas utilizadas, nomeadamente a classificação hierárquica, a classificação não-hierárquica e a análise de componentes principais, foram aqui aplicadas com o intuito de investigar características discriminativas (ou não) da classe a que cada espécie pertence.

Da classificação hierárquica concluiu-se que o mapeamento da distância global entre nucleótidos iguais capturou as características essenciais do ADN das espécies analisadas, no sentido em que permitiu a construção de dendrogramas interpretáveis como árvores filogenéticas, por estarem de acordo com as similaridades esperadas entre as espécies. Assim, e à semelhança dos resultados descritos em [2] para 28 espécies, também os resultados obtidos na análise efectuada nesta dissertação para 46 espécies, que incluem as 28 espécies tratadas em [2], permitiram inferir que a distribuição das primeiras distâncias representa uma possível assinatura genética capaz de permitir a diferenciação entre espécies.

A classificação não-hierárquica e a análise de componentes principais identificaram dois grupos principais de organismos que, de acordo com as espécies que os constituem, correspondem à tradicional divisão entre organismos eucariotas e procariotas.

Confrontou-se a distribuição empírica com o modelo geométrico esperado caso o sequenciamento das quatro letras que constituem o alfabeto genómico do ADN obedecesse à lei de independência estocástica, do que resultou a hipótese de existir um modelo probabilístico teórico mais bem adaptado à distribuição empírica, eventualmente baseado em misturas de distribuições geométricas. Essa hipótese foi investigada, tendo-se concluído, com base em medidas de similaridade, que o modelo de mistura de quatro distribuições geométricas, com os parâmetros estimados a partir do algoritmo EM, foi o que melhor se ajustou à distribuição empírica da maioria das espécies, incluindo todos os animais e plantas. Relativamente às restantes espécies, verificou-se que no caso de algumas bactérias, protozoários e um fungo, a modelação pode ser feita com misturas de duas ou três distribuições geométricas. A qualidade do ajustamento entre os modelos teóricos e a distribuição empírica foi avaliada também com o auxílio do teste de ajustamento do qui-quadrado. Porém, dado o elevado número de observações que constituem o conjunto das distâncias entre nucleótidos por genoma sequenciado, o teste do qui-quadrado conduziu-nos à rejeição da hipótese nula, tal como aconteceria com qualquer outro teste estatístico (tradicional) de ajustamento.

O comportamento não conservativo dos testes de ajustamento face a um conjunto com um número elevado de observações mostra a necessidade de uma investigação conjunta envolvendo a Estatística e Técnicas de Prospecção de dados (*data mining*) com vista ao desenvolvimento de métodos que avaliem a qualidade de ajustamento a modelos teóricos nessas condições. Da breve pesquisa realizada constatou-se a existência de uma lacuna na investigação de métodos adequados. Trata-se, por conseguinte, de uma temática de grande interesse para investigação futura.

Referências bibliográficas

[1] Vera Afreixo. *Sinais Simbólicos e Aplicações em Genómica*. PhD thesis, Universidade de Aveiro, 2008.

[2] Vera Afreixo, Carlos A.C. Bastos, Armando J. Pinho, Sara P. Garcia, and Paulo J.S.G. Ferreira. Genome Analysis with Inter-Nucleotide Distances. *Bioinformatics*, 25(23):3064–3070, 2009.

[3] B. Alberts and A. Johnson *et al. Molecular Biology of The Cell.* Garland Science, 2002, Fourth edition.

[4] O.T. Avery, C.M. MacLeod, and M. McCarty. *Studies of the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types. Induction of Transformation by a Deoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III.* J. Exp. Med., 1944.

[5] Ricardo Baeza-Yates, Joseph Glaz, Henryk Gzyl, Jurgen Husler, and José Luis Palacios. *Recent Advances in Applied Probability.* Springer, 2005.

[6] Pierre-François Baisnée, Steve Hampson, and Pierre Baldi. Why Are Complementary DNA Strands Symmetric? *Bioinformatics*, 18(8):1021–1033, 2002.

[7] A. Blejeck. http://ablejec.nib.si/R/ECPR/I2R.pdf, (consultado em Outubro de 2009).

[8] P.D. Cristea. Conversion of Nucleotides Sequences into Genomic Signals. *J. Cell. Mol. Moed*, 6(2):279–303, 2002.

[9] Ralf Dahm. Friedrich Miescher and the Discovery of DNA. *Developmental Biology*, 278(2):274–288, 2005.

[10] Peter Dalgaard. *Introductory Statistical with R.* Springer, 2008.

[11] Grupo de Ciências Biológicas do Instituto Superior Técnico. http://www.e-escola.pt/topico.asp?id=224&ordem=2, (consultado em Março de 2010).

[12] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[13] E.R. Dougherty, I. Shmulevich, J. Chen, and Z.J. Wang. *Genomic Signal Processing and Statistics*. Hindawi Publ. Corp, 2005.

[14] National Center for Biotchnology Information. http://www.ncbi.nlm.nih.gov/Genomes/, (consultado em Novembro de 2009).

[15] E.W. Forgy. Cluster Analysis of Multivariate Data: Efficiency Versus Interpretability of Classifications. *Biometric Society Meeting, Riverside, California*, 21:768–769, 1965.

[16] Lei Gao, Ji Qi, and Bailin Hao. Simple Markov Subtraction Essentially Improves Prokaryote Phylogeny. *AAPPS Bulletin*, 16(3):3–7, 2006.

[17] Joseph F. Hair, Ronald L. Tatham, Rolph E. Anderson, and william Black. *Multivariate Data Analysis*. Prentice-Hall, Inc, 1998, Fifth edition.

[18] Jiawei Han and Micheline Kamber. *Data Mining - Concepts and Techniques*. Elsevier, 2006.

[19] J.A. Hartigan and M.A. Wong. A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society*, 28(1):100–108, 1979.

[20] A.D. Hershey and M. Chase. Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage. *Journal of General Physiology*, 36:39–56, 1952.

[21] Joint Genome Institute. http://genome.jgi-psf.org/, (consultado em Novembro de 2009).

[22] SAS Institute Inc. *SAS/STAT User's Guide*. Cary,Nc: SAS Institute Inc, 2004.

[23] Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice-Hall, 1998.

[24] T. Kanungo, D.M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A.Y. Wu. Singular Value Decomposition and Principal Component Analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24:881–892, 2002.

[25] Paul H. Kvam and Brani Vidakovic. *Nonparametric Statistics with Applications to Science and Engineering.* Wiley, 2007.

[26] E.L. Lehmann and George Casella. *Theory of Point Estimation.* Springer, 1998, Second Edition.

[27] Sébastien Lê *et al.* FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*, 25(1):1–18, 2008.

[28] J. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. In L. M. Le Cam J. Neyman, editor, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. University of California Press, 1967.

[29] K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis.* Academic Press, 1994.

[30] Geoffrey J. McLachlan and David Peel. *Finite Mixture Models.* Wiley, 2000.

[31] Geoffrey J. McLachlan and Krishnan Thriyambakam. *The EM Algorithm and Extensions.* Wiley, 2008, Second Edition.

[32] F. Miescher. *Letter I; to Wilhelm His; Tübingen, February 26th, 1869. In: W. His et al., Editors, Die Histochemischen Und Physiologischen Arbeiten Von Friedrich Miescher - Aus Dem Wissenschaftlichen Briefwechsel Von F. Miescher.* Leipzig F. C. W. Vogel 1897, 1869.

[33] G.W. Milligan, P. Arabie, L.J. Hubert, and G De Soete. *Clustering And Classification.* World Scientific, 1996.

[34] N. Monteiro, J. Gomes, and J. Xavier. Detection of Statistical Periodicities in DNA by Conflict and Entropy Minimization Methods. *16th European Signal Processing Conference*, pages 25–29, 2008.

[35] A.S.S. Nair and T. Mahalakshmi. Visualization of Genomic Data Using Inter-Nucleotide Distance Signals. *In proceedings of IEEE Genomic Signal Processing*, 2005.

[36] A.S.S. Nair and T. Mahalakshmi. Are Categorical Periodograms and Indicator Sequences of Genomes Spectrally Equivalent? *In Silico Biology*, pages 215–222, 2006.

[37] Ricardo A. Olea. *Geostatistics For Engineers and Earth Scientists*. Kluwer Academic Publishers, 1999.

[38] Genome Project. http://www.fugu-sg.org/, (consultado em Novembro de 2009).

[39] Human Genome Project. http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml, (consultado em Abril de 2010).

[40] Matthias Scholz. *Approaches to Analyse and Interpret Biological Profile Data*. PhD thesis, Potsdam University, 2006.

[41] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Addison-Wesley, 2005.

[42] Michael E. Wall, Andreas Rechtsteiner, and Luís M. Rocha. *A Practical Approach to Microarray Data Analysis*. Kluwer Academic Publishers, 2003.

[43] J.D. Watson and F.H.C. Crick. A Structure for Deoxyribose Nucleic Acid. *Nature*, 171:737–738, 1953.

[44] Xenbase. http://www.xenbase.org/, (consultado em Novembro de 2009).

[45] Derek Young *et al.* Package Mixtools - Tools for Analyzing Finite Mixture Models.

Apêndice A

Resultados complementares

ACP - Variáveis padronizadas

Os valores dos coeficientes de correlação entre as variáveis originais padronizadas e as três primeiras componentes principais encontram-se na Tabela A.1.

Tabela A.1: Valores dos coeficientes de correlação entre as variáveis padronizadas e as três componentes CP1, CP2 e CP3.

