UNIVERSIDADE FEDERAL DO AMAZONAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

PREDIÇÃO E VALIDAÇÃO ESTRUTURAL DE MACROMOLÉCULAS COMPLEXAS COM ESTUDO DE CASO ENVOLVENDO PROTEÍNAS DO SARS-COV-2

CLARICE DE SOUZA SANTOS

Dezembro de 2021

Manaus - AM

CLARICE DE SOUZA SANTOS

PREDIÇÃO E VALIDAÇÃO ESTRUTURAL DE MACROMOLÉCULAS COMPLEXAS COM ESTUDO DE CASO ENVOLVENDO PROTEÍNAS DO SARS-COV-2

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Informática do Instituto de Computação da Universidade Federal do Amazonas (PPGI/IComp, UFAM) como parte dos requisitos necessários à obtenção do título de Mestre em Informática .

Orientadora: Rosiane de Freitas Rodrigues, D.Sc.

Co-orientador: Kelson Mota Teixeira de Oliveira, D.Sc.

Dezembro de 2021

Manaus - AM

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

S237p

Santos, Clarice de Souza
Predição e Validação Estrutural de Macromoléculas Complexas com Estudo de Caso Envolvendo Proteínas do Sars-Cov-2 / Clarice de Souza Santos . 2021
111 f.: il. color; 31 cm.

Orientadora: Rosiane de Freitas Rodrigues
Coorientador: Kelson Mota Teixeira de Oliveira
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. Algoritmos. 2. Branch-and-Prune. 3. Geometria de distâncias. 4. Gráfico de Ramachandran. 5. Modelagem molecular. I. Rodrigues, Rosiane de Freitas. II. Universidade Federal do Amazonas III. Título

FOLHA DE APROVAÇÃO

"Predição e Validação Estrutural de Macromoléculas Complexas com Estudo de Caso envolvendo Proteínas do SARS-CoV-2"

CLARICE DE SOUZA SANTOS

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Rosiane de Freitas Rodrigues

Profa. Rosiane de Freitas Rodrigues - PRESIDENTE

Jonathas Nunes da Silva

Prof. Jonathas Nunes da Silva - MEMBRO EXTERNO

Prof. Juan Gabriel Colonna - MEMBRO INTERNO

Manaus, 30 de Dezembro de 2021

A Deus,

aos professores,

aos colegas de curso e

aos meus familiares.

Agradecimentos

Agradeço primeiramente a Deus, hoje e sempre, pois tenho certeza que Ele me permitiu mais essa benção e esteve sempre ao meu lado em todos os momentos.

A minha orientadora, Rosiane de Freitas Rodrigues, pela oportunidade de enriquecimento profissional e pessoal, dedicação, lições e pela competência com que sempre orientou esta pesquisa.

Aos meus amigos, em especial ao meu amigo e marido Juan Soares Rosas, que me ajudaram diretamente e indiretamente neste trabalho e aos professores e pesquisadores com quem tive a oportunidade de estudar e que deram valiosas sugestões que contribuíram para o enriquecimento deste trabalho.

A minha mãe e tios, Luzamaria Costa de Souza, Luzenice Alcantarino e Zaqueu Alcantarino, ao meu irmão Pablo Luan pelo apoio e constante estímulo.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo suporte financeiro.

Resumo da Dissertação apresentada ao PPGI/IComp/UFAM como parte dos requisitos necessários para a obtenção do grau de Mestre em Informática.

PREDIÇÃO E VALIDAÇÃO ESTRUTURAL DE MACROMOLÉCULAS COMPLEXAS COM ESTUDO DE CASO ENVOLVENDO PROTEÍNAS DO SARS-COV-2

CLARICE DE SOUZA SANTOS

Dezembro/2021

Orientadora: Profa. Rosiane de Freitas Rodrigues, D.Sc.

Co-orientador: Prof. Kelson Mota Teixeira de Oliveira, D.Sc.

A estrutura tridimencional de uma proteína é importante devido a função da proteína estar ligada tanto a sua composição atômica quanto a sua estrutura tridimensional e no caso de um vírus a predição de maneira mais rápida e simples agiliza a criação de vacinas e remédios para combatê-lo. Esta dissertação apresenta aspectos matemático-computacionais e físico-químicos envolvidos na reconstrução da estrutura de proteínas, usando como estudo de caso proteínas do vírus SARS-CoV-2. Para isto, foram implementados os principais algoritmos que resolvem o problema (*Molecular Distance Geometry Problem* - MDGP), propostas e testadas variações de um dos algoritmos e realizada uma visita técnica ao Centro Nacional de Ressonância Magnética Nuclear da UFRJ, onde foi possível analisar o método de obtenção de dados através da Ressonância Magnética Nuclear. Após análise dos métodos implementados, foi identificada a necessidade de uma validação química para as estruturas geradas, uma vez que o cálculo estrutural somente garante a validade matemática dos resultados, então foi criada uma metodologia de reconstrução estrutural que vai desde a busca de dados, criação de instâncias de teste até o cálculo e validação estrutural. Essa metodologia foi utilizada no estudo de caso realizado com proteínas do novo coronavírus, principalmente nas variantes que atingiram o estado do Amazonas.

Palavras-chave: algoritmos, Branch-and-Prune, geometria de distâncias, gráfico de Ramachandran, modelagem molecular.

Abstract of Dissertation presented to PPGI/IComp/UFAM as a partial fulfillment of the requirements for the degree of Master in Informatics.

PREDICTION AND STRUCTURAL VALIDATION OF COMPLEX MACROMOLECULES WITH CASE STUDY INVOLVING SARS-COV-2 PROTEINS

CLARICE DE SOUZA SANTOS

December/2021

Advisor: Prof. Rosiane de Freitas Rodrigues, D.Sc.

Co-advisor: Prof. Kelson Mota Teixeira de Oliveira, D.Sc.

