Estratégia Algorítmica para a Reconstrução e Validação da Estrutura Molecular de Variantes do SARS-CoV-2

Clarice de Souza[1], João Bessa[1], Rosiane de Freitas[1], Micael Oliveira[2], Kelson Mota[2]

[1]Instituto de Computação - Universidade Federal do Amazonas

[2]Laboratório de Química Teórica e Computacional - Universidade Federal do Amazonas

{clarice, joao.bessa, rosiane}@icomp.ufam.edu.br

{micaeloliveira, kelsonmota}@ufam.edu.br

Resumo. *Aspectos matemático-computacionais e físico-químicos envolvidos na reconstrução da estrutura molecular tridimensional de proteínas do SARS-CoV-2 são abordados neste artigo, envolvendo a variante P.1 detectada em pacientes infectados em solo brasileiro, principalmente as do estado do Amazonas. Foi realizado um estudo sobre o impacto teórico da P.1 por intermédio da reconstrução estrutural de proteínas onde a mutagênese foi realizada computacionalmente e com o auxílio da implementação de um algoritmo de enumeração implícita de factibilidade, Branch-and-Prune, cujas soluções foram validadas através do uso do gráfico de Ramachandran. Desta forma, mesmo com a ausência de estruturas cristalográficas caracterizando estas mutações, pôde-se computacionalmente modelar uma estrutura tridimensional onde ao fim realizou-se o alinhamento estrutural com a cristalografia do complexo ACE2-RBD.*

Abstract. *Mathematical-computational and physical-chemical aspects involved in the reconstruction of the three-dimensional molecular structure of SARS-CoV-2 proteins are addressed in this article, involving the P.1 variant detected in patients infected in Brazilian soil, especially those in the state of Amazonas. A study was carried out on the theoretical impact of P.1 through the structural reconstruction of proteins where mutagenesis was performed computationally and with the aid of the implementation of an implicit enumeration algorithm of feasibility, Branch-and-Prune, whose solutions were validated through the use of the Ramachandran chart. In this way, even with the absence of crystallographic structures characterizing these mutations, it was possible to computationally model a three-dimensional structure where, at the end, the structural alignment was performed with the crystallography of the ACE2-RBD complex.*

1. Introdução

A pandemia da COVID-19 (Coronavirus Disease-2019) que assolou o mundo e o está fazendo mudar seus comportamentos como sociedade, é causada por um novo vírus da família *Coronavidae*, o SARS-CoV-2. Embora a comunidade científica reconheça que o vírus surgiu primeiramente nos morcegos, não está comprovado como e onde a transmissão do SARS-CoV-2 para seres humanos ocorreu. Em poucas semanas a doença se

espalhou pela China, atingindo outros países e continentes, tomando as proporções de uma pandemia, conforme decretado pela Organização Mundial de Saúde (OMS) no dia 11 de março de 2020 [Li et al. 2020].

A partir do coronavírus da China, da Ásia, foram observadas duas linhagens: A e B, no Brasil, as linhagens sequenciadas até então foram oriundas da linhagem B. Já foram catalogados no Brasil mais de 427 genomas do SARS-CoV-2, a partir do sequenciamento registrado, foi possível observar como o vírus se multiplica e determinar como se deu a evolução e a disseminação de três grandes linhagens do novo coronavírus em território brasileiro. Descobriu-se também no sequenciamento genético de quatro viajantes do estado do Amazonas, após sua ida ao Japão, uma nova linhagem intitulada B.1.1.28/P.1, este clado é constituído pelo conjunto de 3 (três) mutações na glicoproteína Spike: K417T, E484K e N501Y [Naveca et al. 2021].

Sequenciar o genoma do vírus e reconstruir partes ou completamente sua estrutura molecular propicia um maior entendimento de seu comportamento no organismo hospedeiro, o que auxilia no desenvolvimento de drogas e vacinas de combate ao mesmo. Por outro lado, o processo de reconstruir a estrutura 3D de macromoléculas traz desafios tecnológicos e matemático-computacionais complexos do ponto de vista teórico e prático e até hoje constitui um dos grandes desafios da ciência [Dill and MacCallum 2012]. Deste modo, este trabalho tem como enfoque as cepas do vírus SARS-CoV-2 detectadas no Amazonas, analisando estrategias algorítmicas para determinar e validar a estrutura tridimensional do vírus.

