UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
CENTRO DE CIÊNCIAS EXATAS E DA TERRA
DEPARTAMENTO DE FÍSICA TEÓRICA E EXPERIMENTAL
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA

Distribuição dos tamanhos de DNA humano codificante via teoria da informação

Jonathan Pessoa Correia

Natal-RN
14 de Junho de 2021

Jonathan Pessoa Correia

Distribuição dos tamanhos de DNA humano codificante via teoria da informação

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Física do Departamento de Física Teórica e Experimental da Universidade Federal do Rio Grande do Norte como parte dos requisitos necessários para obtenção do título de Mestre em Física.

Orientador: *Prof. Dr. Raimundo Silva Júnior*

Natal-RN

14 de Junho de 2021

Universidade Federal do Rio Grande do Norte - UFRN
Sistema de Bibliotecas - SISBI
Catalogação de Publicação na Fonte. UFRN - Biblioteca Setorial Prof. Ronaldo Xavier de Arruda - CCET

Correia, Jonathan Pessoa.
Distribuição dos tamanhos de DNA humano codificante via teoria da informação / Jonathan Pessoa Correia. - 2021.
77f.: il.

Dissertação (mestrado) - Universidade Federal do Rio Grande do Norte, Centro de Ciências Exatas e da Terra, Programa de Pós-Graduação em Física. Natal, 2021.
Orientador: Prof. Dr. Raimundo Silva Júnior.

1. Física - Dissertação. 2. Teoria da informação - Dissertação. 3. DNA humano - Dissertação. 4. Entropia Shannon - Dissertação. I. Silva Júnior, Raimundo. II. Título.

RN/UF/CCET CDU 53

Elaborado por Joseneide Ferreira Dantas - CRB-15/324

Agradecimentos

Ao professor Raimundo, pela orientação na pesquisa durante esse período. Sou muito grato pela paciência, pelas oportunidades, pelo apoio e pela cobrança quando necessária.

Ao professor José Ronaldo pelas discussões, pela paciência e por ajudar em partes importantes do trabalho.

Aos colegas Marcones e Polyanna pelo auxilio na parte computacional. Esse trabalho seria impossível sem a contribuição deles.

Aos membros da minha família que mesmo longe nunca mediram esforços para que eu chegasse até aqui. Em especial, agradeço á minha mãe Maria Aparecida e ao meu pai Paulino pelo incentivo, pelo exemplo e pelos valores que me tornaram ser quem sou.

Aos professores do DFTE que contribuíram para minha formação acadêmica.

À Deus por todo cuidado ao longo desses anos e pela força para enfrentar os desafios.

Conteúdo

Lista de Figuras

Resumo

Analisamos as sequências codificantes do DNA do *Homo Sapiens* por meio de um modelo que naturalmente envolve correlações entre as bases nas sequências de DNA dos organismos vivos. O modelo é baseado na otimização da entropia de Shannon, que é o centro de todos os argumentos estatísticos. No presente trabalho, propomos a função de distribuição de dupla exponencial dos comprimentos do DNA medido em pares de bases (pb). Os resultados mostram que as Correlações de Curto Alcance (CCA), sempre presentes nas sequências de DNA codificantes, são apropriadamente capturadas por meio da distribuição dupla exponencial e descreve adequadamente a distribuição de comprimentos cumulativos das bases de DNA. Com base neste modelo, usamos uma função de distribuição cumulativa empírica e o banco de dados de proteínas compilado pelo Projeto Ensembl para mostrar consistência com os dados.

Abstract

We analyze the coding sequence for the Homo Sapiens DNA via a model that naturally embraces correlations among the bases in DNA sequences of living organisms. The model is based on the Shannon entropy's optimization, which is the core of all statistical arguments. On our work , we propose the double-exponential1 distribution function of the length of DNA measured in base pairs (bp). The results show that the Short-Range-Correlations (SRC), always present in coding DNA sequences, are appropriately captured through the double-exponential distribution and adequately describes the cumulative length distribution of DNA bases. Based on this model, we use an Empirical cumulative distribution function and the database of proteins compiled by the Ensembl Project to show consistency with the data.

1. Introdução

O nosso planeta possui uma imensa diversidade quando se trata da vida, com estimativas indicando que haja mais de 10 milhões de espécies vivendo na Terra [1]. Em todas as regiões do planeta é possível encontrar plantas, animais, fungos e diversos outros seres vivos adaptados às mais distintas condições ambientais e climáticas [2].

Embora notavelmente distintas, as criaturas vivas são essencialmente parecidas quando vistas em sua forma mais fundamental: Todas são constituídas por células. Mesmos os indivíduos pluricelulares tiveram sua origem em uma única célula: o zigoto. Tais agregados celulares realizam funções altamente especializadas e conectados por um intrincado sistema de comunicação [3].

O mundo vivo sempre foi objeto de curiosidade e investigação desde o início da civilização. Nos últimos anos houve uma revolução na biologia em grande parte devido ao entendimento dos mecanismos moleculares e celulares centrados nos genes. Os genes são as unidades fundamentais da informação biológica (material genético) e são eles os responsáveis por carregar o conhecimento necessário para, a partir de uma única célula, construir todo um organismo [4].

Antes da grande descoberta que revolucionaria a biologia, tinha-se conhecimento que o material genético precisava possuir algumas propriedades para ser o responsável pelo armazenamento e a transferência da informação biológica para seus descendentes [3, 4]:

- A primeira era que sua estrutura deveria permitir sua replicação fiel para formar outras células.
- Como carrega informação sobre os seres vivos, o material genético deveria ser capaz de decodificar essa informação para produzir proteínas.
- Por fim, deveria permitir que pequenas alterações pudessem ser feitas no material genético (mutações) e passar essas modificações para a próxima geração sem comprometer a estrutura do DNA.

Durante certo tempo acreditou-se que eram as proteínas que carregavam a informação sobre como a vida era formada. Apenas com o experimento de transformação de Frederick Griffith em 1928 e mais tarde, em 1944, com o trabalho de Oswald Avery [5] obteve-se forte evidências de que o material genético era o DNA[1] e não as proteínas [1].

Nesse aspecto, um grande passo foi dado em 1953 quando James Watson e Francis Crick[2] propuseram a estrutura do DNA como sendo uma dupla hélice [7]. Esses cientistas apresentaram uma definição de genes a partir de termos químicos, que permitindo-nos compreender a ação gênica e da hereditariedade em um nível molecular, além de seu modelo possuir todas as propriedades já mencionadas acima e que eram esperadas para o material genético. A compreensão dos mecanismos por trás da hereditariedade nos permitiu obter respostas sobre as questões fundamentais para a biologia e para vida. Sem mencionar, é claro, todas as aplicações que foram possível realizar em diversas áreas do conhecimento da agricultura a medicina [1, 3, 4, 8].

O estudo dos genes sempre atraiu a atenção de várias áreas da ciência. Muitos físicos, inclusive, foram importantes para determinar a estrutura do DNA como o próprio Francis Crick e Rosalind Franklin com seus dados de difração de raios X do DNA [9]. Mais recentemente o desenvolvimento de formulações generalizadas da entropia [10, 11] a partir da extensão da abordagem entrópica de Boltzmann e outros modelos estatísticos como o caminhante aleatório [12, 13], modelo de Ising unidimensional [14] e outras abordagens estatísticas [15, 16, 17, 18, 19] nos permitiram descobrir diversas propriedades do DNA.

Este trabalho tem por objetivo fornecer uma pequena contribuição para essa área rica e vasta, propondo uma função de distribuição para os comprimentos dos pares de base do DNA partindo da entropia de teoria da informação.

A Teoria Matemática da informação é um ramo da Matemática Estatística e da Teoria de Probabilidade que possui conexões com diversos outros domínios científicos como com Sistemas de Comunicação, Transmissão de Dados, Criptografia, Codificação, Mecânica Estatística, Teoria do Ruído, Correção de Erros, Compressão de Dados, e outros [20, 21].

A teoria de informação tem sua origem no artigo do engenheiro americano Claude Shannon intitulado 'A Mathematical Theory of Communication' [22]. Antes disso haviam algumas abordagens teóricas que tratavam de sistemas de comunicação, informação e velocidade de transmissão, todavia foi o trabalho de Shannon que lançou as bases da teoria de informação fornecendo um modelo quantitativo e qualitativo [23, 24]. Outros conceitos importantes também foram desenvolvidos em seu trabalho como a entropia de

[1] Do inglês Deoxyribonucleic acid (Ácido desoxirribonucleico)

[2] O período dessa descoberta é relatada por Watson em [6].

informação $H(X)$ associada a um conjunto de variáveis aleatórias X, a capacidade de transmissão de um canal sem perda de informação e do bit como sendo uma nova forma de enxergar a unidade fundamental de informação [25, 26, 27].

Dividimos esse trabalho da seguinte maneira:

No segundo capítulo tratamos do DNA apresentando sua estrutura, componentes e como este se organiza até formar os cromossomos. Tratamos também de distinguir as partes codificantes (éxons) e não-codificantes (íntrons) do DNA e que serão importantes para o trabalho. O comprimento l de uma dada região é dado pelo números de pares de base (pb) presentes na mesma.

No terceiro capítulo apresentamos os fundamentos da teoria de informação, introduzindo os conceitos de informação, bit, surpresa, entropia e etc. Além de mostrar as principais propriedades e fazer a conexão direta entre entropia $H(X)$ e a informação.

No quarto capítulo, discutimos alguns trabalhos relevantes de biofísica aplicados à genética. Tratamos em especial de alguns modelos e resultados importantes que indicam comportamento de correlação e/ou propõem funções de distribuições do tipo lei de potência para as sequências de DNA.

No quinto capítulo, extremizamos a entropia $H(X)$ de teoria da informação por meio de cálculo variacional para obter uma função de distribuição $p(l)$ para os comprimentos l das regiões de DNA codificante e apresentamos os resultados.

Por fim, no quinto capítulo discutimos as conclusões e apresentamos as perspectivas deste trabalho.

2. Introdução ao DNA

A biologia sofreu uma grande revolução quando os últimos anos da década de 60 nos trouxeram repostas para as grandes questões acerca da vida [3]. A grande responsável por esse desenvolvimento foi a genética, que concentra-se no estudo de todos os aspectos envolvendo as unidades fundamentais da informação biológica: os genes. Sabe-se que todos os seres humanos são formados por trilhões de células, que por sua vez contém estruturas bastante complexas como mitocôndria, Complexo de Golgi, lisossomos, diversas outras organelas e um núcleo. Dentro desse núcleo encontramos as longas cadeias de DNA [8].

Podemos notar uma característica muito interessante da vida quando observamos alguns filhotes de animais. Lembro-me de uma cadelinha vira-lata que minha família possuiu por muitos anos. Ela tinha características da raça basset (comprida e baixinha), preta e com algumas manchas amarelas na pata e no rosto. Quase todos os filhotes dela era baixinhos mesmo depois de atingirem a vida adulta e alguns deles apresentavam as mesmas machas amarelas no rosto e/ou pata. A questão é: Como é possível que características físicas sejam passadas de uma geração para outra? Hoje sabemos que informação acerca das características físicas dos seres vivos estão armazenadas no ácido desoxirribonucleico ou DNA.

2.1 A estrutura do DNA

Apesar de toda a diversidade, é interessante observar que todas as células armazenam informação da mesma forma e que qualquer um desses arquivos podem ser lidos pelo sistema de processamento de qualquer outra célula. O modo como cada uma delas realiza o processo de replicação dessa informação ocorre com diferentes mecanismos para iniciá-lo e interrompê-lo, com diversas velocidades e distintos processos auxiliares porém seguindo os mesmo princípios universais: as informações hereditárias de todas as criaturas vivas são guardadas em moléculas de DNA de fita dupla [1].

É muito importante ressaltar que a estrutura do DNA é altamente complexa e levaram-se anos de estudos e pesquisas para começarmos a compreender de forma mais completa o seu funcionamento. Existe uma variedade volumosa de livros e artigos tratando esse assunto de forma mais ampla e completa[1]. O objetivo deste capítulo é prover ao leitor uma breve introdução ao assunto fornecendo condições de compreender de forma simples a estrutura e o funcionamento do DNA.

Damos o nome de genoma ao conjunto de toda informação contida no DNA. O DNA é uma longa cadeia linear formada a partir de unidades menores que se repetem: os monômeros. Essas subunidades (monômeros) também são chamadas de nucleotídeos e possuem quatro bases - Adenina (A), Timina (T), Citosina (C) e Guanina (G)- que ligando-se entre si para formar uma longa cadeia polimérica. Os nucleotídeos também são classificados em dois grupos distintos chamados de pirimidinas, formadas pelas bases timina e citosina, e purinas, formado por guanina e adenina.

(a) Unidades dos nucleotídeos (b) Fita de DNA

Figura 2.1: Estrutura dos nucleotídeos e da fita de DNA [3].

Os nucleotídeos também são formados por estruturas menores: um açúcar, um grupo fosfato e uma das bases (A, T, C, G) (Veja a Fig. 2.1(a)). É claro que todos esses constituintes possuem estruturas químicas muito complexas. Porém com o objetivo de apresentar de forma simples a estrutura do DNA esquematizamos apenas a ideia de como esses componentes se organizam para formar os nucleotídeos[2]. Cada açúcar de um nucleotídeo liga-se ao outro através do grupo fosfato de modo que os nucleotídeos unem-se entre si como em um quebra cabeças formando uma cadeia linear repetitiva de açúcar e fosfato, com séries de bases projetando-se dela (Veja Fig. 2.1(b)).

A molécula de DNA não é formada apenas por uma fita. Se fosse uma única fita isolada, então não haveria regra quanto qual seria o próximo nucleotídeo a se unir a cadeia, uma vez que todos eles se ligam-se através de uma parte da molécula que é comum a todos os nucleotídeos (a cadeia principal de açúcar-fosfato). Acontece que o DNA é formado

[1] Veja por exemplo, as Refs. [1, 3, 4] nas quais este capítulo foi fortemente baseado.
[2] Para mais detalhes veja [8], [28]

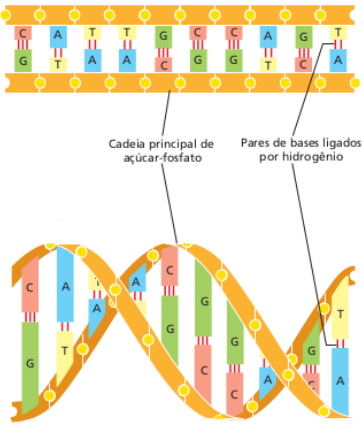

Figura 2.2: Dupla Hélice: Na figura superior vemos a estrutura plana do DNA com uma fita formada por um suporte comum de açúcar-fosfato com as bases projetando-se dela e ligando-se com as outras bases de outra fita, através de pontes de hidrogênio (traços em vermelho). As ligações entre as bases obedecem: A liga-se a T e C liga-se a G. A figura inferior mostra como essa estrutura se torce em torno de si para formar a molécula de DNA tridimensional [3].

.

por duas fitas que se complementam e torcem ao redor de si formando uma dupla-hélice. Para formar essa estrutura as bases que se projetam da primeira fita ligam-se, por pontes de hidrogênio, à sua base correspondente na segunda fita obedecendo uma regra rigorosa definida pelas estruturas complementares das bases: Adenina liga-se a Timina, e Citosina liga-se a Guanina. Como pode ser visto na Fig. 2.2.

Esse mesmo mecanismo é usado no processo de cópia do DNA em que uma da fitas é usada como molde para síntese de outra fita. A nova fita de DNA é produzida e suas bases ligam-se a do molde obedecendo a mesma regra das estruturas complementares: A liga-se a T e C liga-se a G. Essa regra de pareamento das bases garante qual será o novo monômero a ser acrescentado à nova fita criando uma estrutura de dupla hélice composta por sequências complementares das bases. Os nucleotídeos opostos e complementares na cadeia de DNA e que estão conectados por pontes de hidrogênio são chamados de pares de base (pb). Assim Adenina forma um par de base com Timina e Citosina forma um parte de base com Guanina.

Nos organismos que possuem núcleo celular (eucariotos) a maior parte da informação genética está concentrada dentro nos núcleos (DNA intranuclear) embora seja possível encontrar uma pequena fração dos genomas nas mitocôndrias (DNA extranuclear). Estima-se que o genoma humano intranuclear possui cerca de um metro de comprimento.

Uma vez que o tamanho do núcleo é da ordem de mícrons, concluímos que o DNA está compactado de forma muito eficiente para ocupar um espaço muito pequeno. Deve existir algum mecanismo para evitar que a fita se embarace durante o processo de duplicação do DNA.

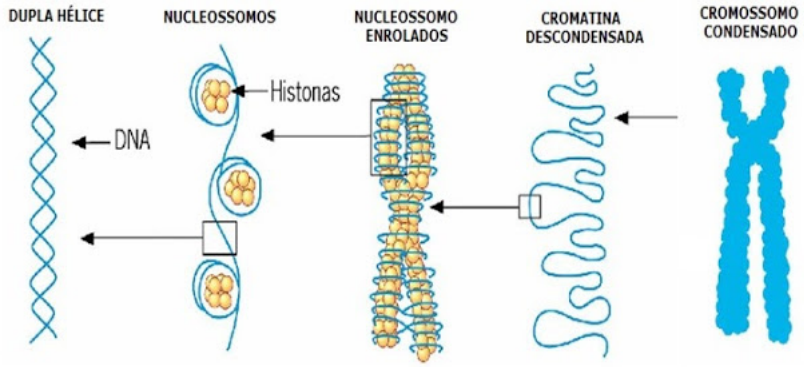

Figura 2.3: Estrutura do Cromossomo: O Cromossomo é o ajuntamento da cromatina que por sua vez é a compactação dos nucleossomos. Os nucleossomos são o agrupamento de fitas de dupla hélices que se enrolam ao redor das histonas [29].

Hoje sabemos que para otimizar o espaço ocupado dentro do núcleo, o DNA se condensa para formar os cromossomos. Essa compactação ocorre com o enrolamento da dupla hélice ao redor de estruturas moleculares chamadas histonas. Se pensarmos na dupla hélice do DNA como um longo fio as estonas podem ser vistas como pequenos novelos no qual o DNA se envolve. O conjunto do DNA mais as histonas é chamado de nucleossomos que também se compactam para formar uma longa cadeia linear denominada cromatina. A cromatina pode ser vista como um longo fio formado pelos nucleossomos e que forma o cromossomo[3]. Um esquema dessa estrutura é mostrado na Fig. 2.3.

Existe ainda outras estruturas complementares que formam os cromossomos como o arcabouço que ajuda a organizar sua forma tridimensional, o centrômero que fixa um ponto para mover o cromossomo durante uma divisão celular, os telômeros e muitos outros. O importante, no entanto, é saber que a dupla fita se enrola nas histonas e esse arranjo formam os nucleossomos que por sua vez se compactam formando uma fita que recebe o nome de cromatina. A cromatina por sua vez é a base para formar o cromossomo.

O DNA nuclear é formado por um número de cromossomos específico para cada espécie. O *homo sapien*, por exemplo, possui 23 cromossomos. Além disso, como na maioria dos mamíferos, os humanos são organismos diploides, ou seja, os núcleos possuem duas cópias completas do genoma, e assim dois conjuntos quase idênticos do cromossomo embora

[3]Um vídeo interessante sobre o assunto pode ser visto em [30].

possa haver pequenas variações na sequência dos nucleotídeos.

2.2 Éxons e íntrons

Até 1977, pensava-se que a fita de DNA era composta por arranjos contínuos de nucleotídeos. Mas ao observar a cadeia de DNA mais atentamente percebia-se que haviam mais nucleotídeos do que aqueles utilizados para produção de proteínas, levando à descoberta da existência de genes interrompidos [31].

Hoje sabemos que no processo de duplicação, a sequência do DNA é completamente transcrita mas partes dela são recortadas e não são utilizadas na produção do RNA. As sequências funcionais do genoma e que armazenam informação são chamadas de genes ou éxons[4] [31]. Na prática são essas faixas que carregam a informação genética dos organismos vivos e que determinam as características físicas, a produção de proteínas e diversas outras funções do organismo. É importante ressaltar que o número e o tamanho dos éxons diferem muito entres os seres vivos. Veja por exemplo a Fig. 2.4. Os humanos possuem cerca de 20.500 éxons enquanto que o milho tem 32.000 [3].

É importante destacar que as regiões codificadoras apresentam trechos não codificadores chamados íntrons[5]. Em alguns casos, a quantidade de íntrons pode aumentar consideravelmente o tamanho de uma cadeia de DNA [3]. O tamanho de uma região de éxon ou íntron é dado pelo número de pares de base (pb) naquela sequência. Considere por exemplo a Fig. 2.1(b). O comprimento dessa região será de 10 pb.

Existem outras regiões no genoma que desempenham outros papéis mas que não discutiremos com mais detalhes como, por exemplo, a sequência espaçadora e a reguladora que fica entre duas faixas codificadoras.

