UNIVERSIDADE FEDERAL DO PARANÁ

JONATHAN ANDRÉ BACK

ANÁLISE E AGRUPAMENTO DE RECEPTORES DE CÉLULA T: USO DE MÉTRICAS DE DISTÂNCIA COM FOCO NAS REGIÕES CDR3

CURITIBA

Jonathan André Back

ANÁLISE E AGRUPAMENTO DE RECEPTORES DE CÉLULA T: USO DE MÉTRICAS DE DISTÂNCIA COM FOCO NAS REGIÕES CDR3

TCC apresentado ao curso de Graduação em Biomedicina, Setor de Ciências Biológicas, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Bacharel em Biomedicina.

Orientadora: Profa. Dra. Larissa Magalhães Alvarenga

Coorientador: Prof. Dr. Mauro Antônio Alves Castro

Coorientador: M.e. Jean Silva de Souza Resende

CURITIBA

RESUMO

Os receptores de células T (TCRs) desempenham um papel fundamental na resposta imunológica, reconhecendo antígenos apresentados por moléculas de MHC e contribuindo para a defesa imunológica adaptativa. Compreender a diversidade e especificidade antigênica dos TCRs é essencial para avanços na pesquisa e em terapias baseadas em células T. Embora a maioria dos estudos se concentre na área de aprendizado de máquina para predição da especificidade dos TCRs, essa abordagem, além de ser de difícil implementação, pode ser substituída por abordagens embasadas em análise de sequências. Este trabalho propôs uma metodologia baseada em alinhamento global, transformação de pontuações de alinhamento em métricas de distância e clusterização hierárquica para analisar sequências de TCRs do IEDB, com foco na região CDR3. Aqui mostramos que o uso da matriz de substituição BLOSUM62 para alinhamento global e a subsequente transformação de seu resultado para uma métrica de distância resultaram em um agrupamento eficaz de sequências de TCR, evidenciando a capacidade de separá-las por sua especificidade antigênica a um patógeno específico. Embora o estudo tenha limitações quanto ao número de sequências e patógenos analisados, os resultados indicam que a metodologia baseada em métricas de distância pode ser uma ferramenta promissora para a análise de TCRs. Em futuros estudos, a ampliação do número de amostras e a inclusão de uma maior diversidade de patógenos poderão aprofundar a compreensão sobre a organização dos TCRs em relação à sua especificidade antigênica.

Palavras-chave: Receptores de Células T; Especificidade Antigênica; Análise de sequências; Métricas de Distância; Alinhamento Global.

ABSTRACT

T-cell receptors (TCRs) play a crucial role in the immune response by recognizing antigens presented by MHC molecules and contributing to adaptive immune defense. Understanding the diversity and antigenic specificity of TCRs is essential for advancements in research and T-cell-based therapies. While most studies focus on machine learning for TCR specificity prediction, this approach, in addition to being difficult to implement, can be replaced by sequence-based methods. This work proposed a methodology based on global alignment, transformation of alignment scores into distance metrics, and hierarchical clustering to analyze TCR sequences from the IEDB, focusing on the CDR3 region. Here we show that the use of the BLOSUM62 substitution matrix for global alignment, followed by the transformation of the results into a distance metric, led to effective clustering of TCR sequences, highlighting the ability to separate them by their antigenic specificity to a specific pathogen. Although the study has limitations regarding the number of sequences and pathogens analyzed, the results indicate that the distance-based methodology can be a promising tool for TCR analysis. Future studies, expanding the number of samples and including a greater diversity of pathogens, could deepen the understanding of TCR organization in relation to its antigenic specificity.

Keywords: T-cell receptors; Antigenic specificity; Sequence analysis; Distance metrics; Global alignment.

SUMÁRIO

1 INTRODUÇÃO

Os receptores de células T (TCRs) são essenciais na resposta imunológica adaptativa, sendo responsáveis por reconhecer antígenos apresentados por moléculas de MHC e mediar a eliminação de células infectadas ou alteradas. Uma análise detalhada do repertório de TCRs possibilita insights sobre a diversidade funcional e a especificidade antigênica desses receptores, aspectos essenciais para o avanço de pesquisas e terapias imunológicas (DAVIS; BJORKMAN, 1988; HUPPA; DAVIS, 2003; ROSSJOHN et al., 2015; TOWNSEND; BODMER, 1989).

Atualmente, a maioria dos dados disponíveis sobre repertórios de TCRs é gerada por meio do método de sequenciamento de TCRs (PAI; SATPATHY, 2021). Este método foca na amplificação por reação em cadeia da polimerase (PCR) da região complementar determinante 3 (CDR3) da cadeia β, considerada de interesse devido à sua maior diversidade e especificidade em comparação às demais cadeias (ROBINS et al., 2009; WANG et al., 2010). Essa diversidade é ampliada pela presença de um segmento gênico D, ausente na cadeia α, permitindo combinações mais variadas. No entanto, a amplificação por PCR, embora eficiente, além de possivelmente trazer vieses no sequenciamento (SHUGAY et al., 2014), restringe as análises principalmente à região CDR3 da cadeia β, limitando a exploração de outras regiões potencialmente relevantes no repertório de receptores (LIU et al., 2016).

Neste trabalho, propomos uma abordagem para análise de sequências aminoacídicas da região CDR3 de TCRs baseada em métricas de distância (DASH et al., 2017; MAYER-BLACKWELL; FIORE-GARTLAND; THOMAS, 2022; MEYSMAN et al., 2023), utilizando a matriz de substituição BLOSUM62 (SONG et al., 2015) (PEARSON, 2013) para calcular pontuações de alinhamento global entre as sequências. Após isso, aplicou-se uma transformação nesse resultado para converter os dados em uma métrica de distância. Diferentemente de metodologias existentes (MAYER-BLACKWELL; FIORE-GARTLAND; THOMAS, 2022; MEYSMAN et al., 2023), a implementação é realizada integralmente em R, com etapas que vão desde a filtragem inicial dos dados até a análise de clusterização hierárquica e visualizações gráficas, como dendrogramas, grafos e mapas de calor.

Para validar a metodologia proposta, utilizou-se dados do IEDB (Immune Epitope Database), uma base de dados que cataloga dados experimentais sobre

TCRs e epítopos de células T estudados em humanos, contendo majoritariamente informações sobre sequências CDR3 de cadeia β dos receptores TCR (VAUGHAN; SETTE, 2013; VITA et al., 2015, 2019). Os dados foram selecionados com foco nos epítopos de dois patógenos com alta representatividade no banco de dados do IEDB, além de sua relevância clínica e imunológica: o SARS-CoV-2, com destaque para a glicoproteína Spike, e o *Mycobacterium tuberculosis*.

O presente trabalho busca evidenciar a eficácia da abordagem de métricas de distância em identificar agrupamentos biológicos relevantes e demonstrar sua viabilidade como ferramenta exploratória no estudo de repertórios de TCRs. A metodologia introduzida busca oferecer uma perspectiva flexível, com potencial para ser expandida em estudos futuros.

1.1 PROBLEMA

O problema central deste trabalho reside na escassez de metodologias que utilizem abordagens baseadas em distância para realizar agrupamento de TCRs por sua especificidade antigênica. (MEYSMAN et al., 2023)

Dessa forma, este estudo propõe-se a abordar essa lacuna ao desenvolver uma metodologia em R que combina o uso de distâncias baseadas na matriz BLOSUM62 com técnicas de visualização, como mapas de calor e grafos, para explorar relações e agrupamentos entre sequências de TCRs.

1.2 OBJETIVOS

1.2.1 Objetivo geral

Desenvolver e validar uma metodologia baseada em métricas de distância para a análise de sequências aminoacídicas de CDR3 de TCRs, visando agrupar elas de acordo com sua especificidade antigênica.

1.2.2 Objetivos específicos

- Utilizar o IEDB como base de dados, para organizar pares de patógenos e sequências de TCR reconhecedores.

- Implementar uma metodologia em R para calcular distâncias entre sequências de CDR3 a partir de algoritmos de alinhamento de sequências.
- Desenvolver visualizações, com ênfase nos mapas de calor e grafos, para interpretar as relações entre sequências.
- Validar a metodologia proposta com os dados provenientes do IEDB.
- Testar a capacidade da metodologia de revelar agrupamentos de TCRs associados à especificidade antigênica.

1.3 JUSTIFICATIVA

A compreensão do sistema imunológico enfrenta desafios crescentes devido à sua complexidade e, aliado aos avanços em bioinformática e tecnologias de sequenciamento, exige o desenvolvimento de ferramentas analíticas que explorem com maior profundidade os repertórios imunológicos (HUDSON et al., 2023; PAI; SATPATHY, 2021). Nesse contexto, investigar a diversidade, estrutura e especificidade antigênica desses receptores é essencial para desvendar novos mecanismos biológicos e identificar potenciais alvos terapêuticos (VUJOVIC et al., 2020; WANG et al., 2010).

