UNIVERSIDADE FEDERAL DE ALFENAS

JOSÉ DOS SANTOS FERNANDES

ANÁLISE MUTACIONAL VIA IMPLEMENTAÇÃO DO ALGORITMO SOMA COM TRANSPORTE

Alfenas/MG

2020

JOSÉ DOS SANTOS FERNANDES

ANÁLISE MUTACIONAL VIA IMPLEMENTAÇÃO DO ALGORITMO SOMA COM TRANSPORTE

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação, pelo Instituto de Ciências Exatas da Universidade Federal de Alfenas. Orientadores: Mariane Moreira de Souza e Anderson José de Oliveira.

Alfenas/MG
2020

Resumo

A modelagem computacional pode ser utilizada para auxiliar no desenvolvimento de áreas correlatas, como a matemática aplicada na modelagem do código genético. Dentre outros, o Algoritmo Soma com Transporte é uma forma de simular e analisar fenômenos mutacionais, que acarretam mudanças durante o processo de síntese proteica. Através da implementação deste algoritmo, é possível caracterizar os rotulamentos associados ao mapeamento do código genético, identificando as influências no processo de análise de mutações. Este trabalho modela e implementa o algoritmo para realização da Soma com Transporte, com o objetivo de se obter uma ferramenta que facilite a análise de fenômenos mutacionais e realização dos cálculos envolvidos. Utilizamos os rotulamentos A, B e C durante a modelagem, analisando o comportamento das sequências, tanto fictícias quanto reais, em especial na geração das proteínas e nos efeitos da mutação em um determinado organismo. Por fim, comparamos os resultados entre os rotulamentos A, B e C, visando detectar os diferentes efeitos que uma mesma sequência gera em um organismo. Os resultados obtidos apontam para a possibilidade de utilização do algoritmo em problemas de bioinformática e biomedicina, gerando uma ferramenta que pode ser facilmente utilizada pelo usuário final ou pelo pesquisador.

Palavras - Chave: Síntese Proteica, Rotulamentos do Código Genético, Mutações, Álgebra, Soma com Transporte.

Abstract

Computational modeling can be used to assist in the development of related areas, such as applied mathematics in the modeling of the genetic code. Among others, the Sum of Transportation Algorithm is a way to simulate and analyze mutational phenomena, which cause changes during the process of protein synthesis. Through the implementation of this algorithm, it is possible to characterize the labeling associated with the mapping of the genetic code, identifying the influences in the mutation analysis process. This work models and implements the Sum of Transportation Algorithm, with the aim of obtaining a tool that facilitates the analysis of mutational phenomena and performing the calculations involved. We use A, B and C labeling during modeling, analyzing the behavior of the sequences, both fictitious and real, especially in the generation of proteins and the effects of mutation in a given organism. Finally, we compare the results between A, B and C labeling, in order to detect the different effects that the same sequence generates in an organism. The results obtained point to the possibility of using the algorithm in problems of bioinformatics and biomedicine, generating a tool that can be easily used by the end-user or a researcher.

Keywords: Protein Synthesis, Genetic Code Labeling, Mutations, Algebra, the Sum of Transportation.

Lista de Ilustrações

Lista de Tabelas

Sumário

1. Introdução

Embora não aparente, a teoria das comunicações e a genética compartilham entre si uma característica: a de transmitir a informação. A teoria das comunicações é programada para enviar mensagens, através de algum meio, como cabos, ondas ou pulsos elétricos, sendo este processo concebido pelo homem. Ao mesmo tempo, a genética envolve processos que necessitam enviar, naturalmente, mensagens hereditárias entre meios, dentro de um mesmo organismo, segundo Battail (2008). Sendo assim, ambas buscam enviar mensagens de um ponto a outro, com a menor quantidade possível de erros ou ruídos, que podem comprometer a informação (tal qual uma mutação genética), configurando uma das bases deste estudo.

Shannon (1948), um importante matemático e criptógrafo americano, foi o precursor da teoria de códigos em sistemas de comunicação, iniciando na década de 70 a aplicação da teoria da informação na análise dos dados genéticos, porém sem obter sucesso. Entretanto, seu estudo motivou posteriormente diversos trabalhos na teoria de comunicações para o genoma, culminando nos estudos de diversos pesquisadores como May (2004). O estudo de May explora os paralelos entre a genética e a transmissão da informação, com a ideia de que os genes são mantidos em forma de sequências de ácidos nucleicos, e tem a função de gerar proteínas. Por meio de estruturas algébricas, pode-se analisar algumas destas relações.

A partir dos estudos de Rocha (2010) e Faria (2011), são propostos modelos que fazem analogias entre o sistema de informação de genes e genomas e o sistema de comunicação digital e, em Oliveira (2012) é apresentada a aplicabilidade de estruturas algébricas, com a representação matemática de padrões biológicos. Assim, pode-se apresentar a aplicação de estruturas algébricas na relação entre sistemas de comunicação criados pelo homem e sistemas de comunicação biológicos. Uma das formas de estabelecer conexões entre essas três áreas (Biologia, Matemática e Computação) é apresentada em Oliveira (2012), por meio do Algoritmo Soma com Transporte, objeto de estudo deste trabalho.

O objetivo deste trabalho é, a partir da implementação do Algoritmo Soma com Transporte, que simula o processo de síntese proteica através da soma de sequências, analisar a automação da realização dos cálculos envolvidos, em especial quando se leva em conta os rotulamentos A, B e C, que exigem etapas adicionais e repetitivas para que se alcance um resultado. Também se deseja analisar o comportamento das sequências durante a geração de proteínas e os efeitos de uma mutação, além de comparar os efeitos dos rotulamentos A, B e C em uma mesma sequência.

Através do Algoritmo Soma com Transporte, é possível analisar os efeitos de uma interferência na mensagem, dentro de um sistema de comunicação biológico, entendendo

como diferenças na representação matemática dos genes e genomas podem criar resultados diferentes, além de trazer, por meio da implementação, uma maior agilidade para processar dados em grandes blocos de informação. Desta forma, é possível identificar as relações entre estas três áreas que, apesar de parecerem distantes, compartilham características em comum que permitem o estudo multidisciplinar de temas pertinentes às áreas de Biologia, Matemática e Computação.

Este trabalho está organizado da seguinte maneira: no Capítulo 2 é apresentada uma revisão de conceitos de Biologia, Álgebra (abordando os conceitos de Grupos e Anéis), Sistemas de Comunicação e Rotulamentos do Código Genético; no Capítulo 3 é apresentado o Algoritmo Soma com Transporte e as técnicas utilizadas para a obtenção de resultados, junto das execuções do algoritmo; as discussões realizadas através da análise de sequências genéticas geradas, fictícias e reais, por meio do Algoritmo Soma com Transporte, são apresentados no Capítulo 4; por fim, no Capítulo 5, são apresentadas as conclusões e propostas futuras de trabalho.

2. Revisão de Conceitos

Este Capítulo apresenta os principais conceitos fundamentais para o entendimento deste trabalho. Na Seção 2.1 serão apresentados os principais elementos de Biologia, como células, DNA, RNA, Síntese Proteica e Mutações; a Seção 2.2 apresenta conceitos de Álgebra, no caso, Grupos e Anéis; na Seção 2.3 serão apresentados os conceitos de Sistemas de Comunicação; por fim, na Seção 2.4, serão apresentados os Rotulamentos do Mapeamento do Código Genético.

2.1. Elementos de Biologia

Nesta Seção serão apresentados conceitos de Biologia que são importantes para o entendimento dos tópicos que serão abordados futuramente, tais como células, e alguns aspectos de sua composição; nucleotídeos e ácidos nucleicos, possibilitando a caracterização do DNA e RNA; também são apresentados os conceitos de proteínas e aminoácidos, além do código genético; por fim, apresentamos os elementos que caracterizam as mutações.

Os conceitos relacionados com os elementos de biologia podem ser encontrados em detalhes em Marzzoco e Torres (2013), Alberts et al. (2010) e Pamphile e Vicentini (2011).

2.1.1. Células

A célula é a menor porção de matéria com vida do organismo, de caráter microscópico e que só podem ser vistas com aparelhos especiais. Ao examinar no microscópio um pedaço de cortiça, Robert Hooke (1665) descobriu algumas estruturas no material, e percebeu que elas eram compostas por compartimentos, os quais foram chamados de células. As principais funções das células são: captar nutrientes, utilizando-os na produção de energia, eliminar resíduos do metabolismo e capacidade de movimentação (em alguns casos).

As células se dividem em dois grupos, procariontes e eucariontes. Células procariontes são caracterizadas pela ausência de núcleo (carioteca) e da maioria das organelas, tais como: mitocôndria, cloroplasto e complexo de Golgi. Neste grupo, estão inclusos as bactérias e as cianobactérias (algas azuis). Por outro lado, as células eucariontes apresentam núcleo e um compartimento onde o material genético fica isolado do citoplasma, sendo células mais complexas que as procarióticas, com membrana nuclear individualizada e vários tipos de organelas.

Exemplos de organismos compostos de células eucariontes são os animais, vegetais, fungos e leveduras, que são seres pluricelulares. Estes dois grupos são exemplificados nas Figuras 1 e 2.

