UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE QUÍMICA ORGÂNICA E INORGÂNICA

BACHARELADO EM QUÍMICA – HABILITAÇÃO EM QUÍMICA INDUSTRIAL

GABRIELE GOMES DINIZ

OBTENÇÃO E ANÁLISE *IN SILICO* DE SEQUÊNCIAS DE AMINOÁCIDOS PRESENTES EM ANTICORPOS HUMANOS

FORTALEZA – CE

2021

GABRIELE GOMES DINIZ

OBTENÇÃO E ANÁLISE *IN SILICO* DE SEQUÊNCIAS DE AMINOÁCIDOS PRESENTES EM ANTICORPOS HUMANOS

Trabalho de conclusão de curso apresentado ao Curso em Química do Centro de Ciências da Universidade Federal do Ceará, como requisito à obtenção do grau Bacharel em Química.

Orientador Profissional: Cássio Pinheiro Oliveira

Orientador Pedagógico: Maria Goretti de Vasconcelos Silva

FORTALEZA

2021

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

D611o Diniz, Gabriele Gomes.
Obtenção e análise in silico de sequências de aminoácidos presentes em anticorpos humanos / Gabriele Gomes Diniz. – 2021.
36 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Ciências, Curso de Ciências Biológicas, Fortaleza, 2021.
Orientação: Profa. Dra. Maria Goretti de Vasconcelos Silva.
Coorientação: Prof. Me. Cássio Pinheiro Oliveira.

1. Anticorpos. 2. Bioquímica. 3. CDRs. 4. In silico. 5. NCBI. I. Título.

CDD 570

GABRIELE GOMES DINIZ

OBTENÇÃO E ANÁLISE *IN SILICO* DE SEQUÊNCIAS DE AMINOÁCIDOS PRESENTES EM ANTICORPOS HUMANOS

Monografia apresentada ao Curso de Química da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Bacharel em Química.

Aprovada em: 13/09/2021.

BANCA EXAMINADORA

_____
Prof. Dr. Maria Goretti de Vasconcelos Silva (Orientador)
Universidade Federal do Ceará (UFC)

_____
Ms. Cássio Pinheiro Oliveira (Orientador)
Fundação Oswaldo Cruz (Fiocruz)

_____
Prof. Dr. Gilvan Pessoa Furtado
Fundação Oswaldo Cruz (Fiocruz)

AGRADECIMENTOS

À MC Consultoria pela oportunidade.

Ao Matthias Coelho Batista, por todo o apoio nesse projeto.

Ao Cássio Pinheiro Oliveira, meu supervisor.

Ao Marcos Lourenzoni por ter viabilizado esse trabalho.

Aos meus pais, por me apoiarem em todas as empreitadas.

RESUMO

Os anticorpos são proteínas que tem a função de identificar corpos estranhos e elicitar resposta imunológica. O estudo das características bioquímicas dos anticorpos foi, e continua sendo, essencial para o desenvolvimento de novas tecnologias que envolvam essa classe de proteínas. Devido a sua função, a diversidade de anticorpos é alta, o que representa uma barreira para o estudo de anticorpos em larga escala. Esse tipo de estudo, entretanto, pode ser uma forma de chegar a conhecimentos importantes que abrangem os anticorpos de uma maneira mais geral. Com o avanço da computação e da bioinformática, se tornou possível realizar estudos em larga escala com métodos *in silico*. O presente trabalho busca, então, obter sequências de cadeias leves e pesadas de anticorpos humanos e realizar análises *in silico* das características bioquímicas dessas cadeias obtidas, em larga escala. A obtenção se deu pela interface do NCBI, portal que reúne diversos bancos de dados, incluindo bancos de sequência de aminoácidos de anticorpos. Esta obtenção foi realizada pela ferramenta Entrez, que realiza buscas textuais. As sequências de aminoácidos obtidas foram então filtradas por meio da linguagem Python, e numeradas pelo programa ANARCI, com o esquema de Kabat. Após a numeração, realizou-se análises de comprimentos das CDRs e a partir disso, selecionou-se duas CDRs para seguir com as análises bioquímicas. Estas análises foram feitas por meio da linguagem Python e com uso da biblioteca MatPlotLib, e estas análises consistiam de frequência de aminoácido por posição, caráter químico por posição, e GRAVY por posição. Obteve-se cerca de 52 mil cadeias leves e 36 mil cadeias pesadas. Observou-se que algumas posições dentro das CDRs estudadas apresentavam frequência maior de certos tipos de aminoácidos, e/ou frequência maior de certo caráter químico, o que indica a possibilidade que o resíduo ou sua natureza química naquela posição tem algum papel na função ou estabilidade dos anticorpos humanos. Observou-se também que não há uma correlação forte entre o GRAVY por posição e o caráter químico por posição. De modo geral, conseguiu-se mapear alguns padrões dentre as sequências de aminoácidos estudadas, o que pode fornecer dados importantes para tecnologias que trabalhem com humanização, enxertos de CDRs e mutagênese de anticorpos humanos.

Palavras-chave: Anticorpos. Bioquímica. CDRs. *In silico*. NCBI.

ABSTRACT

Antibodies are proteins that identify antigens and elicit immune response. The study of antibodies' biochemical characteristics was, and still is, essential to the development of new technologies that employ such proteins. Due to their function, there is a great diversity of antibodies, which represents a hurdle to the study of antibodies in large scale. This type of study, however, may be a way of arriving to important insights that relate to antibodies as a whole. With the advancement of informatics and bioinformatics, it became possible to do such studies in large scale through *in silico* methods. This work seeks to obtain amino acid sequences of light and heavy chains of human antibodies and run in silico analysis of the biochemical characteristics of those sequences. The obtainment of the sequences was through the interface of NCBI, a website that unites several data bases, including aminoacid sequence data banks. The "Entrez" tool, which does textual searches, was used for such obtainment. The obtained sequences were then selected through scripts in Python language and numbered by the software ANARCI according to the Kabat numbering scheme. After that, an analysis of the CDRs lengths was done, and through the results two CDRs were selected to be further analyzed. The biochemical analysis were made through the Python language and the graphs were plotted with the MatPlotLib library; the analyses were: amino acid per position, chemical character per position and GRAVY per position. Around 52 thousand light chain sequences and 36 thousand heavy chain sequences were obtained. Some CDRs positions displayed tendencies in either amino acid type per position and/or chemical character per position, which indicates the possibility that those positions played a role in either function or stability of human antibodies. There didn't seem to be any correlation between the GRAVY per position and the dominant chemical character per position. All in all, it was possible to map certain trends within the studied amino acid sequences, which may provide important data to technologies that deal with humanization, CDR grafting and mutagenesis of human antibodies.

Key words: Antibodies. Biochemistry. CDRs. *In silico*. NCBI.

LISTA DE ILUSTRAÇÕES

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BLAST	Basic Alignment Search Tool
CDR	Região determinante de complementariedade
Fab	Fragmento de ligação com o antígeno
Fc	Fragmento cristalizável
NCBI	Centro Nacional de Informação Biotecnológica
PDB	Protein Data Bank
PIR	Protein Information Resource
PRF	Protein Research Foundation
RefSeq	Reference Sequence Database
WGS	Whole Genome Sequencing

LISTA DE AMINOÁCIDOS

A	Alanina
C	Cisteína
D	Ácido Aspártico
E	Ácido Glutâmico
F	Fenilalanina
G	Glicina
H	Histidina
I	Isoleucina
K	Lisina
L	Leucina
M	Metionina
N	Asparagina
P	Prolina
Q	Glutamina
R	Arginina
S	Serina
T	Treonina
V	Valina
W	Triptofano
Y	Tirosina

SUMÁRIO

1. INTRODUÇÃO

Compreender a bioquímica dos anticorpos é um conhecimento-chave para o desenvolvimento de novas tecnologias na área da imunologia. Os experimentos iniciais com anticorpos se deram por meio de eletroforese, digestão com enzimas e ultracentrifugação (PUNT *et al*, 2018) e eles nos permitiram compreender a sua massa e estruturas básicas.

Entretanto, os anticorpos são variados e sua diversidade torna difícil a tarefa de compreendê-los por completo. O estudo *in silico* de anticorpos tornou-se uma possibilidade cada vez mais atraente com o avanço da computação e da bioinformática.

O desenvolvimento e aperfeiçoamento das plataformas de sequenciamento de ácidos nucleicos e de técnicas como difração de raio X e Ressonância Magnética Nuclear, favoreceram a criação e expansão de diversos bancos de dados biológicos. O grande montante de sequências e estruturas de proteínas obtidas levou também à criação de ferramentas de bioinformática voltadas para a análise desses dados.

Atualmente, é possível obter informações de milhares de anticorpos de um único banco de dados público. Além disso, a diminuição do custo de poder computacional tem viabilizado análises de dados em larga escala.

Diversos grupos já trabalham com essa possibilidade, de uma forma ou de outra. Silva (2016) utilizou métodos *in silico* para analisar sequências de aminoácidos de imunoglobulinas geradas por *phage display*. Rodrigues (2014) estudou a estrutura tridimensional de cerca de 3000 proteínas utilizando estruturas obtidas do abYsis, um banco de dados de proteínas. Carvalho (2019) utilizou a plataforma Yves para se debruçar sobre a estrutura e natureza bioquímica de anticorpos contra HIV-1.