Um fato curioso é que logo após compreender as sequências de DNA os cientistas foram capazes de estimar a quantidade de genes de vários organismos. Inicialmente, a estimativa foi realizada em seres mais simples como a bactéria *Escherichia coli* e chegamos ao resultado de aproximadamente 3.200 genes. A levedura unicelular *Saccharomyces cerevisiase* possui cerca de 6.300, enquanto que a mosca de fruta *Drosophila melanogaster* apresentava 13.600 genes. O próximo passo natural da genética era fazer as mesmas estimativas para o DNA humano. Esperava-se que, devido à complexidade da estrutura cerebral e com um sistema imunológico tão sofisticado, o DNA humano possui-se maior quantidade de genes.

[4] O termo é derivado do inglês **ex**pressed regi**ons**

[5] Do termo **int**ervening regi**ons**

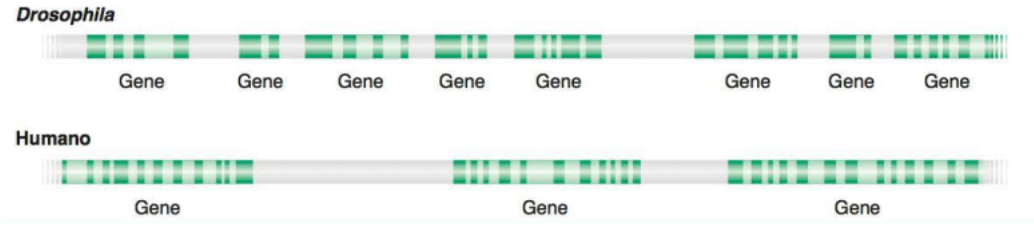

Figura 2.4: Genoma da mosca e do ser humano. As regiões em verde-escuro correspondem ao éxon enquanto que as verde-claro são íntrons. As em branco são regiões reguladoras mais DNA espaçador. Note que éxons e íntrons possuem tamanho e distribuições distintas para as duas espécies [1].

A grande surpresa foi quando liberaram as primeiras estimativas da sequência de genoma humano e constataram que a quantidade de genes era próxima da mais baixa estimava (26 mil). Como pode um organismo tão complexo como o *Homo sapiens* ter apenas o dobro da quantidade de genes de uma mosca? A resposta para essa pergunta está na estrutura descontínua das sequências. As proteínas de organismos superiores são codificadas a partir de fragmentos do DNA: os éxons e íntrons.

2.3 Transcrição

Agora que conhecemos um pouco da estrutura do DNA podemos nos perguntar como essa estrutura de dupla fita formada por uma sequência de nucleotídeos é capaz de gerar as características físicas de um organismo vivo como o seu tamanho, cor dos cabelos, olhos, boca, comportamento e assim por diante.

Antes da descoberta do DNA já sabia-se que os organismos vivos eram formados por proteínas ou material constituído por elas. Existem três tipos de proteínas classificadas de acordo com a função desempenhada por elas no organismo: As estruturais são as que formam, como o nome sugere, as estruturas externas do organismo como os pêlos, unhas, cabelo, ossos e elementos estruturais de dentro da célula, como o citoplasma. As proteínas enzimáticas são responsáveis por catalisar todas as reações químicas dentro das células que vão dar origem as moléculas, ácidos nucleicos, carboidratos e outras proteínas. Por fim, as proteínas reguladoras tem o papel de ativar e desativar a atividade gênica no local e momento adequado. Diante disso, vemos que as proteínas são as estruturas fundamentais para determinar as características dos seres vivos. Logo a chave para compreender de modo mais profundo como funcionam os organismos biológicos está em entender como, a partir da dupla fita de DNA, é possível obter as proteínas.

A transferência de informação contida no gene para o produto gênico (aminoácidos) ocorre em várias etapas. A primeira delas consiste em fazer uma cópia, utilizando um filamento do DNA como molde, para produzir uma estrutura intermediária chamada RNA. A segunda etapa é converter a informação do RNA em uma cadeia de aminoácidos por um processo chamado de tradução.

Nos primórdios da genética, acreditava-se que a informação não era transferida diretamente do DNA para a proteína. Isso devido ao trabalho de Volkin e Astrachan [32] que através do protocolo chamado experimento de pulso-caça indicaram que o DNA era produzido no núcleo mas encontrado no citoplasma onde são sintetizadas as proteínas, ou seja em locais diferentes. Portanto deveria existir um intermediário para fazer o transporte da informação contida no DNA para o citoplasma.

Já mencionamos que a informação de qualquer organismo está codificada na sequência de DNA de forma estática. Essa informação especifica quando, onde e como os produtos gênicos são feitos. Para fazer uso dessa informação é necessário produzir uma cópia do DNA em um processo chamado de **transcrição**. A cópia do DNA é uma cadeia de nucleotídeos de apenas um filamento chamada de RNA. Em especial chamamos essa classe de RNA que codifica a informação para produzir as proteínas e serve de intermediário de transferência da informação do DNA para a proteína de **RNA mensageiro (mRNA)**. Existe uma outra classe, chamada de **RNA funcional** que não codifica a informação mas fornece um tipo de suporte para produção de proteínas. Como por exemplo, o RNA transportador que leva o aminoácido correto para o mRNA no processo de tradução e o RNA ribossômico que é o principal componente dos ribossomos. O RNA é mais flexível do que o DNA e pode formar uma variedade maior de formas moleculares tridimensionais complexas. Outra característica muito importante é que os nucleotídeos do RNA contém as bases, guanina, adenina, citosina mas a base timina é substituída pela uracila (U). Esta última entra no grupo das pirimidinas.

A informação do DNA é transferida para o RNA utilizando o **princípio do pareamento de bases**. É ele o responsável por determinar qual é a sequência de nucleotídeos de um novo filamento de DNA na replicação e do RNA transcrito na transcrição.

Esse procedimento é extremamente complexo pois envolve um grande número de replicações sucessivas, diversas enzimas e o sentido da replicação. Aqui, vamos apenas fornecer uma visão geral de como ocorre esse processo. O primeiro passo consiste na quebra das ligações de hidrogênio entre os pares de base e um dos filamentos é usado como molde para síntese do DNA. Uma enzima chamada **RNA polimerase** se liga ao DNA e se move ao longo dele. Em seguida a RNA polimerase utiliza os ribonucleotídeos, que foram

produzidos em outra parte da célula, para formar pares estáveis com suas bases complementares no molde, isto é, a RNA polimerase vai ligar o ribonucleotídeo G com a base C do DNA. Da mesma maneira a uracila U se liga com a A. Durante esse procedimento o RNA polimerase posiciona cada ribonucleotídeo em oposição à sua base complementar e os alinha para fazer uma molécula crescente de RNA. Conforme a RNA polimerase se move ao longo da cadeia, ela desenrola a dupla hélice de DNA à sua frente, posiciona o ribonucleotídeos correspondente à base complementar e volta a enrolar o DNA que já foi transcrito (Ver Fig. 2.5(a)).

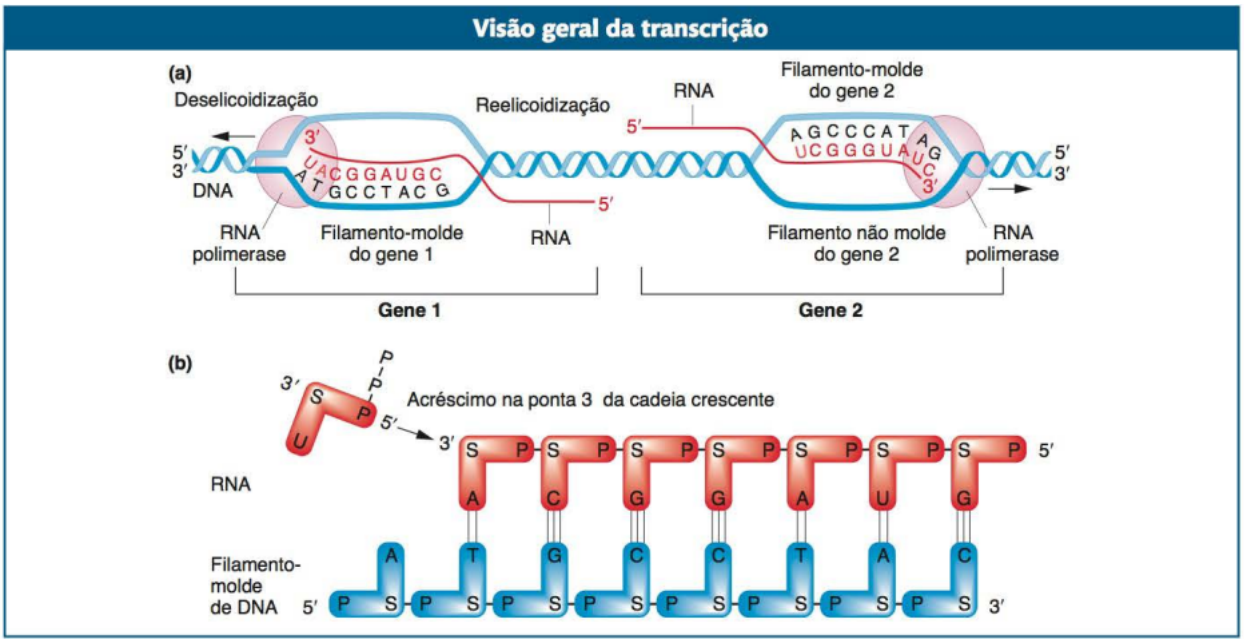

Figura 2.5: Visão geral da transcrição. Na Fig. (a) dois segmentos do DNA são copiados. O gene 1 é transcrito do filamento de baixo. Uma vez que a sequência é lida no sentido de 3′ para 5′ o RNA polimerase migra para a esquerda lendo o filamento molde e sintetizando RNA que é produzido no sentido 3′ a 5′. No gene 2, a parte transcrita é a superior. Nesse caso, como o filamento de cima funciona como molde para o RNA, o sentido de transcrição é da direita para esquerda (seguindo de 3′ para 5′). Conforme a transcrição acontece, a ponta 5′ é deslocada do molde e a bolha de transcrição se fecha atrás da polimerase. (b) olhando com mais detalhes o gene 1, à medida que ocorre a transcrição, o grupo fosfato da extremidade 5′ do ribonucleotídeo que entra liga-se, à ponta 3′ da cadeia crescente de RNA [1].

O DNA, bem como o RNA, possui uma orientação em cada fita que é dada pela posição do átomo de carbono no anel de açúcar do qual o grupo fosfato está ligado. Não vamos entrar em detalhes, basta saber que uma das pontas do RNA é a 5′ e a outra

é 3′. Durante a síntese, a fita do DNA é lida no sentido de 3′ para 5′, isto é, a RNA polimerase se posiciona na fita e se move da direção 3′ para 5′. Após 'ler' qual nucleotídeo se encontra na fita do DNA, a RNA polimerase adiciona a respectiva base complementar do nucleotídeo lido, à fita da RNA que está sendo produzida. Isto significa que a RNA cresce no sentido de 3′ para 5′, pois é nessa direção que as bases complementares são adicionados (Ver Fig. 2.5).

Conforme a RNA polimerase vai lendo os pares de base e se deslocando ao longo da fita, a parte de atrás da fita vai se fechando. Outra observação importante é que como as bases no RNA transcrito e no molde são complementares, então a sequência de nucleotídeos do RNA tem que ser a mesma da fita de DNA que não foi serviu de molde. A diferença está no fato que as T são substituídas por U. Na literatura científica, chamamos os filamentos de DNA que não foram moldes para a produção de mRNA mas que tem a mesma sequência de nucleotídeos (exceto pela troca de T por U) do mRNA produzido de *filamento codificador*.

Após ser transcrito o mRNA é processado, isso ocorre com a remoção e adição de sequências dos nucleotídeos. Logo após, o mRNA se desloca do núcleo da célula para o citoplasma onde será utilizado como molde para produção de proteínas.

2.4 Tradução

Agora precisamos entender como a sequência de DNA dita como são produzidas as proteínas. A resposta mais natural para essa pergunta vem de pensarmos que os nucleotídeos são letras em um código, cuja combinação de letras forma as palavras. Essas letras (A, G, C, U) irão se combinar para formar os **códons** que nada mais são do que os aminoácidos. Estes por sua vezes são as unidades fundamentais das proteínas.

Lembre-se que a molécula de mRNA é lida de uma extremidade para outra e que cada base (A, G, C e U) pode ser encontrada em uma posição. Então se cada aminoácido fosse decodificado por uma das bases poderíamos ter quatro combinações diferentes. Esse número é insuficiente para construir os 20 aminoácidos conhecidos. Por outro lado, se cada aminoácido é formada por um par de bases (CA, UA, GG, UC, ...) então seriam possível formar $4 \times 4 = 16$ aminoácidos. Essas combinações ainda são insuficientes.

A partir dessa discussão podemos concluir que é necessário pelo menos que cada aminoácido seja formado a partir da combinação de uma trinca de bases pois $4 \times 4 \times 4 = 64$ combinações. Por exemplo: CAU, GGG, AGU e etc. Esse vocabulário fornece palavras suficientes para descrever os aminoácidos. Em 1961, obteve-se a prova de que cada códon

possui três bases através do experimento realizado por Sidney Branner [33]. É importante ressaltar que há uma redundância do código genético. Uma vez que temos 20 aminoácidos, ainda sobra 44 combinações entre os nucleotídeos que aparentemente não tem nenhum significado dentro do código. Hoje sabemos que na realidade o código genético é **degenerado**. Isso significa que alguns aminoácidos são especificados por duas ou mais trincas diferentes (Veja Fig. 2.6).

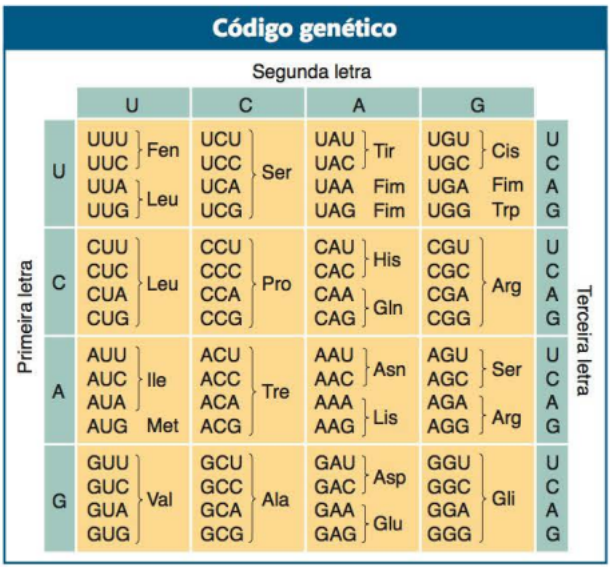

Figura 2.6: Observe que a combinação de cada trinca das bases (códons) formam um aminoácidos [1]. Veja também que alguns aminácidos podem ser formados por mais de um códon.

A produção das proteínas ocorre quando o tRNA (RNA transportador) e as moléculas de mRNA se associam aos ribossomos[6]. O tRNA é responsável por transportar os aminoácidos para a síntese proteica. De forma muito simplificada, o ribossomo funciona como uma 'máquina biológica' extremamente complexa e que lê três bases do mRNA de cada vez e associa a cada uma delas ao aminoácido correspondente. Este por sua vez é posicionado em uma cadeia crescente de aminoácidos. Após o último deles ser adicionado, a cadeia se dobra em um complexa estrutura tridimensional para formar as proteínas.

[6]Veja também [34]

3. Teoria da informação

Informação é um conceito muito abrangente e que carrega uma variedade de significados, indo do uso mais cotidiano ao técnico [35]. Em grande parte das vezes seu uso não leva em consideração os vários significados adquiridos ao longo do tempo. Genericamente, informação sempre esteve ligada às noções de comunicação, dados, conhecimento, padrão, percepção e representação do conhecimento [35].

Mesmo sendo utilizada pelos gregos antigos foi somente no século XX que a ideia de informação ganhou a forma que utilizamos atualmente [27]. Durante muito tempo, inclusive, diversas áreas do conhecimento como a física, biologia, telecomunicações e outras lidavam com processos envolvendo armazenamento e processamento de informação mesmo não tendo uma definição matemática ou mesmo rigorosa para a mesma. Foi somente em 1948, que o artigo 'A matematical Theory of communication' [22] de autoria de Claude Shannon estabeleceu conceitos concretos e mensuráveis para a informação.

Atualmente, na era da internet e da constante troca e produção de conteúdo, os conceitos envolvendo coleta, processamento, tratamento e sigilo de dados são familiares para as grande maioria das pessoas. Antes de prosseguirmos é importante ressaltar que há uma diferença sutil entre informação e dados: A primeira é todo conhecimento que nos possa ser útil para algum propósito específico enquanto que a última trata-se uma combinação entre informação útil e a aquela que não serve aos nosso propósitos (também chamada de ruído) [36]. A capacidade de separar informação útil do ruído é fundamental para transmissão de uma imagem para televisão, uma música para um rádio ou mesmo um vídeo pelo Whatsapp [25].

Na própria natureza, a capacidade dos seres vivos de utilizar os olhos e ouvidos para distinguir informação útil (como a aproximação de um predador ou presa) dos sons produzidos pela floresta é fundamental pra determinar a sobrevivência de um indivíduo [25].

Note que informação, no contexto de análise de dados, irá depender do problema a ser estudado, seja ele biológico, de mercado financeiro, telecomunicações e o tipo de compreensão que deseja-se obter do sistema [25]. Em suma, o que quero descobrir a

respeito de um sistema irá me dizer quais dos dados disponíveis a respeito dele me serão úteis [27].

Nesse capítulo, introduzimos os fundamentos da teoria de informação de Shannon e discutimos conceitos importantes como entropia e sua conexão com informação[1].

3.1 Fundamentos

Imagine que você esteja no portão de entrada da UFRN e deseja chegar ao Departamento de Física. Sendo aluno novo na universidade você não tem nenhuma informação inicial sobre a localização do prédio do departamento mas sabe que percorrerá um longo caminho e haverá momentos que possuirá a liberdade de escolher qual rua leva a seu destino. Já na entrada do campus você se depara com uma rua que forma uma bifurcação e te permite decidir entre duas outras ruas. A Fig. 3.1 mostra um pequeno esquema dessa mapa hipotético da universidade. Como você não tem nenhuma informação sobre onde fica o departamento, a probabilidade de escolher qualquer uma das ruas é a mesma. Porém, um transeunte próximo te indica que seguindo a direita até chegar à Escola de Ciência e Tecnologia (C&T) é a melhor maneira de chegar até o departamento. A dica dada pelo gentil desconhecido corresponde a um bit de informação uma vez que permitiu escolher entre duas possibilidades de igual probabilidade, além de evitar que você tomasse o caminho errado seguindo até o Centro de Convivência. Obedecendo essas instruções você chega o C&T e se depara com outra bifurcação e mais uma vez te dizem para seguir a direita, ganhando dessa forma mais um bit de informação. Após esse percurso, você chega a Departamento de Informática e Matemática Aplicada (DIMAP), onde mais uma vez se depara com uma bifurcação e novamente pede informação. Dessa vez, pra sua alegria, te informam que o departamento é o próximo bloco a esquerda, e que após esse longo caminho você poderá chegar ao seu destino.

Se representarmos as instruções por meio de dígitos binários (0 =esquerda, 1 = direita) podemos indicar o caminho até chegar ao Departamento de Física por meio da sequência (110) que nos especifica que nas duas primeiras bifurcações devemos seguir pela rua à nossa direita enquanto que na última bifurcação seguimos pela esquerda.