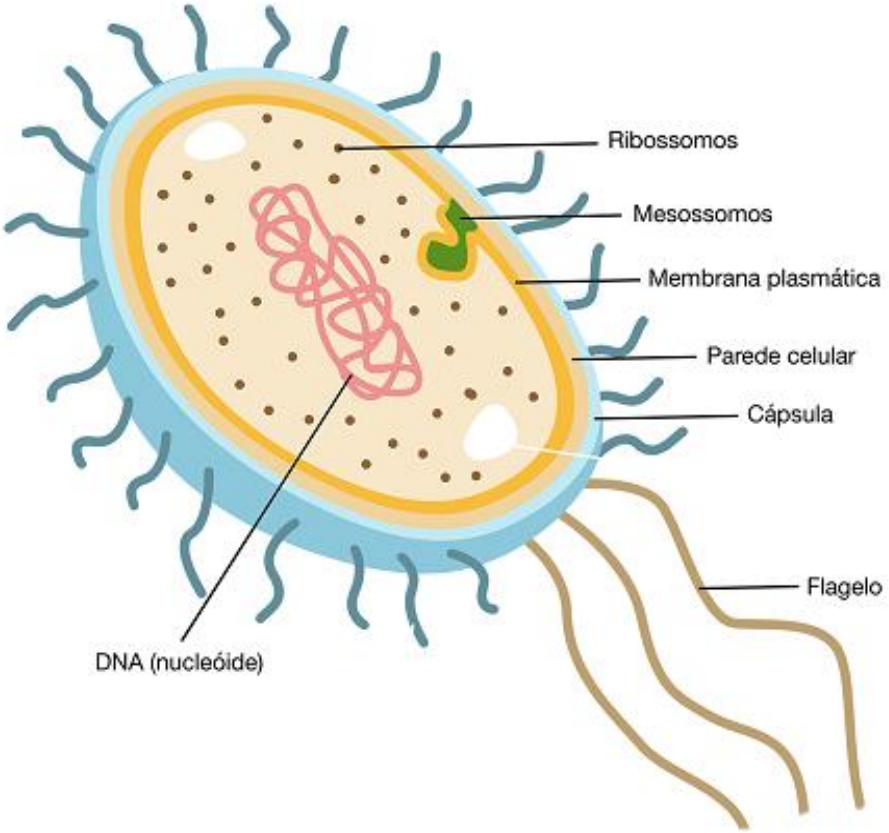

Figura 1: Célula procarionte. Fonte: Duque (2018).

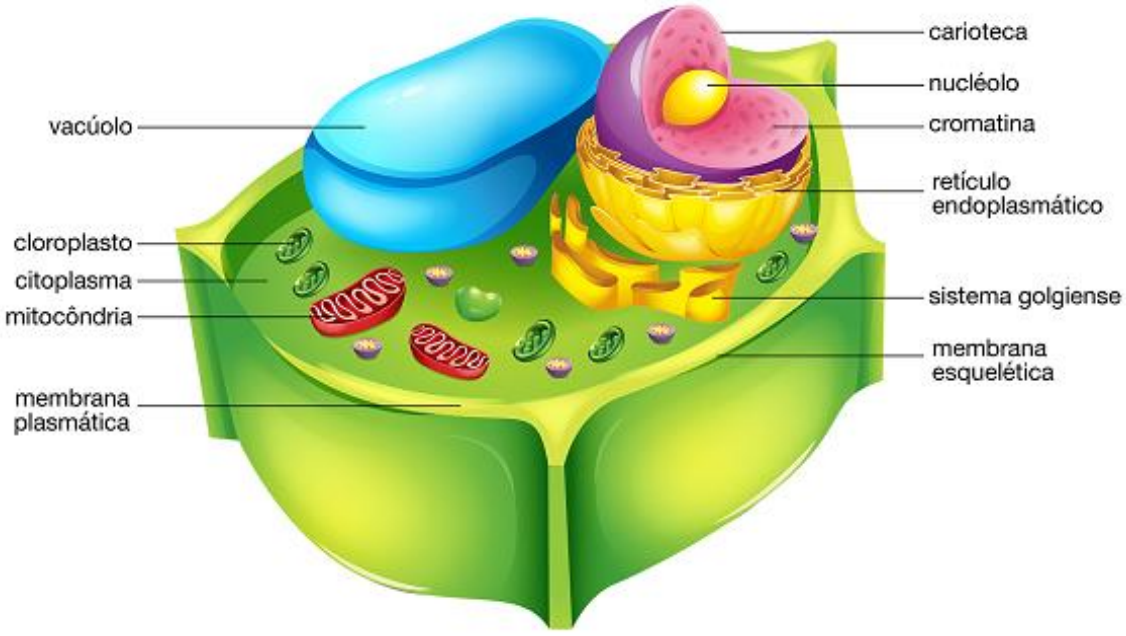

Figura 2: Célula eucarionte. Fonte: Duque (2018).

2.1.2. Nucleotídeos e Ácidos Nucleicos

Os nucleotídeos são subunidades de uma molécula que, quando se ligam, formam os ácidos nucleicos, também conhecidos como DNA e RNA, responsáveis por armazenar e

codificar a informação genética. Tanto os nucleotídeos quanto os ácidos nucleicos são responsáveis por desempenhar funções estruturais e catalisadoras dentro da célula.

O DNA e o RNA estão diretamente envolvidos na transmissão da hereditariedade dos seres e na produção de proteínas, principal constituinte dos seres vivos. O DNA é responsável por realizar a produção do RNA, que, por sua vez, é responsável pela síntese de proteínas. Este processo é conhecido como síntese proteica.

O DNA é composto de três partes: o grupo fosfato, a desoxirribose e a base nitrogenada. Por sua vez, a base nitrogenada é composta pelas bases purinas, que compreende a adenina e a guanina, e pelas bases pirimidinas, representadas pela timina e pela citosina.

As bases nitrogenadas são unidas por pontes de hidrogênio, e respeitam a complementaridade descrita pela regra de Chargaff, onde adenina liga-se a timina, com duas pontes de hidrogênio, e citosina liga-se a guanina, com três pontes de hidrogênio. Esta regra foi descoberta por Erwin Chargaff, em meados da década de 40.

Diferente do DNA, em que existem duas fitas de nucleotídeos ligadas entre si, o RNA é composto por uma única fita. Outra diferença é que, no RNA, temos a uracila no lugar da timina, mas ainda respeitando a regra de Chargaff. Além disso, o RNA pode ser de três tipos: RNA mensageiro (RNAm), RNA transportador (RNAt) e RNA ribossômico (RNAr).

2.1.3. Estrutura do DNA

Descrito por Watson e Crick em 1953, e aceito como o marco da biologia moderna, a estrutura do DNA é formada por duas cadeias de múltiplos nucleotídeos, formando uma dupla hélice, sendo que estas cadeias são antiparalelas. As cadeias de açúcar-fosfato se encontram nas bordas, e as bases nitrogenadas se encontram ao centro da hélice, ligando entre si pelas pontes de hidrogênio, sempre respeitando a regra de Chargaff.

Estas cadeias (também conhecidas como fitas) atuam como molde para a síntese de fitas complementares, através do processo de duplicação, onde as pontes de hidrogênio são quebradas, e cada fita se liga a outras bases nitrogenadas, formando duas estruturas iguais, conforme mostra a Figura 3.

2.1.4. Síntese Proteica e o RNA

Para realizar a síntese proteica, deve ocorrer o processo de transcrição do RNA. Para isso, o DNA produz moléculas de RNA mensageiro (RNAm), que irá migrar para o citoplasma no processo de síntese. Uma das cadeias do DNA serve de molde para produzir este RNAm, ou seja, o RNA formado será uma única fita. Através da enzima RNA polimerase, quebram-se as pontes de hidrogênio, e desta forma, os nucleotídeos livres do RNA se ligarão em uma das fitas, formando o que é chamada de fita ativa. Esta fita se solta de seu molde e migra

para o citoplasma, enquanto as cadeias de DNA se pareiam novamente, reconstituindo a cadeia original. Este processo é ilustrado na Figura 4.

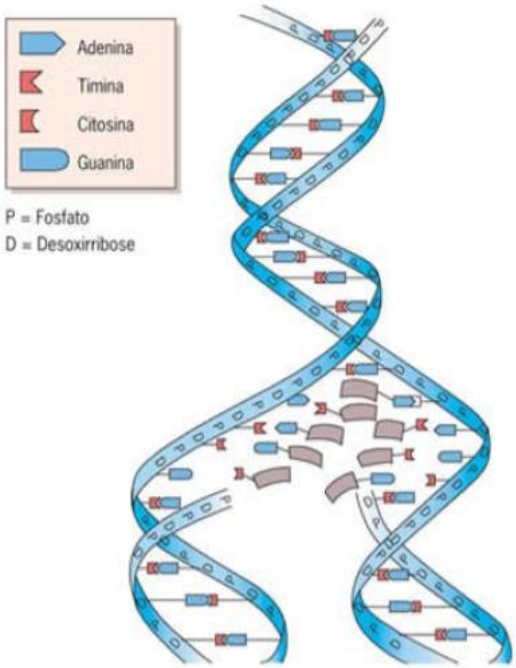

Figura 3: Processo de duplicação do DNA. Fonte: Marques (2014).

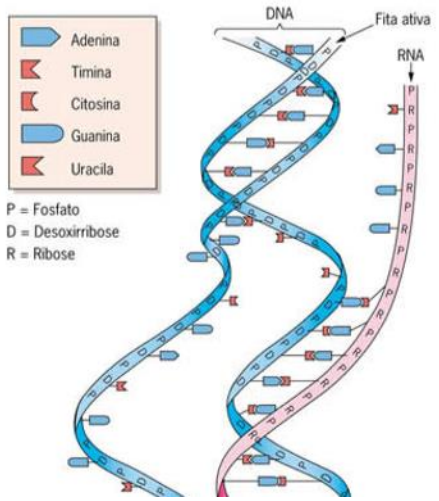

Figura 4: Processo de transcrição do DNA. Fonte: Marques (2014).

Após migrar para o citoplasma, o RNAm é traduzido pelo RNA ribossômico (RNAr), formando os aminoácidos, que por sua vez, serão transportados pelo RNA transportador (RNAt), para serem transformados na proteína desejada. Na Figura 5, é possível visualizar de forma simplificada os processos envolvidos na síntese proteica.

Figura 5: Síntese proteica. Fonte: O autor.

2.1.5. Proteínas e Aminoácidos

As proteínas apresentam-se sob inúmeras formas e tamanhos, e exercem funções metabólicas essenciais ao bom funcionamento da célula. Segundo Torres e Marzzoco (2013), as proteínas representam uma classe especial de moléculas, sendo constituídas por subunidades denominadas aminoácidos.

Os aminoácidos são utilizados na síntese proteica e, por possuírem carbono, são essenciais à vida humana. Um dos critérios para a classificação dos 20 aminoácidos existentes é a hidropaticidade, sendo que 12 são hidrofílicos (capaz de se dissolver na água, ou seja, um composto com afinidade à água, também conhecido como polar) e 8 são hidrofóbicos (não é capaz de se dissolver na água, não possuindo afinidade com ela, também chamado de apolar). Além da característica de hidropaticidade, existem outras características que podem ser utilizadas para a classificação, como carga e polaridade.

