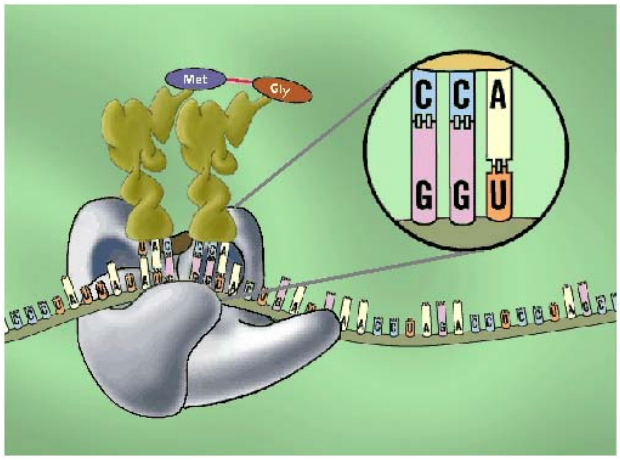
Universidade de Aveiro Departamento de Matemática
2002

Vera Mónica Almeida Afreixo

Análise Estatística da Linguagem Genética



Universidade de Aveiro Departamento de Matemática
2002

Vera Mónica Almeida Afreixo

Análise Estatística da Linguagem Genética

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática, realizada sob a orientação científica da Profª Doutora Adelaide de Fátima Baptista Valente Freitas, Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro.

o júri

presidente	Doutor Helmuth Robert Malonek, Professor Catedrático da Universidade de Aveiro.
vogais	Doutora Paula Manuela Lemos Pereira Milheiro de Oliveira, Professora Associada da Faculdade de Engenharia da Universidade do Porto.
	Doutora Adelaide de Fátima Baptista Valente Freitas, Professora Auxiliar da Universidade de Aveiro (Orientadora).

agradecimentos

À minha orientadora, Professora Doutora Adelaide Freitas, pela sua orientação científica, apoio e disponibilidade.

A todos os professores do Departamento de Matemática da Universidade de Aveiro que de alguma forma me apoiaram na parte escolar e na realização deste trabalho.

Ao Mestre Arnaldo Oliveira pela leitura e sugestões apresentadas.

À minha colega de Mestrado, Maria Helena Silva, pelo saudável espírito com que ultrapassamos a parte escolar.

Aos meus pais, avó e irmão, pelo apoio e carinho oferecidos nos momentos mais difíceis. A eles dedico este trabalho.

resumo

O objectivo principal deste trabalho é analisar a linguagem genética no contexto dos codões, ou seja, da parte codificante dos genes responsável pela produção de proteínas. Concretamente, pretende-se decifrar leis gerais que governem a tradução do mRNA pelo ribossoma. Para esse efeito foram utilizados dados genéticos de duas espécies distintas, que partilham todavia o mesmo ancestral: *Candida albicans* e *Saccharomyces cerevisiae*.

No presente estudo são empregues diferentes metodologias e modelos estatísticos adequados a dados de natureza discreta; nomeadamente, Análise de Tabelas de Contingência, Análise Classificatória, Análise em Componentes Principais, Cadeias de Markov, Análise de Zipf, Critério de Informação Bayesiana e Teoria da Informação. Com as Tabelas de Contingência, averigua-se, do ponto de vista da independência e associação, o comportamento de pares de codões ou nucleótidos, justapostos ou espaçados. As Análises Classificatória e em Componentes Principais permitem estudar, de forma exploratória, a preferência de um codão face ao codão justaposto e aos seus nucleótidos constituíntes. As cadeias de Markov são aplicadas com o objectivo de averiguar a adequação do modelo no sequenciamento dos codões. A Análise de Zipf visa estimar a respectiva lei e averiguar a existência de correlações de longo alcance entre os codões sequenciados. Para estimar a ordem da cadeia de Markov no sequenciamento de codões é usado o Critério de Informação Bayesiana. A Teoria da Informação é aplicada com o intuito de obter valores de entropia no conjunto das sequências de código.

Tudo leva a crer que os textos genéticos são estruturas bem organizadas, em que existe alguma associação entre um dado codão e os símbolos (codões ou nucleótidos) justapostos ou espaçados. Esta associação decresce à medida que o espaçamento aumenta.

abstract

The main aim of this work is to analyse the genetic language at the codon context. In other words, the coding part of the genes responsible for protein production is studied with the goal of deciphering general laws which govern the mRNA translation by the ribosome. For this purpose, it was used genetic data from two species that share the same ancestral: *Candida albicans* e *Saccharomyces cerevisiae*.

In this study different methodologies and statistical models are employed, namely: Contingency Tables, Cluster Analysis, Principal Components Analysis, Markov Chains, Zipf Analysis, Bayesian Information Criterion and Information Theory. With the Contingency Tables, we investigate, from the independency and association point of view, the behaviour of the codon or nucleotide pairs, placed side by side or spaced. The Cluster Analysis and Principal Component Analysis allow studying, in an exploratory way, the preference of a codon relative to its adjacent and its nucleotides. The Markov Chains are applied with the goal of investigate the fitting of the model in the codon sequencing. The Zipf Analysis aims to estimate the respective law and examine the existence of long range correlations among sequencing codons. The Bayesian Information Criterion is applied to estimate the order of the Markov chain in the codon sequencing. Finally, the Information Theory is used to obtain entropy values for the set of code sequences.

As a result of this study, we are inclined to think that genetic texts are well organized structures, with some association between a given codon and contiguous or spaced symbols (codons or nucleotides). That association decreases as the spacing goes by.

Os conhecimentos matemáticos são proposições construídas pelo nosso intelecto de modo a funcionarem sempre como verdadeiras, ou porque são inatas ou porque a matemática foi inventada antes das outras ciências. E a biblioteca foi construída por uma mente humana que pensava de modo matemático, porque sem matemática não se fazem labirintos. E, portanto, trata-se de confrontar as nossas proposições matemáticas com as proposições do construtor, e deste confronto pode surgir ciência, porque é ciência de termos sobre termos. E em todo o caso pára de me arrastar para discussões metafísicas. Que bicho te mordeu hoje?

O nome da rosa. Umberto Eco.

Índice

1 Introdução **1**
- 1.1 Contextualização . . . 1
- 1.2 Conceitos Biológicos . . . 2
- 1.3 Motivação e Objectivos Gerais . . . 6
- 1.4 Organização da Dissertação . . . 7

2 Análise de Tabelas de Contingência **9**
- 2.1 Introdução . . . 9
- 2.2 Nomenclatura . . . 10
- 2.3 Independência e Associação . . . 11
  - 2.3.1 Testes de Ajustamento . . . 11
  - 2.3.2 Independência . . . 12
  - 2.3.3 Associação . . . 15
- 2.4 Análise de Resíduos . . . 16
- 2.5 Aplicação ao Caso em Estudo . . . 17
  - 2.5.1 Determinação dos Graus de Liberdade . . . 17
  - 2.5.2 Independência e Associação de Pares de Codões Justapostos . . . 18
  - 2.5.3 Análise de Resíduos de Pares de Codões Justapostos . . . 20
  - 2.5.4 Independência e Associação de Pares de Codões Espaçados de Cinco codões . . . 25

3 Análise Classificatória **27**
- 3.1 Introdução . . . 27
- 3.2 Nomenclatura . . . 28
- 3.3 Métodos Hierárquicos . . . 31
  - 3.3.1 Dendograma . . . 32
  - 3.3.2 Características do Processo de Agrupamento . . . 33
  - 3.3.3 Escolha do Número de Grupos . . . 34
- 3.4 Aplicação ao Caso em Estudo . . . 34

4 Análise em Componentes Principais **43**
- 4.1 Introdução . . . 43
- 4.2 Nomenclatura . . . 43
- 4.3 Método de Análise em Componentes Principais . . . 43
  - 4.3.1 Procedimento . . . 45
  - 4.3.2 Decomposição da Variância . . . 48
  - 4.3.3 Critérios de Selecção das Componentes . . . 49

4.3.4 Validação da Aplicação da Análise em Componentes Principais 50
4.3.5 Rotação das Componentes Principais 51
4.4 Aplicação ao Caso em Estudo 52

5 Cadeias de Markov **63**
5.1 Introdução 63
5.2 Cadeia de Markov com Espaço de Parâmetros Discreto 63
5.2.1 Definição de Cadeia de Markov 63
5.2.2 Comportamento Limite das Cadeias de Markov 65
5.3 Aplicação ao Caso em Estudo 66

6 Análise das Frequências dos Símbolos **73**
6.1 Introdução 73
6.2 A Gramática das Sequências de Código 74
6.3 Lei de Zipf 74
6.4 Análise de Zipf sobre o n-uplo 75
6.5 Aplicação ao Caso em Estudo 76
6.5.1 Análise de Zipf 77
6.5.2 Análise de Zipf sobre o Par 82
6.5.3 Análise de Zipf sobre o Terno 87
6.5.4 Realce dos n-uplos mais Frequentes 90
6.5.5 Análise Comparativa das Frequências entre as Espécies 91

7 Critério de Informação Bayesiana em Cadeias de Markov de Ordem k **93**
7.1 Introdução 93
7.2 Cadeia de Markov de Ordem k 93
7.3 Critério de Informação Bayesiana (BIC) 94
7.4 Aplicação ao Caso em Estudo 94
7.4.1 Concretização da Notação para Sequências de Codões 94
7.4.2 Concretização para Genes 96

8 Teoria da Informação **99**
8.1 Introdução 99
8.2 Entropia 100
8.3 Conceitos e Propriedades 101
8.4 Aplicação ao Caso em Estudo 102

9 Conclusão **105**

Apêndice A **107**

Apêndice B **111**

Apêndice C **123**

Bibliografia **127**

Índice

1 Introdução **1**
- 1.1 Contextualização . . . 1
- 1.2 Conceitos Biológicos . . . 2
- 1.3 Motivação e Objectivos Gerais . . . 6
- 1.4 Organização da Dissertação . . . 7

2 Análise de Tabelas de Contingência **9**
- 2.1 Introdução . . . 9
- 2.2 Nomenclatura . . . 10
- 2.3 Independência e Associação . . . 11
  - 2.3.1 Testes de Ajustamento . . . 11
  - 2.3.2 Independência . . . 12
  - 2.3.3 Associação . . . 15
- 2.4 Análise de Resíduos . . . 16
- 2.5 Aplicação ao Caso em Estudo . . . 17
  - 2.5.1 Determinação dos Graus de Liberdade . . . 17
  - 2.5.2 Independência e Associação de Pares de Codões Justapostos . . . 18
  - 2.5.3 Análise de Resíduos de Pares de Codões Justapostos . . . 20
  - 2.5.4 Independência e Associação de Pares de Codões Espaçados de Cinco codões . . . 25

3 Análise Classificatória **27**
- 3.1 Introdução . . . 27
- 3.2 Nomenclatura . . . 28
- 3.3 Métodos Hierárquicos . . . 31
  - 3.3.1 Dendograma . . . 32
  - 3.3.2 Características do Processo de Agrupamento . . . 33
  - 3.3.3 Escolha do Número de Grupos . . . 34
- 3.4 Aplicação ao Caso em Estudo . . . 34

4 Análise em Componentes Principais **43**
- 4.1 Introdução . . . 43
- 4.2 Nomenclatura . . . 43
- 4.3 Método de Análise em Componentes Principais . . . 43
  - 4.3.1 Procedimento . . . 45
  - 4.3.2 Decomposição da Variância . . . 48
  - 4.3.3 Critérios de Selecção das Componentes . . . 49

4.3.4 Validação da Aplicação da Análise em Componentes Principais 50
4.3.5 Rotação das Componentes Principais 51
4.4 Aplicação ao Caso em Estudo 52

5 Cadeias de Markov **59**
5.1 Introdução 59
5.2 Cadeia de Markov com Espaço de Parâmetros Discreto 59
5.2.1 Definição de Cadeia de Markov 59
5.2.2 Comportamento Limite das Cadeias de Markov 61
5.3 Aplicação ao Caso em Estudo 62

6 Análise das Frequências dos Símbolos **69**
6.1 Introdução 69
6.2 A Gramática das Sequências de Código 70
6.3 Lei de Zipf 70
6.4 Análise de Zipf sobre o n-uplo 71
6.5 Aplicação ao Caso em Estudo 72
6.5.1 Análise de Zipf 73
6.5.2 Análise de Zipf sobre o Par 78
6.5.3 Análise de Zipf sobre o Terno 83
6.5.4 Realce dos n-uplos mais Frequentes 86
6.5.5 Análise Comparativa das Frequências entre as Espécies 87

7 Critério de Informação Bayesiana em Cadeias de Markov de Ordem k **89**
7.1 Introdução 89
7.2 Cadeia de Markov de Ordem k 89
7.3 Critério de Informação Bayesiana (BIC) 90
7.4 Aplicação ao Caso em Estudo 90
7.4.1 Concretização da Notação para Sequências de Codões 90
7.4.2 Concretização para Genes 92

8 Teoria da Informação **95**
8.1 Introdução 95
8.2 Entropia 96
8.3 Conceitos e Propriedades 97
8.4 Aplicação ao Caso em Estudo 98

9 Conclusão **101**

A **103**

B **107**

C **119**

Bibliografia **123**

Capítulo 1

Introdução

1.1 Contextualização

A Genética tem sido uma das áreas de investigação que sofreu um grande desenvolvimento nas últimas décadas, em particular após a descoberta da estrutura do ácido desoxirribonucleico (DNA - *DeoxyriboNucleic Acid*) e da forma como esta codifica as proteínas. Na realidade na última década, sequenciaram-se mais de 50 genomas. Assim, a uma velocidade exponencial, ficou disponível um grande número de sequências de DNA, tendo o mundo da Genética ficado mergulhado num enorme conjunto de dados sem aparente regularidade.
Recentemente, um pouco por todo o lado, grupos interdisciplinares têm estado a realizar trabalhos de investigação com o objectivo de extrair informação relevante contida no DNA. Nomeadamente, decifrar leis gerais que governem a tradução, pelo ribossoma, do ácido ribonucleico mensageiro (mRNA - *messenger RiboNucleic Acid*). Em Portugal surge na Universidade de Aveiro um grupo interdisciplinar envolvendo matemáticos, engenheiros informáticos, físicos e biólogos com vista a contribuir para responder a esta e outras questões relacionadas com o genoma. O trabalho de investigação está a ser desenvolvido no âmbito do projecto, *New bioinformatics tool for genome analysis unvails new rules governing speed and accuracy of mRNA decoding*, financiado pela Fundação para a Ciência e Tecnologia (FCT).
Nesta dissertação é dado um primeiro contributo para o projecto; são consideradas para o estudo duas espécies: *Candida albicans* e *Saccharomyces cerevisiae*. O estudo efectuado reduz-se à parte do genoma que codifica as proteínas, parte para a qual está descoberta a funcionalidade na estrutura dos seres vivos. No presente trabalho maior atenção será dada ao contexto dos codões onde a investigação é mais escassa.
Primero, foi necessário dominar os conceitos genéticos básicos. Por outro lado, perante a enorme quantidade de dados discretos disponíveis, foi necessário definir objectivos restritos, processar os dados de modo conveniente à concretização desses objectivos, encontrar metodologias estatísticas adequadas e fazer uma interpretação matemática dos dados.
Nesta introdução apresentar-se-á uma contextualização Biológica no sentido de introduzir os conceitos genéticos com os quais se vão trabalhar. Far-se-ão correspondências entre os termos genéticos e possíveis interpretações em terminologia matemática a usar. Definir-se-ão os objectivos gerais propostos. Por fim, apresentar-se-á resumidamente a organização desta dissertação.

1.2 Conceitos Biológicos

A linguagem genética é universal. Todo o ser vivo usa a mesma linguagem genética. Uma das extraordinárias revelações dos anos 50, do século passado, foi a demonstração de que a "infinita" complexidade das estruturas dos seres vivos é devida à simples combinação de 4 pequenas moléculas.
A informação genética está contida no DNA. O DNA encontra-se no interior de todas as células, organizado em estruturas a que se chamam cromossomas.
Na composição do DNA entram quatro bases nitrogenadas chamadas de *nucleótidos*, ou simplesmente, de *bases*. Estas são: adenina (A), guanina (G), timina (T) e citusina (C). Para além dos quatro nucleótidos, o DNA é composto, lateralmente pelo ortofosfato (ácido fosfórico) e pela desoxirribose (ver Figura 1.1).

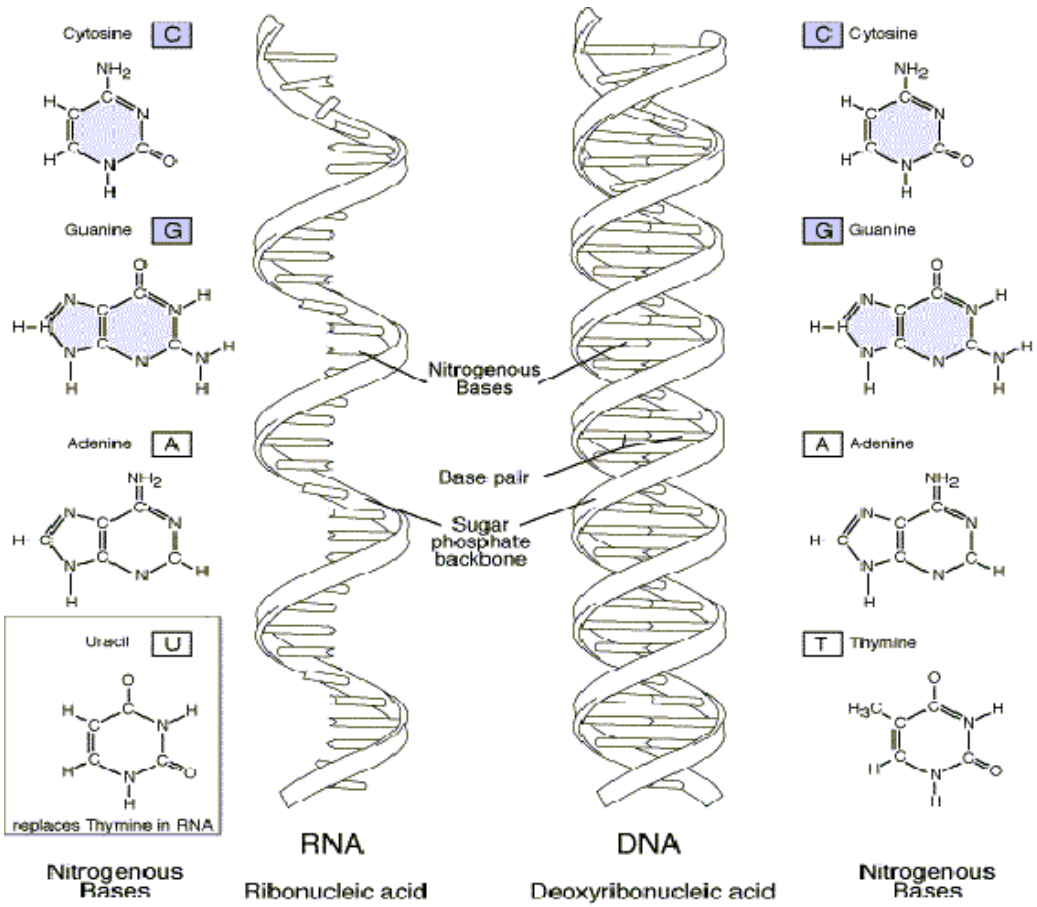


Figura 1.1: Estrutura do DNA e RNA. As unidades lineares de desoxirribose e ortofosfato correspondem ao *sugar phosphate backbone*.

Como é visível na Figura 1.1, a molécula de DNA tem uma estrutura semelhante à de uma escada torcida, formando uma espiral. Os nucleótidos formam os degraus, estando a adenina emparelhada com a timina e a guanina com a citusina constituindo uma dupla sequência de bases.
O constituinte do DNA a que usualmente se dá maior importância é a sequência dos quatro nucleótidos, pois nela está contida informação relativa a características hereditárias, sendo também necessária para a produção contínua de proteínas e consequente sobrevivência dos seres vivos.
Na sequência, os 4 nucleótidos combinam-se e encadeiam-se como as letras do alfabeto ao longo de um texto sem espaços. Assim, faz sentido a atribuição do nome de *linguagem genética ou texto genético* ao texto que constituiu o sequenciamento dos nucleótidos no DNA, (ver exemplo de um extrato dum texto genético no Apêndice C.1).
A sequência de nucleótidos que constitui o DNA é composta por duas partes distintas e complementares, a parte de subsequências codificadas e a de subsequências não codificadas. As subsequências codificadas consistem no conjunto das partes da sequência com significado em termos de produção de proteínas. Para as sequências não codificadas ainda não é conhecida a sua funcionalidade.
Nas subsequências codificadas, a que se irão chamar simplesmente de sequências de código ou código genético, o número total de nucleótidos é um múltiplo de três, em que cada terno de nucleótidos constitui o código de um *aminoácido*, isto é, da unidade proteica na construção de uma proteína. A cada grupo de três nucleótidos que codifica um aminoácido chama-se de *codão*[1]. Existem sessenta e quatro codões distintos ($4^3 = 64$), correspondendo aos sessenta e quatro arranjos possíveis das quatro bases em grupos de três.
Existe uma correspondência, que não é função, entre codões e aminoácidos. O número de aminoácidos usados pelos seres vivos é vinte e o de codões é sessenta e quatro. Os codões TAA, TAG e TGA habitualmente designados na literatura de codões terminais, não codificam aminoácidos, mas sim uma mensagem de terminação da construção da proteína. No entanto, a correspondência entre os sessenta e um codões não terminais e os aminoácidos é uma função não injectiva. Existem vários codões que codificam um mesmo aminoácido, os chamados codões sinónimos. A informação que a chave genética proporciona é inferior à que potencialmente poderia proporcionar (ver Figura 1.2). Observe-se que, no contexto da análise de código, a existência de redundância é uma defesa no sentido de eliminar alguns erros de tradução.
Existem essencialmente dois tipos de células: as eucariontes e as procariontes. A grande diferença entre estes dois tipos de células é o facto das células eucariontes terem núcleo definido e as procariontes não terem núcleo. As células das espécies a estudar, *Candida albicans* e *Saccharomyces cerevisiae*, são eucariontes.
O DNA encontra-se no interior de todas as células. No caso particular das células eucariontes o DNA está contido no interior do núcleo (Figura 1.3). No entanto, a síntese de proteínas dá-se no citoplasma, ocorrendo a transferência.
O mRNA tem a função de transportar a mensagem genética desde o DNA até ao ponto onde a mensagem é traduzida no citoplasma. A molécula de mRNA distingue-se do DNA por ser uma molécula de cadeias simples em que o nucleótido timina é substituído pelo nucleótido uracilo (U) e a desoxirribose pela ribose (ver Figura 1.1).

[1]Conclusão de George Camow em 1954.

Aminoácido		Codões	Aminoácido		Codões
Alanine	ALA	GCT; GCC; GCA; GCG.	Serine	SER	TCT; TCC; TCA; TCG
Valine	VAL	GTT; GTC; GTA; GTG.			AGT; AGC.
Leucine	LEU	CTA; CTG; CTT; CTC;	Threonine	THR	ACT; ACC; ACA; ACG.
		TTA; TTG.	Cysteine	CYS	TGT; TGC.
Isoleucine	ILE	ATT; ATC; ATA.	Glutamine	GLN	CAA; CAG.
Phenylalanin	PHE	TTT; TTC.	Asparagine	ASN	AAT; AAC.
Proline	PRO	CCT; CCC; CCA; CCG.	Histidine	HIS	CAT; CAC.
Methionine	MET	ATG.	Tyrosine	TYR	TAT; TAC.
Aspartic	ASP	GAT; GAC.	Tryptophan	TRP	TGG.
Glutamic	GLU	GAA; GAG.	Arginine	ARG	CGT; CGC; CGA; CGG
Lysine	LYS	AAA; AAG.			AGA; AGG.
Glycine	GLY	GGT; GGC; GGA; GGG.		STOP	TAA; TAG; TGA.

Figura 1.2: Tabela de correspondências entre codões e aminoácidos.

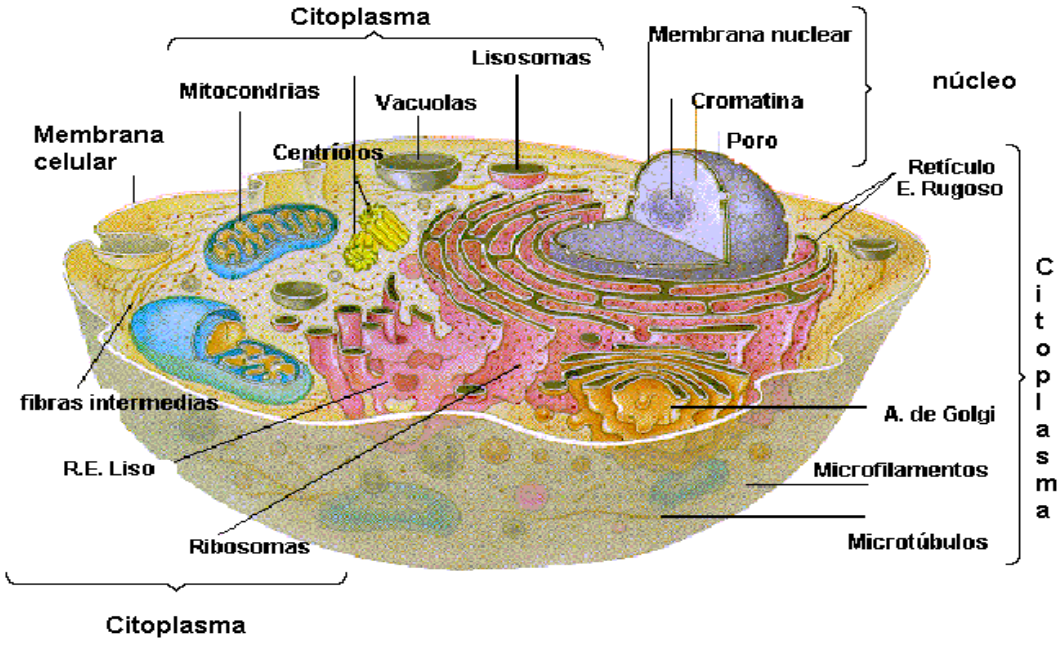


Figura 1.3: Célula eucarionte legendada.

Ainda na produção de proteínas existem os ribossomas. Os ribossomas são pequenos corpúsculos que se encontram distribuídos por todo o citoplasma com a função de descodificar o código genético (ver Figura 1.3). Estes corpúsculos deslocam-se um após o outro ao longo do mRNA cada um deles com uma proteína em fase de síntese, resultante da leitura da informação contida nas moléculas de mRNA (ver Figura 1.4). Cada mRNA pode dar origem a mais do que uma proteína.

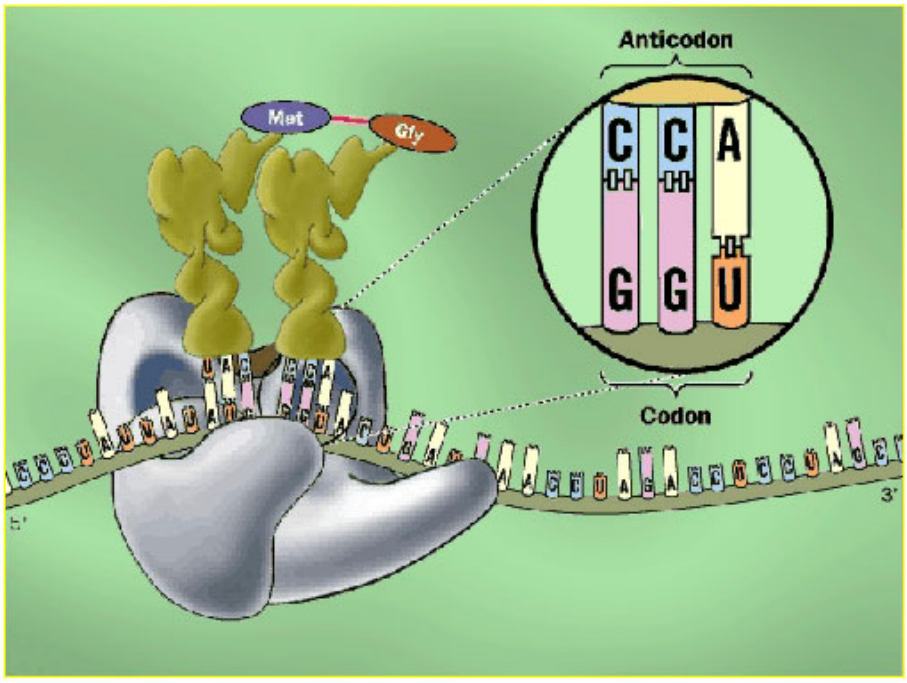


Figura 1.4: Construção da proteína a partir da leitura do sequenciamento dos codões pelo ribossoma.

Iniciada a leitura da molécula de mRNA, e dado um codão fixo da molécula, diz-se que o codão que o antecede está na posição 5' e o codão que o sucede na posição 3'. A leitura da molécula de mRNA pelo ribossoma é feita da posição 5' para a posição 3' (ver Figura 1.5).

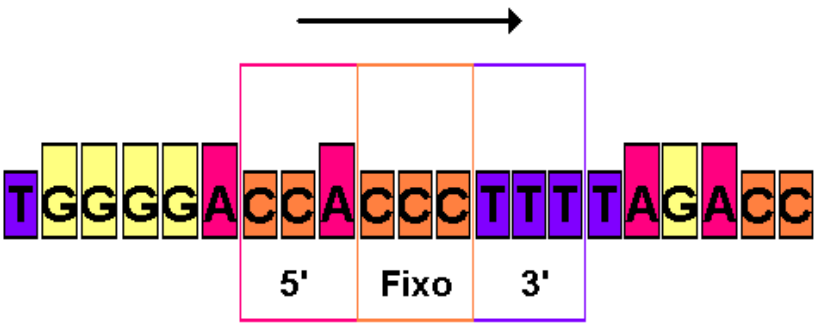


Figura 1.5: Esquema que explicita o sentido da leitura.

As sequências de código, são sequências de codões que começam sempre com o codão ATG (conhecido por codão de iniciação) e terminam com um dos codões terminais.
Cada sequência de código associada a uma dada proteína encontra-se contida num *gene*. Um gene contém parte codificada e não codificada da sequência de DNA. Ao conjunto de todos os genes de uma espécie é chamado de *genoma* dessa espécie.

1.3 Motivação e Objectivos Gerais

Esta dissertação de Mestrado surge integrada num projecto interdisciplinar de investigação com o objectivo geral e global de contribuir para a compreensão da estrutura da linguagem genética ou códigos genéticos.
Uma das questões mais ambiciosas do projecto é a decifração de leis gerais que governam a tradução do mRNA pelo ribossoma. De agora em diante não se fará referência ao mRNA, mas ao DNA, pois a informação disponibilizada consiste em sequências dos nucleótidos A, C, G e T. Na realidade, entre uma cadeia simples de DNA e uma cadeia de mRNA existe uma aplicação bijectiva entre nucleótidos, pelo que as leis que se obtenham para o sequenciamento de codões no DNA têm correspondência imediata para a sequência de codões no mRNA.
A parte de DNA a considerar no presente estudo serão as sequências de código do genoma das espécie *Candida albicans* e *Saccharomyces cerevisiae*, sendo o principal objectivo desta dissertação investigar leis no contexto dos codões. Assim, as sequências de código a considerar constituem sequências discretas de sessenta e quatro codões. Os codões de cada gene estão sequenciados pela ordem de leitura feita pelo ribossoma.
A enorme quantidade de informação contida nas sequências de código foi inicialmente organizada em tabelas de contingência de pares de codões justapostos para os dois tipos de leituras, a 3' e a 5', e em totais de frequências de cada codão.
As Figuras 1.5 e 1.6 em conjunto ilustram o tipo de leitura e a forma como os dados foram extraídos e organizados em tabelas de contingência. Outras contagens foram posteriormente consideradas na análise estatística, tendo em conta a necessidade emergente do estudo e a capacidade de resposta do *software* implementado pelo grupo de informática envolvido no projecto.

5' / Fixo	C_1	...	C_{61}
C_1	$n_{1\,1}$	...	$n_{1\,61}$
⋮	⋮	⋱	⋮
C_{64}	$n_{64\,1}$	...	$n_{64\,61}$

3' / Fixo	C_1	...	C_{64}
C_1	$n_{1\,1}$	...	$n_{1\,64}$
⋮	⋮	⋱	⋮
C_{61}	$n_{61\,1}$	...	$n_{61\,64}$

Figura 1.6: Esquema que explicita o modo como foram feitas as contagens.

No estudo a desenvolver utilizar-se-ão as tabelas de contingência relativas à leitura 3', por parecerem de leitura mais natural e são, por exemplo, de particular interesse na aplicação a

modelos markovianos.
No DNA, as sequências de código do genoma não se encontram sequenciadas. Neste estudo as contagens incluem os codões de todos os genes do genoma. No entanto, as contagens são feitas dentro de cada gene individualmente. Na aplicação de algumas análises será necessário aceitar os dados como provenientes de uma só sequência. Nesse caso, ter-se-á de acautelar as conclusões pela falsa hipótese de trabalho.
O genoma de uma dada espécie é, segundo os biólogos, representativo de todos os indivíduos pertencentes a essa espécie. Partindo deste pressuposto, as conclusões que se tirarem para o genoma de uma espécie são extensivas a todos os indivíduos dessa espécie.
São conhecidas duas regras básicas relativas à identificação de uma sequência de código. Concretamente, qualquer sequência de código tem início com o codão ATG e termina com um dos três codões terminais, TAA, TAG ou TGA. Naturalmente o objectivo fundamental do estudo a desenvolver é o de descobrir outras regras no sequenciamento de codões, para além das duas regras básicas referidas.
Pensa-se que as espécies *Candida albicans* e *Saccharomyces cerevisiae* têm o mesmo ancestral, em que a primeira terá degenerado e a segunda mantido as características genéticas do ancestral. Destas duas espécies conhece-se o sequenciamento completo do genoma; importará para o nosso estudo apenas o sequenciamento de código.
O objectivo geral é, de alguma forma, averiguar possíveis relações entre os codões para cada texto[2] ou de forma geral para o conjunto de textos genéticos. Para tal aplicou-se metodologias estatísticas de análise de dados discretos conhecidas sobre as sequências de codões.

1.4 Organização da Dissertação

Esta dissertação é constituída, para além desta introdução, por mais oito capítulo e três apêndices.
Nos seguintes sete capítulos serão abordados modelos probabilísticos e ferramentas estatísticas apropriadas à análise de dados de natureza qualitativa. No fim de cada um dos capítulos far-se-á uma aplicação desses instrumentos de análise ao problema de interesse no presente trabalho e sobre cada uma das espécies em estudo.
Começar-se-á no Capítulo 2 por analisar a associação existente entre pares justapostos de símbolos (codões e/ou nucleótidos) nas sequências de código no genoma das espécies *Candida albicans* e a *Saccharomyces cerevisiae*. Para tal, realizar-se-á o teste de independência habitual, numa Análise das Tabelas de Contingência e quantificar-se-á a preferência, se existir, da associação face à independência através de uma análise de resíduos.
No Capítulo 3 aplicar-se-á a Análise Classificatória, com vista a verificar, numa análise meramente exploratória, o estabelecimento de grupos de símbolos (codões) com comportamento semelhante ao nível do sequenciamento.
Existem muitos métodos distintos de agrupar as observações podendo obviamente os resultados serem distintos consoante o método utilizado. Na formação dos agrupamentos dos símbolos serão aplicados vários métodos tanto às matrizes dos resíduos ajustados, obtida no capítulo anterior na análise de resíduos, como às matrizes das frequências relativas dos símbolos para ambas as espécies.
Ainda no contexto da análise exploratória surge o Capítulo 4. Neste capítulo realizar-se-á uma Análise em Componentes Principais. O objectivo principal será o de reduzir o número

[2]Entenda-se por texto o conjunto de todas as sequências de código de um dado genoma.

símbolos, correlacionados entre si, identificando grupos com comportamento semelhante. Este método será aplicado tanto à matriz dos resíduos ajustados como dos valores de frequências relativas.
No Capítulo 5 averiguar-se-á a possibilidade de ajuste de comportamentos markovianos de ordem 1 à sequência dos codões. Nesse momento, assumir-se-á a estacionaridade do processo assim como algumas hipóteses de trabalho sobre os dados, as quais são biologicamente aceites.
No Capítulo 6 estudar-se-á, numa abordagem essencialmente exploratória, o comportamento das frequências relativas dos símbolos. Considerar-se-á uma Análise de Zipf sobre os símbolos que constituem a linguagem genética, de modo semelhante ao que é usual realizar no estudo de linguagens correntes, [3]. Esta metodologia permite, por um lado, estimar as leis inerentes às frequências relativas ordenadas e, por outro lado, averiguar a possibilidade de existência de correlações de longo alcance. No fim do capítulo apresentar-se-á uma breve análise exploratória onde se observa a semelhança de comportamento das frequências dos símbolos das duas espécies em estudo.
No sentido de descobrir o "tamanho" das correlações existentes entre os símbolos nas sequências de código, assumindo que estes seguem comportamentos markovianos de ordem k e de confrontar alguns resultados obtidos nos Capítulos 5 e 6, surge o Capítulo 7. No Capítulo 7 apresentar-se-á inicialmente uma pequena abordagem teórica sobre o Critério de Informação Bayesiana (BIC) em Cadeias de Markov de Ordem k. Seguidamente aplicar-se-á o BIC a sequências de código, de genes aleatoriamente seleccionados de ambas as espécies, para estimar a ordem que melhor se ajusta.
No Capítulo 8 abordar-se-ão alguns conceitos da Teoria da Informação aplicáveis ao texto genético e apresentar-se-ão os resultados obtidos da sua aplicação às sequências de código das espécies *Candida albicans* e *Saccharomyces cerevisiae*.
O Capítulo 9 conclui esta dissertação, resumindo os resultados obtidos nas várias abordagens utilizadas. Também refere algumas técnicas que se pensou inicialmente usar mas que, por algum motivo, se vieram a revelar inadequadas ou não foi possível a sua concretização. Finalmente, deixa em aberto algumas ideias e direcções de trabalho para investigação futura.

Capítulo 2

Análise de Tabelas de Contingência

2.1 Introdução

A Análise de Tabelas de Contingência é uma metodologia estatística aplicada a dados de natureza qualitativa[1].
Os indivíduos de uma dada população podem ser classificados em categorias (ou classes) de acordo com diversos critérios. Fixos os critérios, a classificação dos indivíduos consiste em detectar a(s) categoria(s) na qual cada indivíduo se identifica. As categorias a considerar são mutuamente exclusivas e a classificação é exaustiva, isto é, qualquer indivíduo pertence a uma e uma só categoria. Deste modo, os dados consistem nas frequências observadas em cada uma das categorias.
No caso concreto das tabelas a estudar, consideram-se dois critérios em linha o codão fixo e em coluna o codão justaposto obtido por uma leitura 3', de acordo com o esquema da Figura 1.5. Uma vez que aos codões terminais não lhe sucede nenhum codão, a tabela será 61×64. As frequências são, naturalmente, o número de vezes que cada par de codões surge no conjunto total das sequências de código de cada espécie.
As Tabelas de Contingência foram a forma escolhida para organizar e reduzir os dados relativos às sequências de símbolos (codões ou aminoácidos), já que se pretendia numa primeira fase, uma análise global do genoma das espécies no contexto dos pares de símbolos justapostos. A informação contida na tabela permitirá realizar uma análise da associação entre pares de símbolos.
Quando se passa dos símbolos sequenciados para a respectiva Tabela de Contingência perde-se alguma informação. No entanto, é bastante difícil trabalhar com um número tão elevado de símbolos sequenciados que constitui o genoma, da ordem dos 4 000 000 contra $61 \times 64 = 3904$ células da tabela.
Neste capítulo apresentar-se-á uma abordagem teórica às Tabelas de Contingência de margens livres e à análise de resíduos no contexto das Tabelas de Contingência. Seguir-se-á uma aplicação destas metodologias às sequências de código do genoma das duas espécies de interesse no estudo: *Saccharomyces cerevisiae* e *Candida albicans*. E dar-se-á relevo aos testes de ajustamento tendo em conta o objectivo que motiva o projecto.

[1] Pode também ser aplicada a dados de natureza quantitativa desde que discretizados.

2.2 Nomenclatura

Sejam A e B duas características da população subdivididas em r e c categorias, designadas por A_1, ..., A_r e B_1, ..., B_c, respectivamente. Seja $n \geq 1$ o número total de observações ou indivíduos, de uma amostra casual, a classificar.
A tabela que resulta da classificação das n observações nas $r \times c$ categorias cruzadas, diz-se *Tabela de Contingência* $r \times c$ e tem a forma da Tabela 2.1.

	B_1	...	B_j	...	B_c	Total marginal
A_1	n_{11}	...	n_{1i}	...	n_{1c}	$n_{1\cdot}$
...	...	...	...	...	...	...
A_i	n_{i1}	...	n_{ij}	...	n_{ic}	$n_{i\cdot}$
...	...	...	...	...	...	...
A_r	n_{r1}	...	n_{rj}	...	n_{rc}	$n_{r\cdot}$
Total marginal	$n_{\cdot 1}$	...	$n_{\cdot j}$	...	$n_{\cdot c}$	n

Tabela 2.1: Esquema de uma Tabela de Contingência $r \times c$.

Relativamente à Tabela 2.1 tem-se: n_{ij} o número de observações classificadas simultaneamente na categoria A_i de A e B_j de B, $n_{i\cdot}$ o total marginal de observações na categoria A_i e $n_{\cdot j}$ o total marginal de observações na categoria B_j, com $i \in \{1, ..., r\}$ e $j \in \{1, ..., c\}$.
Obviamente que:

$$n_{i\cdot} = \sum_{j=1}^{c} n_{ij}$$

$$n_{\cdot j} = \sum_{i=1}^{r} n_{ij}$$

$$n = \sum_{j=1}^{c}\sum_{i=1}^{r} n_{ij} = \sum_{j=1}^{c} n_{\cdot j} = \sum_{i=1}^{r} n_{i\cdot}.$$

Para $i \in \{1, ..., r\}$ e $j \in \{1, ..., c\}$ sejam:
p_{ij} - probabilidade de uma observação pertencer simultaneamente à i-ésima categoria da variável A e à j-ésima categoria da variável B, isto é, de pertencer à célula (i, j);
$p_{i\cdot}$ - probabilidade marginal de uma observação pertencer à i-ésima categoria da variável A;
$p_{\cdot j}$ - probabilidade marginal de uma observação pertencer à j-ésima categoria da variável B.
Naturalmente ter-se-á que:

$$p_{i\cdot} = \sum_{j=1}^{c} p_{ij}, \qquad p_{\cdot j} = \sum_{i=1}^{r} p_{ij} \qquad e \qquad \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} = 1. \tag{2.1}$$

2.3 Independência e Associação

2.3.1 Testes de Ajustamento

Considerando apenas uma das características, por exemplo, a variável A subdividida em r categorias A_1, ..., A_r e uma distribuição de probabilidade particular conhecida (p_{0i}, com $i \in \{1, 2, ..., r\}$).
Poder-se-á testar uma hipótese do tipo

$$H_0 : p_i = p_{0i}, i \in \{1, ..., r\} \tag{2.2}$$

onde, para simplificação de escrita, p_i denota $p_{i.}$.
Um teste de hipóteses deste tipo é chamado *teste de ajustamento.*
A estatística de teste utilizada nos testes de ajustamento é a estatística obtida por Karl Pearson dada genericamente por:

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - \hat{e}_i)^2}{\hat{e}_i} \tag{2.3}$$

onde $\hat{e}_i$ é a frequência esperada de uma observação estar na classe i sob a validade da hipótese H_0. Nas circuntâncias consideradas, $\hat{e}_i = np_{0i}$.
Para um total de n observações designe-se por N_i a variável aleatória que representa o número de observações na categoria A_i, com $i \in \{1, ..., r\}$. Consequentemente (N_1, N_2, ..., N_r) é um vector aleatório com distribuição multinomial tal que $\sum_{i=1}^{r} N_i = n$ e com função de probabilidade dada por

$$P(N_1 = n_1, N_2 = n_2, ..., N_r = n_r) = \frac{n!}{n_1! n_2! ... n_r!} p_1^{n_1} p_2^{n_2} ... p_r^{n_r} \tag{2.4}$$

com $n_r = n - n_1 - n_2 - ... - n_{r-1}$ e $p_r = 1 - p_1 - p_2 - ... - p_{r-1}$.

Teorema 2.3.1 *Seja (N_1, N_2, ..., N_r) um vector aleatório com distribuição multinomial de parâmetros $n, p_{01}, p_{02}, ..., p_{0r}$. Então a variável aleatória*

$$\chi_a^2 = \sum_{i=1}^{r} \frac{(N_i - np_{0i})^2}{np_{0i}} \tag{2.5}$$

tem assimptoticamente distribuição de um qui-quadrado, com $(r - 1)$ graus de liberdade.

A prova deste teorema pode ser encontrada em Cramér(1946), como é referido em [14].
No Teorema 2.3.1 pressupõe-se $n \to \infty$. Para a aproximação no caso finito ser válida assume-se que as frequências esperadas não sejam "muito pequenas". Na prática a frequência esperada, np_{0i}, da variável aleatória N_i, $i \in \{1, ..., r\}$, nunca deve ser inferior a cinco, sendo preferível tomar dez como limite inferior (ver [14]).
Uma extensão ao teste anterior resulta quando a distribuição de probabilidades em H_0 não se encontra completamente especificada, dependendo de k ($k < r$) parâmetros independentes e desconhecidos. No sentido de ultrapassar esse problema, do desconhecimento da distribuição

de probabilidades, calculam-se as estimativas de máxima verosimilhança dos parâmetros desconhecidos e tomam-se os valores de p_{0i} calculados em função das estimativas. Nesse caso a variável aleatória χ^2_a terá uma distribuição assimptótica de um qui-quadrado com $(r-k-1)$ graus de liberdade.

2.3.2 Independência

Considere-se agora os 2 critérios A e B cruzados numa Tabela de Contingência $r \times s$, com a estrutura da Tabela 2.1.
No desenvolvimento que se segue ter-se-á presente o modelo probabilístico de margens livres[2].
O modelo probabilístico de margens livres consiste no caso mais geral dentro dos modelos probabilísticos para Tabelas de Contingência. Nesse modelo o único valor fixo é o número total de observações, n. E, além disso o vector aleatório $(N_{11}, ..., N_{ij}, ..., N_{rc})$ tem distribuição multinomial, onde N_{ij} é a variável aleatória que representa o número de observações que pertencem à célula (i, j).
Um dos grandes objectivos no estudo de Tabelas de Contingência $r \times s$ é o de averiguar a existência de independência entre as duas variáveis face à possibilidade de existência de associação. A lei multiplicativa das probabilidades perante a independência das variáveis aleatórias A e B pode ser traduzida por:

$$p_{ij} = p_{i\cdot}p_{\cdot j} \quad com \quad i \in \{1, ..., r\} \quad e \quad j \in \{1, ..., c\}. \tag{2.6}$$

Para testar a hipótese de independência H_0,

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j} \quad com \quad i \in \{1, ..., r\} \quad e \quad j \in \{1, ..., c\} \tag{2.7}$$

utiliza-se a estatística de Karl Pearson que, no contexto das Tabelas de Contingência $r \times c$ é dada por:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \tag{2.8}$$

onde $\hat{e}_{ij}$ é a frequência esperada de uma observação estar na célula (i, j), sob a validade de H_0; portanto, $\hat{e}_{ij} = n\hat{p}_{i\cdot}\hat{p}_{\cdot j}$ com $i \in \{1, ..., r\}$ e $j \in \{1, ..., c\}$. As probabilidades marginais $p_{i\cdot}$ e $p_{\cdot j}$ com $i \in \{1, ..., r\}$ e $j \in \{1, ..., c\}$ são desconhecidas sendo estas substituídas pelas estimativas de máxima verosimilhança, $\hat{p}_{i\cdot}$ e $\hat{p}_{\cdot j}$ respectivamente.

Teorema 2.3.2 *Considere-se uma amostra casual de n observações do par (A, B). Seja $(N_{11}, ..., N_{ij}, ..., N_{rc})$ o vector aleatório constituído pelo número de observações em cada célula, distribuídas de acordo com uma multinomial de parâmetros $\{n, p_{ij}\ i \in \{1, ..., r\}, j \in \{1, ..., c\}\}$. Assumindo a independência entre as variáveis aleatórias A e B, as estimativas de máxima verosimilhança das probabilidades marginais são dadas por:*

$$\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n} \quad e \quad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n},$$

[2]Para além do modelo de margens livres existem outros modelos: os de margens fixas para os quais se fixam uma ou ambas as margens (totais marginais).

com $i \in \{1, ..., r\}$ *e* $j \in \{1, ..., c\}$.

Prova 2.3.1 *A distribuição de probabilidades do vector aleatório* $(N_{11}, ..., N_{ij}, ..., N_{rc})$ *é dada de modo análogo à equação 2.4. Assim, dada uma realização do vector aleatório tem-se a função de verosimilhança dada por:*

$$L(p_{11}, p_{12}, ..., p_{rs}) = \frac{n!}{\prod_{i=1}^{r}\prod_{j=1}^{c} n_{ij}!} \prod_{i=1}^{r}\prod_{j=1}^{c} p_{ij}^{n_{ij}}.$$

Logaritmizando a função de verosimilhança obtém-se:

$$ln(L) = ln(\frac{n!}{\prod_{i=1}^{r}\prod_{j=1}^{c} n_{ij}!}) + ln(\prod_{i=1}^{r}\prod_{j=1}^{c} p_{ij}^{n_{ij}}).$$

A primeira parcela do segundo membro é constante, pois não depende dos p_{ij}*'s. Dado que se pretende estudar os maximizantes da função de verosimilhança estudar-se-á apenas a segunda parcela.*
Uma vez que se assume a independência e por (2.1) vem que:

$$ln(\prod_{i=1}^{r}\prod_{j=1}^{c} p_{ij}^{n_{ij}}) = ln(\prod_{i=1}^{r} p_{i\cdot}^{n_{i\cdot}} \prod_{j=1}^{c} p_{\cdot j}^{n_{\cdot j}})$$

$$= ln((1 - \sum_{i=1}^{r-1} p_{i\cdot})^{n_{r\cdot}} (1 - \sum_{j=1}^{c-1} p_{\cdot j})^{n_{\cdot c}} \prod_{i=1}^{r-1} p_{i\cdot}^{n_{i\cdot}} \prod_{j=1}^{c-1} p_{\cdot j}^{n_{\cdot j}})$$

$$= \sum_{i=1}^{r-1} [\frac{n_{r\cdot}}{r-1} ln(1 - \sum_{i=1}^{r-1} p_{i\cdot}) + n_{i\cdot} ln(p_{i\cdot})] + \sum_{j=1}^{c-1} [\frac{n_{\cdot c}}{c-1} ln(1 - \sum_{j=1}^{c-1} p_{\cdot j}) + n_{\cdot j} ln(p_{\cdot j})].$$

Recorrer-se-á ao cálculo dos zeros das derivadas para detectar a existência de máximos.

$$\frac{\partial ln(L)}{\partial p_{i\cdot}} = -\frac{n_{r\cdot}}{r-1} \frac{r-1}{1 - \sum_{i=1}^{r-1} p_{i\cdot}} + \frac{n_{i\cdot}}{p_{i\cdot}}$$

$$\frac{\partial ln(L)}{\partial p_{\cdot j}} = -\frac{n_{\cdot c}}{c-1} \frac{c-1}{1 - \sum_{j=1}^{c-1} p_{\cdot j}} + \frac{n_{\cdot j}}{p_{\cdot j}}$$

donde, de $\frac{\partial ln(L)}{\partial p_{i\cdot}} = 0$*, resulta:*

$$p_{i\cdot} = \frac{n_{i\cdot}}{\frac{n_{r\cdot}}{p_{r\cdot}}}. \tag{2.9}$$

Note-se que $n = \sum_{i=1}^{r} n_{i\cdot}$ *e portanto:*

$$1 = \sum_{i=1}^{r} p_{i\cdot} = \sum_{i=1}^{r-1} \frac{n_{i\cdot}}{\frac{n_{r\cdot}}{p_{r\cdot}}} + p_{r\cdot} = \frac{n - n_{\cdot r}}{\frac{n_{r\cdot}}{p_{r\cdot}}} + p_{r\cdot} = p_{r\cdot}\frac{n}{n_{r\cdot}}.$$

Logo, $p_{r\cdot} = \frac{n_{r\cdot}}{n}$. *Substituindo em (2.9) obtém-se:*

$$p_{i\cdot} = \frac{n_{i\cdot}}{n}.$$

Partindo de $\frac{\partial ln(L)}{\partial p_{\cdot j}} = 0$ *ter-se-ía analogamente que:*

$$p_{\cdot j} = \frac{n_{\cdot}j}{n}.$$

Prova-se que a segunda derivada é negativa e assim se conclui a demonstração.

$\diamondsuit$

Verificadas as condições do Teorema 2.3.2, obtêm-se a estimativa da frequência esperada e_{ij} dada por:

$$\hat{e}_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{n}. \tag{2.10}$$

Se as variáveis não são independentes espera-se que a diferença entre os valores estimados, assumindo a independência, e os valores observados seja grande, caso contrário será pequena. É neste contexto que se aplica o teste de ajustamento do qui-quadrado para Tabelas de Contingência $r \times c$.
A regra de decisão do teste é consequência da usual comparação entre o valor obtido pela estatística de teste (2.8) e o valor teórico do quantil de ordem $(1-\alpha)\cdot 100\%$ da distribuição de probabilidade do qui-quadrado para um nível de significância α do teste. Utilizar-se-á um dos níveis de significância mais usuais, $\alpha = 0.05$. Se o valor observado da estatística de Pearson, para uma dada tabela em estudo, for superior ao valor teórico do quantil, a hipótese de independência será rejeitada e caso contrário não será rejeitada.
Para determinar o valor teórico do quantil ter-se-á de averiguar o número de graus de liberdade (*df - degrees of freedom*) da estatística χ^2. Numa usual Tabela de Contingência $r \times c$, em que se tem n observações independentes, o número de graus de liberdade é dado por:

$$\textit{número de classes} - \textit{número de parâmetros independentes} - 1,$$

à semelhança da estatística do teste de ajustamento com parâmetros desconhecidos e portanto igual a $r \times c - (c-1+r-1) - 1$. O número de parâmetros independentes é $(c-1)+(r-1)$, uma vez que o número total de parâmetros é $r+c$: $p_{1\cdot}$, $p_{2\cdot}$, ..., $p_{r\cdot}$, $p_{\cdot 1}$, $p_{\cdot 2}$, ..., $p_{\cdot c}$, e a soma dos parâmetros em coluna e em linha são respectivamente iguais a 1.

Também se pode definir o número de graus de liberdade como a diferença entre o número de células independentes e o número de parâmetros independentes. O número de células independentes duma Tabela de Contingência $r \times c$ de n observações independentes é $rc - 1$, já que o número de células é $r \times c$ e a soma dos totais das células da tabela é n. Assim, o número de graus de liberdade da estatística de Pearson na Tabela de Contingência $r \times c$ é $rc - 1 - (c-1) - (r-1) = (r-1)(c-1)$.

2.3.3 Associação

As grandes dificuldades do uso da estatística de Pearson é, por um lado, o facto de os graus de liberdade, $(r-1)(c-1)$, dependerem da dimensão da Tabela de Contingência $r \times c$, e, por outro lado, do número total de observações. Este último problema é ilustrado com o seguinte exemplo:

Exemplo - *Considerem-se as seguintes Tabelas de Contingência* 2×2*, a última obtida da primeira por multiplicação das frequências observadas por 2:*

	B_1	B_2	Total marginal
A_1	1	4	5
A_2	7	2	9
Total marginal	8	6	$n = 14$

	B_1	B_2	Total marginal
A_1	2	8	10
A_2	14	4	18
Total marginal	16	12	$n = 28$

Observa-se que as probabilidades de ocorrência de cada categoria são as mesmas, no entanto as estatísticas de Pearson referentes a cada uma das tabelas diferem de um factor multiplicativo 2.
De um modo geral, multiplicando todas as observações por um factor k *a estatística de Pearson vem alterada do factor multiplicativo* k.

$\diamondsuit$

No sentido de contornar estes problemas surgiram as medidas de associação, que não refletem a dimensão da tabela. Os exemplos mais comuns de medidas de associação, baseada na estatística de Pearson, são:

- Coeficiente de contingência de Pearson

$$P = \sqrt{\frac{\chi^2}{\chi^2 + n}}; \tag{2.11}$$

- Coeficiente de Tschuprow

$$T = \sqrt{\frac{\chi^2}{n\sqrt{(r-1)(c-1)}}}; \tag{2.12}$$

- Coeficiente de Cramér

$$C = \sqrt{\frac{\chi^2/n}{min(r-1, c-1)}}. \tag{2.13}$$

O coeficiente de Cramér é a medida de associação adoptada nas aplicações efectuadas neste trabalho, porque para além de ser independente do número de observações, não depende das dimensões da tabela e assume valores que variam entre 0 e 1. Estas propriedades não são verificadas simultaneamente pelos coeficientes de Tschuprow e de Pearson.
O coeficiente de Cramér assume o valor zero em caso de independência completa e o valor um em caso de associação completa.

2.4 Análise de Resíduos

Se o teste realizado, no contexto das Tabelas de Contingência $r \times c$ concluir a rejeição da independência das duas variáveis, terá interesse analisar as classes que provocaram tal rejeição. Assim, no sentido de identificar as categorias responsáveis pelo valor elevado da estatística de Pearson, pode-se realizar uma análise dos *resíduos estandardizados* também conhecidos por resíduos de Pearson, e dados por:

$$r_{ij} = \frac{n_{ij} - \hat{e}_{ij}}{\sqrt{\hat{e}_{ij}}} \tag{2.14}$$

Observe-se que,

$$\sum_{i=1}^{r}\sum_{j=1}^{c} r_{ij}^2 = \chi^2$$

Em particular, utilizam-se os resíduos ajustados dados por:

$$d_{ij} = \frac{r_{ij}}{\sqrt{v_{ij}}} \tag{2.15}$$

onde v_{ij} é uma estimativa da variância de r_{ij} dada por:

$$v_{ij} = (1 - \frac{n_{i\cdot}}{n})(1 - \frac{n_{\cdot j}}{n}).$$

A importância dos resíduos ajustados é realçada pelo Teorema 2.4.1, resultado obtido por Haberman(1973) e que pode ser encontrado, por exemplo, em [7] e [23].