2. Cálculo da Estrutura Tridimensional

Na biologia molecular, muitas pesquisas tem como foco as proteínas e suas propriedades. Proteínas são estruturas fundamentais do sistemas biológicos e são constituídas por uma cadeia linear de aminoácidos que são determinantes na sua estrutura tridimensional [Dong and Wu 2002]. A conformação espacial da proteína oriunda do enovelamento da sequência de aminoácidos é fundamental na função desempenhada pela proteína [Lavor et al. 2011]. Com isso, métodos que descrevem precisamente a conformação espacial de proteínas tem utilizado técnicas computacionais e experimentais que evitem representações irreais da estrutura proteica.

A estrutura tridimensional pode ser obtida teoricamente por Mecânica Quântica e Clássica Molecular, experimentalmente através da Cristalografia de Raios-X e Ressonância Magnética Nuclear (do inglês *Nuclear Magnetic Resonance - NMR)*, e de maneira Híbrida é possível utilizar *Modelagem por Homologia* a partir do sequenciamento. No método de Raios-X as coordenadas atômicas são obtidas com grande precisão a partir da difração de elétrons sobre uma amostra cristalizada da proteína, permitindo, a partir de um refinamento matemático rigoroso, a obtenção da estrutura tridimensional [Newman 2006]. O método de RMN por sua vez é aplicado em estruturas complexas de serem cristalizadas, mas o sinal RMN fornece apenas um pequeno subconjunto de distâncias entre os átomos de uma molécula, pois apenas obtêm distâncias entre pares de átomos que estejam próximos na faixa de 5 a 6 Å [Fidalgo et al. 2012].

Independente do método de elucidação estrutural, o problema é estimar a estrutura completa da molécula, determinando a posição correta no espaço de todos os átomos que a compõe, também conhecido como Problema de Geometria de Distâncias Moleculares -

PGDM (do inglês, *Molecular Distance Geometry Problem* - MDGP) e formulado tradicionalmente como um problema de otimização contínua [Silva and Lavor 2008]. Determinar a estrutura tridimensional quando todas as distâncias entre os átomos são conhecidas é um problema polinomial, mas, partindo-se de um subconjunto incompleto de distâncias (como nos dados por RMN), passa a ser um problema NP-difícil [Silva and Lavor 2008]. O estudo do problema sob o enfoque de Geometria de Distâncias pode contribuir decisivamente na identificação da estrutura proteica mais viável, a partir de um conjunto de distâncias obtidas experimentalmente.

2.1. O Problema de Geometria de Distâncias Moleculares

Considerando uma molécula formada por *n* átomos $a_1, a_2, \ldots, a_n$ da qual são conhecidas um conjunto de distâncias d_{ij} entre pares de átomos a_i e a_j. O Problema de Geometria de Distâncias Moleculares (*Molecular Distance Geometry Problem - MDGP*) pode ser definido como a obtenção de uma configuração tridimensional $x_1, x_2, \ldots, x_n$ para a molécula respeitando as distâncias euclidianas conhecidas [Silva and Lavor 2008].

As coordenadas $x_1, x_2, \ldots, x_n$ podem ser obtidas a partir das distâncias d_{ij} através da resolução do sistema de equações do cálculo de distâncias: $||x_i - x_j|| = d_{ij} \quad \forall (i,j) \in S$, onde S é o conjunto de pares de átomos cuja distância d_{ij} é conhecida, sendo $x_i = (u_i, v_i, w_i)^T$ um vetor de coordenadas com u_i, v_i e w_i sendo a primeira, segunda e terceira coordenadas do átomo i e $||\cdot||$ a norma euclidiana. A estrutura é conseguida através da imersão dos átomos no R^3. Considerando a posição do ponto como $p = (u, v, w)$, imergir um ponto significa encontrar os valores das dimensões u, v e w para o ponto.