2.1.6. Código Genético

As proteínas são polímeros formados através da combinação aleatória de 20 tipos de aminoácidos. Existem 64 possíveis combinações, utilizando as quatro bases nitrogenadas do DNA/RNA agrupadas três-a-três, caracterizando a especificação dos 20 aminoácidos distintos. A síntese proteica é feita em blocos de três nucleotídeos, que são os códons, sendo cada um responsável por um aminoácido específico. A correspondência do códon para um aminoácido é o que denominamos de código genético, com seus respectivos aminoácidos, abreviação e códons correspondentes.

A Tabela 1 apresenta o código genético, e para a mesma, tem-se: Ala: Alanina, Arg: Arginina, Asn: Asparagina, Asp: Ácido Aspártico, Cys: Cisteína, Gln: Glutamina, Glu: Ácido Glutâmico, Gly: Glicina, His: Histidina, Ile: Isoleucina, Leu: Leucina Lys: Lisina, Met: Metionina, Phe: Fenilalanina, Pro: Prolina, Ser: Serina, Thr: Treonina, Trp: Triptofano, Tyr: Tirosina, Val: Valina.

		Segunda Base					
		U	C	A	G		
Primeira Base	U	UUU (Phe)	UCU (Ser)	UAU (Tyr)	UGU (Cys)	U	Terceira Base
		UUC (Phe)	UCC (Ser)	UAC (Tyr)	UGC (Cys)	C	
		UUA (Leu)	UCA (Ser)	UAA (Stop)	UGA (Stop)	A	
		UUG (Leu)	UCG (Ser)	UAG (Stop)	UGG (Trp)	G	
	C	CUU (Leu)	CCU (Pro)	CAU (His)	CGU (Arg)	U	
		CUC (Leu)	CCC (Pro)	CAC (His)	CGC (Arg)	C	
		CUA (Leu)	CCA (Pro)	CAA (Gln)	CGA (Arg)	A	
		CUG (Leu)	CCG (Pro)	CAG (Gln)	CGG (Arg)	G	
	A	AUU (Ile)	ACU (Thr)	AAU (Asn)	AGU (Ser)	U	
		AUC (Ile)	ACC (Thr)	AAC (Asn)	AGC (Ser)	C	
		AUA (Ile)	ACA (Thr)	AAA (Lys)	AGA (Arg)	A	
		AUG (Met)	ACG (Thr)	AAG (Lys)	AGG (Arg)	G	
	G	GUU (Val)	GCU (Ala)	GAU (Asp)	GGU (Gly)	U	
		GUC (Val)	GCC (Ala)	GAC (Asp)	GGC (Gly)	C	
		GUA (Val)	GCA (Ala)	GAA (Glu)	GGA (Gly)	A	
		GUG (Val)	GCG (Ala)	GAG (Glu)	GGG (Gly)	G	

Tabela 1: Código Genético. Adaptado de Oliveira e Palazzo Junior (2012).

2.1.7. Mutações

Uma mudança na sequência de nucleotídeos do material genético de um organismo qualquer é caracterizada como uma mutação. A partir da alteração deste material genético, as mutações são caracterizadas como desfavoráveis (por exemplo, quando um erro na sequência proteica gera uma doença) ou favoráveis (quando, por exemplo, a síntese de uma nova proteína aumenta a resistência do organismo). Da mesma forma, o efeito da mutação é caracterizado utilizando a seguinte escala:

Mutações de pequena escala: afetam um gene em poucos nucleotídeos, podendo ser:

1. Mutação pontual: causada por erros na replicação do DNA, onde ocorre a troca de nucleotídeos, ocorrendo uma transição ou transversão dos mesmos, caracterizados respectivamente pela troca de pirimidina por pirimidina $(C \leftrightarrow U)$ou purina por purina $(A \leftrightarrow G)$ e pela troca de uma pirimidina por uma purina e purina por pirimidina $(C/U \leftrightarrow A/G)$. Essas mutações pontuais são classificadas em três tipos: mutação silenciosa, onde o códon codifica o mesmo aminoácido; mutação missense, de sentido trocado, ou não sinônima, onde o códon codifica um aminoácido distinto; mutação sem sentido, quando codifica-se um códon de parada (STOP), provocando o fim da síntese proteica antes do previsto.

2. Inserção: provoca a adição de um ou mais nucleotídeos na sequência do DNA, sendo ocasionada por erros no processo de replicação de elementos repetitivos.

3. Deleção: uma mutação geralmente irreversível, onde ocorre a remoção de um ou mais nucleotídeos na sequência do DNA.

Mutações de grande escala: afetam muitos nucleotídeos, e podem ser caracterizadas em:

1. Duplicação gênica: ocorrência de cópias de determinada região cromossômica, que aumenta a dosagem de genes dentro da mesma;

2. Deleção de regiões cromossômicas: a ocorrência da mutação leva a perda dos genes de determinadas regiões cromossômicas;

3. Perda de heterozigosidade: leva a deleção de um alelo num organismo que, antes da mutação, possuía dois alelos.

2.2. Elementos de Álgebra Abstrata

Nesta Seção serão apresentados conceitos referentes à Álgebra, no caso, abordando o conceito de Grupos na Subseção 2.2.1, e o conceito de Anéis na Subseção 2.2.2. Os elementos aqui abordados podem ser vistos em mais detalhes nos trabalhos de Lin e Costelo (1983), Domingues e Iezzi (2003) e Garcia e Lequain (2010).

2.2.1. Grupos

Um grupo é caracterizado como um conjunto *G não-vazio* com uma operação $*$ (as operações podem ser de soma ou multiplicação), onde:

$$\gamma: G \times G \to G$$
$$(a, b) \to a * b,$$

munido das seguintes propriedades:

1. associativa: $a * (b * c) = (a * b) * c, \forall a, b, c \in G$,
2. $\exists$ elemento neutro e para a operação $*$, tal que: $e * a = a = a * e, \forall a \in G$,
3. $\forall a \in G, \exists b$ (elemento inverso): $a * b = b * a = e, \forall a, b \in G$.

Além disso, um grupo é chamado de grupo abeliano (ou grupo comutativo) se a operação binária $*$ satisfazer a seguinte condição: para quaisquer a e b em G, $a * b = b * a$.

2.2.2. Anéis

Um conjunto não vazio A, junto a um par de operações binárias: uma adição $(a, b) \rightarrow a + b$ e uma multiplicação $(a, b) \rightarrow a \cdot b$, é denominado anel se as seguintes propriedades forem verificadas:

- **Para a soma:**
  - (i) associatividade: $\forall a, b, c \in A, (a + b) + c = a + (b + c)$;
  - (ii) comutatividade: $\forall a, b \in A, a + b = b + a$;
  - (iii) existe um elemento neutro $0A$ tal que: $\forall a \in A, a + 0A = a = 0A + a$;
  - (iv) existe elemento oposto: $\forall a \in A, \exists (-a) \in A$ tal que $a + (-a) = 0A = (-a) + a$;
- **Para a multiplicação:**
  - (v) associatividade: $\forall a, b, c \in A, a \cdot (b \cdot c) = (a \cdot b) \cdot c$;
  - (vi) distributividade (em relação à adição): $\forall a, b, c \in A, a \cdot (b + c) = a \cdot b + a \cdot c$ e $(a + b) \cdot c = a \cdot c + b \cdot c$.

Um anel (A, +, ·), onde a operação · é comutativa é denominado de anel comutativo, e um anel (A, +, ·) em que o conjunto A é finito é denominado de anel finito e é representado da forma $\mathbb{Z}_m$(m > 1). Sendo um anel finito, suas propriedades *(i)* - *(vi)* podem ser verificadas, através da adição e multiplicação.

2.3. Sistemas de Comunicação

A transmissão da informação é o que leva ao desenvolvimento de algoritmos de detecção, prevenção e correção de erros, mas o grande motivador da área é derivado dos estudos referentes à teoria das comunicações, cujos aspectos básicos são abordados por Shannon (1948).

Considera-se que um sistema de comunicação é constituído de meios físicos, equipamentos, ou até mesmo de um organismo, cuja principal tarefa é transferir dados de uma fonte de informação para um determinado destino de maneira correta, isto é, que a mensagem seja recebida de forma fidedigna à informação original, por meio de um canal de comunicação.

A fonte gera uma mensagem a ser enviada (e que é definida dentro de um alfabeto finito, permitindo um mapeamento das informações) e o emissor deve realizar a conversão dos dados para sinais que sejam adequados ao meio (considerando fatores como distância, acesso e tamanho das informações). Essa conversão ocorre por processos de codificação da mensagem através de sinais, enviados por um meio de transmissão (canal), até que ela chegue a um receptor, que transforma os sinais novamente em dados (ou seja, tanto a fonte quanto o destino precisam conhecer o alfabeto associado às informações). Por sua vez, o

receptor será responsável por levar estes dados ao destinatário, que os consumirá da forma adequada ao escopo do problema/caso. Este esquema é apresentado na Figura 6.

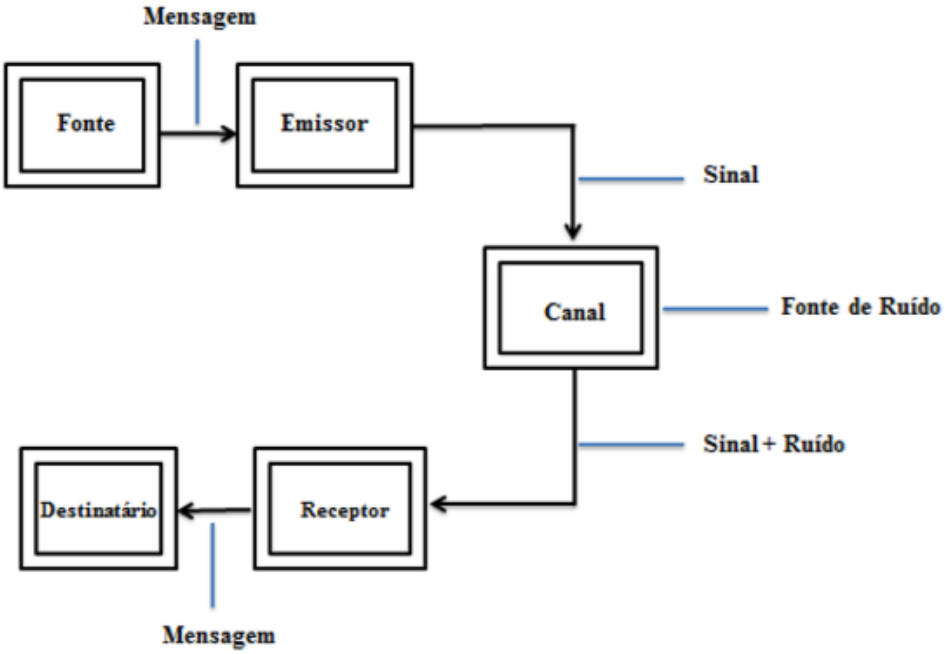

Figura 6: Exemplo de sistema de comunicação. Fonte: Faria e Palazzo Junior (2011).

