Representação Binária de Mutações em Peptídeos Trípticos para Análise Proteogenômica e Identificação de Haplótipos no Câncer

Rafaela Marie M. da Cunha[1], Giullia de Souza Santos[1], Lucas Marques da Cunha[1]

[1]*Universidade Federal de Rondônia, Porto Velho, Brasil (mellorafa442@gmail.com)*

Resumo: A variabilidade genética e a identificação de haplótipos desempenham um papel fundamental na predisposição ao câncer e na caracterização de biomarcadores tumorais. Este estudo apresenta um método computacional inovador para modelar combinações de mutações em peptídeos trípticos, permitindo uma análise detalhada das variações proteicas associadas a polimorfismos genéticos. A abordagem utiliza uma representação binária das mutações, possibilitando a exploração sistemática de todas as combinações possíveis. A metodologia foi aplicada a dados de espectrometria de massas do The Cancer Genome Atlas (TCGA), com foco em genes do complexo HLA, que possuem relevância na resposta imunológica. A ferramenta LDLink (SNPclip) foi empregada para identificar haplótipos e eliminar variantes redundantes com base no desequilíbrio de ligação (LD). Os resultados evidenciaram que peptídeos com múltiplas variações genéticas apresentam forte correlação com padrões haplotípicos específicos, auxiliando na identificação de perfis genéticos associados ao câncer. Além disso, o estudo reforça a importância de considerar padrões haplotípicos na análise proteogenômica, contribuindo para uma melhor compreensão das relações entre variação genética e progressão tumoral. A abordagem proposta representa um avanço na identificação de biomarcadores tumorais e pode apoiar o desenvolvimento de novas estratégias para a medicina de precisão.

Palavras-chave: Variabilidade genética; Haplótipos; Mutações; Câncer; Bioinformática.

INTRODUÇÃO

A identificação precisa de variantes proteicas é crucial para a compreensão dos processos biológicos subjacentes a doenças como o câncer, o desenvolvimento de novas terapias e a validação da variação genética (Da Cunha et al., 2022). No entanto, os bancos de dados de proteínas tradicionais apresentam limitações significativas. Frequentemente, eles não incluem proteínas variantes em sua totalidade ou, quando o fazem, desconsideram a complexidade das combinações de mutações que podem coexistir no mesmo peptídeo. Essa limitação compromete a identificação precisa de variantes coexistentes e restringe o escopo da análise proteômica, uma área vital para a integração de dados genômicos, transcriptômicos e proteômicos.

Métodos existentes, como os propostos por Choi e Paek (2020) e Choong et al. (2020), buscam abordar o problema da identificação de variantes, mas enfrentam desafios relacionados à complexidade computacional ou à potencial perda de combinações raras, porém biologicamente relevantes. Além disso, as abordagens tradicionais de conversão de variantes de nucleotídeo único (SNVs) em variantes de aminoácidos únicos (SAVs) muitas vezes simplificam demais o processo, ignorando a formação de combinações intermediárias que podem ocorrer naturalmente (Makarov et al., 2012; Han et al., 2021). Este cenário é particularmente crítico em tecidos tumorais, onde variantes somáticas e germinativas podem coexistir de forma heterogênea, dando origem a haplótipos complexos e interações biológicas que os modelos convencionais não conseguem capturar (Da Cunha et al., 2022).

Diante deste desafio, este estudo propõe uma nova abordagem computacional baseada em representação binária para modelar e combinar mutações em peptídeos trípticos. O principal objetivo deste trabalho é garantir a geração exaustiva de todas as combinações possíveis de mutações em peptídeos sem redundância desnecessária. Essa estratégia permite a identificação precisa de haplótipos complexos, assegurando que nenhuma variante seja perdida durante a análise proteômica e, consequentemente, impulsionando a identificação de novos biomarcadores tumorais e o desenvolvimento de estratégias de medicina de precisão.