A análise de anticorpos em larga escala pode fornecer informações cruciais para o melhor desenvolvimento de tecnologias que os utilizem, como por exemplo enxertos de CDRs, mutagênese e humanização. Análises bioquímicas de anticorpos em larga escala oferece a mesma possibilidade, e realizar essas análises *in silico* elimina algumas das limitações presentes em experimentos de bancada, como utilização de insumos e dispêndio de muita mão-de-obra humana.

O presente trabalho busca obter sequências de aminoácidos de anticorpos humanos e então analisá-las do ponto de vista bioquímico, utilizando bancos de dados e linguagem de programação para realizar essa obtenção e análise em larga escala, com milhares de sequências de aminoácidos.

2. REVISÃO BIBLIOGRÁFICA

2.1 Imunoglobulinas: definição

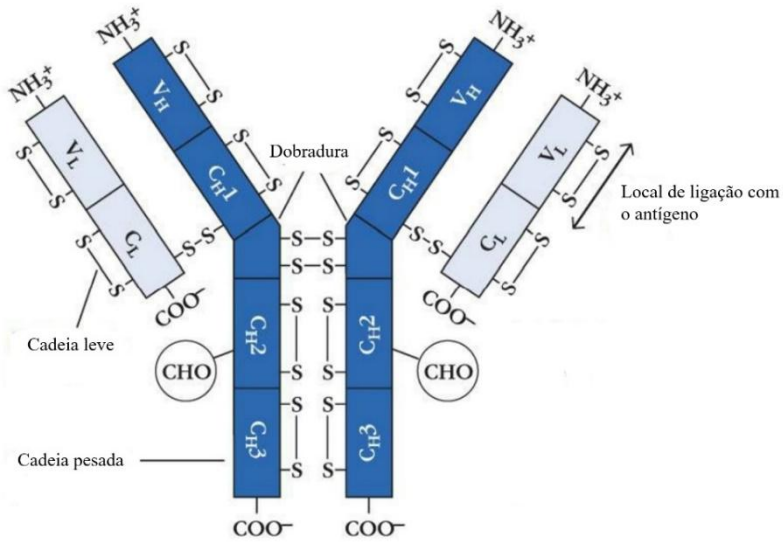
As imunoglobulinas, também chamadas de anticorpos, são dímeros de proteínas encontradas em diversos organismos que tem como função identificar antígenos e elicitar resposta imunológica. Os primeiros experimentos com anticorpos datam de 1890, quando dois cientistas, von Behring e Kitasato, utilizaram sérum do sangue de coelhos imunizados contra tétano em ratos, que foram posteriormente inoculados com bactérias de tétano. O grupo de controle do experimento morreu, mas os ratos tratados, além de sobreviver, não demonstraram sinais de infecção. Esse estudo demonstrou que existem substâncias no sangue capazes de prevenir infecções, e que estas substâncias podem ser transferidas de um corpo para o outro. Essa descoberta levou von Behring a ganhar o Prêmio Nobel de Medicina e Fisiologia em 1901. Até os dias de hoje, esse é um princípio utilizado para diversos tratamentos modernos. Soros antiofídicos, por exemplo, contém imunoglobulinas específicas para veneno de cobras (PUNT et al, 2018).

Ao longo das décadas seguintes, diversos estudos têm sido feitos para identificar os agentes da chamada “imunidade humoral”. Na década de 1930, Elvin Kabat identificou que as chamadas gamma globulinas eram as responsáveis pelo fenômeno da imunidade passiva adquirida, e o termo eventualmente evoluiu para imunoglobulina. Foi também Kabat, juntamente com Tiselius, Pedersen e Heidelberger, que desvendou a natureza química desse grupo de moléculas orgânicas, utilizando eletroforese. (BLACK, 1997).

2.2 Imunoglobulinas: estrutura e CDRs

As imunoglobulinas possuem estruturas muito diversas, mas algo em comum sobre todas as elas é que são constituídas de uma unidade básica, dois pares de cadeia, sendo cada par constituído de uma cadeia leve e uma pesada, ligados entre si por pontes dissulfeto. Alguns anticorpos têm apenas uma unidade, como os lgG e os lgD. Outros ainda possuem duas, como os lgA, ou cinco, como os lgM (SCHROEDER, CAVACINI, 2010).

Figura 1 - Estrutura básica de uma imunoglobulina lgG.



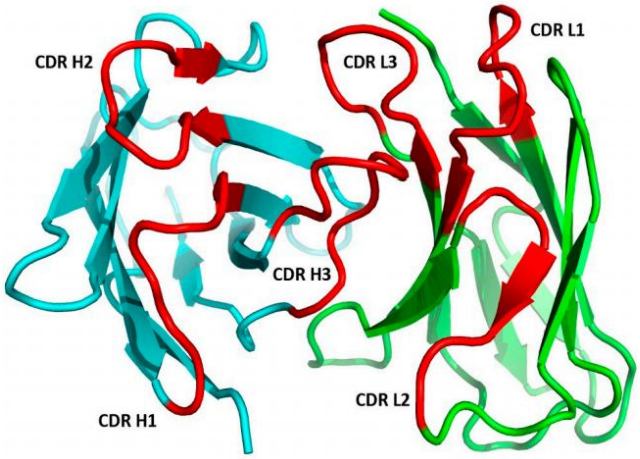
Fonte: Adaptada de Punt *et al* (2018).

A estrutura dos anticorpos pode ser dividida entre dois tipos de fragmentos: os cristalizáveis (Fc) e os de ligação com antígeno (Fab, do inglês *antigen binding fragment*). Esses nomes surgiram por causa do processo de descoberta da estrutura, que foi reconstruída a partir da digestão da molécula com papaína e pepsina. Os fragmentos cristalizáveis espontaneamente formavam cristais, e os de ligação com o antígeno demonstraram ter a mesma capacidade do anticorpo de se ligar ao antígeno, mesmo separados do restante das estruturas. (PUNT *et al*, 2018).

A unidade básica da estrutura de um anticorpo também é separada em domínios constantes (C_H, para as cadeias pesadas e C_L, para as leves) e variáveis (V_H, para as cadeias pesadas e V_L, para as leves). Os domínios constantes apresentam baixa variação da sua sequência de aminoácidos entre os diferentes tipos de anticorpos, mas os variáveis, como fica explícito no nome, apresentam alta variabilidade. Apesar da variação da estrutura das imunoglobulinas, esses domínios sempre estão presentes. (PUNT *et al*, 2018).

Dentro dos domínios variáveis, existem regiões de hipervariabilidade, tanto nas cadeias leves quanto nas pesadas. Essas regiões, chamadas de CDRs (sigla para *complementarity-determining region*, em inglês), começaram a ser estudadas na década de 1970, por Wu e Kabat. Existem três CDRs para cada tipo de cadeia (H1, H2, H3 para as cadeias pesadas, e L1, L2 e L3 para as leves). Estruturalmente, as CDRs estão em loops que ligam duas folhas betas, uma sobre a outra; as regiões dessas folhas fora das CDRs e loops não-hipervariáveis são conhecidas como frameworks. (CHIU *et al*, 2019).

Figura 2 – Exemplo de loops de CDRs, que estão em vermelho. A cadeia pesada está em azul e a leve está em verde.



Fonte: Chiu *et al* (2019).

As CDRs podem ser determinadas a partir de uma sequência de aminoácidos por diversos métodos, e os métodos de numeração de Kabat e Chothia foram os dois primeiros a realizar essa determinação. A numeração de Kabat, porém, foi idealizada a partir das sequências de aminoácidos, enquanto Chothia corrigiu a numeração de Kabat para maior acurácia estrutural, de modo que as CDRs correspondessem aos loops da estrutura dos anticorpos (DUNBAR, DEANE, 2015).

O estudo das CDRs de anticorpos é essencial para compreender como elas definem a capacidade de um anticorpo de se ligar a um antígeno (CHIU *et al*, 2019).

2.3 NCBI

O avanço da tecnologia computacional não gerou apenas numa maior capacidade de compreender os diversos fragmentos das estruturas de anticorpos: ela também se tornou aliada na tarefa de armazenar e catalogar sequências de aminoácidos e estruturas, por meio da bioinformática, que pode ser definida como o uso de computadores para a aquisição, manejo e análise de informação biológica (FRANCO *et al*, 2008 *apud* BROWN, 2000). No âmbito da bioinformática, surgiram as bases de dados, que, dentre outras funções, armazenam sequências de aminoácidos pertencentes a proteínas de maneira geral, entre elas enzimas e anticorpos.

Dentre as bases de dados que armazenam informações sobre proteínas (e, consequentemente, sobre anticorpos), uma de vital importância é o NCBI. O NCBI, ou Centro Nacional de Informação Biotecnológica, é um site que reúne o conteúdo de cerca de 40 bases de

dados, que variam de bases de artigos até biologia molecular, sendo estas bases divididas em sete categorias: literatura, genomas, variação, saúde, genes e expressão de genes, nucleotídeos, proteínas e pequenas moléculas.