Veja que o caminho percorrido até chegar ao DIMAP é um dos quatro equiprováveis caminhos que poderiam ser atingidos após passar por duas bifurcações, ou seja, os dois bits de informação que te foram fornecidos permitem ir a um (DIMAP) dos quatro destinos

[1]Para mais detalhes sobre este capítulo veja [25, 26, 27]

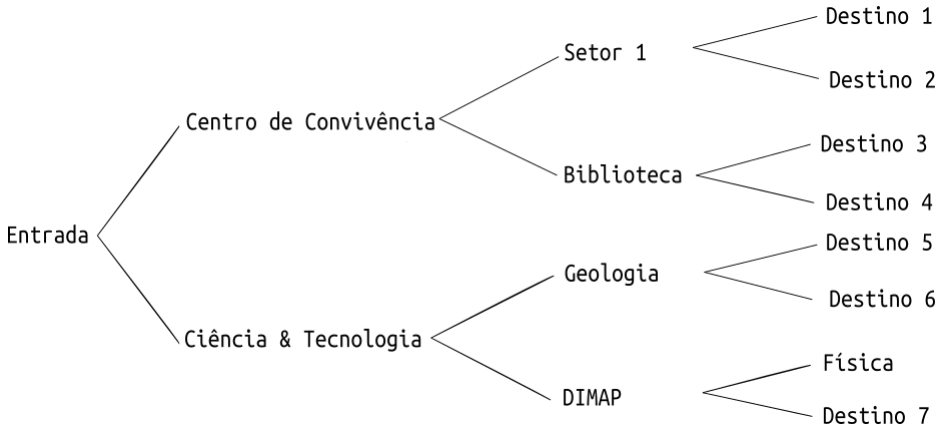

Figura 3.1: Possíveis trajetórias que podem ser seguidas em um mapa hipotético da UFRN. Para um individuo que não conheça a universidade, cada bifurcação requer um bit de informação para decidir pela direção correta.

possíveis (Setor 1, biblioteca, geologia e DIMAP). Partindo da entrada do campus, três bits de informação nos permitem chegar no ponto desejado, o departamento de física, dentre os oito equiprováveis destinos possíveis. Ao decidir por uma direção na primeira bifurcação, metade dos oito possíveis destinos foram eliminados. Da mesma maneira, cada decisão tomada em um bifurcação elimina a metade do número de possíveis destinos que você pode chegar.
Em resumo:

- Um bit de informação nos permite escolher entre duas alternativas equiprováveis.
- Dois bits nos permitem chegar em um dos quatro destinos de igual probabilidade.
- Três bits nos levam a uma alternativa entre 8 equiprováveis

Note que o número de destinos m aumenta com a quantidade de bifurcações n, em outras palavras se possuímos n bits de informação então podemos chegar a m destinos igualmente prováveis. Matematicamente expressamos essa ideia por meio da relação

$$m = 2^n. \tag{3.1}$$

Se por outro lado, sabemos o número m de destinos possíveis, quantos bits (ou quantas bifurcações) são necessários para chegar em um desses m destinos permitidos? Podemos utilizar a função logaritmo na base dois em ambos os lados da equação acima para obter

$$n = \log_2 m \tag{3.2}$$

O logaritmo de m é a potência n à qual 2 deve ser elevado para obter m. Pode parecer pouco que um bit de informação nos permita escolher apenas entre duas alternativas equiprováveis mas a Eq. (3.1) nos diz que dobra-se quantidade de destinos possíveis m sempre que fornecemos mais um bit de informação. Considere por exemplo que tenhamos $n = 10$ bifurcações (ou bits) ao longo de toda a trajetória, de forma que teremos $m = 2^{10} = 1024$ resultados possíveis. Aumentando para $n = 20$ a quantidade de bifurcações teremos $m = 2^{20} = 1048576$ possíveis destinos. Esses 20 bits de informação te permitem escolher um caminho entre mais de 1 milhão de possibilidades.

3.2 Elementos da teoria

Em teoria de informação encontramos um vocabulário que é próprio à teoria e que será introduzido a partir de agora. Primeiramente temos uma **fonte** que produz uma **mensagem** de qualquer natureza podendo ser uma carta, imagem, etc. A mensagem é compactada (ou codificada) por um **codificador** e enviada através de um **canal de comunicação** até chegar ao receptor **(decodificador)** que a decodifica e lê[2]. A mensagem é constituída por uma sequência de k símbolos onde cada um deles corresponde a uma variável aleatória.

$$S = \{s_1, s_2, \ldots, s_k\}. \tag{3.3}$$

Se a mensagem for uma carta então cada letra é um dos k símbolos que a constituem. Cada símbolo k que forma a mensagem é um valor resultante de uma variável aleatória S que pode adotar qualquer valor de um alfabeto com símbolos α diferentes

$$A_s = \{s_1, \ldots, s_\alpha\}. \tag{3.4}$$

Ainda considerando a carta como exemplo, o conjunto α é formado pelas letras do alfabeto, de onde cada simbolo k é retirado. A probabilidade que um símbolo seja gerado pela fonte é dada pela distribuição de probabilidade

$$P(S) = \{p(s_1), \ldots, p(s_\alpha)\}, \tag{3.5}$$

com a normalização:

$$\sum_{i=1}^{\alpha} P(s_i) = 1. \tag{3.6}$$

Nem sempre é possível fazer a transmissão da mensagem de um lado para outro sem que ela seja modificada. Essa perturbação que a mensagem sofre pode ocorrer por diversos fatores como a própria natureza do canal e é chamada de ruído. Se os dados são

[2]Mais adiante, apenas por simplificação, chamaremos todo esse processo de tratamento T

transmitidos pelo canal de modo que a informação que entrou não tenha sido alterada, então dizemos que a informação foi transmitida com sucesso.

Antes do envio, a mensagem é codificada buscando compactar a informação e/ou eliminar redundâncias próprias da mensagem antes de ser passada pelo canal. Lembre-se que nossa mensagem está ordenada em uma sequencia de k símbolos. A codificação pode ser representada por uma função g tal que $x = g(s)$, onde x é a representação da codificação do simbolo s e forma uma sequência de palavras-código,

$$X = \{x_1, \ldots, x_n\}, \tag{3.7}$$

onde cada palavra-código é o valor de uma variável aleatória X o qual pode adotar qualquer um dos m diferentes valores de um código-livro

$$A_x = \{x_1, \ldots, x_m\}. \tag{3.8}$$

A probabilidade de cada palavra-código é definida pela distribuição

$$P(X) = \{p(x_1), ..., p(x_m)\}. \tag{3.9}$$

A fim de tornar mais claras essas definições, considere que tenhamos um dado de quatro lados e realizamos oito lançamentos com ele, obtendo como resultados os valores hipotéticos $S = \{4, 2, 1, 4, 3, 3, 2, 1\}$. Queremos transmitir uma mensagem contendo os oito resultados do lançamento, o conjunto S, para uma pessoa que esteja distante de nós. Note que cada símbolo da mensagem (cada elemento do conjunto S) é o valor resultante de uma variável aleatória que pode adotar qualquer valor entre os quatro valores possíveis do lançamento lançamento do dado tal que $A_s = \{1, 2, 3, 4\}$. Uma vez que os lançamento são independentes, a distribuição de probabilidade $P(S)$ de escolher um dos símbolos gerados pela fonte é uniforme $P(S) = \{1/4, 1/4, 1/4, 1/4\}$.

O passo seguinte consiste em codificar a mensagem. Lembre-se da discussão anteriormente onde vimos que com um bit de informação somos capazes de escolher entre dois valores equiprováveis e com 2 bits podemos representar um entre quatro caminhos de igual probabilidade. Com essa informação podemos representar, em nosso exemplo, qualquer um dos quatro elementos do conjunto S atribuindo a eles um valor formado por dois dígitos binários como mostrado na Tab. 3.1.

Utilizando a tabela podemos codificar nossa mensagem obtendo o conjunto $X = \{11, 01, 00, 11, 10, 10, 01, 00\}$. Esse conjunto é formado é formado pelo valor das variáveis aleatórias do conjunto $A_x = \{00, 01, 10, 11\}$.

Após ser transmitida a mensagem produz como saída o conjunto

$$Y = \{y_1, \ldots, y_m\}, \tag{3.10}$$

onde o valor de saída é uma variável aleatória de Y que pode assumir m diferentes valores

$$A_m = \{y_1, \ldots, y_m\}. \tag{3.11}$$

Se o canal apresentar algum tipo de ruído então o valor de entrada x_j pode ser diferente do valor de saída y_j e cuja probabilidade de cada resultado é dada pela distribuição de probabilidade

$$P(Y) = \{p(y_1), \ldots, p(y_m)\}. \tag{3.12}$$

Símbolo	Palavra-código
1	00
2	01
3	10
4	11

Tabela 3.1: Tabela de codificação da mensagem S.

Uma maneira de medir o quanto de informação que passa pelo canal, é considerando a quantidade de informação por simbolo ou, a forma mais usual, como a quantidade de informação por segundo (bits/s) e denominamos essa grandeza como **taxa de transmissão**. Todo canal possui uma quantidade máxima de informação que pode ser transmitida sem que a mensagem seja alterada durante o processo. Tal limite para transmissão é denominado **capacidade do canal**. Note que a taxa de transmissão da informação é menor ou igual à capacidade do canal mas nunca maior.

3.3 Surpresa e Entropia

Agora que conhecemos os principais elementos da teoria, caminhamos em direção de definir o que é informação. Claude Shannon, em seu trabalho de 1961, estabeleceu uma relação entre probabilidade e informação. Ele observou que para se obter uma definição matemática para informação era necessário que a mesma possuísse algumas propriedades que foram sumarizadas no chamado **Desiderata de Shannon** [27]. Esse conjunto de propriedades exigia que a função que representa informação fosse:

- Contínua: Pois se a probabilidade de um resultado muda então a quantidade informação associada a ele também deve mudar de forma contínua.
- Simétrica: A quantidade de informação de uma sequência de resultados não deve depender da ordem na qual esses resultados ocorrem.
- Com valor máximo: Este valor ocorre quando os resultados de um experimento aleatório são igualmente prováveis
- Aditiva: A informação de um conjunto de resultados pode ser obtida somando a informação de cada um dos resultados.

Começaremos nossa discussão considerando uma moeda viciada que possui probabilidade de 80% de se obter cara como resultado em um lançamento. Diante de uma probabilidade tão alta em se obter um resultado, você não se surpreende ao observar cara após um lançamento. A **surpresa de um evento** está intimamente ligada a probabilidade de sua ocorrência, de forma que quanto mais improvável é o resultado mais você se surpreende ao observá-lo.

Uma maneira de expressar a surpresa em obter um resultado x é 1/probabilidade de x ou $1/p(x)$, dessa maneira a surpresa em se obter um resultado x aumenta se a probabilidade de observá-lo diminui. Dizemos que esse evento cara tem pouca informação. Para satisfazer a condição de aditividade, Shannon notou que a melhor maneira de definir a surpresa de um evento é através do logaritmo de $1/p(x)$ que recebe o nome de **informação de Shannon do evento** x [27].

Se usarmos o logaritmo na base 2 a informação de Shannon $h(x)$ é

$$h(x) = \log \frac{1}{p(x)} = -\log p(x) \tag{3.13}$$

e é medida em bits[3]. A informação de Shannon nos diz o quão surpresos ficamos ao observarmos um evento e varia de forma inversamente proporcional à probabilidade de ocorrência (Veja Fig. 3.2). Como $0 \leq p(x) \leq 1$ então no limite em que $p(x)$ é muito baixa (se aproxima de 0) então $h(x)$ tende ao infinito (é infinitamente surpreendente que um evento impossível aconteça). Por outro lado, se $p(x) = 1$ então $h(x) = 0$ (não é surpreendente que um evento certo se realize.)

Note que, a função logaritmo satisfaz as propriedades de simetria e continuidade além da aditividade já mencionada. Essa foi a conexão entre informação e probabilidade estabelecida por Shannon.

[3] A partir de agora usaremos a função logaritmo na base 2.

Note também que, da forma como está definida, para determinar a surpresa de um evento precisamos saber qual é a distribuição de probabilidades $p(x)$ dos eventos x. Como estamos mais interessados em saber a surpresa de um conjunto de resultados, definimos a surpresa média $H(x)$ de uma variável X cuja a distribuição de probabilidade é $p(X)$ como

$$H(X) \approx \frac{1}{n} \sum_{i=1}^{n} \log \frac{1}{p(x_i)}. \tag{3.14}$$

Uma maneira de interpretar $H(X)$, nesse contexto, é lembrando que geralmente a entropia de uma variável é igual ao logaritmo do número de resultados equiprováveis m:

$$H(x) = \log m \text{ bits}. \tag{3.15}$$

No caso em que $m = 2$ a entropia é igual a um bit com já vimos. Isso significa que a entropia para o lançamento de uma moeda (com dois resultados equiprováveis) é igual a 1 enquanto que o dado de 4 lados (que nos fornecia 4 resultados equiprováveis) a entropia é $H = \log 4 = 2$ bits. Elevando ambos os lados da Eq. (3.15) por dois, temos

$$m = 2^{H(x)}. \tag{3.16}$$

A Eq. (3.16) nos diz que uma variável aleatória cuja entropia é $H(X)$, fornece uma informação de Shannon para escolher entre $m = 2^{H(x)}$ resultados equiprováveis. Se, por exemplo, $H(x) = 3$ então $m = 8$. Esse caso é equivale as trajetórias possíveis que um caminhante tinha e que foi discutido na primeira seção deste capítulo.

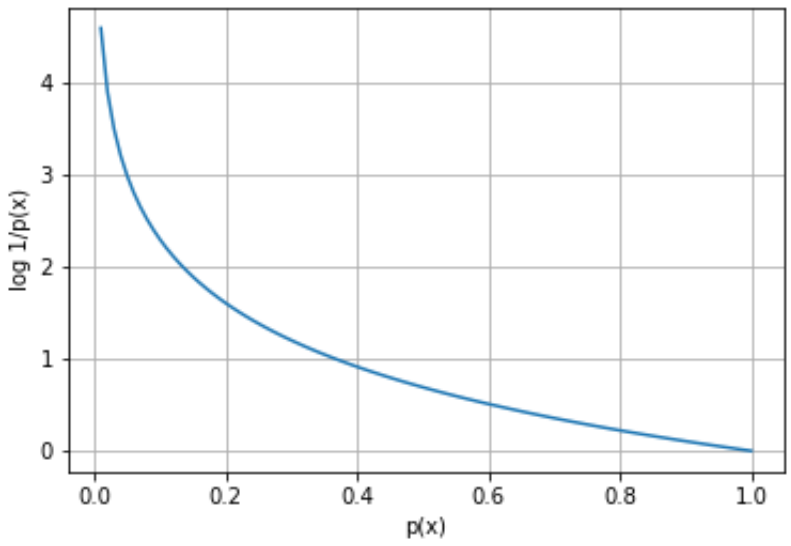

Figura 3.2: Relação de proporcionalidade inversa entre surpresa (informação) e $p(x)$.

A fim de explorar a ideia de entropia, vamos considerar dois exemplos que utilizam a chamada entropia binária em que dois resultados são possíveis com probabilidades p e $q = 1 - p$. O primeiro exemplo é uma moeda não viciada cuja a probabilidade de se obter cara em um lançamento é $p(x_{cara}) = 0.5$. Então a surpresa pra esse resultado é

$$h(x) = \log\left[\frac{1}{p(x_{cara})}\right] = \log\left[\frac{1}{0.5}\right] = 1 \text{ bit.} \tag{3.17}$$

Até então nada de surpreendente considerando nossa discussão anterior e, obviamente, o mesmo resultado é obtido para coroa, uma vez que $p(x_{coroa}) = 0.5$. A fim de determinar a surpresa média de um lançamento de moedas, consideremos que fizemos 100 lançamentos. Como o resultado de cada um deles não interfere em outro dizemos que são eventos independentes e é razoável esperar que em metade deles observa-se cara e na outra metade coroa. Sob essas condições a surpresa média, dada pela Eq. (3.14), é

$$\begin{aligned} H(X) &= \frac{\sum_{j=1}^{50} \log\left[\frac{1}{p(x_{cara})}\right] + \sum_{j=1}^{50} \log\left[\frac{1}{p(x_{coroa})}\right]}{100} \\ &= \frac{50 \times \log\left[\frac{1}{p(x_{cara})}\right] + 50 \times \log\left[\frac{1}{p(x_{coroa})}\right]}{100} \\ &= 0.5 \times \log\left[\frac{1}{0.5}\right] + 0.5 \times \log\left[\frac{1}{0.5}\right] \\ H(X) &= 1 \text{ bit por moeda} \end{aligned} \tag{3.18}$$

Em resumo, como a quantidade de surpresa fornecida ao observarmos o resultado de cada um dos lançamentos é um bit, segue que a informação média $H(X)$ de cada lance também é um.

O segundo exemplo consiste de uma moeda viciada cuja probabilidade de obter coroa é de 90%. Como um dos resultados tem probabilidade muito alta então a quantidade de surpresa associada ao resultado é menor do que para uma moeda justa como no gráfico da Fig. 3.3. Nessa condição é fácil prever que a maioria das vezes em que a moeda for lançada o resultado será cara. A informação de Shannon para cara é

$$h(x_{cara}) = \log\left[\frac{1}{0.9}\right] = 0.15 \text{ bit,} \tag{3.19}$$

enquanto que para coroa

$$h(x_{coroa}) = \log\left[\frac{1}{0.1}\right] = 3.32 \text{ bit.} \tag{3.20}$$

Note que, como esperado, mais informação está associada com o resultado menos provável (coroa nesse caso). Adotaremos o mesmo procedimento anterior para analisar a

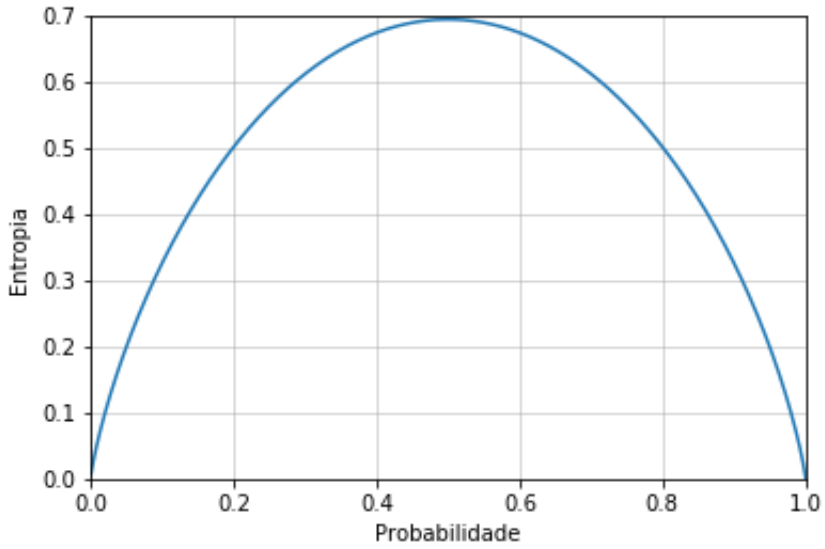

Figura 3.3: Função entropia binária. Perceba que o valor máximo é a quando a probabilidade p= 0.5. Veja que a entropia assume valor máximo quando os resultados são equiprováveis.

entropia: Lançamos 100 vezes a moeda e como os lançamentos são independentes entre si esperamos observar que em 90 vezes o resultado é cara e 10 coroas. Segue que:

$$\begin{aligned} H(X) &= \frac{\sum_{j=1}^{90} \log\left[\frac{1}{p(x_{cara})}\right] + \sum_{j=1}^{10} \log\left[\frac{1}{p(x_{coroa})}\right]}{100} \\ &= \frac{90 \times \log\left[\frac{1}{p(x_{cara})}\right] + 10 \times \log\left[\frac{1}{p(x_{coroa})}\right]}{100}. \end{aligned}$$

Substituindo $p(x_{cara}) = 0.9$ e $p(x_{coroa}) = 0.1$

$$H(X) = 0.9 \times \log\left[\frac{1}{0.9}\right] + 0.10 \times \log\left[\frac{1}{0.1}\right] = 0.489. \tag{3.21}$$

Se você pensar intuitivamente irá concordar que a moeda viciada tem menos surpresa ao se obter um resultado, uma vez que um deles tem alta probabilidade. Se lembrarmos que $p(x_{cara}) = 0.9$ e $p(x_{coroa}) = 0.1$ podemos reescrever (3.18) e (3.21) de forma sucinta

$$H(X) = \sum_{i=1}^{2} p(x_i) \log \frac{1}{p(x_i)}. \tag{3.22}$$

Ou de forma geral, para m resultados

$$H(X) = \sum_{i=1}^{m} p(x_i) \log \frac{1}{p(x_i)}. \tag{3.23}$$

Uma entropia de 0.469 implica que podemos representar uma informação como o lançamento de 1000 moedas utilizando 469 dígitos binários. Nesse exemplo, com H(x)=

0.469 bits, a variável X pode ser representada por

$$m = 2^{H(x)} = 2^{0.469} = 1.38 \tag{3.24}$$

valores de resultados equiprováveis. Parece estranho pensar em m como sendo um número não inteiro mas podemos imaginar uma moeda com entropia H(x) = 0.469 como tendo a mesma entropia de um dado imaginário de 1.38 lados.

3.4 Entropia Relativa e Divergência

A partir de agora vamos generalizar um pouco nossa discussão. A entropia é uma média tomada sobre um conjunto X, conforme a Eq. (3.23). Afim de simplificar a notação, a partir de agora vamos utilizar um conceito muito usando em estatística. A **esperança** de uma variável aleatória $f(X)$ é a média dos valores $f(x)$ de acordo com a distribuição de probabilidades $p(x)$

$$\mathbf{E}f(X) = \sum_x p(x)f(x),$$

para o caso discreto. No caso contínuo, a esperança é uma integral $\int_x f(x)p(x)dx$ que é bem definida e finita se $f(x)$ é integrável com relação à medida de probabilidade $p(x)dx$. Com essa notação podemos reescrever a entropia (3.23) como

$$H(X) = \mathbf{E}\log\frac{1}{p(x)}. \tag{3.25}$$

Considere uma variável aleatória X de distribuição de probabilidade $p(x)$ e que por alguma razão também tenhamos outra distribuição $q(x)$ definida sobre o mesmo conjunto X. A divergência $D(p,q)$ fornece uma medida de desvio ou divergência de $q(x)$ com relação a $p(x)$

$$D(p,q) = \sum_x p(x)\log\frac{p(x)}{q(x)} = \mathbf{E}\log\frac{p(x)}{q(x)} \tag{3.26}$$

A divergência possui uma propriedade muito importante:

- $D(p,q) \geq 0$ com a igualdade sendo válida, se e somente se, $p(x) = q(x)$.

