

Filtro BCH para Validação Pós-MAFFT na Detecção de SNPs do SARS-CoV-2

Marcelo A. C. Fernandes

Resumo— Este trabalho propõe o uso de códigos BCH como filtro algébrico de validação pós-alinhamento para detecção de mutações pontuais (SNPs) na proteína Spike do SARS-CoV-2. As sequências provêm do banco GISAID, onde são alinhadas contra a referência pelo MAFFT, um método heurístico baseado em FFT que não garante alinhamento exato como algoritmos de programação dinâmica. O decodificador BCH atua como segunda camada de controle de qualidade, verificando se o padrão de diferenças entre cada sequência e a referência é compatível com o modelo de Canal Binário Simétrico (BSC), rejeitando sequências cujo alinhamento pode ter falhado silenciosamente. O método dispensa limiares *a priori* de qualidade, delegando ao filtro BCH a rejeição de sequências com densidade de erros excessiva. Resultados sobre 99.509 sequências das variantes Alpha e Delta demonstram recall médio de 0,988, confirmando os SNPs canônicos. O parâmetro estimado do BSC, $\hat{p} \approx 1,3 \times 10^{-3}$, é estável entre códigos e distinto entre variantes, sugerindo seu uso como assinatura de canal.

Palavras-Chave— Códigos BCH, SNP, SARS-CoV-2, filtragem genômica, canal binário simétrico, MAFFT, validação pós-alinhamento.

Abstract— This work proposes BCH error-correcting codes as an algebraic post-alignment validation filter for detecting single nucleotide polymorphisms (SNPs) in the SARS-CoV-2 Spike protein. The sequences are obtained from the GISAID database, where they are aligned against the reference by the MAFFT algorithm, a heuristic FFT-based method that does not guarantee exact alignment as dynamic programming algorithms do. The BCH decoder acts as a second quality-control layer, verifying whether the error pattern between each sequence and the reference is compatible with the Binary Symmetric Channel (BSC) model, rejecting sequences whose alignment may have silently failed. The method eliminates the need for *a priori* quality thresholds, delegating to the BCH filter the rejection of sequences with excessive error density. Results on 99.509 sequences from the Alpha and Delta variants demonstrate an average recall of 0.988, confirming canonical SNPs. The estimated BSC parameter, $\hat{p} \approx 1.3 \times 10^{-3}$, is stable across codes and distinct between variants, suggesting its use as a channel signature.

Keywords— BCH codes, SNP, SARS-CoV-2, genomic filtering, binary symmetric channel, MAFFT, post-alignment validation.

I. INTRODUÇÃO

A vigilância genômica de vírus RNA, como o SARS-CoV-2, depende da identificação precisa de mutações pontuais (SNPs, do inglês *Single Nucleotide Polymorphisms*). Métodos tradicionais baseados em alinhamento de sequências apresentam custo computacional elevado, da ordem de $\mathcal{O}(n^2)$, e são sensíveis a artefatos de sequenciamento, como bases

indefinidas e erros de leitura [1]. Uma dificuldade inerente à análise genômica é distinguir mutações biológicas reais de erros introduzidos pelo processo de medição. Sequências com muitos erros espalhados podem não representar variantes virais legítimas, mas sim falhas instrumentais ou desalinhamentos, comprometendo inferências epidemiológicas.

Na vigilância genômica em larga escala, algoritmos heurísticos como o MAFFT [2] são amplamente utilizados por sua eficiência. O MAFFT emprega a Transformada Rápida de Fourier (FFT) para estimar rapidamente a similaridade entre pares de sequências, utilizando correlações no domínio da frequência sobre representações numéricas dos nucleotídeos, combinadas com correlações baseadas em *k*-mers, para construir alinhamentos múltiplos. No banco GISAID [3], o MAFFT é aplicado para alinhar genomas completos do SARS-CoV-2 contra a sequência de referência WIV04, e a partir desse alinhamento são extraídas as regiões codificantes de todas as proteínas virais, incluindo a Spike, foco deste trabalho. Contudo, por ser heurístico, o MAFFT não garante alinhamento exato como algoritmos de programação dinâmica [4], e desalinhamentos locais podem ocorrer sem serem detectados pelo próprio alinhador.