MATERIAL E MÉTODOS

A pesquisa desenvolveu uma abordagem quantitativa, baseada na análise computacional, para investigar de

forma sistemática as permutações de alterações em peptídeos. Para garantir a cobertura exaustiva de todas as propostas mutacionais possíveis, foram aplicadas técnicas avançadas de modelagem computacional, associadas à representação binária. Essa estratégia permitiu um mapeamento preciso das variações, minimizando redundâncias e otimizando a eficiência no processamento dos dados. A utilização da representação binária facilitou ainda a manipulação de grandes volumes de dados, determinando o custo computacional e promovendo a identificação de padrões relevantes na análise das mutações.

Além disso, foram utilizados dados de espectrometria de massa do estudo de câncer retal do The Cancer Genome Atlas (TCGA). Estes dados desempenharam um papel crucial na identificação de mutações específicas em peptídeos, permitindo a visualização das alterações observadas com variações.

Coleta de Dados

Para a análise das mutações, foram utilizados os seguintes bancos de dados:

- **dbSNP**: Base de dados online que contém informações sobre variações genéticas em humanos e outros organismos, mantida pelo National Center for Biotechnology Information (NCBI). Disponível em: `https://www.ncbi.nlm.nih.gov/snp/`
- **RefSeq**: Fornece sequências referenciais de proteínas, DNA e RNA de diversos organismos. Disponível em: `https://www.ncbi.nlm.nih.gov/refseq/`
- **TCGA**: Um programa de referência em genômica do câncer. Foram utilizados os dados do adenocarcinoma do reto (PDC000111). Disponível em: `https://pdc.cancer.gov/pdc/study/PDC000111`

Filtragem de Dados

Os dados coletados passaram por um rigoroso processo de filtragem para garantir a qualidade e relevância das mutações selecionadas. Foram considerados os seguintes critérios:

- Remoção de variantes de baixa cobertura ou pouco estudadas (Frequência Alélica ≤0.05);
- Seleção de mutações *missense*, relevantes para o estudo de combinações binárias;
- Remoção de peptídeos identificados apenas com uma variante;
- Remoção de peptídeos com múltiplos SNPs de baixa correlação, utilizando o critério de desequilíbrio de ligação (LD) conforme descrito na subseção de Determinação de Haplótipos.

O banco de mutações personalizadas gerado por esta abordagem foi utilizado para realizar a busca contra os espectros experimentais do TCGA, empregando a ferramenta Pattern Lab for Proteomics V (Carvalho et al., 2016).

Método de Combinação Binária

Essa seção apresenta o método computacional baseado em combinações binárias para a análise de mutações em sequências biológicas. Essa abordagem permite explorar sistematicamente todas as possíveis variações de mutações, oferecendo um meio eficiente para a investigação computacional de peptídeos e outras macromoléculas biológicas.

A formulação matemática deste método pode ser expressa por:

$$C = 2^n - 1 \qquad (1)$$

onde C é o número total de combinações possíveis de mutações e n representa o número total de mutações possíveis identificadas para um dado peptídeo. Essa equação determina a quantidade de combinações distintas geradas a partir de um conjunto de mutações, desconsiderando a combinação vazia (peptídeo *wild-type*).

Cada combinação pode ser representada por um vetor binário V, conforme fórmula (2):

$$V = (b_1, b_2, b_3, \ldots, b_n) \; onde \; b_i \in \{0, 1\} \qquad (2)$$

Neste vetor, $b_i = 1$ indica que a mutação M_i foi aplicada ao peptídeo, e $b_i = 0$ indica que a mutação M_i não foi aplicada.

O peptídeo mutado P′ gerado por uma configuração específica de V pode ser expresso como:

$$P' = f(P, V) = P / \{M_i \mid b_i = 1\} \qquad (3)$$

onde P é o peptídeo de referência (*wild-type*) e V é o vetor binário que define quais mutações são aplicadas. A função *f(P,V)* aplica sequencialmente as mutações indicadas pelos bits '1' no vetor *V* à sequência do peptídeo *P*.

