UNIVERSIDADE FEDERAL DE ALFENAS

BIANCA LAPA RIBEIRO

HIPERCUBOS BOOLEANOS APLICADOS NA ANÁLISE DO CÓDIGO GENÉTICO E ENTROPIA APLICADA EM SEQUÊNCIAS DE DNA

ALFENAS/MG
2025

BIANCA LAPA RIBEIRO

HIPERCUBOS BOOLEANOS APLICADOS NA ANÁLISE DO CÓDIGO GENÉTICO E ENTROPIA APLICADA EM SEQUÊNCIAS DE DNA

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Estatística Aplicada e Biometria, pela Universidade Federal de Alfenas. Área de concentração: Estatística Aplicada e Biometria.

Linha de Pesquisa: Biometria.

Orientador: Prof. Dr. Anderson José de Oliveira.

ALFENAS/MG
2025

Sistema de Bibliotecas da Universidade Federal de Alfenas
Biblioteca Central

Ribeiro, Bianca Lapa.
Hipercubos Booleanos Aplicados na Análise do Código Genético e Entropia Aplicada em Sequências de DNA / Bianca Lapa Ribeiro. - Alfenas, MG, 2025.
95 f. : il. -

Orientador(a): Anderson José de Oliveira.
Dissertação (Mestrado em Estatística Aplicada e Biometria) - Universidade Federal de Alfenas, Alfenas, MG, 2025.
Bibliografia.

1. Teoria da Informação. 2. Cadeias de Markov. 3. Código de Gray. 4. Aminoácidos. I. de Oliveira, Anderson José, orient. II. Título.

Ficha gerada automaticamente com dados fornecidos pelo autor.

BIANCA LAPA RIBEIRO

HIPERCUBOS BOOLEANOS APLICADOS NA ANÁLISE DO CÓDIGO GENÉTICO E ENTROPIA APLICADA EM SEQUÊNCIAS DE DNA

O Presidente da banca examinadora abaixo assina a aprovação da Dissertação apresentada como parte dos requisitos para a obtenção do título de Mestre em Estatística Aplicada e Biometria pela Universidade Federal de Alfenas. Área de concentração: Estatística Aplicada e Biometria

Aprovada em: 10 de fevereiro de 2025.

Prof. Dr. Anderson José de Oliveira

Presidente da Banca Examinadora

Instituição: Universidade Federal de Alfenas - UNIFAL-MG

Profa. Dra. Clarice Dias de Albuquerque

Instituição: Universidade Federal do Cariri - Juazeiro do Norte - CE

Profa. Dr. Denismar Alves Nogueira

Instituição: Universidade Federal de Alfenas - UNIFAL-MG

AGRADECIMENTOS

Acima de tudo agradeço a Deus, pela vida, pela conquista alcançada, força, sabedoria e serenidade concedidas ao longo desta jornada. Por estar presente em cada momento, guiando meus passos, iluminando meu caminho, me dando coragem nos desafios e serenidade nas dificuldades.

Aos meus pais, Maria de Fátima e Ronaldo, por todo apoio e incentivo nos momentos bons e difíceis, por desejarem sempre o melhor para mim e pelo esforço que fizeram para que eu pudesse superar cada obstáculo em meu caminho e chegar aqui. À minha irmã Caroline, e aos meus amigos, que sempre estiveram ao meu lado, me apoiando e torcendo por mim.

Ao meu orientador e agora também amigo, Prof. Dr. Anderson José de Oliveira, por todo conhecimento compartilhado, pelos momentos vivenciados juntos, pela pessoa e profissional que é e por ser uma grande inspiração para mim. E, principalmente, agradeço por sempre ter acreditado e depositado sua confiança em mim ao longo desses anos que se iniciaram ainda na graduação.

Aos membros da banca, Denismar, Clarice, Cátia e Eric que com todas suas contribuições, observações e apontamentos foram fundamentais para o aprimoramento deste trabalho.

À Universidade Federal de Alfenas e à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) pelo apoio financeiro.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

E a todos aqueles que, de alguma forma, contribuíram nessa etapa da minha vida.

RESUMO

A modelagem matemática do código genético é um estudo que possibilita dentre outros aspectos a análise, a interpretação e a caracterização de propriedades associadas aos aminoácidos e possíveis interferências em diversas situações, como o caso das mutações genéticas. Diagramas de Hasse, códigos de Gray e hipercubos booleanos representam algumas ferramentas matemáticas que podem ser empregadas nesse estudo. O código genético consiste na associação das trincas encontradas no RNA mensageiro, formadas pelas bases nitrogenadas e os aminoácidos que estão nas proteínas. A partir do mapeamento das bases nitrogenadas com a estrutura algébrica do anel $\mathbb{Z}_4 = \{0, 1, 2, 3\}$, é possível obter 24 permutações, organizadas em três rotulamentos (A, B e C). Ademais, o código genético pode ser representado por um hipercubo booleano 6-dimensional, construído a partir da tabela do código de Gray. Outro aspecto a ser considerado é a entropia da informação, a qual auxilia a identificar padrões, tanto dentro de uma sequência genética específica, quanto entre diferentes sequências genéticas, uma vez que, conhecida a entropia, é possível gerar métodos para que uma mensagem chegue ao destino com confiabilidade. Nesse sentido, é possível aplicar esse conceito da base de informação, o DNA, até a síntese completa de uma proteína. Em Teoria da Informação, a entropia é obtida em relação a uma sequência, um modelo e a distribuição de probabilidades fornecida pelo modelo, sendo utilizadas as Cadeias de Markov. O objetivo deste trabalho é apresentar uma caracterização do código genético por meio de estruturas matemáticas, como o código de Gray e o hipercubo booleano, além de analisar sequências de DNA a partir de elementos estatísticos, em particular, a entropia. A metodologia adotada neste trabalho baseia-se em uma natureza qualitativa e quantitativa, visando um estudo descritivo aplicado. A pesquisa foi dividida em cinco etapas: 1) fundamentação teórica: elementos de Biologia, Álgebra e Teoria da Informação; 2) compreensão das construções dos diagramas de Hasse e dos códigos de Gray, utilizando permutações associadas aos três rotulamentos (A, B e C) do código genético; 3) construção de hipercubos booleanos, a partir de tabelas dos códigos de Gray; 4) análise das construções realizadas; 5) análise das possibilidades de aplicações de elementos da Teoria da Informação associados a problemas biológicos, como a entropia. Com isso, buscou-se compreender as conexões existentes entre Biologia, Álgebra, Geometria e Engenharia, além de analisar as possibilidades de aplicação e possíveis contribuições da Teoria da Informação no estudo e análise do código genético e em sequências de DNA.

Palavras-chave: Teoria da Informação; Cadeias de Markov; Código de Gray; Aminoácidos.

ABSTRACT

Mathematical modeling of the genetic code is a study that enables, among other aspects, the analysis, interpretation and characterization of properties associated with amino acids and possible interferences in various situations, such as genetic mutations. Hasse diagrams, Gray codes and Boolean hypercubes represent some mathematical tools that can be used in this study. The genetic code consists of the association of triplets found in messenger RNA, formed by nitrogenous bases and the amino acids that are in proteins. From the mapping of the nitrogenous bases with the algebraic structure of the ring $\mathbb{Z}_4 = \{0, 1, 2, 3\}$, it is possible to obtain 24 permutations, organized into three labelings (A, B and C). In addition, the genetic code can be represented by a 6-dimensional Boolean hypercube, constructed from the Gray code table. Another aspect to be considered is the entropy of information, which helps to identify patterns, both within a specific genetic sequence and between different genetic sequences, since, once entropy is known, it is possible to generate methods for a message to reach its destination reliably. In this sense, it is possible to apply this concept from an information base, DNA, to the complete description of a protein. In Information Theory, entropy is obtained in relation to a sequence, a model and the probability distribution provided by the model, being used as Markov Chains. The aim of this work is to present a characterization of the genetic code through mathematical structures, such as the Gray code and the Boolean hypercube, in addition to analyzing DNA sequences from statistical elements, in particular, the entropy. The methodology is based on a qualitative and quantitative nature, aiming for descriptive study and applied approach. The research will be divided into five steps: 1) theoretical basis: elements of Biology and Algebra and Information Theory; 2) understanding the constructions of Hasse diagrams and Gray codes, using permutations associated with three labelings (A, B and C) of the genetic code; 3) construction of Boolean hypercubes, from the tables of Gray codes; 4) analysis of the constructions made; 5) analysis of the proposed applications of elements of Information Theory associated with biological problems, such as the entropy. With this, we seek to understand the existing connections among Biology, Algebra, Geometry and Engineering, in addition to analyzing the possibilities of application and possible contributions of Information Theory in the study and analysis of the genetic code in DNA sequences.

Keywords: Information Theory; Markov Chains; Gray Code; Amino acids.

LISTA DE FIGURAS

SUMÁRIO

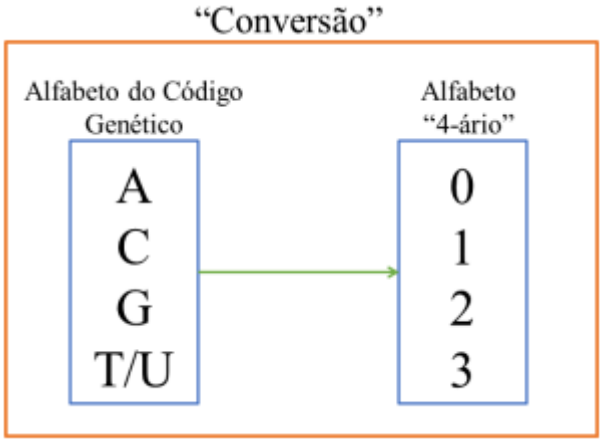
1 INTRODUÇÃO

Na modelagem do código genético são utilizadas estruturas matemáticas com o propósito de analisar, interpretar e caracterizar propriedades associadas aos aminoácidos e possíveis interferências em diversas situações. Algumas das abordagens utilizadas incluem os diagramas de Hasse, como mostram Fernandes e Oliveira (2021), o código de Gray e o hipercubo booleano, conforme apresentado em Jimenez-Montano e La Mora-Basanez (2002).

O código genético consiste na associação das trincas encontradas no RNA mensageiro, formadas pelas bases nitrogenadas e os aminoácidos que estão nas proteínas. As trincas de bases nitrogenadas são denominadas códons. Existem 64 possíveis trincas, agrupadas três-a-três, que correspondem a 20 aminoácidos, os quais são as unidades básicas que formam as proteínas. A correspondência de um determinado aminoácido com um códon se dá por meio do código genético, conforme apresentado em Alberts (2017).

Em Faria (2011) é apresentada a modelagem do DNA por meio de um mapeamento do alfabeto do código genético, que está relacionado com o conjunto de nucleotídeos $N = \{A, C, G, T/U\}$ (A - adenina, C - citosina, G - guanina e T/U - timina/uracila) e o alfabeto 4-ário da estrutura de anel, denotado por $\mathbb{Z}_4 = \{0, 1, 2, 3\}$. As possibilidades de associações do alfabeto biológico com o alfabeto matemático possibilitam a identificação de 24 permutações, que são separadas em três rotulamentos, os quais tem como finalidade determinar qual a melhor associação de cada um dos símbolos do conjunto N com os correspondentes símbolos do conjunto $\mathbb{Z}_4$ e vice-versa. Como a estrutura algébrica do alfabeto do código genético das sequências do DNA é desconhecida, tem-se a conversão realizada na Figura 1.

Figura 1 – O alfabeto do código genético e o alfabeto 4-ário.

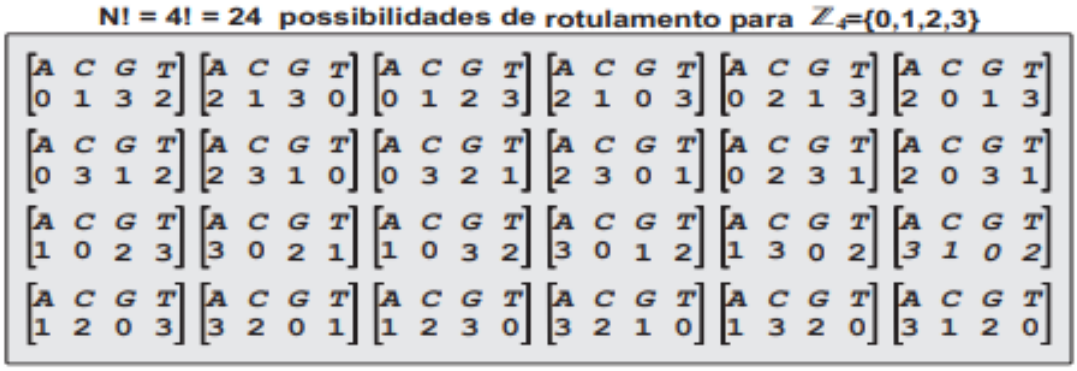


Fonte: Fernandes e Oliveira (2021).

As 24 permutações mencionadas anteriormente são apresentadas na Figura 2.

Assim, cada permutação possui suas próprias características, permitindo organizar esse mapeamento em três conjuntos, chamados de rotulamentos A, B e C (Figura **??**), cada um contendo 8 permutações, conforme a caracterização geométrica associada a cada uma das permutações.

Figura 2 – Possibilidades de permutações entre os elementos dos conjuntos N e $\mathbb{Z}_4$.



Fonte: Faria (2011).

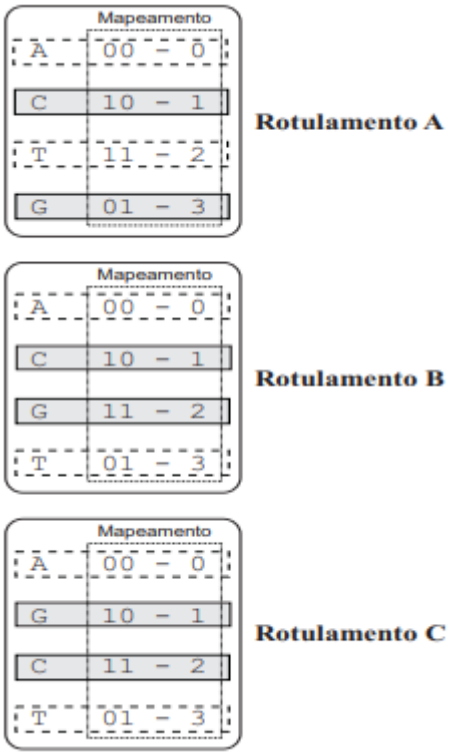
De acordo com Faria (2011), o mapeamento dos rotulamentos decorre da complementaridade biológica $(A-T)$ e $(C-G)$, que procede da complementaridade matemática $(00-11)$ e $(01-10)$ e, consequentemente, sobre os códigos gerados que podem identificar e reproduzir sequências de DNA. Contudo, a complementaridade biológica pode ou não estar casada com a complementaridade matemática, como mostra a Figura 3. Rotulamento A: A complementaridade biológica $(A-T)/(C-G)$ está casada com a complementaridade mátemática $(00-11)/(10-01) \equiv$ maior distância de Hamming $d_H = 1$ entre $A-T$ e $C-G$; rotulamento B: $(A-T)/(C-G)$ não casa com $(00-11)/(10-01) \equiv$ maior distância de Hamming $d_H = 1$ entre $A-T$ e $C-G$; rotulamento C: $(A-T)/(C-G)$ não casa com $(00-11)/(10-01) \equiv$ maior distância de Hamming $d_H = 1$ entre $A-T$ e $C-G$.

No rotulamento A, a combinação entre a complementaridade biológica e a matemática resulta em um mapeamento não-linear. Por outro lado, nos rotulamentos B e C, essa combinação não ocorre. Como a complementaridade biológica deve prevalecer, obtém-se mapeamentos lineares. O mapeamento não-linear é denominado de $\mathbb{Z}_4$-linear e os mapeamentos lineares de $\mathbb{Z}_2 \times \mathbb{Z}_2$-linear e Klein-linear, respectivamente, como mostra a Figura 4.

Diante disso, nota-se que no rotulamento A qualquer um dos nucleotídeos para alcançar o seu complementar necessita caminhar duas arestas, enquanto que nos dois rotulamentos restantes basta caminhar somente uma aresta.

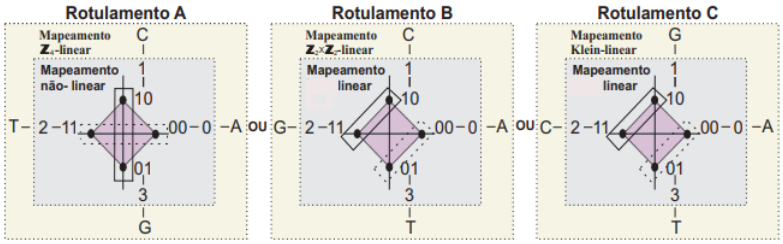
Além disso, uma outra importante estrutura utilizada no processo de modelagem do código genético é o diagrama de Hasse, o qual é composto por 64 códons, dispostos em 7 linhas e que consiste em apresentar os códons do código genético de maneira

Figura 3 – Complementaridade biológica e matemática.



Fonte: Faria (2011).

Figura 4 – Mapeamentos do alfabeto $\mathbb{Z}_4$.



Fonte: Faria (2011).

organizada, com o propósito de analisar as propriedades dos aminoácidos e classificar os códons conforme suas características. Nesse contexto, em Fernandes e Oliveira (2021) é apresentada a construção de reticulados booleanos e diagramas de Hasse (casos primal e dual), a fim de analisar as diferenças e semelhanças físico-químicas dos aminoácidos, por meio da caracterização biológica das construções e do cálculo das médias das distâncias de Hamming entre os códons, os quais poderão ser utilizados no processo de análise de fenômenos mutacionais.

Além das estruturas apresentadas anteriormente, em Hage e Harary (2002) é apresentada a estrutura de um hipercubo booleano e a relação com um código binário, mostrando construções para as dimensões de 1 a 4 do hipercubo. Em Jimenez-Montano,

La Mora-Basanez e Poeschel (2002) é apresentada uma representação do código genético através de um hipercubo booleano 6-dimensional, construído a partir da tabela do código de Gray, em que os vértices de cada cubo são representados pelos códons. O hipercubo representa simultaneamente todo o conjunto de códons e mantém o controle de quais códons são adjacentes. A construção do hipercubo booleano é realizada a partir de todas as ligações, feitas por meio da identificação da diferença de um bit existente de um vértice para o outro, desde a dimensão 1 até a dimensão 6.

Seguindo nesse contexto de aplicações da Matemática na Biologia, em Nussbaum (2008) é apresentada uma visão geral sobre a estrutura do genoma humano e os elementos da genética molecular, com foco no gene β-globina. Além disso, analisa uma sequência de DNA, que será utilizada no capítulo 5 deste trabalho.

Desta forma, um outro aspecto que merece destaque no processo de análise e interpretação de problemas biológicos por meio de estruturas matemáticas é na relação existente entre elementos de Estatística e Biologia. Em Horta (2001), são apresentadas aplicações da Teoria da Informação à Biologia Molecular, em particular, aplicações de estimadores de entropia em sequências de DNA, abrangendo a comparação genômica, além dos limites entre íntrons e éxons. Nessa perspectiva, a entropia está relacionada às distribuições de probabilidades e previsibilidade. A entropia é uma medida de incerteza definida na Teoria na Informação, por Claude Shannon, a partir da necessidade de otimizar a quantidade de informação que pode ser transmitida por meio de canais de comunicação ruidosos, apresentado em Shannon (1948).

Na Teoria da Informação, como afirma Horta (2001), a entropia é calculada em relação a uma sequência, um modelo e a distribuição de probabilidades associada ao modelo, geralmente utilizando Cadeias de Markov. Assim, a entropia pode auxiliar na descoberta de padrões tanto em sequências genéticas específicas, quanto na comparação entre diferentes sequências. O valor da entropia é obtido por meio de cálculos sobre a distribuição de probabilidades dada pela aplicação do modelo à sequência. Desse modo, a incerteza é medida mediante à função logaritmo e a entropia é a média dessa incerteza para uma distribuição de probabilidades.

Dessa maneira, a partir das informações apresentadas anteriormente e questões que ainda estão em aberto acerca da modelagem matemática do código genético, o objetivo deste trabalho é representar o código genético por meio da construção de estruturas matemáticas, como o código de Gray e o hipercubo booleano, construídos a partir de diagramas de Hasse, obtidos de permutações escolhidas dos três rotulamentos do código genético: A, B e C. Além disso, verificar aplicações de elementos da Teoria da Informação, em particular da entropia, na análise e interpretação do código genético, por meio da sequência de DNA β-globina.

Este trabalho está estruturado da seguinte maneira: neste capítulo é apresentada

a introdução. No capítulo 2, são apresentados os conceitos teóricos fundamentais que sustentam este trabalho, abrangendo aspectos específicos de Genética e Álgebra. No capítulo 3 são apresentadas a construção e a análise dos diagramas de Hasse, códigos de Gray e hipercubos booleanos associados aos rotulamentos A, B e C do código genético, por meio das permutações 1203, 3021, 0321, 2103, 1302 e 3120.

No capítulo 4, são apresentados os elementos da Teoria da Informação e Biologia Molecular. No capítulo 5, são apresentados os resultados associados à aplicação de elementos da Teoria da Informação em uma sequência de DNA, incluindo a aplicação da entropia aos Modelos M_1 e M_2 condicionados aos éxons da sequência, além da aplicação desses modelos na sequência completa, explorando ainda os contextos finitos e ordem do modelo. Por fim, no capítulo 6 são apresentadas as considerações finais do trabalho e as propostas para trabalhos futuros.

2 ELEMENTOS DE BIOLOGIA E ÁLGEBRA

Em virtude da interdisciplinaridade desta pesquisa, neste capítulo serão apresentados conceitos relacionados a Biologia e Álgebra, os quais serão utilizados na obteção e análise dos resultados no capítulo 3.

2.1 BIOLOGIA MOLECULAR

Nesta seção serão apresentados os principais elementos de Biologia utilizados neste trabalho. As referências utilizadas foram: Alberts et al. (2017), Faria (2011), Fernandes e Oliveira (2021), Franco e Palazzo Júnior (2015) e Gerônimo e Franco (2008).

2.1.1 Célula

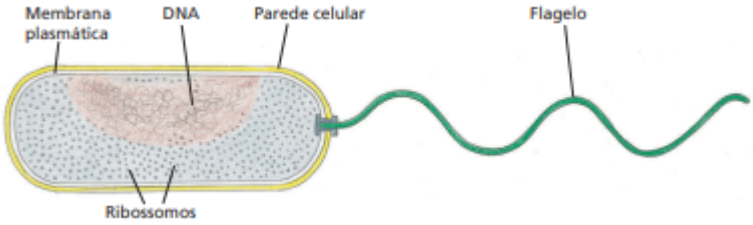
De acordo com Alberts *et al.* (2017), a célula é uma unidade de matéria viva que forma o corpo de todos os seres vivos, sendo capaz, individualmente, de obter energia, crescer e se reproduzir.

A célula possui três componentes básicos: a membrana plasmática, o citoplasma e o material genético. A membrana plasmática envolve a célula, delimitando o citoplasma - um material fluido que contém organelas e outras estruturas celulares - e protege o material genético, que carrega as informações genéticas dos seres vivos.

Diante destes componentes, é possível classificar as células em dois tipos principais: procarionte, onde o material genético fica disperso sobre o citoplasma, e eucarionte, em que o material genético fica separado do citoplasma por meio da carioteca, uma estrutura membranosa. Devido a esta delimitação do material genético, será estudado com mais detalhes a classe celular eucariótica que forma o corpo de todos os multicelulares, como os fungos, as plantas, os animais e os seres humanos.

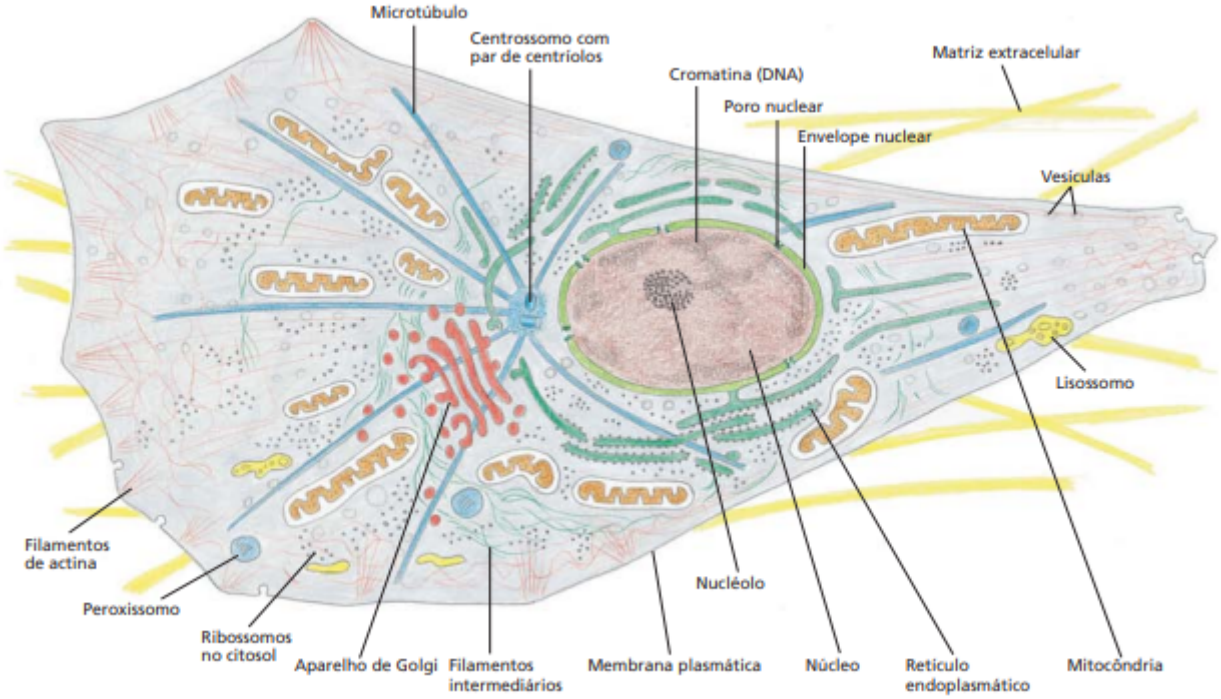
Ainda em Alberts *et al.* (2017), é mencionado que a estrutura de uma célula eucariótica é formada, além dos componentes básicos, por estruturas citoplasmáticas, tais como: as organelas (limitadas por membranas), os ribossomos e os centríolos. Outra estrutura importante desse tipo de célula é o núcleo, formado pela carioteca, onde estão localizados os cromossomos. Esses cromossomos são compostos principalmente por DNA (ácido desoxirribonucleico), uma molécula que contém toda a informação genética de um ser vivo. As Figuras 5 e 6 ilustram, respectivamente, uma célula procarionte e uma célula eucarionte.

Figura 5 – Célula procarionte.



Fonte: Alberts *et al.* (2017).

Figura 6 – Célula eucarionte.



Fonte: Alberts *et al.* (2017).

2.1.2 Nucleotídeos e ácidos nucleicos

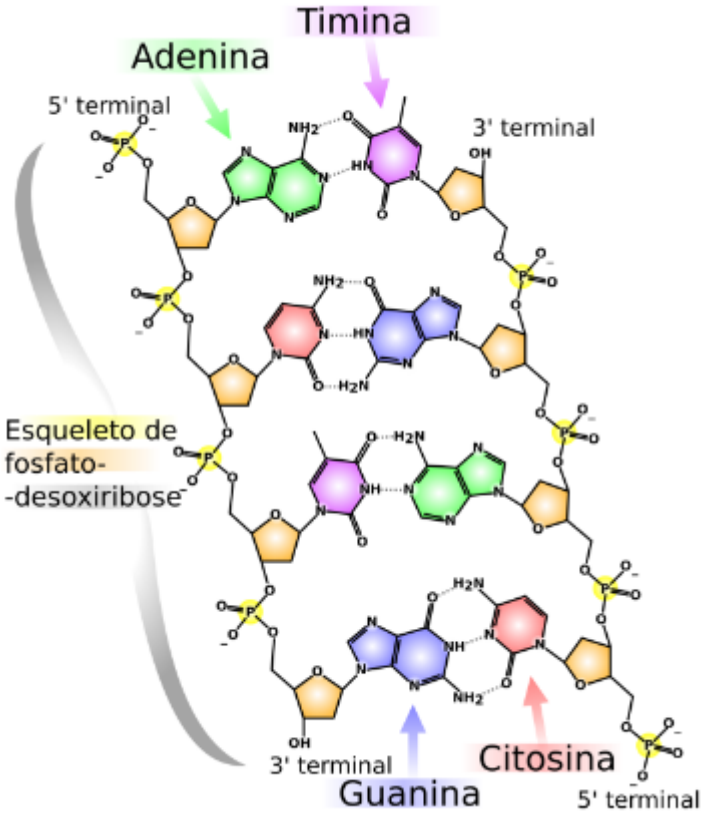
Conforme Alberts *et al.* (2017), as células possuem uma composição química, formada por substâncias inorgânicas, tendo origem mineral, como a água e sais minerais. E por substâncias orgânicas, com origem vegetal ou animal, sendo os carboidratos, lipídios, proteínas e ácidos nucleicos. Esta última é uma importante substância relacionada à transmissão de características hereditárias presente na informação genética do organismo, formadas por nucleotídeos.

