Universidade de Aveiro
2008
Departamento de Electrónica, Telecomunicações e Informática

Vera Mónica Almeida Afreixo

Sinais simbólicos e aplicações em genómica

Universidade de Aveiro
2008
Departamento de Electrónica, Telecomunicações e Informática

Vera Mónica Almeida Afreixo

Sinais simbólicos e aplicações em genómica

tese apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Engenharia Electrotécnica, realizada sob a orientação científica do Prof. Doutor Paulo Jorge dos Santos Gonçalves Ferreira, Professor Catedrático do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro, e do Prof. Doutor Armando José Formoso de Pinho, Professor Associado do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro

Apoio financeiro da FCT e do FSE no âmbito do III Quadro Comunitário de Apoio.

FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR
UNIÃO EUROPEIA
Fundo Social Europeu

aos meus filhos
ao Arnaldo

o júri

presidente

Doutor José Joaquim de Almeida Grácio
Professor catedrático da Universidade de Aveiro

Doutor Paulo Jorge dos Santos Gonçalves Ferreira
Professor catedrático da Universidade de Aveiro

Doutor Armando José Formoso de Pinho
Professor associado da Universidade de Aveiro

Doutor Luís António Serralva Vieira Sá
Professor catedrático da Universidade de Coimbra

Doutor Jorge Salvador Marques
Professor associado da Universidade Técnica de Lisboa

Doutor Aurélio Joaquim de Castro Campilho
professor catedrático da Universidade do Porto

agradecimentos

Em primeiro lugar, gostaria de agradecer aos meus orientadores, o Prof. Doutor Paulo Jorge dos Santos Gonçalves Ferreira e o Prof. Doutor Armando José Formoso de Pinho, pelo facto de me terem ingressado nas áreas de processamento de sinal, proporcionando os meios e contribuído para o aprofundamento dos meus conhecimentos e orientado este trabalho, demonstrando sempre a sua disponibilidade e acompanhamento.

Reconheço o apoio de vários investigadores do grupo de Bioinformática da Universidade de Aveiro. Em particular gostaria de agradecer ao Doutor Manuel Santos, à Doutora Adelaide Freitas, ao Doutor José Luís Oliveira, à Doutora Gabriela Moura, ao Mestre Miguel Pinheiro e à Doutora Laura Carreto por todas as discussões e apoio durante a realização deste trabalho. Agradeço também ao Doutor António Neves e à Doutora Dorabella Santos pelo saudável trabalho de equipa que me proporcionaram.

Do ponto de vista financeiro, gostaria de agradecer à Fundação para a Ciência e a Tecnologia, que suportou este doutoramento através da bolsa de doutoramento SFRH / BD / 17263 / 2004.

Por último, mas de uma forma muito especial, agradeço à minha família todo o apoio que me deu, em particular, ao meu marido Arnaldo, aos meus pais António e Maria Clara e especialmente aos meus filhos que me deram grande parte da força para concluir este doutoramento e a serenidade para transpor os momentos mais dificeis. A eles dedico este trabalho.

palavras-chave

Processamento de sinal; sinais simbólicos; DNA; análise de Fourier; correlação; modelação; localização de genes; compressão de DNA.

resumo

Esta dissertação surge no contexto do processamento de sinais simbólicos com o objectivo específico de contribuir para o conhecimento da estrutura das sequências de DNA.

A localização automática de genes foi um dos problemas biológicos que motivou o desenvolvimento deste trabalho. A compressão de sequências genéticas, quer para reduzir o espaço de armazenamento quer para obtenção de modelos das mesmas, foi outra das motivações.

Com o objectivo de contribuir para melhorar uma das técnicas frequentemente usadas na localização automática de genes são comparadas metodologias de análise espectral para sequências simbólicas. Também se discute a validade de aplicação de metodologias de análise espectral às sequências simbólicas e apresenta-se um novo método baseada na função de autocorrelação simbólica.

Uma característica que usualmente é tomada para identificação de genes é o tamanho da risca espectral que reflecte a periodicidade de período três. Apresenta-se um algoritmo rápido baseado em contadores de símbolos para cálculo de várias riscas espectrais, e em particular da risca de período três. São também enunciadas e analisadas propriedades associadas ao tamanho de algumas riscas e à redundância espectral.

Por último, desenvolve-se uma técnica para compressão de sequências genéticas baseada num modelo de três estados. Em regiões codificantes do DNA esta técnica leva em geral a melhores resultados do que as actuais técnicas de compressão.

keywords

Signal processing; symbolic data; DNA; Fourier analysis; correlation; modeling; gene location; DNA compression.

abstract

This dissertation addresses the problem of processing sequences of symbols, and has the specific aim of contributing to the analysis and modeling of DNA sequences.

This work was partly motivated by the problem of automatic gene location. Another motivation was the compression of genetic sequences, both for the purpose of reducing the required storage and for determining good DNA models.

The main methodologies of spectral analysis of symbolic sequences are compared. The application of spectral analysis methods to the symbolic sequences is discussed and a new method based on the symbolic autocorrelation function is presented.

One feature that is often used in gene identification is the size of the Fourier coefficient that reflects periodicity of period three. A fast algorithm for the calculation of Fourier coefficients, based on symbol counters, was developed. Some properties associated with the size of some spectral coefficients and spectral redundancy are discussed.

Finally, a technique based on a model with three states was developed to compress genetic sequences. In protein-coding regions this technique leads in general to better results than the state-of-the-art DNA compression techniques.

"We have found the secret of life"
Francis Crick

Conteúdo

Capítulo 1

Introdução

A Genética tem sido uma das áreas de investigação que muito se tem desenvolvido nas últimas décadas, em particular após as descobertas da estrutura do ácido desoxirribonucleico (DNA — *DeoxyriboNucleic Acid*) e da forma como são codificadas as proteínas. A estrutura do DNA, tal como actualmente é aceite, foi proposta em 1953 no famoso artigo de James Watson e Francis Crick [156], publicado na revista *Nature*. A função do DNA foi descrita também por Francis Crick por volta de 1957–58 [47]. Estas descobertas provocaram um grande desenvolvimento da biologia molecular. Actualmente, esta ciência dispõe de um enorme conjunto de dados, sendo exemplo disso a informação proveniente da sequenciação do DNA de várias espécies.

Recentemente, nos principais centros de investigação na área das biociências, grupos interdisciplinares têm estado a realizar trabalhos de investigação com o objectivo de extrair informação relevante contida no DNA. De modo geral, qualquer relação entre a estrutura do DNA e as suas funções biológicas tem sido objecto de estudo. A título de exemplo pode referir-se a caracterização funcional de genes em vários organismos modelo e o estudo de leis que governam a tradução pelo ribossoma.

Neste capítulo apresenta-se uma contextualização biológica deste trabalho. Depois discute-se a sua motivação e objectivos gerais. Seguidamente enumeram-se as publicações elaboradas no âmbito deste trabalho e resumem-se as restantes publicações recentemente feitas no contexto da biologia molecular. Por fim, apresenta-se, de modo resumido, a organização desta dissertação.

1.1 Conceitos biológicos

De seguida será feita uma breve descrição de alguns conceitos de biologia molecular, mais concretamente, sobre o código genético e as proteínas. Pretende-se com esta secção efectuar uma contextualização biológica da dissertação, introduzindo os conceitos fundamentais para a sua compreensão e discussão.

1.1.1 O código genético

Uma das extraordinárias revelações dos anos 50 do século passado, foi a demonstração de que a "infinita" complexidade das estruturas de todos os seres vivos codificada no genoma é devida à combinação de apenas quatro moléculas que constituem o DNA (ver, por exemplo, [6]).

Na composição do DNA entram quatro moléculas chamadas *nucleótidos* ou de forma simplista *bases*. Os nucleótidos são compostos por uma pentose (desoxirribose), um grupo fosfato e uma base nitrogenada que no caso do DNA são a adenina ($\mathcal{A}$), a guanina ($\mathcal{G}$), a timina ($\mathcal{T}$) e a citosina ($\mathcal{C}$). As bases nitrogenadas podem ser classificadas de acordo com a sua estrutura em purinas (adenina e a guanina) e pirimidinas (citosina e timina).

A molécula de DNA tem uma estrutura semelhante à de uma escada torcida, formando uma espiral, designada de dupla hélice. Os degraus da escada são a representação das ligações por pontes de hidrogénio que se formam entre bases que se dizem *complementares*: duas, no caso de $\mathcal{A}$ com $\mathcal{T}$, e três, no caso de $\mathcal{C}$ com $\mathcal{G}$ (ver figura 1.1).

A sequência de DNA é representada como combinação de quatro nucleótidos que se encadeiam como as letras do alfabeto ao longo de um texto sem espaços, fazendo sentido a atribuição do nome de *linguagem genética ou texto genético* ao texto formado pelos nucleótidos que constituem o DNA.

Nas sequências dos nucleótidos, estão contidas informações relativas às características hereditárias, assim como a informação para a produção contínua de proteínas e consequente sobrevivência dos seres vivos. A sequência de nucleótidos que constitui o DNA é composta por duas partes: a parte de subsequências de código (regiões codificantes)

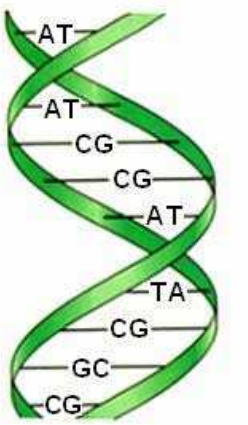

Figura 1.1: Representação esquemática da estrutura do DNA.

e a de subsequências de não-código (regiões não codificantes). As subsequências de código consistem no conjunto das partes da sequência com significado em termos de produção de proteínas. Nas subsequências de código o número total de nucleótidos é um múltiplo de três e cada tripleto de nucleótidos constitui o código de um *aminoácido*, isto é, da unidade proteica na construção de uma proteína.[1] A cada grupo de três nucleótidos que codifica um aminoácido chama-se *codão*. Existem sessenta e quatro codões distintos ($4^3 = 64$), correspondendo aos sessenta e quatro arranjos possíveis dos quatro nucleótidos em grupos de três. As sequências de código são sequências de codões que começam na sua maioria com o codão $\mathcal{ATG}$ (conhecido por codão de iniciação) e terminam com um dos codões terminais $\mathcal{TAA}$, $\mathcal{TAG}$ ou $\mathcal{TGA}$.

Apesar do "alfabeto" do DNA consistir apenas em 4 letras, os textos completos são muito longos. O DNA dos seres humanos, por exemplo, contém cerca de $2,9 \times 10^9$ pares de nucleótidos. No entanto, apenas cerca de 2% do texto completo constitui as regiões codificantes [132].

Cada sequência de código associada a uma dada proteína encontra-se contida num *gene*. Ao conjunto de todos os genes de uma espécie é chamado de *genoma* dessa espécie. Muitos genes, por sua vez, ainda apresentam uma estrutura sequencial alternada de duas partes: as subsequências que constituem código de proteínas — os *exões*; as restantes subsequências — os *intrões*.

[1] Uma sequência consecutiva de n nucleótidos é conhecida como oligonucleótido de comprimento n.

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 $\mathcal{TAA}$, $\mathcal{TAG}$ e $\mathcal{TGA}$ 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 não é bijectiva. 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 tabela 1.1).

É tradicional dividir os seres vivos em dois grandes grupos de acordo com o tipo de células: os eucariotas e os procariotas.[2] A grande diferença entre as células é essencialmente estrutural: as células dos eucariotas têm núcleo individualizado (DNA compartimentado) e com vários organelos membranares; as células procariotas são mais simples, não possuindo núcleo nem organelos membranares (compartimentos intracelulares).

Neste trabalho, quando houver referência a células e a estruturas celulares estas serão de eucariotas a menos que haja especificação do contrário. As células dos eucariotas são constituídas essencialmente por dois espaços intra-celulares: o núcleo e o citoplasma — espaço que circunda o núcleo. A síntese de proteínas dá-se no citoplasma, depois da transcrição do DNA, da remoção das sequências não codificantes no DNA transcrito ("splicing") e da sua migração do núcleo para o citoplasma.

Desde o DNA à construção de proteínas, existem mais duas estruturas genéticas com especial importância: o RNA (*RiboNucleic Acid*) e os ribossomas. O RNA (mais concretamente o mRNA — ácido ribonucleico mensageiro) tem a função de transportar a mensagem genética desde o DNA, no núcleo, até ao ponto onde a mensagem é traduzida no citoplasma. A molécula de RNA distingue-se do DNA por ser uma molécula de cadeias simples em que o nucleótido timina é substituído pelo nucleótido uracilo ($\mathcal{U}$) e o açúcar desoxirribose pela ribose. Os ribossomas são pequenos organelos não membranosos que se encontram distribuídos por todo o citoplasma com a função de traduzir o código genético. O ribossoma liga-se ao mRNA, descodificando a informação contida nesta molécula, através da ligação a cada codão de uma molécula de tRNA (ácido ri-

[2] Actualmente o grupo dos procariotas está subdividido em dois grupos: bacteria e archaea - Mayr (1990). Mas nem sempre é usada a subdivisão eucariotas e procariotas. Outras existem, como por exemplo, a subdivisão em três grupos: bacteria, archaea e eucariotas.

codão	aminoácido	codão	aminoácido	codão	aminoácido	codão	aminoácido
$\mathcal{TTT}$	F Phe	$\mathcal{TCT}$	S Ser	$\mathcal{TAT}$	Y Tyr	$\mathcal{TGT}$	C Cys
$\mathcal{TTC}$	F Phe	$\mathcal{TCC}$	S Ser	$\mathcal{TAC}$	Y Tyr	$\mathcal{TGC}$	C Cys
$\mathcal{TTA}$	L Leu	$\mathcal{TCA}$	S Ser	$\mathcal{TAA}$	X Ter	$\mathcal{TGA}$	X Ter
$\mathcal{TTG}$	L Leu	$\mathcal{TCG}$	S Ser	$\mathcal{TAG}$	X Ter	$\mathcal{TGG}$	W Trp
$\mathcal{CTT}$	L Leu	$\mathcal{CCT}$	P Pro	$\mathcal{CAT}$	H His	$\mathcal{CGT}$	R Arg
$\mathcal{CTC}$	L Leu	$\mathcal{CCC}$	P Pro	$\mathcal{CAC}$	H His	$\mathcal{CGC}$	R Arg
$\mathcal{CTA}$	L Leu	$\mathcal{CCA}$	P Pro	$\mathcal{CAA}$	Q Gln	$\mathcal{CGA}$	R Arg
$\mathcal{CTG}$	L Leu	$\mathcal{CCG}$	P Pro	$\mathcal{CAG}$	Q Gln	$\mathcal{CGG}$	R Arg
$\mathcal{ATT}$	I Ile	$\mathcal{ACT}$	T Thr	$\mathcal{AAT}$	N Asn	$\mathcal{AGT}$	S Ser
$\mathcal{ATC}$	I Ile	$\mathcal{ACC}$	T Thr	$\mathcal{AAC}$	N Asn	$\mathcal{AGC}$	S Ser
$\mathcal{ATA}$	I Ile	$\mathcal{ACA}$	T Thr	$\mathcal{AAA}$	K Lys	$\mathcal{AGA}$	R Arg
$\mathcal{ATG}$	M Met	$\mathcal{ACG}$	T Thr	$\mathcal{AAG}$	K Lys	$\mathcal{AGG}$	R Arg
$\mathcal{GTT}$	V Val	$\mathcal{GCT}$	A Ala	$\mathcal{GAT}$	D Asp	$\mathcal{GGT}$	G Gly
$\mathcal{GTC}$	V Val	$\mathcal{GCC}$	A Ala	$\mathcal{GAC}$	D Asp	$\mathcal{GGC}$	G Gly
$\mathcal{GTA}$	V Val	$\mathcal{GCA}$	A Ala	$\mathcal{GAA}$	E Glu	$\mathcal{GGA}$	G Gly
$\mathcal{GTG}$	V Val	$\mathcal{GCG}$	A Ala	$\mathcal{GAG}$	E Glu	$\mathcal{GGG}$	G Gly

Tabela 1.1: Correspondências entre codões e aminoácidos. Na representação do aminoácido tem-se a letra que representa o aminoácido e o respectivo acrónimo ou "Ter", no caso dos codões de terminação.

bonucleico de transferência) com a sequência complementar (anticodão) transportando o respectivo aminoácido (ver figura 1.2). Cada mRNA, apesar de ter um tempo de vida limitado, pode dar origem a mais do que uma proteína.

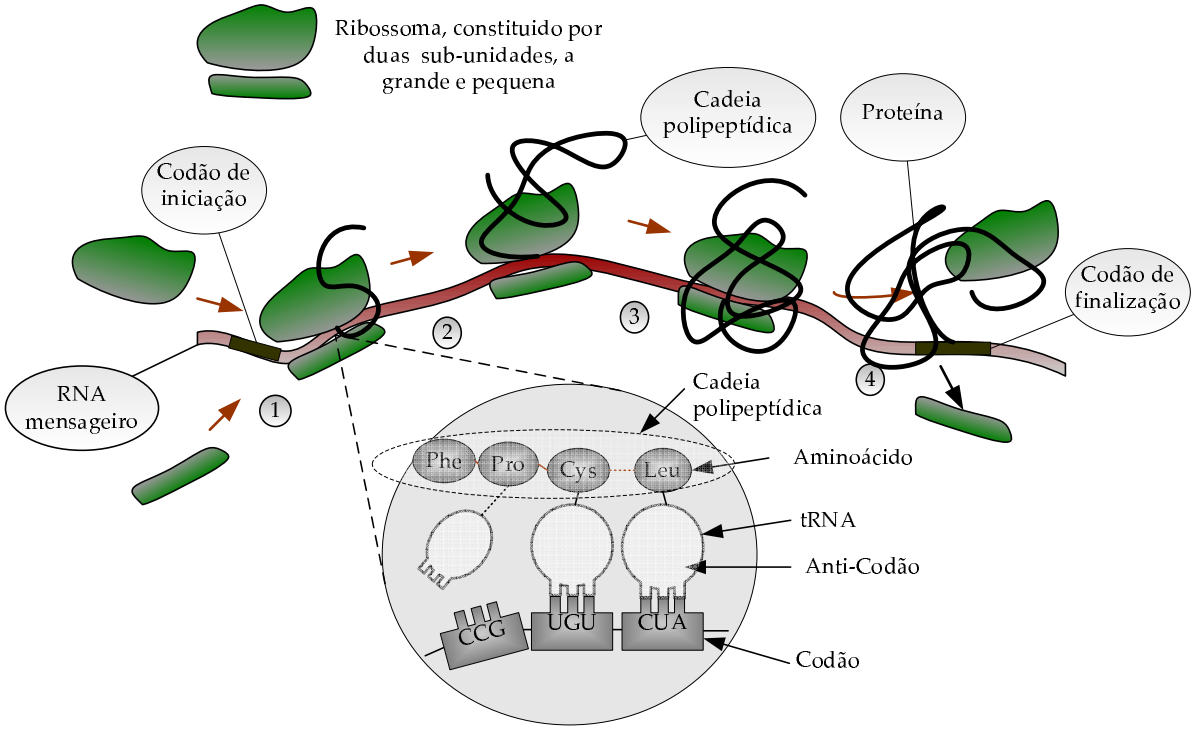

Figura 1.2: Tradução de código genético pelo ribossoma.

1.1.2 As proteínas

As proteínas são moléculas muitos especiais essenciais à vida dos seres vivos. Estas moléculas realizam as mais variadas funções no nosso organismo, desde o transporte de nutrientes e metabólitos à catálise de reacções biológicas. Apesar da complexidade das suas funções, as proteínas têm na sua constituição apenas 20 aminoácidos diferentes. No entanto, a maioria das proteínas são constituídas por mais de 200 aminoácidos (até vários milhares).

Quando se descreve a estrutura das proteínas fala-se em estrutura primária, secundária, terciária e em alguns casos em estrutura quaternária (ver figura 1.3). A estrutura primária consiste apenas na sequência de aminoácidos, sem contar com a orientação

espacial da molécula.

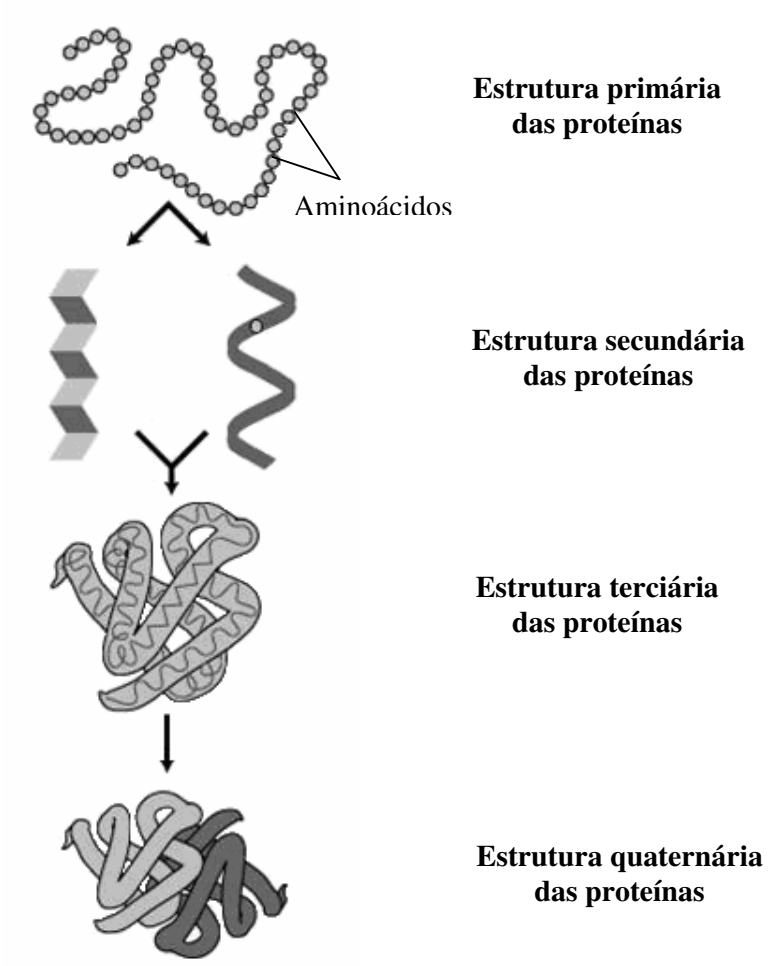

Figura 1.3: Representação esquemática das estruturas das proteínas: primária, secundária, terciária e quaternária. Adaptado de uma figura do National Human Genome Research Institute (NHGRI), disponível em http://www.nhgri.nih.gov/DIR/VIP/.

A estrutura secundária é uma função dos ângulos formados pelas ligações peptídicas que ligam os aminoácidos. A conformação espacial é mantida graças às interacções intermoleculares entre o átomo de azoto, ou nitrogénio, do grupo amina e um átomo de oxigénio do grupo carboxílico do carbono alfa do aminoácido seguinte (alfa designa qual o átomo de carbono que tem os grupos que participam na ligação peptídica). Na maior parte das vezes, estas ligações forçam a proteína a assumir uma forma helicoidal, como uma corda enrolada em torno de um tubo imaginário — esta forma é chamada de hélice

alfa. Outra forma de estrutura secundária é a folha beta, onde dois segmentos da cadeia interagem paralelamente. Os "turns" são o terceiro tipo das estruturas secundárias clássicas, e são responsáveis pela reversão da direcção da cadeia polipeptídica.

A estrutura terciária ocorre quando existem interacções entre as folhas beta e as hélices alfa, consistindo num rearranjo espacial da estrutura secundária. Assim, a forma das proteínas está relacionada com sua estrutura terciária. O que determina a estrutura terciária são as cadeias laterais dos aminoácidos; algumas cadeias são tão longas e hidrofóbicas que perturbam a estrutura secundária helicoidal, provocando a dobra ou "looping" da proteína. Muitas vezes, as partes hidrofóbicas da proteína agrupam-se no interior da proteína dobrada, longe da água e dos iões do meio onde a proteína se encontra, deixando as partes hidrofílicas na superfície exterior da estrutura da proteína. Regiões como local activo e local de regulação são propriedades da estrutura terciária.

Existe também a estrutura quaternária: certas proteínas, tal como a hemoglobina, são compostas por mais de uma unidade polipeptídica (cadeia proteica). A interacção entre várias cadeias peptídicas determina a estrutura quaternária.

As proteínas podem ser simples (constituídas somente por aminoácidos) ou conjugadas (que contêm grupos prostéticos, isto é, outras moléculas, tais como açúcares, iões metálicos, pigmentos, etc.). A hemoglobina e as lipoproteínas são exemplos de proteínas conjugadas.

1.2 Motivação e objectivos gerais da tese

A genómica a uma e duas dimensões, a genómica e proteómica estrutural, e a proteómica e vias biológicas, são três grandes áreas de investigação actual de algumas biociências (em particular da bioinformática).[3] O estudo a desenvolver nesta dissertação surge no contexto da genómica a uma dimensão, em particular no contexto de sequências genéticas.

De modo geral os objectos de estudo deste trabalho são as sequências simbólicas, mas

[3] A bioinformática é uma ciência que estuda técnicas computacionais e matemáticas de análise de estruturas genéticas (DNA, RNA e proteínas).

dado o enfoque biológico desta dissertação as metodologias a apresentar são aplicadas sobre sequências genéticas. Assim, serão exploradas e desenvolvidas metodologias que nos permitam estudar a estrutura de sequências de símbolos, tendo como objectivo principal contribuir para a caracterização das sequências de DNA, aliando sempre que possível uma explicação biológica. Em particular, esta dissertação centra-se no estudo de estruturas de correlação nas sequências de DNA, com o objectivo de contribuir para melhorar as técnicas de localização de genes e compressão de genomas.

No âmbito das técnicas de localização de genes, o estudo incide no coeficiente espectral que reflecte a periodicidade de período três. Neste contexto, o objectivo centra-se em explicitar as razões para a evidência dessa periodicidade, em pesquisar métodos rápidos de cálculo e em estudar a convergência entre as diferentes metodologias que têm sido usadas.

No que diz respeito à compressão de genomas, o estudo centra-se na compressão de regiões codificantes, tendo por objectivo explorar a relevante periodicidade de período três que geralmente caracteriza estas regiões.

1.3 Publicações resultantes deste trabalho

As publicações feitas no âmbito desta tese são quatro: [5], [4], [124], [57]. Em paralelo e conjuntamente com o grupo de bioinformática da Universidade de Aveiro, realizaram-se recentemente alguns trabalhos dos quais estão publicados [116], [123], [60] e [3].

O trabalho [5] encontra-se desenvolvido no capítulo três desta tese. Este trabalho apresenta essencialmente várias metodologias de análise espectral que têm sido aplicadas às sequências de DNA, onde se mostra a relação existente entre as várias metodologias e se dá uma justificação para aplicar a análise espectral às sequências simbólicas (sequências de DNA).

Em [4] realça-se a importância, já reconhecida, do coeficiente espectral que reflecte a periodicidade de período três. Apresenta-se um método rápido de cálculo de algumas riscas espectrais e conjuntamente algumas condições quer necessárias quer suficientes para os valores assumidos por estas riscas espectrais. Este trabalho é também parte

integrante desta tese e encontra-se no início do capítulo quatro.

Em [124] e [57] apresenta-se uma contribuição ao nível dos modelos de compressão de sequências de DNA com o objectivo de melhorar os actuais resultados de compressão destas sequências. Foi dedicada especial importância às regiões codificantes, e a contribuição vai no sentido de melhorar o modelo de compressão tendo em conta os estudos feitos sobre a periodicidade das regiões de código. O capítulo cinco baseia-se essencialmente nestas duas publicações.

Para além dos assuntos abordados nesta dissertação, foram alvo de estudo outros tópicos no contexto da bioinformática que resultaram nas publicações [116], [123], [60] e [3]. Os trabalhos [116], [123] são dirigidos a estudos de associação entre pares de símbolos com o propósito de melhorar o conhecimento sobre a descodificação das sequências de código pelo ribossoma. Em [60] e [3] desenvolveram-se algoritmos de "biclustering", para evidenciar estruturas atípicas num conjunto de dados dispostos de forma matricial.

1.4 Organização da dissertação

Esta dissertação é constituída, para além desta introdução, por mais cinco capítulos:

- **Capítulo 2 – Estado da arte:** resume várias metodologias que estudam a estrutura de correlação das sequências de DNA. É dada especial atenção a metodologias de análise espectral direccionadas para a periodicidade de período três e a modelos de compressão, entropia e informação mútua;

- **Capítulo 3 – Análise espectral de sequências simbólicas:** apresenta comparações e/ou relações de várias metodologias que têm sido aplicadas às sequências de DNA e que têm por base a análise espectral;

- **Capítulo 4 – Distribuição dos símbolos e espectro da sequência:** apresenta métodos rápidos de cálculo de algumas riscas espectrais e explora algumas das suas propriedades em particular da risca que reflecte a peridicidade de período três;

- **Capítulo 5 – Modelo de três estados:** apresenta um novo modelo de compressão, que na maior parte das regiões de código permite obter melhores resultados de compressão comparativamente às metodologias existentes, e são discutidas explicações para os resultados experimentais obtidos;
- **Capítulo 6 – Conclusões e trabalho futuro:** apresenta as conclusões desta dissertação conjuntamente com ideias de trabalho futuro.

