UNIVERSIDADE DE SÃO PAULO

FACULDADE DE CIÊNCIAS FARMACÊUTICAS

Curso de Graduação em Farmácia

ESTUDO DOS IMPACTOS CONFORMACIONAIS EM PRODUTOS DE GENES COM VARIANTES NÃO SINÔNIMAS COM CONFIGURAÇÕES CIS

Camila Hosoe Takase

Trabalho de Conclusão do Curso de Farmácia da Faculdade de Ciências Farmacêuticas da Universidade de São Paulo.

Orientador: Prof. Dr. Michel Satya Naslavsky

Coorientador: Dr. Glaucio Monteiro Ferreira

São Paulo

2025

SUMÁRIO

Pág.

LISTA DE ABREVIATURAS

APOE	Apolipoproteína E
DM	Dinâmica Molecular
GWAS	*Genome-wide Association Studies* (Estudos de Associação Ampla do Genoma)
PDB	*Protein Data Bank*
pLOFs	*Putative loss of function*
PRS	*Perturbation Response Scanning*
SABE	Saúde, Bem-estar e Envelhecimento (Coorte sequenciada)
SKAT	*Sequence Kernel Association Test*
SNP	Polimorfismo de Nucleotídeo Único
SNV	Variante de Nucleotídeo Único
NGS	*Next-Generation Sequencing* (Sequenciamento de Nova Geração)
WES	*Whole-Exome Sequencing* (Sequenciamento de Exomas Completos)
WGS	*Whole-Genome Sequencing* (Sequenciamento de Genomas Completos)

LISTA DE FIGURAS

Figura 14. (A) Gráfico de composição de SSE de cada fragmento ao longo da simulação referente à proteína com a variante rs2297810; (B) Gráfico de cada resíduo e seu SSE ao longo da simulação referente à proteína com a variante rs2297810

Figura 15. Gráfico de composição de SSE da simulação referente à proteína com a variante rs2297809

Figura 16. (A) Gráfico de composição de SSE de cada fragmento ao longo da simulação referente à proteína com a variante rs2297809; (B) Gráfico de cada resíduo e seu SSE ao longo da simulação referente à proteína com a variante rs2297809

Figura 17. Gráfico de composição de SSE da simulação referente à proteína com as variantes rs2297810 e rs2297809

Figura 18. (A) Gráfico de composição de SSE de cada fragmento ao longo da simulação referente à proteína com as variantes rs2297810 e rs2297809; (B) Gráfico de cada resíduo e seu SSE ao longo da simulação referente à proteína com as variantes rs2297810 e rs2297809

RESUMO

TAKASE, C. H. **Estudo dos impactos conformacionais em produtos de genes com variantes não sinônimas com configurações cis**. 2025. no. f. Trabalho de Conclusão de Curso de Farmácia – Faculdade de Ciências Farmacêuticas – Universidade de São Paulo, São Paulo, 2025.

Variantes em regiões gênicas podem ter efeitos funcionais sobre seus produtos com potenciais consequências em fenótipos clínicos. Variantes codificantes não sinônimas, cujo aminoácido é substituído, podem ter impactos na estrutura da proteína e, portanto, em sua função. A análise e avaliação dos efeitos da presença simultânea de duas ou mais variantes codificantes não-sinônimas em cis (no mesmo cromossomo homólogo) quanto à estrutura da proteína formada e consequências fenotípicas para o indivíduo portador são comumente relegadas em comparação à investigação de implicações decorrentes de variantes não sinônimas analisadas de maneira isolada.

O projeto tem por objetivo investigar a ocorrência de variantes codificantes não-sinônimas em cis em uma amostra populacional de brasileiros e seu impacto na estrutura proteica. Foi utilizada uma amostra populacional de brasileiros a partir dos dados de sequenciamento de genomas completos obtidos das coortes SABE (Saúde, Bem-estar e Envelhecimento) como base para a filtragem de variantes não-sinônimas. A partir disso, sabendo da existência de uma abundância de genes com configurações em cis de variantes codificantes em genomas humanos, foram realizadas filtragens para identificação de potenciais variantes em cis. Com base nos genes obtidos pela fase anterior foram buscados aqueles para os quais a estrutura tridimensional da proteína codificada se encontra resolvida e disponível, de forma a ser possível avaliar e comparar os impactos decorrentes da co-ocorrência de variantes em cis com a ocorrência de uma única variante isolada, buscando elucidar seus efeitos e promover um maior foco para esse âmbito.

Foi selecionado o gene *CYP4B1* para o qual foram realizadas simulações de Dinâmica Molecular para a proteína selvagem, com a variante rs2297810, com a variante rs2297809 e com as duas variantes co-ocorrendo. As análises dos resultados demonstraram que a interação entre as variantes rs2297810 e rs2297809 na CYP4B1 ocorre de maneira sinérgica, com o efeito da co-ocorrência sendo maior que a adição de seus efeitos individuais.

Palavras-chave: Variantes não-sinônimas, cis, genes, dobramento de proteínas

ABSTRACT

Variants in gene regions can have functional effects over its products, often with consequences to clinical phenotypes. Coding non-synonymous variants, to which amino acids are substituted, can impact protein structure and, as a consequence, its function. The analysis and assessment of the effects of the presence of simultaneous non-synonymous coding variants in cis configuration (across a single homologous chromosome) concerning the structure of the resulting protein and the phenotypic consequences to the carrier are commonly dismissed in comparison to the study of implications that occur due to isolated non-synonym coding variants.

The project aims to investigate the occurrence of cis non-synonymous coding variants in a Brazilian population sample utilizing the whole-genome sequencing dataset obtained from the SABE (Saúde, Bem-estar e Envelhecimento) cohort as a basis to filter the non-synonym variants. Knowing the existence of a significant abundance of cis configurations of coding variants in diploid human genomes, filters will be used to identify potential cis variants. Based on the genes obtained in the last phase, the ones with resolved and available tridimensional protein structures will be sought, to enable evaluation and comparison with the impacts caused by the co-occurrence of cis variants and the occurrence of a single isolated variant, seeking to elucidate its effects and promote a bigger focus on this sphere.

The *CYP4B1* gene was selected and Molecular Dynamics simulations were ran for the wild type protein, with the rs2297810 variant, with the rs2297809 variant and with both variants co-occurring. Results analysis showed that the interaction between rs2297810 and rs2297809 variants on the CYP4B1 protein is synergic, with the effects of the co-occurrence being bigger than the addition of their individual effects.

Keywords: non-synonymous variants, cis, genes, protein folding

1. INTRODUÇÃO

O estudo de variantes genéticas e respectivas correlações com desfechos podem ser descritos como o efeito direto da variante sobre o contexto a sua volta - muitas vezes o gene - ou como o efeito indireto sobre um determinado fenótipo. No primeiro caso (da inferência de um efeito direto), a identificação de variantes a partir de estudos de sequenciamento permite, por exemplo, a anotação de sua potencial consequência regulatória ou funcional (SUHRE; GIEGER, 2012). Vemos, por exemplo, que a anotação de variantes de potencial perda de função (*putative loss of function* - pLOFs) ou não-sinônimas (com trocas de aminoácido) é realizada de maneira automatizada, após a anotação do contexto da variante no transcrito ser realizada para dados provenientes de ensaios de sequenciamento (BALASUBRAMANIAN et al., 2017). Contudo, esta análise é feita de maneira pontual, variante a variante. A investigação dos efeitos de múltiplas variantes não sinônimas em fase, ou seja, quando mais de uma variante não sinônima ocorre no mesmo cromossomo homólogo (em uma configuração cis), apresenta grande potencial para novas abordagens de inferência de efeito direto.

Atualmente o Sequenciamento de Nova Geração (NGS) permite o sequenciamento de DNA em larga escala com grande precisão e relativa acessibilidade, sendo uma ferramenta que permite a análise de sequências e sua aplicação em projetos em busca de variantes, que por sua vez serão anotadas e utilizadas em estudos de associação com fenótipos (SHEN et al., 2014). Como mencionado acima, as variantes precisam ser anotadas em relação ao seu contexto genômico predito (inferência de efeitos diretos da variante sobre um dos transcritos e sequências protéicas) para que possam ser interpretadas ou filtradas em estudos futuros. Sendo assim, a partir da comparação com um genoma de referência, a presença de variantes num indivíduo também permite a busca da co-ocorrência de variantes em um mesmo indivíduo, em cis ou em trans (OSADA; MIYAGI; TAKAHASHI, 2017). Estas configurações são importantes para, por exemplo, classificar a patogenicidade de variantes associadas a condições recessivas (RICHARDS et al., 2015).

Como mencionado no segundo tipo de descrição (primeiro parágrafo), também é possível medir o efeito indireto de variantes sobre um fenótipo. Dentre os métodos

existentes, os Estudos de Associação Ampla do Genoma (GWAS - *genome-wide association studies*) buscam associações entre polimorfismos genéticos (variantes comuns na população) e determinadas características, que podem ser quantitativas, discretas (qualitativas), sendo variações normais ou ocorrência de uma doença (CORDELL; CLAYTON, 2005). Contudo, essa abordagem tipicamente utiliza-se de microarranjos de DNA contendo sondas para variantes comuns, sem acolher variantes raras. Portanto, a interrogação de genótipos de variantes comuns entre casos e controles permite a análise das frequências e, por fim, a associação estatística entre variantes e fenótipos. Alternativamente, os estudos em famílias acolhem as características menos prevalentes e que co-segregam entre familiares, permitindo por vezes a identificação de variantes de grande efeito. Nos estudos de identificação de genes associadas a estas doenças, observou-se que muitas ocorrem nas regiões codificantes, levando o sequenciamento de exomas completos (WES) a se tornar um dos testes padrão ouro (WRIGHT; FITZPATRICK; FIRTH, 2018).

A partir desse tipo de sequenciamento pode-se diminuir significativamente o tempo necessário para a obtenção de diagnósticos genéticos, além de torná-los mais precisos. Com isso, é possível um melhor entendimento do prognóstico, a melhoria em tratamentos personalizados e uma melhor gestão e monitoramento das condições e indivíduos, de forma que o estudo de variantes genéticas raras é de interesse (RASZEK, 2013).