A aplicação de códigos corretores de erros a sequências de DNA tem sido investigada sob diversas perspectivas: na identificação de sequências como palavras-código de códigos *Bose-Chaudhuri-Hocquenghem* (BCH), revelando SNPs que correspondem a estados ancestrais de nucleotídeos [5]; na avaliação do decodificador PGZ como método de classificação livre de alinhamento sobre $\mathbb{F}_4$ [6]; e na construção de barcodes robustos para sequenciamento de molécula única, com taxas de erro inferiores a 10^{-5} [7]. Esses trabalhos, contudo, focam na análise de sequências individuais ou na construção de barcodes, sem explorar o decodificador BCH como filtro de validação pós-alinhamento sobre conjuntos massivos de sequências genômicas.

Assim, neste artigo, propõe-se uma abordagem baseada em teoria da informação que utiliza códigos BCH [8] como filtro algébrico de validação pós-alinhamento heurístico. O decodificador BCH é aplicado às sequências previamente alinhadas pelo MAFFT para dois fins simultâneos: (i) validação da qualidade do alinhamento, rejeitando sequências cujo padrão de erros é incompatível com o modelo de Canal Binário Simétrico (BSC), o que indica possíveis falhas silenciosas do alinhador, e (ii) detecção direta de SNPs nas sequências validadas, com complexidade $\mathcal{O}(n)$. Esta abordagem cria um *pipeline* de duas camadas: o MAFFT realiza o alinhamento rápido e o BCH valida algebricamente o resultado, combinando eficiência computacional com rigor algébrico. O método é validado sobre a proteína Spike das variantes Alpha (B.1.1.7) e Delta

(B.1.617.2) do SARS-CoV-2, a partir de 99.509 sequências do GISAID [3].

II. DESCRIÇÃO DO PROBLEMA

A. Hipótese

Dada uma sequência viral s de tamanho L , expressa como

$$s = [s_1, s_2, \dots, s_i, \dots, s_L], \quad (1)$$

onde cada i -ésimo elemento s_i pertence a um alfabeto de quatro símbolos correspondentes às bases nitrogenadas, ou seja, $s_i \in \{A, C, G, T\}$. Esta sequência pode sofrer mutações decorrentes de erros introduzidos pela RNA polimerase durante a replicação viral, recombinação genética e pressão seletiva exercida pelo sistema imunológico do hospedeiro [9]. Dentre os tipos de mutação existentes, as mais frequentes em vírus RNA são as substituições de nucleotídeo único, denominadas SNPs, nas quais exatamente um nucleotídeo, s_i , é substituído por outro elemento do alfabeto, mantendo o comprimento L da sequência inalterado [10].

Sob esta perspectiva, uma sequência viral de referência s pode ser interpretada como uma mensagem transmitida por um canal H que introduz alterações pontuais em seus símbolos, produzindo a sequência variante recebida $\hat{s}$. Para que esta analogia seja tratada no arcabouço da teoria da informação, cada símbolo é codificado por $b = 2$ bits conforme o mapeamento $A \mapsto 00$, $C \mapsto 01$, $G \mapsto 10$, $T \mapsto 11$. Assim, a sequência s é convertida em um vetor binário

$$\mathbf{m} = [m_1, m_2, \dots, m_k, \dots, m_{bL}], \quad (2)$$

onde cada k -ésimo elemento satisfaz $m_k \in \{0, 1\}$.

O processo mutacional pode ser modelado como a transmissão do vetor $\mathbf{m}$ por um Canal Binário Simétrico (BSC) de parâmetro p [11], caracterizado como $H(p)$. No BSC, cada bit é invertido de forma independente e identicamente distribuída com probabilidade $p \ll 1$, refletindo o caráter esparsos e decorrelacionado das mutações pontuais em relação ao genoma de referência. A sequência variante corresponde, portanto, ao vetor recebido $\hat{\mathbf{m}} \in \{0, 1\}^{bL}$, e o vetor de erro associado é expresso como

$$\mathbf{e} = \hat{\mathbf{m}} \oplus \mathbf{m}, \quad (3)$$

onde $\oplus$ denota a operação XOR bit a bit. Um SNP que substitui o nucleotídeo s_i por $\hat{s}_i \neq s_i$ se manifesta em $\mathbf{e}$ como um ou dois erros de bit nas posições $2i - 1$ e $2i$, dependendo de quantos bits diferem entre as representações binárias de s_i e $\hat{s}_i$. Sob esta formulação, a identificação de variantes virais reduz-se ao problema clássico de detecção de erros em canais discretos sem memória, para o qual existem soluções algébricas eficientes baseadas em códigos de bloco lineares.