Os nucleotídeos são uma repetição de moléculas menores, compostos por um grupo de fosfato (ácido fosfórico), uma molécula de açúcar e uma base nitrogenada. Estas bases são divididas em bases púricas, que são adenina(A) e guanina(G); bases pirimídicas, que

são citosina(C), timina (T) e uracila(U). Desse modo, podem ser formados dois tipos de ácidos nucleicos: ácido desoxirribonucleico (DNA) e ácido ribonucleico (RNA). Tanto o DNA quanto o RNA são polímeros, isto é, moléculas formadas por várias unidades menores ligadas entre si de modo organizado.

De acordo com Hepp e Nonohay (2016), o DNA funciona como um banco de informações genéticas, uma espécie de código, que transfere essas informações, com o objetivo de garantir a integridade da informação genômica e funcionando como molde para a síntese da molécula de RNA. A molécula de DNA é formada por uma fita dupla de vários nucleotídeos (dupla hélice), que são formados por um grupo de fosfato, uma pentose (desoxirribose, que é um açúcar composto por cinco átomos de carbono, formando uma cadeia fechada) e uma base nitrogenada. As quatro bases nitrogenadas são: adenina e timina (purinas - maiores), citosina e guanina (pirimidinas - menores). A adenina liga-se apenas à timina e a citosina apenas à guanina. As bases nitrogenadas são apresentadas na Figura 7.

Figura 7 – Composição dos nucleotídeos do DNA. Cada uma das quatro bases liga-se à desoxirribose pelo nitrogênio e a um grupo de fosfato para formar os nucleotídeos correspondentes.



Fonte: Hepp e Nonohay (2016).

De acordo com Alberts *et al.* (2017), a molécula de RNA (ácido ribonucleico) é formada a partir da molécula de DNA em um processo chamado de transcrição. Essa

molécula apresenta informações com as quais é possível coordenar a produção de proteínas. Assim sendo, o RNA também participa do fluxo de informações genéticas pelos indivíduos.

O RNA é uma molécula composta por uma única fita de nucleotídeos ligados entre si. A pentose do RNA é sempre a ribose e as quatro bases nitrogenadas são: adenina, guanina, citosina e uracila, sendo a última exclusiva do RNA.

As moléculas de RNA são, geralmente, constituídas de uma única fita que se enrola entre si devido ao emparelhamento entre as bases complementares: a uracila liga-se com a adenina e a guanina liga-se com a citosina, seguindo a Regra de Chargaff. Elas são classificadas de acordo os papéis que desempenham no processo de transferência de informação. Existem três tipos de RNA: RNA mensageiro (RNAm ou mRNA), RNA transportador (RNAt) e RNA ribossômico (RNAr), conforme apresentado em Hepp e Nonohay (2016).

O RNA mensageiro é uma cópia das fitas de DNA, ficando responsável em levar as informações obtidas do DNA até o citoplasma, onde as proteínas serão produzidas. Como o RNA é uma cópia fiel de uma das fitas de DNA, é a partir dessa informação que o RNA mensageiro irá determinar quais são os aminoácidos necessários para a formação de determinada proteína, pois ele possui as trincas (códons) de bases nitrogenadas que definem cada aminoácido.

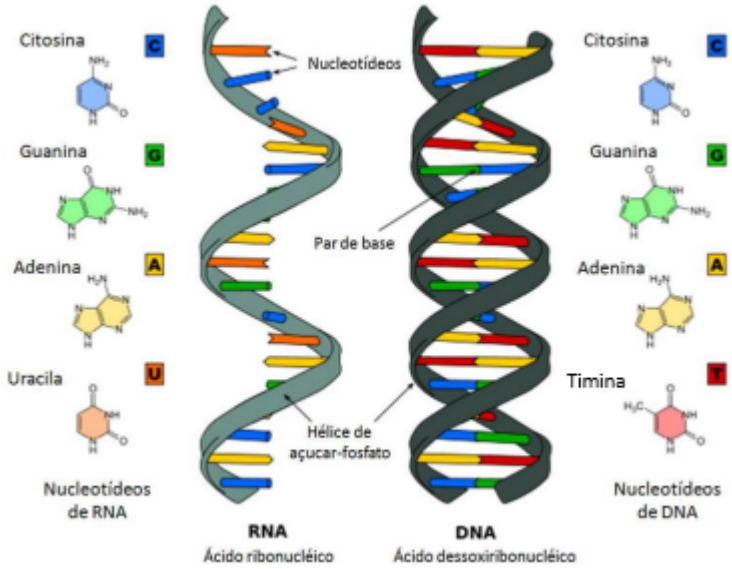
O RNA transportador é encarregado de transportar os aminoácidos que serão usados na formação das proteínas até o ribossomo, enquanto, que o RNA ribossômico faz parte da constituição dos ribossomos. É nos ribossomos que a sequência de bases do RNA mensageiro é interpretada e a proteína, de fato, sintetizada, conforme mostra Alberts *et al.* (2017). A Figura 8 apresenta as diferenças estruturais entre o RNA e o DNA.

2.1.3 A síntese proteica e o código genético

De acordo com Alberts *et al.* (2017), as proteínas são importantes para diversas funções do nosso organismo e a sequência de nucleotídeos que forma uma molécula de DNA representa a informação necessária à produção de todas elas. A síntese proteica pode ser dividida em dois passos principais, a transcrição, que consiste na conversão de DNA em RNA, e a tradução, que consiste na conversão dos nucleotídeos em aminoácidos. O DNA é o responsável por determinar a síntese de proteínas, pois contém as informações que comandam a síntese de RNA, no processo de transcrição.

O RNA se desloca para o citoplasma das células, onde coordena a produção de proteínas por meio do processo de tradução ou síntese proteica. Esse processo, realizado pelos ribossomos, consiste em "decodificar"um RNA mensageiro e conectar os aminoácidos correspondentes a três sequências de bases nitrogenadas. Essas trincas de bases nitrogena-

Figura 8 – Diferenças estruturais entre o DNA e o RNA.



Fonte: Hepp e Nonohay (2016).

das são denominadas códons, cada um codificando um aminoácido específico da proteína. Existem 64 possíveis trincas, agrupadas três-a-três, que correspondem a 20 aminoácidos, sendo que mais de um códon pode corresponder ao mesmo aminoácido. Os aminoácidos são, portanto, as unidades básicas que formam as proteínas. A correspondência de um determinado aminoácido com um códon se dá por meio do código genético, como apresentado na Figura 9.

Figura 9 – Código genético: símbolos, códons e respectivos aminoácidos associados.

O código genético					
1ª posição (extremidade 5′)	2ª posição				3ª posicão (extremidade 3′)
↓	U	C	A	G	↓
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	**PARADA**	**PARADA**	A
	Leu	Ser	**PARADA**	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Os aminoácidos e seus símbolos			Códons
A	Ala	Alanina	GCA GCC GCG GCU
C	Cys	Cisteína	UGC UGU
D	Asp	Ácido aspártico	GAC GAU
E	Glu	Ácido glutâmico	GAA GAG
F	Phe	Fenilalanina	UUC UUU
G	Gly	Glicina	GGA GGC GGG GGU
H	His	Histidina	CAC CAU
I	Ile	Isoleucina	AUA AUC AUU
K	Lys	Lisina	AAA AAG
L	Leu	Leucina	UUA UUG CUA CUC CUG CUU
M	Met	Metionina	AUG
N	Asn	Aspargina	AAC AAU
P	Pro	Prolina	CCA CCC CCG CCU
Q	Gln	Glutamina	CAA CAG
R	Arg	Arginina	AGA AGG CGA CGC CGG CGU
S	Ser	Serina	AGC AGU UCA UCC UCG UCU
T	Thr	Treonina	ACA ACC ACG ACU
V	Val	Valina	GUA GUC GUG GUU
W	Trp	Triptofano	UGG
Y	Tyr	Tirosina	UAC UAU

Fonte: Alberts *et al.* (2017).

Os códons de parada (stop), UAA, UAG e UGA, são utilizados para mostrar a interrupção da síntese de uma proteína e não determinam nenhum aminoácido. Além disso, os códons que determinam um aminoácido são quimicamente semelhantes, distinguindo-se

somente em relação a uma base nitrogenada e, assim, faz com que as células se tornem mais resistentes a mutações, pois o aminoácido codificado é o mesmo, ainda que ocorra a troca da terceira base do códon.

2.2 ÁLGEBRA

Nesta seção serão apresentados os principais elementos algébricos utilizados neste trabalho. As referências foram: Fernandes e Oliveira (2021), Gerônimo e Franco (2008), Hage e Harary (2002), Jimenez-Montano, La Mora-Basanez e Poeschel (2002), Montano, La Mora-Basanez e Poeschel (1996) e Sánchez, Morgado e Grau (2004).

2.2.1 Reticulados Booleanos

Definição 1. Uma relação não-vazia R sobre um conjunto A não-vazio será chamada relação de ordem sobre A, se R é reflexiva, anti-simétrica e transitiva.

a) Reflexiva: $x \preceq x \ \forall \ x \in A$.

b) Anti-simétrica: $x \preceq y$ e $y \preceq x \Rightarrow x = y$.

c) Transitiva: $x \preceq y$ e $y \preceq z \Rightarrow x \preceq z$.

Um conjunto ordenado, é um conjunto sobre o qual se definiu uma relação de ordem. Além disso, se, para quaisquer x, $y \in A$, tivermos xRy ou yRx, a relação será chamada relação de ordem total sobre A. O conjunto A, nesse caso, será chamado conjunto totalmente ordenado.

Quando a está em relação com b, escreveremos $a \preceq b$ e diremos que "a é menor ou igual do que b na relação R". Quando a está em relação com b e, além disso, a é diferente de b, escreveremos $a \prec b$ e diremos que "a é menor do que b".

Definição 2. Uma relação $\preceq$ em A diz-se de ordem parcial (ROP) se satisfaz as propriedades:

a) Reflexiva: $x \preceq x \ \forall \ x \in A$.

b) Transitiva: $x \preceq y$ e $y \preceq z \Rightarrow x \preceq z$.

c) Anti-simétrica: $x \preceq y$ e $y \preceq x \Rightarrow x = y$.

Definição 3. $(A, \preceq)$ é um conjunto parcialmente ordenado (CPO) se $\preceq$ for uma relação de ordem parcial definida num conjunto não-vazio A.

Definição 4. Um reticulado é um conjunto parcialmente ordenado no qual todo par de elementos do conjunto possui simultaneamente soma e produto.

2.2.1.1 Reticulado como Álgebra: Conectivos $\wedge$ e $\vee$

Considere os valores verdade F e V (Falso e Verdadeiro), bem como os conectivos lógicos $\wedge$ e $\vee$. Se considerarmos 0 e 1 ao invés de F e V, para os conectivos lógicos temos que a estrutura $(\{F, V\}, \wedge, \vee)$ constitui um reticulado, ou seja, a conjunção e disjunção (Tabela 1) levam a um reticulado booleano atendendo as suas propriedades. Desse modo, utilizando o estudo da lógica, pode-se verificar que as estruturas $(\{F, V\}, \wedge)$ e $(\{F, V\}, \vee)$ são semigrupos abelianos (pois satisfazem as propriedades associativa, comutativa e as operações são fechadas) e, que os conectivos $\wedge$ e $\vee$ satisfazem à propriedade de absorção, ou seja, $a \vee (a \wedge b) = a$ e $a \wedge (a \vee b) = a$.

Tabela 1 – Operação com os conectivos $\wedge$ (conjunção) e $\vee$ (disjunção).

A	B	$A \wedge B$	$A \vee B$
1	1	1	1
1	0	0	1
0	1	0	1
0	0	0	0

Fonte: Fernandes e Oliveira (2021).

Definição 5. Um reticulado booleano $B(X)$ é um conjunto parcialmente ordenado de elementos com as seguintes propriedades:

a) $B(X)$ contém dois elementos chamados de elementos mínimo e máximo, denotados por 0 e 1, respectivamente, que são limites universais, ou seja: $0 \leq \alpha \leq 1$ para todos os $\alpha \in X$ e satisfaz as propriedades especiais:

   - Interseção: $0 \wedge \alpha = 0$ e $1 \wedge \alpha = \alpha$.
   - União: $0 \vee \alpha = \alpha$ e $1 \vee \alpha = 1$.

b) Para todos os elementos $\alpha \in X$ existe o elemento $\neg\alpha \in X$, chamado complemento do elemento α, de tal forma que:

$$\alpha \wedge \neg\alpha = 0 \text{ e } \alpha \vee \neg\alpha = 1.$$

c) Em $B(X)$, as operações $\wedge$ e $\vee$ satisfazem a lei distributiva, isto é:

$$(\alpha \wedge \beta) \vee (\alpha \wedge \delta) = \alpha \wedge (\beta \vee \delta) \text{ e } (\alpha \vee \beta) \wedge (\alpha \wedge \delta) = \alpha \wedge (\beta \wedge \delta).$$

2.2.2 Diagrama de Hasse

Definição 6. Quando um conjunto possui um número finito de elementos, a relação de ordem possui uma representação gráfica adequada para as suas propriedades. Essa representação é denominada "diagrama de Hasse" ou "diagrama de linha". Esse tipo de representação torna mais evidente o comportamento dos elementos do conjunto dado pela relação.

Considere uma relação de ordem R sobre um conjunto A, a construção se faz da seguinte maneira:

a) Cada elemento do conjunto A é representado por um ponto;

b) Se um elemento x estiver relacionado com um elemento y, representaremos isso por um segmento de reta unindo ambos;

c) A propriedade reflexiva é omitida na representação, ou seja, não colocaremos os laços em torno de cada elemento de A;

d) A propriedade transitiva fica subentendida na representação, ou seja, xRy e yRz, temos xRz, mas o segmento indicativo de xRz não é colocado;

e) A representação será orientada de baixo para cima, ou seja, se xRy, então o elemento x será colocado em uma posição abaixo do elemento y.

Exemplo 1. Considere o conjunto $A = \{a, b, c\}$. O diagrama de Hasse desse conjunto mediante a relação de inclusão é representado conforme a Figura **??**:

2.2.3 Distância de Hamming

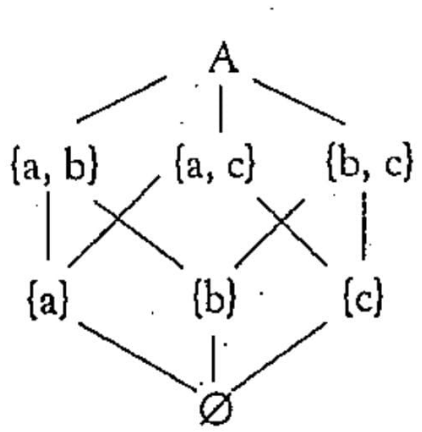
Definição 7. Dadas duas sequências $x = (x_1, x_2, \ldots, x_n)$ e $y = (y_1, y_2, \ldots, y_n)$, onde $x_i, y_i \in \Sigma$ (um alfabeto finito), a distância de Hamming, denotada por $d_H(x, y)$, é definida como:

$$d_H(x, y) = \sum_{i=1}^{n} \delta(x_i, y_i)$$

onde $\delta(x_i, y_i)$ é a função delta de Kronecker, definida como:

$$\delta(x_i, y_i) = \begin{cases} 0 & \text{se } x_i = y_i, \\ 1 & \text{se } x_i \neq y_i. \end{cases}$$

Figura 10 – Diagrama de Hasse.



Fonte: Geronimo e Franco (2008).

Portanto, a distância de Hamming $d_H(x, y)$ é o número de posições em que x e y diferem. Para o caso binário, a distância de Hamming pode ser determinada facilmente pela propriedade de adição módulo-2, pois ela é igual ao número de dígitos "1" contidos no vetor resultante da operação $x \oplus y$.

Exemplo 2. Considere os vetores $x = 10110$ e $y = 10101$. A distância e o peso de Hamming é igual a 2, uma vez que:

$$d_H(x, y) = \omega(x \oplus y) = \omega(10110 \oplus 10101) = \omega(00011)$$
$$d_H(x, y) = 2.$$

2.2.4 Código de Gray

Conforme apresentado em Jimenez-Montano, La Mora-Basanez e Poeschrl (2002), o código de Gray é uma forma de codificação binária em que dois números consecutivos diferem em apenas um bit, reduzindo erros em transições de estado, frequentemente usado em circuitos digitais e sistemas de comunicação.

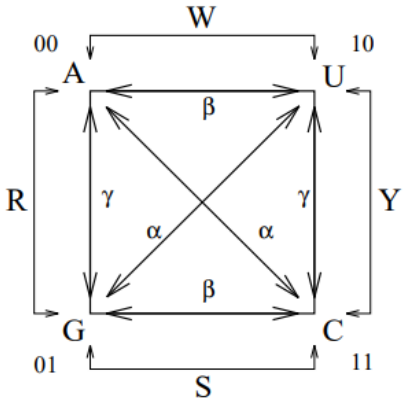
As quatro bases que ocorrem nas macromoléculas de DNA (RNA) definem o alfabeto X correspondente: $\{A, C, G, T\}$ ou $\{A, C, G, U\}$. Cada base é completamente especificada por duas categorizações dicotômicas independentes:

1. de acordo com o tipo químico C : $\{R, Y\}$, onde R : (A, G) são purinas e Y : (C, U) são pirimidinas;

2. de acordo com ligação H : $\{W, S\}$, onde $W : (A, U)$ são bases fracas e $S : (C, G)$ bases fortes.

A Figura 11 mostra essas categorizações das bases, representadas pelos nós, que são pontos de representação no sistema de categorização que organiza as bases em termos de suas propriedades. O primeiro atributo é o caráter químico e o segundo o caráter da ligação de hidrogênio. Estendendo essa associação para triplas de base, cada códon é de uma forma única associado a uma palavra-código que consiste em seis valores de atributo, conforme Figura 12.

Figura 11 – Categorizações das bases. As categorizações das bases são de acordo com o tipo químico C e a ligação H. A representação binária das bases também é mostrada. O primeiro bit é o tipo químico e o segundo o caráter de ligação H. α, β e γ são as transformações das bases.



Fonte: Jimenez-Montano, La Mora-Basanez e Poeschel (2002).

2.2.5 Hipercubo Booleano

De acordo com Jimenez-Montano, La Mora-Basanez e Poeschrl (2002), um hipercubo n-dimensional, denotado por Q_n, consiste em 2^n nós (vértices), cada um endereçado por um número de identificação único de n bits. Existe uma ligação entre dois nós de Q_n se, e somente se, seus endereços de nó diferem em exatamente uma posição de bit. Dois nós em um hipercubo são considerados adjacentes se houver uma ligação presente entre eles.

Nesse sentido, a distância de Hamming entre quaisquer dois nós de cubo é o número de bits que diferem em seus endereços, que são representações binárias dos nós dentro de uma estrutura de cubo, em que cada nó é identificado por uma sequência de bits. O número de saltos necessários para alcançar um nó de outro nó é igual à distância entre os dois nós.

Figura 12 – Código de Gray. Primeira e quarta colunas são mostrados os vetores 6–binários (palavras-código). Segunda e quinta colunas aparecem os códons correspondentes. Finalmente, na terceira e sexta colunas os aminoácidos em notação de letra única. Os dois primeiros dígitos correspondem à primeira base, os dois seguintes à segunda base e os dois últimos à última base, de acordo com a codificação binária das bases da Figura 11.

0	0	0	0	1	1	A	A	C	N
0	0	0	0	1	0	A	A	U	N
0	0	0	0	0	0	A	A	A	K
0	0	0	0	0	1	A	A	G	K
1	0	0	0	0	1	U	A	G	t
1	0	0	0	0	0	U	A	A	t
1	0	0	0	1	0	U	A	U	Y
1	0	0	0	1	1	U	A	C	Y
1	1	0	0	1	1	C	A	C	H
1	1	0	0	1	0	C	A	U	H
1	1	0	0	0	0	C	A	A	Q
1	1	0	0	0	1	C	A	G	Q
0	1	0	0	0	1	G	A	G	E
0	1	0	0	0	0	G	A	A	E
0	1	0	0	1	0	G	A	U	D
0	1	0	0	1	1	G	A	C	D
0	1	1	0	1	1	G	U	C	V
0	1	1	0	1	0	G	U	U	V
0	1	1	0	0	0	G	U	A	V
0	1	1	0	0	1	G	U	G	V
1	1	1	0	0	1	C	U	G	L
1	1	1	0	0	0	C	U	A	L
1	1	1	0	1	0	C	U	U	L
1	1	1	0	1	1	C	U	C	L
1	0	1	0	1	1	U	U	C	F
1	0	1	0	1	0	U	U	U	F
1	0	1	0	0	0	U	U	A	L
1	0	1	0	0	1	U	U	C	L
0	0	1	0	0	1	A	U	G	M
0	0	1	0	0	0	A	U	A	I
0	0	1	0	1	0	A	U	U	I
0	0	1	0	1	1	A	U	C	I

0	0	1	1	1	1	A	C	C	T
0	0	1	1	1	0	A	C	U	T
0	0	1	1	0	0	A	C	A	T
0	0	1	1	0	1	A	C	G	T
1	0	1	1	0	1	U	C	G	S
1	0	1	1	0	0	U	C	A	S
1	0	1	1	1	0	U	C	U	S
1	0	1	1	1	1	U	C	C	S
1	1	1	1	1	1	C	C	C	P
1	1	1	1	1	0	C	C	U	P
1	1	1	1	0	0	C	C	A	P
1	1	1	1	0	1	C	C	G	P
0	1	1	1	0	1	G	C	G	A
0	1	1	1	0	0	G	C	A	A
0	1	1	1	1	0	G	C	U	A
0	1	1	1	1	1	G	C	C	A
0	1	0	1	1	1	G	G	C	G
0	1	0	1	1	0	G	G	U	G
0	1	0	1	0	0	G	G	A	G
0	1	0	1	0	1	G	G	G	G
1	1	0	1	0	1	C	G	G	R
1	1	0	1	0	0	C	G	A	R
1	1	0	1	1	0	C	G	U	R
1	1	0	1	1	1	C	G	C	R
1	0	0	1	1	1	U	G	C	C
1	0	0	1	1	0	U	G	U	C
1	0	0	1	0	0	U	G	A	t
1	0	0	1	0	1	U	G	C	W
0	0	0	1	0	1	A	G	G	R
0	0	0	1	0	0	A	G	A	R
0	0	0	1	1	0	A	G	U	S
0	0	0	1	1	1	A	G	C	S

Fonte: Jimenez-Montano, La Mora-Basanez e Poeschel (2002).

No capítulo 3 deste trabalho serão apresentadas algumas construções envolvendo hipercubos booleanos e serão elencadas mais informações acerca desta importante estrutura matemática.

3 CONSTRUÇÃO E ANÁLISE DE HIPERCUBOS BOOLEANOS ASSOCIADOS AO CÓDIGO GENÉTICO

Neste capítulo serão apresentados os primeiros resultados deste trabalho, por meio da construção dos hipercubos booleanos associados a cada um dos rotulamentos do código genético. A construção do hipercubo booleano toma como referência inicial os reticulados booleanos, os diagramas de Hasse (detalhados em Fernandes e Oliveira (2021)), além das tabelas do código de Gray. Ressalta-se que os resultados apresentados neste capítulo, até onde é de nosso conhecimento representa uma contribuição inédita no processo de modelagem e análise de propriedades associadas ao estudo de propriedades do código genético.

3.1 RESULTADOS OBTIDOS PARA O ROTULAMENTO A

Nesta seção serão apresentados os resultados obtidos utilizando o rotulamento A do código genético, por meio da permutação 1203, para o reticulado primal e, como consequência, a permutação 3021 do rotulamento A foi utilizada para a construção do reticulado dual.

A atribuição $\{0, 1, 2, 3\}$, referente ao anel $\mathbb{Z}_4$, é feita em relação à ordem {A, C, G, U} em N, ou seja, as bases são estabelecidas na seguinte ordem: adenina (A), citosina (C), guanina (G) e por fim, uracila (U).

As Tabelas 2 e 3 apresentam as associações estabelecidas pelas permutações 1203 e 3021, respectivamente primal e dual, referentes ao rotulamento A.

Tabela 2 – Associação estabelecida pela permutação 1203 primal do rotulamento A.

A	**C**	**G**	**U**
1	2	0	3
10	11	00	01

Fonte: do autor.

Tabela 3 – Associação estabelecida pela permutação 3021 dual do rotulamento A.

A	**C**	**G**	**U**
3	0	2	1
01	00	11	10

Fonte: do autor.

Inicialmente, serão apresentadas as tabelas das operações primal e dual para as permutações escolhidas do rotulamento A, utilizando para isso as operações de disjunção e conjunção, da lógica booleana clássica, tanto para o caso de $\mathbb{Z}_2 \times \mathbb{Z}_2$, quanto para o caso do alfabeto biológico N. A partir desses resultados, serão construídos os reticulados booleanos, primal e dual, a fim de estabelecer uma relação entre as bases nitrogenadas e a modelagem matemática associada.

O procedimento a ser utilizado para o caso do rotulamento A será seguido de forma análoga para os rotulamentos B e C, apresentados nas seções seguintes. Este procedimento pode ser encontrado de forma detalhada em Fernandes e Oliveira (2021).

As operações mencionadas anteriormente são apresentadas nas Tabelas 4, 5, 6 e 7.

Tabela 4 – Operação primal $\wedge$ (e) do rotulamento A.

$\wedge$	**00**	**01**	**10**	**11**	$\wedge$	**G**	**U**	**A**	**C**
00	00	00	00	00	**G**	G	G	G	G
01	00	01	00	01	**U**	G	U	G	U
10	00	00	10	10	**A**	G	G	A	A
11	00	01	10	11	**C**	G	U	A	C

Fonte: do autor.

Tabela 5 – Operação primal $\vee$ (ou) do rotulamento A.

$\vee$	**00**	**01**	**10**	**11**	$\vee$	**G**	**U**	**A**	**C**
00	00	01	10	11	**G**	G	U	A	C
01	01	01	11	11	**U**	U	U	C	C
10	10	11	10	11	**A**	A	C	A	C
11	11	11	11	11	**C**	C	C	C	C

Fonte: do autor.

Tabela 6 – Operação dual $\wedge$ (e) do rotulamento A.

$\wedge$	**00**	**01**	**10**	**11**	$\wedge$	**C**	**A**	**U**	**G**
00	00	00	00	00	**C**	C	C	C	C
01	00	01	00	01	**A**	C	A	C	A
10	00	00	10	10	**U**	C	C	U	U
11	00	01	10	11	**G**	C	A	U	G

Fonte: do autor.

O reticulado booleano primal será construído a partir da tabela referente à operação primal, por meio do seguinte procedimento:

a) A base G (guanina) se liga à ela mesma e às bases U (uracila), A (adenina) e C (citosina);

Tabela 7 – Operação dual $\vee$ (ou) do rotulamento A.

$\vee$	**00**	**01**	**10**	**11**	$\vee$	**C**	**A**	**U**	**G**
00	00	01	10	11	**C**	C	A	U	G
01	01	01	11	11	**A**	A	A	G	G
10	10	11	10	11	**U**	U	G	U	G
11	11	11	11	11	**G**	G	G	G	G

Fonte: do autor.

b) A base U (uracila) se liga à ela mesma e à base C (citosina);

c) A base A (adenina) se liga à ela mesma e à base C (citosina);

d) A base C (citosina) se liga à ela mesma.