Variantes raras possuem, por definição, menor ocorrência, e por vezes não possuem seu impacto considerado em análises populacionais (MORRIS; ZEGGINI, 2010). Ou seja, por terem baixa frequência, o risco populacional pode ser considerado baixo. Entretanto, podem apresentar significância funcional e um efeito individual elevado (ZOGHBI et al., 2021). Espera-se tal efeito nas doenças raras que afetam as famílias, mas também não é possível descartar o efeito individual em doenças comuns (SAINT PIERRE; GÉNIN, 2014). Visto que grandes estudos de sequenciamento de genomas completos (WGS) e WES têm sido publicados, estudar o impacto da co-ocorrência de variantes não sinônimas em cis pode ser útil para promover um melhor entendimento sobre o contexto de interação de variantes (FISH; CAPRA; BUSH, 2016). Esse estudo pretende elucidar essa interação, buscando variantes que co-ocorrem em um mesmo gene para um mesmo indivíduo e,

inicialmente, o seu efeito sobre o fenótipo da estrutura protéica (LIU; WATSON; ZHANG, 2015). A identificação de múltiplas variantes raras por gene, a partir de grandes bancos de dados como o gnomAD, permite a determinação de uma métrica para o gene a respeito do potencial impacto de novas variantes detectadas na mesma região (KARCZEWSKI et al., 2020). Esta métrica pode auxiliar na avaliação dos impactos, prevalência e novas intervenções frente às diversas condições. Ainda que a co-ocorrência de variantes esteja sendo implementada no gnomAD, seus impactos combinados ainda não foram explorados (GUDMUNDSSON et al., 2022).

Estudos de agregação de variantes raras, a exemplo do Burden e SKAT, permitem agregar por gene ou via gênica a presença de múltiplas variantes raras e comparar esta métrica entre casos versus controles, testando a associação entre um score e determinada característica (LEE et al., 2014). Além disso, ressalta-se que no Burden, essa concentração de variantes raras em indivíduos caso já engloba as situações em que existe a coocorrência de variantes não-sinônimas em cis, ou seja, quando duas ou mais variantes são contabilizadas para um mesmo indivíduo, este efeito somado é considerado dentro da análise de agregação, ainda que o peso não seja atribuído de maneira diferencial quando as variantes estão isoladas ou em cis, assumindo um efeito aditivo.

A ocorrência simultânea de duas ou mais variantes não-sinônimas com configuração em cis pode apresentar significância relevante e alterar os efeitos fenotípicos. Esse efeito é observado para o produto do gene *APOE*, para o qual a análise individual das variantes não é suficiente para o entendimento do fenótipo. O grande aumento da ocorrência de Alzheimer é verificado a partir da presença da variante ε4, na qual há co-ocorrência de arginina nas posições 112 e 158 (GHEBRANIOUS et al., 2005). Esse efeito potencializador não é observado para o alelo mais comum ε3, que apresenta arginina apenas na posição 158, com cisteína na posição 112, ou para o alelo ε2, que apresenta cisteína nas duas posições e possui efeito protetor contra o Alzheimer (LE GUEN et al., 2022).

Outro exemplo são os receptores β2-adrenérgicos humanos, os quais foram examinados (DRYSDALE et al., 2000) apresentando resultados que indicam que as interações dos múltiplos polimorfismos de nucleotídeo único (SNPs) no mesmo haplótipo afetam o fenótipo biológico e terapêutico. Em adição, foi constatado que

SNPs isolados não apresentaram nenhum poder preditivo. Em outro estudo foram realizadas todas as comparações possíveis de pares no genoma utilizando os 372 genomas com ancestralidade europeia (EUR) do 1000 Genomes, a partir do qual extraíram-se 1.047 genes de “fase alternada”, que possuem pares de mutações idênticas que potencialmente geram perturbações, tanto em cis quanto em trans. Esses genes podem apresentar diferentes haplótipos apesar de idênticos no genótipo mutado, de forma que sua interpretação clínica e funcional frente a mutações pode ser dependente de sua fase no genoma em questão (HOEHE et al., 2014). Outros resultados também corroboram com a noção de que o efeito conjunto de múltiplas variantes nos mesmos genes é significativamente diferente daquele de uma única variante (LIU; WATSON; ZHANG, 2015).

A partir disso, esse estudo busca outros casos em que a co-ocorrência de variantes apresente um significado funcional com base na comparação de casos e controle, elucidando essa interação e suas consequências. Também foi utilizado como base o estudo de 2019 (HOEHE et al., 2019), que evidenciou a existência de uma abundância significativa de configurações em cis de variantes codificantes em genomas humanos diploides, com uma proporção de aproximadamente 60:40 em relação a genes com configurações em trans. Essa abundância de cis foi observada em praticamente todos os genomas em todas as populações. Além disso, esse estudo também constatou que a proporção da ocorrência das configurações é característica dos genes, com a existência de uma grande porção de genes abundantes em cis. Dessa forma, justifica-se a verificação da co-ocorrência de variantes não sinônimas em cis numa população miscigenada (a exemplo dos brasileiros) e no estudo de impacto de algumas variantes comuns em genes com estrutura protéica resolvida e associação prévia com fenótipos comuns.

Com base nos resultados abrem-se possibilidades de aplicação das informações sobre deleteriedade para uso clínico, tanto em relação a diagnósticos como para tratamentos otimizados e individualizados. Sabendo dessas informações também é possível realizar o direcionamento e priorização da cristalização da estrutura de determinadas proteínas para maior elucidação dos efeitos estruturais e funcionais envolvidos em sua deleteriedade. Ademais, pode se tornar um potencial anotador, com o resultado das estruturas voltando para as variantes.

2. OBJETIVO(S)

Objetivo Geral:

Investigar a ocorrência de variantes codificantes não-sinônimas em cis em uma amostra populacional de brasileiros e seu impacto na estrutura protéica.

Objetivos específicos:

→ Filtrar dados de genomas completos da coorte SABE em busca de variantes não sinônimas em heterozigose e homozigose;

→ Identificar e descrever potenciais variantes em cis utilizando métodos de faseamento e filtros de co-ocorrência de genótipos em homozigose e heterozigose;

→ Cruzar os genes e variantes filtradas nas etapas anteriores com bancos de dados e proteínas com estrutura tridimensional disponível e aderentes a uma série de parâmetros de qualidade;

→ Avaliar o impacto das variantes independentes e das variantes co-ocorrendo em cis sobre estruturas protéicas para uma série de parâmetros

3. MATERIAL E MÉTODOS

3.1. Estratégia e algoritmo

Iniciou-se o processo buscando a organização dos dados anotados previamente obtidos a partir do sequenciamento de genomas completos das coortes SABE (Saúde, Bem-estar e Envelhecimento) (NASLAVSKY et al., 2022), que passaram por uma filtragem prévia de forma a conter apenas variantes não sinônimas. Essa base de dados abrange 1.171 idosos não aparentados, com 329.744 das variantes anotadas como sendo codificantes não sinônimas. Esses dados estavam apresentados no formato de uma tabela do Excel, onde cada linha representa uma variante, coordenadas e anotações e nas últimas colunas estão os indivíduos, cujas células verticalmente dispostas e cruzando com as variantes representam os genótipos.

A partir disso, foi desenvolvido um script para o tratamento de dados da tabela, iniciando com a criação de uma etapa para exclusão de informações não relevantes para o projeto, permitindo expressiva diminuição na quantidade de dados, com melhora na organização, visualização e tempo de processamento. Foram mantidas apenas as colunas contendo a anotação de consequência da variante do gene (nesse caso consequência exônica e não-sinônima), o aminoácido referência, o aminoácido alternativo (variante) para cada transcrito, o gene a qual a variante pertence, o identificador da variante na base de polimorfismos de nucleotídeo único (dbSNP, identificadores rs_) e os indivíduos sequenciados (colunas com genótipos).

3.2. Identificação de variantes potencialmente em cis

A atribuição ideal da fase de duas variantes consiste no uso de trios (pais e filho/filha). Na ausência deste tipo de dado, o faseamento estatístico solucionaria este problema com resultados probabilísticos (REF: ShapeIt https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html).

Para este projeto, foi utilizada a filtragem de indivíduos que tenham ao menos um genótipo em homozigose e outro em heterozigose, assim garantindo haplótipos com 2 ou mais alelos não sinônimos (em uma mesma fase), de modo a garantir a existência de variantes em cis, e permitindo a comparação entre casos e controles.

Foram criadas etapas que realizam a contagem da frequência do aparecimento de ‘0/1’ (genótipo relacionada à presença de uma variante em heterozigose) na coluna de cada indivíduo, que caso fosse 2 ou maior é considerada na contagem. Da mesma forma, foram criadas etapas para a contagem da frequência de ‘1/1’ (genótipos referentes à presença de uma variante em homozigose), considerando aqueles com contagem de 1 ou mais.

A frequência de indivíduos que atendem a essas condições foi contabilizada e adicionada ao DataFrame criado. De mesma forma, também se criou uma etapa que contabiliza os indivíduos que atendem a ambas as condições, sendo estes os indivíduos de interesse, que apresentam as variantes codificantes não-sinônimas na configuração em cis.

A partir disso criou-se uma etapa final na qual se gera uma tabela para o gene analisado, com indicação da quantidade de variantes codificantes não sinônimas em heterozigose e em homozigose presentes na quantidade de interesse para cada indivíduo. Essa tabela também contém a quantidade total de variantes na quantidade de interesse de cada tipo, assim como a quantidade de indivíduos que atendem ambas as condições.

Essa filtragem foi empregada em um recorte das 10.000 primeiras variantes da lista proveniente dos dados sequenciados, realizado de forma a diminuir o tamanho do arquivo e possibilitar a análise.

A partir dos resultados obtidos foi possível constatar que a grande maioria dos genes não apresenta variantes em cis segundo os critérios utilizados pelo filtro. Dentre os genes que apresentam variantes em cis, foi realizada a separação daqueles que as possuíam em quantidades significativas para análise posterior, sendo selecionados 45 genes.