Durante essa transmissão, o canal de envio atua como um meio, possibilitando uma transmissão de dados que preserva a natureza e integridade deles. Todavia, o meio é suscetível a falhas e efeitos, que impactam na transmissão e nas características da mensagem. Isso ocorre pelo intermédio de interferências, ruídos e distorções que impedem a passagem dos sinais para o receptor, ou distorcem a mensagem, fazendo com que o destinatário receba uma mensagem alterada, diferente da originalmente enviada.

É possível fazer um mapeamento entre a Biologia Molecular e a Teoria de Comunicações, considerando as seguintes características (vide Figura 7):

1. No sistema de comunicação, quem gera a informação a ser transmitida é o transmissor. Biologicamente, o responsável por esta função é o DNA.

2. Os processos de transcrição e tradução representam os processos de codificação de canal e de modulação. Biologicamente, os processos de transcrição e tradução têm como objetivo a transmissão da informação. Neste processo podem ocorrer erros, ocasionando interferências na informação, por exemplo, a não leitura de um códon, que acarreta uma perda de pareamento do ribossomo, levando a mutações**.** O receptor pode ser modelado como o local para onde a informação está sendo enviada, no sistema de comunicação. Neste estudo, biologicamente, a informação é a proteína.

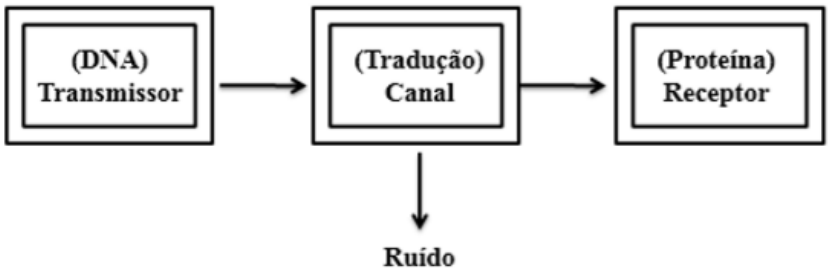

Figura 7: Analogia entre sistemas de comunicação e síntese proteica. Fonte: Faria e Palazzo Junior (2011).

2.4. Rotulamentos do Código Genético

A definição de um rotulamento permite entender a conexão entre as estruturas biológicas e estruturas algébricas. Os rotulamentos do código genético foram definidos inicialmente por Rocha (2010) como um modelo que codifica e decodifica o mecanismo de importação de proteínas mitocondriais, sendo o mesmo realizado através de uma bijeção entre um alfabeto biológico N= {A,C,G,T/U}, onde A, C, G, T/U representam as bases nitrogenadas adenina, citosina, guanina, timina/uracila, respectivamente, e um alfabeto $\mathbb{Z}_4$ = {0,1,2,3}.

O mapeamento $N \rightarrow \mathbb{Z}_4$ é composto de 24 permutações, divididas em 3 rotulamentos: A, B e C. A classificação dos rotulamentos é realizada de acordo com formas geométricas, que produzem um nível de não-linearidade das sequências. No rotulamento A, as permutações geram o mapeamento $\mathbb{Z}_4$ - linear; no B, geram o mapeamento $\mathbb{Z}_2 x \mathbb{Z}_2$- linear; em C, geram um mapeamento Klein - linear. A Figura 8 mostra todas as permutações possíveis.

Faria (2011), por sua vez, apresentou uma modelagem matemática para o problema de Rocha (2010), explicando de forma detalhada a caracterização dos rotulamentos A, B e C, comprovando a existência dos protocolos de comunicação em sequências de DNA. Os rotulamentos, da forma que foram descritos por Faria, abordam a complementaridade algébrica ou biológica de forma interessante, e que proporciona um melhor entendimento da classificação.

No rotulamento A, a complementaridade algébrica $(00-11)/(01-10)$ é associada à complementaridade biológica $(A-T)/(C-G)$. Nos rotulamentos B e C não existe a associação entre complementaridade biológica e algébrica, tomando como base apenas a complementaridade algébrica: no rotulamento B, a união das bases se dá por $(A-G)/(C-T)$, enquanto no rotulamento C, se dá por $(A-C)/(G-T)$, como visto na Figura 9.

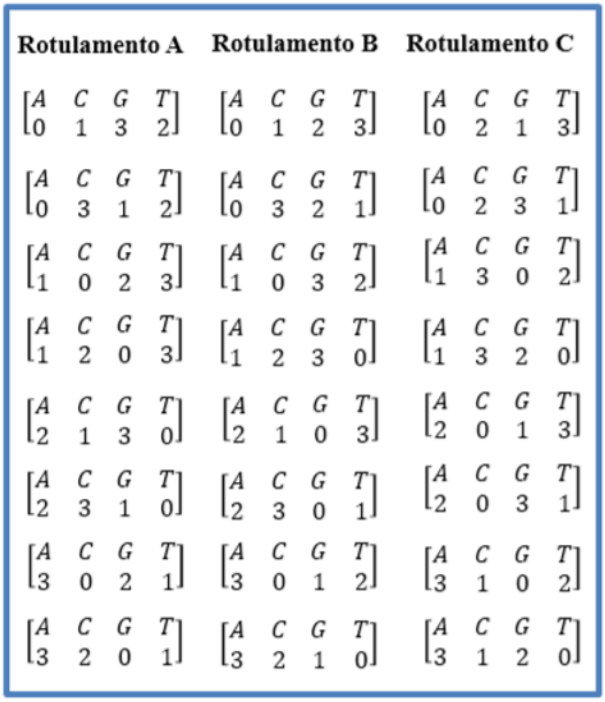

Figura 8: Permutações possíveis nos rotulamentos A, B e C. Fonte: Faria e Palazzo Junior (2011).

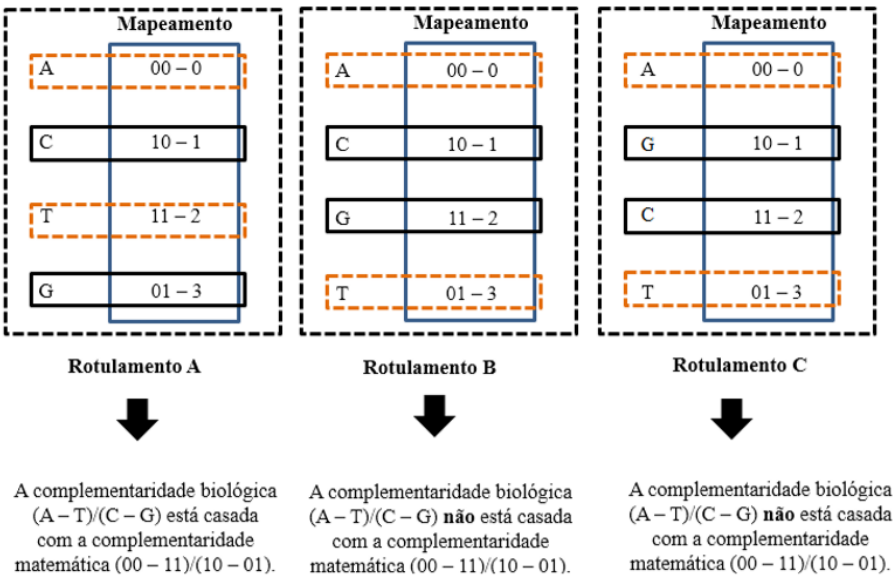

Figura 9: Complementaridade biológica nos rotulamentos A, B e C. Fonte: Faria e Palazzo Junior (2011).

Outro fator levado em conta para a construção dos rotulamentos foi a geometria. No rotulamento A, as complementaridades geram um mapeamento não-linear: dispondo os elementos mapeados nas arestas de um quadrado neste rotulamento, qualquer nucleotídeo precisa caminhar duas arestas para encontrar seu complementar biológico. Em B e C, os mapeamentos são lineares: as arestas são vizinhas e o nucleotídeo encontra seu complementar ao caminhar apenas uma aresta, conforme a Figura 10.

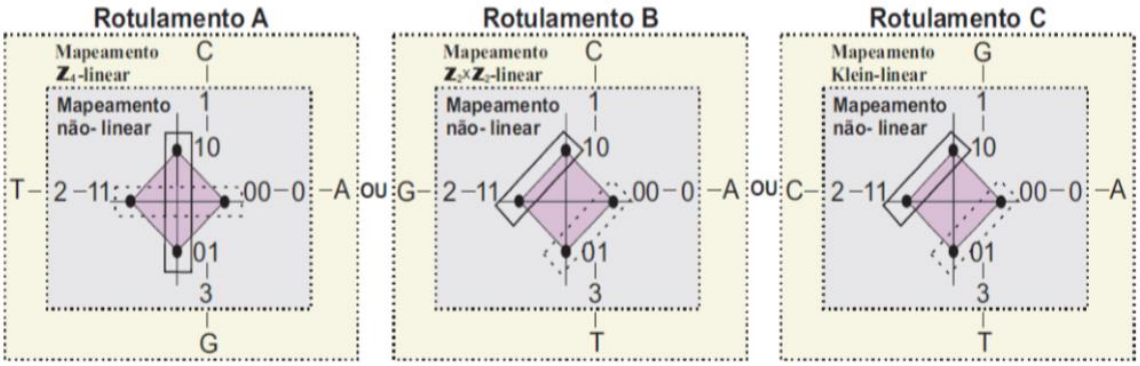

Figura 10: Caracterização geométrica dos rotulamentos A, B e C. Fonte: Faria e Palazzo Junior (2011).