Capítulo 2

Estado da arte

Um dos grandes objectivos que tem motivado a investigação de muitos dos cientistas que aplicam modelos matemáticos a dados genéticos é encontrar correlações de curto, médio e longo alcance em sequências de DNA. Genericamente, a motivação que está por trás deste trabalho é a mesma, embora na prática este seja dedicado a correlações de curto alcance.

Neste capítulo, apresenta-se um resumo de várias metodologias matemáticas que, de modo relevante, foram aplicadas às sequências de DNA. O enfoque é dado às metodologias mais relacionadas com os dois problemas de interesse biológico que motivaram o desenvolvimento desta dissertação: a localização automática de genes e a compressão de sequências genéticas. Também se discute a validação dos pressupostos de alguns métodos, tendo em conta a natureza simbólica e não estacionária das sequências de DNA.

2.1 Contextualização

Nesta secção, pretende-se criar uma perspectiva global sobre a grande quantidade de trabalho que tem sido realizado no contexto do estudo de estruturas de correlação nas sequências de DNA. Neste contexto, apresentam-se de forma resumida diversas metodologias, a sua aplicabilidade e os resultados.

A parceria estabelecida entre a genómica e a modelação matemática tem sido bem sucedida, reflectindo-se no contínuo trabalho que tem sido publicado. A base desta colaboração assenta essencialmente em dois objectivos:

- Justificar matematicamente leis genómicas reconhecidas empiricamente pelos biólogos;
- Justificar biologicamente leis matemáticas encontradas nos genomas.

No entanto, no âmbito do estudo da estrutura das sequências de DNA, têm sido levantados alguns problemas, em particular: a inexistência de conformidade entre resultados e a falta de motivação biológica para o estudo de algumas estruturas de correlação [97]. No contexto do estudo de estruturas de correlação das sequências de DNA, outro problema de natureza mais específica se coloca frequentemente: a aplicabilidade das metodologias às sequências de DNA. Ainda nesta introdução, é apresentada uma lista de várias metodologias aplicadas ao DNA e uma breve discussão sobre a aplicabilidade de algumas delas.

2.1.1 Metodologias aplicadas ao DNA

Segue-se uma listagem de metodologias aplicadas às sequências de DNA, não com o objectivo de fazer uma enumeração exaustiva de todas as metodologias, mas sim de realçar o grande número de diferentes metodologias que têm sido aplicadas a este tipo de dados. Para cada metodologia indicada, são referenciados exemplos de trabalhos onde essa metodologia é usada sobre dados de natureza genética.

1. Análise espectral
   (a) Transformada de Fourier (Discrete Fourier Transform — DFT ou Fast Fourier Transform — FFT) [11, 12, 38, 45, 62, 88, 102, 93, 118, 121, 134, 142, 151, 150, 153, 154]
   (b) Envolvente espectral [137, 138]
   (c) Transformada localizada de Fourier (Short-Time Fourier Transform — STFT) [10, 36, 139, 155]

2. Teoria da informação
   (a) Informação mútua [67, 71, 74, 75, 76, 86, 94, 102, 95, 103]
   (b) Entropia [54, 30, 73, 90]
   (c) Entropia de Rényi [77, 90]
   (d) Decomposição da informação (Information Decomposition — ID) [86, 85]
   (e) Divergência de Jensen-Shannon [30, 126, 29, 98]
   (f) Divergência de Jensen-Rényi [117]
   (g) Informação mútua média (Average Mutual Information — AMI) [66, 67]
   (h) Critério de informação Bayesiana (Bayesian Information Criterion — BIC) [98]
3. Análise de onduletas
   (a) Transformada discreta com onduletas [17, 25, 155, 16, 146, 14, 15]
   (b) Máximos do módulo da transformada com onduletas (Wavelet Transform Modulus Maxima — WTMM) [13]
4. Cadeias de Markov
   (a) Cadeias de Markov de ordem N [50, 2]
   (b) Cadeias de Markov não observáveis (Hidden Markov Models — HMM) [35, 107, 20]
   (c) Cadeias de Markov binárias de ordem N [147]
5. Análise de associação
   (a) Qui-quadrado [64, 61]
   (b) Valor de z [133, 78, 34]
   (c) Método de Monte Carlo [56]
6. Métodos linguísticos
   (a) Análise de Zipf [83, 136, 81, 135]

   (b) Análise de redundância de Shannon de ordem N [112, 136]

   (c) Linguagem regular [95].

7. Análise da percentagem de $\mathcal{C} + \mathcal{G}$

   (a) Análise de variância — Anova [93, 99, 100]

   (b) Teste binomial [100]

8. Análise de correlação (outros métodos)

   (a) Análise fractal [14, 106, 152, 13]

   (b) Função de autocorrelação [72, 71, 83, 153, 154, 18, 152, 94, 74, 27, 143]

   (c) Análise R/S: índice de Hurst/análise de Hurst [158, 106, 158, 159, 16]

   (d) Modelos dinâmicos [7], modelo evolutivo [108, 152, 160]

   (e) Análise factorial [115, 26]

   (f) Análise de caminhos de DNA [121, 82, 13, 35, 25]

   (g) Movimento Browniano fraccional (Fractional Brownian Motion — FBM) [154, 8]

   (h) Ruído Gaussiano fraccional (Fractional Gaussian Noise — FGN) [106]

   (i) Caminhos de Lévy [13, 37, 131]

   (j) Análise de entropia de difusão (Diffusion Entropy Analysis — DEA) [131, 40]

   (k) Análise de expoente de Holder [70]

   (l) Min-max [121]

   (m) Análise de flutuações sem tendências (Detrended Fluctuation Analysis — DFA) [122, 38, 157]

   (n) Dimensão da correlação [159]

   (o) Análise em componentes principais (Principal Component Analysis — PCA) [63]

   (p) Modelos autoregressivos [105, 51]

Algumas destas metodologias podem ser encontradas, de forma agrupada, em alguns artigos de revisão, ver [97, 108].

2.1.2 Validação dos pressupostos

O ajuste perfeito de modelos matemáticos a realidades físicas é pouco frequente. Além disso, quando os modelos pressupõem hipóteses nem sempre é simples confirmá-las. Quando se pretende modelar sequências de DNA estes problemas mantêm-se.

De seguida discutem-se duas características das sequências genéticas que poderão entrar em conflito com os pressupostos de algumas metodologias: a natureza simbólica e o comportamento geralmente não estacionário.

2.1.2.1 Dados simbólicos

As metodologias naturais de análise de dados de natureza simbólica são geralmente metodologias baseadas nas frequências dos símbolos, em que frequências se traduzem, por exemplo, em tabelas de contingência ou em estimativas de probabilidades.

As cadeias de Markov são um caso particular de metodologias adaptadas a dados simbólicos. De notar que uma cadeia de Markov de ordem 1 é um modelo simples, no entanto, é também um modelo pobre para descrever as sequências de DNA [97], embora em 1989 se afirmasse que as cadeias de Markov de ordem 1 geralmente conseguiam descrever as sequências de DNA (ver [49]). Por outro lado, as cadeias de Markov de ordem elevada poderão ser penalizadas por terem muitos parâmetros livres [97]. No entanto em [51] fica a ideia de que algumas sequências poderão ser descritas por modelos de Markov de ordem não superior a trinta. Outra alternativa são os Modelos de Markov não observáveis (HMM), que ultrapassam alguns dos problemas das cadeias de Markov, especificamente a restrição de estacionaridade, continuando a ser modelos que respeitam a natureza simbólica dos dados (ver, por exemplo, [97]).

Outros exemplos de metodologias ajustadas a dados simbólicos e que têm sido aplicadas sobre as sequências de DNA são, por exemplo: os estudos linguísticos; a análise de associação e de contextos enviesados; o cálculo de entropia e informação mútua; a análise estatística do qui-quadrado (ver, por exemplo, [102, 95, 116]).

No entanto, as metodologias para dados de natureza simbólica não exploram alguns tipos de correlações existentes nas sequências de DNA e geralmente não toleram a falta de estacionaridade da sequência [21]. Assim, no sentido de explorar outros tipos de

correlação, surge a necessidade de ajuste de algumas metodologias de dados numéricos às sequências de DNA.

Na aplicação de metodologias de análise de sequências numéricas, a dificuldade devida ao facto das sequências de DNA serem simbólicas tem sido parcialmente ultrapassada por vários autores com o mapeamento em sequências numéricas ou vectoriais. No entanto, não se pode esquecer que os dados genéticos consistem em sequências simbólicas que não apresentam estrutura algébrica, e ao fazer-se o mapeamento de símbolos em números deve-se ter presente que os dados continuam a ser de natureza simbólica.

Naturalmente surgem dúvidas associadas à validade da interpretação numérica e ao tipo de mapeamento que melhor representa uma sequência de DNA, as quais se procuram ultrapassar com base em motivações biológicas. Por exemplo, em [38] são apresentados seis tipos de mapeamento baseados em regras biológicas:

- Regra purina *versus* pirimidina (RY)

$$\mathcal{A}, \mathcal{G} \rightarrow 1 \text{ e } \mathcal{C}, \mathcal{T} \rightarrow -1$$

- Regra da ligação de hidrogénio

$$\mathcal{C}, \mathcal{G} \rightarrow 1 \text{ e } \mathcal{A}, \mathcal{T} \rightarrow -1$$

- Regra híbrida

$$\mathcal{C}, \mathcal{A} \rightarrow 1 \text{ e } \mathcal{G}, \mathcal{T} \rightarrow -1$$

- Regra $\mathcal{A}\bar{\mathcal{A}}$

$$\mathcal{A} \rightarrow 1 \text{ outros casos } \rightarrow -1$$

- Regra $\mathcal{T}\bar{\mathcal{T}}$ (análoga à regra $\mathcal{A}\bar{\mathcal{A}}$)
- Regra $\mathcal{G}\bar{\mathcal{G}}$ (análoga à regra $\mathcal{A}\bar{\mathcal{A}}$)
- Regra $\mathcal{C}\bar{\mathcal{C}}$ (análoga à regra $\mathcal{A}\bar{\mathcal{A}}$)

Um mapeamento também frequentemente usado é o das sequências indicadoras onde a sequência simbólica de DNA é representada por quatro sequências binárias associadas

a cada um dos nucleótidos, ou seja, cada sequência binária indica com um 1 a posição do respectivo símbolo e com um 0 a posição de qualquer outro símbolo (ver, por exemplo, [154]). Feita a interpretação da sequência simbólica como sequência numérica (mapeamento), as metodologias de análise de sequências numéricas podem ser usadas.

No âmbito do estudo de estruturas de correlação de curto alcance de uma sequência é muito frequente o estudo de periodicidades. Para esse efeito, recorre-se frequentemente à análise espectral a qual é uma metodologia para dados de natureza numérica. Naturalmente, na aplicação mais directa desta metodologia às sequências de DNA recorre-se a algum tipo de mapeamento. Na bibliografia da área encontram-se diferentes mapeamentos (ver, por exemplo, [48, 155, 11, 12, 151, 150, 153, 102, 134]), alguns dos quais foram indicados atrás.

Mantendo a natureza simbólica das sequências de DNA, o estudo de periodicidades também poderá ser feito de forma exaustiva por contagem de repetições de símbolos ao longo da sequência [76, 64].[1]

Outras análises, geralmente dedicadas a dados numéricos, têm sido aplicadas às sequências de DNA, por exemplo: análise fractal; modelação por movimento Browniano fraccional ou por ruído Gaussiano fraccional; caminhos de Lévy; modelos dinâmicos; a análise de onduletas, etc.. Também para estes vários tipos de análises surge a necessidade de mapeamentos e consequentemente os riscos de tirar conclusões que deles dependam.

De seguida exemplifica-se um dos possíveis problemas do mapeamento, no âmbito da análise da estrutura harmónica de uma sequência simbólica:

Exemplo 2.1.1. *Considere-se a seguinte sequência simbólica:*

$$s = (\mathcal{ATGCACATGCAC}...)$$

onde o mapeamento

$$\mathcal{A} \mapsto 1,\ \mathcal{T} \mapsto -1,\ \mathcal{G} \mapsto 1,\ \mathcal{C} \mapsto -1,$$

leva a uma sequência numérica de período dois. Por outro lado, o mapeamento

$$\mathcal{A} \mapsto -1,\ \mathcal{T} \mapsto -1,\ \mathcal{G} \mapsto 1,\ \mathcal{C} \mapsto 1,$$

[1] É de notar que em [64] o que é apresentado é um perfil periódico. Não são usadas todas as repetições, mas apenas as que levam à rejeição do teste do χ^2.

leva a uma sequência de período 6.

Este exemplo mostra claramente que algumas das estruturas com relevância harmónica podem estar escondidas ou expostas, dependendo do mapeamento entre símbolos e números.

2.1.2.2 Estacionariedade

A estacionaridade é um requisito comum de várias metodologias, quer das que estudam sequências de natureza simbólica, quer de natureza numérica. Muitos dos modelos matemáticos dedicados ao estudo de sequências de símbolos e/ou de séries temporais assumem que as sequências têm comportamento estacionário. No entanto, as sequências de DNA são geralmente classificadas como sequências simbólicas não estacionárias.

Diz-se que um processo aleatório é estritamente estacionário se a sua distribuição de probabilidade é independente do tempo (ou da posição) e considera-se estacionário de segunda ordem (ou em sentido lato) se a média e a variância são independentes do tempo (ou da posição).

Algumas vezes o conceito de estacionaridade surge associado ao conceito de homogeneidade, mas não são conceitos sinónimos. A noção de processo homogéneo pode surgir associada a cadeias de Markov, dizendo-se homogéneo se as probabilidades de transição são independentes do tempo (ou da posição). Assim, em particular, um processo que não seja homogéneo também não é estacionário, pelo que a falta de estacionaridade das sequências de DNA pode ser justificada pela falta de homogeneidade (distribuição de probabilidades dependente do tempo ou posição).

O estudo da homogeneidade em sequências de DNA nem sempre se refere às probabilidades de transição, mas também à equiprobabilidade temporal de estruturas simbólicas (por exemplo, o teor de $\mathcal{C}+\mathcal{G}$). É no contexto da equiprobilidade que alguns autores classificam as sequências de DNA em homogéneas, heterogéneas de forma simples ou heterogéneas de forma complexa (ver, por exemplo [99, 101, 98, 119, 141]).

Quando as sequências não são estacionárias e se pretende, mesmo assim, aplicar ou "forçar" a aplicação de métodos que, de alguma forma, ficam condicionados por esta característica, têm sido implementadas essencialmente duas abordagens:

- Estudo da sequência por janelas temporais ou blocos;
- Segmentação da sequência em partes homogéneas.

A criação de janelas temporais parece ser a solução mais simples (ver, por exemplo, [155]). Esta técnica poderá reduzir o efeito da falta de estacionaridade e homogeneidade, no entanto não o elimina.

Face a não existir garantia da homogeneidade com as janelas temporais, poderão ser preferidos os métodos que segmentem as sequências em partes homogéneas. As ferramentas de segmentação que têm sido usadas baseiam-se em diferentes modelos matemáticos: a divergência de Jensen-Shannon [104, 30, 29, 126, 98, 65]; métodos baseados na máxima verosimilhança [98]; a análise de variância [100, 99]; min-max [121]; análise de flutuações sem tendências [122, 38, 157]; dimensão da correlação [159], análise de Hurst [106, 158, 159]; análise de onduletas [13, 16, 146]; modelos de Markov não observáveis [97]. De notar que o facto de se ter uma subsequência homogénea não é condição suficiente para garantir que esta seja estacionária.

2.2 Estruturas de correlação no DNA

Em termos genéricos, a descoberta da estrutura das sequências de DNA passa pela descoberta das leis de correlação que descrevem a sequência, as quais podem ser de curto, médio ou longo alcance.

Para contextualizar o trabalho a apresentar nos restantes capítulos descreve-se de seguida e de forma sucinta algumas estruturas de correlação nas sequências de DNA que têm sido estudadas. Relembrando que o objectivo particular desta dissertação é o de melhorar as técnicas de localização automática de genes e de compressão de sequências de DNA.

Tem sido dado grande realce à descoberta de leis de correlação de longo alcance em sequências de DNA. Frequentemente tem sido publicado que a correlação de longo alcance entre símbolos se comporta de acordo com leis de potência, $\frac{1}{f^{\alpha}}$ (ver, por exemplo, [63, 76]). Algumas extensões têm sido feitas: em [109, 41] são estudadas correlações de

longo alcance para as três posições dos codões nas sequências de código em separado, obtendo-se diferentes leis de potência para as três posições.

No entanto, a interpretação da correlação de longo alcance nas sequências de DNA não é fácil (ver, por exemplo, [16]), uma vez que as leis observadas podem ser apenas consequência da existência de replicação assimétrica ou de estrutura de mosaico. Em suma, podem resultar da falta de homogeneidade da sequência (ver [96]).

Alguns autores acreditam que a correlação entre nucleótidos tem crescido com a evolução das espécies, razão pela qual é tão importante para a biologia. É neste contexto que surgem os modelos evolutivos (ver, por exemplo, [160]) e os modelos dinâmicos (ver, por exemplo, [7, 108]). Estes modelos estão associados a fenómenos de mutação, inserção ou apagamento de nucleótidos que podem ocorrer em qualquer posição das sequências de DNA (de forma não determinística).

Ainda relacionado com a evolução das sequências de DNA, actualmente é dado grande interesse a técnicas de pesquisa de subsequências repetidas e estruturas de repetição [36]. Encontram-se subsequências repetidas de diversos tamanhos desde as unidades até às centenas. Alguns autores afirmam que as repetições podem parcialmente justificar a existência de correlações de longo alcance (ver, por exemplo, [73]). Mas o interesse por estruturas repetidas nas sequências de DNA não é recente (ver, por exemplo, [82, 77, 46, 24, 73, 119]).

Existem também muitos estudos dedicados a leis de correlação de curto alcance. Neste contexto, grande parte dos trabalhos são dirigidos a aspectos particulares e não globais da sequência, por exemplo, o estudo de associação entre pares de símbolos e análise de contextos (ver, por exemplo, [4, 34, 56, 61, 78, 133]).

As correlações de curto e médio alcance também têm sido exploradas através de modelos de Markov. Por exemplo, como já foi referido, em [51] mostra-se para as sequências de DNA estudadas o ajuste de um modelo de Markov de ordem não superior a 30.

No contexto do estudo das correlações, um tópico de destaque tem sido a detecção de periodicidades relevantes nas sequências de DNA. Têm sido detectadas periodicidades que reflectem correlações de curto, médio e longo alcance: de 3 bases (a desenvolver na subsecção seguinte); de 6 e 11 bases [153]; de cerca de 10.5 bases [144]; de 10-11 bases [71]; de 3, 10.5, 200 e 400 bases [145]; de 6, 7, 11 e 19 bases [87].

As periodicidades de período entre 10 e 11 têm sido bastante estudadas. Por exemplo, em [72, 71], observa-se que nas sequências de código a periodicidade de períodos entre 10 e 11 se destaca de forma relevante. Também se observa que a periodicidade de período três se destaca de forma ainda mais evidente. A periodicidade de 10-11 é entendida como reflexo da estrutura das proteínas, relacionando-se com a alternância de aminoácidos hidrofóbicos e hidrofílicos nas hélices alfa das sequências de código. Há autores que conseguem diferenciar eucariotas de *archaea* através dos máximos relativos das periodicidades dos genomas, na ordem das dezenas. Os eucariotas geralmente reflectem uma periodicidade relevante de período aproximadamente igual a 11, ao passo que os *archaea* apresentam uma periodicidade relevante para um período ligeiramente mais baixo de aproximadamente 10 (ver, por exemplo, [72, 118]).

A interpretação biológica das restantes periodicidades que de modo relevante têm sido detectadas nas sequências de DNA não tem sido simples. Existem várias tentativas de explicação que muitas vezes não passam de hipóteses pouco fundamentadas. Em [145], por exemplo, refere-se que a existência de periodicidades de período elevado poderá evidenciar uma organização segmentada do genoma.

Em assuntos relacionados com as correlações das sequências de DNA é comum surgir a referência a quatro trabalhos importantes, [134, 102, 121, 154]. Embora não sejam artigos recentes, são pioneiros no estudo de correlações em sequências de DNA.

Em [134] procuram-se periodicidades através da análise espectral. Define-se a transformada de Fourier da sequência dos nucleótidos, existindo a preocupação em efectuar um mapeamento da sequência de DNA invariante a nível da transformada. Este artigo é geralmente referenciado pelo método usado e não pelos resultados ao nível de estruturas de correlação detectadas nas sequências de DNA. No capítulo 3 desta dissertação encontra-se uma discussão comparativa com a metodologia usada neste artigo.

Em [102] procura-se verificar e caracterizar a existência de leis de correlação de longo alcance nas sequências de DNA, usando a informação mútua entre pares de símbolos e espectro de potência sugerido em [134]. Combinando as duas metodologias, verifica-se que as sequências de DNA estudadas apresentam correlação de longo alcance e que esta é devida à existência de estruturas repetitivas ao longo da sequência.

Em [121] a sequência de nucleótidos é mapeada numa sequência numérica de acordo com

a regra purina *vs* pirimidina. Sobre a sequência resultante do mapeamento estudam-se as flutuações das sequências de somas acumuladas, concluindo-se a existência de correlações de longo alcance. Observa-se também que a concentração de purinas e pirimidinas varia ao longo da sequência.

Em [154] apresenta-se uma função de autocorrelação para sequências simbólicas. Através do espectro de potência, conclui-se a existência de correlação de longo alcance nas sequências de DNA e uma relevante periodicidade de período três.

2.2.1 Localização de genes

Um problema já muito estudado, mas ainda actual, é a localização automática dos genes ou regiões de código. Para evidenciar o esforço que tem sido investido na resolução deste problema, seguem-se alguns nomes de programas frequentemente usados na localização de genes, maioritariamente bem sucedidos nos procariotas:

- GISMO [89];
- GlimmerM [111];
- EasyGene [91];
- GenemarkS [31];
- CRITICA [19];
- GeneMark [107];
- GenScan [39];
- GeneScan [142];
- GRAIL [69].

Nas sequências de DNA, os genes (regiões codificantes) encontram-se misturados com as regiões não codificantes. Na detecção automática de genes têm sido exploradas características estruturais dos genes e aplicadas diferentes metodologias.

Algumas técnicas de localização usam genes conhecidos de outras espécies com a mesma função, isto é, genes que codificam a mesma proteína, e estudam índices de semelhança entre o gene conhecido e extractos das sequências onde se pretende localizar o gene (ver, por exemplo, [59, 19]). Apesar dos bons resultados que têm sido obtidos, os métodos baseados apenas nestas características não detectam genes com novas funções (genes que codificam novos tipos de proteínas).

A detecção de genes também pode ser vista como a segmentação da sequência de DNA em duas regiões (codificante e não codificante) e naturalmente os algoritmos genéricos dedicados à segmentação foram e têm sido muito explorados e aplicados neste contexto (sobre este assunto, ver [35], onde se descreve um conjunto de diferentes métodos de segmentação aplicados às sequências de DNA). Em [28] é apresentada uma abordagem computacional para encontrar fronteiras entre as sequências de código e não código usando métodos de segmentação. Em [158] os autores analisam as sequências de DNA através de uma matriz de transição e calculam algumas medidas de complexidade, pretendendo encontrar uma boa medida de complexidade para distinguir claramente diferentes regiões funcionais nas sequências de DNA. Em [21], são aplicadas técnicas de modelação não linear (NM), observa-se que geralmente as regiões de código apresentam sequências com disposição aparentemente aleatória e as regiões de não código apresentam assinaturas determinísticas. Em [67] é usada a informação mútua média (AMI) para diferenciar as sequências que são ou não de código, evidenciando-se duas regiões com distribuições muito diferentes. Em [107] é usada uma modificação do algoritmo de Viterbi para descobrir a sequência escondida que localiza os genes. Ainda com o propósito de diferenciar as regiões codificantes, muitas outras metodologias têm sido aplicadas às sequências de DNA: modelo de Markov interpolado [128]; mapas auto-organizados [110]; divergência de Jensen-Rényi; etc..

Também no contexto da detecção de genes têm sido muito usados os estudos da periodicidade, especialmente a periodicidade de período três, pois grande parte das regiões de código evidenciam essa periodicidade (ver, por exemplo, [11, 12, 21, 67, 88, 108, 118, 142, 144, 150, 151, 149, 148, 153, 154, 155, 158, 159]). Pode-se naturalmente colocar a questão: a que se deve a tendência para a periodicidade de período três se destacar nas sequências de código? Nas sequências de código, completa-se um novo codão de três em três bases, o que poderia sugerir a periodicidade de período três, como se afirma

em [154]. No entanto, esta característica estrutural não garante destaque de qualquer tipo de periodicidade, pois o conjunto dos codões é constituído por todos os arranjos com repetição das quatro bases três a três. Outros trabalhos têm tentado justificar a periodicidade de período três: em [92, 5, 108, 118], demonstra-se que a causa da periodicidade três se evidenciar numa sequência deve-se ao facto da distribuição dos símbolos não ser uniforme pelas três posições dos codões; em [71] afirma-se que a periodicidade de período três se destaca porque existem diferenças nas frequências dos codões (podem ser encontrados contra-exemplos, contudo se para uma dada sequência de código se gerar aleatoriamente outra sequências que mantenham as frequências dos codões a periodicidade de período três toma proporções idênticas às da sequência original, [55]); em [48] conclui-se que o aparecimento da periodicidade de três é devido à dominância do nucleótido $\mathcal{G}$, observando-se que estudando apenas a base $\mathcal{G}$ se tem boas indicações sobre as regiões de período três o que consequentemente facilita a detecção de genes.

Em assuntos relacionados com detecção de periodicidades nas sequências de DNA e localização de genes é comum surgir a referência aos artigos [137, 142, 45, 12, 155]. Este conjunto de artigos descreve diferentes metodologias para estudar o espectro de potência de sequências simbólicas.

Em [137] estuda-se o espectro de potência de uma sequência numérica. Essa sequência numérica é obtida através de um mapeamento genérico de símbolos em números que maximiza a potência espectral. O cálculo do espectro de potência transforma-se num problema de valores próprios.

Em [142] é usado o mapeamento em sequências indicadoras e é utilizada análise espectral sobre as quatro sequências indicadoras. A medida espectral para cada valor da frequência é dada pela soma dos quatro módulos quadráticos dos coeficientes de Fourier correspondente a cada uma das sequências indicadoras.

Em [45] é mostrada a relação de equivalência entre dois métodos de análise espectral para sequências simbólicas: método baseado em sequências indicadoras e método baseado na estrutura de um tetraedro regular. A demonstração que o autor faz é generalizada para uma dimensão arbitrária (ver capítulo 3 desta dissertação para o caso particular das sequências de DNA).

Em [12, 11] é usada a análise espectral no sentido de identificar as regiões de código, sendo para tal efectuados vários mapeamentos das sequências de nucleótidos: em números complexos, em sequências indicadoras e em vectores de dimensão três. A análise espectral é usada para detectar regiões funcionais não só pelo estudo da amplitude mas também pelo estudo da fase.

Em [155] é usada a noção de envolvente espectral definida por [137], de uma forma adaptada a sequências não estacionárias, usando para isso a transformada localizada de Fourier e análise de onduletas sobre as sequências de DNA.

2.2.2 Métodos de compressão de DNA

As bases de dados que contêm os genomas sequenciados não param de crescer. Assim, para o armazenamento das sequências que descrevem os genomas, coloca-se a possibilidade de compressão. Segue-se uma descrição sobre aquilo que tem sido feito na área da compressão de sequências de DNA.

O primeiro método usado para comprimir DNA foi chamado de Biocompress [68]. O algoritmo Biocompress faz uso de uma característica usual encontrada nas sequências de DNA: a ocorrência de repetições invertidas complementares ("complemented inverted repeats"), também conhecidas por palíndromas complementares. O algoritmo de compressão Biocompress foi melhorado resultando no algoritmo Biocompress-2.

Foi proposta outra técnica de compressão baseada em repetições exactas, Cfact [125]. Contrariamente ao Biocompress, o Cfact não explora a potencial redundância dos palíndromas complementares. Na verdade, o Cfact é um algoritmo geral de compressão e não incluiu nenhuma característica particular das sequências de DNA.

A ideia de usar sub-sequências repetidas também foi usada em [42]. Contudo, neste caso são usadas repetições aproximadas conjuntamente com palíndromas complementares. Uma primeira versão do algoritmo foi chamado de GenCompress-1, seguido da segunda versão chamada de GenCompress-2.

Modificações posteriores do GenCompress resultaram em mais dois algoritmos: DNACompress e PatternHunter [43]. DNACompress é considerado mais rápido que o GenCompress. Antes da publicação do DNACompress, surgiu uma técnica baseada em

árvores de contexto (Context Tree Weighting — CTW) e na compressão LZ (Lempel-Ziv), CTW+LZ [114]. Comparando os algoritmos GenCompress e CTW+LZ, o último apresenta alguns ganhos em termos de compressão relativamente ao primeiro [114]. No entanto, resulta em maiores tempos de execução no caso de sequências longas [43].