3.3. Anotação de genes cuja proteína possui estrutura tridimensional resolvida

Com base nos genes classificados como apresentando variantes em cis em quantidades significativas, foram utilizados datasets disponíveis no PDB (https://www.rcsb.org/) para verificar e selecionar aqueles que possuem estrutura tridimensional resolvida. Foram analisados também os dados de qualidade do cristal, tais como classificação, método utilizado para construção do cristal, resolução do cristal, R-value free, R-value work e missing atoms.

A partir disso, as estruturas foram selecionadas com base na resolução, presença de ligante e qualidade e método da cristalização, buscando proteínas completas, além da presença de sua publicação.

Foram eliminados aqueles que constavam “Solution NMR” como método, uma vez que devido à utilização de soluções geralmente apresentam múltiplas conformações, havendo preferência aos métodos de Difração de raio X e Cryo-EM.

Em relação à resolução, buscou-se aquelas estruturas que apresentavam valores menores que 3.0 Å, consequentemente apresentando valores satisfatórios de R-value free e R-value work.

Com isso houve a redução de 45 genes para 9, sendo eles *CLCNKA, CROCC, CYP4B1, INPP5B, MACF1, PADI4, PER3, PRDM2* e *SPEN.*

3.4. Identificação das co-ocorrências mais comuns

Entre os genes selecionados na etapa anterior foi realizada a identificação dos conjuntos de variantes em cis que aparecem em maior quantidade entre os indivíduos, de forma a auxiliar no direcionamento da escolha das variantes que serão analisadas no projeto.

Para isso foi realizada uma varredura das tabelas com os resultados de cada gene, contabilizando as variantes presentes e dessa forma estabelecendo a quantidade de co-ocorrências, assim como a proporção em que aparecem.

Foram consideradas relevantes aquelas que ocorreram mais de 10 vezes. Todos os 9 genes apresentaram no mínimo um conjunto de co-ocorrências que atingia essa quantidade.

3.5. Modelagem das proteínas por homologia

Foram empregadas as ferramentas *Modeller* (https://salilab.org/modeller/) e *Chimera* (https://www.cgl.ucsf.edu/chimera/download.html), integradas com a realização de um blastp para a realização da modelagem por homologia das proteínas provenientes dos 9 genes promissores derivados das etapas anteriores (*CLCNKA, CROCC, CYP4B1, INPP5B, MACF1, PADI4, PER3, PRDM2 e SPEN*). Nessa etapa de modelagem também foram incluídas as regiões faltantes das proteínas, não presentes nas sequências de referência.

3.6. *CYP4B1*

Entre os genes promissores, com variantes que co-ocorrem em cis e que passaram pela modelagem por homologia foi selecionado o gene *CYP4B1* devido ao seu tamanho, à presença de informações sobre as variantes de relevância, interesse no estudo de uma proteína com função não elucidada e qualidade do modelo criado.

Em mamíferos a *CYP4B1* age no metabolismo de endo e xenobióticos, não apresentando papel importante no metabolismo hepático de fase I, uma vez que apresenta majoritária expressão extra-hepática, em especial no pulmão. Contudo, não se observa atividade catalítica desta enzima em humanos, possivelmente devido a uma mutação muito conservada de troca de prolina por serina na posição 427, de forma que sua função fisiológica ainda não foi elucidada. Entretanto, está possivelmente envolvida em diversos tipos de câncer devido a níveis de expressão alterados em tecidos tumorais (RÖDER et al., 2023).

3.6.1. Co-ocorrências mais comuns no gene *CYP4B1*

As variantes que co-ocorrem com maior frequência no gene *CYP4B1*, resultado obtido na etapa 3.4., são os pares rs2297810 e rs4646491 e rs2297810 e rs2297809. Tais variantes apresentam associação descrita com um risco aumentado de desenvolvimento de câncer de pulmão (YANG et al., 2023).

Para as simulações foram selecionadas as variantes rs2297810 e rs2297809.

Rs2297810 corresponde à mudança de um resíduo de metionina por um resíduo de isoleucina na posição 331.

Rs2297809 corresponde à mudança de um resíduo de arginina por um resíduo de cisteína na posição 340.

Figura 1. Proteína CYP4B1, com os resíduos 331 e 340 em evidência e indicados em vermelho

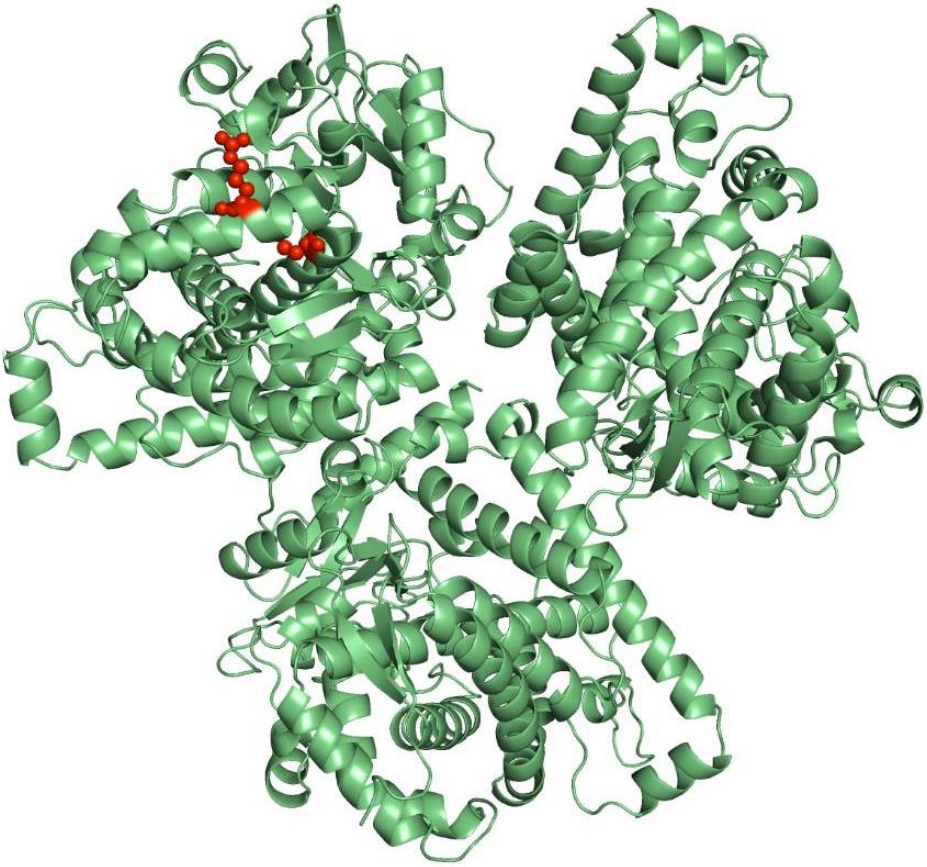
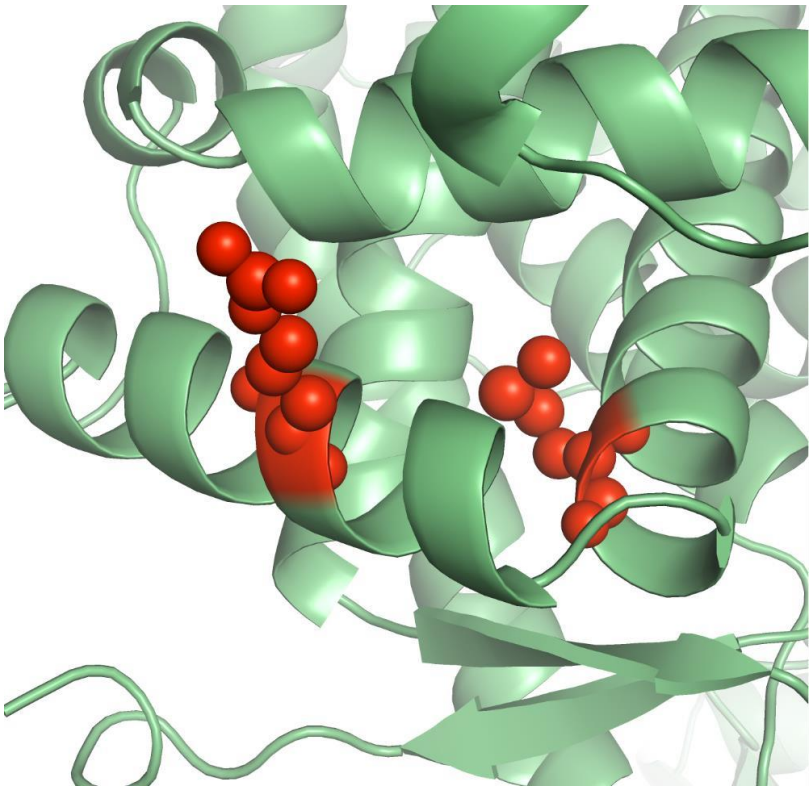


Figura 2. Ampliação da região em que se encontram os resíduos 331 e 340



3.7. Dinâmica Molecular e análise das estruturas proteicas

Existem diferentes exemplos e métodos, com diferentes níveis de sucesso, utilizados para o teste de dobramento de proteínas com mais de uma mutação e sua análise e predição da estabilidade de sua estrutura.

Empregam-se diversas estratégias computacionais (MARABOTTI; SCAFURI; FACCHIANO, 2021), que inferem os efeitos decorrentes das mutações a partir da análise diversos fatores, como o empacotamento da cadeia lateral, perturbações na cadeia lateral junto de funções de energia heurística ou pela estimação da energia livre de dobramento decorrente das mutações com base em potenciais presentes em bancos de dados. Contudo, cabe citar que muitos métodos analisam o efeito de apenas uma mutação e algumas ferramentas comumente empregadas, como o FoldX, apesar de serem capazes de iterativamente gerar mutantes com múltiplas mutações in sílico, diferem da análise do efeito simultâneo de múltiplas mutações (ANDERSSON, 2016).