Apesar dos avanços na caracterização molecular de TCRs, persistem lacunas importantes na literatura, especialmente no que diz respeito à utilização de metodologias baseadas em distância para análise desses repertórios e agrupamento dessas sequências por sua especificidade antigênica (MEYSMAN et al., 2023).

Portanto, ao explorar dados organizados de fontes como o IEDB (VAUGHAN; SETTE, 2013; VITA et al., 2015, 2019) e aplicar diferentes algoritmos, este trabalho busca refinar métodos que combinem métricas de distâncias para avaliar similaridades entre sequências de TCRs a partir de sua região CDR3, trazendo novas opções para exploração dos repertórios de TCRs e agrupamento dessas sequências.

2 REVISÃO DE LITERATURA

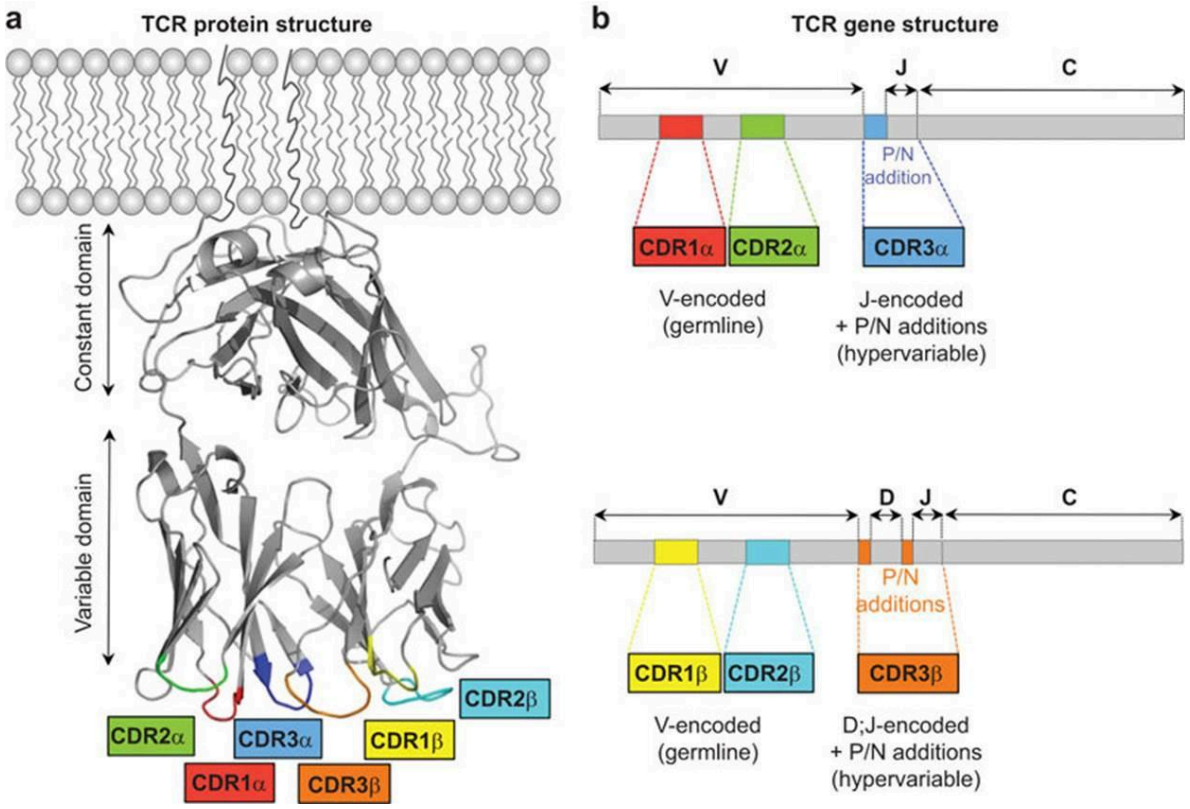
2.1 TCR E MÉTODOS DE SEQUENCIAMENTO

Os TCRs, componentes-chave da imunidade adaptativa, são formados por heterodímeros proteicos, compostos prevalentemente por cadeias α e β, que contêm domínios variáveis (V) e constantes (C) (ATTAF; HUSEBY; SEWELL, 2015), conforme apontado na Figura 1a. A habilidade de reconhecer antígenos é proporcionada pelas três regiões hipervariáveis: CDR1, CDR2 e CDR3, localizadas no domínio V de cada cadeia (HUPPA; DAVIS, 2003). As regiões CDR1 e CDR2 têm a função de estabilizar a interação com regiões conservadas do MHC, enquanto a CDR3 é a principal responsável pelo reconhecimento direto do peptídeo antigênico (ROSSJOHN et al., 2015), sendo formada pela junção das regiões V(D)J durante o rearranjo gênico, como observa-se na Figura 1b. A CDR3β sobressai por apresentar uma diversidade estrutural única, que é gerada por recombinação V(D)J, adição de nucleotídeos nas N-regiões e excisão de bases, tornando-a um marcador crítico da especificidade do TCR (Figura 1b). A diversidade do repertório de TCRs é fundamental para o reconhecimento de uma vasta gama de antígenos (DASH et al., 2017). A região CDR3α resulta da junção V-J com inserções de nucleotídeos, enquanto a CDR3β é formada pela junção V-D-J, incorporando diversidade adicional via inserções aleatórias nas N-regiões (ATTAF; HUSEBY; SEWELL, 2015). Essa variabilidade permite que a CDR3β adote conformações tridimensionais específicas, definindo a complementaridade com antígenos. A CDR3β é considerada uma "assinatura molecular" para identificar clones específicos, como respostas a patógenos ou à progressão tumoral (VUJOVIC et al., 2020). A análise das sequências de TCR pode revelar informações sobre a especificidade do reconhecimento de diferentes epítopos, além de que a interação do TCR com o complexo peptídeo-MHC (pMHC) leva à ativação das células T (TOWNSEND; BODMER, 1989).

O advento das tecnologias de sequenciamento de nova geração (NGS) revolucionou a pesquisa em imunologia, permitindo análises detalhadas de repertórios de TCRs (PAI; SATPATHY, 2021). Essas tecnologias possibilitam a identificação de milhares de sequências de TCR em uma única execução,

proporcionando uma visão abrangente da diversidade clonal presente no repertório imune (WANG et al., 2010).

FIGURA 1 – ESTRUTURA DO TCR E CDR3



FONTE: ATTAF, M.; HUSEBY, E.; SEWELL, A.K. (2015), licenciado sob CC BY 4.0.

LEGENDA: **(a)** Estrutura proteica do TCR, destacando os domínios constante (C) e variável (V) das cadeias α e β. As poções CDR1, CDR2 e CDR3 estão representadas em cores distintas, enfatizando sua localização no domínio variável. **(b)** Organização genética das regiões CDR: a CDR3α resulta da recombinação V-J com adições de nucleotídeos, enquanto a CDR3β é formada pela junção V-D-J, incorporando diversidade adicional via inserções aleatórias. A hipervariabilidade da CDR3β reflete sua centralidade no reconhecimento antigênico.

A abordagem mais amplamente utilizada para explorar repertórios de TCR por meio do NGS é o *bulk* TCR-seq, que envolve o sequenciamento de uma população mista de células T. Um aspecto crítico dessa metodologia é a preparação das bibliotecas de sequenciamento, comumente realizada utilizando a técnica de PCR multiplex (ROBINS et al., 2009). Essa abordagem permite a amplificação simultânea de múltiplos *loci* do TCR, e é projetada para capturar a variabilidade dos rearranjos V(D)J presentes nas sequências de CDR3. A PCR multiplex, por sua alta eficiência e baixo custo, vem sendo um padrão em estudos de *bulk* TCR-seq (PAI;

SATPATHY, 2021). No entanto, é necessário o design de *primers* otimizados para evitar reações inespecíficas, caso contrário há enorme potencial de amplificação desigual de regiões menos representadas (LIU et al., 2016). Além disso, o *bulk* não preserva informações sobre a célula individual, como sua origem funcional (por exemplo, CD4+ ou CD8+), sua expressão conjunta com outras moléculas relevantes e, principalmente, não possibilita a reconstrução do receptor TCR inteiro, pois os dados das cadeias α e β não são pareados, dessa forma são dados conhecidos como *single chain*, limitando inferências sobre a funcionalidade e especificidade do TCR (WANG et al., 2010).