O reticulado booleano dual será construído a partir da tabela referente à operação dual, cujo procedimento é apresentado a seguir:

a) A base C (citosina) se liga à ela mesma e às bases A (adenina), U (uracila) e G (guanina);

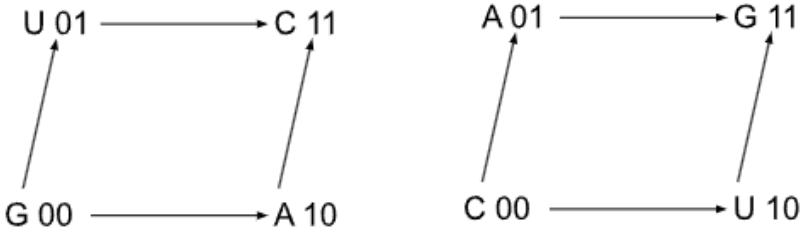
b) A base A (adenina) se liga à ela mesma e à base G (guanina);

c) A base U (uracila) se liga à ela mesma e à base G (guanina);

d) A base G (guanina) se liga à ela mesma.

A Figura 13 apresenta os reticulados booleanos primal e dual relacionados às permutações utilizadas para o rotulamento A.

Figura 13 – Reticulados booleanos primal (1203) e dual (3021) associados ao rotulamento A.



Fonte: Adaptado de Fernandes e Oliveira (2021).

A seguir será apresentada a construção do diagrama de Hasse, tanto para a permutação do caso primal, quanto para a do caso dual. O detalhamento para essas construções podem ser encontrados em Fernandes e Oliveira (2021).

3.1.1 Diagramas de Hasse associados ao rotulamento A

O procedimento apresentado a seguir foi proposto em Fernandes e Oliveira (2021).

a) 1ª linha: elemento máximo, atribuído por 11.

$$\text{CCC} \leftrightarrow 111111$$

b) 2ª linha: um elemento 0 e cinco elementos 1 (seis códons).

CCU $\leftrightarrow$ 111101	CUC $\leftrightarrow$ 110111	UCC $\leftrightarrow$ 011111
ACC $\leftrightarrow$ 101111	CAC $\leftrightarrow$ 111011	CCA $\leftrightarrow$ 111110

c) 3ª linha: dois elementos 0 e quatro elementos 1 (quinze códons).

CUU $\leftrightarrow$ 110101	CAU $\leftrightarrow$ 111001	UCU $\leftrightarrow$ 011101
ACU $\leftrightarrow$ 101101	UUC $\leftrightarrow$ 010111	AUC $\leftrightarrow$ 100111
CGC $\leftrightarrow$ 110011	AAC $\leftrightarrow$ 101011	*GCC* $\leftrightarrow$ 001111
UAC $\leftrightarrow$ 011011	CCG $\leftrightarrow$ 111100	UCA $\leftrightarrow$ 011110
ACA $\leftrightarrow$ 101110	CUA $\leftrightarrow$ 110110	CAA $\leftrightarrow$ 111010

d) 4ª linha: três elementos 0 e três elementos 1 (vinte códons).

UUU $\leftrightarrow$ 010101	AUU $\leftrightarrow$ 100101	UAU $\leftrightarrow$ 011001	AAU $\leftrightarrow$ 101001
GCU $\leftrightarrow$ 001101	CGC $\leftrightarrow$ 110001	AGC $\leftrightarrow$ 100011	GUC $\leftrightarrow$ 000111
UGC $\leftrightarrow$ 010011	GAC $\leftrightarrow$ 001011	CUG $\leftrightarrow$ 110100	ACG $\leftrightarrow$ 101100
CAG $\leftrightarrow$ 111000	UCG $\leftrightarrow$ 011100	GCA $\leftrightarrow$ 001110	CGA $\leftrightarrow$ 110010
UUA $\leftrightarrow$ 010110	AUA $\leftrightarrow$ 100110	UAA $\leftrightarrow$ 011010	AAA $\leftrightarrow$ 101010

e) 5ª linha: quatro elementos 0 e dois elementos 1 (quinze códons).

f) 6ª linha: cinco elementos 0 e um elemento 1 (seis códons).

g) 7ª linha: elemento mínimo, atribuído por 00.

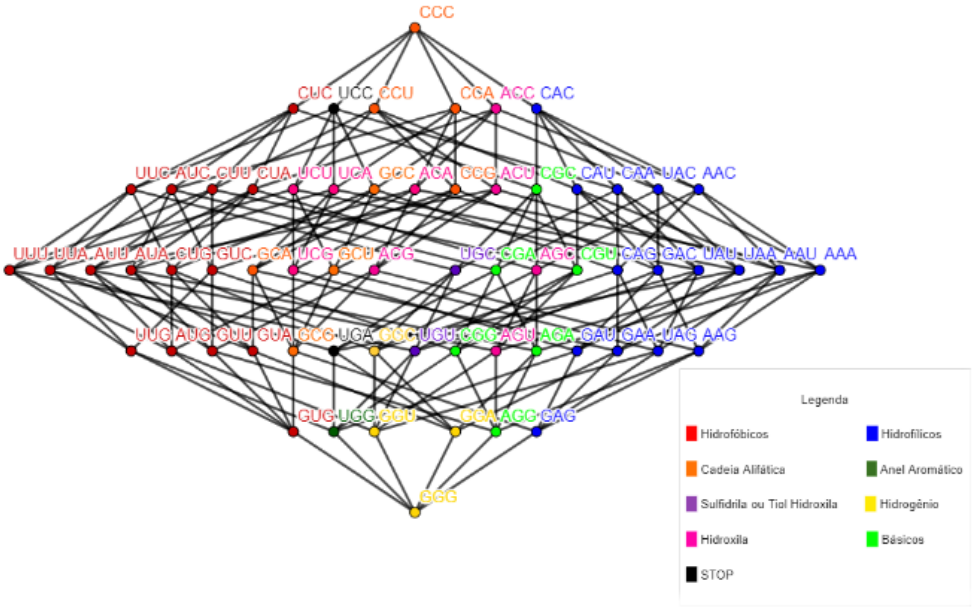
GUU ↔ 000101	GAU ↔ 001001	UGU ↔ 010001
AGU ↔ 100001	GGC ↔ 000011	AUG ↔ 100100
CGG ↔ 110000	UUG ↔ 010100	GCG ↔ 001100
UAG ↔ 011000	AAG ↔ 101000	UGA ↔ 010010
AGA ↔ 100010	GUA ↔ 000110	GAA ↔ 001010

GGU ↔ 000001	GUG ↔ 000100	UGG ↔ 010000
AGG ↔ 100000	GAG ↔ 001000	GGA ↔ 000010

GGG ↔ 000000

Os diagramas de Hasse, primal e dual, referentes às permutações 1203 e 3021, respectivamente, são apresentados nas Figuras 14 e 15. Nas legendas são apresentadas as características de cada um dos códons, por meio das cores escolhidas.

Figura 14 – Diagrama de Hasse primal referente à permutação 1203 do rotulamento A.

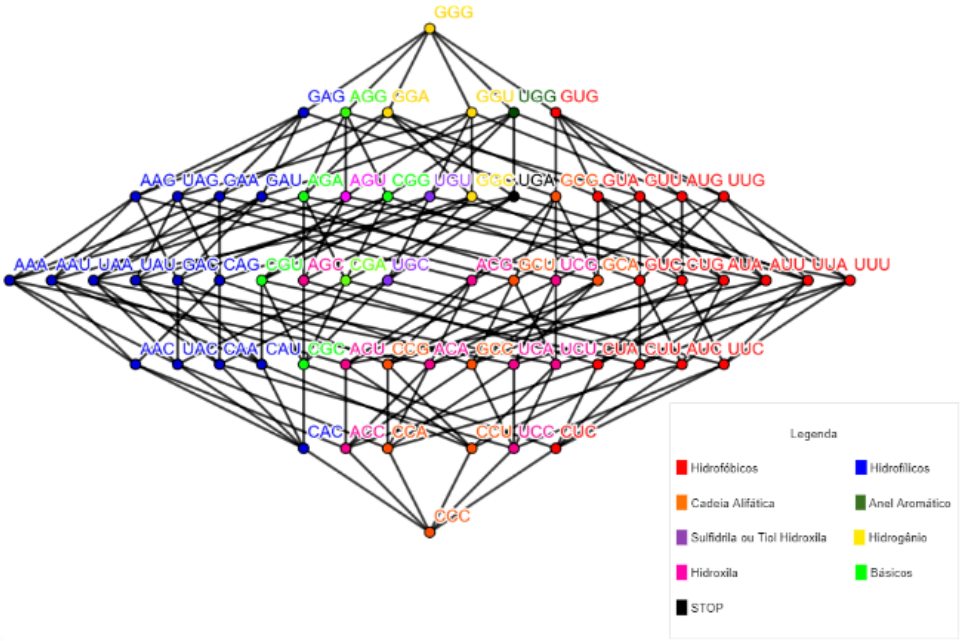


Fonte: adaptado de Fernandes e Oliveira (2021).

Pode-se perceber nas Tabelas 14 e 15 um comportamento muito parecido com os diagramas de Hasse apresentados, uma vez que as cores associadas a cada um dos códons reflete a característica físico-química de cada um dos aminoácidos do código genético.

A seguir serão apresentadas as tabelas do código de Gray referentes às permutações selecionadas do rotulamento A. Essa construção tomou como referência o trabalho de Jimenez-Montano, La Mora-Basanez e Poeschel (2002), mas já representa uma contribuição inédita deste trabalho, dadas as características apresentadas na construção e análises advindas.

Figura 15 – Diagrama de Hasse dual referente à permutação 3021 do rotulamento A.



Fonte: adaptado de Fernandes e Oliveira (2021).

3.1.2 Código de Gray - Rotulamento A

A Tabela 8 apresenta o código de Gray associado à permutação 1203 (primal) do rotulamento A do código genético, na qual os dígitos 00 correspondem à base G, os dígitos 11 correspondem à base C, os dígitos 10 correspondem à base A e os dígitos 01 correspondem à base U.

A Tabela 9 apresenta o código de Gray associado à permutação 3021 (dual) do rotulamento A do código genético, na qual os dígitos 00 correspondem à base C, os dígitos 11 correspondem à base G, os dígitos 10 correspondem à base U e os dígitos 01 correspondem à base A.

A seguir serão apresentadas as construções dos hipercubos booleanos relacionados às permutações escolhidas do rotulamento A, além das conexões dessas estruturas com as tabelas do código de Gray e os diagramas de Hasse construídos anteriormente. Inicialmente será apresentado um detalhamento da estrutura do hipercubo desde a dimensão 1 até a dimensão 6 (foco deste estudo), relacionado ao código genético.

3.1.3 Construção e análise dos hipercubos booleanos do código genético

Nas construções dos hipercubos apresentadas a seguir, todas as ligações entre os vértices são feitas a partir da diferença de um bit de um vértice para o outro e a referência

Tabela 8 – Código de Gray associado à permutação 1203 do código genético.

Binário	Códon	Aminoácido	Binário	Códon	Amioácido
000011	GGC	glicina	000010	GGA	glicina
000000	GGG	glicina	000001	GGU	glicina
100001	AGU	serina	100000	AGG	arginina
100010	AGA	arginina	100011	AGC	serina
110011	CGC	arginina	110010	CGA	arginina
110000	CGG	arginina	110001	CGU	arginina
010001	UGU	cisteína	010000	UGG	triptofano
010010	UGA	STOP	010011	UGC	cisteína
011011	UAC	tirosina	011010	UAA	STOP
011000	UAG	STOP	011001	UAU	tirosina
111001	CAU	histidina	111000	CAG	glutamina
111010	CAA	glutamina	111011	CAC	histidina
101011	AAC	asparagina	101010	AAA	lisina
101000	AAG	lisina	101001	AAU	asparagina
001001	AUG	aspartato	001000	GAU	glutamato
001010	GAG	glutamato	001011	GAA	aspartato
001111	GCC	alanina	001110	GCA	alanina
001100	GCG	alanina	001101	GCU	alanina
101101	ACU	treoina	101100	ACG	treoina
101110	ACA	treoina	101111	ACC	treoina
111111	CCC	prolina	111110	CCA	prolina
111100	CCG	prolina	111101	CCU	prolina
011101	UCU	serina	011100	UCG	serina
011110	UCA	serina	011111	UCC	serina
010111	UUC	leucina	010110	UUA	fenilalanina
010100	UUG	leucina	010101	UUU	fenilalanina
110101	CUU	isoleucina	110100	CUG	metionina
110110	CUC	isoleucina	110111	CGC	isoleucina
100111	AUC	leucina	100110	AUA	leucina
100100	AUG	leucina	100101	AUU	leucina
000101	GUU	valina	000100	GUG	valina
000110	GUA	valina	000111	GUC	valina

Fonte: do autor.

utilizada foi Hage e Harary (2002).

a) No caso da dimensão 1, estão representadas as combinações do alfabeto $\mathbb{Z}_2 = \{0, 1\}$, todas as combinações possíveis de um bit, conforme apresentado na Figura 16.

b) No caso da dimensão 2, a representação geométrica é um quadrado, em que estão representadas as combinações do $\mathbb{Z}_2 \times \mathbb{Z}_2$, ou seja, todas as combinações possíveis de dois bits, conforme apresentado na Figura 17.

c) No caso da dimensão 3, nota-se que a representação geométrica é um cubo, em que cada um dos vértices estão representados pelas combinações do $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$, ou

Tabela 9 – Código de Gray associado à permutação 3021 do código genético.

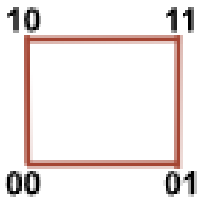
Binário	Códon	Amioácido	Binário	Códon	Aminoácido
000011	CCG	prolina	000010	CCU	prolina
000000	CCC	prolina	000001	CCA	prolina
100001	UCA	serina	100000	UCC	serina
100010	UCU	serina	100011	UCG	serina
110011	GCG	alanina	110010	GCU	alanina
110000	GCC	alanina	110001	GCA	alanina
010001	ACA	treonina	010000	ACC	treonina
010010	ACU	treonina	010011	ACG	treonina
011011	UAC	leucina	011010	UAA	leucina
011000	UAG	leucina	011001	UAU	leucina
111001	AUG	valina	111000	AUU	valina
111010	AUC	valina	111011	AUA	valina
101011	UUG	leucina	101010	UUU	fenilalanina
101000	UUC	leucina	101001	UUA	fenilalanina
001001	CUA	isoleucina	001000	CUC	isoleucina
001010	CUU	isoleucina	001011	CUG	isoleucina
001111	CGG	arginina	001110	CGU	arginina
001100	CGC	arginina	001101	CGA	arginina
101101	UGA	STOP	101100	UGC	cisteína
101110	UGU	cisteína	101111	UGG	triptofano
111111	GGG	glicina	111110	GGU	glicina
111100	GGC	glicina	111101	GGA	glicina
011101	AGA	arginina	011100	AGC	serina
011110	AGU	serina	011111	AGG	arginina
010111	AAG	lisina	010110	AAU	arparagina
010100	AAC	arparagina	010101	AAA	lisina
110101	GAA	glutamato	110100	GAC	aspartato
110110	GAU	aspartato	110111	GAG	glutamato
100111	UAG	STOP	100110	UAU	tirosina
100100	UAC	tirosina	100101	UAA	STOP
000101	CAA	glutamina	000100	CAC	histidina
000110	CAU	histidina	000111	CAG	glutamina

Fonte: do autor.

Figura 16 – Dimensão 1.

Fonte: Adaptado de Jimenez-Montano, La Mora-Basanez e Poeschel (2002).

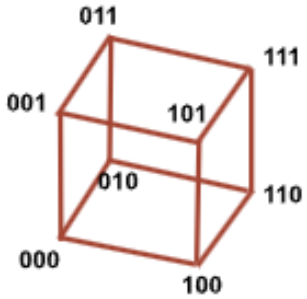
Figura 17 – Dimensão 2.



Fonte: Adaptado de Jimenez-Montano, La Mora-Basanez e Poeschel (2002).

seja, todas as combinações possíveis de três bits, conforme apresentado na Figura 18.

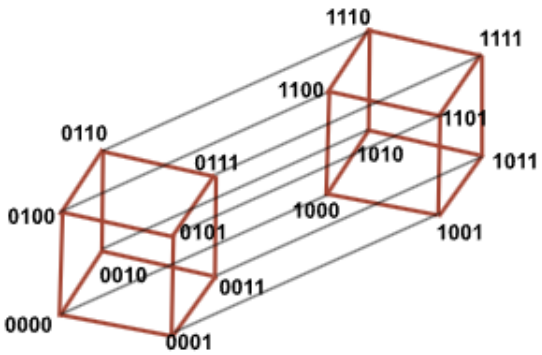
Figura 18 – Dimensão 3.



Fonte: Adaptado de Jimenez-Montano, La Mora-Basanez e Poeschel (2002).

d) No caso da dimensão 4, a representação geométrica é feita por dois cubos, em que cada um dos vértices desses cubos estão representados pelas combinações do $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$, ou seja, todas as combinações possíveis de quatro bits, conforme Figura 19.

Figura 19 – Dimensão 4.

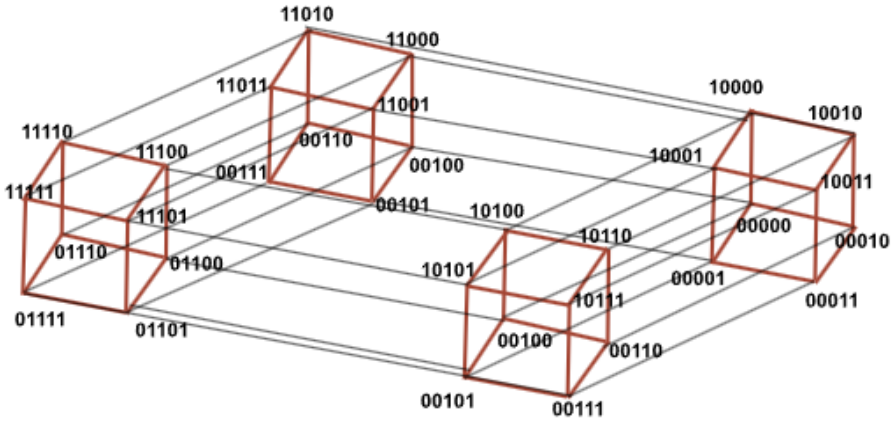


Fonte: Adaptado de Jimenez-Montano, La Mora-Basanez e Poeschel (2002).

e) No caso da dimensão 5, a representação geométrica é dada por quatro cubos, em que cada um dos vértices desses cubos estão representados pelas combinações do

$\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$, ou seja, todas as combinações possíveis de cinco bits, conforme apresentado na Figura 20.

Figura 20 – Dimensão 5.



Fonte: Adaptado de Jimenez-Montano, La Mora-Basanez e Poeschel (2002).

A partir das construções anteriores, será apresentada a representação disponível em Jimenez-Montano, La Mora-Basanez e Poeschel (2002), onde é feita uma associação do hipercubo booleano 6-dimensional com uma representação do código genético, conforme apresentado na Figura 21.

São representados oito cubos, onde cada um dos vértices desses cubos estão representados por $\mathbb{Z}_2^6$, ou seja, $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$, sendo todas as combinações possíveis de seis bits.

Esses elementos não estão representados de maneira aleatória, a organização se dá de modo que a ligação entre os vértices tenha a diferença de um bit.

Em Jimenez-Montano, La Mora-Basanez e Poeschel (2002), é feita a proposta da construção do hipercubo booleano 6-dimensional, associado ao código genético, cujo procedimento, tomando como referência as tabelas do código de Gray é apresentado a seguir.

a) Nas duas primeiras linhas da tabela do código de Gray, cada códon representa o vértice de uma face do cubo. Assim, ligando-se o primeiro códon com o segundo, o segundo com o terceiro, o terceiro com o quarto e o quarto com o primeiro, tem-se a primeira face do cubo;

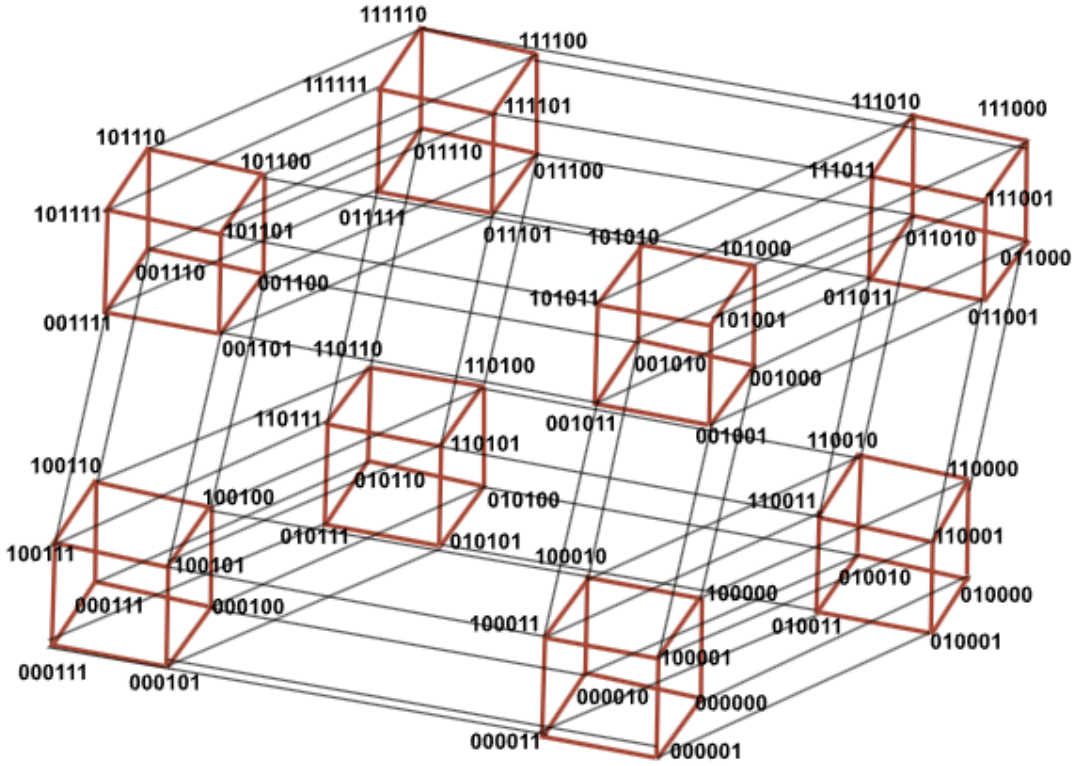
b) Na terceira e quarta linhas da tabela do código de Gray, cada códon representa o vértice de uma face do cubo. Assim, ligando-se o primeiro códon com o segundo, o segundo com o terceiro e o terceiro com o quarto e o quarto com o primeiro, tem-se a segunda face do cubo;

c) Construindo uma aresta que liga os vértices formados pela primeira face com os formados pela segunda face, obtém-se um cubo. Esses vértices devem ser ligados de forma que a diferença entre os códons que os representa seja de uma base;

d) Efetuando o mesmo processo de forma sucessiva para as próximas linhas, quatro a quatro, tem-se os 8 cubos que serão os "vértices"do hipercubo 6-dimensional;

e) Ligando-se esses "vértices", é formado o hipercubo 6-dimensional.

Figura 21 – Dimensão 6.



Fonte: Adaptado de Jimenez-Montado, La Mora-Basanez e Poeschel (2002).

A partir do procedimento para a construção do hipercubo booleano 6-dimensional, a associação com o código genético será feita utilizando, inicialmente, a permutação 1203 do rotulamento A para o caso primal e a permutação 3021 do rotulamento A para o caso dual. Ressalta-se que essas contruções representadas também, até onde é de nosso conhecimento, contribuições inéditas relacionadas ao processo de modelagem do código genético.

Outro ponto a ser destacado, é que pode-se notar que cada um dos elementos refere-se a um elemento da extensão de Galois $GF(2^6)$, outra importante conexão com a parte algébrica, que não faz parte do escopo deste trabalho, mas que poderão ser exploradas em pesquisas futuras.

Os hipercubos apresentados a seguir tomarão como referência os dados da Tabela 10 onde são apresentados a representação vetorial, os códons e os aminoácidos associados a cada códon.

Tabela 10 – Representações do código genético.

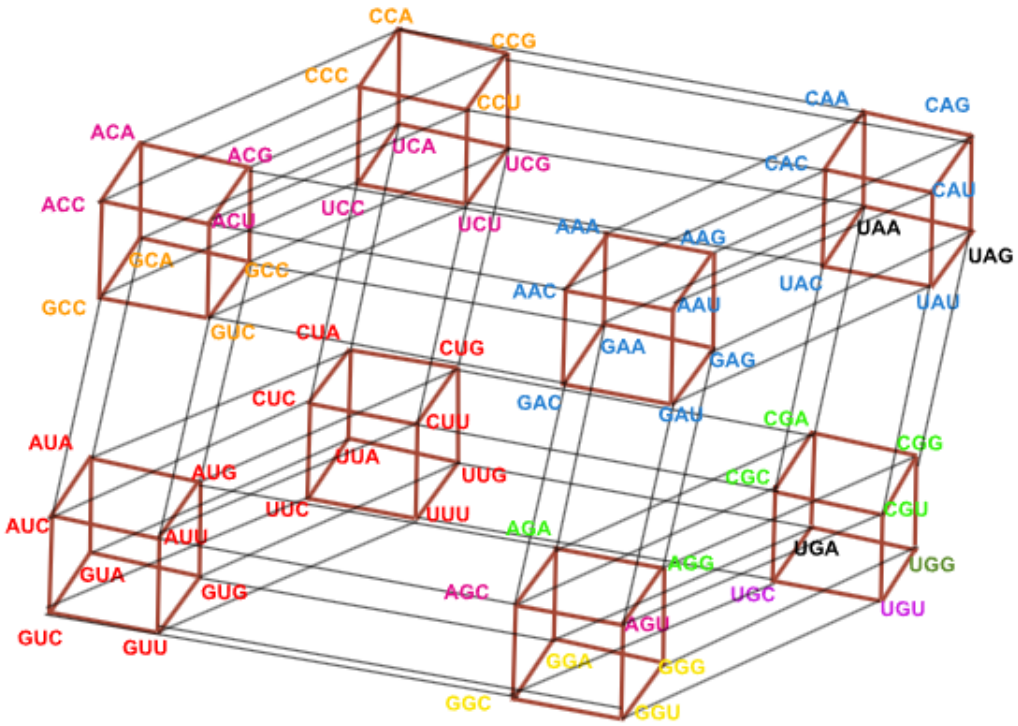
Vetorial	Códon	Aminoácido	Vetorial	Códon	Aminoácido
000000	GGG	glicina	100100	AUG	leucina
100000	AGG	arginina	101001	AAU	asparagina
010000	UGG	triptofano	010010	UGA	STOP
001000	GAG	glutamato	001001	GAU	aspartato
000100	GUG	valina	110100	CUG	leucina
000010	GGA	glicina	011010	UAA	STOP
000001	GGU	glicina	001101	GCU	alanina
110000	CGG	arginina	110110	CUA	leucina
011000	UAG	STOP	011011	UAC	tirosina
001100	GCG	alanina	111101	CCU	prolina
000110	GUA	valina	101110	ACA	serina
000011	GGC	glicina	010111	UUC	fenilalanina
110001	CGU	arginina	111011	CAC	histidina
101000	AAG	lisina	101101	ACU	serina
010100	UUG	leucina	100110	AUA	isoleucina
001010	GAA	glutamato	010011	UGC	cisteína
000101	GUU	valina	111001	CAU	histidina
110010	CGA	arginina	101100	ACG	serina
011001	UAU	tirosina	010110	UUA	leucina
111100	CCG	prolina	001011	GAC	aspartato
001110	GCA	alanina	110101	CUU	leucina
000111	GCC	alanina	100110	AUA	isoleucina
110111	CUC	isoleucina	010101	UUU	fenilalanina
101011	AAC	asparagina	111010	CAA	glutamina
100101	AUU	leucina	011101	UCU	serina
100010	AGA	arginina	111110	CCA	prolina
010001	UGU	cisteína	011111	UCC	serina
111000	CAG	glutamina	111111	CCC	prolina
011100	UCG	serina	101111	ACC	serina
001110	GCA	alanina	100111	AUC	isoleucina
000111	GUC	valina	100011	AGC	serina
110011	CGC	arginina	100001	AGU	serina

Fonte: do autor.