A metodologia adotada para a análise das variantes de aminoácidos utiliza essa representação binária para combinar diferentes mutações em peptídeos, facilitando a análise de todas as possíveis variantes de uma sequência peptídica derivada de digestão enzimática.

Para ilustrar a aplicação dessa metodologia, consideramos um peptídeo derivado da digestão com tripsina, que cliva após resíduos de lisina (K) e arginina (R), desde que não sejam seguidos por prolina (P). O peptídeo de referência (*wild-type*) gerado por essa digestão é: K.VLSPADKTNVK.A.

Mutações Consideradas

A partir de bases de dados genômicos como dbSNP e RefSeq, foram identificadas três variantes possíveis de aminoácidos (*Single Amino Acid Variants - SAVs*) para este peptídeo ilustrativo: V2I (Valina → Isoleucina) - rs123456; D6E (Aspartato → Glutamato) - rs234567; e K11R (Lisina → Arginina) - rs345678.

Codificação Binária das Mutações

Cada mutação é representada por um bit em um vetor binário de três posições. A Tabela 1 ilustra o vetor binário para cada combinação:

Tabela 1. Codificação Binária para as Mutações.

Vetor Binário	Mutações Aplicadas
000	Nenhuma Mutação (*WT*)
001	K11R
010	D6E
011	D6E, K11R
100	V2I
101	V2I, K11R
110	V2I, D6E
111	V2I, D6E, K11R

A equação para determinar o número total de combinações possíveis é dada pela equação (1). Neste exemplo, o resultado é $C = 2^3 – 1 = 7$.

Peptídeos Variantes Gerados

Cada combinação de mutações resulta em uma nova sequência de peptídeo, conforme a Tabela 2:

Tabela 1. Peptídeos Variantes Gerados a partir da Codificação Binária.

Vetor Binário	Mutações Aplicadas
000	K.VLSPADKTNVK.A (WT)
001	K.VLSPADKTNVR.A (K11R)
010	K.VLSPAEKTNVK.A (D6E)
011	K.VLSPAEKTNVR.A (D6E, K11R)
100	K.ILSPADKTNVK.A (V2I)
101	K.ILSPADKTNVR.A (V2I, K11R)
110	K.ILSPAEKTNVK.A (V2I, D6E)
111	K.ILSPAEKTNVR.A (V2I, D6E, K11R)

Determinação de Haplótipos

Para determinar se as variantes identificadas são parte de um mesmo haplótipo, utilizamos a ferramenta LDLink (Machiela e Chanock, 2015), especificamente o módulo SNPclip, que permite a poda de uma lista de variantes com base no desequilíbrio de ligação (LD). O objetivo foi reduzir a redundância de variantes fortemente correlacionadas e identificar aquelas que representam haplótipos distintos.

Utilizamos como referência o banco de dados do 1000 Genomes Project, selecionando o genoma GRCh37. A lista de variantes foi fornecida no formato de identificadores RS (*Reference SNP IDs*) ou coordenadas genômicas. As configurações dos parâmetros no LDLink (SNPclip) foram:

- **População**: Analisamos todas as populações disponíveis no banco de dados, sem restringir a um grupo específico, para uma abrangência maior.
- **Frequência Alélica Mínima (MAF)**: Definimos um limiar de 0.01 (1%), garantindo que apenas variantes com frequência razoável fossem consideradas.
- **Limiar de Desequilíbrio de Ligação (LD)**: Estabelecemos um valor de $R^2 \geq 0.8$ para considerar variantes em LD significativo.

O SNPclip processou a lista de variantes e removeu SNPs redundantes que apresentavam forte correlação (LD elevado). Variantes independentes ou menos correlacionadas foram mantidas, representando haplótipos distintos na região genômica analisada. Este método permitiu identificar variantes únicas que podem ser usadas para análises subsequentes, evitando vieses causados por redundâncias genéticas e facilitando a interpretação dos resultados em estudos de associação genética e na modelagem de mutações combinatórias.