Na categoria proteínas, o NCBI fornece uma base de dados sobre proteínas que retira informações de diversas fontes, como GenBank, o projeto de Sequência e Referência do NCBI (o RefSeq), e também fontes externas, como o UniProtKB/SWISS-Prot, produto da colaboração entre o Instituto Europeu de Bioinformática, o Instituto Suíço de Bioinformática e o Protein Information Resourse (PIR), além do Protein Data Bank (PDB). Porém, é importante reiterar que a maioria dos bancos utilizados traz traduções teóricas de sequências de RNA, a única exceção sendo o PDB. (SAYERS, 2013).

O NCBI tem como principais vantagens ser gratuito, oferecer ferramentas diversas e realizar pesquisas em diversas bases de dados simultaneamente. Além do acesso a essas bases de dados, o NCBI conta também com ferramentas como o BLAST.

O BLAST é uma ferramenta multi-uso cuja principal função é buscar sequências por similaridades. Essas similaridades podem ser entre nucleotídeos, ou entre sequências de aminoácidos. Também é possível, com o BLAST, procurar a semelhança entre uma sequência de nucleotídeos e sequências de aminoácidos, com o BLAST realizando a tradução de nucleotídeos para aminoácidos e vice-versa (MADDEN, 2013).

A ferramenta BlastP é o Blast *Protein-to-Protein*, ou seja, que busca sequências de aminoácidos similares a partir de uma sequência de aminoácido; é possível realizar a busca em diversos bancos de dados simultaneamente, ou apenas em bancos de dados específicos.

Tabela 1 - Bases de dados utilizadas pelo BlastP.

BASES DE DADOS CONSULTADAS PELO BLASTP	
Sigla	Descrição
nr	Todas as traduções não redundantes do GenBank CDS + PDB + SwissProt + PIR + PRF, excluindo as amostras ambientais de projetos de WGS (Whole Genome Sequencing)
refseq_select	NCBI RefSeq sequências de aminoácidos de humanos, ratos e procariontes, restrito aos conjuntos de proteínas do RefSeq Select.
refseq_protein	Sequências de proteínas de referência do NCBI
landmark	Base de dados de referência para o SmartBLAST, incluindo proteomas de 27 genomas em uma grande variedade taxonômica
swissprot	Sequências não-redundantes do UnitProtKB/SwissProt
pataa	Sequências de aminoácidos da divisão de Patente da GenBank
pdb	Sequências do Protein Data Bank (PDB)
env_r	Proteínas de projetos metagenômicos WGS

tsa_nr	Proteínas do Transcriptome Shotgun Assembly

Fonte: BlastP Suite. Disponível em: <https://blast.ncbi.nlm.nih.gov/Blast.cgi>. Acesso em: 25 de julho de 2021.

Entretanto, sequências de aminoácidos também podem ser obtidas pelo Entrez, a ferramenta de busca textual que funciona em todo o site do NCBI. Por meio dela, é possível buscar todo o tipo de conteúdo que o site oferece, incluindo as sequências de aminoácidos, quando se usa o Entrez na categoria “Protein” (OSTELL, 2002).

3. OBJETIVOS

3.1 Gerais

Obter e analisar fatores bioquímicos em larga escala de sequências de aminoácidos de cadeias variáveis de anticorpos

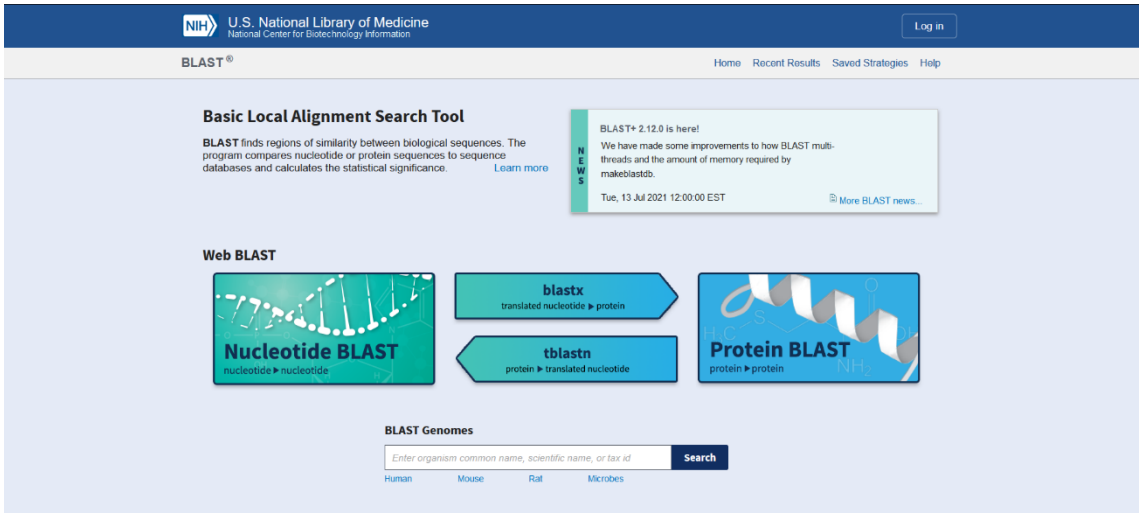
3.2 Específicos

- Obter sequências de aminoácidos dos bancos de dados do NCBI
- Filtrar, de forma automática, as sequências obtidas
- Analisar fatores bioquímicos das sequências de aminoácidos

4. METODOLOGIA

4.1 Obtenção de sequências de domínios variáveis utilizando o BlastP do NCBI

Figura 3 – Página da ferramenta BLAST do NCBI.



Fonte: BLAST Suite. Disponível em: < https://blast.ncbi.nlm.nih.gov/Blast.cgi> Acesso em 02 de agosto de 2021.

As sequências escolhidas para usar como isca foram retiradas manualmente do banco IgBlast, na categoria de germlines humanos. Essas sequências foram classificadas filogeneticamente por meio do Clustal Omega (SIEVERS, HIGGINS, 2014), disponível em https://www.ebi.ac.uk/Tools/msa/clustalo/. Para garantir variedade e relevância dos resultados, selecionou-se as sequências mais distantes da árvore filogenética, e dessas sequências mais distantes, foram selecionadas as mais longas e íntegras.

Para iniciar a busca fazer a retirada dos resultados, empregou-se o Biopython, biblioteca da linguagem Python livre e gratuita voltada para a bioinformática (COCK *et al*, 2009), e uma série de scripts desenvolvidos *in house*. Realizou-se um BLAST no banco nr com as sequências de busca, executando um script blast_search.py, resultando em um documento no formato .xml. Em seguida, removeu-se o elemento “CREATE VIEW” deste documento, que dificultava as etapas posteriores do processamento. A partir do documento .xml gerou-se um arquivo .csv formatado com apenas as informações relevantes por meio do script parse.py. Com o arquivo .csv, gerou-se um arquivo FASTA com o script to_fasta.py.

4.2 Numeração usando o ANARCI

Após obtenção e tratamento, as sequências são filtradas por meio de numeração via ANARCI (DUNBAR, DEANE, 2015), um programa que numerar sequências de aminoácidos relacionadas a anticorpos e células T, disponível em

http://opig.stats.ox.ac.uk/webapps/newsabdab/sabpred/anarci/. O ANARCI é capaz de realizar numerações seguindo os cincos esquemas mais utilizados de numeração de resíduos de anticorpos: Kabat, Chotia, Enhanced Chotia, IMGT e AHo. (DUNBAR, DEANE, 2015). Para o presente trabalho, se utilizou o esquema de numeração de Kabat, visto que o foco está apenas nas sequências de aminoácidos, e não em suas estruturas tridimensionais. Executou-se o ANARCI em terminal Linux, com formato de saída em .csv.

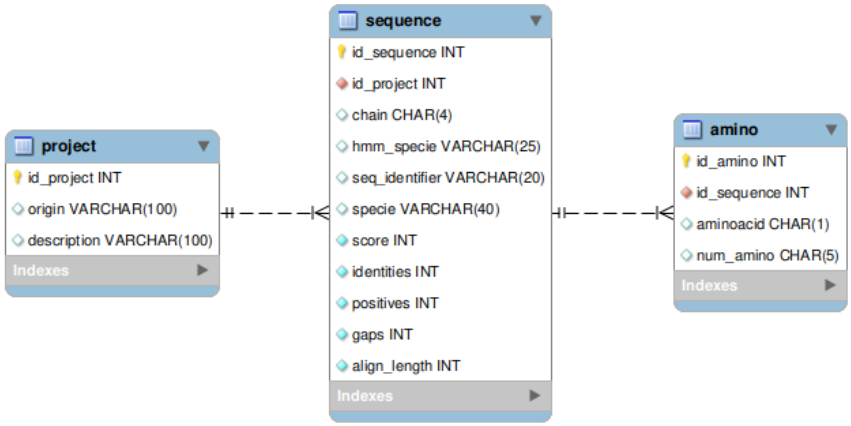
Essa numeração foi realizada para descartar sequências baseado em dois critérios. Caso o ANARCI não tenha conseguido realizar a numeração da sequência, isso é indicativo de que a sequência está fragmentada demais, ou que não é uma sequência de aminoácidos de anticorpo, e por isso deve ser descartada. Uma sequência obtida também será descartada se o ANARCI detectar que ela é de uma cadeia diferente da cadeia da sequência de busca; por exemplo, se a sequência de busca for de cadeia leve, mas a sequência obtida for, segundo o ANARCI, de cadeia pesada.