Para o projeto foi utilizado o software *Desmond* (BOWERS, K. J. et al., 2006), integrado com o ambiente de modelagem molecular *Maestro* (Schrödinger Release 2025-2), que permite a realização de simulações de Dinâmica Molecular (DM) de alta performance com a utilização de GPU, com visualização e análise dos resultados.

As simulações de DM consistem em um método físico baseado em princípios físicos de Newton para análise das interações e movimentos de átomos e moléculas. Para tal é utilizado um campo de força que permite a estimativa de forças entre átomos que interagem e o cálculo de energia total do sistema. Há então a geração de configurações sucessivas do sistema, provendo trajetórias que especificam a posição e velocidade das partículas ao longo do tempo, a partir das quais é possível calcular uma variedade de propriedades (DE VIVO et al., 2016).

4. RESULTADOS

Foram realizadas simulações de DM para a proteína *CYP4B1* em sua forma selvagem, com a variante rs2297810, com a variante rs2297809 e com as duas variantes co-ocorrendo. Todas as simulações apresentam uma trajetória de 200 ns.

4.1. RMSD

O RMSD (*Root Mean Square Deviation*) é utilizado para medir a média de mudança no deslocamento de uma seleção de átomos em determinado segmento em relação a uma referência (geralmente o primeiro segmento), sendo calculado para todos os segmentos (DINDI et al., 2023). Seu monitoramento pode auxiliar na compreensão da estrutura conformacional ao longo da simulação e sua análise pode indicar se houve equilíbrio na simulação e, consequentemente, se é passível de análise.

Pode-se observar que as flutuações representadas na Figura 6 são mais significativas, indicando maiores mudanças conformacionais durante a simulação da proteína com as duas variantes co-ocorrendo em cis.

Figura 3. Gráfico de RMSD da simulação referente à proteína selvagem

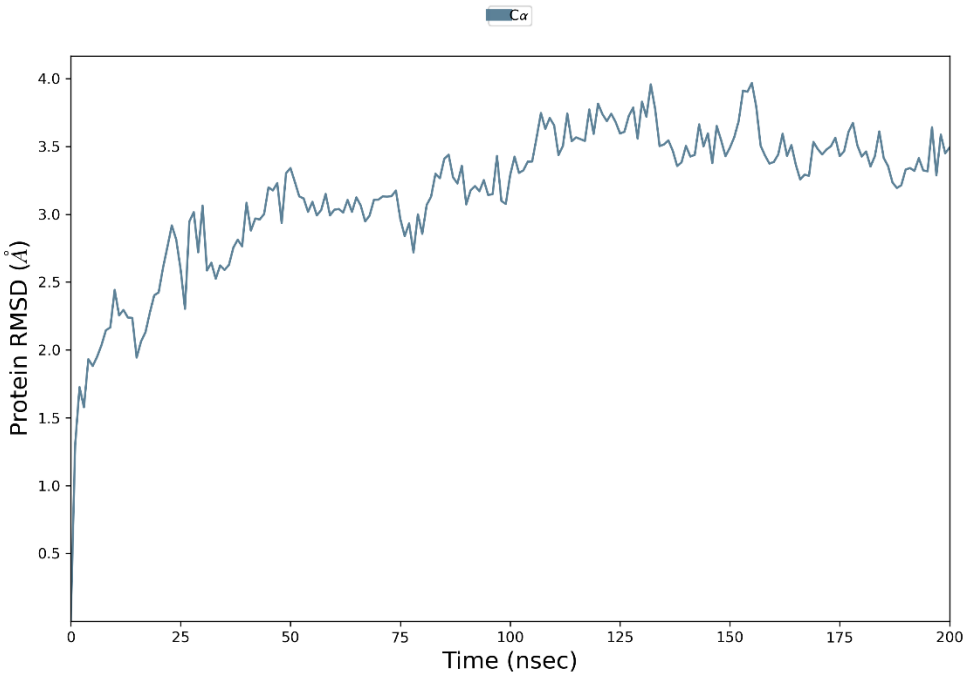


Figura 4. Gráfico de RMSD da simulação referente à proteína com a variante rs2297810

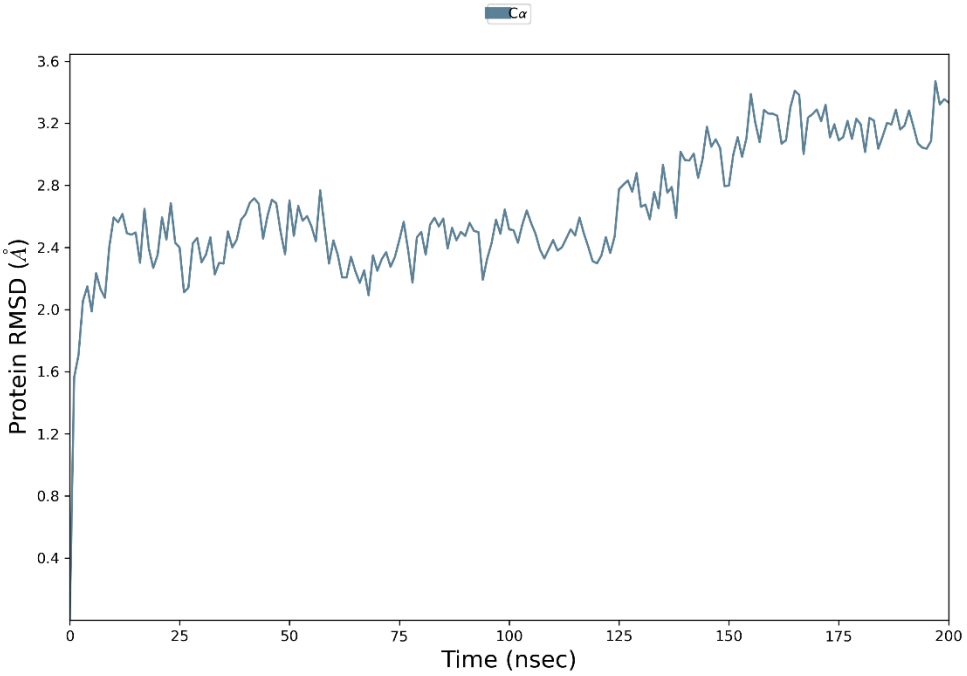


Figura 5. Gráfico de RMSD da simulação referente à proteína com a variante rs2297809

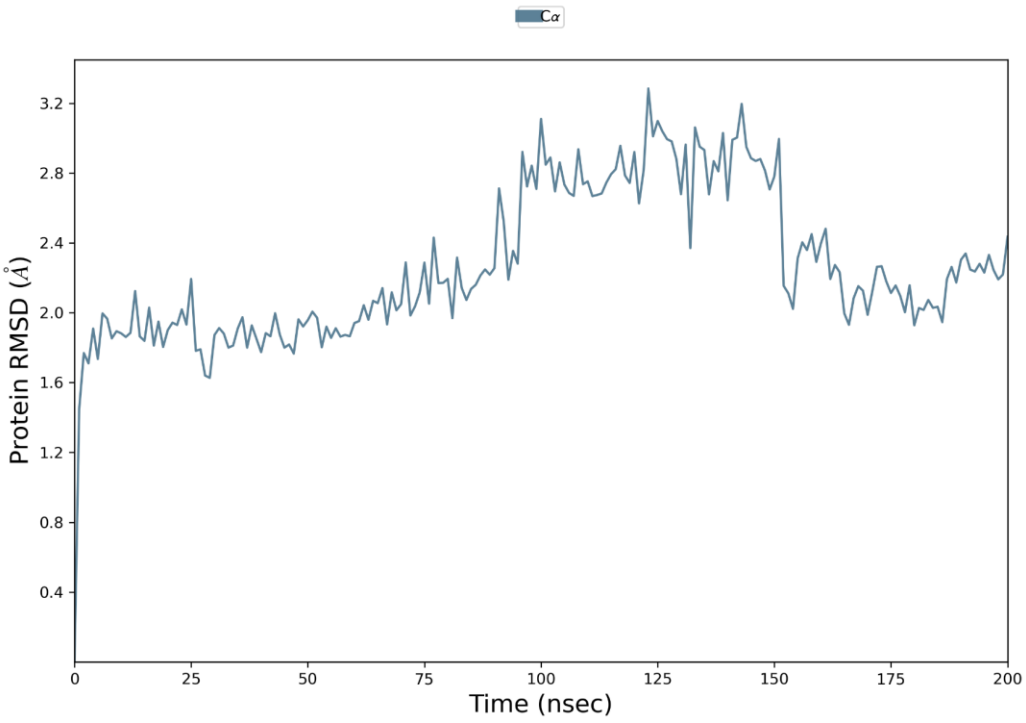
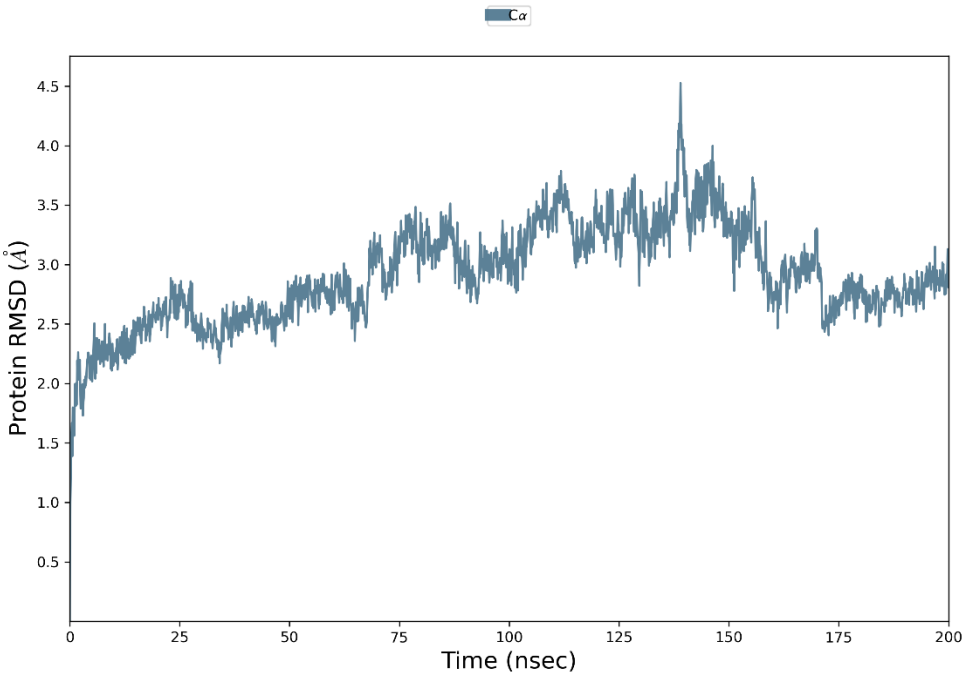


Figura 6. Gráfico de RMSD da simulação referente à proteína com as variantes rs2297810 e rs2297809



4.2. RMSF

O RMSF (*Root Mean Square Fluctuation*) é utilizado para caracterizar mudanças locais ao longo da cadeia proteica, podendo ser calculado para cada resíduo (DINDI et al., 2023). Elementos da estrutura secundária da proteína como alfa-hélices e folhas-beta geralmente possuem maior rigidez em relação a regiões não estruturadas da proteína como loops, apresentando menores flutuações.