Teorema 2.4.1 *Considere-se uma amostra casual de n observações do par* (A, B). *Seja* $(N_{11}, ..., N_{ij}, ..., N_{rc})$ *o vector aleatório constituído pelo número de observações de cada célula, distribuído de acordo com uma multinomial de parâmetros* $\{n,\ p_{ij}\ i \in \{1, ..., r\},\ j \in \{1, ..., c\}\}$. *Assumindo que se verifica a independência entre as variáveis aleatórias A e B então:*

$$d_{ij} \text{ tem assimptoticamente } (n \to \infty) \text{ distribuição de uma normal } N(0, 1).$$

Observe-se que se os resíduos ajustados, d_{ij}, tiverem um comportamento, a nível de distribuição, muito diferente duma distribuição normal $N(0, 1)$ será de rejeitar a independência na Tabela de Contingência $r \times c$. Uma vez que d_{ij} tem assimptoticamente distribuição de uma normal $N(0, 1)$ poder-se-á dizer que para um nível de confiança de 99.73% que as categorias mais responsáveis pela rejeição da hipótese da independência são as correspondentes às células (i, j) tais que $|d_{ij}| \geq 3$ uma vez que $P(-3 < d_{ij} < 3) = 0.9973$. Assim, na prática considera-se que um valor de resíduo ajustado é responsável pela rejeição se em módulo for não inferior a 3.

2.5 Aplicação ao Caso em Estudo

2.5.1 Determinação dos Graus de Liberdade

Genericamente uma sequência de s símbolos é uma sequência de $s-1$ pares de símbolos, uma sequência de $s-2$ ternos de símbolos e assim sucessivamente. Observe-se que os sucessivos pares de símbolos de uma sequência não são independentes, já que o segundo elemento de um dado par é o primeiro elemento do par seguinte. Devido a esta falha de independência a Tabela de Contingência resultante da contagem de pares de símbolos numa sequência, goza da seguinte propriedade (ver [1]):

Propriedade dos totais marginais numa sequência de símbolos - *Seja b o número de categorias que constitui cada uma das duas variáveis numa Tabelas de Contingência* $b \times b$. *A variável coluna refere-se ao primeiro elemento do par e a variável linha ao segundo. Ambas as categorias são representadas por* A_i, $i = 1, 2, ..., b$ *e os totais marginais da categoria* A_i *da variável coluna e linha são* $n_{i\cdot}$ *e* $n_{\cdot i}$, *respectivamente. Tem-se que:*

- *para uma sequência de símbolos em que o símbolo de iniciação é igual ao de terminação os totais marginais são iguais;*
- *para uma sequência em que o símbolo de iniciação é diferente do símbolo de terminação tem-se:*
  - *se o símbolo da categoria i for igual ao símbolo de iniciação,* $n_{i\cdot} - n_{\cdot i} = 1$;
  - *se o símbolo da categoria i for igual ao símbolo de terminação,* $n_{i\cdot} - n_{\cdot i} = -1$;
  - *em qualquer outro caso a diferença* $n_{i\cdot} - n_{\cdot i}$ *é zero.*

$\diamondsuit$

A propriedade anterior sugere que o número de parâmetros independentes seja $(b-1)+1=b$ (ver tabela da Figura 2.1, em que, por exemplo, o número de células a amarelo referem-se ao número de parâmetros independentes), menor que no modelo de margens livres. Este facto pode fazer parecer a existência de menor número de graus de liberdade face ao modelo de margens livres apresentado. Na realidade, o número de graus de liberdade mantém-se.
Observe-se que conhecendo as $b-1$ primeiras colunas e a entrada (b,b) podem deduzir-se os valores das restantes $b-1$ células (na tabela da Figura 2.1 as $b-1$ células correspondem às células a branco). Assim, o número de células independentes é $b^2-(b-1)$ (ver tabela da Figura 2.1 as células a azul claro). E o número de graus de liberdade é a diferença entre o número de células independentes e o número de parâmetros independentes: $b^2-(b-1)-b=(b-1)^2$.
O número de graus de liberdade deste modelo é o mesmo que o do modelo de margens livres[3].

	A_1	...	A_i	...	A_b	Total Marginal
A_1	n_{11}		n_{1i}		n_{1b}	$n_{1.}$
⋮		⋱		⋱		⋮
A_i	n_{i1}		n_{ii}		n_{ib}	$n_{i.}$
⋮		⋱		⋱		⋮
A_b	n_{b1}		n_{bi}		n_{bb}	$n_{b.}$
Total Marginal	$n_{.1}$	...	$n_{.i}$	...	$n_{.b}$	n

Figura 2.1: Tabela de Contingência de pares de símbolos de uma sequência.

Embora o número de graus de liberdade coincidam, a distribuição assimptótica pode ser afectada por amostras não multinomiais. Contudo, na aplicação prática aceitar-se-ão que as amostras são multinomiais.
Os dados em estudo consistem em sequências de símbolos, codões ou aminoácidos, lidas no sentido, de 5' para 3'.

2.5.2 Independência e Associação de Pares de Codões Justapostos

No Apêndice B.1, encontram-se as Tabelas de Contingência 61×64 referentes às espécies *Saccharomyces cerevisiae* e *Candida albicans*. Nessas tabelas, em cada categoria, tem-se o número de observações de cada par de codões possíveis no conjunto das sequências de código - dados qualitativos. Observe-se que não surgem codões após o codão terminal, e portanto o número total de categorias possíveis é de 64×61.
No sentido de averiguar quanto à existência de independência de um codão fixo face ao codão seguinte, aplicar-se-á o teste de ajustamento do qui-quadrado àquelas Tabelas de Contingência. Como as tabelas são constituídas por 64×61 categorias, então o número de graus de liberdade das Tabelas de Contingência em estudo é 3780 $((64-1)\times(61-1))$.

[3]Por curiosidade, refere-se aqui que para os modelos de margens fixas que o número de graus de liberdade é também $(c-1)(r-1)$.

Calcularam-se as estatísticas de Pearson e obtiveram-se valores muito elevados: 137053.773 e 382664.868, para a *Saccharomyces cerevisiae* e *Candida albicans*, respectivamente. Da aplicação do teste de ajustamento resulta claramente a rejeição da hipótese nula (a hipótese de independência), já que o valor teórico do quantil de ordem 0.95 de uma distribuição do qui-quadrado com 3780 graus de liberdade é de 4052.700, valor muito inferior aos valores obtidos para a estatística de Pearson.
Calcularam-se os coeficientes de Cramér e obtiveram-se os seguintes valores numéricos: 0.0019 e 0.0008 para a *Candida albicans* e *Saccharomyces cerevisiae*, respectivamente. Os valores dos coeficientes são valores muito próximos de zero, indicando portanto uma associação muito fraca.
Os codões justapostos têm associação fraca entre si, no entanto a hipótese de serem independentes é rejeitada - esta é uma problemática usual quando se trabalha com "um grande número" de observações.
De seguida, averiguar-se-á quanto à independência e associação entre um codão e os nucleótidos do codão que lhe segue. Ao nucleótido justaposto designar-se-á por Nuc(1) e ao seguinte por Nuc(2) e o último por Nuc(3). Assim, cada uma das 3 tabelas a analisar têm dimensão 61×4 (ver Apêndice B.2).
No sentido de concretizar o objectivo em causa aplicou-se novamente um teste de ajustamento do qui-quadrado e calculou-se o coeficiente de Cramér, para ambas as espécies. Os resultados encontram-se resumidos na Tabelas 2.2 e 2.3.

	Nuc(1)	Nuc(2)	Nuc(3)	df	Codão seguinte	df
Saccharomyces cerevisiae	58379.8	22971.9	10003.9	180	137053.8	3780
Candida albicans	159544.6	64534.7	40745.1	180	382664.9	3780
Valor teórico do quantil	242.880	242.880	242.880		4052.7	

Tabela 2.2: Resultados da estatística χ^2 e valores teóricos do quantil de ordem 0.95 de uma distribuição do qui-quadrado com *df* graus de liberdade para a *Candida albicans* e *Saccharomyces cerevisiae*.

	Nuc(1)	Nuc(2)	Nuc(3)	Codão seguinte	n
Saccharomyces cerevisiae	0.00657	0.00002026	0.002585	0.0008	2961829
Candida albicans	0.015695	0.0000177	0.006348	0.0019	3388506

Tabela 2.3: Resultados do coeficiente de associação de Cramér para a *Candida albicans* e *Saccharomyces cerevisiae*.

Em qualquer uma das situações estudadas os resultados concluiram a rejeição da independência, observando-se contudo valores pequenos de associação.
Relativamente aos nucleótidos do codão justaposto ao codão fixo observam-se valores relativamente distintos para a estatística de teste e para o coeficiente de associação. Todavia o segundo nucleótido do codão justaposto é o que apresenta menor coeficiente de associação

para ambas as espécies em estudo, indiciando provavelmente uma menor dependência entre um codão e o nucleótido central do codão seguinte.

2.5.3 Análise de Resíduos de Pares de Codões Justapostos

Os resultados a apresentar nesta subsecção foram obtidos através do *software SPSS 7.5 for Windows*.
Calculados os resíduos ajustados observam-se muitas células com valor, em módulo, superior a 3. Fez-se a representação gráfica dos valores observados *versus* os quantis de uma distribuição normal $N(0,1)$, através do gráfico *QQ-plot* (ver Figuras 2.2 e 2.3).

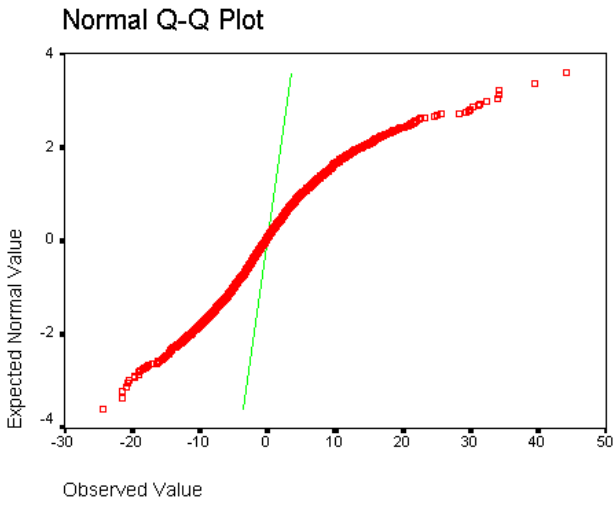


Figura 2.2: *QQ-plot* da distribuição normal $N(0,1)$ face aos valores observados para os resíduos da *Saccharomyces cerevisiae*, numa leitura 3'.

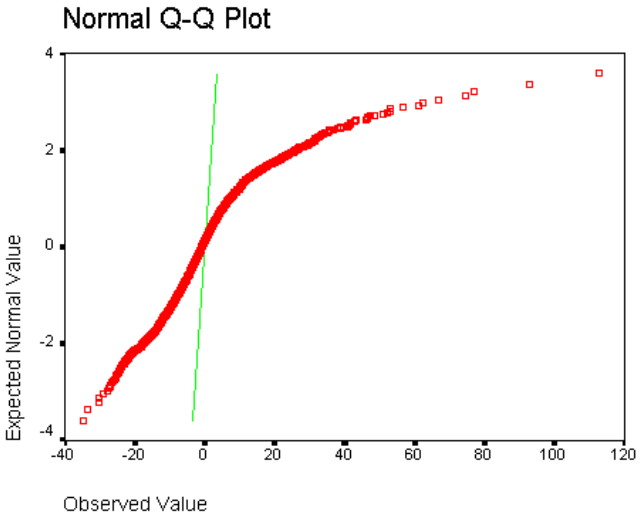


Figura 2.3: *QQ-plot* da distribuição normal $N(0,1)$ face aos valores observados para os resíduos da *Candida albicans*, numa leitura 3'.

Nas Figuras 2.2 e 2.3 observa-se um grande afastamento dos resíduos ajustados relativamente aos pontos da recta representada a verde, a bissectriz dos quadrantes ímpares. Consequentemente em nenhuma das espécies se prevê um ajustamento adequado destes dados à distribuição normal $N(0,1)$. A rejeição de ajuste vem confirmar a rejeição da hipótese nula, a hipótese da independência anteriormente testada.
No entanto, as representações gráficas parecem fazer crer que os resíduos estão distribuídos segundo uma distribuição normal com média aproximadamente 0 e desvios padrão aproximadamente 6 e 10 para as espécies *Saccharomyces cerevisiae* e *Candida albicans*, respectivamente (ver Figuras 2.4 e 2.5).

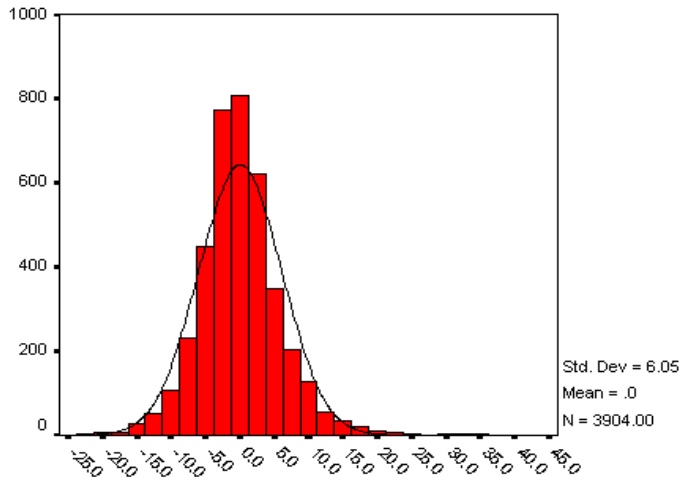


Figura 2.4: Histograma dos resíduos ajustados da *Saccharomyces cerevisiae*, numa leitura 3'.

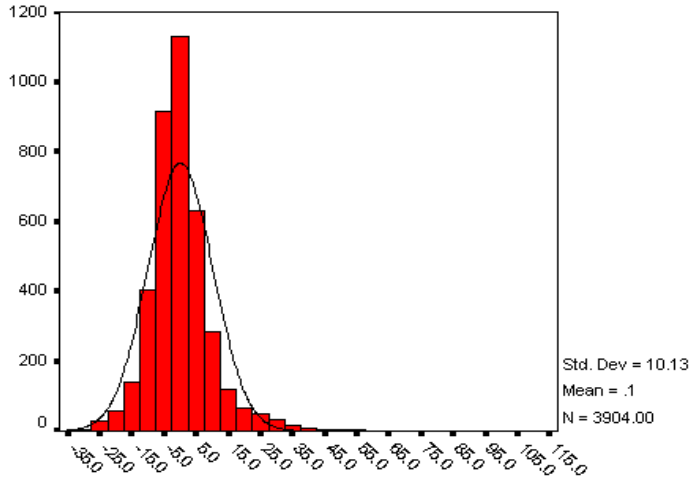


Figura 2.5: Histograma dos resíduos ajustados da *Candida albicans*, numa leitura 3'.

No sentido de averiguar se se ajusta uma distribuição normal aos resíduos ajustados, aplicaram-se ensaios de ajustamento de Kolmogorov-Smirnov. Considerou-se por um lado, a distribuição normal $N(0,1)$ e a distribuição normal de parâmetros estimados pelo conjunto dos dados, por outro lado.

Ensaio de ajustamento de Kolmogorov-Smirnov	$N(0,1)$	$N(0.04, 6.05)$	$N(0.11, 10.13)$	Probabilidade crítica (p-value)
Saccharomyces cerevisiae	0.3428	0.0408	-	0.02
Candida albicans	0.4052	-	0.0642	0.02

Tabela 2.4: Resultado dos ensaios de ajustamento de Kolmogorov-Smirnov para a *Candida albicans* e *Saccharomyces cerevisiae*.

Os resultados obtidos conduziram à rejeição da hipótese dos resíduos seguirem qualquer das distribuições normais consideradas, com probabilidade crítica igual a 0.02 (ver Tabela 2.4). Construiram-se as tabelas dos resíduos para cada uma das espécies, uma vez que se acredita que este novo conjunto de valores possa conter informação sobre as leis que governam o sequenciamento. Na Figura 2.6 apresenta-se um extrato da tabela de resíduos de Pearson da espécie *Saccharomyces cerevisiae*.

	AAA	AAG	AAT	AAC	ACT	ACC	ACA	ACG	CGT
LYS AAA	2,771	34,185	-7,033	5,170	-16,932	9,907	-1,979	2,458	-4,627
LYS AAG	6,207	31,116	-14,621	8,577	-14,081	14,136	2,444	-0,959	-4,795
ASN AAT	4,488	-2,436	15,303	16,944	-7,985	5,741	7,051	0,620	-7,400
ASN AAC	-2,240	2,564	9,030	21,648	-11,755	12,177	3,661	-1,587	-8,241
THR ACT	-6,400	-6,575	-6,375	3,272	13,146	21,837	0,468	-4,913	-3,058
THR ACC	-7,414	-3,286	-3,491	3,017	11,501	18,693	1,869	-2,573	-0,143
THR ACA	2,425	-0,700	-1,641	1,142	4,268	11,829	13,675	1,860	-5,170
THR ACG	-2,448	0,131	-1,855	-1,546	3,058	9,589	4,555	0,670	-2,856
ARG CGT	-2,635	-0,564	-6,385	-1,118	-5,818	-3,303	-4,021	-4,559	19,351
ARG CGC	-3,698	-0,553	-4,058	0,141	-4,062	-3,723	-1,888	-1,680	7,678
ARG CGA	2,380	-1,952	-2,642	-5,824	-4,976	-5,792	-1,130	1,511	3,492
ARG CGG	3,508	1,639	-4,477	-3,044	-2,583	-3,451	-0,233	0,205	1,989
ARG AGA	18,231	12,149	-13,639	-0,636	-11,810	-0,170	12,804	-0,131	-4,900
ARG AGG	17,274	7,952	-5,911	-0,788	-10,152	-2,518	6,792	1,935	-2,845
SER TCT	-8,369	-18,987	-1,773	2,427	13,410	1,872	0,933	-5,132	0,021
SER TCC	-9,319	-14,361	-2,435	2,853	7,316	-1,253	-0,242	-2,838	2,993
SER TCA	1,830	-14,079	3,334	-3,407	10,680	0,377	4,931	2,613	-0,092
SER TCG	-0,596	-7,934	1,097	-1,211	1,863	-2,443	0,309	0,487	-1,019
SER AGT	4,833	-6,030	5,806	9,440	-1,422	1,961	6,197	-1,436	-3,910
SER AGC	4,621	-3,022	4,108	12,568	-5,960	2,860	2,248	1,020	-5,994
ILE ATT	4,504	-6,863	3,812	-0,317	-4,253	9,936	-3,678	-8,531	1,464
ILE ATC	-0,248	-3,589	-0,060	2,301	-1,651	11,451	-8,918	-6,003	-0,310
ILE ATA	10,057	-4,621	-1,171	-1,766	-5,505	5,758	2,617	3,242	-2,972
MET ATG	2,130	-8,255	-2,335	-1,859	-11,172	1,987	1,494	-0,919	-3,767
PHE TTT	4,940	-1,665	-2,525	-4,075	3,392	-7,317	2,982	0,043	-7,917
PHE TTC	0,087	-4,627	-1,288	-1,755	6,530	-2,007	-7,023	-2,466	-3,373
TYR TAT	10,634	-2,586	-3,986	1,628	-6,513	-6,769	-1,030	0,672	4,613
TYR TAC	6,152	3,395	-4,609	4,529	-4,068	-3,061	-5,716	-1,784	5,888

Figura 2.6: Extrato da tabela de resíduos ajustados da *Saccharomyces cerevisiae*.

Observe-se que os resíduos quantificam a preferência de justaposição de símbolos face à independência. Os resíduos respondem em termos de preferências de justaposição, na medida em que os valores positivos informam as categorias preferidas da tabela e os valores negativos informam as que são preteridas, face à independência.
Por exemplo, dos símbolos representados na subtabela de resíduos da Figura 2.6 o codão AAA mostra, face à independência, maior preferência pelo codão AAG.
O grupo de trabalho do Departamento de Electrónica, ao longo do ano lectivo 2001/02, construiu um programa que faz a conversão da matriz dos valores dos resíduos ajustados numa imagem em que cada pixel tem uma côr. A escala de côr varia de vermelho a verde de acordo com a ordem de grandeza dos valores dos resíduos. Para valores de resíduos negativos, nulos ou positivos as cores correspondentes são o vermelho, o preto ou o verde, respectivamente (ver Figuras 2.7 e 2.8).

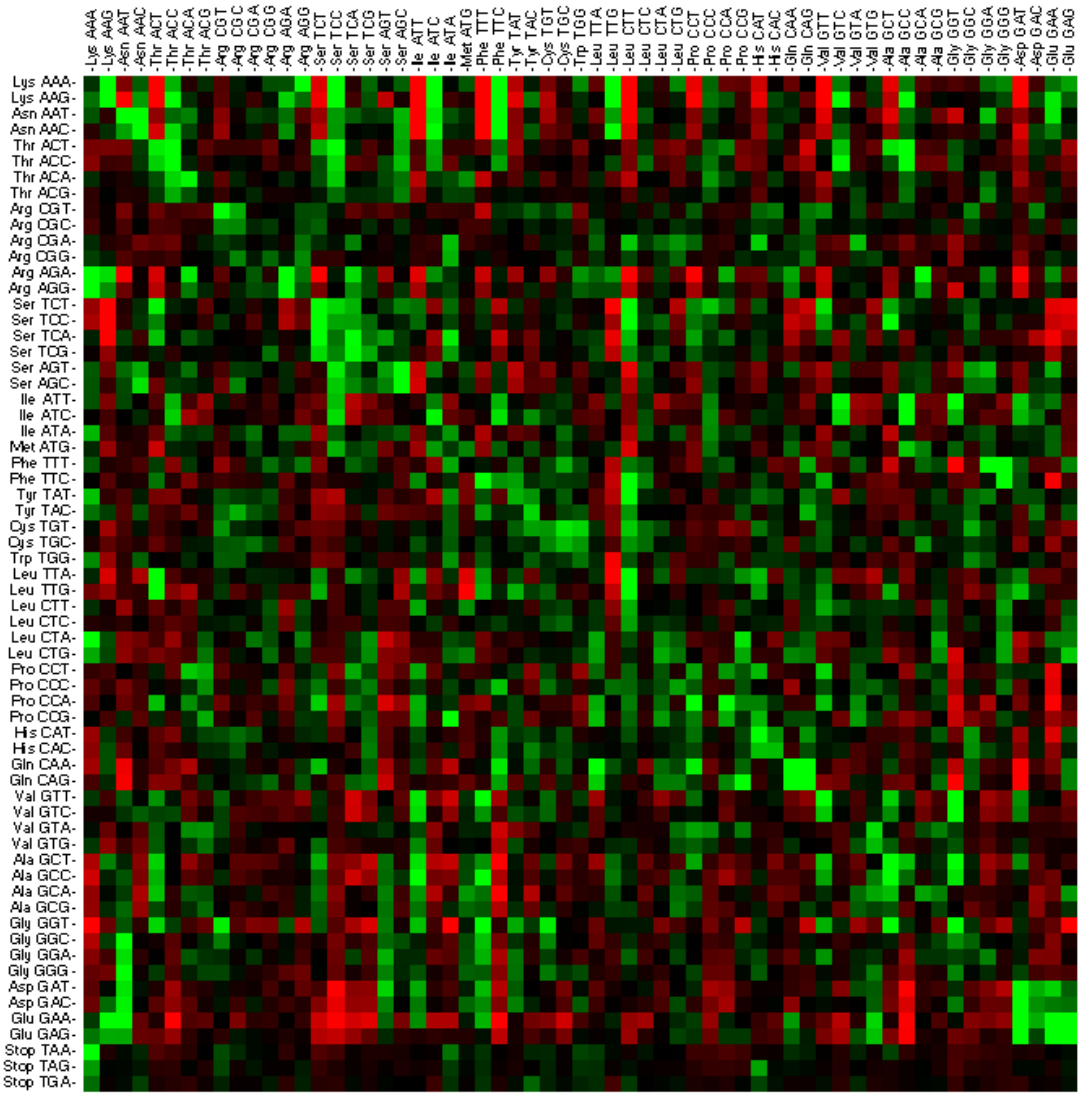


Figura 2.7: Imagem correspondente à matriz dos resíduos da *Saccharomyces cerevisiae*.

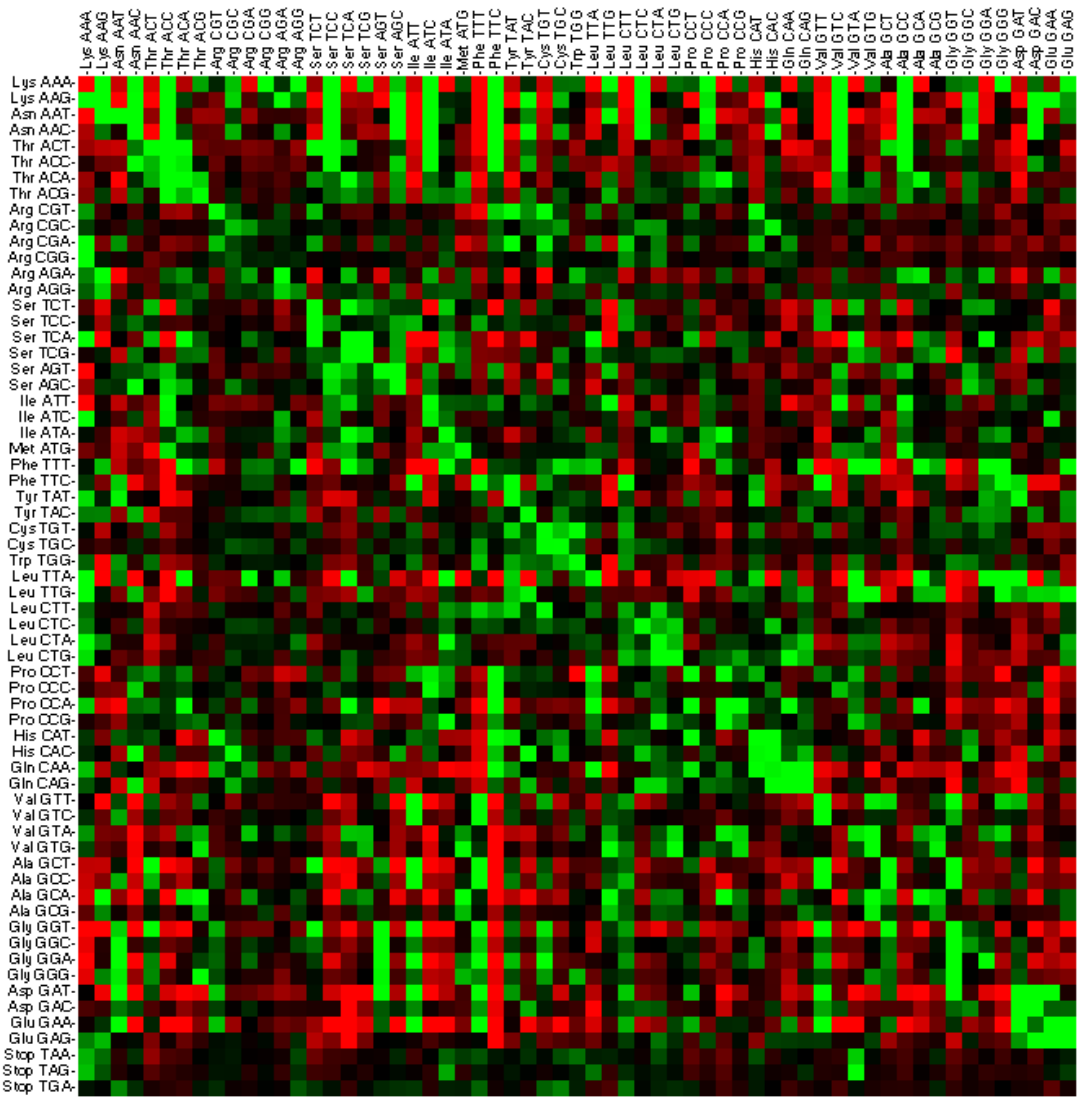


Figura 2.8: Imagem correspondente à matriz dos resíduos da *Candida albicans*.

A imagem correspondente a cada uma das matrizes de resíduos ajustados para a *Saccharomyces cerevisiae* e *Candida albicans* são dadas nas Figuras 2.7 e 2.8, respectivamente. Nelas observam-se que as diagonais têm tendência para tons de verde. Assim, pode-se afirmar que os codões na sua disposição sequencial têm preferência pela justaposição com eles próprios, face à independência. A título de curiosidade é de referir que este tipo de representação foi feito para um conjunto amplo de espécies e esta característica não é regra comum de todas elas.

2.5.4 Independência e Associação de Pares de Codões Espaçados de Cinco codões

Perante o resultado obtido em 2.5.2, a rejeição da independência entre pares de codões justapostos com valores muitos pequenos de associação, pensou-se em averiguar qual seria o comportamento de pares de codões espaçados de 5 símbolos. Em Biologia Molecular tudo leva a crer que a existência de associação entre dois codões justapostos seja grande e que a associação entre dois codões espaçados de cinco codões seja quase nula. Assim, da comparação entre os resultados obtidos para os dois casos poderá ser possível concluir algo mais sobre a associação entre pares de codões justapostos.
Foram feitas contagens dos possíveis pares de codões espaçados de 5 codões, (C_i, C_j), como ilustra o esquema da Figura 2.9.

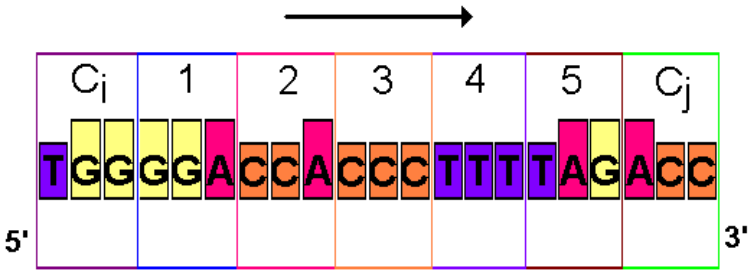


Figura 2.9: Ilustração do par (C_i, C_j) com espaçamento de 5 codões.

Para as Tabelas de Contingência obtidas calcularam-se os valores da estatística χ^2 e do coeficiente de Cramér (ver Tabela 2.5).

	χ^2	Valor teórico do quantil (0.05)	n	Coeficiente de associação de Cramér
Saccharomyces cerevisiae	12391.70	4052.7	596171	0.0003
Candida albicans	18450.48	4052.7	677620	0.0004

Tabela 2.5: Valores obtidos na análise das Tabelas de Contingência de pares de codões espaçados de 5 codões, para a *Candida albicans* e *Saccharomyces cerevisiae*.

Novamente obtém-se a rejeição da independência com valores muitos pequenos de associação. No entanto, comparando os resultados relativos aos pares justapostos (ver Tabelas 2.2 e 2.3) e aos pares espaçados de cinco símbolos (ver Tabela 2.5), averigua-se que os valores das estatísticas são maiores no caso dos codões justapostos. Assim, existe uma maior associação entre os pares de símbolos justapostos do que nos espaçados de 5 símbolos, resultado este coerente com aquilo que é biologicamente aceite. Contudo não se conseguiu concluir estatisticamente se os valores de associação entre os dois tipos de pares de símbolos são significativamente diferentes. Este problema, apesar de interessante para o projecto em desenvolvimento, fica por resolver em virtude de não terem sido encontradas metodologias apropriadas.

Capítulo 3

Análise Classificatória

3.1 Introdução

De modo geral, os métodos de Análise Classificatória encontram-se definidos na literatura especializada. Por exemplo, em [21] os métodos são definidos da seguinte forma:

Dado um conjunto de n indivíduos para os quais existe informação sobre a forma de N variáveis, o método de análise de clusters[1] *procede ao agrupamento dos indivíduos em função da informação existente, de tal modo que os indivíduos pertencentes a um mesmo grupo sejam tão semelhantes quanto possível e sempre mais semelhantes aos elementos do mesmo grupo do que a elementos dos restantes grupos.*

As técnicas a considerar efectuam uma classificação exclusiva e intrínseca, isto é, o resultado depois da aplicação do método define uma partição[2] do conjunto inicial, ou de forma mais explicita resulta numa hierarquia de partições[3]. No entanto existem outras técnicas que resultam num conjunto de grupos que não formam hierarquias de partições do conjunto inicial.
De facto, a Análise Classificatória consiste na construção de uma hierarquia de grupos de indivíduos *semelhantes* face a um conjunto de variáveis.
Na realidade, duas escolhas são necessárias na construção da hierarquia de grupos de indivíduos: a medida de semelhança entre indivíduos e o critério de agregação entre grupos. É conhecido que a utilização de diferentes medidas de semelhança e/ou diferentes critérios de agregação pode levar a resultados distintos. Assim, os resultados a obter podem depender da medida e do critério a utilizar. No entanto, o ideal na análise seria que os resultados segundo diferentes escolhas fossem conduzentes à mesma conclusão independentemente dos métodos utilizados.

[1] A expressão análise de *cluster* é utilizada por vários autores com o mesmo significada de Análise Classificatória.

[2] Partição de E, P(E), é o conjunto de partes de E, disjuntas duas a duas e cuja reunião é igual a E:

$$P(E) = \{A_i | A_i \subset E, i = 1, ..., k; A_i \cap A_j = \emptyset; \bigcup_{1 \leq i \leq k} A_i = E\}$$

[3] $P_0, P_1, ..., P_{i-1}, P_i, ..., P_k$ diz-se uma hierarquia de partições se:

$$\forall A \in P_{i-1} \exists B \in P_i : A \subset B \quad \text{com } i \in \{1, ..., k\} \text{ e } P_0 = E.$$

A Análise Classificatória pode ser feita de duas formas distintas: usando modelos aleatórios ou através de métodos exploratórios.
De acordo com o objectivo geral deste estudo, o de averiguar a existência de leis que regem o código genético, serão aplicados métodos de Análise Classificatória com o objectivo parcial o de tentar identificar grupos de indivíduos semelhantes quanto à preferência de vizínhos justapostos nas sequências de genes.
Os dados a considerar na Análise Classificatória serão tabelas de frequências relativas e tabelas de resíduos ajustados relativas aos pares de codões justapostos resultantes do estudo realizado no capítulo anterior. A característica coluna das tabelas a estudar refere-se ao primeiro codão do par. As suas categorias, de um modo geral no âmbito da Análise Classificatória, designam-se por *indivíduos*. A característica linha refere-se ao segundo codão do par e as componentes (categorias) constituintes desta característica designam-se de *variáveis*. Nas tabelas a estudar contabiliza-se um total de 64 codões (variáveis) e um conjunto de 61 indivíduos: 61 codões (não terminais).
No presente capítulo apresentar-se-á um conjunto de medidas de semelhança e três critérios de agregação (algoritmos), dos quais se farão alguns comentários relativos ao procedimento dos critérios. Apresentar-se-á a usual representação gráfica de hierarquias de partições o diagrama em árvore, também designado por dendograma. Por fim apresentar-se-ão os resultados da aplicação da Análise Classificatória aos 61 indivíduos, os codões não terminais, face às variáveis, os 64 codões possíveis de estar numa segunda posição nos possíveis pares de símbolos.
Os *softwares* utilizados nesta análise foram o *S-Plus 2000*, *SPSS 7.5 for Windows*, o *Cluster* e o *TreeView*.

3.2 Nomenclatura

Os dados sobre os quais se aplica a Análise Classificatória estão normalmente representados sob a forma de uma tabela do tipo:

	V_1	...	V_N
O_1	x_{11}	...	x_{1N}
...	...	...	...
O_n	x_{n1}	...	x_{nN}

sendo,
N - o número de variáveis;
n - o número de indivíduos;
V_i - a i-ésima variável, com $i \in \{1, ..., N\}$;
O_j - a j-ésima observação ou indivíduo, com $j \in \{1, ..., n\}$;
x_{ik} - o valor observado da variável V_k para o indivíduo O_i.

O conjunto de elementos sobre os quais se realizará a hierarquia de partições, corresponde ao conjunto de indivíduos $\{O_j \text{ com } j \in \{1, ..., n\}\}$. Cada uma das colunas, V_i com $i \in \{1, ..., N\}$, constituiu um conjunto que descreve num aspecto cada um dos indivíduos.

Fixe-se que a média de todas as variáveis para o indivíduo O_i é dada por:

$$\bar{x}_i = \frac{1}{N}\sum_{k=1}^{N} x_{ik};$$

E a matriz de covariâncias das N variáveis é dada por:

$$\Sigma = \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu)(X_k - \mu)' \quad \text{com} \quad \mu = \frac{1}{n}\sum_{k=1}^{n} X_k \quad \text{e} \quad X_i = [x_{i1} \ ... \ x_{iN}]'.$$

No contexto da análise multivariada define-se um grande número de medidas de semelhança, no sentido de medir a proximidade entre indivíduos. Apresentar-se-ão de seguida algumas das medidas mais comuns habitualmente consideradas na Análise Classificatória.

Medidas de semelhança

- Coeficiente de correlação de Pearson centrado[4]

$$r_{ij} = \frac{\sum_{k=1}^{N}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{N}(x_{ik} - \bar{x}_i)^2 \sum_{k=1}^{N}(x_{jk} - \bar{x}_j)^2}}. \tag{3.1}$$

  O coeficiente de correlação de Pearson entre dois indivíduos está sempre compreendido entre -1 e 1; quando assume o valor 1 significa que os indivíduos têm o mesmo tipo de comportamento, quando assume o valor 0 significa a existência de independência entre indivíduos, e para -1 significa que têm comportamentos opostos.
  Uma das propriedades desta medida de semelhança é a invariância a transformações de escala. Por exemplo, se se multiplicarem todas as componentes de um indivíduo por um mesmo factor, os valores do coeficiente de correlação relativos a esse indivíduo mantêm-se. O coeficiente de correlação é uma boa medida para associar indivíduos com idênticos perfis.
  O coeficiente de correlação de Pearson não é uma métrica pois não verifica, por exemplo, a desigualdade triângular (a definição de métrica encontra-se no Apêndice A.1).

[4]O Coeficiente de correlação de Pearson não centrado tem a seguinte forma:

$$r_{ij} = \frac{\sum_{k=1}^{N} x_{ik}x_{jk}}{\sqrt{\sum_{k=1}^{N}(x_{ik})^2 \sum_{k=1}^{N}(x_{jk})^2}}.$$

A distância correspondente ao coeficiente de correlação de Pearson obtém-se através de uma pequena alteração deste coeficiente, e é dada por:

$$d_{ij} = \sqrt{0.5(1 - r_{ij})} \text{ - Distância de Pearson.}$$

Observe-se que esta distância assume valores entre zero e um. Dois indivíduos são tanto mais semelhantes quanto mais a sua distância se aproxime de 0.

- Distância Euclideana

$$d_{ij} = \sqrt{\sum_{k=1}^{N} (x_{ik} - x_{jk})^2}. \tag{3.2}$$

Esta medida de semelhança para além de satisfazer as propriedades de uma métrica, é semi-definida positiva, invariante para transformações ortogonais sobre os indivíduos. Contudo não é invariante a mudanças de escala assim como a medida de semelhança que se apresenta de seguida.

- Distância absoluta ou *City-Block Metric*

$$d_{ij} = \sum_{k=1}^{N} |x_{ik} - x_{jk}|. \tag{3.3}$$

- Distância Euclideana Estandartizada ou Distância de Karl Pearson

$$d_{ij} = \sqrt{\sum_{k=1}^{N} \frac{(x_{ik} - x_{jk})^2}{w_k}} \tag{3.4}$$

com

$$w_k = \frac{1}{N} \sum_{l=1}^{N} (x_{lk} - \bar{x}_k)^2.$$

As medida estandardizadas têm a propriedade de ser invariante a mudanças de escala.

- Distância Absoluta Estandartizada

$$d_{ij} = \sum_{k=1}^{N} \frac{|x_{ik} - x_{jk}|}{w_k}. \tag{3.5}$$

Como medida estandartizada é invariante a mudanças de escala.

- Distância Mahalanobis

$$d_{ij} = (X_i - X_j)' \Sigma^{-1} (X_i - X_j). \tag{3.6}$$

Cada medida tem características particulares e consoante a situação em estudo torna-se mais ou menos adequada. Contudo, no estudo de dados multivariados, é na maior parte das vezes complicado identificar a melhor medida de semelhança.

3.3 Métodos Hierárquicos

Os métodos de agrupamento a apresentar são métodos hierárquicos. Estes resultam em hierarquias de partições. Os métodos hierárquicos podem ainda subdividir-se em métodos aglomerativos e divisívos. A aplicação dos métodos aglomerativos determina inicialmente uma partição com tantas partes quanto o número de diferentes indivíduos, enquanto que nos métodos divisívos considera como ponto de partida uma partição com uma só parte, um único conjunto a que pertencem todos os indivíduos.
Dos métodos hierárquicos apenas se estudarão os métodos aglomerativos. Na realidade são os métodos aglomerativos os mais usados e mais divulgados na literatura. Um dos motivos de tal facto é o esforço computacional destes face aos divisívos ser menor para o mesmo tipo de resultados.
Para métodos aglomerativos definem-se vários critérios de agregação. O que distingue os diversos critérios de agregação é a definição de distância entre grupos considerada. Exemplos de critérios de agregação são: *single linkage, complete linkage*, critério da média, critério do centróide e método de Ward.
Não se pode dizer que existe um critério de agregação que seja melhor que os outros. Na prática, o que se faz é utilizar vários critérios e comparar os resultados. Se os resultados de diferentes critérios forem concordantes o resultado final será "mais credível".
De seguida apresentar-se-ão os algoritmos de três dos cinco critérios de agregação referidos, para os quais ter-se-á de fixar uma medida de semelhança entre indivíduos. No estudo a desenvolver apenas se utilizarão distâncias como medidas de semelhanças.
Seja d_{ij} a medida entre os indivíduos O_i e O_j.

- **Análise Classificatória usando agrupamento entre os vizinhos mais próximos.**
  Critério da Ligação Única - Single Linkage

  Seja $D = [d_{ij}]_{(n\times n)}$ a matriz de semelhanças entre os n indivíduos. Tome-se d_0 como a distância máxima admitida entre grupos (critério de paragem). É de referir que vários programas não dão a possibilidade de escolher o valor d_0, como por exemplo o *software* estatístico *S-Plus 2000*, para estas circunstâncias o algoritmo termina quando encontrar a partição definida por uma única classe ou grupo a qual contém todos os indivíduos.

  1. A partir dos n indivíduos, O_1, ..., O_n, construam-se n grupos G_1, ..., G_n em que o grupo G_i contém apenas o elemento O_i, com $i \in \{1, ..., n\}$.
  2. Sem perda da generalidade, seja $d_{12} = min_{i,j} d_{ij} < d_0$ com $i, j = 1, ..., n$. Juntem-se os dois grupos mais semelhantes G_1 e G_2 no grupo G_2*. Os restantes grupos mantêm-se inalterados ($G_i* = G_i$, com $i \in \{3, ..., n\}$). Assim, G_2*, ..., G_n* é a segunda partição da hierarquia de grupos.
  3. Redefina-se a matriz de distâncias entre grupos por $D* = [d_{ij}*]_{((n-1)\times(n-1))}$, com

  $$d_{ij}* = \begin{cases} min(d_{1j}, d_{2j}) & \text{para } i = 2 \text{ e } j = 2, ..., n \\ d_{ij} & \text{para } i, j = 3, ..., n. \end{cases}$$

  Se $min_{i,j}(d_{ij}*) < d_0$ com $i, j = 2, ..., n$, então redefina-se $d_{ij} = d_{(i+1)(j+1)}*$ e $G_i = G_{i+1}*$, com $n = n - 1$ e $i, j = 1, ..., n$, e passe-se ao ponto (2); caso contrário, pára o algoritmo.

Observe-se que os passos (2) e (3) são repetidos, até que $min(d_{ij}*) \geq d_0$ ou a partição construída tenha apenas uma classe que contenha todos os indivíduos.

- **Análise Classificatória usando ligações completas.**
  Critério da Ligação Completa - Complete Linkage

  O algoritmo é semelhante ao anterior, deferindo apenas na definição de distância entre grupos.
  Neste caso, o passo 3 é substituído por:

  - Defina-se uma nova matriz de distâncias entre os grupos G_2*, ..., G_n* por $D* = [d_{ij}*]_{((n-1)\times(n-1))}$ com,

$$d_{ij}* = \begin{cases} max(d_{1j}, d_{2j}) & \text{para } i = 2 \text{ e } j = 2, ..., n \\ d_{ij} & \text{para } i,\ j = 3, ..., n. \end{cases}$$

- **Análise Classificatória usando média do grupo**
  Critério da ligação média - Average Linkage

  Neste caso, o passo 3 é substituído por:

  - Defina-se uma nova matriz de distâncias entre os grupos G_2*, ..., G_n* por $D* = [d_{ij}*]_{(n-1)\times(n-1)}$ com

$$d_{2j}* = \begin{cases} \frac{1}{2}(d_{1j} + d_{2j}) & \text{para } i = 2 \text{ e } j = 2, ..., n \\ d_{ij} & \text{para } i,\ j = 3, ..., n. \end{cases}$$

3.3.1 Dendograma

O processo de agrupamento ou de agregação[5] pode ser representado por um diagrama bi-dimensional em forma de árvore conhecido por *dendograma*. No eixo dos xx estão representados os indivíduos e no eixo dos yy as distâncias (ver Figura 3.1).
O dendograma tem a vantagem de facilitar a visualização do processo de agrupamento nas suas diversas fases, desde os indivíduos separados até à inclusão num só grupo. Na realidade o dendograma é uma representação gráfica da hierarquia de partições.
Os ramos da árvore que constituem o dendograma identificam-se com as $n-1$ ligações entre grupos, onde cada linha horizontal representa uma ligação a que normalmente se chama de *nodo*.
Nos métodos aglomerativos a primeira ligação identifica-se com a menor ramificação e a segunda ligação com a segunda menor ramificação, assim sucessivamente.

[5] Diz-se processo de agrupamento ou de agregação, o processo que determina as semelhanças entre indivíduos através de uma medida de semelhança e efectua o agrupamento dos indivíduos em grupos através de um critério de agregação.

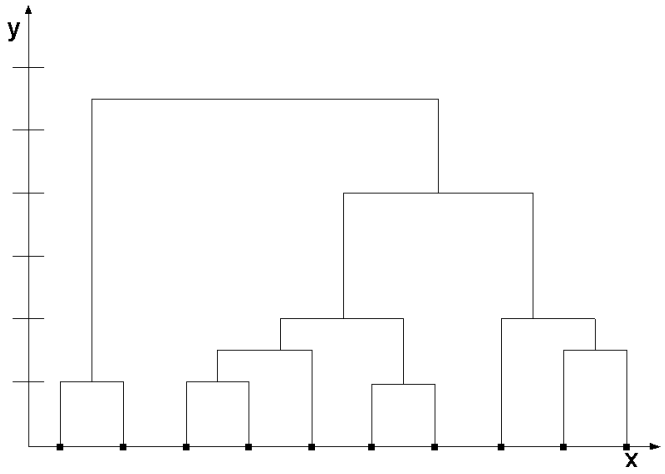


Figura 3.1: Dendograma.

3.3.2 Características do Processo de Agrupamento

Algumas das seguintes propriedades relativas aos métodos aglomerativos encontram-se estabelecidas na literatura da especialidade, por exemplo em [6] e [17]:

- Para os processos aglomerativos anteriormente apresentados, dada uma ligação esta já não se quebra;
- Nos métodos aglomerativos *single linkage* e *complete linkage*, as transformações monótonas sobre as medidas de semelhança entre indivíduos (d_{ij}) não alteram a estrutura do dendograma;
- Se se conhecer o número de grupos de uma dada amostra ou da população em estudo, a partir do dendograma poder-se-á identificar a constituição de cada grupo;
- O critério de *single linkage* produz grupos com efeito de encadeamento: em todos os grupos qualquer elemento é mais "semelhante" a pelo menos um outro elemento do mesmo grupo do que a qualquer elemento de outro grupo;
- O critério de *complete linkage* produz grupos sem efeito de encadeamento: em todos os grupos qualquer elemento é mais semelhante a qualquer elemento do mesmo grupo do que a qualquer elemento de outro grupo;
- Os critérios da média e de *complete linkage* tem tendência para encontrar grupos com representação esférica, mesmo que a estrutura dos dados revele outras formas;
- No critério de *complete linkage* todas as distâncias dentro dos grupos são inferiores a d_0, enquanto que no critério de *single linkage* nem sempre tal se verifica. Observe-se que para o critério de *single linkage* $d_{ij} \leq max(d_{ik}, d_{kj})$, para todo o i, j e k.

3.3.3 Escolha do Número de Grupos

A aplicação dos algoritmos apresentados resulta numa hierarquia de grupos. Se o número de grupos for conhecido à partida, a identificação dos grupos é quase imediata, à custa do dendograma. Caso contrário, a observação do dendograma pode sugerir uma estimativa para número de grupos, mas nem sempre essa escolha é objectiva. A escolha do número de grupos poderá ser ainda menos objectiva se ao utilizar diferentes critérios de agregação forem obtidos dendogramas que sugiram diferentes partições. A determinação do número de grupos é feita cortando horizontalmente o dendograma. A título de exemplo, no dendograma da Figura 3.2 pode-se propôr duas partições: uma correspondente ao traço a vermelho e a outra ao traço verde, com 3 e 2 grupos, respectivamente. A escolha da melhor partição, em particular, do número de grupos óptimo, deverá ser feita mediante o contexto do problema, pelo que o conhecimento prévio da natureza dos dados pode auxiliar nesta decisão.

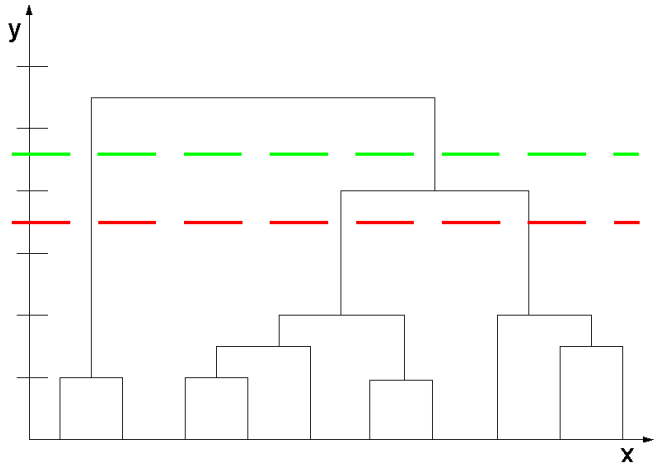


Figura 3.2: Dendograma que realça duas escolhas do números de grupos.

3.4 Aplicação ao Caso em Estudo

Para cada uma das espécies em estudo obtiveram-se as tabela das frequências relativas a partir das tabelas de contingência (dividindo o número de observações n_{ij} da célula (i, j) pelo total marginal $n_{i.}$).
A fim de determinar grupos de codões com comportamento semelhante face à "escolha" do vizinho justaposto realizou-se uma Análise Classificatória usado a tabela das frequências relativas referida e a tabela dos resíduos ajustados obtida a quando do estudo das Tabelas de Contingência no Capítulo 2.
Como já se afirmou, os métodos de agregação existentes podem produzir diferentes resultados. Pode-se observar essa divergência na aplicação aos dados em estudo.
Os critérios de aglomeração, considerados no âmbito desta aplicação da Análise Classificatória, foram o de *Single Linkage*, *Complete Linkage* e *Average Linkage*. As medidas de semelhança usadas foram a euclideana, a euclideana estandartizada e o coeficiente de correlação de Pearson.
Nas Figuras 3.3, 3.4, 3.5 e 3.6 encontram-se alguns dos dendogramas obtidos na análise realizada para as duas espécies.

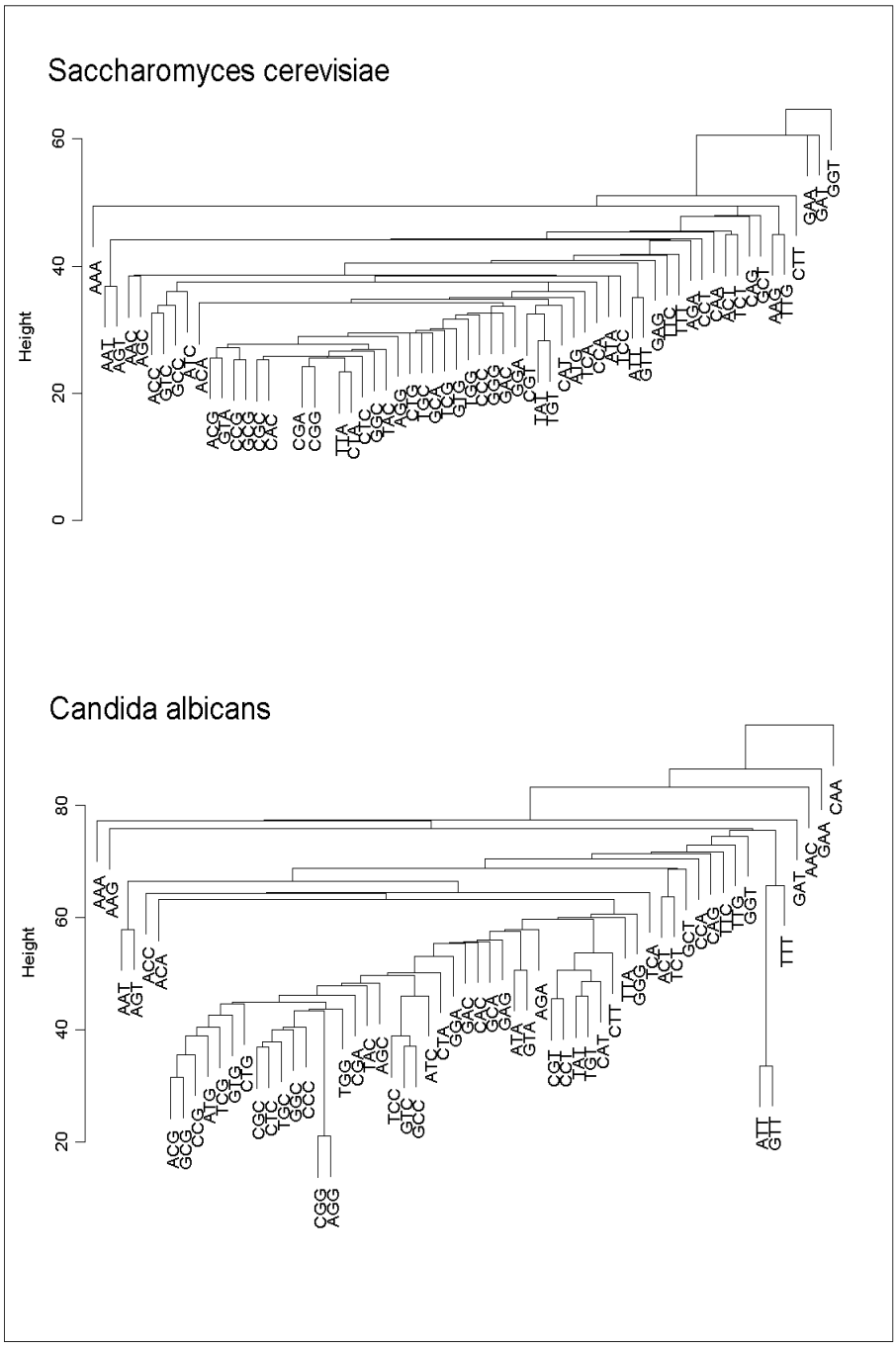


Figura 3.3: Dendograma usando o método de *Single Linkage* e a distância euclideana como medida de semelhança, para ambas as espécies numa leitura 3' com a matriz dos resíduos ajustados.

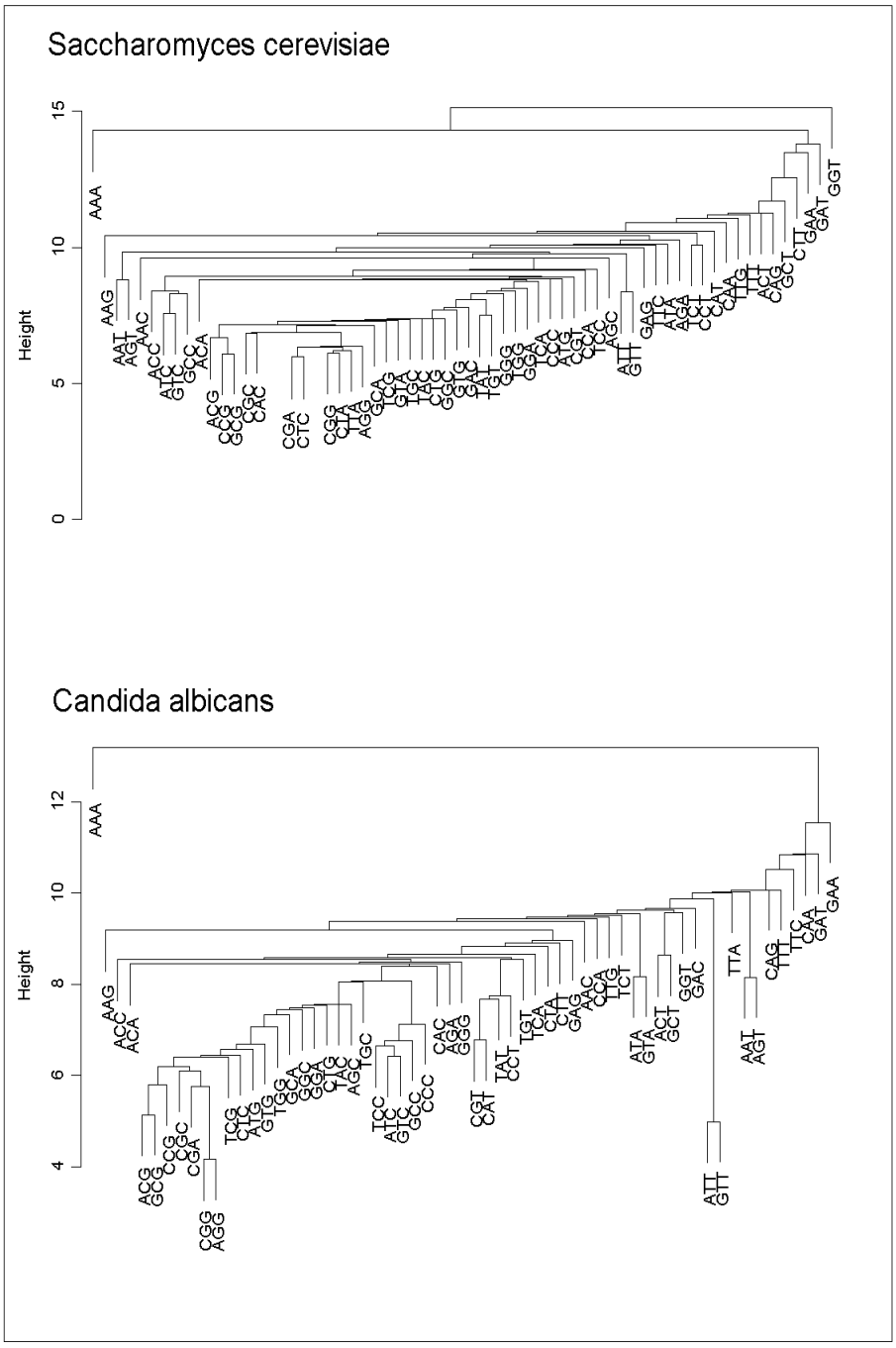


Figura 3.4: Dendograma usando o método de *Single Linkage* e a distância euclideana estandartizada como medida de semelhança, para ambas as espécies numa leitura 3' com a matriz dos resíduos ajustados.