δ	CP1	CP2	CP3	δ	CP1	CP2	CP3
δ_1	0.3088	−0.1863	0.7465	δ_{51}	0.8917	−0.1588	−0.2010
δ_2	0.4737	0.1040	−0.5257	δ_{52}	0.9584	−0.1062	−0.0181
δ_3	−0.4686	0.0977	−0.6556	δ_{53}	0.8759	−0.3839	−0.1176
δ_4	0.0663	0.1616	−0.1632	δ_{54}	0.8692	−0.2312	−0.2509
δ_5	0.0136	−0.1153	−0.0058	δ_{55}	0.7386	−0.0663	0.0576
δ_6	−0.8270	−0.0182	−0.1532	δ_{56}	0.7338	−0.1568	−0.2818
δ_7	−0.1318	−0.3691	0.6257	δ_{57}	0.6006	0.1124	0.0996
δ_8	−0.2875	−0.5144	0.6388	δ_{58}	0.6199	0.2766	0.1760
δ_9	−0.8810	−0.2050	0.0337	δ_{59}	0.6591	0.1452	0.0064
δ_{10}	0.0593	−0.5582	0.7388	δ_{60}	0.7840	−0.2635	−0.2878
δ_{11}	−0.0343	−0.6627	0.6778	δ_{61}	0.5905	0.5555	0.2091
δ_{12}	−0.8044	−0.3170	−0.0061	δ_{62}	0.5468	0.5650	0.2640
δ_{13}	0.4777	−0.5417	0.5918	δ_{63}	0.5295	0.5955	0.2375
δ_{14}	0.5110	−0.5738	0.5190	δ_{64}	0.5067	0.2251	−0.0163
δ_{15}	−0.3163	−0.6115	−0.0349	δ_{65}	0.4529	0.5465	0.2607
δ_{16}	0.6874	−0.5131	0.4237	δ_{66}	0.5034	0.6127	0.2501
δ_{17}	0.7181	−0.5755	0.2821	δ_{67}	0.5356	0.5179	0.0032
δ_{18}	0.3607	−0.6244	−0.2539	δ_{68}	0.5176	0.5100	0.0790
δ_{19}	0.7941	−0.5214	0.2131	δ_{69}	0.3932	0.4699	0.0498

continua na página seguinte

Tabela A.1 – continuação da página anterior

δ	CP1	CP2	CP3	δ	CP1	CP2	CP3
δ_{20}	0.7971	−0.5199	0.1249	δ_{70}	0.4486	0.5104	0.2474
δ_{21}	0.5643	−0.5132	−0.3393	δ_{71}	0.4959	0.1918	−0.1901
δ_{22}	0.8329	−0.4883	0.1072	δ_{72}	0.4248	0.5427	−0.0244
δ_{23}	0.8692	−0.4482	0.0153	δ_{73}	0.3610	0.5392	0.0795
δ_{24}	0.7081	−0.4473	−0.3325	δ_{74}	0.3911	0.6836	0.2819
δ_{25}	0.8762	−0.4375	0.0624	δ_{75}	0.4525	0.6993	0.0110
δ_{26}	0.9108	−0.3815	−0.0030	δ_{76}	0.5315	0.6217	0.0264
δ_{27}	0.8094	−0.3990	−0.2835	δ_{77}	0.4882	0.7185	−0.0065
δ_{28}	0.8978	−0.4175	0.0393	δ_{78}	0.3549	0.4574	−0.2073
δ_{29}	0.8969	−0.4116	−0.0097	δ_{79}	0.5347	0.7008	0.0267
δ_{30}	0.8419	−0.3828	−0.2788	δ_{80}	0.6428	0.4617	0.1131
δ_{31}	0.9045	−0.4056	0.0087	δ_{81}	0.5033	0.5111	0.0515
δ_{32}	0.9127	−0.3793	−0.0233	δ_{82}	0.5491	0.6494	0.0984
δ_{33}	0.8689	−0.3555	−0.2283	δ_{83}	0.5071	0.5257	−0.2228
δ_{34}	0.9126	−0.3746	0.0129	δ_{84}	0.5939	0.5375	0.1210
δ_{35}	0.9202	−0.3441	−0.0544	δ_{85}	0.6558	0.5590	0.1563
δ_{36}	0.9169	−0.2936	−0.2012	δ_{86}	0.5571	0.5962	−0.0903
δ_{37}	0.9188	−0.3670	0.0200	δ_{87}	0.6830	0.4675	0.3073
δ_{38}	0.9059	−0.3853	−0.0203	δ_{88}	0.5604	0.5047	−0.2419
δ_{39}	0.9052	−0.3524	−0.2024	δ_{89}	0.7095	0.4962	0.1425
δ_{40}	0.9162	−0.3523	0.0164	δ_{90}	0.5752	0.6376	−0.1016
δ_{41}	0.9184	−0.3126	0.0117	δ_{91}	0.6559	0.5536	0.0534
δ_{42}	0.9091	−0.3024	−0.1598	δ_{92}	0.6560	0.5682	−0.0325
δ_{43}	0.9139	−0.3514	0.0104	δ_{93}	0.7228	0.4323	0.0288
δ_{44}	0.9016	−0.3526	−0.0109	δ_{94}	0.7212	0.4641	0.0861
δ_{45}	0.9092	−0.2743	−0.1714	δ_{95}	0.6658	0.4432	0.2642
δ_{46}	0.9168	−0.2123	−0.0898	δ_{96}	0.7025	0.3584	0.1338
δ_{47}	0.8916	−0.3115	−0.0863	δ_{97}	0.6272	0.5708	−0.1147
δ_{48}	0.8883	−0.3478	−0.1610	δ_{98}	0.7454	0.5058	0.0753
δ_{49}	0.8854	−0.2944	−0.0782	δ_{99}	0.7171	0.4172	0.0381
δ_{50}	0.7133	−0.0244	−0.2962	δ_{100}	0.5759	0.5787	0.1808

As representações do círculo de correlações e da distribuição das espécies em função das componentes CP1 e CP3, encontram-se na Figura A.1 e na Figura A.2, respectivamente.

As representações do círculo de correlações e da distribuição das espécies em função das componentes CP2 e CP3, encontram-se na Figura A.3 e a Figura A.4, respectivamente.

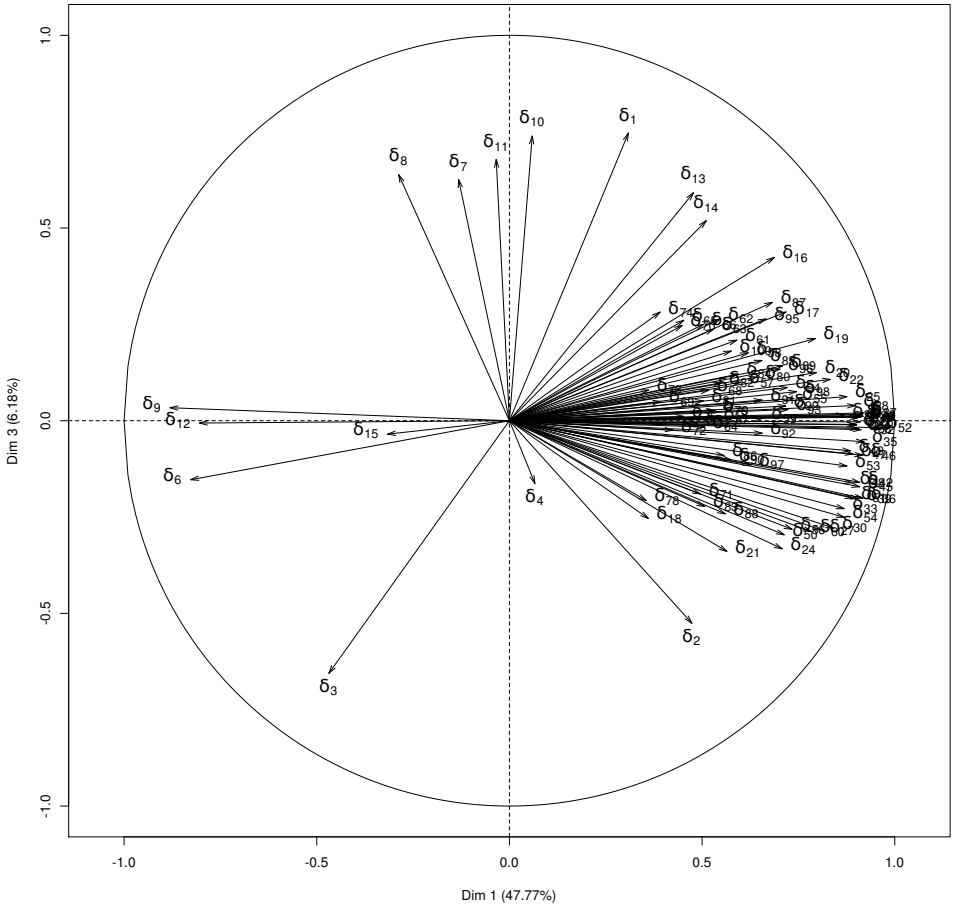


Figura A.1: Círculo das correlações em função das componentes CP1 e CP3.

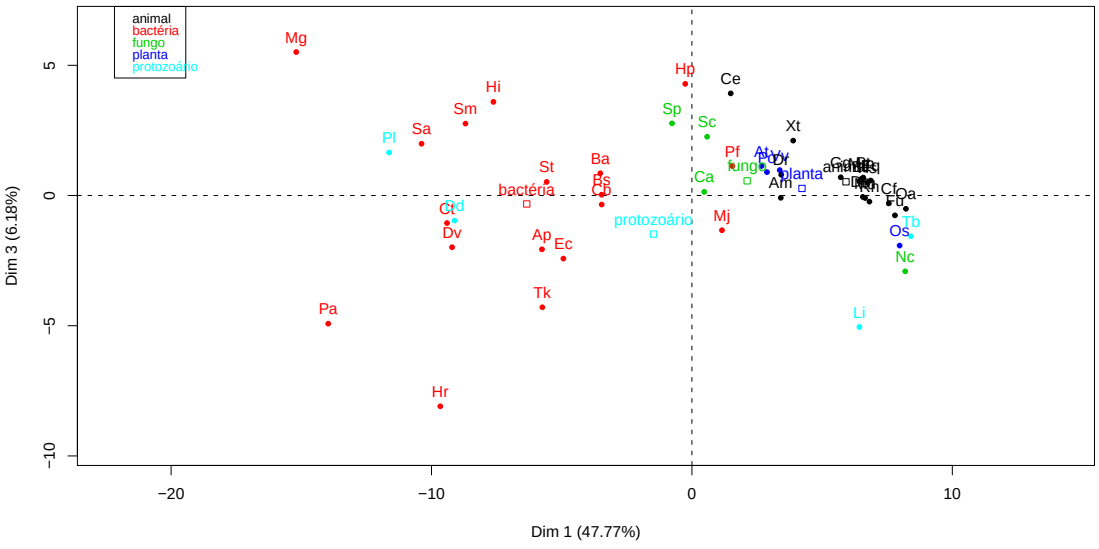


Figura A.2: Representação das espécies entre CP1 e CP3 (variáveis originais padronizadas).

