Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 11, n. 1, 2025.

Trabalho apresentado no XLIII CNMAC, Centro de Convenções do Armação Resort - Porto de Galinhas - PE, 2024

**Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**

# Decodificando a Vida: Explorando os Paralelos entre Códigos Corretores de Erros e a Biologia

Rafaela C. O. Cunha [1], Beatriz Motta[2]
UFJF, Juiz de Fora, MG

O objetivo desse trabalho é apresentar e exemplificar o que foi proposto em [2], seguindo [1].

A Teoria de Códigos é um campo de estudo que é atualmente muito ativo, tanto do ponto de vista teórico quanto tecnológico. Nos mais diversos meios circulam informações que carregam consigo possíveis erros feitos na transmissão. De forma geral, os códigos corretores de erros tem como objetivo a correção de erros que ocorrem durante a transmissão ou armazenamento da informação, e estão presentes na comunicação via satélite, nas comunicações internas de um computador e no armazenamento de dados. Essa teoria surge na década de 1940, no Laboratório Bell de Tecnologia, com os trabalhos de Richard W. Hamming e C. E. Shannon, com objetivo de elaborar mecanismos capazes de permitir uma transmissão confiável de dados através de canais sujeitos a interferências, também denominadas ruídos.

A teoria se baseia no seguinte cenário: buscamos transmitir uma mensagem, a qual denominamos palavra, que consiste numa sequência finita de símbolos, onde tais símbolos são elementos de um alfabeto (um conjunto) finito. Chamamos cada um desses elementos de letras. Por exemplo, tomando como alfabeto o conjunto $\mathbb{Z}_2 = \{0, 1\}$, uma palavra pode ser descrita como um número binário. No caso dos códigos utilizados nesse trabalho, os métodos para melhorar a confiabilidade da transmissão estão intrinsecamente ligados às propriedades dos corpos finitos.

Em um sistema de comunicação digital, a informação é transportada de um transmissor para um receptor por uma sequência de bits passando por um canal de transmissão. Nas células eucarióticas, existe entre outros processos biológicos o transporte de proteínas ($P$) para organelas onde uma informação genética no núcleo se move para o citosol através de intermediários de mRNA (RNA mensageiro), que é posteriormente traduzido em uma proteína precursora contendo uma extensão N-terminal (extremidade da proteína que possui um grupo amino livre) que funciona como uma sequência de direcionamento ($TS$) direcionando a proteína para a organela correspondente. As sequências internas ($IS$) em algumas proteínas precursoras têm a finalidade de sinalizar para os compartimentos submitocondriais corretos de uma organela. Assim, pode-se conceber que um código corretor de erros usado na transmissão de dados através de um canal de ruídos possa ser aplicado à geração de sequências de DNA (como proteínas, por exemplo) em células eucarióticas ou procarióticas. Nesse caso, os conceitos da teoria de códigos podem ser usados adequadamente para modelar os processos de transcrição e tradução.

Uma questão sempre colocada na maioria dos trabalhos relacionados à codificação genômica é: *Existe alguma forma de códigos corretores de erros serem associados à estrutura das sequências de DNA?* Levando em consideração as semelhanças e premissas apresentadas, os autores de [2] propuseram um algoritmo capaz de reproduzir sequências de DNA, associadas a regiões codificadoras de genes, como palavras de códigos corretores de erros. Nesse trabalho é estudamos e exemplificamos essa codificação, conforme [1]. Este resultado permite o uso de simulações computacionais eficientes na análise de processos biológicos como polimorfismos e mutações (erros espontâneos durante a replicação do DNA provocando alterações na sequência dos nucleotídeos) e, por conseguinte,

[1] rafaelaoliveira@ice.ufjf.br
[2] beatriz@ice.ufjf.br

reduzindo o tempo e os materiais gastos em experimentos laboratoriais. A seguir, resumimos o funcionamento da codificação feita.