3.1.4 Hipercubos booleanos associados ao rotulamento A

Os hipercubos booleanos associados às permutações 1203 (primal) e 3021 (dual) do rotulamento A estão apresentados nas Figuras 22 e 23.

Figura 22 – Hipercubo booleano associado à permutação 1203 (primal) do rotulamento A.

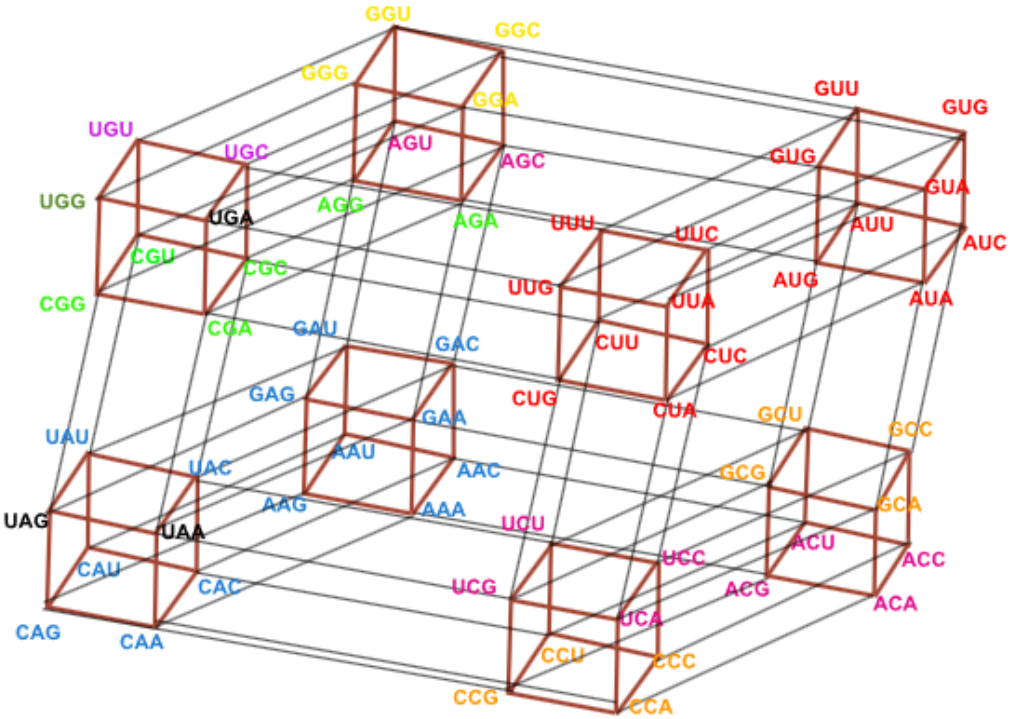


Fonte: do autor.

As cores são representadas de acordo com as propriedades dos aminoácidos. Desta forma, no hipercubo booleano primal pode-se observar que os códons hidrofóbicos (representados pela cor vermelha) são aqueles que não possuem "afinidade"com a água (códons com U na segunda posição) estão separados dos códons hidrofílicos (representados pela cor azul), os quais possuem "afinidade"com a água (códons com A na segunda posição). O códon que codifica um aminoácido hidrofóbico é sempre complementar a um códon que codifica um aminoácido hidrofílico, ou seja, observe por exemplo que (UGU, GUU, UUG, UUC, CUU, UCU) é a imagem de (AGA, GAA, AAG, AAC, CAA, ACA).

No hipercubo booleano dual, pode-se observar que os códons hidrofílicos (códons com A na segunda posição representados em azul e possuem "afinidade"com a água), além de separados, ficaram na parte inferior do hipercubo, com relação dos códons hidrofóbicos (códons com U na segunda posição estão representados em vermelho e não possuem "afinidade"com a água) que estão situados na parte superior do hipercubo. O códon que codifica um aminoácido hidrofílico é sempre complementar a um códon que codifica um aminoácido hidrofóbico, ou seja, observe por exemplo que (ACA, CAA, AAC, AAG, GAA,

Figura 23 – Hipercubo booleano associado à permutação 3021 (dual) do rotulamento A.



Fonte: do autor.

AGA) é a imagem de (UCU, CUU, UUC, UUG, GUU, UGU).

Além disso, pode-se observar nos hipercubos booleanos os aminoácidos hidrogênio, básicos, hidroxila e cadeia alifática. O códon UGG é um anel aromático que codifica o aminoácido triptofano, representado pela cor verde musgo. As trincas GGU, GGC, GGA e GGG, representam o hidrogênio (representadas pela cor amarela), são os códons que codificam o aminoácido glicina, considerado o mais simples e está presente na maioria das proteínas. No hipercubo booleano estão próximos um do outro em uma face de um dos cubos menores. Os códons CCA, CCC, CCG e CCU são classificados como cadeia alifática e codifica o aminoácido prolina, sendo este o mais rígido dos vinte aminoácidos, ou seja, possui uma estrutura quimicamente coesa e rígida.

Pode-se observar também os aminoácidos sulfidrila ou tiol (representados pela cor roxa), os básicos (representados pela cor verde claro). Por fim, o códon STOP (representado pela cor preta), que é utilizado para interromper a proteína antes de seu término.

3.2 RESULTADOS OBTIDOS PARA O ROTULAMENTO B

Nesta seção serão apresentados alguns resultados obtidos com a utilização de permutações associadas ao rotulamento B do código genético.

Foi escolhida a permutação 0321, para o reticulado primal e, como consequência, a permutação 2103 foi utilizada para a construção do reticulado dual.

Ao contrário do rotulamento A, onde foi utilizada a complementaridade biológica das bases, no rotulamento B será considerada a complementaridade algébrica, sendo assim, A - G e C - U. Esse fato toma como referência a proposta de Faria (2011).

As Tabelas 11 e 12 apresentam as associações estabelecidas para as permutações do rotulamento B, relacionadas aos casos primal e dual.

Tabela 11 – Associação estabelecida pela permutação 0321 primal do rotulamento B.

A	**C**	**G**	**U**
0	3	2	1
10	00	01	11

Fonte: do autor.

Tabela 12 – Associação estabelecida pela permutação 2103 dual do rotulamento B.

A	**C**	**G**	**U**
2	1	0	3
01	11	10	00

Fonte: do autor.

As operações primal e dual relacionadas às operações do rotulamento B são apresentadas nas Tabelas 13, 14, 15 e 16.

Tabela 13 – Operação primal $\wedge$ (e) do rotulamento B.

$\wedge$	**00**	**01**	**10**	**11**	$\wedge$	**C**	**G**	**A**	**U**
00	00	00	00	00	**C**	C	C	C	C
01	00	01	00	01	**G**	C	G	C	G
10	00	00	10	10	**A**	C	C	A	A
11	00	01	10	11	**U**	C	G	A	U

Fonte: do autor.

Tabela 14 – Operação primal $\vee$ (ou) do rotulamento B.

$\vee$	**00**	**01**	**10**	**11**	$\vee$	**C**	**G**	**A**	**U**
00	00	01	10	11	**C**	C	G	A	U
01	01	01	11	11	**G**	G	G	U	U
10	10	11	10	11	**A**	A	U	A	U
11	11	11	11	11	**U**	U	U	U	U

Fonte: do autor.

Tabela 15 – Operação dual $\wedge$ (e) do rotulamento B.

$\wedge$	**00**	**01**	**10**	**11**	$\wedge$	**U**	**A**	**G**	**C**
00	00	00	00	00	**U**	U	U	U	U
01	00	01	00	01	**A**	U	A	U	A
10	00	00	10	10	**G**	U	U	G	G
11	00	01	10	11	**C**	U	A	G	C

Fonte: do autor.

Tabela 16 – Operação dual $\vee$ (ou) do rotulamento B.

$\vee$	**00**	**01**	**10**	**11**	$\vee$	**U**	**A**	**G**	**C**
00	00	01	10	11	**U**	U	A	G	C
01	01	01	11	11	**A**	A	A	C	C
10	10	11	10	11	**G**	G	C	G	C
11	11	11	11	11	**C**	C	C	C	C

Fonte: do autor.

O reticulado booleano primal será obtido a partir da tabela primal. Tem-se que:

a) A base C (citosina) se liga à ela mesma e às bases G (guanina), A (adenina) e U (uracila);

b) A base G (guanina) se liga à ela mesma e à base U (uracila);

c) A base A (adenina) se liga à ela mesma e à base U (uracila);

d) A base U (uracila) se liga à ela mesma.

O reticulado booleano dual será obtido a partir da tabela dual. Tem-se que:

a) A base U (uracila) se liga à ela mesma e às bases A (adenina), G (guanina) e C (citosina);

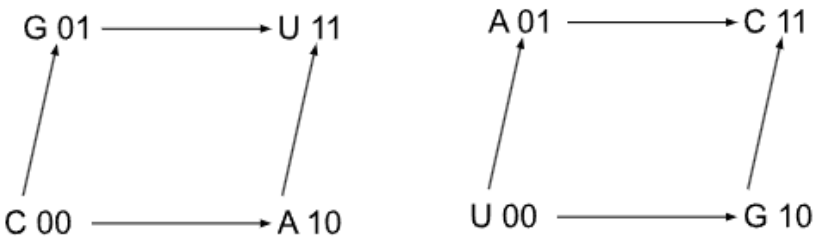
b) A base A (adenina) se liga à ela mesma e à base C (citosina);

c) A base G (guanina) se liga à ela mesma e à base C (citosina);

d) A base C (citosina) se liga à ela mesma.

Os reticulados booleanos primal e dual são apresentados na Figura 24.

Figura 24 – Reticulados booleanos primal (0321) e dual (2103) associados ao rotulamento *B*.



Fonte: Adaptado de Fernandes e Oliveira (2021).

3.2.1 Diagramas de Hasse associados ao rotulamento B

Conforme apresentado para os casos do rotulamento A, no rotulamento B os diagramas de Hasse primal e dual seguirão o seguinte procedimento de construção, proposto em Fernandes e Oliveira (2021).

a) 1ª linha: elemento máximo, atribuído por 11.

$$\text{UUU} \leftrightarrow 111111$$

b) 2ª linha: um elemento 0 e cinco elementos 1 (seis códons).

$$\begin{array}{lll} \text{UUG} \leftrightarrow 111101 & \text{UGU} \leftrightarrow 110111 & \text{GUU} \leftrightarrow 011111 \\ \text{AUU} \leftrightarrow 101111 & \text{UAU} \leftrightarrow 111011 & \text{UUA} \leftrightarrow 111110 \end{array}$$

c) 3ª linha: dois elementos 0 e quatro elementos 1 (quinze códons).

d) 4ª linha: três elementos 0 e três elementos 1 (vinte códons).

e) 5ª linha: quatro elementos 0 e dois elementos 1 (quinze códons).

UGG ↔ 110101	UAG ↔ 111001	GUG ↔ 011101
AUG ↔ 101101	GGU ↔ 010111	AGU ↔ 100111
UCU ↔ 110011	AAU ↔ 101011	CUU ↔ 001111
GAU ↔ 011011	UUC ↔ 111100	GUA ↔ 011110
AUA ↔ 101110	UGA ↔ 110110	UAA ↔ 111010

GGG ↔ 010101	AGG ↔ 100101	GAG ↔ 011001	AAG ↔ 101001
CUG ↔ 001101	UCG ↔ 110001	ACU ↔ 100011	CGU ↔ 000111
GCU ↔ 010011	CAU ↔ 001011	UGC ↔ 110100	AUC ↔ 101100
UAC ↔ 111000	GUC ↔ 011100	CUA ↔ 001110	UCA ↔ 110010
GGA ↔ 010110	AGA ↔ 100110	GAA ↔ 011010	AAA ↔ 101010

CGG ↔ 000101	CAG ↔ 001001	GCG ↔ 010001
ACG ↔ 100001	CCU ↔ 000011	AGC ↔ 100100
UCC ↔ 110000	GGC ↔ 010100	CUC ↔ 001100
GAC ↔ 011000	AAC ↔ 101000	GCA ↔ 010010
ACA ↔ 100010	CGA ↔ 000110	CAA ↔ 001010

CCG ↔ 000001	CGC ↔ 000100	GCC ↔ 010000
ACC ↔ 100000	CAC ↔ 001000	CCA ↔ 000010

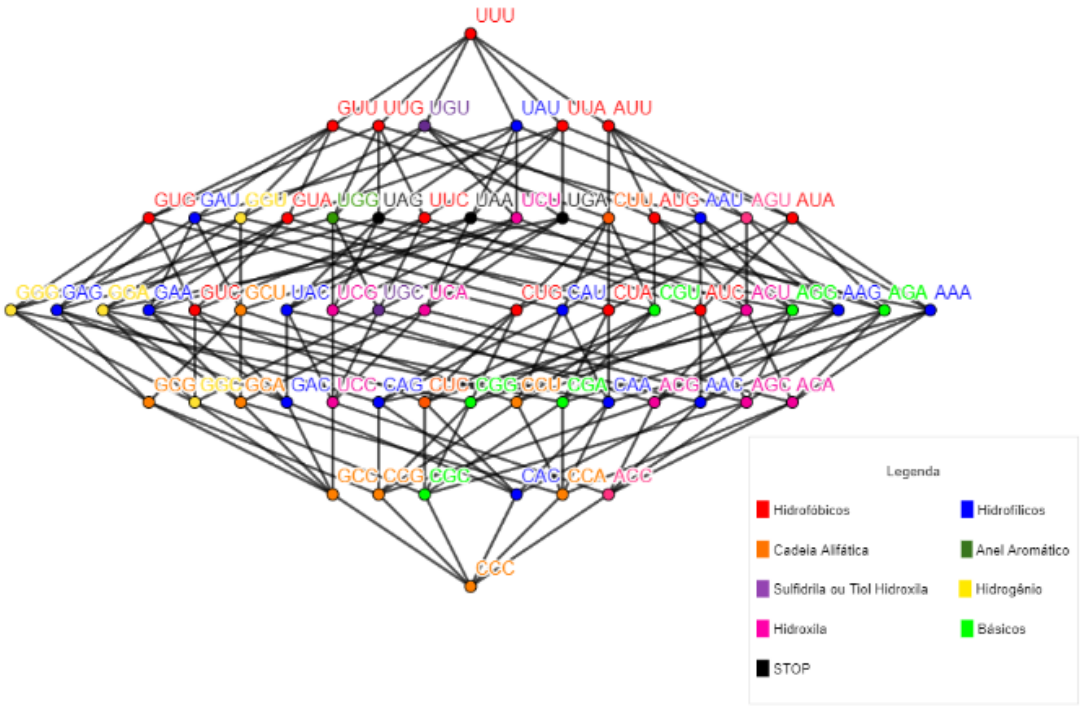
f) 6ª linha: cinco elementos 0 e um elemento 1 (seis códons).

g) 7ª linha: elemento mínimo, atribuído por 00.

CCC ↔ 000000

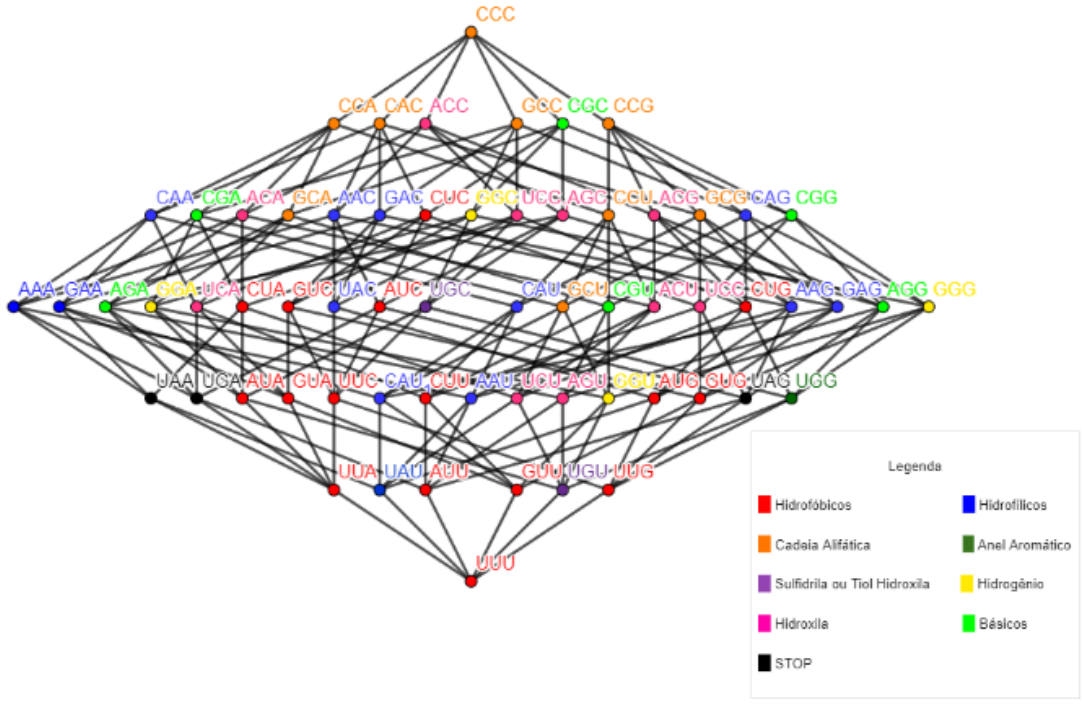
Os diagramas de Hasse, primal e dual, referentes às permutações 0321 e 2103, respectivamente, são apresentados nas Figuras 25 e 26.

Figura 25 – Diagrama de Hasse primal referente à permutação 0321 do rotulamento B.



Fonte: Adaptado de Fernandes e Oliveira (2021).

Figura 26 – Diagrama de Hasse dual referente à permutação 2103 do rotulamento B.



Fonte: Adaptado de Fernandes e Oliveira (2021).

3.2.2 Código de Gray - Rotulamento B

A Tabela 17 apresenta o código de Gray associado à permutação 0321 (primal) do rotulamento B do código genético, na qual os dígitos 00 correspondem à base C, os dígitos 11 correspondem à base U, os dígitos 10 correspondem à base A e os dígitos 01 correspondem à base G.

Na Tabela 18 é apresentado o código de Gray associado à permutação 2103 (dual) do rotulamento B do código genético, na qual os dígitos 00 correspondem à base U, os dígitos 11 correspondem à base C, os dígitos 10 correspondem à base G e os dígitos 01 correspondem à base A.

As cores apresentadas se referem às características e propriedades associadas aos aminoácidos, conforme apresentado no diagramas de Hasse construídos anteriormente. Por meio dessa construção é possível analisar os padrões identificados na construção dos diagramas de Hasse, nas tabelas do código de Gray e posteriormente na construção dos hipercubos booleanos.

Tabela 17 – Código de Gray associado à permutação 0321 do código genético.

Binário	Códon	Aminoácido	Binário	Códon	Aminoácido
000011	CCU	prolina	000010	CCA	prolina
000000	CCC	prolina	000001	CCG	prolina
100001	ACG	treonina	100000	ACC	treonina
100010	ACA	treonina	100011	ACU	treonina
110011	UCU	serina	110010	UCA	serina
110000	UCC	serina	110001	UCG	serina
010001	GCG	alanina	010000	GCC	alanina
010010	GCA	alanina	010011	GCU	alanina
011011	GAU	aspartato	011010	GAA	glutamato
011000	GAC	aspartato	011001	GAG	glutamato
111001	UAG	STOP	111000	UAC	tirosina
111010	UAA	STOP	111011	UAU	tirosina
101011	AAU	asparagina	101010	AAA	lisina
101000	AAC	asparagina	101001	AAG	lisina
001001	CAG	glutamina	001000	CAC	histidina
001010	CAA	glutamina	001011	CAU	histidina
001111	CUU	isoleucina	001110	CUA	isoleucina
001100	CUC	isoleucina	001101	CUG	metionina
101101	AUG	leucina	101100	AUC	leucina
101110	AUA	leucina	101111	AUU	leucina
111111	UUU	fenilalanina	111110	UUA	fenilalanina
111100	UUC	leucina	111101	UUG	leucina
011101	GUG	valina	011100	GUC	valina
011110	GUA	valina	011111	GUU	valina
010111	GGU	glicina	010110	GGA	glicina
010100	GGC	glicina	010101	GGG	glicina
110101	UGG	triptofano	110100	UGC	cisteína
110110	UGA	STOP	110111	UGU	cisteína
100111	AGU	serina	100110	AGA	arginina
100100	AGC	serina	100101	AGG	arginina
000101	CGG	arginina	000100	CGC	arginina
000110	CGA	arginina	000111	CGU	arginina

Fonte: do autor.

3.2.3 Hipercubos booleanos associados ao rotulamento B

Os hipercubos booleanos associados à permutação 0321 (primal) e 2103 (dual) do rotulamento B estão apresentados nas Figuras 27 e 28.

Ao contrário do rotulamento A, não foi utilizada a complementaridade biológica das bases no rotulamento B, mas sim, a complementaridade algébrica, na qual 00 - 11 e 01 - 10, sendo citisona (C) com uracila (U) e guanina (G) e adenina (A). A partir das construções realizadas, é possível observar que se considerar por exemplo CAC, ACC, CCA, CCG, GCC, CGC, sua imagem será UAU, AUU, UUA, UUG, GUU, UGU.

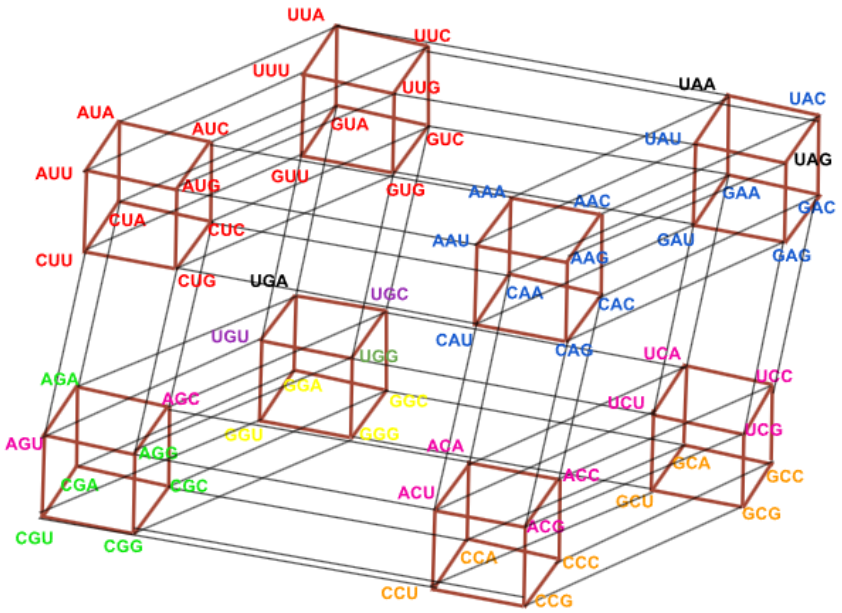
Tabela 18 – Código de Gray associado à permutação 2103 do código genético.

Binário	Códon	Aminoácido	Binário	Códon	Aminoácido
000011	UUC	leucina	000010	UUG	leucina
000000	UUU	fenilalanina	000001	UUA	fenilalanina
100001	GUA	valina	100000	GUU	valina
100010	GUG	valina	100011	GUC	valina
110011	CUC	isoleucina	110010	CUG	metionina
110000	CUU	isoleucina	110001	CUA	isoleucina
010001	AUA	leucina	010000	AUU	leucina
010010	AUG	leucina	010011	AUC	leucina
011011	AGC	serina	011010	AGG	arginina
011000	AGU	serina	011001	AGA	arginina
111001	CGA	arginina	111000	CGU	arginina
111010	CGG	arginina	111011	CGC	arginina
101011	GGC	glicina	101010	GGG	glicina
101000	GGU	glicina	101001	AAG	glicina
001001	UGA	STOP	001000	UGU	cisteína
001010	UGG	triptofano	001011	UGC	cisteína
001111	UCC	serina	001110	UCG	serina
001100	UCU	serina	001101	UCA	serina
101101	GCA	alanina	101100	GCU	alanina
101110	GCG	alanina	101111	GCC	alanina
111111	CCC	prolina	111110	CCG	prolina
111100	CCU	prolina	111101	CCA	prolina
011101	ACA	treonina	011100	ACU	treonina
011110	ACG	treonina	011111	ACC	treonina
010111	AAC	asparagina	010110	AAG	lisina
010100	AAU	asparagina	010101	AAA	lisina
110101	CAA	glutamina	110100	CAU	histidina
110110	CAG	glutamina	110111	CAC	histidina
100111	GAC	aspartato	100110	GAG	glutamato
100100	GAU	aspartato	100101	GAA	glutamato
000101	UAA	STOP	000100	UAU	tirosina
000110	UAG	STOP	000111	UAC	tirosina

Fonte: do autor.

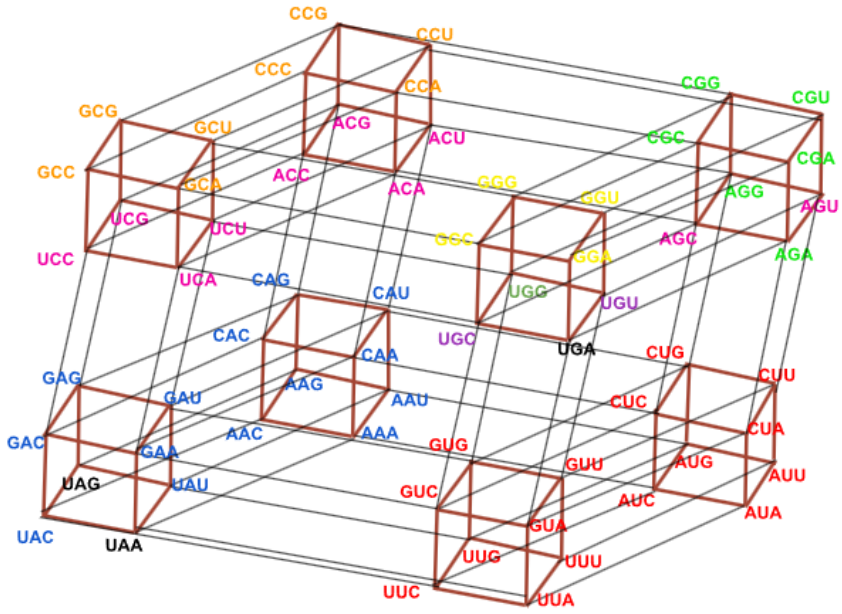
De acordo com a legenda, diferente do diagrama de Hasse do rotulamento A, o códon que codifica um aminoácido hidrofóbico (códons com U na segunda posição, representados em vermelho) não é complementar a um códon que codifica um aminoácido hidrofílico (códons com A na segunda posição, representados em azul). Os códons hidrofílicos e hidrofóbicos ficaram separados. Os códons que codificam os aminoácidos hidrogênio, básicos, anel aromático e STOP ficaram próximos. Além disso, o códon UGG (verde) que codifica o aminoácido triptofano ficou ao lado do códon UGA (preto), sendo este utilizado para interromper a proteína antes de seu término.

Figura 27 – Hipercubo booleano associado à permutação 0321 do código genético.



Fonte: do autor.

Figura 28 – Hipercubo booleano associado à permutação 2103 do código genético.



Fonte: do autor.

3.3 RESULTADOS OBTIDOS PARA O ROTULAMENTO *C*

Nesta seção serão apresentados alguns resultados obtidos para o rotulamento C do código genético.

Para o estudo deste rotulamento, será utilizada a permutação 1302, para o reticulado

primal e a permutação 3120 para o reticulado dual.

Neste caso, assim como no rotulamento B, não foi utilizada a complementaridade biológica das bases, mas sim, a complementaridade algébrica, na qual 00 - 11 e 01 - 10. Desta forma, as bases nitrogenadas referentes a essa complementaridade algébrica são: A - C e G - U, ou vice-versa. As Tabelas 19 e 20 apresentam as associações estabelecidas para os casos primal e dual das permutações escolhidas.