B. Ruídos não naturais

O modelo BSC pode descrever o processo mutacional biológico de forma precisa, porém a sequência recebida $\hat{\mathbf{m}}$ obtida experimentalmente não resulta apenas das mutações naturais. O processo de sequenciamento de nova geração

(*Next-Generation Sequencing* - NGS) introduz três fontes adicionais de perturbação que desviam $\hat{\mathbf{m}}$ do modelo BSC puro, chamadas aqui de ruídos não-naturais.

A primeira fonte são os erros de leitura do sequenciador, nos quais a i -ésima base identificada $\hat{s}_i$ difere da base real por imprecisão instrumental. Esses erros são quantificados pelo escore de qualidade de *Phred* $Q = -10 \log_{10} P_e$, onde P_e é a probabilidade de erro de leitura [1]. No domínio binário, cada erro de leitura equivale a uma perturbação adicional $\boldsymbol{\eta}_r \in \{0, 1\}^{bL}$, de modo que o vetor efetivamente observado passa a ser

$$\hat{\mathbf{m}} = \mathbf{m} \oplus \mathbf{e} \oplus \boldsymbol{\eta}_r, \quad (4)$$

onde $\mathbf{e}$ são os erros biológicos (SNPs) e $\boldsymbol{\eta}_r$ representa os erros instrumentais, tipicamente com $P(\eta_{r,k} = 1) \ll p$, para cada k -ésimo bit. O erro de leitura é muito menos provável que uma mutação biológica (ou o erro do canal p). Um escore *Phred* típico de Q30 corresponde a $P_e = 10^{-3}$, da mesma ordem de $p \approx 1,3 \times 10^{-3}$, e sequenciadores modernos frequentemente atingem Q40 ($P_e = 10^{-4}$), tornando o ruído de leitura desprezível frente às mutações reais.

A segunda fonte são as bases indefinidas, introduzidas quando o sequenciador não consegue determinar com confiança qual base ocupa a posição i . O padrão IUPAC define um conjunto de códigos de ambiguidade para representar estas incertezas: N indica total indefinição (qualquer base), enquanto R (A|G), Y (C|T), S (C|G), W (A|T), K (G|T) e M (A|C) representam ambiguidade entre duas bases, e B (C|G|T), D (A|G|T), H (A|C|T) e V (A|C|G) entre três [1]. No domínio binário, cada código de ambiguidade nas posições de bit $2i - 1$ e $2i$ introduz uma perturbação cuja probabilidade depende do número de bases compatíveis $n_c \in \{2, 3, 4\}$: a probabilidade de erro é $(n_c - 1)/n_c$, que vale $1/2$ para ambiguidade entre duas bases, $2/3$ entre três e $3/4$ para N. Em todos os casos, o símbolo é parcialmente ou totalmente apagado, caracterizando um ruído de apagamento (*erasure*) de intensidade variável, representado por $\boldsymbol{\eta}_a \in \{0, 1\}^{bL}$, com $P(\eta_{a,k} = 1) = (n_c - 1)/n_c$ nas posições afetadas e zero nas demais. Incorporando esta segunda fonte, o vetor observado passa a ser

$$\hat{\mathbf{m}} = \mathbf{m} \oplus \mathbf{e} \oplus \boldsymbol{\eta}_r \oplus \boldsymbol{\eta}_a, \quad (5)$$

onde $\boldsymbol{\eta}_r$ e $\boldsymbol{\eta}_a$ são estatisticamente independentes entre si e de $\mathbf{e}$, com $P(\eta_{r,k} = 1) \ll p \leq P(\eta_{a,k} = 1)$ para as posições indefinidas.

A terceira e mais crítica fonte são os erros estruturais decorrentes do processo de montagem (*assembly*) das leituras curtas (*reads*). As sequências NGS são obtidas em fragmentos de comprimento tipicamente entre 100 e 300 nucleotídeos [1], que são posteriormente alinhados contra um genoma de referência para reconstituir a sequência completa. Este processo pode introduzir inserções e deleções (*indels*), fazendo com que $\hat{s}$ possua comprimento $\hat{L} \neq L$. No domínio binário, um *indel* de δ nucleotídeos desloca todas as posições subsequentes de 2δ bits, produzindo um padrão de erros em rajada (*burst error*) que compromete a correspondência posicional entre $\mathbf{m}$ e $\hat{\mathbf{m}}$. Formalmente, o vetor observado na presença de *indels* pode ser modelado como