2.2 IEDB (IMMUNE EPITOPE DATABASE)

O *Immune Epitope Database and Analysis Resource* (IEDB) foi desenvolvido como parte de uma iniciativa financiada pelo Instituto Nacional de Alergia e Doenças Infecciosas (NIAID) dos Estados Unidos (VAUGHAN; SETTE, 2013; VITA et al., 2015, 2019). O IEDB foi concebido como um repositório abrangente de dados relacionados a epítopos imunológicos, com o objetivo de impulsionar a pesquisa em imunologia, vacinas e doenças infecciosas (VAUGHAN; SETTE, 2013). Este recurso centraliza informações experimentais e computacionais sobre epítopos reconhecidos por células B e T, bem como epítopos apresentados por moléculas do MHC. Sua criação responde à necessidade crescente de integrar dados imunológicos gerados por diversos estudos em uma interface acessível e unificada (VITA et al., 2015, 2019).

O banco de dados armazena informações detalhadas para mais de 2,2 milhões de epítopos imunológicos, organizados para facilitar a exploração e a análise (VITA et al., 2019). Entre os dados disponíveis, destacam-se as sequências de epítopos reconhecidos por células T e apresentados por moléculas MHC, bem como os epítopos reconhecidos por células B, que incluem descrições detalhadas de epítopos lineares e conformacionais. Além disso, o IEDB contém dados relacionados à afinidade de ligação de peptídeos às moléculas MHC, especificidade antigênica, organismos de origem e condições experimentais (VAUGHAN; SETTE, 2013). Esses dados são coletados a partir de literatura revisada por pares, submissões de usuários e predições computacionais (VITA et al., 2015). IEDB também oferece uma série de ferramentas analíticas que vão além de sua função como repositório. Entre

as funcionalidades mais relevantes, estão as buscas avançadas, que permitem filtrar dados por parâmetros específicos como organismo de origem, tipo de molécula MHC e método experimental. Além disso, o banco de dados permite a exportação de conjuntos de dados personalizados, facilitando a integração com *pipelines* de bioinformática.

O impacto científico do IEDB é evidente em sua contribuição para o avanço da imunologia. Durante a pandemia de COVID-19, foi amplamente utilizado para mapear epítopos na proteína Spike do SARS-CoV-2, acelerando a criação de imunoterapias e vacinas (CHEN et al., 2020; GRIFONI et al., 2020). Além disso, o IEDB tem sido fundamental em estudos de autoimunidade, ajudando a desvendar os mecanismos de tolerância imunológica e as bases moleculares de doenças autoimunes (VAUGHAN; SETTE, 2013).

Neste estudo, o IEDB foi utilizado como uma fonte confiável para a obtenção de dados sobre epítopos conhecidos e suas interações com receptores de células T (TCRs). Os dados extraídos foram fundamentais para testar e validar as abordagens metodológicas desenvolvidas, como cálculos de distâncias e análises de clusterização de sequências CDR3.

2.3 ALINHAMENTO DE SEQUÊNCIAS E AS MATRIZES BLOSUM

O alinhamento de sequências biológicas é uma técnica central em bioinformática, amplamente utilizada para comparar sequências de DNA, RNA ou proteínas com o objetivo de identificar similaridades funcionais, estruturais e evolutivas (MCGINNIS; MADDEN, 2004; WALLACE; BLACKSHIELDS; HIGGINS, 2005; WATERMAN, 1984). Em particular, o alinhamento de sequência proteína-proteína e a comparação precisa dessas sequências é essencial para inferir funções biológicas, prever estruturas tridimensionais e investigar relações filogenéticas (SIEVERS; HIGGINS, 2018).

Um dos elementos críticos em alinhamentos de proteínas é a utilização de matrizes de pontuação ou substituição, que quantificam a probabilidade de substituições entre diferentes aminoácidos com base em dados biológicos reais (SONG et al., 2015). Essas matrizes incorporam a propensão evolutiva de certos aminoácidos serem trocados por outros ao longo do tempo, levando em consideração propriedades físico-químicas e a importância funcional dos resíduos.

Entre as matrizes mais amplamente utilizadas, destacam-se as séries PAM (Point Accepted Mutation) e BLOSUM (BLOcks SUbstitution Matrix) (PEARSON, 2013).

A matriz BLOSUM foi desenvolvida por Henikoff e Henikoff em 1992 como uma alternativa às matrizes PAM, focando na análise de blocos de sequências conservadas em proteínas (HENIKOFF; HENIKOFF, 1992). A matriz BLOSUM62, em particular, tornou-se o padrão de escolha para muitos algoritmos de alinhamento, como o BLASTP, devido à sua robustez e ampla aplicabilidade. O nome "BLOSUM62" refere-se ao uso de clusters de sequências com pelo menos 62% de identidade para calcular as probabilidades de substituição de aminoácidos (PEARSON, 2013).

O desenvolvimento técnico da matriz BLOSUM62 envolve várias etapas importantes. Inicialmente, sequências de proteínas são agrupadas em blocos baseados em alinhamentos múltiplos de regiões conservadas. Cada bloco representa um conjunto de sequências relacionadas que compartilham uma função ou estrutura comum. Em seguida, pares de aminoácidos em cada bloco são contados, e as frequências observadas de substituições são comparadas às frequências esperadas em uma distribuição aleatória. O logaritmo da razão entre a frequência observada e esperada (log-odds score) é então calculado para cada par de aminoácidos, gerando os valores que compõem a matriz (HENIKOFF; HENIKOFF, 1992; SONG et al., 2015).

A BLOSUM62 é particularmente eficaz para alinhamentos em um intervalo moderado de similaridade, sendo ideal para sequências com aproximadamente 20-62% de identidade (PEARSON, 2013). Além disso, a matriz incorpora implicitamente aspectos evolutivos, como a maior probabilidade de substituições entre aminoácidos com propriedades semelhantes, como tamanho, carga ou hidrofobicidade (HENIKOFF; HENIKOFF, 1992).

No contexto do BLASTP, a matriz serve como base para pontuar correspondências e substituições entre aminoácidos durante o alinhamento, permitindo a identificação de regiões conservadas e potenciais domínios funcionais em proteínas desconhecidas (MCGINNIS; MADDEN, 2004). Essa abordagem é fundamental para tarefas como anotação funcional de proteínas, detecção de homologia e estudos comparativos de genomas (SONG et al., 2015).

A escolha da BLOSUM62 em ferramentas de bioinformática se deve ao seu equilíbrio entre sensibilidade e especificidade, uma vez que matrizes com números

mais altos, como a BLOSUM80, são adequadas para alinhamentos de sequências altamente conservadas, enquanto matrizes com números mais baixos, como a BLOSUM45, são usadas para identificar relações mais distantes. A BLOSUM62, por sua vez, representa um ponto intermediário, sendo ideal para a maioria das aplicações práticas em biologia molecular (PEARSON, 2013).

2.4 MÉTODOS PARA AGRUPAMENTO DE TCRs

O agrupamento de TCRs com base em sua especificidade antigênica é um dos desafios centrais na imunologia computacional, pois a ligação de um TCR a um epítopo depende de interações complexas entre CDRs e os complexos peptídeo-MHC (HUDSON et al., 2023; LEE et al., 2020). Essa tarefa é fundamental para entender respostas imunológicas específicas, prever reações imunológicas em infecções e terapias baseadas em células T, bem como estudar a diversidade e especificidade dos repertórios imunológicos (PAI; SATPATHY, 2021). Atualmente, os métodos de agrupamento de TCRs em relação a epítopos específicos podem ser amplamente divididos em duas categorias principais: métodos baseados em distância e métodos baseados em características (MEYSMAN et al., 2023).

Os métodos baseados em distância fundamentam-se no cálculo de uma métrica que quantifica a similaridade entre as sequências de TCRs. Uma das ferramentas mais proeminentes nessa abordagem é o *TCRdist (DASH et al., 2017; MAYER-BLACKWELL; FIORE-GARTLAND; THOMAS, 2022)*, que mede a distância entre TCRs considerando propriedades estruturais e químicas das sequências. A forma mais simples de mensurar essa distância é calcular o número de aminoácidos divergentes entre duas sequências de CDR3, considerando que um menor número de discrepâncias sugere maior similaridade funcional (VUJOVIC et al., 2020). Em modelos mais elaborados, como o *TCRdist,* aplica-se matrizes de substituição baseadas na BLOSUM62 para calcular a distância entre sequências, após isso o agrupamento considera um limiar de distância pré-definido para julgar se dois TCRs são suficientemente similares para pertencerem ao mesmo cluster de forma que, quanto maior a distância, menor a confiabilidade na previsão de especificidade compartilhada (MAYER-BLACKWELL; FIORE-GARTLAND; THOMAS, 2022). Métodos baseados em distância têm a vantagem de não depender de conhecimento

prévio sobre os epítopos, sendo, portanto, aplicáveis a TCRs ainda não caracterizados (MEYSMAN et al., 2023).