Tabela 19 – Associação estabelecida pela permutação 1302 primal do rotulamento C.

A	**C**	**G**	**U**
1	3	0	2
11	00	01	10

Fonte: do autor.

Tabela 20 – Associação estabelecida pela permutação 3120 dual do rotulamento C.

A	**C**	**G**	**U**
3	1	2	0
00	11	10	01

Fonte: do autor.

Analogamente ao que foi apresentado para os rotulamentos A e B, nas Tabelas 21, 22, 23 e 24 são apresentadas as operações primal e dual do rotulamento C.

Tabela 21 – Operação primal $\wedge$ (e) do rotulamento C.

$\wedge$	**00**	**01**	**10**	**11**	$\wedge$	**C**	**G**	**U**	**A**
00	00	00	00	00	**C**	C	C	C	C
01	00	01	00	01	**G**	C	G	C	G
10	00	00	10	10	**A**	C	C	U	U
11	00	01	10	11	**U**	C	G	U	A

Fonte: do autor.

Tabela 22 – Operação primal $\vee$ (ou) do rotulamento C.

$\vee$	**00**	**01**	**10**	**11**	$\vee$	**C**	**G**	**U**	**A**
00	00	01	10	11	**C**	C	G	U	A
01	01	01	11	11	**G**	G	G	A	A
10	10	11	10	11	**A**	U	A	U	A
11	11	11	11	11	**U**	A	A	A	A

Fonte: do autor.

Tabela 23 – Operação dual $\wedge$ (e) do rotulamento C.

$\wedge$	**00**	**01**	**10**	**11**	$\wedge$	**A**	**U**	**G**	**C**
00	00	00	00	00	**U**	A	A	A	A
01	00	01	00	01	**A**	A	U	A	U
10	00	00	10	10	**G**	A	A	G	G
11	00	01	10	11	**C**	A	U	G	C

Fonte: do autor.

Tabela 24 – Operação dual $\vee$ (ou) do rotulamento C.

$\vee$	**00**	**01**	**10**	**11**	$\vee$	**A**	**U**	**G**	**C**
00	00	01	10	11	**U**	A	U	G	C
01	01	01	11	11	**A**	U	U	C	C
10	10	11	10	11	**G**	G	C	G	C
11	11	11	11	11	**C**	C	C	C	C

Fonte: do autor.

O reticulado booleano primal será obtido a partir da tabela primal. Tem-se que:

a) A base C (citosina) se liga à ela mesma e às bases G (guanina), U (uracila) e A (adenina);

b) A base G (guanina) se liga à ela mesma e à base A (adenina);

c) A base U (uracila) se liga à ela mesma e à base A (adenina);

d) A base A (adenina) se liga à ela mesma.

O reticulado booleano dual será obtido a partir da tabela dual. Tem-se que:

a) A base A (adenina) se liga à ela mesma e às bases U (uracila), G (guanina) e C (citosina);

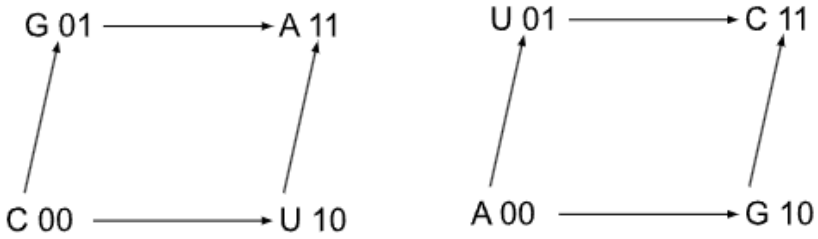
b) A base U (uracila) se liga à ela mesma e à base C (citosina);

c) A base G (guanina) se liga à ela mesma e à base C(citosina);

d) A base C (citosina) se liga à ela mesma.

A Figura 29 apresenta os reticulados booleanos primal e dual associados às permutações selecionadas para o rotulamento C.

Figura 29 – Reticulados booleanos primal (1302) e dual (3120) associados ao rotulamento C.



Fonte: Adaptado de Fernandes e Oliveira (2021).

3.3.1 Diagramas de Hasse associados ao rotulamento C

O procedimento para a construção dos diagramas de Hasse associados ao rotulamento C segue processo análogo ao apresentado nos casos dos rotulamentos A e B:

a) 1ª linha: elemento máximo, atribuído por 11.

$$AAA \leftrightarrow 111111$$

b) 2ª linha: um elemento 0 e cinco elementos 1 (seis códons).

$$\begin{array}{lll} AAG \leftrightarrow 111101 & AGA \leftrightarrow 110111 & GAA \leftrightarrow 011111 \\ UAA \leftrightarrow 101111 & AUA \leftrightarrow 111011 & AAU \leftrightarrow 111110 \end{array}$$

c) 3ª linha: dois elementos 0 e quatro elementos 1 (quinze códons).

$$\begin{array}{lll} AGG \leftrightarrow 110101 & AUG \leftrightarrow 111001 & GAG \leftrightarrow 011101 \\ UAG \leftrightarrow 101101 & GGA \leftrightarrow 010111 & UGA \leftrightarrow 100111 \\ ACA \leftrightarrow 110011 & UUA \leftrightarrow 101011 & CAA \leftrightarrow 001111 \\ GUA \leftrightarrow 011011 & AAC \leftrightarrow 111100 & GAU \leftrightarrow 011110 \\ UAU \leftrightarrow 101110 & AGU \leftrightarrow 110110 & AUU \leftrightarrow 111010 \end{array}$$

d) 4ª linha: três elementos 0 e três elementos 1 (vinte códons).

GGG $\leftrightarrow$ 010101	UGG $\leftrightarrow$ 100101	GUG $\leftrightarrow$ 011001	UUG $\leftrightarrow$ 101001
CAG $\leftrightarrow$ 001101	ACG $\leftrightarrow$ 110001	UCA $\leftrightarrow$ 100011	CGA $\leftrightarrow$ 000111
GCA $\leftrightarrow$ 010011	CUA $\leftrightarrow$ 001011	AGC $\leftrightarrow$ 110100	UAC $\leftrightarrow$ 101100
AUC $\leftrightarrow$ 111000	GAC $\leftrightarrow$ 011100	CAU $\leftrightarrow$ 001110	ACU $\leftrightarrow$ 110010
GGU $\leftrightarrow$ 010110	UGU $\leftrightarrow$ 100110	GUU $\leftrightarrow$ 011010	UUU $\leftrightarrow$ 101010

e) 5ª linha: quatro elementos 0 e dois elementos 1 (quinze códons).

CGG $\leftrightarrow$ 000101	CUG $\leftrightarrow$ 001001	GCG $\leftrightarrow$ 010001
UCG $\leftrightarrow$ 100001	CCA $\leftrightarrow$ 000011	UGC $\leftrightarrow$ 100100
ACC $\leftrightarrow$ 110000	GGC $\leftrightarrow$ 010100	$CAC \leftrightarrow$ 001100
GUC $\leftrightarrow$ 011000	UUC $\leftrightarrow$ 101000	GCU $\leftrightarrow$ 010010
UCU $\leftrightarrow$ 100010	CGU $\leftrightarrow$ 000110	CUU $\leftrightarrow$ 001010

f) 6ª linha: cinco elementos 0 e um elemento 1 (seis códons).

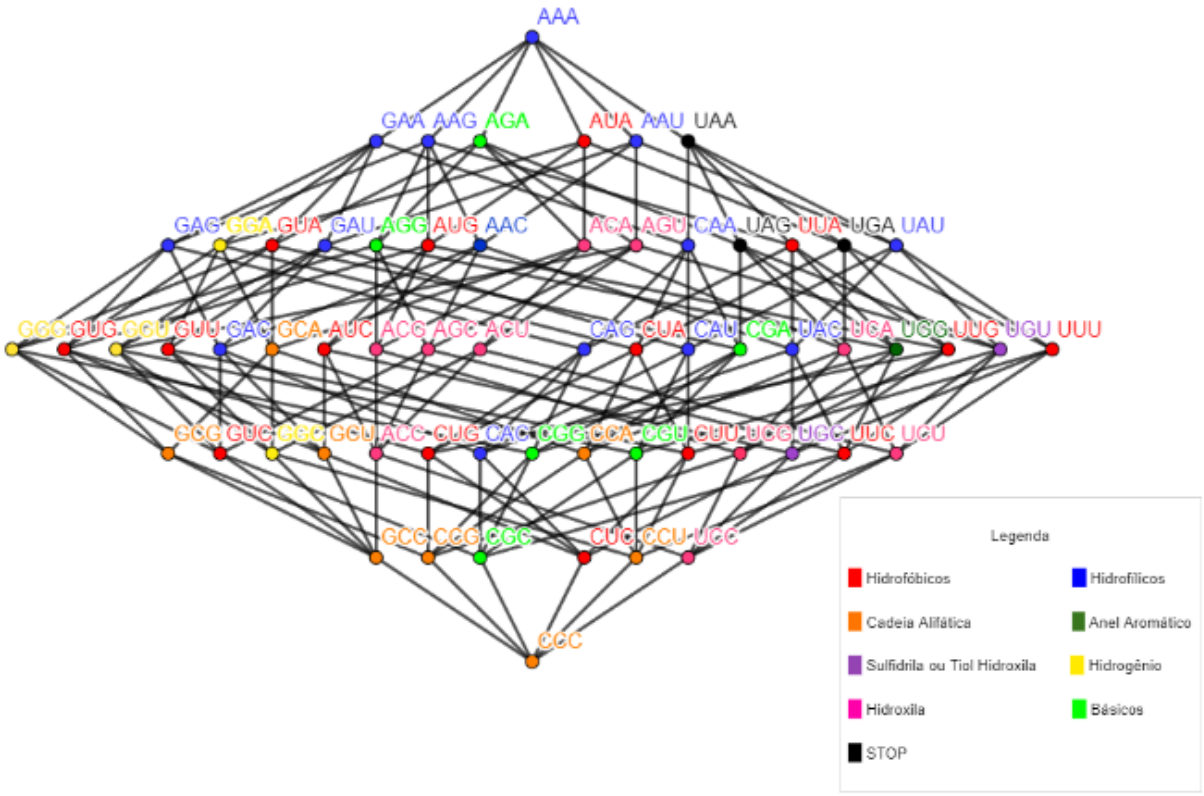
CCG $\leftrightarrow$ 000001	CGC $\leftrightarrow$ 000100	GCC $\leftrightarrow$ 010000
UCC $\leftrightarrow$ 100000	CUC $\leftrightarrow$ 001000	CCU $\leftrightarrow$ 000010

g) 7ª linha: elemento mínimo, atribuído por 00.

CCC $\leftrightarrow$ 000000

Os diagramas de Hasse, primal e dual, referentes às permutações 1302 e 3120, respectivamente, são apresentados nas Figuras 30 e 31.

Figura 30 – Diagrama de Hasse primal referente à permutação 1302 do rotulamento C.



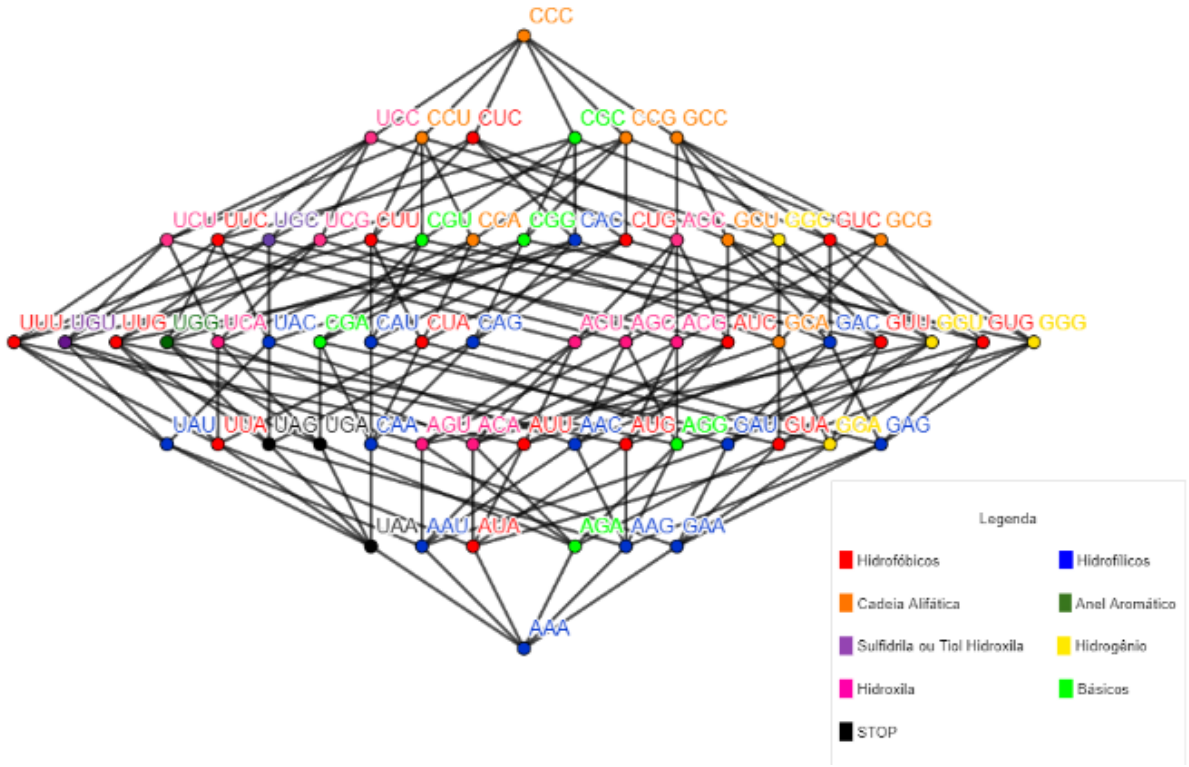
Fonte: Adaptado de Fernandes e Oliveira (2021).

3.3.2 Código de Gray - Rotulamento C

A Tabela 25 apresenta o código de Gray associado à permutação 1203 (primal) do rotulamento C do código genético, na qual os dígitos 00 correspondem à base C, os dígitos 11 correspondem à base A, os dígitos 01 correspondem à base G e os dígitos 10 correspondem à base U.

A Tabela 26 apresenta o código de Gray associado à permutação 3120 (dual) do rotulamento C do código genético, na qual os dígitos 00 correspondem à base A, os dígitos 11 correspondem à base C, os dígitos 10 correspondem à base G e os dígitos 01 correspondem à base U.

Figura 31 – Diagrama de Hasse dual referente à permutação 3120 do rotulamento C.



Fonte: Adaptado de Fernandes e Oliveira (2021).

Da mesma forma que nos casos apresentados para os rotulamentos A e B, pode-se perceber uma separação dos códons, de acordo com suas características físico-químicas, mostrando mais uma vez uma interessante classificação dos códons, por meio de suas propriedades, a partir da modelagem matemática.

3.3.3 Hipercubos booleanos associados ao rotulamento C

Os hipercubos booleanos associados à permutação 1302 (primal) e 3120 (dual) do rotulamento C estão apresentados nas Figuras 32 e 33.

Tabela 25 – Código de Gray associado à permutação 1302 do código genético.

Binário	Códon	Aminoácido	Binário	Códon	Aminoácido
000011	CCA	prolina	000010	CCU	prolina
000000	CCC	prolina	000001	CCG	prolina
100001	UCG	serina	100000	UCC	serina
100010	UCU	serina	100011	UCA	serina
110011	ACA	treonina	110010	ACU	treonina
110000	ACC	treonina	110001	ACG	treonina
010001	GCG	alanina	010000	GCC	alanina
010010	GCU	alanina	010011	GCA	alanina
011011	GUA	valina	011010	GUU	valina
011000	GUC	valina	011001	GUG	valina
111001	AUG	leucina	111000	AUC	leucina
111010	AUU	leucina	111011	AUA	leucina
101011	UUA	fenilalanina	101010	UUU	fenilalanina
101000	UUC	leucina	101001	UUG	leucina
001001	CUG	metionina	001000	CUC	isoleucina
001010	CUU	isoleucina	001011	CUA	isoleucina
001111	CAA	glutamina	001110	CAU	histidina
001100	CAC	histidina	001101	CAG	glutamina
101101	UAG	STOP	101100	UAC	tirosina
101110	UAU	tirosina	101111	UAA	STOP
111111	AAA	lisina	111110	AAU	asparagina
111100	AAC	asparagina	111101	AAG	lisina
011101	GAG	glutamato	011100	GAC	aspartato
011110	GAU	aspartato	011111	GAA	glutamato
010111	GGA	glicina	010110	GGU	glicina
010100	GGC	glicina	010101	GGG	glicina
110101	AGG	arginina	110100	AGC	serina
110110	AGU	serina	110111	AGA	arginina
100111	UGA	STOP	100110	UGU	cisteína
100100	UGC	cisteína	100101	UGG	triptofano
000101	CGG	arginina	000100	CGC	arginina
000110	CGU	arginina	000111	CGA	arginina

Fonte: do autor.

Pode-se perceber que por meio das construções apresentadas, o rotulamento A segue uma complementaridade biológica, em que as cores representam propriedades químicas e funcionais dos aminoácidos. A organização dos códons no hipercubo segue as interações e características biológicas dos aminoácidos. O hipercubo associado ao rotulamento A representa uma organização mais estruturada e com maior concentração de certos grupos em regiões específicas, como os grupos hidrofóbicos nas bordas e os grupos hidroxila e básicos nas regiões centrais. Essa configuração pode ser mais típica de proteínas com uma estrutura bem definida e funções internas de interação. Os grupos hidrofóbicos ficam nas bordas no rotulamento A, visto que em proteínas solúveis em água, cadeias hidrofóbicas tendem a se "esconder"no interior, mas em contextos de membrana (ou no hipercubo),

Tabela 26 – Código de Gray associado à permutação 3120 do código genético.

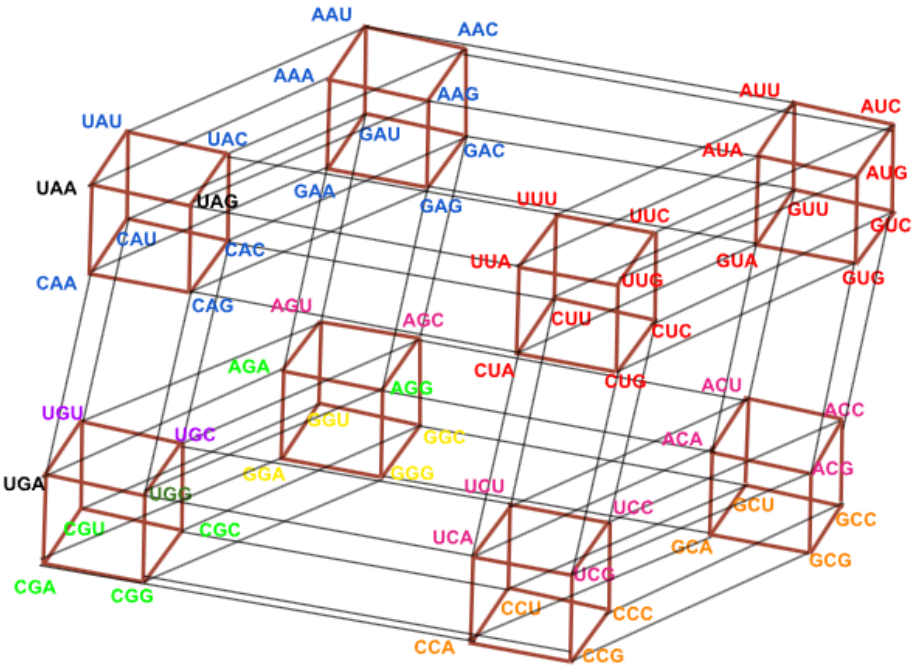
Binário	Códon	Aminoácido	Binário	Códon	Aminoácido
000011	AAC	asparagina	000010	AAG	lisina
000000	AAA	lisina	000001	AAU	asparagina
100001	GAU	aspartato	100000	GAA	glutamato
100010	GAG	glutamato	100011	GAC	aspartato
110011	CAC	histidina	110010	CAG	glutamina
110000	CAA	glutamina	110001	CAU	histidina
010001	UAU	tirosina	010000	UAA	STOP
010010	UAG	STOP	010011	UAC	tirosina
011011	UGC	cisteína	011010	UGG	triptofano
011000	UGA	STOP	011001	UGU	cisteína
111001	CGU	arginina	111000	CGA	arginina
111010	CGG	arginina	111011	CGC	arginina
101011	GGC	glicina	101010	GGG	glicina
101000	GGA	glicina	101001	GGG	glicina
001001	AGU	serina	001000	AGA	arginina
001010	AGG	serina	001011	AGC	arginina
001111	ACC	treonina	001110	ACG	treonina
001100	ACA	treonina	001101	ACU	treonina
101101	GCU	alanina	101100	GCA	alanina
101110	GCG	alanina	101111	GCC	alanina
111111	CCC	prolina	111110	CCG	prolina
111100	CCA	prolina	111101	CCU	prolina
011101	UCU	serina	011100	UCA	serina
011110	UCG	serina	011111	UCC	serina
010111	UUC	leucina	010110	UUG	leucina
010100	UUA	fenilalanina	010101	UUU	fenilalanina
110101	CUU	isoleucina	110100	CUA	isoleucina
110110	CUG	metionina	110111	CUC	isoleucina
100111	GUC	valina	100110	GUG	valina
100100	GUA	valina	100101	GUU	valina
000101	AUU	leucina	000100	AUA	leucina
000110	AUG	leucina	000111	AUC	leucina

Fonte: do autor.

podem aparecer nas bordas por motivos de empacotamento.

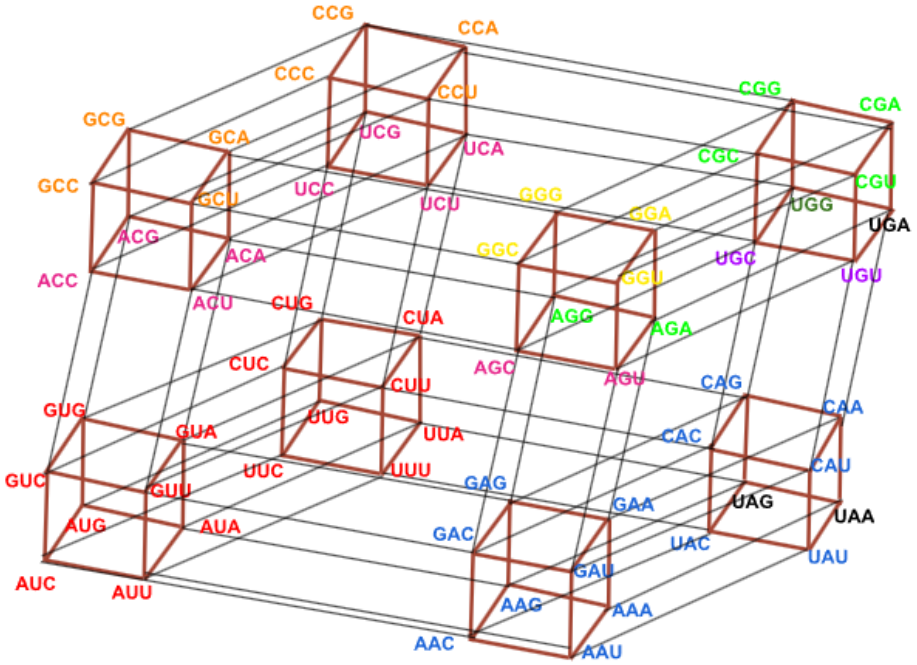
Os rotulamentos B e C seguem uma complementaridade algébrica, em que as propriedades dos códons são distribuídas de acordo com o contexto algébrico. Os hipercubos associados aos rotulamentos B e C mostram uma distribuição mais equilibrada e dispersa de suas propriedades, ou seja, as características físico-químicas dos aminoácidos estão organizadas de forma menos agrupada e mais homogênea no espaço analisado, sugerindo uma proteína mais flexível ou que interage com o ambiente de forma mais dinâmica. A distribuição uniforme das cores, como os grupos básicos e hidroxila, pode indicar a capacidade dessas proteínas de interagir tanto com o meio interno quanto com outras

Figura 32 – Hipercubo booleano associado à permutação 1302 do código genético.



Fonte: do autor.

Figura 33 – Hipercubo booleano associado à permutação 3120 do código genético.



Fonte: do autor.

moléculas externas. A dispersão nos rotulamentos B e C ocorre uma vez que, abordagem algébrica prioriza simetria ou otimização de espaço, não a biologia direta, levando a uma distribuição menos intuitiva, mas funcional.

Nos hipercubos associados aos três rotulamentos, há uma clara separação entre as

regiões hidrofóbicas e hidrofílicas, refletindo a organização típica das proteínas na célula, mas com variações nas áreas de flexibilidade e estabilidade estrutural. A distribuição dos grupos aromáticos, sulfidrila, hidroxila e outros sugere que, apesar de os três rotulamentos compartilharem semelhanças fundamentais em sua organização, os hipercubos B e C têm uma dinâmica mais dispersa nas interações químicas, enquanto o A tem uma configuração mais concentrada e estruturada. Assim, o rotulamento A é ideal para modelar proteínas com funções especializadas e estruturas rígidas, enquanto os rotulamentos B e C capturam o comportamento de proteínas flexíveis e adaptativas, essenciais para processos dinâmicos. Com isso, os hipercubos servem como ferramentas visuais para entender como a organização dos códons influencia a função proteica.

Dessa maneira, foi possível estabelecer uma análise das características e propriedades associadas aos aminoácidos e as conexões existentes entre as construções dos diagramas de Hasse, códigos de Gray e hipercubos booleanos para permutações escolhidas dos três rotulamentos associados ao código genético.

4 ELEMENTOS DE TEORIA DA INFORMAÇÃO E SUAS RELAÇÕES COM BIOLOGIA MOLECULAR

Nesta seção serão apresentados os principais elementos de Teoria da Informação utilizados neste trabalho, juntamente com as relações envolvendo Biologia Molecular. As referências utilizadas foram: Fernandes Jr. e Vargas Jr. (2011), Horta (2001), Lin e Costello Jr. (1983) e Shannon (1948).

De acordo com Horta (2001), uma das propriedades essenciais da informação é a sua capacidade de ser transmitida. No campo da computação, a expressão da informação é frequentemente referida como mensagem, geralmente representada por uma sequência de símbolos.

Uma mensagem pode ser abordada sob três perspectivas principais:

a) Informação sintática: enfatiza os símbolos individuais da mensagem e suas relações;

b) Informação semântica: foca no significado da mensagem;

c) Informação pragmática: relaciona-se com a aplicação ou efeito da mensagem.

Neste contexto, o termo "informação"refere-se especificamente à dimensão sintática, uma vez que corresponde à análise da forma de expressão da informação. Neste trabalho, um dos objetivos é analisar a relação da Teoria da Informação com a Biologia, desse modo, é possível realizar com a aplicação da Teoria da Informação à Biologia Molecular uma análise sintática das sequências de nucleotídeos ou aminoácidos que compõem um segmento de DNA ou proteína. A partir dessa análise sintática, é possível fazer inferências sobre o significado biológico dessas sequências. Outra abordagem é analisar sequências cujos significados semânticos são conhecidos, realizando medições para identificar padrões nas informações semânticas.

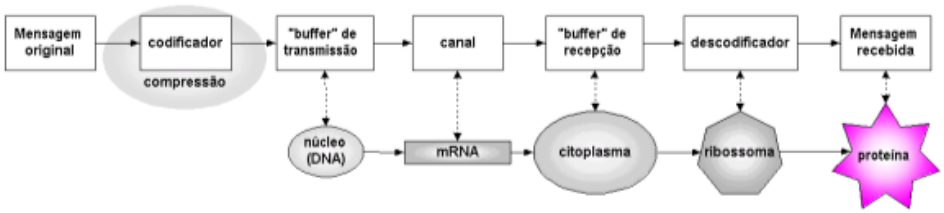
A Teoria da Informação, particularmente a Teoria Matemática da Comunicação, foi inicialmente desenvolvida por Claude Shannon na década de 1940 com o intuito de aprimorar os sistemas de telefonia. A entropia foi introduzida por Rudolf Clausius para a Física em 1864, enquanto Shannon desenvolveu a Teoria da Informação em 1949. O principal objetivo de Shannon era otimizar a quantidade de informação transmitida através de um canal de comunicação imperfeito, ou seja, um canal que pode introduzir erros nas mensagens.