$$\hat{\mathbf{m}} = \mathbf{P}_\delta \mathbf{m} \oplus \mathbf{e} \oplus \boldsymbol{\eta}_r \oplus \boldsymbol{\eta}_a, \quad (6)$$

onde $\mathbf{P}_\delta$ é uma matriz de permutação que representa o deslocamento de bits induzido pelos *indels*. Quando $\delta \neq 0$, o vetor de erro resultante $\mathbf{e}' = \hat{\mathbf{m}} \oplus \mathbf{m}$ deixa de refletir apenas os SNPs e passa a incluir erros correlacionados em bloco, invalidando a hipótese do BSC. A presença de *indels* obriga o uso de algoritmos de alinhamento de sequências para restabelecer a correspondência posicional entre $\hat{s}$ e a referência s . Algoritmos clássicos como Smith-Waterman e Needleman-Wunsch [4] resolvem este problema de forma exata por programação dinâmica, porém com complexidade $\mathcal{O}(L^2)$, proibitiva para milhões de sequências. Implementações heurísticas como BWA e Bowtie2 [4] reduzem o custo via indexação por transformada de Burrows-Wheeler, mas introduzem aproximações. Na prática, o GISAID utiliza o MAFFT [2] com a opção `-addfragments`, que emprega correlação via FFT e *6-mers*, tornando-o escalável porém sujeito a desalinhamentos locais. O método proposto explora esta limitação: o filtro BCH, aplicado sobre as sequências já alinhadas, identifica algebricamente aquelas com correspondência posicional correta (erros esparsos, compatíveis com o BSC) e rejeita aquelas com desalinhamentos residuais (erros em rajada que excedem a capacidade de correção t).

C. Codificação de Canal como Solução

O modelo completo da Eq. (6) revela que a sequência recebida $\hat{\mathbf{m}}$ é corrompida por três classes de ruído de naturezas distintas: erros aleatórios esparsos e (SNPs biológicos), perturbações instrumentais η_r (erros de leitura e bases indefinidas) e erros em rajada correlacionados induzidos por $\mathbf{P}_\delta$ (indels). A presença simultânea destas três componentes torna inviável a detecção direta de SNPs sobre $\hat{\mathbf{m}}$ sem uma etapa prévia de seleção. No contexto deste trabalho, as sequências utilizadas já foram submetidas a alinhamento heurístico pelo MAFFT [2] antes da extração das regiões codificantes. O MAFFT busca restabelecer a correspondência posicional comprometida pelos indels, reduzindo o efeito de $\mathbf{P}_\delta$. Contudo, conforme discutido na Seção anterior, o MAFFT pode introduzir desalinhamentos locais sutis que não são detectados pelo próprio alinhador. Estes desalinhamentos residuais manifestam-se como erros em rajada no vetor de diferenças $\mathbf{e}' = \hat{\mathbf{m}} \oplus \mathbf{m}$, de forma análoga ao efeito de $\mathbf{P}_\delta$.

A abordagem proposta consiste em utilizar o decodificador BCH(n, k, t) como filtro algébrico de validação pós-alinhamento sobre $\hat{\mathbf{m}}$, aplicado bloco a bloco: sequências em que todos os blocos possuem no máximo t erros são aceitas (padrão compatível com o BSC, alinhamento correto), enquanto sequências com blocos que excedem t erros são rejeitadas (desalinhamentos residuais ou artefatos). O critério de aceitação reduz o problema da Eq. (6) ao caso $\hat{\mathbf{m}} = \mathbf{m} \oplus \mathbf{e} \oplus \eta_r$, no qual a detecção de SNPs por comparação direta é confiável. A escolha do BCH é motivada pela correção exata de até t erros por bloco com complexidade $\mathcal{O}(n \log n)$ [8].

III. PROPOSTA

A. Pré-processamento

O pré-processamento restringe o conjunto de sequências candidatas a um subconjunto sobre o qual o filtro BCH pode

ser aplicado de forma consistente. As sequências candidatas $\hat{s}$ correspondem às regiões codificantes extraídas por coordenadas fixas a partir do alinhamento produzido pelo MAFFT. Como o MAFFT não garante correspondência posicional exata em todas as regiões, duas condições são impostas sobre cada sequência candidata $\hat{s}$.