4.3 Inserção das sequências no banco de dados

As sequências que não foram descartadas no processo de numeração seguem para um banco de dados, gerenciado pelo MySQL. Criou-se 3 tabelas: project, sequence e amino. A tabela project armazena os dados gerais dos conjuntos de sequências obtidos a cada busca, enquanto a tabela sequence lida com os metadados das sequências. A tabela amino armazena os aminoácidos das sequências.

A inserção das sequências no banco de dados foi realizada pelo script db_insert.py. Das sequências com mesmo *seq_identifier* e mesma cadeia apenas uma foi mantida.

Figura 4 – Esquematização do banco de dados para as sequências obtidas via BlastP.



Fonte: Próprio autor.

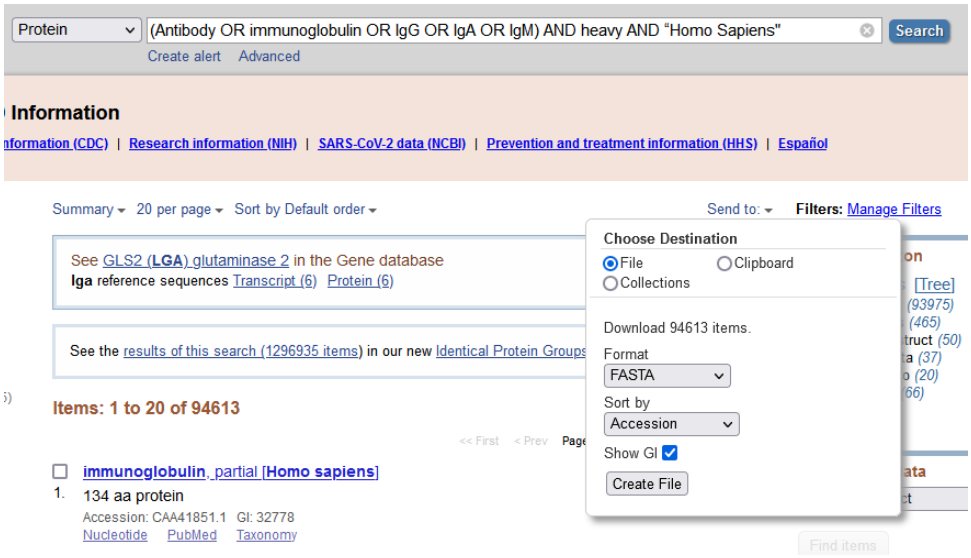
Ao fazer as primeiras análises dos dados obtidos, percebeu-se que os germlines humanos usados como sequência de busca não apresentam a CDR H3 completa, e, portanto, as sequências obtidas apresentavam a mesma característica. Percebeu-se que a busca por sequências de anticorpos via BLAST apresentou duas limitações cruciais: os resultados não apresentavam sequências maiores que a sequência isca, apenas sequências menores (Query Coverage de 100% para baixo), e os resultados mais distantes por similaridade eram sequências incompletas, fragmentadas, que eram descartadas na fase de numeração, e implicavam em um maior trabalho computacional de tentar numerar sequências que não podiam ser numeradas.

Como forma de resolver superar as limitações da busca via BLAST, optou-se por buscar as sequências utilizando o Entrez, a ferramenta de busca textual do NCBI.

4.4 Obtenção de sequências de aminoácidos de anticorpos por meio do Entrez

Não foi necessário utilizar um script para a obtenção das sequências via Entrez. A pesquisa textual foi feita aplicando os operadores booleanos da ferramenta, que permitem a combinação ou a exclusão de termos na hora da busca. Executou-se uma busca com os seguintes termos: "(Antibody OR immunoglobulin OR lgG OR lgA OR lgM) AND heavy AND "Homo Sapiens" e também uma busca "(Antibody OR immunoglobulin OR lgG OR lgA OR lgM) AND light AND "Homo Sapiens". A primeira busca teve o objetivo de captar sequências de cadeia pesada, e a segunda teve o objetivo de captar sequências de cadeia leve. Em seguida, todos os resultados foram convertidos em um único arquivo FASTA por meio da interface, selecionando a opção Send to, e File, e por fim, escolhendo o formato FASTA e organizando os resultados por Accession.

Figura 5 – Conversão dos resultados obtidos em um arquivo pela interface do NCBI.



Fonte: NCBI. Acesso em: 10 de agosto de 2021.

4.5 Seleção das sequências obtidas via Entrez

Considerando que uma fração considerável das sequências obtidas não puderam ser numeradas na tentativa anterior de obtenção, realizou-se uma etapa de seleção das sequências utilizando um script em Python chamado filter.py.

Esse script mede a distância entre dois resíduos de cisteína, presentes ao início da CDR1 e ao final da FR3, conceito utilizado no trabalho desenvolvido por Silva (2016); Silva mediu as distâncias entre os resíduos de cisteína das cadeias leves e pesadas em um script na linguagem Perl. Para anticorpos humanos, essa distância entre os dois resíduos de cisteína possui valor máximo e mínimo, conforme determinado por estudos da UCL (ABHINANDAN, MARTIN, 2008): para as cadeias pesadas a distância varia entre 51 e 84 resíduos, e para as leves, varia entre 56 e 85 resíduos. Esses parâmetros foram aplicados pelo filter.py para descartar sequências que não obedecessem a esse padrão.

Também como desenvolvido pelo trabalho de Silva, o filter.py detectou a existência da sequência de resíduos "WGXG" para V_H e "FGXG" para V_L, onde X representa qualquer aminoácido, localizados ao fim da CDR3, e assim o script mediu o tamanho da CDR3. Conforme também estabelecido pela UCL, os tamanhos das CDRs variam entre 5 a 30 resíduos para cadeia pesadas, e entre 5 a 15 resíduos para as cadeias leves, e estes foram os parâmetros utilizados pelo script. Assim, selecionou-se as sequências de aminoácidos que de fato devem representar anticorpos.

4.6 Numeração das sequências obtidas via Entrez

Este passo se deu da mesma maneira do que foi executado para a obtenção de sequências via BlastP, utilizando o ANARCI e numerando as sequências de acordo com o esquema de Kabat.

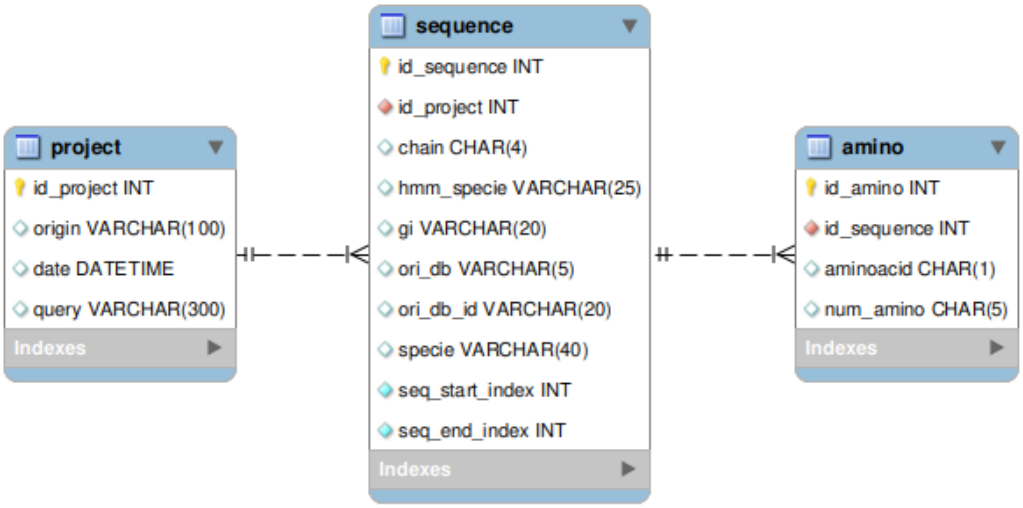
4.7 Inserção das sequências obtidas via Entrez no banco de dados

Novamente, esta inserção ocorreu de maneira análoga à inserção das sequências obtidas via BlastP. A diferença entre o banco de dados para as sequências do BlastP e as do Entrez está nas tabelas project e sequence.

Na tabela project, temos origem do volume de dados (origin), data de entrada (date) e termos usados na busca de obtenção (query), enquanto no banco de dados para as sequências do BlastP possuía banco de origem das sequências (origin) e sequência de aminoácidos usados como isca (description).