Demonstração. Vamos supor que o logaritmo seja natural. Nesse caso, a função logaritmo possui uma reta tangente no ponto $x = 1$. Expressamos essa ideia por meio da desigualdade $\log x \leq x - 1$ com a igualdade válida se, e somente se, $x = 1$. Desse maneira,

$$D(p,q) = \mathbf{E}\log\frac{p(x)}{q(x)} = -\mathbf{E}\log\frac{q(x)}{p(x)} \leq \mathbf{E}\left[\frac{q(x)}{p(x)} - 1\right]$$

$$-D(p,q) \leq \sum_x p(x)\left[\frac{q(x)}{p(x)} - 1\right] = \sum_x [q(x) - p(x)].$$

Os termos do lado direito são $\sum_x q(x) - \sum_x p(x) = 1 - 1 = 0$ e portanto $D(p,q) \geq 0$, para todo x com a igualdade válida somente para $\frac{q(x)}{p(x)} = 1$. □

Em teoria da informação, estudamos os princípios matemáticos utilizados para compreensão dos sistemas de comunicação por meio de um tratamento entre a entrada e saída. O tratamento T, que consiste em codificar, enviar e decodificar uma mensagem, é uma descrição probabilística que modela toda a transformação e transmissão de dados.

Em um bloco de tratamento T, a saída $y \in Y$ é uma variável aleatória que depende aleatoriamente da entrada $x \in X$. A dependência entre elas é caracterizada pela distribuição de probabilidade condicional:

$$p(y|x) = \frac{p(x,y)}{p(x)}, \tag{3.27}$$

onde $p(x,y)$ é a probabilidade conjunta de se obter x e y enquanto que $p(x)$ é a probabilidade associada à entrada x.

A expressão (3.27) nos diz qual a probabilidade de se obter o valor de saída y dado que o valor de entrada foi x. Assim é possível obter a distribuição de probabilidade da saída $p(y)$ para cada valor da entrada x através das probabilidades condicionais $p(y|x)$

$$p(y) = \sum_x p(x,y) = \sum_x p(y|x)p(x) \tag{3.28}$$

O tratamento T é então definido se conhecermos os alfabetos de entrada X, os de saída Y e as distribuições condicionais $p(y|x)$.

Consideremos um tratamento T que modela um processo de transmissão de dados de entrada X e saída Y. Definimos a quantidade $I(X,Y)$ e denominamos de **informação mútua**. Essa grandeza mede a quantidade média de informação que se obtém sobre o conjunto de entrada X quando se conhece as variáveis de saída Y. Sendo assim, se os conjuntos de entrada e saída forem independentes entre si então o conhecimento de Y não nos fornece nenhuma informação sobre X, ou seja, $I(X,Y) = 0$. Por outro lado, se X depende fortemente de Y a informação que se obtêm é muito mais relevante.

Uma vez que as variáveis aleatórias X e Y são independentes quando sua distribuição de probabilidades conjunta $p(x,y)$ é o produto das distribuições $p(x,y) = p(x)p(y)$, escrevemos a informação mútua $I(X,Y)$ como sendo a divergência $D(p,q)$ da distribuição de probabilidade $q(x,y) = p(x)p(y)$ com a relação a $p(x,y)$

$$I(X,Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{3.29}$$

chamamos a expressão (3.29) de informação mútua ou informação de Shannon entre X e Y.

Como foi definida a partir da divergência, a informação de Shannon conserva a mesma propriedade de positividade, $I(X,Y) \geq 0$ com a igualdade sendo verdadeira se, e somente se, X e Y forem independentes. Uma propriedade adicional da informação de Shannon é a simetria em X e Y, $I(X,Y) = I(Y,X)$.

Podemos confirmar nossa intuição de que $I(X,Y)$ mede uma informação se reescrevermos a (3.29) utilizando a probabilidade condicional (3.27)

$$I(X,Y) = \sum_{x,y} p(x|y)p(y) \log \frac{p(x|y)}{p(x)} = \sum_{y} p(y) \sum_{x} p(x|y) \log \frac{p(x|y)}{p(x)},$$

$$I(X,Y) = \mathbf{E}_y D(p(x|y), p(x)). \tag{3.30}$$

Observe que a soma realizada sobre x é a divergência entre a distribuição de probabilidade conhecendo y, isto é probabilidade condicional $p(x|y)$, e aquela quando não é conhecido o valor de y, $p(x)$. Sendo assim, $I(X,Y)$ mede o desvio que produz o conhecimento de uma saída y sobre a distribuição de entrada X. Em seguida, calcula-se a média desses desvios sobre y.

3.5 Informação mútua e Entropia

Vamos agora estabelecer a relação entre a informação mútua e entropia. Para isso escrevemos:

$$I(X,Y) = \mathbf{E} \log \frac{p(X,Y)}{p(X)p(Y)} \tag{3.31}$$

$$= \mathbf{E} \log \frac{1}{p(X)} + \mathbf{E} \log \frac{1}{p(Y)} - \mathbf{E} \log \frac{1}{p(X,Y)}$$

Cada termo da soma acima corresponde respectivamente à entropia da variável X, $H(X)$, da Y, $H(Y)$ e a *entropia conjunta* do par (X,Y), de maneira que obtemos a relação:

$$I(X,Y) = H(X) + H(Y) - H(X,Y). \tag{3.32}$$

Como $I(X,Y) = H(X) + H(Y) - H(X,Y) \geq 0$ nos diz que o par (X,Y) é globalmente menos incerto que se tomarmos X e Y separadamente.

$$H(X) + H(Y) \geq H(X,Y) \tag{3.33}$$

A diferença é a informação mútua entre X e Y.

Uma consequência mais interessante pode ser observada se isolarmos na expressão $I(X,Y)$, dada por (3.31), os termos dependentes de apenas uma variável. Isso pode ser

realizado se relembrarmos que $p(X,Y) = p(X|Y)p(Y)$

$$I(X,Y) = \mathbf{E}\log\frac{p(X|Y)}{p(X)}$$

$$I(X,Y) = \mathbf{E}\log\frac{1}{p(X)} - \mathbf{E}\log\frac{1}{p(X|Y)}. \tag{3.34}$$

O primeiro termo é a entropia da variável X, $H(X)$, e o segundo termo é chamado de *entropia condicional* de X conhecendo Y e é

$$H(X|Y) = \mathbf{E}\log\frac{1}{p(X,Y)} = \sum_{x,y} p(x,y)\log\frac{1}{p(x|y)}. \tag{3.35}$$

Utilizando mais uma vez a Eq. (3.27) podemos reescrevemos a entropia condicional como

$$H(X|Y) = \sum_y p(y)\sum_x p(x|y)\log\frac{1}{p(x|y)}$$

$$H(X|Y) = \mathbf{E}_y H(X|Y=y) \tag{3.36}$$

onde $H(X|Y=y) = \sum_x p(x|y)\log\frac{1}{p(x|y)}$ é a entropia da variável aleatória X conhecendo um dado valor $Y = y$. Observe que $H(X|Y=y)$ é uma função de y de maneira que se tomamos a sua esperança sobre os valores de y obtemos a entropia condicional $H(X|Y)$. Resumidamente, a informação mútua $I(X,Y)$ pode ser escrita em termos da entropia do conjunto X e a entropia condicional $H(X|Y)$

$$I(X,Y) = H(X) - H(X|Y). \tag{3.37}$$

A entropia condicional (3.36) é uma média sobre Y da quantidade $H(X|Y=y)$ e mede a incerteza sobre X conhecendo um valor especifico $Y = y$. Note que se $H(X|Y=y) \geq 0$ então $H(X|Y) \geq 0$. A entropia condicional $H(X|Y)$ mede a incerteza média que permanece sobre X uma vez que Y é conhecido. Uma vez que $H(X)$ mede a informação média sobre X e $H(X|Y)$ mede a incerteza sobre X após o conhecimento de Y fica claro, avaliando a (3.37), que é possível avaliar a informação mútua como sendo a diferença entre a incerteza associada ao conjunto X e a incerteza em X dado que temos alguma informação de Y. Como $I(X,Y) = I(Y,X)$ podemos trocar os papéis entre X e Y

$$I(X,Y) = H(Y) - H(Y|X). \tag{3.38}$$

Podemos observar algumas propriedades. Como $I(X,Y) = H(X) - H(X|Y) \geq 0$ então $H(X) \geq H(X|Y)$, isso significa que a incerteza sobre X diminui quando temos algum conhecimento sobre Y. É importante ressaltar que esse resultado é válido na média. A incerteza sobre X diminui quando realizamos a média sobre todo o conjunto Y.

A incerteza média permanece inalterada quando $H(X) = H(X|Y)$, isto é quando X e Y são independentes. Isso não é uma surpresa uma vez que o conhecimento de uma variável aleatória Y não tem influência sobre a incerteza de X.

A entropia condicional $H(X|Y) = 0$ se, e somente se, $H(X|Y = y) = 0$ para todo y. Uma vez $H(X|Y)$ é uma média em Y tomada sobre a incerteza $H(X|Y = y)$. Isso significa que X é uma função determinística de Y, do tipo $X = f(Y)$. Em outras palavras, dado um elemento de X então o valor de saída $Y = y$ é certo. Em resumo:

- $H(X|Y) \geq 0$ com a igualdade válida se, e somente se, $X = f(Y)$

No caso particular que $X = Y$ então $H(X|Y) = 0$. Portanto:

$$I(X, X) = H(X), \tag{3.39}$$

é chamada de informação intrínseca ou autoinformação trazida, em média, pelo conhecimento de X sobre ele mesmo. Vimos anteriormente que a informação mútua, dada por (3.37), é uma medida de informação obtida a partir da diferença entre entropias. Aqui, novamente, encontramos a interpretação entropia, $H(X)$, como sendo a medida de informação a partir da informação mútua $I(X, X)$.

A autoinformação carrega a propriedade de sempre ser positiva ou nula $H(X) \geq 0$ uma vez que $I(X, X)$ sempre o é. Com a igualdade sendo satisfeita se, e somente se X é independente de si mesma. Isto é, se X é uma variável determinística.

3.6 Diagrama de Venn

Ao longo deste capítulo vimo que, fornecendo as variáveis aleatórias de entrada X e de saída Y, conseguimos obter diversas medidas de informação $H(X), H(Y), H(X, Y), H(X|Y)$, $H(Y|X), I(X, Y)$ e estabelecer várias relações entre elas:

- $I(X, Y) = H(X) - H(X|Y)$
- $I(X, Y) = H(Y) - H(Y|X)$
- $I(X, Y) = H(X) + H(Y) - H(X, Y)$

De onde é possível obter:

$$H(X, Y) = H(X|Y) + I(X, Y) + H(Y|X). \tag{3.40}$$

Para finalizar essa parte da dissertação, vamos apresentar uma maneira de visualizar essas relações por meio de um diagrama que é mostrado na Fig. 3.6. Essa representação

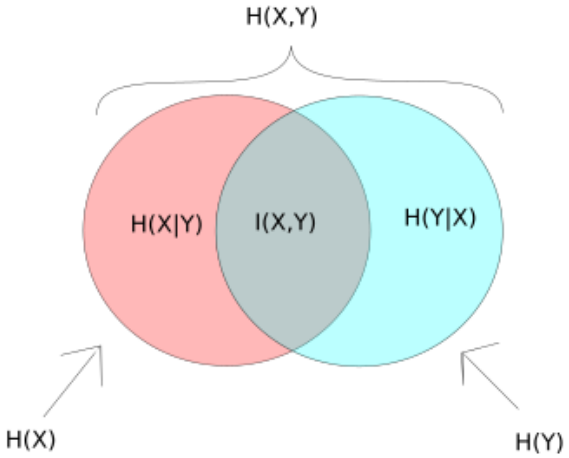

Figura 3.4: Diagrama de Venn. A informação mútua corresponde a intersecção da informação em X com a em Y.

é chamada de **diagrama de Venn** e nela indicamos as incertezas $H(X)$ e $H(Y)$ como sendo duas regiões cuja intersecção é a informação mútua $I(X,Y)$ e a união é a entropia condicional $H(X,Y)$. A partir do diagrama é possível encontrar as igualdades e desigualdades que obtivemos ao longo da seção sem precisar estabelecê-las a partir das definições.

Podemos obter por exemplo:

- $I(X,Y) = I(Y,X)$
- $H(X,Y) = H(X|Y) + H(Y|X) + I(X,Y)$
- $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$
- $H(X) = H(X|Y) + I(X,Y)$,

e as desigualdades

- $H(X) \geq H(X|Y)$
- $H(Y) \geq H(Y|X)$
- $H(X,Y) \geq H(X) + H(Y)$.

Utilizando o diagrama de Venn podemos observar alguns casos particulares:

Se os conjuntos X e Y são independentes então não há uma intersecção entre os conjuntos $H(X)$ e $H(Y)$ e, portanto, $I(X,Y) = 0$.

Se os conjuntos X e Y são equivalentes, ou seja $H(X) = H(Y)$, então $H(X|Y) = H(Y|X) = 0$.

Se X é uma função de $Y = f(X)$, então o conjunto $H(Y)$ está contido dentro de $H(X)$ e portanto $H(Y|X) = 0$. A reciproca é verdadeira para Y sendo uma função de $X = g(Y)$ com $H(X)$ contido em $H(Y)$.

4. A Física e o DNA

O estudo dos genes sempre chamou a atenção de pesquisadores de diversas áreas. Prova disso é que durante a descoberta da estrutura do DNA, tínhamos o envolvimento de físicos como Francis Crick e Rosalind Franklin, químicos e outros [37]. Esse interesse multidisciplinar deve-se em grande parte pela relevância que a genética possui, no sentido de trazer respostas importantes sobre a vida, a existência e o impacto delas na sociedade humana. Outro fator relevante está relacionado à complexidade dos genes. O compartilhamento de informações e técnicas de outros campos do conhecimento, como a difração e o raio x, foi o que nos permitiu fazer as descobertas que seriam impossíveis sem um trabalho multidisciplinar.

Este capítulo tem por objetivo mostrar algumas contribuições da física para a genética. Para isso, reunimos alguns artigos relevantes nesse campo e discutimos os aspectos importantes dos modelos propostos por esses trabalhos e apresentamos os resultados obtido por eles. É importante salientar que este capítulo não tem a pretensão de discutir em detalhes os artigos citados mas apenas fornecer uma visão geral e os principais resultados de cada um deles.

4.1 Correlação de longo alcance em sequências de nucleotídeos

Um dos primeiros modelos utilizados para analisar as propriedades estatísticas da sequência de nucleotídeos foi desenvolvido em 1992 por Peng[1]. Nesse modelo, chamado de *DNA walk*, os autores construíram um mapeamento da sequência de DNA em uma caminhada aleatória e a partir desse mapeamento estabeleceram uma medida quantitativa de correlação entre os nucleotídeos ao longo da cadeia de DNA [12].

No modelo unidimensional do caminhante aleatório clássico, um indivíduo pode ca-

[1] Está seção é baseada no artigo [12]

minhar uma unidade de comprimento u indo para cima ($u(i) = +1$) ou para baixo ($u(i) = -1$) ao longo de uma reta, a cada passo i da caminhada. No caso de uma caminhada correlacionada, cada passo depende da história (memória) do caminhante, isto é, a direção para o passo seguinte irá depender de passos anteriores. No modelo de *DNA walk*, se uma pirimidina for encontrada na posição i ao longo da cadeia de DNA então o caminhante dará um passo para cima $u(i) = +1$ e se ocorrer uma purina na posição i então o passo será dado na direção de baixo $u(i) = -1$.

O *DNA walk* nos fornece uma representação gráfica para cada gene e nos permite visualizar diretamente a correlação na sequência de nucleotídeos. O passo seguinte da análise é quantificar essa correlação. Para isso estabelece-se o chamado 'deslocamento da rede' de um caminhante após l passos e que corresponde à soma das unidades de caminhada $u(i)$ para cada passo i

$$y(l) = \sum_{i=1}^{l} u(i). \tag{4.1}$$

Outra quantidade estatística importante para qualquer caminhada é a flutuação quadrática média $F(l)$ sobre a média do deslocamento

$$F(l) = \overline{[\Delta y(l) - \overline{\Delta y(l)}]^2} = \overline{[\Delta y(l)]^2} - [\overline{\Delta y(l)}]^2, \tag{4.2}$$

onde a quantidade $\Delta y(l)$ é definida por:

$$\Delta y(l) = y(l_0 + l) - y(l_0). \tag{4.3}$$

A flutuação média (4.2) é definida em termos da média quadrática e do quadrado da média. As barras indicam média sobre todas as posições l_0 no gene.

Podemos relacionar a flutuação média quadrática $F(l)$ com a função de autocorrelação $C(l) = \overline{u(l_0)u(l_0 + l)} - [\overline{u(l_0)}]^2$ através da relação $F^2(l) = \sum_{i=1}^{l} \sum_{j=1}^{l} C(j-i)$. É importante ressaltar que a flutuação em $F(l)$ é substancialmente reduzida em comparação com a flutuação em $C(l)$ uma vez que essas grandezas se relacionam por meio de uma soma. A consequência disso é que a medida de $F^2(l)$ leva a uma caracterização mais confiável da do modelo de *DNA walk* do que a medida de $C(l)$.

O cálculo de $F(l)$ pode identificar três tipos de comportamento:

- Se a sequência de nucleotídeos é aleatória, então $C(l) = 0$ na média, indicando que não há correlação na sequência. Exceto $C(l = 0) = 1$ e então $F(l) \sim l^{1/2}$.

- Caso haja correlação local ao longo de uma faixa característica R então $C(l) \sim \exp(-l/R)$.

- Se não houver comprimento característico, isto é se a correlação apresentar "alcance infinito", a propriedade de escala $C(l)$ não será uma exponencial mas uma lei de potência e a flutuação será descrita por

$$F(l) \sim l^{\alpha}. \tag{4.4}$$

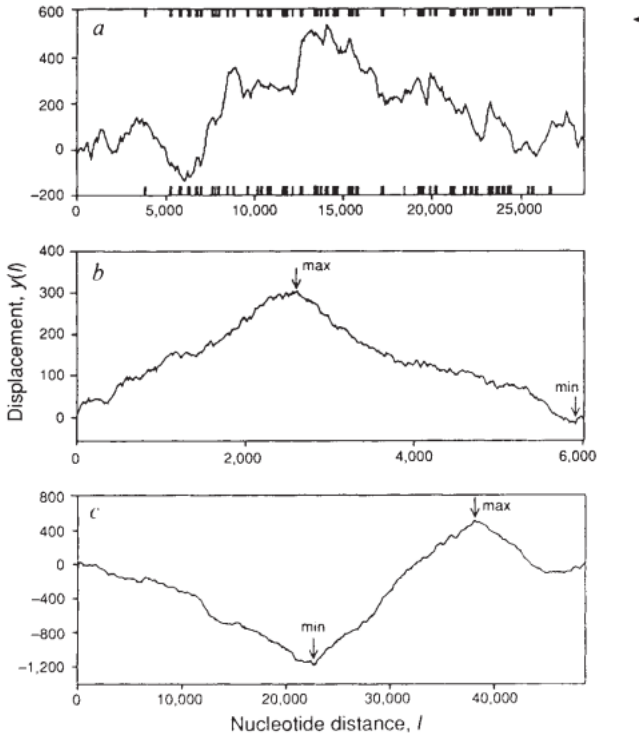

Figura 4.1: Representação do DNA *walk* de uma sequência humana: (a) Cadeia pesada de miosina β-cardíaca rica em íntrons; (b) seu cDNA e (c) sequência de DNA de bacteriófago sem íntron. Observe que as flutuações são mais complexas para genes que contém íntrons (a) do que para as sequências sem íntrons (b) e (c). As barras mais grossas em (a) indicam as regiões codificantes do gene. Podemos perceber que a representação do *DNA walk* não é capaz de distinguir as concentrações de pirimidinas e purinas. Os pontos de mínimo e máximo são indicados por setas. Esses picos indicam a concentração do conteúdo de nucleotídeos: pirimidinas ou purinas [12].

Para observar como o modelo funciona, os autores consideram sequências gênicas sob três condições diferente: Uma contendo íntrons, outra com cDNA[2] e outra sem íntrons. O resultado da aplicação do *DNA walk* em cada uma dessas sequências pode ser vista na Fig.