A primeira condição exige $\hat{L} = L$, ou seja, a sequência candidata deve possuir o mesmo comprimento da referência após o processo de montagem dos *reads*. Esta restrição não elimina os indels internos nem o desalinhamento local que eles introduzem, mas garante que a representação binária $\hat{\mathbf{m}} \in \{0, 1\}^{bL}$ tenha a mesma dimensão que $\mathbf{m}$, condição necessária para que o decodificador BCH opere bloco a bloco sobre as posições correspondentes. Sequências com $\hat{L} \neq L$ são descartadas. A segunda condição limita a densidade de bases indefinidas. Seja n_N o número de posições i para as quais $\hat{s}_i \notin \{A, C, G, T\}$; exige-se que

$$\frac{n_N}{L} < \epsilon, \quad (7)$$

onde ϵ é um limiar de qualidade. Sequências que não satisfazem (7) são descartadas por apresentarem densidade excessiva de ruído de *erasure*. As posições indefinidas remanescentes, em número inferior a ϵL , são substituídas pela base correspondente na referência, $\hat{s}_i \leftarrow s_i$, evitando que ambiguidades IUPAC introduzam erros artificiais adicionais no vetor $\hat{\mathbf{m}}$.

As sequências que satisfazem ambas as condições formam o conjunto de entrada do filtro BCH. Estas sequências ainda podem conter erros de leitura η_r , apagamentos residuais η_a , indels locais e mutações biológicas e, conforme a Eq. (6). É o filtro BCH, descrito na subseção seguinte, que separa as sequências cujo padrão de erros é compatível com o modelo BSC daquelas dominadas por rajadas e artefatos. Quando $\epsilon \approx P_e \ll p$, os ruídos instrumentais tornam-se desprezíveis frente à taxa de mutação biológica, e o vetor $\hat{\mathbf{m}}$ aproxima-se do modelo BSC ideal com $\hat{\mathbf{m}} \approx \mathbf{m} \oplus \mathbf{e}$.

B. Codificação e Filtragem BCH

A sequência de referência é dividida em blocos de k bits e codificada com um código BCH(n, k, t), onde t é a capacidade de correção e $g(x)$ é o polinômio gerador:

$$c(x) = m(x) \cdot g(x). \quad (8)$$

Para cada sequência candidata, o decodificador BCH (algoritmo PGZ [11]) é aplicado bloco a bloco comparando os dados da candidata com a paridade da referência e retorna o número de erros detectados (`cnumerr`). Uma sequência é aceita como *sequência natural* quando $0 \leq \text{cnumerr} \leq B \cdot t$, indicando padrão de erros esparsos por bloco consistente com mutações pontuais, onde B é o número de blocos. Quando `cnumerr` = -1, o decodificador falha em algum bloco, indicando densidade de erros superior a t por bloco, típica de artefatos de sequenciamento, e a sequência é descartada. Este critério constitui um filtro algébrico de qualidade genômica independente de metadados do sequenciador (e.g., *score Phred*).

C. Detecção de SNPs por Comparação Direta

Para as sequências aceitas pelo filtro BCH, a detecção de SNPs é realizada por comparação ponto a ponto entre $\hat{\mathbf{m}}$ e $\mathbf{m}$, com complexidade $\mathcal{O}(L)$. Esta simplicidade é consequência direta da filtragem: ao garantir que o padrão de erros de $\hat{\mathbf{m}}$ é esparsa e compatível com o BSC, o filtro BCH confirma que o alinhamento heurístico do MAFFT produziu correspondência posicional correta para aquela sequência, tornando a comparação direta confiável e não ambígua sem necessidade de um segundo alinhamento por programação dinâmica. Formalmente, as posições dos SNPs são obtidas diretamente do vetor de erro $\mathbf{e} = \hat{\mathbf{m}} \oplus \mathbf{m}$, conforme a Eq. (3), identificando os índices i para os quais $s_i \neq \hat{s}_i$. Cada diferença nucleotídica é então convertida para a posição do aminoácido correspondente, obtida como $\lceil i/3 \rceil$, permitindo a caracterização direta das mutações no espaço proteico.