Em contraste, os métodos baseados em características utilizam abordagens de aprendizado supervisionado para identificar padrões comuns em TCRs que ligam epítopos conhecidos (HUDSON et al., 2023). Esses métodos dependem de dados de treinamento que contêm TCRs anotados para epítopos específicos. O objetivo principal é construir um modelo preditivo que capte as características subjacentes que distinguem os TCRs de um epítopo em particular (MEYSMAN et al., 2023). As características podem incluir propriedades físico-químicas dos aminoácidos, posições específicas na sequência CDR3 ou interações estruturais previstas.

Os modelos baseados em características utilizam algoritmos avançados de aprendizado de máquina, como redes neurais, *support vector machines* (SVMs) ou florestas aleatórias, para ajustar as predições às especificidades do epítopo. Após o treinamento, esses modelos são aplicados para prever a especificidade antigênica de TCRs não vistos anteriormente, com base nas características aprendidas. Apesar de serem potencialmente mais poderosos, os métodos baseados em características apresentam limitações, como a dependência de grandes volumes de dados de treinamento anotados e a dificuldade em interpretar os padrões aprendidos, o que pode dificultar a validação experimental (HUDSON et al., 2023; MEYSMAN et al., 2023; NIELSEN et al., 2024).

Comparações entre os dois tipos de abordagem mostram que métodos baseados em distância, como o *TCRdist*, podem atingir desempenhos comparáveis aos melhores métodos baseados em características em determinados contextos (MEYSMAN et al., 2023; NIELSEN et al., 2024). Isso ressalta a relevância dos métodos baseados em distância como linha de base para novos algoritmos, além de destacar sua simplicidade e interpretabilidade como vantagens em aplicações práticas.

3 MATERIAL E MÉTODOS

3.1 OBTENÇÃO DOS DADOS DO IEDB

Os dados utilizados, contendo informação sobre os receptores TCR e seus pares antigênicos, foram obtidos através do portal online do IEDB (Immune Epitope Database) (VITA et al., 2019), disponível em iedb.org. O IEDB foi selecionado como fonte devido à sua ampla cobertura de epitopos imunogênicos e dados padronizados derivados de ensaios experimentais. Conforme ilustrado na Figura 2, os dados foram selecionados de forma a incluir apenas ensaios de células T realizados em hospedeiros humanos, sem filtros quanto a epitopos (lineares ou não-lineares), patógenos ou doenças. Após a seleção, os dados foram exportados no formato CSV (Comma-separated value) para serem carregados e manipulados em R.

FIGURA 2 – PORTAL DO IEDB

FONTE: IEDB (2024)

LEGENDA: Portal do IEDB na internet por onde é possível baixar os dados disponíveis em seu repositório. A imagem demonstra como foi feita a seleção dos dados a serem exportados.

O arquivo CSV final exportado é composto por 215399 observações e 70 colunas, contendo informações sobre a caracterização experimental do receptor TCR e informações sobre o seu par antigênico.

3.2 CARREGAMENTO E MANIPULAÇÃO DOS DADOS

Os dados exportados do portal online do IEDB, no formato CSV, foram carregados utilizando a linguagem de programação R. Inicialmente, todas as colunas presentes no arquivo foram avaliadas a fim de determinar quais seriam relevantes para o escopo do trabalho. Dessa forma, foram selecionadas especificamente as colunas que continham informações sobre o patógeno, antígeno, epítopo, a sequência aminoacídica do CDR3 e o tipo de cadeia (alfa ou beta) dos respectivos TCRs, reduzindo o arquivo de 70 para 5 colunas.

Após o carregamento inicial, os dados foram submetidos a uma etapa de filtragem para criar dois grupos distintos baseados no patógeno de origem das amostras. O primeiro grupo foi composto por epítopos relacionados ao *SARS-CoV-2*, enquanto o segundo grupo incluiu epítopos associados ao *Mycobacterium tuberculosis*. Essa escolha, inicialmente arbitrária, foi fundamentada na disponibilidade de informações completas para cada patógeno nos dados selecionados (VAUGHAN; SETTE, 2013), além da relevância imunológica e clínica desses patógenos: o primeiro devido à pandemia recente e seu impacto global, e o segundo por ser o agente causador da tuberculose, uma das doenças infecciosas mais prevalentes no mundo (COHEN et al., 2019).

Para fins de análise inicial e validação da metodologia, foram amostradas e combinadas 5 sequências aminoacídicas de CDR3 para cada grupo, conforme mostrado no Quadro 1. Essas sequências foram selecionadas da seguinte forma: para a seleção das sequências relacionadas ao *SARS-CoV-2*, foi realizada uma filtragem específica para a glicoproteína Spike. Primeiramente, as sequências foram agrupadas com base no epítopo correspondente e a frequência de cada epítopo foi calculada. O epítopo mais frequente foi então considerado como representativo da glicoproteína Spike e, após isso, foram amostradas aleatoriamente 5 sequências de CDR3 associadas a este epítopo. Para as sequências relativas ao *Mycobacterium tuberculosis*, o mesmo processo de agrupamento foi aplicado para identificar a molécula de origem mais frequente para este patógeno, com amostragem de 5

sequências CDR3 associadas ao epítopo predominante. Todas regiões CDR3 amostradas são originadas da cadeia beta de seu respectivo TCR, uma vez que a maioria dos TCRs nesse conjunto de dados possui informação incompleta para a cadeia alfa.

QUADRO 1 – SEQUÊNCIAS CDR3 AMOSTRADAS PARA CADA GRUPO

epitope sequence	source molecule	source organism	chain type	CDR3aa
KLPDDFTGCV	Spike glycoprotein	SARS-CoV2	beta	ASSSYRDRVYSPLH
KLPDDFTGCV	Spike glycoprotein	SARS-CoV2	beta	ASSSFGTGGYEQY
KLPDDFTGCV	Spike glycoprotein	SARS-CoV2	beta	ASRESLLAGGPDTQY
KLPDDFTGCV	Spike glycoprotein	SARS-CoV2	beta	ASSQGLWGAQETQY
KLPDDFTGCV	Spike glycoprotein	SARS-CoV2	beta	ASSLGSVTLGALNSALH
VMATRRNVL	Uncharacterized protein MT1568	Mycobacterium tuberculosis	beta	ASSIRGQRGYT
VMATRRNVL	Uncharacterized protein MT1568	Mycobacterium tuberculosis	beta	ASSLSGGIDTQY
VMATRRNVL	Uncharacterized protein MT1568	Mycobacterium tuberculosis	beta	ASGPLSGEQY
VMATRRNVL	Uncharacterized protein MT1568	Mycobacterium tuberculosis	beta	ASRTPHVTEAF
VMATRRNVL	Uncharacterized protein MT1568	Mycobacterium tuberculosis	beta	ASSQVVSTDTQY

FONTE: O autor (2024).

A decisão de limitar o número de grupos e sequências nesta etapa inicial do trabalho foi tomada para facilitar a construção e visualização dos resultados obtidos, bem como para reduzir a complexidade computacional durante o desenvolvimento da metodologia. No entanto, a perspectiva futura do trabalho é expandir tanto o número de grupos de patógenos quanto a quantidade de sequências analisadas, permitindo uma abordagem mais abrangente e representativa do repertório de TCRs.

3.3 ALINHAMENTO DAS SEQUÊNCIAS E CÁLCULO DE DISTÂNCIA

Para avaliar a similaridade entre as sequências CDR3 dos receptores, foi utilizado o alinhamento global com a matriz de substituição BLOSUM62 (HENIKOFF; HENIKOFF, 1992), sendo escolhida por ser adequada para a análise de proteínas com moderado nível de divergência evolutiva (PEARSON, 2013). O alinhamento global das sequências foi conduzido utilizando a função *pairwiseAlignment* do pacote *Biostrings* em uma abordagem "todos contra todos", garantindo que todas as

sequências fossem comparadas entre si. As pontuações de similaridade resultantes foram calculadas em R. Para cada par de sequências $\boldsymbol{s_i}$ e $\boldsymbol{s_j}$, a pontuação de alinhamento foi extraída diretamente do alinhamento, resultando em uma matriz de similaridade ***S***, onde cada elemento $S(i,j)$ representa a pontuação de alinhamento entre as sequências ***i*** e ***j***. A diagonal principal da matriz foi preenchida com as pontuações de alinhamento das sequências com elas mesmas, correspondendo ao alinhamento perfeito.