Pode-se observar que na simulação com a variante rs2297810 há a presença de picos maiores na região entre os resíduos 240 a 260 e próximo ao resíduo 410 indicando maior flutuação em comparação à simulação com a proteína selvagem. Já na simulação com a variante rs2297809 os picos na região entre os resíduos 240 a 260 são menores em relação à proteína selvagem, mas também apresenta aumento próximo ao resíduo 410.

A simulação com as duas variantes demonstra similarmente um leve aumento próximo ao resíduo 410. Contudo, a maior diferença é observada entre os resíduos 240 a 260, havendo um pico mais definido ultrapassando os 6,4 Å. A partir disso, é possível inferir que essa região apresenta mudanças estruturais consequentes das

mutações, possivelmente havendo um efeito sinérgico entre elas, vez que há aumento de flutuação na região mesmo com a variante rs2297809 por si só apresentando sua diminuição, de forma que esse aumento não é resultado de apenas um efeito aditivo de seus impactos.

Corroborando com essa hipótese, esse mesmo efeito é observado nos gráficos de RMSD, nos quais se observa que a média de flutuação da simulação com a variante rs2297809 é menor em relação às simulações da proteína selvagem e da proteína com a variante rs2297810, enquanto a simulação com as duas proteínas visivelmente apresenta variações mais significativas.

Figura 7. Gráfico de RMSF da simulação referente à proteína selvagem

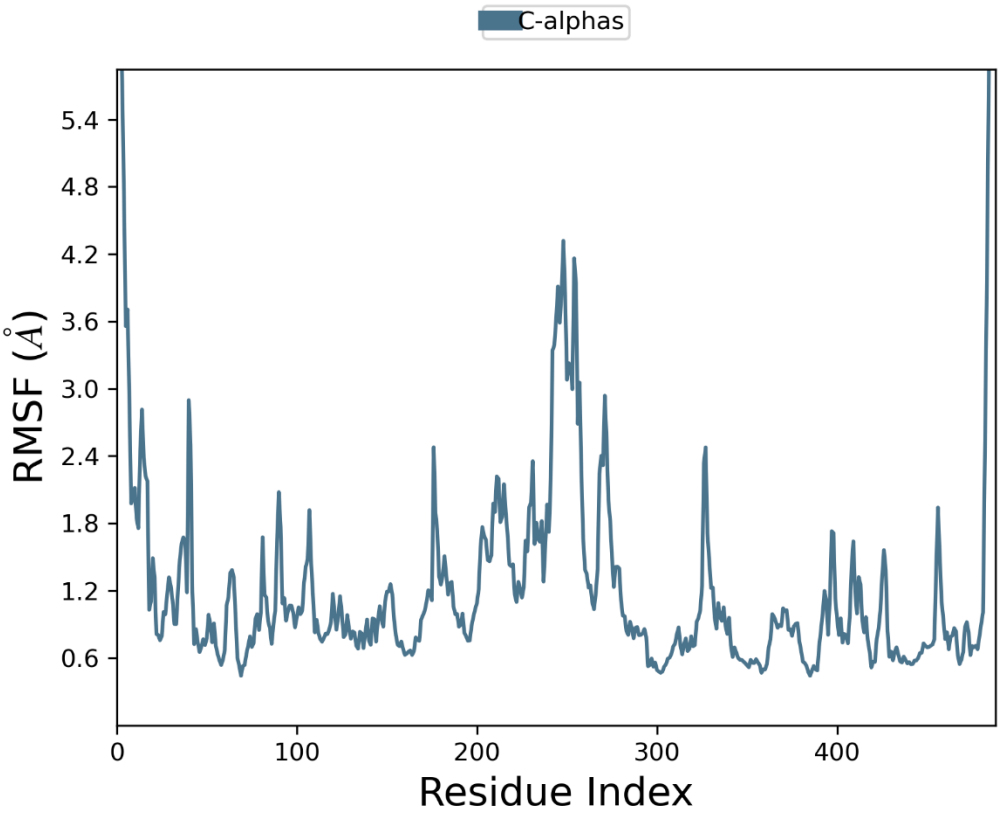


Figura 8. Gráfico de RMSF da simulação referente à proteína com a variante rs2297810

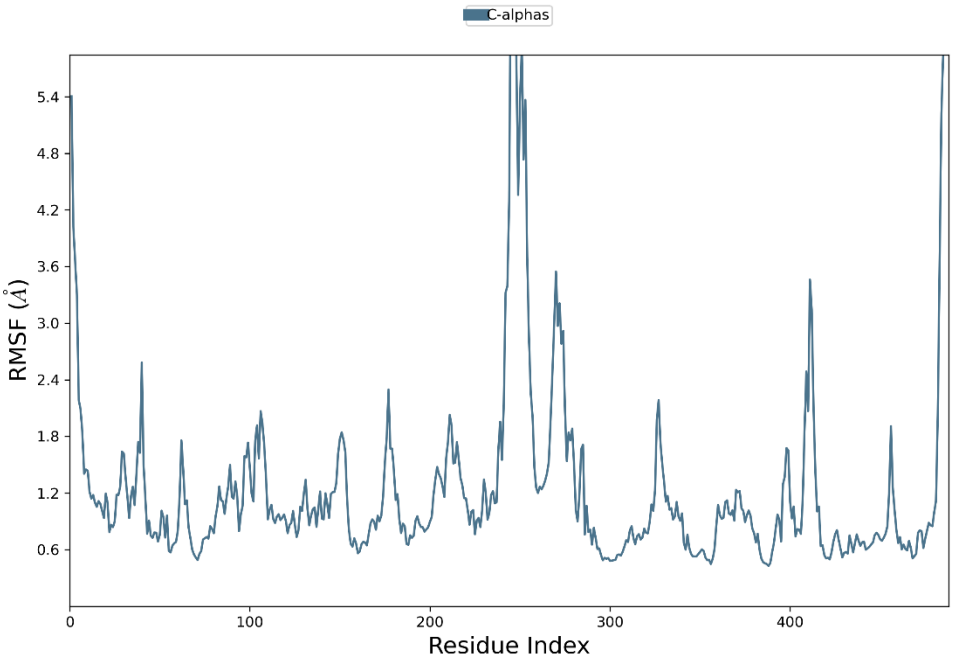


Figura 9. Gráfico de RMSF da simulação referente à proteína com a variante rs2297809

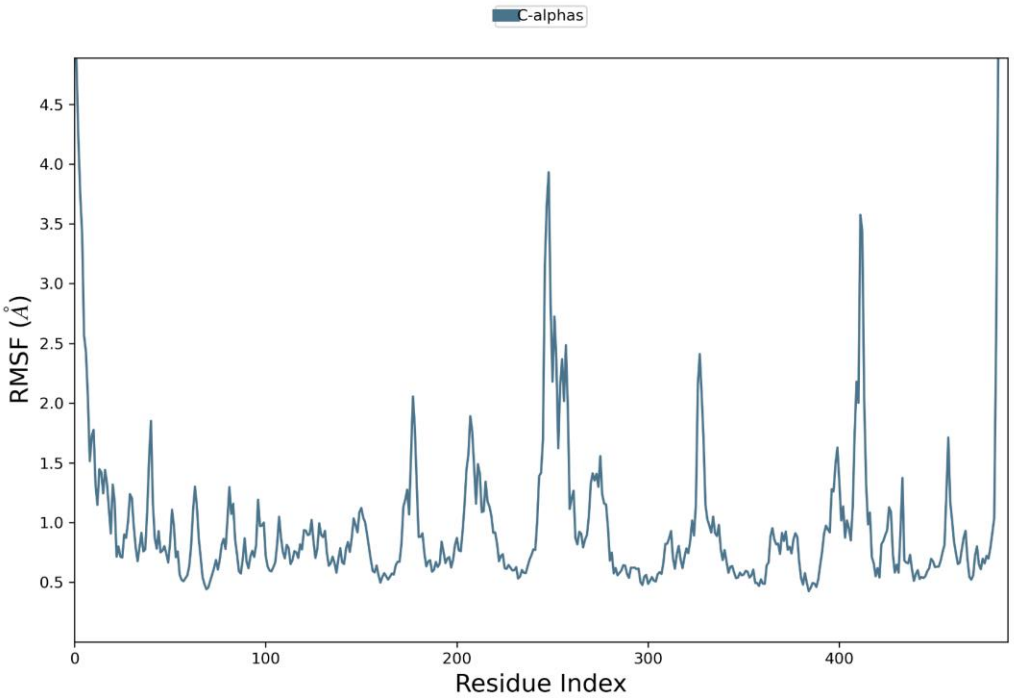
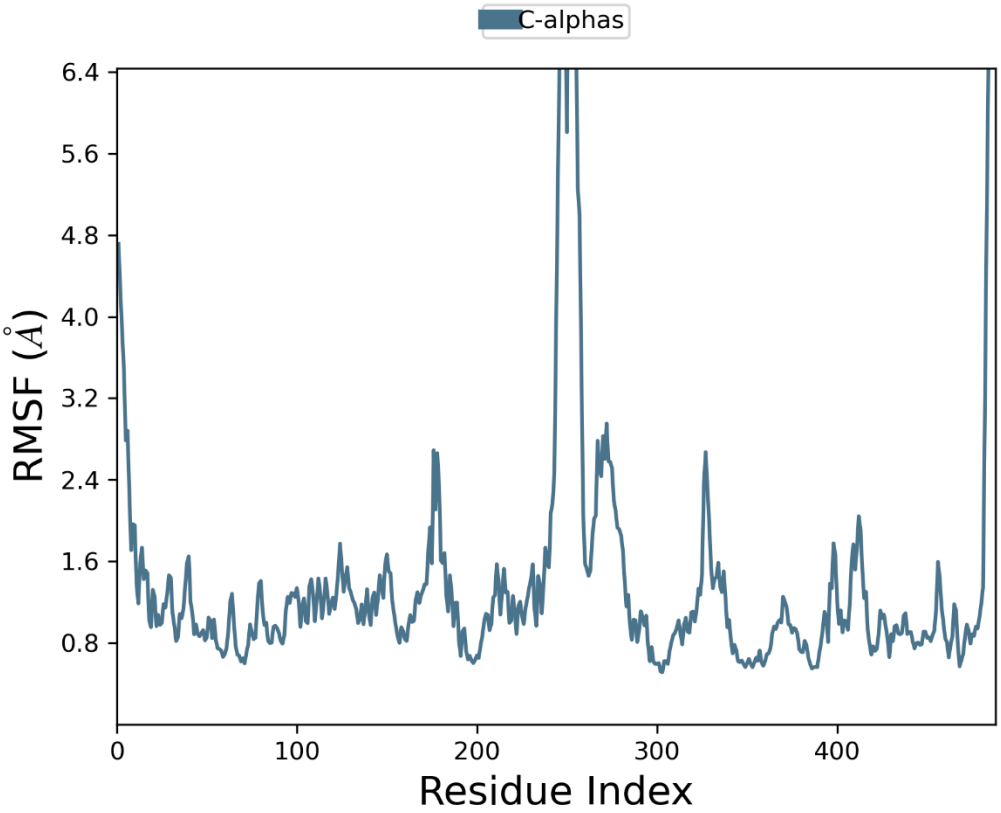