RESULTADOS E DISCUSSÃO

Os Polimorfismos de Nucleotídeo Único (SNPs) desempenham um papel crucial na variação genética entre indivíduos, e sua associação com diferentes doenças tem sido amplamente investigada. Estudos anteriores identificaram genes do complexo HLA e SNPs específicos ligados a diversas condições, como a pancreatite autoimune tipo 1 e o carcinoma nasofaríngeo (Fujibayashi et al., 2016; Tian et al., 2015; Zeng et al., 2025). Além disso, genes como HLA-A, HLA-C, HLA-B, HP, ASAHI e TNN foram implicados na predisposição e progressão do câncer colorretal e do câncer retal (Michelakos et al., 2022; Kovcić et al., 1994; Vijayan et al., 2024; Liu et al., 2025). Esses achados reforçam a importância da caracterização detalhada de variantes genéticas na

compreensão da suscetibilidade a doenças e na identificação de alvos terapêuticos.

No presente estudo, investigamos a correlação entre diferentes SNPs por meio da análise de desequilíbrio de ligação (LD), utilizando a ferramenta LDLink. A Tabela 3 apresenta uma amostra representativa de SNPs, peptídeos de referência, peptídeos mutados e genes associados encontrados em nossa análise de dados do TCGA, com destaque para aqueles SNPs que exibiram um valor de $R^2 \geq 0.8$ (em negrito), indicando um alto grau de correlação. Este resultado corrobora a literatura, uma vez que muitos SNPs do complexo HLA exibem fortes padrões de LD, o que pode influenciar a predisposição genética a diversas doenças e a formação de haplótipos funcionais. A identificação dessas variantes altamente correlacionadas é essencial para estratégias de predição de neoantígenos e imunoterapia personalizada, pois permite selecionar alvos mutacionais que melhor representam a diversidade genética de um paciente.

Nosso estudo propõe um novo método computacional baseado em representação binária para analisar combinações de mutações em peptídeos. Essa abordagem se destaca por considerar não apenas SNPs isolados, mas também suas combinações em regiões de alto LD, permitindo uma modelagem mais realista das variações proteicas. A capacidade de capturar essas interações complexas é crucial para a identificação de neoantígenos tumorais e o avanço da proteogenômica aplicada à imunoterapia.

Tabela 3. Relação de SNPs, Peptídeos de Referência, Peptídeos Mutados e Genes Associados.