The three-dimensional structure of a protein is important because the function of the protein is linked to both its atomic composition and its three-dimensional structure, and in the case of a virus, prediction in a faster and simpler way speeds up the creation of vaccines and medicines to fight it . This dissertation presents mathematical-computational and physical-chemical aspects involved in the reconstruction of the three-dimensional molecular structure of proteins, using proteins from the SARS-CoV-2 virus as a case study. For this, the main algorithms that solve the Molecular Distance Geometry Problem (MDGP) were implemented, variations of one of the algorithms were proposed and tested, and a technical visit to the National Center for Nuclear Magnetic Resonance at UFRJ was carried out, where it was possible to analyze the method of obtaining data through Nuclear Magnetic Resonance. After analyzing the implemented methods, the need for chemical validation for the generated structures was identified, since the structural calculation only guarantees the mathematical validity of the results, so a structural reconstruction methodology was created, ranging from data search, creation from test instances to calculation and structural validation. This methodology was used in the case study carried out with proteins of the new coronavirus, mainly in the variants that reached the state of Amazonas.

Keywords: algorithms, branch-and-Prune, distance geometry, molecular modeling, Ramachandran plot.

Sumário

Lista de Figuras **x**

Lista de Tabelas **xiv**

1 Introdução **1**

1.1 Questão de Pesquisa e Objetivos 5

1.2 Metodologia 6

1.3 Organização da Dissertação 7

2 Fundamentos Teóricos **8**

2.1 Complexidade Computacional 8

2.2 Teoria dos Grafos 9

2.3 Otimização Combinatória 10

2.3.1 Modelagem Matemática 12

2.4 Geometria de Distâncias 13

3 Reconstrução da Estrutura Tridimensional de Proteínas **15**

3.1 O Problema de Geometria de Distâncias Moleculares 15

3.1.1 Classificação do Problema Segundo o Conjunto de Distâncias . . 16

3.1.2 Estratégias de Resolução 18

3.1.3 Trabalhos Relacionados 30

3.2 Modelagem Molecular de Moléculas Complexas 31

3.2.1 Bioquímica das Proteínas 31

3.3 Os Desafios da Predição Estrutural 36

3.3.1 Métodos Utilizados na Determinação Estrutural de Proteínas . . . 37

3.4 Validação Bioquímica - Gráfico de Ramachandran 46

3.5 Ferramentas de Reconstrução Estrutural 47

4 Estudo de Caso sobre SARS-CoV-2 **50**

4.1 Aspectos Bioquímicos do Vírus 51

4.2 Coronavírus Humano 53

4.2.1 SARS-CoV 53

4.2.2 MERS-CoV 54

4.2.3 SARS-CoV-2 55

4.3 Linhagens Emergentes do Vírus SARS-CoV-2 56

4.4 Vacinas Administradas Contra a COVID-19 58

4.5 Trabalhos Relacionados 60

5 Resultados **62**

5.1 FASE I - Imersão e Manipulação do Problema 63

5.1.1 Algoritmo de Geração das Instâncias de Teste 63

5.1.2 Algoritmo proposto - BP com Quatro Esferas 65

5.1.3 Variações de Podas Testadas no Algoritmo Branch and Pune . . . 66

5.1.4 Comparação dos Algoritmos 66

5.1.5 Assinalamento da Proteína TRX1 71

5.2 FASE II - Validação Bioquímica e Estudo de caso 74

5.2.1 Metodologia Fase II 75

5.2.2 Estudo de Caso - Proteínas do Vírus SARS-CoV-2 76

6 Considerações Finais **92**

Referências **96**

A Tabela de deslocamento **106**

B Assinalamento CCPNmr Analysis **107**

B.1 Assinalamento do backbone com tripla ressonância 107

B.2 Assinalamento Cadeia Lateral 107

Lista de Figuras

1.1 Exemplos de moléculas. 1

2.1 Representação gráfica de um grafo não-direcionado planar com 7 vértices e 10 arestas. 10

2.2 Exemplo de grafo completo com 4 vértices. 10

2.3 Região factível (interseção entre restrições). 13

3.1 Interseção de três esferas. 18

3.2 Possibilidade de posição de v (SILVA; LAVOR; OCHIAND, 2008). . . . 18

3.3 Quinto ponto determinado a partir das distâncias para os outros quatro. . . 20

3.4 Adição do terceiro átomo. 21

3.5 Árvore binária resultante do branch-and-prune. 28

3.6 Ordem HC (LAVOR, 2018). 30

3.7 Estrutura do aminoácido Glicina (Gly), o menos complexo dentre todos os aminoácidos, contendo apenas um único hidrogênio como cadeia lateral. 32

3.8 Aminoácidos com cadeias laterais apolares. 33

3.9 Aminoácidos com cadeias laterais polares sem carga. 34

3.10 Aminoácidos com cadeias laterais polares com carga positiva. 34

3.11 Aminoácidos com cadeias laterais polares com carga negativa. 34

3.12 Ligação peptídica. 35

3.13 Classificação geral dos métodos de obtenção de estrutura terciária de uma proteína. 38

3.14 Metodologia da Cristalografia de raio X. 39

3.15 Experimento de RMN. 40

3.16 Metodologia da RMN. 41

3.17 Espectro HNCO HSQC. 42

3.18 Assinalamento com CBCA(CO)NNH. 43
3.19 Assinalamento com CBCA(CO)NNH. 43
3.20 Assinalamento com CBCA(CO)NNH. 44
3.21 Assinalamento com HNCA e HNCOCA. 44
3.22 Dificuldade em distinguir entre um H_β e um H_δ. 45
3.23 Assinalamento cadeia lateral HCCH TOCSY COSY. 46
3.24 Assinalamento com HCCH TOCSY COSY. 47
3.25 Regiões do gráfico de Ramachandran tendo como exemplo a validação para a estrutura ORF8 (PDB ID: 7JX6) obtida com o módulo "Structure Assessment" na plataforma MolProbity (WILLIAMS et al., 2018). 48

