

## MUSTER - SISTEMA DE MINERAÇÃO DE CONTATOS EM COMPLEXOS PROTEÍNA-LIGANTES: ESTUDO DE CASO EM COVID-19

Cláudio G. Matos Júnior <sup>1</sup> (IC), Carlos Henrique da Silveira (PQ) <sup>2</sup>, João Paulo R. Romanelli (PQ) <sup>3</sup>  
<sup>1,2,3</sup> Universidade Federal de Itajubá

**Palavras-chave:** antivirais, inteligência artificial, simulação, mineração de dados

### Introdução

O presente trabalho faz parte do projeto de pesquisa “Previsão *in silico* de candidatos a fármacos antivirais como terapêutica auxiliar da COVID-19” financiado pela VALE S.A através do convênio Unifei-Vale de Iniciação em Engenharia.

A COVID-19 é uma doença pandêmica grave que vitimou milhões de pessoas [2]. Ainda que cerca de 80% dos casos sejam assintomáticos ou apresentem sintomas leves que não demandem intervenções hospitalares, certos indivíduos podem desenvolver uma resposta inflamatória exacerbada, principalmente nos pulmões, provocando pneumonia severa, só tratáveis em unidades de tratamento intensivo (UTI). Dessa forma, projetar fármacos antivirais eficazes é de extrema importância.

Assim, este trabalho tem como objetivo projetar novos candidatos a fármacos como terapêutica auxiliar no combate à COVID-19, podendo ser aplicado a outros vírus de potencial pandêmico, através de técnicas de inteligência artificial (IA), mineração de dados e simulações moleculares. Nesta etapa da pesquisa o *script* *muster* foi aprimorado para mineração de contatos em complexos proteína-alvo do Sars-Cov-2.

### Metodologia

Inicialmente, através da Interface de Programação de Aplicação (API) de dados do *Protein Data Bank* (PDB), foi criado um banco de dados local com os arquivos .pdb com ligantes de interesse da Sars-Cov-2 obtidos pelo aluno de Iniciação Científica do grupo de pesquisa Wellington G. Silva. Um PDB é um banco de dados 3D de proteínas e ácidos nucleicos.

Obtida a lista de PDBs e criado o banco de dados, foi desenvolvido um *script* na linguagem de programação R para obtenção dos identificadores dos ligantes de interesse, obtidos anteriormente, através dos PDBs. Esse *script* faz a leitura da tabela de dados com o nome e ligante de interesse dos PDBs e devolve uma tabela com o número de identificação dos ligantes com estrutura adequada para ser utilizada no *muster*.

O *muster* realiza, basicamente, cinco etapas:

importação de ligantes e alvos, filtragem, alinhamento dos complexos ligantes-alvo, cálculo das áreas de contato e clusterização conforme densidade dos contatos.

Na primeira etapa o *script* realiza a importação dos arquivos .pdb previamente inseridos no diretório local de PDBs a partir da tabela de dados inseridos inicialmente.

Na segunda etapa é realizada uma filtragem nos arquivos PDBs, realizando adequações em suas estruturas, necessárias para a execução dos passos seguintes.

Na terceira etapa é realizado o alinhamento dos complexos proteína-alvo utilizando o software PyMOL através de um *script* Python. O PyMOL é um sistema de visualização molecular de código aberto e é possível, através de comandos pré-inseridos em um *script* R, realizar a rotação e alinhamento das moléculas.

Na quarta e quinta etapa são aplicados algoritmos de aprendizado de máquina que tentam compor métricas sobre as similaridades de contatos atômicos.

### Resultados e discussão

Para este estudo de caso foram utilizados os PDBs 7NF5 (*Crystal Structure Of MG-132 Covalently Bound To The Main Protease*) e 6XQU (*Room-Temperature X-Ray Crystal Structure Of Sars-Cov-2 Main Protease*) que tem como ligantes, respectivamente, o PHQ-Leu-Leu-Leu-aldehyde que é um peptídeo modificado e o Boceprevir que é um inibidor clínico da protease da hepatite C, em estudo de reposicionamento contra COVID-19.

Na Figura 1 é possível observar o alinhamento dos complexos proteína-alvo realizados pelo *script* *muster*, após as etapas de importação, filtragem e alinhamento dos PDBs.

Após as etapas quatro e cinco é possível observar os resultados, que serão abordados aqui de duas maneiras.

Na Figura 2 é possível observar o grafo dos ligantes em que as arestas do grafo representam a área

de contato entre a proteína e o ligante.