Para converter a matriz de similaridade ***S*** em uma matriz de distância ***D***, aplicou-se a seguinte transformação em cada um dos elementos da matriz ***S***: $D(i,\ j) = max(S(,\ j)) - S(i,\ j)$, onde $max(S(,\ j))$ é o maior valor na coluna ***j*** da matriz ***S*** e $D(i,\ j)$ é o elemento equivalente a $S(i,\ j)$ após a transformação. Essa abordagem normaliza as distâncias em relação à similaridade máxima de cada coluna, assegurando que os valores de ***D*** sejam não negativos. Para garantir que a matriz ***D*** seja simétrica, a região abaixo da diagonal foi preenchida durante a computação para otimização do processo e a região acima da diagonal foi espelhada para completar a matriz posteriormente.

A transformação das pontuações de alinhamento em distâncias é uma prática já descrita na literatura em análises baseadas em distância, onde esse método calcula matrizes de distância entre sequências e as utiliza para identificar padrões e agrupamentos baseados em semelhanças funcionais e estruturais entre os TCRs (DASH et al., 2017; MAYER-BLACKWELL; FIORE-GARTLAND; THOMAS, 2022). De maneira semelhante, a transformação aplicada e o uso da matriz ***D*** no presente trabalho busca extrair relações entre as sequências CDR3 dos TCRs, além de permitir que o método subsequente de clusterização hierárquica utilize a matriz ***D*** como entrada, pois ele baseia-se em métricas de dissimilaridade para agrupar os objetos.

3.4 CLUSTERIZAÇÃO HIERÁRQUICA E VISUALIZAÇÃO GRÁFICA

A clusterização hierárquica foi realizada com o objetivo de agrupar as sequências CDR3. Essa abordagem é amplamente utilizada em dados biológicos devido à sua capacidade de lidar com dados de dissimilaridade, permitindo uma análise sem a necessidade de especificar previamente o número de clusters (MURTAGH; CONTRERAS, 2012). No presente trabalho, utilizou-se a matriz de

distância ***D*** previamente gerada como entrada para a função *hclust* do R, que implementa o algoritmo de clusterização hierárquica aglomerativa. Executou-se o algoritmo utilizando os parâmetros padrão, atribuindo cada sequência ao seu próprio cluster e, de forma iterativa, combinando os dois clusters mais próximos em cada etapa até formar um único cluster contendo todas as sequências, por meio do método de aglomeração *complete linkage*.

O resultado da clusterização foi um objeto dendrogramático, referente às sequências representadas na matriz de distância. A representação gráfica do dendrograma foi possível através do uso da função plot do R, que gera visualizações tradicionais, nas quais os ramos representam os agrupamentos formados em cada estágio do algoritmo.

Além da visualização convencional, o dendrograma resultante foi convertido em um grafo utilizando o pacote *TreeAndLeaf (CARDOSO et al., 2022)*, uma ferramenta que oferece uma representação alternativa da estrutura hierárquica. Nesta etapa, os nós do grafo foram customizados com atributos visuais, como cores associadas aos patógenos de origem, ampliando a capacidade de interpretar os agrupamentos. Após isso, o grafo gerado foi visualizado com o pacote *RedeR (CASTRO et al., 2012)*.

Por fim, para explorar em conjunto a matriz de distância e os resultados da clusterização hierárquica, foi gerado um mapa de calor utilizando o pacote *pheatmap*. Este gráfico combina os dendrogramas das linhas e colunas com a matriz de distância, utilizando uma escala de cores que varia do branco ao vermelho para representar os valores de distância. A integração do mapa de calor com os dendrogramas destacou os agrupamentos hierárquicos e facilitou a identificação de padrões e associações relevantes entre as sequências.

4 APRESENTAÇÃO DOS RESULTADOS

4.1 ALINHAMENTO E TRANSFORMAÇÃO EM VALORES DE DISTÂNCIA

Os resultados do alinhamento global entre as sequências aminoacídicas das regiões CDR3 foram representados graficamente nas Figuras 3 e 4, correspondendo à matriz de similaridade e à matriz de distâncias, respectivamente. A Figura 3 apresenta a matriz de similaridade, construída a partir do alinhamento global de todas as sequências, utilizando a matriz BLOSUM62 como base para o cálculo das pontuações. Na Figura 3, os valores positivos indicam substituições conservativas, refletindo similaridades entre aminoácidos que são mais frequentemente trocados em estruturas evolutivamente relacionadas, enquanto valores negativos representam substituições não conservativas, sinalizando maior incompatibilidade entre os aminoácidos alinhados (SONG et al., 2015). Essas pontuações derivam das probabilidades de ocorrência de substituições específicas em alinhamentos globais de proteínas (HENIKOFF; HENIKOFF, 1992).

FIGURA 3 – MATRIZ DE SIMILARIDADE DAS SEQUÊNCIAS

	ASSSYRDRVYSPLH	ASSSFGTGGYEQY	ASRESLLAGGPDTQY	ASSQGLWGAQETQY	ASSLGSVTLGALNSALH	ASSIRGQRGYT	ASSLSGGIDTQY	ASGPLSGEQY	ASRTPHVTEAF	ASSQVVSTDTQY
ASSSYRDRVYSPLH	73	0	0	0	0	0	0	0	0	0
ASSSFGTGGYEQY	5	69	0	0	0	0	0	0	0	0
ASRESLLAGGPDTQY	-16	3	76	0	0	0	0	0	0	0
ASSQGLWGAQETQY	-1	11	21	75	0	0	0	0	0	0
ASSLGSVTLGALNSALH	-4	-14	-14	-5	79	0	0	0	0	0
ASSIRGQRGYT	-6	4	-24	-15	-26	55	0	0	0	0
ASSLSGGIDTQY	-15	10	22	17	-13	-4	59	0	0	0
ASGPLSGEQY	-17	1	-13	-11	-21	-8	2	52	0	0
ASRTPHVTEAF	-18	-11	-16	-22	-25	-7	-7	-1	57	0
ASSQVVSTDTQY	-12	3	8	16	-22	-10	27	1	-5	57

FONTE: O autor (2024).

LEGENDA: Resultado do alinhamento global, representado como uma matriz de similaridade entre as sequências aminoacídicas da região CDR3 de cada receptor TCR.

A Figura 4, por sua vez, ilustra a matriz de distâncias, derivada da transformação da matriz de similaridade, produzindo um conjunto de dados que permitem a representação quantitativa das distâncias entre sequências. Valores baixos na matriz de distâncias correspondem a sequências mais similares, enquanto valores elevados refletem maior divergência entre elas. Essa transformação foi crucial para viabilizar a aplicação de análises subsequentes baseadas em distância, como clusterização hierárquica e análise de agrupamentos.

FIGURA 4 – MATRIZ DE DISTÂNCIAS DAS SEQUÊNCIAS

	ASSSYRDRVYSPLH	ASSSFGTGGYEQY	ASRESLLAGGPDTQY	ASSQGLWGAQETQY	ASSLGSVTLGALNSALH	ASSIRGQRGYT	ASSLSGGIDTQY	ASGPLSGEQY	ASRTPHVTEAF	ASSQVVSTDTQY
ASSSYRDRVYSPLH	0	68	89	74	77	79	88	90	91	85
ASSSFGTGGYEQY	68	0	66	58	83	65	59	68	80	66
ASRESLLAGGPDTQY	89	66	0	55	90	100	54	89	92	68
ASSQGLWGAQETQY	74	58	55	0	80	90	58	86	97	59
ASSLGSVTLGALNSALH	77	83	90	80	0	105	92	100	104	101
ASSIRGQRGYT	79	65	100	90	105	0	59	63	62	65
ASSLSGGIDTQY	88	59	54	58	92	59	0	57	66	32
ASGPLSGEQY	90	68	89	86	100	63	57	0	53	51
ASRTPHVTEAF	91	80	92	97	104	62	66	53	0	62
ASSQVVSTDTQY	85	66	68	59	101	65	32	51	62	0

FONTE: O autor (2024).