Nessa perspectiva, quando um canal de comunicação não é perfeito, conforme apresentado em Lin e Costello Jr. (1983), uma maneira do receptor detectar e corrigir erros é garantir que a mensagem contenha redundância. Como a capacidade máxima do canal é fixa, a presença de redundância diminui a quantidade efetiva de informação transmitida.

Um dos principais objetivos de Shannon era determinar as taxas máximas teóricas de compressão de dados, o que significa remover redundâncias de maneira que a mensagem ocupe o menor espaço possível, mantendo um nível de redundância suficiente para detectar e corrigir erros que poderiam comprometer a decodificação da mensagem. Para resolver isso, Shannon introduziu o conceito de entropia.

Diante disso, na Figura 34 é apresentado um diagrama de blocos com os componentes típicos envolvidos na transmissão de mensagens entre um emissor e um receptor. Também estabelece um paralelo entre o sistema de comunicação e a síntese de proteínas a partir do DNA. No "sistema de comunicação genético", a mensagem original e o codificador são desconhecidos, pois as sequências codificadoras estão armazenadas no núcleo das células. Sempre que necessário, a célula ativa o núcleo para gerar "cópias"dessas sequências de código, o RNAm (canal), que é então decodificado pelos ribossomos em proteínas, conforme apresentado em Horta (2001).

Figura 34 – Esquema de transmissão de mensagens.



Fonte: Horta (2001).

A entropia está associada à Teoria da Informação, que fundamenta muitos conceitos essenciais da compressão de dados. Como alguns estudos utilizam técnicas de compressão para avaliar a entropia do DNA, a Teoria da Informação é fundamental.

Segundo Mohammed (2010), existe a dificuldade em distinguir o conteúdo informativo das áreas codificantes (éxons) e não codificantes (íntrons) em uma sequência de DNA é um dos principais desafios da genômica. Os íntrons contêm quase a mesma quantidade de dados que os éxons, desafiando a ideia de que eles são desprovidos de informação. Com isso, é importante avaliar a entropia dos íntrons e éxons de uma sequência de DNA.

Na teoria da informação, conforme Horta (2001), calcular a entropia envolve associar uma sequência, um modelo e a entropia da distribuição de probabilidades gerada pelo modelo aplicado à sequência. Quando o modelo gerador de uma mensagem é conhecido, a entropia pode ser calculada de maneira mais simples.

Nesse sentido, a entropia de uma mensagem gerada por um modelo depende da probabilidade de ser gerada por esse modelo. A incerteza é medida usando a função logarítmica, e a entropia é a média dessa incerteza para uma distribuição de probabilidades. Dessa maneira, a entropia está relacionada à distribuição de probabilidades e à previsi-

bilidade. Shannon definiu a entropia considerando uma cadeia gerada por um processo de Markov. Como o modelo gerador do DNA é desconhecido, são utilizados modelos da teoria da informação. A entropia do DNA não está completamente caracterizada, mas é explorada para identificar propriedades específicas. A entropia é utilizada para tentar identificar características do DNA e inferir sobre seu significado (Horta, 2001).

Além disso, o DNA é suscetível a mutações genéticas, que podem ser minimizadas pela estrutura rígida das bases nitrogenadas e pela redundância do código genético. As mutações podem alterar aminoácidos ou códons de parada, afetando a proteína resultante. Se a mutação prosseguir, ela pode ser transmitida aos descendentes e representar um ganho evolutivo; se não for viável, não será transmitida. Estimadores de entropia são usados para identificar padrões na estrutura primária do DNA, como regiões íntron/éxon. A entropia do material genético expresso em sequências de proteínas tende a ser maior do que a descartada, o que contraria a teoria atual, que sugere que as sequências descartadas ("íntrons") deveriam tolerar mais mudanças aleatórias do que as sequências retidas ("éxons").

4.1 COMPRESSÃO

Quando a informação é representada de uma maneira específica, torna-se possível transmiti-la. Sob essa ótica, codificar significa transformar a forma como a mensagem é representada, sem alterar seu significado original. Já a decodificação é o processo inverso, que restaura a mensagem à sua representação inicial. Um código nada mais é do que um método que traduz uma forma de expressão em outra. Um exemplo disso é a representação de textos em computadores, onde o bit (a menor unidade de informação) identifica apenas sinais elétricos, que podem estar ligados (1) ou desligados (0), conforme Horta (2001).

De acordo com Horta (2001), na computação, a transmissão de dados envolve custos. Para minimizá-los, utiliza-se um tipo específico de codificação chamado compressão, que consiste em reformular a mensagem para que ela ocupe menos espaço. Esse processo também pode ser chamado de codificação, enquanto sua reversão é a decodificação. O transmissor é quem codifica (ou compacta) os dados, e o receptor é quem os decodifica (ou descompacta).

De acordo com Lin e Costello Jr. (1983), um alfabeto é um conjunto finito de símbolos que podem ser usados para formar uma mensagem. A compressão visa otimizar a mensagem. A previsibilidade da mensagem influencia diretamente a eficácia da compressão. Quanto mais previsível for a mensagem, melhor será a compressão. O modelo é uma ferramenta utilizada para definir a distribuição de probabilidades da mensagem, isto é, sua previsibilidade.

4.2 CADEIAS DE MARKOV

De acordo com Fernandes Jr. e Vargas Jr. (2011), os processos de Markov constituem um tipo especial de processo estocástico que possui a propriedade de que as probabilidades associadas com o processo num dado instante do futuro dependem somente do estado presente, sendo, portanto, independentes dos eventos no passado. Os processos markovianos são caracterizados pelo que se designa como "falta de memória".

Definição 8. Um processo estocástico é uma coleção de variáveis randômicas $(X(t))$ indexadas por um parâmetro t pertecente a um conjunto T, em que $X(t)$ representa o estado do sistema no parâmetro (geralmente tempo) t.

Os processos estocásticos podem ser classificados como:

a) Em relação ao Estado $X(t)$:

   1. Estado Discreto: definido sobre um conjunto enumerável ou finito;
   2. Estado Contínuo: caso contrário.

b) Em relação ao Tempo (Parâmetro) t:

   1. Tempo Discreto: é finito ou enumerável;
   2. Tempo Contínuo: caso contrário.

Exemplo 3. Classificação dos processos estocásticos:

- İndice pluviométrico diário: Estado Contínuo e Tempo Discreto;

- Número de dias chuvosos: Estado Discreto e Tempo Discreto;

- Número de usuários em uma fila de banco em um determinado instante: Estado Discreto e Tempo Contínuo.

Definição 9. Um processo estocástico é dito ser um processo markoviano se:

$$P\left\{X(t_{k+1}) \leq x_{k+1} | X(t_k) = x_{k-1}, ..., X(t_1) = x_1, X(t_0) = x_0\right\},$$

$$P\left\{X(t_{k+1}) \leq x_{k+1} | X(t_k) = x_k\right\}$$

para $t_0 \leq t_1 \leq ... \leq t_k \leq t_{k+1} = 0, 1, ...$
e toda sequência $k_0, k_1, ..., k_{t-1}, k_1, k_{t+1}$.

A expressão apresentada anteriormente pode ser compreendida como: a probabilidade de qualquer evento futuro, dado qualquer evento passado e o estado presente $X(t_k) = x_k$, é independente do evento passado e depende somente do estado presente. Assim, um processo estocástico é dito ser um processo markoviano se o estado futuro depende apenas do estado presente e não dos estados passados.

Definição 10. Um processo markoviano é uma Cadeia de Markov quando as variáveis randômicas $X(t)$ estão definidas em um espaço de estados discretos E. Quando o tempo é discreto, a Cadeia de Markov é dita ser uma Cadeia de Markov em Tempo Discreto, ou seja,

$$P\left\{X(t_{k+1}) = x_{k+1} | X(k) = x_k, X(k-1) = x_{k-1}, ..., X(1) = x_1, X(0) = x_0\right\}$$

$$P\left\{X(k+1) = x_{k+1} | X(k) = x_k\right\},$$

para toda sequência $0, 1, ..., k-1, k, k+1$.

4.3 ENTROPIA DE UMA MENSAGEM

Pode ocorrer de alguma mensagem ser gerada a partir de algum método específico, o modelo gerador, o qual é responsável por estabelecer a forma, a estrutura e as características das mensagens geradas. As mensagens criadas desta forma refletem o modelo gerador se forem suficientemente grandes (Horta, 2001).

De acordo com Shannon (1948), ao caracterizar sua medida de entropia por meio de axiomas, postulou que essa dependia exclusivamente das probabilidades associadas aos eventos em análise. Sua interpretação considerou a ocorrência de fenômenos de natureza probabilística: antes da condução de um experimento, existe uma incerteza inerente à manifestação de um evento; após a execução do experimento, parte dessa incerteza é dissipada, e, conforme Shannon, obtém-se, assim, informação. Desse modo, a incerteza a priori pode ser convertida em uma quantidade mensurável de informação a posteriori, mediante a resolução do processo probabilístico.

Assim, a entropia de uma mensagem criada por um modelo gerador é uma função da probabilidade de ela ser gerada por esse modelo. Se M é o modelo gerador, S é alguma mensagem obtida através deste modelo e $P(S)$ é a probabilidade de S ser gerada por M então a entropia de $H(S)$ é definida por (Horta, 2001):

$$H(S) = -\log_2 P(S) \text{ bits.} \tag{1}$$

A probabilidade de uma mensagem obtida a partir de um modelo considera a probabilidade de cada símbolo na cadeia, o tamanho da cadeia e a probabilidade do símbolo de início que possui um valor especial, se o modelo gerador não for ergódico. Mensagens equiprováveis possuem o mesmo valor de entropia.

4.4 ENTROPIA DE UMA DISTRIBUIÇÃO DE PROBABILIDADES

De acordo com Horta (2001), a informação e a incerteza são conceitos que descrevem um processo de escolha de um ou mais elementos a partir de um conjunto. A incerteza é avaliada utilizando a função logarítmica. A base do logaritmo determina a unidade de medida da incerteza. Se for utilizada a base 10, a incerteza será expressa em número de dígitos. Por outro lado, ao utilizar a base 2, a incerteza será medida em bits. Shannon (1948) definiu a entropia da distribuição de probabilidades $P = \{P_1, P_2, ..., P_n\}$ a quantidade:

$$H(P) = -\sum_{\sigma}^{n} P_\sigma \log_2 P_\sigma. \tag{2}$$

A entropia é a medida da quantidade da informação de Shannon associada a uma distribuição de probabilidades P. Na Teoria da Informação, essa distribuição está relacionada a uma mensagem ou a uma parte de uma mensagem (sequência). Quando a entropia é calculada com base na distribuição de probabilidades dos símbolos de um alfabeto, a unidade de medida é bits por símbolo.

4.5 ENTROPIA DO PROCESSO DE MARKOV

O cálculo da entropia de uma cadeia que segue o processo de Markov é feito, de acordo com Horta (2001), pelo produto da probabilidade de cada estado do processo pela entropia do estado. A entropia de cada estado reflete sua incerteza. Considere $H(e)$ a entropia do estado e, expressa por:

$$H(e) = -\sum_{\sigma \in \Sigma} (\sigma | s_1 ... s_k) \log_2 P(s_1 ... s_k), \tag{3}$$

onde $s_1 ... s_k$ representa algum estado e. A entropia de cadeias que seguem alguma ordem do processo de Markov é calculada pela expressão:

$$H = \sum H(e) \cdot P(e), \tag{4}$$

onde $P(e)$ é a probabilidade de ocorrer o estado e na cadeia.

4.6 ENTROPIA CONDICIONAL

Sejam X e Y variáveis aleatórias. Suponha que X assume valores sobre o alfabeto de origem $X = \{x_1, ..., x_i\}$ e Y sobre o alfabeto de reconstrução $Y = \{y_1, ..., y_j\}$. Seja ainda $P(x_i) = P(X = x_i)$ e $P(y_i) = P(Y = y_i)$. Nesta seção foi apresentado que a entropia da distribuição de probabilidades é dada pela equação de Shannon:

$$H(X) = -\sum_{k=1}^{i} P(x_k) \log_2 P(x_k) \text{ e } H(Y) = -\sum_{k=1}^{j} P(y_k) \log_2 P(y_k).$$

A entropia condicional $H(X|Y)$ pode ser interpretada como a quantidade de incerteza restante sobre a variável X, dada a reconstrução da variável Y. Matematicamente, a entropia condicional $H(X|Y)$ é expressa por:

$$H(X|Y) = -\sum H(X|Y) \cdot P(e). \tag{5}$$

em que $P(e)$ é o estado condicional.

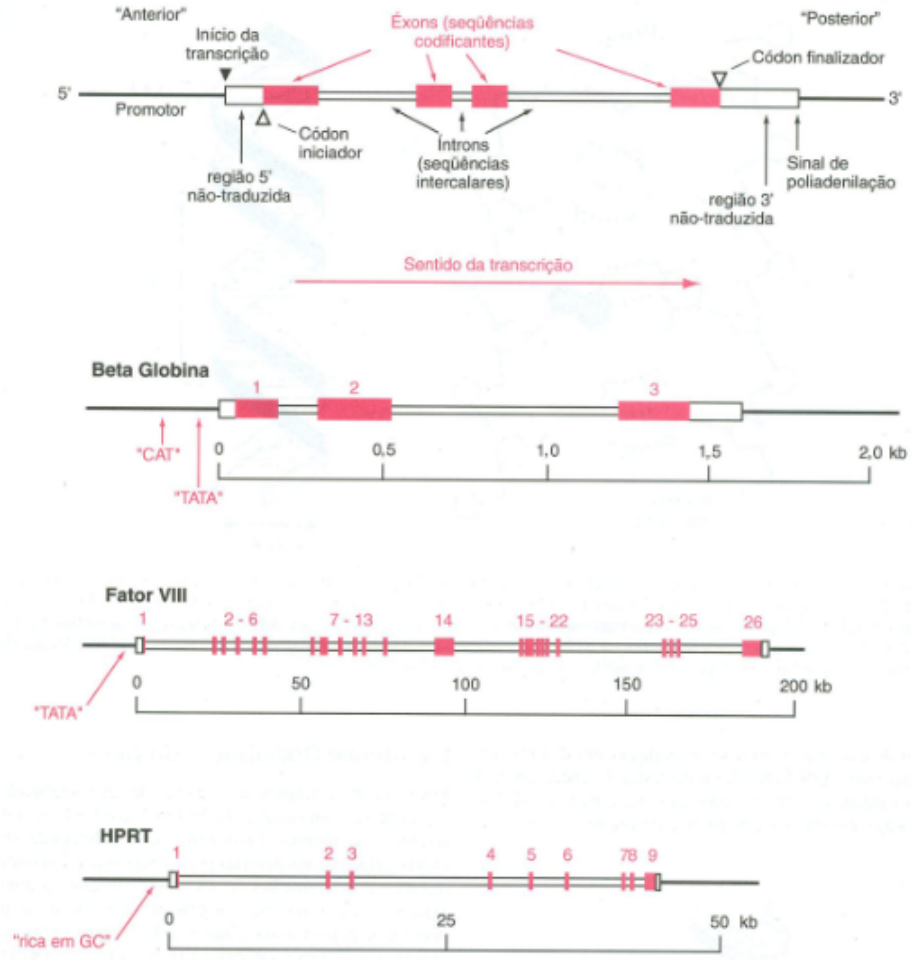
4.7 GENE

Um gene pode ser compreendido como uma sequência de DNA cromossômico necessária para a produção de um produto funcional, seja um polipeptídeo ou uma molécula funcional de RNA. A grande maioria dos genes é interrompida por uma ou mais regiões não codificantes. Estas sequências intercalares, chamadas íntrons, são inicialmente transcritas em RNA no núcleo, mas não estão presentes no RNAm final no citoplasma. Assim, a informação das sequências intrônicas normalmente não é representada no produto proteico final. Os íntrons alternam-se com sequências codificantes, ou éxons, que finalmente codificam a sequência de aminoácidos da proteína (Nussbaum, 2008).

Segundo Nussbaum (2008), um gene inclui não só as sequências realmente codificantes, mas também as sequências nucleotídicas adjacentes necessárias para a expressão apropriada do gene, ou seja, para a produção de uma molécula normal de RNAm, na quantidade, local e no momento corretos, durante o desenvolvimento ou durante o ciclo celular, como mostra a Figura 35. As sequências adjacentes de nucleotídeos fornecem os sinais moleculares de "início"e "fim"para a síntese do RNAm transcrito do gene.

Genes frequentemente fazem parte de grupos chamados "famílias de genes", caracterizados por semelhanças em suas sequências de DNA ou nas proteínas que produzem. Um exemplo importante é a família de genes das hemoglobinas, que inclui genes nos cromossomos 16 e 11, derivados de uma duplicação de um gene ancestral há cerca de 500

Figura 35 – Estrutura geral de um gene humano típico. Os exemplos de três genes humanos de importância médica são apresentados na parte inferior da figura. Os éxons individuais são numerados. Mutações diferentes no gene de β-globina causam uma variedade de hemoglobinopatias importantes. As mutações no gene de fator VIII causam hemofilia A. As mutações no gene de hipoxantina fosforibosiltransferase (HPRT) levam à síndrome de Lesch-Nyhan.



Fonte: Nussbaum (2008).

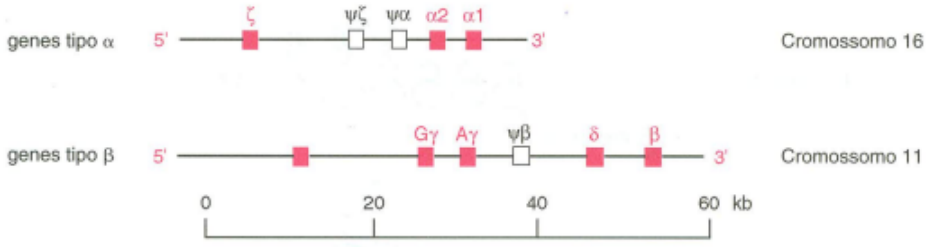
milhões de anos (Figura 36). Esses genes codificam diferentes cadeias de globina usadas em vários estágios de desenvolvimento, desde o embrião até a idade adulta. Dentro de cada grupo de genes, as sequências são mais semelhantes entre si do que com genes de outros grupos, indicando uma evolução por duplicações genéticas ao longo dos últimos 100 milhões de anos, conforme apresenta Nussbaum (2008).

Os genes de globina mantêm um padrão de éxons e íntrons bastante conservado, embora as sequências dentro dos íntrons tenham mudado mais do que as codificantes. Alguns genes de globina não produzem RNA ou proteínas e são chamados de pseudogenes. Esses pseudogenes são vestígios de genes que já foram funcionais, mas perderam a função devido a mutações. Podem se formar por duplicação genética ou retrotransposição, onde

uma cópia de RNAm é inserida no genoma.

A maior família de genes no genoma humano é a superfamília de imunoglobulinas, que inclui centenas de genes envolvidos em reconhecimento celular e funções do sistema imunológico e nervoso. Exemplos incluem genes para imunoglobulinas e receptores de células T, além de genes para moléculas de adesão celular em tecidos neurais.

Figura 36 – Organização cromossômica de dois grupos de genes globina humana. Os genes funcionais são indicados em rosa e os pseudogenes são indicados pelos boxes vazados.



Fonte: Nussbaum (2008).

4.7.0.1 O Gene de β-Globina

O gene de β-globina é uma proteína formada por 146 aminoácidos, composto por três éxons e dois íntrons, como mostrado na Figura 35.

As sequências de DNA necessárias para iniciar a transcrição do gene de β-globina estão localizadas no promotor, cerca de 200 pares de bases antes do início da transcrição. A Figura 37 mostra a sequência de DNA, a sequência de RNAm correspondente e a sequência dos primeiros 10 aminoácidos, mostrando como essas três camadas de informação estão relacionadas. O filamento 3' para 5' do DNA é o que serve como molde e é transcrito, enquanto o filamento 5' para 3' do DNA, que não é transcrito, corresponde diretamente à sequência 5' para 3' do RNAm (sendo idêntico, exceto pela substituição de T por U). Este filamento não transcrito é geralmente o que aparece na literatura científica e bancos de dados, conforme detalhamento apresentado em Nussbaum (2008).

O promotor do gene de β-globina, como muitos outros, é composto por pequenos elementos funcionais que interagem com proteínas específicas chamadas fatores de transcrição. Estes fatores regulam a transcrição e, no caso dos genes de globina, garantem que a expressão ocorra apenas em células eritróides, onde a hemoglobina é produzida. Uma parte crucial do promotor é o "TATA box", uma região rica em adeninas e timinas localizada 25 a 30 pares de bases antes do início da transcrição. O TATA box é fundamental para a localização precisa do início da transcrição, que no gene de β-globina está cerca de 50 pares de bases antes do início da tradução. Portanto, o gene possui cerca de 50 pares de

Figura 37 – Estrutura da sequência de nucleotídeos da ponta 5' do gene humano de β-globina no braço curto do cromossomo 11. A transcrição do filamento 3' para 5' (inferior) começa no ponto indicado para produzir o RNAm de β-globina. A matriz de leitura traducional é determinada pelo códon iniciador AUG (***); os códons subsequentes que especificam aminoácidos são indicados em rosa. As outras duas matrizes potenciais não são usadas.

Fonte: Nussbaum (2008).

bases transcritos, mas não traduzidos. Em outros genes, essa região 5' não traduzida (5' UTR) pode ser maior e conter íntrons Nussbaum (2008).

Outra região importante é o "CAT box"(ou CCAAT), que está algumas dezenas de pares de bases antes do início da transcrição. Alterações nessas sequências reguladoras podem reduzir significativamente a transcrição, evidenciando sua importância para a expressão normal do gene. Muitas mutações nessas áreas foram encontradas em pacientes com β-talassemia (Nussbaum, 2008).

Nem todos os promotores têm os elementos TATA e CAT boxes. Genes expressos continuamente em vários tecidos, conhecidos como genes de manutenção, geralmente não possuem essas sequências e, em vez disso, têm promotores ricos em citosinas e guaninas, situados em regiões chamadas ilhas de CpG. Essas sequências ricas em CG funcionam como pontos de ligação para fatores de transcrição específicos, conforme (Nussbaum, 2008).

De acordo com Nussbaum (2008), para que o gene de β-globina seja expresso corretamente durante o desenvolvimento, é essencial uma região reguladora localizada mais distante, conhecida como região controladora de locus (LCR), situada antes do gene da α-globina, como mostrado na Figura 36. Essa LCR é crucial para uma expressão adequada e em alto nível do gene. Desse modo, mutações que alteram ou removem essas sequências reguladoras ou a LCR comprometem ou impedem a expressão do gene da β-globina.

Em relação à recomposição do RNA, o transcrito primário de RNA do gene da β-globina possui dois éxons, de aproximadamente 100 e 850 pares de bases, que precisam ser unidos para formar um RNAm funcional. Esse processo de união é muito eficiente, com 95% dos transcritos de β-globina sendo processados corretamente. A recomposição é guiada por sequências específicas nas extremidades 5' e 3' dos íntrons. A sequência 5' contém nove nucleotídeos, incluindo o dinucleotídeo GT, que é praticamente invariável entre diferentes genes. A sequência 3' tem cerca de doze nucleotídeos, com os dinucleotídeos

AG sendo essenciais para a recomposição normal. Os pontos de corte dos íntrons não afetam diretamente a sequência do RNAm. Em alguns casos, como o íntron 1 do gene da β-globina, o íntron corta um códon específico.

A importância médica da recomposição do RNA é evidente porque mutações nas sequências conservadas dos limites íntron/éxon frequentemente prejudicam o processamento do RNA, diminuindo a quantidade de RNAm funcional da β-globina. Mutações nos dinucleotídeos GT ou AG geralmente impedem a remoção adequada dos íntrons. Diversas mutações nos sítios de corte foram identificadas em pacientes com β-talassemia, conforme Nussbaum (2008).

5 CÁLCULOS E ANÁLISES DAS ENTROPIAS ASSOCIADAS A SEQUÊNCIAS DE DNA

Neste capítulo serão apresentados os resultados associados à aplicação de elementos da Teoria da Informação, em particular, os cálculos e as análises das entropias associadas à sequência de nucleotídeos no gene humano completo de β-globina. As principais referências utilizadas para o desenvolvimento deste capítulo foram Nussbaum (2008) e Horta (2001).

A sequência de nucleotídeos no gene humano completo de β-globina representa apenas $0,000067\%$ da sequência de todo o genoma humano e foi selecionada como forma de analisar os cálculos associados às entropias da mesma. De acordo com Nussbaum (2008) essa sequência apresenta como principais características do gene de β-globina os elementos de sequência promotora conservados, os limites de íntrons e éxons, os sítios de corte do RNA e os códons iniciadores e finalizadores, conhecidos por mutações ocorridas no gene de β-globina, o que torna relevante seu estudo, buscando assim uma conexão com possíveis estudos associados a mutações que podem ocorrer em pesquisas futuras.

Vale ressaltar que a utilização de estimadores de entropia em cadeias de DNA visa identificar "padrões de comportamento"na sua estrutura primária. A partir desses padrões, é possível localizar regiões do DNA com significado específico, como o limite entre íntron e éxon, ou ainda descobrir padrões que possam indicar a presença de significados biológicos. Também, os padrões de comportamento buscam correlações entre DNAs distintos, bem como uma possível compreensão do código genético.

5.1 TEORIA DA INFORMAÇÃO EM SEQUÊNCIAS DE DNA

A distribuição de probabilidades obtida a partir de um modelo fornece a previsibilidade de uma mensagem. Normalmente, uma mensagem é escrita sobre um alfabeto Σ. O tamanho de uma mensagem (n) é o número de símbolos que ocorrem nela.

Para o caso biológico, considere como alfabeto $\Sigma = \{\text{A, C, G, U}\}$, onde A, C, G, U, correspondem às bases adenina, citosina, guanina e uracila, respectivamente.

Seja a mensagem associada à sequência de nucleotídeos no gene humano completo β-globina, extraída de Nussbaum (2008):

$$S=\text{ "Éxon 1, Éxon 2, Éxon 3",}$$

sendo que:

Éxon 1: ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGG
GCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGg,

Éxon 2: CTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGG
ATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCA
AGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGG
GCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTG
AGAACTTCAGG,

Éxon 3: CTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAA
GAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGG
CTAATGCCCTGGCCCACAAGTATCACTAAGCT,

a qual possui 447 símbolos.

Os cálculos apresentados a seguir seguem os modelos propostos por Horta (2001).

5.2 ENTROPIA ASSOCIADA AO MODELO M_1

Seja M_1 o modelo associado ao processo de contagem das ocorrências de cada símbolo do alfabeto e obtenção da probabilidade de cada símbolo. Seja $P(x)$ a probabilidade de ocorrer o símbolo x em S. Utilizando o modelo M_1 para o cálculo da distribuição de probabilidades, tem-se para S:

$$P(A) = \frac{88}{447}; P(C) = \frac{115}{447}; P(G) = \frac{137}{447}; P(U) = \frac{107}{447},$$

em que $P(A), P(C), P(G)$ e $P(U)$, representam as probabilidades associadas à ocorrência de adenina, citosina, guanina e uracila, respectivamente.

De acordo com Horta (2001), a entropia para M_1 é calculada diretamente pela expressão de Shannon (2).

Logo:

$$H(M_1) = -\left(P_A \log_2 P_A + P_C \log_2 P_C + P_G \log_2 P_G + P_U \log_2 P_U\right)$$

$$H(M_1) = -\left(\tfrac{88}{447} \log_2 \tfrac{88}{447} + \tfrac{115}{447} \log_2 \tfrac{115}{447} + \tfrac{137}{447} \log_2 \tfrac{137}{447} + \tfrac{107}{447} \log_2 \tfrac{107}{447}\right)$$

$$H(M_1) = -\left[(-0,4616) + (-0,5039) + (-0,5229) + (-0,4937)\right]$$

$$H(M_1) = 1,9821.$$

O cálculo da entropia para S a partir da distribuição de probabilidades fornecida pelo modelo M_1 resultou em $H(M_1) = 1,9821$ bits/símbolo.