3. Modelagem e Implementação do Algoritmo Soma com Transporte

Neste Capítulo será apresentada a modelagem e implementação do Algoritmo Soma com Transporte, demonstrando seu funcionamento, características e procedimentos que foram executados para sua utilização de forma satisfatória, possibilitando a análise dos resultados obtidos. Também serão expostos os dados utilizados nas simulações com sequências fictícias (com o intuito de realizar testes) e sequências reais.

Para todos os casos analisados, foi utilizado um computador equipado com um processador Intel i5-4200U e 8 GB de memória RAM, instalado com as versões mais atuais do Windows 10 e do Python, além de IDEs que facilitaram as execuções sequenciais.

3.1. Modelagem Algébrica

A modelagem algébrica do código genético, como visto em Oliveira (2012), tem o intuito de identificar propriedades, características e implicações biológicas, associando os cálculos aos códons do código genético, como na soma de códons e na separação de códons pela paridade.

Através da modelagem algébrica, associada aos rotulamentos anteriormente vistos, é possível identificar estruturas de grupos no contexto biológico e matemático, o que pode ser utilizado como ferramenta em análises mutacionais. Por exemplo, um organismo afetado por uma bactéria pode sofrer alterações em sua estrutura genética, causada pela alteração de um ou mais códons, podendo ser tanto uma mutação de pequena escala, como uma mutação de grande escala.

Para realizar a análise mutacional, podemos utilizar algoritmos que aplicam a modelagem algébrica no contexto biológico, entre eles, a Soma com Transporte, descrito por Oliveira (2012), e o SMG (Sanchez, Morgado e Grau), descrito por Sanchez et.al (2005).

O funcionamento do Algoritmo Soma com Transporte se dá através dos seguintes passos:

Passo 1 - Especificar o rótulo utilizado.

Passo 2 - Somar as terceiras bases em $\mathbb{Z}_4$. Se o valor encontrado for superior a 3, transporte 1 para a próxima soma.

Passo 3 - Somar as primeiras bases em $\mathbb{Z}_4$, adicionando 1, caso a soma do passo 2 seja superior a 3. Se o valor encontrado for superior a 3, transporte 1 para a próxima soma.

Passo 4 - Somar as segundas bases em $\mathbb{Z}_4$, adicionando 1, caso a soma anterior seja superior a 3.

Este algoritmo foi inspirado no algoritmo SMG (Sanchez, Morgado e Grau), o qual segue os seguintes passos:

Passo 1 - as bases correspondendo a terceira posição são adicionadas de acordo com a tabela soma.
Passo 2 - se a base resultante da operação soma é anterior à base adicionada (a ordem no conjunto de bases), então o novo valor é escrito e a base C é adicionada à próxima posição.
Passo 3 - as outras bases são adicionadas de acordo com a tabela soma, passo 2, indo da primeira para a segunda base.

$+$	A	C	G	U
A	A	C	G	U
C	C	G	U	A
G	G	U	A	C
U	U	A	C	G

$+$	U	G	C	A
U	U	G	C	A
G	G	C	A	U
C	C	A	U	G
A	A	U	G	C

Figura 11: Tabelas soma para os casos Primal e Dual. Fonte: Oliveira (2012).

O algoritmo de Soma com Transporte, proposto por Oliveira, busca diminuir a complexidade em relação a algoritmos semelhantes utilizados em outras abordagens, como o SMG. Isso é feito através do uso da operação soma entre códons com uma associação ao mapeamento $N \rightarrow \mathbb{Z}_4$ nos rotulamentos A, B e C.

3.2. Implementação do Algoritmo

A descrição do Algoritmo Soma com Transporte e da estrutura necessária para complementar seu funcionamento é exposta através do pseudocódigo 01.

```
1. inicioAlgoritmo
2.    determine um rotulamento, dentre A, B ou C;
3.    traduza as sequências, dado o rotulamento, para a base Z4;
4.    enquanto as sequências não forem completamente executadas,
repita, trinca por trinca:
5.       somar a terceira base;
6.       se a soma for superior a 3, transportar 1 para a próxima
soma;
7.       somar a primeira base;
8.       se a soma for superior a 3, transportar 1 para a próxima
soma;
9.       somar a segunda base;
10.      se a soma for superior a 3 e ainda existirem trincas,
transportar 1 para a próxima soma;
11.   fimEnquanto;
12.   traduza a soma, dado o rotulamento, para a base nitrogenada;
13.   retorne o resultado da soma;
14.   fimAlgoritmo
```

Pseudocódigo 01: Algoritmo soma com transporte. Fonte: Oliveira (2012).

Apesar do Algoritmo Soma com Transporte ser bem definido, alguns pontos necessitam ser trabalhados para que os resultados estejam dentro do esperado, sem que os recursos da máquina sejam esgotados. Para isso, foram criadas funções que padronizam a entrada e saída do algoritmo, e removem informações indesejadas ou que estejam fora do escopo do trabalho.

Além disso, foi implementada uma estrutura de controle para que as trincas não sejam processadas de forma incorreta, tal como a associação entre posições indesejadas, ou que as operações continuem mesmo após a descoberta de uma instrução de parada. Para garantir que qualquer tamanho de sequência seja processada em tempo satisfatório, foram utilizadas hashmaps para a realização dos mapeamentos e associações, o que também simplificou o funcionamento da Soma com Transporte.

O funcionamento do pseudocódigo que simula a Soma com Transporte é ilustrado na Figura 12.

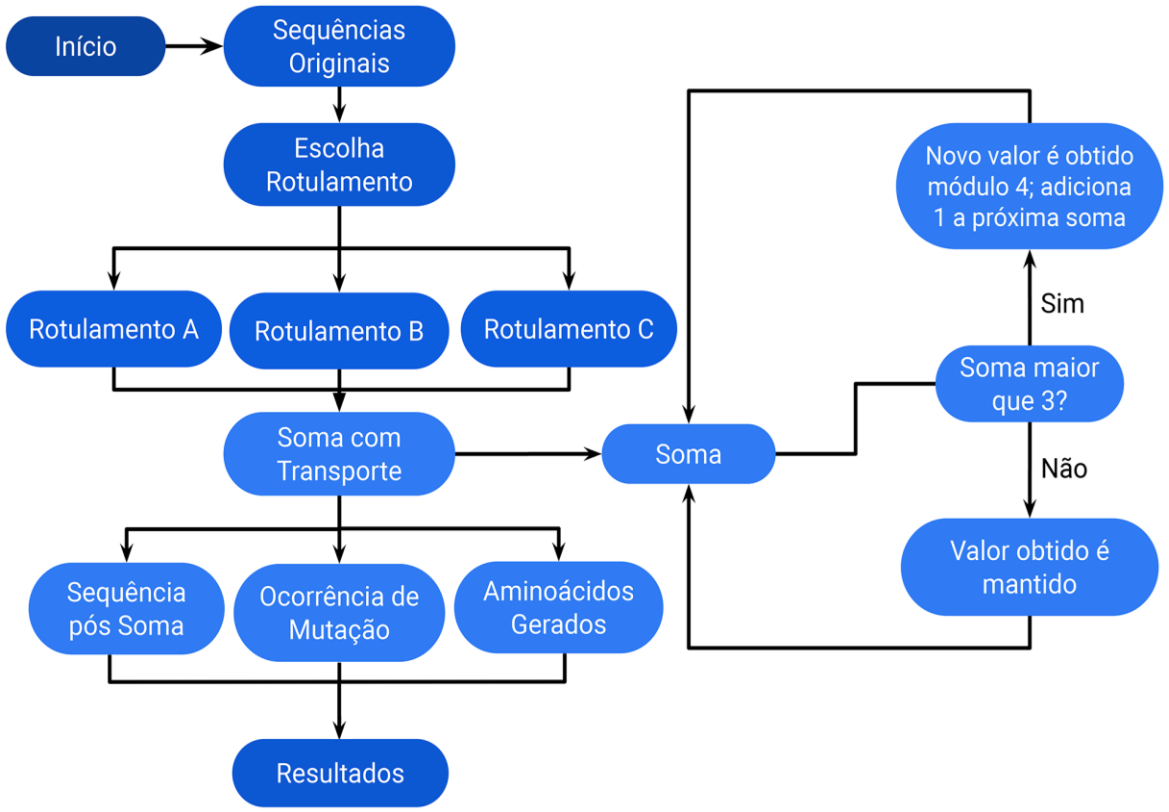

Figura 12: Diagrama de funcionamento do pseudocódigo.

3.3. Execuções com Sequências Fictícias

Para testar a eficácia do algoritmo na utilização do algoritmo Soma com Transporte, foram realizados os seguintes testes, manualmente:

Sequência 1	GUG
Sequência 2	AAU
Rotulamento A	AGU
Rotulamento B	UUC
Rotulamento C	CAG

Tabela 2: Execução da simulação com 1 trinca.

Sequência 1	GUGGCG
Sequência 2	AAUCUA
Rotulamento A	AGUAAG
Rotulamento B	UUC**UAG**
Rotulamento C	CAGAUC

Tabela 3: Execução da simulação com 2 trincas.

Na execução com uma trinca, todos os rotulamentos apresentam o mesmo resultado e nenhuma das sequências apresenta fatores que impedem a síntese proteica. Já na execução com duas trincas, apesar dos rotulamentos apresentarem resultado idêntico a uma simulação manual, podemos notar que, para o rotulamento B, foi codificada uma trinca com a instrução STOP, que corresponde a uma instrução que interrompe a síntese proteica e, como nenhuma das sequências originais possuía essa instrução, a ocorrência de STOP mostra a ocorrência de uma mutação com perda de sentido. Desta forma, o algoritmo deve desprezar qualquer informação que se localiza após UAG.

Sequência 1	GUGGCGCGGAAAAAGCCCCACUGUUAUCGAUUCGCGGAGGGG
Sequência 2	AAUCUACUCACCACAACGGCUGCCGCAUUAUACAGACGAAGC
Rotulamento A	AGUAAGGCCACUACGUUAAUGCCGCUGGCAAGGGAGCAAAGC
Rotulamento B	UUC**UAG**
Rotulamento C	CAGAUCCCGCUUCUUACGGAUCAUGUGCGCAAGCACGUU**UAG**

Tabela 4: Execução da simulação com 14 trincas.