SNP ID	Peptídeo Referência	Peptídeo Mutado	Gene
rs9260138, rs2231004, **rs1136690**, **rs9260139**	GYYNQSEAGSHTIQIMYGCDVGSDGR	GYYNQSEAGSHTIQIMYGCDVGSDGRVgLGsvR	HLA-A
rs1136690, rs1136688, **rs9260140**	GYYNQSEAGSHTIQIMYGCDVGSDGR	VhLGiLhGYYNQSEAGSHTIQIMYGCDVGSDGR	HLA-A
rs1136690, **rs9260140**, **rs1136689**	GYYNQSEDGSHTIQIMYGCDVGPDGR	nLlGYYNQSEDGSHTIQIMYGCDVGPDGR	HLA-A
rs2308525, **rs1050451**, rs2308527, rs1131151	ALLLLLSGGLALTETWACSHSMR	sLiLLLSGaLALTETWACSHSMk	HLA-C
rs2308525, **rs1050451**, rs2308527, rs41549413	ALLLLLSGGLALTETWACSHSMR	sLiLLLSGaLALiETWACSHSMR	HLA-C
rs1071645, **rs3753115**	STYPPSGPTVFPAVIRAPVPGLLGNFPGPFEEEMK	STYPPSGPTVFPAiIRAPmPGLLGNFPGPFEEEMK	ASAH1
rs2308525, **rs1050451**, rs1131151, rs2308527, rs41549413	ALLLLLSGGLALTETWACSHSMR	sLiLLLSGaLALiETWACSHSMk	HLA-C
rs2308525, **rs1050451**, rs2308527, rs41549413	ALLLLLSGGLALTETWACSHSMR	sLvLLLSGaLALsETWACSHSMR	HLA-C
rs1131165, **rs1131156**, rs1131159	TVLLLLSAALALTETWAGSHSMR	TVLLLLleAmALTETWAGSHSMR	HLA-B
rs9260139, **rs9260140**, **rs1136689**	GYYNQSEDGSHTIQIMYGCDVGPDGR	TqlGYYNQSEDGSHTIQIMYGCDVGPDGR	HLA-A
rs2308525, **rs1050451**, rs2074493, rs2308527, rs41549413	ALLLLLSGGLALTETWACSHSMR	sLvLLLSGaLALiETWAr	HLA-C
rs9260139, rs1136688, **rs9260140**	GYYNQSEDGSHTIQIMYGCDVGPDGR	AdLGTphGYYNQSEDGSHTIQIMYGCDVGPDGR	HLA-A
rs199926732, **rs200877317**	TEGDGVYTLNNEK	TEGDGVYTLNNk, TEGDGVYTLNdk	HP
rs1059506, rs1059509	DYIALNEDLR	DYIAvNEDLR, DYIALk	HLA-A
rs1136690, **rs9260139**,	GYYNQSEAGSHTIQIMYGCDVGSDGR	VhLGsphGYYNQSEAGSHTIQIMYGCDVGSDGR	HLA-A

rs1136688, **rs9260140** **rs1136690**, **rs9260139**, **rs9260140**, **rs1136689**	GYYNQSEDGSHTIQIMYGCDVGPDGR	nplGYYNQSEDGSHTIQIMYGCDVGPDGR	HLA-A
rs1131165, rs1050462, **rs1131156**, rs1131159	TVLLLLSAALALTETWAGSHSMR	TlLLLLweAmALTETWAGSHSMR	HLA-B
rs9260139, rs2231004, rs1136688, **rs1136690**, **rs9260139**	GYYNQSEAGSHTIQIMYGCDVGSDGR	VDLGTqRGYYNQSEAGSHTIQIMYGCDVGSDGR, VnLGTqR	HLA-A
rs16866380, **rs16866378**	SPEPSHPK, NVYSLEIR	SPEPSHlK, NaYSLEIR	TTN
rs1050451, **rs2308525**, rs2074493, rs2074493, rs2308527, rs41549413	ALLLLLSGGLALTETWACSHSMR	sLvLLLSGGLALiETWAgSHSMR, sLvLLLSGaLALiETWAgSHSMR	HLA-C
rs199926732, **rs200877317**	TEGDGVYTLNNEK	TEGDGVYTLNdk	HP
rs1131165, **rs1131156**, rs1131159	TVLLLLSAALALTETWAGSHSMR	TVLLLLweAvALTETWAGSHSMR	HLA-B
rs2308525, **rs1050451**, rs2074493, rs2308527, rs41549413	ALLLLLSGGLALTETWACSHSMR	sLiLLLSGaLALiETWAgSHSMR	HLA-C
rs1131165, **rs1131156**, rs1131159	TVLLLLSAALALTETWAGSHSMR	TVLLLLwgAvALTETWAGSHSMR	HLA-B
rs1071645, **rs3753115**	STYPPSGPTVFPAVIR, APVPGLLGNFPGPFEEEMK	STYPPSGPTVFPAiIR, APmPGLLGNFPGPFEEEMK	ASAH1
rs1136690, rs2231004, rs1136688, **rs9260139**	GYYNQSEAGSHTIQIMYGCDVGSDGR	GYYNQSEAGSHTIQIMYGCDVGSDGRVhLGiLR	HLA-A
rs1136690, rs2231004, rs1136688, **rs9260139**	GYYNQSEAGSHTIQIMYGCDVGSDGR	GYYNQSEdGSHTIQIMYGCDVGpDGRAdLGsvR	HLA-A

Comparação entre Métodos de Análise de Mutações Proteicas

A avaliação comparativa entre os métodos de Combinação Binária de Mutações em Peptídeos Trípticos (proposto), MinProtMaxVP (Choong et al., 2020) e MutCombinator (Choi e Paek, 2020) revelou diferenças significativas em seus objetivos, eficiência computacional e aplicabilidade biológica.