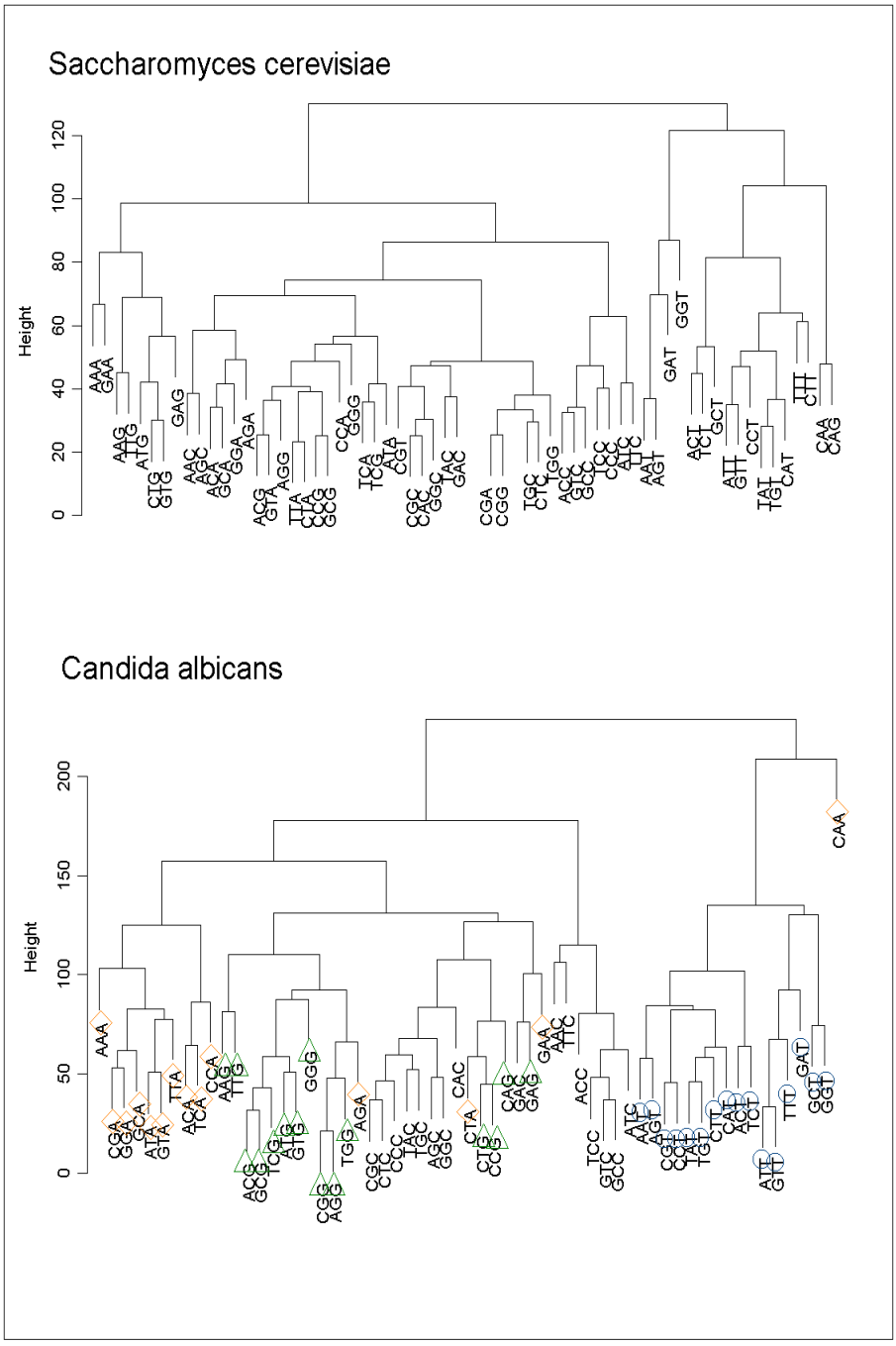


Figura 3.5: Dendograma usando o método de *Complete Linkage* e a distância euclideana como medida de semelhança, para ambas as espécies numa leitura 3' com a matriz dos resíduos ajustados.

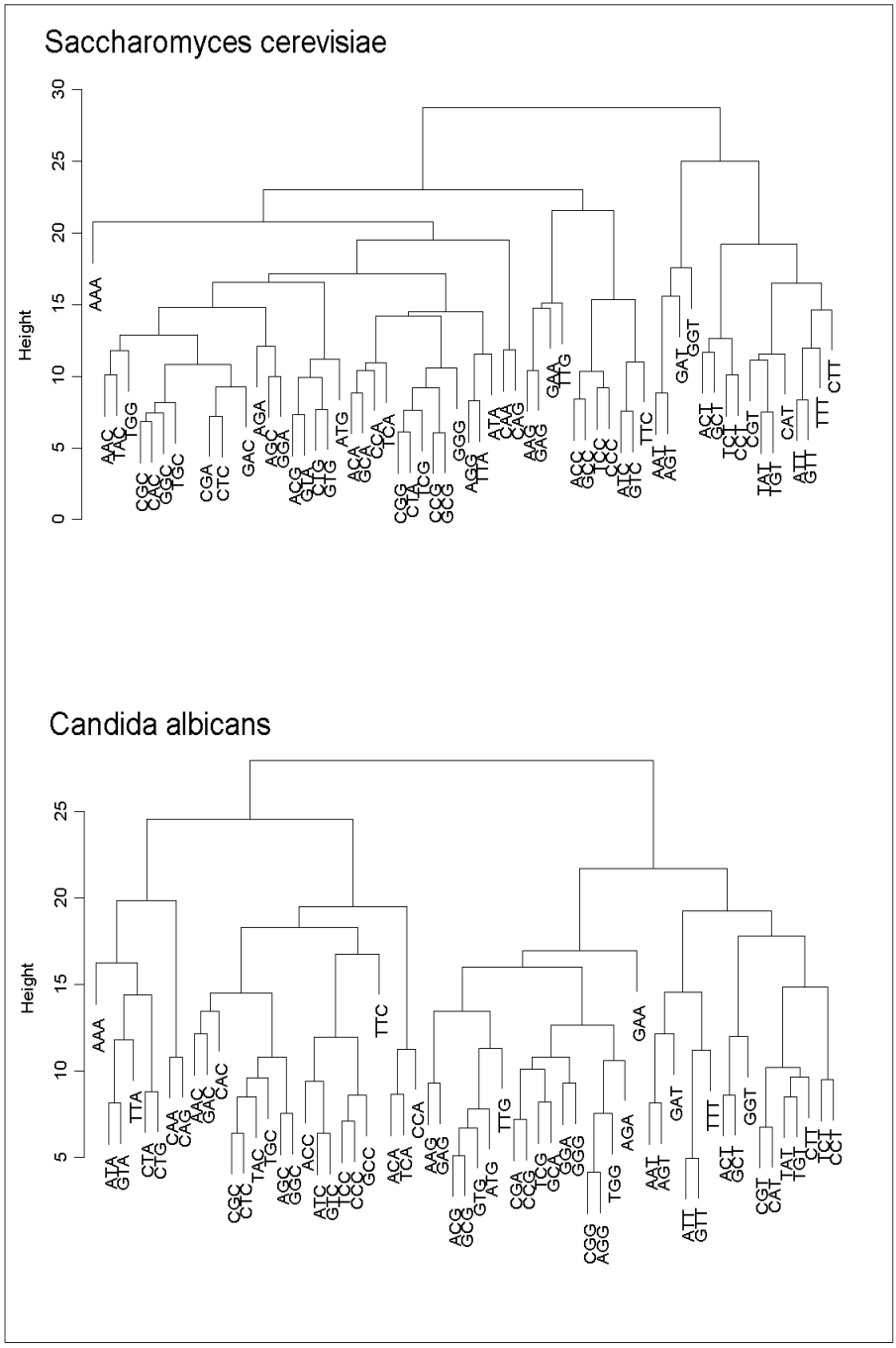


Figura 3.6: Dendograma usando o método de *Complete Linkage* e a distância euclideana estandartizada como medida de semelhança, para ambas as espécies numa leitura 3' com a matriz dos resíduos ajustados.

Tal como se pode concluir da análise das Figuras 3.3, 3.4, 3.5 e 3.6, os dendogramas obtidos dependem das medidas de semelhança e dos critérios de agregação utilizados, não permitindo objectividade na identificação da partição óptima. Para além dos dendogramas apresentados foram obtidos outros por aplicação de outros critérios de aglomeração e medidas de semelhança, resultando num conjunto de dendogramas distintos entre si e dos anteriores.
O *S-Plus 2000* apresenta apenas duas métricas como medida de semelhança: a euclideana e a de *Manhattan*[6], com a possibilidade de estandartizar. No sentido de utilizar também o coeficiente de correlação de Pearson como medida de semelhança na construção das partições, recorreu-se a outros *softwares*: o *Cluster* e *TreeView*[7].
Da aplicação conjunta do *Cluster* e *TreeView* obtém-se com o auxílio de um programa de desenho não só o dendograma dos indivíduos como também o das variáveis e uma matriz de cores.
À custa da aplicação conjunta dos dois programas fez-se uma Análise Classificatória às tabelas de resíduos ajustados usando *Single Linkage*, o coeficiente de correlação de Pearson. Nas Figuras 3.7 e 3.8 apresentam-se os dendogramas obtidos para a *Candida albicans* e *Saccharomyces cerevisiae*, respectivamente. Observa-se uma hierarquia de partições sendo visível a formação de quatro grupos. Averigua-se também que dentro de cada um dos quatro grupos o nucleótido com que terminam os codões (indivíduos) é quase sempre o mesmo. Este facto podería levar à formulação da seguinte hipótese:

O comportamento da sequência que se segue a um dado codão fixo é marcada pelo nucleótido terminal desse codão.

Contudo, o conjunto dos dendogramas apresentados para outras medidas de semelhança e para outros critérios de agregação não evidenciam a hipótese anterior. Apenas na Figura 3.6 para a *Candida albicans* assinala-se o nucleótido terminal e averigua-se um grupo de codões bem definido que terminam no nucleótido T. Poder-se-á daqui, quando muito, presumir que no sequenciamento de codões, ao considerar um codão fixo que termine no nucleótido T o sequenciamento que se lhe segue tem semelhante comportamento.
Observe-se que o comportamento é relativo à matriz dos resíduos ajustados, valores estes associados à preferência ou preterência de justaposição face à independência de justaposição entre símbolos.

[6] A definição usada pelo *help* do programa *S-Plus 2000* é: *manhattan distances are the sum of absolute differences*. Portanto é a conhecida distância absoluta ou *City-Block Metric*.
[7] Software de Michael Eisen.

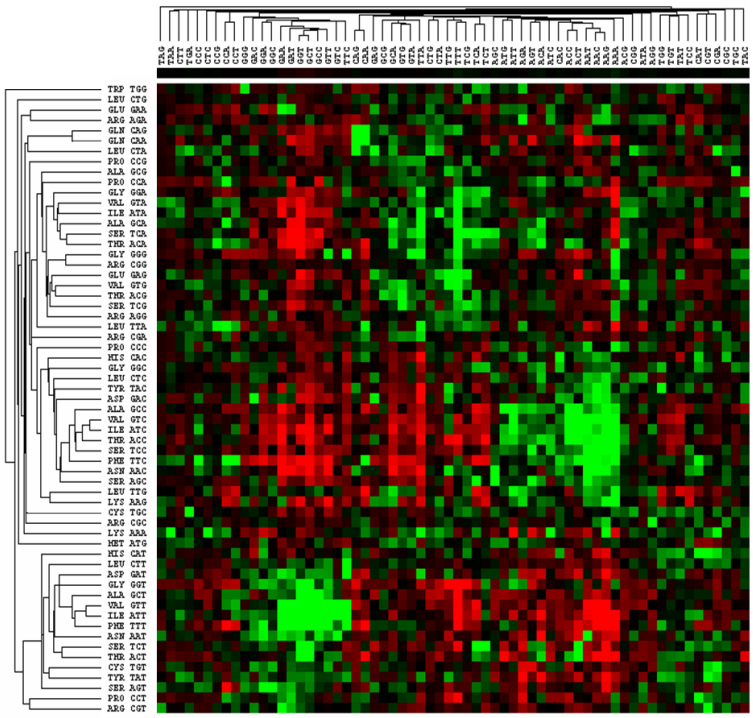


Figura 3.7: Dendograma usando o coeficiente de correlação de Pearson e o método de *Single Linkage*, para a *Candida albicans* numa leitura 3' com a matriz dos resíduos ajustados.

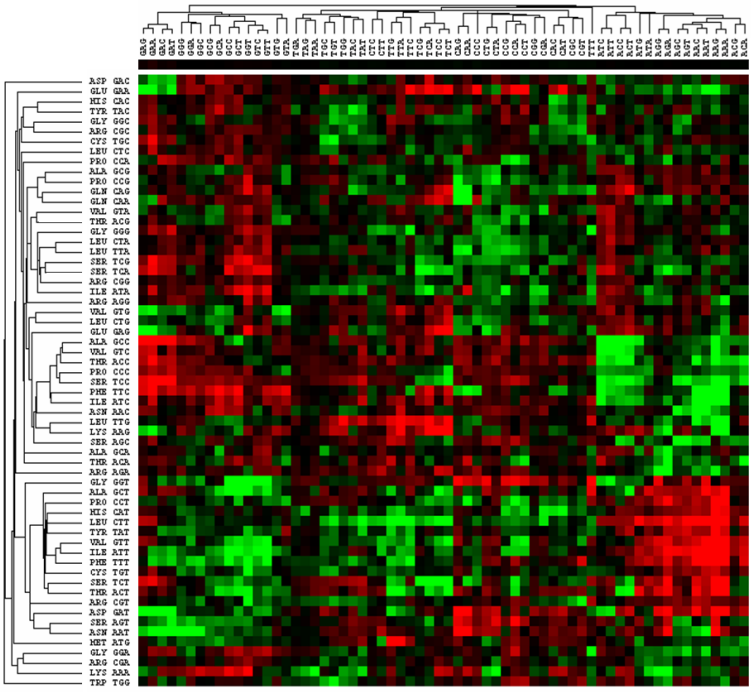


Figura 3.8: Dendograma usando o coeficiente de correlação de Pearson e o método de *Single Linkage*, para a *Saccharomyces cerevisiae* numa leitura 3' com a matriz dos resíduos ajustados.

No sentido de extrair mais informação dos dados utilizaram-se, em vez da matriz dos resíduos ajustados, as tabelas de frequências relativas. As frequências relativas a utilizar correspondem às frequências com que aparece um dado par de codões face à frequência total de cada indivíduos (primeiro codão do par de codões justapostos). As frequências relativas medem novamente as preferência e as preterências de justaposição só que, neste caso, relativamente às frequências totais de cada indivíduo.
Neste novo contexto aplicaram-se novamente os vários critérios de agregação e medidas de semelhança atrás consideradas. Apresentam-se aqui os resultados de um só processo de agregação (Figura 3.9), já que ocorre situação análoga à dos resíduos: diferindo um dos parâmetros do processo de agregação obtêm-se dendogramas muito distintos entre si, não sendo no seu conjunto conclusivos.
Assim, da observação do conjunto de todos os dendogramas patentes neste capítulo, não se pode identificar como característica de divisão dos grupos o nucleótido de terminação dos codões. No entanto, realizando uma observação mais minuciosa averigua-se na maior parte dos dendogramas uma pequena tendência de agrupamento de símbolos com a mesma terminação. Surgindo a hipótese de que:

Numa sequência de codões o nucleótido terminal de um dado codão tem forte influência na escolha do codão que lhe sucede.

No entanto, parece que o nucleótido terminal só por si não marca a distribuição da sequência de codões que lhe segue, outros factores parecem contribuir. Considerando a análise realizada no Capítulo 2 é de crer que:

A sequência de codões tem comportamento Markoviano.

ou que:

Dado um codão fixo a "escolha" do codão que lhe sucede poderá ser determinado não só pelo nucleótido terminal do codão fixo, como também pelo nucleótido intermédio e inicial com níveis diferentes de influência, sendo o nucleótido intermédio menos determinante na escolha.

Para a leitura 5' também foi feito um estudo análogo e foram obtidos o mesmo tipo de resultados, neste caso face ao primeiro nucleótido do codão fixo. Assim, dispensa-se a apresentação do conjunto de dendogramas referente à leitura 5'.

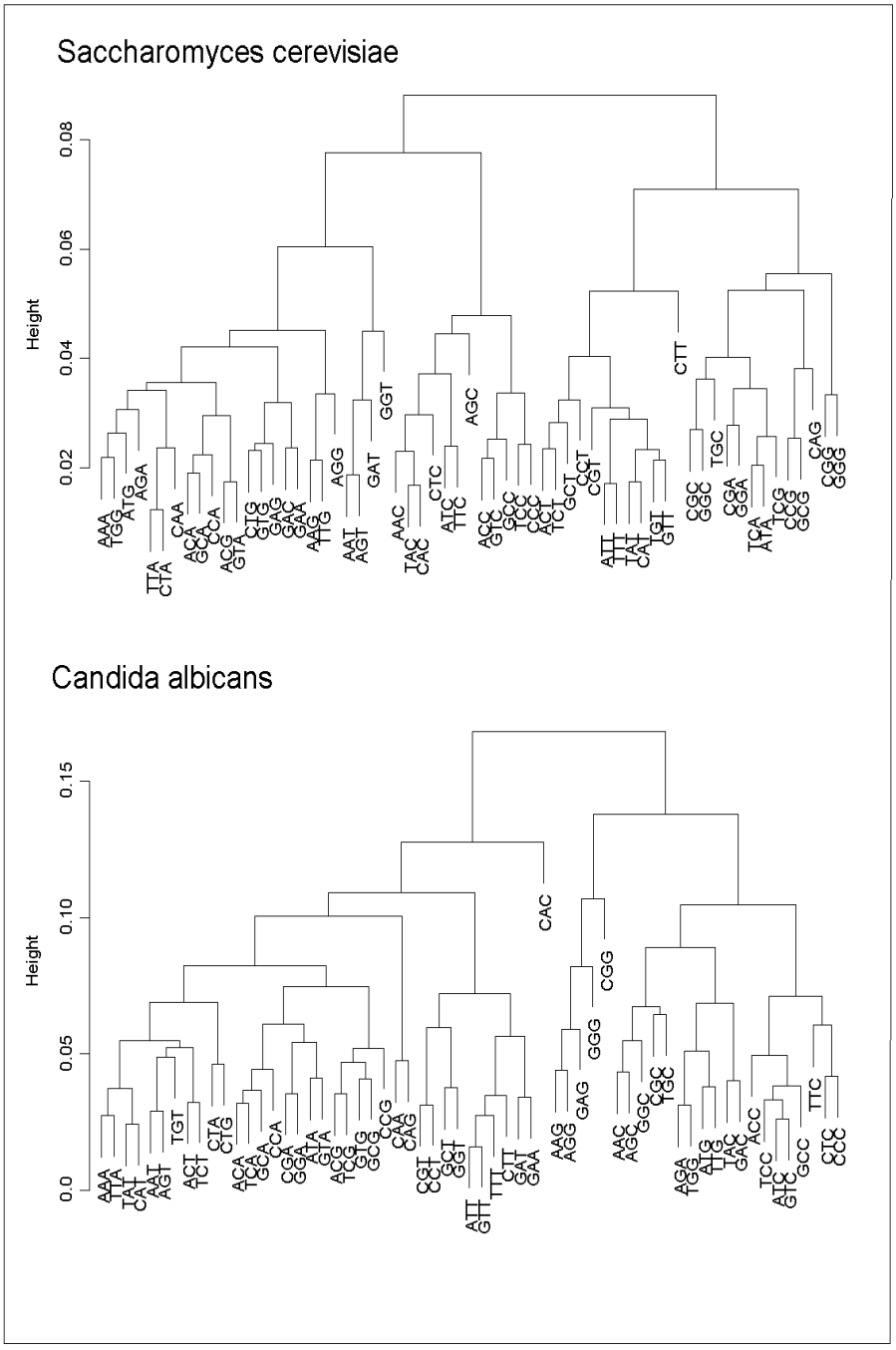


Figura 3.9: Dendograma usando o método de *Complete Linkage* e a distância euclideana como medida de semelhança, para a ambas as espécies numa leitura 3' com a matriz das frequências relativas.

Capítulo 4

Análise em Componentes Principais

4.1 Introdução

A Análise em Componentes Principais é uma das técnicas exploratórias mais populares da análise multivariada e tem por objectivo descrever um conjunto de variáveis correlacionadas através de um conjunto de menor número de variáveis, sem perda "significativa" da informação contida nas variáveis de partida.
A técnica de Análise em Componentes Principais foi definida primeiro por Karl Pearson (1901) que na prática usou-a no máximo para três variáveis iniciais. A técnica foi mais tarde desenvolvida por Hotelling (1933).
Neste capítulo apresentar-se-á a definição de componentes principais, assim como um conjunto de resultados e conceitos inerentes. Para além do método exploratório, apresenta-se o teste de esfericidade de Bartlett para validação da aplicabilidade da Análise em Componentes Principais. Por fim, apresentar-se-ão os resultados da aplicação desta análise às tabelas de resíduos ajustados e frequências relativas dos pares de codões justapostos.
O *software* utilizado foi *SPSS 7.5 for Windows* e o *S-Plus 2000*.

4.2 Nomenclatura

Os dados sobre os quais se aplica a Análise em Componentes Principais estão usualmente organizados em matrizes de dados semelhantes às consideradas na Análise Classificatória, pelo que toda a nomenclatura introduzida na secção 3.2 é novamente considerada.

4.3 Método de Análise em Componentes Principais

A Análise em Componentes Principais é um método estatístico que permite transformar um conjunto de variáveis iniciais correlacionadas entre si, num conjunto de variáveis não correlacionadas, chamadas *componentes principais*.

Definição 4.3.1 *As componentes principais CP_1, ..., CP_p de um conjunto de N variáveis para o qual se tem um conjunto de n observações, consiste num conjunto de p variáveis não correlacionadas entre si, com $p \leq N$, tais que:*

$$\begin{aligned} CP_1 &= a_{11}V_1 + a_{12}V_2 + ... + a_{1N}V_N = a_1'V \\ &\vdots \\ CP_p &= a_{p1}V_1 + a_{p2}V_2 + ... + a_{pN}V_N = a_p'V \end{aligned}$$

com a_i=$[a_{i1}\ a_{i2}\ ...\ a_{iN}]'$, $i \in \{1,\ ...,\ p\}$ e $V = [V_1\ ...\ V_N]'$.
Cada uma das combinações lineares anteriores é uma combinação linear estandardizada, ou seja:

$$a_j'a_j = \sum_{i=1}^{N} a_{ji}^2 = a_{j1}^2 + a_{j2}^2 + ... + a_{jN}^2 = 1, \text{ para } j=1, ..., N.$$

Os coeficientes a_{11}, a_{12}, ..., a_{1N} são definidos de forma a maximizar a variância da variável CP_1, tal que $a_{11}^2 + a_{12}^2 + ... + a_{1N}^2 = 1$. Fixa a componente CP_1, constroi-se a variável CP_2 de forma análoga a CP_1, garantindo a ausência de correlação entre as duas componentes. As restantes $p-2$ componentes principais são construídas sucessivamente e de forma a não serem correlacionadas entre si.

Como consequência directa da definição tem-se:

$$var(CP_1) \geq var(CP_2) \geq ... \geq var(CP_p).$$

Pode-se calcular tantas componentes quantas as variáveis iniciais, mas o interesse está em considerar um número reduzido de componentes que expliquem, na medida do possível, a variabilidade dos dados iniciais.
A variabilidade dos dados iniciais é traduzida pela matriz de covariâncias. Contudo, para um grande conjunto de observações e de variáveis, a variabilidade traduzida numa matriz envolve maior esforço computacional e não é de fácil interpretação. Usualmente, no contexto da Análise em Componentes Principais, reduz-se aquela informação a uma medida de variação total dos dados iniciais dada pelo traço da matriz de covariâncias.
Muitas vezes, em lugar de utilizar as componentes principais definem-se outras componentes com a mesma estrutura das componentes principais, com a pequena diferença de serem corrigidas pela média, a estas componentes chamam-se de *scores*. Portanto, quando se utilizam os *scores*, em lugar de V_i considera-se $V_i - \mu_i$ na determinação das componentes principais.
Uma Análise em Componentes Principais pode ser vantajosa se as variâncias de grande número de componentes forem "pequenas" e existir um conjunto pequeno de componentes

com "grande" variância. Se tal acontecer algumas componentes podem ser desprezadas sem perdas significativas em termos de variância. Assim, os dados seriam descritos por um menor número de variáveis que explicariam, uma percentagem elevada de variabilidade das variáveis iniciais.
Mas nem sempre a aplicação da Análise em Componentes Principais se revela de grande utilidade. Basta para isso conduzir a muitas componentes cuja variância não é desprezável. Nesse caso, não seria possível reduzir "significativamente" o número de variáveis, facto este que acontece quando as variáveis iniciais são "pouco" correlacionadas. Para o caso das variáveis serem "significativamente" correlacionadas já o sucesso da Análise em Componentes Principais pode ser grande.

4.3.1 Procedimento

Sejam λ_1, ..., λ_N os valores próprios da matriz de covariâncias Σ das variáveis V_1, V_2, ..., V_N ordenados por ordem decrescente.
Para o cálculo dos vectores próprios associados a cada valor próprio recorre-se ao seguinte resultado clássico da Álgebra Linear, adaptado à presente situação:

Teorema 4.3.1 *Seja I a matriz identidade. Se λ_j, com $j \in \{1,...,N\}$, é um valor próprio da matriz Σ, então existe uma solução não nula v_j da equação $[\Sigma - \lambda_j I]v_j = 0$. Ao vector v_j=[v_{i1} v_{i2} ... v_{iN}]' chama-se de vector próprio associado a λ_j.*

Fixe-se que v_j é o vector próprio estandartizado, da matriz de covariâncias Σ, associado a λ_j e ordenados por ordem decrescente de valor próprio. Diz-se que v_j é um vector estandartizado se $v_j' v_j = 1$, com $j \in \{1,...,N\}$.
Verificar-se-á que o vector a_i da Definição 4.3.1, coincide com o vector próprio v_i.

Teorema 4.3.2 *Qualquer combinação linear estandardizada de V_1, V_2, ..., V_N tem variância não superior a λ_1 - o maior valor próprio de Σ.*

Prova 4.3.1
Suponha-se sem perda da generalidade que V_1, V_2, ..., V_N são linearmente independentes.
Seja α uma combinação linear estandardizada de V_1, V_2, ..., V_N dada por:

$$\alpha = b_1 V_1 + b_2 V_2 + ... + b_N V_N = b'V \tag{4.1}$$

com $b'b = 1$, $b = [b_1 \; ... \; b_N]'$ e $V = [V_1 \; ... \; V_N]'$.
Sejam Σ a matriz de covariâncias e v_1, v_2, ..., v_N os vectores próprios estandardizados de dimensão N associados aos valores próprios λ_1, λ_2, ..., λ_N, respectivamente.
Os N vectores próprios que resultam da matriz de covariâncias constituem uma base em $\mathbb{R}^N$. Consequentemente, o vector b pode ser escrito como combinação linear dos vectores que constituem a base:

$$b = c_1 v_1 + c_2 v_2 + ... + c_N v_N. \tag{4.2}$$

Por (4.1) e pelo teorema da decomposição espectral, (ver Mardia [2], pag 469-470) resulta:

$$var(\alpha) = b'\Sigma b = b'(\sum_{i=1}^{N} \lambda_i v_i v_i')b.$$

Tendo em conta (4.2) e que $v_i'v_j = \begin{cases} 1 & se\ i = j \\ 0 & se\ i \neq j \end{cases}$ *vem,*

$$var(\alpha) = \sum_{i=1}^{N} \lambda_i b' v_i v_i' b = \sum_{i=1}^{N} \lambda_i (c_1 v_1' + c_2 v_2' + ... + c_N v_N') v_i v_i' (c_1 v_1 + c_2 v_2 + ... + c_N v_N) =$$

$$= \sum_{i=1}^{N} \lambda_i c_i c_i = \sum_{i=1}^{N} \lambda_i c_i^2.$$

De seguida provar-se-á que $\sum_{i=1}^{N} c_i^2 = 1$*:*

$$b'b = 1$$

$$(c_1 v_1' + c_2 v_2' + ... + c_N v_N')(c_1 v_1 + c_2 v_2 + ... + c_N v_N) = 1$$

$$c_1^2 v_1' v_1 + \ ... \ + c_N^2 v_N' v_N = 1$$

$$\sum_{i=1}^{N} c_i^2 = 1.$$

λ_1 *é o maior dos valores próprios. Portanto,*

$$var(\alpha) = \sum_{i=1}^{N} \lambda_i c_i^2 \leq \lambda_1 \sum_{i=1}^{N} c_i^2 = \lambda_1.$$

Como se pretendia.

♢

Corolário 4.3.1 *Seja* $\alpha = \beta' V$ *uma combinação linear estandardizada de* $V_1, V_2, ..., V_N$ *que tem variância máxima. Então,* $var(\alpha) = \lambda_1$ *e* $\beta = v_1$*, sendo* v_1 *o vector próprio associado a* λ_1*.*

Prova 4.3.2
Do seguimento da demonstração do Teorema 4.3.2 observa-se que se a variância de alguma componente for λ_1 *então essa componente tem variância máxima.*
Tendo em conta o Teorema 4.3.1 e propriedades da variância e das matrizes, tem-se que $var(v_1'V) = v_1'\Sigma v_1 = v_1'\lambda_1 I v_1 = \lambda_1 v_1' I v_1 = \lambda_1$*. Assim, a variância é máxima quando* $\beta = v_1$*.*

♢

Do Corolário 4.3.1 conclui-se que o vector a_1 da Definição 4.3.1 coincide com o vector v_1. Tendo em conta que as componentes são construídas de forma a não serem correlacionadas, de modo análogo, prova-se que $a_i = v_i$ para todo $i \in \{1, ..., N\}$. Para que as componentes principais não sejam correlacionadas entre si, o vector constituído pelos coeficientes da combinação linear das componentes (4.2) não pode ser escrito à custa dos vectores próprios que constituem os coeficientes da combinação linear das componentes até então contruídas. Portanto, o vector constituído pelos coeficientes da combinação linear da segunda componente não se escreve à custa de v_1, resultando que a variância máxima dessa componente é λ_2. Para as outras componentes o raciocínio é idêntico.

Propriedades 4.3.1
Sejam v_1, ..., v_N os vectores próprios de Σ associados aos N valores próprios ordenados $\lambda_1 > \lambda_2 > ... > \lambda_N$, respectivamente. Então,

- $var(CP_j) = \lambda_j$ *, com j=1, ..., N;*
- $\sum_{i=1}^{N} var(CP_i) = tr\Sigma$;
- $\prod_{i=1}^{N} var(CP_i) = |\Sigma|$.

Prova 4.3.3

- $var[CP_j] = var[v_j'V] = v_j'\Sigma v_j$.

  Pelo Teorema 4.3.1 $\Sigma v_j = \lambda_j I v_j$, *assim vem que:* $var[CP_j] = v_j'(\lambda_j I v_j) = \lambda_j(v_j' I v_j)$=$\lambda_j$.

- *Observando-se que o traço de uma matriz é a soma dos valores próprios dessa matriz, pela propriedade anterior fica evidente a igualdade.*

- *Observando-se que o determinante de uma matriz quadrada é igual ao produto dos valores próprios e novamente pela primeira propriedade obtém-se a igualdade.*

$\diamondsuit$

Assim, tendo em conta os resultados obtidos, o algoritmo de determinação das componentes baseia-se no cálculo matricial sobre a matriz de covariâncias. A partir da matriz de covariâncias calculam-se os valores e vectores próprios associados, determinando as componentes principais e as respectivas variâncias.

4.3.2 Decomposição da Variância

Seja v a matriz dos vectores próprios, $v = [v_1\ v_2\ ...\ v_N]'$ e CP o vector das componentes principais, $CP = [CP_1\ CP_2\ ...\ CP_N]'$. Observe-se que $CP = vV$.
Defina-se $var(CP)$ como a matriz de covariâncias do vector CP. Assim, $var(CP) = v\Sigma v'$. Observe-se que $var(CP)$ é uma matriz diagonal $N \times N$ em que a diagonal é constituída pelos N valores próprios. Assim, a variância da i-ésima componente principal corresponde ao i-ésimo elemento da diagonal da matriz $var(CP)$.

Definição 4.3.2 *O traço da matriz de covariâncias do vector CP, a soma de todos os valores da diagonal da matriz $var(CP)$, chama-se de variação total das componentes principais.*

Assim, atendendo às propriedades do traço de uma matriz, resulta que a variação total das componentes é igual à variação total das variáveis iniciais. De facto,

$$tr(var(CP)) = tr(v\Sigma v') = tr(\Sigma v'v) = tr(\Sigma) \tag{4.3}$$

Para averiguar o quanto a i-ésima componente explica a variação total, calcula-se naturalmente o quociente entre a variação da componente (variância) e a variação total,

$$\frac{\lambda_i}{\sum_{i=1}^{N} \lambda_i}.$$

Para averiguar o quanto as primeiras p componentes explicam a variação total, calcula-se o quociente entre a variação dessas p componentes e a variação total, Pe (ver equação 4.4). A este quociente chamar-se-á de percentagem de explicação das primeiras p componentes principais.

$$Pe = \frac{\sum_{i=1}^{p} \lambda_i}{\sum_{i=1}^{N} \lambda_i} \tag{4.4}$$

4.3.3 Critérios de Selecção das Componentes

O grande interesse da Análise em Componentes Principais é converter o conjunto das variáveis iniciais num menor número de variáveis (as componentes principais) que expliquem aproximadamente a variação total do conjunto das variáveis iniciais.
Se algumas variáveis iniciais forem linearmente dependentes entre si, no cálculo dos valores próprios obter-se-ão valores nulos. Assim, a matriz de covariâncias terá característica $m < N$ e a variação total do conjunto inicial será completamente explicado pelas m primeiras componentes principais.
Contudo, a dependência linear não é muito frequente, situação em que existem variáveis redundantes e por isso, sem interesse. O que é usual ocorrer, e situação de particular interesse para a aplicação de uma Análise em Componentes Principais, é a dependência linear "aproximada" entre algumas variáveis. Tal corresponde à existência de valores próprios "muito pequenos" em que o retirar das componentes associadas a esses valores não implica perda "significativa" de informação em termos de variação total. A vantagem de retirar componentes é essencialmente reduzir a dimensionalidade do espaço inicial das variáveis definindo um conjunto menor de variáveis e portanto, de interpretação, espera-se que, mais clara.
Porém põe-se o problema de escolher o número de componentes principais a considerar. A decisão do número de componentes depende da percentagem de explicação (Pe) desejada. Assim, existem vários critérios práticos, empíricos, que fazem a selecção do número de componentes a extrair, dos quais se apresentam dois dos mais usuais na literatura (ver, por exemplo, [21] e [12]):

- Critério: Faça-se a representação gráfica da percentagem de variação total explicada por cada componente e por ordem decrescente. Quando os valores representados ficarem aproximadamente sobre numa recta quase paralela ao eixo das abcissas, são de excluir as componentes correspondentes.

  Aquela representação gráfica é chamada de *scree plot*. A utilidade desse gráfico, na prática, justifica-se porque apresenta, na maior parte das vezes, uma nítida separação dos grandes valores próprios face aos pequenos.

- Critério: Incluir o número mínimo de componentes por forma a explicar pelo menos 70% da variação total ($Pe \geq 70\%$).

  A percentagem de variação total a considerar não é unanimamente escolhida por diversos autores. Na maior parte das vezes, é escolhido mediante o contexto em estudo. 70% é o limiar inferior normalmente utilizado. Todavia, alguns autores sugerem 90% (ver [12]). O valor de 70% pareceu ser o valor mais interessante perante a existência de um grande número de variáveis com é o caso em estudo (61 variáveis).

Assim, o algoritmo para cálculo das componentes principais pode ser descrito através dos seguintes passos:

1. Ler os dados e calcular a correspondente matriz de covariâncias das variáveis iniciais;

2. Encontrar os N valores próprios da matriz de covariâncias, proceder à sua ordenação, $\lambda_1 > \lambda_2 > ... > \lambda_N$, e cálcular os vectores próprios associados;

3. Escrever as componentes principais definido os coeficientes da i-ésima componente como sendo os elementos do i-ésimo vector próprio. Identificar a variância da i-ésima componente como sendo o i-ésimo maior valor próprio;

4. Desprezar as componentes cuja explicação da variação total seja pequena, por forma a garantir que as componentes consideradas completem a percentagem mínima de explicação (Pe) desejada.

4.3.4 Validação da Aplicação da Análise em Componentes Principais

Antes da aplicação da Análise em Componentes Principais deve-se averiguar se as variáveis não são correlacionadas entre si. No caso de não existir correlação a aplicação da Análise em Componentes Principais não tem interesse, pois as componentes principais coincidem com as variáveis iniciais.
Existem testes estatísticos para testar a não correlação das variáveis no âmbito da Análise em Componentes Principais. Por exemplo, o teste de esfericidade de Bartlett (ver [21] e [12]). Nele é testada a hipótese H_0 de os valores próprios serem todos iguais entre si.

A hipótese nula é dada por:

$$H_0\text{: } \lambda_1 = \lambda_2 = ... = \lambda_N.$$

A estatística de teste é dada por T:

$$T = -[n - 1 - \frac{1}{6}(2N + 5)] \sum_{i=1}^{N} \ln \lambda_i. \tag{4.5}$$

Esta estatística tem distribuição assimptótica de um qui-quadrado com $(\frac{N(N-1)}{2})$ graus de liberdade.

4.3.5 Rotação das Componentes Principais

As componentes principais são ordenadas de acordo com a importância decrescente da variância explicada. De forma geral, a primeira componente corresponde a um "factor geral" e é responsável pela maior parte da variação total. As outras componentes distribuem, entre si, a restante variação de acordo com a definição de componentes principais.
Todavia, existem técnicas associadas à rotação das variáveis que permitem redistribuir a variância das primeiras p componentes principais (o número de componentes através do qual se obtém a percentagem de explicação Pe desejada). O principal objectivo da aplicação da rotação às componentes principais é o de tentar encontrar um padrão na estrutura dos dados que seja "mais fácil de interpretar".

Seja,

$$A = \begin{bmatrix} a_{11} & \dots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{p1} & \dots & a_{pN} \end{bmatrix}$$

a matriz em que a i-ésima linha corresponde aos coeficientes da i-ésima componente principal, com p número de componentes retidas. O ideal da rotação no âmbito desta análise, é actuar sobre a matriz A transformando-a numa matriz com estrutura simplificada $B_{(p\times N)}$,

$$B = A \cdot G. \tag{4.6}$$

Diz-se que B é uma matriz com estrutura simplificada se:

- cada coluna tiver pelo menos um zero;
- cada linha tiver pelo menos p zeros;
- os coeficientes nulos de cada par de linhas correspondem a conjuntos distintos de variáveis.

Contudo, a obtenção de uma matriz com estrutura simplificada não se afigura de tarefa fácil. Os métodos conhecidos de rotação sobre a matriz A encontram a matriz B prenchendo apenas alguns requisitos de estrutura simplificada, mas não todos.
Os métodos de rotação mais simples são os ortogonais (mantêm a ortogonalidade entre os eixos), como o *QUARTIMAX*, o *VARIMAX* e o *EQUIMAX* todos eles implementados no *SPSS 7.5 for Windows*. Também é possível efectuar rotações oblíquas, como são exemplo os métodos *OBLIMIN* e *PROMAX*.
Nos métodos ortogonais encontra-se a matriz G (equação 4.6) tal que maximiza a seguinte soma:

$$\sum_{i=1}^{p} [\sum_{j=1}^{N} b_{ij}^4 - \frac{c}{N} (\sum_{j=1}^{N} b_{ij}^2)^2] \tag{4.7}$$

onde c é uma constante que varia conforme o método adoptado e b_{ij} os elementos da matriz

B.
O método mais popular é o *VARIMAX* com $c = 1$ (Kaiser-1958). Este método ortogonal resulta num pequeno número de componentes em que os coeficientes das variáveis são significativamente diferentes de zero. O resultado final das p componentes depois desta rotação é uma combinação das N variáveis iniciais, em que os coeficientes da combinação são escolhidos por forma a maximizar a variação entre o conjunto de valores "próximos" de zeros e o conjunto de valores "significativamente" diferentes de zero.

4.4 Aplicação ao Caso em Estudo

De seguida apresentar-se-ão os resultados da aplicação da Análise em Componentes Principais aos dados relativos aos resíduos ajustados e frequências relativas dos pares de codões não terminais das duas espécies de interesse.
Mas antes aplicar-se-á o teste de esfericidade de Bartlett. O número de variáveis (N) é 61 e o número de indivíduos (n) também é 61[1]. Assim, o número de graus de liberdade associados à estatística de teste T é 1830 sendo o valor do quantil de ordem 0.95 da distribuição do qui-quadrado de 1930.048. Os valores obtidos para a estatística de teste T (ver equação 4.5) são os que se apresentam na Tabela 4.1.

Espécie / Tipo de dado	T
Candida albicans / resíduos	5634.546
Candida albicans / frequências relativas	6963.612
Saccharomyces cerevisiae / resíduos	6832.452
Saccharomyces cerevisiae / frequências relativas	6991.646

Tabela 4.1: Valores observados para a estatística T no teste de esfericidade de Bartlett.

Da aplicação do teste vem que a hipótese das variáveis não estarem correlacionadas é rejeitada. Acrescentando ao estudo a rejeição da independência entre as variáveis, já que duas variáveis independentes não são correlacionadas.
Efectuando uma Análise em Componentes Principais, à matriz dos resíduos ajustados e à matriz das frequências relativas obtiveram-se os resultados da variação total das componentes principais (ver Tabelas 4.2 e 4.3) e os correspondentes *scree plots* (ver Figuras 4.1 e 4.2).

[1]Consideraram-se neste estudo apenas os codões não terminais, esta opcção é justificada pelo facto dos três codões terminais terem características muito distintas dos restantes codões em termos biológicos. No entanto foi feito o estudo considerando estes codões, obtendo-se o esperado distancimento dos codões terminais em relação aos restantes; resultado sem grande interesse no âmbito do estudo a desenvolver.

Total Variance Explained							
Sacch-resíduos					Sacch-fr		
	Initial Eigenvalues				Initial Eigenvalues		
Component	Total	% of Variance	Cumulative %	Component	Total	% of Variance	Cumulative %
1	694.9801827	29.53691299	29.53691299	1	0.000320126	31.17259317	31.17259317
2	393.0248271	16.70369949	46.24061248	2	0.000176717	17.20799939	48.38059256
3	323.7747721	13.76054672	60.0011592	3	0.000155126	15.10550539	63.48609795
4	173.1920926	7.360727538	67.36188674	4	6.55716E-05	6.385090861	69.87118881
5	139.3505528	5.92244967	73.28433641	5	5.17128E-05	5.035574968	74.90676378
6	83.41710374	3.545257543	76.82959395	6	4.82809E-05	4.701393025	79.60815681
7	80.09756729	3.404176025	80.23376998	7	3.04643E-05	2.966491625	82.57464843
8	77.79857327	3.306467935	83.54023791	8	2.74612E-05	2.67405434	85.24870277
9	54.48792713	2.315756914	85.85599483	9	2.1175E-05	2.061929738	87.31063251
10	42.85490397	1.821349157	87.67734398	10	1.95886E-05	1.907458811	89.21809132
11	37.04829229	1.574566028	89.25191001	11	1.55086E-05	1.510164026	90.72825535
12	26.70498257	1.134971566	90.38688158	12	1.17272E-05	1.141947121	91.87020247
13	24.13614417	1.025794991	91.41267657	13	9.29835E-06	0.905435418	92.77563789
14	21.73398009	0.923702135	92.3363787	14	7.83342E-06	0.762786143	93.53842403
15	18.30047231	0.777776794	93.1141555	15	6.8632E-06	0.668310178	94.20673421
16	15.086196	0.64116887	93.75532437	16	6.59191E-06	0.641892792	94.848627
17	14.43944027	0.613681514	94.36900588	17	5.35567E-06	0.521513268	95.37014027
18	13.81736011	0.587242879	94.95624876	18	5.11779E-06	0.49834886	95.86848913
19	11.99922895	0.509971637	95.4662204	19	4.45099E-06	0.433418567	96.30190769
20	9.996901526	0.424871986	95.89109239	20	3.84293E-06	0.374208964	96.67611666
21	9.24285925	0.392824913	96.2839173	21	3.17792E-06	0.309452632	96.98556929
22	8.477013676	0.360276194	96.64419349	22	2.97843E-06	0.290026953	97.27559624
23	7.580913937	0.322191626	96.96638512	23	2.92145E-06	0.284478835	97.56007508
24	7.047659118	0.299528101	97.26591322	24	2.57935E-06	0.251166925	97.811242
25	6.821870797	0.289932014	97.55584523	25	2.4478E-06	0.238356229	98.04959823
26	5.525924672	0.234853828	97.79069906	26	2.23047E-06	0.217193856	98.26679209
27	5.057368021	0.214939998	98.00563906	27	1.94638E-06	0.189530606	98.45632269
28	4.486448209	0.190675696	98.19631476	28	1.70968E-06	0.166481575	98.62280427
29	4.120632105	0.175128378	98.37144313	29	1.60974E-06	0.156750203	98.77955447
30	4.104152519	0.174427989	98.54587112	30	1.36415E-06	0.132834901	98.91238937
31	3.284216944	0.139580426	98.68545155	31	1.2871E-06	0.125332504	99.03772188
32	3.142235642	0.133546169	98.81899772	32	1.22236E-06	0.119028621	99.1567505
33	2.708732561	0.115122129	98.93411985	33	1.08778E-06	0.105923279	99.26267378
34	2.649122393	0.112588675	99.04670852	34	9.92153E-07	0.096611776	99.35928555
35	2.612519092	0.111033021	99.15774154	35	8.24825E-07	0.080318086	99.43960364
36	2.330462112	0.099045496	99.25678704	36	7.52148E-07	0.073241113	99.51284475
37	2.174617156	0.092422028	99.34920907	37	7.37491E-07	0.071813783	99.58465854
38	1.998415604	0.084933397	99.43414246	38	5.65291E-07	0.055045725	99.63970426
39	1.911303886	0.081231117	99.51537358	39	5.53295E-07	0.053877576	99.69358184
40	1.739369076	0.073923825	99.5892974	40	5.41715E-07	0.052749977	99.74633181
41	1.522761109	0.064717906	99.65401531	41	4.16202E-07	0.040528033	99.78685985
42	1.122964729	0.047726413	99.70174172	42	3.5845E-07	0.034904342	99.82176419
43	1.074148705	0.045651714	99.74739344	43	2.66488E-07	0.025949542	99.84771373
44	0.961567057	0.040866953	99.78826039	44	2.65507E-07	0.025853999	99.87356773
45	0.883364626	0.037543321	99.82580371	45	2.29927E-07	0.022389331	99.89595706
46	0.807222267	0.034307243	99.86011096	46	1.94699E-07	0.01895897	99.91491603
47	0.698539975	0.029688205	99.88979916	47	1.85055E-07	0.018019907	99.93293594
48	0.664717938	0.028250756	99.91804992	48	1.82998E-07	0.01781958	99.95075552
49	0.532117414	0.022615186	99.9406651	49	1.68119E-07	0.016370727	99.96712625
50	0.42695008	0.018145535	99.95881064	50	1.22766E-07	0.011954409	99.97908065
51	0.344696321	0.01464972	99.97346036	51	7.8241E-08	0.007618784	99.98669944
52	0.209931843	0.00892218	99.98238254	52	4.3822E-08	0.004267202	99.99096664
53	0.153321003	0.006516199	99.98889874	53	3.68629E-08	0.003589553	99.99455619
54	0.10435318	0.004435048	99.99333379	54	2.56323E-08	0.002495971	99.99705216
55	0.076203911	0.003238694	99.99657248	55	1.31816E-08	0.001283573	99.99833574
56	0.041944128	0.001782641	99.99835512	56	9.6717E-09	0.000941791	99.99927753
57	0.028136938	0.00119583	99.99955095	57	5.2705E-09	0.00051322	99.99979075
58	0.007794309	0.000331261	99.99988221	58	1.86284E-09	0.000181395	99.99997214
59	0.002731033	0.00011607	99.99999828	59	2.86003E-10	2.78498E-05	99.99999999
60	4.04167E-05	1.71772E-06	100	60	7.22542E-14	7.03582E-09	100
61	-7.90127E-15	-3.35807E-16	100	61	-5.5048E-23	-5.36034E-18	100

Tabela 4.2: Variação total explicada pelas componentes principais na *Saccharomyces cerevisiae*, quando se utiliza a matriz de resíduos ajustados (tabela do lado esquerdo) e a matriz das frequências relativas (tabela do lado direito) na Análise em Componentes Principais.

Total Variance Explained							
Candida-resíduos				Cand-fr			
	Initial Eigenvalues				Initial Eigenvalues		
Component	Total	% of Variance	Cumulative %	Component	Total	% of Variance	Cumulative %
1	2201.444294	33.48688879	33.48688879	1	0.000319968	31.20057987	31.20057987
2	1104.534701	16.8014384	50.2883272	2	0.000177281	17.28692228	48.48750215
3	677.8320152	10.31072436	60.59905156	3	0.00015326	14.94465183	63.43215398
4	556.4385021	8.464167951	69.06321951	4	6.56263E-05	6.399331008	69.83148499
5	353.7975244	5.381729798	74.44494931	5	5.18521E-05	5.056189207	74.88767419
6	273.3598707	4.158166353	78.60311566	6	4.81062E-05	4.690917803	79.578592
7	193.80147	2.947977513	81.55109317	7	3.08771E-05	3.010877319	82.58946931
8	173.4652485	2.638636599	84.18972977	8	2.74004E-05	2.671862508	85.26133182
9	149.5725655	2.275197188	86.46492696	9	2.1179E-05	2.065200213	87.32653204
10	122.3198479	1.860647192	88.32557415	10	1.93622E-05	1.888037897	89.21456993
11	101.5228724	1.5442976	89.86987175	11	1.50031E-05	1.462972795	90.67754273
12	83.94362846	1.276893973	91.14676572	12	1.18413E-05	1.154665532	91.83220826
13	69.57720321	1.058361582	92.20512731	13	9.36168E-06	0.912872989	92.74508125
14	59.62514402	0.906977557	93.11210486	14	7.98169E-06	0.778308069	93.52338932
15	55.76304309	0.848229877	93.96033474	15	6.94787E-06	0.677498115	94.20088743
16	46.63729523	0.709415143	94.66974988	16	6.58919E-06	0.642523307	94.84341074
17	41.17995728	0.626401792	95.29615167	17	5.48585E-06	0.53493397	95.37834471
18	35.04031719	0.533009719	95.82916139	18	5.16451E-06	0.503599817	95.88194453
19	29.43034556	0.447674607	96.276836	19	4.43195E-06	0.432166565	96.31411109
20	26.30222013	0.400091668	96.67692767	20	3.69537E-06	0.360341303	96.67445239
21	23.97362189	0.364670599	97.04159827	21	3.18512E-06	0.310586695	96.98503909
22	19.76392111	0.300635465	97.34223373	22	3.05554E-06	0.297951032	97.28299012
23	18.81718395	0.286234336	97.62846807	23	2.88253E-06	0.28108003	97.56407015
24	17.50979907	0.266347278	97.89481535	24	2.62656E-06	0.256120607	97.82019076
25	14.67970473	0.223297787	98.11811313	25	2.32424E-06	0.226640769	98.04683153
26	13.77402645	0.209521219	98.32763435	26	2.22167E-06	0.216638718	98.26347024
27	11.86664818	0.180507465	98.50814182	27	1.95522E-06	0.190656699	98.45412694
28	10.21852388	0.155437307	98.66357912	28	1.72076E-06	0.167794239	98.62192118
29	10.03178975	0.152596833	98.81617596	29	1.57908E-06	0.153978312	98.77589949
30	9.314145129	0.141680506	98.95785646	30	1.36346E-06	0.132952985	98.90885248
31	8.051183379	0.122469182	99.08032565	31	1.26111E-06	0.12297276	99.03182524
32	6.752160706	0.102709324	99.18303497	32	1.24221E-06	0.121129837	99.15295508
33	6.159118671	0.093688367	99.27672334	33	1.06238E-06	0.103594348	99.25654942
34	5.738687717	0.087293055	99.36401639	34	9.9649E-07	0.097169442	99.35371887
35	5.095636155	0.077511387	99.44152778	35	8.24158E-07	0.080365069	99.43408393
36	4.437491355	0.067500132	99.50902791	36	7.61812E-07	0.074285543	99.50836948
37	3.895084056	0.059249397	99.56827731	37	7.36739E-07	0.071840698	99.58021018
38	3.787465795	0.05761238	99.62588969	38	5.67322E-07	0.055320516	99.63553069
39	3.335285394	0.050734116	99.6766238	39	5.66148E-07	0.055206063	99.69073676
40	3.17159898	0.048244229	99.72486803	40	5.51968E-07	0.053823326	99.74456008
41	2.728661172	0.041506557	99.76637459	41	4.19275E-07	0.040884168	99.78544425
42	2.351158927	0.035764247	99.80213884	42	3.62505E-07	0.035348487	99.82079274
43	2.13315352	0.032448096	99.83458693	43	2.73843E-07	0.026702848	99.84749558
44	1.989893787	0.030268925	99.86485586	44	2.51983E-07	0.024571283	99.87206687
45	1.585358878	0.024115412	99.88897127	45	2.33177E-07	0.022737509	99.89480438
46	1.490836799	0.022677606	99.91164888	46	2.05477E-07	0.020036442	99.91484082
47	1.276936969	0.019423906	99.93107278	47	1.84606E-07	0.018001279	99.9328421
48	0.964687805	0.014674182	99.94574696	48	1.81317E-07	0.017680515	99.95052261
49	0.855272822	0.013009834	99.9587568	49	1.71553E-07	0.016728413	99.96725103
50	0.787987113	0.01198633	99.97074313	50	1.19065E-07	0.01161023	99.97886126
51	0.50384515	0.007664153	99.97840728	51	7.6577E-08	0.00746715	99.98632841
52	0.445608348	0.006778294	99.98518557	52	4.92839E-08	0.004805761	99.99113417
53	0.363509948	0.005529469	99.99071504	53	3.67391E-08	0.003582491	99.99471666
54	0.248882754	0.003785837	99.99450088	54	2.45712E-08	0.002395982	99.99711264
55	0.175813128	0.002674351	99.99717523	55	1.29973E-08	0.001267385	99.99838002
56	0.085528156	0.001300997	99.99847623	56	8.15241E-09	0.000794955	99.99917498
57	0.059366018	0.000903036	99.99937926	57	6.89181E-09	0.000672032	99.99984701
58	0.031806399	0.000483818	99.99986308	58	1.23934E-09	0.00012085	99.99996786
59	0.006541216	9.95006E-05	99.99996258	59	3.28276E-10	3.20108E-05	99.99999987
60	0.002459875	3.7418E-05	100	60	1.3068E-12	1.27428E-07	100
61	-2.40583E-14	-3.65958E-16	100	61	5.54583E-21	5.40783E-16	100

Tabela 4.3: Variação total explicada pelas componentes principais na *Candida albicans*, quando se utiliza a matriz de resíduos ajustados (tabela do lado esquerdo) e a matriz das frequências relativas (tabela do lado direito) na Análise em Componentes Principais.

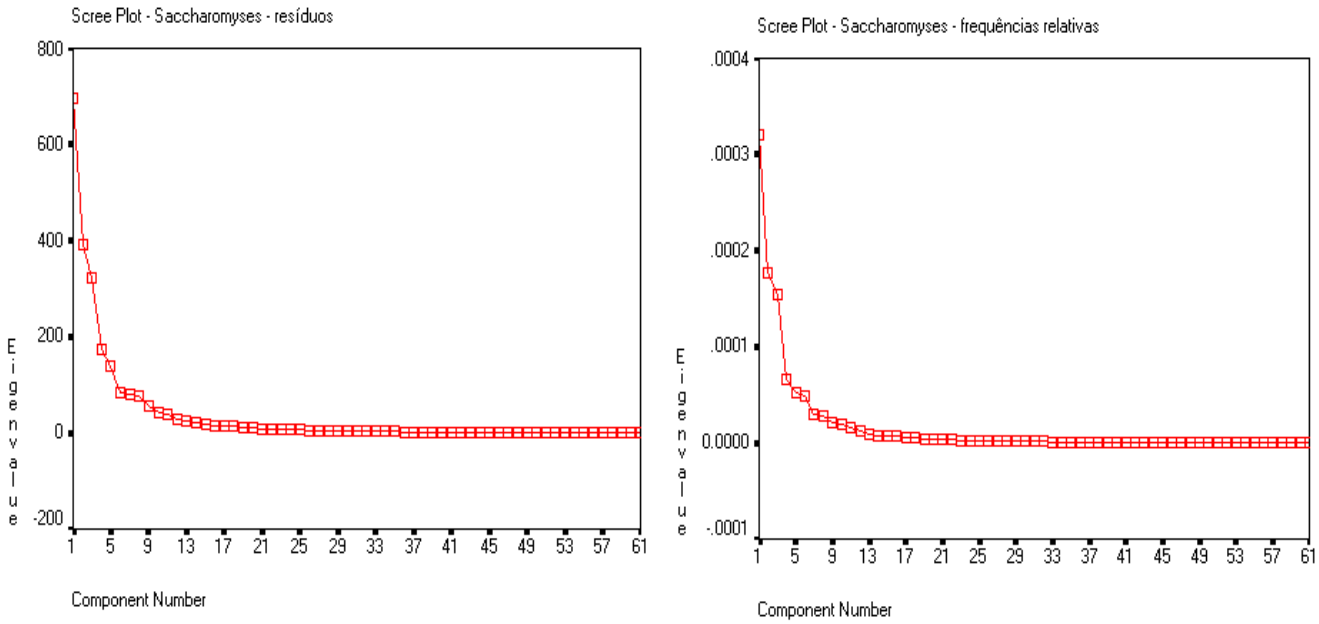


Figura 4.1: *Scree plot* da *Saccharomyces cerevisiae*, primeiro gráfico com base na matriz dos resíduos e segundo gráfico com base na matriz das frequências relativas.