O MDGP pode ser classificado em duas formas [Fidalgo et al. 2012]: **Conjunto completo de distâncias** - todas as distâncias entre quaisquer pares de átomos são conhecidas, [Dong and Wu 2002] apresentaram um algoritmo polinomial que resolve este problema através de sistemas lineares; **Conjunto arbitrário de distâncias** - são conhecidas somente algumas distâncias entre átomos da molécula, essa versão é NP-completo para imersão em uma dimensão $MDGP_1$ e NP-difícil para imersão em dimensões maiores que 1 $MDGP_k$ para $k > 1$ [Liberti et al. 2011],[Maculan et al. 2010].

Existem vários algoritmos para resolver o MDGP, a grande maioria deles resolve a versão contínua clássica do problema, como o *Geometric Build-up*. Entretanto [Lavor et al. 2011] propuseram um modelo discreto, o *Discretizable Molecular Distance Geometry Problem* (DMDGP). Para esta versão a principal forma de resolução é o *Branch and Prune* (BP) que é um método de enumeração implícita, que enumera todas as possíveis posições dos átomos e descarta as inválidas.

2.2. Gráfico de Ramachandran

O gráfico de Ramachandran descreve os ângulos de torção $\phi - \psi$ do backbone proteico, fornecendo uma visão geral da conformação de uma proteína. Os ângulos de torção ϕ (phi) e ψ (psi) são definidos para cada um dos resíduos de aminoácidos. São ângulos que definem rotação, sendo que o ϕ define a rotação em torno da ligação $C_\alpha - N$ do resíduo, e ψ define a rotação em torno da ligação $C_\alpha - C$ do mesmo resíduo [Berg et al. 2012]. Vários valores são proibidos por impedimento estérico entre os átomos do backbone e as cadeias laterais dos aminoácidos e somente alguns valores reproduzem com razoável precisão a conformação espacial da proteína, sendo necessário descartar resultados que se

mostrem proibidos ou irreais. Teoricamente para validar se a conformação das proteínas é a melhor possível, o uso do diagrama de Ramachandran tem sido considerado muito útil, uma vez que testa a qualidade das estruturas tridimensionais [Nelson and Cox 2017].

Neste trabalho demos o enfoque à variante P.1 do vírus SARS-CoV-2 detectada inicialmente no Amazonas, onde analisamos estratégias algorítmicas para determinar e validar a estrutura tridimensional da região RBD da proteína Spike. Desta forma, utilizou-se uma implementação do algoritmo Branch-and-Prune que a partir da simulação de dados RMN pôde reconstruir a estrutura contendo a variante e que foi posteriormente validada pelo diagrama de Ramachandran.

3. Metodologia e Experimentos

Devido a indisponibilidade de informações mais precisas das mutações, como por exemplo, as estruturas cristalográficas com as mutações que compõem a linhagem P.1 no estado do Amazonas, foi adotada a metodologia mostrada na Figura 1 para o cálculo estrutural das variantes.

Primeiramente é realizada uma busca do arquivo original sem mutação da proteína do vírus, no banco de dados de proteínas RCSB PDB [Berman et al. 2000]. É então aplicada a mutação *in silico* utilizando o auxílio do software PyMol 2.3 [Schrödinger, LLC 2015] com o módulo *"Mutagenesis"*, tendo como rotâmero o de menor tensão estérica definido automaticamente pelo software. Na terceira etapa foram geradas instâncias para o problema simulando um conjunto arbitrário de distância, com distâncias menores que 6 Å, simulando assim os dados da NMR. Para robustecer os testes, foram geradas instâncias utilizando somente os átomos do backbone e algumas utilizando backbone e cadeia lateral.