4.1 Estrutura geral do novo coronavírus. 53
4.2 Comparação entre as estruturas cristalográficas do vírus MERS-CoV onde foi apresentado o complexo RBD com o receptor celular DDP4 (PDB ID: 4L72) (WANG et al., 2013), enquanto ao lado direito da figura encontra-se o complexo ACE2-RBD (PDB ID: 6M0J) (WANG et al., 2020) referente ao vírus SARS-CoV-2. A visualização foi por intermédio da versão gratuita do software ICM-MolBrowser 3.8.7. 55
4.3 Mecanismo de reconhecimento molecular entre a glicoproteína Spike do vírus SARS-CoV-2 e os receptores ACE2 nos alvéolos pulmonares (COGNITION, 2020). 55
4.4 Gráfico da quantidade de ocorrências de infecção pela COVID-19 no Brasil e no mundo. 56
4.5 Diagrama resumindo algumas vacinas já aprovadas em estudos clínicos de Fase III. Os diagramas estruturais das vacinas foram obtidos na plataforma BioRender. 59
4.6 Estrutura tridimensional do SARS-CoV-2 M^{pro}, em duas visões diferentes (ZHANG et al., 2020). 61
4.7 Estruturas químicas dos inibidores de $\alpha-$cetoamida 11r, 13a, 13b e 14b. Os círculos coloridos destacam as modificações de uma etapa de desenvolvimento para a próxima(ZHANG et al., 2020). 61

5.1 Geração das instâncias de teste. 64

5.2 Interseção de quatro esferas. 65
5.3 Comparação das etapas de geração dos átomos nos algoritmos BP e BP4. 70
5.4 Estruturas geradas pelos algoritmos comparados com a original do banco PDB para o hormônio pancreático contido na instância 1PPT. 71
5.5 Alinhamento sequencial de Trx1 e proteínas estruturalmente homólogas de diferentes organismos (PINHEIRO et al., 2007). 72
5.6 Fluxo para predição da estrutura 3D de uma proteína realizado no CNRMN 73
5.7 Fluxo do cálculo estrutural das variantes encontradas no Amazonas. . . . 75
5.8 Etapas adotadas para minimização estrutural das estruturas ORF3a e ORF8 contendo as variantes no Amazonas até o passo final de alinhamento estrutural com a respectiva estrutura sem quaisquer mutações. . . . 78
5.9 Diagramas de Ramachandran referente às soluções previstas pelo algoritmo para ORF7 (PDB ID: 6W37). Todas as imagens foram construídas com o auxílio da plataforma MolProbity. 81
5.10 Sobreposição estrutural entre as estruturas cristalográficas sem quaisquer mutações da ORF3a (PDB ID: 6XDC) e ORF8 (PDB ID: 7JX6) e as respectivas variantes G196V e L84S. 82
5.11 Diagramas de Ramachandran referente às soluções previstas pelo algoritmo para ORF8 (PDB ID: 7JX6). Todas as imagens foram construídas com o auxílio da plataforma MolProbity. 83
5.12 Diagramas de Ramachandran referente às soluções previstas pelo algoritmo para ORF3a (PDB ID: 6XDC). Todas as imagens foram construídas com o auxílio da plataforma MolProbity. 83
5.13 Diagramas de Ramachandran para as soluções obtidas no algoritmo BP para a cepa P.1. As estruturas de referência onde a mutagênese e a respectiva reconstrução foram aplicadas foram os complexos ACE2-RBD (PDB ID: 6M0J) e Spike-Anticorpo (PDB ID: 7BWJ). Todas as imagens foram construídas com o auxílio da plataforma MolProbity. 85
5.14 Alinhamento estrutural no software Schrödinger Maestro 2020-4 entre soluções válidas e inválidas contendo P.1 em comparação com a respectiva estrutura cristalográfica do complexo ACE2-RBD (PDB ID: 6M0J). . . . 86

5.15 Comparativo entre as interações químicas formadas no complexo ACE2-RBD (PDB ID: 7DF4) em consequência das mutações K417N e K417T referentes às linhagens P.1 e P.2, respectivamente. Todos os diagramas foram gerados na plataforma DynaMut2 (RODRIGUES; PIRES; ASCHER, 2021). As linhas tracejadas em Verde representam os contatos hidrofóbicos, as linhas em Vermelho correspondem às ligações de Hidrogênio e por fim a cor Laranja refere-se às interações polares. 87
5.16 Diagramas de Ramachandran referente às soluções previstas pelo algoritmo para (PDB ID: 6WTT). Todas as imagens foram construídas com o auxílio da plataforma MolProbity. 90
5.17 Diagramas de Ramachandran referente às soluções previstas pelo algoritmo para (PDB ID: 6XS6). Todas as imagens foram construídas com o auxílio da plataforma MolProbity. 91
5.18 Diagramas de Ramachandran referente às soluções previstas pelo algoritmo para (PDB ID: 6YWK). Todas as imagens foram construídas com o auxílio da plataforma MolProbity. 91

A.1 Tabela de Deslocamento - fornecida no CNRMN. 106

B.1 Assinalamento C_α e C_β. 108
B.2 Assinalamento C_α e C_β. 109
B.3 Assinalamento C_α e C_β. 109
B.4 Assinalamento da cadeia lateral. 110
B.5 Assinalamento da cadeia lateral. 111
B.6 Assinalamento da cadeia lateral. 111

Lista de Tabelas

4.1 Origem e classificação dos 7 (sete) Coronavírus que afetam seres humanos. As imagens foram obtidas na plataforma BioRender. 54

5.1 Tabela de comparação dos Algoritmos. Molécula - nome da proteína do PDB, |V| - número de átomos, |E| - quantidade de distâncias conhecidas, Enum - número de átomos fixados, Tsoma - soma total das distâncias entre todos os átomos fixados. 67

5.2 Tabela de comparação de número de átomos gerados pelo BP e BP4 e visitados pelo GBU e BP4 com conjunto de distâncias menores ou iguais a 10Å. 68

5.3 Tabela de comparação entre os algoritmos GBU, BP e BP4 com conjunto de distâncias menores ou iguais a 10Å. 69

5.4 Informações cristalográficas das estruturas estudadas nesta dissertação (NELSON et al., 2020; KERN et al., 2020; NELSON; HALL; FREMONT, 2020; HANKE et al., 2020; XU et al., 2021). 76