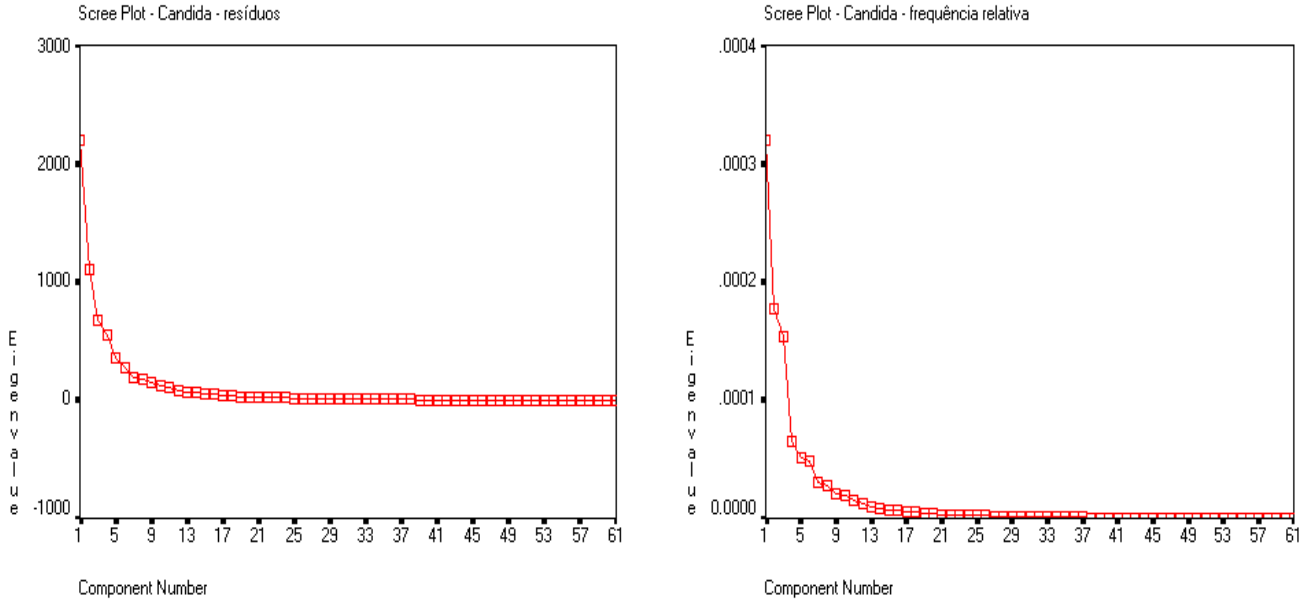


Figura 4.2: *Scree plot* da *Candida albicans*, primeiro gráfico com base na matriz dos resíduos e segundo gráfico com base na matriz das frequências relativas.

Tendo em conta os critérios que determinam a percentagem de variação explicada, o número de componentes a considerar são cinco. Seguem-se explicitados os coeficientes da combinação linear das variáveis iniciais que definem as 5 componentes principais que explicam $73,3\%$ e $74,4\%$ da variação total das variáveis na matriz dos resíduos na *Saccharomyces cerevisae* e *Candida albicans*, respectivamente (ver Tabela 4.4).

Saccharomyces	C1	C2	C3	C4	C5
AAA	-9.71216	-0.97609	-1.13907	-2.04466	-0.90311
AAC	-6.42911	-2.05314	0.123779	0.532781	0.357354
AAG	-10.2734	-1.90755	0.05517	-3.01056	-0.68347
AAT	-6.6315	-1.05179	0.06394	2.59715	0.355373
ACA	-2.89036	-0.83143	2.124102	1.833113	-1.49154
ACC	-1.42897	-4.30107	3.018222	-0.67619	-0.79229
ACG	-1.65813	-0.5164	1.226377	1.070312	-0.83742
ACT	-1.84881	-4.51091	4.448723	-0.31928	-1.2988
AGA	-6.61923	1.313006	-0.49243	1.309173	-0.79125
AGC	-3.65894	-0.88038	0.954305	2.324517	-0.02285
AGG	-4.49747	2.051158	-0.41143	1.135191	-0.20146
AGT	-4.10851	-1.28148	1.033411	2.816339	-0.39655
ATA	-2.9234	0.231126	-0.14391	1.931504	0.610169
ATC	-1.17617	-4.04511	1.123335	-1.66209	2.34019
ATG	-3.25803	-1.05311	0.692534	0.563894	0.57927
ATT	-1.28009	-4.98824	0.389469	-0.96243	2.007275
CAA	-0.13412	4.217841	0.62291	-5.33655	-1.50977
CAC	0.158521	1.358706	0.987074	-1.99635	1.088318
CAG	-1.99197	6.420558	-0.37653	-3.91601	-0.94498
CAT	0.561655	2.365758	1.059735	-2.11718	0.678235
CCA	2.231625	2.710818	2.900133	-1.05808	-2.57961
CCC	1.033439	2.019247	1.571754	-1.1217	-0.98661
CCG	1.843583	3.627185	0.522979	0.725284	-1.18622
CCT	2.461832	2.88176	1.253781	0.536847	-1.3302
CGA	0.455265	2.997385	0.582921	0.499135	-0.01044
CGC	-0.0766	1.151641	0.559624	-0.9421	0.278532
CGG	-0.44946	2.216243	-0.28543	-0.19617	0.0557
CGT	0.547391	0.450981	0.75338	-2.28036	1.204892
CTA	-1.22433	3.892536	-0.84536	-0.53557	-0.39167
CTC	0.715369	1.257366	0.221676	-1.0446	1.297235
CTG	-1.38964	4.353693	-1.01979	-0.46318	-0.21955
CTT	1.435358	1.551849	0.485464	-1.16006	0.61304
GAA	2.104126	-0.38106	-8.91243	1.288381	-1.10922
GAC	2.159347	-1.22364	-4.81343	0.690566	1.230333
GAG	-0.0864	1.189787	-6.12782	1.485556	-0.51295
GAT	3.261177	-1.3319	-5.31427	1.898674	0.909784
GCA	2.986949	0.198055	-1.10671	0.296642	-3.68214
GCC	4.124732	-2.90361	-1.0323	-1.86944	-1.9183
GCG	2.471434	0.701576	-1.23951	0.097109	-2.16855
GCT	4.861913	-3.93941	-0.9352	-2.04703	-3.10555
GGA	2.834387	0.147416	-0.83097	1.304349	-0.19445
GGC	2.239587	-0.78068	-1.28994	0.180555	0.161021
GGG	2.569617	-0.75738	-0.99921	1.238268	0.309076
GGT	4.904293	-6.12943	0.247454	-1.98764	0.613334
GTA	1.317148	0.464275	-0.46651	0.971655	-2.06715
GTC	3.016342	-3.35036	-0.90594	-1.46583	0.253593
GTG	2.167231	0.242188	-0.44815	0.027139	-1.64929
GTT	4.258889	-4.10806	-1.03844	-1.20516	0.328984
TAC	0.308468	1.719588	-0.56678	-1.39214	3.036543
TAT	1.473699	2.714195	-0.57367	0.664501	3.211509
TCA	3.09398	1.446451	4.114306	3.352991	-1.67145
TCC	3.559882	-0.61819	4.618715	0.825641	-0.22662
TCG	1.435909	1.737233	2.512852	2.186952	-0.72237
TCT	4.467283	-1.28517	5.546467	2.216179	-0.34977
TGC	0.837452	1.174423	0.870396	0.007963	1.960145
TGG	0.877094	1.016178	0.570751	0.986425	2.437829
TGT	1.267196	0.228856	1.366567	-0.13562	3.311754
TTA	4.200701	2.31967	1.717935	0.938144	0.632291
TTC	3.20792	1.255652	1.599102	-0.34351	3.230964
TTG	4.449821	0.806655	1.074602	-0.80631	1.355554
TTT	-0.2521	3.959924	0.333569	2.204206	1.595959

Candida	C1	C2	C3	C4	C5
AAA	12.1143	-9.41004	-4.25567	5.444561	-3.04716
AAC	15.82147	-4.86494	2.533375	-0.89608	3.32077
AAG	15.78147	-2.24169	1.658755	4.122618	-0.65331
AAT	11.66299	-7.93838	0.47602	-0.36782	2.850155
ACA	6.690921	3.153163	-5.90839	-4.42885	-3.14907
ACC	9.112747	-4.41161	-0.77242	-4.85159	0.381439
ACG	4.881506	0.250418	-2.5422	-0.45565	-0.82092
ACT	5.865563	-3.74386	-2.74285	-7.90438	-1.25915
AGA	3.377136	0.869655	-2.52432	-0.05879	-2.74053
AGC	5.000928	-0.81701	-0.63317	-0.87209	1.300786
AGG	3.133541	1.014625	-1.32385	1.980068	-0.37807
AGT	3.315041	-3.20516	-2.38061	-2.48508	-0.34789
ATA	1.773712	3.117873	-3.5202	0.669601	-0.77608
ATC	3.464914	0.062104	-0.08532	1.222119	0.340274
ATG	3.424278	0.850626	-2.68391	0.475566	-0.67869
ATT	3.831995	-2.68664	-3.65758	-0.29187	-0.60211
CAA	0.545046	7.252634	14.88668	-2.35439	-6.51692
CAC	5.008987	-0.31019	4.861983	-0.23926	0.716433
CAG	4.523238	4.368064	7.193972	1.20564	-3.74156
CAT	0.089251	-1.8162	6.160373	-3.18192	1.416659
CCA	3.142186	7.150827	1.57569	-6.17646	-4.29299
CCC	3.974544	0.013942	1.241979	-2.1598	1.789028
CCG	2.264237	2.913017	-0.08289	-1.37414	-0.85087
CCT	0.443395	1.455764	1.955634	-5.46324	0.874311
CGA	-1.85702	1.958991	4.059484	-1.7373	1.823604
CGC	1.325181	0.721579	1.805955	-0.35786	0.693821
CGG	0.39706	1.51851	1.430355	0.60516	0.667359
CGT	-1.38718	-0.0342	3.759718	-2.10787	1.364076
CTA	0.889974	4.956279	1.153434	3.212223	0.019694
CTC	1.584181	1.19418	1.232864	1.15253	0.848991
CTG	0.862773	4.275802	2.234915	1.753947	-2.11091
CTT	-0.93431	1.507082	1.152303	0.501369	1.846753
GAA	-10.7673	-3.23623	2.879901	6.659608	-3.41951
GAC	-1.31128	-4.21338	1.966929	2.823746	1.559725
GAG	0.277791	-0.46256	1.615708	5.839758	-0.86145
GAT	-10.8963	-8.28511	3.694881	3.452937	1.474243
GCA	-4.0242	4.680569	-5.65173	2.147871	-5.39755
GCC	-7.5246	-4.74905	-0.64948	-0.33615	-0.36103
GCG	-1.53627	0.703496	-1.67132	1.786238	-1.69523
GCT	-12.4054	-5.00213	-1.55774	-3.61221	-3.73008
GGA	-8.24787	-2.2708	-0.93189	-1.15976	0.151213
GGC	-4.2393	-3.21618	-0.13001	-0.45738	0.678177
GGG	-4.21978	-2.35877	-1.90339	0.04901	-0.60149
GGT	-12.7296	-9.2029	-0.53388	-5.25789	-2.16078
GTA	-2.73446	5.085121	-4.02788	1.718384	-2.28634
GTC	-4.43139	-2.21008	-0.47405	0.656968	0.220443
GTG	-2.34275	2.6457	-3.5885	2.370991	-2.64597
GTT	-9.93954	-2.77435	-2.36434	0.915569	-1.25242
TAC	1.902315	0.937529	0.9881	2.429899	1.663284
TAT	-4.71274	0.878853	2.447739	2.394043	3.346968
TCA	-2.32172	8.584624	-3.16148	-5.18726	2.294126
TCC	-1.38204	-1.03353	0.620468	-1.92705	3.369316
TCG	2.178009	4.024794	-2.79893	0.031926	0.932829
TCT	-4.86314	2.307308	-2.06369	-4.47452	3.049602
TGC	-0.62572	0.128534	0.37179	0.594099	2.027201
TGG	-3.78502	-0.30997	1.697036	-1.06775	3.000849
TGT	-3.53575	-0.04726	0.579729	-1.01689	3.144337
TTA	-6.85414	10.83538	-2.24947	-1.0482	4.513733
TTC	-6.37704	0.898037	0.420151	1.421608	4.659159
TTG	0.546227	4.293244	0.747354	5.328479	2.266272
TTT	2.438224	10.36434	-4.63081	4.535542	3.372147

Tabela 4.4: As 5 primeiras componentes principais para as duas espécies, no âmbito do estudo dos resíduos.

Se as variáveis forem suficientemente correlacionados com poucas componentes consegue-se extrair a estrutura de agrupamento dos indivíduos. Tendo em conta as conclusões obtidas na Análise Classificatória (Capítulo 3), passou-se às representações gráficas dos indivíduos (linhas) em função das novas variáveis (componentes principais), considerando nos eixos as componentes principais duas a duas e dividindo o conjunto dos codões (indivíduos) em 4 grupos de símbolos de acordo com a sua composição a nível das bases (Figura 4.3).

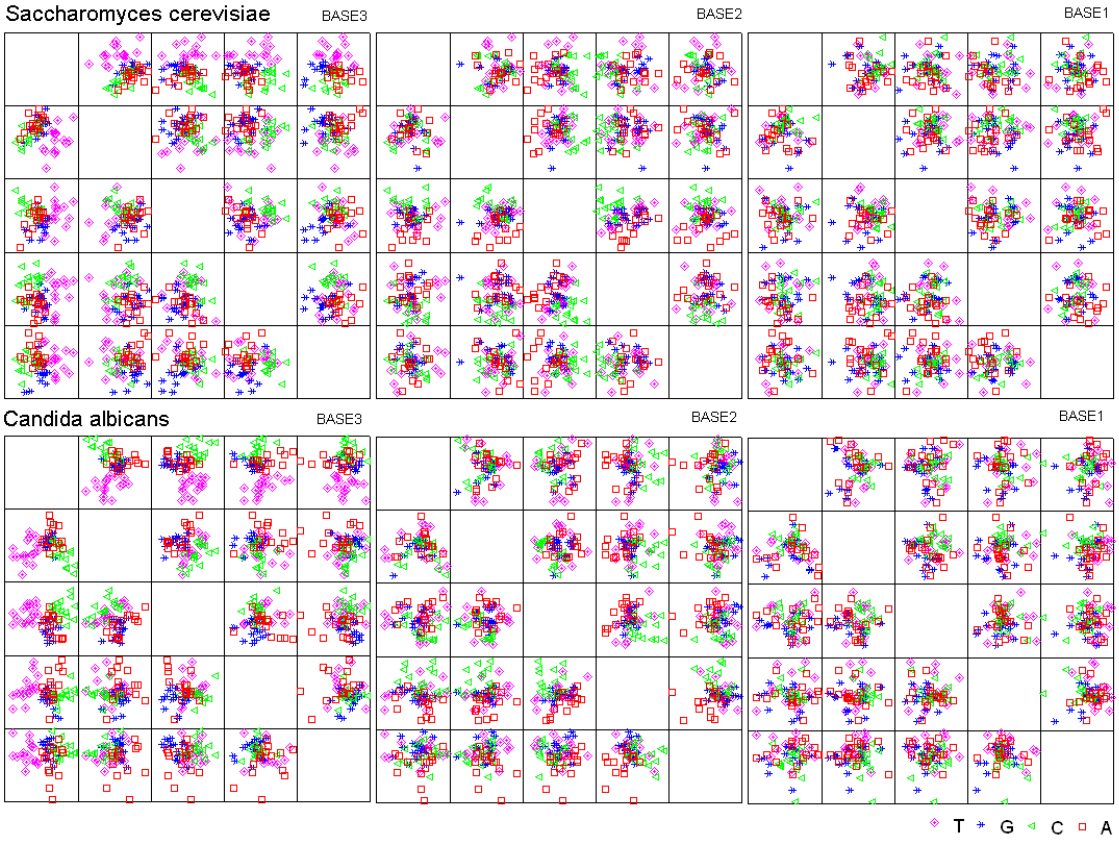
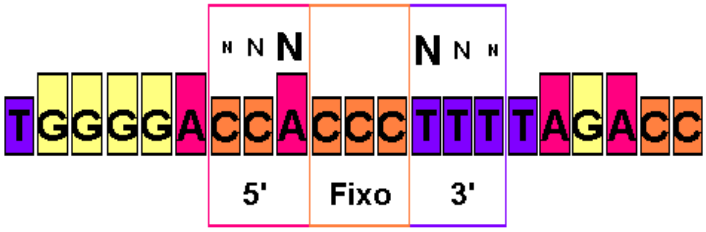


Figura 4.3: Representações gráficas dos codões, referenciando a forma do nucleótido terminal (base3), médio (base2) e inicial (base1), para as espécies *Saccharomyces cerevisiae* e *Candida albicans*.

Da observação desses gráficos averigua-se que existe uma tendência para separar os codões em 4 grupos de acordo com o nucleótido de terminação do codão. Observa-se que a formação de grupos de símbolos é gradualmente menos definida caso se considere a marcação do nucleótido final, intermédio ou inicial. Esta análise vem assim vincar a ideia de que:
Dado um codão fixo, a distribuição da sequência de codões justapostos é influenciada pelos três nucleótidos que compõem esse codão fixo, dando uma maior influência com a proximidade do símbolo-nucleótido ao codão fixo.
Por fim aplicou-se a rotação ortogonal VARIMAX. Contudo os resultados obtidos com a ro-

tação não alteram significativamente a representação das variáveis e a nível da descoberta de padrão não parece existir avanço de mais informação.
Da Análise das Tabelas de Contingência verificou-se, através dos coeficientes de associação dos nucleótidos do codão justaposto, que o menor valor de associação se observa para o nucleótido intermédio. As três metodologias estatísticas até agora consideradas são unânimes em concluir que, dado um codão, o nucleótido mais próximo do codão justaposto é o nucleótido que lhe é mais associado face aos restantes dois nucleótidos.
Assim, das análises realizadas, é de crer que:
Dado um codão fixo a preferência pelo codão que lhe sucede poderá ser determinada pelo codão anterior em que os três nucleótidos constituíntes desse codão têm níveis de influência eventualmente distintos.
De um modo esquemático, os níveis de influência podem ser representados pelos tamanhos da letra N sobre cada nucleótido, como ilustra a seguinte figura:



Capítulo 5

Cadeias de Markov

5.1 Introdução

Dado um codão fixo numa sequência de codões de um determinado gene, pensa-se que o posicionamento desse codão poderá ser explicado, em parte, pelo tipo de codões que se encontram nas posições anteriores adjacentes e que o "peso de explicação" estará relacionado com as distâncias desses codões ao codão fixo. Pressupondo que o codão anterior adjacente tem maior "peso da explicação", parece natural a modelação do sequenciamento dos codões por modelos probabilísticos Markovianos. Neste capítulo averiguar-se-á a adequação do sequenciamento de codões no genoma por uma cadeia de Markov.
Para tal comparar-se-ão as frequências observadas de cada um dos codões e as respectivas estimativas das probabilidades sob a validade de um modelo de cadeia de Markov homogénea e estritamente estacionária.

5.2 Cadeia de Markov com Espaço de Parâmetros Discreto

5.2.1 Definição de Cadeia de Markov

Antes de se aplicar as Cadeias de Markov ao estudo das sequências de genes, apresentar-se-á um conjunto de definições e conceitos a utilizar nesse estudo.

Definição 5.2.1 *Um processo estocástico* $\{X_n, n \in \mathbb{N}_0\}$ *é uma cadeia de Markov com espaço de parâmetros discreto e espaço de estados finito,* $E = \{C_1, ..., C_N\}$*, se:*

$$\forall_{n\in\mathbb{N}_0} \quad \forall_{C_j\in E} \qquad P(X_{n+1} = C_j | X_0, ..., X_n) = P(X_{n+1} = C_j | X_n). \tag{5.1}$$

Em particular, se as probabilidades dadas em (5.1) não dependem de n diz-se que a cadeia de Markov é *homogénea* no tempo e pode-se escrever:

$$P(X_{n+1} = C_j | X_n = C_i) = p_{ij} \tag{5.2}$$

Assim, a probabilidade p_{ij} é a probabilidade do sistema transitar do estado C_i para C_j num passo.

Chama-se *matriz das probabilidades de transição da cadeia*, à matriz $N \times N$,

$$P = [p_{ij}]_{\ i=1,\ 2,\ ...,\ N\ e\ j=1,\ 2,\ ...,\ N}.$$

A partir desta altura quando se escreve cadeia de Markov, entender-se-á cadeia de Markov homógenea, cujas probabilidades de transição satisfazem as seguintes condições:

1. $p_{ij} \geq 0$, com $i, j \in \{1,\ 2,\ ...,\ N\}$;
2. $\sum_{j=1}^{N} p_{ij} = 1$, com $i \in \{1,\ 2,\ ...,\ N\}$.

Um resultado importante, neste contexto, é o facto da distribuição de probabilidades da cadeia de Markov $\{X_n, n \in \mathbb{N}_0\}$ ficar completamente especificada com o conhecimento da matriz P e da distribuição de X_0 (chamada *distribuição inicial da cadeia*).

Teorema 5.2.1 *Dada a matriz das probabilidades de transição de uma cadeia de Markov $\{X_n, n \in \mathbb{N}_0\}$, P, e a distribuição de probabilidades da variável aleatória X_0, com $P(X_0 = C_i) = a_{(C_i)}$. Então:*

$$P(X_0 = C_{i_0},\ X_1 = C_{i_1},\ ...,\ X_m = C_{i_m}) \ = \ a_{(C_{i_0})}\, p_{i_0 i_1}\, ...\, p_{i_{m-1} i_m},$$

para todo $m \in \mathbb{N}_0$ e $C_{i_0},\ C_{i_1},\ ...,\ C_{i_m} \in E$.

Definição 5.2.2 *A probabilidade do processo, $\{X_n, n \in \mathbb{N}_0\}$, chegar ao estado C_j em n passos tendo partido do estado C_i é denotada por p_{ij}^n a que se chama de probabilidade de transição em n passos e tem-se que:*

$$p_{ij}^n = P(X_{n+m} = C_j | X_m = C_i)$$

com $m \in \mathbb{N}_0$ e $C_i, C_j \in E$.

Definição 5.2.3

- *Um estado C_j é acessível a partir do estado C_i se $p_{ij}^n > 0$ com $n \in \mathbb{N}_0$ e $i, j \in \{1, 2, ..., N\}$;*
- *Dois estados comunicam entre si se são acessíveis a partir um do outro;*
- *Se todos os estados comunicam entre si a cadeia de Markov diz-se irredutível.*

Definição 5.2.4 *O período de um estado C_i é o máximo divisor comum ($d(C_i)$) de todos os inteiros $n \geq 1$, para os quais $p_{ii}^n > 0$ (se $p_{ii}^n = 0$ para $n \geq 1$ convencionou-se $d(C_i) = 0$). Quando $d(C_i) = 1$ a cadeia de Markov é chamada de aperiódica.*

5.2.2 Comportamento Limite das Cadeias de Markov

Dada uma cadeia de Markov com matriz das probabilidades de transição P chama-se *distribuição estacionária* à distribuição de probabilidades $\pi = \{\pi_i, i = 1,\ 2,\ ...,\ N\}$, que satisfaz a condição,

$$\pi_k = \sum_{j=1}^{N} \pi_j p_{jk}, \quad \forall k \in \{1, 2, ..., N\}.$$

Na forma matricial tem-se que:

$$\Pi = \Pi \cdot P \tag{5.3}$$

onde Π representa o vector $[\pi_1\ \ ...\ \ \pi_N]$.

Repare-se que da condição (5.3) resulta recursivamente que: $\Pi = (\Pi \cdot P) \cdot P = \Pi \cdot P^2 = = (\Pi \cdot P) \cdot P^2 = ... = \Pi \cdot P^n$, ou seja:

$$\pi_k = \sum_{j=1}^{N} \pi_j p_{jk}^n, \quad \forall k \in \{1, 2, ..., N\}, \quad \forall n \in \mathbb{N}. \tag{5.4}$$

No contexto das Cadeias de Markov são conhecidos vários resultados relativos ao comportamento limite da cadeia. Porém, aqui apenas se apresentará um resultado, útil na aplicação do estudo a desenvolver.

Teorema 5.2.2 *Numa cadeia de Markov homogénea, irredutível e aperiódica, com um número finito de estados (N) e distribuição estacionária $\pi = \{\pi_1, ..., \pi_N\}$, o sistema de $N+1$ equações*

$$\begin{cases} \pi_k = \displaystyle\sum_{j=1}^{N} \pi_j p_{jk} \quad k = 1,\ 2,\ ...,\ N \\ \displaystyle\sum_{j=1}^{N} \pi_j = 1 \end{cases} \tag{5.5}$$

tem solução única, estritamente positiva e tal que:

$$\pi_k = lim_{n \to \infty} p_{ik}^n,\ \forall i \in \{1, 2, ..., N\},\ com\ k = 1, 2, ..., N.$$

Em geral, o resultado anterior, é apresentado na literatura sob um aspecto mais geral, com espaço de estados infinito numerável. A prova pode ser encontrada em [22], por exemplo.
Se a distribuição de probabilidades inicial coincidir com a distribuição estacionária $\{\pi_k,\ k = 1, ..., N\}$, resulta que:

Para todo $m \in \mathbb{N}_0$ e $C_{i_0},\ ...,\ Ci_m \in E$ pelo Teorema 5.2.1

$$P(X_0 = C_{i_0}, X_1 = C_{i_1}, ..., X_m = C_{i_m}) = \pi_{i_0} p_{i_0 i_1} ... p_{i_{m-1} i_m},$$

$$P(X_n = C_{i_0}, ..., X_{n+m} = C_{i_m}) = P(X_n = C_{i_0}) p_{i_0 i_1} ... p_{i_{m-1} i_m},\ \forall_{n \in \mathbb{N}_0}.$$

Dada a hipótese de que a probabilidade inicial coincide com a distribuição estacionária,

$$P(X_n = C_{i_0}) = \sum_{j=1}^{N} a_{(C_j)} p^n_{ji_0} = \sum_{j=1}^{N} \pi_{i_0} p^n_{ji_0} = \pi_{i_0},$$

pelo que os vectores $(X_0,\ ...,\ X_m)$ e $(X_n,\ ...,\ X_{n+m})$ são identicamente distribuídos.

Ao processo, $\{X_n, n \in \mathbb{N}_0\}$ que verifica esta condição designa-se processo estocástico estritamente estacionário.

5.3 Aplicação ao Caso em Estudo

Apresentar-se-á uma aplicação ao caso dos codões, pelo que se tem um total de 64 indivíduos[1], correspondendo aos 64 elementos do espaço de estados (E).
No sentido de estabelecer um modelo probabilístico para o sequenciamento dos codões na *Saccharomyces cerevisiae* e na *Candida albicans*, e dada a natureza do problema propõe-se para cada uma das espécies uma cadeia de Markov para o processo $\{X_n, n \in \mathbb{N}_0\}$, onde X_n representa o codão na n-ésima posição no sequenciamento dos codões no genoma da espécie em causa.
Para definir esta cadeia alguns pressupostos serão admitidos. Nomeadamente,

- a estacionaridade do processo;
- o sequenciamento continuado dos codões sobre todos os genes aleatoriamente sequenciados.

Para especificar a distribuição de probabilidades da cadeia de Markov a propor interessará estudar a matriz e a distribuição inicial da cadeia em estudo.
Tem-se como ponto de partida deste trabalho as tabelas de contingência, consideradas no Capítulo 2 e que se encontram no Apêndice B.1, e as frequências de cada um dos codões no

[1] Também se efectuou uma abordagem para o caso dos aminoácidos em que se obtiveram resultados semelhantes aos que serão apresentados no contexto dos codões.

genoma.
Como resultado teórico utilizou-se o Teorema 5.2.2, para o qual é necessário ter uma matriz das probabilidades de transição irredutível e aperiódica. Não especificando a estratégia, o Teorema 5.2.2 poderá permitir estimar as probabilidades dos símbolos (codões) como as probabilidades limite (o conjunto solução do sistema perante a estacionaridade).
Para cada uma das espécies e a partir da tabela de contingência dos pares de codões pode-se obter uma matriz cuja soma das linhas é 1, através do cálculo de frequências relativas. Contudo, deste modo não resultaria uma matriz com a forma específica da matriz do Teorema 5.2.2 desejável para este estudo.
Por um lado, a matriz de frequências absolutas tem 3 linhas de zeros, relativos aos codões terminais[2], consequentemente não levará a uma matriz das probabilidades de transição irredutível. Por outro lado, tendo em conta a Definição 5.2.4 os codões terminais provocariam que a cadeia não fosse aperiódica uma vez que:

$$d(C_{terminal}) = 0 \ (\neq 1).$$

Repare-se que o codão de iniciação é também motivo da não aplicabilidade do Teorema 5.2.2:

$$P(X_1 = Cj | X_0 \neq ATG) = 0$$

$$P(X_{n+1} = Cj | X_n \neq ATG) \neq 0$$

com $n \in \mathbb{N}$ e $C_j \in E$.
Assim, as sequências de codões não constituem uma cadeia de Markov homogénea.
Como consequência da estrutura da sequência de codões, propôs-se então que os dados a considerar sejam as frequências de pares do sequenciamento dos codões não terminais e não iniciais[3]. Como resultado desta alteração tem-se novos totais marginais, a partir dos quais se calculou a matriz, P^*, de frequências relativas face aos totais marginais das linhas da matriz (a soma em cada linha é 1). A matriz obtida, P^*, tomou-se como uma estimativa da matriz das probabilidades de transição, a qual é irredutível e aperiódica, ou seja, está nas condições do Teorema 5.2.2. Aplicou-se o Teorema 5.2.2 e obteve-se a solução do sistema. O algoritmo para a determinação da solução do sistema (5.5) foi implementado no *MATLAB 6*, fazendo uso das potencialidades de cálculo matricial e encontra-se no Apêndice A.2.
Estima-se a distribuição inicial da cadeia como sendo a distribuição estacionária do processo. Tal implica que se assuma que o processo do sequenciamento de codões dentro de cada gene entre a posição inicial e a posição final, excluindo estas duas, seja estritamente estacionário. Assim, as probabilidades de cada estado do processo serão dadas pelas probabilidades limite, a solução do sistema.
Resumindo, o modelo teórico proposto para o sequenciamento de codões nos genes (excluindo a posição inicial e terminal) é definido por uma cadeia de Markov com matriz de probabilidades de transição P^* e distribuição inicial Π coincidente com a distribuição estacionária da cadeia.
No sentido de analisar a adequação do modelo teórico aos dados proceder-se-á a uma comparação entre frequências observadas de cada codão com a sua probabilidade estimada assumindo o modelo teórico acima descrito.
Da representação gráfica conjunta das frequências observadas e das probabilidades estimadas de cada codão, Figuras 5.1 e 5.2, observa-se relativa proximidade entre estes valores.

[2] Um codão terminal de um dado gene não tem influência sobre o codão inicial de outro gene.

[3] Atenção que o ATG continua na constituição das sequências sem codões terminais nem iniciais. Eliminou-se o codão de iniciação em cada gene a fim de evitar a não homogenea da cadeia.

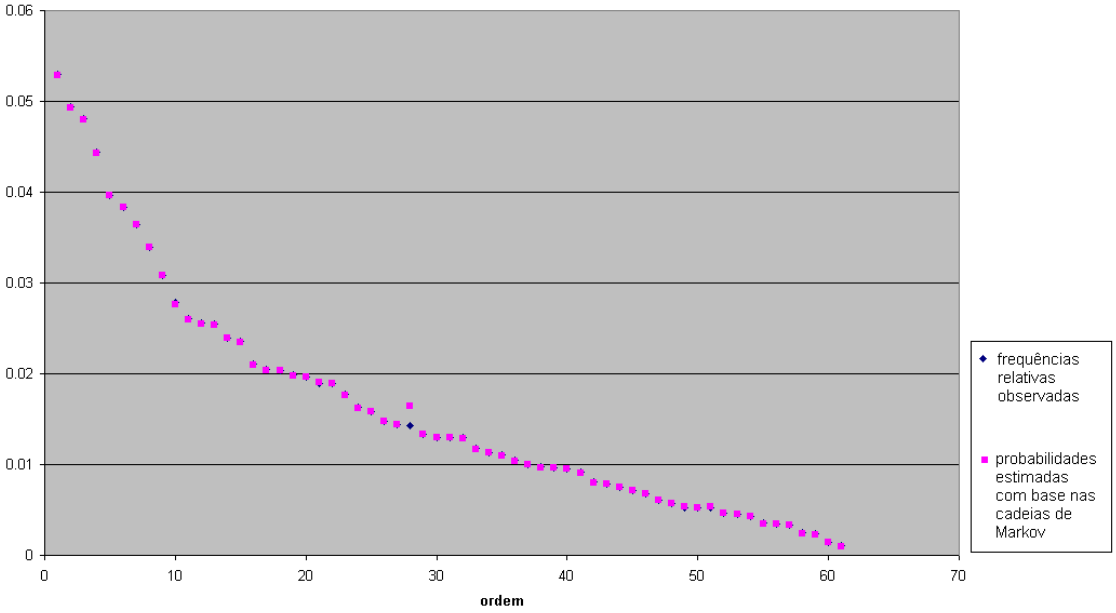


Figura 5.1: Frequências relativas/probabilidades para os codões na *Candida albicans*.

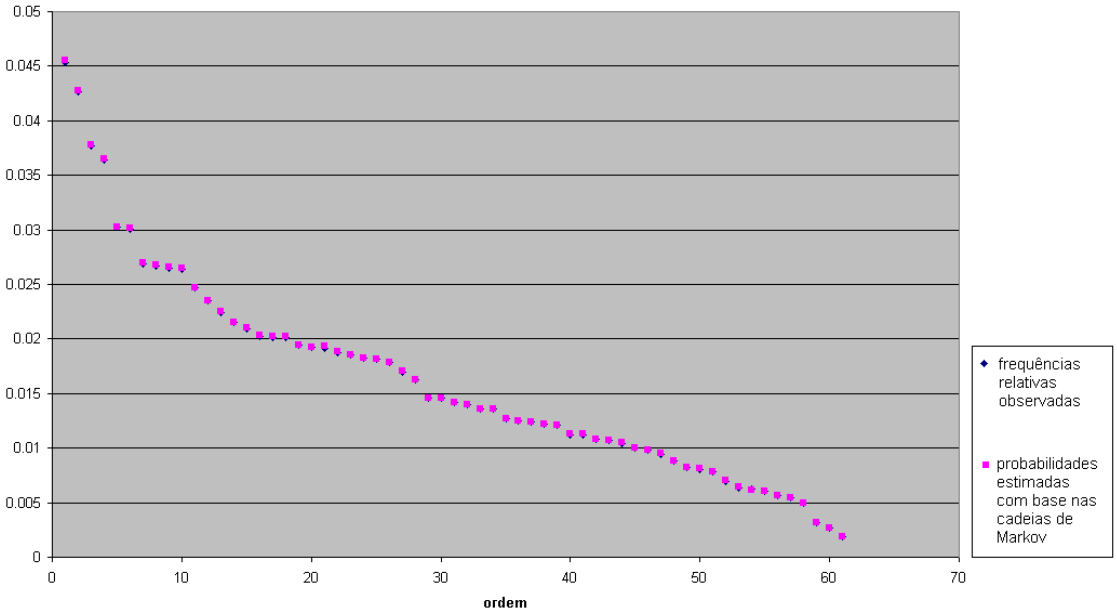


Figura 5.2: Frequências relativas/probabilidades para os codões na *Saccharomyces cerevisiae*.

Nesta altura será de questionar se as diferenças observadas são significativamente diferentes. Por outras palavras, testar a seguinte hipótese nula:

$$H_0 : \pi_j = \frac{n_j^*}{n},$$

com n_j^* a frequência observada do codão C_j no conjunto de todos os genes e dentro de cada gene entre o codão inicial e o codão terminal, excluindo estes dois, n o número total de observações, N o número de indivíduos e $j \in \{1, ..., N\}$.
Da aplicação de testes de ajustamento do qui-quadrado obtiveram-se os seguintes resultados:

Espécie	χ^2	df	valor do quantil 0.95
Candida albicans	938.2207	60	79.08
Saccharomyces cerevisiae	32.1871	60	79.08

Tabela 5.1: Resultados da estatística χ^2 e valores teóricos.

Com estes resultados, para o caso de *Candida albicans*, conclui-se que existe diferenças significativas entre os dados observados e a probabilidade estimada assumindo o modelo teórico, uma vez que para 60 graus de liberdade[4], o valor teórico do quantil de ordem 0.95 da distribuição do qui-quadrado, é de aproximadamente 80 e o valor da estatística é muito superior. O mesmo já não acontece com a *Sacchamyces cerevisiae*, já que as diferenças não são significativas.
No sentido de identificar os responsáveis pelo desajuste na espécie *Candida albicans*, observaram-se as parcelas constituintes do qui-quadrado relativas a cada codão (ver Tabela 5.2), por outras palavras observaram-se os quadrados dos resíduos de Pearson.
Na Tabela 5.2 os valores das células da coluna *p-CM* referem-se aos valores esperados assumindo a cadeia de Markov, os valores da coluna *fr* são as frequências relativas observadas e os valores da coluna *teste* as parcelas da estatística de teste ou os quadrados dos resíduos de Pearson.
Da análise dos resultados da Tabela 5.2, conclui-se que o grande responsável pelo desajuste existente na espécie *Candida albicans* é o ATG. Observa-se menos vezes o codão ATG do que seria esperado se se assumisse o modelo teórico. Sabe-se que o codão ATG é o codão de iniciação da parte codificada da sequência de cada gene, mas também surge ao longo do sequenciamento como codão não terminal. No entanto, foram retirados das contagens todos os codões de iniciação, averiguando-se nestas condições que o ATG é o grande responsável pelo desajuste do modelo teórico.
Para o caso da *Saccharomyces* é aceite a hipótese nula, existe ajuste. Portanto, não será de afastar a hipótese de que o modelo considerado de cadeia de Markov explique a sequência do conjunto dos codões nesta espécie.
De modo geral pode-se concluir:

A cadeia de Markov proposta adequa-se às frequências de quase todos os codões no genoma.

[4]O número de graus de liberdade é $60 = 61 - 1$, ver Capítulo 2.

aminoácido/ codão	ORDEM	fr	p-CM	teste
LYS AAA	1	0.053004	0.0529	0.683753
GLU GAA	2	0.049375	0.0493	0.389849
ASN AAT	3	0.048083	0.048	0.488889
ASP GAT	4	0.044391	0.0443	0.634674
ILE ATT	5	0.039625	0.0396	0.054638
LEU TTA	6	0.038303	0.0383	0.000649
GLN CAA	7	0.03644	0.0364	0.145393
LEU TTG	8	0.033906	0.0339	0.003775
PHE TTT	9	0.030851	0.0308	0.286987
SER TCA	10	0.027822	0.0276	6.007761
VAL GTT	11	0.026059	0.026	0.44449
THR ACT	12	0.025611	0.0255	1.616464
TYR TAT	13	0.025427	0.0254	0.094869
PRO CCA	14	0.023936	0.0239	0.182992
GLY GGT	15	0.023543	0.0235	0.260168
ALA GCT	16	0.021062	0.021	0.627202
SER TCT	17	0.020528	0.0204	2.718493
ARG AGA	18	0.020402	0.0204	0.000613
THR ACA	19	0.019895	0.0198	1.525802
LYS AAG	20	0.019594	0.0196	0.006807
ASN AAC	21	0.018978	0.019	0.08659
ILE ATA	22	0.018908	0.0189	0.010438
SER AGT	23	0.01769	0.0176	1.562172
ALA GCA	24	0.016251	0.0162	0.538528
HIS CAT	25	0.015831	0.0158	0.205645
PHE TTC	26	0.01478	0.0148	0.093851
GLY GGA	27	0.014432	0.0144	0.234409
MET ATG	28	0.014327	0.0164	883.8204
ILE ATC	29	0.01333	0.0133	0.23127
GLU GAG	30	0.01303	0.013	0.226771
PRO CCT	31	0.013021	0.013	0.117298
ASP GAC	32	0.012923	0.0129	0.13984
THR ACC	33	0.011736	0.0117	0.37992
LEU CTT	34	0.011295	0.0113	0.006229
VAL GTG	35	0.011021	0.011	0.13001
VAL GTA	36	0.01048	0.0104	2.080897
TRP TGG	37	0.009991	0.01	0.026117
ALA GCC	38	0.009764	0.0097	1.428439
TYR TAC	39	0.009605	0.0096	0.009423
CYS TGT	40	0.009535	0.0095	0.440122
SER TCC	41	0.009145	0.0091	0.761695
VAL GTC	42	0.008063	0.008	1.68228
GLN CAG	43	0.007869	0.0079	0.411066
GLY GGG	44	0.007495	0.0075	0.013544
SER TCG	45	0.007181	0.0071	3.126746
LEU CTA	46	0.00683	0.0068	0.440355
ARG CGT	47	0.006066	0.0061	0.637151
HIS CAC	48	0.00572	0.0057	0.238063
LEU CTG	49	0.005278	0.0053	0.30799
SER AGC	50	0.005262	0.0052	2.468793
ARG CGA	51	0.005253	0.0053	1.417794
PRO CCC	52	0.004678	0.0047	0.340594
GLY GGC	53	0.004516	0.0045	0.192791
THR ACG	54	0.004316	0.0043	0.191258
LEU CTC	55	0.003533	0.0035	1.042992
ARG AGG	56	0.003409	0.0034	0.079788
PRO CCG	57	0.003341	0.0033	1.724287
ALA GCG	58	0.00246	0.0024	4.996733
CYS TGC	59	0.002354	0.0023	4.288838
ARG CGG	60	0.001406	0.0014	0.093257
ARG CGC	61	0.001042	0.001	5.822041
total				938.2207

Tabela 5.2: Parcelas da estatística de teste das probabilidades dos codões no genoma assumindo a cadeia de Markov com as frequências observadas, para o genoma da *Candida albicans*.

Para terminar esta abordagem ainda se analisou as frequências dos codões na posição adjacente ao ATG de iniciação, no sentido de averiguar se a rejeição na espécie Candida albicans é essencialmente resultado do comportamento destes codões. Aos codões na posição adjacente ao codão de iniciação chamar-se-ão de *codões quase iniciais*. Assim, interessa averiguar se a rejeição é devida à possível falta de estacionaridade do processo provocado pelos codões quase iniciais. Averiguar-se-á conjuntamente os resultados da espécie *Saccharomyces cerevisiae* servindo estes, em terminologia laboratorial, como "experiência de controlo".
Os resultados estão na Tabela 5.3, em que as células das colunas *fa(inicial)* referem-se às frequências absolutas de cada um dos codões na posição quase inicial e as células da coluna *fa(CM)* refere-se à frequência esperada assumindo a cadeia de Markov, $n\pi_j$ $j \in \{1, ..., 61\}$[5].
Da observação comparativa para as duas espécies, não se pode concluir que os codões da posição quase inicial sejam os responsáveis pelo desajuste da cadeia de Markov na *Candida albicans*, uma vez que na *Saccharomyces cerevisiae* para resultados idênticos desta estatística de teste relativamente aos codões quase iniciais tinha sido obtido ajuste (ver Tabela 5.1).
Observe-se que, em qualquer das espécies, não há ajuste entre os valores de frequências teóricos e os valores de frequências observados dos codões nas posições quase iniciais. Assim, a hipótese de estacionaridade estrita do processo a partir dos codões quase iniciais assumida no modelo não é válida. Contudo, não se pode afirmar que a distribuição inicial seja o factor responsável pela não aceitação do ajustamento anteriormente efectuado.
De qualquer forma, o facto de se estar a estudar resultados relativos ao conjunto dos genes[6] não tendo garantia que genes diferentes tenham distribuições semelhante no sequenciamento dos codões que os compõem, não permite que este estudo seja muito conclusivo em relação ao comportamento em sequência, apenas conclusivo na sua globalidade.
Este estudo não põe de parte a possibilidade do comportamento das sequências de cada gene ser ou não Markoviano, já que o que foi apresentado refere-se ao conjunto de todas sequências de código contidas no genoma.

[5]Observe-se que π_j refere-se aos valores da coluna p-CM da Tabela 5.2.

[6]Observe-se que quando se escreve conjunto dos genes, entenda-se o conjunto das sequências de código de cada gene.

Saccharomyces	fa(CM)	fa(inicial)	fa(CM)-fa(inicial)	teste	Candida	fa(CM)	fa(inicial)	fa(CM)-fa(inicial)	teste
LYS AAA	270.1964	212	58.1964	12.534664	LYS AAA	447.9043	355	92.9043	19.270208
LYS AAG	191.2839	163	28.2839	4.1821554	LYS AAG	165.9532	177	-11.0468	0.7353386
ASN AAT	230.4245	198	32.4245	4.5626581	ASN AAT	406.416	339	67.416	11.182919
ASN AAC	155.9311	126	29.9311	5.7452987	ASN AAC	160.873	106	54.873	18.716914
THR ACT	127.5226	175	-47.4774	17.67611	THR ACT	215.9085	225	-9.0915	0.3828259
THR ACC	78.9125	81	-2.0875	0.0552214	THR ACC	99.0639	72	27.0639	7.3937598
THR ACA	115.5279	160	-44.4721	17.119394	THR ACA	167.6466	339	-171.3534	175.14216
THR ACG	51.7666	66	-14.2334	3.913521	THR ACG	36.4081	55	-18.5919	9.4940067
ARG CGT	40.4032	36	4.4032	0.4798672	ARG CGT	51.6487	20	31.6487	19.393329
ARG CGC	17.0451	16	1.0451	0.0640791	ARG CGC	8.467	6	2.467	0.7188011
ARG CGA	20.2016	22	-1.7984	0.1600983	ARG CGA	44.8751	11	33.8751	25.571473
ARG CGG	11.9947	10	1.9947	0.3317155	ARG CGG	11.8538	5	6.8538	3.9628283
ARG AGA	132.573	78	54.573	22.464697	ARG AGA	172.7268	157	15.7268	1.4319274
ARG AGG	59.9735	32	27.9735	13.047708	ARG AGG	28.7878	39	-10.2122	3.6226814
SER TCT	148.3555	403	-254.6445	437.08404	SER TCT	172.7268	460	-287.2732	477.78278
SER TCC	89.6446	211	-121.3554	164.28355	SER TCC	77.0497	170	-92.9503	112.13228
SER TCA	121.2096	305	-183.7904	278.68181	SER TCA	233.6892	683	-449.3108	863.88329
SER TCG	55.5544	147	-91.4456	150.52449	SER TCG	60.1157	171	-110.8843	204.52774
SER AGT	92.1698	241	-148.8302	240.32198	SER AGT	149.0192	280	-130.9808	115.1259
SER AGC	63.13	142	-78.87	98.534404	SER AGC	44.0284	62	-17.9716	7.3356835
ILE ATT	190.6526	86	104.6526	57.445672	ILE ATT	335.2932	249	86.2932	22.208969
ILE ATC	107.9523	59	48.9523	22.198023	ILE ATC	112.6111	77	35.6111	11.261327
ILE ATA	117.4218	82	35.4218	10.685443	ILE ATA	160.0263	185	-24.9737	3.8973949
MET ATG	118.6844	96	22.6844	4.3357173	MET ATG	138.8588	125	13.8588	1.3831773
PHE TTT	169.1884	148	21.1884	2.6535406	PHE TTT	260.7836	300	-39.2164	5.8973265
PHE TTC	114.8966	69	45.8966	18.333857	PHE TTC	125.3116	95	30.3116	7.3320674
TYR TAT	121.8409	51	70.8409	41.188411	TYR TAT	215.0618	138	77.0618	27.613091
TYR TAC	92.1698	46	46.1698	23.127428	TYR TAC	81.2832	60	21.2832	5.5727949
CYS TGT	51.1353	28	23.1353	10.467174	CYS TGT	80.4365	56	24.4365	7.4237757
CYS TGC	31.565	22	9.565	2.8984389	CYS TGC	19.4741	21	-1.5259	0.1195624
TRP TGG	66.2865	37	29.2865	12.939272	TRP TGG	84.67	38	46.67	25.724447
LEU TTA	167.2945	72	95.2945	54.281771	LEU TTA	324.2861	176	148.2861	67.806691
LEU TTG	167.9258	95	72.9258	31.669775	LEU TTG	287.0313	193	94.0313	30.804603
LEU CTT	80.1751	77	3.1751	0.1257405	LEU CTT	95.6771	120	-24.3229	6.183334
LEU CTC	35.9841	25	10.9841	3.3528823	LEU CTC	29.6345	38	-8.3655	2.3614905
LEU CTA	85.8568	69	16.8568	3.3096005	LEU CTA	57.5756	106	-48.4244	40.727713
LEU CTG	67.5491	60	7.5491	0.8436665	LEU CTG	44.8751	86	-41.1249	37.688103
PRO CCT	85.8568	114	-28.1432	9.2251249	PRO CCT	110.071	94	16.071	2.3464586
PRO CCC	44.191	28	16.191	5.932169	PRO CCC	39.7949	29	10.7949	2.9282613
PRO CCA	113.0027	104	9.0027	0.7172272	PRO CCA	202.3613	218	-15.6387	1.2085756
PRO CCG	34.7215	25	9.7215	2.7218744	PRO CCG	27.9411	36	-8.0589	2.3243848
HIS CAT	88.382	51	37.382	15.811069	HIS CAT	133.7786	79	54.7786	22.430307
HIS CAC	49.2414	31	18.2414	6.7574982	HIS CAC	48.2619	21	27.2619	15.399543
GLN CAA	170.451	102	68.451	27.48907	GLN CAA	308.1988	131	177.1988	101.88039
GLN CAG	78.2812	51	27.2812	9.5075685	GLN CAG	66.8893	32	34.8893	18.198176
VAL GTT	135.7295	146	-10.2705	0.7771573	VAL GTT	220.142	180	40.142	7.3197307
VAL GTC	71.3369	82	-10.6631	1.5938694	VAL GTC	67.736	58	9.736	1.3993991
VAL GTA	77.0186	107	-29.9814	11.671003	VAL GTA	88.0568	117	-28.9432	9.5132781
VAL GTG	68.1804	62	6.1804	0.5602394	VAL GTG	93.137	106	-12.863	1.7764881
ALA GCT	127.5226	207	-79.4774	49.533629	ALA GCT	177.807	218	-40.193	9.0855661
ALA GCC	76.3873	89	-12.6127	2.0825478	ALA GCC	82.1299	79	3.1299	0.1192778
ALA GCA	102.9019	162	-59.0981	33.940923	ALA GCA	137.1654	264	-126.8346	117.28188
ALA GCG	39.1406	56	-16.8594	7.2620085	ALA GCG	20.3208	45	-24.6792	29.972389
GLY GGT	142.0425	133	9.0425	0.5756503	GLY GGT	198.9745	151	47.9745	11.567073
GLY GGC	61.8674	65	-3.1326	0.1586164	GLY GGC	38.1015	23	15.1015	5.9854678
GLY GGA	71.3369	69	2.3369	0.0765537	GLY GGA	121.9248	101	20.9248	3.5911255
GLY GGG	38.5093	32	6.5093	1.1002793	GLY GGG	63.5025	57	6.5025	0.66584
ASP GAT	238.6314	207	31.6314	4.1928492	ASP GAT	375.0881	249	126.0881	42.385266
ASP GAC	128.1539	105	23.1539	4.1832756	ASP GAC	109.2243	74	35.2243	11.359664
GLU GAA	287.2415	237	50.2415	8.7877564	GLU GAA	417.4231	224	193.4231	89.627277
GLU GAG	122.4722	102	20.4722	3.4220907	GLU GAG	110.071	86	24.071	5.2639936
total	6313	6313		1979.7039		8467	8467		2895.4425

Tabela 5.3: Parcelas da estatística de teste, **teste**, das frequências esperadas assumindo a cadeia de Markov, **fa(CM)**, com as frequências observadas dos codões na posição adjacente ao ATG de iniciação, **fa(inicial)**, para o genoma da *Saccharomyces cerevisiae* e da *Candida albicans*.

Capítulo 6

Análise das Frequências dos Símbolos

6.1 Introdução

A Análise de Zipf surge ligada ao estudo de linguagens[1], especificamente no âmbito do estudo de linguagens correntes[2]. Mas, também é extensível a outro tipo de linguagens, como por exemplo, a linguagem binária ou a linguagem genética, objecto do presente estudo. De um modo geral, na Análise de Zipf efectua-se a contagem do número de vezes que cada palavra distinta surge no texto e um dos objectivos é averiguar a lei que define o comportamento das frequências ordenadas dessas palavras no texto.
É óbvio que para estudar uma linguagem é necessário reconhecer o alfabeto e as palavras que a constituem. Uma linguagem pode ser de um de dois tipos: linguagem com palavras de comprimento variável ou linguagem com palavras de comprimento fixo.
A linguagem a estudar consiste em sequências de símbolos em que as palavras têm um comprimento fixo de n símbolos, permitindo assim um estudo baseado na Análise de Zipf sobre o n-uplo.
A linguagem genética, objecto deste estudo, consiste na sequência de símbolos, os codões. Cada codão é uma sequência de três nucleótidos também chamados de bases. Como existem 4 bases distintas, tem-se sessenta e quatro codões diferentes, correspondendo ao número de combinações de quatro bases três a três. Assim, o alfabeto de textos de sequências de codões é constituido por sessenta e quatro símbolos ou codões distintos.
Também se poderá considerar a linguagem genética como sequências de aminoácidos. Observe-se que existe uma correspondência não injectiva entre codões e aminoácidos: os codões de terminação TAA, TAG e TGA não codificam aminoácidos, apenas codificam a paragem da produção de aminoácidos na construção das proteínas. O alfabeto de textos de sequências de aminoácidos é constituido por vinte símbolos ou aminoácidos distintos.
Neste capítulo far-se-á a apresentação da Lei de Zipf e generalização, no âmbito da Análise de Zipf sobre o n-uplo. Averiguar-se-á também a possibilidade da sequência dos codões seguir um comportamento Markoviano nas espécies em estudo.

[1] Um vasto leque de diferentes tipos de linguagens pode ser encontrado no seguinte endereço http://linkage.rockefeller.edu/wli/zipf/.
[2] Português, Inglês, ...

6.2 A Gramática das Sequências de Código

Pode-se fazer uma correspondência entre a gramática de textos em geral e uma possível gramática para o texto genético (ver [8]). Entenda-se por texto genético o conjunto das sequências de símbolos, podem os símbolos ser os nucleótidos, os codões ou os aminoácidos.
O alfabeto é um conjunto de letras (símbolos). No texto, as letras estão combinadas por forma a construir frases. Assim, a linguagem é um conjunto de frases compostas por um conjunto de letras do alfabeto.
As linguagens formais são definidas através de uma gramática. A gramática é um conjunto de regras sintácticas, que descrevem a construção das frases.
Uma gramática pode ser definida pelo seguinte quadrúplo ordenado (Σ, I, P, S) , onde:

- S o conjunto dos símbolos iniciais;
- Σ o conjunto dos símbolos terminais;
- I o conjunto dos símbolos não terminais;
- P o conjunto das regras.

Um texto genético adapta-se a uma gramática deste tipo, definindo a seguinte correspondência:

- $S = \{ATG\}$;
- $\Sigma = \{TAA, TAG, TGA\}$;
- $I = \{$AAA, AAC, ..., TTG, TTT$\}\backslash\Sigma$;
- P desconhecido.

Neste caso, desconhece-se o conjunto P, sendo a sua descoberta (pelo menos parcial) o objectivo deste estudo.

6.3 Lei de Zipf

George Zipf (1902-1950) dedicou-se ao estudo de linguagens correntes e propôs leis que caracterizavam o seu comportamento. De seguida, citar-se-ão alguns resultados de Zipf, extensões propostas por diversos autores e algumas generalizações de resultados que para este estudo pareceram convenientes (ver, por exemplo, em [3], [5], [9] ou [25]).
Dado um texto, denote-se por $W(R)$ a frequência relativa da R-ésima palavra mais frequente no texto. A Análise de Zipf consiste, numa primeira fase, na aplicação sequencial dos três passos seguintes:

1. Contagem do número de vezes que cada uma das palavras distintas surge no texto;
2. Cálculo da frequência relativa de cada palavra;
3. Ordenação das frequências $W(R)$ por ordem decrescente.

Assim, no estudo de um texto com N palavras distintas, usando a Análise de Zipf, determinam-se inicialmente as frequências relativas ordenadas $W(1) \geq W(2) \geq ... \geq W(N)$.
Zipf concluiu que, nas linguagens correntes, a ordem R das frequências das palavras ajusta-se razoavelmente a uma lei de proporcionalidade inversa da frequência $W(R)$. Com esta motivação surge a *Lei de Zipf* dada por:

$$R \cdot W(R) = K, \tag{6.1}$$

para alguma constante K. A constante K pode variar de linguagem para linguagem e também com os diferentes modos de escrita de uma mesma linguagem.
Como é referido em [9] Zipf ainda propôs uma generalização da lei anterior dada pela lei de potência:

$$W(R) = KR^a \tag{6.2}$$

com $K \in \mathbb{R}^+$ e $a \in \mathbb{R}^-$.
De acordo com diversos autores (por exemplo [5]) a equação (6.2) é também chamada de Lei de Zipf. De notar que a equação (6.2) é uma generalização da equação (6.1), uma vez que para $a = -1$ as equações coincidem.
Descobrir a Lei de Zipf associada a determinada linguagem consiste na determinação do valor do expoente a da equação (6.2). Para descobrir o valor de a, numa determinada linguagem geralmente recorre-se ao *gráfico de Zipf*, que consiste na representação gráfica do logaritmo da frequência em função do logaritmo da ordem, ou seja, uma simples linearização da equação (6.2). Assim, a determinação do expoente a passa a ser a estimação do declive da recta que melhor se ajusta aos pontos do gráfico. A partir desta facilmente se deduz a Lei de Zipf:

$$\ln(W(R)) = a \ln(R) + b$$

$$W(R) = R^a \cdot e^b.$$

Portanto,

$$W(R) \sim R^a.$$

Onde o símbolo $\sim$ refere-se à lei de proporcionalidade directa.