A etapa de reconstrução estrutural foi subdividida em cálculo estrutural e validação estrutural. Utilizou-se o algoritmo *Branch-and-Prune* para o cálculo estrutural das variante P.1 do vírus. As soluções válidas e inválidas foram originadas em pares a partir da árvore binária *Branch-and-Prune*. Como o algoritmo de predição estrutural pode gerar várias possíveis estruturas matematicamente válidas como solução, mas não garante a validação química de cada estrutura gerada, uma vez que utiliza somente as informações de distâncias entre átomos para imersão dos vértices no plano, foi acrescentada uma etapa de validação físico-química. Para essa validação estrutural foi utilizado o Gráfico de Ramachandran, que verifica a conformação das proteínas (Seção 2.2).

A plataforma adotada para gerar os diagramas de Ramachandran foi MolProbity (`https://swissmodel.expasy.org/assess`) com o módulo *"Structure Assessment"* que foi implementado na plataforma SWISS-Model [Waterhouse et al. 2018] pela qual foi possível obter a porcentagem de aminoácidos pertencentes às regiões favoráveis e com restrições estéricas. As variantes do SARS-CoV-2 escolhidas para teste nesse trabalho, são mostradas na Tabela 1. Após a reconstrução estrutural comparamos a energia potencial das soluções obtidas com auxílio do plugin NAMDEnergy no software VMD 1.9.4a.51 e analisados após a minimização estrutural em 10^4 interações pelo método do gradiente conjugado no algoritmo NAMD3 [Phillips et al. 2020].

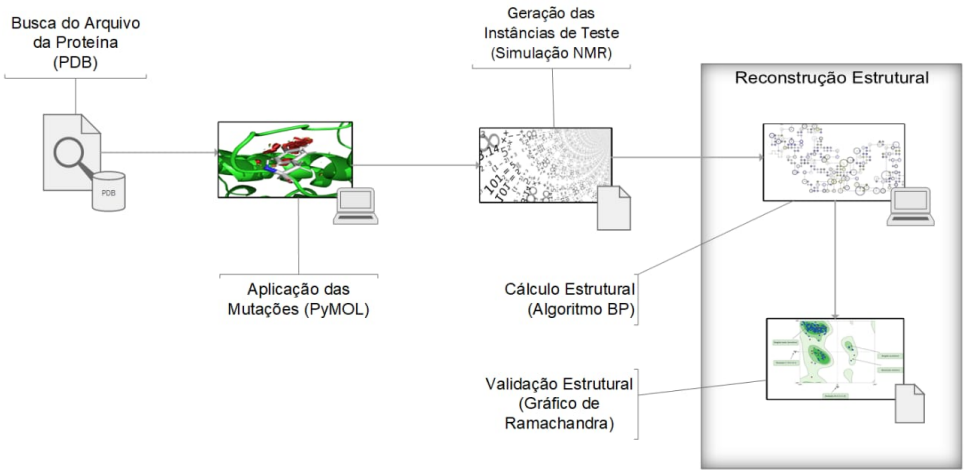

Figura 1. Fluxo do cálculo estrutural das variantes encontradas no Amazonas.

Tabela 1. Informações das estruturas estudadas neste trabalho [Wang et al. 2020, Ju et al. 2020].

Proteína	Massa (kDa)	Átomos	Resíduos	Resolução	Região
6M0J	97,14	6571	832	2,45 Å	Complexo ACE2-RBD
7BWJ	71,63	4820	662	2,85 Å	Anticorpo-antígeno

3.1. Estudo da linhagem amazonense B.1.1.28 de clado P.1

Foram calculadas e analisadas todas as possibilidades de estrutura da árvore de geração que o algoritmo BP constrói, todas as estruturas geradas respeitaram as restrições matemáticas do problema, mas algumas delas foram consideradas quimicamente inválidas quando feriam as restrições físico-químicas das proteínas ao serem gerados os gráfico de Ramachandran.