D. Estimativa do Parâmetro BSC

O filtro BCH, além de selecionar as sequências naturais, fornece para cada uma delas o valor de `cnumerr`, que corresponde ao número total de erros de bit detectados em todos os B blocos. Sob o modelo BSC de parâmetro p , o número esperado de erros por sequência é $\mu = B \cdot k \cdot p$. O parâmetro p pode, portanto, ser estimado a partir das sequências aceitas pelo filtro como

$$\hat{p} = \frac{\overline{\text{cnumerr}}}{B \cdot k}, \quad (9)$$

onde $\overline{\text{cnumerr}}$ é a média de `cnumerr` sobre o conjunto de sequências naturais. Esta estimativa fecha o ciclo do modelo proposto: o mesmo filtro que seleciona as sequências fornece os dados necessários para caracterizar o canal genômico. A capacidade do canal BSC estimado é então calculada como

$$C = 1 + \hat{p} \log_2 \hat{p} + (1 - \hat{p}) \log_2 (1 - \hat{p}), \quad (10)$$

fornecendo uma medida da capacidade de conservação da informação genômica da variante analisada.

IV. METODOLOGIA

Para a validação do método proposto, foram utilizadas sequências da proteína Spike do SARS-CoV-2 obtidas do banco de dados GISAID [3], que reúne sequências genômicas virais compartilhadas por laboratórios de todo o mundo. O arquivo de sequências nucleotídicas da Spike (`spikenuc.fasta`), contendo 17.352.969 entradas, foi utilizado como fonte de dados. Após o alinhamento, as regiões codificantes da Spike (posições 21.563–25.384 da referência) são extraídas por coordenadas fixas, e os *gaps* de alinhamento são removidos, resultando em sequências de comprimentos variáveis. Este arquivo foi combinado com o arquivo de metadados correspondente via identificador de isolado, resultando em um dataset inicial de 17.314.847 sequências com informações completas de anotação.

É importante observar que, embora as sequências do `spikenuc.fasta` já tenham passado pelo alinhamento heurístico do MAFFT, o arquivo final não constitui um alinhamento múltiplo, pois os *gaps* são removidos após a extração,

e as sequências resultantes possuem comprimentos diferentes. A condição $\hat{L} = L$ imposta no pré-processamento (Seção III-A) seleciona, portanto, as sequências cuja região Spike, após alinhamento pelo MAFFT e extração por coordenadas, apresenta o mesmo comprimento da referência, indicando ausência de inserções ou deleções líquidas nesta região. Foram mantidas apenas sequências de genomas completos (>29.000 bp), alta cobertura ($<1\%$ de bases indeterminadas) e sem indels não verificados [3]. Aplicando o filtro para variantes VOC (*Variants of Concern*), o dataset foi reduzido para 4.831.991 sequências, das quais foram selecionadas Alpha (B.1.1.7, 958.810 seq.) e Delta (B.1.617.2, 3.250.710 seq.) por serem as de maior representatividade.

O pré-processamento descrito na Seção III foi aplicado com $L = 3822$ bases. A condição $\hat{L} = L$ reduziu o conjunto para 14.338 sequências Alpha e 85.171 sequências Delta, totalizando 99.509 sequências. Adotou-se $\epsilon = 1$, ou seja, nenhuma sequência foi descartada por *N-content*; todas as bases indefinidas foram substituídas pela base correspondente na referência Wuhan-Hu-1 (NC_045512.2), delegando ao filtro BCH a rejeição de sequências cuja densidade de erros exceda a capacidade de correção t por bloco. A análise *a posteriori* justifica essa escolha: a mediana do *N-content* é nula e 75% das sequências apresentam valores inferiores a $2,355 \times 10^{-3}$, confirmando que a grande maioria possui predominantemente bases definidas. Sobre este conjunto pré-processado foram aplicados os dois códigos BCH descritos na Seção III-B: BCH(63, 51, 2) com $B = 150$ blocos e BCH(255, 223, 4) com $B = 35$ blocos. O parâmetro BSC foi estimado pela Eq. (9) e a validação dos SNPs detectados foi realizada por comparação com as mutações anotadas no espaço de aminoácidos disponibilizadas pelo GISAID para cada sequência.

V. RESULTADOS

A Tabela I resume os resultados da filtragem descrita na Seção III-B para os dois códigos testados sobre o conjunto de 99.509 sequências pré-processadas conforme a Seção III-A. O BCH(63, 51, 2) aceitou 95,8% das sequências Alpha e 95,2% das Delta, enquanto o BCH(255, 223, 4) aceitou frações ligeiramente maiores (99,3% e 98,6%, respectivamente), uma vez que blocos maiores distribuem melhor o efeito de erros esparsos, reduzindo a probabilidade de que um único bloco exceda a capacidade de correção t .