[2] DNA complementar (cDNA) é o DNA sintetizado a partir de uma molécula de RNA mensageiro, cujos íntrons já foram removidos

4.1. A primeira sequência (contendo íntrons) apresentava um contorno muito irregular sugerindo correlações de longo alcance (Veja Fig. 4.1(a)). Para confirmar essa hipótese é preciso obter o coeficiente α. Para isso, esboçou-se o gráfico log × log de $F(l) \times l$. Como $F(l)$ é uma função do tipo lei de potência do comprimento l (na forma da Eq. (4.4)) então o gráfico $log \times log$ será uma reta e o expoente α pode ser interpretado como o coeficiente angular de reta no gráfico. Utilizando esse procedimento os autores obtiveram que $\alpha = 0.67 \pm 0.01$.

As outras duas sequências que são livres de íntrons, apresentavam um contorno bem regular sugerindo correlação de curto alcance. Em quase todas as sequências livres de íntrons estudadas,as regiões ricas de purinas (comparado com a concentração média ao longo de toda a fita) eram alternas por regiões ricas em pirimidinas. Essa alternância entre as regiões explica as porções de 'sobe' e 'desce' do *DNA walk* e que podem ser observadas em 4.1 (a) e 4.1(b).

Levando em consideração o fato de que concentrações de purinas e pirimidinas não são constantes ao longo da fita de nucleotídeos, os autores particionaram a cadeia pesada de miosina β-cardíaca em segmentos demarcados por um deslocamento máximo e mínimo e analisaram a flutuação em cada um deles. A Fig. 4.2(a) mostra $F(l) \times l$ para essa fita e o correspondente cDNA cujo coeficiente encontrado foi $\alpha = 0.49 \pm 0.01$. Para garantir que não introduziram nenhum viés nessa porção de máximo-mínimo, os autores utilizaram o mesmo procedimento de partição para analisar todo o gene e observaram que não houve mudança no valor de α.

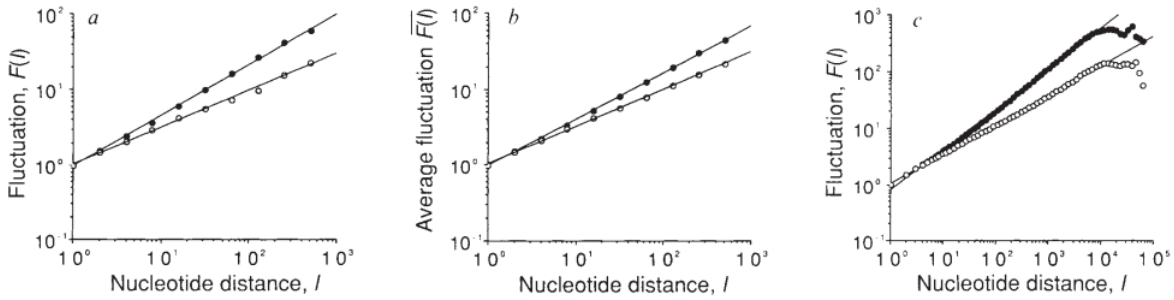

Figura 4.2: Gráfico log × log para: (a) flutuação $F(l)$ como função de l para cadeia pesada da miosina β-cardíaca humana (•, $\alpha \approx 0.67$) e o respectivo cDNA (∘, $\alpha \approx 0.49$). (b) A média de $F(l)$ sobre os grupos A (•, $\alpha \approx 0.62$) e grupo B (∘, $\alpha \approx 0.49$). (c) uma correlação ainda maior para a região cromossômica da β-globina humana contendo íntron contendo 73376 nucleotídeos (•, $\alpha \approx 0.71$) e uma cadeia de Markov padrão (∘, $\alpha \approx 0.52$) para a mesma sequência de nucleotídeos exibindo o expoente $\alpha = 1/2$ [12].

Para verificar se o comportamento de escalonamento é universal, os autores analisaram uma variedade de sequências de genomas e de cDNA. Para isso consideraram dois grupos: o primeiro (grupo A) contendo genes sem íntrons e elementos regulatórios genômicos não transcritos e o segundo (grupo B) composto de cDNA e sequências sem íntrons (Veja a Fig. 4.2(b)). A conclusão deles foi que correlações de longo alcance, onde $\alpha > 1/2$, são uma característica de genes que contém íntrons e de elementos regulatórios genômicos não transcritos. Obteve-se o valor médio de $\alpha = 0.61 \pm 0,03$ para o grupo A. Por outro lado, para sequências de cDNA e sem íntrons (grupo B), obteve-se $\alpha = 1/2$ indicando que não há correlação de longo alcance.

Outro resultado observado pelos autores é a quebra da linearidade que ocorre nos gráficos 4.2(a) e 4.2(b) e que ocorre partir de 1000 devido ao aumento do erro estatístico. De fato, a partir de $l = 1000$ ocorre uma 'queda' dos dados com relação a reta da Fig. 4.2(c). Esse comportamento é típico de toda análise fractal. Na Fig. 4.2(c), a título de comparação uma cadeia de Markov padrão é exibida e corrobora com o resultado de $\alpha = 1/2$.

O trabalho *DNA walk* é importante pois mostrou que cálculo de $F(l)$ nos fornece um método quantitativo para distinguir, apenas por meio de suas propriedades estatísticas, genes com íntrons de sequências com poucos íntrons. Em resumo, a relevância deste artigo está na introdução do método que exibe correlação em sequências de nucleotídeos e na definição de uma medida quatitativa do grau de correlação derivada da representação do *DNA walk*. Os autores concluíram que as sequências de nucleotídeos com concentração de íntrons apresentam correlação de longo alcance. A escala quantitativa da correlação é do tipo lei de potência e é observada em muitos fenômenos que possuem autossimilaridade.

4.2 Análise de DNA humano por meio de estatísticas de lei de potência

Um trabalho mais recente envolvendo o estudo das propriedades estatísticas do DNA foi realizado por Costa[3]. Nele, os autores utilizaram o formalismo estatístico de Kaniadakis para descrever correlações do tipo lei de potência nas sequências de DNA. Esse formalismo faz parte de um conjunto de generalizações da estatística de Maxwell-Boltzann e recebem o nome de entropias ou estatísticas generalizadas e são, inclusive, utilizadas para modelar diversos tipos de sistemas físicos, até mesmo os ditos sistemas complexos.

[3] Está seção é baseado no artigo [10]

A entropia de Kaniadakis depende de um parâmetro κ e fornece uma distribuição de lei de potência ao invés de exponencial.

No artigo, o modelo foi construído assumindo que em um dado volume do espaço V, cada proteína tem um comprimento pertencente ao intervalo $[\vec{l}, \vec{l} + \vec{dl}]$. Assume-se também que a distribuição de probabilidade de um comprimento l_i pode estar dentro de um intervalo $[l_i, l_i + dl_i], i = x, y, z$ podendo ser decomposta em suas componentes cartesianas e as distribuições das componentes são independentes entre si. Isso permite escrever a distribuição dos comprimento como

$$F(l)d^3l = f(l_x)f(l_y)f(l_z)dl_x dl_y dl_z, \tag{4.5}$$

com $\vec{l} = l_x\hat{i} + l_y\hat{j} + l_z\hat{z}$ e $l = \sqrt{l_x^2 + l_y^2 + l_z^2}$. Introduzindo as funções exponencial $\exp_\kappa$ e logaritmo $\ln_\kappa$ generalizadas no formalismo de Kaniadakis

$$\begin{aligned} \exp_\kappa(x) &= [\sqrt{1 + (\kappa x)^2} + \kappa x]^{\frac{1}{\kappa}} \\ \log_\kappa &= \frac{x^\kappa - x^{-\kappa}}{2\kappa}, \end{aligned} \tag{4.6}$$

onde κ é o parâmetro não aditivo. Aqui as expressões conhecidas para exponencial e logaritmo são recuperados quando $\kappa \longrightarrow 0$. Com essas funções generalizadas é possível reescrever (4.5) como

$$F(l)d^3l = \exp_\kappa\{\ln_\kappa[f(l_x)] + \ln_\kappa[f(l_y)] + \ln_\kappa[f(l_z)]\}dl_x dl_y dl_z. \tag{4.7}$$

Para determinar as funções de distribuição F e f, se aplica o logaritmo generalizado em (4.7) e deriva-se com respeito a l_i

$$\frac{\partial \ln_\kappa[F(l)]}{\partial l_i} = \frac{\partial \ln_\kappa[f(l_i)]}{\partial l_i} \tag{4.8}$$

que fornece

$$\frac{\partial \ln_\kappa[F(l)]}{\partial F(l)}\frac{dF(l)}{dl}\frac{1}{l} = \frac{1}{l_i}\frac{\partial \ln_\kappa[f(l_i)]}{\partial l_i}. \tag{4.9}$$

O lado esquerdo da Eq. (4.9) é uma constante independente do índice usado. De forma que pode-se escrever

$$\Phi_\kappa(l) = \frac{1}{l}\frac{\partial \ln_\kappa[F(l)]}{\partial F(l)}\frac{\partial F(l)}{\partial l}, \tag{4.10}$$

$$\Phi_\kappa(l) = \frac{1}{l_i}\frac{\partial \ln_\kappa[f(l_i)]}{\partial l_i}. \tag{4.11}$$

A Eq. (4.9) é satisfeita se todos os componentes são iguais a uma constante $\Phi_\kappa(l)$. Escolhe-se $\Phi_\kappa(l) = -\frac{2}{\sigma_k^2}$. O sinal negativo é usado para garantir a normalização da função

e o fator $\frac{2}{\sigma_\kappa^2}$ por conveniência matemática. O novo parâmetro introduzido σ_κ corresponde à largura da distribuição no formalismo de Kaniadakis. Com essa escolha para Φ_κ obtemos, por substituição em (4.11),

$$f(l_i) = \exp_\kappa\left(-\frac{l_i^2}{\sigma_\kappa^2}\right), \tag{4.12}$$

que permite obter a distribuição $F(l)$ como

$$F(l) = \exp_\kappa\left(-\frac{l^2}{\sigma_\kappa^2}\right). \tag{4.13}$$

A probabilidade de encontrar um comprimento de tamanho l dentro do intervalo $[l, l+dl]$ é:

$$F(l) = \int f(l) d^3l, \tag{4.14}$$

onde com $f(l)$ determinado, $d^3l = 4\pi l^2 \sin\theta d\theta d\phi dl$ e resolvendo a integral (4.14) obtemos como resultado para a distribuição $F_\kappa(l)$

$$F_\kappa(l) = 4\pi l^2 \exp_\kappa\left(-\frac{l^2}{\sigma_\kappa^2}\right). \tag{4.15}$$

Os autores assumem que a distribuição (4.15) pertence à mesma classe de universalidade investigada anteriormente no mesmo contexto de não aditividade. Essa afirmação é baseada no otimização universal a qual emerge entre esses diferentes sistemas complexos. Então a distribuição dos comprimentos das moléculas no formalismo de Kaniadakis é dada por

$$\phi_k(l) = l \exp_\kappa\left(-\frac{l^2}{\sigma_\kappa^2}\right). \tag{4.16}$$

A fim de testar a validade da distribuição (4.16), os autores utilizaram o conjunto de dados fornecido pelo projeto Ensembl [38]. Na análise, considerou-se apenas o parte codificante do DNA (éxons) e o tamanho de cada sequência é dado pelo número de pares de base (bp). Uma dificuldade encontrada é a presença de flutuações nas distribuições dos tamanhos e que foi contornada analisando as probabilidades acumuladas ao invés da função (4.16). Sendo assim a distribuição de probabilidade que irá descrever a ocorrência dos comprimentos é dada por

$$\phi_\kappa(x) = 1 - \left[\exp_\kappa\left(-\frac{x^2}{\sigma_\kappa^2}\right)\left(\sqrt{1+\kappa^2\frac{x^4}{\sigma_\kappa^4}} + \kappa^2\frac{x^2}{\sigma_\kappa^2}\right)\right]. \tag{4.17}$$

O primeiro passo consistiu em extrair os comprimentos l dos éxons e utilizar a função (4.16) para descrever a probabilidade cumulativa dos comprimentos da sequência. Os valores para κ e σ_κ foram determinados de tal maneira que função (4.16) tivesse o melhor

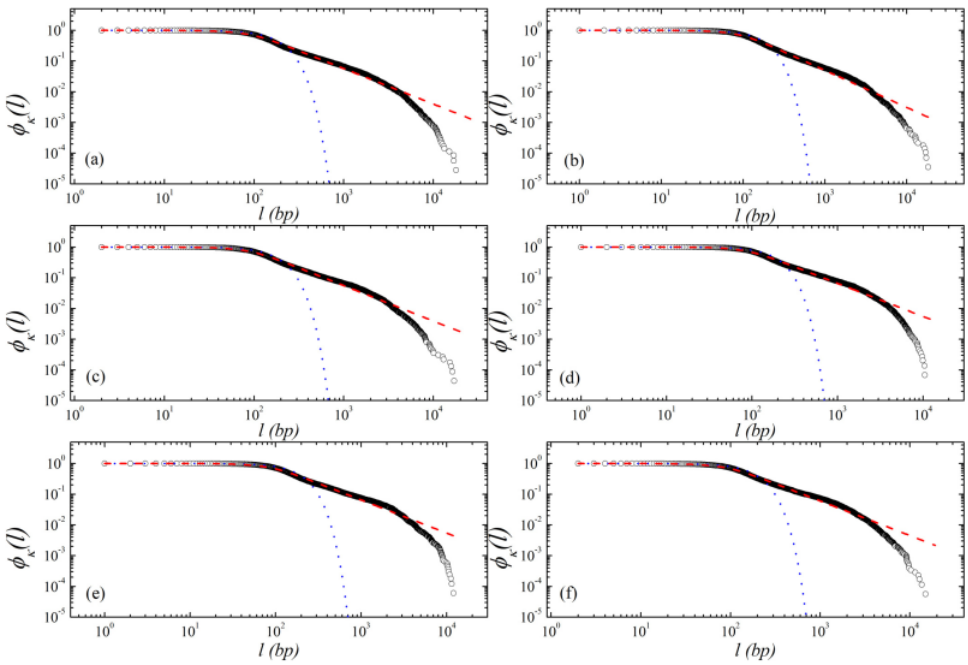

Figura 4.3: Ajuste da função de distribuição (4.16) usando o formalismo de Kaniadakis (em vermelho) e função gaussiana (em azul) para os cromossomos de 1 a 6 [10].

ajuste aos dados dos comprimentos, (Ver Fig. 4.3). A fim de comparação, a função gaussiana também foi plotada. Visualmente, o que se observou, para todos os cromossomos, é que a função (4.16) captura muito melhor o comportamento da distribuição de comprimentos enquanto que a exponencial tinha um ajuste muito ruim aos dados especialmente na região onde a curvatura da distribuição muda.

A segunda parte consistiu em calcular a função de distribuição acumulada (4.17) utilizando o algoritmo de Gauss-Newton. Por fim, estimou-se os intervalos de confiança de 90% para os melhores ajustes dos parâmetros κ e σ utilizando a amostragem *bootstrap*. No artigo, os autores exibiram uma tabela (exibida na Fig. 4.4) com os valores determinados para κ, σ, o primeiro, segundo e terceiro quartil do conjunto de dados para cada um dos 24 cromossomos. Além disso exibiram as elipses de confiança para alguns deles.

O resultado principal desse artigo foi mostrar que para a parte codificante dos cromossomos os valores de melhor ajuste para κ estão no intervalo $0.525 \leq \kappa \leq 0.659$ enquanto que o para σ tem-se $103.70 \leq \sigma \leq 114.00$. Outro resultado muito importante foi obtido pela análise das elipses de confiança, mostrando que a função exponencial simples não é capaz de capturar o comportamento da distribuição de probabilidade dos comprimentos da região codificante.

Crm[1]	$N_{bp}^{(2)}$	$Q1^{(3)}$	$Q2^{(4)}$	$Q3^{(5)}$	$\kappa^{(6)}$	$\sigma^{(7)}$	RMSE[8]	$\delta^{(9)}$
1	35504	96	144	246	$0.636^{+0.002}_{-0.002}$	$112.20^{+0.872}_{-0.782}$	0.014	0.002
2	28236	90	138	239	$0.626^{+0.002}_{-0.002}$	$109.80^{+0.716}_{-0.644}$	0.013	0.006
3	22819	92	141	244	$0.636^{+0.002}_{-0.002}$	$109.30^{+0.851}_{-0.801}$	0.014	0.008
4	14750	94	142	258	$0.656^{+0.002}_{-0.002}$	$107.20^{+0.911}_{-1.021}$	0.017	0.006
5	16632	94	142	257	$0.656^{+0.002}_{-0.002}$	$107.40^{+0.965}_{-0.867}$	0.017	0.002
6	17663	95	144	258	$0.643^{+0.002}_{-0.002}$	$111.50^{+0.902}_{-0.910}$	0.015	0.004
7	18265	92	140	244	$0.633^{+0.002}_{-0.002}$	$110.40^{+0.920}_{-0.857}$	0.015	0.009
8	13785	92	141	256	$0.649^{+0.002}_{-0.002}$	$107.40^{+0.866}_{-0.893}$	0.015	0.007
9	13584	94	141	239	$0.640^{+0.003}_{-0.003}$	$109.40^{+1.208}_{-1.032}$	0.019	0.002
10	14041	93	139	233	$0.646^{+0.003}_{-0.002}$	$105.50^{+1.032}_{-1.233}$	0.002	0.002
11	22584	94	144	268	$0.638^{+0.001}_{-0.001}$	$114.00^{+0.613}_{-0.588}$	0.011	0.009
12	21774	92	140	243	$0.634^{+0.002}_{-0.002}$	$109.80^{+0.716}_{-0.726}$	0.013	0.006
13	6275	92	137	240	$0.646^{+0.004}_{-0.004}$	$106.70^{+1.706}_{-1.501}$	0.025	0.003
14	12615	89	144	274	$0.640^{+0.001}_{-0.001}$	$111.80^{+0.583}_{-0.501}$	0.009	0.008
15	14484	91	138	234	$0.631^{+0.002}_{-0.002}$	$109.00^{+1.038}_{-0.992}$	0.018	0.002
16	18182	92	142	247	$0.631^{+0.002}_{-0.002}$	$112.20^{+0.791}_{-0.727}$	0.013	0.007
17	23728	91	141	249	$0.631^{+0.001}_{-0.001}$	$111.60^{+0.645}_{-0.679}$	0.012	0.005
18	6423	94	144	259	$0.643^{+0.002}_{-0.003}$	$111.50^{+1.191}_{-1.084}$	0.018	0.003
19	24261	87	139	263	$0.641^{+0.001}_{-0.001}$	$107.80^{+0.468}_{-0.548}$	0.009	0.005
20	8537	92	140	243	$0.635^{+0.003}_{-0.003}$	$109.80^{+1.283}_{-1.125}$	0.020	0.004
21	4019	92	144	242	$0.618^{+0.004}_{-0.004}$	$115.34^{+1.271}_{-1.399}$	0.020	0.007
22	8425	92	142	252	$0.626^{+0.002}_{-0.003}$	$113.60^{+1.035}_{-0.970}$	0.016	0.002
X	12558	92	138	249	$0.659^{+0.002}_{-0.003}$	$103.70^{+1.164}_{-1.106}$	0.019	0.009
Y	1779	89	120	190	$0.525^{+0.012}_{-0.013}$	$117.80^{+2.706}_{-2.502}$	0.035	0.000

Figura 4.4: Principais características dos cromossomos analisados. As colunas são a identificação do cromossomo (1); tamanho da amostra (2); primeiro (3), segundo (4) e terceiro (5) quartis do conjunto de dados; parâmetros de melhor ajuste e seus respectivos 95% intervalos de confiança (6 e 7); erro padrão residual e a tolerância de convergência alcançada do ajuste (8 e 9) [10].

4.3 Uma descrição alternativa de correlação de lei de potência em sequências de DNA

Por fim, vamos discutir um trabalho mais recente produzido por [39] e que seguiu os passos semelhantes aos desenvolvidos no artigo da seção anterior. Este trabalho, também influenciou na escolha da função de probabilidade dos comprimentos dos pares de base que utilizamos nesta dissertação.

Nos últimos anos a entropia de Boltzmann- Gibbs é generalizada na tentativa de investigar as propriedades de sistemas complexos. Uma dessas extensões é a entropia não-aditiva de Tsallis. Neste trabalho, utilizou-se o formalismo de Tsallis para investigar as propriedades de correlação estatística das moléculas de DNA. Mais especificamente, os autores modelaram o comportamento de correção de lei de potência (CLP) das sequências dos nucleotídeos codificante (éxons) do DNA humano. Este mecanismo é apropriadamente

descrito através da distribuição dos comprimentos dos pares de bases (pb).