5.5 Comparação entre as variantes P.1 e P.2 em relação aos valores médios de alguns parâmetros resultantes da dinâmica molecular na faixa de 18*ns* para ACE2-RBD (PDB ID: 6M0J) que quantificam mudanças estruturais no Spike Região RBD na interação com ACE2. 86

5.6 Resumos dos testes realizados para validar a metodologia utilizada. As porcentagens de resíduos nas regiões favoráveis do gráfico de Ramachandran são apresentadas para ambas as soluções validadas e invalidadas pelos testes. 90

Lista de Algoritmos

1 Algoritmo Dong Wu 20

2 Algoritmo GBU 26

3 BranchAndPrune(G, v, U, x^I, X) 27

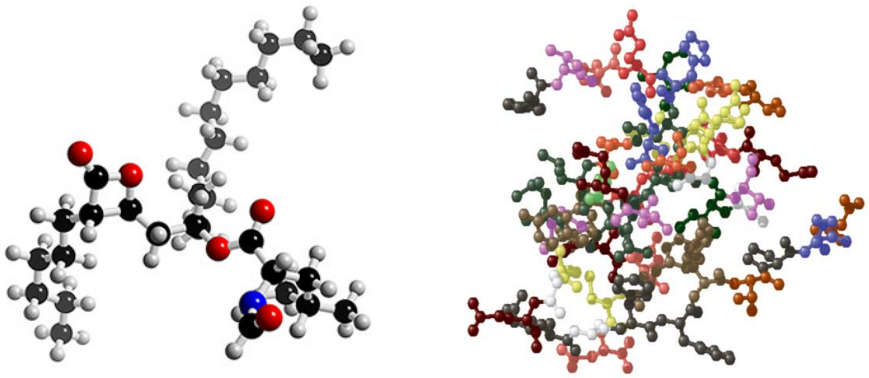
4 Algoritmo BP com quatro esferas 65

Capítulo 1

Introdução

Pesquisas envolvendo o estudo e a determinação de propriedades de moléculas têm se intensificado desde a descoberta do código genético e do fluxo da informação biológica. Surgiram, então, vários métodos para determinar as sequências que constituem essas moléculas, onde a utilização de ferramentas computacionais tornou-se essencial para armazenar e analisar a grande quantidade de sequências e outras informações biológicas. Assim, surgiram as áreas de Biotecnologia e Bioinformática que reúnem vários campos de pesquisa, principalmente: Biologia, Ciência da Computação, Química, Física, Matemática, dentre outras.

Uma molécula é a menor partícula de uma substância, conservando suas propriedades químicas e demais características. A molécula é um grupo de átomos, iguais ou diferentes, que se mantêm unidos e que não podem ser separados sem afetar ou destruir as propriedades das substâncias, exemplos de moléculas podem ser visto na Figura 1.1.



(a) Molécula Orlistate $C_{29}H_{53}NO_5$ (OLIVEIRA, 2007)

(b) Proteína (PROTEíNAS, 2011)

Figura 1.1: Exemplos de moléculas.

Muitas pesquisas em biologia têm como foco as propriedades e atividades das proteínas, que são moléculas fundamentais dos sistemas biológicos constituídas por uma cadeia linear de aminoácidos. No caso das proteínas, a ordem de tais aminoácidos determina a estrutura tridimensional da proteína (DONG; WU, 2002). Essa estrutura é importante porque a função da proteína está ligada tanto a sua composição atômica quanto a sua estrutura tridimensional (LAVOR et al., 2011a).

A estrutura tridimensional pode ser obtida experimentalmente através da Ressonância Magnética Nuclear (RMN) ou Cristalografia de Raio-X, e teoricamente através da minimização da energia potencial ou por simulação de dinâmica molecular (SILVA; LAVOR; OCHIAND, 2008). A RMN fornece apenas um pequeno subconjunto de distâncias entre os átomos de uma molécula, pois só obtêm distâncias entre pares de átomos que estejam próximos (5 a 6 angstrons) (FIDALGO et al., 2012; DONG; WU, 2002).

O problema de estimar a estrutura completa da molécula, determinando a posição no espaço de todos os átomos que a compõe é chamado de Problema de Geometria de Distâncias Moleculares (do inglês, *Molecular Distance Geometry Problem* - MDGP) e formulado tradicionalmente como um problema de otimização contínua (SILVA; LAVOR; OCHIAND, 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, passa a ser um problema NP-difícil (SILVA; LAVOR; OCHIAND, 2008). Uma solução viável, segundo o modelo teórico envolvido, para este problema pode ser obtida em tempo razoável, para alguns casos, por meio do uso de técnicas de otimização combinatória.

Definir a estrutura geométrica das proteínas não é trivial, pois em geral se apresentam como macromoléculas de difícil identificação. A estimativa da estrutura completa da proteína é lenta, pois sua modelagem tridimensional é complexa podendo resultar em muitas estruturas viáveis, mas completamente distantes da real geometria da macromolécula (MOTA, 2013).

O estudo do problema sobre o enfoque de Geometria de Distâncias contribui na identificação da estrutura da proteína, pois é possível obter experimentalmente um conjunto de distâncias entre os átomos da molécula e calcular teoricamente por Geometria de Distâncias a estrutura tridimensional viável para os átomos.

Como potenciais contribuições, do estudo do problema sobre o enfoque de Geometria de Distâncias, além da elucidação completa da estrutura de proteínas, cita-se (MOTA, 2013):

- Reconhecimento real de sítios de atividade de proteínas e seus receptores.
- Conformação real de proteínas em estados não-cristalizados.
- Estudos aprofundados da relação solvente-estrutura geométrica, o papel do solvente no estudo de sítios ativos de proteínas.
- Formulação de novas drogas com perfil genético específico.

Sendo assim, a pesquisa em torno do problema MDGP apresenta como motivação todas estas contribuições citadas acima, que envolvem grandes desafios teóricos matemáticos e computacionais, bem como o desafio de tentar contribuir no entendimento e aproveitamento das informações obtidas em laboratório e em pesquisas de outras áreas envolvidas, como a físico-química.