No caso da simulação com 14 trincas, entre as codificações, podemos perceber que, além do rotulamento B, o rotulamento C codifica uma trinca com a instrução STOP. Logo, ela terá o mesmo comportamento de B, e desprezará as informações posteriores a UAG.

Sequência 1	GUGGCGCGGAAAAAGCCCCACUGUUAUCGAUUCGCGGAGGGGAGUUAUCCCCCACCGCUGUUU
Sequência 2	AAUCUACUCACCACAACGGCUGCCGCAUUAUACAGACGAAGCAGGGGGUGUUAUUACUUACCA
Rotulamento A	AGUAAGGCCACUACGUUAAUGCCGCUGGCAAGGGAGCAAAGCCUUUAUACAGCGAUAGAGGGG
Rotulamento B	UUC**UAG**
Rotulamento C	CAGAUCCCGCUUCUUACGGAUCAUGUGCGCAAGCACGUU**UAG**

Tabela 5: Execução da simulação com 21 trincas.

Nesta execução, foram utilizados 63 nucleotídeos por Base, formando 21 trincas. Com estas bases fictícias, apenas o rotulamento A foi capaz de chegar ao fim da codificação sem gerar uma instrução STOP. Por gerar esta instrução, os rotulamentos B e C sofrem uma mutação com perda de sentido e deleção, já que, além de codificar uma instrução de parada, remove parte dos nucleotídeos das sequências originais, afetando a síntese proteica e modificando as proteínas que deveriam ser geradas.

3.4 Execuções com Sequências Reais

Nesta Seção serão apresentados resultados de alguns casos particulares analisados, criados a partir da combinação entre uma sequência e uma lista de sequências, que possuem a mesma quantidade de nucleotídeos da sequência inicial. Para 255 nucleotídeos, serão analisadas as somas entre a sequência ***O.Sativae*** e as sequências ***A.Thaliana***, ***A.marina***, ***A.denitrificans***, ***A.hospitalis***, e para 1023 nucleotídeos, as somas entre ***A.pasteurianus*** e

as sequências ***B.subtilis***, ***A.marina***, ***B.marismortui*** e ***H.sapiens***. Estas sequências foram obtidas por meio de resultados gerados em Faria e Palazzo Júnior (2011).

Com o intuito de facilitar a visualização dos dados, visto que as sequências podem possuir até 85 aminoácidos (para 255 nucleotídeos), ou até 341 aminoácidos (para 1023 nucleotídeos), foi gerado um alfabeto que associa a abreviação do aminoácido, seu nome real, e sua correspondência nos resultados que serão apresentados na Tabela 7.

Ala	Alanina	A
Arg	Arginina	B
Asn	Asparagina	C
Asp	Ácido Aspártico	D
Cys	Cisteina	E
Gln	Glutamina	F
Glu	Ácido Glutâmico	G
Gly	Glicina	H
His	Histidina	I
Ile	Isoleucina	J
Leu	Leucina	K
Lys	Lisina	L
Met	Metionina	M
Phe	Fenilalanina	N
Pro	Prolina	O
Ser	Serina	P
Thr	Treonina	Q
Trp	Triptofano	R
Tyr	Tirosina	S
Val	Valina	T

Tabela 6: Dicionário de aminoácidos, abreviações e suas correspondências.

A partir da execução do algoritmo anteriormente descrito, e utilizando as sequências desejadas, obtivemos os resultados disponibilizados nos anexos 1 a 24 deste trabalho. Deste resultado, podemos comparar o tamanho da sequência original e o tamanho da sequência obtida através da Soma com Transporte, além de visualizar a ocorrência de mutações. As Tabelas 8 a 15 comparam os aminoácidos das sequências originais, e os obtidos pela Soma com Transporte, levando em conta os diferentes rotulamentos:

O. sativae	HKTRNGQ@JBKQCJ@@NL@EKTPJRNGALN@IA@GCMPNF@@Q@AJKQPJFQFKSKQLJPIALQEFCKAKQLNHLASKHILOCFO
A. thaliana	MQLBGSCPFOGMKGHALPJHAHAAQJAPAHAAJHJHCTNPPKJIPTABCOPKALFPNHSAJKHNAKQGAJAKNAOMMANKJKNTN
Rotulamento A	M@
Rotulamento B	PIKOGAOHKKCFLKOBNPKAOPT@
Rotulamento C	PNIKDHTKRPPOLFTTLMDTOPHOCJ@

Tabela 7: Comparação entre as sequências *O. sativae*, *A. thaliana* e as sequências obtidas através da Soma com Transporte, com 255 nucleotídeos, levando em conta as regras dos rotulamentos A, B e C.

O. sativae	HKTRNGQ@JBKQCJ@@NL@EKTPJRNGALN@IA@GCMPNF@@Q@AJKQPJFQFKSKQLJPIALQEFCKAKQLNHLASKHILOCFO
A. marina	MMQQKKPNKADKPCIKHKARRTGJLQKPOJEQSNNHONKJBLGAGAAKNHSTGDKGAGFAFCJTAIJFBEIOOIKQJECGMGSPE
Rotulamento A	M@
Rotulamento B	PHSLKIKDHKLFBKJLHKBPHKHBEQIKOEPAHDGEIPKTJRABPDOQPPOGSLGLHGTTSAJ@
Rotulamento C	PCOLKDDL@

Tabela 8: Comparação entre as sequências *O. sativae*, *A. marina* e as sequências obtidas através da Soma com Transporte, com 255 nucleotídeos, levando em conta as regras dos rotulamentos A, B e C.

O. sativae	HKTRNGQ@JBKQCJ@@NL@EKTPJRNGALN@IA@GCMPNF@@Q@AJKQPJFQFKSKQLJPIALQEFCKAKQLNHLASKHILOCFO
A. denitrificans	MPGACJBKGEKBOACDHRGFOQHGGTBGAKLAAHNQHHFAALAKHKHALHDBQTBBRJHHDPAJOSAARAKKEDNHCKHFJRLLD
Rotulamento A	MLHOTBFTTNDSPHEOKHPCLCPTSPCTAQEKHIBOAOTJN@
Rotulamento B	PQFICKJCKSRTADROHEAKQ@
Rotulamento C	PORDL@

Tabela 9: Comparação entre as sequências *O. sativae*, *A. denitrificans* e as sequências obtidas através da Soma com Transporte, com 255 nucleotídeos, levando em conta as regras dos rotulamentos A, B e C.

O. sativae	HKTRNGQ@JBKQCJ@@NL@EKTPJRNGALN@IA@GCMPNF@@Q@AJKQPJFQFKSKQLJPIALQEFCKAKQLNHLASKHILOCFO
A. hospitalis	MPKTFKTGLTALLSCJLTCPKOCHTJJKTLCDJHSTFJAATBCTSSTBSKQLCGASJJILKCGGTJGRJKGGLKDGQLAKLJODT
Rotulamento A	MPONTBHOJAIOITENNTDODIAJK@
Rotulamento B	PATHFLSOJAQOILJNNTQ@
Rotulamento C	PQTHFHOHGBMKHCQDLFBS@

Tabela 10: Comparação entre as sequências *O. sativae*, *A. hospitalis* e as sequências obtidas através da Soma com Transporte, com 255 nucleotídeos, levando em conta as regras dos rotulamentos A, B e C.

A. pasteurianus	TBOBTOKPDIPHTBCAKHITFQPJOGTKCIAOCSPBBEOQBIAOGPBCKHTRQKFIAOBTBDAOHHBPDOBDPKRKASHAOAKBPKPIBIBKAHHTHQ AFHGHTEBPPPOOABKPKEKDGBBABHTIKCQBKPKHJBOIFQOGPBJBBIJAGKLKGPKPDBODABKTOIAAPKOKOABBFQMHHQKCMOPAQBB TKBSJKBOAFPTJDTKDBKBOKKBOTHHBFMBODHSLOMNIPBBOAPPKSHNKPGJIHOMFTBCAAKKOKPDOTFAIHBBHCKBPPIGPCBLQJCFO BOTKPPKBAEKCHGHFBGOSATLNTTNDBOGGQIJPQPCPBBOAAQBQQN
B. subtilis	MJJPSLEOCEHPDMANDPGQHPKPEPPEHBFDCJGPKOLGCJAABNPDDGALGSFELCEHATKJQGAGQQAQQEPNEHHAAJKADBKPHIKAOAL TJONQJPLFGAGFANBLRELLHKKQOBHNMPADBJLPJQHMSJONRMNDKCPGTFTBACEQBTIFSGGHDSJEQGQGINGANBDJCKDSKLJOTDA PGLMLDGKMDLKGOSPSGGKLDNFQASKAHSJAGLSCSQDGGKNOBALGLJPPSJDPSJIPQNPHSQPTCTBDLIJIQLCTCPNSTKKOTRMTPSDS GBAGIJNAMCHFQHLTTHLOOJPBHLTAARNPHJAHHQNKAKLKTPKMMHHHN
Rotulamento A	QPBDEPJSQHOBHQAIBTQHCTCRDCJJHBKKFKHKAQOAPKHBLAGQBSSHAB@
Rotulamento B	@
Rotulamento C	SPKCGLBBPOISPHKAHKAKTHPAEKPPNKGGHCNCA@

Tabela 11: Comparação entre as sequências *A. pasteurianus*, *B. subtilis* e as sequências obtidas através da Soma com Transporte, com 1023 nucleotídeos, levando em conta as regras dos rotulamentos A, B e C.