A tabela sequence do Entrez possui campo para a origem da sequência (ori_db), identificador da sequência no banco de origem (ori_db_id) e a posição de início e fim do alinhamento feito pelo ANARCI (seq_start_index e seq_end_index). Entretanto, não contém os campos referentes dados de alinhamento presentes na tabela sequence do banco do dados para as sequências retiradas do BlastP (score, identities, positives, gaps, align_length).

Figura 6 – Esquematização do banco de dados para as sequências obtidas via Entrez.



Fonte: Próprio autor.

4.8 Análises: classificação de CDRs, aminoácido por posição da CDR, caracterização química por posição e GRAVY.

As análises foram realizadas por meio de busca nesse banco de dados construído com as sequências obtidas pelo Entrez. Os dados de resposta foram tratados e convertidos em gráficos por meio da linguagem Python e da biblioteca MatPlot Lib.

5. RESULTADOS E DISCUSSÃO

5.1 Sequências obtidas via BlastP

Tabela 2 – Sequências de cadeia leve e cadeia pesada obtidas via BlastP

	Sequências Numeradas
V_H	**140149**
V_L	**64865**

Fonte: Próprio autor.

O banco nr do BlastP possui mais de 400 milhões de sequências, como indicado no site do BLAST, mas o número de sequências obtidas pelo script foi consideravelmente menor do que isso. Em primeiro lugar, há a questão de que, dessas 400 milhões de sequências, apenas uma fração não determinada é de anticorpos, pois o banco é de proteínas de uma maneira geral. Em segundo lugar, a resposta do BlastP atingia um limite máximo, mesmo que, na teoria, os resultados de uma busca por similaridade possam ser virtualmente infinitas, indo da sequência mais similar para a menos similar no banco de dados. Outra questão a se observar aqui é que o número de sequências numeradas é menor do que o número de sequências obtidas no total, pois uma fração do que foi obtido diretamente consistia em sequências fragmentadas que não podiam ser numeradas pelo ANARCI e, portanto, não podiam ser reconhecidas como anticorpos.

Porém, o que se observou nessas sequências obtidas por BlastP é que elas possuíam a CDR H3 incompletas, assim como os germlines utilizados como sequência de busca. Por isso, essas sequências obtidas não são as ideais para o presente trabalho. Seria possível a utilização de sequências com a CDR H3 completa; entretanto, a grande variabilidade da CDR H3 limitaria a representatividade dos dados obtidos. E como já foi exposto na metodologia, a busca por meio de BlastP apresentou certas limitações práticas que resultaram no uso de outra metodologia para a obtenção das sequências dos bancos do NCBI.

5.2 Sequências obtidas via Entrez

Tabela 3 – Sequências obtidas pelo Entrez após filtagrem

	Sequências após Filtragem
V_H	**56407**
V_L	**32853**

Fonte: Próprio autor.

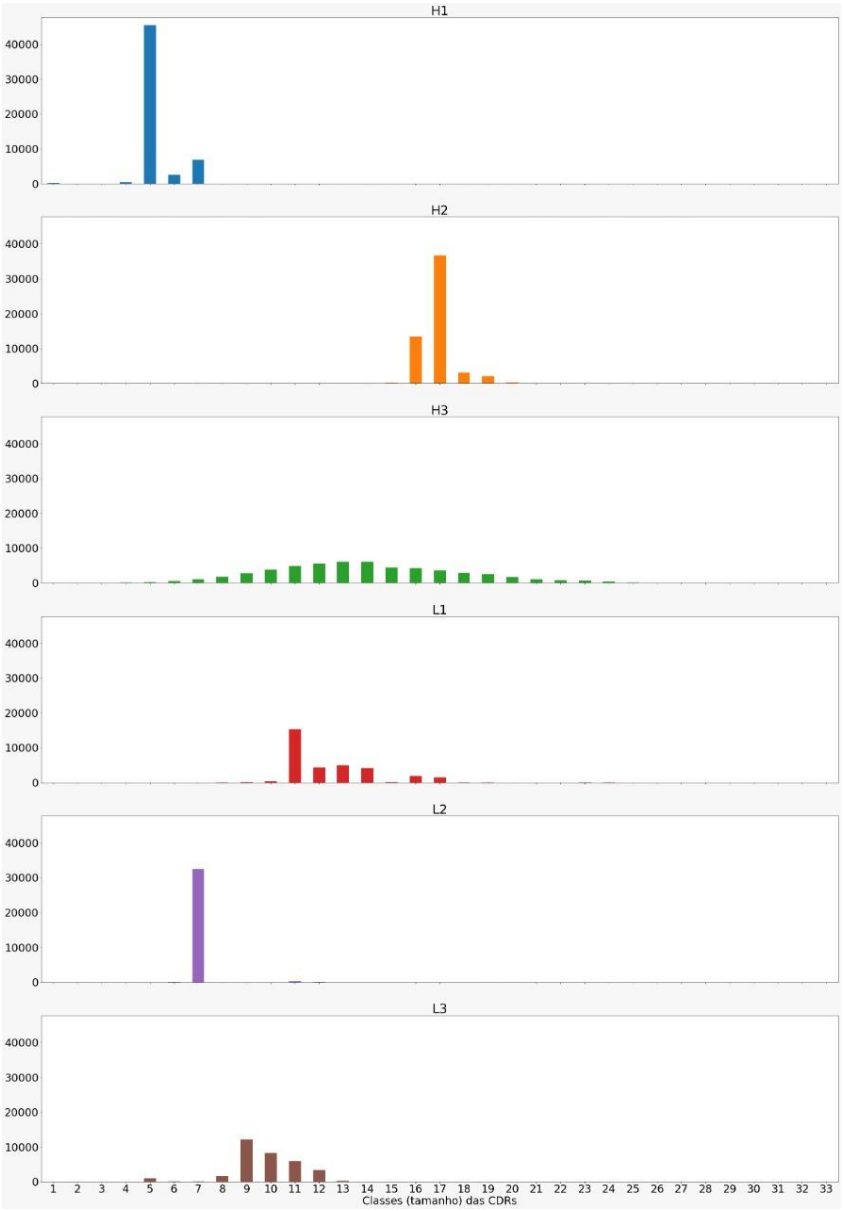
Assim como no caso do BlastP, as sequências obtidas de fato por meio do Entrez representam um número maior que as sequências utilizáveis para análises, pois buscou-se filtrar as sequências incompletas e que não possuíssem os resíduos canônicos. A obtenção feita pelo Entrez

já limitava os resultados da busca aos anticorpos humanos (ou que fossem textualmente classificados como tal), e por essa razão o número de sequências obtidas foi menor que a do BlastP.

De toda forma, o NCBI busca em vários bancos de dados e não há uma nomenclatura sistematizada para os anticorpos registrados nesses bancos, tornando necessário um estudo maior de como as sequências são cadastradas textualmente para otimizar o processo de obtenção.

5.3 Classificação das CDRs das sequências obtidas via Entrez

Gráfico 1 – Classificação das CDRs das sequências obtidas por tamanho. O eixo horizontal representa os tamanhos da CDRs, e o eixo vertical representa a quantidade de sequências para a classe (tamanho).



Fonte: Próprio autor.

A classificação das CDRs das sequências obtidas pelo Entrez demonstra que as etapas de filtragem e numeração foram bem-sucedidas: o perfil da distribuição de comprimento da CDRs é

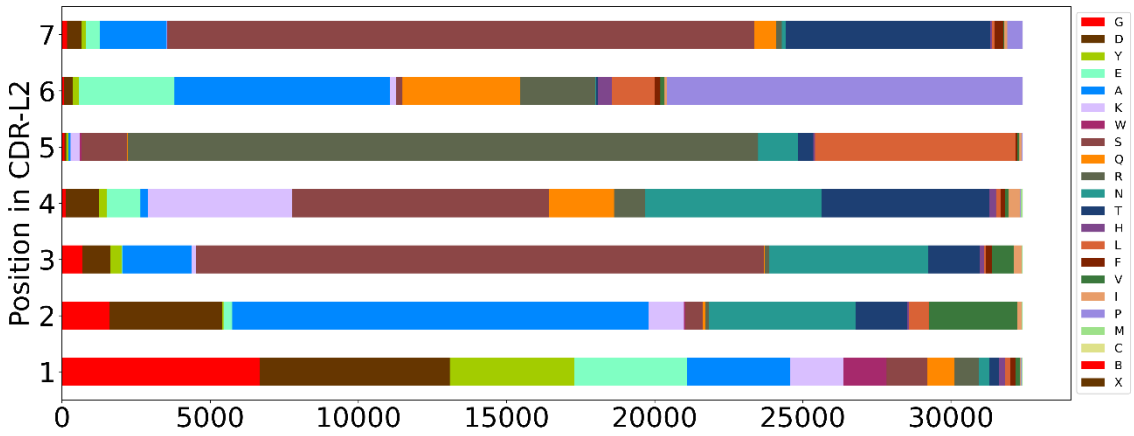
bastante similar ao encontrado em trabalhos anteriores, como o desenvolvido por Wu, Johnson e Kabat (1993), que classificou por tamanho as CDR H3 de 177 sequências de aminoácidos humanas. A escala do presente trabalho é consideravelmente maior, mas o perfil se manteve, o que indica que dentre os anticorpos humanos, existe uma tendência de distribuição por tamanho da CDR H3, que pode ser visualizada em análises feitas em diversas escalas.