1. **Método de Combinação Binária (Proposto)**: Caracteriza-se pela geração **exaustiva** de todas as combinações possíveis de mutações dentro de peptídeos trípticos. Essa estratégia garante a consideração de todas as variações relevantes, o que é crucial para a identificação de neoantígenos tumorais e para o estudo da heterogeneidade tumoral. A restrição a peptídeos trípticos, que são as unidades detectadas pela espectrometria de massas, torna o método altamente direcionado e compatível com as tecnologias proteômicas.
2. **MinProtMaxVP**: Busca um equilíbrio entre abrangência e eficiência computacional ao minimizar a quantidade de sequências necessárias para cobrir todas as variantes proteicas possíveis. Embora reduza redundâncias, essa abordagem pode, por sua própria natureza de minimização, excluir combinações raras de mutações que podem ser biologicamente relevantes e não foca exclusivamente em peptídeos trípticos, o que pode incluir sequências não prioritárias para espectrometria de massas (Choong et al., 2020).
3. **MutCombinator**: Adota uma abordagem baseada em grafos para combinar mutações e analisar interações entre variantes,

permitindo o estudo de eventos mutacionais complexos. No entanto, sua aplicação demanda maior capacidade computacional e pode não ser ideal para espectrometria de massas, pois não se concentra intrinsecamente na detecção de peptídeos mutados (Choi e Paek, 2020).

Haplótipos e Captura de Variantes Combinatórias

Os haplótipos são conjuntos de variantes genéticas herdadas juntas (Machiela e Chanock, 2015). Desse modo, considerá-los na análise é um fator essencial na análise de mutações proteicas, pois certas mutações ocorrem agrupadas em padrões específicos, e esse aspecto é crucial para modelar a heterogeneidade tumoral. Alguns haplótipos podem conferir vantagens adaptativas às células cancerígenas, favorecendo sua expansão e resistência a tratamentos.

Dentre os métodos analisados, o método proposto já é intrinsecamente compatível com a análise de haplótipos, pois considera todas as combinações possíveis dentro dos peptídeos trípticos. Assim, desde que os SNPs de um haplótipo estejam na análise inicial, suas combinações serão automaticamente incluídas. Já os métodos MinProtMaxVP e MutCombinator não priorizam explicitamente haplótipos em suas descrições. O MinProtMaxVP pode eliminar algumas combinações relevantes ao minimizar redundâncias, enquanto o MutCombinator, por sua abordagem em grafos, poderia ser ajustado para incluir haplótipos, mas essa funcionalidade não é detalhada como um foco primário em sua pesquisa atual.

Análise de Compromissos (Trade-offs) e Limitações

A abordagem de combinação binária proposta, embora poderosa em sua exaustividade, apresenta inerentemente um compromisso entre abrangência e custo computacional. À medida que o número de mutações em um peptídeo aumenta linearmente, o número de combinações possíveis cresce exponencialmente (2^n). Para peptídeos com um grande número de SNVs, a geração e processamento de todas as combinações podem se tornar computacionalmente intensivos.

No entanto, este método é especificamente direcionado a peptídeos trípticos, que são fragmentos menores (tipicamente 6-30 aminoácidos) gerados por digestão enzimática e detectados por espectrometria de massas. O número de SNVs que podem ocorrer dentro de um único peptídeo tríptico é, na prática, limitado, o que mitiga em parte o problema do crescimento exponencial. Além disso, a filtragem inicial de dados com base na frequência alélica e no desequilíbrio de ligação (LD) ajuda a reduzir o espaço de busca, focando em variantes mais relevantes e correlacionadas.