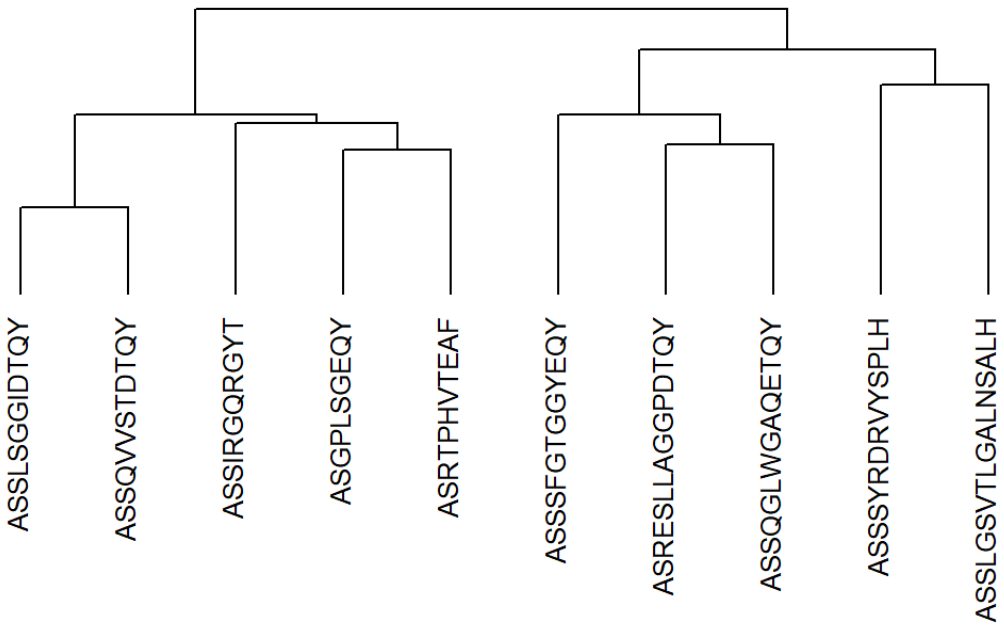
LEGENDA: Resultado da transformação da matriz de similaridade em uma matriz de distâncias, seguindo o cálculo descrito na metodologia.

Houve a tendência de observar as menores distâncias entre sequências pertencentes ao mesmo grupo, podendo sugerir tendências de conservação em regiões específicas do CDR3 (SONG et al., 2015). Dessa forma, os resultados apresentados estão em concordância com o esperado para abordagens baseadas em distância de sequências, como previamente descrito na literatura (DASH et al., 2017; MAYER-BLACKWELL; FIORE-GARTLAND; THOMAS, 2022).

4.2 CLUSTERIZAÇÃO HIERÁRQUICA

O resultado da clusterização hierárquica das sequências aminoacídicas da região CDR3, conduzida conforme descrito na metodologia e realizado a partir da matriz de distâncias entre as sequências, está representado no dendrograma da Figura 5. O dendrograma revela dois agrupamentos principais, que correspondem aos grupos de sequências associadas a diferentes patógenos. Especificamente, observa-se que o agrupamento à direita do dendrograma compreende as sequências relacionadas ao reconhecimento do patógeno SARS-CoV-2, enquanto o agrupamento à esquerda concentra as sequências associadas ao *Mycobacterium tuberculosis*.

FIGURA 5 – DENDROGRAMA RESULTANTE DA CLUSTERIZAÇÃO HIERÁRQUICA



FONTE: O autor (2024).

LEGENDA: Dendrograma resultante da clusterização hierárquica das sequências CDR3 feita pelo algoritmo *hclust*, é possível visualizar dois *clusters* bem definidos, sendo que cada *cluster* é referente a um dos grupos definidos inicialmente: À direita observa-se o cluster das sequências as quais reconhecem o patógeno SARS-CoV-2 e à esquerda as sequências que reconhecem o patógeno *Mycobacterium tuberculosis*.

Os agrupamentos observados atenderam à expectativa inicial, que era a formação de clusters distintos entre os grupos associados a diferentes patógenos,

alinhando-se ao esperado para abordagens baseadas em distância. Embora ainda sejam necessárias análises complementares, é possível identificar tendências relevantes. Por exemplo, a separação clara entre sequências associadas a diferentes patógenos sugere que as distâncias calculadas entre as sequências CDR3 foram capazes de capturar diferenças e similaridades que podem estar relacionadas ao reconhecimento de antígenos distintos, conforme observado anteriormente na literatura (MAYER-BLACKWELL; FIORE-GARTLAND; THOMAS, 2022). Essa tendência, embora preliminar, destaca o potencial de expandir a metodologia em estudos futuros para mais sequências e mais grupos distintos.

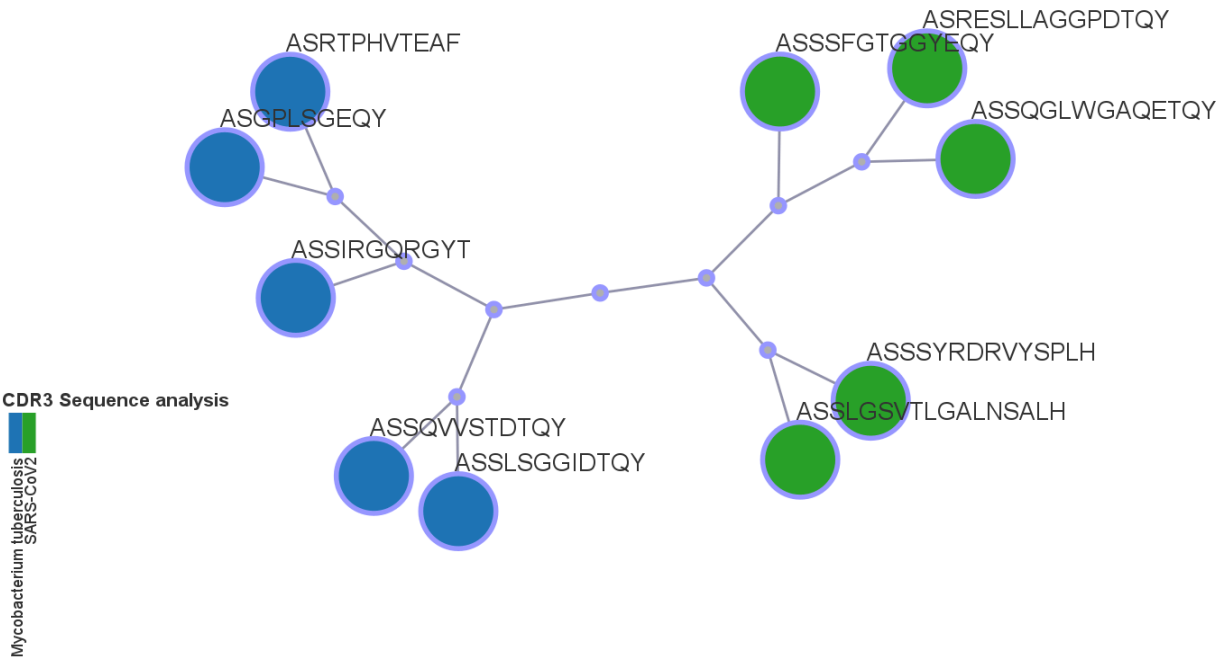
4.3 VISUALIZAÇÃO GRÁFICA: GRAFO E MAPA DE CALOR

Os resultados da visualização gráfica das sequências CDR3 foram representados através de duas abordagens distintas: um grafo e um mapa de calor, com o objetivo de explorar visualmente o dendrograma gerado anteriormente e os padrões de agrupamento das sequências, baseado nas distâncias calculadas entre elas.

A primeira visualização foi construída a partir do dendrograma da Figura 5, onde cada nó do grafo representa uma sequência, e as arestas conectam sequências com base na similaridade ou dissimilaridade de suas regiões CDR3. A Figura 6 ilustra o grafo gerado, onde é possível observar melhor os agrupamentos das sequências, com clusters de sequências que estão fortemente conectados entre si, indicando maior similaridade.

O grafo gerado confirma a tendência observada anteriormente no dendrograma, onde as sequências reconhecedoras do SARS-CoV-2 e do *Mycobacterium tuberculosis* tendem a se agrupar em regiões distintas, com sequências mais semelhantes. Esse agrupamento visual ajuda a validar a análise hierárquica anterior, fornecendo uma confirmação visual de que as distâncias entre as sequências refletem corretamente as diferenças entre os grupos.

FIGURA 6 - GRAFO DAS SEQUÊNCIAS CDR3



FONTE: O autor (2024).