6.4 Análise de Zipf sobre o n-uplo

A Análise de Zipf sobre o n-uplo é uma extensão à Análise de Zipf. Na Análise de Zipf sobre o n-uplo averiguar-se-á a distribuição de frequências das palavras constituídas por n símbolos consecutivos, em que n é o comprimento de cada uma das palavras consideradas no texto.
Quando $n = 1$ chama-se simplesmente Análise de Zipf. Para $n \geq 2$, introduz-se essa informação no título da análise designando-a de Análise de Zipf sobre o n-uplo. Em particular, também quando $n = 2$, dir-se-á Análise de Zipf sobre o par e quando $n = 3$, Análise de Zipf sobre o terno.

Assim, no caso concreto das sequências de codões e para uma Análise de Zipf tem-se 64 codões. No caso de uma Análise de Zipf sobre o par tem-se 61×64 palavras, os pares de codões possíveis numa leitura $3'$. Observe-se que com o crescimento de n o número de palavras cresce exponencialmente, $61^{n-1} \times 64$.
A Análise de Zipf sobre o n-uplo, para além de averiguar a existência de uma lei entre as frequências relativas das palavras de comprimento n e a sua ordem, explora também a possibilidade de existência de correlações de longo alcance, na medida em que explora se o conhecimento do passado (curto ou longo) é suficiente para explicar o conhecimento do presente.
Numa sequência, a contagem das palavras com n símbolos ($n \geq 2$) é uma contagem encadeada, isto é, um símbolo que esteja na última posição de uma palavra está na penúltima posição da palavra seguinte, e assim por diante. O número de palavras numa sequência com um total de L símbolos, é $L_n = L - n + 1$.
De acordo com [11] a Análise de Zipf sobre n-uplos só fará sentido quando:

$$L > 10 \cdot S_n, \tag{6.3}$$

sendo S_n o número de palavras distintas com n símbolos. Esta condição pretende garantir que cada palavra distinta no texto tenha frequência "significativa" face ao total de palavras no texto.
Têm sido estudados, por muitos autores[3], vários tipos de linguagens tendo associado a cada uma dessas linguagens determinada Lei de Zipf. Contudo, o ajuste não é ideal e, em alguns casos os autores fazem referência à existência de um conjunto de pontos "mal comportados", normalmente localizados nas caudas da distribuição de frequências (ver [5][4]).
No artigo [11] utiliza-se a Análise de Zipf sobre o n-uplo no sentido de averiguar se nas sequências de bases não se verificam correlações de longo alcance[5]. De acordo com a inequação (6.3) a análise feita no âmbito das bases, é válida para valores de n grandes. Como resultado do estudo o autor concluiu assim que as correlações existentes entre as bases seriam de curto alcance. Seguindo este tipo de abordagem aplicar-se-á a Análise de Zipf sobre o n-uplo no contexto dos codões.

6.5 Aplicação ao Caso em Estudo

Para averiguar a existência de uma lei que defina a distribuição de frequências com que os codões ou aminoácidos aparecem no texto genético aplicar-se-á a Análise de Zipf sobre o n-uplo a cada uma das espécies: *Saccharomyces cerevisiae* e *Candida albicans*.
Os símbolos do alfabeto a considerar serão os codões e os aminoácidos, separadamente. Observe-se que o conjunto de codões é constituído por 64 elementos ($S_1 = 64$) e o conjunto de aminoácidos por 21 elementos[6] ($S_1 = 21$).

[3] no endereço http://linkage.rockefeller.edu/wli/zipf/ encontra-se uma lista de aplicação da Análise de Zipf em muitos contextos distintos, como também um grande conjunto de referências bibliográficas.

[4] Em [5] é referido que o melhor ajuste dá-se para o conjunto das ordens (R) tal que: $R < S_{n-1}$.

[5] No sentido de averiguar quanto às possíveis correlações o autor supôs, por um lado, que os dados verificavam um comportamento markoviano de primeira ordem e por outro considerou a Análise de Zipf de ordem superior a 2, concretamente a de ordem 6.

[6] Na verdade o conjunto dos aminoácidos tem 20 elementos, mas considera-se, por abuso, que os codões terminais são código de um imaginário aminoácido de terminação na construção da proteína.

Far-se-á, em primeiro lugar, e em paralelo um estudo da linguagem genética no contexto dos codões e dos aminoácidos recorrendo à Análise de Zipf. A seguir far-se-ão as análises de Zipf sobre o par e sobre o terno de codões. O número total de codões, L, que constitui o texto genético da espécie *Saccharomyces cerevisiae* é de 2 968 093 e da espécie *Candida albicans* é de 3 397 032.
Observe-se que para $n = 4$ o segundo membro da inequação (6.3) é igual a $10 \cdot 64^4 = 16\ 777\ 260$, pelo que não se verifica a desigualdade. Assim, o estudo só se justifica para $n = 1, 2, 3$. Para $n = 3$, tem-se $S_3 = 238\ 144$. Obviamente é difícil trabalhar com um número tão grande de objectos (palavras) distintos, o esforço computacional é muito elevado, pois envolve trabalhar com algoritmos de ordenação.
Seguindo a abordagem considerada em [11], averiguar-se-á o ajustamento do texto genético no contexto dos codões a uma cadeia de Markov[7], tendo em conta que a análise de Zipf sobre o n-uplo só é válida para $n \leq 3$.
Na cadeia de Markov a propôr, $\{X_n, n \in \mathbb{N}_0\}$, o espaço de estados é constituído pelos 61 codões não iniciais. A distribuição de probabilidade inicial dos símbolos é dada pelas frequências relativas obtidas directamente das contagens, a que se designará por Π_0. A matriz das probabilidades de transição coincide com a matriz P^* do capítulo anterior referente a codões não inicial e não terminais, em que cada elemento coincide com a frequência relativa de um par de codões para o primeiro codão do par fixo.
No caso de $n = 1$ os resultados da Análise de Zipf e os estimados pelas cadeias de Markov, pela forma como foram definidos, coincidem.
No caso de $n = 2$ e $n = 3$ comparam-se os resultados da Análise de Zipf do par e do terno com os estimados pelas cadeias de Markov, respectivamente. O objectivo será analisar a adequação do modelo de cadeia de Markov escolhido, aos dados.
Sob o modelo de curto alcance considerado, o modelo de Markov com P^* e Π_0, se não se verificar ajuste entre a distribuição de frequências observadas dos ternos e a distribuição dos ternos sob a validade do modelo teórico admitido, poder-se-á por a hipótese de existirem correlações de longo alcance.
Partindo das contagens dos n-uplos de símbolos (codões ou aminoácidos), procedeu-se à ordenação das respectivas frequências. Mostrar-se-ão as correspondentes representações gráficas por serem de mais fácil interpretação. Note-se que no caso dos aminoácidos, apenas se apresentam os resultados da Análise de Zipf, pois para além de ser uma abordagem semelhante à que se apresentará para o caso dos codões, também não foram disponibilizadas, em tempo útil, as respectivas contagens de pares e ternos de aminoácidos.

6.5.1 Análise de Zipf

Em primeiro lugar, fez-se o estudo da linguagem das palavras com um único símbolo ($n = 1$), a Análise de Zipf. Apresentam-se quatro gráficos de Zipf (logaritmo da ordem R *versus* logaritmo da frequência da palavra com ordem R), para cada uma das espécies relativos aos codões e aos aminoácidos, respectivamente (ver Figuras 6.1, 6.2, 6.3 e 6.4).

[7]Embora nada seja dito em relação ao tipo de cadeia de Markov, no âmbito da Biologia, tudo leva a crer que a sequência de codões seja uma cadeia homogénea e estritamente estacionária.

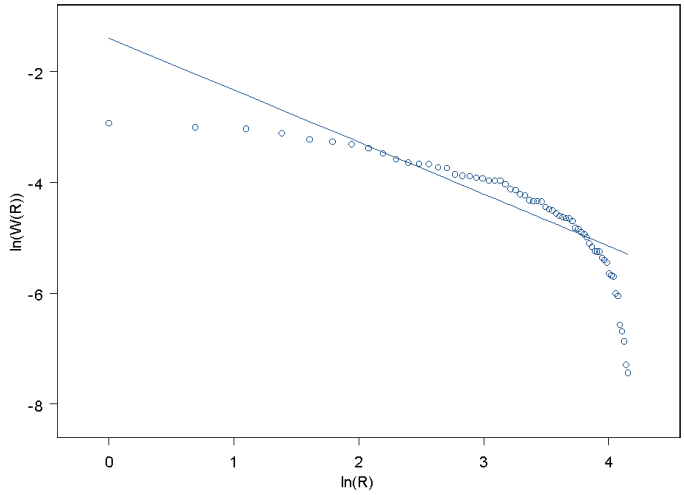


Figura 6.1: Gráfico de Zipf relativo aos codões da espécie *Candida albicans*.

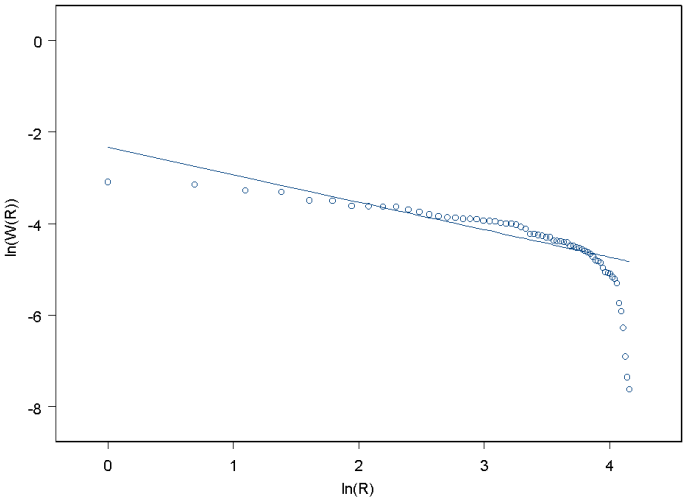


Figura 6.2: Gráfico de Zipf relativo aos codões da espécie *Saccharomyces cerevisiae*.

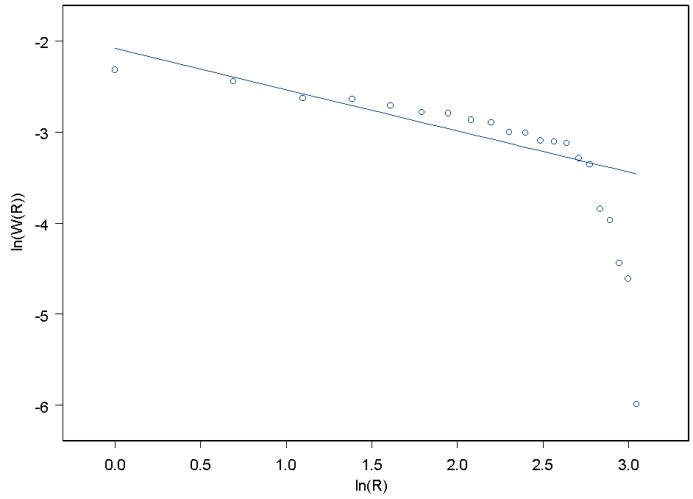


Figura 6.3: Gráfico de Zipf relativo aos aminoácidos da espécie *Candida albicans*.

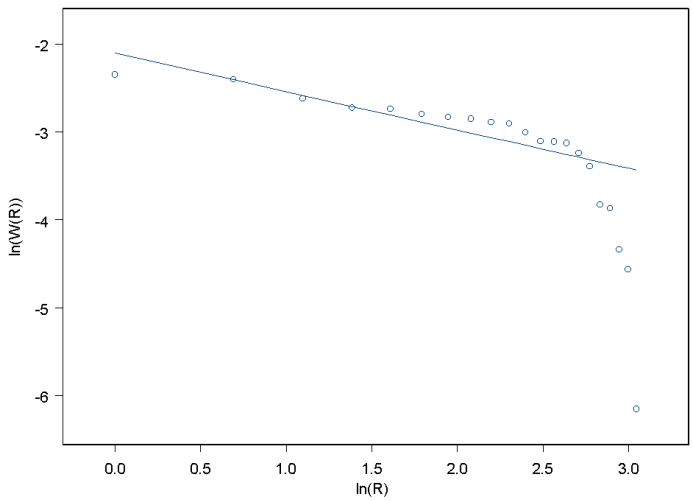


Figura 6.4: Gráfico de Zipf relativo aos aminoácidos da espécie *Saccharomyces cerevisiae*.

Conjuntamente com o gráfico de Zipf traçou-se a recta robusta[8] para estimar a Lei de Zipf e obtiveram-se as estimativas apresentadas na Tabela 6.1. Com o auxílio do programa *S-Plus 2000* determinaram-se as estimativas LTS (*Least Trimmed Squares*).

Espécie (Figura)	Declive da recta	Ordenada na origem
Candida albicans / codões (ver Figura 6.1)	-0.9516	-1.3089
Saccharomyces cerevisiae / codões (ver Figura 6.2)	-0.5635	-2.4126
Candida albicans / aminoácidos (ver Figura 6.3)	-0.4532	-2.0810
Saccharomyces cerevisiae / aminoácidos (ver Figura 6.4)	-0.4366	-2.1052

Tabela 6.1: Estimativas LTS dos parâmetros da recta robusta que, de acordo com o método LTS, melhor se ajusta aos pontos do gráfico de Zipf, para as espécies *Saccharomyces cerevisiae* e *Candida albicans*.

Desta análise conclui-se que as leis de Zipf que melhor se ajustam às linguagens de codões e de aminoácidos das duas espécies em estudo têm expoentes aproximadamente iguais a -0.95, -0.56, -0.45, -0.43 para as situações referentes às Figuras 6.1, 6.2, 6.3 e 6.4 respectivamente. Contudo, a distribuição empírica dos pontos no caso dos codões não parece ter um comportamento rectilíneo mesmo à excepção das caudas. No sentido de reduzir o desajuste sugere-se seguir uma metodologia aproximada à Análise de Zipf trabalhando desta forma com o logaritmo da frequência relativa em função da ordem. Os gráficos obtidos estão apresentados nas Figuras 6.5 e 6.6.

[8] Os modelos de regressão robusta são úteis para detectar relações lineares quando a variação aleatória dos dados não é normal ou quando contém *outliers*. O método de estimação robusto da recta de regressão utilizado foi o conhecido na literatura inglesa por *Least Trimmed Squares (LTS)*.
Para ilustrar brevemente este método desenvolvido por Rousseeuw (1984), considere-se os pares de valores observados (x_i, y_i) e seja y_i' o valor estimado de y_i por LTS, com $i \in \{1, ..., n\}$. Os valores estimados y_i' são determinados por forma a minimizar a soma aparada do quadrado dos resíduos, ou seja,

$$\sum_{j=1}^{h} (r_j)^2$$

onde r_j representa a diferença entre um valor observado e a estimativa, para $j \in \{1, ..., n\}$ com $r_1 < r_2 < ...r_n$. O número de resíduos a considerar é representado por h e é determinado com base no número de valores aparados.

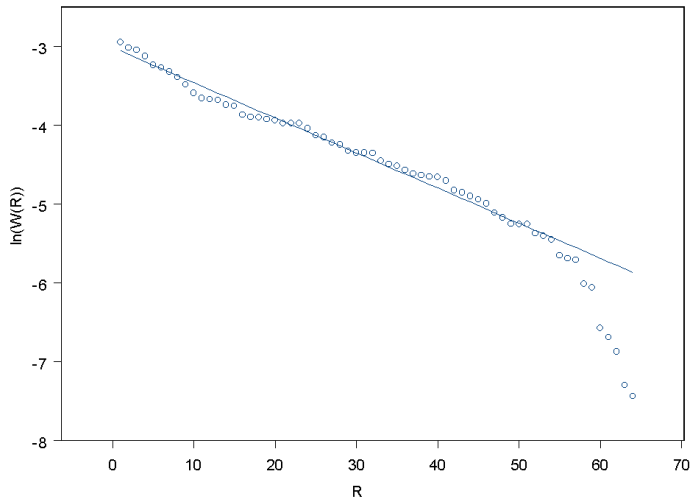


Figura 6.5: Gráfico da ordem *versus* logaritmo da frequência relativa dos codões da espécie *Candida albicans*.

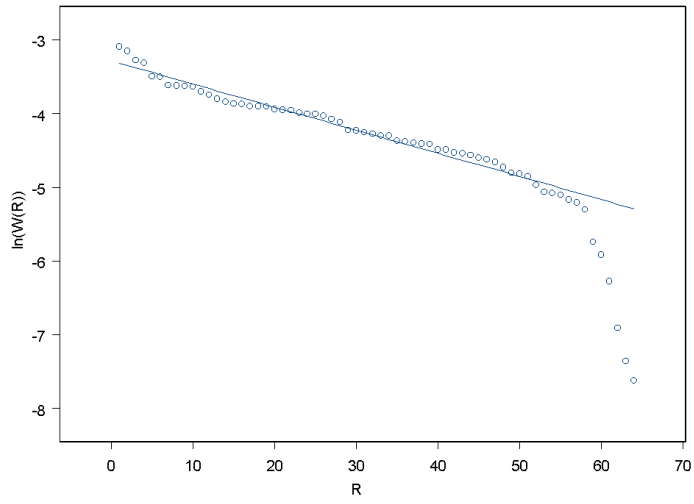


Figura 6.6: Gráfico da ordem *versus* logaritmo da frequência relativa dos codões da espécie *Saccharomyces cerevisiae*.

Recorrendo novamente à aplicação *S-Plus 2000* e utilizando o método LTS obtiveram-se as seguintes estimativas dos parâmetros da recta robusta, para os pares de codões em ambas as espécies.

Espécie (Figura)	Declive da recta	Ordenada na origem
Candida albicans (Figura 6.5)	-0.0434	-3.0410
Saccharomyces cerevisiae (Figura 6.6)	-0.0300	-3.3252

No caso de se aceitar o ajuste destas recta às respectivas distribuições de frequências empíricas, a lei que caracteriza os textos genéticos no contexto dos codões será da forma:

$$W(R) = k(e^R)^a \tag{6.4}$$

com $k \in \mathbb{R}^+$ e $a \in \mathbb{R}^-$.
A equação (6.4) resulta como consequência imediata do ajuste dos pontos de coordenadas $(R, \ln W(R))$ a uma recta de declive a e ordenada na origem $ln(k)$.
Assim, de acordo com os valores obtidos acima as leis que caracterizam as rectas das Figuras 6.5 e 6.6 correspondem às relações $W(R) \sim e^{-0.0434R}$ e $W(R) \sim e^{-0.03R}$, respectivamente.
Para a maioria dos textos genéticos acredita-se que é possível estimar uma lei para a frequência relativa dos codões à custa da Análise de Zipf. Essa lei pode não coincidir com a Lei de Zipf, como no caso presente para os textos da *Saccharomyces cerevisiae* e da *Candida albicans* no contexto dos codões.

6.5.2 Análise de Zipf sobre o Par

Na Análise de Zipf sobre o par cada palavra é constituída por dois codões justapostos (dois símbolos do alfabeto). Neste contexto, a leitura das palavras, na sequência dos codões é feita de forma a existir encadeamento, isto é, o último símbolo de uma palavra coincide com o primeiro símbolo da palavra seguinte.
A título de exemplo, na Figura 6.7 é ilustrada a contagem das cinco palavras possíveis, $P1$, $P2$, $P3$, $P4$, $P5$, no sequenciamento de seis codões para uma Análise de Zipf sobre o par.

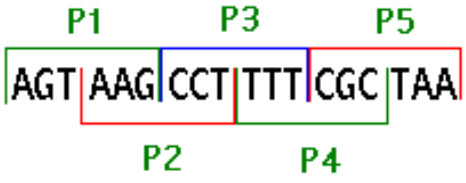


Figura 6.7: Ilustração da determinação de pares de codões na sequência.

Construiram-se os gráficos das frequências relativas ordenadas para as duas espécies em estudo no contexto da Análise de Zipf sobre o par (ver Figuras 6.8 e 6.9). Conjuntamente para cada uma das espécies estimou-se uma lei das frequências relativas de pares de símbolos no contexto

dos codões.
Tendo em vista a aplicação do modelo de cadeias de Markov homogéneas e estritamente estacionárias, neste estudo foram também retirados os pares de codões que provocam o início e a paragem da construção da proteína.

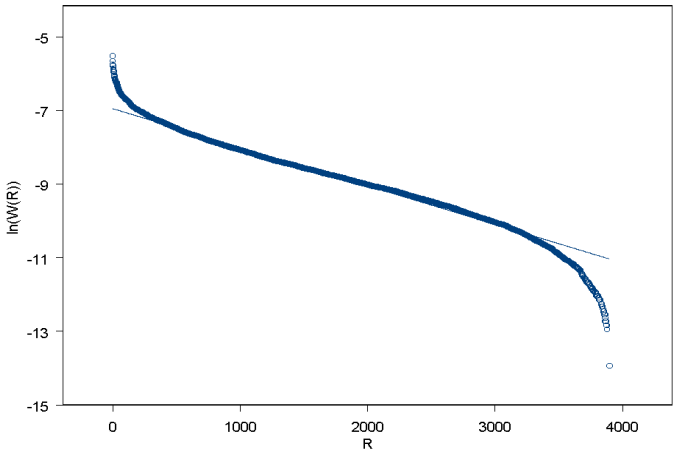


Figura 6.8: Gráfico da ordem *versus* logaritmo da frequência relativa dos codões da espécie *Candida albicans*, relativo à Análise de Zipf sobre o par.

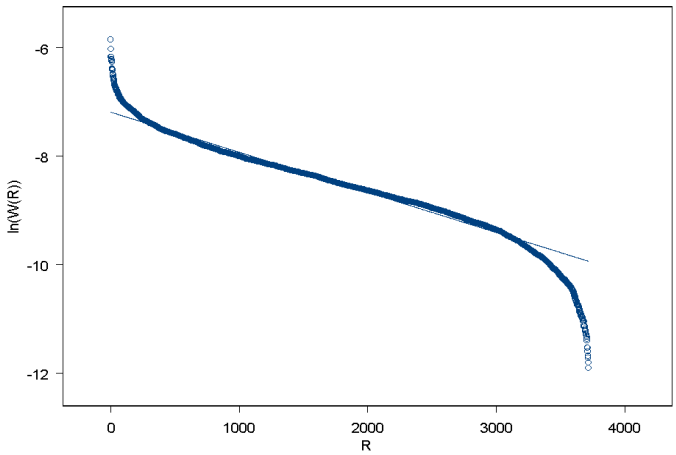


Figura 6.9: Gráfico da ordem *versus* logaritmo da frequência relativa dos codões da espécie *Saccharomyces cerevisiae*, relativo à Análise de Zipf sobre o par.

A estimativa dos parâmetros da recta robusta obtida pelo método LTS ajustada aos gráficos das Figuras 6.8 e 6.9 apresentam-se na seguinte tabela:

Espécie (Figura)	Declive da recta	Ordenada na origem
Candida albicans (ver Figura 6.8)	-0.0010	-6.9545
Saccharomyces cerevisiae (ver Figura 6.9)	-0.0007	-7.1907

Assim, as leis que caracterizam as rectas das Figuras 6.8 e 6.9 correspondem às relações $W(R) \sim e^{-0.001R}$ e $W(R) \sim e^{-0.0007R}$, respectivamente.
No âmbito da Análise de Zipf sobre o par poder-se-á ainda averiguar, em certa parte, a adequação do ajuste de um modelo de cadeia de Markov.
Na cadeia de Markov, $\{X_n, n \in \mathbb{N}_0\}$, X_n refere-se ao codão na n-ésima posição na sequência de codões não inicial e não terminais. Se se assumir que a cadeia de Markov é homogénea e estritamente estacionária tem-se que: $P(X_{n+1} = C_i | X_n = C_j) = P(X_1 = C_i | X_0 = C_j)$, $P(X_n = C_i) = P(X_0 = C_i)$ e $P(X_n = C_i, X_{n+1} = C_j) = P(X_0 = C_i, X_1 = C_j)$ $\forall_{n \in \mathbb{N}_0}$ $\forall_{C_i, C_j \in E}$. Consequentemente,

$$\begin{aligned} P(X_n = C_i, X_{n+1} = C_j) &= P(X_n = C_i) P(X_{n+1} = C_j | X_n = C_i) = \\ &= P(X_m = C_i) P(X_{m+1} = C_j | X_m = C_i), \quad \forall_{n,m \in \mathbb{N}_0} \quad \forall_{C_i, C_j \in E}. \end{aligned} \tag{6.5}$$

Pelo que a equação (6.5) pode-se escrever na seguinte forma simplificada:

$$P((C_i, C_j)) = P(C_i) P(C_j | C_i) \tag{6.6}$$

onde $P((C_i,C_j))$ representa a probabilidade de ocorrência do par de codões justapostos (C_i,C_j), $P(C_i)$ a probabilidade de ocorrência do codão C_i e $P(C_j|C_i)$ a probabilidade do codão C_j condicionada a que o codão anterior adjacente seja C_i.
Considere-se a frequência relativa do par de codões justapostos (C_i,C_j) no conjunto das sequências de codões observada, de uma dada espécie, como a estimativa de $P((C_i,C_j))$. De modo análogo, estima-se $P(C_i)$ pela frequência relativa do codão C_i no conjunto das sequências de codões da espécie, a i-ésima componente do vector que define a distribuição de probabilidade inicial Π_0. A estimativa de $P(C_j|C_i)$ a utilizar coincide com o elemento da célula (i,j) da matriz das probabilidades de transição P^* considerada no capítulo anterior, com $i,\ j \in \{1, ..., 61\}$.
As sequências de codões não se encontram justapostas não entrando para a contagem os pares de codões que hipoteticamente poderiam unir todas as sequências de codões numa só sequência. Perante este problema averiguar-se-á a possibilidade de não se verificar a igualdade da equação (6.6), por uso das referidas estimativas.
Nas Figuras 6.10 e 6.11, apresentam-se para a *Candida albicans* e para a *Saccharomyces cerevisiae* as representações gráficas conjuntas dos valores estimados para $P((C_i,C_j))$ e para $P(C_i)P(C_j|C_i)$ assumindo aquele modelo de Markov.

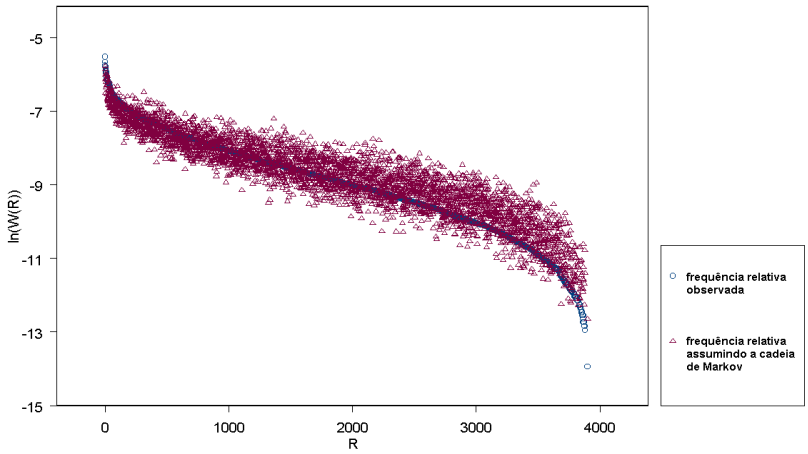


Figura 6.10: Representação gráfica do logaritmo da frequência relativa dos pares de codões justapostos (Análise de Zipf sobre o par), conjuntamente com a do logaritmo da frequência relativa assumindo uma cadeia de Markov *versus* R, para a *Candida albicans*.

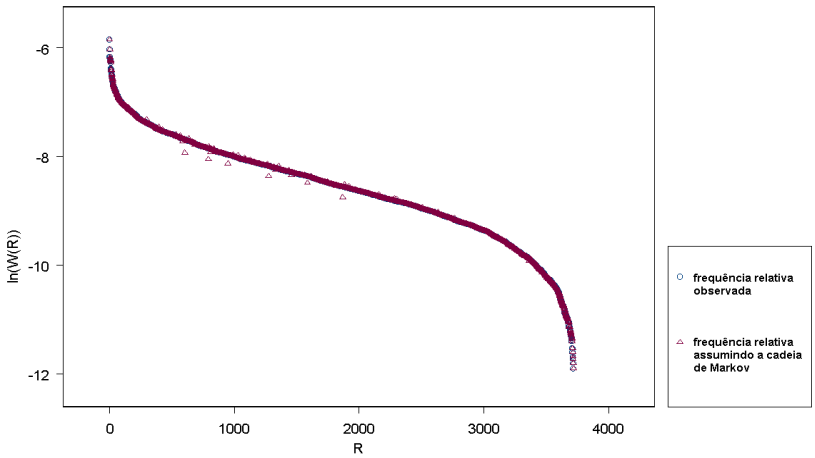


Figura 6.11: Representação gráfica do logaritmo da frequência relativa dos pares de codões justapostos (Análise de Zipf sobre o par), conjuntamente com a do logaritmo da frequência relativa assumindo uma cadeia de Markov *versus* R, para a *Saccharomyces cerevisiae*.

Nas Figura 6.10 e 6.11 observa-se graficamente discrepâncias de valores, sendo estas discrepâncias mais frequentes na *Candida albicans* (ver Figura 6.10). Aplicou-se um teste de ajustamento do qui-quadrado às duas situações anteriores para testar a hipótese:

$$H_0 : \frac{n_{ij}}{n} = p^*_{ij} \text{ com } i, j \in \{1, ..., 61\},$$

onde $\frac{n_{ij}}{n}$ é a frequência relativa observada do par (C_i, C_j) e p^*_{ij} é o valor resultante para $P(C_i)P(C_j|C_i)$ quando se assume a cadeia de Markov com distribuição inicial Π_0 e matriz de probabilidades de transição P^*.
O valor do quantil de ordem 0.95 de uma distribuição de um qui-quadrado com $61 \times 61 - 1 = 3720$ graus de liberdade é de 3862.33. Donde do teste de hipótese, para a espécie *Candida albicans* representada na Figura 6.10, conclui-se o desajuste entre os valores observados e os valores estimados assumindo o modelo probabilístico associado à cadeia de Markov, já que, o valor da estatística de teste χ^2 é de 468911.0335. Para a espécie *Saccharomyces cerevisiae* não é de rejeitar o ajuste ao nível de 5% de significância, já que o valor da estatística de teste do χ^2 é de $287,64 < 3862.33$.
Face ao desajuste e a título de curiosidade, no caso da espécie *Candida albicans*, ainda se ordenaram as estimativas da probabilidade dos pares de codões com base no modelo probabilístico associado à cadeia de Markov e representou-se conjuntamente com esta distribuição a distribuição de frequências relativas ordenadas (ver Figura 6.12).

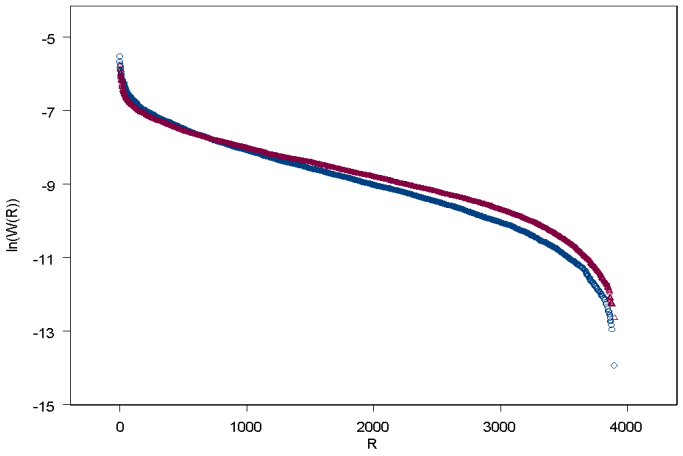


Figura 6.12: Representação gráfica do logaritmo da frequência relativa dos pares de codões justapostos (Análise de Zipf sobre o par), conjuntamente com a do logaritmo da frequência relativa ordenadas assumindo uma cadeia de Markov *versus* R, para a *Candida albicans*.

Todavia, efectuando um teste de ajustamento conclui-se que as diferenças entre as distribuições ordenadas de frequências dos codões observados e estimados assumindo a cadeia de Markov, são significativas, ao nível de 5% de significância.
Perante os resultados obtidos, pode-se afirmar que para o sequenciamento dos codões no genoma como um todo, na *Saccharomyces cerevisiae* não é de rejeitar a adequação do modelo probabilístico associado à cadeia de Markov, face à *Candida albicans* em que há rejeição do modelo proposto.

6.5.3 Análise de Zipf sobre o Terno

Para $n = 3$ o número de palavras distintas a estudar, no âmbito dos codões, é muito elevado: cerca de duzentas e cinquenta mil palavras!
As contagens foram feitas para ambas as espécies em estudo. Nas Figuras 6.13 e 6.14 apresentam-se os gráfico das frequências relativas das palavras em função da sua ordem.

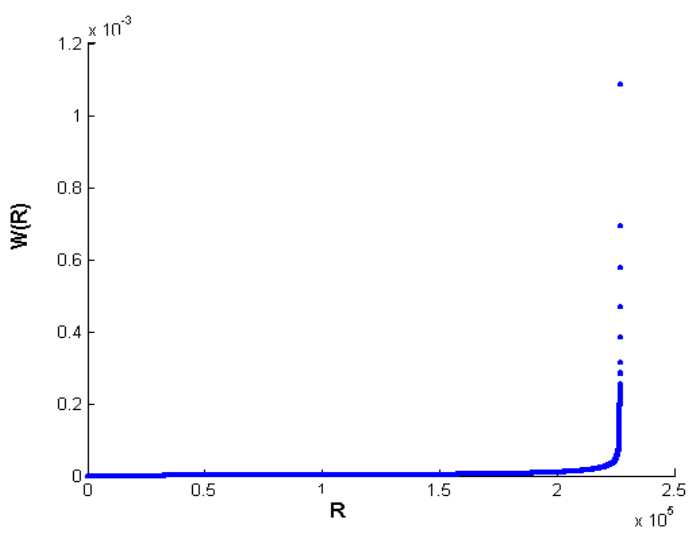


Figura 6.13: Gráfico da frequência relativas dos ternos de codões da espécie *Candida albicans* em função da ordem.

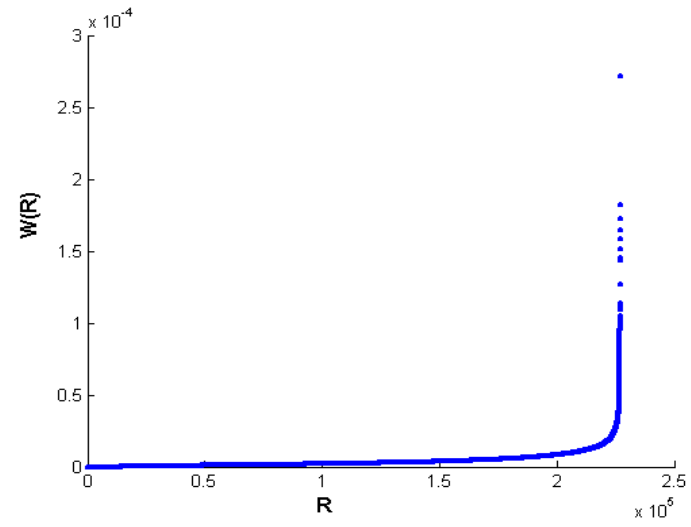


Figura 6.14: Gráfico da frequência relativas dos ternos de codões da espécie *Saccharomyces cerevisiae* em função da ordem.

Neste caso não se estimou a Lei de Zipf associada ou leis idênticas, porque por um lado, o esforço computacional é grande e, por outro, o conhecimento de leis à custa da Análise de Zipf não parece contribuir, neste estudo, para o conhecimento do comportamento dos símbolos em sequência.
Ainda no sentido de averiguar quanto à adequação do modelo de cadeia de Markov estacionária e homogénea com Π_0 a distribuição de probabilidade inicial e P^* a matriz de probabilidades de transição, far-se-á um estudo análogo ao da subsecção anterior para o caso dos ternos de codões.
Se se assumir a cadeia de Markov e em analogia com a equação (6.6) poder-se-á escrever de forma simplificada:

$$P((C_i, C_j, C_k)) = P(C_i)P(C_j|C_i)P(C_k|C_j) \tag{6.7}$$

onde $P((C_i, C_j, C_k))$ representa a probabilidade de ocorrência do terno de codões justapostos (C_i, C_j, C_k).
Assim, no seguimento da metodologia utilizada, representam-se em conjunto os dados a partir da Análise de Zipf sobre o terno, a frequência relativa do terno como estimativa de $P((C_i, C_j, C_k))$, com os valores provenintes de assumir da cadeia de Markov em que a matriz das probabilidades de transição contém a estimativa de $P(C_j|C_i)$ e de $P(C_k|C_j)$ e a distribuição de probabilidade inicial permite estimar $P(C_i)$ (ver Figuras 6.15 e 6.16).

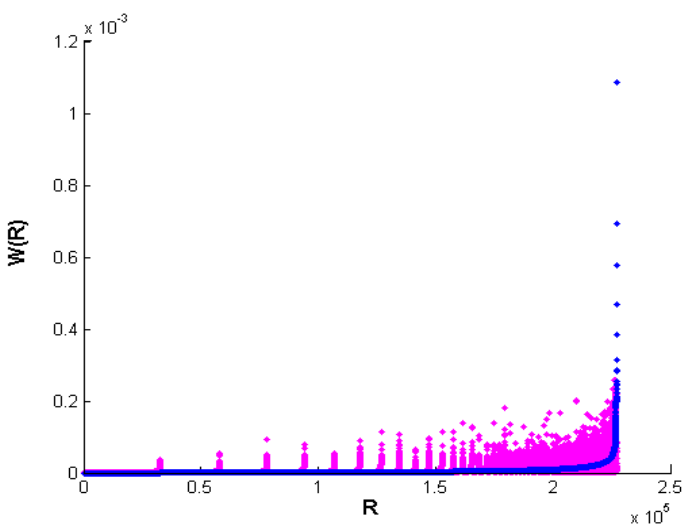


Figura 6.15: Gráfico conjunto do logaritmo da frequência relativa dos ternos de codões em função da ordem, com o logaritmo da sua frequência relativa assumindo a cadeia de Markov, para a *Candida albicans*.

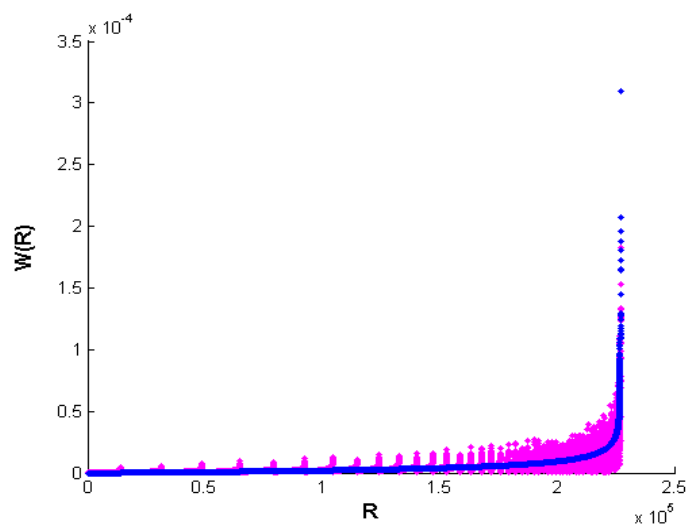


Figura 6.16: Gráfico conjunto do logaritmo da frequência relativa dos ternos de codões em função da ordem, com o logaritmo da sua frequência relativa assumindo a cadeia de Markov, para a *Saccharomyces cerevisiae*.

Por observação das Figuras 6.15 e 6.16 parece não haver ajuste entre as duas distribuições de frequências consideradas. Aplicou-se um teste de ajustamento do qui-quadrado às duas situações anteriores para testar a hipótese:

$$H_0 : \frac{n_{ijk}}{n} = p^*_{ijk} \text{ com } i, j, k \in \{1, ..., 61\},$$

onde $\frac{n_{ijk}}{n}$ é a frequência relativa observada do terno (C_i, C_j, C_k) e p^*_{ijk} é o valor resultante para $P(C_i)P(C_j|C_i)P(C_k|C_j)$ quando se assume a cadeia de Markov com distribuição inicial Π_0 e matriz de probabilidades de transição P^*.
Ao calcular a estatística de teste, χ^2, obtém-se para a *Candida albicans* o valor 1.5197×10^7 e para a *Saccharomyces cerevisiae* o valor 7.8587×10^5. Contudo, existe um grande número de ternos de codões com frequência inferior a 5, o que na prática não é aconselhado para aplicação do teste de ajustamento do qui-quadrado (ver Capítulo 2).
Se o número de observações fosse significativo, não se aceitaría o ajustamento entre as frequências observadas e as esperadas assumindo a cadeia de Markov, para ambas as espécies.
No entanto, assumindo que o modelo probabilístico não se rejeite aquando do estudo feito para $n = 2$ e seja rejeitado para $n = 3$ os dados podem sugerir a existência de correlações de longo alcance no sequenciamento dos codões. Como foi o caso do conjunto das sequências de código da espécie *Saccharomyces cerevisiae*.
Nesta altura surgiría uma nova questão: quão longa é a correlação entre codões? Esta questão, apesar de interessante para o projecto em desenvolvimento, fica por resolver em virtude de não terem sido encontradas metodologias apropriadas.

Não se rejeita a hipótese da correlação entre codões sequenciados ser de longo alcance!

6.5.4 Realce dos n-uplos mais Frequentes

Das Figuras 6.13 e 6.14 é visível que alguns ternos de codões são muito mais frequentes face aos restantes. Na Tabela 6.2 registaram-se os primeiros dez ternos de codões mais frequentes.

Candida albicans		*Saccharomyces cerevisiae*	
3-uplo	fr. absoluta	3-uplo	fr. absoluta
CAA CAA CAA	3667	GAT GAT GAT	613
GAA GAA GAA	2337	GAT GAA GAT	580
AAT AAT AAT	1947	GAT GAA GAA	533
GAT GAT GAT	1580	AAT AAT AAT	511
AAC AAC AAC	1299	GAA GAT GAT	489
GAA GAA GAT	1061	CAA CAA CAA	484
GAT GAA GAT	960	GAA GAT GAA	427
GAT GAA GAA	953	GAT GAT GAT	384
GAA GAT GAT	852	AAT GAA GAA	379
GAA GAT GAA	833	GAA GAA AAA	378
Total de ternos	3369709	Total de ternos	2955798

Tabela 6.2: Ternos de codões mais frequentes.

De seguida apresenta-se o mesmo tipo de realce para os codões e pares de codões mais frequentes, Tabelas 6.3 e 6.4.

2-uplo	fr. absoluta	1-uplo	fr. absoluta
GAA GAA	13439	AAA	178774
CAA CAA	11716	GAA	166537
GAT GAA	10702	AAT	162174
ATT ATT	10417	GAT	149726
GAT GAT	10293	ATT	133651
TTG AAA	9484	TTA	129190
AAT GAA	9202	CAA	122906
ATT GAA	9084	TTG	114361
AAT GAT	9056	TTT	104057
ATT GAT	8966	TCA	93839
Total de pares	3388475	Total	3397032

Tabela 6.3: Codões e pares de codões mais frequentes na *Candida albicans*.

2-uplo	fr. absoluta	1-uplo	fr. absoluta
GAA GAA	8509	GAA	134712
GAT GAA	7118	AAA	126732
GAA AAA	6188	GAT	111895
GAT GAT	6170	AAT	108149
AAA GAA	5964	AAG	89806
GAA GAT	5865	ATT	89369
AAG AAA	5861	CAA	79845
AAT GAA	5833	TTT	79264
GAA AAT	5760	TTG	78755
AAA AAA	5591	TTA	78325
Total de pares	2961829	Total	2974737

Tabela 6.4: Codões e pares de codões mais frequentes na *Saccharomyces cerevisiae*.

Em forma de conclusão, relativamente às Tabelas 6.2, 6.3 e 6.4, pode afirmar-se que, para ambas as espécies, os ternos de codões mais frequentes, justapostos no sequenciamento, são constituídos por codões "parecidos". Neste contexto, dois codões dizem-se parecidos se diferem no máximo de um nucleótido.
Observe-se que para a espécie *Candida albicans* o codão mais frequente é AAA, contudo não é relativamente muito frequente observar dois codões AAA justapostos ou espaçados de um codão, Tabelas 6.2 e 6.3.
Observe-se ainda que, de modo geral, a frequência dos codões terminal são as menores do conjunto das frequências dos codões. No entanto, no genoma da *Candida albicans*, o codão CGC tem frequência inferior a um dos codões terminais.

6.5.5 Análise Comparativa das Frequências entre as Espécies

As espécies em estudo provêm de um ancestral comum. Comparando as distribuições das frequências relativas tanto no contexto dos aminoácidos como dos codões observa-se um padrão semelhante (ver Figuras 6.17 e 6.18).
Nos gráficos das Figuras 6.17 e 6.18 encontram-se representadas as frequências relativas de ambas as espécies, as circunferências a azul representam as frequências relativas da *Candida albicans* e os triângulos vermelhos as frequências relativas da *Saccharomyces cerevisiae*.
Observa-se que na *Candida albicans* os valores de frequências relativas para codões que terminam em A e T são, em geral, mais elevados que na *Saccharomyces cerevisiae*. A situação inverte-se para os codões que terminam em G e C.
No entanto, no caso dos aminoácidos o comportamento, em termos de frequências relativas, das duas espécies é relativamente semelhante.
Com base nos resultados obtidos conclui-se que:

A biforcação de uma espécie nas duas espécies distintas Candida albicans e Saccharomyces cerevisiae manteve praticamente inalterada a mesma proporção de produção de aminoácidos.

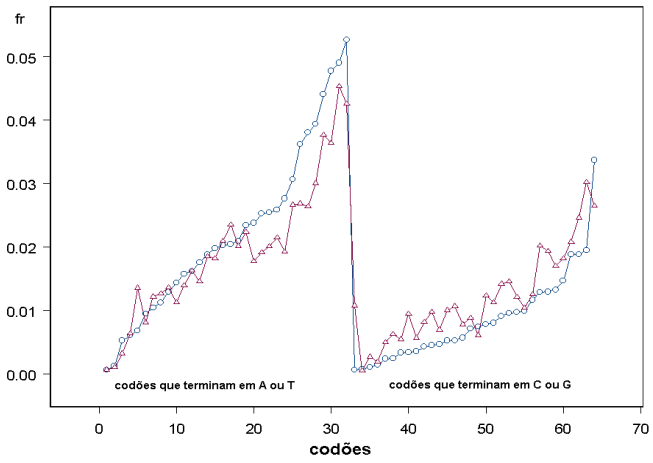


Figura 6.17: Representação conjunta das frequências relativas dos codões das espécies *Candida albicans* e *Saccharomyces cerevisiae*.

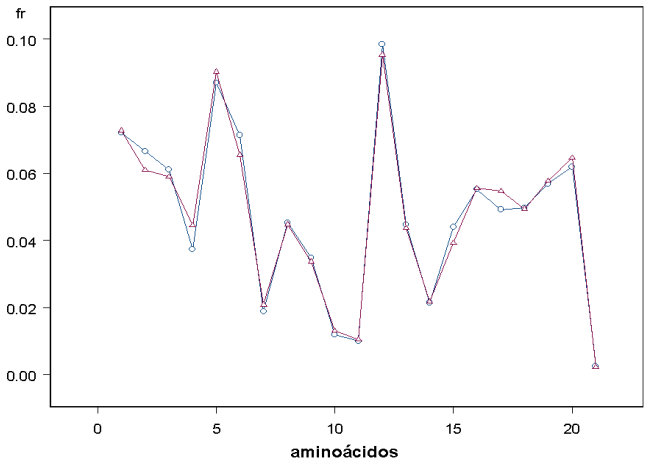


Figura 6.18: Representação conjunta das frequências relativas dos aminoácidos das espécies *Candida albicans* e *Saccharomyces cerevisiae*.

Capítulo 7

Critério de Informação Bayesiana em Cadeias de Markov de Ordem k

7.1 Introdução

Um dos grandes objectivos do presente trabalho é o de dar respostas quanto ao comportamento padrão dos codões na sequência de códigos do texto genético.
Pelos resultados obtidos nos Capítulo 5 e 6 em que não se rejeita a existência de correlações de longo alcance (análise de Zipf), nem a hipótese da cadeia ser de Markov de ordem 1, parece ainda aceitável assumir para o sequenciamento dos codões não terminais nem iniciais um modelo de Markov, de ordem não necessariamente 1 mas k, com k a definir.
Neste tópico ir-se-á supôr que o sequenciamento dos codões nos genes segue uma cadeia de Markov de ordem k e o objectivo será o de estimar o valor de k para o qual a respectiva cadeia de Markov melhor se ajuste à sequência. O método de estimação a considerar será o Critério de Informação Bayesiana.
Esta metodologia envolve grande esforço computacional no cálculo de estimativas para sequências longas. Assim, a aplicação prática será feita sobre genes aleatoriamente escolhidos estimando-se a ordem que melhor se ajusta a cada um deles.

7.2 Cadeia de Markov de Ordem k

Diz-se que um processo estocástico $\{X_n, n \in \mathbb{N}_0\}$ é uma cadeia de Markov de ordem k se a distribuição de probabilidade de uma determinada observação dada a sequência das observações anteriores é igual à distribuição de probabilidade dessa observação dadas as k últimas observações anteriores. Concretamente,

$$P(X_n = x_n | X_{n-1}, X_{n-2}, ..., X_0) = P(X_n = x_n | X_{n-1}, X_{n-2}, ..., X_{n-k}) \quad \forall_{n \in \mathbb{N}_0} \quad \forall_{x_n \in E} \quad (7.1)$$

com E o espaço de estados.

Se o modelo ajustado à sequência dos codões fosse uma cadeia de Markov de ordem k, tal significaria que em termos de probabilidade para uma dada sequência de codões, o codão fixo dependeria, em probabilidade, apenas dos codões presentes até à k-ésima posição que lhe antecede. Observe-se que numa cadeia de Markov de ordem 1, dada a sequência de codões até

à $(i-1)$-ésima posição, ter-se-ía, em termos probabilísticos que o codão presente na posição i depende, não de toda a sequência de codões nas posições anteriores, mas apenas do codão presente na posição $i-1$.
É de notar que para uma sequência de codões independentes a cadeia de Markov que se ajusta tem ordem 0, mas o recíproco não se verifica.
O problema que se coloca é o de averiguar a ordem da cadeia de Markov que melhor se ajusta às sequências dos codões, supondo que a sequência tem comportamento Markoviano. Para resolução deste problema recorrer-se-á ao trabalho desenvolvido em [26] onde se define uma extensão do método da máxima verosimilhança que permite a estimação da ordem da cadeia de Markov que melhor se ajuste à sequência em estudo. Esse método é conhecido habitualmente na literatura por, *Critério de Informação Bayesiana* (BIC - *Bayesian Information Criterion*).

7.3 Critério de Informação Bayesiana (BIC)

O Critério de Informação Bayesiana (BIC) é um método que, a partir de uma amostra, estima o número de parâmetros do modelo. Neste caso específico, de uma sequência de símbolos, estima a ordem da cadeia de Markov que melhor se ajusta à sequência, no sentido da máxima verosimilhança.
Este método foi inicialmente introduzido por Schwarz num contexto geral de estimação do número de parâmetros de modelos gerais. No caso particular do modelo ser uma cadeia de Markov de ordem k desconhecido, o critério traduz-se na descoberta do valor de k que minimiza $BIC(k)$ dado pela equação 7.2.

$$BIC(k) = -\ln L(k) + (A-1) \times A^k \ln n_k \tag{7.2}$$

sendo A o cardinal do espaço de estados, L a função de verosimilhança assumindo que a cadeia é de Markov de ordem k e n_k é o número de "palavras" de tamanho $k+1$ de uma trajectória com n símbolos. Observe-se que $n_k = n - k$.
Um dos teoremas relacionados com este critério, e referenciado por vários autores como por exemplo em [4], é o teorema da consistência do estimador de k dado pelo BIC.

Teorema 7.3.1 *Para qualquer processo de Markov estritamente estacionário e irredutível,*

$$arg(min(BIC(k))) \tag{7.3}$$

é um estimador quase certo da ordem do processo.

7.4 Aplicação ao Caso em Estudo

7.4.1 Concretização da Notação para Sequências de Codões

Seja o processo estocástico $\{X_n, n \in \mathbb{N}_0\}$, onde X_n representa o codão na n-ésima posição no sequenciamento dos codões no genoma da espécie em causa. O espaço de estados, E, é o

conjunto dos codões não terminais; portanto,

$$E = \{AAA(C_1), AAC(C_2), ..., TTT(C_{61})\} \backslash \{TAA(C_{62}), TAG(C_{63}), TGA(C_{64})\}. \qquad (7.4)$$

Assim, $A = 61$ em (7.2), uma vez que existem 61 codões não terminais distintos.
A matriz das probabilidades de transição de uma cadeia de Markov de ordem 1 terá, para os 61 estados possíveis, 60×61 graus de liberdade. No caso da ordem da cadeia ser k ter-se-ão 60×61^k graus de liberdade, já que a dimensão da matriz é 61×61^k. No caso geral o factor $(A-1) \times A^k$ em (7.2), representa o número de graus de liberdade associados à matriz das probabilidades de transição de uma cadeia de Markov de ordem k em que o espaço de estados tem A elementos.
Na Figura 7.1 encontram-se, a título de exemplo, duas tabelas correspondentes às ordens 1 e 2, respectivamente, em que o número de rectângulos a verde constitui o número de parâmetros independentes e logo o número de graus de liberdade.

a)

	C_1	...	C_{61}	total
C_1	$p_{1,1}$	...	$p_{1,61}$	1
. . .	. . .	. . .	. . .	. . .
C_{61}	$p_{61,1}$	...	$p_{61,61}$	1

b)

	C_1	...	C_{61}	total
$C_{1,1}$	$p_{1,1,1}$	...	$p_{1,1,61}$	1
. . .	. . .	. . .	. . .	. . .
$C_{1,61}$	$p_{1,61,1}$	...	$p_{1,61,61}$	1
. . .	. . .	. . .	. . .	. . .
$C_{61,61}$	$p_{61,61,1}$	...	$p_{61,61,61}$	1

Figura 7.1: Matriz das probabilidades de transição realçando o número de parâmetros independentes para uma cadeia de Markov de ordem k (**a)** $k = 1$, **b)** $k = 2$).

Para um número fixo de estados, observe-se que quanto maior for a ordem da cadeia de Markov, maior é o número de graus de liberdade.
Assumindo que a sequência dos codões se identifica com uma cadeia de Markov de ordem k, a distribuição de probabilidade do codão na posição n dada a sequência de codões que lhe antecede, define-se apenas à custa dos k codões anteriores de modo que:

$$P(X_n = C_{i_n} | X_{n-1} = C_{i_{n-1}}, ..., X_0 = C_{i_0}) =$$

$$= P(X_n = C_{i_n} | X_{n-1} = C_{i_{n-1}}, ..., X_{n-k} = C_{i_{n-k}}), \quad \forall_{C_{i_0}, ..., C_{i_n} \in E} \ \forall_{n,\, k \in \mathbb{N}} \ \ n \geq k. \qquad (7.5)$$

Assumindo o comportamento Markoviano, o objectivo é estimar a ordem da cadeia de Markov que melhor se ajuste à sequência dos codões para tal recorrer-se-á ao critério BIC, sendo a estimativa dada pelo valor da expressão (7.3).
A trajectória (a amostra) a considerar na estimação do k, será uma parte codificada do sequenciamento, mais precisamente de um gene, escolhido aleatoriamente no genoma. Em cada uma das trajectórias a aplicar o critério, retirar-se-á o codão de iniciação e o codão de terminação.
Sob a hipótese de que a cadeia de Markov é de ordem k, a função de verosimilhança relativa ao sequenciamento do gene observado é o produto das probabilidades dos primeiros k codões com as probabilidades dos sucessivos grupos de $k+1$ codões. Portanto, a expressão da verosimilhança para cadeias de Markov de ordem k é dada por:

$$L(k) = P(c_{i_1}, c_{i_2}, ..., c_{i_k})\Pi[P(c_{i_{k+1}}|c_{i_1}, c_{i_2}, ..., c_{i_k})]^{n_{c_{i_1},c_{i_2},...,c_{i_{k+1}}}}, \tag{7.6}$$

onde o produtório é sobre todas as possíveis combinações dos $k+1$ codões sucessivos e tem-se $L(k) > 0$.
Nesta altura surge o problema de que estimativa usar para as distribuições de probabilidades consideradas em (7.6). Se tivermos em conta as frequências obtidas para cada gene, o comprimento não é suficiente para averiguar correlações de longo alcance. Por outro lado, se se considerarem as frequências obtidas a nível do genoma tem-se o problema de assumir que genes diferentes estão distribuídos de acordo com a mesma cadeia de Markov, o que não parece biologicamente correcto. No presente trabalho calcular-se-á (7.6) para $k = 0$, 1 e 2 assumindo que a distribuição de probabilidades é estimada de dois modos distintos: pelos valores de frequências relativas da trajectória (gene) e a partir dos resultados obtidos utilizando as frequências relativas no genoma.
Observe-se também que para $k = 2$,

$$P(X_t = C_{i_0}, X_{t-1} = C_{i_1}|X_{t-1} \neq C_{i_1}, X_{t-2} = C_{i_2}) = 0.$$

Assim sendo, é possível reduzir uma cadeia de Markov de ordem 2 a uma cadeia de ordem 1 onde o espaço de estados coincide com o produto cartesiano $E \times E$. Raciocínio análogo pode ser feito para cadeias de Markov de ordem superior a 2.

7.4.2 Concretização para Genes

Os cálculos, apesar de conceptualmente simples são bastante morosos, razão pela qual foi desenvolvido o programa *BICGene* usando a linguagem C++, construído no âmbito desta dissertação. Este programa, com base num ficheiro de texto contendo um único gene, produz um ficheiro de saída contendo os resultados da aplicação do Critério de Informação Bayesiana. Neste estudo usaram-se apenas a dois genes aleatoriamente selecionados designados por, gene $C1$ e gene $S1$ das espécies *Candida albicans* e *Saccharomyces cerevisiae*, respectivamente (Tabela do Apêndice C.1 e Tabela do Apêndice C.2 respectivamente).
Assumindo o comportamento Markoviano para o sequenciamento dos codões nos genes estimar-se-á, à custa do BIC, a ordem da cadeia de Markov que melhor se ajusta à sequência de codões quando se eliminam o codão terminal e inicial de cada um dos genes, $C1$ e $S1$ respectivamente.