5.3 ENTROPIA ASSOCIADA AO MODELO M_2

Seja M_2 o modelo aplicado pelo processo de consideração da existência de três estados vinculados à mensagem, conforme apresentado em Horta (2008).

No primeiro estado, são considerados os símbolos que ocorrem nas posições 146, 149, 152, 155, 158, 161, 164, 167, 170, 173, 176, 179, 182, 185, 188, 191, 194, 197, 200, 203, 206, 209, 212, 215, 218, 221, 224, 227, 230, 233, 236, 368, 371, 374, 377, 380, 383, 386, 389, 392, 395, 398, 401, 404, 407, 410, 413, 416, 419, 422, 425, 428, 431, 434, 437, 440, 443, 446, 449, 452, 455, 458, 461, 464, 467, 470, 473, 476, 479, 482, 485, 488, 491, 494, 497, 500, 503, 506, 509, 512, 515, 518, 521, 524, 527, 530, 533, 536, 539, 542, 545, 548, 551, 554, 557, 560, 563, 566, 569, 572, 575, 578, 581, 584, 587, 590, 1441, 1444, 1447, 1450, 1453, 1456, 1459, 1462, 1465, 1468, 1471, 1474, 1477, 1480, 1483, 1486, 1489, 1492, 1495, 1498, 1501, 1504, 1507, 1510, 1513, 1516, 1519, 1522, 1525, 1528, 1531, 1534, 1537, 1540, 1543, 1546, 1549, 1552, 1555, 1558, 1561, 1564, 1567 e 1570 de S.

No segundo estado, são considerados os símbolos que ocorrem nas posições 147, 150, 153, 156, 159, 162, 165, 168, 171, 174, 177, 180, 183, 186, 189, 192, 195, 198, 201, 204, 207, 210, 213, 216, 219, 222, 225, 228, 231, 234, 237, 369, 372, 375, 378, 381, 384, 387, 390, 393, 396, 399, 402, 405, 408, 411, 414, 417, 420, 423, 426, 429, 432, 435, 438, 441, 444, 447, 450, 453, 456, 459, 562, 465, 468, 471, 474, 477, 480, 483, 486, 489, 492, 495, 498, 501, 504, 507, 510, 513, 516, 519, 522, 525, 528, 531, 534, 537, 540, 543, 546, 549, 552, 555, 558, 561, 564, 567, 570, 573, 576, 579, 582, 585, 588, 1442, 1445, 1448, 1451, 1454, 1457, 1460, 1463, 1466, 1469, 1472, 1475, 1478, 1481, 1484, 1487, 1490, 1493, 1496, 1499, 1502, 1505, 1508, 1511, 1514, 1517, 1520, 1523, 1526, 1529, 1532, 1535, 1538, 1541, 1544, 1547, 1550, 1553, 1556, 1559, 1562, 1565, 1568 e 1571 de S.

No terceiro estado, somente são considerados os símbolos que ocorrem nas posições 148, 151, 154, 157, 160, 163, 166, 169, 172, 175, 178, 181, 184, 187, 190, 193, 196, 199, 202, 205, 208, 211, 214, 217, 220, 223, 226, 229, 232, 235, 238, 370, 373, 376, 379, 382, 385, 388, 391, 394, 397, 400, 403, 406, 409, 412, 415, 418, 421, 424, 427, 430, 433, 436, 439, 442, 445, 448, 451, 454, 457, 460, 463, 466, 469, 472, 475, 478, 481, 484, 487, 490, 493, 496, 499, 502, 505, 508, 511, 514, 517, 520, 523, 526, 529, 532, 535, 538, 541, 544, 547, 550, 553, 556, 559, 562, 565, 568, 571, 574, 577, 580, 583, 586, 589, 1443, 1446, 1449, 1452, 1455, 1458, 1461, 1464, 1467, 1470, 1473, 1476, 1479, 1482, 1485, 1488, 1491, 1494, 1497, 1500, 1503, 1506, 1509, 1512, 1515, 1518, 1521, 1524, 1527, 1530, 1533, 1536, 1539, 1542, 1545, 1548, 1551, 1554, 1557, 1560, 1563, 1566, 1569 e 1572 de S.

O tamanho da mensagem S é 447, sendo assim, cada estado ocorre exatamente 149 vezes.

As probabilidades de ocorrência de cada símbolo em cada estado estão apresentadas na Tabela 27.

Tabela 27 – Probabilidades de ocorrer cada símbolo em cada estado.

Estado	A	C	G	U
1	$P(1,\text{A}) = \frac{23}{149}$	$P(1,\text{C}) = \frac{32}{149}$	$P(1,\text{G}) = \frac{55}{149}$	$P(1,\text{U}) = \frac{39}{149}$
2	$P(2,\text{A}) = \frac{33}{149}$	$P(2,\text{C}) = \frac{42}{149}$	$P(2,\text{G}) = \frac{41}{149}$	$P(2,\text{U}) = \frac{33}{149}$
3	$P(3,\text{A}) = \frac{32}{149}$	$P(3,\text{C}) = \frac{41}{149}$	$P(3,\text{G}) = \frac{41}{149}$	$P(3,\text{U}) = \frac{35}{149}$

O cálculo da entropia de cada estado é dado por:

$$H(e) = -\sum_{\sigma \in \Sigma} P(e,\sigma) \log_2 P(e,\sigma), \tag{6}$$

em que $H(e)$ corresponde a entropia do estado e.

Dessa maneira,

$$\begin{aligned} H(1) &= -[(P(1,A)\log_2 P(1,A)) + (P(1,C)\log_2 P(1,C)) + \\ &\quad +(P(1,G)\log_2 P(1,G)) + (P(1,U)\log_2 P(1,U))] \end{aligned}$$

$$\begin{aligned} H(1) &= -\left[\left(\frac{23}{149}\log_2\frac{23}{149}\right) + \left(\frac{32}{149}\log_2\frac{32}{149}\right) + \left(\frac{55}{149}\log_2\frac{55}{149}\right) + \right. \\ &\quad \left. + \left(\frac{39}{149}\log_2\frac{39}{149}\right)\right] \end{aligned}$$

$$H(1) = -[(-0,4161) + (-0,4766) + (-0,5307) + (-0.5061)]$$

$$H(1) = 1,9295.$$

$$\begin{aligned} H(2) &= -[(P(2,A)\log_2 P(2,A)) + (P(2,C)\log_2 P(2,C)) + \\ &\quad +(P(2,G)\log_2 P(2,G)) + (P(2,U)\log_2 P(2,U))] \end{aligned}$$

$$\begin{aligned} H(2) &= -\left[\left(\frac{33}{149}\log_2\frac{33}{149}\right) + \left(\frac{42}{149}\log_2\frac{42}{149}\right) + \left(\frac{41}{149}\log_2\frac{41}{149}\right) + \right. \\ &\quad \left. + \left(\frac{33}{149}\log_2\frac{33}{149}\right)\right] \end{aligned}$$

$$H(2) = -[(-0,4817) + (-0,5149) + (-0,5122) + (-0,4817)]$$

$$H(2) = 1,9905.$$

$$\begin{aligned} H(3) &= -[(P(3,A)\log_2 P(3,A)) + (P(3,C)\log_2 P(3,C)) + \\ &\quad +(P(3,G)\log_2 P(3,G)) + (P(3,U)\log_2 P(3,U))] \end{aligned}$$

$$\begin{aligned} H(3) &= -\left[\left(\frac{32}{149}\log_2\frac{32}{149}\right) + \left(\frac{41}{149}\log_2\frac{41}{149}\right) + \left(\frac{41}{149}\log_2\frac{41}{149}\right) + \right. \\ &\quad \left. + \left(\frac{35}{149}\log_2\frac{35}{149}\right)\right] \end{aligned}$$

$$H(3) = -[(-0,4766) + (-0,5122) + (-0,5122) + (-0,4909)]$$

$$H(3) = 1,9919.$$

Como $H(M_2)$ é calculado por (4):

e tem-se que $P(e) = P(1) = P(2) = P(3) = \frac{149}{447} = 0,3333$, temos:

$$H(M_2) = [(P(1)H(1)) + (P(2)H(2)) + (P(3)H(3))]$$

$$H(M_2) = [(0,3333 \cdot 1,9295) + (0,3333 \cdot 1,9905) + (0,3333 \cdot 1,9919)]$$

$$H(M_2) = [(0,6431) + (0,6634) + (0,6639)]$$

$$H(M_2) = 1,9704.$$

O cálculo da entropia para S a partir da distribuição de probabilidades fornecida pelo modelo M_2 resultou em $H(M_2) = 1,9704$ bits/símbolo.

Dessa maneira, a escolha no modelo interfere diretamente na estimativa da entropia da mensagem. A entropia da mensagem utilizando M_2 foi menor que a entropia da mesma mensagem utilizando M_1. O conhecimento da mensagem possibilita a escolha de um modelo mais adequado e com isto se reduz significativamente sua estimativa da entropia.

5.4 ANÁLISE DA SEQUÊNCIA COMPLETA

A partir da análise dos éxons da sequência, será apresentada a seguir, a análise da sequência completa, como uma maneira de comparar os resultados obtidos em cada modelo.

Dado um alfabeto $\Sigma = \{A, C, G, U\}$ e a mensagem:

$S =$ "Sequência de nucleotídeos no gene humano completo de β-globina",

em que a Sequência de nucleotídeos no gene humano completo de β-globina é expressa por:

5'. . . . agccacaccctagggttggccaatctactcccaggagcagggagggcaggagccagggctgggcataaaagtcaggg
cagagccatctattgcttACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAAC
AGACACC ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTG
TGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG gttggtat
caaggttacaagacaggtttaaggagaccaatagaaactgggcatgtggagacagagaagactcttgggtttctgataggcact
gactctctctctgcctattggtctattttcccaccctt**ag** CTGCTGGTGGTCTACCCTTGGACCCAGAG
GTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCC
TAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGG
CTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGT
GACAAGCTGCACGTGGATCCTGAGAACTTCAGG gtgagtctatgggaccttgatgttttctttccc
cttcttttctatggttaagttcatgrcataggaaggggagaagtaacagggtacagtttagaatgggaaacagacgaatgattgc
atcagtgtggaagtctcaggatcgttttagtttcttttatttgctgttcataacaattgttttcttttgtttaattcttgctttctttttt
tttcttctccgcaatttttactattatacttaatgccttaacattgtgtataacaaaaggaaatatctctgagatacattaagtaact
taaaaaaaaactttacacagtctgcctagtacattactatttggaatatatgtgtgcttatttgcatattcataatgtccctacttta
ttttcttttatttttaattgatacataatcattatacatatttatgggttaaagtgtaatgttttaatatgtgtacacatattgaccaa
atcagggtaattttgcatttgtaattttaaaaaatgctttcttctttttaatatactttttttgtttatcttatttctaatactttccctaat
ctctttctttcagggcaataatgatacaatgtatcatgcctctttgcaccattctaaagaataacagtgataatttctgggttaagg
caatagcaatatttctgcatataaatatttctgcatataaattgtaactgatgtaagaggtttcatattgctaatagcagctacaat
ccagctaccattctgcttttatttatggttgggataaggctggattattctgagtccaagctaggccctttgctaatcatattcat
acctcttatcttcctcccac**ag** CTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACT
TTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCT
GGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAAGCTCGCTTTCTTGCTGTC
CAATTTCTATTAAAGGTTCCTTTGTTCCCTAAFTCCAACTACTAAACTGGGGGA
TATTATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTTC
ATTGCaatgatgtatttaaattatttctgaatattttactaaaaagggaatgtgggaggtcagtgcatttaaaacataaagaa
atgatgagctgttcaaaccttgggaaaatacactatatcttaaactccatgaaagaaggtgaggctgcaaccagctaatgcaca
ttggcaacagcccctgatgcctatgccttattcatccctcagaaaaggattcttgtagaggcttga . . . . 3'

com 1929 símbolos.

A sequência de nucleotídeos do gene humano completo de β-globina é apresentada no sentido biológico 5'- 3' do gene. As letras maiúsculas representam sequências que correspondem ao RNAm final, sendo os éxons representados pela cor rosa. As letras minúsculas indicam íntrons e sequências flanqueadoras, sendo representados pela cor preta. O códon iniciador ATG (AUG no RNAm) e o códon finalizador TAA (UAA no RNAm)

são mostrados em vermelho. Os dinucleotídeos GT e AG, importantes para recomposição do RNA e nas junções íntron/éxon, são mostrados em azul.

5.4.1 Cálculos para o Modelo M_1

Seja M_1 o modelo aplicado pelo processo de contagem das ocorrências de cada símbolo do alfabeto e obtenção da probabilidade de cada símbolo pela divisão do contador do símbolo pelo tamanho da mensagem. Seja $P(x)$ a probabilidade de ocorrer o símbolo x em S. Utilizando o modelo M_1 para o cálculo da distribuição de probabilidades, tem-se:

$$P(\mathrm{A}) = \tfrac{511}{1929};\ P(\mathrm{C}) = \tfrac{384}{1929};\ P(\mathrm{G}) = \tfrac{404}{1929};\ P(\mathrm{U}) = \tfrac{630}{1929}.$$

A entropia para M_1 é calculada diretamente pela expressão de Shannon (2):

$$H(M_1) = -\left(P_A \log_2 P_A + P_C \log_2 P_C + P_G \log_2 P_G + P_U \log_2 P_U\right)$$

$$H(M_1) = -\left(\frac{511}{1929}\log_2\frac{511}{1929} + \frac{384}{1929}\log_2\frac{384}{1929} + \frac{404}{1929}\log_2\frac{404}{1929} + \frac{630}{1929}\log_2\frac{630}{1929}\right)$$

$$H(M_1) = -\left[(-0,5077) + (-0,4636) + (-0,4724) + (-0,5273)\right]$$

$$H(M_1) = 1,9710.$$

O cálculo da entropia para S a partir da distribuição de probabilidades fornecida pelo modelo M_1 resultou em $H(M_1) = 1,9710$ bits/símbolo.

5.4.2 Cálculos para o Modelo M_2

Seja M_2 o modelo aplicado pelo processo de consideração da existência de três estados vinculados à mensagem.

No primeiro estado, são considerados os símbolos que ocorrem nas posições $1, 4, 7, 10, 13, 16, 19, 22, 25, 28, 31, 34, 37, 40, 43, 46, 49, 52, 55, 58, 61, 64, 67, 70, 73, 76, 79, 82, 85, 88, 91, 94, 97, 100, 103, 106, 109, 112, 115, 118, 121, 124, 127, 130, 133, 136, 139, 142, 145, 148, 151, 154, 157, 160, 163, 166, 169, 172, 175, 178, 181, 184, 187, 190, 193, 196, 199, 202, 205, 208, 211, 214, 217, 220, 223, 226, 229, ...$ de S.

No segundo estado, são considerados os símbolos que ocorrem nas posições $2, 5, 8, 11, 14, 17, 20, 23, 26, 29, 32, 35, 38, 41, 44, 47, 50, 53, 56, 59, 62, 65, 68, 71, 74, 77, 80, 83, 86, 89, 92,$

$95, 98, 101, 104, 107, 110, 113, 116, 119, 122, 125, 128, 131, 134, 137, 140, 144, 146, 149, 152, 155, 158, 161, 164, 167, 170, 173, 176, 179, 182, 185, 188, 191, 194, 197, 200, 203, 206, 209, 212, 215, 218, 221, 224, 227, 230, ...$ de S.

No terceiro estado, somente são considerados os símbolos que ocorrem nas posições $3, 6, 9, 12, 15, 18, 21, 24, 27, 30, 33, 36, 39, 42, 45, 48, 51, 54, 57, 60, 63, 66, 69, 72, 75, 78, 81, 84, 87, 90, 93, 96, 99, 102, 105, 108, 111, 114, 117, 120, 123, 126, 129, 132, 135, 138, 141, 144, 147, 150, 153, 156, 159, 162, 165, 168, 171, 174, 177, 180, 183, 186, 189, 192, 195, 198, 201, 204, 207, 210, 213, 216, 219, 222, 225, 228, ...$ de S.

O tamanho da mensagem S é 1929, assim cada estado ocorre exatamente 643 vezes.

As probabilidades de ocorrência de cada símbolo em cada estado estão apresentadas na Tabela 28.

Tabela 28 – Probabilidades de ocorrer cada símbolo em cada estado.

Estado	A	C	G	U
1	$P(1,A)=\frac{165}{643}$	$P(1,C)=\frac{131}{643}$	$P(1,G)=\frac{138}{643}$	$P(1,U)=\frac{209}{643}$
2	$P(2,A)=\frac{165}{643}$	$P(2,C)=\frac{128}{643}$	$P(2,G)=\frac{129}{643}$	$P(2,U)=\frac{221}{643}$
3	$P(3,A)=\frac{181}{643}$	$P(3,C)=\frac{125}{643}$	$P(3,G)=\frac{122}{643}$	$P(3,U)=\frac{215}{643}$

O cálculo da entropia de cada estado é dado por:

$$H(e) = -\sum_{\sigma\in\Sigma} P(e,\sigma)\log_2 P(e,\sigma),$$

em que $H(e)$ corresponde a entropia do estado.

Desse modo,

$$\begin{aligned} H(1) &= -[(P(1,A)\log_2 P(1,A)) + (P(1,C)\log_2 P(1,C)) + \\ &\quad +(P(1,G)\log_2 P(1,G)) + (P(1,U)\log_2 P(1,U))] \end{aligned}$$

$$\begin{aligned} H(1) &= -\left[\left(\frac{165}{643}\log_2\frac{165}{643}\right) + \left(\frac{131}{643}\log_2\frac{131}{643}\right) + \left(\frac{138}{643}\log_2\frac{138}{643}\right) + \right. \\ &\quad \left. + \left(\frac{209}{643}\log_2\frac{209}{643}\right)\right] \end{aligned}$$

$$H(1) = -[(-0,5035) + (-0,4676) + (-0,4765) + (-0,5270)]$$

$$H(1) = 1,9746.$$

$$\begin{aligned} H(2) &= -[(P(2,A)\log_2 P(2,A)) + (P(2,C)\log_2 P(2,C)) + \\ &\quad +(P(2,G)\log_2 P(2,G)) + (P(2,U)\log_2 P(2,U))] \end{aligned}$$

$$\begin{aligned} H(2) &= -\left[\left(\frac{165}{643}\log_2\frac{165}{643}\right) + \left(\frac{128}{643}\log_2\frac{128}{643}\right) + \left(\frac{129}{643}\log_2\frac{129}{643}\right) + \right. \\ &\quad \left. + \left(\frac{221}{643}\log_2\frac{221}{643}\right)\right] \end{aligned}$$

$$H(2) = -[(-0,5035) + (-0,4636) + (-0,4649) + (-0,5296)]$$

$$H(2) = 1,9616.$$

$$\begin{aligned} H(3) &= -[(P(3,A)\log_2 P(3,A)) + (P(3,C)\log_2 P(3,C)) + \\ &\quad +(P(3,G)\log_2 P(3,G)) + (P(3,U)\log_2 P(3,U))] \end{aligned}$$

$$\begin{aligned} H(3) &= -\left[\left(\frac{181}{643}\log_2\frac{181}{643}\right) + \left(\frac{125}{643}\log_2\frac{125}{643}\right) + \left(\frac{122}{643}\log_2\frac{122}{643}\right) + \right. \\ &\quad \left. + \left(\frac{215}{643}\log_2\frac{215}{643}\right)\right] \end{aligned}$$

$$H(3) = -[(-0,5148) + (-0,4593) + (-0,4550) + (-0,5285)]$$

$$H(3) = 1,9576.$$

Como $H(M_2)$ é calculado por:

$$H(M_2) = \sum_{e=1}^{3} P(e)H(e)$$

e tem-se que $P(e) = P(1) = P(2) = P(3) = \frac{643}{1929} = 0,3333$, temos:

$$H(M_2) = [(P(1)H(1)) + (P(2)H(2)) + (P(3)H(3))]$$

$$H(M_2) = [(0,3333 \cdot 1,9746) + (0,3333 \cdot 1,9616) + (0,3333 \cdot 1,9576)]$$

$$H(M_2) = [(0,6581) + (0,6538) + (0,6525)]$$

$$H(M_2) = 1,9644.$$

O cálculo da entropia para S a partir da distribuição de probabilidades fornecida pelo modelo M_2 resultou em $H(M_2) = 1,9644$ bit/símbolo.

Nesse caso, a entropia da mensagem utilizando M_2 também foi menor que a entropia da mesma mensagem utilizando M_1.

Este modelo pode ser caracterizado como modelo adaptativo, uma vez que há um modelo inicial que existe tanto no transmissor quanto no receptor e, à medida que o código é transmitido, alterna da mesma forma o modelo, havendo um custo para sua manutenção. Nesse sentido, há o envio, símbolo a símbolo, para que tanto o transmissor quanto o receptor possam atualizar suas distribuições de probabilidades. Além disso, uma das características de um modelo adaptativo é que dado um conjunto de classes condicionais com k membros, um alfabeto de λ símbolos e uma mensagem de tamanho n, os modelos adaptativos possuem um limite para o tamanho da representação da mensagem em função de k, λ e n, conforme apresentado em Horta (2001).

A Tabela 29 apresenta uma comparação dos valores obtidos para as entropias relacionadas aos Modelos M_1 e M_2, para os casos envolvendo os éxons na sequência e a sequência completa.

Tabela 29 – Comparação entre os modelos utilizados no cálculo da entropia.

	Modelo M_1	Modelo M_2
Éxons	1,9827	1,9704
Sequência completa	1,9710	1,9644

5.5 CONTEXTOS FINITOS E ORDEM DO MODELO

Conforme apresentado em Horta (2001), no modelo de contextos baseados em sequências, a probabilidade do próximo símbolo é determinada pela sequência de alguns dos símbolos imediatamente anteriores. Este modelo também é conhecido como modelo de Markov, uma vez que os contextos podem ser vistos como os estados do processo de Markov. A ordem do modelo refere-se ao número de símbolos anteriores considerados para determinar a probabilidade do próximo símbolo. Existem duas exceções a esta regra: a ordem -1, usada quando não há informações sobre a mensagem, resultando em uma distribuição equiprovável entre os símbolos do alfabeto, e a ordem 0, onde a distribuição

de probabilidades é obtida pela contagem das ocorrências de cada símbolo na mensagem. Desse modo, nesses dois casos específicos, os símbolos anteriores não são considerados. O número de contextos para um modelo varia proporcionalmente ao tamanho do alfabeto e exponencialmente com a ordem do modelo.

Dessa maneira, quando se utiliza um modelo de ordem n, os primeiros n símbolos da mensagem não podem ser codificados por este, pois não possuem símbolos anteriores. Neste contexto, tanto o codificador quanto o decodificador precisam ter um modelo inicial para representar os primeiros n símbolos da mensagem, conforme apresentado em Horta (2001).

Seja o alfabeto Σ ={A, C, G, U} e $\lambda = 4$.

No modelo de ordem -1, como $\lambda = 4$, tem-se: P(A)= P(C)= P(G)= P(U)= 0,25.

Calculando a entropia: $H(P) = -\sum_{\sigma \in \Sigma} P_\sigma \log_2 P_\sigma$.

Assim,

$$H(S) = -[(0,25 \log_2 0,25) + (0,25 \log_2 0,25) + (0,25 \log_2 0,25) + (0,25 \log_2 0,25)]$$

$$H(S) = -[(0,5) + (0,5) + (0,5) + (0,5)]$$

$$H(S) = 2 \text{ bits/símbolo.}$$

No modelo de ordem 0, tem-se as seguintes probabilidades:

$$P(A) = \frac{88}{447}; P(C) = \frac{115}{447}; P(G) = \frac{137}{447}; P(U) = \frac{107}{447}.$$

O cálculo da entropia no modelo de ordem 0 é apresentado a seguir.

$$H(S) = -\left[\left(\frac{88}{447}\log_2\frac{88}{447}\right) + \left(\frac{115}{447}\log_2\frac{115}{447}\right) + \left(\frac{137}{447}\log_2\frac{137}{447}\right) + \left(\frac{107}{447}\log_2\frac{107}{447}\right)\right]$$

$$H(S) = -[(-0,4616) + (-0,5039) + (-0,5229) + (-0,4937)]$$

$$H(S) = 1,9821 \text{ bit/símbolo.}$$

No modelo de ordem 1, tem-se as seguintes probabilidades:

$$P(A|A) = \frac{23}{88}; P(C|A) = \frac{24}{88}; P(G|A) = \frac{28}{88}; P(U|A) = \frac{13}{88};$$
$$P(A|C) = \frac{30}{115}; P(C|C) = \frac{34}{115}; P(G|C) = \frac{5}{115}; P(U|C) = \frac{45}{115};$$
$$P(A|G) = \frac{25}{137}; P(C|G) = \frac{36}{137}; P(G|G) = \frac{43}{137}; P(U|G) = \frac{31}{137};$$
$$P(A|U) = \frac{9}{107}; P(C|U) = \frac{20}{107}; P(G|U) = \frac{60}{107}; P(U|U) = \frac{17}{107};$$

A partir do modelo de ordem 1, é necessário calcular a entropia de cada contexto, a qual é calculada pela expressão a seguir, antes do cálculo da entropia do modelo.

$$H(S|s) = -\sum_{\sigma \in \Sigma} P(\sigma|s) \log_2 P(\sigma|s). \tag{7}$$

Assim, para o primeiro contexto, tem-se que:

$$\begin{aligned} H(S|A) &= -[(P(A|A)\log_2 P(A|A) + (P(C|A)\log_2 P(C|A) + (P(G|A)\log_2 P(G|A) + \\ &\quad +(P(U|A)\log_2 P(U|A)] \end{aligned}$$

$$H(S|A) = -\left[\left(\frac{23}{88}\log_2\frac{23}{88}\right) + \left(\frac{24}{88}\log_2\frac{24}{88}\right) + \left(\frac{28}{88}\log_2\frac{28}{88}\right) + \left(\frac{13}{88}\log_2\frac{13}{88}\right)\right]$$

$$H(S|A) = -[(-0,5060) + (-0,5112) + (-0,5257) + (-0,4076)]$$

$$H(S|A) = 1,9505.$$

Para o segundo contexto, tem-se que:

$$\begin{aligned} H(S|C) &= -[(P(A|C)\log_2 P(A|C) + (P(C|C)\log_2 P(C|C) + (P(G|C)\log_2 P(G|C) + \\ &\quad +(P(U|C)\log_2 P(U|C)] \end{aligned}$$

$$\begin{aligned} H(S|C) &= -\left[\left(\frac{30}{115}\log_2\frac{30}{115}\right) + \left(\frac{34}{115}\log_2\frac{34}{115}\right) + \left(\frac{5}{115}\log_2\frac{5}{115}\right) + \right. \\ &\quad \left. + \left(\frac{45}{115}\log_2\frac{45}{115}\right)\right] \end{aligned}$$

$$H(S|C) = -[(-0,5057) + (-0,5198) + (-0,1967) + (-0,5297)]$$

$$H(S|C) = 1,7519.$$

Para o terceiro contexto, tem-se que:

$$\begin{aligned} H(S|G) &= -[(P(A|G)\log_2 P(A|G) + (P(C|G)\log_2 P(C|G) + (P(G|G)\log_2 P(G|G) + \\ &\quad +(P(U|G)\log_2 P(U|G)] \end{aligned}$$

$$\begin{aligned} H(S|G) &= -\left[\left(\frac{25}{137}\log_2\frac{25}{137}\right) + \left(\frac{36}{137}\log_2\frac{36}{137}\right) + \left(\frac{43}{137}\log_2\frac{43}{137}\right) + \right. \\ &\quad \left. + \left(\frac{31}{137}\log_2\frac{31}{137}\right)\right] \end{aligned}$$

$$H(S|G) = -[(-0,4478) + (-0,5066) + (-0,5247) + (-0,4851)]$$

$$H(S|G) = 1,9642.$$

Por fim, para o quarto contexto, tem-se que:

$$\begin{aligned} H(S|U) &= -[(P(A|U)\log_2 P(A|U) + (P(C|U)\log_2 P(C|U) + (P(G|U)\log_2 P(G|U) + \\ &\quad +(P(U|U)\log_2 P(U|U)] \end{aligned}$$

$$\begin{aligned} H(S|U) &= -\left[\left(\frac{9}{107}\log_2\frac{9}{107}\right) + \left(\frac{20}{107}\log_2\frac{20}{107}\right) + \left(\frac{60}{107}\log_2\frac{60}{107}\right) + \right. \\ &\quad \left. + \left(\frac{17}{107}\log_2\frac{17}{107}\right)\right] \end{aligned}$$

$$H(S|U) = -[(-0,3004) + (-0,4522) + (-0,4680) + (-0,4217)]$$

$$H(S|U) = 1,6423.$$

Após o cálculo da entropia de cada contexto, a entropia para o modelo de ordem 1 é calculada pela expressão (4):

$$\begin{aligned} H(S) &= \left[\left(\frac{88}{447}\cdot 1,9108\right) + \left(\frac{115}{447}\cdot 1,7431\right) + \left(\frac{137}{447}\cdot 1,9363\right) \right. \\ &\quad \left. + \left(\frac{107}{447}\cdot 1,6423\right)\right] \end{aligned}$$

$$H(S) = [(0,1969\cdot 1,9505) + (0,2573\cdot 1,7519) + (0,3065\cdot 1,9642) + (0,2394\cdot 1,6423)]$$

$$H(S) = [(0,3840) + (0,4508) + (0,6020) + (0,3932)]$$

$$H(S) = 1,8300 \text{ bit/símbolo.}$$

No modelo de ordem 2, tem-se as seguintes probabilidades:

$$P(A|AA) = \frac{3}{23}; P(C|AA) = \frac{5}{23}; P(G|AA) = \frac{13}{23}; P(U|AA) = \frac{2}{23};$$
$$P(A|AC) = \frac{5}{24}; P(C|AC) = \frac{8}{24}; P(G|AC) = \frac{3}{24}; P(U|AC) = \frac{8}{24};$$
$$P(A|AG) = \frac{6}{28}; P(C|AG) = \frac{3}{28}; P(G|AG) = \frac{10}{28}; P(U|AG) = \frac{9}{29};$$
$$P(A|AU) = 0; P(C|AU) = \frac{5}{13}; P(G|AU) = \frac{7}{13}; P(U|AU) = \frac{1}{13};$$
$$P(A|CC) = \frac{7}{34}; P(C|CC) = \frac{10}{34}; P(G|CC) = \frac{1}{34}; P(U|CC) = \frac{16}{34};$$
$$P(A|CA) = \frac{9}{30}; P(C|CA) = \frac{13}{30}; P(G|CA) = \frac{6}{30}; P(U|CA) = \frac{2}{30};$$
$$P(A|CG) = 0; P(C|CG) = 0; P(G|CG) = \frac{1}{5}; P(U|CG) = \frac{4}{5};$$
$$P(A|CU) = \frac{4}{45}; P(C|CU) = \frac{8}{45}; P(G|CU) = \frac{26}{45}; P(U|CU) = \frac{7}{45};$$
$$P(A|GG) = \frac{6}{43}; P(C|GG) = \frac{18}{43}; P(G|GG) = \frac{6}{43}; P(U|GG) = \frac{13}{43};$$
$$P(A|GA) = \frac{8}{25}; P(C|GA) = \frac{4}{25}; P(G|GA) = \frac{8}{25}; P(U|GA) = \frac{5}{25};$$
$$P(A|GC) = \frac{11}{36}; P(C|GC) = \frac{10}{36}; P(G|GC) = 0; P(U|GC) = \frac{15}{36};$$
$$P(A|GU) = \frac{1}{31}; P(C|GU) = \frac{4}{31}; P(G|GU) = \frac{22}{31}; P(U|GU) = \frac{4}{31};$$
$$P(A|UU) = \frac{3}{17}; P(C|UU) = \frac{3}{17}; P(G|UU) = \frac{6}{17}; P(U|UU) = \frac{5}{17};$$
$$P(A|UA) = \frac{3}{9}; P(C|UA) = \frac{2}{9}; P(G|UA) = \frac{1}{9}; P(U|UA) = \frac{3}{9};$$
$$P(A|UC) = \frac{8}{20}; P(C|UC) = \frac{6}{20}; P(G|UC) = \frac{1}{20}; P(U|UC) = \frac{5}{20};$$
$$P(A|UG) = \frac{12}{60}; P(C|UG) = \frac{16}{60}; P(G|UG) = \frac{25}{60}; P(U|UG) = \frac{7}{60}.$$

A entropia de cada contexto é calculada por (1).