Alguns resultados apresentados por Rodrigues (2014) foram similares quanto a algumas outras CDRs, como a L1, L2 e a L3, mesmo considerando que a numeração foi feita segundo o esquema de Chothia, utilizado por este esquema ser mais correto do ponto de vista estrutural, visto que o trabalho se debruça sobre as estruturas tridimensionais dos anticorpos. O número total de sequências estudadas por Rodrigues foi entre 2500 e 3000.

A partir da Figura 10 é possível determinar quais das CDRs seriam as mais proveitosas para realizar as demais análises do trabalho. Uma ampla maioria das sequências de VL obtidas possuíam a CDR L2 de classe 7, assim como uma grande parte das sequências de VH possuíam a CDR H1 de classe 5, tornando essas CDRs estratégicas para realizar análises que sejam representativas de uma boa parte das sequências de cadeias leves e pesadas obtidas.

5.4 Aminoácido por posição da CDR L2

Gráfico 2 – Distribuição de aminoácido por posição da CDR L2. A barra horizontal representa a totalidade das sequências, e trechos de cada cor representam a fração ocupada pelo tipo de aminoácido associado a cor.



Fonte: Próprio autor.

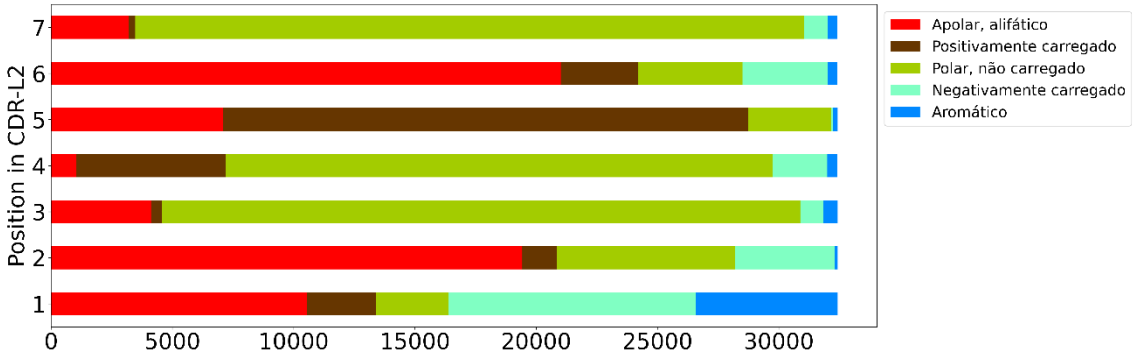
As CDRs são as regiões de um anticorpo com maior variabilidade nos aminoácidos, mas uma análise em larga escala evidencia a abundância de determinados aminoácidos em algumas posições. Por exemplo, a abundância de ácido aspártico (de símbolo D) na posição 7 e 3, e a de arginina na posição 5, o que indica que a natureza dos aminoácidos nessas determinadas posições é importante para a função ou estabilidade dos anticorpos. Kabat e Wu (1977) teorizaram que

existem, dentro das CDRs, resíduos com função estrutural e que por isso podem ser vistos com mais frequência em certas posições dentro da CDR.

A variabilidade maior nas posições 1, 2 e 6 indica que os aminoácidos nessas posições terão algum papel na especificidade do anticorpo, algo também teorizado por Kabat e Wu (1977), o que explicaria não haver uma predominância evidente de nenhum aminoácido, embora alguns sejam mais recorrentes do que outros.

5.5 Caracterização Química da CDR L2 de classe 7

Gráfico 3 – Distribuição de classe de aminoácido por posição da CDR L2. A barra horizontal representa a totalidade das sequências, e trechos de cada cor representam a fração ocupada na posição pelo caráter químico associado a cor.



Fonte: Próprio autor.

Por meio dessa análise das cargas e natureza dos resíduos em cada posição da CDR L2 de classe 7, é possível perceber algumas tendências. O resíduo na posição 7 para a maior parte das sequências possui natureza polar, apesar de não ser carregado. O resíduo nas posições 2 e 6 tende a ser não-polar, alifático. Na posição 5, o que predomina são resíduos positivamente carregados, enquanto nas posições 3 e 4 a predominância é de resíduos polares. Na posição 1, não há uma tendência clara sobre a natureza do resíduo.

Observando essas tendências, é importante ressaltar o papel das interações intramoleculares e ligações fracas para as conformações de proteínas de maneira geral, para qual os anticorpos não são exceção (BERKOWITZ, HOUDE, 2014). Uma grande maioria dos resíduos na posição 7 serem polares apontam para a possibilidade de que a polaridade do resíduo nesse local é importante para os anticorpos humanos de uma maneira mais geral, e a mesma observação vale para as outras tendências observadas nessa análise. Sobre a posição 7, é digno de nota que ela pode ser considerada framework, a depender do sistema de numeração utilizado para classificar os resíduos.

Essa análise também é interessante pois evidencia que pode haver predominância de carga ou natureza numa posição, mesmo que não haja predominância de um aminoácido em

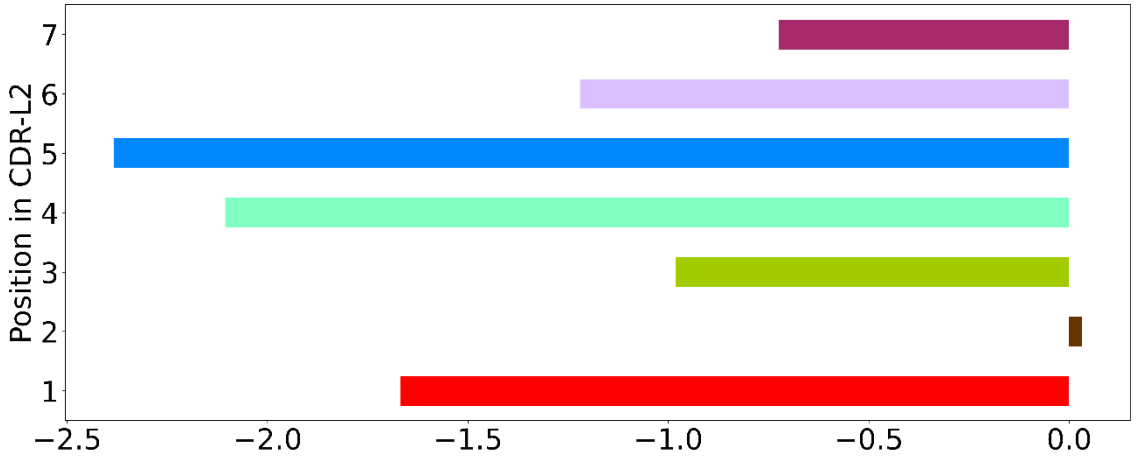
específico. Na posição 4, aminoácidos polares predominam, mas não há predominância de nenhum aminoácido, como visto na análise anterior; algo semelhante ocorre na posição 6. Esta compreensão é importante quando estamos falando de processos que necessitam a substituição de aminoácidos em determinadas posições, , como, por exemplo, síntese de anticorpos biespecíficos por meio de mutagênese das CDRs (CHOI *et al*, 2013).

Apenas na posição 1 é que não há predominância de nenhuma natureza, o que reforça a possibilidade dessa posição ter papel na especificidade dos anticorpos de maneira mais generalizada.

Outra observação a ser feita sobre essa análise é que o caráter aromático aparece com baixa frequência em todas as posições desta CDR; isso aponta para a possibilidade de que a substituição de qualquer um dos aminoácidos por um aminoácido de caráter aromático precise ser analisada com cuidado, pois eles não ocorrem naturalmente com muita frequência na CDR L2.

5.6 Média Geral de Hidropatia (GRAVY) Por Posição da CDR L2

Gráfico 4 – GRAVY por posição da CDR L2. O eixo horizontal representa o valor de GRAVY, e o vertical, a posição.



Fonte: Próprio autor.

A predição da estrutura de um anticorpo por meio apenas da sequência de aminoácidos é, como se poderia imaginar, limitada, mas o estudo do caráter hidrofóbico ou hidrofílico dos diferentes segmentos de uma sequência de aminoácidos pode ajudar na distinção de quais resíduos são internos ou externos (considerando-se um solvente aquoso), e na determinação de quais resíduos estariam envolvidos na interação com membranas fosfolipídicas (KYTE, DOLITTLE, 1982). Quanto mais negativo o GRAVY (Great Average of Hydropathy), mais hidrofílico é o resíduo ou a proteína em questão.