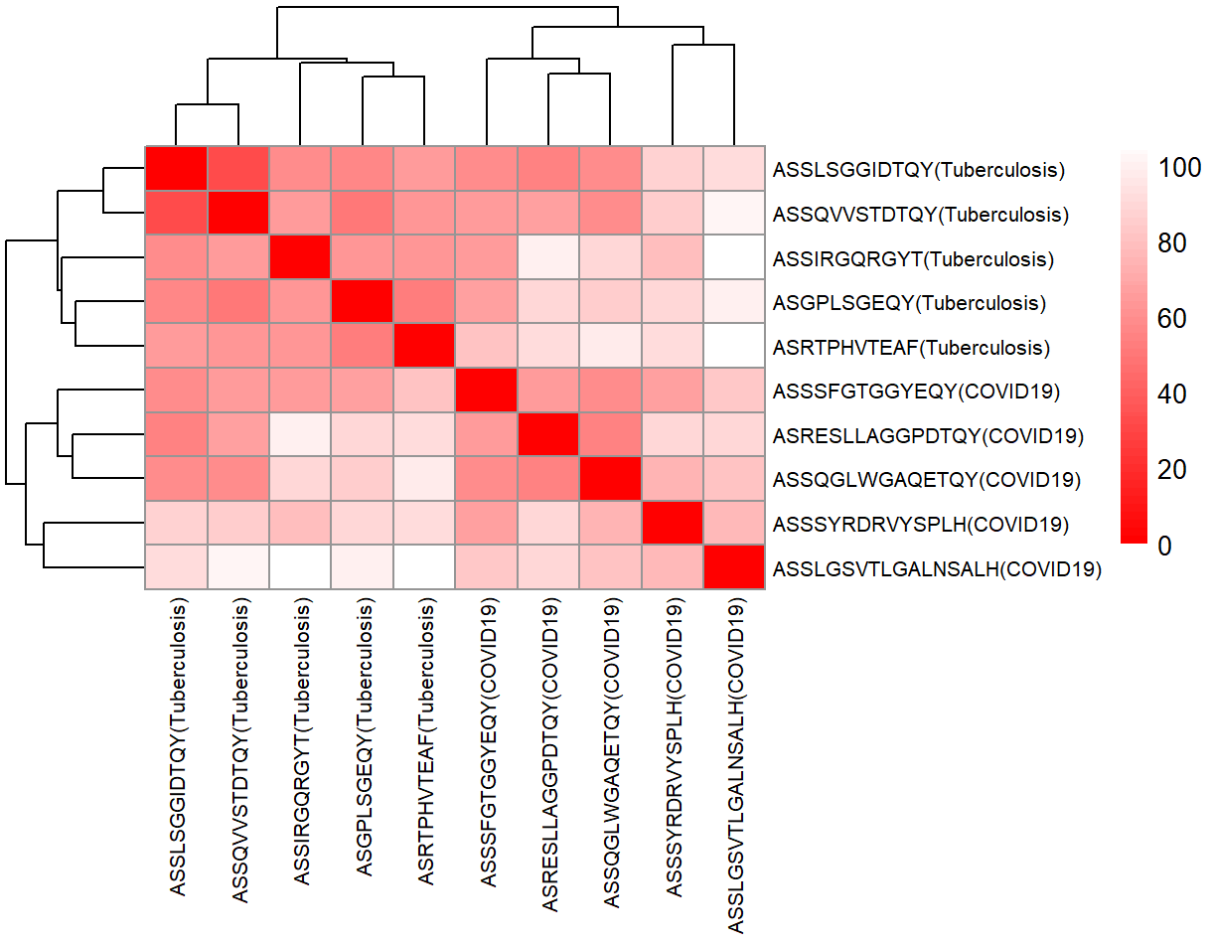
LEGENDA: Grafo gerado a partir das distâncias de similaridade entre as sequências CDR3. Os clusters de sequências são evidentes pela proximidade entre os nós, e as diferentes cores dos nós correspondem a diferentes patógenos.

A segunda visualização foi realizada utilizando um mapa de calor, como mostrado na Figura 7, onde as distâncias entre as sequências CDR3 são representadas por cores, variando do vermelho (indicado para baixa distância) ao branco (representando maior distância). A organização das sequências no mapa de calor foi feita de acordo com os resultados da clusterização hierárquica, de modo que as sequências com maior similaridade estão dispostas proximamente e são refletidas por cores mais próximas do vermelho. A Figura 6 mostra dois grandes blocos de sequências, com um bloco concentrado em torno das sequências reconhecedoras do SARS-CoV-2 e o outro em torno das sequências associadas ao *Mycobacterium tuberculosis*.

Essa visualização facilita a identificação dos agrupamentos, de forma que os clusters formados por sequências mais similares são facilmente discerníveis pelas cores mais quentes, enquanto as sequências mais distantes são indicadas pelas cores frias. Os resultados do mapa de calor reforçam as tendências observadas no

dendrograma e no grafo, proporcionando uma visualização adicional de que a metodologia baseada em distância foi capaz de agrupar sequências CDR3, separadas pelo seu reconhecimento a patógenos diferentes.

FIGURA 7 - MAPA DE CALOR COM DENDROGRAMA DAS SEQUÊNCIAS CDR3



FONTE: O autor (2024).

LEGENDA: Mapa de calor representando as distâncias de similaridade entre as sequências CDR3. As sequências do mesmo grupo antigênico formam blocos de alta similaridade, evidenciando a correspondência entre a visualização do mapa de calor e o dendrograma.

No entanto, apesar de as sequências agruparem de acordo com os patógenos, é possível observar uma baixa distância entre algumas sequências de Tuberculose e COVID, o que sugere um certo grau de sobreposição entre esses dois grupos. Tal fenômeno é totalmente plausível e esperado, pois a região CDR3, embora desempenhe um papel crucial no reconhecimento antigênico, pode não ser suficiente, por si só, para garantir uma separação absoluta entre os grupos de patógenos, conforme descrito anteriormente (DASH et al., 2017;

MAYER-BLACKWELL; FIORE-GARTLAND; THOMAS, 2022; MEYSMAN et al., 2023; NIELSEN et al., 2024; VUJOVIC et al., 2020). A variação natural das sequências CDR3 e as possíveis interações complexas entre os TCRs e seus antígenos podem contribuir para essa sobreposição. Apesar dessa limitação, o resultado permanece satisfatório, uma vez que a separação geral entre os grupos ainda é claramente observada, o que valida a abordagem baseada em distância e demonstra que a metodologia adotada é eficaz para identificar as principais diferenças entre os grupos antigênicos.

5 CONSIDERAÇÕES FINAIS

Este trabalho apresentou uma abordagem metodológica para a análise de TCRs baseada em dados provenientes do IEDB, empregando estratégias de alinhamento global, transformação de similaridades em métricas de distância e clusterização hierárquica. Os resultados obtidos não apenas confirmam a eficácia de tais métodos em destacar padrões biológicos relevantes, mas também reforçam a aplicabilidade de análises baseadas em distância para a compreensão das relações entre sequências de TCRs.

A utilização do alinhamento global com a matriz de substituição BLOSUM62 mostrou-se uma escolha adequada, permitindo calcular similaridades entre sequências de forma consistente. A transformação das pontuações de similaridade em distâncias para viabilizar a aplicação de métodos de clusterização hierárquica também funcionou de maneira eficaz, demonstrando uma separação clara entre grupos de receptores associados a diferentes patógenos. O agrupamento hierárquico evidenciou tendências de similaridade entre os receptores, com sequências relacionadas ao mesmo patógeno formando clusters bem definidos. Este resultado reforça a hipótese de que as regiões CDR3 possivelmente possuem assinaturas características que refletem especificidade antigênica, como já descrito na literatura (HUDSON et al., 2023; LEE et al., 2020).

Contudo, é fundamental reconhecer as suas limitações na metodologia empregada pois, ao focar-se no alinhamento global das regiões CDR3 e na subsequente clusterização hierárquica, não se considerou a normalização dos

CDR3s pelo seu comprimento. Essa falha pode ter introduzido vieses na análise, uma vez que sequências mais longas tendem a acumular maiores pontuações de similaridade, independentemente da sua real relação funcional. Para mitigar essa limitação, recomenda-se incorporar uma etapa de normalização que ajuste as pontuações de alinhamento em função do comprimento das sequências.

Ademais, a utilização de representações gráficas, como dendrogramas e grafos, complementada por mapas de calor, contribuiu significativamente para a interpretação dos resultados. O uso do pacote *TreeAndLeaf* para conversão de dendrogramas em grafos, aliado à integração com o pacote RedeR, mostrou-se particularmente promissor, permitindo uma análise mais intuitiva e interativa dos agrupamentos hierárquicos.

É importante destacar as limitações inerentes ao escopo deste estudo. Primeiramente, o número reduzido de sequências e patógenos analisados, embora justificado pela necessidade de reduzir a complexidade computacional, limita a generalização dos achados. Além disso, a análise poderia ser enriquecida pela avaliação de outras características dos TCRs, como propriedades físico-químicas dos aminoácidos, posições específicas na sequência CDR3, ou interações estruturais previstas. A inclusão dessas características adicionais permitiria uma descrição mais completa e precisa do repertório de TCRs, potencialmente melhorando a capacidade de prever a especificidade antigênica. Em estudos futuros, a ampliação do número de amostras e a inclusão de uma maior diversidade de patógenos poderão fornecer uma visão mais abrangente sobre a organização dos agrupamentos de TCRs a partir de sua especificidade antigênica.