Em [113] também foram propostos os métodos Dna1, Dna2 e Dna3. Estes métodos, apesar de mais rápidos, têm em muitos casos pior desempenho que o DNACompress.

Em [140] foi introduzido o método NMLComp onde é usada codificação baseada no modelo de máxima verosimilhança normalizada para regressão discreta (Normalized Maximum Likelihood – NML). Esta técnica de compressão de DNA explora regularidades escondidas (por exemplo, repetições aproximadas). Este algoritmo foi melhorado por integração de mais métodos com vários parâmetros resultando num método de compressão, GeNML [84], com melhor desempenho que o primeiro.

Mais recentemente, foi proposto um novo algoritmo, DNAPack, que usa a distância de Hamming para repetições e palíndromas complementares, e também o CTW ou códigos aritméticos de ordem dois para regiões não repetitivas [22]. O DNAPack apresenta ganhos relativamente ao DNACompress [22].

Apesar de muitas tentativas de compressão dos genomas, sequências de quatro bases, os resultados têm sido muito fracos, obtendo-se valores próximos de dois *bits* por símbolo. Acredita-se assim que no contexto da compressão, ainda poderá haver muito a fazer. Obviamente que quanto melhor for o conhecimento da estrutura de correlação das sequências, melhor se poderão comprimir.

2.3 Discussão

As sequências de DNA são simbólicas e geralmente não são sequências estacionárias. Para que se obtenham resultados que descrevam as sequências de DNA estes pressupostos têm de ser tomados em conta. Observa-se que muitos dos trabalhos desenvolvidos nestas áreas têm tido em atenção esses pressupostos, mas muito ainda há a fazer.

Tendo por base os trabalhos desenvolvidos no contexto da correlação evidenciam-se três tópicos sobre os quais assenta a motivação desta dissertação:

- Existem muitas metodologias de análise espectral que não têm aparente relação;
- A periodicidade de período três têm especial importância na detecção de genes;
- Devido à falta de modelos de compressão que se ajustem à natureza dos dados, os resultados de compressão são geralmente fracos.

Capítulo 3

Análise espectral de sequências simbólicas

Neste capítulo, são apresentados e discutidos alguns métodos de análise de Fourier aplicados a dados simbólicos, direccionados para as sequências de DNA, em particular para as sequências de nucleótidos. Serão considerados métodos baseados em sequências indicadoras e em tetraedros regulares, métodos de autocorrelação simbólica e métodos de envolvente espectral, em que para cada frequência se optimiza o mapeamento de símbolos em números, dando ênfase a qualquer característica periódica dos dados. Será também discutida a equivalência ou a relação entre esses métodos.

Mostra-se que é possível definir uma função de autocorrelação de dados simbólicos, assumindo apenas que se pode comparar quaisquer dois símbolos e decidir se são iguais ou distintos. Esta autocorrelação é uma sequência numérica e a sua transformada de Fourier pode também ser obtida através da soma dos quadrados dos módulos da transformada de Fourier das sequências indicadoras (sequências de zeros/uns que indicam a posição dos símbolos). Discute-se ainda outra interpretação que pode ser dada através do conceito de envolvente espectral: dentro de todos os mapeamentos possíveis entre símbolos e números, existe um que maximiza a energia espectral para cada frequência que conduz à soma dos quadrados dos módulos da transformada de Fourier das sequências indicadoras.

Partes do conteúdo deste capítulo foram publicadas em [4].

3.1 Introdução

A análise através de transformações lineares e não lineares são ferramentas naturais para dar algumas respostas quanto à existência de correlações. No entanto, a natureza simbólica dos dados apresenta-se como um problema/desafio (ver capítulo anterior). Por exemplo a análise de Fourier aplica-se a dados com determinada estrutura algébrica, como por exemplo os grupos (comutativos ou não comutativos). Mas nos dados simbólicos essa estrutura não está presente, uma vez que não faz sentido definir operações algébricas com os símbolos. No sentido de ultrapassar esta dificuldade, têm sido apresentadas algumas abordagens que são aparentemente distintas e que não têm explicitamente relação entre elas (ver, por exemplo, [134, 154, 137]).

Com o conteúdo deste capítulo pretende-se colmatar a falta de análises comparativas entre diferentes metodologias de análise espectral de sequências simbólicas e validar a aplicação de metodologias numéricas a dados simbólicos.

Dado o particular interesse por aplicações sobre dados genéticos, para discussão considera-se o caso das sequências de DNA, descritas pelo alfabeto dos nucleótidos ($\Gamma = \{\mathcal{A}, \mathcal{C}, \mathcal{G}, \mathcal{T}\}$). Contudo, a análise comparativa a apresentar pode ser aplicada e/ou alargada a outros alfabetos finitos.

3.1.1 Conceitos básicos

De seguida são apresentadas definições e propriedades da transformada de Fourier discreta e autocorrelação, dois dos conceitos fundamentais para a análise em frequência de sequências numéricas. Como se verá, mediante certas adaptações, continuam a ser úteis para sequências simbólicas.

3.1.1.1 Transformada de Fourier discreta

A transformada de Fourier discreta (Discrete Fourier Transform — DFT) $X = (X_j)_{0 \leq j < N}$ de uma sequência numérica $x = (x_k)_{0 \leq k < N}$ [1] é dada por

$$X = Fx,$$

onde F é a matriz de Fourier $N \times N$ com elementos

$$F_{ab} = e^{-i\frac{2\pi}{N}ab}, \qquad a, b = 0, 1, ..., N-1,$$

e i representa a unidade imaginária. De outra forma tem-se que a DFT ou coeficiente espectral de x, na frequência a/N, é dada por

$$X_a = \sum_{b=0}^{N-1} x_b e^{-i\frac{2\pi}{N}ab}, \qquad a = 0, 1, ..., N-1.$$

Observe-se que

$$X_0 = N\overline{x},$$

onde $\overline{x}$ representa o valor médio da sequência x.

A transformada inversa é dada por

$$x_c = \frac{1}{N} \sum_{d=0}^{N-1} X_d e^{i\frac{2\pi}{N}cd}, \qquad c = 0, 1, ..., N-1.$$

O produto interno entre dois vectores x e y de dimensão N é definido por

$$\langle x, y \rangle = \sum_{k=0}^{N-1} \bar{x}_k y_k = x'y,$$

onde $\bar{x}_k$ representa o conjugado do complexo x_k e x' é o conjugado transposto do vector x. Tem-se

$$\langle x, y \rangle = \frac{1}{N} \langle X, Y \rangle .$$

Em particular tem-se a igualdade de Parseval,

$$\langle x, x \rangle = \|x\|^2 = \frac{1}{N}\|X\|^2.$$

[1] É de notar que as sequências são entendidas como vectores coluna.

Seja $y_k = x_{k+m}$, para $0 \leq k, m < N$ e considerando extensões módulo N de x,[2] tem-se que

$$Y_a = X_a e^{i\frac{2\pi}{N}ma}, \qquad a = 0, 1, ..., N-1. \tag{3.1}$$

A DFT é uma ferramenta poderosa no estudo da periodicidade. Para estudar a periodicidade da sequência numérica $x = (x_k)_{0\leq k<N}$, é frequentemente usado o espectro de potência $(|X_a|^2)_{0\leq a<N}$,[3] ou simplesmente a sequência das amplitudes $(|X_a|)_{0\leq a<N}$.

3.1.1.2 Autocorrelação

A autocorrelação de $x = (x_k)_{0\leq k<N}$ é a correlação entre $(x_k)_{0\leq k<N}$ e $(x_{k+m})_{0\leq k<N}$, com $0 \leq m < N$ e considerando a extensão módulo N de x. A função de autocorrelação de x é usualmente definida por

$$r_m^{\diamond} = \frac{\sum_{j=0}^{T}(x_j - \overline{x})(x_{j+m} - \overline{x})}{\sum_{j=0}^{T}(x_j - \overline{x})^2}.$$

Frequentemente T é considerado igual a $N - m - 1$ ou a $N - 1$. Neste último caso, a autocorrelação designa-se por autocorrelação cíclica.

A autocorrelação é tão mais forte quanto mais próximo o seu valor absoluto estiver de 1 ($r_m^{\diamond}$ varia entre -1 e 1).

Por ser uma função de correlação muito usada no contexto do estudo de sequências simbólicas, a função de autocorrelação (ou autocovariância)[4] cíclica em sequências

[2]Por extensões módulo N de x entende-se

$$x_{k+N} \equiv x_k, \ \forall k \in \{0, \ldots, N-1\}.$$

[3]Também é comum encontrar como sinónimo de espectro de potência os termos: energia e potência espectral.

[4]A autocovariância é vulgarmente definida por

$$c_m = \frac{1}{T}\sum_{j=0}^{T}(x_j - \bar{x})(x_{j+m} - \bar{x}),$$

simbólicas será também aqui considerada. Esta função é definida como uma forma simplificada de $r_m^{\diamond}$ e é dada por

$$r_m = \sum_{k=0}^{N-1} x_k x_{k+m},$$

considerando extensões módulo N de x, $m \in \{0, 1, \ldots, N-1\}$ e onde x é uma sequência numérica, frequentemente uma sequência de zeros e uns.[5] É de notar que neste caso esta função de autocorrelação não subtrai o valor médio.

A partir de agora, por função de autocorrelação cíclica dever-se-á entender r_m. Seguem-se algumas propriedades da função de autocorrelação cíclica em sequências numéricas:

- A função de autocorrelação cíclica é uma função par, pois $r_m = r_{N-m}$.
- Considerando $x = (x_k)_{0 \leq k < N}$ e $y = (x_{k+m})_{0 \leq k < N}$, com $0 \leq m < N$, tem-se que

$$r_m = \sum_{k=0}^{N-1} x_k x_{k+m} = \langle x, y \rangle = \frac{1}{N} \langle X, Y \rangle \tag{3.2}$$

  onde X e Y são as DFT's de x e y respectivamente.

- A média dos quadrados, $\overline{x^2}$, pode ser determinada indirectamente através da autocorrelação: $r_0 = N\overline{x^2}$. Para uma sequência de zeros e uns tem-se que $\overline{x^2} = \overline{x}$.
- A função de autocorrelação atinge o valor máximo para $m = 0$. Esse valor pode ser atingido para outros valores de m quando a função é periódica. Assim, $r_0 \geq |r_m|$.[6]

onde $T = N - 1$ no caso cíclico, caso contrário $T = N - m - 1$.

[5] Para uma sequência de zeros e uns, a média ($\bar{x}$), é a proporção de uns da sequência e a variância da sequência é dada por $\bar{x}(1 - \bar{x})$. Deste modo

$$r_m = N(r_m^{\diamond}\bar{x}(1 - \bar{x}) + \bar{x}^2).$$

[6] Isto é uma consequência da desigualdade de Cauchy-Schwarz,

$$|\langle x, y \rangle| \leq \|x\| \|y\|.$$

Obtém-se igualdade quando x é proporcional a y: $x = \alpha y$, para α constante.

- A autocorrelação de uma sequência periódica é também uma sequência periódica com o mesmo período.

Outras considerações

Genericamente, uma sequência simbólica definida sobre o alfabeto Γ será denotada por $s = (s_k)_{0 \leq k < N}$. Sempre que necessárias serão usadas extensões módulo N de s.

3.2 Comparação de metodologias

De seguida efectua-se uma revisão comparativa dos métodos de análise das sequências simbólicas através da análise de Fourier, concretamente em aplicações a sequências de nucleótidos. Assim apresenta-se uma perspectiva unificadora de algumas ferramentas de análise de Fourier usadas na análise de sequências de DNA nomeadamente, métodos baseados em sequências indicadoras, autocorrelação, envolvente espectral e na estrutura do tetraedro regular.

3.2.1 Sequências indicadoras

A qualquer sequência simbólica $(s_k)_{0 \leq k < N}$ definida sobre o alfabeto $\Gamma = \{\mathcal{A}, \mathcal{C}, \mathcal{G}, \mathcal{T}\}$ podem ser associadas quatro sequências indicadoras numéricas, ou seja quatro sequências binárias u_k^a, u_k^c, u_k^g, e u_k^t, com $0 \leq k < N$. Cada uma destas sequências identifica a posição do símbolo correspondente através da unidade, isto é,

$$u_k^a = \begin{cases} 1, & s_k = \mathcal{A} \\ 0, & s_k \neq \mathcal{A} \end{cases}, \tag{3.3}$$

$$u_k^c = \begin{cases} 1, & s_k = \mathcal{C} \\ 0, & s_k \neq \mathcal{C} \end{cases},$$

$$u_k^g = \begin{cases} 1, & s_k = \mathcal{G} \\ 0, & s_k \neq \mathcal{G} \end{cases},$$

$$u_k^t = \begin{cases} 1, & s_k = \mathcal{T} \\ 0, & s_k \neq \mathcal{T} \end{cases},$$

com $0 \leq k < N$.

O espectro total das sequências simbólicas é neste caso definido como a soma dos quadrados das DFTs das sequências indicadoras (ver, por exemplo, [151, 12]), ou seja,

$$S(j) = |U_j^a|^2 + |U_j^c|^2 + |U_j^g|^2 + |U_j^t|^2 \tag{3.4}$$

com $j \in \{0, 1, \ldots, N-1\}$. Este espectro total é usado muitas vezes com pouca ou nenhuma explicação. Intuitivamente, a solução parece razoável: o espectro total como soma dos espectros de cada uma das sequências indicadoras. Desta forma não é necessário definir operações algébricas sobre os símbolos e não é necessário definir mapeamentos de símbolos para números. Contudo, as interpretações teóricas e significado desta solução não parecem ser claros.

Com os métodos de autocorrelação pode-se obter uma visão satisfatória, pelo menos mais intuitiva, do que a dada pelas sequências indicadoras. Mostra-se de seguida que a formulação de espectro total (3.4) resulta da transformada de Fourier da autocorrelação simbólica.

3.2.2 Autocorrelação simbólica

Claramente que o modo mais simples e mais directo de obter uma sequência numérica a partir da sequência de DNA é mapear os símbolos em números para posteriormente processar a sequência numérica obtida. Assim, pode-se aplicar indirectamente, sobre as sequências simbólicas, a transformada de Fourier ou o cálculo da autocorrelação (neste caso numérica). Em [153, 92], por exemplo, define-se a sequência numérica, seguida da autocorrelação e sobre o resultado a transformada de Fourier. Isto tem desvantagens, pois consoante o mapeamento de símbolos em números pode-se revelar ou esconder alguma informação. Além disso, não existe um claro significado bioquímico para a estrutura aritmética resultante do mapeamento.

No entanto, a função de autocorrelação pode ser definida directamente da sequência

dos símbolos, $(s_k)_{0\leq k<N}$, bastando considerar

$$r_k = \sum_{j=0}^{N-1} d(s_j, s_{j+k}),$$

onde para quaisquer dois símbolos x e y

$$d(x,y) = \begin{cases} 1, & x = y \\ 0, & x \neq y. \end{cases} \tag{3.5}$$

Observe-se que da forma como a autocorrelação está definida, o resultado da sua aplicação sobre uma sequência de símbolos é uma sequência numérica. Por outro lado, tem-se que esta função é simétrica em relação ao centro da sequência, isto é:

- para N par tem-se $r_{i+\frac{N}{2}} = r_{-i+\frac{N}{2}},\ i \in \{0,\ldots,\frac{N}{2}\}$;[7]
- para N ímpar tem-se $r_{i+\frac{N+1}{2}} = r_{-i+\frac{N+1}{2}},\ i \in \{0,\ldots,\frac{N+1}{2}\}$.

Esta função de autocorrelação resulta intuitivamente numa melhor solução e além disso evita-se o mapeamento. A função de autocorrelação assim definida está relacionada com a medida de correlação para a igualdade dos símbolos introduzida por [154].

É de observar que a autocorrelação r_k é uma sequência numérica que pode ser reescrita à custa das quatro funções indicadoras, uma vez que

$$u_k^a = d(s_k, \mathcal{A}),\ u_k^c = d(s_k, \mathcal{C}),\ u_k^g = d(s_k, \mathcal{G}) \text{ e } u_k^t = d(s_k, \mathcal{T}).$$

Consequentemente,

$$r_k = r_k(u^a) + r_k(u^c) + r_k(u^g) + r_k(u^t),$$

[7]

$$\begin{aligned} r_{-i+\frac{N}{2}} &= \sum_{j=0}^{N-1} d(s_j, s_{j-i+\frac{N}{2}}) = \sum_{j=0}^{N-1} d(s_{j-i+\frac{N}{2}}, s_j) \\ &= \sum_{k=0}^{N-1} d(s_k, s_{k+i-\frac{N}{2}}) = \sum_{k=0}^{N-1} d(s_k, s_{k+i+\frac{N}{2}}) \\ &= r_{i+\frac{N}{2}}. \end{aligned}$$

onde

$$r_k(u^a) = \sum_{j=0}^{N-1} u_j^a u_{j+k}^a$$

e de modo semelhante para $r_k(u^c)$, $r_k(u^g)$ e $r_k(u^t)$.

Teorema 3.2.1. *Seja $(R_j)_{0\leq j<N}$ a DFT de $(r_k)_{0\leq k<N}$. Então tem-se que*

$$R_j = |U_j^a|^2 + |U_j^c|^2 + |U_j^g|^2 + |U_j^t|^2,\ j \in \{0, \ldots, N-1\}.$$

Demonstração. Aplicando as igualdades (3.2) e (3.1), tem-se

$$r_k(u^a) = \frac{1}{N} \sum_{j=0}^{N-1} |U_j^a|^2 e^{i\frac{2\pi}{N}kj}$$

e obviamente vem que

$$r_k = \frac{1}{N} \sum_{j=0}^{N-1} (|U_j^a|^2 + |U_j^c|^2 + |U_j^g|^2 + |U_j^t|^2) e^{i\frac{2\pi}{N}kj}$$

ou seja,

$$R_j = |U_j^a|^2 + |U_j^c|^2 + |U_j^g|^2 + |U_j^t|^2,\ j \in \{0, \ldots, N-1\}.$$

□

Concluiu-se que a DFT da autocorrelação simbólica é a soma dos módulos quadráticos das DFTs das sequências indicadoras. Por outras palavras, obtém-se a igualdade com o espectro total definido à custa das sequências indicadoras, ver equação (3.4).

3.2.3 Envolvente espectral

Nesta secção verifica-se que a soma dos quadrados das DFTs das quatro sequências indicadoras também está relacionado com o valor da DFT da sequência numérica, que resulta do mapeamento de símbolos em números, construída de forma a maximizar a magnitude quadrática em cada valor da frequência.

O conceito de envolvente espectral foi introduzido em [137], e em [155] foi discutida uma variante para dados não estacionários. Por ser uma abordagem mais simples, segue-se inicialmente o trabalho desenvolvido em [155], considerando o caso estacionário

(Envolvente espectral – Abordagem I). A relação entre os trabalhos [137] e [155] será discutida no final desta subsecção (Envolvente espectral – Abordagem II).

O mapeamento dos nucleótidos em números

$$\mathcal{A} \mapsto a,\ \mathcal{C} \mapsto c,\ \mathcal{G} \mapsto g \text{ e } \mathcal{T} \mapsto t$$

pode descrever-se por

$$z = uw$$

onde

$$u = \left[\begin{array}{cccc} u^a & u^c & u^g & u^t \end{array} \right]$$

uma matriz $N \times 4$ e

$$w = \left[\begin{array}{c} a \\ c \\ g \\ t \end{array} \right]$$

um vector de pesos.

Considere-se a DFT de z:

$$Z = Fz = Fuw = Uw,$$

onde U é uma matriz $N \times 4$ obtida por aplicação das DFTs sobre as sequências indicadoras,

$$U = Fu = \left[\begin{array}{cccc} Fu^a & Fu^c & Fu^g & Fu^t \end{array} \right] = \left[\begin{array}{cccc} U^a & U^c & U^g & U^t \end{array} \right].$$

Denotando por U_j a j-ésima linha de U, pode-se escrever

$$Z_j = U_j w,\ \text{para } j \in \{0, \ldots, N-1\}.$$

Tem-se ainda

$$|Z_j|^2 = w' U_j' U_j w = |aU_j^a + cU_j^c + gU_j^g + tU_j^t|^2,\ j \in \{0, \ldots, N-1\}. \tag{3.6}$$

3.2.3.1 Envolvente espectral – Abordagem I

A ideia sublinhada por [137] e [155] é a de ajustar um mapeamento de símbolos para números de tal forma que $|Z_j|$ seja máximo. Para cada frequência j/N, selecciona-se o vector w normalizado ($\|w\| = 1$) que maximiza $|Z_j|^2$. Isto é, tem-se o seguinte problema de maximização

$$\max_{\|w\|=1} |Z_j|^2 = \max_{\|w\|=1} (w' U_j' U_j w),\ j \in \{0, \ldots, N-1\}.$$

Observe-se que para $\|w\| = 1$ tem-se que $w' U_j' U_j w$ coincide com um quociente de Rayleigh.[8] Os quocientes de Rayleigh de uma matriz estão intimamente relacionados com os valores próprios[9] dessa matriz (ver teorema de Rayleigh–Ritz [80]), o que permite concluir o seguinte resultado.

Teorema 3.2.2.

$$\max_{\|w\|=1} (w' U_j' U_j w) = \lambda_{\max}(U_j' U_j),\ j \in \{0, \ldots, N-1\}.$$

Demonstração. Observe-se que $U_j' U_j$ é uma matriz hermítica ($(U_j' U_j)' = U_j' U_j$) e que $w' U_j' U_j w$ é um quociente de Rayleigh uma vez que $\|w\| = 1$. Usando a consequência do teorema de Rayleigh–Ritz enunciada no teorema 3.2.3 o resultado fica evidente.

Teorema 3.2.3. *Seja A uma matriz hermítica. O máximo do quociente de Rayleigh,*

$$R(w) = \frac{x' A x}{x' x}, \qquad x \neq 0$$

é $\lambda_{\max}(A)$, o maior valor próprio de A.

□

[8] O quociente de Rayleigh, R, de uma matriz A, é definido por

$$R(x) = \frac{x' A x}{x' x}, \qquad x \neq 0,$$

com A uma matriz hermítica ($A' = A$).

[9] Um vector v não nulo (de um espaço vectorial) diz-se vector próprio de uma matriz A se existir um escalar λ tal que

$$Av = \lambda v.$$

Chama-se a λ o valor próprio associado ao vector próprio v.

Assim conclui-se que

$$\max_{||w||=1} |Z_j|^2 = \lambda_{\max}(U_j' U_j).$$

O resultado seguinte apresenta outra interpretação do máximo de $|Z_j|^2$.

Teorema 3.2.4. *Seja $f(w') = w' U_j' U_j w$, com $j \in \{0, \ldots, N-1\}$ e onde w é um vector normalizado. O ponto de máximo de f é $\left(\frac{U_j}{\|U_j\|},\ R_j\right)$.*

Demonstração. Tem-se que

$$f(w') = w' U_j' U_j w = |U_j w|^2 = |a U_j^a + c U_j^c + g U_j^g + t U_j^t|^2.$$

Pela desigualdade de Cauchy-Schwarz tem-se que

$$f(w') \leq ||w'||^2 ||U_j||^2$$

e o máximo é atingido para $w' = \alpha U_j$, com α constante. Como w é normalizado tem-se que

$$w' = \frac{U_j}{\|U_j\|}.$$

O valor máximo é dado por

$$\begin{aligned} \max_{||w||=1} |U_j w|^2 &= \max_{||w||=1} w' U_j' U_j w \\ &= \frac{U_j}{\|U_j\|} U_j' U_j \left(\frac{U_j}{\|U_j\|}\right)' \\ &= |U_j^a|^2 + |U_j^c|^2 + |U_j^g|^2 + |U_j^t|^2. \end{aligned}$$

□

Como resultado dos dois teoremas anteriores tem-se que

$$\lambda_{\max}(U_j' U_j) = |U_j^a|^2 + |U_j^c|^2 + |U_j^g|^2 + |U_j^t|^2 = R_j.$$

É de observar que no caso não estacionário, considerado em [155], a matriz $U_j' U_j$ é calculada para diversos blocos de dados. Assim, o cálculo dos maiores valores próprios ainda envolve alguma complexidade computacional. Com esta interpretação, pode-se reduzir o cálculo de valores próprios ao cálculo do espectro total de cada segmento.

Mapeamento genérico

Existem vários trabalhos em que também podem ser encontrados vários mapeamentos de símbolos em números (ver, por exemplo, [11]). De seguida, comparam-se os resultados até agora obtidos com o espectro da sequência numérica obtida de forma geral por mapeamento dos símbolos em números. Resulta da aplicação da desigualdade de Cauchy-Schwarz à expressão dada em (3.6) a seguinte desigualdade

$$\begin{aligned} |Z_j|^2 &= |aU_j^a + cU_j^c + gU_j^g + tU_j^t|^2 \leq \\ &\leq (|a|^2 + |c|^2 + |g|^2 + |t|^2)(|U_j^a|^2 + |U_j^c|^2 + |U_j^g|^2 + |U_j^t|^2). \end{aligned}$$

Para este mapeamento genérico não parece existir relação de igualdade entre os espectros. Conclui-se, apenas, a existência de um majorante para o espectro da sequência obtida por mapeamento arbitrário entre símbolos e números.[10]

No entanto, como já foi escrito na demonstração do teorema 3.2.4 a igualdade obtém-se para $(a, c, g, t) = \alpha(U_j^a, U_j^c, U_j^g, U_j^t)$, com α constante. E se se considerar $\alpha = \dfrac{1}{\|U_j\|}$ tem-se o espectro total

$$|Z_j|^2 = R_j.$$

3.2.3.2 Envolvente espectral – Abordagem II

O conceito de envolvente espectral, introduzido em [137], depende do conceito de valor próprio generalizado. No sentido de continuar a discussão introduz-se a seguinte notação: $\lambda(B, C)$ denota o valor próprio generalizado relativo ao problema $Bx = \lambda Cx$ (para denotar o valor próprio de B continua-se a usar a notação $\lambda(B)$ e naturalmente tem-se que $\lambda(B, I) = \lambda(B)$).

Nesta abordagem calculam-se os pesos w que maximizam a potência espectral para cada frequência, relativamente à variabilidade total, V, de uma sequência definida

[10] É de notar, que já anteriormente foi observado que consoante o mapeamento de símbolos em números utilizado, algumas das estruturas com relevância harmónica podem ser mais ou menos evidenciadas (ver, no capítulo anterior, o exemplo 2.1.1).

num alfabeto de n símbolos (será concretizado $n = 4$ dado o objectivo de comparar metodologias). Portanto, o objectivo é encontrar o valor máximo de

$$\frac{|Z_j|^2}{w'Vw} = \frac{w'U_j'U_jw}{w'Vw}, \tag{3.7}$$

com V a matriz de covariâncias de uma distribuição multinomial.[11]

Este problema pode-se reduzir ao cálculo do valor próprio generalizado máximo, pois pretende-se calcular o maior α tal que

$$\frac{w'U_j'U_jw}{w'Vw} = \alpha$$

ou seja, pretende-se calcular $\lambda_{\max}(U_j'U_j, V)$ (consequência do teorema de Rayleigh–Ritz).

Levanta-se a questão sobre a igualdade de valores próprios máximos, ou seja,

$$\lambda_{\max}(U_j'U_j, V) =^{?} \lambda_{\max}(U_j'U_j)$$

A igualdade anterior verificar-se-ia se V fosse a matriz identidade, mas V é uma matriz de covariâncias de uma distribuição multinomial o que elimina essa possibilidade.[12]

Segue-se um resultado que apresenta duas desigualdades que relacionam os dois valores próprios máximos.

Teorema 3.2.5. *Seja* $A = \begin{pmatrix} I_3 \\ 0 \end{pmatrix}$, *onde* I_3 *é a matriz identidade* 3×3. *Tem-se que*

$$\frac{\lambda_{\max}(U_j'U_j, V)}{\lambda_{\max}((A'VA)^{-T})} + |U_j{}^t|^2 \leq \lambda_{\max}(U_j'U_j) \leq \frac{\lambda_{\max}(U_j'U_j, V)}{\lambda_{\min}((A'VA)^{-T})} + |U_j{}^t|^2$$

[11] Seja z uma observação de uma variável multinomial de quatro símbolos $\mathcal{A}$, $\mathcal{C}$, $\mathcal{G}$ e $\mathcal{T}$ com probabilidade de ocorrência P_a, P_c, P_g e P_t respectivamente. A matriz de covariâncias é dada por

$$V = \begin{bmatrix} P_a(1-P_a) & -P_aP_c & -P_aP_g & -P_aP_t \\ -P_cP_a & P_c(1-P_c) & -P_cP_g & -P_cP_t \\ -P_gP_a & -P_gP_c & P_g(1-P_g) & -P_gP_t \\ -P_tP_a & -P_tP_c & -P_tP_g & P_t(1-P_t) \end{bmatrix}.$$

[12] Observe-se que para $w' = \dfrac{U_j}{\|U_j\|}$ geralmente a potência espectral dada por (3.7) não coincide com o espectro total da sequência ($S(j) = \lambda_{\max}(U_j'U_j)$).