Seguindo a ideia da seção anterior, podemos reescrever a distribuição dos comprimentos (4.5) como

$$F(l)d^3l = \exp_q \left\{ ln_{q*}[f(l_x)] + ln_{q*}[f(l_y)] + ln_{q*}[f(l_z)] \right\}, \tag{4.18}$$

onde $\bar{l} = l_x\hat{i} + l_y\hat{j} + l_z\hat{z}$, com $l^2 = l_x^2 + l_y^2 + l_z^2$, $q* \to 2 - q$. O índice q é chamado parâmetro de correlação da lei de potência (CLP). Tomando a diferencial parcial do q- logaritmo com respeito a cada uma das componentes (como na seção anterior) podemos mostrar que

$$f(l_i) = A_q \left[1 - (q-1)\frac{l_i^2}{\sigma_q^2} \right]^{\frac{1}{q-1}} \quad i = x, y, z \tag{4.19}$$

e

$$F(l) = B_q \left[1 - (1-q)\frac{l^2}{\sigma_q^2} \right]^{\frac{1}{q-1}}. \tag{4.20}$$

Onde $A_q = f(q)\sigma_q^{1/2}$ e $B_q = g(q)\sigma_q^{1/q}$ são fatores de normalização. Além disso, $f(q)$ e $g(q)$ são funções de q e σ_q é a largura das distribuições (4.19) e (4.20). A exponencial deformada e sua inversa é definida como:

$$\exp_q(x) = [1 + (1-q)x]^{\frac{1}{1-q}} \tag{4.21}$$

$$\ln_q(x) = \frac{x^{1-q} - 1}{1-q} (\forall q, x > 0). \tag{4.22}$$

Observe que as Eqs. (4.19) e (4.20) podem ser escritas em termos da função exponencial deformada. No limite em que $q = 1$, a função exponencial é recuperada de maneira que:

$$f(l_i) = A \exp\left(\frac{-l_i^2}{\sigma} \right), i = x, y, z \tag{4.23}$$

$$F(l) = A^3 \exp\left(\frac{l^2}{\sigma} \right). \tag{4.24}$$

Como as distribuições (4.19) e (4.20) apresentam o mesmo comportamento matemático, podemos construir uma função que descreve a distribuição dos comprimentos dos nucleotídeos:

$$\phi_q(l) = B_q l \left[1 - (1-q)\frac{l^2}{\sigma_q^2} \right]^{\frac{1}{1-q}} = B_q l \exp_q\left(-\frac{l^2}{\sigma_q^2} \right). \tag{4.25}$$

Para testar a viabilidade da distribuição de lei de potência proposta, os autores utilizaram os pares de bases codificantes disponíveis em bancos de dados públicos como o Ensembl [38].

Um problema encontrado na análise dos comprimentos dos cromossomos é a presença de flutuações na distribuição dos tamanhos das sequências de DNA. Para contornar

este problema, os autores analisaram a distribuição acumulada, obtida pela normalização $\phi_q(x > l) + \phi_q(x < l) = 1$. De fato, a abordagem usou a distribuição cumulativa $\phi_q(x > l)$ para todos os cromossomos. Visualmente se observou que a função proposta (4.25) é capaz de se ajustar melhor ao conjunto de dados dos comprimentos, quando comparado com a função de distribuição exponencial (Ver Fig. 4.5). Claramente, os resultados mostram que a distribuição tipo lei de potência tem um melhor ajuste aos dados, comparado a exponencial. Na Eq. (4.25), o parâmetro q ajusta a função $\phi_q(l)$ ao conjunto de dados dos comprimentos dos nucleotídeos. A Fig. 4.5 mostra o comportamento de $\phi_q(l)(x > l)$ em função do comprimento l, os dados associados aos comprimentos das cadeias para cada cromossomo e a função de distribuição exponencial.

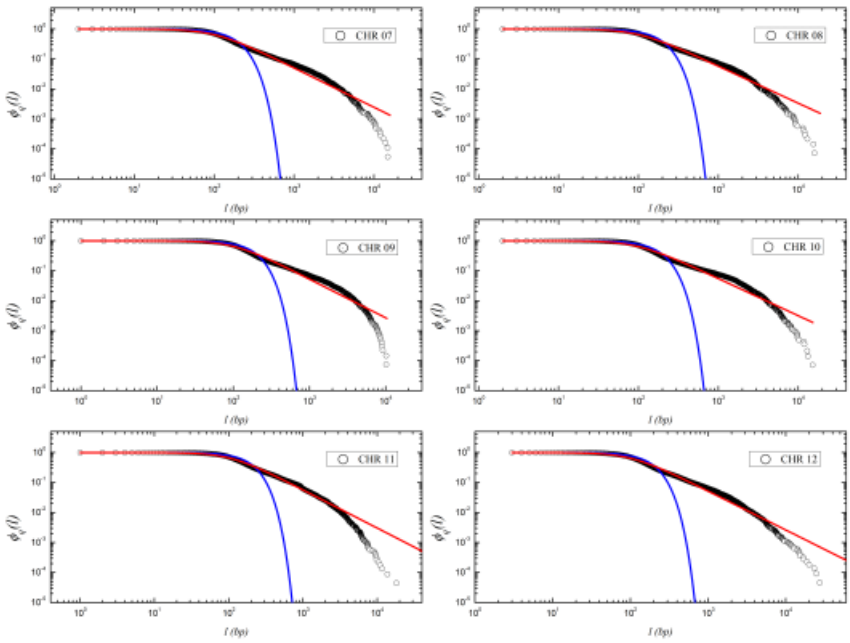

Figura 4.5: Função de distribuição acumulada para os comprimentos dos cromossomos 7 a 12. Observe que a distribuição de lei de potência (em vermelho) se ajusta melhor ao conjunto de dados do que a distribuição exponencial (em azul), especialmente na região em que os comprimentos são grandes [39].

A fim de testar a viabilidade estatística do modelo representado pela distribuição de lei de potência (4.25), o ajuste produzido foi confrontado com outra distribuição, definida a partir da uma soma de duas exponenciais, da forma

$$D = \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2} \left[-\frac{1}{\lambda_1} \exp(-\lambda_1 l) - \frac{1}{\lambda_2} \exp(-\lambda_2 l) \right] + 1 \tag{4.26}$$

onde λ_1 e λ_2 são parâmetros ajustáveis e l é o comprimentos dos nucleotídeos. A comparação entre as duas distribuições (4.25) e (4.26) é mostrada na Fig. 4.1.

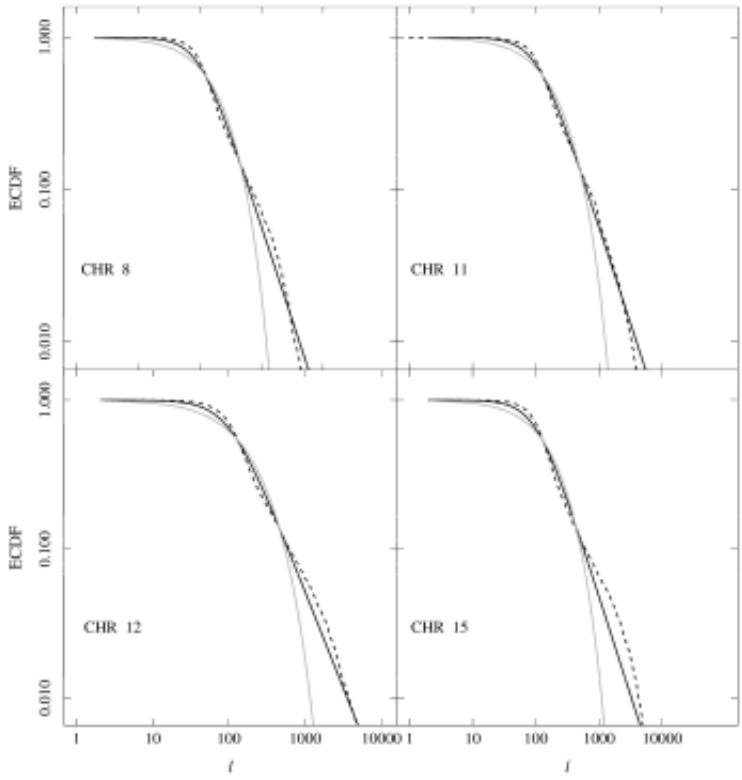

Figura 4.6: Função de distribuição empírica acumulada (ECDF) *versus* l, para quatro cromossomos escolhidos aleatoriamente. As linhas pontilhadas representam os dados, a linha preta indica a distribuição de lei de potência e a linha cinza representa a soma de exponenciais [39].

A ferramenta estatística chamada inferência Bayesiana é usada para comparar modelos. Tecnicamente, a análise utiliza o modelo teórico, o conjunto de dados e uma informação (prior). Considerando esta análise, num problema da estimativa de parâmetro, o ponto inicial para análise Bayesiana é calcular a probabilidade posterior de um conjunto

$$P(\Theta|D,M) = \frac{L(D|\Theta,M)\pi(\Theta|M)}{\xi(D|M)}, \tag{4.27}$$

onde $P(\Theta|D,M)$ é a distribuição posterior, $L(D|\Theta,M)$ é a *likelihood*, $\pi(\Theta|M)$ é a distribuição prior e $\xi(D|M)$ é a evidência Bayesiana.

A comparação entre os modelos foi realizada utilizando o fator de Bayes, definido por

$$B_{ij} = \frac{\xi_i}{\xi_j}, \tag{4.28}$$

onde ξ_j é a evidência do modelo base, que para nosso caso é a distribuição de lei de potência, e ξ_i é a evidência do modelo que queremos comparar (no caso, a dupla exponencial). Para quantificar qual dos modelos é mais favorável para descrever um conjunto de dados,

$\|lnB_{ij}\|$	Interpretação
< 1	Inconclusivo
1	Fraco
$2,5$	Moderado
5	Forte

Tabela 4.1: Escala de Jeffrey. A partir do valor do logaritmo do fator de Bayes podemos dizer se modelo baseado na entropia de Tsallis é inconclusivo, fraco, moderado ou forte com relação ao modelo de comparação de dupla exponencial.

utilizamos a interpretação do fator de Bayes dado pela escala de Jeffrey.

Os resultados relevantes desse artigo são que o valor do parâmetro de correlação de lei de potência q estava no intervalo $[1.44; 1.61]$. Além disso, utilizando a escala de Jeffrey, os resultados da análise Bayesiana mostraram que a dupla exponencial é desfavorável em relação ao modelo baseado na distribuição tipo lei de potência. Portanto, o modelo da distribuição tipo lei de potência, tem um melhor ajuste aos dados dos comprimentos, quando comparado ao modelo da distribuição de dupla exponencial.

5. Distribuição de tamanhos de DNA: Modelo entrópico

Desde a descoberta de Francis Crick e James Watson, inúmeras pesquisas foram realizadas a fim de estudar o genoma humano em seus mais diversos aspectos. As sequências genômicas são um exemplo dos chamados sistemas complexos. Isso significa, entre outras coisas, que partes dos sistema interagem entre si de maneira não linear [40]. Diversos modelos baseados principalmente em física estatística foram empregados para tratar a crescente quantidade de dados produzidos à medida que mais partes do genoma eram sequenciadas [10, 41, 42]. Em especial, o processo caminhada aleatória [12] de DNA foi utilizado para capturar a interação de ocorre na sequência.

A possibilidade de eventos associados à correlação entre as sequências do DNA de muitos organismos é estudado por diversos grupos. Com o intuito do reunir e fornecer uma fonte centralizada de informação para esses pesquisadores, o European Bioinformatics Institute [43] juntamente com o Wellcome Trust Sanger Institute [44] lançaram o projeto Ensembl como um conjunto de dados completo acerca da sequência do genoma humano [45]. Os dados para análise desta dissertação foram retirados do banco de dados do projeto Ensembl [38].

Nesta dissertação analisamos a distribuição de probabilidade acumulada dos comprimentos dos pares de base das sequências de DNA humano. A função de distribuição $p(l)$ é obtida a partir da otimização da entropia de teoria da informação e o comprimento da sequência, dado por l, consiste do número de pares de base (pb) na sequência. Para testar a validade do modelo, utilizamos os dados do projeto Ensembl [38], de onde extraímos as sequências codificantes do DNA. Em seguida, determinamos os valores dos parâmetros livres de $p(l)$ que melhor se ajustam à curva dos dados de comprimento. Além disso, para dar maior robustez à análise estatística construímos a elipse de confiança para os parâmetros livres de $p(l)$. A linguagem computacional utilizada para análise foi R [46].

5.1 Função de distribuição de probabilidades

Lembremos que a entropia é dada por (3.23) e que para o caso contínuo pode ser escrita como

$$S = -\int_0^\infty p(l)\log\left[p(l)\right]dl. \tag{5.1}$$

Quando extremizada, a função acima deve estar sujeita aos seguintes vínculos:

$$\int_0^\infty p(l)dl = 1, \tag{5.2}$$

$$\int_0^\infty p(l)ldl = L, \tag{5.3}$$

onde L é o valor do comprimento médio de uma cadeia do DNA.

A fim de obter a função de distribuição de probabilidades mais provável faremos uso de uma ferramenta matemática própria pra determinar máximos e mínimos de um funcional[1]: O cálculo variacional[2]. A entropia sujeita aos vínculos (5.2) e (5.3) pode ser escrita como

$$S = -\int_0^\infty [p(l)]\log p(l)dl - \lambda_0\int_0^\infty p(l)dl - \lambda\int_0^\infty p(l)ldl, \tag{5.4}$$

onde os coeficientes λ e λ_0 são os multiplicadores de Lagrange. O problema de extremização consiste em resolver $\delta S = 0$

$$\delta S = \delta\left[-\int_0^\infty p(l)\log\left[p(l)\right]dl - \lambda_0\int_0^\infty p(l)dl - \lambda\int_0^\infty p(l)ldl\right] = 0.$$

Em cálculo das variações, o símbolo δ denota a variação de um funcional que possui dependência de uma função e de sua derivada. Uma vez que a entropia de informação S possui apenas dependência da função de probabilidades p e não da derivada de p, a maximização de S corresponderá a uma derivada com relação a p

$$\begin{aligned}\delta S = \frac{\partial}{\partial p}\left[-\int_0^\infty p(l)\log\left[p(l)\right]dl - \lambda_0\int_0^\infty p(l)dl - \lambda\int_0^\infty p(l)ldl\right]\delta p = 0.\\ = -\frac{\partial}{\partial p}\int_0^\infty \left[p(l)\log\left[p(l)\right] + \lambda_0 p(l) + \lambda l p(l)\right]dl\delta p = 0.\\ = \int_0^\infty \left[-\log\left[p(l)\right] - 1 - \lambda_0 - \lambda l\right]\delta p dl = 0.\end{aligned}$$

Para satisfazer essa equação o integrando deve ser igual a zero

$$-\log\left[p(l)\right] - 1 - \lambda_0 - \lambda l = 0,$$

[1] Um funcional é uma função real cujo domínio é um espaço de funções

[2] Uma ótima referência sobre esse tema e que o aborda da maneira sucinta é o livro do Nivaldo Lemos, Mecânica Analítica [47].

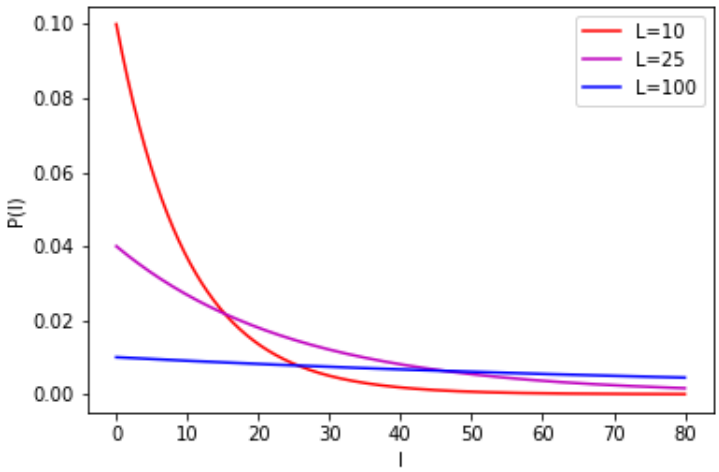

Figura 5.1: Distribuição de probabilidade $P(l)$ para diferentes valores de L. Observe que para $L = 100$, dentro do intervalo de comprimentos mostrado no gráfico, a probabilidade é melhor distribuída do que para valores menores. Por outro lado em $L = 10$, a função $P(l)$ possui maiores valores para $l < 25$ e é quase zero para todos os comprimentos maiores.

que nos fornece:

$$p(l) = \exp(-1 - \lambda_0 - \lambda l). \tag{5.5}$$

Aplicando as condições de vínculo (5.2) e (5.3) obtemos os valores de λ e λ_0

$$p(l) = \frac{1}{L} \exp(-\frac{l}{L}). \tag{5.6}$$

A Eq. (5.6) nos diz a probabilidade de encontrar uma sequência cromossômica de tamanho l. Veja que a mesma equação mostra que dado dois comprimentos, l_1 e l_2, tal que $l_1 < l_2$ então $p(l_1) > p(l_2)$. Em outras palavras, a probabilidade de ocorrência de comprimentos maiores é menor que a de comprimentos menores. Há uma relação inversa entre a probabilidade (5.6) e o comprimento l associado a ela.

Outro ponto importante é a forte dependência do parâmetro L, como observado na Fig. 5.1. Note que quanto maior for L mais lentamente a probabilidade "decai", indicando que comprimentos maiores, em relação ao valor médio, possuem probabilidade de ocorrência relevante. Isso faz sentido se pensarmos que um valor grande para L significa que há enorme variedade nos tamanhos do cromossomo e os comprimentos maiores terão probabilidade de ocorrência tão significativa quanto os menores. O valor L possui informação do quão bem distribuídos são os tamanhos dos cromossomos.

Alguns trabalhos apontam que a função exponencial não é uma boa candidata para o ajuste desses dados em relação a outras funções de distribuição [10]. Nossa proposta

consiste em utilizar uma soma de exponenciais da forma

$$P(l) = A\left[\exp(-k_2 l) + \exp(-k_1 l)\right], \tag{5.7}$$

onde k_1 e k_2 são constantes ajustáveis a um conjunto de dados e A é uma constante de normalização que pode ser determinada pela condição $\int_0^\infty P(l)dl = 1$, tal que

$$P(l) = \frac{k_1 k_2}{k_1 + k_2}\left[\exp(-k_2 l) + \exp(-k_1 l)\right]. \tag{5.8}$$

A função (5.8) conserva uma característica importante: Os menores comprimentos possuem maior probabilidade de ocorrência, isto é dado $l_1 < l_2$ então $P(l_1) > P(l_2)$. Aqui, k_1 e k_2 carregam a informação do quão bem distribuídos são os tamanhos das sequências do DNA, de modo que a função de distribuição proposta (5.8) depende de dois parâmetros ajustáveis e que serão relevantes para obter $P(l)$. A fim de ilustrar essa declaração, imagine o caso extremo em que um dos parâmetros é muito maior que o outro, digamos que $k_1 << k_2$. Mesmo nessa condição, a exponencial dependente de k_1 não pode ser desprezada.

A título de comparação vamos analisar as exponenciais das Eqs. (5.6) e (5.8). Na primeira, o argumento da exponencial é inverso ao valor médio L. No segundo caso, o argumento da dupla exponencial é proporcional aos parâmetros de ajuste k_1 e k_2. Da discussão a respeito da informação que L carrega acerca da distribuição de probabilidades, somos levados a crer que quanto menores forem os valores dos parâmetros k_1 e k_2 mais provável é a ocorrência de comprimentos menores das sequências, enquanto que se k_1 e k_2 assumirem valores maiores então será favorecida a ocorrência de comprimentos maiores.

A constante de normalização $A = k_1 k_2/(k_1 + k_2)$ se ajusta conforme os parâmetros, assumindo valor maior quanto maior k_1 e/ou k_2. Podemos entender que k_1 e k_2 distribuem entre si a informação sobre a distribuição dos comprimentos. Note que a exponencial simples é recuperada quando $k_1 = k_2$.

Um problema que emerge dessa análise são as flutuações das probabilidades de ocorrência dos comprimentos [48]. Uma solução é utilizar as distribuições de probabilidade acumuladas, uma vez que além de descrever completamente uma distribuição de dados, ainda oferecem melhor descrição e visualização dos dados analisados.