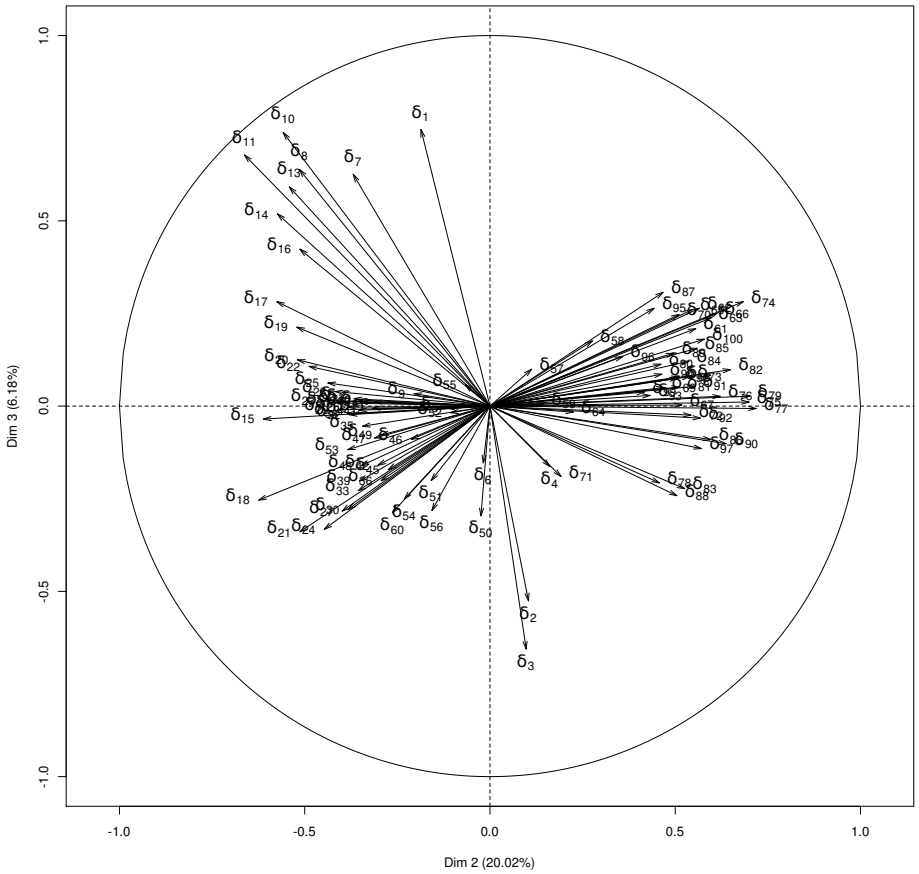


Figura A.3: Círculo das correlações em função das componentes CP2 e CP3.

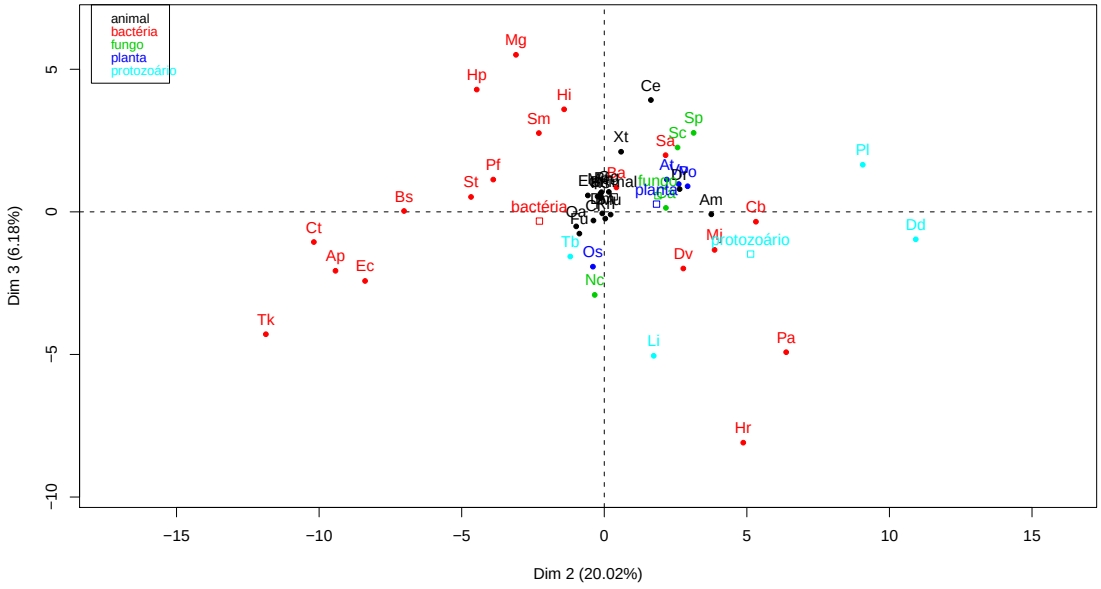


Figura A.4: Representação das espécies entre CP2 e CP3 (variáveis originais padronizadas).

ACP - Variáveis apenas centradas

Os valores dos coeficientes de correlação entre as variáveis originais centradas e as três primeiras componentes principais encontram-se na Tabela A.2.

Tabela A.2: Valores dos coeficientes de correlação entre as variáveis centradas e as três componentes CP1, CP2 e CP3.

δ	CP1	CP2	CP3	δ	CP1	CP2	CP3
δ_1	0.2235	0.2773	0.0134	δ_{51}	0.6616	0.5982	0.1749
δ_2	0.4541	0.1737	−0.2120	δ_{52}	0.7635	0.5733	0.1350
δ_3	−0.4060	−0.2747	0.0165	δ_{53}	0.5677	0.7816	−0.0401
δ_4	0.0755	−0.1158	0.3456	δ_{54}	0.6124	0.6454	0.0897
δ_5	−0.0919	0.0797	0.4689	δ_{55}	0.5840	0.4392	0.2215
δ_6	−0.7487	−0.3877	0.1529	δ_{56}	0.5325	0.5243	0.0701
δ_7	−0.2631	0.1940	0.1029	δ_{57}	0.5348	0.2040	0.2134
δ_8	−0.4637	0.2473	0.1179	δ_{58}	0.6296	0.0750	0.3359
δ_9	−0.8565	−0.2717	0.0274	δ_{59}	0.6002	0.2151	0.1796
δ_{10}	−0.1693	0.4456	0.0171	δ_{60}	0.5305	0.6308	0.0321
δ_{11}	−0.3057	0.4868	0.0911	δ_{61}	0.7401	−0.1954	0.5235
δ_{12}	−0.8456	−0.1523	0.0893	δ_{62}	0.7080	−0.2324	0.5444
δ_{13}	0.1900	0.6452	0.0628	δ_{63}	0.7094	−0.2653	0.5201
δ_{14}	0.1996	0.6899	0.0908	δ_{64}	0.5360	0.0509	0.2293
δ_{15}	−0.5568	0.3244	0.1291	δ_{65}	0.6234	−0.2566	0.4504
δ_{16}	0.3775	0.7384	0.0740	δ_{66}	0.7190	−0.3121	0.4154
δ_{17}	0.3655	0.8139	0.0926	δ_{67}	0.6974	−0.1960	0.3804
δ_{18}	−0.0004	0.6869	0.2173	δ_{68}	0.6292	−0.1775	0.5611
δ_{19}	0.4480	0.8098	0.1288	δ_{69}	0.5698	−0.2350	0.2531
δ_{20}	0.4451	0.8169	0.1381	δ_{70}	0.6144	−0.2350	0.3481
δ_{21}	0.2237	0.7085	0.1788	δ_{71}	0.5002	0.0837	0.1475
δ_{22}	0.4879	0.8126	0.1271	δ_{72}	0.6587	−0.2705	−0.1611
δ_{23}	0.5358	0.8015	0.1220	δ_{73}	0.5238	−0.2834	0.3855
δ_{24}	0.3747	0.7321	0.1794	δ_{74}	0.6488	−0.4224	0.4699
δ_{25}	0.5521	0.7969	0.0838	δ_{75}	0.6968	−0.3905	0.3877
δ_{26}	0.6059	0.7714	0.0858	δ_{76}	0.7395	−0.2810	0.3504
δ_{27}	0.4933	0.7452	0.1047	δ_{77}	0.7549	−0.4030	0.3026
δ_{28}	0.5788	0.7992	0.0680	δ_{78}	0.5023	−0.1987	−0.0076
δ_{29}	0.5772	0.7958	0.0789	δ_{79}	0.7989	−0.3696	0.2583
δ_{30}	0.5251	0.7578	0.1018	δ_{80}	0.8165	−0.0966	−0.0719
δ_{31}	0.5914	0.7970	0.0396	δ_{81}	0.7063	−0.2206	−0.0044
δ_{32}	0.6053	0.7799	0.0705	δ_{82}	0.7998	−0.3200	0.1666
δ_{33}	0.5634	0.7463	0.0900	δ_{83}	0.7182	−0.1981	−0.3085
δ_{34}	0.6102	0.7742	0.0589	δ_{84}	0.7913	−0.1779	0.0575

continua na página seguinte

Tabela A.2 – continuação da página anterior

δ	CP1	CP2	CP3	δ	CP1	CP2	CP3
δ_{35}	0.6276	0.7571	0.0583	δ_{85}	0.8631	−0.1771	0.0753
δ_{36}	0.6394	0.7145	0.0887	δ_{86}	0.7849	−0.2606	−0.1008
δ_{37}	0.6217	0.7746	0.0262	δ_{87}	0.8357	−0.0783	0.1582
δ_{38}	0.5977	0.7880	0.0219	δ_{88}	0.7653	−0.1673	−0.2718
δ_{39}	0.6021	0.7667	0.0357	δ_{89}	0.8981	−0.0989	−0.2110
δ_{40}	0.6189	0.7645	0.0582	δ_{90}	0.8359	−0.2874	−0.1926
δ_{41}	0.6459	0.7353	0.0084	δ_{91}	0.8602	−0.1657	0.0083
δ_{42}	0.6328	0.7194	0.0515	δ_{92}	0.8678	−0.1672	−0.1377
δ_{43}	0.6127	0.7636	0.0883	δ_{93}	0.8504	−0.0110	−0.0426
δ_{44}	0.6139	0.7636	−0.0343	δ_{94}	0.8816	−0.0388	−0.2205
δ_{45}	0.6351	0.7009	0.0834	δ_{95}	0.8502	−0.0690	−0.2615
δ_{46}	0.6693	0.6505	0.1473	δ_{96}	0.8342	0.0267	−0.3047
δ_{47}	0.6198	0.7312	−0.0448	δ_{97}	0.8433	−0.1818	−0.2004
δ_{48}	0.5911	0.7506	0.0127	δ_{98}	0.9354	−0.0746	−0.2648
δ_{49}	0.6069	0.7059	0.1024	δ_{99}	0.8784	−0.0088	−0.3988
δ_{50}	0.5778	0.4069	0.0783	δ_{100}	0.8217	−0.2489	−0.1006

A representação da distribuição das espécies em função das componentes CP1 e CP3 encontra-se na Figura A.5.