Demonstração. Para relacionar os valores próprios generalizados com os valores próprios de uma matriz existe o teorema de Ostrowski [79] que apresenta duas desigualdades:

Teorema 3.2.6. *Dadas duas matrizes B e C, tal que $C^{-T}(=(C')^{-1})$ está definida,*

$$\lambda_{\max}(B)\lambda_{\min}(C^{-T}) \leq \lambda_{\max}(B,C) \leq \lambda_{\max}(B)\lambda_{\max}(C^{-T}).$$

Coloca-se o problema de aplicabilidade do teorema 3.2.6 a $\lambda_{\max}(U_j'U_j, V)$, uma vez que V não é uma matriz invertível. Esse problema pode ser parcialmente ultrapassado com o desenvolvimento apresentado em [137]:

$$\max \frac{w'U_j'U_jw}{w'Vw} = \max \frac{y'A'U_j'U_jAy}{y'A'VAy}$$

para $w = Ay + te_4$, $y = \begin{bmatrix} a \\ c \\ g \end{bmatrix}$ e $e_4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$.

Ou seja,

$$\lambda_{\max}(U_j'U_j, V) = \lambda_{\max}(A'U_j'U_jA, A'VA),$$

onde $A'VA$ é uma matriz invertível.

Aplicando o teorema 3.2.6 a $\lambda_{\max}(A'U_j'U_jA, A'VA)$, obtém-se

$$\lambda_{\max}(A'U_j'U_jA)\lambda_{\min}((A'VA)^{-T}) \leq \lambda_{\max}(U_j'U_j, V) \leq \lambda_{\max}(A'U_j'U_jA)\lambda_{\max}((A'VA)^{-T}).$$

Dado que

$$\lambda_{\max}(A'U_j'U_jA) = \lambda_{\max}(U_j'U_j) - |{U_j}^t|^2,$$

tem-se

$$0 \leq \lambda_{\min}((A'VA)^{-T}) \leq \frac{\lambda_{\max}(U_j'U_j, V)}{\lambda_{\max}(U_j'U_j) - |{U_j}^t|^2} \leq \lambda_{\max}((A'VA)^{-T})$$

ou seja,

$$\frac{\lambda_{\max}(U_j'U_j, V)}{\lambda_{\max}((A'VA)^{-T})} \leq \lambda_{\max}(U_j'U_j) - |{U_j}^t|^2 \leq \frac{\lambda_{\max}(U_j'U_j, V)}{\lambda_{\min}((A'VA)^{-T})}.$$

□

Desta forma relacionam-se os coeficientes do espectro total das sequências simbólicas, $\lambda_{\max}(U_j'U_j) = R_j$, como os coeficientes propostos por Stoffer,

$$\lambda_{\max}(U_j'U_j, V).$$

3.2.4 Redução da dimensão

Em [134] é apresentada ainda outra forma de ver o espectro total, a qual é baseada na estrutura de um tetraedro regular. Em [45] encontra-se uma generalização para sequências simbólicas num alfabeto de n símbolos e neste caso é usada a noção topológica de simplexo (generalização do conceito de triângulo).

As quatro sequências indicadoras são obviamente redundantes, uma vez que

$$u^a + u^c + u^g + u^t = 1, \tag{3.8}$$

o que naturalmente tem consequência sobre o espectro total. Uma vez que é válida a igualdade (3.8), então em termos de transformada de Fourier tem-se

$$U_j^a + U_j^c + U_j^g + U_j^t = \begin{cases} N, & j = 0 \\ 0, & j \neq 0 \end{cases}.$$

O espectro total pode ser obtido a partir de três DFTs, em vez de quatro. De facto, é possível trabalhar com três sequências não redundantes (x, y e z) em vez de quatro (u^a, u^c, u^g e u^t) (ver, por exemplo, [12, 108, 134]). Embora possa ser usado outro, um mapeamento em vectores com apenas três componentes que não perde informação da sequência de símbolos é (ver, por exemplo, [12])

$$\begin{aligned}
\mathcal{A} &\mapsto \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}, \\
\mathcal{C} &\mapsto \begin{pmatrix} -\frac{\sqrt{2}}{3} & \frac{\sqrt{6}}{3} & -\frac{1}{3} \end{pmatrix}, \\
\mathcal{G} &\mapsto \begin{pmatrix} -\frac{\sqrt{2}}{3} & -\frac{\sqrt{6}}{3} & -\frac{1}{3} \end{pmatrix}, \\
\mathcal{T} &\mapsto \begin{pmatrix} \frac{2\sqrt{2}}{3} & 0 & -\frac{1}{3} \end{pmatrix}.
\end{aligned}$$

A ligação com as sequências indicadoras é dada por

$$x = \frac{\sqrt{2}}{3}(-u^c - u^g + 2u^t), \tag{3.9}$$

$$y = \frac{\sqrt{6}}{3}(u^c - u^g), \tag{3.10}$$

$$z = \frac{1}{3}(3u^a - u^c - u^g - u^t), \tag{3.11}$$

$$\mathbf{1} = u^a + u^c + u^g + u^t.$$

A relação entre o método que se baseia nas sequências indicadoras, o espectro total, e o que se baseia na redução da dimensão (método baseado na noção topológica de simplexo) é mostrada em [45], para um número arbitrário de símbolos. No caso presente das sequências de nucleótidos, tem-se o tetraedro como simplexo ($n = 4$) o que resulta na seguinte relação de igualdade:

Teorema 3.2.7. *Sejam X, Y e Z as DFT de $(x_k)_{0\leq k<N}$ (equação 3.9), $(y_k)_{0\leq k<N}$ (equação 3.10) e $(z_k)_{0\leq k<N}$ (equação 3.11), respectivamente. Então tem-se que*

$$\begin{aligned} &3(|U_j^a|^2 + |U_j^c|^2 + |U_j^g|^2 + |U_j^t|^2) = \\ &= \begin{cases} 4(|X_j|^2 + |Y_j|^2 + |Z_j|^2), & j \neq 0 \\ 4(|X_j|^2 + |Y_j|^2 + |Z_j|^2) - N, & j = 0 \end{cases}. \end{aligned}$$

Demonstração. Segue-se de perto a demonstração apresentada em [45], com algumas adaptações à notação usada e ao caso particular em estudo, sequência simbólica de quatro símbolos.

Considere-se

$$v = \begin{bmatrix} x \\ y \\ z \\ 0 \end{bmatrix}, \; R = \begin{bmatrix} 0 & -\frac{\sqrt{6}}{6} & -\frac{\sqrt{6}}{6} & \frac{\sqrt{6}}{3} \\ 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 \\ \frac{\sqrt{3}}{2} & -\frac{\sqrt{3}}{6} & -\frac{\sqrt{3}}{6} & -\frac{\sqrt{3}}{6} \\ c & c & c & c \end{bmatrix},$$

c uma constante real não nula e

$$t = \frac{1}{4}\begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}.$$

Tem-se que

$$v = aR(u' - t'), \tag{3.12}$$

onde v é obtido de u por uma translação seguida de uma transformação R. Pretende-se que R seja uma rotação, e para tal tem-se $c = -\frac{3}{2}$ (R é uma rotação se o seu determinante é igual a um). Usa-se $a = \frac{2}{\sqrt{3}}$, escolhido de forma a que o simplexo (o tetraedro) seja normalizado (mapeamento dos símbolos em vectores normalizados).

Note-se que a informação dada pela última equação do sistema 3.12 em [45] é desprezada, pois simplesmente descreve que os vectores indicadores estão sobre o hiperplano

$$1 = u_k^a + u_k^c + u_k^g + u_k^t,$$

com $k \in \{0, \ldots, N-1\}$.

Recorrendo ao conceito de DFT tem-se que

$$V_j = \sum_{k=0}^{N-1} v'_k e^{-i\frac{2\pi}{N}kj}, \qquad j = 0, 1, ..., N-1$$

e o espectro total associado a v é dado por

$$|V_j|^2 = |X_j|^2 + |Y_j|^2 + |Z_j|^2.$$

Por outro lado, tem-se que

$$|V_j|^2 = \frac{4}{3}\left(|U_j^a|^2 + |U_j^c|^2 + |U_j^g|^2 + |U_j^t|^2\right) \text{ para } j \neq 0,$$

e para $j = 0$ tem-se

$$\begin{aligned} |V_0|^2 &= \frac{4}{3}\left(|U_0^a|^2 + |U_0^c|^2 + |U_0^g|^2 + |U_0^t|^2\right) - \frac{4}{3}N\left(\frac{1}{4^2} + \frac{1}{4^2} + \frac{1}{4^2} + \frac{1}{4^2}\right) \\ &= \frac{4}{3}\left(|U_0^a|^2 + |U_0^c|^2 + |U_0^g|^2 + |U_0^t|^2\right) - \frac{N}{3}. \end{aligned}$$

□

Assim obtém-se outra visão do espectro total relativo às sequências simbólicas.

3.3 Conclusões

Foram discutidos vários métodos de análise de Fourier para dados simbólicos, dando ênfase ao caso das sequências de DNA para o alfabeto dos nucleótidos. No entanto, as metodologias são extensíveis a qualquer outro alfabeto finito.

Consideraram-se métodos de análise espectral baseados nas sequências indicadoras, na estrutura do tetraedro, na autocorrelação simbólica e vectorial e métodos semelhantes à análise espectral, em que para cada frequência se optimiza o mapeamento de símbolos para números, de forma a dar ênfase a qualquer característica periódica. Foi discutida a relação entre os métodos e identificados os casos de equivalência entre os mesmos.

Mostrou-se que é possível definir a função de autocorrelação de dados simbólicos sem recorrer a mapeamentos: basicamente só se precisa de reconhecer se dois símbolos são ou não iguais. A autocorrelação simbólica é uma sequência numérica, e a sua transformada de Fourier resulta no espectro de dados simbólicos. Por outro lado, mostrou-se que este espectro pode ser obtido através da soma dos quadrados dos módulos das transformadas de Fourier das sequências indicadoras (sequências zero/um indicando a posição dos símbolos), solução que tem sido usada por diversos autores mas sem qualquer justificação rigorosa ou intuitiva.

Foi explorado o conceito de envolvente espectral, que dá outra interpretação do espectro: entre todos os mapeamentos entre símbolos e números há um que maximiza a energia espectral para cada frequência e que também conduz ao espectro de dados simbólicos.

Por fim, apresentou-se a relação entre a potência espectral das quatro sequências indicadoras e a potência espectral de três sequências que reduzem a redundância existente entre as quatro sequências indicadoras, usando a estrutura vectorial de um tetraedro regular.

De todos os método estudados, o que usa as sequências indicadoras parece ser o mais simples, em que o espectro total pode ser entendido como o espectro de máxima energia espectral ou como a transformada da autocorrelação simbólica.

Capítulo 4

Distribuição dos símbolos e espectro da sequência

Neste capítulo estuda-se a relação entre o módulo dos coeficientes espectrais das sequências de nucleótidos, ou de qualquer outra sequência simbólica, e a distribuição dos símbolos ao longo da sequência. Determinam-se condições necessárias e suficientes para que alguns módulos dos coeficientes espectrais sejam zero. Apresenta-se um algoritmo rápido de cálculo de elementos do espectro de potência de uma sequência simbólica. Finalmente mostra-se que o espectro de uma sequência de símbolos contém informação redundante.

Parte dos resultados discutidos neste capítulo foram alvo da publicação [5].

4.1 Introdução

Considere-se novamente uma sequência simbólica, s, de tamanho N com elementos

$$s_0, s_1, \ldots, s_{N-1},$$

onde cada símbolo s_i pertence a um alfabeto finito Γ. Embora os resultados discutidos neste capítulo sejam extensíveis a outros alfabetos, apenas se apresentam resultados relativos ao alfabeto dos nucleótidos $\Gamma = \{\mathcal{A}, \mathcal{C}, \mathcal{G}, \mathcal{T}\}$.

Há um grande interesse em métodos que detectem e revelem a estrutura do DNA (correlações de curto, médio e longo alcance). Tendo em conta os dados disponíveis, o seu volume e o interesse das aplicações, é obvia a motivação para a resolução deste problema.

Como se viu no capítulo anterior, é possível utilizar a análise de Fourier sobre as sequências de DNA. Uma das soluções é baseada nas sequências indicadoras de s, u^a, u^c, u^g e u^t (ver (3.3)). O espectro de Fourier ou espectro total da sequência s é definido em termos do espectro individual das quatro sequências indicadoras [118, 12, 153, 142, 154], como já foi discutido anteriormente (ver capítulo 3). Recordando, $S(k)$ $(0 \leq k < N)$ pode ser dado por

$$S(k) = |U_k^a|^2 + |U_k^c|^2 + |U_k^g|^2 + |U_k^t|^2, \tag{4.1}$$

com

$$U_k^b = \sum_{j=0}^{N-1} u_j^b e^{-i2\pi jk/N}, \qquad b \in \{a, c, g, t\}. \tag{4.2}$$

Um dos objectivos deste capítulo é compreender melhor o comportamento das riscas espectrais,[1] face à distribuição dos nucleótidos na sequência. Por exemplo, directamente da expressão dos coeficientes espectrais (ver (4.2)) resulta que:

- U_0^b, com $b \in \{a, c, g, t\}$, reflecte o número total de ocorrências do símbolo na sequência, uma vez que

$$U_0^b \;=\; \sum_{j=0}^{N-1} u_j^b,$$

- $U_{(N/2)}^b$, com $b \in \{a, c, g, t\}$, coincide com a diferença entre o número de ocorrências do símbolo nas posições pares e o número de ocorrências nas posições ímpares,

[1] Entenda-se por risca espectral em k, o módulo quadrático do coeficiente espectral em k. No caso particular das sequências simbólicas é dado por $S(k)$.

uma vez que

$$\begin{aligned} U^b_{(N/2)} &= \sum_{j=0}^{N-1} u^b_j e^{-i2\pi jN/(2N)} \\ &= \sum_{j=0}^{N-1} u^b_j e^{-i\pi j} \\ &= u^b_0 - u^b_1 + u^b_2 - u^b_3 + \ldots - u^b_{N-1}. \end{aligned}$$

É natural perguntar acerca das relações entre outros coeficientes espectrais e a distribuição de símbolos ao longo da sequência, e é precisamente essa questão que está na base dos resultados contidos neste capítulo.

Será apresentado um algoritmo rápido para calcular riscas espectrais baseado em contadores de símbolos. Concretamente, será discutida uma nova forma de cálculo dos coeficientes espectrais da forma $U^b_{(N/m)}$, com $b \in \{a, c, g, t\}$. Será explorada a relação entre esses coeficientes espectrais e a distribuição de frequência do símbolo $\mathcal{B}$ ao longo da sequência. Em particular, será discutida com maior destaque a risca $S(N/3)$ (risca com especial importância na localização de regiões codificantes).

Aqui também será discutida uma forma simples de obter condições necessárias e suficientes para que alguns valores do espectro total de uma sequência seja zero. Algumas contribuições foram dadas em [92], mas o trabalho a ser apresentado vai diferir deste em muitos aspectos.

Finalmente, concluiu-se que os coeficientes espectrais, U^b_k com $k \in \{0, 1, \ldots, N-1\}$, não são independentes. Por exemplo, se existirem n símbolos $\mathcal{B}$ ($\mathcal{B} \in \Gamma$) numa sequência de comprimento total N, então os N elementos de U^b podem ser determinados a partir de qualquer subconjunto contíguo de cardinalidade n, através de uma recursão linear.

A notação e terminologia usada neste capítulo é a mesma do capítulo anterior.

4.2 Contadores de símbolos

Suponha-se que é válida a factorização $N = nm$ e considere-se

$$y^b_{(v)} = \sum_{j=0}^{n-1} u^b_{(v+jm)}, \qquad \text{com } 0 \leq v < m \text{ e } b \in \{a, c, g, t\}. \tag{4.3}$$

Ou seja, divide-se a sequência original de comprimento N em n sequências de comprimento m

$$u^b = (\underbrace{u^b_1 u^b_2 \dots u^b_m}_{1} \dots \dots \dots \underbrace{u^b_{N-m+1} u^b_{N-m+2} \dots u^b_N}_{n}),$$

e de seguida, por adição agrupam-se os dados de acordo com a expressão (4.3) no sentido de obter uma sequência de comprimento m.

Observe-se que $y^b_{(v)}$ representa o número de ocorrências do símbolo $\mathcal{B}$ na v-ésima posição das n subsequências em que a sequência original se subdivide, ou seja o número de ocorrências de $\mathcal{B}$ nas posições

$$v, v + m, v + 2m, \dots, v + (n - 1)m,$$

das respectivas n subsequências.

As sequências y^a, y^c, y^g e y^t serão denominadas de contadores dos símbolos $\mathcal{A}$, $\mathcal{C}$, $\mathcal{G}$ e $\mathcal{T}$, respectivamente.

Segue-se um exemplo de uma sequência para a qual se apresentam os respectivos contadores de um símbolo.

Exemplo 4.2.1. *Considere-se a sequência*

$$s = (\mathcal{B}\mathcal{X}\mathcal{X}\mathcal{X}\mathcal{B}\mathcal{X}\mathcal{B}\mathcal{X}\mathcal{X}\mathcal{X}\mathcal{B}\mathcal{X})$$

com $\mathcal{B}, \mathcal{X} \in \Gamma$ *e onde* $\mathcal{X}$ *representa qualquer outro símbolo diferente de* $\mathcal{B}$*. Tem-se* $N = 12$ *e*

$$u^b = (1\ 0\ 0\ 0\ 1\ 0\ 1\ 0\ 0\ 0\ 1\ 0).$$

Os possíveis contadores do símbolo $\mathcal{B}$ *para a sequência dada são seis (ver colunas da tabela 4.1), pois as factorizações possíveis de* 12 *são* $12 \times 1 = 6 \times 2 = 4 \times 3 = 3 \times 4 = 2 \times 6 = 1 \times 12$ *(=nm).*

	$m=1$	$m=2$	$m=3$	$m=4$	$m=6$	$m=12$
$y^b_{(0)}$	4	4	2	2	2	1
$y^b_{(1)}$	–	0	2	0	0	0
$y^b_{(2)}$	–	–	0	2	0	0
$y^b_{(3)}$	–	–	–	0	0	0
$y^b_{(4)}$	–	–	–	–	2	1
$y^b_{(5)}$	–	–	–	–	0	0
$y^b_{(6)}$	–	–	–	–	–	1
$y^b_{(7)}$	–	–	–	–	–	0
$y^b_{(8)}$	–	–	–	–	–	0
$y^b_{(9)}$	–	–	–	–	–	0
$y^b_{(10)}$	–	–	–	–	–	1
$y^b_{(11)}$	–	–	–	–	–	0

Tabela 4.1: Exemplo de contadores de símbolos de uma sequência que apresenta como sequência indicadora do símbolo $\mathcal{B}$, $u^b = (1\ 0\ 0\ 0\ 1\ 0\ 1\ 0\ 0\ 0\ 1\ 0)$. As colunas correspondem aos seis possíveis contadores do símbolo b.

De seguida será estabelecida uma relação entre a DFT de comprimento m do contador de símbolos y^b e a DFT de u^b, com $b \in \{a, c, g, t\}$.

Teorema 4.2.1. *Seja $N = nm$. Considere-se k um inteiro fixo, tal que $0 \leq k < m$. Então:*

$$Y^b_{(k)} = U^b_{(nk)},$$

com $b \in \{a, c, g, t\}$.

Demonstração.

$$\begin{aligned} Y^b_{(k)} &= \sum_{j=0}^{m-1} y^b_{(j)} e^{-i2\pi jk/m} \\ &= \sum_{j=0}^{m-1}\sum_{u=0}^{n-1} u^b_{(j+um)} e^{-i2\pi jk/m} \\ &= \sum_{j=0}^{m-1}\sum_{u=0}^{n-1} u^b_{(j+um)} e^{-i2\pi (j+um)k/m} \\ &= \sum_{p=0}^{N-1} u^b_p e^{-i2\pi pk/m} \\ &= \sum_{p=0}^{N-1} u^b_p e^{-i2\pi pnk/N} = U^b_{(nk)}. \end{aligned}$$

A terceira igualdade deve-se à propriedade trigonométrica: $e^{i\theta+2k\pi} = e^{i\theta}$, $k \in \mathbb{Z}$. E a quinta igualdade é válida uma vez que $m = \frac{N}{n}$. □

A ideia de reduzir o cálculo da DFT de comprimento $N = mn$ ao cálculo da DFTs de comprimento m de somas de blocos tem por base [32], resultado parcialmente redescoberto em [33] (ver também [58]).

O teorema seguinte relaciona o espectro de potência de uma sequência simbólica (de comprimento N) e os módulos quadráticos das DFTs dos contadores de símbolos y^a, y^c, y^g e y^t (de comprimento N/n).

Teorema 4.2.2. *Seja $N = nm$. Considere-se k um inteiro fixo, tal que $0 \leq k < m$. Tem-se*

$$S(nk) = |Y^a_{(k)}|^2 + |Y^c_{(k)}|^2 + |Y^g_{(k)}|^2 + |Y^t_{(k)}|^2, \tag{4.4}$$

e $S(nk)$ anula-se se e só se $Y^a_{(k)}$, $Y^c_{(k)}$, $Y^g_{(k)}$ e $Y^t_{(k)}$ são todos nulos.

Demonstração. A equação (4.4) resulta imediatamente de (4.1)

$$S(nk) = |U^a_{(nk)}|^2 + |U^c_{(nk)}|^2 + |U^g_{(nk)}|^2 + |U^t_{(nk)}|^2$$

e da transformação descrita anteriormente $Y^b_{(k)} = U^b_{(nk)}$ com $b \in \{a, c, g, t\}$. Obviamente o coeficiente é nulo se e só se $Y^a_{(k)}$, $Y^c_{(k)}$, $Y^g_{(k)}$ e $Y^t_{(k)}$ forem todos nulos. □

O cálculo de $S(nk)$ é muito mais eficiente quando feito através dos contadores de símbolos, e portanto usando DFTs de comprimento N/n, do que quando é feito directamente, à custa de DFTs de comprimento N (confirmar com o exemplo 4.2.1).

4.2.1 Risca $S(N/3)$

Como já foi referido, a evidência de periodicidade de período três tem sido frequentemente usada na detecção de regiões codificantes nas sequências de DNA, sendo assim é óbvia a motivação para a procura de algoritmos rápidos de cálculo de $S(N/3)$.

Apresenta-se de seguida uma expressão para a risca espectral $S(N/3)$ como função dos contadores de símbolos.

Teorema 4.2.3. *Seja $N = 3n$. Então a risca espectral $S(N/3)$ é dada por*

$$S(N/3) = F(y^a) + F(y^c) + F(y^g) + F(y^t), \tag{4.5}$$

onde

$$F(y^b) = \left[y^b_{(0)} - \frac{y^b_{(1)} + y^b_{(2)}}{2} \right]^2 + \frac{3}{4}[y^b_{(1)} - y^b_{(2)}]^2, \tag{4.6}$$

com $b \in \{a, c, g, t\}$.

Demonstração. Tem-se que $n = N/3$. Em (4.4) considere-se $k = 1$ e obtém-se

$$S(N/3) = |Y^a_{(1)}|^2 + |Y^c_{(1)}|^2 + |Y^g_{(1)}|^2 + |Y^t_{(1)}|^2. \tag{4.7}$$

Tem-se também que

$$Y^b_{(1)} = y^b_{(0)} + y^b_{(1)} w + y^b_{(2)} w^2,$$

onde $w = e^{-i2\pi/3}$ e $b \in \{a, c, g, t\}$.

Logo

$$\begin{aligned}
|Y^b_{(1)}|^2 &= \left| y^b_{(0)} + y^b_{(1)} w + y^b_{(2)} w^2 \right|^2 \\
&= \left| y^b_{(0)} - y^b_{(1)} \frac{1}{2} - y^b_{(2)} \frac{1}{2} + i(y^b_{(1)} \frac{-\sqrt{3}}{2} + y^b_{(2)} \frac{\sqrt{3}}{2}) \right|^2 \\
&= \left[y^b_{(0)} - \frac{y^b_{(1)} + y^b_{(2)}}{2} \right]^2 + \frac{3}{4}[y^b_{(1)} - y^b_{(2)}]^2 \\
&= F(y^b).
\end{aligned}$$

□

Teorema 4.2.4. *Seja $N = 3n$. Então $S(N/3)$ é uma função simétrica dos contadores de símbolos.*

Demonstração. Para simplificar a notação, seja $x = y^b_{(0)}$, $y = y^b_{(1)}$ e $z = y^b_{(2)}$, com $b \in \{a, c, g, t\}$. Observe-se que

$$\begin{aligned} F(y^b) = F(x, y, z) &= \left[x - \frac{y+z}{2}\right]^2 + \frac{3}{4}[y-z]^2 \\ &= x^2 + y^2 + z^2 - xy - xz - yz, \end{aligned}$$

e que é invariante sobre as permutações de x, y e z, ou seja $F(x, y, z) = F(x, z, y) = F(z, x, y) = F(y, x, z) = F(z, y, x) = F(y, z, x)$. □

O número de operações aritméticas necessárias para calcular (4.5) é $O(1)$,[2] isto é, independente de N. O cálculo dos contadores de símbolos pode ser feito aquando da leitura dos dados, uma vez que apenas requer o incremento de contadores apropriados. O resultado final é um procedimento de cálculo aproximadamente $O(1)$, isto é, praticamente independente de N. Observe-se que o cálculo da FFT ("Fast Fourier Transform") de comprimento N é um processo $O(N \log N)$, e o cálculo de um coeficiente espectral requer $O(N)$ operações aritméticas.

No contexto da distribuição dos contadores introduz-se a seguinte terminologia: um contador de símbolos, y^b, é *uniformemente distribuído* nas m posições, quando $y^b_{(v_1)} = y^b_{(v_2)}$, $\forall_{v_1, v_2 \in \{0,\ldots,m-1\}}$ com $b \in \{a, c, g, t\}$ (ver exemplo 4.2.2).

Exemplo 4.2.2. *A seguinte sequência simbólica*

$$s = (\mathcal{AAA}\ \mathcal{CCC}\ \mathcal{CCC}\ \mathcal{GGG}\ \mathcal{TTT}\ \mathcal{ACG}\ \mathcal{GAC}\ \mathcal{CGA})$$

é uniformemente distribuída nas $m = 3$ posições, pois

$$\begin{aligned} y^a_{(0)} = y^a_{(1)} = y^a_{(2)} &= 2; \\ y^c_{(0)} = y^c_{(1)} = y^c_{(2)} &= 3; \\ y^g_{(0)} = y^g_{(1)} = y^g_{(2)} &= 2; \\ y^t_{(0)} = y^t_{(1)} = y^t_{(2)} &= 1. \end{aligned}$$

[2] Diz-se que $g = O(f)$ se existir uma constante c tal que $|g(x)| \leq c|f(x)|$ quando $x \to \infty$.

Observe-se que surge directamente de (4.6) uma condição necessária e suficiente para anular $S(N/3)$ (condição que, de outro modo, já foi estudada em [92]).

Teorema 4.2.5. *Seja $N = 3n$. A risca espectral $S(N/3)$ é zero se e só se todos os contadores de símbolos y^a, y^c, y^g e y^t forem uniformemente distribuídos.*

Demonstração. Usando (4.7) vem que:

$$\begin{aligned} S(N/3) = 0 &\Leftrightarrow |Y^a_{(1)}|^2 + |Y^c_{(1)}|^2 + |Y^g_{(1)}|^2 + |Y^t_{(1)}|^2 = 0 \\ &\Leftrightarrow |Y^a_{(1)}|^2 = |Y^c_{(1)}|^2 = |Y^g_{(1)}|^2 = |Y^t_{(1)}|^2 = 0 \end{aligned}$$

Por definição de DFT tem-se que

$$Y^b_{(1)} = y^b_{(0)} + y^b_{(1)} w + y^b_{(2)} w^2,$$

onde $w = e^{-i2\pi/3}$ e $b \in \{a, c, g, t\}$.

Por um lado, se a distribuição é uniforme ($y^b_{(0)} = y^b_{(1)} = y^b_{(2)}$) implica que $Y^b_{(1)} = 0$, uma vez que $1 + w + w^2 = 0$.

Por outro lado, supondo que $|Y^b_{(1)}|^2 = 0$, facilmente se mostra que $y^b_{(0)} = y^b_{(1)} = y^b_{(2)}$, uma vez que

$$|Y^b_{(1)}|^2 = \left[y^b_{(0)} - \frac{y^b_{(1)} + y^b_{(2)}}{2} \right]^2 + \frac{3}{4}[y^b_{(1)} - y^b_{(2)}]^2 = 0$$

o que implica

$$y^b_{(1)} = y^b_{(2)}$$

e

$$y^b_{(0)} = \frac{y^b_{(1)} + y^b_{(2)}}{2} = y^b_{(1)},$$

provando o que se pretendia. □

Para além da conclusão dada pelo teorema anterior, o método de cálculo de $S(N/3)$ baseado em (4.5) permite estabelecer relações entre a distribuição dos nucleótidos e o valor da risca espectral $S(N/3)$.

Uma vez que $S(N/3)$ é dado pela soma de quatro módulos quadrados (ver (4.1)), estudou-se isoladamente cada uma das parcelas $|U^b_k|^2 = |Y^b_{(1)}|^2$, com $b \in \{a, c, g, t\}$.