Em primeiro lugar, apresentam-se os resultados obtidos por uso das frequências relativas no gene como estimativas das probabilidades. Nas Tabelas 7.1 e 7.2 apresentam-se os resultados dos cálculos das estimativas do logaritmo da verosimilhança e do BIC, para as ordens 0, 1 e 2.

k	n_k	$\ln L(k)$	$BIC(k)$
0	457	-120.709	488.19
1	456	-775.581	23183.9
2	455	-216.181	1.37×10^{11}

Tabela 7.1: Estimativas da ordem da cadeia de Markov geradas pelo gene $C1$ da espécie *Candida albicans*, utilizando estimativas das probabilidades obtidas à custa do gene.

k	n_k	$\ln L(k)$	$BIC(k)$
0	712	-264.5	658.585
1	711	-1459.21	25493.2
2	710	-182.053	1.47×10^{11}

Tabela 7.2: Estimativas da ordem da cadeia de Markov geradas pelo gene $S1$ da espécie *Saccharomyces cerevisiae*, utilizando estimativas das probabilidades obtidas à custa do gene.

De seguida apresentam-se os resultados obtidos por uso das frequências relativas no genoma como estimativas das probabilidades.
Nas Tabelas 7.3 e 7.4 apresentam-se os resultados dos cálculos das estimativas do logaritmo da verosimilhança e do BIC.

k	n_k	$\ln L(k)$	$BIC(k)$
0	457	-1746.292	1930.098
1	456	-1725.010	12933.181
2	455	-1759.739	685213.612

Tabela 7.3: Estimativas da ordem da cadeia de Markov geradas pelo gene $C1$ da espécie *Candida albicans*, utilizando estimativas do genoma.

Os resultados obtidos tanto por uso de estimativas com base no genoma como com base no gene foram concordantes. Em qualquer dos genes, a ordem estimada pelo BIC é 0!

k	n_k	$\ln L(k)$	$BIC(k)$
0	712	-2837.425	3034.467
1	711	-2812.715	14829.725
2	710	-2857.123	735737.652

Tabela 7.4: Estimativas da ordem da cadeia de Markov geradas pelo gene $S1$ da espécie *Saccharomyces cerevisiae*, utilizando estimativas do genoma.

Comparando os resultados obtidos apenas para as ordens 1 e 2, estima-se melhor ajuste da cadeia de Markov de ordem 1 face à cadeia de ordem 2. Contudo, como se pode averiguar dos resultados apresentados, é melhor assumir a cadeia de Markov de ordem zero. No entanto não se verifica a independência para o conjunto das sequências de código do genoma.
No Capítulo 2 efectuaram-se testes para a independência relativos ao genoma e não se averiguou quanto à independência de codões justapostos dentro de cada gene. Perante a análise feita classificaram-se também cada um destes genes em tabelas de contingência de pares de codões, no sentido de aplicar um teste de ajustamento quanto à independência entre codões justapostos. Nestas tabelas verifica-se que a frequência de indivíduos na maior parte das categorias é inferior a 5 indivíduos. Todavia, se não se atender ao facto das frequências serem muito pequenas o teste de ajustamento do qui-quadrado resulta na independência para os genes $C1$ e $S1$ em estudo!
Tal pode não causar espanto se se recordar que a associação observada entre codões justapostos é fraca.

Capítulo 8

Teoria da Informação

8.1 Introdução

A Teoria da Informação e em particular a Teoria Matemática da Comunicação, foi desenvolvida inicialmente por Claude Shannon na década de quarenta do século passado, [24], com o objectivo de optimizar o sistema telefónico. A motivação para o seu desenvolvimento foi a necessidade de maximizar a quantidade de informação que pode ser transmitida por um canal de comunicação imperfeito, isto é, um canal que pode introduzir erros nas mensagens transmitidas.
Perante um canal de comunicação imperfeito, uma das formas do receptor poder detectar e corrigir erros, durante a descodificação da mensagem, é esta conter redundância. Como a capacidade máxima do canal é fixa, a existência de redundância faz com que a quantidade de informação efectiva e transmitida seja inferior à capacidade do canal.
Um dos objectivos importantes de Shannon era o de determinar as taxas teóricas máximas para compressão dos dados; entenda-se por *compressão* a remoção de redundância da mensagem. O ideal seria codificar a mensagem de forma a ocupar o mínimo espaço possível, mas contendo um valor de redundância mínimo que ainda assim permitisse detectar e corrigir os erros que possam comprometer seriamente ou deturpar completamente a descodificação da mensagem. Para resolver esta questão, Shannon desenvolveu o conceito de *entropia*.
Na Figura 8.1 é ilustrado um diagrama de blocos com os intervenientes típicos envolvidos no envio de mensagens entre um emissor e um receptor. Na mesma figura é estabelecido um paralelismo entre um sistema de comunicação e a produção de proteínas desde o DNA.
No "sistema de comunicação genético" é desconhecida a mensagem original e o codificador, uma vez que as sequências de código já estão armazenadas no núcleo das células de todos os indivíduos. Sempre que necessário a célula activa o núcleo por forma a produzirem "cópias" de partes das sequências de código, o mRNA (canal), a serem descodificadas pelo ribossoma em proteínas.
O objectivo geral é a decifração de leis gerais da sequência de código (mensagem codificada). Utilizar-se-á a Teoria Matemática da Comunicação como nova metodologia no sentido de contribuir para esse estudo. Assim, neste capítulo apresentar-se-ão, sobre dados discretos, um conjunto de medidas probabilísticas no contexto da Teoria da Informação. Seguir-se-á o cálculo dessas medidas ao conjunto das sequências de código no genoma da *Candida albicans* e da *Saccharomyces cerevisiae*.

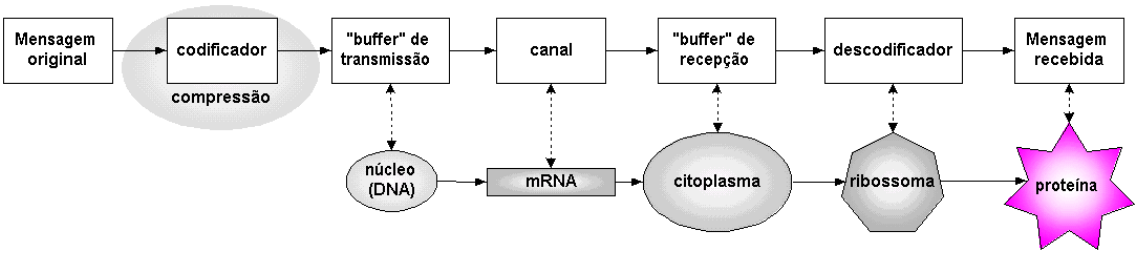


Figura 8.1: Esquema de transmissão de mensagens.

8.2 Entropia

A entropia de uma variável aleatória pode ser interpretada como o grau de não informação dado pela observação da variável. Quanto "mais casual" for a variável maior será a entropia e menor o conhecimento sobre a variável.
O problema põe-se em medir o não conhecimento sobre um conjunto de acontecimentos possíveis de uma dada variável.
Para uma dada experiência aleatória, se a probabilidade de um dado acontecimento for 0.999 é quase certo que o acontecimento ocorrerá. Se a probabilidade de um dado acontecimento for 0.001 é quase certo que o acontecimento não ocorrerá. A incerteza é máxima quando a probabilidade do acontecimento for 0.5.
Contudo, o problema não se põe com um só acontecimento, mas com um conjunto de acontecimentos possíveis e mutuamente exclusivos que podem ocorrer sobre um mesmo espaço de probabilidades. Para medir o desconhecimento que se tem sobre o comportamento de uma variável aleatória poder-se-á recorrer à entropia.
Seja $p(x)$ a função de probabilidade da variável aleatória X, com X definida sobre um espaço de estados discreto E. A entropia (H) da variável aleatória discreta X, é definida por:

$$H(X) = -\sum_{x \in E} p(x) \log_2 p(x) \tag{8.1}$$

A entropia também pode ser descrita como um valor esperado sobre a forma de:

$$H(X) = \mathbf{E}\{-\log_2 p(X)\}.$$

Assim, a entropia pode ser vista como a incerteza média de uma variável aleatória.
A entropia também pode ser vista como o comprimento médio da mensagem necessária para transmitir o resultado da variável, o que normalmente é medido em *bits*, justificando o uso do logaritmo de base 2. Em geral, uma codificação óptima envia uma mensagem de probabilidade p com comprimento $[-\log_2 p]$, sendo utilizados menos *bits* nos resultados mais provaveis e mais *bits* nos resultados menos provaveis.

Para uma dada experiência aleatória, com espaço de estados discreto $E = \{x_1, x_2, ..., x_N\}$, dir-se-á que uma variável possui resultados equiprováveis quando assume qualquer dos possíveis estados com igual probabilidade, ou seja, $p(x_i) = \frac{1}{N}$, com $i = 1, 2, ..., N$. Por outro lado, o valor máximo de entropia é atingido aquando da "total casualidade". Assim, da equação 8.1 resulta que a entropia máxima é dada por:

$$H_M = \log_2 N \tag{8.2}$$

Portanto, uma medida que poderá dar informação parcial sobre a complexidade das sequência de codões será a entropia, no sentido em que quanto maior for a entropia "mais casual" é a variável codão na sequência de código.

8.3 Conceitos e Propriedades

De seguida apresentar-se-ão alguns conceitos do contexto da entropia, que pareceram de utilidade na aplicação aos textos genéticos em estudo.
Seja (X, Y) um par de variáveis aleatórias discretas, com espaço de estados $E_1 \times E_2$ e com distribuição de probabilidade conjunta dada por $p(x, y)$.

Definição 8.3.1 *A entropia conjunta do par (X, Y) é a incerteza média sobre o par de variáveis aleatórias e é dada por*[1]*:*

$$H(X, Y) = -\sum_{x \in E_1} \sum_{y \in E_2} p(x, y) \log_2 p(x, y). \tag{8.3}$$

Definição 8.3.2 *A entropia condicional de Y, dada uma variável aleatória discreta X, é a informação extra, em média, necessária para "comunicar" com Y dado que o "receptor" conhece X e é dada por:*

$$H(Y|X) = -\sum_{x \in E_1} \sum_{y \in E_2} p(x, y) \log_2 p(x|y). \tag{8.4}$$

onde $p(x|y)$ denota a distribuição de probabilidade de x condicionada a y.

Como consequência imediata das definições anteriores resultam algumas propriedades, nomeadamente:

[1] A expressão da entropia no âmbito da análise de n-uplos tem o seguinte aspecto, para Y a variável símbolo:

$$H(n) = -\sum_{i}^{C^n} p(Y = a_i) \log_2 p(Y = a_i),$$

com C o número de letras do alfabeto do texto em estudo e a_i a i-ésima palavra das C^n.

Propriedades 8.3.1 $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$.

Desta propriedade resulta que:

$$H(X) - H(X|Y) = H(Y) - H(Y|X).$$

Definição 8.3.3 *Chama-se informação mútua entre X e Y ao valor $I(X,Y)$ definido por:*

$$I(X,Y) = H(X) - H(X|Y). \tag{8.5}$$

Propriedades 8.3.2 $I(X,Y) = H(X) + H(Y) - H(X,Y)$.

Observe-se que se assumir que duas variáveis são independentes facilmente se prova que a informação mútua é zero.

Definição 8.3.4 *A redundância da variável aleatória discreta X pode ser definida por $R(X)$ e é, em percentagem, dada por:*

$$R(X) = (1 - \frac{H(X)}{H_M}) \times 100. \tag{8.6}$$

8.4 Aplicação ao Caso em Estudo

No âmbito da situação em estudo, calculou-se a entropia da variável aleatória X_n que representa o codão na posição n e da variável par de codões justapostos para as duas espécies em estudo. Também foram calculados os correspondentes valores de entropia máxima e de redundância.
Os resultados obtidos encontram-se na Tabela 8.1.

Espécie	Variável(Y)	$H(Y)$	H_M	$R(Y)$
Candida albicans	codões	3.843	4.159	7.6
Saccharomyces cerevisiae	codões	3.958	4.159	4.8
Candida albicans	par de codões	7.606	8.270	8.0
Saccharomyces cerevisiae	par de codões	7.885	8.270	4.7

Tabela 8.1: *Tabela de resultados de entropia e redundância.*

Da Tabela 8.1 averigua-se que o desconhecimento que se tem sobre qualquer das variáveis em estudo é muito grande, mas contudo não é total, pois a entropia das variáveis não é máxima. A redundância é pequena, inferior a 10%, observando-se redundância superior na espécie *Candida albicans*.
A título de pequena conclusão poder-se-á dizer que:

O comportamento da variável codão em sequência não é de todo "casual", mas o desconhecimento sobre o seu comportamento é grande.

A linguagem em estudo é definida na realidade pelo código genético, que consiste na mensagem codificada sobre a constituição de proteínas.
Entendendo os dados em estudo, as sequências de codões, como código e tendo-se em conta os resultados obtidos, poder-se-á conjecturar que:

A mensagem genética parece estar construída de forma a ocupar o mínimo espaço possível, contendo, todavia uma redundância mínima necessária para a detecção e correção de erros.

Espécie	$I(X,Y)$
Candida albicans	0.08
Saccharomyces cerevisiae	0.031

Tabela 8.2: *Tabela de resultados de informação mútua.*

Os resultados de informação mútua estão apresentados na Tabela 8.2. Os resultados obtidos são muito próximos de zero, no entanto são diferentes. Este resultado vem mais uma vez confirmar a falha da independência entre dois codões justaposto na sequência já obtida através de outros métodos anteriormente aplicados.

Capítulo 9

Conclusão

Ao longo desta dissertação foram apresentados alguns resultados da aplicação de várias metodologias estatísticas no estudo do genoma de duas espécies distintas: a *Saccharomyces cerevisiae* e a *Candida albicans*. O objectivo principal consistia em retirar regras, a partir do genoma entendido como sequência de codões.
Neste estudo foram utilizados os seguintes métodos e/ou modelos:

- Análise de Tabelas de Contingência;
- Análise Classificatória;
- Análise em Componentes Principais;
- Cadeias de Markov;
- Análise de Zipf;
- Critério de Informação Bayesiana;
- Teoria de Informação;

Através das Tabelas de Contingência de pares de codões do genoma, e usando a estatística de teste χ^2, rejeitou-se a hipótese de independência entre pares no sequenciamento e obtiveram-se valores pequenos de associação. As Análises Classificatória e em Componentes Principais evidenciaram que, em geral, cada codão fixo depende do codão que o antecede (ou precede) podendo ser o nucleótido que se encontra justaposto o mais marcante. No entanto, por desconhecimento de testes apropriados, estas hipóteses não foram validadas.
Compararam-se os valores observados das frequências de cada codão com os resultados obtidos assumindo cadeias de Markov de ordem 1, homogéneas, irredutíveis, aperiódicas e estritamente estacionárias e, de modo geral, averiguou-se a existência de ajuste. No entanto, não se provou se a sequência de codões é bem modelada por uma cadeia de Markov de ordem 1.
Fez-se uma análise das frequências dos símbolos (codões ou aminoácidos) ordenadas por ordem decrescente de frequências aplicando a Análise de Zipf, tanto para estimar leis da distribuição das frequências ordenadas como para explorar a existência de correlações de longo alcance, averigando-se a possibilidade de existência deste tipo de correlações.
O Critério de Informação Bayesiana permitiu averiguar, para a sequência de código de dois genes, que a ordem da cadeia de Markov que melhor se ajusta é a ordem 0. No entanto, não põe de parte a possibilidade de existência de correlações de curto ou longo alcance nas

sequências de código do genoma.
Através da Teoria da Informação calcularam-se valores da redundância e informação mútua, resultando em valores de redundância com percentagem inferior a 10% e em valores de informação mútua pequena mas não nula. Tal análise permitiu confirmar, mais uma vez, a não independência entre pares de codões justapostos.
Dos métodos usados não se obtiveram muitas respostas objectivas. As metodologias aplicadas, no sentido de averiguar se o sequenciamento dos codões é ou não Markoviano e de ordem superior a 1, resultaram em respostas vagas. A resposta quanto à independência foi praticamente concludente embora tudo aponte para uma fraca associação.
Para ambas as espécies estudadas rejeitou-se a independência dos codões no sequenciamento no genoma e não se rejeitou a existência de correlações de alcance superior a um entre os codões. No entanto, existem genes para os quais se ajusta melhor uma cadeia de Markov de ordem 0, do que uma cadeia de ordem superior!
O estudo feito parece também fazer crer que o código genético é um código com redundância suficiente para minimizar os erros de tradução.
Durante a procura de métodos estatísticos apropriados à análise de dados de natureza discreta e sequenciados, e motivados pela natureza do problema do tipo eventualmente Markoviano, estudaram-se, para além dos métodos apresentados, os Modelos de Markov Escondidos (HMM - Hidden Markov Models). No entanto, este método foi posteriormente abandonado pois a sua aplicação não pareceu apropriada ao conjunto dos dados fornecidos inicialmente pela equipa envolvida no projecto. Também foi colocada em questão a adequação desse modelo à natureza do problema e dados existentes.
Também se tentaram aplicar os passeios aleatórios e a análise espectral. No entanto, da forma perliminar como estes métodos foram aplicados, não permitiram obter conclusões direccionadas para os objectivos do projecto.
Um dos problemas colocados neste estudo foi o facto da estatística χ^2 não ser invariante a transformações de escala. O ideal seria encontrar uma estatística para testar a independência que não dependesse da dimensão da amostra.
Seria também útil encontrar métodos para averiguar a influência dos seis nucleótidos que estão mais próximos de um dado codão fixo, permitindo testar as hipóteses que resultaram da Análise Classificatória e da Análise em Componentes Principais, assim como outros testes que permitissem a avaliar o ajustamento de cadeias de Markov ao sequenciamento dos codões, aminoácidos e/ou nucleótidos.
O trabalho descrito nesta dissertação constitui uma das primeiras abordagens que se conhece, no contextos dos codões, no estudo de genomas, usando métodos estatísticos. Os resultados obtidos e as questões deixadas em aberto podem constituir uma boa base de trabalho futuro.

Apêndice A

Apêndice A.1
Métrica:
Seja P e Q dois pontos. A distância entre P e Q, $d(P,Q)$, é um número real que verifica as seguintes condições:

1. *simetria* $\hookrightarrow d(P,Q) = d(Q,P)$
2. *não negatividade* $\hookrightarrow d(P,Q) \geq 0$
3. *exactidão* $\hookrightarrow d(P,Q) = 0$ *sse* $P = Q$
4. *desigualdade triangular* $\hookrightarrow d(P,Q) \leq d(P,R) + d(R,Q)$

Apêndice A.2
Cálculo das probabilidades de cada indivíduo (codão ou aminoácido), assumindo a cadeia de Markov de ordem 1 como modelo teórico. Suponha-se que se tem N indivíduos.

O sistema a resolver (Teorema 5.2.2) é da forma:

$$\begin{cases} \pi_k = \sum_{j=1}^{N} \pi_j p_{jk} \quad k=1,2,\ldots,\ N \\ \sum_{j=1}^{N} \pi_j = 1 \end{cases}$$

O objectivo foi o de transformar este sistema de modo a ser escrito da forma $B\Pi = b$, com $\Pi = [\pi_1 \ldots \pi_N]'$ e B uma matriz $N \times N$ invertível. De seguida apresenta-se a justificação e os passos do algoritmo utilizado na resolução dos quatro sistemas em estudo:

O sistema de $N+1$ equações acima é equivalente a:

$$\begin{cases} \pi_k - \Pi'[p_{1k} \ldots p_{Nk}]' = 0 \quad k=0,1,2,\ldots,\ N \\ \Pi'[1 \ldots 1] = 1 \end{cases}$$

Seja $P = [p_{ij}]$. Então, tem-se:

$$\begin{cases} \Pi'(I - P) = [0 \ldots 0] \\ \Pi'[1 \ldots 1] = 1 \end{cases}$$

$$\begin{cases} (I - P)'\Pi = [0 \ldots 0]' \\ [1 \ldots 1]'\Pi = 1 \end{cases}$$

O sistema tem N incógnitas e N+1 equações. Uma das equações é redundante. Retire-se a penúltima equação.
Na prática, o que se faz é substituir a matriz P por uma matriz P^{**} que resulta da substituição da última coluna de P por a coluna $[-1 \ ... \ -1 \ 0]'$. *Note-se que a última coluna da matriz* $I - P^{**}$ *é* $[1 \ ... \ 1 \ 1]' = [0 \ ... \ 0 \ 1]' - [-1 \ ... \ -1 \ 0]'$.
Assim, o sistema de equações é equivalente a:

$$(I - P^{**})'\Pi = [0 \ ... \ 0 \ 1]'$$

Portanto, o procedimento para cálculo das probabilidades consiste na substituição da última coluna da matriz das probabilidades de transição por uma coluna de elementos iguais a -1 à excepção do último elemento que é 0. Faça-se a diferença entre a identidade e a matriz obtida, determine-se a transposta da matriz diferença e calcule-se a sua inversa. Por fim, faça-se o produto da matriz inversa pela matriz constituida por zeros à excepção do último elemento que é um. O resultado é uma matriz coluna cujos elementos são a solução do sistema.

Apêndice B

Apêndice B.1

	3' AAA	3' AAG	3' AAT	3' AAC	3' ACT	3' ACC	3' ACA	3' ACG	3' CGT	3' CGC	3' CGA	3' CGG	3' AGA	3' AGG	3' TCT	3' TCC	3' TCA	3' TCG	3' AGT	3' AGC	3' ATT	3' ATC	3' ATA	3' ATG	3' TTT	3' TTC	3' TAT	3' TAC	3' TGT	3' TGC	3' TGG	3' TTA
LYS AAA	7172	4737	6495	2842	4003	2011	3061	571	1349	231	1561	488	4046	731	3392	1663	6065	1239	2274	681	5873	3087	3403	2598	5533	2770	5313	1988	1719	382	1995	8677
LYS AAG	7041	2757	3982	1470	1268	614	1287	340	328	62	223	105	1902	397	901	386	1194	506	1167	404	2542	892	1133	971	2496	766	1732	808	383	111	308	1475
ASN AAT	7568	1984	10417	2985	3449	1636	2115	437	979	137	1078	207	2325	397	3400	1573	4482	814	3049	918	5696	1974	2442	2118	4292	2921	4937	1242	1541	349	1930	6674
ASN AAC	4152	2275	4980	4377	1896	1296	1343	376	305	85	248	97	1111	235	1107	732	1529	597	1265	699	2325	960	874	842	1764	545	1584	852	513	177	573	1105
THR ACT	3772	670	3635	1025	3974	1323	2097	328	535	65	415	56	1548	199	2192	843	2897	607	1804	449	2993	863	1357	918	2320	1022	2185	490	936	172	1036	3232
THR ACC	3266	1487	3161	1657	2118	2046	1489	369	114	36	103	25	953	161	361	294	637	166	972	386	2155	829	827	966	405	249	491	297	206	52	223	523
THR ACA	2974	1434	2993	1427	2294	1142	3229	456	227	59	235	79	1654	289	1584	514	2576	607	1258	528	2642	989	1627	1176	2766	847	1203	538	530	137	482	3303
THR ACG	1009	225	478	196	257	182	485	174	53	10	37	23	327	90	322	122	508	170	242	85	534	164	393	257	609	176	350	130	136	30	103	489
ARG CGT	1141	269	727	287	411	153	330	54	386	35	215	44	269	39	374	153	545	84	257	81	628	224	282	263	377	275	536	163	216	55	236	626
ARG CGC	327	164	211	146	41	35	72	23	47	24	35	16	106	25	55	32	97	27	54	50	73	32	76	47	82	42	74	47	39	22	33	72
ARG CGA	546	269	604	293	306	130	286	40	114	54	110	33	529	80	265	142	632	115	223	114	520	215	363	306	692	327	448	226	175	66	195	1157
ARG CGG	759	123	185	80	53	20	69	12	19	12	20	12	85	32	83	39	158	51	54	20	117	54	106	41	228	105	124	79	48	22	58	177
ARG AGA	3789	1881	3145	1593	1631	701	1205	234	491	98	391	110	2964	410	1092	464	1909	356	1074	376	2756	1203	1589	1306	2248	925	1541	761	562	147	708	3181
ARG AGG	1221	308	473	227	156	68	195	59	46	11	65	25	240	87	209	73	308	98	183	76	322	148	276	179	606	171	370	174	119	47	113	392
SER TCT	3163	627	2880	869	2359	771	1674	298	630	79	534	92	925	135	3112	1050	3176	613	1365	368	2488	640	1151	798	1509	883	1516	438	730	139	744	2410
SER TCC	2906	1226	2528	1213	1238	630	750	198	173	43	113	45	626	135	525	411	813	288	845	313	1465	552	663	693	726	332	445	251	198	53	192	719
SER TCA	3743	1443	4091	1610	2905	1063	2886	564	464	96	524	118	2358	356	2742	875	5116	1162	2069	625	4395	1162	2402	1900	3974	1337	1827	689	665	146	699	5037
SER TCG	1698	287	906	262	581	258	624	155	89	12	57	31	450	124	695	244	1151	552	501	163	1042	329	688	431	961	323	551	294	185	79	198	830
SER AGT	2490	597	2886	841	1428	623	921	173	354	70	386	86	809	143	1412	623	1690	324	1786	433	2068	620	1167	763	1876	854	1593	375	637	123	717	2446
SER AGC	1260	618	1457	959	529	372	507	145	65	27	47	29	410	111	422	278	690	230	697	398	617	227	333	255	466	143	403	230	156	59	177	375
ILE ATT	6124	1116	5288	1300	2726	1145	1499	343	820	93	750	131	2591	289	2886	1561	2644	608	2031	424	4840	1511	2087	1840	2763	2762	3623	1101	1453	372	1567	4655
ILE ATC	4335	2007	4015	1866	1913	1236	1068	298	239	54	164	41	1196	199	414	408	634	196	1031	331	2465	913	766	978	862	393	734	461	283	86	335	819
ILE ATA	2428	1320	3287	1193	1732	753	1302	382	330	88	392	137	1064	288	1699	707	2173	701	977	322	2754	1027	1708	1329	2865	998	1649	595	666	189	622	3185
MET ATG	3614	1300	3126	967	1341	599	1631	408	189	37	120	61	1475	289	1317	522	1980	595	1332	331	2750	769	1326	1685	2152	635	1583	580	598	153	570	1704
PHE TTT	4128	849	3754	972	1773	859	1383	254	298	60	352	102	1695	254	2246	1220	2515	557	1521	406	4362	1244	1928	1762	2356	3306	2876	887	1094	377	1375	4130
PHE TTC	4389	1665	3990	1592	1908	958	1137	262	498	86	359	86	1094	164	558	486	634	232	1133	286	2302	712	851	801	1680	894	915	586	511	122	468	1243
TYR TAT	3847	970	3944	1021	1771	756	950	214	833	89	751	123	1200	167	1924	893	2069	472	1111	269	3362	1122	1194	1287	2790	1896	3082	847	1146	260	1120	3904
TYR TAC	1804	1325	1847	1077	981	568	595	221	303	64	160	58	642	109	447	330	510	280	501	215	1490	455	482	558	1039	271	974	578	372	122	431	467
CYS TGT	1268	302	1128	373	638	247	425	94	402	66	339	84	214	46	749	319	813	191	272	86	1388	394	595	547	1173	663	807	268	757	196	417	1470
CYS TGC	539	279	418	267	159	131	165	62	71	32	52	13	152	47	155	94	188	112	180	95	304	139	178	141	439	142	245	141	169	99	166	232
TRP TGG	2034	660	1680	544	610	266	694	140	151	16	102	49	950	160	653	220	1003	234	549	143	1232	409	738	578	1354	440	986	368	610	123	623	1544
LEU TTA	4897	2149	5244	1863	3373	1273	2115	501	984	110	885	176	2797	345	2935	1207	4716	1151	1942	378	5378	1666	2025	2320	4661	1755	3399	1269	1375	202	1314	6967
LEU TTG	9484	2603	6696	2230	3026	962	2656	710	579	109	343	167	2471	495	1545	627	1961	895	2502	561	5937	1613	2240	2057	3369	1166	2069	1081	901	257	589	3040
LEU CTT	1324	324	1301	425	645	284	553	116	310	60	362	111	483	99	1016	514	1098	298	472	116	1071	328	560	460	766	947	1075	501	550	160	536	1264
LEU CTC	969	555	893	549	277	182	305	87	70	43	69	27	199	54	176	137	178	107	230	97	441	224	233	205	315	172	158	118	90	54	91	173
LEU CTA	860	446	873	518	392	191	482	138	120	64	198	74	280	109	381	217	582	230	343	145	625	332	666	398	731	323	521	251	181	78	186	802
LEU CTG	1121	385	755	264	306	128	396	88	79	21	61	20	264	76	313	101	486	184	276	89	621	183	433	334	528	193	301	136	83	52	84	387
PRO CCT	2552	432	2074	479	1115	359	1024	145	379	58	325	63	629	87	1038	468	1576	355	782	196	1416	381	854	435	670	430	1312	349	376	97	481	1164
PRO CCC	1476	525	1172	508	450	278	491	114	81	22	68	15	250	63	186	101	222	88	339	147	751	279	461	423	396	133	252	86	89	27	82	281
PRO CCA	3162	1415	3125	1408	2320	928	2458	418	348	72	351	97	1848	232	1566	503	2481	571	1372	413	3354	997	2085	1464	2980	1092	1553	690	430	106	614	3476
PRO CCG	732	138	301	131	179	73	282	51	45	13	30	16	157	35	175	68	401	130	111	55	363	138	243	155	402	163	274	102	80	29	82	455
HIS CAT	2295	547	2266	658	1056	379	746	104	635	92	463	97	827	131	1110	488	1556	317	718	181	1849	564	852	719	1546	1048	1796	531	676	150	646	2546
HIS CAC	1094	781	1101	771	457	281	429	114	135	80	95	50	398	105	278	144	270	104	309	126	817	250	293	273	365	117	293	209	124	54	127	189
GLN CAA	4531	2415	4267	1946	2407	1154	2288	391	765	148	842	211	2572	358	1935	756	3396	705	1748	617	3944	1692	2178	2007	3784	1562	3079	1075	961	254	1166	5064
GLN CAG	1574	562	1327	483	364	147	437	126	119	32	107	41	421	119	368	176	579	225	354	122	773	278	500	386	1150	297	785	364	200	44	143	687
VAL GTT	3799	787	2789	898	1947	755	1010	228	588	61	389	86	1537	213	2211	1086	1911	473	1425	273	3093	872	1160	1140	1427	1918	2461	827	1021	216	941	3128
VAL GTC	2263	1246	2349	1088	1377	735	764	214	112	18	72	21	665	125	311	230	368	143	714	218	1594	427	498	614	493	185	390	296	141	51	132	421
VAL GTA	860	687	1114	559	754	390	715	222	202	52	192	62	596	152	1004	358	1439	421	444	176	1094	387	680	755	1861	585	926	408	380	99	479	2149
VAL GTG	2286	620	1391	496	760	245	946	226	157	29	80	44	645	159	615	245	724	383	534	168	1779	462	935	705	1790	396	750	337	342	101	289	1144
ALA GCT	3663	714	2725	781	2321	717	1233	204	533	45	316	60	1570	136	1761	638	1635	350	1465	243	2580	657	1071	871	1129	731	1840	580	821	129	743	2135
ALA GCC	2935	1280	2554	1136	1550	805	874	249	135	21	82	20	955	143	366	219	429	189	839	190	2307	603	759	1111	741	256	445	197	201	33	165	642
ALA GCA	2083	1030	2115	949	1471	603	1588	262	256	40	207	66	1532	225	1266	475	1845	515	945	286	2217	719	1012	981	2527	865	1130	511	533	145	492	3362
ALA GCG	438	83	209	97	138	70	199	62	16	9	18	10	135	51	169	73	307	135	89	29	337	103	165	172	407	148	239	123	118	30	89	392
GLY GGT	4590	1037	3667	1230	2345	826	1218	240	463	50	298	60	1855	141	1599	625	1764	235	1486	260	3401	946	1293	1230	1045	825	1913	638	880	144	999	2158
GLY GGC	1008	510	1032	539	463	202	413	116	117	24	56	15	353	76	459	159	426	138	429	181	493	168	256	209	268	112	374	230	171	44	233	298
GLY GGA	1248	558	1718	794	1094	476	1053	152	239	43	168	52	1316	144	1273	440	1925	334	950	347	1617	636	1110	865	2462	922	1557	645	605	139	545	3449
GLY GGG	2478	462	1231	441	401	146	524	83	51	9	40	19	391	64	548	228	713	226	291	99	1064	365	670	459	1956	673	822	358	283	82	190	1437
ASP GAT	6969	1777	6891	1793	2939	1296	1992	388	842	78	666	146	2369	379	3055	1356	3648	734	2232	600	6242	1979	2842	2205	4505	2749	4939	1348	1388	263	1639	6921
ASP GAC	2582	1630	2761	1486	1114	705	769	228	255	44	111	45	764	177	854	451	823	455	763	320	1763	546	661	690	1305	315	975	580	265	96	367	569
GLU GAA	8038	4218	7281	3232	4156	2013	3024	648	797	106	766	206	3971	595	2650	1169	4171	952	2969	792	6516	2796	3174	2848	4819	1824	4051	1500	1273	268	1577	6929
GLU GAG	2963	1095	2445	737	735	369	743	246	146	29	119	60	652	221	782	315	870	394	816	257	1834	530	856	656	2234	534	1362	583	280	78	263	1307
STOP TAA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STOP TAG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STOP TGA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	178771	66085	162178	64010	86381	39584	67101	14555	20460	3513	17717	4743	68812	11497	69238	30845	93836	24221	59666	17747	133651	44960	63770	55481	104055	49849	85758	32397	32161	7939	33698	129189
	2E+05	66087	2E+05	64010	86381	39585	67102	14556	20460	3513	17717	4743	68813	11498	69239	30846	93839	24221	59667	17747	1E+05	44961	63773	63985	1E+05	49850	85761	32397	32161	7940	33699	1E+05
	AAA	AAG	AAT	AAC	ACT	ACC	ACA	ACG	CGT	CGC	CGA	CGG	AGA	AGG	TCT	TCC	TCA	TCG	AGT	AGC	ATT	ATC	ATA	ATG	TTT	TTC	TAT	TAC	TGT	TGC	TGG	TTA

Figura B.1: Tabela de contingência da *Candida albicans*.

	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'
	TTG	CTT	CTC	CTA	CTG	CCT	CCC	CCA	CCG	CAT	CAC	CAA	CAG	GTT	GTC	GTA	GTG	GCT	GCC	GCA	GCG	GGT	GGC	GGA	GGG	GAT	GAC	GAA	GAG	TAA	TAG	TGA
LYS AAA	7240	2250	801	2030	1305	2135	775	4969	559	2966	1078	7297	1570	4633	1523	2200	2127	3124	1291	2279	333	3311	452	1606	889	7886	2111	8725	2643	363	234	115
LYS AAG	2279	714	236	608	345	488	204	1117	166	828	380	1932	533	1001	408	521	586	995	459	1618	266	862	262	894	654	2226	764	3398	1150	135	75	40
ASN AAT	4991	1791	474	1046	699	1972	580	3041	361	2804	655	6099	871	4462	1623	1450	1642	3466	1954	2228	378	4487	1107	3238	1648	9056	2380	9202	2078	169	91	166
ASN AAC	2635	668	282	402	271	992	494	1741	260	1102	764	2324	783	699	318	268	351	428	354	384	75	813	235	505	301	1899	931	1878	848	102	42	52
THR ACT	2543	619	155	307	261	1387	349	2355	303	1523	357	2681	399	2486	760	710	896	2921	1126	1516	226	3394	438	1492	833	4061	973	4366	895	25	24	41
THR ACC	606	366	112	143	164	584	216	1026	160	616	270	1068	283	759	232	220	353	353	278	273	57	551	103	263	227	1273	470	1014	493	21	13	13
THR ACA	1966	610	211	430	337	1083	376	2495	297	641	307	1592	425	1491	407	874	735	963	407	1496	157	827	232	915	478	1711	642	2385	705	52	33	22
THR ACG	526	116	45	113	52	138	68	323	94	99	48	240	73	507	111	269	321	236	125	505	89	315	74	261	281	321	122	527	154	19	6	12
ARG CGT	521	273	70	96	75	301	103	563	55	567	155	1006	168	526	125	154	163	540	271	316	45	683	140	387	203	1220	320	1335	276	30	18	20
ARG CGC	73	34	25	32	34	47	29	57	25	92	85	152	79	23	18	28	22	18	15	18	3	40	8	37	23	97	43	65	53	7	4	1
ARG CGA	515	188	92	137	102	139	56	543	49	217	131	856	165	430	165	330	203	324	162	427	30	292	109	344	110	640	262	871	220	21	5	7
ARG CGG	185	52	28	43	32	23	24	93	16	58	21	89	44	86	35	54	49	46	29	87	23	48	26	57	24	114	60	176	65	5	4	3
ARG AGA	2203	640	205	487	279	560	222	1489	170	966	468	2299	482	1682	531	857	704	1377	645	1121	108	1596	291	882	363	2577	820	3036	781	76	28	37
ARG AGG	481	112	54	139	80	30	50	151	44	127	60	241	79	258	93	143	124	163	114	245	44	164	80	146	72	348	124	446	178	29	15	19
SER TCT	2037	596	213	294	288	1002	334	1723	234	1292	289	2644	430	1726	611	574	719	2086	750	1052	167	2390	417	1098	531	3111	843	2924	602	33	24	58
SER TCC	817	360	129	169	168	319	203	569	102	521	232	919	279	399	159	152	191	253	117	155	25	505	86	186	147	958	366	743	262	18	7	22
SER TCA	2668	1138	320	614	621	1425	465	3037	376	1035	386	3127	661	1952	513	1310	1143	1379	481	1730	190	1250	260	1152	590	2270	726	3068	709	84	59	34
SER TCG	1078	261	120	152	141	278	115	656	118	214	75	516	180	603	201	419	430	364	195	547	90	400	74	322	180	528	184	773	203	27	9	16
SER AGT	1786	622	169	401	263	608	244	767	125	973	295	1690	327	1697	468	618	579	1565	684	897	119	2266	570	1354	729	3493	900	3214	774	68	40	47
SER AGC	492	125	64	73	57	133	104	222	41	283	193	474	196	140	50	70	76	95	61	69	20	203	70	130	91	524	258	441	234	32	14	19
ILE ATT	3942	1635	409	669	536	2142	552	2636	338	2380	584	4363	702	5248	1637	1190	1408	4491	2417	1908	354	4598	775	2537	1171	8966	2024	9084	1695	125	87	104
ILE ATC	1231	392	183	217	150	895	524	1188	230	826	341	1319	368	518	287	150	233	301	257	180	47	379	104	209	153	1357	559	1040	490	57	18	16
ILE ATA	2299	933	307	950	430	1086	437	1518	435	733	354	1663	530	1263	384	700	715	905	425	850	173	677	179	577	277	2315	738	2010	777	127	104	48
MET ATG	2034	885	211	521	414	677	268	1284	228	730	290	1702	373	1682	505	877	1062	1293	665	1598	284	1451	265	1113	692	2649	715	2783	811	108	28	47
PHE TTT	3167	1567	456	640	555	984	322	799	187	1147	299	2648	503	4259	1360	1124	1232	3398	1756	1754	267	3874	699	2471	973	6643	1638	6719	1362	180	84	103
PHE TTC	1499	767	229	258	214	1091	663	1936	274	1529	463	2753	499	428	207	116	149	264	177	138	26	546	74	255	131	997	319	841	257	72	31	24
TYR TAT	3440	1330	316	461	339	1265	377	1608	234	1807	426	3565	548	2381	796	612	914	1965	1030	1022	240	2295	446	1592	743	4285	1052	4163	902	111	67	54
TYR TAC	1530	421	150	143	165	473	175	668	135	513	401	1171	417	645	224	141	242	386	213	215	43	539	117	312	173	1276	614	1060	466	57	23	13
CYS TGT	1058	680	177	222	132	491	168	604	86	639	177	1078	182	1099	294	278	345	786	423	367	65	1093	243	769	373	1511	441	1256	299	43	17	42
CYS TGC	275	91	45	50	48	83	43	97	22	216	113	250	76	66	26	26	62	42	25	23	13	78	16	60	35	144	79	122	59	23	6	20
TRP TGG	994	375	166	281	116	137	159	643	68	525	187	998	190	779	329	454	362	604	295	570	83	806	151	396	409	1569	369	1435	368	81	30	27
LEU TTA	4531	1685	434	1012	767	2528	862	4015	567	2273	709	6450	1075	2794	876	1306	1369	2502	1048	2082	277	2439	307	1786	795	4977	1091	5451	1440	181	146	51
LEU TTG	3612	1166	418	732	483	1049	469	2099	360	1311	563	2630	739	3191	1050	1510	1865	2557	1284	2716	457	2830	548	1626	906	4976	1522	5351	1446	164	53	34
LEU CTT	1148	620	208	381	348	511	193	707	139	708	240	1302	285	1256	364	454	558	958	568	726	141	1145	236	734	341	2124	582	2282	567	52	38	52
LEU CTC	296	166	171	104	117	155	107	235	54	244	157	394	208	136	88	76	74	97	63	66	15	113	52	76	48	374	219	351	151	12	10	8
LEU CTA	642	449	163	585	404	389	153	685	196	348	194	985	340	441	182	308	350	391	183	503	94	240	104	238	120	854	376	884	406	50	22	20
LEU CTG	570	260	153	232	431	140	67	429	102	278	121	750	207	498	133	350	393	297	150	411	72	267	101	265	131	753	184	983	267	27	14	18
PRO CCT	1024	391	118	163	217	901	171	1630	185	1127	195	2096	345	1147	316	446	497	1094	386	685	121	1238	219	725	348	2398	506	2592	475	17	10	35
PRO CCC	459	157	87	80	123	147	50	299	47	275	134	591	193	326	103	133	248	128	68	137	14	156	41	104	78	663	194	610	261	7	4	6
PRO CCA	2372	828	249	625	532	1610	319	4346	550	975	428	3300	760	2087	505	1216	980	1578	421	1886	180	1497	209	1002	441	2126	699	4095	899	43	22	23
PRO CCG	503	125	34	105	130	199	59	531	88	129	48	428	115	323	77	241	256	187	65	272	55	159	43	211	88	305	120	606	153	8	5	2
HIS CAT	2007	665	205	366	326	946	233	1408	149	1760	540	2614	441	1299	411	528	435	976	514	793	134	1252	352	976	363	2273	611	2540	503	101	45	39
HIS CAC	626	190	158	141	141	248	154	503	87	590	661	1046	385	389	122	120	109	210	128	168	27	319	77	201	99	747	297	778	305	12	10	11
GLN CAA	4480	1220	420	1024	1117	1402	457	3862	410	2100	823	11716	2549	2763	790	1354	1327	2195	801	2233	269	2323	369	1500	629	4619	1294	6639	1721	163	73	45
GLN CAG	1168	363	143	331	255	267	130	738	138	488	275	2211	903	432	149	288	306	350	152	467	121	278	114	298	222	855	309	1029	399	33	28	9
VAL GTT	2730	898	219	373	384	1364	392	1937	233	1390	372	2469	430	3987	1136	866	940	3223	1549	1403	262	3424	506	1628	764	5790	1427	5701	1243	89	46	47
VAL GTC	771	184	89	84	95	385	245	541	82	409	198	667	188	434	200	138	157	274	202	149	29	400	66	158	140	1042	433	788	331	19	12	1
VAL GTA	1784	524	144	345	244	672	234	1038	238	494	214	1176	313	729	234	607	470	618	257	699	87	380	93	343	234	903	319	1084	435	118	92	24
VAL GTG	1749	394	162	354	240	355	193	732	192	374	170	770	237	1365	339	789	999	722	380	993	216	609	171	503	410	1182	387	1758	557	49	20	15
ALA GCT	1814	790	146	254	218	1035	212	1813	178	1211	246	2477	344	2362	624	678	714	4007	1331	1466	213	3295	435	1236	654	3441	792	3862	665	64	38	29
ALA GCC	1020	361	86	100	109	315	153	579	66	489	184	784	224	723	170	180	295	480	246	263	39	579	77	255	146	956	342	909	339	17	6	10
ALA GCA	2282	589	140	249	211	753	241	1768	194	509	198	1551	375	1307	344	789	694	1200	489	1798	158	790	162	729	336	1366	476	2109	646	57	35	24
ALA GCG	454	75	22	47	49	73	29	182	31	72	29	134	59	283	74	197	217	150	87	267	62	123	43	124	98	193	53	287	93	17	2	9
GLY GGT	1529	705	152	249	151	720	239	1349	118	1354	381	2216	250	2968	882	868	777	2918	1169	1322	160	5109	767	1989	1068	4485	1194	4060	690	49	26	30
GLY GGC	322	97	41	53	37	122	85	229	27	364	186	515	147	215	78	89	107	169	75	109	20	437	137	215	115	599	299	521	220	7	7	6
GLY GGA	1699	557	158	281	152	495	164	937	94	633	267	1528	250	1226	346	685	390	822	344	938	63	1212	272	1060	339	1595	555	2112	531	57	18	20
GLY GGG	1369	198	78	166	65	69	72	276	30	219	65	286	90	947	233	455	345	345	177	476	62	310	69	186	112	634	165	700	238	13	8	15
ASP GAT	5676	1416	319	680	510	1847	532	2957	293	2106	503	4487	676	4485	1248	1415	1596	3303	1620	2157	325	3529	789	2302	1272	10293	2879	10702	2299	178	83	77
ASP GAC	1921	399	116	175	168	597	267	953	155	656	369	1434	421	809	235	228	342	425	272	369	80	514	161	326	237	2557	1342	2414	1105	26	17	19
GLU GAA	6464	1465	438	1154	747	1312	460	3125	344	2021	705	5887	1165	4185	1242	1699	1700	3160	1406	2829	349	3366	467	1963	1036	8762	2379	13439	3549	213	124	51
GLU GAG	2228	559	190	463	356	345	118	911	140	526	233	1264	411	858	245	445	523	553	297	850	162	340	141	357	203	1879	694	2798	1204	66	47	20
STOP TAA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STOP TAG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STOP TGA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	114361	38097	11915	23036	17800	43919	15779	80732	11269	53392	19292	122904	26540	87892	27196	35347	37171	71040	32933	54811	8294	79406	15231	48676	25277	149726	43587	166535	43947	4209	2301	1992
	1E+05	38098	11916	23036	17802	43919	15779	80733	11269	53396	19293	1E+05	26541	87892	27196	35348	37171	71041	32933	54812	8296	79406	15232	48676	25278	1E+05	43588	2E+05	43947	4209	2301	1992
	TTG	CTT	CTC	CTA	CTG	CCT	CCC	CCA	CCG	CAT	CAC	CAA	CAG	GTT	GTC	GTA	GTG	GCT	GCC	GCA	GCG	GGT	GGC	GGA	GGG	GAT	GAC	GAA	GAG	TAA	TAG	TGA

Figura B.2: Tabela de contingência da *Candida albicans* (continuação).

	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'
	AAA	AAG	AAT	AAC	ACT	ACC	ACA	ACG	CGT	CGC	CGA	CGG	AGA	AGG	TCT	TCC	TCA	TCG	AGT	AGC	ATT	ATC	ATA	ATG	TTT	TTC	TAT	TAC	TGT	TGC	TGG	TTA
LYS AAA	5691	4194	4899	3000	2236	1296	2418	961	729	274	451	290	3660	1777	2537	1411	2522	1098	2046	1430	4076	2139	2812	2468	3651	2303	2934	2098	1039	593	1587	3454
LYS AAG	5861	4281	3130	2329	1534	1014	1606	739	555	233	254	189	2392	1074	1262	771	1155	573	1094	811	2353	1387	1475	1348	2312	1447	1615	1425	492	340	682	1800
ASN AAT	4148	2457	4857	3116	1887	1226	1897	832	521	223	297	140	1633	836	2456	1439	2226	987	1799	1216	3461	1833	1947	1919	2749	1907	1895	1394	761	465	1053	2703
ASN AAC	3400	2604	3512	2702	1597	1005	1374	563	440	199	146	102	1509	671	1821	1128	1283	616	1369	1069	2188	1323	1290	1302	1776	1267	1463	1210	615	329	906	1524
THR ACT	1721	1223	1814	1033	1651	1057	1227	557	267	110	123	85	843	327	1899	1057	1505	570	830	456	1623	964	926	752	1726	1297	930	752	401	209	619	2168
THR ACC	1966	1585	1557	1277	1335	862	979	470	185	63	56	41	773	304	928	498	724	284	586	427	1442	914	835	746	764	623	534	470	266	148	376	854
THR ACA	2213	1730	2272	1464	1104	724	1406	538	269	123	158	99	1555	662	1305	760	1196	484	959	594	1482	653	1081	1058	1552	766	1004	629	427	288	616	1373
THR ACG	1115	710	904	562	383	260	482	209	98	52	91	47	507	259	452	293	524	222	328	260	506	293	518	436	650	391	481	321	162	136	252	591
ARG CGT	674	456	495	289	320	233	248	119	329	105	87	47	298	140	443	315	360	153	210	107	600	314	293	282	327	280	447	370	250	138	266	544
ARG CGC	369	259	276	191	125	107	169	67	118	56	37	43	165	91	207	105	151	74	151	145	194	131	181	148	162	115	195	219	83	71	128	203
ARG CGA	407	233	406	254	147	120	161	96	54	36	29	24	267	129	191	85	169	82	171	95	262	141	257	227	281	144	219	194	89	90	104	280
ARG CGG	301	180	238	111	55	40	101	38	35	27	22	29	94	78	137	86	122	93	79	39	123	69	100	105	221	114	151	107	44	43	48	168
ARG AGA	2871	2367	2549	1838	1248	735	1200	485	401	164	184	131	2601	1021	1134	601	1149	501	970	674	1875	1029	1332	1236	1764	952	1088	787	450	234	616	1692
ARG AGG	1776	1017	875	639	498	290	516	221	232	109	147	107	741	341	541	288	549	273	322	253	715	344	639	470	925	489	491	400	154	134	259	745
SER TCT	2332	1516	2129	1301	1779	992	1420	602	530	258	259	102	937	491	2761	1560	2101	913	950	618	1837	1067	999	1144	1562	1144	1044	778	433	260	566	2165
SER TCC	2437	1824	1858	1302	1490	799	1069	458	283	107	144	55	920	399	1354	815	987	511	932	649	1633	1036	912	1056	1103	721	537	439	247	159	319	1030
SER TCA	2339	1685	2305	1415	1199	582	1050	516	263	143	278	108	1611	718	1697	1058	1786	810	1008	748	1316	694	1046	1179	1734	994	1135	729	444	305	592	1786
SER TCG	1302	915	1043	663	432	255	499	250	123	90	81	43	633	306	614	482	644	373	401	330	604	355	541	590	783	491	519	436	175	119	291	786
SER AGT	1868	934	1709	1072	833	639	898	410	196	108	121	72	647	348	1115	628	1080	420	847	583	1189	603	897	637	1089	819	705	486	367	229	381	1177
SER AGC	1451	859	1406	987	817	575	689	355	146	104	100	61	615	328	890	460	646	310	752	579	878	488	671	515	675	439	524	336	217	140	259	561
ILE ATT	2576	1714	2252	1491	1641	959	1225	549	578	220	246	135	1346	583	2382	1466	1739	710	909	549	2554	1524	1384	1333	2284	1929	1586	1148	763	536	990	2766
ILE ATC	3167	2257	2368	1861	1311	847	1010	431	261	143	124	74	1307	523	931	692	704	320	910	533	1783	1245	1004	1198	1065	872	684	644	399	267	463	1069
ILE ATA	2369	1526	2197	1434	1010	542	1132	527	294	141	312	202	1241	602	1495	880	1472	676	745	568	1357	780	1276	1126	1811	1072	1312	943	551	353	761	1550
MET ATG	2875	1800	2349	1584	1140	820	1385	584	341	121	152	110	1317	563	1355	835	1149	567	1022	690	1845	932	1112	1397	1646	907	967	793	434	308	551	1146
PHE TTT	2170	1240	1945	1321	1376	1006	1060	592	274	162	225	159	1218	550	2006	1205	1560	707	878	630	2403	1309	1114	1192	2169	2048	1686	1446	668	456	960	2469
PHE TTC	3255	2289	2600	1854	1299	902	922	453	387	163	159	85	1182	527	923	662	893	335	863	599	2017	1285	941	1169	1379	1231	875	842	414	271	572	1335
TYR TAT	1948	1217	1748	1128	867	571	894	490	418	153	181	104	902	478	1431	836	1021	508	540	362	1729	1018	884	1048	1730	1312	1406	896	569	328	603	1649
TYR TAC	1905	1542	1627	1222	902	609	838	360	362	136	103	82	892	453	923	549	666	282	531	341	1227	942	686	927	1103	852	1057	920	511	241	555	952
CYS TGT	813	462	612	452	422	290	336	182	185	68	88	41	327	184	587	360	479	218	222	170	776	470	440	411	732	514	475	320	307	198	279	693
CYS TGC	573	419	457	285	272	184	251	104	145	50	59	47	268	158	386	254	276	123	203	169	509	276	367	337	580	320	366	267	258	161	190	400
TRP TGG	1589	1008	1184	778	467	325	482	235	105	47	86	40	850	380	703	395	666	287	349	254	869	427	601	562	936	630	682	546	432	256	436	786
LEU TTA	3050	2355	2530	1816	1568	982	1501	665	519	268	335	178	1860	848	1815	1135	1635	798	1195	781	2049	1161	1322	1496	2083	1331	1320	988	697	408	877	2186
LEU TTG	4991	3661	3323	2559	1543	1108	1242	630	468	206	234	135	1972	854	1170	829	1022	427	1183	849	2546	1329	1459	1601	1662	1030	1132	912	450	310	439	1457
LEU CTT	806	488	722	559	569	331	423	234	255	104	240	102	365	164	1284	919	994	467	251	190	911	469	449	387	1208	1340	1092	740	495	296	483	1573
LEU CTC	783	540	725	447	395	296	310	157	87	59	76	27	259	136	447	290	316	178	317	212	562	359	394	295	547	431	368	247	206	138	161	397
LEU CTA	1604	1141	1462	924	654	416	736	287	297	103	194	95	938	419	923	698	842	450	578	418	905	576	764	728	1039	665	807	582	393	209	481	1229
LEU CTG	1619	1012	1165	841	468	294	458	269	175	111	183	103	546	289	478	326	509	274	373	301	810	483	693	543	748	565	524	461	226	170	250	692
PRO CCT	1200	725	1198	728	704	570	689	337	332	124	191	64	437	206	1175	726	1018	449	443	250	1065	631	616	584	867	714	777	508	242	150	367	1142
PRO CCC	1040	776	849	543	526	360	534	231	108	51	48	25	375	167	720	360	445	204	370	261	794	496	527	505	617	416	308	254	124	75	174	592
PRO CCA	1996	1769	1940	1272	1306	702	1077	435	284	120	133	87	1422	560	1453	705	1196	489	844	529	1568	765	1049	1011	1585	876	1055	690	278	228	646	1496
PRO CCG	561	418	510	367	218	106	316	118	80	52	54	43	244	124	299	187	321	125	180	123	454	242	360	290	428	270	362	303	64	50	137	588
HIS CAT	1300	803	1235	769	689	393	663	250	351	147	263	88	707	308	1117	617	871	427	437	296	1211	768	710	655	1151	854	956	566	450	215	488	1426
HIS CAC	975	800	901	655	497	293	485	193	227	83	70	58	573	195	534	360	460	172	314	243	745	426	419	414	482	442	539	371	220	148	281	481
GLN CAA	3085	2498	3044	1922	1504	810	1414	654	505	191	304	173	2035	942	1511	776	1196	577	1324	763	2300	1244	1601	1715	1925	1294	1547	1142	622	288	891	2279
GLN CAG	1558	935	1300	750	425	276	573	247	327	151	160	118	630	356	540	318	588	291	399	295	888	529	682	564	1266	717	871	757	249	218	337	1016
VAL GTT	1907	1308	1731	1177	1339	797	860	369	555	185	146	92	939	346	1970	1171	1368	555	665	464	1973	1086	909	987	1509	1491	1333	1037	628	401	681	1901
VAL GTC	1679	1493	1299	981	947	698	704	257	216	94	52	38	740	315	714	499	452	208	562	366	1429	850	577	746	806	612	444	504	287	156	298	703
VAL GTA	1524	951	1315	812	649	385	699	361	291	114	228	90	780	385	803	603	759	362	450	353	816	432	736	683	1116	585	757	518	273	222	381	780
VAL GTG	1490	1050	974	764	507	373	579	258	160	92	85	88	641	299	520	362	463	263	391	313	728	405	586	596	815	460	504	382	213	174	263	577
ALA GCT	1927	1384	1615	1177	1635	864	1023	437	438	158	145	117	990	374	1750	1250	1226	546	671	438	1904	1024	771	990	1163	895	977	782	479	208	562	1652
ALA GCC	1800	1763	1603	1009	1167	759	800	402	268	83	50	39	711	357	939	540	527	261	593	408	1788	1008	712	959	891	604	567	402	349	155	277	753
ALA GCA	1861	1362	1848	1186	851	468	939	371	293	138	124	65	1517	637	1181	697	935	467	798	573	1244	716	769	1113	1491	812	962	682	431	365	465	1294
ALA GCG	702	570	519	422	306	141	305	152	92	64	34	34	302	139	397	246	379	169	208	143	421	217	286	341	575	351	421	307	128	105	187	527
GLY GGT	2610	2008	1895	1526	1499	975	1096	468	434	124	98	43	1316	350	1656	960	1187	464	731	600	2619	1449	1211	1425	1016	1056	1192	1039	587	353	736	1602
GLY GGC	1136	851	1036	660	617	357	543	223	284	107	62	51	555	277	798	530	636	239	595	394	794	445	608	481	651	408	681	540	376	185	429	623
GLY GGA	1643	1054	1576	920	677	367	738	394	169	96	93	61	888	395	795	447	580	329	715	449	941	452	734	723	1354	632	811	469	327	196	431	893
GLY GGG	974	638	557	419	285	182	337	159	129	51	64	33	332	159	446	302	301	156	216	187	379	229	317	328	980	600	324	287	132	95	170	558
ASP GAT	3781	2171	3626	2144	1820	1153	1661	763	640	245	271	110	1625	708	2661	1517	1972	938	1365	917	3983	2326	2175	2127	3251	2216	2331	1443	772	394	1177	3324
ASP GAC	2381	1972	2077	1617	1300	793	987	408	515	190	149	94	1134	509	1458	799	831	495	918	665	1979	1218	1173	1172	1429	1075	1283	1062	536	285	682	1382
GLU GAA	6188	4859	5760	3712	2806	1514	2494	1172	810	295	429	294	3217	1380	2482	1438	1930	884	2375	1473	4453	2232	2744	2784	3337	1798	2551	1863	928	471	1363	3244
GLU GAG	2879	1935	2113	1484	869	553	905	421	351	188	156	118	1012	507	881	481	803	445	774	565	1543	822	917	906	1814	930	1164	972	312	252	483	1425
STOP TAA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STOP TAG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STOP TGA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	126422	89593	107921	73117	59717	37082	54005	24315	18784	7982	9471	5557	62113	28006	69783	42042	57036	26179	43211	29747	89218	50374	54831	55467	79047	53860	56847	43145	23896	14725	30866	78200
	1E+05	89593	1E+05	73117	59717	37082	54005	24315	18784	7982	9471	5557	62113	28006	69783	42042	57036	26179	43211	29747	89218	50374	54831	61731	79047	53860	56847	43145	23896	14725	30866	78200
	AAA	AAG	AAT	AAC	ACT	ACC	ACA	ACG	CGT	CGC	CGA	CGG	AGA	AGG	TCT	TCC	TCA	TCG	AGT	AGC	ATT	ATC	ATA	ATG	TTT	TTC	TAT	TAC	TGT	TGC	TGG	TTA

Figura B.3: Tabela de contingência da *Saccharomyces cerevisiae*.