A representação da distribuição das espécies em função das componentes CP2 e CP3 encontra-se na Figura A.6.

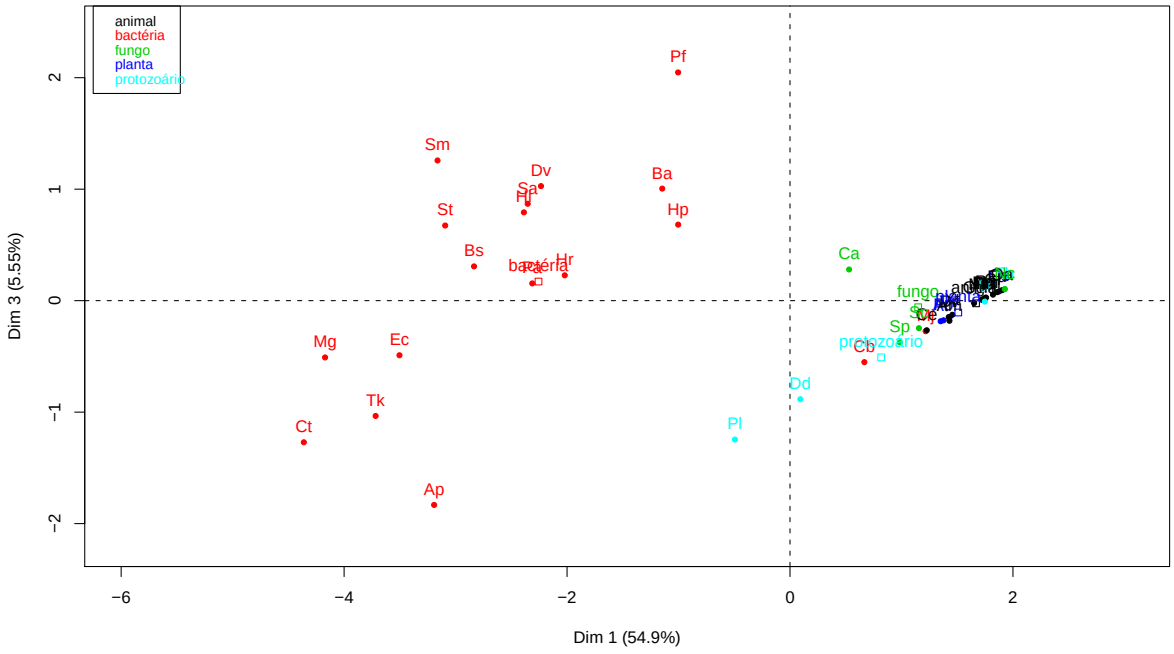


Figura A.5: Representação das espécies entre CP1 e CP3 (variáveis originais apenas centradas).

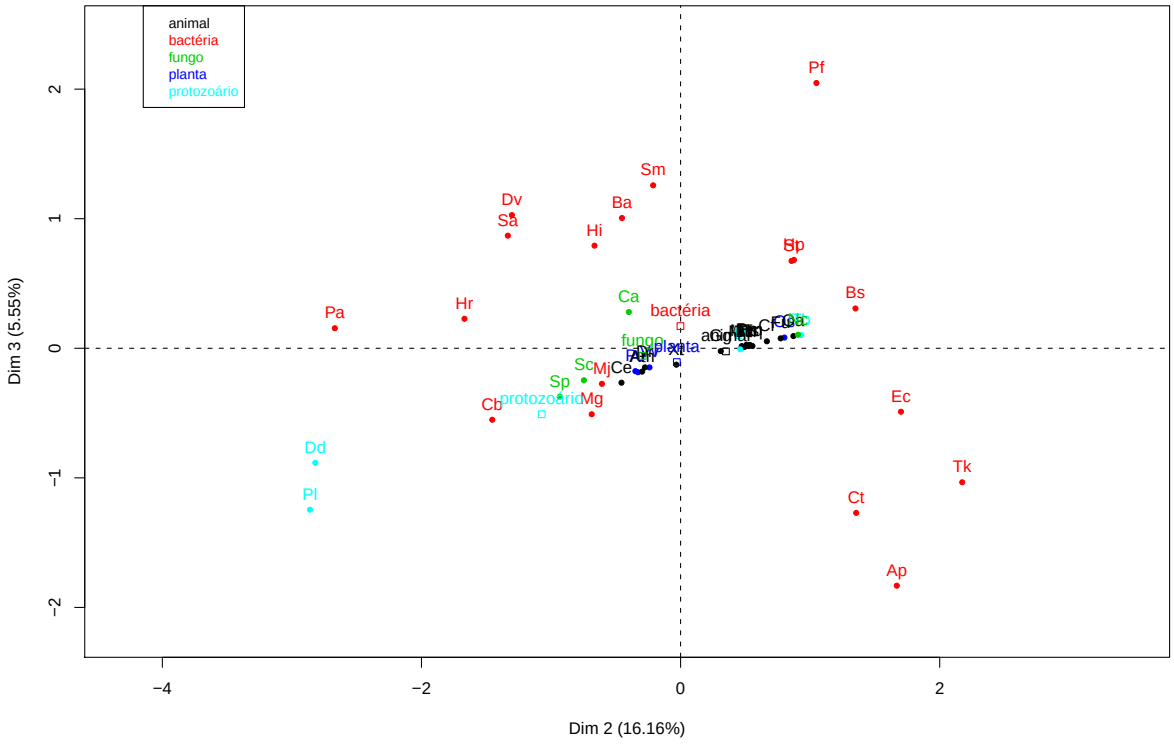


Figura A.6: Representação das espécies entre CP2 e CP3 (variáveis originais apenas centradas).

ACP - Variáveis não padronizadas

Os valores dos coeficientes de correlação entre as variáveis originais não padronizadas e as duas primeiras componentes principais encontram-se na Tabela A.3.

Tabela A.3: Valores dos coeficientes de correlação entre as variáveis não padronizadas e as duas componentes CP1 e CP2.

δ	CP1	CP2	δ	CP1	CP2	δ	CP1	CP2
δ_1	−0.2246	0.0378	δ_{35}	−0.6943	0.1950	δ_{69}	−0.5609	−0.5000
δ_2	−0.4559	−0.1687	δ_{36}	−0.7099	0.1681	δ_{70}	−0.6178	−0.5057
δ_3	0.4024	0.0820	δ_{37}	−0.6843	0.2024	δ_{71}	−0.5308	−0.1998
δ_4	−0.1152	−0.0472	δ_{38}	−0.6637	0.2306	δ_{72}	−0.6164	−0.6392
δ_5	0.0279	0.2169	δ_{39}	−0.6743	0.2243	δ_{73}	−0.5395	−0.4593
δ_6	0.7508	0.2130	δ_{40}	−0.6872	0.2075	δ_{74}	−0.6398	−0.6591
δ_7	0.2427	0.3150	δ_{41}	−0.7056	0.1579	δ_{75}	−0.6972	−0.6544
δ_8	0.4423	0.4716	δ_{42}	−0.6994	0.1678	δ_{76}	−0.7432	−0.6083
δ_9	0.8695	0.3309	δ_{43}	−0.6850	0.2173	δ_{77}	−0.7419	−0.7204
δ_{10}	0.1493	0.4118	δ_{44}	−0.6736	0.1940	δ_{78}	−0.5031	−0.4262
δ_{11}	0.2752	0.5384	δ_{45}	−0.7098	0.1684	δ_{79}	−0.7822	−0.7346
δ_{12}	0.8471	0.4188	δ_{46}	−0.7444	0.1189	δ_{80}	−0.7887	−0.6016
δ_{13}	−0.2283	0.3482	δ_{47}	−0.6807	0.1718	δ_{81}	−0.6764	−0.6122
δ_{14}	−0.2443	0.3811	δ_{48}	−0.6607	0.2165	δ_{82}	−0.7745	−0.7189
δ_{15}	0.5215	0.5987	δ_{49}	−0.6822	0.1903	δ_{83}	−0.6826	−0.6234
δ_{16}	−0.4252	0.3073	δ_{50}	−0.6379	−0.0017	δ_{84}	−0.7661	−0.6297
δ_{17}	−0.4261	0.3821	δ_{51}	−0.7420	0.0994	δ_{85}	−0.8396	−0.6690
δ_{18}	−0.0792	0.5576	δ_{52}	−0.8271	−0.0048	δ_{86}	−0.7552	−0.6872
δ_{19}	−0.5153	0.3404	δ_{53}	−0.6361	0.2464	δ_{87}	−0.8251	−0.5684
δ_{20}	−0.5148	0.3508	δ_{54}	−0.6898	0.1517	δ_{88}	−0.7260	−0.6365
δ_{21}	−0.3037	0.4343	δ_{55}	−0.6458	0.0193	δ_{89}	−0.8559	−0.6777
δ_{22}	−0.5587	0.3230	δ_{56}	−0.6022	0.1132	δ_{90}	−0.7886	−0.7645
δ_{23}	−0.6082	0.2896	δ_{57}	−0.5814	−0.1209	δ_{91}	−0.8317	−0.6697
δ_{24}	−0.4564	0.3598	δ_{58}	−0.6732	−0.2609	δ_{92}	−0.8338	−0.6878
δ_{25}	−0.6179	0.2654	δ_{59}	−0.6495	−0.1502	δ_{93}	−0.8369	−0.5453
δ_{26}	−0.6714	0.2166	δ_{60}	−0.6042	0.1855	δ_{94}	−0.8493	−0.6150
δ_{27}	−0.5679	0.2829	δ_{61}	−0.7636	−0.5206	δ_{95}	−0.7974	−0.6488
δ_{28}	−0.6454	0.2514	δ_{62}	−0.7282	−0.5287	δ_{96}	−0.7990	−0.5534
δ_{29}	−0.6456	0.2534	δ_{63}	−0.7249	−0.5580	δ_{97}	−0.8052	−0.6889
δ_{30}	−0.6032	0.2764	δ_{64}	−0.5576	−0.2542	δ_{98}	−0.8922	−0.6881
δ_{31}	−0.6560	0.2386	δ_{65}	−0.6364	−0.5076	δ_{99}	−0.8306	−0.6236
δ_{32}	−0.6725	0.2241	δ_{66}	−0.7141	−0.6276	δ_{100}	−0.7739	−0.7307
δ_{33}	−0.6382	0.2405	δ_{67}	−0.7081	−0.5187			
δ_{34}	−0.6743	0.2127	δ_{68}	−0.6760	−0.4054			