Uma limitação da fase atual do trabalho é que, embora a metodologia tenha sido demonstrada com dados do TCGA e a análise de LD tenha sido realizada, os resultados apresentados focam na proposição e na viabilidade teórica da abordagem. É necessário um estudo experimental mais abrangente e comparativo para validar a eficácia do método na detecção de peptídeos mutados em amostras reais de espectrometria de massas e para quantificar seu desempenho em relação a abordagens existentes em termos de sensibilidade, especificidade e tempo de processamento em larga escala.

CONCLUSÃO

Neste estudo, propusemos e descrevemos em detalhes uma nova abordagem computacional baseada em combinação binária de mutações em peptídeos trípticos. O objetivo principal foi permitir uma modelagem mais detalhada e exaustiva das variações proteicas, crucial para a identificação de haplótipos e para o estudo da heterogeneidade tumoral, especialmente relevante para a identificação de biomarcadores e neoantígenos em câncer.

Demonstramos a capacidade da abordagem de gerar sistematicamente todas as combinações possíveis de mutações, um diferencial importante em relação a métodos existentes como MinProtMaxVP e MutCombinator, que podem não garantir essa exaustividade ou são mais intensivos computacionalmente para a finalidade específica de análise de peptídeos por espectrometria de massas. A análise de desequilíbrio de ligação (LD) revelou que certos SNPs apresentam $R^2 \geq 0.8$, indicando alta correlação entre variantes que podem influenciar a predisposição genética a doenças, reforçando a importância de considerar variantes combinatórias.

Os resultados obtidos indicam o potencial teórico de incluir haplótipos e padrões de LD na análise de mutações para aprimorar significativamente a predição de neoantígenos e contribuir para o avanço da imunoterapia personalizada. A aplicação da abordagem proposta na proteogenômica e espectrometria de massas promete uma detecção de mutações relevantes de forma mais abrangente e precisa.

Para trabalhos futuros, diversas direções podem ser exploradas:

- **Validação Experimental Abrangente**: Realizar experimentos com conjuntos de dados de espectrometria de massas de maior escala, incluindo dados de pacientes, para validar empiricamente a capacidade do método de identificar peptídeos mutados em contextos biológicos complexos.
- **Comparação Quantitativa**: Conduzir uma análise comparativa rigorosa do desempenho do método proposto contra abordagens de última geração, avaliando métricas como sensibilidade, especificidade, precisão e recall.
- **Otimização Computacional**: Desenvolver estratégias de otimização algorítmica para reduzir o custo computacional do modelo, especialmente para cenários com um número muito elevado de mutações potenciais por peptídeo, sem comprometer sua abrangência.
- **Integração com Ferramentas Existentes**: Explorar a integração do método com pipelines proteogenômicos existentes para facilitar sua adoção e aplicação por outros pesquisadores.
- **Aplicação Clínica**: Integrar a abordagem com dados clínicos para validar seu potencial na identificação de biomarcadores tumorais prognósticos ou preditivos de resposta a terapias, e no desenvolvimento de novas estratégias terapêuticas direcionadas.

AGRADECIMENTOS

Os autores agradecem o apoio da Coordenação de Inovação e Transferência de Tecnologia e do Programa Institucional de Bolsas de Iniciação Científica (PIBIC/CNPq). Suas contribuições foram fundamentais para a conclusão deste trabalho. Adicionalmente, ferramentas baseadas em inteligência artificial, como o modelo de linguagem Gemini, foram utilizadas como apoio na revisão gramatical e ortográfica deste manuscrito. Ressalta-se que o uso dessas ferramentas teve caráter exclusivamente auxiliar, sem impacto sobre o conteúdo intelectual do trabalho.

REFERÊNCIAS

Carvalho, P., Lima, D., Leprevost, F. et al. Integrated analysis of shotgun proteomic data with PatternLab for proteomics 4.0. Nat Protoc 11, 102–117 (2016).