O primeiro teste realizado foi para a instância do complexo ACE2-RBD e gerou duas estruturas tridimensionais como soluções (estrutura X,Y,Z), que respeitam as restrições matemáticas de distância. Contudo, ao serem gerados os gráficos de Ramachandran para as duas soluções, foi verificado que a primeira foi totalmente invalidada porque gerou um gráfico com praticamente todos os pontos em regiões erradas. A segunda entretanto, gerou um gráfico de Ramachandran que respeitava as regiões válidas, sendo assim uma estrutura aceitável.

A estrutura para o complexo ACE2-RBD (PDB ID: 6M0J) [Wang et al. 2021] contendo as mutações da linhagem P.1, considerada quimicamente válida apresentou um total de $97,06\%$ dos aminoácidos na região sem impedimentos estéricos, isto portanto reflete a grande capacidade de reconstrução estrutural por parte do algoritmo implementado neste trabalho. Na ausência de mutações o número de resíduos na região mais favorável aumentou para $97,08\%$. Consequentemente as mutações contribuem para o surgimento de conflitos estéricos na estrutura. Por fim, ao analisar o mesmo complexo a solução mais consistente apresentou $2,92\%$ dos ângulos de torção em condições marginais enquanto $0,38\%$ em condições de total impedimento estérico.

O teste para o anticorpo-antígeno (PDB ID: 7BWJ) [Ju et al. 2020] seguiu o

padrão do primeiro, geraram duas estruturas 3D como solução, contudo após validação uma delas mostrou-se inconsistente. A estrutura reconstruída apresentou $93,97\%$ dos aminoácidos em uma região sem impedimentos estéricos. Desta forma ainda que o algoritmo não seja capaz de remover os conflitos estéricos induzidos pela mutagênese *in silico*, a reconstrução estrutural aproximou-se de forma satisfatória da estrutura cristalográfica.

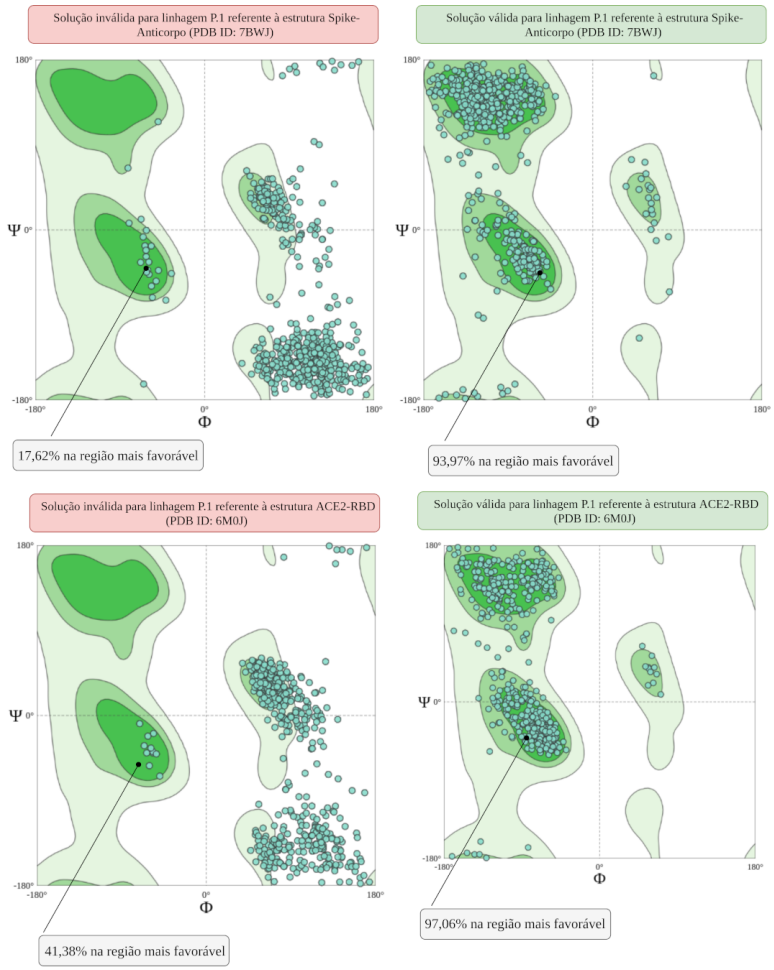

Figura 2. Diagramas de Ramachandran para as soluções obtidas no algoritmo implementado para a linhagem P.1.