	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'
	TTG	CTT	CTC	CTA	CTG	CCT	CCC	CCA	CCG	CAT	CAC	CAA	CAG	GTT	GTC	GTA	GTG	GCT	GCC	GCA	GCG	GGT	GGC	GGA	GGG	GAT	GAC	GAA	GAG	TAA	TAG	TGA
LYS AAA	3028	1700	670	2351	1687	1607	725	2071	697	1420	754	2900	1417	2871	1382	1568	1474	2086	1101	1709	567	2092	902	1155	626	4447	2267	5964	2763	289	126	132
LYS AAG	1991	1101	478	1361	1074	1094	562	1601	408	1354	814	2616	1177	1773	972	911	732	2102	1144	1481	611	1760	971	898	405	3406	2055	4920	2117	73	45	53
ASN AAT	2707	1072	557	1189	896	1471	573	1683	643	1373	687	2219	848	2650	1311	1305	1071	2267	1470	1892	817	2636	1535	1632	1023	4845	2789	5833	2506	113	55	73
ASN AAC	1838	868	418	804	646	918	426	1172	303	1000	576	1880	844	1585	966	760	626	1193	685	871	282	1538	705	752	428	2807	1485	3025	1211	90	42	70
THR ACT	2213	616	283	711	475	859	421	1402	362	692	371	1446	544	1566	763	856	771	1598	984	1302	447	1759	558	674	423	2177	1044	2626	980	26	17	29
THR ACC	764	430	143	323	296	391	236	673	156	368	249	799	308	826	496	454	404	745	451	546	199	926	274	313	164	1085	539	1111	502	17	8	14
THR ACA	1090	567	242	653	487	1033	422	1199	334	795	392	1098	591	955	490	856	650	831	536	1020	381	1091	519	722	369	1916	935	2244	968	38	27	20
THR ACG	603	368	152	416	363	505	271	495	241	407	217	622	345	448	259	435	363	400	238	482	222	420	268	319	193	865	468	941	462	24	6	12
ARG CGT	564	235	122	250	157	180	107	228	78	329	174	520	228	508	262	283	225	383	263	294	115	688	277	245	154	728	402	853	319	36	20	18
ARG CGC	145	103	60	94	96	89	63	121	40	173	99	242	120	153	50	66	66	100	55	115	25	121	86	85	54	280	121	234	170	14	2	4
ARG CGA	221	129	49	201	99	110	50	128	20	181	62	225	151	168	68	124	94	126	81	170	43	112	93	135	47	341	212	384	190	11	7	5
ARG CGG	176	109	53	111	104	80	65	77	36	72	52	139	116	70	37	64	48	68	56	56	29	51	37	56	20	157	93	211	102	6	2	6
ARG AGA	1495	550	242	837	624	699	287	918	273	814	493	1520	706	1170	578	746	528	1192	714	933	330	1564	621	719	286	2300	1144	2805	1150	43	28	25
ARG AGG	646	430	200	517	362	403	235	547	190	431	205	838	429	419	255	346	306	584	328	488	195	381	232	252	95	892	525	1178	527	26	19	25
SER TCT	2255	857	378	873	638	1064	629	1639	432	1056	540	2324	820	1607	850	852	919	1761	1003	1352	539	1749	655	789	429	2451	1146	2554	1000	33	18	31
SER TCC	987	447	192	489	365	401	233	512	123	476	317	1074	423	821	399	440	371	658	386	496	164	805	284	348	170	1001	495	1026	516	20	13	15
SER TCA	1532	731	266	856	674	1032	467	1219	343	796	377	1530	845	830	359	721	546	892	398	915	345	958	504	731	326	1810	882	2021	840	53	25	39
SER TCG	753	372	163	505	408	426	274	554	206	485	303	832	438	410	168	344	260	267	175	376	162	300	191	216	117	747	341	765	394	18	8	12
SER AGT	1143	450	188	326	308	409	219	443	156	484	255	858	280	934	523	592	440	1049	655	887	371	1174	537	696	377	2089	1084	2194	902	33	19	29
SER AGC	533	296	125	250	201	307	179	341	100	386	215	709	295	512	263	369	285	555	414	413	130	638	363	374	209	1160	601	1116	524	23	9	19
ILE ATT	2498	1435	516	1137	843	1534	702	1970	686	1385	769	2593	1055	2462	1377	1214	1194	2313	1556	1767	706	2459	916	1072	646	3889	2084	4533	1661	86	37	56
ILE ATC	1126	598	339	628	464	555	310	933	225	683	330	1177	565	864	533	431	416	707	456	560	225	1016	437	434	287	1804	1000	1843	902	47	18	14
ILE ATA	1327	723	264	866	726	781	451	1063	564	701	463	1095	605	798	427	641	594	752	490	811	379	710	427	505	285	1893	1039	2028	1002	101	43	50
MET ATG	1078	997	346	928	839	742	439	892	264	748	381	1630	615	1420	738	803	680	1379	857	1264	493	1432	701	807	476	2401	1442	2872	1133	68	32	37
PHE TTT	2535	986	481	881	705	988	638	1126	401	1039	591	2157	900	2322	1131	1019	851	1913	1434	1506	597	2446	1065	1177	590	3522	1897	4164	1566	103	53	69
PHE TTC	1503	659	363	684	501	751	579	1004	228	972	579	1921	892	862	510	370	349	565	410	468	159	1129	383	377	196	1596	856	1697	714	52	23	25
TYR TAT	1883	747	343	694	590	853	355	1128	373	873	465	1556	574	1301	744	514	714	1186	710	955	399	1601	617	797	459	2367	1383	2524	1063	49	21	40
TYR TAC	1255	463	256	553	507	405	208	742	146	701	467	1352	588	910	545	427	437	707	426	450	179	1064	382	431	253	1626	925	1694	726	56	20	56
CYS TGT	767	360	185	330	272	391	245	365	113	352	227	584	216	642	359	284	260	560	308	379	175	833	348	325	225	870	470	863	421	34	18	22
CYS TGC	347	236	136	186	172	172	123	214	64	269	158	488	211	276	132	196	127	185	133	166	70	331	141	151	114	406	222	378	177	13	6	11
TRP TGG	720	366	191	452	397	256	178	380	98	391	266	805	322	747	360	427	274	568	334	454	195	667	300	278	244	1183	748	1267	613	51	25	28
LEU TTA	2034	1077	502	1354	1081	1249	568	1741	641	1065	631	2865	1311	1334	742	817	830	1254	847	1207	511	1521	625	847	436	2694	1457	3482	1501	105	53	66
LEU TTG	1608	761	341	962	744	966	686	1396	386	1112	621	2315	1022	1733	948	955	911	1810	1150	1438	564	1810	695	774	421	2892	1583	3673	1540	83	35	24
LEU CTT	1533	682	342	600	436	632	349	843	283	622	324	1050	408	813	412	426	454	794	497	617	282	875	414	530	252	1406	721	1380	504	32	22	39
LEU CTC	423	215	113	200	162	210	140	273	67	221	128	308	188	300	171	186	156	239	161	186	75	401	138	143	91	662	301	524	290	21	7	11
LEU CTA	1070	515	274	722	620	654	348	873	343	597	328	1265	646	638	379	465	420	755	397	675	250	703	335	446	206	1340	704	1803	770	40	17	27
LEU CTG	768	424	209	536	456	396	348	510	211	486	288	966	526	559	365	505	416	533	474	646	289	606	350	424	193	1189	740	1514	702	32	13	24
PRO CCT	1069	509	213	450	432	754	363	1139	322	708	385	1508	567	1041	478	688	545	1073	508	786	301	877	394	450	301	1597	749	1968	853	20	11	18
PRO CCC	496	216	134	230	160	226	98	324	63	303	129	422	205	447	227	373	266	359	172	247	70	328	153	161	65	530	287	686	263	6	1	9
PRO CCA	1440	550	243	656	502	939	310	1484	448	695	397	1353	840	987	462	675	526	877	425	833	251	1102	454	474	224	1641	776	2486	1119	23	12	24
PRO CCG	413	252	97	404	295	291	160	409	110	283	160	552	416	314	162	294	192	258	131	282	90	167	193	208	102	542	303	734	366	3	1	11
HIS CAT	1371	543	259	535	388	804	288	926	378	917	514	1046	444	899	439	457	422	795	541	621	282	1011	429	549	330	1535	825	1762	715	36	60	45
HIS CAC	629	295	153	200	211	255	152	465	116	467	321	655	302	531	257	242	207	330	215	266	95	524	218	216	139	854	473	894	372	20	13	18
GLN CAA	2281	855	360	1182	1061	1018	361	1326	516	1064	545	3255	1897	1488	774	977	935	1502	843	1121	475	1668	713	746	431	2866	1312	4102	1744	101	44	45
GLN CAG	1041	521	251	613	517	468	246	578	257	693	366	1916	1380	571	314	425	393	587	391	558	280	512	377	444	176	1256	745	1554	749	46	20	20
VAL GTT	2003	1064	434	698	553	1208	532	1488	359	934	522	1644	596	1884	1011	711	681	1670	1020	1119	414	1787	605	929	399	2479	1311	2627	1009	62	28	45
VAL GTC	898	521	197	339	296	389	231	610	102	480	289	731	289	754	380	329	282	616	364	523	128	803	276	342	182	1005	660	1132	422	22	9	11
VAL GTA	875	503	225	480	446	599	321	838	319	510	217	844	477	666	337	472	479	583	357	731	300	570	331	442	249	1510	731	1556	817	47	18	33
VAL GTG	703	457	214	485	402	436	303	579	215	383	210	799	415	717	393	630	536	650	521	726	251	578	328	412	291	1376	776	1891	881	28	11	13
ALA GCT	1961	834	307	701	446	810	463	1207	236	747	504	1762	597	1686	871	751	773	2371	1216	1295	462	1961	624	632	290	1992	997	2646	930	87	22	30
ALA GCC	837	500	218	429	303	316	205	567	90	414	202	856	317	836	412	348	331	1009	556	548	152	880	250	235	113	799	410	940	338	21	15	20
ALA GCA	1232	498	270	658	484	758	323	904	175	705	321	1229	574	807	406	688	484	1006	539	1036	294	1064	502	605	257	1721	908	2198	890	43	24	32
ALA GCG	535	311	149	399	305	283	179	384	99	335	159	573	318	324	164	328	228	423	253	423	154	296	228	246	97	606	361	720	341	13	5	13
GLY GGT	1642	747	332	819	404	595	276	789	159	748	467	1367	359	2226	1270	901	839	1877	1157	1038	411	2972	845	843	547	2623	1583	3046	1077	36	10	24
GLY GGC	576	370	195	361	303	341	221	449	77	566	342	910	453	591	288	317	267	628	372	412	110	718	374	375	181	941	519	1062	441	11	6	16
GLY GGA	757	388	179	447	378	382	203	367	120	409	322	786	374	486	235	340	276	456	263	485	161	595	335	511	235	1021	529	1266	605	18	9	14
GLY GGG	629	359	172	386	316	300	195	325	139	271	160	429	208	257	116	200	170	246	187	265	99	224	177	176	92	429	262	597	301	8	2	4
ASP GAT	3538	1385	626	1161	933	1592	635	1714	644	1225	672	2000	854	2786	1508	1318	1290	2609	1408	1792	742	2614	1143	1301	775	6170	3265	7118	3079	98	57	50
ASP GAC	1693	780	334	663	607	876	415	1026	281	873	500	1470	707	1320	584	673	587	1037	609	728	227	1212	538	496	297	2705	1481	3204	1286	47	31	35
GLU GAA	3406	1521	687	2095	1552	1297	563	1506	513	1368	809	3237	1530	2900	1391	1557	1415	2663	1400	1930	763	2727	1066	1252	695	5865	3060	8509	3631	133	53	78
GLU GAG	1408	840	432	971	888	756	397	904	207	920	480	1822	898	1031	525	630	599	1061	686	1188	408	797	526	627	280	2402	1367	3510	1764	62	34	41
STOP TAA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STOP TAG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STOP TGA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	78592	37484	16863	40182	31657	40306	20551	52784	16179	41327	23151	79684	36586	63696	33328	36096	31985	59783	35911	48290	18442	66354	28965	33301	18027	111709	59880	134452	57441	2950	1435	1879
	78592	37484	16863	40182	31657	40306	20551	52784	16179	41327	23151	79684	36586	63696	33328	36096	31985	59783	35911	48290	18442	66354	28965	33301	18027	1E+05	59880	1E+05	57441	2950	1435	1879
	TTG	CTT	CTC	CTA	CTG	CCT	CCC	CCA	CCG	CAT	CAC	CAA	CAG	GTT	GTC	GTA	GTG	GCT	GCC	GCA	GCG	GGT	GGC	GGA	GGG	GAT	GAC	GAA	GAG	TAA	TAG	TGA

Figura B.4: Tabela de contingência da *Saccharomyces cerevisiae* (continuação).

Apêndice B.2

		3' A1	3' T1	3' C1	3' G1	3' A2	3' T2	3' C2	3' G2	3' A3	3' T3	3' C3	3' G3
Lys	AAA	44903	28802	19743	32974	45063	37634	25042	18683	38947	40370	22345	24760
Lys	AAG	32428	16036	14871	26258	37218	22515	17657	12203	28457	26937	16753	17446
Asn	AAT	35064	22983	14392	35482	39135	29177	23766	15843	30790	36600	21741	18790
Asn	AAC	27478	15978	10742	18919	27851	19181	15237	10848	19950	24690	15493	12984
Thr	ACT	17004	15418	8767	18528	17396	17710	16898	7713	17585	20408	11359	10365
Thr	ACC	16058	7272	4717	9035	12375	10314	9477	4916	10437	12368	7670	6607
Thr	ACA	19495	11575	8462	14473	18316	13955	13273	8461	16832	17562	9537	10074
Thr	ACG	7732	5120	4680	6783	8450	6782	5679	3404	7574	7367	4441	4933
Arg	CGT	5078	4531	3176	5999	6340	5246	3839	3359	5231	6472	3758	3323
Arg	CGC	2769	1878	1554	1781	2964	1967	1613	1438	2250	2530	1674	1528
Arg	CGA	3373	2172	1548	2378	3477	2735	1779	1480	2953	2977	1774	1767
Arg	CGG	1751	1524	1127	1155	2038	1672	1139	708	1634	1690	1029	1204
Arg	AGA	23931	12559	8843	16780	22503	16650	12399	10561	19475	19768	11093	11777
Arg	AGG	9657	5964	5382	7003	10268	7808	6146	3784	9330	8293	4926	5457
Ser	TCT	20114	17614	12399	19656	21042	19147	20547	9047	20660	23570	13179	12374
Ser	TCC	18774	9257	5641	8370	13748	12002	10456	5836	11919	14046	8323	7754
Ser	TCA	19311	14719	9928	13078	18687	15264	14309	8776	17981	18150	9921	10984
Ser	TCG	9119	6504	5303	5253	9209	7533	6009	3428	8108	7721	4836	5514
Ser	AGT	14114	9720	4873	14504	14972	11315	10212	6712	12746	14639	8628	7198
Ser	AGC	11965	6041	3815	7926	10601	7061	7181	4904	8347	10159	6268	4973
Ile	ATT	22589	20976	15804	29849	28859	26450	21905	12004	25710	30426	17742	15340
Ile	ATC	21755	9305	7409	11905	19350	13625	10207	7192	14432	15939	10509	9494
Ile	ATA	18412	14397	9241	12781	18741	15328	13025	7737	16492	17130	10264	10945
Met	ATG	21415	11873	9545	18898	21690	16814	14165	9062	18437	20218	11894	11182
Phe	TTT	19994	20130	11723	27200	25800	23625	18095	11527	21998	26631	16820	13598
Phe	TTC	22157	11135	9927	10641	21017	15158	10353	7332	15085	17291	11489	9995
Tyr	TAT	15824	14282	9407	17334	19218	16900	12577	8152	15736	19366	11221	10524
Tyr	TAC	14904	9988	7071	11182	16428	12042	8282	6393	11793	14119	8921	8312
Cys	TGT	6569	5993	4012	7322	7189	7485	5400	3822	6362	8316	4974	4244
Cys	TGC	4832	3958	2730	3205	4695	4597	2977	2456	3997	5051	3011	2666
Trp	TGG	10360	7467	4380	8659	11358	8744	6013	4751	9191	9805	6035	5835
Leu	TTA	25179	17531	15385	20105	27223	21399	18117	11461	24373	23970	14242	15615
Leu	TTG	30750	12590	12355	22897	31354	20047	16367	10824	23997	24561	15156	14878
Leu	CTT	7318	12517	7272	10377	10876	12035	9518	5055	10367	12611	8007	6499
Leu	CTC	6187	4178	2474	4024	5750	4911	3740	2462	4387	5892	3631	2953
Leu	CTA	12550	9472	7874	10286	14030	11009	9301	5842	12623	12260	7356	7943
Leu	CTG	10164	6060	5928	9505	12078	8772	6483	4324	9857	9160	6326	6314
Pro	CCT	10383	9253	8061	12609	13502	11044	10914	4846	12320	13362	7491	7133
Pro	CCC	8354	4801	2742	4654	6622	6496	4939	2494	6019	6825	4016	3691
Pro	CCA	18235	12196	9041	13312	18064	14381	12930	7409	16377	17094	8954	10359
Pro	CCG	4631	3552	3658	4338	5871	5055	3421	1832	5341	4692	2906	3240
His	CAT	11184	10640	7891	11612	13473	12088	9652	6114	11907	14101	7910	7409
His	CAC	8128	5160	4030	5833	8612	6134	4878	3527	6439	7715	4800	4197
Gln	CAA	26855	16519	14613	21697	30168	22271	15604	11641	24769	24743	13338	16834
Gln	CAG	10407	8295	8552	9332	14886	10308	6623	4769	11361	10873	6691	7661
Val	GTT	16847	16183	11010	19656	19705	18895	16229	8867	17343	22729	13244	10380
Val	GTC	13633	6623	4874	8198	11439	9719	7442	4728	8946	11272	7147	5963
Val	GTA	11331	8132	6502	10131	12604	9631	8669	5192	10949	11411	6543	7193
Val	GTG	9954	5751	5323	10957	11934	8704	7006	4341	9984	9409	6070	6522
Ala	GCT	17224	13590	9472	19497	18146	16729	16791	8117	16845	21033	11778	10127
Ala	GCC	15739	7158	4857	8157	11356	10929	8838	4788	9317	12819	7221	6554
Ala	GCA	16253	11113	7519	13405	16514	12966	10944	7866	14832	15658	8906	8894
Ala	GCG	5174	4358	3718	5192	6372	5461	4293	2316	5634	5645	3489	3674
Gly	GGT	21778	13560	7761	23255	21586	19558	14607	10603	17983	23417	14012	10942
Gly	GGC	9572	6705	5092	7596	10155	7278	6553	4979	8076	9987	5937	4965
Gly	GGA	12666	8062	4774	7799	11812	9215	6764	5510	9803	11102	6094	6302
Gly	GGG	5698	4994	3537	3798	5866	5996	3924	2241	5273	5534	3621	3599
Asp	GAT	32345	25739	14707	38918	40034	33947	23621	14107	31361	38830	21612	19906
Asp	GAC	20303	13113	9480	16984	22696	16669	12260	8255	16379	20298	12165	11038
Glu	GAA	49163	25959	18506	40824	53128	37116	25355	18853	40543	43843	23772	26294
Glu	GAG	18205	11507	10328	17401	23806	15681	10965	6989	17852	17628	10660	11301
Stop	TAA	0	0	0	0	0	0	0	0	0	0	0	0
Stop	TAG	0	0	0	0	0	0	0	0	0	0	0	0
Stop	TGA	0	0	0	0	0	0	0	0	0	0	0	0

Figura B.5: Tabela de contingência da *Saccharomyces cerevisiae* codões/nucleótidos.

		3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'	3'
		A1	T1	C1	G1	A2	T2	C2	G2	A3	T3	C3	G3
Lys	AAA	53585	48688	31365	45134	63421	56051	37471	21831	63569	62257	23686	29260
Lys	AAG	28169	13595	8269	16054	29211	16973	11799	8102	24520	21913	8226	11428
Asn	AAT	49411	39572	22795	50399	62539	44196	31887	23556	53325	65698	22508	20647
Asn	AAC	29006	13909	10808	10287	28893	14309	13594	7214	18018	22340	13101	10551
Thr	ACT	26955	20566	11767	27093	27081	21442	24444	13413	28811	36965	10410	10195
Thr	ACC	22832	4557	5276	6920	15870	8899	10427	4388	11829	14475	7528	5753
Thr	ACA	26112	17160	9404	14425	19062	20411	19677	7952	25859	22823	8764	9656
Thr	ACG	5098	3708	1532	4218	3997	4682	3798	2079	5517	4713	1688	2638
Arg	CGT	5415	4229	4112	6704	8218	4678	4298	3266	7315	8022	2610	2513
Arg	CGC	1482	707	813	511	1646	713	594	560	1182	1027	653	651
Arg	CGA	4824	4988	2986	4919	5774	5742	3646	2555	7159	5577	2544	2437
Arg	CGG	1810	1369	586	978	1986	1391	826	540	1981	1337	654	771
Arg	AGA	25857	16238	9347	17371	24233	20796	13284	10500	25312	23928	9450	10123
Arg	AGG	4218	3224	1314	2742	4420	3578	2007	1493	4256	3686	1570	1986
Ser	TCT	20491	18472	10675	19601	21675	16937	20401	10226	23433	28792	8674	8339
Ser	TCC	15981	5817	4344	4704	12874	7694	6596	3682	9523	11458	5090	4775
Ser	TCA	33592	27114	14407	18723	25531	30486	26416	11406	36223	33582	10775	13259
Ser	TCG	8499	7193	3015	5513	6707	8009	6623	2882	8926	8160	2928	4207
Ser	AGT	17749	14611	7380	19927	20557	16397	12203	10510	18965	25264	7892	7545
Ser	AGC	8896	4186	2133	2532	7576	3563	3918	2689	5152	6315	3494	2786
Ile	ATT	35154	30253	18740	49504	48563	37132	28250	19706	42966	55890	18667	16127
Ile	ATC	24618	6947	7131	6264	19793	10557	9790	4821	13358	16624	8004	6974
Ile	ATA	21866	18618	10323	12965	19823	21847	15280	6823	20695	23862	8696	10519
Met	ATG	22943	14606	7990	18445	21359	19212	14691	8723	21883	23755	7512	10835
Phe	TTT	27134	26474	10919	39530	34689	33439	20274	15654	32569	42142	15861	13484
Phe	TTC	23244	9955	11705	4946	20898	12350	10764	5837	16100	19137	7855	6758
Tyr	TAT	23185	24065	14072	24438	30637	26144	16790	12189	27104	35291	11586	11780
Tyr	TAC	12870	7444	5417	6666	13603	8423	6240	4131	8440	12207	5574	6176
Cys	TGT	8007	8983	5528	9642	9779	10516	6466	5400	9773	13613	4535	4239
Cys	TGC	3256	2506	1302	876	2977	2264	1414	1285	2177	2860	1487	1416
Trp	TGG	11287	9280	4163	8969	11904	10151	6379	5265	11950	12510	4175	5064
Leu	TTA	38258	35860	24532	30540	42614	39536	31152	15886	46736	48490	15040	18923
Leu	TTG	45944	21363	13217	33836	42718	33450	23273	14918	38958	43605	14460	17338
Leu	CTT	8561	10015	6486	13036	13130	10733	8467	5767	12335	14528	5726	5509
Leu	CTC	5501	2095	2321	1999	5362	2991	2241	1321	3449	3935	2433	2099
Leu	CTA	6798	5217	5347	5674	7928	7401	5207	2500	8128	7279	3474	4155
Leu	CTG	5719	3477	3351	5255	6546	5699	3670	1887	6613	5755	2076	3358
Pro	CCT	12960	9402	8364	13193	16959	9669	11253	6038	16513	17458	4728	5220
Pro	CCC	7727	2419	2369	3264	6951	4440	2820	1568	5216	5568	2258	2737
Pro	CCA	26999	18522	15390	19821	24701	24842	22135	9055	32397	27750	9039	11547
Pro	CCG	3134	2879	2095	3161	3595	3703	2815	1156	4704	3347	1218	2000
His	CAT	13893	14602	10940	13961	19522	15366	10909	7598	18651	20823	6957	6964
His	CAC	7599	2933	4664	4096	9080	4300	3602	2310	5749	6572	3629	3343
Gln	CAA	34515	28498	29066	30826	49010	34727	24661	14508	49408	39512	14158	19828
Gln	CAG	7973	6256	6542	5769	11626	7507	4785	2623	9710	8473	3215	5142
Val	GTT	21926	20532	11585	33849	30518	24271	19984	13119	27447	37038	12507	10900
Val	GTC	14911	3964	3380	4941	11729	6384	6049	3034	7606	10609	4681	4300
Val	GTA	9585	12127	6144	7492	9702	12592	9146	3908	12065	12099	4509	6675
Val	GTG	12357	8949	4483	11380	11384	13602	7927	4258	12723	13120	4281	7046
Ala	GCT	20951	14437	9878	25775	23443	16774	19114	11710	23502	31316	8307	7916
Ala	GCC	18290	4936	3708	5999	12797	9463	6822	3851	9773	13041	4718	5401
Ala	GCA	18018	16064	7347	13383	15130	18288	14626	6768	20703	18964	6643	8502
Ala	GCG	2377	2712	855	2352	2128	3142	2034	991	2918	2691	1019	1668
Gly	GGT	25765	14460	8755	30426	27780	19179	16847	15599	25308	35059	10328	8711
Gly	GGC	6438	3274	2115	3405	6558	2853	3212	2609	4554	5829	2550	2299
Gly	GGA	14078	16090	6018	12490	14066	16555	10604	7451	17887	18055	6548	6186
Gly	GGG	9169	8921	1733	5454	8210	10480	4376	2211	8826	9369	3260	3823
Asp	GAT	42893	38559	18058	50215	57923	44790	28442	18571	50362	60012	19352	19999
Asp	GAC	16959	9038	6185	11406	20355	10242	8527	4464	12213	16012	7325	8038
Glu	GAA	56271	38036	20698	51531	66564	48000	31768	20204	60433	58785	20797	26522
Glu	GAG	15145	11383	5870	11549	18317	13818	7830	3982	14674	15714	5350	8209
Stop	TAA	0	0	0	0	0	0	0	0	0	0	0	0
Stop	TAG	0	0	0	0	0	0	0	0	0	0	0	0
Stop	TGA	0	0	0	0	0	0	0	0	0	0	0	0

Figura B.6: Tabela de contingência da *Candida albicans* codões/nucleótidos.

Apêndice B.3

	3' AAA	3' AAG	3' AAT	3' AAC	3' ACT	3' ACC	3' ACA	3' ACG	3' CGT	3' CGC	3' CGA	3' CGG	3' AGA	3' AGG	3' TCT	3' TCC	3' TCA	3' TCG	3' AGT	3' AGC	3' ATT	3' ATC	3' ATA	3' ATG	3' TTT	3' TTC	3' TAT	3' TAC	3' TGT	3' TGC	3' TGG
AAA	0,04	0,027	0,036	0,016	0,022	0,011	0,017	0,003	0,008	0,001	0,009	0,003	0,023	0,004	0,019	0,009	0,034	0,007	0,013	0,004	0,033	0,017	0,019	0,015	0,031	0,016	0,03	0,011	0,01	0,002	0,011
AAG	0,107	0,042	0,06	0,022	0,019	0,009	0,02	0,005	0,005	9E-04	0,003	0,002	0,029	0,006	0,014	0,006	0,018	0,008	0,018	0,006	0,039	0,014	0,017	0,015	0,038	0,012	0,026	0,012	0,006	0,002	0,005
AAT	0,047	0,012	0,064	0,018	0,021	0,01	0,013	0,003	0,006	8E-04	0,007	0,001	0,014	0,002	0,021	0,01	0,028	0,005	0,019	0,006	0,035	0,012	0,015	0,013	0,027	0,018	0,031	0,008	0,01	0,002	0,012
AAC	0,065	0,036	0,078	0,069	0,03	0,02	0,021	0,006	0,005	0,001	0,004	0,002	0,017	0,004	0,017	0,011	0,024	0,009	0,02	0,011	0,036	0,015	0,014	0,013	0,028	0,009	0,025	0,013	0,008	0,003	0,009
ACT	0,044	0,008	0,042	0,012	0,046	0,015	0,024	0,004	0,006	8E-04	0,005	6E-04	0,018	0,002	0,025	0,01	0,034	0,007	0,021	0,005	0,035	0,01	0,016	0,011	0,027	0,012	0,025	0,006	0,011	0,002	0,012
ACC	0,082	0,038	0,08	0,042	0,054	0,052	0,038	0,009	0,003	9E-04	0,003	6E-04	0,024	0,004	0,009	0,007	0,016	0,004	0,025	0,01	0,055	0,021	0,021	0,024	0,01	0,006	0,012	0,008	0,005	0,001	0,006
ACA	0,044	0,021	0,045	0,021	0,034	0,017	0,048	0,007	0,003	9E-04	0,004	0,001	0,025	0,004	0,024	0,008	0,038	0,009	0,019	0,008	0,039	0,015	0,024	0,018	0,041	0,013	0,018	0,008	0,008	0,002	0,007
ACG	0,069	0,015	0,033	0,013	0,018	0,013	0,033	0,012	0,004	7E-04	0,003	0,002	0,023	0,006	0,022	0,008	0,035	0,012	0,017	0,006	0,037	0,011	0,027	0,018	0,042	0,012	0,024	0,009	0,009	0,002	0,007
CGT	0,056	0,013	0,036	0,014	0,02	0,008	0,016	0,003	0,019	0,002	0,011	0,002	0,013	0,002	0,018	0,008	0,027	0,004	0,013	0,004	0,031	0,011	0,014	0,013	0,018	0,013	0,026	0,008	0,011	0,003	0,012
CGC	0,093	0,047	0,06	0,042	0,012	0,01	0,021	0,007	0,013	0,007	0,01	0,005	0,03	0,007	0,016	0,009	0,028	0,008	0,015	0,014	0,021	0,009	0,022	0,013	0,023	0,012	0,021	0,013	0,011	0,006	0,009
CGA	0,031	0,015	0,034	0,017	0,017	0,007	0,016	0,002	0,006	0,003	0,006	0,002	0,03	0,005	0,015	0,008	0,036	0,007	0,013	0,006	0,029	0,012	0,021	0,017	0,039	0,018	0,025	0,013	0,01	0,004	0,011
CGG	0,16	0,026	0,039	0,017	0,011	0,004	0,015	0,003	0,004	0,003	0,004	0,003	0,018	0,007	0,018	0,008	0,033	0,011	0,011	0,004	0,025	0,011	0,022	0,009	0,048	0,022	0,026	0,017	0,01	0,005	0,012
AGA	0,055	0,027	0,046	0,023	0,024	0,01	0,018	0,003	0,007	0,001	0,006	0,002	0,043	0,006	0,016	0,007	0,028	0,005	0,016	0,005	0,04	0,018	0,023	0,019	0,033	0,013	0,022	0,011	0,008	0,002	0,01
AGG	0,107	0,027	0,041	0,02	0,014	0,006	0,017	0,005	0,004	1E-03	0,006	0,002	0,021	0,008	0,018	0,006	0,027	0,009	0,016	0,007	0,028	0,013	0,024	0,016	0,053	0,015	0,032	0,015	0,01	0,004	0,01
TCT	0,046	0,009	0,042	0,012	0,034	0,011	0,024	0,004	0,009	0,001	0,008	0,001	0,013	0,002	0,045	0,015	0,046	0,009	0,02	0,005	0,036	0,009	0,017	0,012	0,022	0,013	0,022	0,006	0,011	0,002	0,011
TCC	0,094	0,04	0,082	0,039	0,04	0,02	0,024	0,006	0,006	0,001	0,004	0,001	0,02	0,004	0,017	0,013	0,026	0,009	0,027	0,01	0,048	0,018	0,022	0,023	0,024	0,011	0,014	0,008	0,006	0,002	0,006
TCA	0,04	0,015	0,044	0,017	0,031	0,012	0,031	0,006	0,005	0,001	0,006	0,001	0,025	0,004	0,029	0,009	0,055	0,012	0,022	0,007	0,047	0,012	0,026	0,02	0,042	0,014	0,02	0,007	0,007	0,002	0,007
TCG	0,07	0,012	0,037	0,011	0,024	0,011	0,026	0,006	0,004	5E-04	0,002	0,001	0,019	0,005	0,029	0,01	0,048	0,023	0,021	0,007	0,043	0,014	0,028	0,018	0,04	0,013	0,023	0,012	0,008	0,003	0,008
AGT	0,042	0,01	0,048	0,014	0,024	0,01	0,015	0,003	0,006	0,001	0,006	0,001	0,014	0,002	0,024	0,01	0,028	0,005	0,03	0,007	0,035	0,01	0,02	0,013	0,032	0,014	0,027	0,006	0,011	0,002	0,012
AGC	0,071	0,035	0,082	0,054	0,03	0,021	0,029	0,008	0,004	0,002	0,003	0,002	0,023	0,006	0,024	0,016	0,039	0,013	0,039	0,023	0,035	0,013	0,019	0,014	0,026	0,008	0,023	0,013	0,009	0,003	0,01
ATT	0,046	0,008	0,04	0,01	0,02	0,009	0,011	0,003	0,006	7E-04	0,006	1E-03	0,019	0,002	0,022	0,012	0,02	0,005	0,015	0,003	0,036	0,011	0,016	0,014	0,021	0,021	0,027	0,008	0,011	0,003	0,012
ATC	0,097	0,045	0,089	0,042	0,043	0,028	0,024	0,007	0,005	0,001	0,004	9E-04	0,027	0,004	0,009	0,009	0,014	0,004	0,023	0,007	0,055	0,02	0,017	0,022	0,019	0,009	0,016	0,01	0,006	0,002	0,007
ATA	0,038	0,021	0,052	0,019	0,027	0,012	0,021	0,006	0,005	0,001	0,006	0,002	0,017	0,005	0,027	0,011	0,034	0,011	0,015	0,005	0,043	0,016	0,027	0,021	0,045	0,016	0,026	0,009	0,01	0,003	0,01
ATG	0,057	0,02	0,049	0,015	0,021	0,009	0,026	0,006	0,003	6E-04	0,002	1E-03	0,023	0,005	0,021	0,008	0,031	0,009	0,021	0,005	0,043	0,012	0,021	0,026	0,034	0,01	0,025	0,009	0,009	0,002	0,009
TTT	0,04	0,008	0,036	0,009	0,017	0,008	0,013	0,002	0,003	6E-04	0,003	1E-03	0,016	0,002	0,022	0,012	0,024	0,005	0,015	0,004	0,042	0,012	0,019	0,017	0,023	0,032	0,028	0,009	0,011	0,004	0,013
TTC	0,088	0,033	0,08	0,032	0,038	0,019	0,023	0,005	0,01	0,002	0,007	0,002	0,022	0,003	0,011	0,01	0,013	0,005	0,023	0,006	0,046	0,014	0,017	0,016	0,034	0,018	0,018	0,012	0,01	0,002	0,009
TAT	0,045	0,011	0,046	0,012	0,021	0,009	0,011	0,003	0,01	0,001	0,009	0,001	0,014	0,002	0,022	0,01	0,024	0,006	0,013	0,003	0,039	0,013	0,014	0,015	0,033	0,022	0,036	0,01	0,013	0,003	0,013
TAC	0,056	0,041	0,057	0,033	0,03	0,018	0,018	0,007	0,009	0,002	0,005	0,002	0,02	0,003	0,014	0,01	0,016	0,009	0,016	0,007	0,046	0,014	0,015	0,017	0,032	0,008	0,03	0,018	0,012	0,004	0,013
TGT	0,039	0,009	0,035	0,012	0,02	0,008	0,013	0,003	0,013	0,002	0,011	0,003	0,007	0,001	0,023	0,01	0,025	0,006	0,008	0,003	0,043	0,012	0,019	0,017	0,037	0,021	0,025	0,008	0,024	0,006	0,013
TGC	0,068	0,035	0,053	0,034	0,02	0,017	0,021	0,008	0,009	0,004	0,007	0,002	0,019	0,006	0,02	0,012	0,024	0,014	0,023	0,012	0,039	0,018	0,023	0,018	0,056	0,018	0,031	0,018	0,021	0,013	0,021
TGG	0,061	0,02	0,047	0,016	0,018	0,008	0,021	0,004	0,004	5E-04	0,003	0,001	0,028	0,005	0,019	0,007	0,03	0,007	0,016	0,004	0,037	0,012	0,022	0,017	0,04	0,013	0,029	0,011	0,018	0,004	0,019
TTA	0,038	0,017	0,041	0,014	0,026	0,01	0,016	0,004	0,008	9E-04	0,007	0,001	0,022	0,003	0,023	0,009	0,037	0,009	0,015	0,003	0,042	0,013	0,016	0,018	0,036	0,014	0,026	0,01	0,011	0,002	0,01
TTG	0,083	0,022	0,058	0,02	0,027	0,008	0,022	0,006	0,005	1E-03	0,003	0,001	0,022	0,004	0,014	0,005	0,017	0,008	0,022	0,005	0,052	0,014	0,02	0,018	0,03	0,01	0,018	0,009	0,008	0,002	0,005
CTT	0,035	0,009	0,034	0,011	0,017	0,007	0,015	0,003	0,008	0,002	0,01	0,003	0,013	0,003	0,027	0,014	0,029	0,008	0,012	0,003	0,028	0,009	0,015	0,012	0,02	0,025	0,028	0,013	0,014	0,004	0,014
CTC	0,082	0,047	0,075	0,046	0,023	0,015	0,026	0,007	0,006	0,004	0,006	0,002	0,017	0,005	0,015	0,012	0,015	0,009	0,019	0,008	0,037	0,019	0,02	0,017	0,027	0,014	0,013	0,01	0,008	0,005	0,008
CTA	0,037	0,019	0,038	0,023	0,017	0,008	0,021	0,006	0,005	0,003	0,009	0,003	0,012	0,005	0,017	0,009	0,025	0,01	0,015	0,006	0,027	0,014	0,029	0,017	0,032	0,014	0,023	0,011	0,008	0,003	0,008
CTG	0,063	0,022	0,043	0,015	0,017	0,007	0,022	0,005	0,004	0,001	0,003	0,001	0,015	0,004	0,018	0,006	0,027	0,01	0,016	0,005	0,035	0,01	0,024	0,019	0,03	0,011	0,017	0,008	0,005	0,003	0,005
CCT	0,058	0,01	0,047	0,011	0,025	0,008	0,023	0,003	0,009	0,001	0,007	0,001	0,014	0,002	0,024	0,011	0,036	0,008	0,018	0,004	0,032	0,009	0,019	0,01	0,015	0,01	0,03	0,008	0,009	0,002	0,011
CCC	0,094	0,033	0,074	0,032	0,029	0,018	0,031	0,007	0,005	0,001	0,004	1E-03	0,016	0,004	0,012	0,006	0,014	0,006	0,022	0,009	0,048	0,018	0,029	0,027	0,025	0,008	0,016	0,005	0,006	0,002	0,005
CCA	0,039	0,018	0,039	0,017	0,029	0,012	0,03	0,005	0,004	9E-04	0,004	0,001	0,023	0,003	0,019	0,006	0,031	0,007	0,017	0,005	0,042	0,012	0,026	0,018	0,037	0,014	0,019	0,009	0,005	0,001	0,008
CCG	0,065	0,012	0,027	0,012	0,016	0,006	0,025	0,005	0,004	0,001	0,003	0,001	0,014	0,003	0,016	0,006	0,036	0,012	0,01	0,005	0,031	0,012	0,022	0,014	0,036	0,014	0,024	0,009	0,007	0,003	0,007
CAT	0,043	0,01	0,043	0,012	0,02	0,007	0,014	0,002	0,012	0,002	0,009	0,002	0,016	0,002	0,021	0,009	0,029	0,006	0,013	0,003	0,035	0,011	0,016	0,014	0,029	0,02	0,034	0,01	0,013	0,003	0,012
CAC	0,057	0,041	0,057	0,04	0,024	0,015	0,022	0,006	0,007	0,004	0,005	0,003	0,021	0,005	0,014	0,007	0,014	0,005	0,016	0,007	0,042	0,013	0,015	0,014	0,019	0,006	0,015	0,011	0,006	0,003	0,007
CAA	0,037	0,02	0,035	0,016	0,02	0,009	0,019	0,003	0,006	0,001	0,007	0,002	0,021	0,003	0,016	0,006	0,028	0,006	0,014	0,005	0,032	0,014	0,018	0,016	0,031	0,013	0,025	0,009	0,008	0,002	0,01
CAG	0,059	0,021	0,05	0,018	0,014	0,006	0,017	0,005	0,004	0,001	0,004	0,002	0,016	0,004	0,014	0,007	0,022	0,009	0,013	0,005	0,029	0,011	0,019	0,015	0,043	0,011	0,03	0,014	0,008	0,002	0,005
GTT	0,043	0,009	0,032	0,01	0,022	0,009	0,012	0,003	0,007	7E-04	0,004	1E-03	0,018	0,002	0,025	0,012	0,022	0,005	0,016	0,003	0,035	0,01	0,013	0,013	0,016	0,022	0,028	0,009	0,012	0,002	0,011
GTC	0,084	0,046	0,086	0,04	0,051	0,027	0,028	0,008	0,004	7E-04	0,003	8E-04	0,024	0,005	0,011	0,008	0,014	0,005	0,026	0,008	0,059	0,016	0,018	0,023	0,018	0,007	0,014	0,011	0,005	0,002	0,005
GTA	0,024	0,02	0,032	0,016	0,021	0,011	0,02	0,006	0,006	0,001	0,005	0,002	0,017	0,004	0,029	0,01	0,041	0,012	0,013	0,005	0,031	0,011	0,019	0,022	0,053	0,017	0,026	0,012	0,011	0,003	0,014
GTG	0,062	0,017	0,038	0,013	0,02	0,007	0,026	0,006	0,004	8E-04	0,002	0,001	0,017	0,004	0,017	0,007	0,02	0,01	0,014	0,005	0,048	0,012	0,025	0,019	0,048	0,011	0,02	0,009	0,009	0,003	0,008
GCT	0,052	0,01	0,038	0,011	0,033	0,01	0,017	0,003	0,008	6E-04	0,004	8E-04	0,022	0,002	0,025	0,009	0,023	0,005	0,021	0,003	0,036	0,009	0,015	0,012	0,016	0,01	0,026	0,008	0,012	0,002	0,01
GCC	0,089	0,039	0,078	0,035	0,047	0,024	0,027	0,008	0,004	6E-04	0,002	6E-04	0,029	0,004	0,011	0,007	0,013	0,006	0,026	0,006	0,07	0,018	0,023	0,034	0,023	0,008	0,014	0,006	0,006	0,001	0,005
GCA	0,038	0,019	0,039	0,017	0,027	0,011	0,029	0,005	0,005	7E-04	0,004	0,001	0,028	0,004	0,023	0,009	0,034	0,009	0,017	0,005	0,041	0,013	0,019	0,018	0,046	0,016	0,021	0,009	0,01	0,003	0,009
GCG	0,053	0,01	0,025	0,012	0,017	0,008	0,024	0,007	0,002	0,001	0,002	0,001	0,016	0,006	0,02	0,009	0,037	0,016	0,011	0,004	0,041	0,012	0,02	0,021	0,049	0,018	0,029	0,015	0,014	0,004	0,011
GGT	0,058	0,013	0,046	0,016	0,03	0,01	0,015	0,003	0,006	6E-04	0,004	8E-04	0,023	0,002	0,02	0,008	0,022	0,003	0,019	0,003	0,043	0,012	0,016	0,016	0,013	0,01	0,024	0,008	0,011	0,002	0,013
GGC	0,066	0,034	0,068	0,035	0,03	0,013	0,027	0,008	0,008	0,002	0,004	1E-03	0,023	0,005	0,03	0,01	0,028	0,009	0,028	0,012	0,032	0,01	0,017	0,014	0,019	0,007	0,025	0,015	0,011	0,003	0,015
GGA	0,026	0,011	0,035	0,016	0,023	0,01	0,022	0,003	0,005	9E-04	0,003	0,001	0,027	0,003	0,026	0,009	0,04	0,007	0,02	0,007	0,033	0,013	0,023	0,018	0,051	0,019	0,032	0,013	0,012	0,003	0,011
GGG	0,098	0,018	0,049	0,017	0,016	0,006	0,021	0,003	0,002	4E-04	0,002	8E-04	0,015	0,003	0,022	0,009	0,028	0,009	0,012	0,004	0,042	0,014	0,027	0,018	0,077	0,027	0,033	0,014	0,011	0,003	0,008
GAT	0,047	0,012	0,046	0,012	0,02	0,009	0,013	0,003	0,006	5E-04	0,004	1E-03	0,016	0,003	0,02	0,009	0,024	0,005	0,015	0,004	0,042	0,013	0,019	0,015	0,03	0,018	0,033	0,009	0,009	0,002	0,011
GAC	0,059	0,037	0,063	0,034	0,026	0,016	0,018	0,005	0,006	0,001	0,003	0,001	0,018	0,004	0,02	0,01	0,019	0,01	0,018	0,007	0,041	0,013	0,015	0,016	0,03	0,007	0,022	0,013	0,006	0,002	0,008
GAA	0,048	0,025	0,044	0,019	0,025	0,012	0,018	0,004	0,005	6E-04	0,005	0,001	0,024	0,004	0,016	0,007	0,025	0,006	0,018	0,005	0,039	0,017	0,019	0,017	0,029	0,011	0,024	0,009	0,008	0,002	0,009
GAG	0,067	0,025	0,056	0,017	0,017	0,008	0,017	0,006	0,003	7E-04	0,003	0,001	0,015	0,005	0,018	0,007	0,02	0,009	0,019	0,006	0,042	0,012	0,02	0,015	0,051	0,012	0,032	0,013	0,006	0,002	0,006

Figura B.7: Estimativa da matriz das probabilidades de transição da *Candida albicans* codões/nucleótidos.

	3' TTA	3' TTG	3' CTT	3' CTC	3' CTA	3' CTG	3' CCT	3' CCC	3' CCA	3' CCG	3' CAT	3' CAC	3' CAA	3' CAG	3' GTT	3' GTC	3' GTA	3' GTG	3' GCT	3' GCC	3' GCA	3' GCG	3' GGT	3' GGC	3' GGA	3' GGG	3' GAT	3' GAC	3' GAA	3' GAG	TOTAL
AAA	0,049	0,041	0,013	0,004	0,011	0,007	0,012	0,004	0,028	0,003	0,017	0,006	0,041	0,009	0,026	0,009	0,012	0,012	0,018	0,007	0,013	0,002	0,019	0,003	0,009	0,005	0,044	0,012	0,049	0,015	1
AAG	0,022	0,035	0,011	0,004	0,009	0,005	0,007	0,003	0,017	0,003	0,013	0,006	0,029	0,008	0,015	0,006	0,008	0,009	0,015	0,007	0,025	0,004	0,013	0,004	0,014	0,01	0,034	0,012	0,052	0,017	1
AAT	0,041	0,031	0,011	0,003	0,006	0,004	0,012	0,004	0,019	0,002	0,017	0,004	0,038	0,005	0,028	0,01	0,009	0,01	0,021	0,012	0,014	0,002	0,028	0,007	0,02	0,01	0,056	0,015	0,057	0,013	1
AAC	0,017	0,041	0,01	0,004	0,006	0,004	0,016	0,008	0,027	0,004	0,017	0,012	0,036	0,012	0,011	0,005	0,004	0,006	0,007	0,006	0,006	0,001	0,013	0,004	0,008	0,005	0,03	0,015	0,029	0,013	1
ACT	0,037	0,029	0,007	0,002	0,004	0,003	0,016	0,004	0,027	0,004	0,018	0,004	0,031	0,005	0,029	0,009	0,008	0,01	0,034	0,013	0,018	0,003	0,039	0,005	0,017	0,01	0,047	0,011	0,051	0,01	1
ACC	0,013	0,015	0,009	0,003	0,004	0,004	0,015	0,005	0,026	0,004	0,016	0,007	0,027	0,007	0,019	0,006	0,006	0,009	0,009	0,007	0,007	0,001	0,014	0,003	0,007	0,006	0,032	0,012	0,026	0,012	1
ACA	0,049	0,029	0,009	0,003	0,006	0,005	0,016	0,006	0,037	0,004	0,01	0,005	0,024	0,006	0,022	0,006	0,013	0,011	0,014	0,006	0,022	0,002	0,012	0,003	0,014	0,007	0,026	0,01	0,036	0,011	1
ACG	0,034	0,036	0,008	0,003	0,008	0,004	0,01	0,005	0,022	0,006	0,007	0,003	0,017	0,005	0,035	0,008	0,019	0,022	0,016	0,009	0,035	0,006	0,022	0,005	0,018	0,019	0,022	0,008	0,036	0,011	1
CGT	0,031	0,026	0,013	0,003	0,005	0,004	0,015	0,005	0,028	0,003	0,028	0,008	0,049	0,008	0,026	0,006	0,008	0,008	0,026	0,013	0,015	0,002	0,033	0,007	0,019	0,01	0,06	0,016	0,065	0,014	1
CGC	0,021	0,021	0,01	0,007	0,009	0,01	0,013	0,008	0,016	0,007	0,026	0,024	0,043	0,023	0,007	0,005	0,008	0,006	0,005	0,004	0,005	9E-04	0,011	0,002	0,011	0,007	0,028	0,012	0,019	0,015	1
CGA	0,065	0,029	0,011	0,005	0,008	0,006	0,008	0,003	0,031	0,003	0,012	0,007	0,048	0,009	0,024	0,009	0,019	0,011	0,018	0,009	0,024	0,002	0,017	0,006	0,019	0,006	0,036	0,015	0,049	0,012	1
CGG	0,037	0,039	0,011	0,006	0,009	0,007	0,005	0,005	0,02	0,003	0,012	0,004	0,019	0,009	0,018	0,007	0,011	0,01	0,01	0,006	0,018	0,005	0,01	0,005	0,012	0,005	0,024	0,013	0,037	0,014	1
AGA	0,046	0,032	0,009	0,003	0,007	0,004	0,008	0,003	0,022	0,002	0,014	0,007	0,033	0,007	0,024	0,008	0,012	0,01	0,02	0,009	0,016	0,002	0,023	0,004	0,013	0,005	0,038	0,012	0,044	0,011	1
AGG	0,034	0,042	0,01	0,005	0,012	0,007	0,003	0,004	0,013	0,004	0,011	0,005	0,021	0,007	0,023	0,008	0,013	0,011	0,014	0,01	0,021	0,004	0,014	0,007	0,013	0,006	0,03	0,011	0,039	0,016	1
TCT	0,035	0,029	0,009	0,003	0,004	0,004	0,014	0,005	0,025	0,003	0,019	0,004	0,038	0,006	0,025	0,009	0,008	0,01	0,03	0,011	0,015	0,002	0,035	0,006	0,016	0,008	0,045	0,012	0,042	0,009	1
TCC	0,023	0,027	0,012	0,004	0,005	0,005	0,01	0,007	0,018	0,003	0,017	0,008	0,03	0,009	0,013	0,005	0,005	0,006	0,008	0,004	0,005	8E-04	0,016	0,003	0,006	0,005	0,031	0,012	0,024	0,009	1
TCA	0,054	0,028	0,012	0,003	0,007	0,007	0,015	0,005	0,032	0,004	0,011	0,004	0,033	0,007	0,021	0,005	0,014	0,012	0,015	0,005	0,018	0,002	0,013	0,003	0,012	0,006	0,024	0,008	0,033	0,008	1
TCG	0,034	0,045	0,011	0,005	0,006	0,006	0,012	0,005	0,027	0,005	0,009	0,003	0,021	0,007	0,025	0,008	0,017	0,018	0,015	0,008	0,023	0,004	0,017	0,003	0,013	0,007	0,022	0,008	0,032	0,008	1
AGT	0,041	0,03	0,01	0,003	0,007	0,004	0,01	0,004	0,013	0,002	0,016	0,005	0,028	0,005	0,029	0,008	0,01	0,01	0,026	0,011	0,015	0,002	0,038	0,01	0,023	0,012	0,059	0,015	0,054	0,013	1
AGC	0,021	0,028	0,007	0,004	0,004	0,003	0,008	0,006	0,013	0,002	0,016	0,011	0,027	0,011	0,008	0,003	0,004	0,004	0,005	0,003	0,004	0,001	0,011	0,004	0,007	0,005	0,03	0,015	0,025	0,013	1
ATT	0,035	0,03	0,012	0,003	0,005	0,004	0,016	0,004	0,02	0,003	0,018	0,004	0,033	0,005	0,039	0,012	0,009	0,011	0,034	0,018	0,014	0,003	0,034	0,006	0,019	0,009	0,067	0,015	0,068	0,013	1
ATC	0,018	0,027	0,009	0,004	0,005	0,003	0,02	0,012	0,026	0,005	0,018	0,008	0,029	0,008	0,012	0,006	0,003	0,005	0,007	0,006	0,004	0,001	0,008	0,002	0,005	0,003	0,03	0,012	0,023	0,011	1
ATA	0,05	0,036	0,015	0,005	0,015	0,007	0,017	0,007	0,024	0,007	0,012	0,006	0,026	0,008	0,02	0,006	0,011	0,011	0,014	0,007	0,013	0,003	0,011	0,003	0,009	0,004	0,036	0,012	0,032	0,012	1
ATG	0,027	0,032	0,014	0,003	0,008	0,006	0,011	0,004	0,02	0,004	0,011	0,005	0,027	0,006	0,026	0,008	0,014	0,017	0,02	0,01	0,025	0,004	0,023	0,004	0,017	0,011	0,042	0,011	0,044	0,013	1
TTT	0,04	0,031	0,015	0,004	0,006	0,005	0,009	0,003	0,008	0,002	0,011	0,003	0,026	0,005	0,041	0,013	0,011	0,012	0,033	0,017	0,017	0,003	0,037	0,007	0,024	0,009	0,064	0,016	0,065	0,013	1
TTC	0,025	0,03	0,015	0,005	0,005	0,004	0,022	0,013	0,039	0,006	0,031	0,009	0,055	0,01	0,009	0,004	0,002	0,003	0,006	0,004	0,003	5E-04	0,011	0,001	0,005	0,003	0,02	0,006	0,017	0,005	1
TAT	0,046	0,04	0,016	0,004	0,005	0,004	0,015	0,004	0,019	0,003	0,021	0,005	0,042	0,006	0,028	0,009	0,007	0,011	0,023	0,012	0,012	0,003	0,027	0,005	0,019	0,009	0,05	0,012	0,049	0,011	1
TAC	0,014	0,047	0,013	0,005	0,004	0,005	0,015	0,005	0,021	0,004	0,016	0,012	0,036	0,013	0,02	0,007	0,004	0,007	0,012	0,007	0,007	0,001	0,017	0,004	0,01	0,005	0,039	0,019	0,033	0,014	1
TGT	0,046	0,033	0,021	0,006	0,007	0,004	0,015	0,005	0,019	0,003	0,02	0,006	0,034	0,006	0,034	0,009	0,009	0,011	0,025	0,013	0,011	0,002	0,034	0,008	0,024	0,012	0,047	0,014	0,039	0,009	1
TGC	0,029	0,035	0,012	0,006	0,006	0,006	0,011	0,005	0,012	0,003	0,027	0,014	0,032	0,01	0,008	0,003	0,003	0,008	0,005	0,003	0,003	0,002	0,01	0,002	0,008	0,004	0,018	0,01	0,015	0,007	1
TGG	0,046	0,03	0,011	0,005	0,008	0,003	0,004	0,005	0,019	0,002	0,016	0,006	0,03	0,006	0,023	0,01	0,014	0,011	0,018	0,009	0,017	0,002	0,024	0,004	0,012	0,012	0,046	0,011	0,043	0,011	1
TTA	0,054	0,035	0,013	0,003	0,008	0,006	0,02	0,007	0,031	0,004	0,018	0,006	0,05	0,008	0,022	0,007	0,01	0,011	0,019	0,008	0,016	0,002	0,019	0,002	0,014	0,006	0,039	0,008	0,042	0,011	1
TTG	0,027	0,032	0,01	0,004	0,006	0,004	0,009	0,004	0,018	0,003	0,011	0,005	0,023	0,006	0,028	0,009	0,013	0,016	0,022	0,011	0,024	0,004	0,025	0,005	0,014	0,008	0,044	0,013	0,047	0,013	1
CTT	0,033	0,03	0,016	0,005	0,01	0,009	0,013	0,005	0,019	0,004	0,019	0,006	0,034	0,008	0,033	0,01	0,012	0,015	0,025	0,015	0,019	0,004	0,03	0,006	0,019	0,009	0,056	0,015	0,06	0,015	1
CTC	0,015	0,025	0,014	0,014	0,009	0,01	0,013	0,009	0,02	0,005	0,021	0,013	0,033	0,018	0,011	0,007	0,006	0,006	0,008	0,005	0,006	0,001	0,01	0,004	0,006	0,004	0,031	0,018	0,03	0,013	1
CTA	0,035	0,028	0,02	0,007	0,025	0,018	0,017	0,007	0,03	0,009	0,015	0,008	0,043	0,015	0,019	0,008	0,013	0,015	0,017	0,008	0,022	0,004	0,01	0,005	0,01	0,005	0,037	0,016	0,039	0,018	1
CTG	0,022	0,032	0,015	0,009	0,013	0,024	0,008	0,004	0,024	0,006	0,016	0,007	0,042	0,012	0,028	0,007	0,02	0,022	0,017	0,008	0,023	0,004	0,015	0,006	0,015	0,007	0,042	0,01	0,055	0,015	1
CCT	0,027	0,023	0,009	0,003	0,004	0,005	0,021	0,004	0,037	0,004	0,026	0,004	0,048	0,008	0,026	0,007	0,01	0,011	0,025	0,009	0,016	0,003	0,028	0,005	0,017	0,008	0,055	0,012	0,059	0,011	1
CCC	0,018	0,029	0,01	0,006	0,005	0,008	0,009	0,003	0,019	0,003	0,017	0,009	0,037	0,012	0,021	0,007	0,008	0,016	0,008	0,004	0,009	9E-04	0,01	0,003	0,007	0,005	0,042	0,012	0,039	0,017	1
CCA	0,043	0,029	0,01	0,003	0,008	0,007	0,02	0,004	0,054	0,007	0,012	0,005	0,041	0,009	0,026	0,006	0,015	0,012	0,02	0,005	0,023	0,002	0,019	0,003	0,012	0,005	0,026	0,009	0,051	0,011	1
CCG	0,04	0,045	0,011	0,003	0,009	0,012	0,018	0,005	0,047	0,008	0,011	0,004	0,038	0,01	0,029	0,007	0,021	0,023	0,017	0,006	0,024	0,005	0,014	0,004	0,019	0,008	0,027	0,011	0,054	0,014	1
CAT	0,048	0,038	0,012	0,004	0,007	0,006	0,018	0,004	0,026	0,003	0,033	0,01	0,049	0,008	0,024	0,008	0,01	0,008	0,018	0,01	0,015	0,003	0,024	0,007	0,018	0,007	0,043	0,011	0,048	0,009	1
CAC	0,01	0,033	0,01	0,008	0,007	0,007	0,013	0,008	0,026	0,005	0,031	0,034	0,054	0,02	0,02	0,006	0,006	0,006	0,011	0,007	0,009	0,001	0,017	0,004	0,01	0,005	0,039	0,015	0,04	0,016	1
CAA	0,041	0,037	0,01	0,003	0,008	0,009	0,011	0,004	0,031	0,003	0,017	0,007	0,096	0,021	0,023	0,006	0,011	0,011	0,018	0,007	0,018	0,002	0,019	0,003	0,012	0,005	0,038	0,011	0,054	0,014	1
CAG	0,026	0,044	0,014	0,005	0,013	0,01	0,01	0,005	0,028	0,005	0,018	0,01	0,084	0,034	0,016	0,006	0,011	0,012	0,013	0,006	0,018	0,005	0,011	0,004	0,011	0,008	0,032	0,012	0,039	0,015	1
GTT	0,036	0,031	0,01	0,002	0,004	0,004	0,016	0,004	0,022	0,003	0,016	0,004	0,028	0,005	0,045	0,013	0,01	0,011	0,037	0,018	0,016	0,003	0,039	0,006	0,019	0,009	0,066	0,016	0,065	0,014	1
GTC	0,015	0,028	0,007	0,003	0,003	0,003	0,014	0,009	0,02	0,003	0,015	0,007	0,024	0,007	0,016	0,007	0,005	0,006	0,01	0,007	0,005	0,001	0,015	0,002	0,006	0,005	0,038	0,016	0,029	0,012	1
GTA	0,061	0,051	0,015	0,004	0,01	0,007	0,019	0,007	0,03	0,007	0,014	0,006	0,033	0,009	0,021	0,007	0,017	0,013	0,018	0,007	0,02	0,002	0,011	0,003	0,01	0,007	0,026	0,009	0,031	0,012	1
GTG	0,031	0,047	0,011	0,004	0,01	0,006	0,01	0,005	0,02	0,005	0,01	0,005	0,021	0,006	0,037	0,009	0,021	0,027	0,019	0,01	0,027	0,006	0,016	0,005	0,014	0,011	0,032	0,01	0,047	0,015	1
GCT	0,03	0,026	0,011	0,002	0,004	0,003	0,015	0,003	0,026	0,003	0,017	0,003	0,035	0,005	0,033	0,009	0,01	0,01	0,057	0,019	0,021	0,003	0,046	0,006	0,017	0,009	0,049	0,011	0,054	0,009	1
GCC	0,02	0,031	0,011	0,003	0,003	0,003	0,01	0,005	0,018	0,002	0,015	0,006	0,024	0,007	0,022	0,005	0,005	0,009	0,015	0,007	0,008	0,001	0,018	0,002	0,008	0,004	0,029	0,01	0,028	0,01	1
GCA	0,061	0,042	0,011	0,003	0,005	0,004	0,014	0,004	0,032	0,004	0,009	0,004	0,028	0,007	0,024	0,006	0,014	0,013	0,022	0,009	0,033	0,003	0,014	0,003	0,013	0,006	0,025	0,009	0,039	0,012	1
GCG	0,047	0,055	0,009	0,003	0,006	0,006	0,009	0,004	0,022	0,004	0,009	0,004	0,016	0,007	0,034	0,009	0,024	0,026	0,018	0,011	0,032	0,007	0,015	0,005	0,015	0,012	0,023	0,006	0,035	0,011	1
GGT	0,027	0,019	0,009	0,002	0,003	0,002	0,009	0,003	0,017	0,001	0,017	0,005	0,028	0,003	0,037	0,011	0,011	0,01	0,037	0,015	0,017	0,002	0,064	0,01	0,025	0,013	0,057	0,015	0,051	0,009	1
GGC	0,02	0,021	0,006	0,003	0,003	0,002	0,008	0,006	0,015	0,002	0,024	0,012	0,034	0,01	0,014	0,005	0,006	0,007	0,011	0,005	0,007	0,001	0,029	0,009	0,014	0,008	0,039	0,02	0,034	0,014	1
GGA	0,071	0,035	0,011	0,003	0,006	0,003	0,01	0,003	0,019	0,002	0,013	0,005	0,031	0,005	0,025	0,007	0,014	0,008	0,017	0,007	0,019	0,001	0,025	0,006	0,022	0,007	0,033	0,011	0,043	0,011	1
GGG	0,057	0,054	0,008	0,003	0,007	0,003	0,003	0,003	0,011	0,001	0,009	0,003	0,011	0,004	0,038	0,009	0,018	0,014	0,014	0,007	0,019	0,002	0,012	0,003	0,007	0,004	0,025	0,007	0,028	0,009	1
GAT	0,046	0,038	0,009	0,002	0,005	0,003	0,012	0,004	0,02	0,002	0,014	0,003	0,03	0,005	0,03	0,008	0,009	0,011	0,022	0,011	0,014	0,002	0,024	0,005	0,015	0,009	0,069	0,019	0,072	0,015	1
GAC	0,013	0,044	0,009	0,003	0,004	0,004	0,014	0,007	0,022	0,004	0,015	0,008	0,033	0,01	0,019	0,005	0,005	0,008	0,01	0,006	0,008	0,002	0,012	0,004	0,007	0,005	0,059	0,031	0,055	0,025	1
GAA	0,042	0,039	0,009	0,003	0,007	0,004	0,008	0,003	0,019	0,002	0,012	0,004	0,035	0,007	0,025	0,007	0,01	0,01	0,019	0,008	0,017	0,002	0,02	0,003	0,012	0,006	0,053	0,014	0,081	0,021	1
GAG	0,03	0,051	0,013	0,004	0,011	0,008	0,008	0,003	0,021	0,003	0,012	0,005	0,029	0,009	0,02	0,006	0,01	0,012	0,013	0,007	0,019	0,004	0,008	0,003	0,008	0,005	0,043	0,016	0,064	0,027	1

Figura B.8: Estimativa da matriz das probabilidades de transição da *Candida albicans* codões/nucleótidos (continuação).