A. pasteurianus	TBOBTOKPDIPHTBCAKHITFQPJOGTKCIAOCSPBBEOQBIAOGPBCKHTRQKFIAOBTBDAOHHBPDOBDPKRKASHAOAKBPKPIBIBKAHHTHQ AFHGHTEBPPPOOABKPKEKDGBBABHTIKCQBKPKHJBOIFQOGPBJBBIJAGKLKGPKPDBODABKTOIAAPKOKOABBFQMHHQKCMOPAQBB TKBSJKBOAFPTJDTKDBKBOKKBOTHHBFMBODHSLOMNIPBBOAPPKSHNKPGJIHOMFTBCAAKKOKPDOTFAIHBBHCKBPPIGPCBLQJCFO BOTKPPKBAEKCHGHFBGOSATLNTTNDBOGGQIJPQPCPBBOAAQBQQN
A. marina	MGQGPJGAJTJDKHQPKQLTKSSFFHFSINNAFPPPTKGKDIBLSFFIAAIOFQLQIAJRQGHAHSKTHBHAPFOSGGTDKPFQLAQPAJALTKAAJHG JLDBTCPDLAJLKGCTJKKKOKLGBLGSGQKLQAJTLAKSHNHNCHLQQLNRTHBNCJFSGHSHJQQJPQGGBAANATNHFLDKQJHAMCFHDKQO CMPFQKTHRHMPLTKGGTFSQNGPDTSHALETSPSCFBBQCKLLKJOAGLTGNTPAPJDBAKPPPRSFJGBLKDPPFAKMDATQJSTQHHPPFKRB LFKLHTSHBBKCEKDHJABGJAGQNOFKDLDOMJKBMADASKQJBHMTAA
Rotulamento A	QFKHIBPQBBESOBQBSQIPTPMKPDBPITFBIKMGKAKBHGQPOB@
Rotulamento B	@
Rotulamento C	SJPNAMB@

Tabela 12: Comparação entre as sequências *A. pasteurianus*, *A. marina* e as sequências obtidas através da Soma com Transporte, com 1023 nucleotídeos, levando em conta as regras dos rotulamentos A, B e C.

A. pasteurianus	TBOBTOKPDIPHTBCAKHITFQPJOGTKCIAOCSPBBEOQBIAOGPBCKHTRQKFIAOBTBDAOHHBPDOBDPKRKASHAOAKBPKPIBIBKAHHTHQ AFHGHTEBPPPOOABKPKEKDGBBABHTIKCQBKPKHJBOIFQOGPBJBBIJAGKLKGPKPDBODABKTOIAAPKOKOABBFQMHHQKCMOPAQBB TKBSJKBOAFPTJDTKDBKBOKKBOTHHBFMBODHSLOMNIPBBOAPPKSHNKPGJIHOMFTBCAAKKOKPDOTFAIHBBHCKBPPIGPCBLQJCFO BOTKPPKBAEKCHGHFBGOSATLNTTNDBOGGQIJPQPCPBBOAAQBQQN
B.marismortui	MDBKOOASDAQOHDHAKGAAJGGBKTDJDPHBSBQCTAPTKBLNAQRABDFIHJPPOGDJDDDKEBFSABDKABADDBDDJPOGQABBSNASTBP NKQRATSGHKJOQCOALQCIAGHOKOQDGQGQDFFSRQQBDBDAJEAQAQABTDLAHGPDDJDBQAASBDFAKTNKKASPHABPAGKTATPDDG GBCHKBRBITCKGAHQMFTNHLCBQBGPAOJKDDAKBOKBBRLFKBGODGCGATNOBKDCAALAKDOQOPJQQFPABCJKADKEGRPGSGNGD OKLOIHABBHKHBGJSBGCOFKAFDTKBILPJGQQIGHSAFGAALBQBDGACGJJPDQG
Rotulamento A	QIGPINCRPKNQTHHESSJSKJKPHRH@
Rotulamento B	@
Rotulamento C	SMBPMOMAPCBHQOOQIJGJB@

Tabela 13: Comparação entre as sequências *A. pasteurianus*, *B. marismotui* e as sequências obtidas através da Soma com Transporte, com 1023 nucleotídeos, levando em conta as regras dos rotulamentos A, B e C.

A. pasteurianus	TBOBTOKPDIPHTBCAKHITFQPJOGTKCIAOCSPBBEOQBIAOGPBCKHTRQKFIAOBTBDAOHHBPDOBDPKRKASHAOAKBPKPIBIBKAHHTHQ AFHGHTEBPPPOOABKPKEKDGBBABHTIKCQBKPKHJBOIFQOGPBJBBIJAGKLKGPKPDBODABKTOIAAPKOKOABBFQMHHQKCMOPAQBB TKBSJKBOAFPTJDTKDBKBOKKBOTHHBFMBODHSLOMNIPBBOAPPKSHNKPGJIHOMFTBCAAKKOKPDOTFAIHBBHCKBPPIGPCBLQJCFO BOTKPPKBAEKCHGHFBGOSATLNTTNDBOGGQIJPQPCPBBOAAQBQQN
H.sapiens	MGGOFPDOPTGOOKPFGQNPDKRLKKOGCCTKPOKOPFAMDDKMKPODDJGFRNQGDOHODGAOBMOGAAOBTAOAOAAOQOAAOAOAOP ROKPPPTOPFLQSFHPSHNBKHNKIPHQALPTQEQSPOAKCLMNEFKALQEOTFKRTDPQOOOHQBTBAMAJSLFPFIMQGTTBBEOIIGBEPDPD HKAOOFIKJBTGHCKBTGSKDDBCQNBIPTTTOSGOOGTHPDEQQJISCSMECPPEMHHMCBBOJKQJJQKGDPPHCKKHBCPNGTBTEAEOHBD BBQGGGCKBLLHGOIIGKOOHPQLBAKOCCQPPPOFOLLLOKDHGSNQKFDFQPNFLGCE
Rotulamento A	QFKSBTBFLBKQFPBMPDBIQBOJ@
Rotulamento B	@
Rotulamento C	STDBPPBENHTHTQACHEEOAJSFTPJHBPNKTSHBRKQOOQOJHGBPPBPOCS@

Tabela 14: Comparação entre as sequências *A. pasteurianus*, *H. sapiens* e as sequências obtidas através da Soma com Transporte, com 1023 nucleotídeos, levando em conta as regras dos rotulamentos A, B e C.

4. Resultados e Discussões

Neste Capítulo serão analisados os resultados obtidos durante as execuções apresentadas ao longo do Capítulo 3, através da análise das possíveis mutações e da comparação entre rotulamentos para uma mesma sequência.

Através das Tabelas 7 a 14, nota-se que para todas as somas entre sequências e os rotulamentos definidos, foram gerados códons que levam à interrupção do processo de síntese proteica, levando a uma mutação sem sentido que geram a instrução STOP (definida como o símbolo '@' no dicionário da Tabela 4.2.1). A ocorrência de tal mutação gera sequências menores que as originais, logo, menos aminoácidos e proteínas são codificados, gerando um novo organismo que herda partes do código das sequências originais, mas possui novas características, que podem ser positivas ou negativas, dependendo da análise dos efeitos dessa mutação.

Um resumo dos resultados permite uma melhor visualização das diferenças entre as execuções, conforme apresentado nas Figuras 13 e 14.

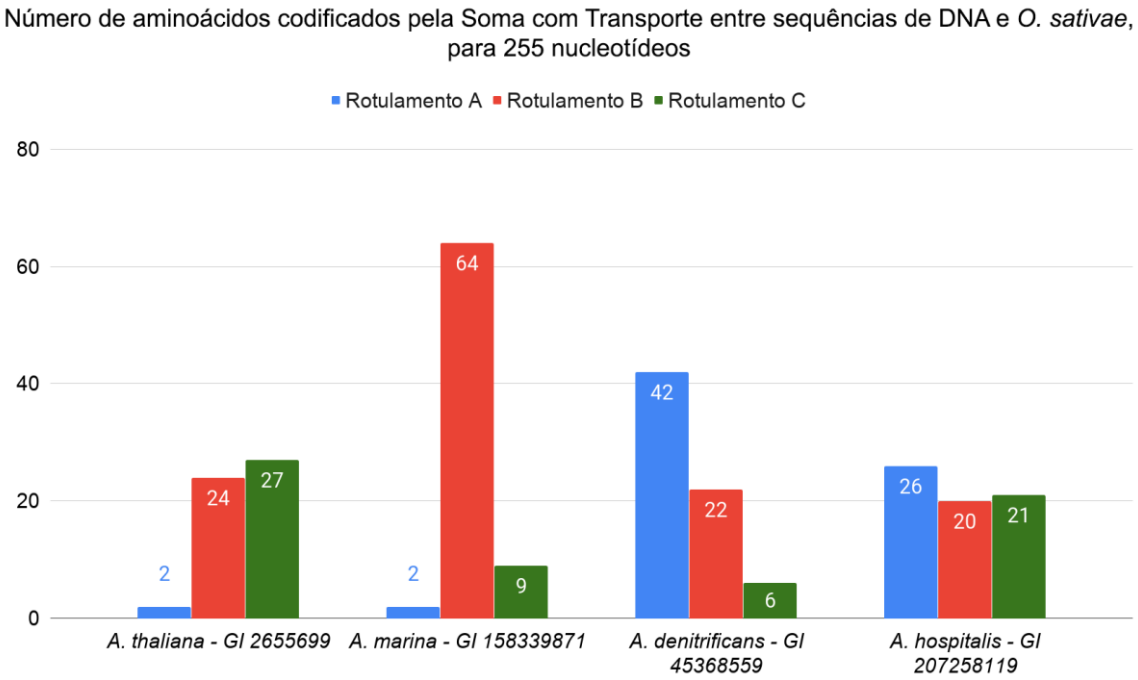

Figura 13: Comparação entre rotulamentos, referente ao número de aminoácidos codificados entre as sequências, para 255 nucleotídeos.

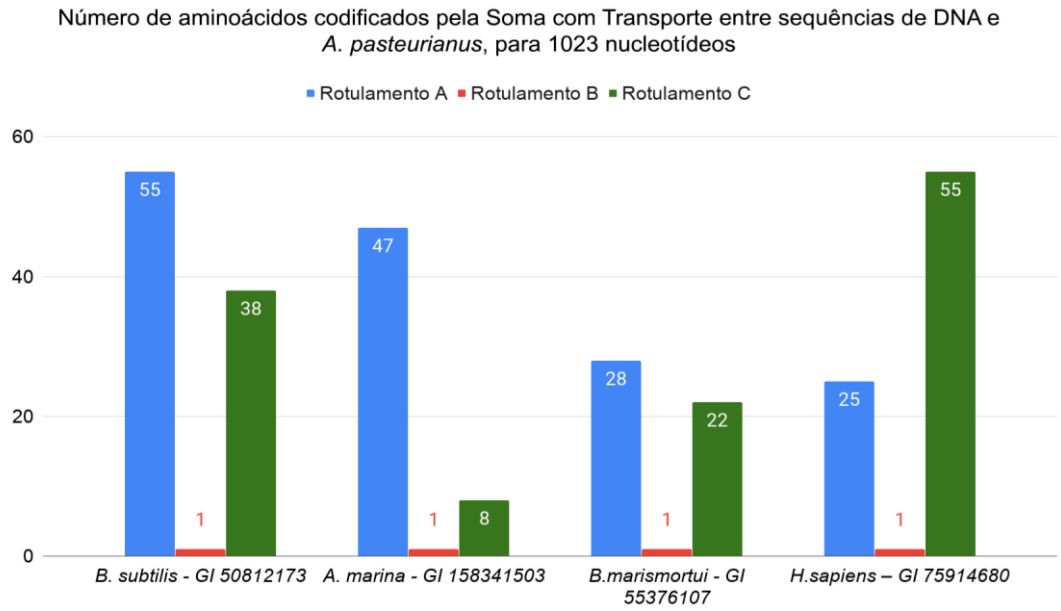

Figura 14: Comparação entre rotulamentos, referente ao número de aminoácidos codificados entre as sequências, para 1023 nucleotídeos.