Por fim, realizamos a sobreposição estrutural (ver Figura 3) entre o complexo ACE2-RBD (PDB ID: 6M0J) [Wang et al. 2020] com as mutações P.1 inseridas computacionalmente e reconstruídas via algoritmo, e alinhadas com a estrutura cristalográfica da P.1 (PDB ID: 7NXC) [Dejnirattisai et al. 2021]. Sendo assim, o RMSD para a solução válida entre Cα foi de $2,8084$Å. Por outro lado, a solução inválida teve um aumento drástico para $21,610$Å. Estes resultados indicam portanto a validade e consistência do algoritmo. Por último, a estrutura válida via diagrama de Ramachandran para o complexo ACE2-RBD (PDB ID: 6M0J) também apresentou uma energia potencial total mais favorável e estável de $-27912,2kcal \cdot mol^{-1}$ (ver Tabela 2) enquanto a estrutura inválida de $-17684,4kcal \cdot mol^{-1}$. Logo podemos concluir que as estruturas mais consistentes são também aquelas que possuem a mínima energia potencial.

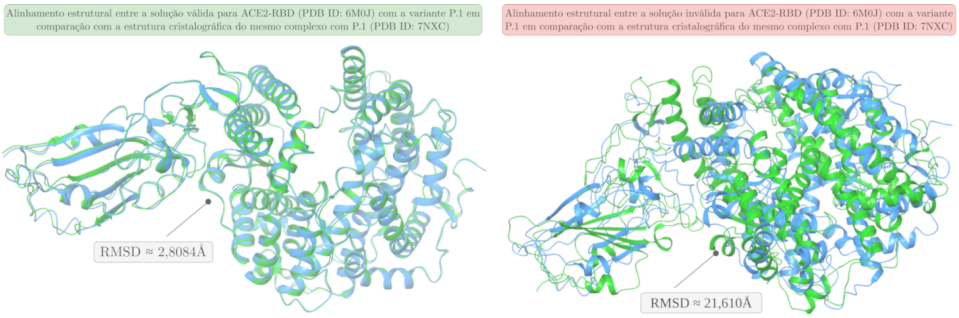

Figura 3. Alinhamento estrutural no software Schrödinger Maestro 2020-4 entre soluções válidas e inválidas contendo P.1 (PDB ID: 6M0J) em comparação com a respectiva estrutura cristalográfica do complexo ACE2-RBD com a P.1 (PDB ID: 7NXC).

Tabela 2. Comparativo entre os valores de energia potencial entre as soluções válidas e inválidas ao complexo ACE2-RBD.

Solução	E_{VdW}	E_{elec}	E_{MM}
Solução válida	$-5170,03 kcal \cdot mol^{-1}$	$-22742,2 kcal \cdot mol^{-1}$	$-27912,2 kcal \cdot mol^{-1}$
Solução inválida	$-3591,29 kcal \cdot mol^{-1}$	$-14093,1 kcal \cdot mol^{-1}$	$-17684,4 kcal \cdot mol^{-1}$

4. Considerações Finais

A metodologia desenvolvida neste trabalho mostrou-se eficaz em reconstruir a região RBD da proteína Spike para a variante P.1 do SARS-CoV-2, a partir da mutagênese *in silico* e cristalografia já existentes. Mediante a utilização do algoritmo BP e as validações de Ramachandran constatamos a grande consistência das reconstruções estruturais. Sendo assim, percebe-se cada vez mais que o desafio de reconstruir proteínas vem sendo solucionado com o grande auxílio da ciência da computação, pois os princípios físico-químicos que regem o enovelamento ainda não são totalmente compreendidos. Um ponto interessante identificado ao testar as instâncias com o backbone e a cadeia lateral foi que apesar do Branch-and-Prune ter sido projetado para encontrar a estrutura do backbone, foi possível reconstruir toda a estrutura testada, sendo uma de suas soluções válidas de acordo com as restrições matemáticas e físico-químicas. Contudo, mais testes estão sendo realizados para validar essa constatação observada. Dentre as perspectivas futuras, pretendemos estruturar uma estratégia algorítmica que tenha como entrada a sequência de aminoácidos da proteína e que, apesar da alta complexidade computacional, possa predizer sua estrutura terciária. Além disso, pretendemos adicionar uma função objetivo baseada em um campo de força clássico para criar restrições físico-químicas ao longo da reconstrução.