Algoritmos como o TRUST 4 (SONG et al., 2021), desenvolvido para a reconstrução de sequências de TCR a partir de dados de RNA-seq convencionais que não foram preparados especificamente para análises de repertórios imunológicos, utilizam abordagens de mapeamento e montagem de alta precisão para identificar sequências de TCR diretamente de dados de transcriptoma. Essa capacidade torna o TRUST4 uma solução para análises de dados de RNA-seq, expandindo significativamente as possibilidades de estudo de repertórios TCR.

Apesar dessas limitações, os resultados obtidos neste estudo fornecem uma base sólida para o desenvolvimento de metodologias mais avançadas para a análise de TCRs. Ao reconhecer e abordar as fragilidades do presente trabalho, podemos

direcionar esforços futuros para aprimorar a nossa compreensão da complexa relação entre TCRs e antígenos nos mais diversos contextos.

5.1 RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Estudos futuros podem expandir a análise realizada no presente trabalho, que utilizou exclusivamente a região CDR3 da cadeia β, para integrar informações tanto da cadeia α quanto da β. A literatura aponta que a combinação das duas cadeias melhora significativamente o desempenho na predição de epítopos, dado que ambas participam do reconhecimento antigênico no contexto do heterodímero completo (MEYSMAN et al., 2023; NIELSEN et al., 2024). Para isso, recomenda-se a utilização de dados de sequenciamento de TCRs em nível de célula única, o que permitiria uma análise detalhada dessas interações (WANG et al., 2010).

Além disso, métodos futuros podem explorar a inclusão de informações adicionais das regiões CDR1 e CDR2, que, embora menos variáveis que a CDR3, desempenham um papel crucial no reconhecimento do epítopo e no contato com o complexo TCR-MHC (DAVIS; BJORKMAN, 1988; HUPPA; DAVIS, 2003; TOWNSEND; BODMER, 1989). Essa abordagem seria particularmente útil para capturar a contribuição combinada das regiões CDR1, CDR2 e CDR3, além de facilitar comparações entre o uso direto das sequências de aminoácidos.

Dessa forma, a combinação de dados pareados das cadeias α e β com informações detalhadas das regiões CDR1, CDR2 e CDR3 pode oferecer uma visão mais abrangente das interações TCR-epitopo, permitindo avaliar a eficácia de diferentes abordagens e contribuir para o desenvolvimento de métodos mais robustos e precisos.

REFERÊNCIAS

ATTAF, M.; HUSEBY, E.; SEWELL, A. K. αβ T cell receptors as predictors of health and disease. **Cellular & Molecular Immunology**, v. 12, n. 4, p. 391–399, Jul. 2015.

CARDOSO, M. A. et al. TreeAndLeaf: an R/Bioconductor package for graphs and trees with focus on the leaves. **Bioinformatics**, v. 38, n. 5, p. 1463–1464, 7 Feb. 2022.

CASTRO, M. A. A. et al. RedeR: R/Bioconductor package for representing modular structures, nested networks and multiple levels of hierarchical associations. **Genome Biology**, v. 13, n. 4, p. R29, 24 Apr. 2012.

CHEN, H.-Z. et al. Bioinformatics analysis of epitope-based vaccine design against the novel SARS-CoV-2. **Infectious diseases of poverty**, v. 9, n. 1, p. 88, 10 Jul. 2020.

COHEN, A. et al. The global prevalence of latent tuberculosis: a systematic review and meta-analysis. **The European Respiratory Journal**, v. 54, n. 3, 12 Sep. 2019.

DASH, P. et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. **Nature**, v. 547, n. 7661, p. 89–93, 6 Jul. 2017.

DAVIS, M. M.; BJORKMAN, P. J. T-cell antigen receptor genes and T-cell recognition. **Nature**, v. 334, n. 6181, p. 395–402, 4 Aug. 1988.

GRIFONI, A. et al. A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. **Cell Host & Microbe**, v. 27, n. 4, p. 671- 680.e2, 8 Apr. 2020.

HENIKOFF, S.; HENIKOFF, J. G. Amino acid substitution matrices from protein blocks. **Proceedings of the National Academy of Sciences of the United States of America**, v. 89, n. 22, p. 10915–10919, 15 Nov. 1992.

HUDSON, D. et al. Can we predict T cell specificity with digital biology and machine learning? **Nature Reviews. Immunology**, v. 23, n. 8, p. 511–521, Aug. 2023.

HUPPA, J. B.; DAVIS, M. M. T-cell-antigen recognition and the immunological synapse. **Nature Reviews. Immunology**, v. 3, n. 12, p. 973–983, Dec. 2003.

LEE, C. H. et al. Predicting Cross-Reactivity and Antigen Specificity of T Cell Receptors. **Frontiers in Immunology**, v. 11, p. 565096, 22 Oct. 2020.

LIU, X. et al. Systematic comparative evaluation of methods for investigating the tcrβ repertoire. **Plos One**, v. 11, n. 3, p. e0152464, 28 Mar. 2016.

MAYER-BLACKWELL, K.; FIORE-GARTLAND, A.; THOMAS, P. G. Flexible Distance-Based TCR Analysis in Python with tcrdist3. **Methods in Molecular Biology**, v. 2574, p. 309–366, 2022.

MCGINNIS, S.; MADDEN, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. **Nucleic Acids Research**, v. 32, n. Web Server issue, p. W20-5, 1 Jul. 2004.

MEYSMAN, P. et al. Benchmarking solutions to the T-cell receptor epitope prediction problem: IMMREP22 workshop report. **ImmunoInformatics**, p. 100024, Feb. 2023.

MURTAGH, F.; CONTRERAS, P. Algorithms for hierarchical clustering: an overview.

Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, v. 2, n. 1, p. 86–97, Jan. 2012.

NIELSEN, M. et al. Lessons learned from the IMMREP23 TCR-epitope prediction challenge. **ImmunoInformatics**, v. 16, p. 100045, Dec. 2024.

PAI, J. A.; SATPATHY, A. T. High-throughput and single-cell T cell receptor sequencing technologies. **Nature Methods**, v. 18, n. 8, p. 881–892, Aug. 2021.

PEARSON, W. R. Selecting the Right Similarity-Scoring Matrix. **Current Protocols in Bioinformatics**, v. 43, p. 3.5.1-9, 2013.

ROBINS, H. S. et al. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. **Blood**, v. 114, n. 19, p. 4099–4107, 5 Nov. 2009.

ROSSJOHN, J. et al. T cell antigen receptor recognition of antigen-presenting molecules. **Annual Review of Immunology**, v. 33, p. 169–200, 2015.

SHUGAY, M. et al. Towards error-free profiling of immune repertoires. **Nature Methods**, v. 11, n. 6, p. 653–655, Jun. 2014.

SIEVERS, F.; HIGGINS, D. G. Clustal Omega for making accurate alignments of many protein sequences. **Protein Science**, v. 27, n. 1, p. 135–145, Jan. 2018.

SONG, D. et al. Parameterized BLOSUM matrices for protein alignment. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, v. 12, n. 3, p. 686–694, 2015.

SONG, L. et al. TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. **Nature Methods**, v. 18, n. 6, p. 627–630, Jun. 2021.

TOWNSEND, A.; BODMER, H. Antigen recognition by class I-restricted T lymphocytes. **Annual Review of Immunology**, v. 7, p. 601–624, 1989.

VAUGHAN, K.; SETTE, A. The Immune Epitope Database (IEDB): A resource for immunological data related to infectious disease, autoimmunity and allergy. (P3024). **The Journal of Immunology**, v. 190, n. 1_Supplement, p. 114.13-114.13, 1 May 2013.

VITA, R. et al. The immune epitope database (IEDB) 3.0. **Nucleic Acids Research**, v. 43, n. Database issue, p. D405-12, Jan. 2015.

VITA, R. et al. The Immune Epitope Database (IEDB): 2018 update. **Nucleic Acids Research**, v. 47, n. D1, p. D339–D343, 8 Jan. 2019.

VUJOVIC, M. et al. T cell receptor sequence clustering and antigen specificity. **Computational and structural biotechnology journal**, v. 18, p. 2166–2173, 5 Aug. 2020.

WALLACE, I. M.; BLACKSHIELDS, G.; HIGGINS, D. G. Multiple sequence alignments. **Current Opinion in Structural Biology**, v. 15, n. 3, p. 261–266, Jun. 2005.

WANG, C. et al. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. **Proceedings of the National Academy of Sciences of the United States of America**, v. 107, n. 4, p. 1518–1523, 26 Jan. 2010.

WATERMAN, M. S. Efficient sequence alignment algorithms. **Journal of Theoretical Biology**, v. 108, n. 3, p. 333–337, 7 Jun. 1984.