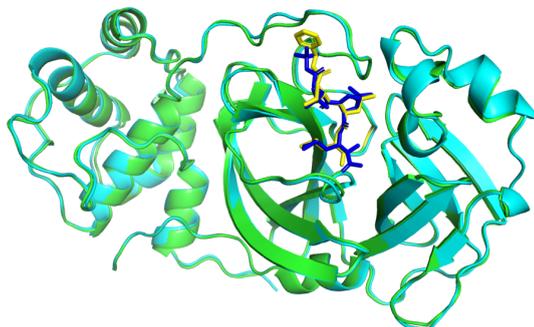


Figura 1 – Alinhamento da *main protease* (Mpro) do vírus Sars-Cov-2 e os ligantes PHQ-LLL-aldehyde (em amarelo) e Boceprevir (em azul).

Os átomos da Mpro estão coloridos em azul, em vermelho estão os átomos do ligante e em verde as águas cristalográficas.

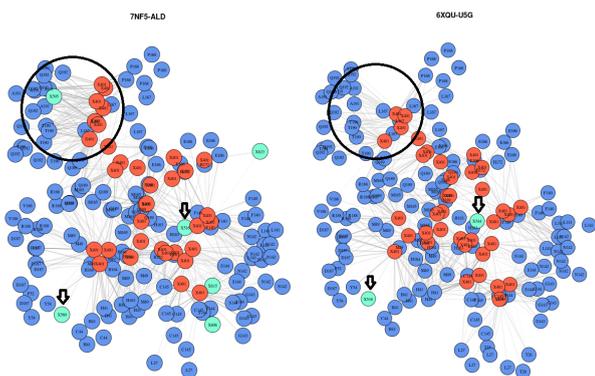


Figura 2 – Grafos de contatos de complexos ligante-alvo com os átomos do ligante (em vermelho), átomos do alvo (em azul) e águas cristalográficas (em ciano).

Um ponto importante a ser destacado é a diferença entre a densidade de contatos na mesma região para cada ligante, destacadas por um círculo na figura. É possível observar uma maior densidade de contatos para o PHQ-Leu-Leu-Leu-aldehyde. Também é possível observar a presença de uma água cristalográfica na mesma região, diferentemente do Boceprevir. Estes dois fatores podem ser diferenciais para o sucesso do ligante.

Outro ponto a se destacar é a presença de águas cristalográficas na mesma região nos dois ligantes indicadas pelas setas na imagem, fato que também ainda está em investigação.

Os resultados também podem ser observados através de uma nuvem de *tags* contendo os aminoácidos em que o tamanho da *tag* está associada à interação da alvo-ligante, ou seja, quanto maior a *tag* mais interações o aminoácido do ligante faz interações com o alvo. Na Figura 3 é possível observar a nuvem de *tags* geradas para os PDBs 7NF5 e 6XQU.

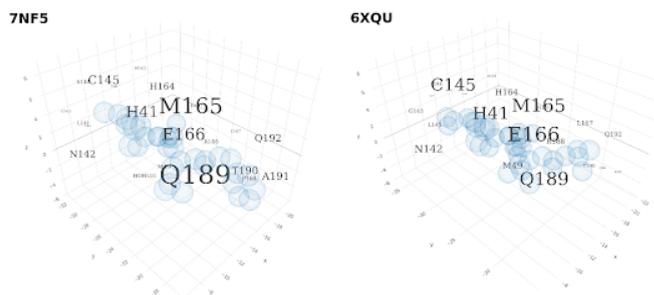


Figura 3 – Nuvem de tags para os PDBs 7NF5 e 6XQU.

É possível observar que, para o PDB 7NF5 a *tag* do aminoácido Q189 é maior em relação ao 6XQU, dessa forma apresenta uma maior densidade de contatos. Já no 6XQU é possível observar que a *tag* do aminoácido E166 é maior em relação ao 7NF5, indicando maior possibilidade de interação.

## Conclusões

Pode-se concluir que o *muster* pode ser uma ferramenta útil na análise dos contatos em complexos de proteínas ligantes, já que fornece uma grande gama de informações ao usuário, podendo ser útil na análise para a Sars-Cov-2 ou para outra doença de potencial pandêmico.

## Agradecimento

Agradeço primeiramente a VALE S.A. pela concessão da bolsa de Iniciação Científica que permitiu a realização do projeto de pesquisa. Agradeço também aos professores Carlos e João pela disponibilidade, paciência e vontade de ensinar.

## Referências

- [1] Mohammed J. Zaki, Wagner Meira, Jr., Data Mining and Machine Learning: Fundamental Concepts and Algorithms, 2nd Edition, Cambridge University Press, March 2020. ISBN: 978-1108473989.
- [2] Painel Covid-19 do centro de ciência e engenharia de sistemas (csse) da Johns Hopkins University, EUA. <https://gisanddata.maps.arcgis.com/apps/dashboards/bda7594740fd40299423467b48e9ecf6> Acessado em 27/10/2022.