No ano de 2020, a importância de se predizer de maneira mais rápida e simples a estrutura trimensional de proteínas foi evidênciada com o aparecimento de um novo vírus da família *Coronaviridae* que se espalhou em poucas semanas pelo mundo dando origem a uma pandemia de magnitude não vista desde a gripe espanhola em 1918. A pandemia batizada de COVID-19 (Coronavirus Disease, identificada no final de 2019) é causada pelo vírus SARS-CoV-2, em grande medida semelhante ao vírus da SARS-CoV.

Apesar dos primeiros casos terem sido detectados oficialmente em dezembro de 2019, na cidade de Wuhan, na China, sendo associados ao mercado de frutos do mar da cidade, sua origem é ainda desconhecida. Embora seja reconhecido pela comunidade científica que o vírus surgiu primeiramente em um morcego, 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.

No final de dezembro de 2020, já contava em termos globais com mais 70 milhões de casos confirmados, sendo mais de 1,5 milhão de mortos e cerca de 45 milhões de casos recuperados. Na mesma época no Brasil, havia mais de 6,5 milhões de casos confirmados e

quase 180 mil mortos e, apesar da taxa de transmissão ter sido de 3 (uma pessoa infectava outras três) e ter caído para cerca de 1,6, ainda se mantinha alta e com taxa de mortalidade (CNS, 2020). Atualmente o Brasil apresenta aproximadamente mais de 222,1 milhões de casos e 640 mil mortes.

A COVID-19 não é a primeira pandemia que a humanidade enfrenta, basta citar exemplos históricos, como a peste negra e a gripe espanhola. Entretanto, está sendo considerada como o maior problema de saúde pública de importância internacional já registrado, incluído a uma lista de outros cinco recentes, que foram: o H1N1 (Influenza A ou gripe suína), com registros iniciais no México e América do Norte (2009 e retorno mais forte em 2018); a Poliomelite, na Ásia Central, Oriente Médio e África Central (2014); o Ebola na África (2014); a epidemia do Zika vírus que impactou o Brasil e Américas (2016); e, o surto de Ebola, altamente letal, na República Democrática do Congo (2019) (OPAS, 2020).

No Brasil, estudos apontam que o vírus já estaria circulando pelo país desde o final de janeiro ou início de fevereiro de 2020, portanto, antes das medidas de contenção e isolamento social implantadas cerca de um mês depois, em meados de março, e cujo contágio se espalhou pelo país durante a grande festa popular do Carnaval brasileiro.

Entretanto, o primeiro caso da COVID-19 no Brasil (identificado como BR1), e que também foi o primeiro da América Latina, teve diagnóstico molecular confirmado pelo Instituto Adolfo Lutz, em 26 de fevereiro de 2020. O paciente se infectou durante uma visita à região de Lombardia, no norte da Itália, entre os dias 9 e 21 deste mês. Esta foi a região mais fortemente afetada pela pandemia. O genoma completo deste vírus foi sequenciado e disponibilizado à comunidade científica apenas 2 dias depois, 28 de fevereiro de 2020, por pesquisadores do Instituto Adolfo Lutz, em conjunto com o Instituto de Medicina Tropical da Faculdade de Medicina da Universidade de São Paulo e com a Universidade de Oxford (FMUSP, 2020).

Atualmente mais de 427 genomas do SARS-CoV-2 já foram mapeados no Brasil. A partir do sequenciamento registrado é possível observar como o vírus se multiplica, a partir de quais proteínas produz.

Sequenciar o genoma do vírus e reconstruir sua estrutura molecular completa ou partes dela, então, propicia um maior entendimento de seu comportamento e ação 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 moléculas complexas traz desafios tecnológicos e matemático-computacionais interessantes do ponto de vista teórico e prático. Deste modo, nesta pesquisa foram utilizadas algumas proteínas de variantes do vírus SARS-CoV-2 detectadas no Brasil como estudo de caso, analisando sua entrada no país, onde são predominantes, seus aspectos de complexidade usando os algoritmos implementados durante o mestrado para determinar a estrutura tridimensional do vírus.

Esta pesquisa de mestrado aconteceu em duas etapas. A *Fase I* ocorreu nos anos de 2012 e 2013, este período foi utilizado para o aprofundamento na teoria sobre o MDGP, implementação e testes com algoritmos da literatura, juntamente com possíveis adaptações destes algoritmos para fins de realização de experimentos computacionais. Devido problemas pessoais e familiares, precisei abandonar o programa e não concluí o mestrado. Mas após alguns anos foi possível a retomada dos estudos e reinício do mestrado, começando então a *Fase II* onde foi iniciada uma investigação sobre a validade química das estruturas construídas pelos algoritmos e um estudo de caso utilizando variações de proteínas do vírus responsável pela pandemia da COVID-19, principalmente as variantes encontrados no Amazonas.

1.1 Questão de Pesquisa e Objetivos

O objetivo geral inicial desta pesquisa de mestrado foi investigar o Problema de Geometria de Distâncias Moleculares sob o enfoque de Otimização Combinatória, com ênfase em Teoria do Grafos e Programação Matemática, de tal forma a identificar novas propriedades ou propor novos modelos teóricos que o resolvam de maneira mais precisa e eficiente.

No decorrer da pesquisa e análise do problema de geometria de distâncias moleculares e dos algoritmos que o resolvem, foi possível identificar algumas questões em aberto: existe uma maneira de validar o processo? Todos os resultados obtidos são quimicamente válidos? Ou seja, todas as estruturas tridimensionais matematicamente válidas geradas são possíveis de existir quimicamente? Essas perguntas nortearam esta pesquisa. Os objetivos específicos foram:

- Elencar (ou Caracterizar) o processo real de obtenção dos dados das proteínas e outros tipos de moléculas, de forma a serem utilizados como base no desenvolvimento de modelos e algoritmos.

- Implementar (ou Elaborar) modelos em grafos, de programação inteira ou algoritmo para o MDGP, que promovam estratégias mais eficientes de resolução.
- Realizar experimentos computacionais com as estratégias algorítmicas propostas para fins de comparação com a literatura.
- Propor uma metodologia, um fluxo de análise e validação das estruturas geradas pelos algoritmos existentes na literatura.
- Realizar estudo de caso utilizando proteínas do SARS-CoV-2 para validação do fluxo metodológico proposto.