Figura 10. Gráfico de RMSF da simulação referente à proteína com as variantes rs2297810 e rs2297809



4.3. SSE

Elementos da estrutura secundária (SSE) da proteína como alfa-hélices e folhas-beta são monitorados ao longo da simulação.

Houve um aumento no SSE total nas simulações com a variante rs2297810 (46,31%), com a variante rs2297809 (49,80%) e com as duas variantes (48.07%) em relação à simulação com a proteína selvagem (44,71%), observando-se mais regiões de alfa-hélice.

Pode-se observar novamente, inclusive com a variante rs2297809, diferenças na região entre os resíduos 240 a 260, havendo nela um crescimento significativo em relação à SSE, sendo outra confirmação de que as variantes apresentam impacto na estrutura desse segmento da proteína, ainda que não sendo a região em que se encontram.

Figura 11. Gráfico de composição de SSE da simulação referente à proteína selvagem

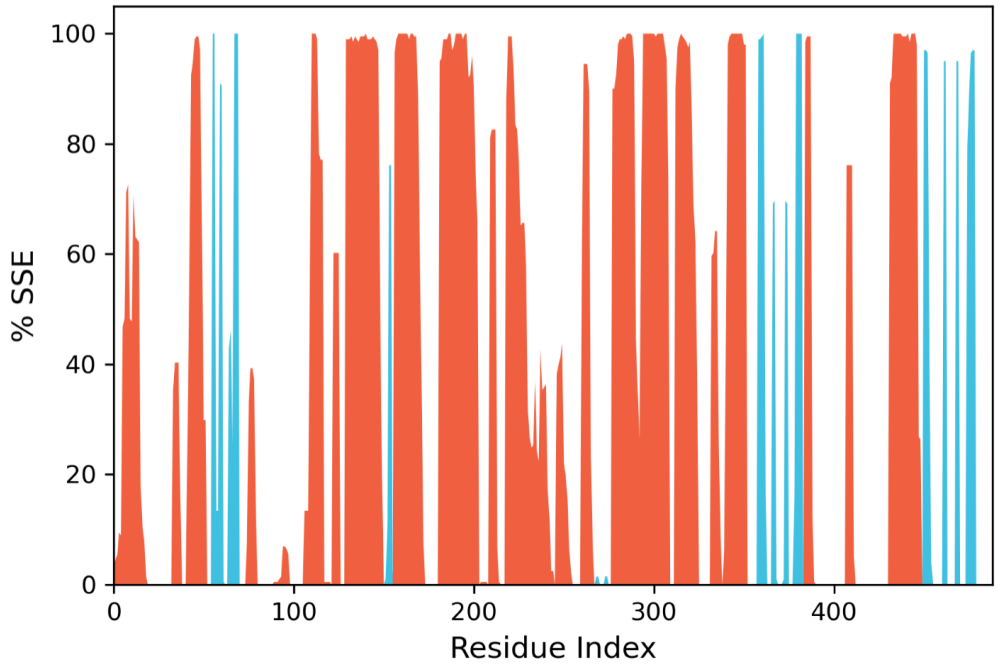


Figura 12. (A) Gráfico de composição de SSE de cada fragmento ao longo da simulação referente à proteína selvagem; (B) Gráfico de cada resíduo e seu SSE ao longo da simulação referente à proteína selvagem

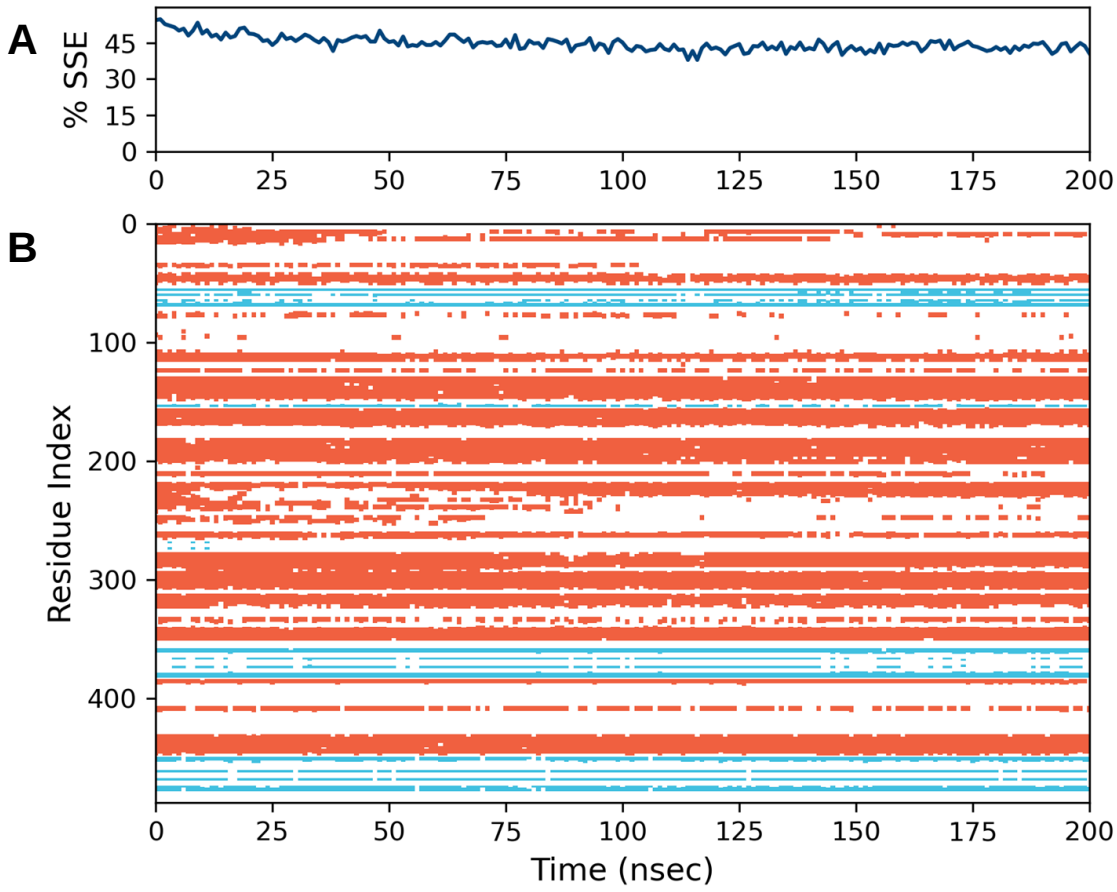


Figura 13. Gráfico de composição de SSE da simulação referente à proteína com a variante rs2297810

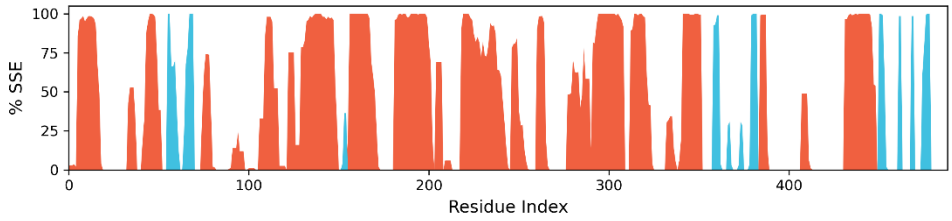


Figura 14. (A) Gráfico de composição de SSE de cada fragmento ao longo da simulação referente à proteína com a variante rs2297810; (B) Gráfico de cada resíduo e seu SSE ao longo da simulação referente à proteína com a variante rs2297810