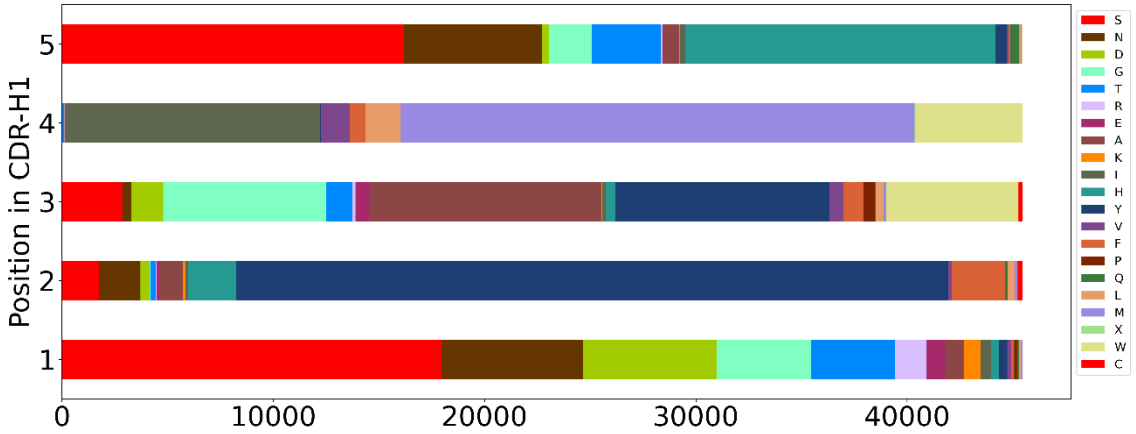
Com exceção da posição 2, todas as demais posições apresentaram GRAVY negativo, ou seja, apresentam caráter hidrofílico. Considerando que estamos analisando uma CDR, esse resultado está dentro do esperado.

Contudo, o caráter hidrofóbico da posição 2 indica que, dentro da estrutura, ela possivelmente estaria em uma posição mais afastada do meio externo.

Curiosamente, não parece haver correlação entre o GRAVY e a natureza predominante de cada posição, embora o senso comum da química nos leva a pensar que deveria existir uma relação entre as análises; por exemplo, uma relação entre a predominância de polaridade, ou ausência dela, e o GRAVY.

5.7 Aminoácido por posição da CDR H1

Gráfico 5 – Distribuição de aminoácido por posição na CDR H1. A barra horizontal representa a totalidade das sequências, e trechos de cada cor representam a fração ocupada pelo tipo de aminoácido associado a cor.



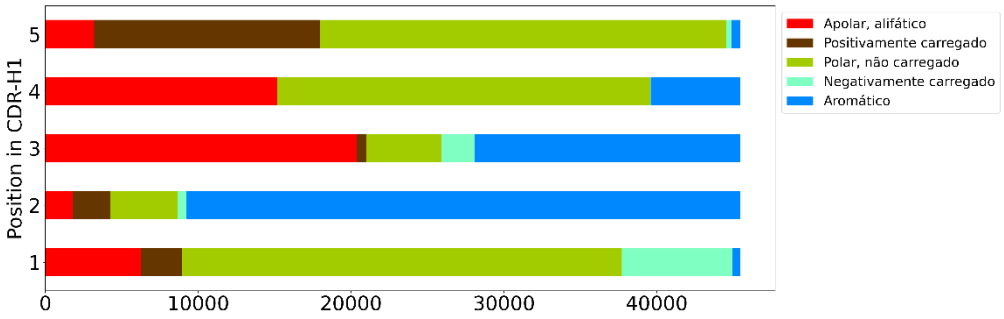
Fonte: Próprio autor.

Na CDR H1, a única posição que tem uma predominância absoluta de um tipo de aminoácido é a 2, sendo ela majoritariamente ocupada por tirosina (representada por Y). Em menor grau, temos a arginina (representada por R) ocupando a posição 4 em pouco mais da metade das sequências estudadas. Mais uma vez, essa tendência indica a possibilidade de que os aminoácidos nessa posição desempenham um papel na função e estabilidade dos anticorpos de uma maneira mais ampla.

A serina ocupa a posição 1 e a 5 numa quantidade considerável das sequências estudadas, porém não são a maioria absoluta; ambas as posições apresentam alta variabilidade de resíduos de aminoácidos. Na posição 3 a variabilidade dos aminoácidos é ainda maior, indicando que os resíduos nessa posição têm papel essencial para a especificidade dos anticorpos.

5.8 Caracterização química da CDR H1

Gráfico 6 – Distribuição de classe de aminoácido por posição da CDR H1. A barra horizontal representa a totalidade das sequências, e trechos de cada cor representam a fração ocupada na posição pelo caráter químico associado a cor.



Fonte: Próprio autor.

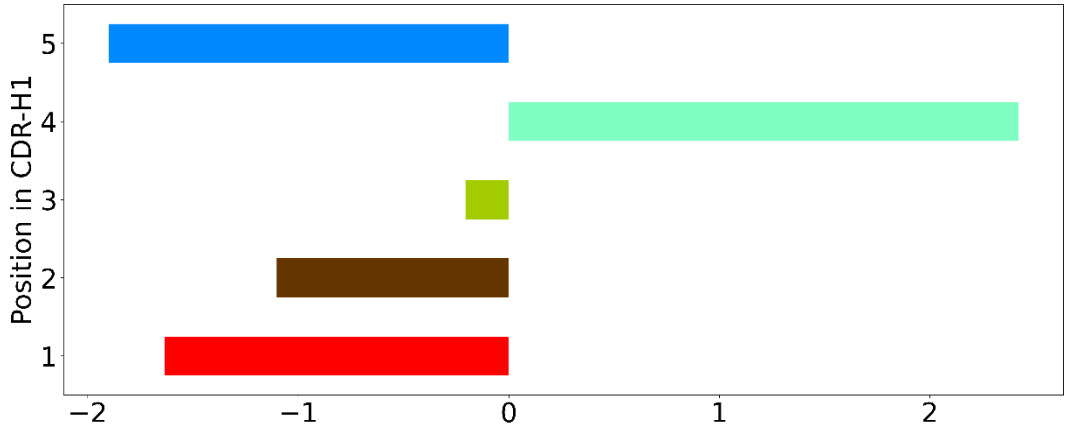
Como esperado, a posição 2 contém caráter predominantemente aromático. A tirosina, aminoácido dominante nessa posição, possui um anel aromático em sua estrutura molecular. Interessantemente, o caráter aromático que foi tão pouco presente na análise da CDR L2 surge com maior frequência na CDR H1, não apenas na posição 2.

Mas como já observado nas análises para a CDR L2, algumas posições com uma variação maior de aminoácidos possuem, ainda assim, um caráter químico predominante. Por exemplo, o caráter da posição 1 tende a ser polar, e, em menor grau, a posição 4 e 5 também apresentam essa tendência, embora na posição 4 o segundo caráter mais predominante seja o apolar, e na posição 5 o segundo caráter mais presente é positivamente carregado. Na posição 3, temos a predominância do caráter apolar e do aromático. Novamente, há uma indicação que o caráter dos aminoácidos nessa posição tem um papel importante, pois mesmo em posições com variabilidade maior, existe uma tendência para o caráter existente na posição.

É necessário um estudo mais aprofundado para compreender com mais certeza o papel do caráter químico em cada uma dessas posições, mas duas possibilidades que podem ser levantadas desde já é que esse caráter tenha papel na conformação do anticorpo, ou na sua especificidade.

5.9 Média Geral de Hidropatia (GRAVY) por posição da CDR H1

Gráfico 7 – GRAVY por posição da CDR H1. O eixo horizontal representa o valor de GRAVY, e o vertical, a posição.



Fonte: Próprio autor.

Assim como a CDR L2, a CDR H1 possui caráter majoritariamente hidrofílico, em graus variados. A posição 3, por exemplo, está muito próxima do zero, enquanto a 5 e a 1 estão mais distantes. E também como na CDR L2, existe uma posição com caráter hidrofóbico, embora a posição 4 da CDR H1 tenha caráter hidrofóbico bem maior que o da posição 2 da CDR L2.

É interessante observar que há a predominância do caráter polar na posição 4, que é fortemente hidrofóbica.

6. Conclusão

A análise em larga escala de sequências de aminoácidos pode ser essencial para um melhor entendimento da estrutura e função de anticorpos. O uso de banco de dados aliado a ferramentas computacionais torna possível a obtenção e estudo de milhares de sequências de aminoácidos simultaneamente com relativamente pouco uso de mão-de-obra humana.

A obtenção automatizada de sequências de aminoácidos em bancos do NCBI pode ser feita de várias formas, das quais duas foram testadas. A busca de acordo com similaridade por meio do BlastP mostrou-se não ser a forma ideal para o trabalho, por isso se utilizou o Entrez, ferramenta de busca textual do NCBI. A obtenção pelo Entrez resultou em cerca de 32 mil sequências de cadeia leve e 56 mil de cadeia pesada após filtragem.

A distribuição de comprimento das CDRs mostrou que apesar dessas regiões apresentarem hipervariabilidade, existem padrões nos comprimentos das diferentes CDRs. Resultados obtidos em estudos de menor escala chegaram à alguns resultados semelhantes aos obtidos nesse trabalho. A partir desse gráfico, concluiu-se também que a CDR mais representativa de todas as sequências estudadas seriam as CDR L2 de classe 7 e CDR H1 de classe 5, que foram utilizadas para as demais análises.