1.2 Metodologia

Apesar da revisão bibliográfica ter sido realizada para a obtenção do ferramental teórico necessário para estudar as classes de problemas de geometria de distâncias moleculares com profundidade, foi identificado um caráter de metodologia de pesquisa de natureza aplicada, uma vez que foram investigados principalmente trabalhos relacionados a classe de problemas com conjunto de poucas distâncias conhecidas, onde métodos conhecidos da literatura foram reimplementados e novos de resolução foram propostos, além de realizados experimentos computacionais massivos envolvendo classes diversas de proteínas e um estudo de caso específico com macromoléculas presentes no SARS-CoV-2.

Quanto aos objetivos, o tipo de pesquisa realizada foi de caráter exploratório, já que com a pesquisa principalmente foi conseguida maior familiaridade com o problema, devido a realização das pesquisas bibliográficas e do estudo de caso específico.

Em relação a abordagem, a pesquisa é classificada como qualitativa com procedimentos de pesquisa bibliográfica e experimental, além de poder ser encaixada em pesquisa de campo e estudo de caso, uma vez que foi realizada uma visita técnica ao Centro Nacional de Ressonância Magnética Nuclear (CNRMN) para o entendimento e identificação do método de obtenção de dados de RMN, bem como um estudo de caso envolvendo o estudo mais aprofundado de algumas das proteínas do vírus causador do surto da COVID-19.

Toda pesquisa foi registrada na forma de relatórios técnicos, resumos, artigos e, por fim, na dissertação de Mestrado. A divulgação da pesquisa foi feita através de seminários, participação em eventos nacionais e internacionais de relevância.

1.3 Organização da Dissertação

A presente dissertação encontra-se dividida em cinco capítulos, listados a seguir:

- No Capítulo 1, é vista a introdução composta pelo contexto, definição do problema, motivação, objetivos e método de pesquisa proposto.

- No Capítulo 2, sobre Fundamentos Teóricos, são apresentadas definições de teoria da computação, complexidade computacional, otimização combinatória e geometria de distâncias.

- No Capítulo 3, sobre a Reconstrução da Estrutura Tridimensional de Proteínas, são apresentados os métodos utilizados na determinação da estrutura de proteínas e a definição formal do Problema de Geometria de Distâncias Moleculares, sua classificação segundo o conjunto de distâncias, uma versão discreta para o problema e os métodos de resolução. Apresenta-se também conceitos de bioquímica das proteínas, os desafios da predição estrutural, métodos de validação química e algumas ferramentas de reconstrução.

- No Capítulo 4, mostra um estudo de caso sobre SARS-CoV-2 e a Pandemia COVID-19, são apresentadas a definição de vírus e o surgimento do coronavírus humano, bem como sua classificação e alguns tipos de vacinas que estão sendo desenvolvidas para combatê-lo.

- No capítulo 5, sobre os Resultados, apresenta-se os trabalhos realizados durante a pesquisa, os algoritmos implementados para o caso em que todas as distâncias são previamente conhecidas e para o caso onde apenas um subconjunto das distâncias é conhecido. São apresentados também os métodos trabalhados, a análise comparativa feita dos algoritmos e os experimentos computacionais realizados para proteínas em geral e para o estudo de caso sobre algumas proteínas do vírus SARS-CoV-2, além de um resultado extra sobre assinalamento de proteína realizado durante uma visita técnica ao Centro Nacional de Ressonância Magnética Nuclear.

- E por fim, no Capítulo 6, sobre as Considerações Finais, são feitos os comentários gerais sobre a pesquisa, resumindo as contribuições obtidas, bem como os trabalhos futuros.

Capítulo 2

Fundamentos Teóricos

Neste capítulo são apresentados conceitos que formam a base para o desenvolvimento deste trabalho. Nas seções seguintes são apresentados os conceitos de complexidade computacional, teoria dos grafos, otimização combinatória, modelagem matemática e geometria de distâncias.

2.1 Complexidade Computacional

A teoria da complexidade computacional tem como objetivo classificar problemas computacionais de acordo com sua dificuldade. Todo problema é tratado como sendo um conjunto de parâmetros que definem as instâncias e um conjunto de propriedades que configuram as restrições do problema a serem satisfeitas. Assim, entre instâncias de um mesmo problema, as únicas variações estão nos conjuntos dos parâmetros.

Em relação ao número de computações necessárias pare se obter a solução ótima, os problemas podem ser classificados em três grandes classes: problemas P, NP e Intratáveis. Em (CORMEN et al., 2002) temos que:

- Problemas P (*Polinomial Time*) são problemas que podem ser resolvidos em tempo polinomial. Ou seja, o esforço computacional cresce polinomialmente em função do tamanho da instância, são resolvidos no tempo $O(n^k)$ sendo k uma constante e n o tamanho da entrada. São também conhecidos como problemas tratáveis e existem algoritmos eficientes para resoluções dos problemas desta classe.

- Problemas NP (*Nondeterminisc Polinomial Time*) são problemas "verificáveis"em tempo polinomial. Dada uma solução para o problema, pode-se verificar em tempo

polinomial se a solução satisfaz o problema de decisão, mas o esforço computacional de encontrar uma solução cresce exponencialmente em função do tamanho da instância.

- Problemas intratáveis configuram problemas onde, assim como os NP, o esforço computacional cresce exponencialmente em função do tamanho da instância, com a garantia de que a solução encontrada seja a solução ótima para o problema.

Além disso, existem duas classes adicionais denominadas NP-Difíceis e NP-Completos, sendo que os problemas que possuem uma versão de Otimização estão incluídos na classe de problemas NP-Difíceis. Os problemas NP-Difíceis e NP-Completos são problemas que não pertecem à classe P, mas possuem a dificuldade característica dos problemas da classe NP.

Um problema é NP-Difícil se todos os problemas da classe NP forem polinomialmente redutíveis a ele. Ou seja, resolvendo um problema NP-Difícil em tempo polinomial, todos os problemas da classe NP também serão resolvidos em tempo polinomial. E os problemas NP-Completos englobam todos os problemas NP que também são NP-Difíceis.