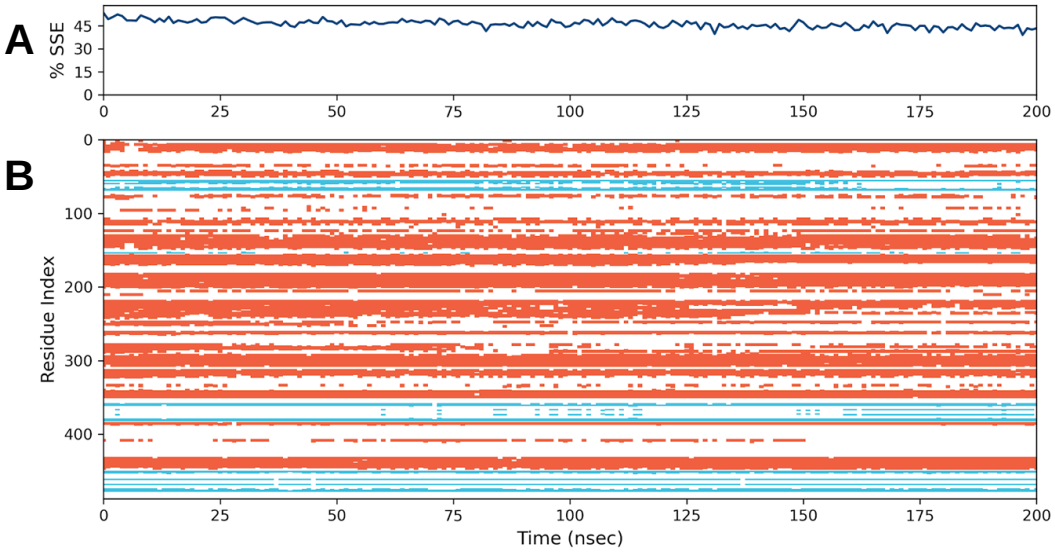


Figura 15. Gráfico de composição de SSE da simulação referente à proteína com a variante rs2297809

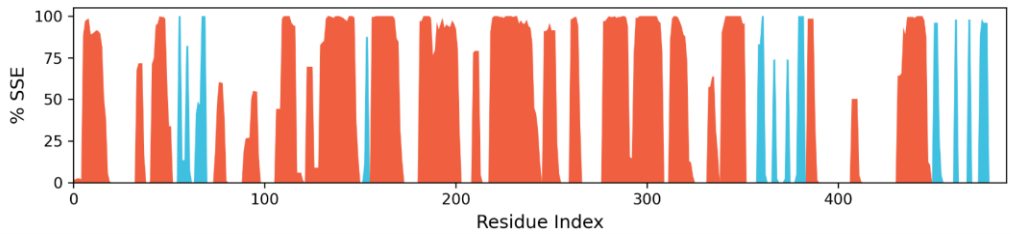


Figura 16. (A) Gráfico de composição de SSE de cada fragmento ao longo da simulação referente à proteína com a variante rs2297809; (B) Gráfico de cada resíduo e seu SSE ao longo da simulação referente à proteína com a variante rs2297809

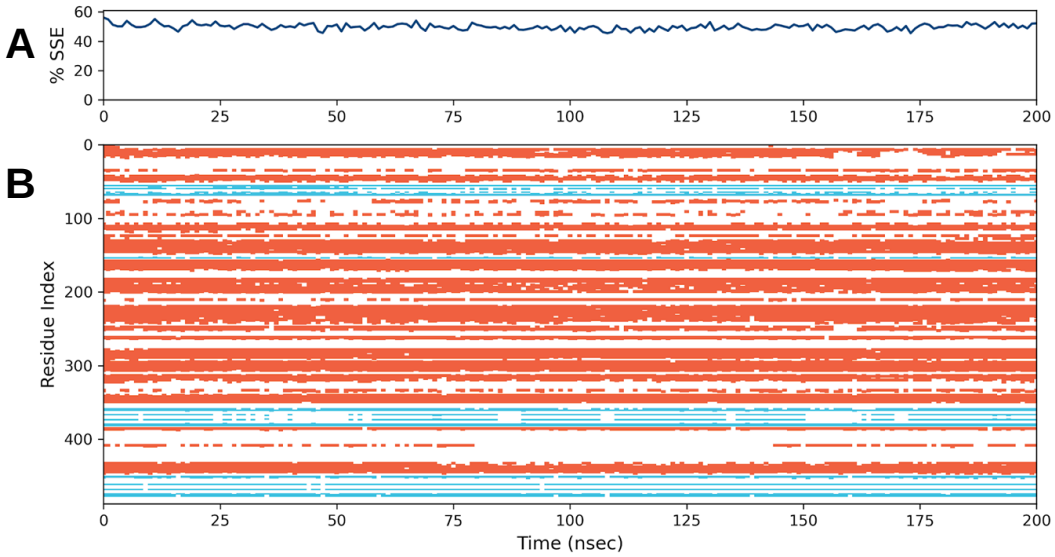


Figura 17. Gráfico de composição de SSE da simulação referente à proteína com as variantes rs2297810 e rs2297809

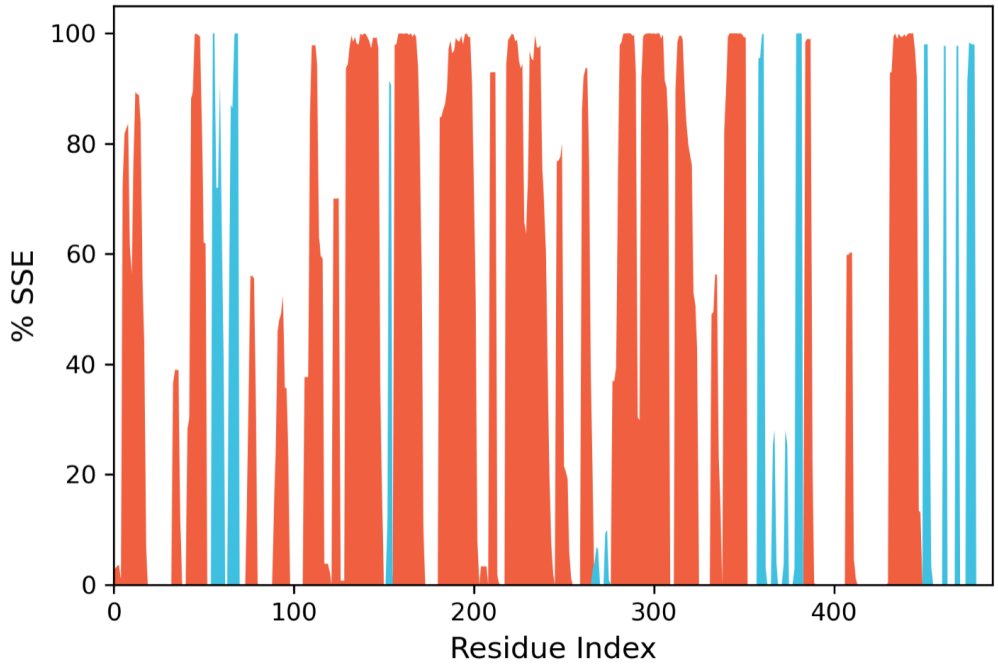
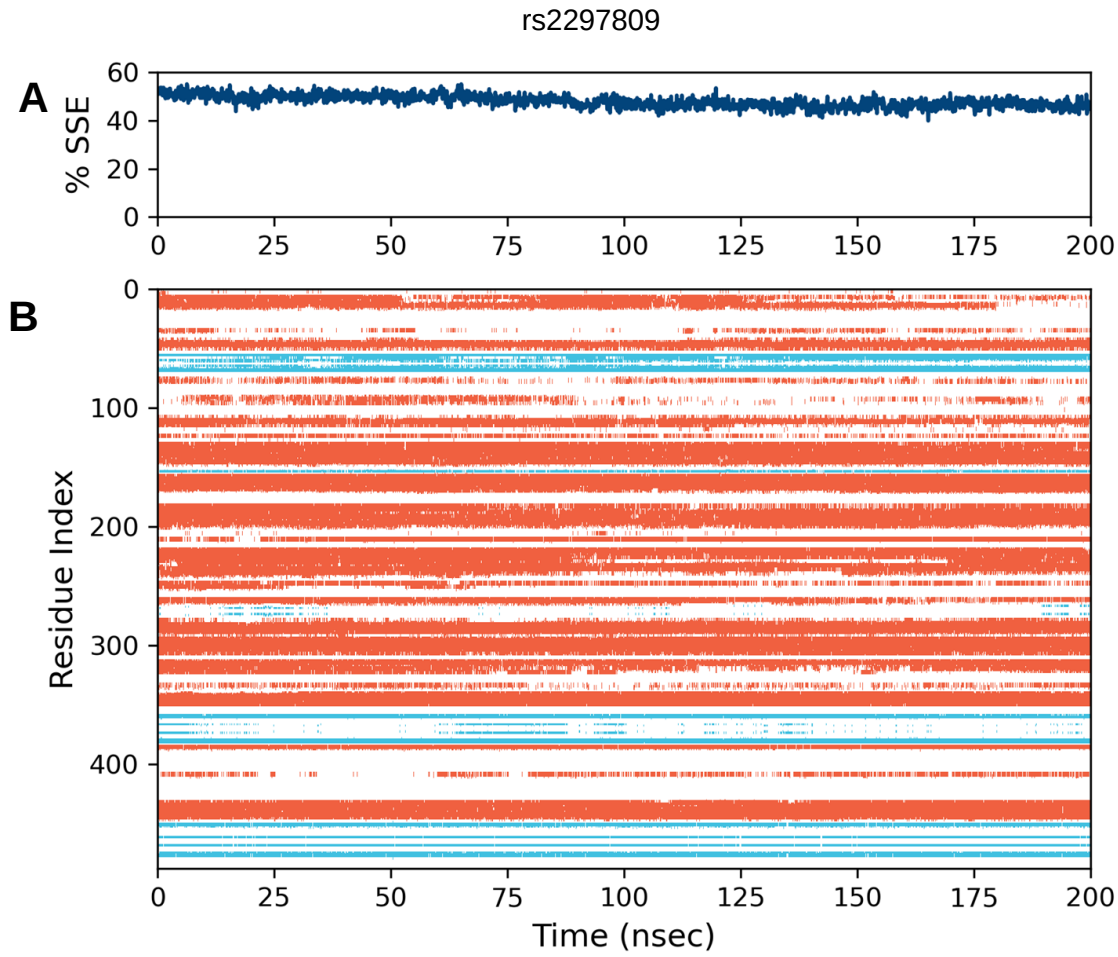


Figura 18. (A) Gráfico de composição de SSE de cada fragmento ao longo da simulação referente à proteína com as variantes rs2297810 e rs2297809; (B) Gráfico de cada resíduo e seu SSE ao longo da simulação referente à proteína com as variantes rs2297810 e



5. DISCUSSÃO

Os resultados e análises obtidos evidenciaram mais um exemplo de caso em que a co-ocorrência de variantes apresenta um efeito diferenciado em relação aos efeitos das ocorrências individuais das variantes e o efeito esperado resultante de apenas sua adição.