As distribuições de probabilidade acumulada consistem em, dado um comprimento l, determinar a probabilidade de se obter um comprimento menor ou igual a l[3]. Nessa

[3]Esse problema pode ser invertido, ou seja, dado um comprimento l, determinar a probabilidade de se obter um comprimento maior que l. Nesse caso, os limites da integração variam de $[l, \infty]$

condição, a distribuição acumulada associada à Eq. (5.8) é

$$\phi(l) = P(l < l') = \int_0^l P(l')dl' \tag{5.9}$$

$$= \int_0^l \frac{k_1 k_2}{k_1 + k_2} \left[\exp(-k_2 l') + \exp(-k_1 l')\right] dl'. \tag{5.10}$$

Que nos leva a

$$= \frac{k_1 k_2}{k_1 + k_2} \int_0^l \left[\exp(-k_2 l') + \exp(-k_1 l')\right] dl'$$

$$= \frac{k_1 k_2}{k_1 + k_2} \left[-\frac{1}{k_1}\exp(-k_1 l') - \frac{1}{k_2}\exp(-k_2 l')\right]_0^l$$

fornecendo

$$\phi(l) = -\frac{k_1 k_2}{k_1 + k_2}\left[\frac{1}{k_1}\exp(-k_1 l) + \frac{1}{k_2}\exp(-k_2 l)\right] + 1. \tag{5.11}$$

5.2 Ajuste estatístico e resultados

Os dados utilizados para testar a viabilidade do modelo foram extraídos do projeto Ensembl [38]. A partir dos arquivos disponíveis, extraímos os comprimentos l das regiões codificante (éxon) de todos os cromossomos e, em seguida, calculamos as funções de distribuição acumulada correspondentes a cada cromossomo através da linguagem de programação R. Armazenamos os dados dos comprimentos e as respectivas funções acumuladas em arquivos de texto. O conjunto dos comprimentos dos pares de base e das probabilidades acumuladas obtidas pelo programa, formam nossa base de dados e os detalhes de como esse procedimento foi realizado podem ser vistos no apêndice **??**.

Logo após, fizemos o ajuste das constante k_1 e k_2, da função (5.11), ao conjunto de dados dos comprimentos e da distribuição acumulada que foram descritos no parágrafo anterior. Os valores para o ajuste dos parâmetros k_1 e k_2 que foram obtidos são exibidos na Tab. 5.1 para todos os cromossomos.

Nos gráficos 5.2, 5.4, 5.6, e 5.8, esboçamos o conjunto dos dados juntamente com a função de distribuição acumulada (5.11) para os cromossomos $Y, X, 1$ e 15, utilizando para k_1 e k_2 os valores que melhor se ajustam ao conjunto de dados (e que se encontram na Tab. 5.1). Observamos que, visualmente, a função (5.11) consegue se ajustar aos dados.

Os valores numéricos dos ajustes tiveram uma variação de $0,005668 \leq k_1 \leq 0,008624$ e $0,002776 \leq k_2 \leq 0,004209$. Como foi discutido na seção anterior, valores pequenos para os parâmetros de ajuste indicam que há maior probabilidade de ocorrência de comprimentos

menores. Sendo assim, para quantidades tão pequenas ($\sim 10^{-3}$) esperamos uma concentração massiva dos comprimentos menores. Esse comportamento pode ser observado nas figuras abaixo em que a probabilidade dos comprimentos menores é muito maior e cresce de forma quase imperceptível para o restantes dos comprimentos. Nesse sentido, para os dados codificantes da sequência do genoma humano as correlações estatísticas de curto alcance são capturadas pelo modelo de dupla exponencial. Mesmo apresentando apenas alguns cromossomos (Ver Figs. 5.2-5.9) ressaltamos que esse comportamento ocorreu para todos os cromossomos, incluindo os cromossomos X e Y.

Possivelmente a concentração massiva de pequenos comprimentos nas regiões codificantes esteja no fato de sequências longas não sejam necessárias para produção de proteínas uma vez que apenas uma trinca (códon) é suficiente pra construir um aminoácido. Outro ponto que vale destacar é que os cromossomos sexuais X e Y apresentaram valores extremos para os parâmetros de ajuste. O menor valor para k_1 $(0,005668)$ foi obtido pelo cromossomo Y enquanto que o maior foi obtido por X $(0,008624)$. Para o k_2, obtivemos outro resultado: O menor valor de ajuste foi de X $(\ 0,002776)$ e o maior foi obtido pelo cromossomo Y $(0,004209)$.

Outra análise realizada sobre k_1 e k_2 foi através das elipses de confiança [49]. As elipses são um conjunto de pontos em um espaço bidimensional, geralmente representado por uma elipse centrada em torno dos melhores valores para o ajuste de k_1 e k_2 [50].

As elipses de confiança (ou regiões de confiança) são ferramentas estatísticas que permitem interpretar uma grande quantidade de dados de forma visual e simples por meio de gráficos. Essas regiões são construídas com o objetivo de estimar o verdadeiro valor de um parâmetro de interesse. Geralmente esses intervalos estão associados a um nível de confiança. Esse nível de confiança nos diz a taxa de sucesso de um procedimento usado para construir o intervalo de confiança [51]. Contextualizando para o problema em obter os valores de ajuste: Dado que obtivemos os valores das constantes k_1 e k_2 com 95% de confiança, isso significa que se realizarmos novamente o ajuste da função (5.11) ao conjunto de dados, temos a probabilidade de 95% de que o novo valor obtido esteja dentro do intervalo que é observado na Tab. 5.1.

Para construir as elipses de confiança utilizamos o **método de *bootstrap*** na linguagem R. Para o conjunto de comprimentos l realizamos o ajuste da função (5.11) e obtivemos os valores de k_1 e k_2. Escolhemos aleatoriamente um dos comprimentos l e o eliminamos; em seguida sorteamos outro comprimento e o duplicamos para manter o conjunto de dados com a mesma quantidade de elementos. A partir desse novo conjunto calculamos novamente os valores de ajuste para k_1 e k_2. Repetimos esse procedimento

1000 vezes para obter 1000 pares k_1, k_2. O conjunto dos 1000 pares foram esboçados em um gráfico ($k_1 \times k_2$). Desenhamos as elipses com $68\%, 95\%$ e 99% de confiança; Isso significa que se realizarmos o ajuste da função (5.11) várias vezes com um conjunto com a mesma quantidade de dados então em uma certa porcentagem das vezes ($68\%, 95\%$ e 99% respectivamente) a região de confiança irá conter os valores para k_1 e k_2. Para plotarmos os gráficos da elipse, normalizamos os parâmetros pela diferença entre o máximo e mínimo valor do conjunto. Os códigos de ajuste, *bootstrap*, marcação dos dados e da função além da construção das elipses de confiança podem ser vistos no apêndice **??**.

Analisando as elipses, observamos que os cromossomos apresentavam comportamento de correlação estatística entre os parâmetros k_1 e k_2. Na maioria deles, podemos observar que aumentando k_1 também aumentamos k_2. Alguns de forma mais dispersa como os cromossomos 1 e 15 (Figs. 5.7 e 5.9), outros com valores mais concentrados como o cromossomo X (Fig. 5.5) mas todos apresentando uma relação de proporcionalidade entre eles. A exceção foi observada no cromossomo Y. Nele, observamos um comportamento de anti-correlação estatística entre os valores de k_1 e k_2. Isso significa que os valores maiores para k_2 estão associados à valores pequenos de k_1 e vice-versa. Observe a elipse de confiança correspondente ao cromossomo Y (Ver Fig. 5.3). Em um primeiro momento, não temos nenhuma explicação para esse fenômeno. O que distingue o cromossomo Y dos demais é que o mesmo é o responsável pela determinação do sexo nos seres humanos, sendo necessária maior investigação para determinar se os parâmetros de ajuste conseguem capturar alguma característica biológica das sequências dos cromossomos.

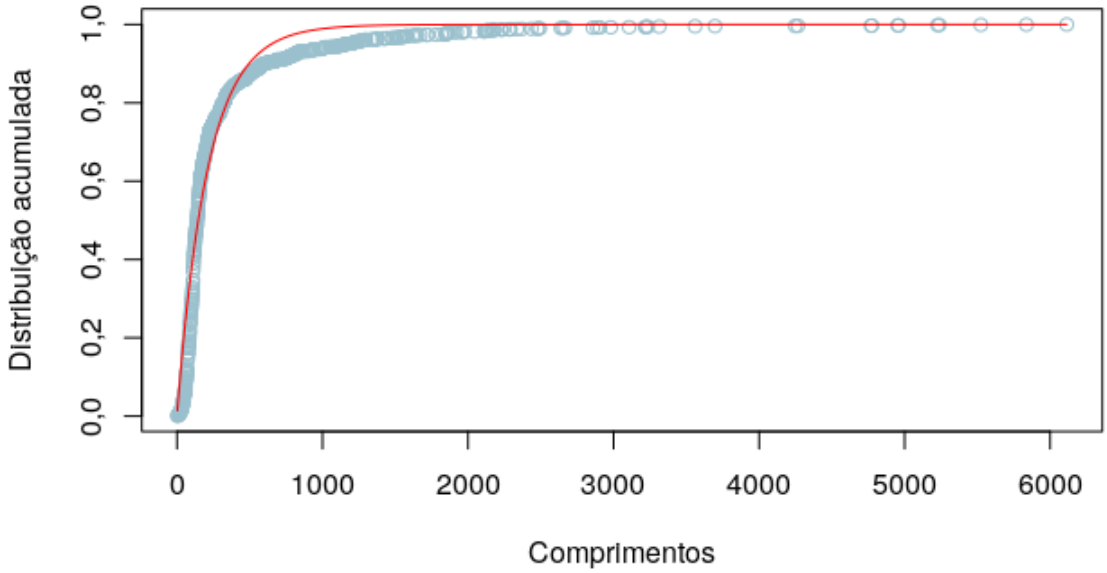

Figura 5.2: Função de distribuição acumulada versus comprimentos ($\phi(l) \times l$) dos pares de base para o cromossomo Y. Em azul, temos os dados e, em vermelho, a função (5.11), com os valores para k_1 e k_2 mostrados na Tab. 5.1.

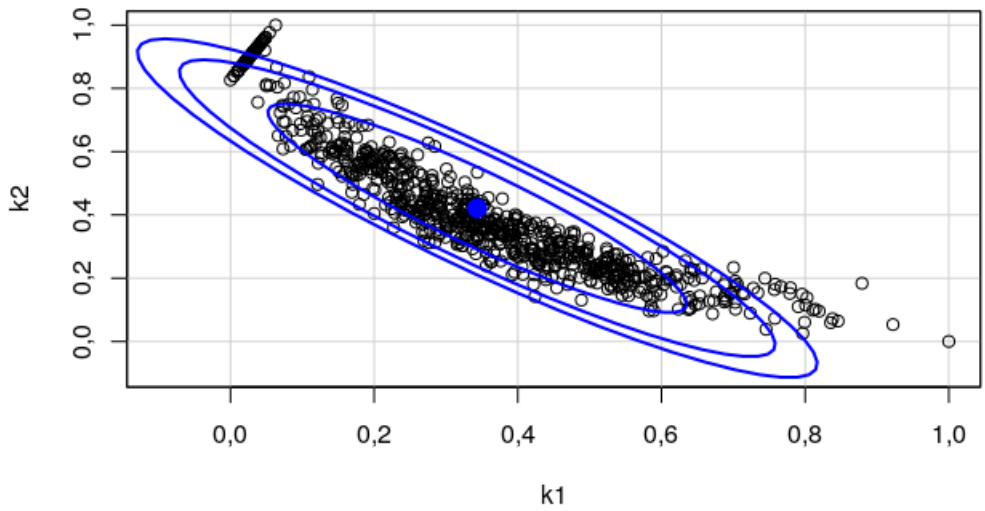

Figura 5.3: Elipses de confiança para as constantes $k1 \times k_2$, obtidas para o cromossomo Y, com intervalo de confiança de $68\%, 95\%$ e 99%. A Elipse de confiança indica um comportamento de anti-correlação estatística entre k_1 e k_2, ou seja, quando aumentamos k_1 os valores para k_2 diminuem e vice-versa.

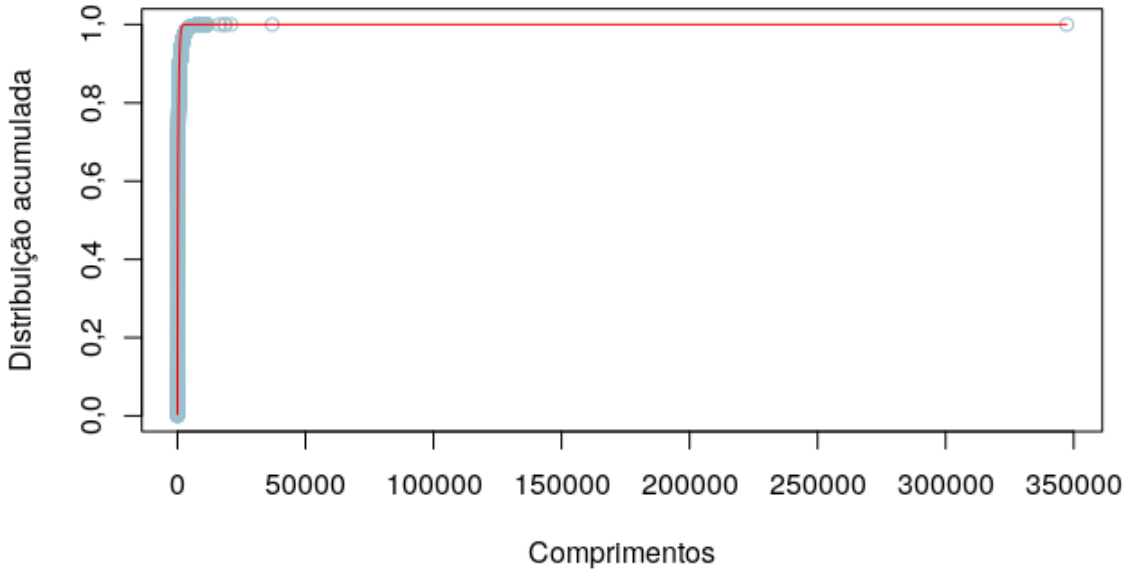

Figura 5.4: Função de distribuição acumulada versus comprimentos ($\phi(l) \times l$) dos pares de base para o cromossomo X. Em azul, temos os dados e, em vermelho, a função (5.11), com os valores para k_1 e k_2 mostrados na Tab. 5.1.

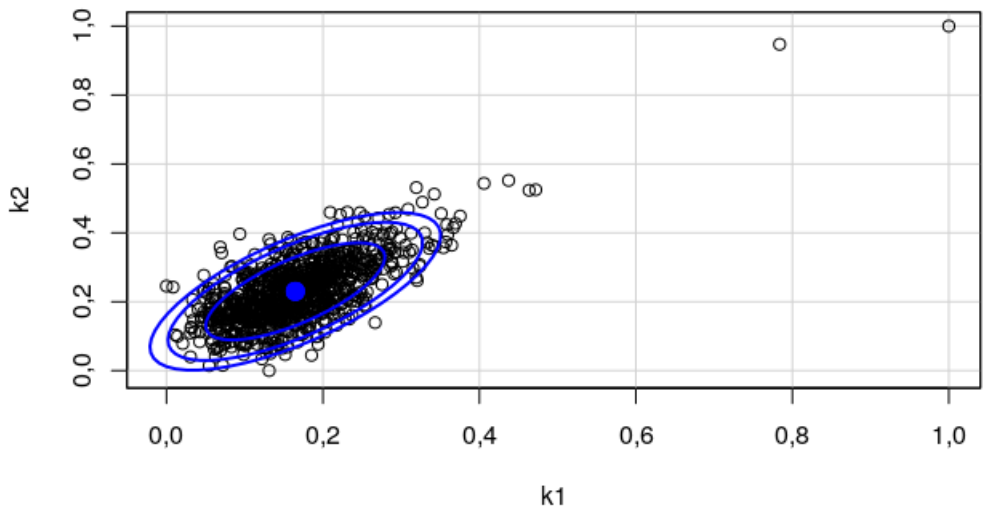

Figura 5.5: Elipses de confiança para as constantes $k1 \times k_2$, obtidas para o cromossomo X, com intervalo de confiança de $68\%, 95\%$ e 99%. O cromossomo X apresenta uma grande concentração dos parâmetros em uma região do gráfico e um comportamento linear entre k_1 e k_2 muito nítido entre eles.

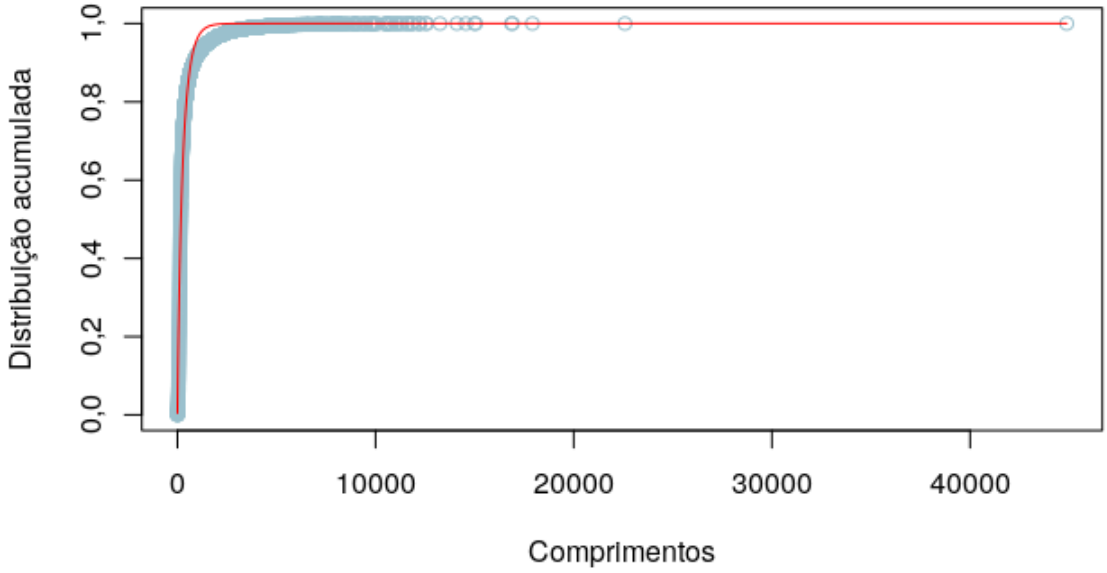

Figura 5.6: Função de distribuição acumulada versus comprimentos ($\phi(l) \times l$) dos pares de base para o cromossomo 1. Em azul, temos os dados e, em vermelho, a função (5.11), com os valores para k_1 e k_2 mostrados na Tab. 5.1.

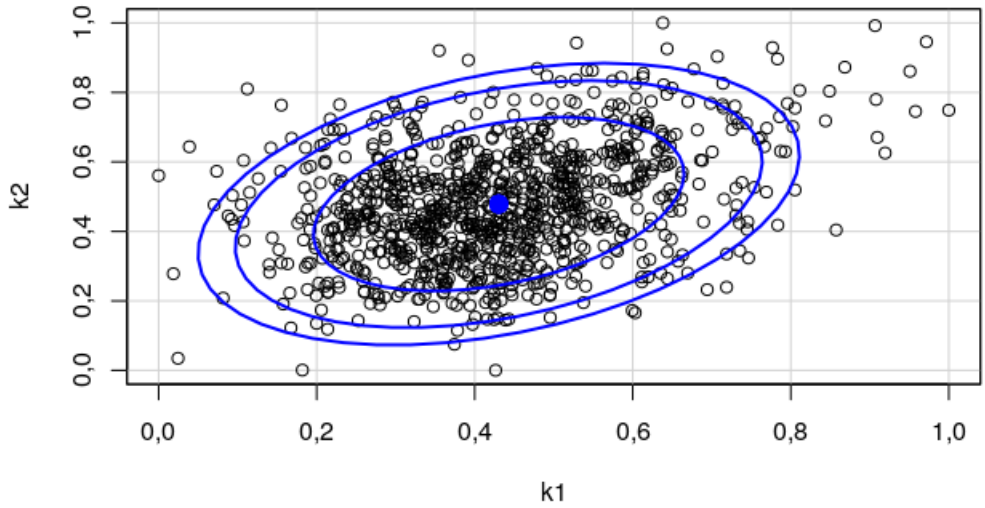

Figura 5.7: Elipses de confiança para as constantes $k1 \times k_2$, obtidas para o cromossomo 1, com intervalo de confiança de $68\%, 95\%$ e 99%. Observe que os pontos são dispersos mas há um crescimento de k_1 quando k_2 cresce.