A análise de aminoácido por posição e natureza do resíduo também por posição dessas CDRs revelou algumas tendências claras. Em algumas posições, observou-se que havia predominância de certa natureza de resíduo apesar de não haver predominância de nenhum aminoácido em específico.

A média geral de hidropatia (GRAVY) mostrou que não há correlação entre a polaridade ou a falta dela na posição e a afinidade ou aversão por água.

7. Referências Bibliográficas

ABHINANDAN, K. R.; MARTIN, A. C. R. Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. **Molecular Immunology**, v. 45, n. 14, p. 3832–3839, 2008. Disponível em: https://doi:10.1016/j.molimm.2008.05.0. Acesso em: 02 de agosto de 2021.

AHMAD, Z. A.; YEAP, S. K.; ALI, A. M.; HO, W. Y.; ALITHEEN, N. B. M.; HAMID, M. scFv Antibody: Principles and Clinical Application. **Clinical and Developmental Immunology**, p 1-15, 2012. Disponível em: https://doi:10.1155/2012/980250. Acesso em: 04 de junho de 2021.

BLACK, C. A. A brief history of the discovery of the immunoglobulins and the origin of the modern immunoglobulin nomenclature. **Immunology and Cell Biology**, *v. 75, n. 1, p. 65–68, 1997.* Disponível em: https://doi:10.1038/icb.1997.10. Acesso em: 05 de junho de 2021.

CARVALHO, B. M. **Desenvolvimento de estratégias de análise de interações antígeno-anticorpo e visualização de anticorpos em larga escala.** Tese (Doutorado em Bioinformática) — Universidade Federal de Minas Gerais, Minas Gerais, 2019.

CHEN, C.; HUANG, H.; WU, C. H. Protein Bioinformatics Databases and Resources. **Methods in Molecular Biology**, p. 3–39, 2017. Disponível em: https://doi:10.1007/978-1-4939-6783-4_1. Acesso em: 10 de junho de 2021.

CHIU, M. L.; GOULET D. R.; TEPLYAKOV A.; GILLILAND, G. L. Antibody Structure and Function: The Basis for Engineering Therapeutics. **Antibodies**, v. 8, n. 4, p. 55, 2019. Disponível em: https://doi:10.3390/antib8040055. Acesso em: 20 de junho de 2021.

CHOI, B. D.; GEDEON, P. C.; KUAN, C.-T.; SANCHEZ-PEREZ, L.; ARCHER, G. E.; BIGNER, D. D.; SAMPSON, J. H. Rational design and generation of recombinant control reagents for bispecific antibodies through CDR mutagenesis. **Journal of Immunological Methods**, vol. 395, n. 1-2, p; 14–20, 2013. Disponível em: https://doi:10.1016/j.jim.2013.06.003. Acesso em: 28 de agosto de 2021.

CHOTHIA, C.; LESK, A. M. Canonical structures for the hypervariable regions of immunoglobulins. **Journal of Molecular Biology**, vol. 196, n. 4, p. 901–917, 1987. Disponível em: https://doi:10.1016/0022-2836(87)90412-8. Acesso em: 15 de junho de 2021.

COCK P. A.; ANTAO T.; CHANG J. T.; CHAPMAN B. A.; COX C. J.; DALKE A.; FRIEDBERG I.; HAMELRYCK T.; KAUFF F.; WILCZYNSKI B.; HOON M. J. L. Biopython: freely available Python tools for computational molecular biology and bioinformatics. ***Bioinformatics***, v. 25, p. 1422-1423, 2009. Disponível em: https://doi.org/10.1093/bioinformatics/btp163. Acesso em: 03 de agosto de 2021.

DUNBAR J.; DEANE C. M. ANARCI: antigen receptor numbering and receptor classification. **Bioinformatics**, v. 32, n. 2, p. 298-300. Disponível em: https://doi:10.1093/bioinformatics/btv552. Acesso em: 10 de agosto de 2021.

FRANCO, M. L. ; CEDIEL, J. F.; PAYAN, C. Breve historia de la bioinformática. **Colomb. Med.**, v. 39, n. 1, p. 117-120, 2008. Disponível em: <http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1657-95342008000100015&lng=en&nrm=iso>. Acesso em: 24 de junho de 2021.

GONZÁLEZ-PECH, R. A.; STEPHENS, T. G.; CHAN, C. X. Commonly misunderstood parameters of NCBI BLAST and important considerations for users. **Bioinformatics**, 2018. Disponível em: https://doi:10.1093/bioinformatics/bty1018. Acesso em: 10 de junho de 2021.

KABAT E. A.; WU T. T.; BILOFSKY H. Unusual distributions of amino acids in complementarity determining (hypervariable) segments of heavy and light chains of immunoglobulins and their possible roles in specificity on antibody-combining sites. **Journal of Biological Chemistry**, v. 252, n.19, p. 6609-6616, 1977. Disponível em: https://bit.ly/3mymTfU. Acesso em 25 de agosto de 2021.

KABAT, E. A.; WU, T. T. Attempts to locate complementarity-determining residues in the variable positions of light and heavy chains. **Annals of the New York Academy of Sciences**, v. 190, n. 1, p. 382–393, 1971. Disponível em: https://doi:10.1111/j.1749-6632.1971.t. Acesso em: 05 de agosto de 2021.

KABAT, E. A.; WU, T. T.; BILOFSKY, H. Attempts to locate residues in complementarity-determining regions of antibody combining sites that make contact with antigen. **Proceedings of the National Academy of Sciences**, v. 73, n. 2, p. 617–619, 1976. Disponível em: https://doi:10.1073/pnas.73.2.617. Acesso em: 15 de junho de 2021.

KYTE, J.; DOOLITTLE, R. F. A simple method for displaying the hydropathic character of a protein. **Journal of Molecular Biology**, v. 157, n. 1, p. 105–132, 1982. Disponível em: https://doi:10.1016/0022-2836(82)90515-0. Acesso em: 05 de agosto de 2021.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. **The NCBI Handbook**. 2. ed. Maryland: NCBI. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK143764/. Acesso em: 25 de julho de 2021.

NCBI PROTEIN RESOURCES. In: SAYERS, E. **The NCBI Handbook.** Maryland: NCBI, 2013. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK169830/. Acesso em: 02 de agosto de 2021.

PADLAN, E. A. Anatomy of the antibody molecule. **Molecular Immunology**, v. 31, n.3, p. 169–217, 1994. Disponível em: https://doi:10.1016/0161-5890(94)90001-9. Acesso em: 04 de junho de 2021.

RECOGNITION AND RESPONSE. In: PUNT, J.; STRANFORD S. A.; JONES, P. P.; OWEN, J. A. **Kuby Immunology**, ed. 8. New York: Macmillan Education, 2019. p. 195-200.

RODRIGUES, F. N. **Análise das estruturas de fragmentos de anticorpos VH e VL com potencial para aplicação in silico.** 2014. Trabalho de conclusão de curso (Bacharelado em Biotecnologia) – Centro de Ciências, Universidade Federal do Ceará, Fortaleza, 2014.

SCHROEDER, H. W.; CAVACINI, L. Structure and function of immunoglobulins. **Journal of Allergy and Clinical Immunology**, v. 125, n.2, p. 41–52, 2010. Disponível em: https://doi:10.1016/j.jaci.2009.09.046. Acesso em: 02 de junho de 2021.

SCHROEDER, H. W.; CAVACINI, L. Structure and function of immunoglobulins. **Journal of Allergy and Clinical Immunology**, v. 125, n. 2, p. S41–S52, 2010. Disponível em: https://doi:10.1016/j.jaci.2009.09.046. Acesso em: 28 de agosto de 2021.

SIEVERS, F.; HIGGINS, D. G. Clustal Omega. **Current Protocols in Bioinformatics**, *2014*. Disponível em: https://doi:10.1002/0471250953.bi0313s. Acesso em: 15 de agosto de 2021.

SILVA, H. M. **Método in silico para análise de sequências de imunoglobulinas produzidas por tecnologia de phage display**. Dissertação (Mestrado em Biologia Molecular) — Universidade de Brasília, Brasília, 2016.

THE BLAST SEQUENCE ANALYSIS TOOL. In: MADDEN, T. **The NCBI Handbook.** Maryland: NCBI, 2013. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK169830/. Acesso em: 02 de agosto de 2021.

THE COMPLEXITY OF PROTEIN STRUCTURE AND THE CHALLENGES IT POSES IN DEVELOPING BIOPHARMACEUTICALS. *In:* BERKOWITZ, S. A.; HOUDE, D. J. **Biophysical Characterization of Proteins in Developing Biopharmaceuticals**, Elsevier, 2014. p. 3–26.

THE ENTREZ SEARCH AND RETRIEVAL SYSTEM. In: OSTELL, J. **The NCBI Handbook**. Maryland: NCBI, 2013. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK184582/. Acesso em: 02 de agosto de 2021.

WU, T. T.; JOHNSON, G.; KABAT, E. A. Length distribution of CDRH3 in antibodies. **Proteins: Structure, Function, and Genetics**, v. 16, n. 1, p. 1–7, 1993. Disponível em: https://doi:10.1002/prot.340160102. Acesso em: 21 de junho de 2021.