O estudo de SNPs atualmente é expressivo e uma importante ferramenta para a contribuição na área de genética molecular. Ainda assim, as pesquisas e análises são indiscutivelmente voltadas para avaliações individuais, mesmo quando investigando múltiplas variantes. Nos estudos da *CYP4B1* e das variantes empregadas neste projeto (YANG et al.,2023; YU et al., 2022) seus impactos fenotípicos são avaliados separadamente, havendo a construção de uma correlação entre elas apenas em relação à frequência genotípica.

A utilização de SNPs para diagnósticos e terapias já é alvo de discussão (ALLEMAILEM et al., 2021) e aplicação (AHMED et al., 2020; FORDE et al., 2023) e o entendimento da interação entre diferentes variantes que o mesmo indivíduo pode portar permitiria grandes avanços e potencialização da eficácia desses métodos, já auxiliados pela maior facilidade de realização de sequenciamentos de DNA atualmente.

Da mesma forma que mutações em uma região da proteína podem gerar consequências estruturais em regiões distantes devido a interações no plano tridimensional, interações com ligantes ou mesmo com outros componentes do sistema - como visto com as mudanças na região entre os resíduos 240 a 260 decorrentes de variações nas posições 331 e 340 - é de grande relevância a investigação do todo para a elucidação das consequências finais.

6. CONCLUSÃO

A partir dos resultados das análises é possível inferir que a presença das variantes apresenta consequências estruturais, gerando maiores mudanças conformacionais e mudanças na estrutura secundária da proteína. Essas alterações se mostraram aumentadas nas simulações com a co-ocorrência das variantes, as quais demonstraram que a interação entre as variantes rs2297810 e rs2297809 na CYP4B1 ocorre de maneira sinérgica, com o efeito da co-ocorrência sendo maior que a adição de seus efeitos individuais.

7. REFERÊNCIAS

ANDERSSON, E. et al. **Assessing How Multiple Mutations Affect Protein Stability Using Rigid Cluster Size Distributions**. [s.l: s.n.].

ATILGAN, C.; ATILGAN, A. R. **Perturbation-Response Scanning Reveals Ligand Entry-Exit Mechanisms of Ferric Binding Protein. PLOS Computational Biology**, v. 5, n. 10, p. e1000544, 23 out. 2009.

BALASUBRAMANIAN, S. et al. **Determining the impact of putative loss-of-function variants in protein-coding genes**. bioRxiv, , 7 fev. 2017.

CORDELL, H. J.; CLAYTON, D. G. **Genetic association studies**. Lancet (London, England), v. 366, n. 9491, p. 1121–1131, 24 set. 2005.

DEHGHANPOOR, R. et al. **Predicting the Effect of Single and Multiple Mutations on Protein Structural Stability**. Molecules (Basel, Switzerland), v. 23, n. 2, p. E251, 27 jan. 2018.

DRYSDALE, C. M. et al. **Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness**. Proceedings of the National Academy of Sciences of the United States of America, v. 97, n. 19, p. 10483–10488, 12 set. 2000.

FISH, A. E.; CAPRA, J. A.; BUSH, W. S. **Are Interactions between cis-Regulatory Variants Evidence for Biological Epistasis or Statistical Artifacts?** The American Journal of Human Genetics, v. 99, n. 4, p. 817–830, 6 out. 2016.

GHEBRANIOUS, N. et al. **Detection of ApoE E2, E3 and E4 alleles using MALDI-TOF mass spectrometry and the homogeneous mass-extend technology**. Nucleic Acids Research, v. 33, n. 17, p. e149, 1 set. 2005.

HOEHE, M. R. et al. **Multiple haplotype-resolved genomes reveal population patterns of gene and protein diplotypes.** Nature Communications, v. 5, n. 1, p. 5569, 26 nov. 2014.

HOEHE, M. R. et al. **Significant abundance of cis configurations of coding variants in diploid human genomes**. Nucleic Acids Research, v. 47, n. 6, p. 2981–2995, 8 abr. 2019.

LE GUEN, Y. et al. **Association of Rare APOE Missense Variants V236E and R251G With Risk of Alzheimer Disease.** JAMA neurology, v. 79, n. 7, p. 652–663, 1 jul. 2022.

LEE, S. et al. **Rare-Variant Association Analysis: Study Designs and Statistical Tests**. American Journal of Human Genetics, v. 95, n. 1, p. 5–23, 3 jul. 2014.

LIU, M.; WATSON, L. T.; ZHANG, L. **Predicting the combined effect of multiple genetic variants**. Human Genomics, v. 9, n. 1, p. 18, 30 jul. 2015.

LOS, B. et al. **Effects of PCSK9 missense variants on molecular conformation and biological activity in transfected HEK293FT cells**. Gene, v. 851, p. 146979, 17 out. 2022.

MARABOTTI, A.; SCAFURI, B.; FACCHIANO, A. **Predicting the stability of mutant proteins by computational approaches: an overview.** Briefings in Bioinformatics, v. 22, n. 3, p. bbaa074, 1 maio 2021.

MORRIS, A. P.; ZEGGINI, E. **An evaluation of statistical approaches to rare variant analysis in genetic association studies.** Genetic Epidemiology, v. 34, n. 2, p. 188–193, 2010.

NASLAVSKY, M. S. et al. **Whole-genome sequencing of 1,171 elderly admixed individuals from Brazil**. Nature Communications, v. 13, n. 1, p. 1004, 4 mar. 2022.

OSADA, N.; MIYAGI, R.; TAKAHASHI, **A. Cis- and Trans-regulatory Effects on Gene Expression in a Natural Population of Drosophila melanogaster.** Genetics, v. 206, n. 4, p. 2139–2148, ago. 2017.

RASZEK, M. M. **Application of Exome Sequencing to Mendelian Disorders and the Emergence of Personalised Medicine.** Em: eLS. [s.l.] John Wiley & Sons, Ltd, 2013.

RICHARDS, S. et al. **Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.** Genetics in medicine: official journal of the American College of Medical Genetics, v. 17, n. 5, p. 405–424, maio 2015.

SAINT PIERRE, A.; GÉNIN, E. **How important are rare variants in common disease?** Briefings in Functional Genomics, v. 13, n. 5, p. 353–361, 1 set. 2014.

SHEN, L. et al. ngs.plot: **Quick mining and visualization of next-generation sequencing data by integrating genomic databases.** BMC Genomics, v. 15, n. 1, p. 284, 15 abr. 2014.

SUHRE, K.; GIEGER, C. **Genetic variation in metabolic phenotypes: study designs and applications.** Nature Reviews Genetics, v. 13, n. 11, p. 759–769, nov. 2012.

WRIGHT, C. F.; FITZPATRICK, D. R.; FIRTH, H. V. **Paediatric genomics: diagnosing rare disease in children.** Nature Reviews Genetics, v. 19, n. 5, p. 253–268, maio 2018.

ZHANG, S. et al. **ProDy 2.0: Increased Scale and Scope after 10 Years of Protein Dynamics Modelling with Python.** Bioinformatics (Oxford, England), p. btab187, 5 abr. 2021.

ZOGHBI, A. W. et al. **High-impact rare genetic variants in severe schizophrenia.** Proceedings of the National Academy of Sciences, v. 118, n. 51, p. e2112560118, 21 dez. 2021.

BOWERS, K. J. et al. **Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters.** p. 43–43, 28 fev. 2006.

DE VIVO, M. et al. R**ole of Molecular Dynamics and Related Methods in Drug Discovery.** Journal of Medicinal Chemistry, v. 59, n. 9, p. 4035–4061, 12 maio 2016.

DINDI, U. M. R. et al. **In-silico and in-vitro functional validation of imidazole derivatives as potential sirtuin inhibitor.** Frontiers in Medicine, v. 10, p. 1282820, 2023.

YANG, Y. et al. **Missense variants in CYP4B1 associated with increased risk of lung cancer among Chinese Han population.** World Journal of Surgical Oncology, v. 21, n. 1, p. 352, 1 dez. 2023.

YU, S. et al. **Case-control study on CYP4B1 gene polymorphism and susceptibility to gastric cancer in the chinese Han population**. BMC Medical Genomics, v. 15, n. 1, p. 223, 1 dez. 2022.

ALLEMAILEM, K. S. et al. **Single nucleotide polymorphisms (SNPs) in prostate cancer: its implications in diagnostics and therapeutics**. American Journal of Translational Research, v. 13, n. 4, p. 3868, 2021.

AHMED, Z. et al. **Human gene and disease associations for clinical-genomics and precision medicine research.** Clinical and Translational Medicine, v. 10, n. 1, p. 297–318, 1 mar. 2020.

FORDE, B. M. et al. **Clinical Implementation of Routine Whole-genome Sequencing for Hospital Infection Control of Multi-drug Resistant Pathogens.** Clinical Infectious Diseases, v. 76, n. 3, p. e1277–e1284, 8 fev. 2023.

Camila Takase

São Paulo, 18 de maio de 2025

Aluna: Camila Hosoe Takase

São Paulo, 18 de maio de 2025

Orientador: Prof. Dr. Michel Satya Naslavsky