Desse modo, tem-se que:

$$\begin{aligned} H(S|AA) &= -[(P(A|AA)\log_2 P(A|AA) + (P(C|AA)\log_2 P(C|AA) + \\ &\quad +(P(G|AA)\log_2 P(G|AA) + (P(U|AA)\log_2 P(U|AA)] \end{aligned}$$

$$H(S|AA) = -\left[\left(\frac{3}{23}\log_2\frac{3}{23}\right)+\left(\frac{5}{23}\log_2\frac{5}{23}\right)+\left(\frac{13}{23}\log_2\frac{13}{23}\right)+\left(\frac{2}{23}\log_2\frac{2}{23}\right)\right]$$

$$H(S|AA) = -[(-0,3833)+(-0,4786)+(-0,4652)+(-0,3064)]$$

$$H(S|AA) = 1,6335.$$

$$\begin{aligned} H(S|AC) &= -[(P(A|AC)\log_2 P(A|AC)+(P(C|AC)\log_2 P(C|AC)+ \\ &\quad +(P(G|AC)\log_2 P(G|AC)+(P(U|AC)\log_2 P(U|AC)] \end{aligned}$$

$$H(S|AC) = -\left[\left(\frac{5}{24}\log_2\frac{5}{24}\right)+\left(\frac{8}{24}\log_2\frac{8}{24}\right)+\left(\frac{3}{24}\log_2\frac{3}{24}\right)+\left(\frac{8}{24}\log_2\frac{8}{24}\right)\right]$$

$$H(S|AC) = -[(-0,4715)+(-0,5283)+(-0,3750)+(-0,5283)]$$

$$H(s_{|AC}) = 1,9031.$$

$$\begin{aligned} H(S|AG) &= -[(P(A|AG)\log_2 P(A|AG)+(P(C|AG)\log_2 P(C|AG)+ \\ &\quad +(P(G|AG)\log_2 P(G|AG)+(P(U|AG)\log_2 P(U|AG)] \end{aligned}$$

$$H(S|AG) = -\left[\left(\frac{6}{28}\log_2\frac{6}{28}\right)+\left(\frac{3}{28}\log_2\frac{3}{28}\right)+\left(\frac{10}{28}\log_2\frac{10}{28}\right)+\left(\frac{9}{28}\log_2\frac{9}{28}\right)\right]$$

$$H(S|AG) = -[(-0,4762)+(-0,3452)+(-0,5305)+(-0,5263)]$$

$$H(S|AG) = 1,8782.$$

$$\begin{aligned} H(S|AU) &= -[(P(A|AU)\log_2 P(A|AU)+(P(C|AU)\log_2 P(C|AU)+ \\ &\quad +(P(G|AU)\log_2 P(G|AU)+(P(U|AU)\log_2 P(U|AU)] \end{aligned}$$

$$H(S|AU) = -\left[(0)+\left(\frac{5}{13}\log_2\frac{5}{13}\right)+\left(\frac{7}{13}\log_2\frac{7}{13}\right)+\left(\frac{1}{13}\log_2\frac{1}{13}\right)\right]$$

$$H(S|AU) = -[(0) + (-0,5302) + (-0,4809) + (-0,2846)]$$

$$H(S|AU) = 1,2957.$$

$$\begin{aligned} H(S|CC) &= -[(P(A|CC)\log_2 P(A|CC) + (P(C|CC)\log_2 P(C|CC) + \\ &\quad +(P(G|CC)\log_2 P(G|CC) + (P(U|CC)\log_2 P(U|CC)] \end{aligned}$$

$$H(S|CC) = -\left[\left(\frac{7}{34}\log_2\frac{7}{34}\right) + \left(\frac{10}{34}\log_2\frac{10}{34}\right) + \left(\frac{1}{34}\log_2\frac{1}{34}\right) + \left(\frac{16}{34}\log_2\frac{16}{34}\right)\right]$$

$$H(S|CC) = -[(-0,4694) + (-0,5193) + (-0,1496) + (-0,5117)]$$

$$H(s_{|CC}) = 1,6500.$$

$$\begin{aligned} H(S|CA) &= -[(P(A|CA)\log_2 P(A|CA) + (P(C|CA)\log_2 P(C|CA) + \\ &\quad +(P(G|CA)\log_2 P(G|CA) + (P(U|CA)\log_2 P(U|CA)] \end{aligned}$$

$$H(S|CA) = -\left[\left(\frac{9}{30}\log_2\frac{9}{30}\right) + \left(\frac{13}{30}\log_2\frac{13}{30}\right) + \left(\frac{6}{30}\log_2\frac{6}{30}\right) + \left(\frac{2}{30}\log_2\frac{2}{30}\right)\right]$$

$$H(S|CA) = -[(-0,5211) + (-0,5228) + (-0,4644) + (-0.2604)]$$

$$H(S|CA) = 1,7687.$$

$$\begin{aligned} H(S|CG) &= -[(P(A|CG)\log_2 P(A|CG) + (P(C|CG)\log_2 P(C|CG) + \\ &\quad +(P(G|CG)\log_2 P(G|CG) + (P(U|CG)\log_2 P(U|CG)] \end{aligned}$$

$$H(S|CG) = -\left[(0) + (0) + \left(\frac{1}{5}\log_2\frac{1}{5}\right) + \left(\frac{4}{5}\log_2\frac{4}{5}\right)\right]$$

$$H(S|CG) = -[(0) + (0) + (-0,4644) + (-0,2575)]$$

$$H(S|CG) = 0,7219.$$

$$\begin{aligned} H(S|CU) &= -[(P(A|CU)\log_2 P(A|CU) + (P(C|CU)\log_2 P(C|CU) + \\ &\quad +(P(G|CU)\log_2 P(G|CU) + (P(U|CU)\log_2 P(U|CU)] \end{aligned}$$

$$H(S|CU) = -\left[\left(\frac{4}{45}\log_2\frac{4}{45}\right) + \left(\frac{8}{45}\log_2\frac{8}{45}\right) + \left(\frac{26}{45}\log_2\frac{26}{45}\right) + \left(\frac{7}{45}\log_2\frac{7}{45}\right)\right]$$

$$H(S|CU) = -[(-0,3104) + (-0,4430) + (-0,4573) + (-0,4176)]$$

$$H(S|CU) = 1,6283.$$

$$\begin{aligned} H(S|GG) &= -[(P(A|GG)\log_2 P(A|GG) + (P(C|GG)\log_2 P(C|GG) + \\ &\quad +(P(G|GG)\log_2 P(G|GG) + (P(U|GG)\log_2 P(U|GG)] \end{aligned}$$

$$H(S|GG) = -\left[\left(\frac{6}{43}\log_2\frac{6}{43}\right) + \left(\frac{18}{43}\log_2\frac{18}{43}\right) + \left(\frac{6}{43}\log_2\frac{6}{43}\right) + \left(\frac{13}{43}\log_2\frac{13}{43}\right)\right]$$

$$H(S|GG) = -[(-0,3965) + (-0,5259) + (-0.3965) + (-0,5218)]$$

$$H(S|GG) = 1,8407.$$

$$\begin{aligned} H(S|GA) &= -[(P(A|GA)\log_2 P(A|GA) + (P(C|GA)\log_2 P(C|GA) + \\ &\quad +(P(G|GA)\log_2 P(G|GA) + (P(U|GA)\log_2 P(U|GA)] \end{aligned}$$

$$H(S|GA) = -\left[\left(\frac{8}{25}\log_2\frac{8}{25}\right) + \left(\frac{4}{25}\log_2\frac{4}{25}\right) + \left(\frac{8}{25}\log_2\frac{8}{25}\right) + \left(\frac{5}{25}\log_2\frac{5}{25}\right)\right]$$

$$H(S|GA) = -[(-0,5060) + (-0,4230) + (-0,5060) + (-0,4644)]$$

$$H(S|GA) = 1,8994.$$

$$\begin{aligned} H(S|GC) &= -[(P(A|GC)\log_2 P(A|GC) + (P(C|GC)\log_2 P(C|GC) + \\ &\quad +(P(G|GC)\log_2 P(G|GC) + (P(U|GC)\log_2 P(U|GC)] \end{aligned}$$

$$H(S|GC) = -\left[\left(\frac{11}{36}\log_2\frac{11}{36}\right) + \left(\frac{10}{36}\log_2\frac{10}{36}\right) + (0) + \left(\frac{15}{31}\log_2\frac{15}{36}\right)\right]$$

$$H(S|GC) = -[(-0,5226) + (-0,5133) + (0) + (-0,6111)]$$

$$H(S|GC) = 1,6470.$$

$$\begin{aligned} H(S|GU) &= -[(P(A|GU)\log_2 P(A|GU) + (P(C|GU)\log_2 P(C|GU) + \\ &\quad +(P(G|GU)\log_2 P(G|GU) + (P(U|GU)\log_2 P(U|GU)] \end{aligned}$$

$$H(S|GU) = -\left[\left(\frac{1}{31}\log_2\frac{1}{31}\right) + \left(\frac{4}{31}\log_2\frac{4}{31}\right) + \left(\frac{22}{31}\log_2\frac{22}{31}\right) + \left(\frac{4}{31}\log_2\frac{4}{31}\right)\right]$$

$$H(S|GU) = -[(-0,1598) + (-0,3812) + (-0,3511) + (-0,3812)]$$

$$H(S|GU) = 1,2733.$$

$$\begin{aligned} H(S|UU) &= -[(P(A|UU)\log_2 P(A|UU) + (P(C|UU)\log_2 P(C|UU) + \\ &\quad +(P(G|UU)\log_2 P(G|UU) + (P(U|UU)\log_2 P(U|UU)] \end{aligned}$$

$$H(S|UU) = -\left[\left(\frac{3}{17}\log_2\frac{3}{17}\right) + \left(\frac{3}{17}\log_2\frac{3}{17}\right) + \left(\frac{6}{17}\log_2\frac{6}{17}\right) + \left(\frac{5}{17}\log_2\frac{5}{17}\right)\right]$$

$$H(S|UU) = -[(-0,4416) + (-0,4416) + (-0,5303) + (-0,5193)]$$

$$H(S|UU) = 1,9328.$$

$$\begin{aligned} H(S|UA) &= -[(P(A|UA)\log_2 P(A|UA) + (P(C|UA)\log_2 P(C|UA) + \\ &\quad +(P(G|UA)\log_2 P(G|UA) + (P(U|UA)\log_2 P(U|UA)] \end{aligned}$$

$$H(S|UA) = -\left[\left(\frac{3}{9}\log_2\frac{3}{9}\right) + \left(\frac{2}{9}\log_2\frac{2}{9}\right) + \left(\frac{1}{9}\log_2\frac{1}{9}\right) + \left(\frac{3}{9}\log_2\frac{3}{9}\right)\right]$$

$$H(S|UA) = -[(-0,5283) + (-0,4822) + (-0,3522) + (-0,5283)]$$

$$H(S|UA) = 1,8910.$$

$$\begin{aligned} H(S|UC) &= -[(P(A|UC)\log_2 P(A|UC) + (P(C|UC)\log_2 P(C|UC) + \\ &\quad +(P(G|UC)\log_2 P(G|UC) + (P(U|UC)\log_2 P(U|UC)] \end{aligned}$$

$$H(S|UC) = -\left[\left(\frac{8}{20}\log_2\frac{8}{20}\right) + \left(\frac{6}{20}\log_2\frac{6}{20}\right) + \left(\frac{1}{20}\log_2\frac{1}{20}\right) + \left(\frac{5}{20}\log_2\frac{5}{20}\right)\right]$$

$$H(S|UC) = -[(-0,5288) + (-0,5211) + (-0,2161) + (-0,5000)]$$

$$H(S|UC) = 1,7660.$$

$$\begin{aligned} H(S|UG) &= -[(P(A|UG)\log_2 P(A|UG) + (P(C|UG)\log_2 P(C|UG) + \\ &\quad +(P(G|UG)\log_2 P(G|UG) + (P(U|UG)\log_2 P(U|UG)] \end{aligned}$$

$$H(S|UG) = -\left[\left(\frac{12}{60}\log_2\frac{12}{60}\right) + \left(\frac{16}{60}\log_2\frac{16}{60}\right) + \left(\frac{25}{60}\log_2\frac{25}{60}\right) + \left(\frac{7}{60}\log_2\frac{7}{60}\right)\right]$$

$$H(S|UG) = -[(-0,4644) + (-0,5085) + (-0,5263) + (-0,3616)]$$

$$H(S|UG) = 1,8608.$$

Após o cálculo da entropia de cada contexto, a entropia do modelo é calculada por:

$$H(S) = \sum_{s \in S} P(s) \cdot H(s).$$

$$\begin{aligned} H(S) \;=\; & \left[\left(\frac{23}{447} \cdot 1,7961\right) + \left(\frac{24}{447} \cdot 1,8779\right) + \left(\frac{28}{447} \cdot 1,7347\right) + \left(\frac{13}{447} \cdot 1,0408\right) + \right. \\ & + \left(\frac{34}{447} \cdot 1,4787\right) + \left(\frac{30}{447} \cdot 1,6806\right) + \left(\frac{5}{447} \cdot 0,8113\right) + \left(\frac{45}{447} \cdot 1,7698\right) + \\ & + \left(\frac{43}{447} \cdot 1,8407\right) + \left(\frac{25}{447} \cdot 1,9491\right) + \left(\frac{36}{447} \cdot 1,5745\right) + \left(\frac{31}{446} \cdot 1,2235\right) + \\ & \left. + \left(\frac{17}{447} \cdot 1,9328\right) + \left(\frac{9}{447} \cdot 1,8910\right) + \left(\frac{20}{447} \cdot 1,5715\right) + \left(\frac{60}{447} \cdot 1,8068\right)\right] \end{aligned}$$

$$\begin{aligned} H(S) \;=\; & [(0,0514 \cdot 1,7961) + (0,0537 \cdot 1,8779) + (0,0626 \cdot 1,7347) + \\ & + (0,0291 \cdot 1,0408) + (0,0761 \cdot 1,4787) + (0,0671 \cdot 1,6806) + \\ & + (0,0112 \cdot 0,8113) + (0,1007 \cdot 1,7698) + (0,0962 \cdot 1,8407) + \\ & + (0,0559 \cdot 1,9491) + (0,0805 \cdot 1,6470) + (0,0693 \cdot 1,2235) + \\ & + (0,0380 \cdot 1,9328) + (0,0201 \cdot 1,8910) + (0,0447 \cdot 1,7660) + \\ & + (0,1342 \cdot 1,8608)] \end{aligned}$$

$$H(S) = 1,3238 \text{ bit/símbolo.}$$

A Tabela 30 a seguir apresenta uma comparação entre os valores obtidos para as ordens do modelo (de -1 a 2), considerando apenas os éxons da sequência.

Tabela 30 – Comparação entre as ordens do modelo utilizadas no cálculo da entropia.

Ordem do modelo	Valor da entropia (bits/símbolo)
Ordem -1	2
Ordem 0	1,9827
Ordem 1	1,8300
Ordem 2	1,3238

Pode-se observar até a ordem 2, que à medida que se aumenta a ordem do modelo, o valor da entropia diminui, uma vez que modelos de ordem maior capturam mais dependências e padrões no conjunto de dados.

6 CONSIDERAÇÕES FINAIS

Este trabalho apresenta algumas aplicações de estruturas matemáticas no contexto do código genético, em particular, o código de Gray e o hipercubo booleano, a partir de construções realizadas de diagramas de Hasse, por meio de permutações escolhidas dos três rotulamentos (A, B e C) associados ao código genético. Essas estruturas matemáticas facilitam a visualização dos códons, permitindo a análise das propriedades dos aminoácidos e a classificação dos códons com base em suas características, tornando-se uma ferramenta valiosa na análise de diversos fenômenos biológicos.

A partir das construções realizadas, foi possível notar que o rotulamento A segue uma abordagem biológica, com as cores representando propriedades e características físico-químicas dos aminoácidos aos códons. Sua organização é mais estruturada, com grupos hidrofóbicos nas bordas e hidroxila e básicos no centro, refletindo proteínas com uma estrutura definida e funções específicas. Já os rotulamentos B e C adotam uma abordagem algébrica, com uma distribuição mais equilibrada e dispersa das propriedades, ou seja, estas não estão concentradas em regiões específicas, sugerindo proteínas mais flexíveis ou com maior interação com o ambiente. Nos hipercubos construídos para os três rotulamentos, pode-se perceber uma separação entre as regiões hidrofóbicas e hidrofílicas, enquanto para o rotulamento A tem uma configuração mais concentrada e estruturada que nos rotulamentos B e C, mostram maior dinâmica nas interações químicas.

Outro resultado relevante dessa dissertação é obtido do ponto de vista de aplicação de Teoria da Informação em sequências de DNA, em que foi realizado a análise da sequência de nucleotídeos no gene humano completo de β-globina, apresentada em Nussbaum (2008). Nesse sentido, foram realizados cálculos para obter a entropia associada a dois modelos (M_1 e M_2), comparando e analisando os éxons da sequência, bem como a sequência completa selecionada. Também, foram analisados os contextos finitos e a ordem do modelo, neste caso utilizado somente os éxons da sequência.

Assim, foi possível observar que tanto para os éxons, quanto para a sequência completa, a escolha do modelo interfere diretamente na estimativa da entropia da mensagem. A entropia da mensagem obtida utilizando M_2 foi menor que a entropia da mesma mensagem utilizando M_1 para ambos os casos. O conhecimento da mensagem possibilita a escolha de um modelo mais adequado e com isso se reduz significativamente sua estimativa da entropia.

Ademais, o modelo pode ser classificado como adaptativo, pois há um modelo inicial presente tanto no transmissor quanto no receptor e, à medida que o código é transmitido, o modelo também se ajusta, acarretando em custos para sua manutenção. Assim, a transmissão ocorre símbolo por símbolo, permitindo que tanto o transmissor

quanto o receptor atualizem suas distribuições de probabilidades.

Ressalta-se que a proposta desse trabalho foi. apresentar um estudo inicial envolvendo a entropia para modelos de DNA e deixa-se como proposta para trabalhos futuros a análise dos resultados obtidos e as possíveis interferências dos mesmos em casos práticos envolvendo sequências de DNA.

Por fim, destaca-se a importante conexão existente entre elementos de Álgebra, Geometria, Biologia, Engenharia e Teoria da Informação, ilustrando a interdisciplinaridade do estudo apresentado aqui.

6.1 EVENTOS CIENTÍFICOS

Com a realização deste trabalho foram apresentados diversos trabalhos em importantes eventos científicos, como forma de divulgar os resultados obtidos, buscar o estabelecimento de parcerias e trocas de informações e saberes. Os eventos e trabalhos apresentados foram:

- VIII Workshop de Matemática e Matemática Aplicada (WMMA) - São João del-Rei - Construção do Hipercubo Booleano Associado ao Rotulamento A do Código Genético;
- II Reunião Mineira de Matemática (RMM) - Belo Horizonte - Construção do Diagrama de Hasse Associado ao Rotulamento A do Código Genético;
- IV Workshop do PPGEAB - Alfenas - Hipercubo Booleano Associado ao Rotulamento B do Código Genético;
- XLIII Congresso Nacional de Matemática Aplicada e Computacional (CNMAC) - Porto de Galinhas - Caracterização de Propriedades dos Aminoácidos por meio do Hipercubo Booleano;
- X Simpósio Integrado Biomas do Brasil: diversidade, saberes e tecnologias sociais - Alfenas - Teoria da Informação no Estudo de Sequências de DNA;
- IX Workshop de Matemática e Matemática Aplicada (WMMA) - Ouro Preto - Teoria da Informação no Estudo de Sequências de DNA.

6.2 PROPOSTAS PARA TRABALHOS FUTUROS

A seguir serão apresentados algumas propostas para trabalhos futuros:

- Realizar um estudo mais aprofundado de elementos de Teoria de Informação, como o caso da entropia e da informação mútua, para sequências de DNA e as possibilidades de análise de mutações relacionadas a essas medidas;
- Realizar análises dos hipercubos booleanos construídos neste trabalho e suas implicações em situações envolvendo o estudo dos aminoácidos e possíveis interferências do modelo construído para a realização de estudos de mutações genéticas.;
- Buscar conexões entre a estrutura dos hipercubos do código genético e a quântica.

REFERÊNCIAS

ALBERTS, B. *et al.* **Biologia molecular da célula**. Artmed Editora, 2017.

FARIA, L. C. B. **Existências de códigos corretores de erros e protocolos de comunicaçao em sequências de DNA**. 2011. 296f. Tese (Doutorado em Engenharia Elétrica) - Universidade Estadual de Campinas, Campinas, SP, 2011.

Fernandes Jr., D. P. ; Vargas Jr., V. **Conceitos e simulação de Cadeias de Markov**. 2011.

FERNANDES, R. S.; OLIVEIRA, A. J. Análise das propriedades físico-químicas dos aminoácidos por meio das distâncias de Hamming associadas ao rotulamento C do código genético. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**, v. 8, n. 1, 2021.

FERNANDES, R. S.; OLIVEIRA, A. J. Caracterização das propriedades dos aminoácidos por meio do diagrama de Hasse associado ao rotulamento A do código genético. **Brazilian Electronic Journal of Mathematics**, Uberlândia, v. 2, n. 4, p. 81-100, 2021.

FRANCO, L. A. L.; PALAZZO JÚNIOR, R. Análise do splicing alternativo do Gene Hint-1 através do código BCH associado. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**, Lavras, v. 3, n. 1, 2015.

GERÔNIMO, J. R.; FRANCO, V. S. **Fundamentos de matemática**: uma introdução à lógica matemática, teoria dos conjuntos, relações e funções. Maringá: Eduem, 2008.

HAGE, P.; HARARY, F. Què es un hipercubo? Un código binario para relaciones de parentesco. In: J. G. Mendieta; S. Schmidt. **Análisis de redes: aplicaciones en ciencias sociales**, 1^a Ed. UNAM, 2002, p.15-22.

HEPP, D.; NONOHAY, J. S. A importância das técnicas e análises de DNA. **ScientiaTec**, v. 3, n. 2, p. 114-124, 2016.

HORTA, J. O. C. **Estimadores de entropia para sequências de DNA**. Tese de Doutorado. Universidade de São Paulo-USP, São Paulo, 136f, 2001.

JIMENEZ-MONTANO, M. A; LA MORA-BASANEZ, C. R. and POESCHEL, T. On the Hypercube Structure of the Genetic Code, **arXiv preprint cond-mat/0204044**, 2002, 1-14. DOI: 10.48550/arXiv.cond-mat/0204044.

MAGOSSI, J. C.; ABREU, P. H. C.; BARROS, A. C. C.; PAVIOTTI, J. R. A Medida de Informação de Shannon: Entropia. **Revista Brasileira de História da Matemática**,

São Paulo, v. 21, n. 41, p. 45–72, 2021. DOI: 10.47976/RBHM2021v20n4145-72. Disponível em: https://rbhm.org.br/index.php/RBHM/article/view/337.

MOHAMMED, R. Information analysis of DNA sequences. **Louisiana State University and Agricultural Mechanical College**, 2010.Disponível em: https://doi.org/10.48550/arXiv.1010.4205.

JIMENEZ-MONTANO, M. A.;LA MORA-BASANEZ, C. R.; POSCHEL, T. The Hypercube Structure of the Genetic Code Explains Conservative and Non-Conservative Aminoacid Substitutions in vivo and in vitro. **BioSystems**, v. 39, n. 2, pág. 117-125, 1996.

LIN, S.; COSTELO JR, D. J. **Error Control Coding**: Fundamentals and Applications. 2. ed. Englewood Cliffs: Prentice Hall, 1983.

NUSSBAUM, R. **O Genoma Humano:** Estrutura e Função dos Genes e Cromossomos. Thompson e Thompson Genética Médica. Elsevier Brasil, 2008.

SÁNCHEZ, R.; MORGADO, E.; GRAU, R. The Genetic Code Boolean Lattice. **preprint arXiv q-bio/0412034**, 2004.

SHANNON, C. E. A Mathematical Theory of Communication, **The Bell System Technical Journal**, New York, v. 27, n.3, p. 379-423. July. 1948.