Teorema 4.2.6. *Faça-se coincidir os contadores de símbolos $y^b_{(0)}$, $y^b_{(1)}$ e $y^b_{(2)}$ com os três eixos ortogonais x, y, z, com $b \in \{a, c, g, t\}$. Assim, para $|U^b_{(N/3)}|^2 = v$ com v fixo, tem-se geometricamente que x, y e z estão sobre um cilindro com eixo $x = y = z$ e raio $r = \sqrt{2v/3}$.*

Demonstração. Tem-se que $F(y^b)$ $(= |Y^b_{(1)}|^2 = |U^b_{(N/3)}|^2)$ é dado por

$$\begin{aligned} F(x,y,z) &= \left[x - \frac{y+z}{2}\right]^2 + \frac{3}{4}[y-z]^2 \\ &= x^2 + y^2 + z^2 - xy - yz - xz \\ &= \frac{1}{2}\left[(x-y)^2 + (y-z)^2 + (z-x)^2\right], \end{aligned}$$

ou, de forma matricial, por

$$F(x,y,z) = \begin{bmatrix} x & y & z \end{bmatrix} \underbrace{\begin{bmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 1 & -1/2 \\ -1/2 & -1/2 & 1 \end{bmatrix}}_{A} \begin{bmatrix} x \\ y \\ z \end{bmatrix}.$$

Como A é uma matriz simétrica, então admite representação diagonal. Os elementos da diagonal coincidem com os valores próprios da matriz A: 3/2, 3/2, 0.

Seja

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \underbrace{\begin{bmatrix} -1 & -1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}}_{B} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix},$$

onde as colunas de B são os vectores próprios de A.

Reduzindo a forma quadrática à forma canónica (diagonal), vem que

$$\frac{3}{2}(X^2 + Y^2) = v. \tag{4.8}$$

A equação (4.8) reflecte a equação de um cilindro circular com raio $\sqrt{2v/3}$ e eixo

$$X = 0 \wedge Y = 0.$$

Como

$$\begin{cases} X = 0 \\ Y = 0 \end{cases} \Leftrightarrow \begin{cases} x = Z \\ y = Z \\ z = Z \end{cases},$$

o eixo do cilindro é $x = y = z$. □

Dada a importância que a risca $S(N/3)$ tem na detecção de genes, segue-se uma discussão sobre os valores de $S(N/3)$ a partir dos quais se pode afirmar que existe evidência de uma periodicidade de período três.[3] Assim, uma questão se levanta: Quando é que a risca espectral $S(N/3)$ é máxima ou significativamente elevada?

Dado que $F(y^b)$, com $b \in \{a, c, g, t\}$, se pode escrever como

$$F(y^b) = |U^b_{(N/3)}|^2 = \frac{1}{2}\left[(y^b_{(0)} - y^b_{(1)})^2 + (y^b_{(1)} - y^b_{(2)})^2 + (y^b_{(2)} - y^b_{(0)})^2\right], \tag{4.9}$$

e para α, β constantes reais tem-se

$$(\alpha - \beta)^2 \leq \alpha^2 + \beta^2,$$

onde

$$\max_{\alpha,\beta}\{(\alpha - \beta)^2\} = \max\{\alpha^2, \beta^2\},$$

então

$$0 \leq F(y^b) \leq \frac{1}{2}\left[(y^b_{(0)})^2 + (y^b_{(1)})^2 + (y^b_{(2)})^2\right] \leq \frac{3}{2}\left[\frac{N}{3}\right]^2.$$

Contudo, o interesse centra-se em encontrar o máximo valor para $F(y^b)$.

Teorema 4.2.7. *O máximo de* $|U^b_{(N/3)}|^2$ *é* $(N/3)^2$*, com* $b \in \{a, c, g, t\}$.

Demonstração. Como já foi referido anteriormente, $F(y^b) = v$ pode ser visto como um cilindro de eixo $y^b_{(0)} = y^b_{(1)} = y^b_{(2)}$ e de raio $\sqrt{2v/3}$. Como $0 \leq y^b_{(0)}, y^b_{(1)}, y^b_{(2)} \leq N/3$ é uma região que define um cubo, então $F(y^b) = v$ é um cilindro que passa pelo menos por um ponto do cubo e o eixo do cilindro é a diagonal do cubo que une os pontos $(0, 0, 0)$ e $\left(\frac{N}{3}, \frac{N}{3}, \frac{N}{3}\right)$. O objectivo passa a ser descobrir o cilindro de maior raio que

[3] O facto de $S(N/3)$ não ser nulo não é suficiente para garantir a evidência da periodicidade de período três, pois pode-se ter $S(N/3) \neq 0$ e não existir evidência desta risca relativamente às restantes.

esteja nas condições anteriormente descritas. Os pontos que estão mais distantes da diagonal referida são apenas seis, correspondendo aos restantes seis vértices do cubo (ver tabela 4.2), e estes estão à distância $\sqrt{\frac{2}{3}\left(\frac{N}{3}\right)^2}\left(v=\left(\frac{N}{3}\right)^2\right)$. Assim, concluiu-se que as seis situações representadas na tabela 4.2 são as únicas a garantir o máximo de $F(y^b)$, isto é, $(N/3)^2$.

$F(y^b)$	$y^b_{(0)}$	$y^b_{(1)}$	$y^b_{(2)}$
$(N/3)^2$	N/3	0	0
$(N/3)^2$	0	N/3	0
$(N/3)^2$	0	0	N/3
$(N/3)^2$	N/3	N/3	0
$(N/3)^2$	N/3	0	N/3
$(N/3)^2$	0	N/3	N/3

Tabela 4.2: Casos de máximo de $F(y^b)$, com $b \in \{a, c, g, t\}$.

□

No sentido de encontrar extremos para $S(N/3) = F(y^a) + F(y^c) + F(y^g) + F(y^t)$ e tendo em conta que as variáveis estão sujeitas a três condições:

$$y^a_{(0)} + y^c_{(0)} + y^g_{(0)} + y^t_{(0)} = N/3;$$
$$y^a_{(1)} + y^c_{(1)} + y^g_{(1)} + y^t_{(1)} = N/3;$$
$$y^a_{(2)} + y^c_{(2)} + y^g_{(2)} + y^t_{(2)} = N/3,$$

poderia-se pensar em aplicar o método dos multiplicadores de Lagrange, mas o resultado da aplicação do método resulta na seguinte solução degenerada (os multiplicadores todos nulos):

$$y^a_{(0)} = y^a_{(1)} = y^a_{(2)};$$
$$y^c_{(0)} = y^c_{(1)} = y^c_{(2)};$$
$$y^g_{(0)} = y^g_{(1)} = y^g_{(2)};$$
$$y^t_{(0)} = y^t_{(1)} = y^t_{(2)};$$

$$y^a_{(0)} + y^c_{(0)} + y^g_{(0)} + y^t_{(0)} = N/3.$$

O extremo obtido é um mínimo $S(N/3) = 0$ (distribuição uniforme dos símbolos pelas três posições), que nada de relevante acrescenta ao nosso estudo.

Como consequência do teorema 4.2.7 e das condições a que os contadores de símbolos estão sujeitos, tem-se que o máximo valor de $S(N/3)$ é $3(N/3)^2$ (as três primeiras situações da tabela 4.2), quando há repetição sucessiva de um terno constituído por três nucleótidos distintos. Nos casos em que se tem ausência de dois dos símbolos (repetições sucessivas de um terno constituído por exactamente dois nucleótidos distintos) o máximo é $2(N/3)^2$ (as três últimas situações da tabela 4.2).

As sequência genéticas não têm obviamente comprimentos iguais e os valores de $S(N/3)$ dependem desses comprimentos. No sentido de encontrar valores de $S(N/3)$ comparáveis entre as várias sequências, vai ser explorada uma normalização que retire a dependência explicita do comprimento da sequência.

Teorema 4.2.8. *Sejam $b \in \{a, c, g, t\}$ e $P_i(b)$ a probabilidade de encontrar o símbolo $\mathcal{B}$ na i-ésima posição e $i \in \{0, 1, 2\}$. Então:*

$$F(y^b) = \left(\frac{N}{3}\right)^2 \left(\left[P_0(b) - \frac{P_1(b) + P_2(b)}{2} \right]^2 + \frac{3}{4}[P_1(b) - P_2(b)]^2 \right). \tag{4.10}$$

Demonstração. É de notar que se pode obter o cálculo dos contadores à custa das probabilidades dos símbolos condicionadas às posições:

$$y^b_{(i)} = \frac{N}{m} P_i(b),\ i \in \{0, 1, \ldots, m-1\} \text{ e } b \in \{a, c, g, t\}.$$

Logo, para $m = 3$, vem

$$y^b_{(i)} = \frac{N}{3} P_i(b),\ i \in \{0, 1, 2\},$$

o que obviamente leva à conclusão do teorema. □

O valor de $F(y^b)$ normalizado, com $b \in \{a, c, g, t\}$, será denotado por $F_{norm}(y^b)$ e é dado por

$$\frac{F(y^b)}{(N/3)^2}.$$

Tem-se que $F_{norm}(y^b) \in [0, 1]$.

A normalização de $S(N/3)$ será denotada por $S_{norm}(N/3)$ e será dada por:

$$\begin{aligned} S_{norm}(N/3) &= F_{norm}(y^a) + F_{norm}(y^c) + F_{norm}(y^g) + F_{norm}(y^t) \\ &= \frac{S(N/3)}{(N/3)^2}. \end{aligned}$$

O valor da risca espectral normalizada, $S_{norm}(N/3)$, varia entre zero e três.

4.2.1.1 Resultados de $S(N/3)$ em alguns genes

De seguida apresenta-se um pequeno estudo exploratório sobre o comportamento da risca espectral $S(N/3)$ para dados reais, consistindo em: genes das espécies *Saccharomyces cerevisiae* e *Escherichia coli* (ver tabela 4.3 e tabela 4.4 respectivamente). Em particular, apresentam-se os valores normalizados de $S(N/3)$ (segunda coluna das tabelas), a indicação dos genes que apresentam $S(N/3)$ como componente espectral máxima (terceira coluna das tabelas) e o número de nucleótidos que constitui o gene (quarta coluna das tabelas).

Nas tabelas 4.3 e 4.4 confirma-se que a maioria dos genes apresentam como risca espectral máxima $S(N/3)$, mas também se observa que existem genes que não apresentam esse máximo. É de realçar que em todos os genes estudados $S_{norm}(N/3)$ é inferior a 0,65 que é um valor muito inferior ao máximo (três quando há repetição periódica de três símbolos diferentes ou eventualmente dois quando há repetição periódica de ternos constituídos por dois e apenas dois símbolos diferentes).

4.2.1.2 $S(N/3)$ em sequências geradas aleatoriamente

A risca espectral $S(N/3)$ pode ser escrita à custa das probabilidades dos símbolos, equação (4.10), pelo que dada uma distribuição de probabilidades podemos obter o valor da risca.

Pretende-se averiguar que valores de probabilidades (ou que valor de $S(N/3)$) levam a sequências aleatórias que reflectem um máximo da risca espectral $S(N/3)$. Para esse efeito, foram geradas aleatoriamente sequências de símbolos baseadas em diferentes valores de probabilidades e calculados os correspondentes espectros de potência.

Sem perda da generalidade, estudou-se apenas $F_{norm}(y^b)$, com $b \in \{a, c, g, t\}$. Suponha-se que os quatro símbolos apresentam iguais frequências relativas. Então,

$$\frac{P_0(b) + P_1(b) + P_2(b)}{3} = 0{,}25.$$

Assumindo equiprobabilidade dos quatro símbolos,[4] calcularam-se valores $F_{norm}(y^b)$ para diferentes valores das probabilidades relativas às três posições da sequência (ver tabela 4.5). Fixados os valores das probabilidades $(P_0(b), P_1(b), P_2(b))$ geraram-se sequências aleatórias com cem e mil símbolos e registou-se para cada um dos casos a percentagem de vezes que a risca espectral $S(N/3)$ atinge o valor máximo (ver quinta e sexta colunas da tabela 4.5). Pode-se observar que o valor máximo de $F_{norm}(y^b)$ é 0,5625$<$ 1. Por outro lado, verificou-se que a risca espectral era máxima, em todos os casos estudados, para $F_{norm}(y^b) \in [0{,}0900; 0{,}5625]$.

Para as sequências geradas aleatoriamente pode-se indicar valores muito inferiores a três de $S_{norm}(N/3)$ que ainda reflectem $S(N/3)$ como máxima risca espectral:

$$4(0{,}09)(N/3)^2 = 0{,}36(N/3)^2$$

para sequências com cem bases.

Note-se que $S(N/3) = 0{,}36(N/3)^2$ pode ser obtido para os seguintes valores de probabilidades dos símbolos: $P_0(a) = 0{,}15$, $P_1(a) = 0{,}15$, $P_2(a) = 0{,}45$, $P_0(c) = 0{,}15$, $P_1(c) = 0{,}15$, $P_2(c) = 0{,}45$, $P_0(g) = 0{,}05$, $P_1(g) = 0{,}35$, $P_2(g) = 0{,}35$, $P_0(t) = 0{,}05$, $P_1(t) = 0{,}35$, $P_2(t) = 0{,}35$ (ver tabela 4.5). No entanto, este menor valor para a máxima risca espectral ainda está muito acima dos valores obtidos em alguns genes (ver tabela 4.3 e 4.4).

O menor valor que se pretende fixar, ou seja, o limite inferior a partir do qual $S(N/3)$ é máximo, parece depender do comprimento da sequência (ver na tabela 4.5 a diferença de resultados entre as sequências de comprimento $N = 100$ e as de comprimento $N = 1000$). Observa-se que quanto maior for o comprimento, menor é o limite a partir do qual se considera $S(N/3)$ elevado. Nas sequências aleatórias de cem símbolos o limite mínimo a partir do qual todas as sequências reflectem $F(y^b)$ como máximo é cerca de 0,09 e nas sequências aleatórias de mil símbolos o limite mínimo é de cerca de 0,02 (ver quinta e sexta colunas da tabela 4.5).

[4]Dois resultados dizem-se equiprováveis se têm iguais probabilidades de ocorrência.

4.2.2 Outras riscas espectrais

De seguida vão ser estudadas, em particular, mais duas riscas espectrais: $S(N/4)$ e $S(N/6)$. O objectivo é apresentar expressões das riscas espectrais baseadas nos contadores de símbolos e caracterizar a distribuição de frequência dos contadores que levam a situações de riscas espectrais nulas.

Tal como para $S(N/3)$, a expressão que define a risca espectral $S(N/4)$ é igualmente uma consequência imediata do teorema 4.2.2 (ver demonstração do teorema 4.2.3).

Teorema 4.2.9. *Seja $N = 4n$. A risca espectral $S(N/4)$ é dada por*

$$S(N/4) = F(y^a) + F(y^c) + F(y^g) + F(y^t),$$

onde

$$F(y^b) = [y^b_{(0)} - y^b_{(2)}]^2 + [y^b_{(1)} - y^b_{(3)}]^2,\ b \in \{a, c, g, t\}.$$

Demonstração. Fixe-se $k = 1$ e $n = N/4$ em (4.4). Isto leva de imediato a que

$$S(N/4) = |Y^a_{(1)}|^2 + |Y^c_{(1)}|^2 + |Y^g_{(1)}|^2 + |Y^t_{(1)}|^2, \tag{4.11}$$

onde

$$Y^b_{(1)} = y^b_{(0)} + y^b_{(1)}w + y^b_{(2)}w^2 + y^b_{(3)}w^3,$$

e $w = e^{-i2\pi/4}$.

$$\begin{aligned} |Y^b_{(1)}|^2 &= |y^b_{(0)} + y^b_{(1)}i - y^b_{(2)} - y^b_{(3)}i|^2 \\ &= [y^b_{(0)} - y^b_{(2)}]^2 + [y^b_{(1)} - y^b_{(3)}]^2 \\ &= F(y^b). \end{aligned}$$

□

Teorema 4.2.10. *Seja $N = 4n$. A risca espectral $S(N/4)$ é zero se e só se*

$$y^b_{(0)} = y^b_{(2)} \text{ e } y^b_{(1)} = y^b_{(3)},$$

para todo o $b \in \{a, c, g, t\}$.

Demonstração. Considerando que $F(y^b) = [y^b_{(0)} - y^b_{(2)}]^2 + [y^b_{(1)} - y^b_{(3)}]^2 = 0$, o resultado é evidente. □

Em [92], o caso $S(N/4) = 0$, com $y^b_{(0)} = y^b_{(2)} \neq y^b_{(1)} = y^b_{(3)}$ e $b \in \{a, c, g, t\}$, é chamado de "hidden periodicity" (periodicidade escondida)[5] de período quatro. Note-se que na realidade esta é uma periodicidade de período dois.

Recorde-se que $U^b_{(N/2)}$, com $b \in \{a, c, g, t\}$, consiste numa diferença entre o número total de ocorrências de $\mathcal{B}$ nas posições pares e ímpares. Em termos do contador de símbolos para $N = 4n$, isto pode ser escrito da seguinte forma:

$$U^b_{(N/2)} = \left[y^b_{(0)} + y^b_{(2)}\right] - \left[y^b_{(1)} + y^b_{(3)}\right].$$

Uma vez que por hipótese $U^b_{(N/4)} = 0$, as quantidades $y^b_{(0)} + y^b_{(2)} = 2y^b_{(0)}$ e $y^b_{(1)} + y^b_{(3)} = 2y^b_{(1)}$ em geral são diferentes entre si. Sendo assim tem-se que

$$U^b_{(N/2)} \neq 0.$$

Portanto, a periodicidade escondida de período quatro referida em [92] é exposta através da análise espectral como uma periodicidade de período dois.

Teorema 4.2.11. *Seja $N = 6n$. A risca espectral $S(N/6)$ é dada por*

$$S(N/6) = F(y^a) + F(y^c) + F(y^g) + F(y^t),$$

onde

$$\begin{aligned} F(y^b) &= \left[y^b_{(0)} + \frac{y^b_{(1)}}{2} + \frac{y^b_{(5)}}{2} - \left(\frac{y^b_{(2)}}{2} + y^b_{(3)} + \frac{y^b_{(4)}}{2}\right)\right]^2 \\ &+ \frac{3}{4}\left[y^b_{(1)} + y^b_{(2)} - (y^b_{(4)} + y^b_{(5)})\right]^2, \end{aligned}$$

com $b \in \{a, c, g, t\}$.

Demonstração. Considerando que $F(y^b) = |Y^b_{(1)}|^2$, com $k = 1$ e $n = N/6$, e aplicando o teorema 4.2.2 (como na demonstração do teorema 4.2.9) o resultado é evidente. □

[5]Entenda-se por periodicidade escondida de período T a periodicidade que não pode ser detectada por análise do coeficiente espectral correspondente à frequência $1/T$.

Teorema 4.2.12. *Seja $N = 6n$. A risca espectral $S(N/6)$ assume o valor zero se e só se*

$$y^b_{(0)} - y^b_{(3)} = y^b_{(4)} - y^b_{(1)} = y^b_{(2)} - y^b_{(5)},$$

para todo o $b \in \{a, c, g, t\}$.

Demonstração.

$$S(N/6) = 0 \Leftrightarrow \begin{cases} y^b_{(0)} + \frac{y^b_{(1)}}{2} + \frac{y^b_{(5)}}{2} = \frac{y^b_{(2)}}{2} + y^b_{(3)} + \frac{y^b_{(4)}}{2} \\ y^b_{(1)} + y^b_{(2)} = y^b_{(4)} + y^b_{(5)} \end{cases}$$

$$\Leftrightarrow \; y^b_{(0)} - y^b_{(3)} = y^b_{(4)} - y^b_{(1)} = y^b_{(2)} - y^b_{(5)}.$$

□

Se os pares de contadores espaçados de três posições $(y^b_{(j)}, y^b_{(j+3)})$, com $b \in \{a, c, g, t\}$ e $j \in \{0, 1, 2\}$, forem uniformemente distribuídos, a risca espectral $S(N/6)$ é igual a zero. Em particular, se os símbolos nas seis posições forem uniformemente distribuídos, então $S(N/6) = 0$.

4.3 Dependência dos coeficientes espectrais

Apesar de não directamente relacionados com os anteriores, os resultados apresentados nesta secção visam também ajudar a compreender a estrutura dos coeficientes espectrais das sequências simbólicas.

Mostra-se que existe dependência linear entre os coeficientes espectrais de uma sequência de nucleótidos. Para tal serão usados polinómios localizadores de erros e códigos (para mais desenvolvimento sobre este tema ver [129]).

Para efeitos de simplificação considera-se apenas a sequência indicadora de um símbolo u^b, com $b \in \{a, c, g, t\}$, e a sua DFT U^b.

Teorema 4.3.1. *Seja u^b, com $b \in \{a, c, g, t\}$, uma sequência indicadora de comprimento N com n elementos iguais a um, posicionados em $\{j_0, j_1, \ldots, j_{n-1}\}$. Então, os*

N coeficientes espectrais U^b dados por (4.2) satisfazem

$$U^b_{(\ell+n)} = -\sum_{k=0}^{n-1} h_k U^b_{(\ell+k)},$$

onde h_k são os coeficientes do polinómio

$$P(z) = \sum_{k=0}^{n} h_k z^k,$$

determinado por $h_n = 1$ e

$$P\left(e^{-i2\pi j_p/N}\right) = 0, \qquad (0 \leq p < n).$$

Demonstração. Considerem-se as n equações

$$P\left(e^{-i2\pi j_p/N}\right) = \sum_{k=0}^{n} h_k e^{-i2\pi j_p k/N} = 0, \qquad (0 \leq p < n).$$

A multiplicação de cada uma delas por

$$u^b_{(j_p)} e^{-i2\pi j_p \ell/N}$$

conduz a

$$\sum_{k=0}^{n} h_k u^b_{(j_p)} e^{-i2\pi j_p (k+\ell)/N} = 0, \qquad (0 \leq p < n).$$

Somando as n equações e tendo em conta que

$$u^b_p = 0 \text{ para } p \in \{0, \ldots, N-1\} \backslash \{j_0, j_1, \ldots, j_{n-1}\},$$

vem que

$$\begin{aligned}
&\sum_{p=0}^{n-1} \sum_{k=0}^{n} h_k u^b_{(j_p)} e^{-i2\pi j_p (k+\ell)/N} = \\
&= \sum_{k=0}^{n} h_k \sum_{p=0}^{n-1} u^b_{(j_p)} e^{-i2\pi j_p (\ell+k)/N} \\
&= \sum_{k=0}^{n} h_k \sum_{p=0}^{N-1} u^b_p e^{-i2\pi p(\ell+k)/N} \\
&= \sum_{k=0}^{n} h_k U^b_{(\ell+k)}.
\end{aligned}$$

Mostra-se que

$$\sum_{k=0}^{n} h_k U^b_{(\ell+k)} = 0,$$

De forma equivalente tem-se,

$$h_n U^b_{(\ell+n)} + \sum_{k=0}^{n-1} h_k U^b_{(\ell+k)} = 0$$

Uma vez que $h_n = 1$, obtém-se o resultado pretendido,

$$U^b_{(\ell+n)} = -\sum_{k=0}^{n-1} h_k U^b_{(\ell+k)}.$$

□

4.4 Conclusões

Neste capítulo foi apresentada a relação entre as riscas espectrais $S(k)$ e a distribuição dos nucleótidos em certas subsequências. Relativamente ao trabalho [92] foi discutida uma maneira muito mais simples de obter condições necessárias e suficientes para que uma risca espectral seja zero, usando uma transformação em somas por blocos. Foi apresentado um procedimento computacional para calcular $S(k)$, e em particular $S(N/3)$, risca espectral de grande importância na identificação de regiões de código.

Foram discutidos limites superiores e inferiores relativos ao coeficiente espectral $S(N/3)$. Foi feito um estudo exploratório sobre a distribuição de probabilidade que leva a que $S(N/3)$ seja máxima.

Finalmente, foi mostrado que os coeficientes espectrais não são independentes entre si. Se existirem n símbolos numa sequência de comprimento total N, então os N elementos de U^b com $b \in \{a, c, g, t\}$ podem ser determinados por exemplo a partir de $U^b_0, U^b_1, \ldots, U^b_{(n-1)}$, através de recursão linear. Esta dependência pode ser usada, por exemplo, para verificar e também para corrigir erros numa sequência indicadora dada.

Referência	$S_{norm}(N/3)$	$S(N/3)$ é a risca máxima do espectro	N
gi—6319247:335-649	0,02	0	315
gi—6319247:c2169-1807	0,17	1	363
gi—6319247: c9017-7236	0,03	1	1782
gi—6319247:10092-10400	0,06	0	309
gi—6319247:c11952-11566	0,18	1	387
gi—6319247:12047-12427	0,22	1	381
gi—6319247:c13744-13364	0,05	0	381
gi—6319247:21526-21852	0,02	0	327
gi—6319247:c27969-24001	0,21	1	3969
gi—6319247:31568-32941	0,07	1	1374
gi—6319247:33449-34702	0,05	1	1254
gi—6319247:35156-36304	0,07	1	1149
gi—6319247:36510-37148	0,06	1	639
gi—6319247:37465-38973	0,03	1	1509
gi—6319247:c39047-38697	0,03	0	351
gi—6319247:39260-41803	0,03	1	2544
gi—6319247:42177-42719	0,03	0	543
gi—6319247:c45022-42881	0,06	1	2142
gi—6319247:45899-48250	0,05	1	2352
gi—6319247:48564-51752	0,02	1	3189
gi—6319247:c52597-51857	0,08	1	741
gi—6319247:c54791-52803	0,02	1	1989
gi—6319247:c56859-54991	0,04	1	1869
gi—6319247:c57387-57031	0,05	0	357
gi—6319247:c57798-57490	0,05	0	309
gi—6319247:c58485-57952	0,06	1	534
gi—6319247:c61054-58697	0,05	1	2358
gi—6319247:61318-62565	0,06	1	1248
gi—6319247:c61610-61233	0,08	1	378
gi—6319247:62842-65406	0,02	1	2565

Tabela 4.3: Resultados relativos à risca $S(N/3)$ de alguns genes da *Saccharomyces cerevisiae*.

Referência	$S_{norm}(N/3)$	$S(N/3)$ é a risca máxima do espectro	N
Genoma completo	4,2E-05	0	2315355
NC-000913.2—:190-255	0,64	1	66
NC-000913.2—:337-2799	0,07	1	2463
NC-000913.2—:2801-3733	0,05	1	933
NC-000913.2—:3734-5020	0,08	1	1287
NC-000913.2—:5234-5530	0,08	1	297
NC-000913.2—:c6459-5683	0,08	1	777
NC-000913.2—:c7959-6529	0,05	1	1431
NC-000913.2—:8238-9191	0,08	1	954
NC-000913.2—:9306-9893	0,06	1	588
NC-000913.2—:c10494-9928	0,10	1	567
NC-000913.2—:c11356-10643	0,06	1	714
NC-000913.2—:10725-11315	0,07	1	591
NC-000913.2—:c11786-11382	0,07	1	405
NC-000913.2—:12163-14079	0,13	1	1917
NC-000913.2—:14168-15298	0,08	1	1131
NC-000913.2—:15445-16557	0,04	1	1113
NC-000913.2—:c16177-15869	0,04	0	309
NC-000913.2—:c16960-16751	0,14	1	210
NC-000913.2—:17489-18655	0,07	1	1167
NC-000913.2—:18715-19620	0,04	1	906
NC-000913.2—:c20508-20233	0,04	0	276
NC-000913.2—:c21078-20815	0,09	1	264
NC-000913.2—:21181-21399	0,03	0	219
NC-000913.2—:21407-22348	0,06	1	942
NC-000913.2—:22391-25207	0,08	1	2817
NC-000913.2—:25207-25701	0,08	1	495
NC-000913.2—:25826-26275	0,09	1	450
NC-000913.2—:26277-27227	0,10	1	951
NC-000913.2—:27293-28207	0,06	1	915
NC-000913.2—:28374-29195	0,11	1	822

Tabela 4.4: Resultados relativos à risca $S(N/3)$ de alguns genes da *Escherichia coli* da estirpe K12.

$P_0(b)$	$P_1(b)$	$P_2(b)$	$F_{norm}(y^b)$	% de max $N = 100$	% de max $N = 1000$
0	0	0,75	$0,5625$	100	100
0	0,05	0,7	$0,4575$	100	100
0	0,1	0,65	$0,3625$	100	100
0,05	0,05	0,65	$0,3600$	100	100
0,05	0.1	0,6	$0,2775$	100	100
0,1	0,1	0,55	$0,2025$	100	100
0,05	0,15	0,55	$0,2100$	100	100
0	0,25	0,5	$0,1875$	100	100
0,05	0,2	0,5	$0,1575$	100	100
0	0,35	0,4	$0,1425$	100	100
0,05	0,25	0,45	$0,1200$	100	100
0,1	0,2	0,45	$0,0975$	100	100
0,05	0,35	0,35	$0,0900$	100	100
0,15	0,15	0,45	$0,0900$	100	100
0,1	0,25	0,4	$0,0675$	97	100
0,1	0,3	0,35	$0,0525$	87	100
0,15	0,25	0,35	$0,03$	54	100
0,15	0,3	0,3	$0,0225$	39	100
0,175	0,25	0,325	$0,0169$	20	100
0,2	0,25	0,3	$0,0075$	8	86
0,25	0,25	0,25	0	1	0

Tabela 4.5: Resultados de $F(y^b)$ normalizado, dadas as probabilidades. Percentagem de vezes que $F(y^b)$ é máximo em sequências geradas aleatoriamente com $N = 100$ e $N = 1000$, para diferentes distribuições de probabilidades.