	3' AAA	3' AAG	3' AAT	3' AAC	3' ACT	3' ACC	3' ACA	3' ACG	3' CGT	3' CGC	3' CGA	3' CGG	3' AGA	3' AGG	3' TCT	3' TCC	3' TCA	3' TCG	3' AGT	3' AGC	3' ATT	3' ATC	3' ATA	3' ATG	3' TTT	3' TTC	3' TAT	3' TAC	3' TGT	3' TGC	3' TGG
AAA	0,044	0,033	0,039	0,024	0,018	0,01	0,019	0,008	0,006	0,002	0,004	0,002	0,028	0,014	0,02	0,011	0,02	0,009	0,016	0,011	0,032	0,017	0,022	0,02	0,029	0,018	0,023	0,017	0,008	0,005	0,013
AAG	0,066	0,048	0,035	0,026	0,017	0,011	0,018	0,008	0,006	0,003	0,003	0,002	0,027	0,012	0,014	0,009	0,013	0,006	0,012	0,009	0,026	0,016	0,016	0,015	0,026	0,016	0,018	0,016	0,006	0,004	0,008
AAT	0,039	0,023	0,045	0,029	0,018	0,011	0,018	0,008	0,005	0,002	0,003	0,001	0,015	0,008	0,023	0,013	0,021	0,009	0,017	0,011	0,032	0,017	0,018	0,018	0,026	0,018	0,018	0,013	0,007	0,004	0,01
AAC	0,047	0,036	0,048	0,037	0,022	0,014	0,019	0,008	0,006	0,003	0,002	0,001	0,021	0,009	0,025	0,015	0,018	0,008	0,019	0,015	0,03	0,018	0,018	0,018	0,024	0,017	0,02	0,017	0,008	0,005	0,012
ACT	0,029	0,021	0,03	0,017	0,028	0,018	0,021	0,009	0,004	0,002	0,002	0,001	0,014	0,005	0,032	0,018	0,025	0,01	0,014	0,008	0,027	0,016	0,016	0,013	0,029	0,022	0,016	0,013	0,007	0,004	0,01
ACC	0,053	0,043	0,042	0,034	0,036	0,023	0,026	0,013	0,005	0,002	0,002	0,001	0,021	0,008	0,025	0,013	0,02	0,008	0,016	0,012	0,039	0,025	0,023	0,02	0,021	0,017	0,014	0,013	0,007	0,004	0,01
ACA	0,041	0,032	0,042	0,027	0,02	0,013	0,026	0,01	0,005	0,002	0,003	0,002	0,029	0,012	0,024	0,014	0,022	0,009	0,018	0,011	0,027	0,012	0,02	0,02	0,029	0,014	0,019	0,012	0,008	0,005	0,011
ACG	0,046	0,029	0,037	0,023	0,016	0,011	0,02	0,009	0,004	0,002	0,004	0,002	0,021	0,011	0,019	0,012	0,022	0,009	0,014	0,011	0,021	0,012	0,021	0,018	0,027	0,016	0,02	0,013	0,007	0,006	0,01
CGT	0,036	0,024	0,026	0,015	0,017	0,012	0,013	0,006	0,018	0,006	0,005	0,003	0,016	0,007	0,024	0,017	0,019	0,008	0,011	0,006	0,032	0,017	0,016	0,015	0,017	0,015	0,024	0,02	0,013	0,007	0,014
CGC	0,046	0,033	0,035	0,024	0,016	0,013	0,021	0,008	0,015	0,007	0,005	0,005	0,021	0,011	0,026	0,013	0,019	0,009	0,019	0,018	0,024	0,016	0,023	0,019	0,02	0,014	0,024	0,028	0,01	0,009	0,016
CGA	0,043	0,025	0,043	0,027	0,016	0,013	0,017	0,01	0,006	0,004	0,003	0,003	0,028	0,014	0,02	0,009	0,018	0,009	0,018	0,01	0,028	0,015	0,027	0,024	0,03	0,015	0,023	0,021	0,009	0,01	0,011
CGG	0,054	0,032	0,043	0,02	0,01	0,007	0,018	0,007	0,006	0,005	0,004	0,005	0,017	0,014	0,025	0,016	0,022	0,017	0,014	0,007	0,022	0,012	0,018	0,019	0,04	0,021	0,027	0,019	0,008	0,008	0,009
AGA	0,046	0,038	0,041	0,03	0,02	0,012	0,019	0,008	0,006	0,003	0,003	0,002	0,04	0,016	0,018	0,01	0,019	0,008	0,016	0,011	0,03	0,017	0,021	0,02	0,028	0,015	0,018	0,013	0,007	0,004	0,01
AGG	0,064	0,036	0,031	0,023	0,018	0,01	0,018	0,008	0,008	0,004	0,005	0,004	0,027	0,012	0,019	0,01	0,02	0,01	0,012	0,009	0,026	0,012	0,023	0,017	0,033	0,018	0,018	0,014	0,006	0,005	0,009
TCT	0,033	0,022	0,031	0,019	0,026	0,014	0,02	0,009	0,008	0,004	0,004	0,001	0,013	0,007	0,04	0,022	0,03	0,013	0,014	0,009	0,026	0,015	0,014	0,016	0,022	0,016	0,015	0,011	0,006	0,004	0,008
TCC	0,058	0,043	0,044	0,031	0,035	0,019	0,025	0,011	0,007	0,003	0,003	0,001	0,022	0,01	0,032	0,019	0,024	0,012	0,022	0,015	0,039	0,025	0,022	0,025	0,026	0,017	0,013	0,01	0,006	0,004	0,008
TCA	0,041	0,028	0,04	0,025	0,021	0,01	0,018	0,009	0,005	0,003	0,005	0,002	0,028	0,013	0,03	0,019	0,031	0,014	0,018	0,013	0,023	0,012	0,018	0,021	0,03	0,017	0,02	0,013	0,008	0,005	0,01
TCG	0,05	0,035	0,04	0,025	0,017	0,01	0,019	0,01	0,005	0,003	0,003	0,002	0,024	0,012	0,023	0,018	0,025	0,014	0,015	0,013	0,023	0,014	0,021	0,023	0,03	0,019	0,02	0,017	0,007	0,005	0,011
AGT	0,043	0,022	0,04	0,025	0,019	0,015	0,021	0,01	0,005	0,003	0,003	0,002	0,015	0,008	0,026	0,015	0,025	0,01	0,02	0,014	0,028	0,014	0,021	0,015	0,025	0,019	0,016	0,011	0,009	0,005	0,009
AGC	0,049	0,029	0,047	0,033	0,028	0,019	0,023	0,012	0,005	0,004	0,003	0,002	0,021	0,011	0,03	0,015	0,022	0,01	0,025	0,019	0,03	0,016	0,023	0,017	0,023	0,015	0,018	0,011	0,007	0,005	0,009
ATT	0,029	0,019	0,025	0,017	0,018	0,011	0,014	0,006	0,006	0,002	0,003	0,002	0,015	0,007	0,027	0,016	0,02	0,008	0,01	0,006	0,029	0,017	0,016	0,015	0,026	0,022	0,018	0,013	0,009	0,006	0,011
ATC	0,063	0,045	0,047	0,037	0,026	0,017	0,02	0,009	0,005	0,003	0,002	0,001	0,026	0,01	0,019	0,014	0,014	0,006	0,018	0,011	0,035	0,025	0,02	0,024	0,021	0,017	0,014	0,013	0,008	0,005	0,009
ATA	0,043	0,028	0,04	0,026	0,018	0,01	0,021	0,01	0,005	0,003	0,006	0,004	0,023	0,011	0,027	0,016	0,027	0,012	0,014	0,01	0,025	0,014	0,023	0,021	0,033	0,02	0,024	0,017	0,01	0,006	0,014
ATG	0,048	0,03	0,039	0,026	0,017	0,013	0,022	0,009	0,006	0,002	0,002	0,002	0,022	0,01	0,017	0,011	0,015	0,008	0,014	0,01	0,032	0,016	0,019	0,024	0,027	0,015	0,017	0,014	0,007	0,005	0,009
TTT	0,028	0,016	0,025	0,017	0,017	0,013	0,013	0,008	0,003	0,002	0,003	0,002	0,015	0,007	0,025	0,015	0,02	0,009	0,011	0,008	0,03	0,017	0,014	0,015	0,028	0,026	0,021	0,018	0,008	0,006	0,012
TTC	0,061	0,043	0,048	0,034	0,024	0,017	0,017	0,008	0,007	0,003	0,003	0,002	0,022	0,01	0,017	0,012	0,013	0,006	0,016	0,011	0,038	0,024	0,018	0,022	0,026	0,023	0,016	0,016	0,008	0,005	0,011
TAT	0,034	0,021	0,031	0,02	0,015	0,01	0,016	0,009	0,007	0,003	0,003	0,002	0,016	0,008	0,025	0,015	0,018	0,009	0,01	0,006	0,03	0,018	0,016	0,018	0,03	0,023	0,025	0,016	0,01	0,006	0,011
TAC	0,044	0,036	0,038	0,028	0,021	0,012	0,019	0,008	0,008	0,003	0,002	0,002	0,021	0,011	0,021	0,013	0,015	0,007	0,012	0,008	0,029	0,022	0,016	0,022	0,026	0,02	0,025	0,021	0,012	0,006	0,013
TGT	0,034	0,019	0,026	0,019	0,018	0,012	0,014	0,008	0,008	0,003	0,004	0,002	0,014	0,008	0,025	0,015	0,02	0,009	0,009	0,007	0,033	0,02	0,018	0,017	0,031	0,022	0,02	0,013	0,013	0,008	0,012
TGC	0,039	0,029	0,031	0,019	0,019	0,013	0,017	0,007	0,01	0,003	0,004	0,003	0,018	0,011	0,026	0,017	0,019	0,008	0,014	0,012	0,035	0,019	0,025	0,023	0,039	0,022	0,025	0,018	0,018	0,011	0,013
TGG	0,052	0,033	0,038	0,025	0,015	0,011	0,016	0,008	0,003	0,002	0,003	0,001	0,028	0,012	0,023	0,013	0,021	0,009	0,011	0,008	0,028	0,014	0,02	0,018	0,03	0,02	0,019	0,018	0,014	0,008	0,014
TTA	0,039	0,03	0,032	0,023	0,02	0,013	0,019	0,009	0,007	0,003	0,004	0,002	0,024	0,011	0,023	0,015	0,021	0,01	0,015	0,01	0,026	0,015	0,017	0,019	0,027	0,017	0,017	0,013	0,009	0,005	0,011
TTG	0,064	0,045	0,042	0,033	0,02	0,014	0,016	0,008	0,006	0,003	0,003	0,002	0,025	0,011	0,015	0,011	0,013	0,005	0,015	0,011	0,032	0,017	0,019	0,02	0,021	0,013	0,014	0,012	0,006	0,004	0,006
CTT	0,022	0,013	0,019	0,015	0,015	0,009	0,011	0,006	0,007	0,003	0,006	0,003	0,01	0,004	0,034	0,025	0,027	0,012	0,007	0,005	0,024	0,013	0,012	0,01	0,032	0,036	0,029	0,02	0,013	0,008	0,013
CTC	0,047	0,032	0,043	0,027	0,023	0,018	0,018	0,009	0,005	0,004	0,005	0,002	0,015	0,008	0,027	0,017	0,019	0,011	0,019	0,013	0,033	0,021	0,023	0,018	0,033	0,026	0,021	0,015	0,012	0,008	0,01
CTA	0,04	0,028	0,036	0,023	0,016	0,01	0,018	0,007	0,007	0,003	0,005	0,002	0,023	0,01	0,023	0,017	0,021	0,011	0,014	0,01	0,023	0,014	0,019	0,018	0,026	0,017	0,02	0,015	0,01	0,005	0,012
CTG	0,051	0,032	0,037	0,027	0,015	0,009	0,014	0,009	0,006	0,004	0,006	0,003	0,017	0,009	0,015	0,01	0,016	0,009	0,012	0,01	0,026	0,015	0,022	0,017	0,024	0,018	0,017	0,015	0,007	0,005	0,008
CCT	0,03	0,018	0,03	0,018	0,017	0,014	0,017	0,008	0,008	0,003	0,005	0,002	0,011	0,005	0,029	0,018	0,025	0,011	0,011	0,006	0,026	0,016	0,015	0,015	0,022	0,018	0,019	0,013	0,006	0,004	0,009
CCC	0,051	0,038	0,041	0,026	0,026	0,018	0,026	0,011	0,005	0,002	0,002	0,001	0,018	0,008	0,035	0,018	0,022	0,01	0,018	0,013	0,039	0,024	0,026	0,025	0,03	0,02	0,015	0,012	0,006	0,004	0,008
CCA	0,038	0,034	0,037	0,024	0,025	0,013	0,02	0,008	0,005	0,002	0,003	0,002	0,027	0,011	0,028	0,013	0,023	0,009	0,016	0,01	0,03	0,015	0,02	0,019	0,03	0,017	0,02	0,013	0,005	0,004	0,012
CCG	0,035	0,026	0,032	0,023	0,013	0,007	0,02	0,007	0,005	0,003	0,003	0,003	0,015	0,008	0,018	0,012	0,02	0,008	0,011	0,008	0,028	0,015	0,022	0,018	0,026	0,017	0,022	0,019	0,004	0,003	0,008
CAT	0,032	0,019	0,03	0,019	0,017	0,01	0,016	0,006	0,009	0,004	0,006	0,002	0,017	0,007	0,027	0,015	0,021	0,01	0,011	0,007	0,029	0,019	0,017	0,016	0,028	0,021	0,023	0,013	0,011	0,005	0,012
CAC	0,042	0,035	0,039	0,028	0,022	0,013	0,021	0,008	0,01	0,004	0,003	0,003	0,025	0,008	0,023	0,015	0,02	0,007	0,014	0,011	0,032	0,018	0,018	0,018	0,021	0,019	0,023	0,016	0,01	0,006	0,012
CAA	0,039	0,031	0,038	0,024	0,019	0,01	0,018	0,008	0,006	0,002	0,004	0,002	0,026	0,012	0,019	0,01	0,015	0,007	0,017	0,01	0,029	0,016	0,02	0,022	0,024	0,016	0,019	0,014	0,008	0,004	0,011
CAG	0,043	0,026	0,036	0,021	0,012	0,008	0,016	0,007	0,009	0,004	0,004	0,003	0,017	0,01	0,015	0,009	0,016	0,008	0,011	0,008	0,024	0,014	0,019	0,015	0,035	0,02	0,024	0,021	0,007	0,006	0,009
GTT	0,03	0,021	0,027	0,019	0,021	0,013	0,013	0,006	0,009	0,003	0,002	0,001	0,015	0,005	0,031	0,018	0,022	0,009	0,01	0,007	0,031	0,017	0,014	0,016	0,024	0,023	0,021	0,016	0,01	0,006	0,011
GTC	0,05	0,045	0,039	0,029	0,028	0,021	0,021	0,008	0,006	0,003	0,002	0,001	0,022	0,009	0,021	0,015	0,014	0,006	0,017	0,011	0,043	0,026	0,017	0,022	0,024	0,018	0,013	0,015	0,009	0,005	0,009
GTA	0,042	0,026	0,037	0,023	0,018	0,011	0,019	0,01	0,008	0,003	0,006	0,003	0,022	0,011	0,022	0,017	0,021	0,01	0,013	0,01	0,023	0,012	0,02	0,019	0,031	0,016	0,021	0,014	0,008	0,006	0,011
GTG	0,047	0,033	0,031	0,024	0,016	0,012	0,018	0,008	0,005	0,003	0,003	0,003	0,02	0,009	0,016	0,011	0,014	0,008	0,012	0,01	0,023	0,013	0,018	0,019	0,026	0,014	0,016	0,012	0,007	0,005	0,008
GCT	0,032	0,023	0,027	0,02	0,027	0,014	0,017	0,007	0,007	0,003	0,002	0,002	0,017	0,006	0,029	0,021	0,021	0,009	0,011	0,007	0,032	0,017	0,013	0,017	0,019	0,015	0,016	0,013	0,008	0,003	0,009
GCC	0,05	0,049	0,042	0,028	0,033	0,021	0,022	0,011	0,007	0,002	0,001	0,001	0,02	0,01	0,026	0,015	0,015	0,007	0,017	0,011	0,05	0,028	0,02	0,027	0,025	0,017	0,016	0,011	0,01	0,004	0,008
GCA	0,039	0,028	0,038	0,025	0,018	0,01	0,019	0,008	0,006	0,003	0,003	0,001	0,031	0,013	0,025	0,014	0,019	0,01	0,017	0,012	0,026	0,015	0,016	0,023	0,031	0,017	0,02	0,014	0,009	0,008	0,01
GCG	0,038	0,031	0,028	0,023	0,017	0,008	0,017	0,008	0,005	0,003	0,002	0,002	0,016	0,008	0,022	0,013	0,021	0,009	0,011	0,008	0,023	0,012	0,016	0,019	0,031	0,019	0,023	0,017	0,007	0,006	0,01
GGT	0,039	0,03	0,029	0,023	0,023	0,015	0,017	0,007	0,007	0,002	0,001	6E-04	0,02	0,005	0,025	0,014	0,018	0,007	0,011	0,009	0,04	0,022	0,018	0,021	0,015	0,016	0,018	0,016	0,009	0,005	0,011
GGC	0,039	0,029	0,036	0,023	0,021	0,012	0,019	0,008	0,01	0,004	0,002	0,002	0,019	0,01	0,028	0,018	0,022	0,008	0,021	0,014	0,027	0,015	0,021	0,017	0,023	0,014	0,024	0,019	0,013	0,006	0,015
GGA	0,049	0,032	0,047	0,028	0,02	0,011	0,022	0,012	0,005	0,003	0,003	0,002	0,027	0,012	0,024	0,013	0,017	0,01	0,021	0,013	0,028	0,014	0,022	0,022	0,041	0,019	0,024	0,014	0,01	0,006	0,013
GGG	0,054	0,035	0,031	0,023	0,016	0,01	0,019	0,009	0,007	0,003	0,004	0,002	0,018	0,009	0,025	0,017	0,017	0,009	0,012	0,01	0,021	0,013	0,018	0,018	0,054	0,033	0,018	0,016	0,007	0,005	0,009
GAT	0,034	0,019	0,033	0,019	0,016	0,01	0,015	0,007	0,006	0,002	0,002	1E-03	0,015	0,006	0,024	0,014	0,018	0,008	0,012	0,008	0,036	0,021	0,02	0,019	0,029	0,02	0,021	0,013	0,007	0,004	0,011
GAC	0,04	0,033	0,035	0,027	0,022	0,013	0,017	0,007	0,009	0,003	0,002	0,002	0,019	0,009	0,024	0,013	0,014	0,008	0,015	0,011	0,033	0,02	0,02	0,02	0,024	0,018	0,021	0,018	0,009	0,005	0,011
GAA	0,046	0,036	0,043	0,028	0,021	0,011	0,019	0,009	0,006	0,002	0,003	0,002	0,024	0,01	0,018	0,011	0,014	0,007	0,018	0,011	0,033	0,017	0,02	0,021	0,025	0,013	0,019	0,014	0,007	0,004	0,01
GAG	0,05	0,034	0,037	0,026	0,015	0,01	0,016	0,007	0,006	0,003	0,003	0,002	0,018	0,009	0,015	0,008	0,014	0,008	0,014	0,01	0,027	0,014	0,016	0,016	0,032	0,016	0,02	0,017	0,005	0,004	0,008

Figura B.9: Estimativa da matriz das probabilidades de transição da *Saccharomyces cerevisiae* codões/nucleótidos.

	3' TTA	3' TTG	3' CTT	3' CTC	3' CTA	3' CTG	3' CCT	3' CCC	3' CCA	3' CCG	3' CAT	3' CAC	3' CAA	3' CAG	3' GTT	3' GTC	3' GTA	3' GTG	3' GCT	3' GCC	3' GCA	3' GCG	3' GGT	3' GGC	3' GGA	3' GGG	3' GAT	3' GAC	3' GAA	3' GAG	TOTAL
AAA	0,027	0,024	0,014	0,005	0,019	0,013	0,013	0,006	0,016	0,006	0,011	0,006	0,023	0,011	0,023	0,011	0,012	0,012	0,017	0,009	0,014	0,005	0,017	0,007	0,009	0,005	0,035	0,018	0,047	0,022	1
AAG	0,02	0,022	0,012	0,005	0,015	0,012	0,012	0,006	0,018	0,005	0,015	0,009	0,029	0,013	0,02	0,011	0,01	0,008	0,024	0,013	0,017	0,007	0,02	0,011	0,01	0,005	0,038	0,023	0,055	0,024	1
AAT	0,025	0,025	0,01	0,005	0,011	0,008	0,014	0,005	0,016	0,006	0,013	0,006	0,021	0,008	0,024	0,012	0,012	0,01	0,021	0,014	0,018	0,008	0,024	0,014	0,015	0,01	0,045	0,026	0,054	0,023	1
AAC	0,021	0,025	0,012	0,006	0,011	0,009	0,013	0,006	0,016	0,004	0,014	0,008	0,026	0,012	0,022	0,013	0,01	0,009	0,016	0,009	0,012	0,004	0,021	0,01	0,01	0,006	0,038	0,02	0,041	0,017	1
ACT	0,036	0,037	0,01	0,005	0,012	0,008	0,014	0,007	0,024	0,006	0,012	0,006	0,024	0,009	0,026	0,013	0,014	0,013	0,027	0,016	0,022	0,007	0,029	0,009	0,011	0,007	0,036	0,018	0,044	0,016	1
ACC	0,023	0,021	0,012	0,004	0,009	0,008	0,011	0,006	0,018	0,004	0,01	0,007	0,022	0,008	0,022	0,013	0,012	0,011	0,02	0,012	0,015	0,005	0,025	0,007	0,008	0,004	0,029	0,015	0,03	0,014	1
ACA	0,025	0,02	0,011	0,004	0,012	0,009	0,019	0,008	0,022	0,006	0,015	0,007	0,02	0,011	0,018	0,009	0,016	0,012	0,015	0,01	0,019	0,007	0,02	0,01	0,013	0,007	0,036	0,017	0,042	0,018	1
ACG	0,024	0,025	0,015	0,006	0,017	0,015	0,021	0,011	0,02	0,01	0,017	0,009	0,026	0,014	0,018	0,011	0,018	0,015	0,016	0,01	0,02	0,009	0,017	0,011	0,013	0,008	0,036	0,019	0,039	0,019	1
CGT	0,029	0,03	0,013	0,007	0,013	0,008	0,01	0,006	0,012	0,004	0,018	0,009	0,028	0,012	0,027	0,014	0,015	0,012	0,02	0,014	0,016	0,006	0,037	0,015	0,013	0,008	0,039	0,021	0,046	0,017	1
CGC	0,025	0,018	0,013	0,008	0,012	0,012	0,011	0,008	0,015	0,005	0,022	0,012	0,03	0,015	0,019	0,006	0,008	0,008	0,013	0,007	0,014	0,003	0,015	0,011	0,011	0,007	0,035	0,015	0,029	0,021	1
CGA	0,03	0,023	0,014	0,005	0,021	0,01	0,012	0,005	0,014	0,002	0,019	0,007	0,024	0,016	0,017	0,007	0,013	0,01	0,013	0,009	0,018	0,005	0,012	0,01	0,014	0,005	0,036	0,022	0,041	0,02	1
CGG	0,03	0,032	0,02	0,01	0,02	0,019	0,014	0,012	0,014	0,006	0,013	0,009	0,025	0,021	0,013	0,007	0,012	0,009	0,012	0,01	0,01	0,005	0,009	0,007	0,01	0,004	0,028	0,017	0,038	0,018	1
AGA	0,027	0,024	0,009	0,004	0,013	0,01	0,011	0,005	0,015	0,004	0,013	0,008	0,025	0,011	0,019	0,009	0,012	0,009	0,019	0,012	0,015	0,005	0,025	0,01	0,012	0,005	0,037	0,018	0,045	0,019	1
AGG	0,027	0,023	0,015	0,007	0,019	0,013	0,014	0,008	0,02	0,007	0,015	0,007	0,03	0,015	0,015	0,009	0,012	0,011	0,021	0,012	0,017	0,007	0,014	0,008	0,009	0,003	0,032	0,019	0,042	0,019	1
TCT	0,031	0,032	0,012	0,005	0,013	0,009	0,015	0,009	0,024	0,006	0,015	0,008	0,033	0,012	0,023	0,012	0,012	0,013	0,025	0,014	0,019	0,008	0,025	0,009	0,011	0,006	0,035	0,016	0,037	0,014	1
TCC	0,025	0,024	0,011	0,005	0,012	0,009	0,01	0,006	0,012	0,003	0,011	0,008	0,026	0,01	0,02	0,01	0,01	0,009	0,016	0,009	0,012	0,004	0,019	0,007	0,008	0,004	0,024	0,012	0,024	0,012	1
TCA	0,031	0,027	0,013	0,005	0,015	0,012	0,018	0,008	0,021	0,006	0,014	0,007	0,027	0,015	0,015	0,006	0,013	0,01	0,016	0,007	0,016	0,006	0,017	0,009	0,013	0,006	0,032	0,015	0,036	0,015	1
TCG	0,03	0,029	0,014	0,006	0,019	0,016	0,016	0,01	0,021	0,008	0,019	0,012	0,032	0,017	0,016	0,006	0,013	0,01	0,011	0,007	0,014	0,006	0,011	0,007	0,008	0,004	0,029	0,013	0,029	0,015	1
AGT	0,027	0,027	0,01	0,004	0,008	0,007	0,009	0,005	0,01	0,004	0,011	0,006	0,02	0,006	0,022	0,012	0,014	0,01	0,024	0,015	0,021	0,009	0,027	0,012	0,016	0,009	0,048	0,025	0,051	0,021	1
AGC	0,019	0,018	0,01	0,004	0,008	0,007	0,01	0,006	0,011	0,003	0,013	0,007	0,024	0,01	0,017	0,009	0,012	0,01	0,019	0,014	0,014	0,004	0,021	0,012	0,013	0,007	0,039	0,02	0,038	0,018	1
ATT	0,031	0,028	0,016	0,006	0,013	0,009	0,017	0,008	0,022	0,008	0,016	0,009	0,029	0,012	0,028	0,015	0,014	0,013	0,026	0,017	0,02	0,008	0,028	0,01	0,012	0,007	0,044	0,023	0,051	0,019	1
ATC	0,021	0,022	0,012	0,007	0,012	0,009	0,011	0,006	0,019	0,004	0,014	0,007	0,023	0,011	0,017	0,011	0,009	0,008	0,014	0,009	0,011	0,004	0,02	0,009	0,009	0,006	0,036	0,02	0,037	0,018	1
ATA	0,028	0,024	0,013	0,005	0,016	0,013	0,014	0,008	0,019	0,01	0,013	0,008	0,02	0,011	0,015	0,008	0,012	0,011	0,014	0,009	0,015	0,007	0,013	0,008	0,009	0,005	0,035	0,019	0,037	0,018	1
ATG	0,019	0,018	0,017	0,006	0,016	0,014	0,011	0,007	0,014	0,004	0,013	0,006	0,028	0,01	0,023	0,012	0,013	0,011	0,021	0,014	0,02	0,008	0,023	0,012	0,013	0,008	0,04	0,024	0,048	0,019	1
TTT	0,031	0,032	0,013	0,006	0,011	0,009	0,013	0,008	0,014	0,005	0,013	0,007	0,027	0,011	0,029	0,014	0,013	0,011	0,024	0,018	0,019	0,008	0,031	0,014	0,015	0,007	0,045	0,024	0,053	0,02	1
TTC	0,025	0,028	0,012	0,007	0,013	0,009	0,014	0,011	0,019	0,004	0,018	0,011	0,036	0,017	0,016	0,009	0,007	0,006	0,011	0,008	0,009	0,003	0,021	0,007	0,007	0,004	0,03	0,016	0,032	0,013	1
TAT	0,029	0,033	0,013	0,006	0,012	0,01	0,015	0,006	0,02	0,007	0,015	0,008	0,027	0,01	0,023	0,013	0,009	0,013	0,021	0,013	0,017	0,007	0,028	0,011	0,014	0,008	0,042	0,024	0,044	0,019	1
TAC	0,022	0,029	0,011	0,006	0,013	0,012	0,009	0,005	0,017	0,003	0,016	0,011	0,031	0,014	0,021	0,013	0,01	0,01	0,016	0,01	0,01	0,004	0,025	0,009	0,01	0,006	0,038	0,022	0,039	0,017	1
TGT	0,029	0,032	0,015	0,008	0,014	0,011	0,016	0,01	0,015	0,005	0,015	0,01	0,025	0,009	0,027	0,015	0,012	0,011	0,024	0,013	0,016	0,007	0,035	0,015	0,014	0,009	0,037	0,02	0,036	0,018	1
TGC	0,027	0,024	0,016	0,009	0,013	0,012	0,012	0,008	0,015	0,004	0,018	0,011	0,033	0,014	0,019	0,009	0,013	0,009	0,013	0,009	0,011	0,005	0,023	0,01	0,01	0,008	0,028	0,015	0,026	0,012	1
TGG	0,026	0,023	0,012	0,006	0,015	0,013	0,008	0,006	0,012	0,003	0,013	0,009	0,026	0,01	0,024	0,012	0,014	0,009	0,018	0,011	0,015	0,006	0,022	0,01	0,009	0,008	0,038	0,024	0,041	0,02	1
TTA	0,028	0,026	0,014	0,006	0,017	0,014	0,016	0,007	0,022	0,008	0,014	0,008	0,037	0,017	0,017	0,01	0,01	0,011	0,016	0,011	0,015	0,007	0,02	0,008	0,011	0,006	0,035	0,019	0,045	0,019	1
TTG	0,019	0,02	0,01	0,004	0,012	0,009	0,012	0,009	0,018	0,005	0,014	0,008	0,03	0,013	0,022	0,012	0,012	0,012	0,023	0,015	0,018	0,007	0,023	0,009	0,01	0,005	0,037	0,02	0,047	0,02	1
CTT	0,042	0,041	0,018	0,009	0,016	0,012	0,017	0,009	0,023	0,008	0,017	0,009	0,028	0,011	0,022	0,011	0,011	0,012	0,021	0,013	0,017	0,008	0,023	0,011	0,014	0,007	0,038	0,019	0,037	0,013	1
CTC	0,024	0,025	0,013	0,007	0,012	0,01	0,012	0,008	0,016	0,004	0,013	0,008	0,018	0,011	0,018	0,01	0,011	0,009	0,014	0,01	0,011	0,004	0,024	0,008	0,008	0,005	0,039	0,018	0,031	0,017	1
CTA	0,031	0,027	0,013	0,007	0,018	0,015	0,016	0,009	0,022	0,009	0,015	0,008	0,032	0,016	0,016	0,009	0,012	0,01	0,019	0,01	0,017	0,006	0,018	0,008	0,011	0,005	0,033	0,018	0,045	0,019	1
CTG	0,022	0,024	0,013	0,007	0,017	0,014	0,013	0,011	0,016	0,007	0,015	0,009	0,031	0,017	0,018	0,012	0,016	0,013	0,017	0,015	0,02	0,009	0,019	0,011	0,013	0,006	0,038	0,023	0,048	0,022	1
CCT	0,028	0,027	0,013	0,005	0,011	0,011	0,019	0,009	0,028	0,008	0,018	0,01	0,037	0,014	0,026	0,012	0,017	0,014	0,027	0,013	0,02	0,007	0,022	0,01	0,011	0,007	0,04	0,019	0,049	0,021	1
CCC	0,029	0,024	0,011	0,007	0,011	0,008	0,011	0,005	0,016	0,003	0,015	0,006	0,021	0,01	0,022	0,011	0,018	0,013	0,017	0,008	0,012	0,003	0,016	0,007	0,008	0,003	0,026	0,014	0,033	0,014	1
CCA	0,028	0,027	0,01	0,005	0,012	0,01	0,018	0,006	0,028	0,008	0,013	0,008	0,026	0,016	0,019	0,009	0,013	0,01	0,017	0,008	0,016	0,005	0,021	0,009	0,009	0,004	0,031	0,015	0,047	0,021	1
CCG	0,036	0,026	0,016	0,006	0,025	0,018	0,018	0,01	0,025	0,007	0,018	0,01	0,034	0,026	0,019	0,01	0,018	0,012	0,016	0,008	0,017	0,006	0,01	0,012	0,013	0,006	0,034	0,019	0,045	0,023	1
CAT	0,035	0,033	0,013	0,006	0,013	0,009	0,02	0,007	0,022	0,009	0,022	0,012	0,025	0,011	0,022	0,011	0,011	0,01	0,019	0,013	0,015	0,007	0,025	0,01	0,013	0,008	0,037	0,02	0,043	0,017	1
CAC	0,021	0,027	0,013	0,007	0,009	0,009	0,011	0,007	0,02	0,005	0,02	0,014	0,028	0,013	0,023	0,011	0,01	0,009	0,014	0,009	0,012	0,004	0,023	0,009	0,009	0,006	0,037	0,02	0,039	0,016	1
CAA	0,029	0,029	0,011	0,005	0,015	0,013	0,013	0,005	0,017	0,006	0,013	0,007	0,041	0,024	0,019	0,01	0,012	0,012	0,019	0,011	0,014	0,006	0,021	0,009	0,009	0,005	0,036	0,017	0,052	0,022	1
CAG	0,028	0,029	0,014	0,007	0,017	0,014	0,013	0,007	0,016	0,007	0,019	0,01	0,052	0,038	0,016	0,009	0,012	0,011	0,016	0,011	0,015	0,008	0,014	0,01	0,012	0,005	0,034	0,02	0,043	0,021	1
GTT	0,03	0,032	0,017	0,007	0,011	0,009	0,019	0,008	0,023	0,006	0,015	0,008	0,026	0,009	0,03	0,016	0,011	0,011	0,026	0,016	0,018	0,007	0,028	0,01	0,015	0,006	0,039	0,021	0,041	0,016	1
GTC	0,021	0,027	0,016	0,006	0,01	0,009	0,012	0,007	0,018	0,003	0,014	0,009	0,022	0,009	0,023	0,011	0,01	0,008	0,019	0,011	0,016	0,004	0,024	0,008	0,01	0,005	0,03	0,02	0,034	0,013	1
GTA	0,022	0,024	0,014	0,006	0,013	0,012	0,017	0,009	0,023	0,009	0,014	0,006	0,023	0,013	0,019	0,009	0,013	0,013	0,016	0,01	0,02	0,008	0,016	0,009	0,012	0,007	0,042	0,02	0,043	0,023	1
GTG	0,018	0,022	0,014	0,007	0,015	0,013	0,014	0,009	0,018	0,007	0,012	0,007	0,025	0,013	0,022	0,012	0,02	0,017	0,02	0,016	0,023	0,008	0,018	0,01	0,013	0,009	0,043	0,024	0,059	0,028	1
GCT	0,028	0,033	0,014	0,005	0,012	0,007	0,014	0,008	0,02	0,004	0,013	0,008	0,03	0,01	0,028	0,015	0,013	0,013	0,04	0,02	0,022	0,008	0,033	0,01	0,011	0,005	0,033	0,017	0,044	0,016	1
GCC	0,021	0,023	0,014	0,006	0,012	0,008	0,009	0,006	0,016	0,003	0,012	0,006	0,024	0,009	0,023	0,011	0,01	0,009	0,028	0,016	0,015	0,004	0,025	0,007	0,007	0,003	0,022	0,011	0,026	0,009	1
GCA	0,027	0,026	0,01	0,006	0,014	0,01	0,016	0,007	0,019	0,004	0,015	0,007	0,026	0,012	0,017	0,008	0,014	0,01	0,021	0,011	0,021	0,006	0,022	0,01	0,013	0,005	0,036	0,019	0,046	0,018	1
GCG	0,029	0,029	0,017	0,008	0,022	0,017	0,015	0,01	0,021	0,005	0,018	0,009	0,031	0,017	0,018	0,009	0,018	0,012	0,023	0,014	0,023	0,008	0,016	0,012	0,013	0,005	0,033	0,02	0,039	0,019	1
GGT	0,024	0,025	0,011	0,005	0,012	0,006	0,009	0,004	0,012	0,002	0,011	0,007	0,021	0,005	0,034	0,019	0,014	0,013	0,028	0,017	0,016	0,006	0,045	0,013	0,013	0,008	0,04	0,024	0,046	0,016	1
GGC	0,022	0,02	0,013	0,007	0,012	0,01	0,012	0,008	0,016	0,003	0,02	0,012	0,031	0,016	0,02	0,01	0,011	0,009	0,022	0,013	0,014	0,004	0,025	0,013	0,013	0,006	0,033	0,018	0,037	0,015	1
GGA	0,027	0,023	0,012	0,005	0,013	0,011	0,011	0,006	0,011	0,004	0,012	0,01	0,024	0,011	0,015	0,007	0,01	0,008	0,014	0,008	0,015	0,005	0,018	0,01	0,015	0,007	0,031	0,016	0,038	0,018	1
GGG	0,031	0,035	0,02	0,01	0,021	0,018	0,017	0,011	0,018	0,008	0,015	0,009	0,024	0,012	0,014	0,006	0,011	0,009	0,014	0,01	0,015	0,005	0,012	0,01	0,01	0,005	0,024	0,015	0,033	0,017	1
GAT	0,03	0,032	0,012	0,006	0,01	0,008	0,014	0,006	0,015	0,006	0,011	0,006	0,018	0,008	0,025	0,014	0,012	0,012	0,023	0,013	0,016	0,007	0,023	0,01	0,012	0,007	0,055	0,029	0,064	0,028	1
GAC	0,023	0,028	0,013	0,006	0,011	0,01	0,015	0,007	0,017	0,005	0,015	0,008	0,025	0,012	0,022	0,01	0,011	0,01	0,017	0,01	0,012	0,004	0,02	0,009	0,008	0,005	0,045	0,025	0,054	0,022	1
GAA	0,024	0,025	0,011	0,005	0,016	0,012	0,01	0,004	0,011	0,004	0,01	0,006	0,024	0,011	0,022	0,01	0,012	0,011	0,02	0,01	0,014	0,006	0,02	0,008	0,009	0,005	0,044	0,023	0,063	0,027	1
GAG	0,025	0,025	0,015	0,008	0,017	0,015	0,013	0,007	0,016	0,004	0,016	0,008	0,032	0,016	0,018	0,009	0,011	0,01	0,019	0,012	0,021	0,007	0,014	0,009	0,011	0,005	0,042	0,024	0,061	0,031	1

Figura B.10: Estimativa da matriz das probabilidades de transição da *Saccharomyces cerevisiae* codões/nucleótidos (continuação).

Apêndice C

Apêndice C.1

Gene C1 da espécie *Candida albicans*
ATG GGTA AAGAAAAAAC TCACGTTAAC GTTGTTGTTA TTGGTCACGT
CGATTCCGGT AAATCTACTA CCACCGGTCA CTTAATTTAC AAGTGTGGTG GTATCGATAA
AAGAACCATT GAAAAATTCG AAAAAGAAGC TGCTGAATTG GGTAAAGGTT CTTTCAAATA
CGCTTGGGTC TTGGACAAAT TGAAGGCTGA AAGAGAAAGA GGTATCACCA TTGATATTGC
TTTGTGGAAA TTCGAAACTC CAAAATACCA CGTTACCGTC ATTGATGCTC CAGGTCACAG
AGATTTCATC AAGAATATGA TCACTGGTAC TTCTCAAGCT GATTGTGCTA TTTTGATTAT
TGCTGGTGGT ACTGGTGAAT TCGAAGCCGG TATTTCTAAG GATGGTCAAA CCAGAGAACA
CGCTTTGTTG GCTTACACTT TGGGTGTCAA ACAATTGATT GTTGCTGTCA ACAAGATGGA
CTCTGTCAAA TGGGACAAAA ACAGATTTGA AGAAATCATC AAGGAAACCT CCAACTTCGT
CAAGAAGGTT GGTTACAACC CAAAGACTGT TCCATTCGTT CCAATCTCTG GTTGGAATGG
TGACAACATG ATTGAACCAT CCACCAACTG TCCATGGTAC AAGGGTTGGG AAAAGGAAAC
CAAATCCGGT AAAGTTACTG GTAAGACCTT GTTAGAAGCT ATTGACGCTA TTGAACCACC
AACCAGACCA ACCGACAAAC CATTGAGATT GCCATTGCAA GATGTTTACA AGATTGGTGG
TATTGGTACT GTGCCAGTCG GTAGAGTTGA AACTGGTATC ATCAAAGCCG GTATGGTTGT
TACTTTCGCC CCAGCTGGTG TTACCACTGA AGTCAAGTCC GTTGAAATGC ATCACGAACA
ATTGGCTGAA GGTGTTCCAG GTGACAATGT TGGTTTCAAC GTTAAGAACG TTTCCGTTAA
AGAAATTAGA AGAGGTAACG TTTGTGGTGA CTCCAAGAAC GATCCACCAA AGGGTTGTGA
CTCTTTCAAT GCCCAAGTCA TTGTTTTGAA CCATCCAGGT CAAATCTCTG CTGGTTACTC
TCCAGTCTTG GATTGTCACA CTGCCCACAT TGCTTGTAAA TTCGACACTT TGGTTGAAAA
GATTGACAGA AGAACTGGTA AGAAATTGGA AGAAAATCCA AAATTCGTCA AATCCGGTGA
TGCTGCTATC GTCAAGATGG TCCCAACCAA ACCAATGTGT GTTGAAGCTT TCACTGACTA
CCCACCATTA GGTAGATTCG CTGTCAGAGA TATGAGACAA ACCGTTGCTG TTGGTGTCAT
CAAATCTGTT GAAAAATCCG ACAAAGCTGG TAAAGTTACC AAGGCTGCTC AAAAAGCTGC
TAAGAAA **TAA**

Tabela C.1: Sequência de codões do gene $C1$ da espécie *Candida albicans*.

Apêndice C.2

Gene S1 da espécie *Saccharomyces cerevisiae*
ATG TCGCCCTCT GCCGTACAATCA TCAAAACTAGAA GAACAGTCAAGT GAAATTGACAAG
TTGAAAGCAAAA ATGTCCCAGTCT GCCGCCACTGCG CAGCAGAAGAAG GAACATGAGTAT
GAACATTTGACT TCGGTCAAGATC GTGCCACAACGG CCCATCTCAGAT AGACTGCAGCCC
GCAATTGCTACC CACTATTCTCCA CACTTGGACGGG TTGCAGGACTAT CAGCGCTTGCAC
AAGGAGTCTATT GAAGACCCTGCT AAGTTCTTCGGT TCTAAAGCTACC CAATTTTTAAAC
TGGTCTAAGCCA TTCGATAAGGTG TTCATCCCAGAC CCTAAAACGGGC AGGCCCTCCTTC
CAGAACAATGCA TGGTTCCTCAAC GGCCAATTAAAC GCCTGTTACAAC TGTGTTGACAGA
CATGCCTTGAAG ACTCCTAACAAG AAAGCCATTATT TTCGAAGGTGAC GAGCCTGGCCAA
GGCTATTCCATT ACCTACAAGGAA CTACTTGAAGAA GTTTGTCAAGTG GCACAAGTGCTG
ACTTACTCTATG GGCGTTCGCAAG GGCGATACTGTT GCCGTGTACATG CCTATGGTCCCA
GAAGCAATCATA ACCTTGTTGGCC ATTTCCCGTATC GGTGCCATTCAC TCCGTAGTCTTT
GCCGGGTTTTCT TCCAACTCCTTG AGAGATCGTATC AACGATGGGGAC TCTAAAGTTGTC
ATCACTACAGAT GAATCCAACAGA GGTGGTAAAGTC ATTGAGACTAAA AGAATTGTTGAT
GACGCGCTAAGA GAGACCCCAGGC GTGAGACACGTC TTGGTTTATAGA AAGACCAACAAT
CCATCTGTTGCT TTCCATGCCCCC AGAGATTTGGAT TGGGCAACAGAA AAGAAGAAATAC
AAGACCTACTAT CCATGCACACCC GTTGATTCTGAG GATCCATTATTC TTGTTGTATACG
TCTGGTTCTACT GGTGCCCCCAAG GGTGTTCAACAT TCTACCGCAGGT TACTTGCTGGGA
GCTTTGTTGACC ATGCGCTACACT TTTGACACTCAC CAAGAAGACGTT TTCTTCACAGCT
GGAGACATTGGC TGGATTACAGGC CACACTTATGTG GTTTATGGTCCC TTACTATATGGT
TGTGCCACTTTG GTCTTTGAAGGG ACTCCTGCGTAC CCAAATTACTCC CGTTATTGGGAT
ATTATTGATGAA CACAAAGTCACC CAATTTTATGTT GCGCCAACTGCT TTGCGTTTGTTG
AAAAGAGCTGGT GATTCCTACATC GAAAATCATTCC TTAAAATCTTTG CGTTGCTTGGGT
TCGGTCGGTGAG CCAATTGCTGCT GAAGTTTGGGAG TGGTACTCTGAA AAAATAGGTAAA
AATGAAATCCCC ATTGTAGACACC TACTGGCAAACA GAATCTGGTTCG CATCTGGTCACC
CCGCTGGCTGGT GGTGTTACACCA ATGAAACCGGGT TCTGCCTCATTC CCCTTCTTCGGT
ATTGATGCAGTT GTTCTTGACCCT AACACTGGTGAA GAACTTAACACC AGCCACGCAGAG
GGTGTCCTTGCC GTCAAAGCTGCA TGGCCATCATTT GCAAGAACTATT TGGAAAAATCAT
GATAGGTATCTA GACACTTATTTG AACCCTTACCCT GGCTACTATTTC ACTGGTGATGGT
GCTGCAAAGGAT AAGGATGGTTAT ATCTGGATTTTG GGTCGTGTAGAC GATGTGGTGAAC
GTCTCTGGTCAC CGTCTGTCTACC GCTGAAATTGAG GCTGCTATTATC GAAGATCCAATT
GTGGCCGAGTGT GCTGTTGTCGGA TTCAACGATGAC TTGACTGGTCAA GCAGTTGCTGCA
TTTGTGGTGTTG AAAAACAAATCT AGTTGGTCCACC GCAACAGATGAT GAATTACAAGAT
ATCAAGAAGCAT TTGGTCTTTACT GTTAGAAAAGAC ATCGGGCCATTT GCCGCACCAAAA
TTGATCATTTTA GTGGATGACTTG CCCAAGACAAGA TCCGGCAAAATT ATGAGACGTATT
TTAAGAAAAATC CTAGCAGGAGAA AGTGACCAACTA GGCGACGTTTCT ACATTGTCAAAC
CCTGGCATTGTT AGACATCTAATT GATTCGGTCAAG TTG**TAA**

Tabela C.2: Sequência de codões do gene $S1$ da espécie *Saccharomyces cerevisea*.

Bibliografia

[1] Avery, Peter J. and Henderson, Daniel A. (1999). Fitting Markov chain models to discrete state series such as DNA sequences. Newcastle University, UK. *Applied Statistic*, 48, pp. 53-61.

[2] Bishop, Yvonne M. M. (1998). *Discrete Multivariate Analysis. Theory and Practice.* MIT Press. Cambridge, England.

[3] Cohen, A., Mantegna R. N. and S. Havlin (1996). *Can Zipf Analyses and Entropy Distinguish Between Artificial and Natural Language Texts?.*

[4] Csiszár, Imre, Shields, Paul C. (1999). Consistency of th BIC order estimator. *Electronic Research Announcements of the American Mathematical Society.* Volume5, Pages 123-127.

[5] Czirók, Andras (1995). Correlations in binary sequences and a generalized Zipf analysis. *Physical Review E*, Volume 52, Number 1, Pages 446-452.

[6] Everitt, Brian S. (1993). *Cluster Analysis.* Trird edition, Arnold.

[7] Everitt, Brian S. (1997). *The Analysis of Contingency Tables* . John Wiley and Sons Inc., New York.

[8] Gasperin, Caroline Varaschin and Lima, Vera Lúcia Strube (2001). Fundamentos do Processamento Estatístico da Linguagem Natural. *Relatório Técnico.* Porto Alegre: PPGCC-PUCRS.

[9] Kotz, Samuel and Johnson, Norman L. (1982). *Encyclopedia of Statistical Sciences.* Campbell B. Read Associate Edition. Wiley Interscience.

[10] Manly, Bryan F. J. (1994). *Multivariate Statistical Methods. A Primer.* Chapman and Hall.

[11] Mantegna, R. N. (1995). Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. *Physical Review E*, Volume 52, Number 3.

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

[13] Murteira, Bento José Ferreira (1990). *Probabilidades e Estatística.* Volume 1, segunda edição revista. McGraw-Hill.

[14] Murteira, Bento José Ferreira (1990). *Probabilidades e Estatística.* Volume 2, segunda edição revista. McGraw-Hill.

[15] Murteira, Bento José Ferreira (1993). *Análise Exploratória de Dados. Estatística Descritiva.* McGraw-Hill.

[16] Murteira, Bento J. F., Müler, Daniel A., Turkman K. Feridun (1993). *Análise de Sucessões Cronológicas.* McGraw-Hill.

[17] Jain, Anil K. and Dubes, Richard C. (1988). *Algorithms for Clustering Data.* Prentice Hall.

[18] Papoulis, Athanasios (1991). *Probability Random Variables and Stochastic Processes.* McGraw-Hill.

[19] Parzen, Emanuel (1999). *Stochastic Processes.* Classics in Applies Mathematics.

[20] Poosala, Viswanath (1995). Zipf's Law. *Technical report.* University of Wisconsin, Madison.

[21] Reis, Elizabeth (2001). *Estatística Multivariada Aplicada.* Edições Sílabo, segunda edição.

[22] Resnick, Sidney I. (1992). *Adventures in Stochastic Processes.* Biskhäuser.

[23] Santner, Thomas J. (1989). *The Statistical Analysis of Discrete Data.* Springer-Verlag.

[24] Shannon, Claude E. (1949). Communication Theory of Secrecy Systems. *Bell System Technical Journal*, Vol. 28 , pp. 656-715.

[25] Schumacker, Randall E. and Ph. D. (2002). A comparison of OLS to LTS and MM Robust Regression in S-PLUs. *Southwest Educational Research Association.* February, 14-16, Austin, Texas.

[26] Weir, Bruce S. (1996). *Genetic Data Analysis II.* Sunderland, MA.