Embora o código genético tenha um alfabeto próprio, é desejável, como já vimos, que o alfabeto de um código corretor de erros tenha certa estrutura algébrica. Em geral, a identificação de uma estrutura algébrica com o alfabeto genético é um problema em aberto. Aqui, vamos usar $\mathbb{F}_4$ como o alfabeto e, como o código genético deve ser convertido para o alfabeto, e vice-versa, segue-se que essa conversão deve levar em consideração todos os possibilidades de associar os elementos do conjunto $N = \{A, C, G, T\}$, onde A é adenina, C é citosina, G é guanina e T é timina, com os elementos do conjunto $\mathbb{F}_4 = \{0, 1, a, a^2 = b\}$. Chamamos esta associação de ***rotulagem***, tendo 24 permutações envolvidas.

O objetivo dessa rotulagem é determinar qual permutação corresponde à palavra-código com a sequência de DNA fornecida. Em seguida, a fim de corresponder o comprimento da sequência de DNA ao comprimento da palavra-código, devemos encontrar o grau da extensão, denotado por $r$, usando a igualdade $n = 4^r - 1$, onde $n$ é o comprimento da sequência de DNA.

Consideraremos, então, $C$ um código BCH com parâmetros $[n, k, d_H]_4$ sobre $\mathbb{F}_{4^r}$, com $n = 4^r - 1$. Temos:

- $n$ é o comprimento das palavras do código, ou seja, o comprimento das sequências de DNA;
- $k$ é a dimensão do código como um subespaço vetorial de $\mathbb{F}_4^r$, isto é, o comprimento da sequência de informação de entrada responsável por gerar a sequência de DNA.

Lembramos, ainda, que $d_H$ é a distância mínima do código e que $r$ é o grau do polinômio primitivo da extensão $\mathbb{F}_{4^r}|\mathbb{F}_4$.

Agora, como gerar o código BCH? Para cada valor de $r$ (o grau da extensão), há muitos polinômios primitivos $p(x)$ a serem considerados e existe um polinômio gerador $g(x)$ do código BCH que corresponde a cada $p(x)$. A complexidade computacional adicional na solução deste problema vem do fato de que quanto maior o grau da extensão, maior o número de $p(x)$ a serem considerados na construção do código. Nem todo polinômio mônico irredutível é um polinômio primitivo. Para cada um dos $p(x)$ que sejam, devemos encontrar a extensão de $\mathbb{F}_4[x]/\langle p(x)\rangle$ correspondente, o grupo de unidades da extensão e, usando o elemento primitivo, devemos construir o polinômio gerador $g(x)$ do código BCH.

Ainda, sabendo que o número de palavras do código geradas cresce exponencialmente com a dimensão do código, em vez de gerar todas as palavras do código e comparar com a sequência de DNA, as 24 permutações são aplicadas àquela sequência de DNA, e essas sequências são consideradas como "possíveis palavras-código" $v$. Então, para determinar quais são, de fato, palavras-código, a relação $vH^t = 0$ é utilizada. No caso em que temos $vH^t = 0$, a palavra está no código. Caso contrário, devemos verificar o que aconteceria se em cada posição tivéssemos um (dos três outros possíveis) nucleotídeo diferente em cada posição na sequência de DNA, para cada permutação, e novamente usamos a relação $vH^t$ para verificar se $v$ é uma palavra do código.

# Referências

[1] R. C. O. Da Cunha e B. Motta. **Códigos corretores de erros e um exemplo de aplicação na biologia**. Trabalho de Conclusão de Curso de Bacharelado em Matemática - Universidade Federal de Juiz de Fora, Juiz de Fora. 2023.

[2] L. C. B. Faria, A. S. L. Rocha, J. H. Kleinschmidt, R. Palazzo e M.C. Silva-Filho. "DNA sequences generated by BCH codes over GF (4)". Em: **Electronics letters** 46.3 (2010), pp. 203–204.