Capítulo 5

Modelo de três estados

A tarefa de sequenciar o código genético de todas as espécies, incluindo aquelas que já se extinguiram, implica armazenar uma enorme quantidade de informação. Actualmente, existem em bases de dados distribuídas por todo o mundo, *gigabytes* de informação correspondente a sequências de nucleótidos, bem como de aminoácidos. Isto explica a necessidade de algoritmos de compressão que optimizem e racionalizem o armazenamento e a comunicação da informação genética.

Se o DNA fosse uma sequência de símbolos puramente aleatória (símbolos uniformemente distribuídos) então a entropia seria máxima e a melhor maneira de a representar seria usando dois *bits* por cada um dos quatro símbolos. No entanto, existem regularidades, isto é, propriedades específicas da sequência estatisticamente comprováveis, que provam a existência de entropia redutível, num grau ainda por descobrir, que abrem caminho a uma investigação que conduzirá a um melhor conhecimento do código genético, podendo conduzir a descobertas filogenéticas e necessariamente à compressão da informação. A compressibilidade das sequências de DNA não é linear. Algumas, denotam grande entropia e pouco melhor se consegue que os dois *bits*/base, noutras, os melhores algoritmos chegam a ganhar 40%, obtendo a marca de 1.6 *bits*/base [9]. Os resultados são variáveis consoante a complexidade do organismo. De facto, os seres eucariotas possuem um maior número de regularidades, logo possuem códigos com maior grau de compressibilidade [1].

Parte do trabalho desenvolvido neste capítulo foi publicado em [124].

5.1 Introdução

O estudo de algoritmos para compressão de dados tem geralmente dois propósitos:

- a necessidade de armazenamento eficiente;
- a necessidade de transmissão eficiente.

Em geral, directamente relacionado com a técnica de compressão, existe um modelo que reproduz a fonte de informação a ser comprimida. Independentemente da compressão, este modelo pode ter interesse por revelar propriedades estatísticas dos dados.

No caso do DNA, um dos objectivos é encontrar métodos eficientes capazes de reduzir o espaço de armazenamento de dados genéticos que estão continuamente a ser gerados. Por exemplo, o genoma humano tem cerca de 3 000 milhões de pares de bases [127], e o genoma do trigo tem cerca de 16 000 milhões [52].

Por outro lado, também se pretende descobrir como funciona o código genético e que estrutura possui. A criação de bons modelos para descrever o DNA é uma forma de alcançar esse conhecimento.

Sabe-se que as regiões codificantes têm propriedades específicas. Em particular, as regiões não codificantes são mais fáceis de comprimir do que as regiões codificantes, uma vez que para comprimir as regiões não codificantes usa-se o facto de existirem alguns tipos de repetições. Geralmente estas características não se verificam nas regiões codificantes [53]. O trabalho a apresentar neste capítulo centra-se na compressão de DNA nas regiões codificantes.

De acordo com o que foi apresentado nos capítulos anteriores, na parte codificante do DNA é usual verificar-se a existência de periodicidade de período três nos nucleótidos. No contexto da compressão de DNA esta característica ainda não foi explorada [144, 148]. Neste capítulo, pretende-se explorar a periodicidade de período três das regiões de código no contexto da compressão. Nesse sentido, é proposto um modelo de contexto finito em três estados. Cada um dos estados é seleccionado periodicamente de acordo com a periodicidade de período três e cada estado é implementado usando um modelo de contexto finito. Comparando o resultado entre o modelo de três estados que varia

ciclicamente com o modelo de contexto finito simples, o primeiro é melhor a descrever os dados. O modelo a apresentar tem outra característica interessante: a entropia relativa a cada um dos três estados pode ser estimada individualmente, e o resultado pode ser interpretado em termos do código genético e das características biológicas do organismo em estudo.

Devido às propriedades genéticas do código, a entropia associada à primeira, segunda e terceira base pode variar de acordo com a distribuição dos codões sinónimos. Por exemplo, a variação da entropia associada à terceira base do codão dado que se conhecem as primeiras duas bases está compreendida entre as duas seguintes situações extremas:

- a sequência de DNA apresenta tal preferência por um dos sinónimos que nenhum dos outros aparece na sequência;
- a sequência não apresenta preferência por nenhum dos sinónimos, possuindo os sinónimos frequências semelhantes entre si na sequência.

No primeiro caso, a entropia do terceiro nucleótido é zero, pois o nucleótido é determinado pelo conhecimento dos dois anteriores (ver tabela 1.1). No segundo caso, a entropia do terceiro nucleótido é semelhante à de qualquer ocorrência na terceira posição do nucleótido que ainda constitua um sinónimo.

O modelo de três estados está apto a detectar variações de entropia nas regiões codificantes ao longo das três posições (estados) dos codões 3n, 3n+1 e 3n+2. Os resultados obtidos confirmam que a entropia difere nos três estados, e a variação não é a mesma de organismo para organismo. As diferenças podem ser motivadas por várias causas, nomeadamente: um grande número de organismos tem mais $\mathcal{C} + \mathcal{G}$ do que $\mathcal{A} + \mathcal{T}$, pois a ligação entre $\mathcal{C}$ e $\mathcal{G}$ é a mais difícil de quebrar; as preferências do codão variam significativamente de acordo com a espécie.

5.1.1 Conceitos básicos

A entropia de uma variável aleatória pode ser interpretada como o grau de informação que se tem sobre a variável. Quanto mais uniforme for a distribuição de probabilidades

da variável maior será a entropia e menos previsível é o valor de qualquer concretização dessa variável (ver, por exemplo, [44]).

Para uma dada experiência aleatória onde só pode ocorrer sucesso ou insucesso (experiência de Bernoulli), 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.

O problema geralmente não se coloca com uma experiência de Bernoulli, mas com um conjunto de vários acontecimentos possíveis e mutuamente exclusivos que podem ocorrer sobre um mesmo espaço de probabilidades (variável aleatória multinomial). Para medir o conhecimento ou desconhecimento que se tem sobre o comportamento de uma variável aleatória pode-se recorrer à entropia.

Seja $P(x)$ a função de probabilidade,[1] com x definida sobre um espaço de estados discreto Γ.[2] A entropia (H) da variável aleatória discreta X, é definida por

$$H(X) = -\sum_{x\in\Gamma} P(x)\log_2 P(x). \tag{5.1}$$

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

$$H(X) = \mathbf{E}\{-\log_2 P(X)\}, \tag{5.2}$$

e é geralmente interpretada como a incerteza média da variável aleatória. Também directamente de (5.2) a entropia pode ser vista como o número médio de *bits* necessário para codificar um elemento do espaço de estados. Ou seja, a entropia consiste no comprimento médio da mensagem necessária para transmitir o resultado da variável, o que normalmente é medido em *bits*. Em geral, uma codificação óptima envia uma mensagem de probabilidade p com comprimento $-\log_2 p$, sendo utilizados menos *bits* nos resultados mais prováveis e mais *bits* nos resultados menos prováveis.

Numa dada experiência aleatória, com $\Gamma = \{x_1, x_2, \ldots, x_N\}$ um espaço de estados discreto, o valor máximo de entropia é atingido aquando da equiprobabilidade de re-

[1] Entenda-se por $P(x)$ a probabilidade da variável aleatória X assumir a concretização x.

[2] O espaço de estados de uma variável é o conjunto de valores que essa variável pode assumir.

sultados. De notar que uma experiência aleatória apresenta resultados equiprováveis quando assume qualquer um dos possíveis estados com igual probabilidade, ou seja,

$$P(x_i) = \frac{1}{N}, \text{ com } i = 1, 2, \ldots, N.$$

Assim resulta que a entropia máxima é dada por

$$H_M = \log_2 N. \tag{5.3}$$

Dada a definição de entropia e tendo por objectivo a análise de sequências simbólicas, esta medida poderá fornecer informação parcial sobre a complexidade das sequências.

De seguida, apresentam-se mais dois conceitos de entropia associados a duas variáveis aleatórias: entropia conjunta e entropia condicionada.

Seja (X, Y) um par de variáveis aleatórias discretas, com espaço de estados $\Gamma_1 \times \Gamma_2$. A distribuição de probabilidade conjunta será denotada por $P(x, y)$ e a distribuição de probabilidade condicionada de X dado Y será denotada por $P(x|y)$ para $(x, y) \in \Gamma_1 \times \Gamma_2$.

A *entropia conjunta* do par (X, Y) consiste na incerteza média do par de variáveis aleatórias, ou seja

$$H(X, Y) = -\sum_{x \in \Gamma_1} \sum_{y \in \Gamma_2} P(x, y) \log_2 P(x, y). \tag{5.4}$$

A *entropia condicionada* de X a Y, é a informação extra, em média, necessária para "comunicar" X dado que o "receptor" conhece Y e é dada por

$$H(X|Y) = -\sum_{x \in \Gamma_1} \sum_{y \in \Gamma_2} P(x, y) \log_2 P(x|y). \tag{5.5}$$

Também podemos escrever

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

A entropia pode medir o grau de informação entre um par de variáveis aleatórias, mas para medir directamente a informação de um par de variáveis aleatórias existem

outras medidas. Uma medida frequentemente usada é a informação mútua. Dadas duas variáveis aleatórias discretas, X e Y, a informação mútua entre elas é dada por

$$I(X,Y) = \sum_{x\in\Gamma_2}\sum_{y\in\Gamma_2} P(x,y)\log_2\frac{P(x,y)}{P(x)P(y)}.$$

Tendo em conta a definição de informação mútua pode-se recordar as seguintes igualdades básicas usando medidas de entropia,

$$\begin{aligned} I(X,Y) &= H(Y) - H(Y|X) \\ &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y). \end{aligned}$$

5.2 Os modelos de compressão

A maioria dos actuais modelos de compressão de DNA têm em conta características da sequência, como por exemplo sub-sequências de repetições exactas ou aproximadas e palíndromas. Estas são características que têm resultado em ganhos significativos em termos de compressão. No entanto, outros aspectos podem ser considerados, tais como as características das regiões codificantes do DNA.

Dado que as regiões codificantes apresentam uma destacada periodicidade de período três, que poderá estar associada à estrutura dos codões (ternos de bases), surge a ideia de estrutura ciclo-estacionaria de período três associada a estas regiões. Assim, no sentido de explorar e confirmar a existência da estrutura ciclo estacionária, apresenta-se uma medida de entropia que tem em conta esta estrutura

$$H^c(X) = \frac{H_0(X) + H_1(X) + H_2(X)}{3},$$

H^c será denominada de entropia ciclo-estacionária, e

$$H_i(X) = -\sum_{x\in\{a,c,g,t\}} P_i(x)\log_2(P_i(x))$$

a entropia associada à i-ésima base do codão com $i \in \{0,1,2\}$.

Depois de analisadas algumas sequências de código (genes), observámos ganhos da entropia ciclo-estacionária relativamente à entropia global da sequência dos nucleótidos (ver alguns exemplos na tabela 5.2). Assim confirma-se que as regiões codificantes reflectem uma estrutura diferente nas três posições das bases dos codões e pode-se inferir a existência de estrutura ciclo-estacionária de período três. No entanto, é de observar que os valores de qualquer tipo de entropia são pouco inferiores aos dois *bits*/base (ver tabela 5.2).

A evidente periodicidade de período três e os resultados de entropia ciclo estacionária (ver tabela 5.2) nas regiões codificantes sugerem para descrição destas regiões um modelo de compressão que se subdivida em três modelos diferentes que poderão eventualmente estar relacionados entre si.

O modelo de compressão a criar baseia-se nos modelos de contexto finito, pelo que se segue uma breve descrição deste tipo de modelos. No fim desta secção apresenta-se uma subsecção com um modelo de compressão para regiões codificantes que tem por base os modelos de contexto finito e incorpora a ciclo-estacionaridade.

5.2.1 Modelo de contexto finito

Considere-se uma fonte de informação geradora de símbolos de um alfabeto Γ de dimensão $|\Gamma|$. No instante t, a sequência gerada pela fonte de informação é

$$x^t = x_1 x_2 \dots x_t.$$

Num modelo de contexto finito (cadeia de Markov de ordem superior ou igual a um) a probabilidade de surgir um dado símbolo do alfabeto depende apenas de um número finito de símbolos gerado anteriormente, M (modelo de contexto finito de ordem M) [23, 130]. Na figura 5.1 exemplifica-se um modelo de contexto finito de ordem cinco.

Para o instante t, o conjunto ordenado de variáveis aleatórias,

$$X_{t-M+1}, \dots, X_{t-1}, X_t,$$

será representado por C^t e a C^t chamamos contexto aleatório associado à posição t. O número de diferentes possibilidades para C^t é $|\Gamma|^M$.

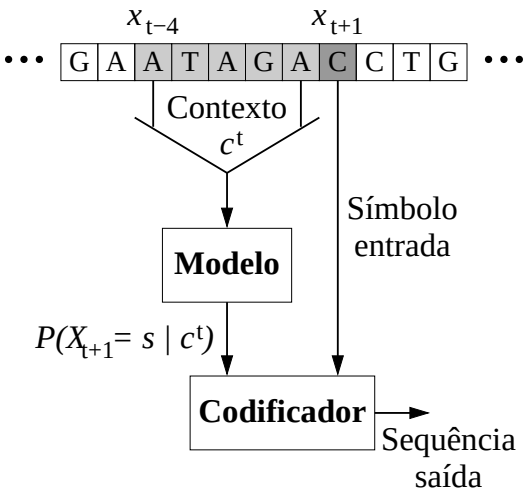

Figura 5.1: Modelo de contexto finito. A probabilidade associada à variável aleatória X_{t+1} é condicionada pelas M saídas anteriores. Neste exemplo, $M = 5$.

Na prática, a probabilidade do próximo símbolo, X_{t+1}, ser $s \in \Gamma$ dado que ocorreu c^t, é estimada por[3]

$$P(X_{t+1} = s|c^t) = \frac{n(s, c^t) + \delta}{\sum_{a\in\Gamma} n(a, c^t) + |\Gamma|\delta},$$

onde $n(s, c^t)$ é um contador que representa o número de vezes que, no passado, a fonte de informação gerou o símbolo s depois de gerar c^t. O parâmetro $\delta > 0$, além de ajustar o estimador, evita gerar probabilidades zero quando um símbolo é codificado pela primeira vez. No nosso caso, usámos $\delta = 1$, que pode ser visto como uma inicialização de todos os contadores a um. Naturalmente, estes contadores são actualizados em cada instante que um símbolo é codificado.

Na tabela 5.1, apresenta-se uma tabela que mostra como um modelo de contexto finito de ordem M é tipicamente implementado. Esta tabela refere-se à actualização de contagens para a t-ésima saída no alfabeto dos nucleótidos e apresenta as frequências com que surge cada uma das palavras de comprimento $M + 1$ até à t-ésima posição (inclusivé).

O codificador a que se refere a figura 5.1 é um codificador aritmético. Se a sequência for estacionária, é conhecido que o codificador aritmético gera débitos binários médios

[3]Por comodidade a notação usada para o valor real, estimador e estimativa de uma determinada probabilidade será a mesma.

	$\mathcal{A}$	$\mathcal{C}$	$\mathcal{G}$	$\mathcal{T}$	Total
c_1^t	$n(\mathcal{A}, c_1^t)$	$n(\mathcal{C}, c_1^t)$	$n(\mathcal{G}, c_1^t)$	$n(\mathcal{T}, c_1^t)$	$\sum_{a\in\Gamma} n(a, c_1^t)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$c_{4^M}^t$	$n(\mathcal{A}, c_{4^M}^t)$	$n(\mathcal{C}, c_{4^M}^t)$	$n(\mathcal{G}, c_{4^M}^t)$	$n(\mathcal{T}, c_{4^M}^t)$	$\sum_{a\in\Gamma} n(a, c_{4^M}^t)$

Tabela 5.1: Tabela de frequências de pares de palavras, na t-ésima actualização.

idênticos aos valores de entropia condicionada do modelo [23, 130]. No nosso caso, a entropia do modelo, isto é, o número médio de *bits* por símbolo depois de codificar N símbolos, é dada por

$$H_N = -\frac{1}{N}\sum_{t=0}^{N-1} E_t \ \text{ bps}, \tag{5.6}$$

com

$$E_t = \log_2 P(X_{t+1} = s|c^t)$$

e onde "bps" significa *bits* por símbolo e refere-se à unidade de medida.

Uma vez que vamos trabalhar sobre bases de DNA, vamos substituir "bps" por "bpb", no sentido de representar *bits* por base.

Num modelo de contexto finito coloca-se o problema de estimar a ordem do modelo (M). Para cada conjunto de dados podemos escolher M de forma exaustiva, tentando minimizar a entropia. Podemos também usar estimativas baseadas em critérios conhecidos, como o critério de informação bayesiana (BIC). O BIC é um método que, de modo geral, permite estimar o número de parâmetros do modelo. No contexto de uma sequência de símbolos, estima a ordem do modelo (ordem da cadeia de Markov) que melhor se ajusta à sequência, no sentido da máxima verosimilhança. No caso particular do modelo ser de contexto finito de ordem M, com M a determinar (ou cadeia de Markov de ordem M, com M desconhecido), o critério traduz-se na descoberta do valor de M que minimiza $BIC(M)$ dado por

$$BIC(M) = -\ln L(M) + \frac{(|\Gamma| - 1) \times |\Gamma|^M}{2} \ln n_M, \tag{5.7}$$

sendo L a função de verosimilhança[4] assumindo o modelo de contexto finito de ordem M e n_M o número de subsequências (palavras) de tamanho $M+1$ de uma sequência com N símbolos. Observe-se que $n_M = N - M$.

5.2.2 Modelo de três estados

O modelo de três estados é diferente do modelo de contexto finito da figura 5.1, uma vez que incluiu três estados internos, como se apresenta esquematicamente na figura 5.2. Cada estado é seleccionado periodicamente, de três em três, e para cada um dos três casos é usado um modelo de contexto finito semelhante ao descrito anteriormente. Este modelo de três estados explora a periodicidade de período três que geralmente se destaca nas regiões codificantes do DNA.

Com este modelo, as probabilidades não dependem apenas dos M últimos símbolos, mas também do valor de t mod 3, que é usado para selecção do estado. Neste caso, o estimador da probabilidade é dado por

$$P(X_{t+1} = s|c_t) = \frac{n_i(s, c^t) + \delta}{\sum_{a\in\Gamma} n_i(a, c^t) + |\Gamma|\delta}, \quad i = t \bmod 3.$$

Por outras palavras, no modelo de três estados temos três conjuntos diferentes de contadores, um para cada estado. Além disso, apenas os contadores associados ao estado escolhido são actualizados. De notar que, no sentido de simplificar o modelo, não se requer o conhecimento correcto da fase de leitura (do inglês *reading frame*). Contudo, depois de se escolher uma dada posição para o modelo, a fase de leitura correspondente é mantida. Se quisermos calcular a entropia associada a cada uma das três posições das bases dentro dos codões, precisamos de conhecer a posição da base correspondente a cada estado do modelo. Para os casos considerados, vão ser

[4] A expressão que define a função de verosimilhança para $M \geq 1$ é

$$L(M) = P(x_0, x_1, \ldots, x_{M-1}) \prod_{j=M}^{N-1} P(x_j|x_{j-M}, \ldots, x_{j-1})$$

e para $M = 0$ é

$$L(M) = \prod_{j=M}^{N-1} P(x_j).$$

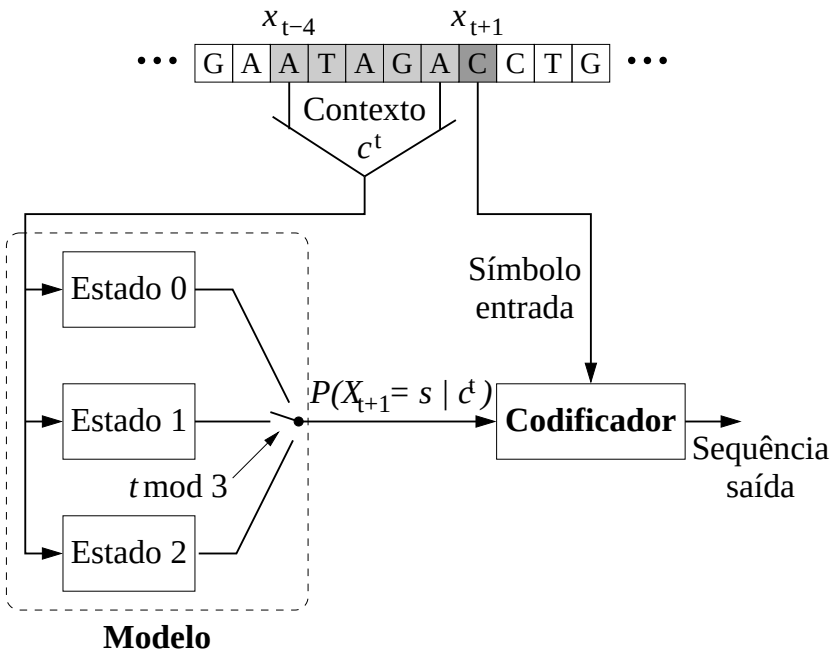

Figura 5.2: Modelo de três estados. A probabilidade associada à variável aleatória, X_{t+1}, é condicionada pelas M saídas anteriores e pelo valor de t mod 3.

usados exemplos em que a aplicação do modelo se inicia na primeira base de um codão. Consequentemente, o estado zero corresponde à primeira base do codão, o estado um à segunda base e o estado dois à terceira base.

A complexidade computacional do modelo de três estados, em termos de necessidade de memória, está directamente relacionada com a implementação do modelo de contexto finito. Podemos observar que para implementar um modelo de contexto finito de ordem M de um alfabeto de tamanho $|\Gamma|$ precisamos de um modelo probabilístico de tamanho $|\Gamma|^M$ que requer $|\Gamma|^{M+1}$ contadores. Por exemplo, se $M = 6$, temos 4^7 contadores e são necessários cerca de trinta e dois *Kbytes* (considerando dois *bytes* por cada contador).

5.3 Resultados experimentais

Os resultados a apresentar são relativos aos dados de alguns ficheiros "ffn" recolhidos do NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/). As espécies seleccionadas foram as seguintes: *Haemophilus influenzae*, *Escherichia coli K12*, *Schizosaccharomyces pombe*,

Saccharomyces cerevisiae e *Arabidopsis thaliana*.

Observou-se que alguns genes contêm símbolos diferentes das letras do alfabeto dos nucleótidos $\{\mathcal{A}, \mathcal{C}, \mathcal{G}, \mathcal{T}\}$ e em alguns casos o comprimento do gene não é múltiplo de três (inconsistência com a estrutura dos codões). No sentido de ultrapassar estas dificuldades, foram eliminados do estudo os genes que têm essas características.

Nas tabelas 5.3, 5.4 e 5.5 apresentam-se os resultados da compressão, em *bits* por base, obtidos de quatro modos diferentes: modelo de contexto finito de três estados, modelo de contexto finito de um só estado, DNACompress [43] (denotado na tabela por "DnaC") e o método de Manzini – Dna3 [113]. A razão para escolher o DNACompress e o Dna3, deve-se ao facto de estas serem as técnicas mais usadas nesta área e com implementações disponíveis. É de observar que os valores obtidos através de (5.6) são muito próximos dos valores obtidos através do codificador aritmético. Todos os valores de compressão apresentados nas tabelas 5.3, 5.4 e 5.5 correspondem a código real gerado pelo codificador aritmético.

Para cada sequência, o comprimento do contexto, M, é escolhido entre as 7 primeiras ordens de modo a que tenha o melhor resultado em termos de compressão. Nos dados das tabelas 5.3, 5.4 e 5.5, pode-se verificar que existe uma tendência para as sequências maiores apresentarem maior valor de M. Também foram determinadas estimativas para o valor da ordem através do BIC. Os valores são geralmente concordantes. Para as sequências das tabelas 5.3, 5.4 e 5.5, obteve-se através do BIC estimativas para M iguais aos valores indicados nas tabelas, com excepção do segundo cromossoma da espécie *Schizosaccharomyces pombe* cuja estimativa pelo BIC foi de três e não quatro.[5]

É visível nas tabelas 5.3, 5.4 e 5.5 que o modelo de contexto finito de três estados é sempre melhor que o modelo de contexto finito de um só estado (ver os totais nas colunas 8 e 10 das tabelas), o que confirma a hipótese sobre a potencial vantagem do conhecimento da periodicidade de período três, no sentido da compressão. Com o modelo de contexto finito de três estados obtiveram-se melhores resultados de compressão relativamente a algumas das melhores técnicas de compressão de DNA da actualidade,

[5]Obter estimativas através do BIC para o número de parâmetros que são inferiores às estimativas que resultam em melhor valor de compressão não é surpreendente, pois o BIC pondera a máxima verosimilhança com a complexidade do modelo (o número de parâmetros do modelo).

o que é confirmado pelos resultados de aplicação dos métodos sobre as sequências de código das espécies *Haemophilus influenzae*, *Escherichia coli K12* e *Schizosaccharomyces pombe* (ver tabela 5.3). Para a espécie *Saccharomyces cerevisiae*, o modelo de três estados nem sempre é melhor, sendo os resultados dependentes do cromossoma (ver tabela 5.4). Finalmente, para a *Arabidopsis thaliana*, o modelo de contexto finito de três estados levou a piores resultados que o DNACompress e o Dna3 (ver tabela 5.5).

O DNACompress e Dna3, tais como muitas das técnicas de compressão para DNA, baseiam-se em repetições de subsequências (exactas e aproximadas). Uma possível razão para que o modelo de três estados seja melhor nuns casos e pior noutros, é o facto das subsequências apresentarem diferentes graus de repetição. Obviamente que para sequências que apresentem uma estrutura repetitiva, o modelo de três estados não será tão bom como os modelos que têm em conta esse tipo de estrutura. Isto é o que acontece com o genoma da *Arabidopsis thaliana*. Na verdade, é frequente as plantas apresentarem repetição de DNA [120]. Na última coluna das tabelas 5.3, 5.4 e 5.5 é mostrada a percentagem de bases que foram codificadas por Dna3 usando como referência subsequências passadas. Como se pode observar, é na espécie *Arabidopsis thaliana* que as percentagens de repetição de todos os cromossomas está acima dos 10%. De modo geral, quando a percentagem de bases codificada usando a estratégia de repetição é superior a 3%, o modelo de três estados não consegue capturar toda a estrutura relevante dos dados. Portanto, se o objectivo for melhorar o desempenho dos métodos de compressão, então o modelo de três estados tem de ser complementado com os métodos disponíveis no sentido de explorar os padrões de repetição.

Os valores de débito binário médio ao longo dos três estados é outra informação de interesse que se consegue extrair das tabelas 5.3, 5.4 e 5.5. Para os resultados apresentados, o estado zero corresponde à primeira base do codão, o estado um corresponde à segunda base do codão e o estado dois corresponde à ultima base do codão. Os valores denotados por "bpb0", "bpb1" e "bpb2" indicam o número médio de *bits* necessários para o codificador representar a primeira, segunda e terceira bases do codão, respectivamente.

Para as espécies *Haemophilus influenzae*, *Schizosaccharomyces pombe* e *Arabidopsis thaliana*, a primeira base é a mais difícil de comprimir, segue-se a segunda e finalmente

a terceira (ver tabela 5.3). Do ponto de vista da teoria da informação, isto significa que a primeira base é a que contém mais informação do codão, seguida da segunda base e só depois da terceira. Esta conclusão pode não ser surpreendente, tendo em conta a natureza degenerativa do código genético (num grupo de sinónimos há codões que tendem a ser pouco usados face a outros).

Por outro lado, muitos aminoácidos podem ser representados por mais de um tripleto, sendo para alguns deles a terceira base irrelevante, o que parece contrariar a tendência observada nas espécies referidas na tabela 5.3. No entanto, para alguns cromossomas da *Saccharomyces cerevisiae*, verifica-se que a segunda base parece comportar menos informação que a terceira (ver tabela 5.4).

Embora o resultado seja marginal, na *Escherichia coli K12* a segunda base parece comportar pelo menos tanta informação como a primeira (ver tabela 5.3). Actualmente, não se tem explicação para este comportamento, mas acreditamos na existência de motivação biológica.

Contudo, como foi discutido anteriormente, a dependência do organismo não é totalmente inesperada (as preferências de codões variam largamente de espécie para espécie) e também é conhecido que muitos organismos apresentam maior teor de $\mathcal{C}+\mathcal{G}$ do que $\mathcal{A}+\mathcal{T}$, devido a ser mais difícil quebrar as ligações entre $\mathcal{C}$ e $\mathcal{G}$, o que pode levar a preferências na escolha de um codão num conjunto de sinónimos.