Através dos gráficos, pode-se notar algumas particularidades das mutações, e como cada rotulamento afeta a síntese proteica, lembrando que, para 255 nucleotídeos, podem ser gerados até 85 aminoácidos, enquanto 1023 nucleotídeos geram no máximo 341 aminoácidos.

Executando a implementação para 255 nucleotídeos, destaca-se o rotulamento B, que retorna até 64 aminoácidos antes de atingir uma codificação de parada. Entretanto, para 1023 nucleotídeos, os rotulamentos A e C se destacam, enquanto o rotulamento B obtém uma instrução de parada logo na primeira codificação.

Outra análise interessante para o problema leva em conta a porcentagem de sucesso na síntese proteica, conforme mostra a Figura 15.

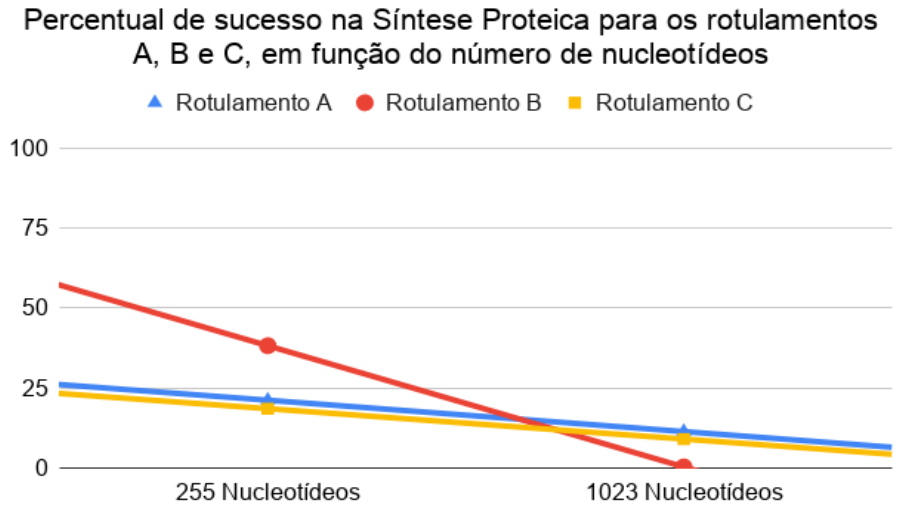

Figura 15: Comparação entre as porcentagens de sucesso da síntese proteica para 255 e 1023 nucleotídeos, levando em conta os rotulamentos A, B e C.

Para obter o percentual de sucesso, criou-se uma razão entre a média do número de aminoácidos codificados e o número total que pode ser obtido para 255 e 1023 nucleotídeos, e então, este número foi multiplicado por 100, a fim de obter o resultado desejado. Desta forma podemos visualizar a performance dos rotulamentos para sequências de diferentes tamanhos.

Traçando uma reta entre os pontos destacados na figura 15, nota-se que o desempenho dos rotulamentos tende a variar conforme o tamanho das sequências. Por exemplo, o rotulamento B varia de 38% para 0,29% de sucesso na síntese. Isso pode ser explicado pelas características das sequências com as quais estamos trabalhando e que, quando passam pela Soma com Transporte, geram a instrução de parada logo na primeira ocorrência, em alguns casos. Entretanto, os outros rotulamentos mostraram um comportamento um pouco mais consistente quando variamos o tamanho das sequências, dentro do mesmo conjunto de dados.

Outro fator a ser destacado é como uma execução de 255 nucleotídeos retorna um resultado melhor do que uma de 1023. Isso se deve por alguns fatores, como a aleatoriedade das sequências e a ocorrência de instruções de parada, que podem aparecer após a soma das sequências. Este comportamento é esperado, demonstrando como uma mutação pode surgir, em um processo trivial para os seres vivos, mas que, se não for controlada ou contida por meio de correções, causa mudanças que podem ser benéficas ou maléficas.

É importante destacar que as combinações de nucleotídeos após a soma são muitas, e quanto maior a sequência, mais sequências podem ser obtidas. Logo, por esse grande número de sequências que podem ser obtidas, é difícil encontrar uma que já foi mapeada e caracterizada.

5. Conclusões e Sugestões para Trabalhos Futuros

Como visto, tanto a genética quanto a teoria de comunicações utilizam processos de transferência de informação, e são objetos de estudo a muito tempo por muitos pesquisadores, explorando as relações entre essas áreas e buscando uma representação matemática que explique estes processos biológicos. É importante lembrar que durante o envio das mensagens, o canal pode sofrer interferências que comprometem os dados originais. Por meio deste trabalho, podemos mostrar a interdisciplinaridade deste tópico, que aborda conceitos de Biologia, Matemática e Computação, com o intuito de explicar o que pode ocorrer durante a síntese proteica.

Ao utilizarmos métodos computacionais para simular o processo de síntese, ganhamos velocidade para realizar cálculos repetitivos e que podem acarretar problemas na realização de operações manuais. Além disso, podemos explorar em mais detalhes os resultados obtidos em cada execução, e comparar fatores que influenciam nas sequências, como o rotulamento e a quantidade de nucleotídeos contidos.

Um trabalho interdisciplinar permite o estudo de uma ampla gama de conceitos que, apesar de não parecerem próximos, possuem interessantes paralelos. Também fica evidente como uma simples informação pode influenciar um organismo, seja de forma positiva ou negativa, através de diferentes sequências ou pela ocorrência das mutações e interferências.

Os objetivos propostos neste estudo foram alcançados e, com o uso da ferramenta, podemos determinar quais são os efeitos de uma interferência no processo de síntese proteica. Desta forma, este estudo interdisciplinar pode auxiliar em estudos genéticos e na detecção de erros, reduzindo tempos de simulação e simplificando as operações necessárias que, por sua vez, facilitam o trabalho de cientistas.

A partir deste trabalho, foi gerado um resumo na forma de pôster, selecionado e apresentado no Encontro Acadêmico de Modelagem Computacional, realizado no Laboratório Nacional de Computação Científica (LNCC), em fevereiro de 2020, na cidade de Petrópolis, RJ. O mesmo consta nos Anais do evento, e mostra a relevância do tópico para a comunidade científica.

A fim de dar prosseguimento a este estudo, apresentamos algumas sugestões, tais como: o estudo aprofundado dos reais efeitos das sequências geradas em um organismo; a utilização de sequências maiores e mais complexas, contendo mais nucleotídeos; a comparação dos resultados das sínteses com sequências já existentes e documentadas, e que constam em bancos de dados.

Referências

ALBERTS, B. et al. **Biologia Molecular da Célula**. 5. ed. Porto Alegre: Artmed, 2010.

DUQUE, N. **Células Eucariontes.** 2018. Disponível em: <https://www.estudopratico.com.br/celulas-eucariontes>. Acesso em: 25 jun. 2020.

BATTAIL, G. **An Outline of Informational Genetics**. [S.l.]: Morgan & Claypool Publishers, 2008.

DOMINGUES, H. H.; IEZZI, G. **Álgebra Moderna.** 4. ed. São Paulo: Atual, 2003.

FARIA, L. C. B.; PALAZZO JÚNIOR, R. **Existências de Códigos Corretores de Erros e Protocolos de Comunicação em Sequências de DNA.** 2011. 147 f. Tese (Doutorado em Engenharia Elétrica) - Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas, Campinas, 2011.

GARCIA, A.; LEQUAIN, Y. **Elementos de Álgebra.** 5. ed. Rio de Janeiro: IMPA, 2010.

LIN, S.; COSTELO JR, D. J. **Error Control Coding:** Fundamentals and Applications. New Jersey: Prentice Hall, Englewood Clisffs, 1983.

MARZZOCO, A.; TORRES, B.B. **Bioquímica Básica**. 3. ed. [S.l]: Guanabara-Koogan, 2013.

MAY, E. E. Towards a Biological Coding Theory Discipline. **New Thesis**, v. 1, n. 1, p. 19-38, 2004.

MAY, et al. An error-correcting code framework for genetic sequence analysis. **Journal of The Franklin Institute**, v. 341, n. 23, p. 89-109, 2004.

OLIVEIRA, A. J.; PALLAZO JÚNIOR, R. **Análise Algébrica dos Rotulamentos Associados ao Mapeamento do Código Genético.** 2012. 105 f. Dissertação (Mestrado em Engenharia Elétrica) - Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas, Campinas, 2012.

PAMPHILE, J. A. (Org.); VICENTINI, V. E. P. (Org.). **Genética.** Maringá: EDUEM, 2011.

ROCHA, A. S. L.; PALAZZO JÚNIOR, R.; SILVA-FILHO, M.C. **Modelo de sistema de comunicações digital para o mecanismo de importação de proteínas**

mitocondriais através de códigos corretores de erros. 2010. 155 f. Tese (Doutorado em Engenharia Elétrica) - Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas, Campinas, 2010.

SHANNON, C. E. A mathematical theory of communication. **The Bell System Technical Journal**, v. 27, n. 3, p. 379-423, 1948.

MARQUES, N. **Duplicação do DNA. 2014.** Disponível em: <http://bionel.blogspot.com.br/p/duplicacao-do-dna.html>. Acesso em: 25 jun. 2020.