2.2 Teoria dos Grafos

Um grafo é um par ordenado $G = (V, E)$, onde V é um conjunto de vértices e E é um conjunto de arestas, onde $n = |V|$ (cardinalidade do conjunto V) e $m = |E|$ (cardinalidade do conjunto E). Cada elemento e no conjunto E é um par (i, j) que indica que o vértice i é ligado ao vértice j (ou seja, são adjacentes, e a aresta e incide em i e j). O grafo é dito não-direcionado quando os pares que representam as arestas não são ordenados, isto é, $(i, j) = (j, i)$.

A representação gráfica de um grafo consiste em pontos distintos do plano associados a cada vértice e, para cada aresta (i, j), uma linha conectando os pontos correspondentes aos vértices i e j. Se for possível efetuar uma representação gráfica do um grafo G sem que as arestas se cruzem, diz-se que G é planar (SZWARCFITER, 1986). Um exemplo é dado na Figura 2.1.

O grau de um vértice v do grafo equivale à quantidade de arestas que incidem em v. O grau máximo do grafo, denotado por $\Delta(G)$, é o valor do maior grau dentre todos os

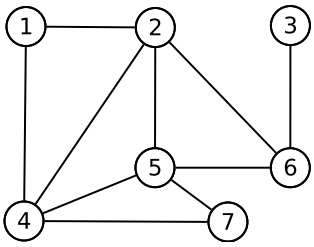


Figura 2.1: Representação gráfica de um grafo não-direcionado planar com 7 vértices e 10 arestas.

vértices de G. De maneira similar, o grau mínimo, denotado por $\delta(G)$, é definido como o valor do menor grau de G.

Um grafo completo K_n de ordem n é aquele em que cada vértice é vizinho de todos os demais, ou seja, cada par de vértices é adjacente, um exemplo de k_4 é mostrado na Figura 2.2. O grau de todos os vértices é exatamente igual a $n-1$ e a quantidade de arestas é dada então por $m = \frac{n(n-1)}{2}$.

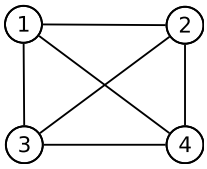


Figura 2.2: Exemplo de grafo completo com 4 vértices.

Uma clique de um grafo $G = (V,E)$ é um subconjunto $V' \subseteq V$ cujo subgrafo induzido $G[V']$ é completo. Uma clique máxima de um grafo G é aquela com a maior cardinalidade dentre todas as cliques de G, e o número da clique $\omega(G)$ é a cardinalidade da clique máxima de G. Em sentido oposto, um conjunto estável de um grafo $G = (V,E)$ é um subconjunto $V' \subseteq V$ em que nenhum par de vértices é adjacente, para qualquer vértice de V'. Um conjunto estável máximo é aquele com a maior cardinalidade dentre os conjuntos estáveis de um grafo, e o número de estabilidade de um grafo G, $\alpha(G)$, é o tamanho do conjunto estável máximo de G.

2.3 Otimização Combinatória

De acordo com (FREITAS, 1996) otimização é o processo de fazer um projeto, sistema ou tomada de decisão tão eficientemente quanto possível. Significa maximizar ou minimizar uma função, ou seja, encontrar a **solução ótima** do problema. Otimização Combinatória

trata problemas de maximização e minimização de funções de variáveis, sujeita a restrições de igualdade, desigualdade e integralidade das variáveis.

Problemas de Otimização Combinatória são caracterizados pela presença de um enorme espaço discreto de busca, onde, dentre inúmeras alternativas, se deseja maximizar ou minimizar um certo valor, geralmente fornecido por uma função objetivo. Tais problemas podem ser classificados segundo a Teoria da NP-Completude (GAREY; JOHNSON, 1979).

Pode-se formular um problema de otimização como (GOLDBARG; PACCA; LUNA, 2005):

$$
\begin{aligned}
Otimizar \quad z &= \sum_{j=1}^{n} c_j x_j \\
sujeito \quad a: & \\
\sum_{j=1}^{n} a_{ij} x_j &\geq b_i \quad i = 1, 2, ..., p \\
\sum_{j=1}^{n} a_{ij} x_j &= b_i \quad i = p+1, p+2, ..., m \\
x_i &\geq 0, \quad j = 1, 2, ..., q \\
x_i &\in \mathbb{R}, \quad j = q+1, q+2, ..., n
\end{aligned}
$$

Os problemas de Otimização Combinatória podem ser resolvidos através de métodos exatos ou métodos aproximados. Os métodos exatos são algoritmos que podem levar um tempo excessivamente longo de execução, dependendo da instância a ser resolvida e ainda, resultam em uma codificação muito mais complexa do que as implementações de métodos aproximados.

Já os métodos aproximados determinam soluções de boa qualidade num curto espaço de tempo, em situações nas quais é muito difícil encontrar soluções ótimas para instâncias de problemas NP-Difíceis, extrapolando a capacidade de armazenamento necessária até mesmo pelo melhor algoritmo exato. As heurísticas e metaheurísticas consistem em métodos aproximados de resolução de problemas de Otimização Combinatória.

A heurística consiste em métodos e regras para se descobrir ou assistir ao processo de resolução de problemas. As heurísticas possuem uma concepção simples onde são incorporadas as partes mais importantes do problema. As metaheurísticas são procedimentos de resolução de problemas de otimização que contem outras heurísticas em seus

procedimentos internos, fazendo uso da aleatoriedade para obter boas soluções.

Como problemas de Otimização estão:

- Caixeiro Viajante, que consiste na procura de um circuito que possua a menor distância, começando numa cidade qualquer, entre várias, visitando cada cidade precisamente uma vez e regressando à cidade inicial.

- Problema da mochila onde é necessário preencher uma mochila com objetos de diferentes pesos e valores, sendo o objetivo preencher a mochila com o maior valor possível, não ultrapassando o peso máximo (CORMEN et al., 2002).