Resultados da medida de Kullback-Liebler

Na tabela A.4 encontram-se os resultados referentes à medida de similaridade Kullback-Liebler.

Tabela A.4: Resultados da aplicação da medida Kullback-Liebler às seguintes distribuições: *Modelo*(p), mistura de quatro distribuições geométricas com os parâmetros iniciais do algoritmo EM, $Mgeom(\Psi^{(0)})$, e mistura de quatro distribuições geométricas com parâmetro obtidos pelo algoritmo EM, $Mgeom(\Psi^{(EM)})$, para cada uma das espécies em estudo.

Esp.	Modelo (p)	$Mgeom(\Psi^{(0)})$	$Mgeom(\Psi^{(EM)})$	Esp.	Modelo (p)	$Mgeom(\Psi^{(0)})$	$Mgeom(\Psi^{(EM)})$
Ap	0.0059	0.0062	**0.0037**	*Cf*	0.0160	0.0174	**0.0003**
Hr	0.0222	0.0288	**0.0200**	*Eq*	0.0109	0.0119	**0.0001**
Mj	0.0131	0.0306	**0.0015**	*Gg*	0.0066	0.0075	**0.0002**
Pf	0.0099	0.0111	**0.0011**	*Am*	0.0154	0.0304	**0.0002**
Tk	0.0079	0.0079	**0.0043**	*Dm*	0.0090	0.0098	**0.0002**
Ba	0.0047	0.0102	**0.0015**	*Mu*	0.0130	0.0142	**0.0006**
Bs	0.0058	0.0060	**0.0013**	*Ce*	0.0156	0.0214	**0.0004**
Ct	0.0049	0.0057	**0.0010**	*Rn*	0.0124	0.0134	**0.0007**
Cb	0.0072	0.0333	**0.0022**	*Xt*	0.0064	0.0078	**0.0001**
Dv	0.0075	0.0106	**0.0067**	*Hs*	0.0116	0.0130	**0.0002**
Ec	0.0033	0.0033	**0.0024**	*Mm*	0.0112	0.0126	**0.0003**
Hi	0.0091	0.0113	**0.0012**	*Pt*	0.0112	0.0127	**0.0002**
Hp	0.0235	0.0257	**0.0009**	*Dr*	0.0096	0.0149	**0.0005**
Mg	0.0167	0.0282	**0.0051**	*Fu*	0.0111	0.0112	**0.0007**
Pa	0.0166	0.0240	**0.0150**	*Oa*	0.0121	0.0122	**0.0005**
Sa	0.0038	0.0127	**0.0027**	*Dd*	0.0307	0.1053	**0.0050**
Sm	0.0083	0.0116	**0.0030**	*Li*	0.0213	0.0230	**0.0133**
St	0.0051	0.0066	**0.0016**	*Pl*	0.0165	0.1137	**0.0073**
At	0.0056	0.0110	**0.0003**	*Tb*	0.0148	0.0149	**0.0011**
Os	0.0128	0.0132	**0.0005**	*Ca*	0.0092	0.0193	**0.0025**
Po	0.0119	0.0226	**0.0001**	*Nc*	0.0126	0.0126	**0.0018**
Vv	0.0125	0.0215	**0.0001**	*Sc*	0.0049	0.0075	**0.0010**
Bt	0.0086	0.0095	**0.0001**	*Sp*	0.0053	0.0098	**0.0009**

Distribuição empírica *vs* distribuições teóricas

Na Figura A.7 são apresentadas as distribuições empíricas e as distribuições teóricas *Modelo(p)* e $Mgeom(\Psi^{(EM)})$ para a sequência de distâncias global das espécies *At*, *Os*, *Po*, *Vv*. A linha a azul diz respeito à distribuição *Modelo(p)* e a linha a verde diz repeito à distribuição $Mgeom(\Psi^{(EM)})$.

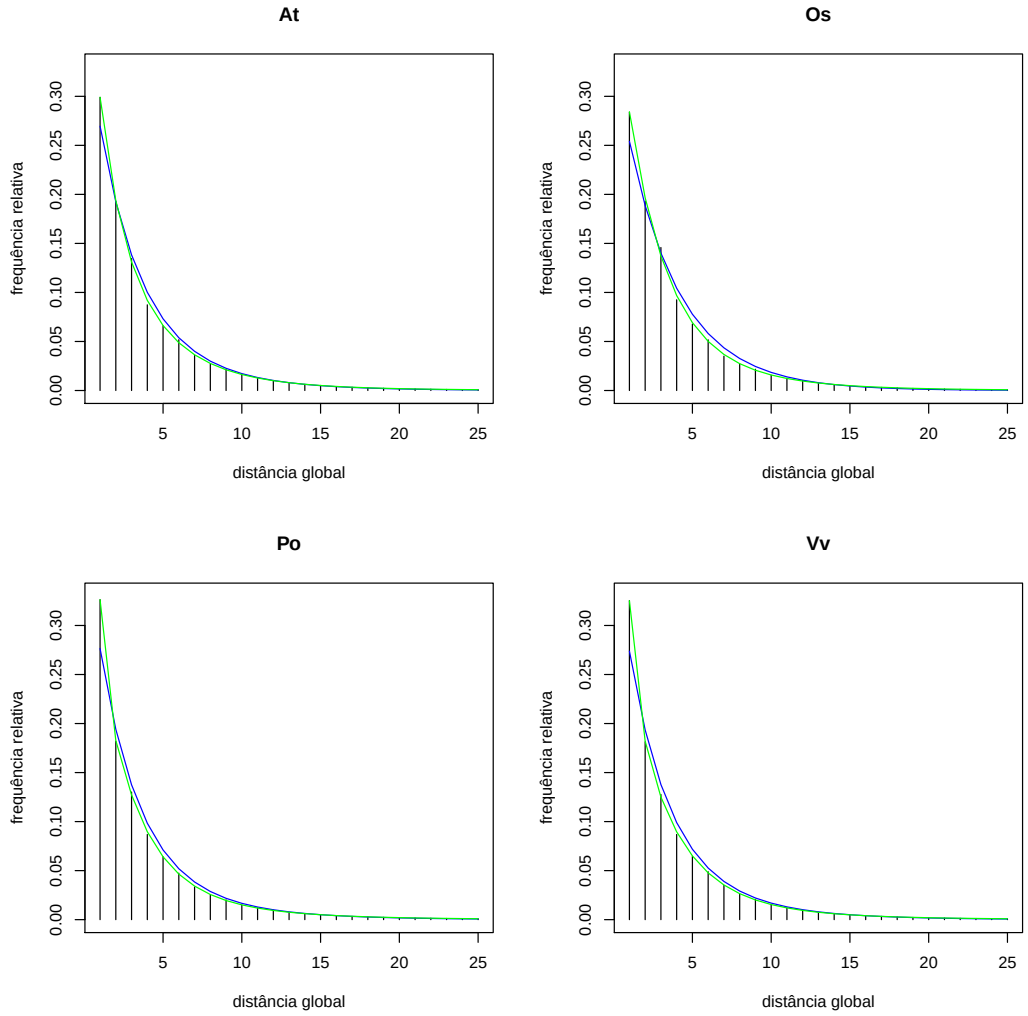


Figura A.7: Distribuição empírica (gráfico de barras) *vs* Distribuições teóricas *Modelo(p)* (linha azul) e $Mgeom(\Psi^{(EM)})$ (linha verde), das espécies *At*, *Os*, *Po* e *Vv*.

Apêndice B

Código R

A seguir apresenta-se o código R desenvolvido para suportar a análise estatística efectuada nesta dissertação, o qual inclui várias funções e *scripts*. O código R apresentado foi desenvolvido e testado na versão 2.11.0 do R, com as seguintes packages adicionadas à instalação base: FactoMineR (Versão 1.12) e Hmisc (Versão 3.7-0). O sistema operativo utilizado foi o Linux Fedora 12.