Referências

Berg, J. M., Tymoczko, J. L., Stryer, L., and Gatto, G. J. (2012). *Biochemistry*. New York.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The Protein Data Bank. *Nucleic acids research*, 28(1):235–242.

Dejnirattisai, W., Zhou, D., Supasa, P., et al. (2021). Antibody evasion by the P.1 strain of SARS-CoV-2. *Cell*.

Dill, K. A. and MacCallum, J. L. (2012). The protein-folding problem, 50 years on. *Science*, 338(6110):1042–1046.

Dong, Q. and Wu, Z. (2002). A linear-time algorithm for solving the molecular distance geometry problem with exact inter-atomic distances. *Journal of Global Optimization*, 22:365–375.

Fidalgo, F., Maioli, D., Abreu, E., and Lavor, C. (2012). Uma formulação numérica para resolução de problemas de geometria de distâncias moleculares. *Simpósio Brasileiro de Pesquisa Operacional*.

Ju, B., Zhang, Q., Ge, J., Wang, R., Sun, J., Ge, X., et al. (2020). Human neutralizing antibodies elicited by SARS-CoV-2 infection. *Nature*, 584(7819):115–119.

Lavor, C., Liberti, L., Maculan, N., and Mucherino, A. (2011). The discretizable molecular distance geometry problem. *Comput Optim Appl*.

Li, Q., Guan, X., Wu, P., et al. (2020). Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *New England Journal of Medicine*, 382(13):1199–1207.

Liberti, L., Lavor, C., Masson, B., and Mucherino, A. (2011). Polynomial cases of the discretizable molecular distance geometry problem.

Maculan, N., Lavor, C., Lee, J., Liberti, L., Mucherino, A., Souza, M., and Xavier, A. E. (2010). The molecular distance geometry problem. *Elavio*.

Naveca, F., Nascimento, V., Souza, V., et al. (2021). Phylogenetic relationship of SARS-CoV-2 sequences from Amazonas with emerging Brazilian variants harboring mutations E484K and N501Y in the Spike protein.

Nelson, D. L. and Cox, M. M. (2017). *Lehninger Principles of Biochemistry*. Macmillan Higher Education, New York, 7 edition.

Newman, J. (2006). A review of techniques for maximizing diffraction from a protein crystal *in stilla*. *Acta Crystallographica Section D*, 62(1):27–31.

Phillips, J. C., Hardy, D. J., Maia, J. D. C., Stone, J. E., Ribeiro, J. V., Bernardi, R. C., Buch, R., et al. (2020). Scalable molecular dynamics on CPU and GPU architectures with NAMD. *The Journal of Chemical Physics*, 153(4):044130.

Schrödinger, LLC (2015). The PyMOL molecular graphics system, version 2.3.0.

Silva, W. and Lavor, C.and Ochiand, L. S. (2008). Cálculo de estruturas de proteínas. *Simpósio Brasileiro de Pesquisa Operacional*.

Wang, P., Nair, M. S., Liu, L., Iketani, S., Luo, Y., Guo, Y., et al. (2021). Antibody resistance of sars-cov-2 variants b.1.351 and b.1.1.7.

Wang, X., Lan, J., Ge, J., Yu, J., Shan, S., et al. (2020). Crystal structure of 2019-nCoV spike receptor-binding domain bound with ACE2. *Nature*, 581(7807):215–220.

Waterhouse, A., Bertoni, M., Bienert, S., et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research*, 46(1):296–303.