Choi, S. and Paek, E. (2020). Mutcombinator: identification of mutated peptides allowing combinatorial mutations using nucleotide-based graph search. Bioinformatics, 36(Supplement_1):i203-i209.

Choong, W.-K., Wang, J.-H., and Sung, T.-Y. (2020). Minprotmaxvp: Generating a minimized number of protein variant sequences containing all possible variant peptides for proteogenomic analysis. Journal of Proteomics, 223:103819.

Da Cunha, L. M., Terrematte, P., Fiuza, T. D. S., Da Silva, V. L., Kroll, J. E., De Souza, S. J., and De Souza, G. A. (2022). dbpepvar: a novel cancer proteogenomics database. IEEE Access, 10:90982-90994.

Fujibayashi, S., Sasajima, J., Goto, T., Tanaka, H., Kawabata, H., Fujii, T., Nakamura, K., Chiba, A., Yanagawa, N., Moriichi, K., Fujiya, M., and Kohgo, Y. (2016). A high-throughput sequence analysis of japanese patients revealed 11 candidate genes associated with type 1 autoimmune pancreatitis susceptibility. Biochem Biophys Rep, 6:76-81.

Han, Q., Yang, Y., Wu, S., Liao, Y., Zhang, S., Liang, H., Cram, D. S., and Zhang, Y. (2021). Cruxome: a powerful tool for annotating, interpreting and reporting genetic variants. BMC genomics, 22(1):407.

Kovcić, V., Jelić, S., Filipović, I., and Tomasević, Z. (1994). Koncentracije haptoglobina i alfa-jedan-antitripsina kao markera bioloske evolucije kod bolesnika s karcinomom digestivnog trakta. Srp Arh Celok Lek, 122(11-12):311-313.

Liu, H., Liu, J., Guan, X., Zhao, Z., Cheng, P., Chen, H., Jiang, Z., and Wang, X. (2025). Titin gen mutations enhance radiotherapy efficacy via modulation of tumour immune microenvironment in rectum adenocarcinoma. Clinical and Translational Medicine, 15(1):e70123.

Machiela, M. J. and Chanock, S. J. (2015). Ldlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics, 31(21):3555-3557.

Makarov, V., O'Grady, T., Cai, G., Lihm, J., Buxbaum, J. D., and Yoon, S. (2012). Ann-tools: a comprehensive and versatile annotation toolkit for genomic variants. Bioinformatics, 28(5):724-725.

Michelakos, T., Kontos, F., Kurokawa, T., Cai, L., Sadagopan, A., Krijgsman, D., Weichert, W., Durrant, L. G., Kuppen, P. J., Ferrone, C. R., et al. (2022). Differential role of hla-a and hla-b, c expression levels as prognostic markers in colon and rectal cancer. Journal for immunotherapy of cancer, 10(3):e004115.

Tian, W., Zhu, F. M., Wang, W. Y., Cai, J. H., Zhang, W., Li, L. X., Liu, K. L., Jin, H. K., and Wang, F. (2015). Sequence-based typing of hla-a gene in 930 patients with nasopharyngeal carcinoma

in hunan province, southern china. Tissue Antigens, 86(1):15-20.

Vijayan, Y., James, S., Viswanathan, A., Aparna, J. S., Bindu, A., Namitha, N. N., Anantharaman, D., Babu Lankadasari, M., and Harikumar, K. B. (2024). Targeting acid ceramidase enhances antitumor immune response in colorectal cancer. Journal of Advanced Research, 65:73-87.

Zeng, Y., Luo, C. L., Lin, G. W., Li, F., Bai, X., Ko, J. M., Xiong, L., Liu, Y., He, S., Jiang, J. X., et al. (2025). Whole-exome sequencing association study reveals genetic effects on tumor microenvironment components in nasopharyngeal carcinoma. J Clin Invest, 135(1):e182768.