Carregamento dos dados

```
# Packages
library(Hmisc)
library(FactoMineR)
# Dados: distâncias entre nucleótidos, distâncias global entre nucleótidos e matriz
# dos erros relativos
dataFrame_nucleotidos = read.csv("distnucleotidos.csv",header = TRUE)
dataFrame_distanciaGlobal = read.csv("distglobal.csv",header = TRUE)
dataFrame_erros = read.csv("erros100dist.csv",header=FALSE)

d = numeric(101)
for(i in 1:101){
  d[i] = paste("d",i,sep="")
}
colnames(dataFrame_erros) = c(" ",d)
rownames(dataFrame_erros) = dataFrame_erros[,1]
erros = dataFrame_erros[,-1]
```

Sumário de estatísticas e representação das caixas de bigodes

```
# Script usado para gerar os valores da Tabela 2.3 e os gráficos da Figura 2.1
wtd.quantile.DataFrame =
    function(dataFrame, colunasConsideradas, header=FALSE) {
  numeroDeColunasDaDataFrame = length(dataFrame[, colunasConsideradas])
  nomesDasColunasDaDataFrame =
      names(dataFrame[2:numeroDeColunasDaDataFrame])
  stats = matrix(nrow=numeroDeColunasDaDataFrame-1,ncol=5)
  n = vector()
  out = vector()
  group = vector()
  bxpDados = list()
  listaDeResultados = list()
  matrizDeResultados = matrix(nrow=numeroDeColunasDaDataFrame-1,ncol=7)

  for (i in 2:numeroDeColunasDaDataFrame) {
    numeroDeObservacoesDaColunaCorrente = NROW(na.omit(dataFrame[[i]]))
    quantisParciaisDaColunaCorrente = wtd.quantile(as.numeric(dataFrame[,1]
        [1:numeroDeObservacoesDaColunaCorrente]), as.numeric(dataFrame[,i]
        [1:numeroDeObservacoesDaColunaCorrente]))[2:4]
    quantisCompletosDaColunaCorrente =
        wtd.quantile(as.numeric(dataFrame[,1]
        [1:numeroDeObservacoesDaColunaCorrente]),
        as.numeric(dataFrame[,i][1:numeroDeObservacoesDaColunaCorrente]))
    attr(quantisParciaisDaColunaCorrente,"names")= NULL

    limiteSuperiorParaOutliers = quantisParciaisDaColunaCorrente[3] +
        (1.5*(quantisParciaisDaColunaCorrente[3]
        -quantisParciaisDaColunaCorrente[1]))
    limiteInferiorParaOutliers = quantisParciaisDaColunaCorrente[1] -
        (1.5*(quantisParciaisDaColunaCorrente[3]
        -quantisParciaisDaColunaCorrente[1]))

    outliersSuperiores =
        na.omit(dataFrame[1:numeroDeObservacoesDaColunaCorrente,1]
        [dataFrame[1:numeroDeObservacoesDaColunaCorrente,1]>
        limiteSuperiorParaOutliers])
    outliersInferiores =
        na.omit(dataFrame[1:numeroDeObservacoesDaColunaCorrente,1]
```

```
        [dataFrame[1:numeroDeObservacoesDaColunaCorrente,1]<
        limiteInferiorParaOutliers])

    hingeSuperior =
        max(na.omit(dataFrame[1:numeroDeObservacoesDaColunaCorrente,1]
        [dataFrame[1:numeroDeObservacoesDaColunaCorrente,1]<=
        limiteSuperiorParaOutliers]))
    hingeInferior =
        min(na.omit(dataFrame[1:numeroDeObservacoesDaColunaCorrente,1]
        [dataFrame[1:numeroDeObservacoesDaColunaCorrente,1]>=
        limiteInferiorParaOutliers]))

    stats[i-1,]=c(hingeInferior,quantisParciaisDaColunaCorrente,
        hingeSuperior)
    n = c(n,numeroDeObservacoesDaColunaCorrente)
    out = c(out,outliersInferiores,outliersSuperiores)
    group = c(group,rep((i-1),length.out = (length(outliersInferiores) +
        length(outliersSuperiores))))

    matrizDeResultados[i-1,]=c(quantisCompletosDaColunaCorrente,
        wtd.mean(as.numeric(dataFrame[,1]
        [1:numeroDeObservacoesDaColunaCorrente]),
        as.numeric(dataFrame[,i][1:numeroDeObservacoesDaColunaCorrente])),
        sqrt(wtd.var(as.numeric(dataFrame[,1]
        [1:numeroDeObservacoesDaColunaCorrente]),as.numeric(dataFrame[,i]
        [1:numeroDeObservacoesDaColunaCorrente]))))
}
if(header == TRUE) {
  colnames(matrizDeResultados) =
      c("Min","1Q.","Med","3Q","Max","Media","Desvio Padrao")
}
bxpDados$stats = t(stats)
bxpDados$n = n
bxpDados$out = out
bxpDados$group = group
bxpDados$names = nomesDasColunasDaDataFrame

listaDeResultados$bxpDados = bxpDados
listaDeResultados$matrizDeResultados = matrizDeResultados
return(listaDeResultados)
```

```
}
quartis=wtd.quantile.DataFrame(dataFrame_distanciaGlobal, header=TRUE)
# Sumário de estatísticas
quartis$matrizDeResultados
rownames(quartis$matrizDeResultados) =
    c(colnames(dataFrame_distanciaGlobal[2:47]))
round(quartis$matrizDeResultados,2)

# Representação gráfica das caixas de bigodes
coresDosBoxPlots =
    c(rep("red",18),rep("blue",4),rep("black",16),rep("cyan",4),
    rep("green",4))
bxp(quartis$bxpDados, log="y", border = coresDosBoxPlots, medlty = 1,
    medlwd=2.5, xlab = "Especies", ylab="Distancia
    global", cex.axis=0.8, ylim=c(1,5000), las=2)
```

Representação das distribuições empírica e da sequência de distâncias entre nucleótidos iguais

```
# Script usado para gerar os gráficos da Figura 2.2
especie = dataFrame_nucleotidos[,70:73] # Espécie St
numeroLinhas =NROW(na.omit(especie))
componentesIniciais = vector()
# Frequência relativa de cada nucleótido
for (j in 1:4){
  componentesIniciais[j] = sum(as.numeric(especie[1:numeroLinhas,j]))
}
N = sum(as.numeric(componentesIniciais))
componentesIniciais = round((componentesIniciais/N), 4)
nome = c("A","C","G","T")
x = c(0:25)
# Representação da distribuição empírica e da distribuição d^x
par(mfrow=c(2,2))
for(j in 1:4){
  n = sum(as.numeric(especie[1:numeroLinhas,j]))
  frespecie = especie[1:25,j] / n
  plot(frespecie, ylim=c(0,0.4),type="h",xlab="",ylab="",main=nome[j])
```

```
  y = dgeom(x, componentesIniciais[j])
  lines(x+1,y,type="l",lwd=1,col="blue")
}
```

Função massa de probabilidade da mistura de distribuições geométricas

```
funcaoMassaDaMistura = function(x,parametros, theta){
  sum(dgeom(x,parametros)*theta)
}
```

Representação de uma mistura de duas distribuições geométricas e das suas componentes

```
n = 200 # Dimensão da amostra
pesos = c(0.4,0.6)
parametros = c(0.3,0.5)
# Recolher uma amostra com reposição de acordo com os pesos
k = sample(1:2, size=n, replace=TRUE, prob=pesos)
# Gerar os valores correspondentes às distribuições
rate = parametros[k]
x = rgeom(n,prob=rate)
x=sort(x)
dim(x) = length(x)
# Cálculo da função massa de probabilidade da mistura
y = apply(x,1,funcaoMassaDaMistura, parametros=parametros,theta=pesos)
plot(x, y, type="l", ylim=c(0,0.6), lwd=3, col="blue" , xlab="", ylab=" ")
# Gerar o gráfico de cada distribuição geométrica individualmente
for(j in 1:2){
  y = apply(x, 1, dgeom, parametros[j])
  lines(x,y)
}
```

Representação dos Dendrogramas

```
# Matriz dos dados sem a coluna da classificação
erros.dendrograma = erros[,-101]
# Matriz de similaridades (distância euclidiana)
dendrograma.distancia = dist(erros.dendrograma, method="euclidian",
    upper=TRUE)
# Critérios de agregação: "complete linkage" e método "Ward"
hc = hclust(dendrograma.distancia, method="complete")
hc = hclust(dendrograma.distancia, method="Ward")
dendrograma1 = as.dendrogram(hc)

# Função para colorir as folhas do dendrograma
dendroCol = function(dend, vectorDasLabels, vectorDosGruposDasLabels,
    vectorDaListaDeGrupos, vectorDaListaDeCores) {
  if(is.leaf(dend)) {
    atributosAnteriores = attributes(dend)
    listaDeCorrespondenciaDeCores = list()
    numeroDeGrupos = length(vectorDaListaDeGrupos)
    for (i in 1:numeroDeGrupos) {
      listaDeCorrespondenciaDeCores[vectorDaListaDeGrupos[i]] =
          vectorDaListaDeCores[i]
    }
    numeroDeLabels = length(vectorDasLabels)
      for (j in 1:numeroDeLabels) {
        if (vectorDasLabels[j] == atributosAnteriores$label) {
          attr(dend, "nodePar") = c(atributosAnteriores$nodePar,
              list(lab.col=listaDeCorrespondenciaDeCores
              [vectorDosGruposDasLabels[j]][[1]]))
        break
        }
      }
  }
  return(dend)
}
# Dados de entrada da função "dendroCol"
vectorDasLabels = rownames(erros)
vectorDosGruposDasLabels = as.vector(erros[,101])
vectorDaListaDeGrupos =
    c("bacteria","planta","animal","protozoario","fungo")
```

```
vectorDaListaDeCores = c("red","blue","black","cyan","green")
# Representação do dendrograma
dendrogramaColorido = dendrapply(dendrograma1,dendroCol,vectorDasLabels,
    vectorDosGruposDasLabels,vectorDaListaDeGrupos, vectorDaListaDeCores)
plot(dendrogramaColorido, pch="",ylab="Similaridades")
```

ACP - Variáveis padronizadas

```
acpPadronizada = PCA(erros, scale.unit=TRUE, ncp=100,
    quali.sup=101,graph=FALSE)
# Tabela com os valores próprios e percentagem de variação total
acpPadronizada.eig = round(acpPadronizada[1]$eig,4)
# Tabela com os vectores próprios (uso da função prcomp)
acpPadronizada.prcomp = prcomp(erros[,1:100], retx=TRUE, scale=TRUE,
    center=TRUE)
acpPadronizada.loadings = round(acpPadronizada.prcomp$rotation[,1:5],4)
# Correlação entre as variáveis e as três primeiras componentes principais
acpPadronizada.correlacao = round(acpPadronizada[2]$var$cor[,1:3],4)
# Valores do cosseno quadrado
acpPadronizada.cosseno = round(acpPadronizada[2]$var$cos2[,1:4],4)

# Representação gráfica dos primeiros 15 valores próprios
acpPadronizada.barplot = acpPadronizada.eig$eig[1:15]
n = length(acpPadronizada.barplot)
barplot(acpPadronizada.barplot,ylab="Variancia",ylim=c(0,50),
    cex.lab=1.3,cex.axis=1.2, space=0.3,names.arg = paste("CP", 1:n, sep =
    ""))
# Representação gráfica do círculo das correlações (variáveis)
plot(acpPadronizada,choix = "var",title=" ",axes = c(1,2),lim.cos2.var = 0)
# Representação gráfica dos indivíduos
plot(acpPadronizada,choix = "ind",habillage=101,axes = c(1, 2),title=" ")
```