Na tentativa de explicar alguns comportamentos aparentemente sem explicação e realçar as características encontradas pelo modelo de três estados, calcularam-se valores de informação mútua ou indirectamente valores de entropia condicionada (ver tabelas 5.6 e 5.7). Estudou-se a informação mútua entre as três posições de uma sequência de código, supondo válida a factorização $N = nm$ e $n, m \in \mathbb{N}$, referida no capítulo anterior.

Definiu-se como medida de informação mútua relativa entre duas subsequências de bases associadas a duas posições do codão, v_1 e v_2, de uma mesma sequência

$$I(v_1, v_2) = \sum_{k_1,k_2\in\{a,c,g,t\}} P_{v_1,v_2}(k_1,k_2)\log_2\frac{P_{v_1,v_2}(k_1,k_2)}{P_{v_1}(k_1)P_{v_2}(k_2)},$$

com $P_{v_1,v_2}(k_1,k_2)$ as probabilidades conjuntas dos símbolos k_1 e k_2 condicionadas às posições v_1 e v_2 para $v_1, v_2 \in \{0,1,\ldots,m-1\}$,[6] respectivamente.

[6] À imagem do que foi feito no capítulo anterior, também podemos estabelecer uma relação entre

Definiu-se entropia condicionada da base da posição v_1 do codão ao conhecimento da base da posição v_2 do codão por

$$H(v_1|v_2) = H(v_1) - I(v_1, v_2),$$

com $v_1, v_2 \in \{0, 1, 2\}$.

Na tabela 5.6 apresentam-se resultados médios de informação mútua sobre um conjunto de genes. O cálculo da informação mútua é feito por gene, para os 9 arranjos (com repetição) possíveis com as três posições de um codão (ou seja $m = 3$). De notar que a medida de informação mútua goza de simetria e naturalmente a tabela 5.6 apresenta seis e não nove resultados de informação mútua. Finalmente calcula-se o valor médio sobre todos os genes que constituem determinada sequência, $\overline{I(v_1, v_2)}$. Observe-se que os valores $\overline{I(0,0)}$, $\overline{I(1,1)}$ e $\overline{I(2,2)}$, coincidem com a entropia média da sequência da respectiva posição.

Também na tabela 5.7 são apresentados resultados médios de entropia condicionada sobre o conjunto de genes que constitui cada sequência da tabela, onde a entropia média condicionada é dada por

$$\overline{H(v_1|v_2)} = \overline{H(v_1)} - \overline{I(v_1, v_2)} = \overline{I(v_1, v_1)} - \overline{I(v_1, v_2)},$$

com $v_1, v_2 \in \{0, 1, 2\}$.

Entre os resultados obtidos nas tabelas 5.6 e 5.7 e os resultados obtidos com o modelo de três estados, tabelas 5.3, 5.4 e 5.5, podem ser estabelecidas algumas ligações. Na

as probabilidades condicionadas e os contadores de símbolos.

Considere-se um contador conjunto de símbolos dado por:

$$y^{k_1,k_2}_{(v_1,v_2)} = \sum_{l=0}^{n-1} u^{k_1}_{(v_1+lm)} u^{k_2}_{(v_2+lm)}$$

com $k_1, k_2 \in \{a, c, g, t\}$, $v_1, v_2 \in \{0, 1, \ldots, m-1\}$ e u^k a sequência indicadora do símbolo k. Este contador conjunto conta o número de vezes que numa sequência de tamanho N surge um k_1 na posição v_1 e simultaneamente um k_2 na posição v_2. Assim pode-se escrever os contadores conjuntos de símbolos à custa das probabilidades conjuntas:

$$y^{k_1,k_2}_{(v_1,v_2)} = \frac{N}{m} P_{v_1,v_2}(k_1, k_2).$$

tabela 5.7 observa-se que para as espécies estudadas os valores mais baixos de entropia média condicionada são obtidos no caso da segunda base do codão condicionada ao conhecimento da primeira. Assim, na maior parte das espécies estudadas, podemos reparar que é mais fácil comprimir a segunda base do codão dado o conhecimento da primeira do que qualquer outra alternativa estudada. Por outras palavras, a primeira e a segunda base do codão constituem o par com maior informação mútua. O modelo de três estados geralmente não reflecte esta característica. Contudo, o estudo com o modelo de três estados é de maior alcance (profundidade) e a informação da terceira base pode ser dada parcialmente, por exemplo, por bases do codão anterior. No entanto, no caso *Saccharomyces cerevisiae*, os valores elevados obtidos para informação mútua entre a primeira e a segunda base são visíveis no modelo de três estados: os valores de entropia para a segunda base são mais baixos do que para as restantes bases (ver tabela 5.4).

Para a espécie *Escherichia coli K12* a informação mútua entre as duas últimas posições num gene, $I(1,2)$, é geralmente superior a qualquer outra. Esta característica também é visível no modelo de três estados: os valores de entropia para a terceira base são mais baixos do para as restantes bases (ver tabela 5.5).

De modo geral, observa-se que a informação mútua é maior entre a primeira e a segunda base, diminui da segunda para a terceira e da primeira para a terceira diminui ainda mais. Este resultado reflecte que quanto maior for a distância entre as bases menor é a informação mútua entre elas e que a informação mútua entre pares consecutivos decresce ao longo do codão.

Geralmente, os valores de entropia dados pela tabela 5.7 são inferiores aos obtidos pelos modelos das tabelas 5.3, 5.4 e 5.5, pois os primeiros são valores calculados de forma estática e os outros de forma adaptativa.

5.4 Conclusões

Neste capítulo, foi estudado o comportamento em termos de compressão do modelo de contexto finito de três estados sobre regiões codificantes de DNA em que se reconhece uma tendência para existir periodicidade três. Concluiu-se que o modelo de três estados

dá sempre melhor resultado que o modelo de contexto finito de apenas um estado para as regiões de código estudadas.

Mostrou-se também que, do ponto de vista da compressão, existe vantagem em ter em conta a periodicidade de período três nas regiões codificantes do DNA. Para alguns organismos testados, o modelo de contexto finito de três estados leva a melhores resultados que as melhores técnicas de compressão de DNA da actualidade. No entanto, realça-se mais uma vez que a técnica de compressão de DNA apresentada não é uma técnica completa de compressão, pois explora apenas uma propriedade de uma certa região de DNA. Uma das características mais relevantes do modelo de três estados é apresentar a possibilidade de analisar a informação distribuída pelas três bases do codão.

Os resultados obtidos com o modelo de três estados, em termos de débito binário ao longo dos três estados, foram confrontados com os resultados de informação mútua média e entropia condicionada média para pares de bases relativos a posições diferentes do codão.

Em termos de compressão, os resultados obtidos não são muito bons, uma vez que os débitos binários se situam pouco abaixo dos dois *bits* por símbolo. No entanto, este modelo ajudou a entender melhor a estrutura das regiões codificantes do DNA.

Referência	H	H^c	Referência	H	H^c
U00096:4638511-4639197, lasT	1,997	1,950	U00096:4637971-4638111, yjjY	1,971	1,878
U00096:c4637875-4637159, arcA	1,998	1,938	U00096:4635747-4637099, creD	1,990	1,960
U00096:4634265-4635689, creC	1,991	1,954	U00096:4633576-4634265, creB	1,990	1,950
U00096:4633090-4633563, creA	1,995	1,952	U00096:c4632879-4632010, rob	1,997	1,965
U00096:4631366-4632013, gpmB	1,985	1,961	U00096:c4631323-4630802, yjjX	1,983	1,930
U00096:4630329-4630655, trpR	1,989	1,945	U00096:4628275-4630239, slt	1,991	1,958
U00096:c4628091-4626424, yjjK	1,990	1,932	U00096:4624863-4626116, nadR	1,995	1,941
U00096:4623481-4624863, sms	1,977	1,933	U00096:4622464-4623432, serB	1,984	1,928
U00096:c4622358-4621714, smp	1,985	1,937	U00096:c4621686-4620670, lplA	1,981	1,939
U00096:4619338-4620669, yjjJ	1,992	1,972	U00096:4618452-4619171, deoD	1,998	1,908
U00096:4617172-4618395, deoB	1,994	1,938	U00096:4615798-4617120, deoA	1,991	1,925
U00096:4614892-4615671, deoC	1,986	1,912	U00096:c4614634-4613084, yjjI	1,996	1,965
U00096:c4613112-4612249, yjjW	1,981	1,944	U00096:4611194-4611829, yjjV	1,988	1,926
U00096:4609980-4611053, yjjU	1,994	1,967	U00096:4608965-4609570, osmY	1,980	1,869
U00096:4606983-4608572, prfC	1,994	1,924	U00096:4606215-4606892, yjjG	1,996	1,946
U00096:4605754-4606200, rimI	1,996	1,956	U00096:4605372-4605785, holD	1,981	1,958
U00096:c4605269-4604238, yjjT	1,988	1,949	U00096:c4603232-4602444, fhuF	1,997	1,963
U00096:4601729-4602406, bglJ	1,989	1,974	U00096:4601046-4601771, yjjQ	1,992	1,981
U00096:c4600490-4599657, yjjP	1,999	1,973	U00096:c4599519-4599193, yjjB	1,982	1,933

Tabela 5.2: Resultados relativos à entropia e à entropia ciclo-estacionária de alguns genes da *Escherichia coli*.

Haemophilus influenzae												
Referência	Sequência	Bases	Três estados					Um estado		*DnaC*	*Dna3*	
			M	bpb0	bpb1	bpb2	bpb	*M*	bpb	bpb	bpb	Repetição
GI:16271976	—	1 505 271	4	1,918	1,834	1,684	**1,812**	5	1,889	1,902	1,895	0,9%

Escherichia coli K12												
Referência	Sequência	Bases	Três estados					Um estado		*DnaC*	*Dna3*	
			M	bpb0	bpb1	bpb2	bpb	*M*	bpb	bpb	bpb	Repetição
GI:49175990	—	4 083 231	5	1,897	1,898	1,750	**1,848**	6	1,917	1,920	1,913	1,3%

Schizosaccharomyces pombe												
Referência	Sequência	Bases	Três estados					Um estado		*DnaC*	*Dna3*	
			M	bpb0	bpb1	bpb2	bpb	*M*	bpb	bpb	bpb	Repetição
GI:19113674	Chr-I	2 996 109	4	1,961	1,884	1,820	**1,889**	4	1,939	1,918	1,921	1,3%
GI:19111836	Chr-II	2 399 394	4	1,962	1,887	1,818	**1,889**	4	1,940	1,915	1,916	1,7%
GI:19075172	Chr-III	1 169 991	3	1,961	1,889	1,833	**1,895**	4	1,943	1,925	1,930	1,2%

Tabela 5.3: Resultados da compressão, em *bits* por base (bpb), obtidos para três organismos: *Haemophilus influenzae*, *Escherichia coli K12* e *Schizosaccharomyces pombe*. Para o modelo de três estados, as colunas "bpb0", "bpb1" e "bpb2" indicam débitos binários médios por estado, a coluna"bpb" indica débitos binários médios totais. A coluna "*M*" indica a ordem do modelo para a qual se obtiveram os melhores resultados. As colunas "*DnaC*" e "*Dna3*" mostram os resultados de compressão usando os métodos *DNACompress* e *Dna3* respectivamente.

Saccharomyces cerevisiae												
Referência	Sequência	Bases	Três estados					Um estado		*DnaC*	*Dna3*	
			M	bpb0	bpb1	bpb2	bpb	*M*	bpb	bpb	bpb	Repetição
GI:50593113	Chr-I	143 157	2	1,937	1,882	1,909	1,911	3	1,954	**1,884**	1,910	3,4%
GI:50593115	Chr-II	605 184	3	1,936	1,869	1,886	**1,897**	3	1,942	1,912	1,918	1,7%
GI:42759850	Chr-III	217 332	2	1,946	1,874	1,908	**1,911**	3	1,951	1,918	1,923	1,8%
GI:50593138	Chr-IV	1 129 605	3	1,931	1,856	1,882	1,890	4	1,936	**1,846**	1,853	5,0%
GI:7276232	Chr-V	391 086	3	1,935	1,872	1,894	1,901	3	1,947	**1,883**	1,894	3,3%
GI:42742172	Chr-VI	183 702	2	1,938	1,863	1,904	**1,904**	3	1,949	1,932	1,939	1,0%
GI:50593213	Chr-VII	784 707	3	1,935	1,861	1,882	**1,893**	3	1,939	1,897	1,902	2,2%
GI:50882583	Chr-VIII	402 792	3	1,938	1,873	1,896	**1,903**	3	1,946	1,907	1,915	2,1%
GI:6322016	Chr-IX	310 041	3	1,938	1,869	1,900	**1,903**	3	1,947	1,933	1,942	0,9%
GI:42742252	Chr-X	557 103	3	1,935	1,866	1,892	**1,899**	3	1,943	1,907	1,914	1,8%
GI:50593424	Chr-XI	478 620	3	1,935	1,855	1,893	**1,895**	3	1,940	1,938	1,942	0,3%
GI:42742286	Chr-XII	784 695	3	1,936	1,862	1,893	1,898	3	1,942	**1,863**	1,872	4,1%
GI:44829554	Chr-XIII	693 291	3	1,934	1,859	1,889	1,894	3	1,940	**1,886**	1,892	3,0%
GI:50593505	Chr-XIV	576 585	3	1,937	1,869	1,893	**1,900**	3	1,944	1,930	1,934	0,9%
GI:42742309	Chr-XV	785 568	3	1,937	1,865	1,887	**1,897**	3	1,941	1,901	1,917	2,2%
GI:50593503	Chr-XVI	687 666	3	1,937	1,862	1,887	1,896	3	1,941	1,889	**1,880**	3,6%
GI:6226515	MT	24 429	2	1,814	1,767	1,305	1,643	3	1,747	**1,466**	1,511	16,6%

Tabela 5.4: Resultados da compressão, em *bits* por base (bpb), para a *Saccharomyces cerevisiae*.

Arabidopsis thaliana												
Referência	Sequência	Bases	Três estados					Um estado		*DnaC*	*Dna3*	
			M	bpb0	bpb1	bpb2	bpb	*M*	bpb	bpb	bpb	Repetição
GI:42592260	Chr-I	9 595 494	5	1,925	1,904	1,882	1,904	6	1,939	**1,725**	1,743	12,0%
GI:30698031	Chr-II	5 474 178	4	1,926	1,906	1,886	1,906	6	1,942	**1,710**	1,737	12,0%
GI:30698537	Chr-III	7 183 863	5	1,925	1,904	1,886	1,905	6	1,941	**1,736**	1,762	10,8%
GI:30698542	Chr-IV	5 572 038	4	1,926	1,905	1,888	1,906	6	1,942	**1,708**	1,740	12,1%
GI:30698605	Chr-V	8 462 424	5	1,924	1,902	1,883	1,903	6	1,939	**1,736**	1,759	10,9%

Tabela 5.5: Resultados da compressão, em *bits* por base (bpb), para a *Arabidopsis thaliana*.

Haemophilus influenzae							
Referência	Sequência	$\overline{I(0,0)}$	$\overline{I(1,1)}$	$\overline{I(2,2)}$	$\overline{I(0,1)}$	$\overline{I(1,2)}$	$\overline{I(0,2)}$
GI:16271976	—	1,936	1,909	1,837	**0,160**	0,136	0,038
Escherichia coli K12							
Referência	Sequência	$\overline{I(0,0)}$	$\overline{I(1,1)}$	$\overline{I(2,2)}$	$\overline{I(0,1)}$	$\overline{I(1,2)}$	$\overline{I(0,2)}$
GI:49175990	—	1,923	1,942	1,948	0,089	**0,130**	0,005
Schizosaccharomyces pombe							
Referência	Sequência	$\overline{I(0,0)}$	$\overline{I(1,1)}$	$\overline{I(2,2)}$	$\overline{I(0,1)}$	$\overline{I(1,2)}$	$\overline{I(0,2)}$
GI:19113674	Chr-I	1,953	1,919	1,875	**0,123**	0,067	0,027
GI:19111836	Chr-II	1,955	1,920	1,875	**0,125**	0,067	0,025
GI:19075172	Chr-III	1,956	1,922	1,887	**0,123**	0,070	0,028
Saccharomyces cerevisiae							
Referência	Sequência	$\overline{I(0,0)}$	$\overline{I(1,1)}$	$\overline{I(2,2)}$	$\overline{I(0,1)}$	$\overline{I(1,2)}$	$\overline{I(0,2)}$
GI:50593113	Chr-I	1,922	1,906	1,939	**0,121**	0,065	0,041
GI:50593115	Chr-II	1,925	1,912	1,940	**0,139**	0,056	0,038
GI:42759850	Chr-III	1,929	1,907	1,937	**0,133**	0,056	0,042
GI:50593138	Chr-IV	1,921	1,900	1,933	**0,143**	0,054	0,036
GI:7276232	Chr-V	1,923	1,910	1,938	**0,138**	0,052	0,035
GI:42742172	Chr-VI	1,920	1,901	1,937	**0,130**	0,055	0,036
GI:50593213	Chr-VII	1,926	1,906	1,938	**0,141**	0,053	0,036
GI:50882583	Chr-VIII	1,928	1,908	1,941	**0,139**	0,054	0,037
GI:6322016	Chr-IX	1,928	1,906	1,943	**0,135**	0,055	0,038
GI:42742252	Chr-X	1,927	1,906	1,937	**0,131**	0,056	0,038
GI:50593424	Chr-XI	1,925	1,901	1,934	**0,141**	0,053	0,035
GI:42742286	Chr-XII	1,929	1,906	1,938	**0,138**	0,057	0,040
GI:44829554	Chr-XIII	1,924	1,901	1,938	**0,141**	0,053	0,035
GI:50593505	Chr-XIV	1,924	1,905	1,939	**0,139**	0,054	0,034
GI:42742309	Chr-XV	1,923	1,902	1,934	**0,141**	0,053	0,037
GI:50593503	Chr-XVI	1,930	1,907	1,939	**0,141**	0,053	0,038
Arabidopsis thaliana							
Referência	Sequência	$\overline{I(0,0)}$	$\overline{I(1,1)}$	$\overline{I(2,2)}$	$\overline{I(0,1)}$	$\overline{I(1,2)}$	$\overline{I(0,2)}$
GI:42592260	Chr-I	1,944	1,941	1,948	**0,116**	0,053	0,028
GI:30698031	Chr-II	1,941	1,940	1,944	**0,117**	0,056	0,031
GI:30698537	Chr-III	1,942	1,940	1,949	**0,115**	0,053	0,028
GI:30698542	Chr-IV	1,943	1,942	1,946	**0,115**	0,052	0,028
GI:30698605	Chr-V	1,944	1,940	1,946	**0,119**	0,054	0,028

Tabela 5.6: Informação mútua média para as três posições das sequências de código em várias espécies.

Haemophilus influenzae							
Referência	Sequência	$\overline{H(0\|1)}$	$\overline{H(0\|2)}$	$\overline{H(1\|0)}$	$\overline{H(1\|2)}$	$\overline{H(2\|0)}$	$\overline{H(2\|1)}$
GI:16271976	—	1,78	1,90	1,75	1,77	1,80	**1,70**
Escherichia coli K12							
Referência	Sequência	$\overline{H(0\|1)}$	$\overline{H(0\|2)}$	$\overline{H(1\|0)}$	$\overline{H(1\|2)}$	$\overline{H(2\|0)}$	$\overline{H(2\|1)}$
GI:49175990	—	1,83	1,92	1,85	**1,81**	1,94	1,82
Schizosaccharomyces pombe							
Referência	Sequência	$\overline{H(0\|1)}$	$\overline{H(0\|2)}$	$\overline{H(1\|0)}$	$\overline{H(1\|2)}$	$\overline{H(2\|0)}$	$\overline{H(2\|1)}$
GI:19113674	Chr-I	1,83	1,93	**1,80**	1,85	1,85	1,81
GI:19111836	Chr-II	1,83	1,93	**1,80**	1,85	1,85	1,81
GI:19075172	Chr-III	1,83	1,93	**1,80**	1,85	1,86	1,82
Saccharomyces cerevisiae							
Referência	Sequência	$\overline{H(0\|1)}$	$\overline{H(0\|2)}$	$\overline{H(1\|0)}$	$\overline{H(1\|2)}$	$\overline{H(2\|0)}$	$\overline{H(2\|1)}$
GI:50593113	Chr-I	1,80	1,88	**1,78**	1,84	1,90	1,87
GI:50593115	Chr-II	1,79	1,89	**1,77**	1,86	1,90	1,88
GI:42759850	Chr-III	1,80	1,89	**1,77**	1,85	1,90	1,88
GI:50593138	Chr-IV	1,78	1,89	**1,76**	1,85	1,90	1,88
GI:7276232	Chr-V	1,78	1,89	**1,77**	1,86	1,90	1,89
GI:42742172	Chr-VI	1,79	1,88	**1,77**	1,85	1,90	1,88
GI:50593213	Chr-VII	1,78	1,89	**1,76**	1,85	1,90	1,89
GI:50882583	Chr-VIII	1,79	1,89	**1,77**	1,85	1,90	1,89
GI:6322016	Chr-IX	1,79	1,89	**1,77**	1,85	1,91	1,89
GI:42742252	Chr-X	1,80	1,89	**1,77**	1,85	1,90	1,88
GI:50593424	Chr-XI	1,78	1,89	**1,76**	1,85	1,90	1,88
GI:42742286	Chr-XII	1,79	1,89	**1,77**	1,85	1,90	1,88
GI:44829554	Chr-XIII	1,78	1,89	**1,76**	1,85	1,90	1,88
GI:50593505	Chr-XIV	1,79	1,89	**1,77**	1,85	1,90	1,88
GI:42742309	Chr-XV	1,78	1,89	**1,76**	1,85	1,90	1,88
GI:50593503	Chr-XVI	1,79	1,89	**1,77**	1,85	1,90	1,89
Arabidopsis thaliana							
Referência	Sequência	$\overline{H(0\|1)}$	$\overline{H(0\|2)}$	$\overline{H(1\|0)}$	$\overline{H(1\|2)}$	$\overline{H(2\|0)}$	$\overline{H(2\|1)}$
GI:42592260	Chr-I	**1,83**	1,92	**1,83**	1,89	1,92	1,89
GI:30698031	Chr-II	**1,82**	1,91	**1,82**	1,88	1,91	1,89
GI:30698537	Chr-III	1,83	1,91	**1,82**	1,89	1,92	1,90
GI:30698542	Chr-IV	**1,83**	1,91	**1,83**	1,89	1,92	1,89
GI:30698605	Chr-V	1,83	1,92	**1,82**	1,89	1,92	1,89

Tabela 5.7: Entropia condicionada média para as três posições das sequências de código em várias espécies.

Capítulo 6

Conclusões e trabalho futuro

Ao longo deste trabalho apresentaram-se metodologias originais para o estudo de algumas estruturas de correlação em sequências simbólicas, em particular sequências de nucleótidos.

Neste capítulo apresentam-se as principais conclusões deste trabalho e são mencionadas algumas direcções de trabalho futuro.

6.1 Conclusões

Nesta dissertação contribuiu-se para melhorar metodologias de localização de genes em sequências de DNA e compressão de genomas. Mais concretamente:

- Compararam-se vários tipos de métodos de análise espectral que têm sido aplicados a sequências simbólicas, em particular às sequências de DNA. Ainda com o propósito de análise espectral de sequências simbólicas foi proposto outro método de análise baseado numa função de autocorrelação natural para este tipo de dados, designada por autocorrelação simbólica. A autocorrelação simbólica não recorre a mapeamentos e consiste numa sequência numérica, onde a sua transformada de Fourier coincide com o espectro de dados simbólicos. Por outro lado, mostrou-se que este espectro pode ser obtido através da soma dos quadrados dos módulos das transformadas de Fourier das sequências indicadoras (sequências

zero/um indicando a posição dos símbolos).

Foi explorado o conceito de envolvente espectral, em que são usados vários mapeamento de símbolos para números de modo a que a energia espectral seja máxima para cada valor da frequência. O espectro obtido também conduz ao espectro de dados simbólicos.

Concluiu-se a equivalência de resultados entre alguns métodos sem aparente relação, nomeadamente entre os métodos baseados em sequências indicadoras, em tetraedros regulares, na autocorrelação simbólica e na envolvente espectral.

De todos os método estudados, o que usa as sequências indicadoras é o mais simples, em que o espectro total, pode ser entendido como o espectro de máxima energia espectral ou como a transformada da autocorrelação simbólica.

- Apresentaram-se métodos rápidos para cálculo de certas riscas espectrais, evidenciando algumas propriedades que relacionam o tamanho da risca com a distribuição dos símbolos em determinadas posições da sequência.

  Reduziu-se o cálculo da DFT de comprimento $N = mn$, correspondente ao espectro total, ao cálculo das DFTs de comprimento m de somas de blocos. Ou seja $S(nk) = |Y_k^a|^2 + |Y_k^c|^2 + |Y_k^g|^2 + |Y_k^t|^2$, com $0 \leq k < m$.

  Dado o interesse da risca espectral $S(N/3)$ na localização de genes, foi dada especial importância ao estudo desta risca. Concluiu-se que, para uma sequência de comprimento múltiplo de três ($N = 3n$), a risca espectral pode ser dada por $S(N/3) = |Y_1^a|^2 + |Y_1^c|^2 + |Y_1^g|^2 + |Y_1^t|^2$ com

  $$|Y_1^b| = \left[x - \frac{y+z}{2}\right]^2 + \frac{3}{4}[y-z]^2,$$

  onde x, y e z contam o número de símbolos $\mathcal{B}$ nas três posições dos n blocos. Feito o cálculo dos contadores de símbolos, que pode ser feito aquando da leitura dos dados, o número de operações aritméticas necessárias para calcular $S(N/3)$ é $O(1)$. Comparativamente o cálculo da FFT ("Fast Fourier Transform") de comprimento N é um processo $O(N \log N)$, e o cálculo de um coeficiente espectral requer $O(N)$ operações aritméticas.

  Discutiram-se algumas razões relativas à distribuição dos símbolos para a evidência da risca espectral $S(N/3)$. Também foram apresentadas expressões de cálculo de

outras riscas espectrais, em particular das riscas $S(N/4)$ e $S(N/6)$. E mostrou-se que o espectro de uma sequência de símbolos contém informação redundante.

- Desenvolveu-se uma técnica de compressão baseada na periodicidade de período três, especialmente dedicada a regiões codificantes, designada por modelo de três estados. Nestas regiões, em alguns genomas, foram obtidos melhores resultados do que os conseguidos pelas melhores técnicas de compressão da actualidade. A compressão de sequências genéticas, assim como o estudo da entropia e informação mútua, foram também usados com o propósito de obter modelos ou simplesmente indicar características particulares que descrevam as sequências de DNA.

  Concluímos que o modelo de três estados dá sempre melhor resultado que o modelo de contexto finito de apenas um estado para as regiões de código estudadas. Uma das características mais relevantes do modelo de três estados é apresentar a possibilidade de analisar a informação distribuída pelas três bases do codão.

  O modelo de contexto finito de três estados apresentado ajudou a entender melhor a estrutura das regiões codificantes do DNA. No entanto, a técnica de compressão apresentada não é uma técnica completa de compressão de DNA, pois explora apenas uma propriedade de uma certa região de DNA. Genericamente as técnicas actuais de compressão não resultam em bons resultados de compressão, uma vez que os débitos binários se situam pouco abaixo dos dois *bits* por símbolo.

6.2 Trabalho futuro

Pretende-se continuar a pesquisa de estruturas de correlação nas sequências de DNA, criando ou melhorando métodos de análise de sequências simbólicas. Em particular, pretende-se pesquisar métodos que tenham em conta a natureza simbólica e o comportamento não estacionário dos dados genéticos.

Para um futuro próximo, existe o objectivo de encontrar métodos de compressão que apresentem nitidamente melhores resultados do que os melhores métodos actuais. Para tal, na linha do que foi feito no capítulo 5, pode ser ainda mais explorada a redundância

estatística da sequência usando a redundância associada ao conhecimento que actualmente se tem ou se possa adquirir sobre a estrutura do DNA.

Finalmente, na colaboração com o grupo de Bioinformática da Universidade de Aveiro, pretende-se desenvolver ou aplicar modelos que vão ao encontro das problemáticas estudadas pelo grupo, em particular ao nível dos modelos matemáticos para análise de *microarrays*.

Bibliografia

[1] C. Acquisti, P. Allegrini, P. Bogani, E. Catanese, L. Fronzoni, P. Grigolini, G. Mersi, and L. Palatella. In the search for the low-complexity sequences in prokaryotic and eukaryotic genomes: How to derive a coherent picture from global and local entropy measures. *Chaos, Solitons and Fractals*, 20(1):127–137, 2004.

[2] Vera Afreixo and Adelaide Freitas. Uma análise estatística das sequências de DNA. *Literacia e Estatística — Actas do X Congresso Anual da Sociedade Portuguesa de Estatística, Edições SPE*, pages 91–98, 2003.

[3] Vera Afreixo, Adelaide V. Freitas, M. Pinheiro, José L. Oliveira, G. Moura, and Manuel A. S. Santos. Exploiting a biclustering algorithm in ORFeome analysis. In *2007 VLDB Workshop on Data Mining in Bioinformatics*, Vienna, September 2007.