TABELA I
RESULTADOS DE FILTRAGEM BCH SOBRE A PROTEÍNA SPIKE.

Variante	Total	Naturais	%Nat.	<code>cnumerr</code>
BCH(63, 51, 2)				
Alpha (B.1.1.7)	14.338	13.733	95,8%	9,88
Delta (B.1.617.2)	85.171	81.080	95,2%	9,58
BCH(255, 223, 4)				
Alpha (B.1.1.7)	14.338	14.242	99,3%	10,20
Delta (B.1.617.2)	85.171	84.011	98,6%	9,68

As sequências aceitas pelo filtro foram submetidas à detecção de SNPs por comparação direta conforme a Seção III-C, e os resultados foram validados contra as anotações de mutação (coluna *AASubstitution*) do GISAID. A Tabela II apresenta as métricas obtidas. O recall de 1,000 para a Alpha indica que o

método recuperou todos os SNPs canônicos desta variante, enquanto o recall de 0,975 para a Delta reflete a presença de sub-variantes AY.* cujos SNPs adicionais não constam do conjunto canônico. Os falsos positivos são biologicamente explicáveis: mutações sinônimas como Spike_Y144Y e Spike_I68I alteram o nucleotídeo sem modificar o aminoácido e, por convenção de anotação, são omitidas das anotações de mutação na coluna *AASubstitution*, embora corretamente detectadas pela comparação binária da Eq. (3). Os SNPs mais frequentes confirmam os marcadores canônicos de cada variante [9]: N501Y, D614G e P681H para a Alpha; L452R, T478K e P681R para a Delta. A estabilidade das métricas entre os dois códigos, com variações inferiores a 0,001 em precision e recall, indica que o desempenho da detecção é robusto em relação à escolha dos parâmetros (n, k, t) .

TABELA II

VALIDAÇÃO DOS SNPs DETECTADOS VS. ANOTAÇÕES DO GISAID.

Variante	Precision	Recall	F1
BCH(63, 51, 2)			
Alpha (B.1.1.7)	0,818	1,000	0,899
Delta (B.1.617.2)	0,805	0,975	0,879
BCH(255, 223, 4)			
Alpha (B.1.1.7)	0,817	1,000	0,899
Delta (B.1.617.2)	0,806	0,976	0,880

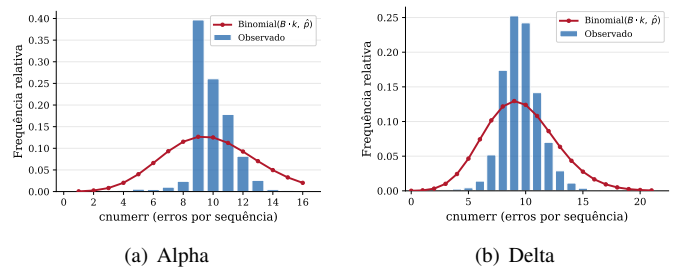
A partir do valor de $\overline{\text{cnnumerr}}$ reportado na Tabela I, o parâmetro do canal BSC foi estimado pela Eq. (9) para cada variante e código, conforme apresentado na Tabela III. Os valores de $\hat{p}$ são consistentes entre os dois códigos, com diferenças inferiores a 2%, e distinguem claramente Alpha de Delta: $\hat{p}_{\text{Alpha}} \approx 1,30 \times 10^{-3}$ contra $\hat{p}_{\text{Delta}} \approx 1,25 \times 10^{-3}$, sugerindo que cada variante constitui uma realização diferente do BSC modelado na Seção II-A. A capacidade $C \approx 0,986$ bits/uso de canal, calculada pela Eq. (10), indica que o canal genômico opera próximo do limite teórico de conservação da informação.

TABELA III

PARÂMETROS BSC ESTIMADOS POR VARIANTE (IC 95%).