ACP - Variáveis centradas

```
acpCentrada = PCA(erros, scale.unit=FALSE, ncp=100, quali.sup=101,
    graph=FALSE)
# Tabela com os valores próprios e percentagem de variação total
acpCentrada.eig = round(acpCentrada[1]$eig,4)
# Tabela com os vectores próprios
acpCentrada.prcomp = prcomp(erros[,1:100],retx=TRUE,
    scale=FALSE,center=TRUE)
acpCentrada.loadings = round(acpCentrada.prcomp$rotation[,1:5],4)
# Correlação entre as variáveis e as três primeiras componentes principais
acpCentrada.correlacao = round(acpCentrada[2]$var$cor[,1:3],4)

# Representação gráfica dos primeiros 15 valores próprios
acpCentrada.barplot = acpCentrada.eig$eig[1:15]
n = length(acpCentrada.barplot)
barplot(acpCentrada.barplot, ylab="Variancia", ylim=c(0,5), cex.lab=1.3,
    cex.axis=1.2, space=0.3, names.arg = paste("CP", 1:n, sep = ""))
# Representação gráfica dos indivíduos
plot(acpCentrada, choix = "ind", habillage=101, axes = c(1, 2), title=" ")
```

ACP - Variáveis não padronizadas

```
acpNaoPadronizada = prcomp(erros[,1:100],scale=FALSE, center=FALSE)
# Desvio padrão e proporção de variância explicada
summary(acpNaoPadronizada)
# Tabela com os vectores próprios
acpNaoPadronizada.loadings = round(acpNaoPadronizada$rotation[,1:2],4)
# Scores
acpNaoPadronizada.scores = acpNaoPadronizada$x[,1:2]
# Correlação entre as variáveis e as duas primeiras componentes principais
acpNaoPadronizada.correlacao = cor(erros[,-101],acpNaoPadronizada.scores)

# Representação gráfica dos indivíduos
plot(acpNaoPadronizada.scores,cex=0.5,col=as.numeric(erros[,101]),pch=19,
    xlab="Dim1 (91.2%)", ylab="Dim2 (2.9%)")
```

```
abline(h = 0,lty=2)
abline(v = 0,lty=2)
text(acpNaoPadronizada.scores,labels= rownames(erros),pos=1,offset = -0.9,
    col=as.numeric(erros[,101]))
```

Decomposição em valores singulares (DVS)

```
valores = svd(erros[,1:100])
# Valores singulares
valoresSingulares = round(valores$d,4)
# Valores singulares à esquerda
valoresSingularesEsquerda = round(valores$u,4)
# Valores singulares à direita
valoresSingularesDireita = round(valores$v,4)

# Percentagem de variância explicada pelas CPs
varianciaCP = (valores$d)^2
soma = sum(varianciaCP)
n = length(varianciaCP)
percentagemVarianciaTotal = vector()
for(i in 1:n){
  percentagemVarianciaTotal[i] = (varianciaCP[i] / soma) * 100
}
round(percentagemVarianciaTotal,2)
```

K-means

```
# K-means - aplicado às dez primeiras variáveis
cl = kmeans(erros[,-(11:101)], 2,  algorithm ="Lloyd")
cl$size
```

```
# Representação dos indivíduos
plot(x,col=cl$cluster, pch=16)
points(cl$centers, col = 9:10, pch = 8, cex=2)

# Kmeans - aplicado aos scores das componente CP1 e CP2
cl = kmeans(acpNaoPadronizada.scores, 2)
# Representação dos indivíduos com identificação
plot(acpNaoPadronizada.scores, pch = 16,asp = 1,cex = 0.5,col=cl$cluster,
    xlab="Dim1 (91.2%)",ylab="Dim2 (2.9%)")
text(acpNaoPadronizada.scores,rownames(erros),col=cl$cluster)
points(cl$centers,col = 1:2,pch = 8,cex=1)
```

Algoritmo EM - dados categorizados

```
# A função EMmisturas retorna as estimativas do vector dos parâmetros de
# uma mistura finita de g distribuições geométricas, implementando o algoritmo EM
# para dados categorizados. Recebe os seguintes parâmetros de entrada:
#   - um vector y com a frequência absoluta das distâncias
#   - um vector pesosIniciais de dimensão g-1
#   - um vector componentesIniciais de dimensão g (contém os parâmetros iniciais
#     das componentes)
#   - o número g de componentes da mistura
#   - o critério de paragem do algoritmo, epson
#   - o número de espécies n

EMmisturas = function(y,pesosIniciais,componentesIniciais,g,epson=1e-5,n){
L = length(y)
N = sum(as.numeric(y))
indicesDaAmostra = c(1:L)
pesosIniciais = c(pesosIniciais,1-sum(pesosIniciais))

pesosCorrentes = vector(mode="numeric",g)
pesosSeguintes = vector(mode="numeric",g)   # Pesos da mistura
logverosimilhancacorrente= vector()
componentesCorrentes = vector(mode="numeric",g)
componentesSeguintes = vector(mode="numeric",g)   # Componentes da mistura
```

```
pesosSeguintes = pesosIniciais
componentesSeguintes = componentesIniciais
k = 1
verosimilhancaAnterior = 0
verosimilhancaCorrente = 1

repeat {
  matriz_Z_Corrente = matrix(0,g,L)
  pesosCorrentes = pesosSeguintes
  componentesCorrentes = componentesSeguintes
  for(m in 1:g) {
    for(j in 1:L) {
      z_mj_k_Numerador = 0
      z_mj_k_Numerador = pesosCorrentes[m]*componentesCorrentes[m]*(1 -
          componentesCorrentes[m])^(j-1)
      z_mj_k_Denominador = 0
      for(h in 1:g){
        z_mj_k_Denominador = z_mj_k_Denominador + (pesosCorrentes[h] *
            componentesCorrentes[h] * (1 - componentesCorrentes[h])^(j -
            1))
      }
      z_mj_k = z_mj_k_Numerador / z_mj_k_Denominador
      matriz_Z_Corrente[m,j] = z_mj_k   # Probabilidade a posteriori
    }
  }

  for(m in 1:g) {
    pesosSeguintes[m]=0
    for(j in 1:L) {
      pesosSeguintes[m] = pesosSeguintes[m] + matriz_Z_Corrente[m,j] *
          y[j] / N
    }
  }
  for(m in 1:g) {
    componentesSeguintes[m]=0
    for(j in 1:L) {
      componentesSeguintes[m] = componentesSeguintes[m] +
          ((matriz_Z_Corrente[m,j] * j * y[j]) / (pesosSeguintes[m] * N))
    }
```

```
    componentesSeguintes[m] = 1 / componentesSeguintes[m]
  }
  verosimilhancaCorrente = 0
    for(j in 1:L) {
      somatorioInterior = 0
      for(m in 1:g) {
        somatorioInterior = somatorioInterior + (pesosCorrentes[m] *
            componentesCorrentes[m] * (1 - componentesCorrentes[m])^(j-1))
      }
      verosimilhancaCorrente = verosimilhancaCorrente + y[j] *
          log(somatorioInterior)
    }

  criterio = (verosimilhancaCorrente - verosimilhancaAnterior)
  if((k >= 2) & ( abs(criterio) < epson )) {
    vectorDeSaida = c(k-1, pesosCorrentes, componentesCorrentes)
    return(vectorDeSaida)
    break
  }
  verosimilhancaAnterior = verosimilhancaCorrente
  k = k + 1
  logverosimilhancacorrente[k] = verosimilhancaCorrente
  cat("Iteracao ",k-1,"-", "Pesos:",pesosSeguintes,"\n")
  cat("Iteracao ",k-1,"-", "Componentes:",componentesSeguintes,"\n")
  cat("Iteracao ",k-1,"-", "Criterio:",  criterio,"\n")
}
}
```

Estimativas iniciais para aplicação do algoritmo EM (mistura de 4 geométricas) às 46 espécies

```
# Script usado para gerar os dados da Tabela 4.4
componentesIniciais = matrix(0,46, 4)
rownames(componentesIniciais) =
    c(colnames(dataFrame_distanciaGlobal[2:47]))
k = 2
```

```
for (p in 1:46){
 colunaEspecieInicio = k
 colunaEspecieFim = colunaEspecieInicio + 3
 especie=dataFrame_nucleotidos[,colunaEspecieInicio:colunaEspecieFim]
 numeroLinhas =NROW(na.omit(especie))
 for (j in 1:4){
    componentesIniciais[p,j] = sum(as.numeric(especie[1:numeroLinhas,j]))
  }
  componentesIniciais[p,] =
      componentesIniciais[p,]/sum(as.numeric(componentesIniciais[p,]))
  k = k + 4
}
componentesIniciais = round(componentesIniciais,4)
```

Estimativas dos parâmetros da mistura de 4 geométricas obtidas via algoritmo EM às 46 espécies

```
# Script usado para gerar os dados da Tabela 4.5
estimativasEM = matrix(0,46,9)
for (p in 1:46){
  coluna = p+1
  numeroLinhasSemNa =NROW(na.omit(dataFrame_distanciaGlobal[,coluna]))
  y = dataFrame_distanciaGlobal[1:numeroLinhasSemNa,coluna]
  g = 4
  pesosIniciais = c(0.25,0.25,0.25)
  estimativasEM[p,] = EMmisturas(y,pesosIniciais,componentesIniciais[p,],g,
      epson=1e-5, p)
}
```