[4] Vera M. A. Afreixo, Paulo J. S. G. Ferreira, and Dorabella M. S. Santos. Fourier analysis of symbolic data: A brief review. *Digital Signal Processing*, 14(6):523–530, November 2004.

[5] Vera M. A. Afreixo, Paulo J. S. G. Ferreira, and Dorabella M. S. Santos. The spectrum and symbol distribution of nucleotide. *Physical Review E*, 70(3):031910, September 2004.

[6] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell. 4th ed.* Garland Publishing, 2002.

[7] Paolo Allegrini, M. Barbi, Paolo Grigolini, and Bruce J. West. Dynamical model for DNA sequences. *Physical Review E*, 52(5):5281–5296, 1995.

[8] Paolo Allegrini, Marco Buiatti, Paolo Grigolini, and Bruce J. West. Fractional brownian motion as a nonstationary process: An alternative paradigm for DNA sequences. *Physical Review E*, 57(4):4558–4567, April 1998.

[9] Lloyd Allison, Timothy Edgoose, and Trevor I. Dix. Compression of strings with approximate repeats. In *Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology*, pages 8–16, Montréal, Québec, Canada, June 1998.

[10] Dimitris Anastassiou. Digital signal processing of biomolecular sequences. Technical report, Department of electrical engineering, 2000.

[11] Dimitris Anastassiou. Frequency-domain analysis of biomolecular sequences. *Bioinformatics*, 16(12):1073–1081, July 2000.

[12] Dimitris Anastassiou. Genomic signal processing. *IEEE Signal Processing Magazine*, 18(4):8–20, July 2001.

[13] A. Arnéodo, E. Bacry, P.V. Graves, and J.F. Muzy. Characterizing long-range correlations in DNA sequences from wavelet analysis. *Physical Review Letters*, 74(16):3293–3296, April 1995.

[14] A. Arnéodo, Y. d'Aubenton Carafa, B. Audit, E. Bacry, J.F. Muzy, and C. Thermes. Nucleotide composition effects on the long-range correlations in human genes. *The European Physical Jounal B*, 1:259–263, February 1998.

[15] A. Arnéodo, Y. d'Aubenton Carafa, B. Audit, E. Bacry, J.F. Muzy, and C. Thermes. What can we learn with wavelets about DNA sequences? *Physica A*, 249:439–448, January 1998.

[16] B. Audit, C. Thermes, C. Vaillant, Y. d'Aubenton Carafa, J.F. Muzy, and A. Arneodo. Long-range correlations in genomic DNA: A signature of the nucleosomal structure. *Physical Review Letters*, 86(11):2471–2474, March 2001.

[17] Benjamin Audit, Cédric Vaillant, Alain Arnéodo, Yves D'Aubenton-Carafa, and Claude Thermes. Wavelet analysis of DNA bending profiles reveals structural constraints on the evolution of genomic sequences. *Journal of Biological Physics*, 30:33–81, 2004.

[18] Mark Ya. Azbel. Universality in a DNA statistical structure. *Physical Review Letters*, 75(1):168–171, 1995.

[19] J. H. Badger and G. J. Olsen. CRITICA: Coding region identification tool invoking comparative analysis. *Molecular Biology and Evolution*, 16:512–524, 1999.

[20] Pierre Baldi, Yves Chauvin, Tim Hunkapiller, and Marcella A. Mcclureii. Hidden Markov models of biological primary sequence information. *Proceedings of the National Academy of Sciences of the U.S.A.*, 21:1059–1063, 1994.

[21] P.J. Barral, A. Hasmy, J. Jiménez, and A. Marcano. Nonlinear modeling technique for the analysis of DNA chains. *Physical Review E*, 61(2):1812–1815, February 2000.

[22] Behshad Behzadi and Fabrice Le Fessant. DNA compression challenge revisited. In *Proceedings of CPM*, 2005.

[23] Timothy C. Bell, John G. Cleary, and Ian H. Witten. *Text Compression*. Prentice Hall advanced reference series computer science, 1990.

[24] Gary Benson and Michael S. Waterman. A method for fast database search for all k-nucleotide repeats. *Nucleic Acids Research*, 22(22):4828–4836, 1994.

[25] John A. Berger, Sanjit K. Mitra, Marco Carli, and Alessandro Neri. Visualization and analysis of DNA sequences using DNA walks. *Journal of the Franklin Institute*, 341:37–53, 2004.

[26] Pedro Bernaola-Galván and Pedro Carpena. Comment on "factorial moments analyses show a characteristic length scale in DNA sequences". *Physical Review Letters*, 88(21):219803, 2002.

[27] Pedro Bernaola-Galván, Pedro Carpena, Ramón Ramón-Roldán, and Jose Oliver. Study of statistical correlations in DNA sequences. *Gene*, 300:105–115, 2002.

[28] Pedro Bernaola-Galván, Ivo Grosse, Pedro Carpena, and José L. Oliver. Finding borders between coding and noncoding DNA regions by an entropic segmentation method. *Physical Review Letters*, 85(6):1342–1345, 2000.

[29] Pedro Bernaola-Galván, José Oliver, and Ramón Ramón-Roldán. Decomposition of DNA sequence complexity. *Physical Review Letters*, 83(16):3336–3339, 1999.

[30] Pedro Bernaola-Galván, Ramón Ramón-Roldán, and José Oliver. Compositional segmentation and long-range fractal correlations in DNA sequences. *Physical Review E*, 53(5):5181–5189, May 1996.

[31] J. Besemer, A. Lomsadze, and M. Borodovsky. GeneMarkS: A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Research*, 29(12):2607–2618, 2001.

[32] Charles G. Boncelet Jr. A rearranged DFT algorithm requiring $n^2/6$ multiplications. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pages 1658–1659, December 1986.

[33] D. Bourchard, R. Schmidt, and H. Waller. The block sum transformation — A new special discrete Fourier transformation method for the analysis of vibrations. *Mechanical Systems and Signal Processing*, 6(5):483–489, 1992.

[34] S. Boycheva, G. Chkodrov, and I. Ivanov. Codon pairs in genome of *Escherichia coli*. *Bioinformatics*, 19:987–998, 2002.

[35] Jerome V. Braun and Hans-Georg Muller. Statistical methods for DNA sequence segmentation. *Statistical Science*, 13(2):142–162, 1998.

[36] Marc Buchner and Suparerk Janjarasjitt. Dectetion and visualization of tandem repeats in DNA sequences. *IEEE Transactions on Signal Processing*, 51(9):2280–2287, 2003.

[37] Sergey V. Buldyrev, Ary L. Goldberger, Shlomo Havlin, Chung-Kang Peng, Michael Simons, and H. Eugene Stanley. Generalized Lévy-walk model for DNA nucleotide sequences. *Physical Review E*, 47(6):4514–4523, June 1993.

[38] S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, and M.E. Matsa. Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. *Physical Review E*, 51(5):5084–5091, May 1995.

[39] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268:78–94, 1997.

[40] Shi-Min Cai, Pei-Ling Zhou, Hui-Jie Yang, Tao Zhou, Bing-Hong Wang, and Fang-Cui Zhao. Diffusion entropy analysis on the stride interval flutuation of human gait. *Physica A*, 375:687–692, 2007.

[41] S. Cebrat, M.R. Dudek, A. Gierlik, and M. Kowalczuk. Effect of replication on the third base of codons. *Physica A*, 265:78–84, 1999.

[42] X. Chen, S. Kwong, and M. Li. A compression algorithm for DNA sequences. *IEEE-EMB Special Issue on Bioinformatics*, 20(4):61–66, 2001.

[43] X. Chen, M. Li, B Ma, and J. Tromp. DNACompress: Fast and effective DNA sequence compression. *Bioinformatics*, 18(12):1696–1698, December 2002.

[44] Thomas M. Cover and Joy A. Thomas. *Elements of information theory.* John Wiley and Sons, 2006.

[45] Eivind Coward. Equivalence of two Fourier methods for biological sequences. *Journal of Mathematical Biology*, 36:64–70, 1997.

[46] Eivind Coward and Finn Drablos. Detecting periodic patterns in biological sequences. *Bioinformatics*, 14(6):498–507, 1998.

[47] Francis Crick. Central dogma of molecular biology. *Nature*, 227:561–563, 1970.

[48] Paul Dan Cristea. Large scale features in DNA genomic signals. *Signal Processing*, 83:871–888, 2003.

[49] R.N. Curnow and T.B.L. Kirkwood. Statistical analysis of deoxyribonucleic acid sequence data — A review. *Jounal of Royal Statistical Society. Series A*, 152:199–220, 1989.

[50] N. Dasgupta, S. Lin, and L. Carin. Sequential modeling for identifying CpG island locations in human genome. *IEEE Signal Processing Letters*, 9(12):407–409, 2001.

[51] M. Dehnert, W.E. Helm, and Hutt M.-Th. A discrete autoregressive process as a model for short-range correlations in DNA sequences. *Physica A*, 327:535–553, 2003.

[52] C. Dennis and C. Surridge. *A. thaliana* genome. *Nature*, 408:791, 2000.

[53] N. V. Dokholyan, S. V. Buldyrev, S. Havlin, and H. E. Stanley. Distribution of base pair repeats in coding and noncoding DNA sequences. *Physical Review Letters*, 79(25):5182–5185, 1997.

[54] W. Ebeling, A. Neiman, and T. Poschel. Dynamic entropies, long-range correlations and fluctations in complex linear structures. In *Proceedings of Coherent Approach to Fluctuations, World Scientific*, 1995.

[55] Stephen T. Eskesen, Frank N. Eskesen, Brian Kinghorn, and Anatoly Ruvinsky. Periodicity of DNA in exons. *BMC Molecular Biology*, 5(12), 2004.

[56] Alexei Fedorov, Serge Saxonov, and Walter Gilbert. Regularities of context-dependent codon bias in eukaryotic genes. *Nucleic Acids Research*, 30(5):1192–1197, 2002.

[57] Paulo J. S. G. Ferreira, António J. R. Neves, Vera Afreixo, and Armando J. Pinho. Exploring three-base periodicity for DNA compression and modeling. In *ICASSP 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, May 2006.

[58] Paulo Jorge S. G. Ferreira. Letter to the editor. A note on the block sum transformation. *Mechanical Systems and Signal Processing*, 7(2):191–192, 1993.

[59] Gloria R. Franco, Mark D. Adams, M. Bento Soares, Andrew J. G. Simpson, J. Craig Venter, and Sergio D. J. Pena. Identification of new schistosoma mansoni genes by the EST strategy using a directional cDNA library. *Gene*, 152:141–147, 1995.

[60] Adelaide V. Freitas, Miguel Pinheiro, Vera Afreixo, Júlia Duarte, José L. Oliveira, G. Moura, and Manuel Santos. A median-based iterative signature algorithm. In *Statistics for Data Mining, Learning and Knowledge Extraction, IASC 2007*, Aveiro, August 2007.

[61] Anders Fuglsang. Patterns of context-dependent codon biases. *Biochemical and Biophysical Research Communications*, 304:86–90, 2003.

[62] Atsushi Fukushima, Toshimichi Ikemura, Makoto Kinouchi, Taku Oshima, Yoshihiro Kudo, Hirotada Mori, and Shigehiko Kanaya. Periodicity in prokaryotic and eukaryotic genomes identified by power spectrum analysis. *Gene*, 300:203–211, 2002.

[63] J.B. Gao, Yinhe Cao, and Jae-Min Lee. Principal component analysis of $1/f^{\alpha}$ noise. *Physics Letters A*, 314:392–400, 2003.

[64] Derek Gatherer and Neil R. McEwan. Analysis of sequence periodicity in *E. coli* proteins: Empirical investigation of the "duplication and divergence" theory of protein evolution. *Journal of Molecular Evolution*, 57:149–158, 2003.

[65] Ivo Grosse, Pedro Bernaola-Galván, Pedro Carpena, Ramón Ramón-Roldán, Jose Oliver, and H. Eugene Stanley. Analysis of simbolic sequences using the Jensen-Shannon divergence. *Physical Review E*, 65:041905, 2002.

[66] Ivo Grosse, Sergey V. Buldyrev, and Eugene Stanley. Average mutual information of coding and noncoding DNA. In *Pacific Symposium on Biocomputing*, 2002.

[67] Ivo Grosse, Hanspeter Herzel, Sergey V. Buldyrev, and Eugene Stanley. Species independence of mutual information in coding and noncoding DNA. *Physical Review E*, 61(5):5624–5629, May 2000.

[68] S. Grumbach and F. Tahi. Compression of DNA sequences. In *Proceedings of the Data Compression conference*, pages 340–350, Snowbird, Utah, 1993.

[69] X. Guan, R.J. Mural, J.R. Einstein, R.C. Mann, and E.C. Uberbacher. GRAIL: An integrated artificial intelligence system for generecognition and interpretation.

In *Proceedings of the Eighth Conference on Artificial Intelligence for Applications*, pages 9–13, Monterey, CA, USA, 1992.

[70] Sabyasachi Guharay, Brian R. Hunt, James A. Yorke, and Owen R. White. Correlations in DNA sequences across the three domains of life. *Physica D*, 146:388–396, 2000.

[71] H. Herzel, E.N. Trifonov, O. Weiss, and I. Grosse. Interpreting correlations in biosequences. *Physica A*, 249:449–459, 1998.

[72] H. Herzel, O. Weiss, and E.N. Trifonov. 10-11 bp periodicities in complete genomes reflect protein struture and DNA folding. *Bioinformatics*, 15(3):187–193, 1999.

[73] Hanspeter Herzel, Werner Ebeling, and Armin O. Schmitt. Entropies of biosequences: The role of repeats. *Physical Review E*, 50(6):5061–5071, 1994.

[74] Hanspeter Herzel and Ivo Grosse. Measuring correlations in symbol sequences. *Physica A*, 216:518–542, 1995.

[75] Hanspeter Herzel and Ivo Grosse. Correlations in DNA sequences: The role of protein coding segments. *Physical Review E*, 55(1):800–810, January 1997.

[76] Dirk Holste and Ivo Grosse. Repeats and correlations in human DNA sequences. *Physical Review E*, 67:061913–1–7, 2003.

[77] Dirk Holste, Ivo Grosse, and Hanspeter Herzel. Statistical analysis of the DNA sequence of human chromosome 22. *Physical Review E*, 64:041917, 2001.

[78] S. Hooper and O. Berg. Detection of genes with atypical nucleotide sequence in microbial genomes. *Jounal of Molecular Evolution*, 54:365–375, 2002.

[79] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1990.

[80] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge : University Press, 1985.

[81] I. Kanter and Kessler. Markov processes: Linguistics and Zipf's law. *Physical Review Letters*, 74(22):4559–4562, May 1995.

[82] S. Karlin and V. Brendel. Patchiness and correlations in DNA sequences. *Science*, 259:677–680, 1993.

[83] Olga V. Kirillova. Comparative statistical analysis of bacteria genomes in "word" context. *Physica A*, 290:453–463, 2001.

[84] Gergely Korodi and Ioan Tabus. An efficient normalized maximum likelihood algorithm for DNA sequence compression. *ACM Transactions on Information Systems*, 23:3–34, 2005.

[85] Eugene V. Korotkov, Maria A. Korotkova, and Kudryashov N. A. Information decomposition method to analyze symbolical sequences. *Physics Letters A*, 312:198–210, June 2003.

[86] Eugene V. Korotkov, Maria A. Korotkova, F. E. Frenkel, and Kudryashov N. A. The informational concept of searching for periodicity in symbol sequences. *Molecular Biology*, 37(3):436–451, 2003.

[87] Eugene V. Korotkov, Maria A. Korotkova, Valentina M. Rudenko, and K.G. Skryabin. Latent periodicity regions in amino acid sequences. *Molecular Biology*, 33(4):611–617, 1999.

[88] Daniel Kotlar and Yizhar Lavner. Gene prediction by spectral rotation measure: A new method for identifying protein-coding regions. *Genome research*, 13:1930–1937, 2003.

[89] Lutz Krause, Alice C. McHardy, Tim W. Nattkemper, Alfred Puhler, Jens Stoye, and Folker Meyer. GISMO — Gene identification using a support vector machine for ORF classification. *Nucleic Acids Research*, 35(2):540–549, January 2007.

[90] A. Krishnamachari, Vijnan moy Mandal, and Karmeshu. Study of DNA binding sites using the Rényi parametric entropy measure. *Jornal of Theoretical Biology*, 227:429–436, 2004.

[91] Thomas Schou Larsen and Anders Krogh. EasyGene — A prokaryotic gene finder that ranks orfs by statistical significance. *BMC Bioinformatics*, 4(21), 2003.

[92] Weijiang Lee and Liaofu Luo. Periodicity of base correlation in nucleotide sequence. *Physical Review E*, 56(1):848–851, 1997.

[93] Wemtian Li, Gustavo Stolovitzky, Pedro Bernaola-Galván, and José L. Oliver. Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes. *Genome Research*, 8:916–928, 1998.

[94] Wentian Li. Mutual information functions versus correlation functions. *Journal of Statistical Physics*, 60:823–837, 1990.

[95] Wentian Li. Generating non-trivial long-range correlations and $1/f$ spectra by replication and mutation. *International Journal of Bifurcation and Chaos*, 2:137–154, 1992.

[96] Wentian Li. The complexity of DNA. *John Wiley and Sons. Inc.*, 3(2):33–37, 1997.

[97] Wentian Li. The study of correlation strutures of DNA sequences: A critical review. *Computers Chemistry*, 21(4):257–271, 1997.

[98] Wentian Li. Delineating relative homogeneous G+C domains in DNA sequences. *Gene*, 276:57–72, 2001.

[99] Wentian Li. Are isochore sequences homogeneous? *Gene*, 300(5):129–139, 2002.

[100] Wentian Li, Pedro Bernaola-Galván, Pedro Carpena, and Jose Oliver. Isochores merit the prefix 'iso'. *Computational Biology and Chemistry*, 27:5–10, 2003.

[101] Wentian Li, Pedro Bernaola-Galván, Fatameh Haghighi, and Ivo Grosse. Applications of recursive segmentation to the analysis of DNA sequences. *Computers and Chemistry*, 26:491–510, 2002.

[102] Wentian Li and Kunihiko Kaneko. Long-rang correlation and partial $1/f^{\alpha}$ spetrum in non-coding DNA sequence. *Europhysics Letters*, 17:655–660, 1992.

[103] Wentian Li, Thomas G. Marr, and Kunihiko Kaneko. Understanding long-range correlations in DNA sequences. *Physica D*, 75:392–416, 1994.

[104] Jianhua Lin. Divergence measure based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[105] S.R.C. Lopes and M.A. Nunes. Long memory analysis in DNA sequences. *Physica A*, 361(2):569–588, March 2006.

[106] Xin Lu, Zhirong Sun, Huimin Chen, and Yanda Li. Characterizing self-similarity in bacteria DNA sequences. *Physical Review E*, 58(3):3578–3584, September 1998.

[107] Alexander Lukashin and Mark Borodovsky. GeneMark.hmm: New solutions for gene finding. *Nucleic acids Research*, 26(4):1107–1115, 1998.

[108] Liaofu Luo and Weijiang Lee. Statistics correlation of nucleotides in DNA sequence. *Physical Review E*, 58(1):861–871, July 1998.

[109] P. Mackiewicz, A. Gierlik, M. Kowalczuk, D. Szczepanik, M.R. Dudek, and S. Cebrat. Mechanisms generating long-range correlation in nucleotide composition of the *Borrelia burgdorferi* genome. *Physica A*, 273:103–115, 1999.

[110] Shaun Mahony, James O McInerney, Terry J Smith, and Aaron Golden. Gene prediction using the self-organizing map: Automatic generation of multiple gene models. *BMC Bioinformatics*, 5(1):23–32, 2004.

[111] William H. Majoros, Mihaela Pertea, Corina Antonescu, and Steven L. Salzberg. GlimmerM, Exonomy and Unveil: Three ab initio eukaryotic genefinders. *Nucleic Acids Research*, 31(13):3601–3604, 2003.

[112] R.N. Mantegna, S.V. Buldyrev, A.L. Golberger, S. Havlin, C.K. Peng, M. Simons, and H.E. Stanley. Linguistic features of noncoding DNA sequences. *Physical Review Letters*, 73(23):3169–3172, December 1994.

[113] Giovanni Manzini and Marcella Rastero. A simple and fast DNA compressor. *Software — Practice and Experience*, 34:1397–1411, 2004.

[114] T. Matsumoto, K. Sadakane, and H. Imai. Biological sequence compression algorithms. *Genome Informatics*, 11:43–52, 2000.

[115] A.K. Mohanty and A.V.S.S. Narayana Rao. Factorial moments analyses show a characteristic length scale in DNA sequences. *Physical Review Letters*, 84(8):1832–1835, February 2000.

[116] Gabriela Moura, Miguel Pinheiro, Raquel M. Silva, Isabel M. Miranda, Vera M. A. Afreixo, GD Gaspar Dias, Adelaide Freitas, José Luís Oliveira, and Manuel Santos. Comparative context analysis of codon pairs on an ORFeome scale. *Genome Biology*, 6(3):R28(14), 2005.

[117] Daniel Nicorici and Jaakko Astola. Segmentation of DNA into coding and non-coding regions based on recursive entropic segmentation and stop-codon statistics. *EURASIP Journal on Applied Signal Processing*, 1:81–91, 2004.

[118] Su-Long Nyeo, I-Ching Yang, and Ahi-Hao Wu. Spectral classification of archaeal and bacterial genomes. *Journal of Biological Systems*, 10(3):233–241, April 2002.

[119] José Oliver, Ramón Ramón-Roldán, Javier Pérez, and Pedro Bernaola-Galván. Segment: Identifying compositional domains in DNA sequences. *Bioinformatics*, 15(12):974–979, 1999.

[120] S. Ouyang and C. R. Buell. The TIGR plant repeat databases: A collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Research*, 32:D360–D363, 2004.

[121] C.-K. Peng, S.V. Byldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H.E. Stanley. Long-range correlations in nucleotide sequences. *Nature*, 356:168–170, 1992.

[122] C.K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, and A.L. Goldberger. Mosaic organization of DNA nucleotides. *Physical Review E*, 49(2):1685–1689, 1994.

[123] Miguel Pinheiro, Vera Afreixo, Gabriela Moura, Adelaide Freitas, Manuel A. Santos, and José L. Oliveira. Statistical, computational and visualization methodologies to unveil gene primary structure features. *Methods of Information in Medicine*, 45:163–168, 2006.

[124] Armando J. Pinho, António J. R. Neves, Vera Afreixo, Carlos A. C. Bastos, and Paulo J. S. G. Ferreira. A three-state model for DNA. Protein-coding regions. *IEEE Transactions on Biomedical Engineering*, 53(11):2148–2155, November 2006.

[125] E. Rivals, J.-P. Delahaye, M. Dauchet, and O. Delgrange. A guaranteed compression scheme for repetitive DNA sequences. In *Proceedings of the Data Compession Conference DCC*, page 453, Snowbird, Utah, 1996.

[126] Ramón Román-Roldán, Pedro Bernaola-Galván, and José L. Oliver. Sequence compositional complexity of DNA through an entropic segmentation method. *Physical Review Letters*, 80(6):1344–1347, 1998.

[127] L. Rowen, G. Mahairas, and L. Hood. Sequencing the human genome. *Science*, 278:605–607, October 1997.

[128] Steven L. Salzberg, Arthur L. Delcher, and Owen White. Microbial gene identification using interpolated Markov models. *Nucleic Acids Reseach*, 26(2):544–548, 1998.

[129] Dorabella Santos. *Signal reconstruction in structures with two channels.* PhD thesis, Aveiro University, 2007.

[130] K. Sayood. *Introduction to data compression.* Morgan Kaufmann, 2000.

[131] Nicola Scafetta, Vito Latora, and Paolo Grigolini. Lévy statistics in coding and non-coding nucleotide sequences. *Physics Letters A*, 299:565–570, July 2002.

[132] Svetlana A. Shabalina and Nikolay A. Spiridonov. The mammalian transcriptome and the function of non-coding DNA sequences. *Genome Biology*, 5:105–42, 2004.

[133] Atul A. Shah, Michael C. Giddings, Jasmin B. Parvaz, Raymond F. Gesteland, John F. Atkins, and Ivaylo P. Ivanov. Computational identification of putative programmed translational frameshift sites. *Bioinformatics*, 18(8):1046–1053, 2002.

[134] B. D. Silverman and R. Linsker. A measure of DNA periodicity. *Journal of Theoretical Biology*, 118:295–300, 1986.

[135] A. Som, S. Chattopadhyay, J. Chakrabarti, and D. Bandyopadhyay. Codon distributions in DNA. *Physical Review E*, 63:051908, 2001.

[136] H.E. Stanley, S.V. Buldyrev, A.L. Goldberger, and S. Havlin. Scaling features of noncoding DNA. *Physica A*, 273:1–18, 1999.

[137] David S. Stoffer, David E. Tyler, and Andrew J. McDougall. Spectral analysis for categorical time-series: Scaling and the spectral envelope. *Biometrika*, 80(3):611–622, 1993.

[138] David S. Stoffer, David E. Tyler, and David A. Wendt. The spectral envelope and its applications. *Statistical Science*, 15(3):224–253, 2000.

[139] David Sussillo, Anshul Kundaje, and Dimitris Anastassiou. Spectrogram analysis of genomes. *EURASIP Journal on Applied Signal Processing*, 1:29–42, 2004.

[140] I. Tabus, G. Korodi, and J Rissanen. DNA sequence compression using the normalized maximum likelihood model for discrete regression. In *Proceedings of the Data compression conference DCC*, pages 253–262, Snowbird, Utah, 2003.

[141] Sébastien Tempel, Mathieu Giraud1, Dominique Lavenier, Israel-César Lerman, Anne-Sophie Valin, Ivan Coueé, Abdelhak El Amrani, and Jacques Nicolas. Domain organization within repeated DNA sequences: application to the study of a family of transposable elements. *Bioinformatics*, 22(16):1948–1954, June 2006.

[142] Shrish Tiwari, S. Ramachandran, Alok Bhattacharya, Sudha Bhattacharya, and Ramakrishna Ramaswamy. Prediction of probable genes by Fourier analysis of genomic sequences. *CABIOS*, 13(3):263–270, 1997.

[143] Masaru Tomita, Masahiko Wada, and Yukihiro Kawashima. Apa dinucleotide periodicity in prokaryote eukaryote, and organelle genomes. *Journal of Molecular Biology*, 49:182–192, 1999.

[144] Edward N. Trifonov and Joel L. Sussman. The pich of chromatin DNA is reflected in its nucleotide sequence. *Biochemistry*, 77(7):3816–3820, 1980.

[145] E.N. Trifonov. 3-, 10.5-, 200- and 400-base periodicities in genome sequences. *Physica A*, 249:511–516, 1998.

[146] A.A. Tsonis, P. Kumar, J.B. Elsner, and P.A. Tsonis. Wavelet analysis of DNA sequences. *Physical Review E*, 53(2):1828–1834, February 1996.

[147] O.V. Usatenko and V.A. Yampol'skii. Binary n-step markov chains and long-range correlated systems. *Physical Review Letters*, 90(11):110601, March 2003.

[148] P. P. Vaidyanathan. Genomics and proteomics: A signal processor's tour. *IEEE Circuits and systems magazine*, 4:6–29, 2004.

[149] P. P. Vaidyanathan and Byung-Jun Yoon. The role of signal-processing concepsts in genomics and proteomics. *Journal of the Franklin Institute, special issue on Genomics*, 2004.

[150] P.P. Vaidyanathan and Byung-Jun Yoon. Digital filters for gene prediction applications. In *Proceedings 36th Asilomar Conference on Signals Systems and Computers*, Monterey, CA, November 2002.

[151] P.P. Vaidyanathan and Byung-Jun Yoon. Gene and exon prediction using allpass-based filters. In *Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, Raleigh, NC, October 2002.

[152] M. de Sousa Vieira and H.J. Herrmann. A growth model for DNA evolution. *Europhysics Letters*, 33(5):409–414, 1996.

[153] Maria de Sousa Vieira. Statistics of DNA sequences: A low-frequency analysis. *Physical Review E*, 60(1):5932–5937, November 1999.

[154] Richard F. Voss. Evolution of long-rang fractal correlations and $1/f$ noise in DNA base sequences. *Physical Review Letters*, 68(25):3805–3808, 1992.

[155] Wei Wang and Don H. Johnson. Computing linear transforms of symbolic signals. *IEEE Transactions on Signal Processing*, 50(3):628–634, March 2002.

[156] James Watson and Francis Crick's. A struture for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.

[157] Zu-Guo Yu, V.V. Anh, and Bin Wang. Correlation property of length sequences based on global struture of the complete genome. *Physical Review E*, 63:011903, 2000.

[158] Zu-Guo Yu and Guo-Yi Chen. Rescaled range and transition matrix analysis of DNA sequences. *Communications in Theoretical Physics*, 33(4):673–678, 1999.

[159] Zu-Guo Yu and Bin Wang. A time series model of CDS sequences in complete genome. *Chaos, Solitons, and Fractals*, 12(3):34–46, 2001.

[160] G.F. Zebende, P.M.C. de Oliveira, and T.J.P. Penna. Long-range correlations in computer diskettes. *Physical Review E*, 57(3):3311–3314, March 1998.