Variante	$\hat{p}$	IC 95%	C (bits/bit)
BCH(63, 51, 2)			
Alpha	$1,292 \times 10^{-3}$	$[1,289; 1,295] \times 10^{-3}$	0,9857
Delta	$1,252 \times 10^{-3}$	$[1,250; 1,253] \times 10^{-3}$	0,9861
BCH(255, 223, 4)			
Alpha	$1,306 \times 10^{-3}$	$[1,302; 1,310] \times 10^{-3}$	0,9856
Delta	$1,240 \times 10^{-3}$	$[1,239; 1,242] \times 10^{-3}$	0,9862

A Fig. 1 ilustra a distribuição observada do cnnumerr para Alpha e Delta com BCH(63, 51, 2). Em ambos os casos, a curva teórica Binomial($B \cdot k, \hat{p}$) captura corretamente a média observada (Alpha: $\mu_{\text{BSC}} \approx 9,86$ vs. observado 9,88; Delta: $\mu_{\text{BSC}} \approx 9,56$ vs. observado 9,58), mas subestima a concentração da distribuição. Esta discrepância é esperada, pois o filtro BCH condiciona a amostra às sequências com no máximo $t = 2$ erros por bloco, produzindo uma distribuição truncada com variância menor que a binomial irrestrita. A concordância entre a média predita e a observada reforça a adequação do modelo BSC.


 Fig. 1. Distribuição observada do cnnumerr (barras azuis) vs. Binomial teórica BSC (curva vermelha) com BCH(63, 51, 2).

VI. CONCLUSÕES

Foi proposta uma metodologia baseada em códigos BCH para validação pós-alinhamento heurístico e detecção de SNPs na proteína Spike do SARS-CoV-2. O filtro BCH atua como segunda camada de controle de qualidade sobre sequências previamente alinhadas pelo MAFFT, rejeitando aquelas cujo padrão de erros é incompatível com o modelo BSC e validando as demais para detecção direta de SNPs com complexidade linear. A validação sobre 99.509 sequências das variantes Alpha e Delta demonstrou recall médio de 0,988 e confirmou os SNPs canônicos de ambas as variantes. O parâmetro de canal $\hat{p} \approx 1,3 \times 10^{-3}$, estável entre códigos e distinto entre variantes, sugere que $\hat{p}$ pode funcionar como assinatura de canal para caracterização de variantes virais. Como trabalho futuro, pretende-se investigar a aplicação do filtro BCH diretamente a sequências brutas não alinhadas, bem como a extensão do método a outros vírus RNA e a exploração de códigos sobre $\mathbb{F}_4$.

REFERÊNCIAS

- [1] D. W. Mount and D. W. Mount, *Bioinformatics: sequence and genome analysis*. Cold spring harbor laboratory press Cold Spring Harbor, NY, 2001, vol. 564.
- [2] K. Katoh and D. M. Standley, “MAFFT multiple sequence alignment software version 7: improvements in performance and usability,” *Molecular biology and evolution*, vol. 30, no. 4, pp. 772–780, 2013.
- [3] S. Elbe and G. Buckland-Merrett, “Data, disease and diplomacy: GISAID’s innovative contribution to global health,” *Global Challenges*, vol. 1, no. 1, pp. 33–46, 2017.
- [4] K. Reinert, B. Langmead, D. Weese, and D. J. Evers, “Alignment of next-generation sequencing reads,” *Annual Review of Genomics and Human Genetics*, vol. 16, no. Volume 16, 2015, pp. 133–151, 2015.
- [5] M. M. Brandão, L. Spoladore, L. C. Faria, A. S. Rocha, M. C. Silva-Filho, and R. Palazzo, “Ancient DNA sequence revealed by error-correcting codes,” *Scientific reports*, vol. 5, no. 1, p. 12051, 2015.
- [6] M. M. Arruda, F. M. De Assis, and T. A. De Souza, “Is BCH code useful to DNA classification as an alignment-free method?” *IEEE Access*, vol. 9, pp. 68 552–68 560, 2021.
- [7] W. Chen, P. Wang, L. Wang, D. Zhang, M. Han, M. Han, and L. Song, “Low-complexity and highly robust barcodes for error-rich single molecular sequencing,” *3 Biotech*, vol. 11, no. 2, p. 78, 2021.
- [8] R. E. Blahut, *Algebraic codes for data transmission*. Cambridge university press, 2003.
- [9] E. C. Holmes, *The evolution and emergence of RNA viruses*. Oxford University Press, 2009.
- [10] A. S. Luring and R. Andino, “Quasispecies theory and the behavior of RNA viruses,” *PLOS Pathogens*, vol. 6, no. 7, pp. 1–8, 07 2010.
- [11] S. Lin and D. Costello, *Error Control Coding: Fundamentals and Applications*, ser. Computer applications in electrical engineering series. Pearson-Prentice Hall, 2004.