Funções usadas no cálculo das medidas de similaridade entre distribuições

```
# Função para o cálculo da medida de similaridade S1
medidaDistancia = function(L, frequenciaObservada, frequenciaEsperada){
  somaNumerador = 0
  somaDenominador = 0
  for (d in 1:L){
    somaNumerador = somaNumerador + abs(frequenciaObservada[d] -
        frequenciaEsperada[d])
    somaDenominador = somaDenominador + (abs(frequenciaObservada[d]) +
        abs(frequenciaEsperada[d]))
  }
  medidaDistancia = 1 - (somaNumerador / somaDenominador)
  return(medidaDistancia)
}
# Função para o cálculo da medida de Kullback-Liebler
kullbackLiebler = function(L,frequenciaObservada, frequenciaEsperada){
  soma = 0
  for(i in 1:L){
    soma = soma + frequenciaObservada[i] * (log(frequenciaObservada[i] /
        frequenciaEsperada[i]))
  }
  return(soma)
}
# Função massa de probabilidade do modelo
funcaoMassaModelo = function(L,parametrosDoModelo){
  probabilidadeModelo = vector(mode="numeric",numeroLinhas)
  for (d in 1:L){
    soma = 0
      for (i in 1:4){
        soma = soma + (parametrosDoModelo[i])^(2) * (1 -
            parametrosDoModelo[i])^(d - 1)
      }
      probabilidadeModelo[d] = soma
  }
  return(probabilidadeModelo)
}
```

Cálculo das medidas de similaridade entre distribuições para as 46 espécies

```
# Script usado para gerar os valores da Tabela 4.6 e da Tabela A.4
medidaEntreDistribuicoes = matrix(0,46,6)
rownames(medidaEntreDistribuicoes) =
    c(colnames(dataFrame_distanciaGlobal[2:47]))
for (p in 1:46){
  coluna =p+1
  numeroLinhasSemNa = NROW(na.omit(dataFrame_distanciaGlobal[,coluna]))
  distanciaGlobal = dataFrame_distanciaGlobal[1:numeroLinhasSemNa,coluna]
  distanciaGlobal = distanciaGlobal[distanciaGlobal > 0]
  L =  length(distanciaGlobal)

  # Frequência relativa da distância global para uma dada espécie
  frequenciaObservadaDosDados = distanciaGlobal /
      sum(as.numeric(distanciaGlobal))
  parametrosDoModelo = componentesIniciais[p,]
  frequenciaEsperadaModelo = funcaoMassaModelo(L,parametrosDoModelo)

  pesosIniciais = c(0.25,0.25,0.25,0.25)
  x = c(0:(L-1))
  dim(x) = length(x)

  # Cálculo da probabilidade da mistura: valores iniciais + valores EM
  frequenciaEsperadaDaMistura4ValoresIniciais =
      apply(x,1,funcaoMassaDaMistura,parametros=componentesIniciais[p,],
      theta=pesosIniciais)
  frequenciaEsperadaDaMistura4EM = apply(x,1,funcaoMassaDaMistura,
      parametros=as.matrix(estimativasEM[p,6:9]),
      theta=as.matrix(estimativasEM[p,2:5]))

  # Cálculo das medidas de similaridades S^1 e Kullback-Liebler
  medidaDistanciaModelo = medidaDistancia(L,frequenciaObservadaDosDados,
      frequenciaEsperadaModelo)
  medidaEntreDistribuicoes[p,1] = medidaDistanciaModelo
  medidaDistanciaModelo4Inicial =
      medidaDistancia(L,frequenciaObservadaDosDados,
      frequenciaEsperadaDaMistura4ValoresIniciais)
  medidaEntreDistribuicoes[p,2] = medidaDistanciaModelo4Inicial
```

```
    medidaDistanciaModelo4EM = medidaDistancia(L,frequenciaObservadaDosDados,
        frequenciaEsperadaDaMistura4EM)
    medidaEntreDistribuicoes[p,3] = medidaDistanciaModelo4EM
    medidaKullbackLieblerModelo = kullbackLiebler(L,
        frequenciaObservadaDosDados, frequenciaEsperadaModelo)
    medidaEntreDistribuicoes[p,4] = medidaKullbackLieblerModelo
    medidaKullbackLieblerModelo4Inicial = kullbackLiebler(L,
        frequenciaObservadaDosDados,
        frequenciaEsperadaDaMistura4ValoresIniciais)
    medidaEntreDistribuicoes[p,5] = medidaKullbackLieblerModelo4Inicial
    medidaKullbackLieblerModelo4EM = kullbackLiebler(L,
        frequenciaObservadaDosDados, frequenciaEsperadaDaMistura4EM)
    medidaEntreDistribuicoes[p,6] = medidaKullbackLieblerModelo4EM
}
medidaEntreDistribuicoes = round(medidaEntreDistribuicoes,4)
```

Teste de ajustamento do qui-quadrado

```
# Aplicação do teste de ajustamento à espécie St
coluna = 19
numeroLinhasSemNa = NROW(na.omit(dataFrame_distanciaGlobal[,coluna]))
frequenciaSuperior = 0
distanciaGlobal= dataFrame_distanciaGlobal[1:numeroLinhasSemNa,coluna]
distanciaGlobal = distanciaGlobal[distanciaGlobal > frequenciaSuperior]
frequenciaObservada = distanciaGlobal

# Cálculo das probabilidades da mistura de geométricas
L = length(frequenciaObservada)
x = c(0:(L-1))
dim(x) = length(x)
N = sum(as.numeric(frequenciaObservada))
# Estimativas obtidas pelo algoritmo EM
pesos = c(0.84874, 0.09261, 0.05866)
componentes = c(0.25982, 0.13727, 1)
pk = apply(x,1,funcaoMassaDaMistura, parametros=componentes, theta=pesos)
```

```
# Teste do qui-quadrado
chisq.test(frequenciaObservada, p =pk, rescale.p=TRUE)

# Cálculo do valor da estatística do teste qui-quadrado
frequenciaEsperada = N * pk
Q = 0
for(k in 1:L){
  Q = Q + ((frequenciaObservada[k] - frequenciaEsperada[k])^2) /
      frequenciaEsperada[k]
}
print(Q)
# Cálculo do quantil
qchisq(0.01,(L-1))
```

Representação gráfica das distribuições: empírica, modelo e mistura de 4 geométricas (via EM)

```
# Script usado para gerar os gráficos da Figura 4.8 e da Figura A.7
numeroDeLinhas = 100
numeroDeColunas = 46
par(mfrow=c(2,2))
nomeDasEspecies = c("Ap","Hr","Mj","Pf","Tk","Ba","Bs","Ct","Cb","Dv","Ec",
    "Hi","Hp","Mg","Pa","Sa","Sm","St", "At","Os","Po",
   "Vv","Bt","Cf","Eq","Gg","Am", "Dm","Mu","Ce","Rn","Xt","Hs","Mm",
   "Pt","Dd","Li","Pl","Tb","Dr","Fu","Oa","Ca","Nc","Sc","Sp")

# Primeira posição de cada conjunto sequencial de quatro espécies
# no vector nomeDasEspecies
p = 19
for(i in p:(p+3)){
  coluna = i + 1
  colunaCorrente = dataFrame_distanciaGlobal[,coluna]
  colunaCorrente[is.na(colunaCorrente)] = 0
  distanciaGlobal = sum(as.numeric(colunaCorrente))
  frDistanciaGlobal = colunaCorrente / distanciaGlobal
  parametrosDoModelo = componentesIniciais[i,]
  frequenciaEsperadaModelo =
      funcaoMassaModelo(numeroDeLinhas,parametrosDoModelo)
```

```
  x = c(0:(25-1))
  dim(x) = length(x)
  frequenciaEsperadaDaMistura4EM = apply(x,1,funcaoMassaDaMistura,
      parametros=as.matrix(estimativasEM[i,6:9]),
      theta=as.matrix(estimativasEM[i,2:5]))

  plot(x+1,frDistanciaGlobal[1:25],type="h",ylim=c(0,0.33),xlab="distancia
      global", ylab="frequencia relativa", main= nomeDasEspecies[i])
  lines(x+1, frequenciaEsperadaModelo[1:25], col = "blue")
  lines(x+1, frequenciaEsperadaDaMistura4EM[1:25], col = "green")
}
```

Matriz dos erros relativos

```
# Função para o cálculo da matriz dos erros relativos
matrizErros = function(L, frequenciaObservadaDosDados,
    frequenciaEsperadaDoModelo) {
  erro = vector(mode="numeric",L)
  for (k in 1:L){
    erro[k] = (frequenciaObservadaDosDados[k] -
        frequenciaEsperadaDoModelo[k]) / frequenciaObservadaDosDados[k]
  }
  return(erro)
}

componentesIniciais = matrix(0,46, 4)
rownames(componentesIniciais) =
    c(colnames(dataFrame_distanciaGlobal[2:47]))
k=2
for (p in 1:46){
  colunaEspecieInicio = k
  colunaEspecieFim = colunaEspecieInicio + 3
  especie=dataFrame_nucleotidos[,colunaEspecieInicio:colunaEspecieFim]
  numeroLinhas =NROW(na.omit(especie))
  for (j in 1:4){
    componentesIniciais[p,j] = sum(as.numeric(especie[1:numeroLinhas,j]))
```

```
    }
    componentesIniciais[p,] =
        componentesIniciais[p,]/sum(as.numeric(componentesIniciais[p,]))
    k = k + 4
}
componentesIniciais

numeroDeLinhas = 100
numeroDeColunas = 46
matrizDosErros = matrix(0,numeroDeLinhas,numeroDeColunas)
for (p in 1:numeroDeColunas){
    coluna = p + 1
    colunaCorrente = dataFrame_distanciaGlobal[,coluna]
    colunaCorrente[is.na(colunaCorrente)] = 0
    distanciaGlobal = sum(as.numeric(colunaCorrente))
    frDistanciaGlobal = colunaCorrente / distanciaGlobal
    parametrosDoModelo = componentesIniciais[p,]
    frequenciaEsperadaModelo = funcaoMassaModelo(numeroDeLinhas,
        parametrosDoModelo)

    matrizDosErros[,p] = matrizErros(numeroDeLinhas, frDistanciaGlobal,
        frequenciaEsperadaModelo)
}
# Conversão dos valores NA a zero
matrizDosErros[is.infinite(matrizDosErros)] = 0
matrizDosErros = t(matrizDosErros)
rownames(matrizDosErros) = c(colnames(dataFrame_distanciaGlobal[2:47]))
matrizDosErros
```