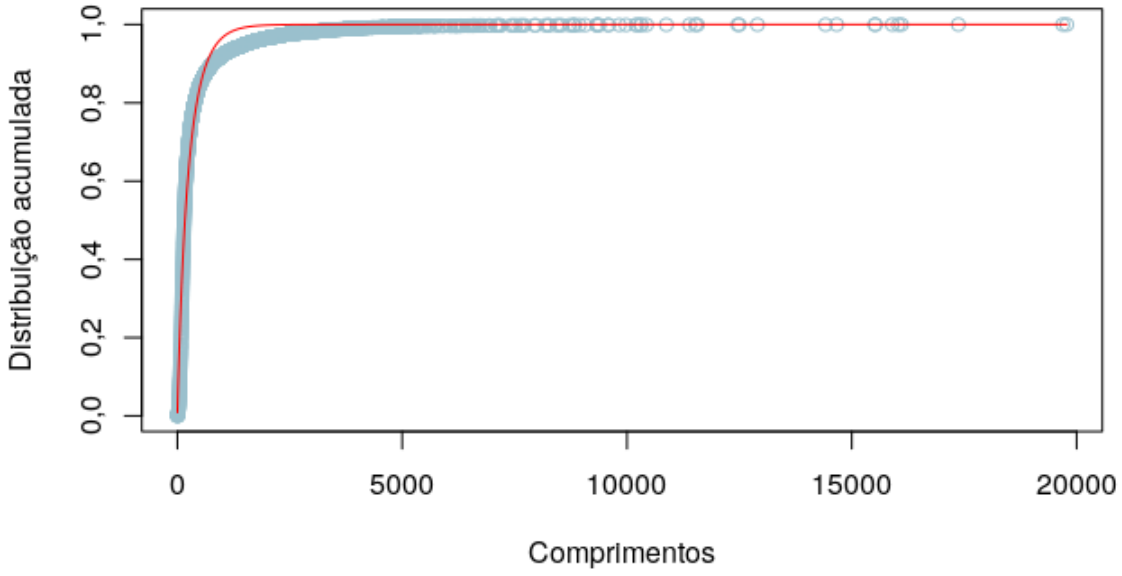

Figura 5.8: Função de distribuição acumulada versus comprimentos ($\phi(l) \times l$) dos pares de base para o cromossomo 15. Em azul, temos os dados e, em vermelho, a função (5.11), com os valores para k_1 e k_2 mostrados na Tab. 5.1.

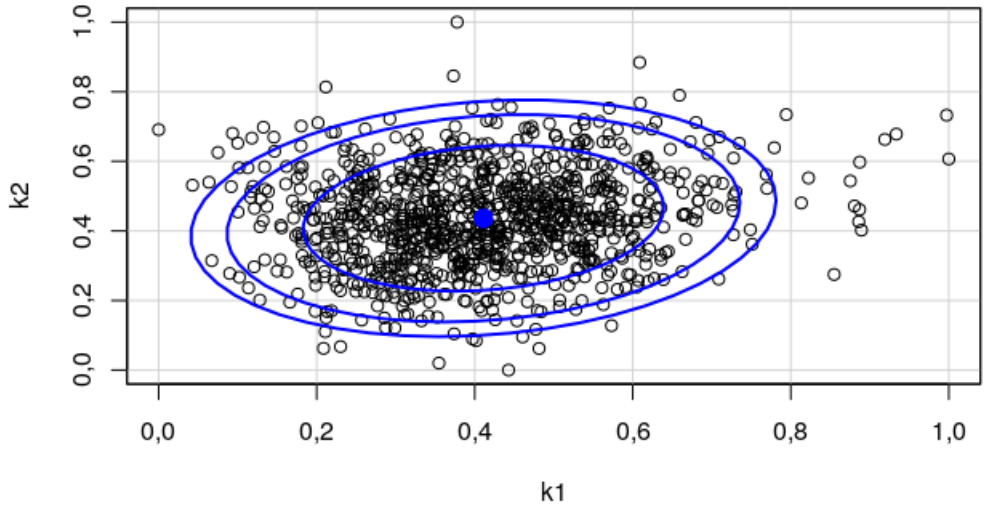

Figura 5.9: Elipses de confiança para as constantes $k1 \times k_2$, obtidas para o cromossomo 15, com intervalo de confiança de $68\%, 95\%$ e 99%.

Crm	N	Q1	Q2	Q3	k_1	k_2
01	3419	855,5	1750	3019,5	$0,00799^{+4,1850\times10^{-4}}_{-4,1850\times10^{-4}}$	$0,002938^{+1,1760\times10^{-5}}_{-1,1760\times10^{-5}}$
02	3096	774,8	1601	2818,5	$0,008155^{+4,015\times10^{-4}}_{-4,015\times10^{-4}}$	$0,003005^{+1,267\times10^{-5}}_{-1,267\times10^{-5}}$
03	2872	718,8	1483	2662.8	$0,008068^{+4,449\times10^{-4}}_{-4,449\times10^{-4}}$	$0,002996^{+1,288\times10^{-5}}_{-1,288\times10^{-5}}$
04	2525	632	1353	2498	$0,008158^{+5,332\times10^{-4}}_{-5,332\times10^{-4}}$	$0,002870^{+1,481\times10^{-5}}_{-1,481\times10^{-5}}$
05	2755	689,5	1472	2625,5	$0,008104^{+4,751\times10^{-4}}_{-4,751\times10^{-4}}$	$0,002813^{+1,364\times10^{-5}}_{-1,364\times10^{-5}}$
06	2709	678	1414	2519	$0,008092^{+5,0265\times10^{-4}}_{-5,0265\times10^{-4}}$	$0,002814^{+1,399\times10^{-4}}_{-1,399\times10^{-4}}$
07	2574	644,2	1338,5	2480,8	$0,008001^{+4,545\times10^{-4}}_{-4,545\times10^{-4}}$	$0,002958^{+1,299\times10^{-4}}_{-1,299\times10^{-4}}$
08	2418	606,2	1260,5	2304,8	$0,008156^{+4,892\times10^{-4}}_{-4,545\times10^{-4}}$	$0,002871^{+1,350\times10^{-5}}_{-1,350\times10^{-5}}$
09	2339	585,5	1235	2325	$0,008043^{+4,944\times10^{-4}}_{-4,944\times10^{-4}}$	$0,002921^{+1,405\times10^{-4}}_{-1,405\times10^{-5}}$
10	2427	607,5	1301	2346,5	$0,008361^{+6,131\times10^{-4}}_{-6,131\times10^{-4}}$	$0,002927^{+1,596\times10^{-5}}_{-1,596\times10^{-5}}$
11	2739	685,5	1404	2527.,5	$0,008103^{+4,043\times10^{-4}}_{-4,043\times10^{-4}}$	$0,002923^{+1,230\times10^{-5}}_{-1,230\times10^{-5}}$
12	2606	654,2	1359	2416,8	$0,00809^{+4,00\times10^{-4}}_{-4,00\times10^{-4}}$	$0,003110^{+1,235\times10^{-5}}_{-1,235\times10^{-5}}$
13	1827	458,5	1006	1989,5	$0,008256^{+6,344\times10^{-4}}_{-6,344\times10^{-4}}$	$0,002843^{+1,694\times10^{-5}}_{-1,694\times10^{-5}}$
14	2193	549	1145	2187	$0,007883^{+3,851\times10^{-4}}_{-3,851\times10^{-4}}$	$0,002944^{+1,203\times10^{-5}}_{-1,203\times10^{-5}}$
15	2357	592	1245	2351	$0,008103^{+4,767\times10^{-4}}_{-4,767\times10^{-4}}$	$0,003074^{+1,423\times10^{-5}}_{-1,423\times10^{-5}}$
16	2421	607	1263	2212	$0,007863^{+3,847\times10^{-4}}_{-3,847\times10^{-4}}$	$0,003069^{+1,275\times10^{-5}}_{-1,275\times10^{-5}}$
17	2612	653,8	1347,5	2339.5	$0,007929^{+3,758\times10^{-4}}_{-3,758\times10^{-4}}$	$0,003098^{+1,210\times10^{-5}}_{-1,210\times10^{-5}}$
18	1755	439,5	978	1984,5	$0,007628^{+4,747\times10^{-4}}_{-4,747\times10^{-4}}$	$0,002935^{+1,478\times10^{-5}}_{-1,478\times10^{-5}}$
19	2675	671,5	1402	2448	$0,008118^{+3,884\times10^{-4}}_{-3,884\times10^{-4}}$	$0,003141^{+1,226\times10^{-5}}_{-1,226\times10^{-5}}$
20	1919	480,5	1062	2039,5	$0,007884^{+4,844\times10^{-4}}_{-4,844\times10^{-4}}$	$0,002959^{+1,481\times10^{-5}}_{-1,481\times10^{-5}}$
21	1465	371	796	1575	$0,00783^{+5,577\times10^{-4}}_{-5,577\times10^{-4}}$	$0,002813^{+1,552\times10^{-5}}_{-1,552\times10^{-5}}$
22	1831	460,5	991	1918,5	$0,007804^{+4,538\times10^{-4}}_{-4,538\times10^{-4}}$	$0,002988^{+1,482\times10^{-5}}_{-1,482\times10^{-5}}$
X	2386	598,2	1266,5	2369,8	$0,008624^{+7,905\times10^{-4}}_{-4,538\times10^{-4}}$	$0,002776^{+1,978\times10^{-5}}_{-1,978\times10^{-5}}$
Y	655	179,5	375	825,5	$0,005668^{+5,892\times10^{-4}}_{-5,892\times10^{-4}}$	$0,004209^{+2,345\times10^{-4}}_{-2,345\times10^{-4}}$

Tabela 5.1: Principais características do conjunto de dados e valores de melhor ajuste para k_1 e k_2. A coluna Crm identifica o cromossomo; N corresponde ao tamanho da amostra; Q1 é o primeiro, Q2 segundo e Q3 o terceiro quartil do conjunto de dados; As colunas k_1 e k_2 identificam, respectivamente, os melhores ajustes dos parâmetro k_1 e k_2 com confiança de 95%.

6. Conclusão

Nesta dissertação, utilizamos a máxima entropia de informação para obter uma função de distribuição de probabilidade para descrever a ocorrência dos comprimentos nas sequências codificantes do DNA humano. Ao longo dos capítulos apresentamos uma breve introdução de genética, de teoria de informação, surpresa e entropia. A partir desta última foi possível construir a distribuição de probabilidade $P(l)$. Para testar a viabilidade do modelo utilizamos os dados de acesso público disponíveis em bancos genéticos construídos com a finalidade de armazenar DNA de espécies já sequenciadas e disponibilizá-los para os pesquisadores e sociedade em geral.

O estudo das propriedades estatísticas das sequências genômicas não é novo, alguns trabalhos inclusive remotam a década de 90. Nesta dissertação, abordamos em um dos capítulos alguns estudos vanguardistas nesta área e constatamos descobertas relevantes realizadas por eles como a existência de correlação entre os segmentos das sequências. O diferencial dessa dissertação é que a partir da entropia da informação damos origem à função de probabilidade que depende de dois parâmetros ajustáveis k_1 e k_2. Os valores numéricos dos ajustes tiveram uma variação de $0,005668 \leq k_1 \leq 0,008624$ e $0,002776 \leq k_2 \leq 0,004209$ e contam com confiança de 95%. Observamos também que os valores extremos para os parâmetros de ajuste foram obtidos pelos cromossomos sexuais X e Y.

Discutimos também a implicação dos resultados obtidos para k_1 e k_2 e concluímos que valores pequenos ($\sim 10^{-3}$) para os parâmetros implicam que há probabilidade muito alta de ocorrência dos menores comprimentos e extremamente baixa para os demais. Uma possível explicação para o fato dos éxons não apresentarem comprimentos tão longos esteja no fato de sequências longas não sejam necessárias para produção de proteínas uma vez que apenas uma trinca (códon) é suficiente pra construir um aminoácido que por sua vez dará origem às proteínas.

Observamos também uma correlação estatística entre os parâmetros de ajuste k_1 e k_2. Vimos que em 22 cromossomos mais o cromossomo sexual X, o crescimento de um dos parâmetros implica no aumento do outro. A exceção foi observada no cromossomo Y, onde

se observou um comportamento de anticorrelação estatística entre os parâmetros, isto é, o crescimento de um dos parâmetros implica no decrescimento do outro. É necessária uma investigação mais profunda para determinar a razão desse comportamento e se os parâmetros são capazes de descrever alguma característica biológica das sequências dos cromossomos.

Construímos as elipses de confiança para k_1 e k_2 e representamos graficamente os intervalos de confiabilidade de $68\%, 95\%$ e 99%. As elipses são uma técnica estatística que nos permite representar visualmente um diagrama de dispersão dos resultados associados a uma região de confiança.

Como possível extensão, pode-se utilizar a mesma abordagem para descrever distribuição de comprimentos das sequências do DNA não codificantes que são uma grande região da sequência gênica mas que pouco se sabe a respeito de sua função. Em outra vertente, pode ser a introdução do modelo baseado na exponencial *stretched*, que é utilizado no estudo de propriedades estatísticas do DNA (Ver, por exemplo a Ref. [52]). Além disso, outra possibilidade, é considerar a viabilidade do modelo para as sequências genômicas de outros seres vivos.

7. Bibliografia

[1] A.J. Griffiths, S.R. Wessler, S. Carroll, J. Doebley, *Introdução à Genética. 10ª edição.* Rio de Janeiro, Guanabara Koogan, 2013.

[2] 'Adaptação', https://escola.britannica.com.br/artigo/adapta%C3%A7%C3%A3o/480526, Accesso: 2019-12-20.

[3] B. Alberts, *Biologia Molecular da Célula*, Av. Jerônimo de Ornelas, 670 – Santana, 90040-340, Porto Alegre RS: Artmed Editora S.A., 2017.

[4] P. Snustad, M.J. Simmons, P.A. Motta, *Fundamentos de Genética.*, Grupo Gen-Guanabara Koogan,2000.

[5] O. T. Avery, C. M. MacLeod and M. McCarty, 'Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III', *The Journal of experimental medicine*, vol. 79, no. 2, pp. 137–158, 1944.

[6] J. D. Watson, *A dupla hélice,Jorge Zahar Editor Ltda.* , 2014.

[7] J. D. Watson and F. H. Crick, *1953. A structure for deoxyribose nucleic acid*, University of Chicago Press, 2010.

[8] A. Zaha, H.B. Ferreira and L.M. Passaglia, *Biologia Molecular Básica-5.* Artmed Editora, 2014.

[9] A. Klug, 'Rosalind Franklin and the discovery of the structure of DNA',*Nature*, vol. 219, no. 5156, pp. 808–810, 1968.

[10] M.O Costa, R. Silva, D.H.A.L. Anselmo and J.R.P. Silva, 'Analysis of human DNA through power-law statistics', *Phys. Rev. E*, vol. 99, Feb 2019.

[11] T. Oikonomou, A. Provata and U. Tirnakli, "Nonextensive statistical approach to non-coding human DNA", *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 11, pp. 2653 - 2659, 2008.

[12] C. K. Peng, S. Buldyrev, A. Goldberger, S. Havlin, F. Sciortino, M Simons and H. Stanley, "Long Range Correlations in Nucleotide Sequence", *Nature*, vol. 356, pp. 168-70 04 1992.

[13] W. Li and K. Kaneko, "Long-Range Correlation and Partial 1/f Spectrum in a Non-coding DNA Sequence", *Europhysics Letters*, vol. 17, pp. 655–660, Feb 1992.

[14] A. Colliva, R. Pellegrini, A.Testori and M. Caselle, "Ising-model description of long-range correlations in DNA sequences", *Phys. Rev. E*, vol. 91, p. 052703, May 2015.

[15] H. Stanley, S. Buldyrev, A. Goldberger and S. Havlin, C.K. Peng, M. Simons, "Scaling features of noncoding DNA", *Physica A: Statistical Mechanics and its Applications*, vol. 273, no.1, pp. 1 - 18, 1999.

[16] R. F. Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences", *Phys. Rev. Lett.*, vol. 68, pp. 3805–3808, Jun 1992.

[17] A. K. Mohanty andA. V. S. S. Narayana Rao, "Factorial Moments Analyses Show a Characteristic Length Scale in DNA Sequences", *Phys. Rev. Lett.*, vol. 84, pp. 1832–1835, Feb 2000.

[18] M. V. Koroteev and J. Miller, "Scale-free duplication dynamics: A model for ultra-duplication", *Phys. Rev. E*, vol. 84, pp. 061919, Dec 2011.

[19] S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M.E. Matsa, C. -K. and Peng, M. Simons, and H. E. Stanley, "Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis", *Phys. Rev. E*, vol. 51, pp. 5084–5091, May 1995.

[20] "Teoria da Informação". https://www.ensinoeinformacao.com/teoria-da-informcao. Accesso: 2020-04-30.

[21] "Information theory". https:https://www.britannica.com/science/information-theory. Accesso: 2020-06-11.

[22] C. E. Shannon, "A mathematical theory of communication", *Bell system technical journal*, vol. 27, no.3, pp.379–423, 1948.

[23] R. V. L. Hartley, "Transmission of Information", *Bell System Technical Journal* vol. 7, no. 3, pp. 535-563.

[24] W. G. Tuller, "Theoretical Limitations on the Rate of Transmission of Information", *Proceedings of the IRE*, vol. 37, pp. 468-478, 1949.

[25] James V. Stone, *Information Theory: A tutorial introduction*, Sebtel Press, 2015.

[26] Oliver Rioul, *Teoria da Informação e da Codificação*, Editora da Universidade de Brasília, 2018.

[27] T. M. Cover and J. A. Thomas, *Elements of information theory*, John Wiley & Sons, 2012.

[28] D. A. Moreira, *Propriedades termo-eletrônicas da molécula do DNA*, PhD. thesis, Universidade Federal do Rio Grande do Norte, 2008.

[29] "Cromossomos", https://slideplayer.com.br/slide/386963/, Accesso: 2019-09-16.

[30] "O que é um Cromossomo? Como Funciona? Genética - Vídeo Animado", https://www.youtube.com/watch?v=UBfInkTvqt8, Acesso: 2020-05-20.

[31] "Splicing", http://labs.icb.ufmg.br/lbcd/prodabi3/grupos/grupo1/splicing.htm, Accesso: 2020-05-15.

[32] E. Volkin and L. Astrachan, "Phosphorus incorporation in Escherichia coli ribonucleic acid after infection with bacteriophage T2", *Virology*, vol. 2, no. 2, pp. 149 - 161, 1956.

[33] S. Kaplan, A.O.W. Stretton and S. Brenner, "Amber suppressors: Efficiency of chain propagation and suppressor specific amino acids", *Journal of Molecular Biology*, vol. 14, no. 2, 1965.

[34] "Do DNA à proteína. Mecanismo da transcrição e tradução do DNA, em 3D", https://www.youtube.com/watch?v=H_82vthxqLk, Accesso: 2021-01-10.

[35] "Conceito de Informação", https://https://conceito.de/informacao, Accesso: 2019-05-04.

[36] "Data and information", https://www.computerhope.com/issues/ch001629.htm, Accessed: 2020-02-15.

[37] E. B. Sturtevant and A. G. Lewis, *A History of Genetics*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 2001.

[38] "Ensembl Genome Database Project", https://www.ensembl.org/index.html, Acesso: 2019-04-30.

[39] R. Silva, D.H.A.L. Anselmo, J.R.P. Silva, W. da Silva, M.O Costa, "An alternative description of power law correlations in DNA sequences", *Physica A: Statistical Mechanics and its Applications*,vol. 545, pp.123735, 2020.

[40] I. Gleria, R. Matsushita, Raul and S. Da Silva, "Sistemas Complexos, criticalidade e leis de potência", *Revista Brasileira de Ensino de Física*, vol. 26, no. 2, pp. 99 - 108, 2004.

[41] G. Kaniadakis , "Maximum Entropy Principle and Power-Law Tailed Distributions", *European Physical Journal B*, vol. 70, 04 2009.

[42] R. Hanel, S. Thurner and M. Gell-Mann ,"Generalized entropies and the transformation group of superstatistics", *Proceedings of the National Academy of Sciences*, vol. 108, no. 16, pp. 639-6394, 2011.

[43] "European Bioinformatics Institute", https://www.ebi.ac.uk/, Acesso: 2020-06-04.

[44] "Wellcome Trust Sanger Institute", https://www.sanger.ac.uk/, Accesso: 2020-06-04.

[45] "Ensembl genome database project", https://en.wikipedia.org/wiki/Ensembl_genome_database_project, Accesso: 2020-06-04.

[46] "The R Project for Statistical Computing", https://www.r-project.org/, Accesso: 2019-05-25.

[47] N. A. Lemos, *Mecânica Analítica*, Editora Livraria da Física, 2007.

[48] T. Oikonomou and A. Provata, "Non-extensive trends in the size distribution of coding and non-coding DNA sequences in the human genome", *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 50, no. 1-2, pp. 259-264, 2006.

[49] L. G. Morettin, *Estatística básica: probabilidade e inferência: volume único*, Pearson Prentice Hall, 2010.

[50] "Confidence Region", https://en.wikipedia.org/wiki/Confidence_region, Acesso: 2020-06-03.

[51] M. F. Triola, *Elementary Statistics, (Technology Update)*, Pearson Education, 2010.

[52] M .Hillebrand, G. Kalosakas, C. Skokos, and A. R.Bishop, "Distributions of bubble lifetimes and bubble lengths in DNA", *Phys. Rev. E*, vol 102, pp. 062114, Dec 2020.