- Problemas de escalonamento, que consistem na alocação de determinados recursos a determinadas atividades em função do tempo, que envolve um processo de tomada de decisão que visa otimizar um ou mais critérios de medida de desempenho (FREITAS, 2009).

2.3.1 Modelagem Matemática

No processo de modelagem matemática existem os seguintes elementos básicos (GOLDBARG; PACCA; LUNA, 2005):

- **Parâmetros do modelo** - são os dados do problema que serão utilizados como entrada do modelo.

- **Variáveis de decisão** - são as variáveis (incógnitas) do modelo, seus valores modificam diretamente o valor geral do problema.

- **Função Objetivo** - é a função que representa matematicamente o objetivo do problema, mede a eficiência de todas as possíveis soluções.

- **Restrições** - são os limitantes entre as variáveis de decisão e os dados do problema, indicam as limitações como disponibilidade de material, necessidades a serem atendidas.

- **Solução Factível** - é uma solução viável que respeita todas as restrições do problema.

- **Região Factível** - é a região formada por todas as soluções factíveis, ou seja, é a interseção das restrições do problema como mostrado na Figura 2.3.

- **Solução Ótima** - é a melhor solução factível de acordo com a função objetivo.

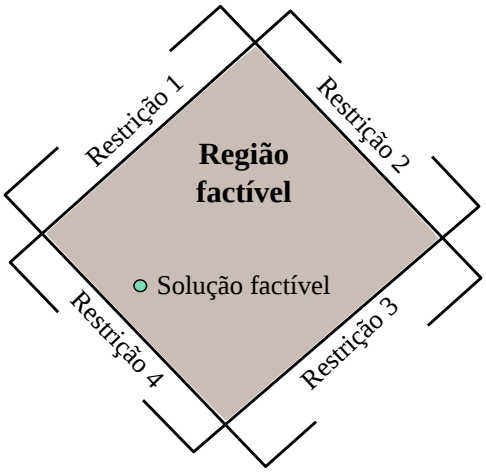


Figura 2.3: Região factível (interseção entre restrições).

2.4 Geometria de Distâncias

Geometria de distância (*Distance Geometry* - DG) é o estudo de problemas de conjuntos de pontos através de distâncias fornecidas entre pares de pontos do conjunto. A geometria de distância é importante em áreas onde os valores de distância entre pontos são determinadas ou considerados, como por exemplo em cartografia e física.

O problema de geometria distância (*Distance Geometry Problem* - DGP) preocupa-se em encontrar as coordenadas de um conjunto de pontos a partir das distâncias entre pares de pontos.

Existem várias aplicações reais para o DGP, onde pode-se destacar as duas a seguir:

- Localização em redes de sensores - o problema de localizar os sensores em redes de telecomunicações é definido como encontrar as posições de todos os sensores tendo algumas posições de sensores chamados âncoras e algumas distâncias entre os sensores.

- Determinação da estrutura espacial de moléculas - o problema de determinar as coordenadas dos átomos de uma estrutura molecular através das distâncias entre

pares de átomos obtidas com técnicas experimentais é conhecido na literatura como Problema de Geometria de Distâncias Moleculares (*Molecular Distance Geometry Problem* - MDGP), e será abordado em detalhes no próximo capítulo.

Capítulo 3

Reconstrução da Estrutura Tridimensional de Proteínas

Neste capítulo são avaliados os métodos e ferramentas que têm sido utilizados para descrever e determinar a estrutura tridimensional de uma proteína, enfocando os métodos experimentais, teóricos/computacionais e híbridos. É apresentado o Problema de Geometria de Distâncias Moleculares, através da sua definição formal, sua classificação em duas categorias dependendo do conjunto de distâncias de entrada, uma formulação discreta do problema e alguns métodos de resolução.

3.1 O Problema de Geometria de Distâncias Moleculares

Considerando uma molécula formada por n átomos $a_1, a_2, ..., a_n$ 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 para a molécula respeitando as distâncias euclidianas conhecidas, ou seja, encontrar as posições $x_1, x_2, ..., x_n$ para os átomos respeitando as distâncias conhecidas (SILVA; LAVOR; OCHIAND, 2008).

As coordenadas $x_1, x_2, ..., x_n$ podem ser obtidas a partir das distâncias d_{ij} através da resolução do sistema de equações formado pelas equações de cálculo de distâncias:

$$||x_i - x_j|| = d_{ij} \quad \forall (i, j) \in S, \tag{3.1}$$

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 $||.||$ a norma euclidiana.

O problema de determinar as coordenadas dos átomos de uma molécula a partir de distâncias entre pares de átomos pode ser investigado colocando os átomos em um espaço métrico qualquer respeitando as distâncias definidas (SIT, 2010). Neste trabalho o problema será estudado apenas no espaço $3D$ euclidiano. Dependendo do conjunto de distâncias fornecidas e do espaço métrico o MDGP pode ter uma solução, várias soluções ou não ter solução válida.

A estrutura tridimensional é conseguida através da imersão dos átomos no $\mathbb{R}^3$. Considerando a posição do ponto como $p = (u, v, w)$, imergir um ponto significa encontra os valores das dimensões u, v e w para o ponto.

Na prática o conjunto de distâncias pode conter erros, pois as distâncias são conseguidas a partir de experimentos, como NMR e cristalografia, ou estimativas teóricas (SIT, 2010). Por isso, uma forma mais prática de definir o problema é utilizando limites inferiores e superiores (do inglês *lower and upper bounds*) no conjunto de distâncias, neste caso a restrição do problema fica:

$$l_{ij} \leq ||x_i - xj|| \leq u_{ij} \quad \forall (i, j) \in S. \tag{3.2}$$

O conjunto de distâncias pode conter para cada distância entre átomos valores exatos ou limites. No primeiro caso o problema é dito ser de distâncias exatas e no segundo ser de distâncias inexatas ou faixas de distância (SIT, 2010). Esses dois tipos podem conter as distâncias entre todos os átomos ou somente entre alguns deles.

3.1.1 Classificação do Problema Segundo o Conjunto de Distâncias

O MDGP pode ser classificado, dependendo do esparsamento do conjunto de distâncias, em duas formas (FIDALGO et al., 2012):

- Conjunto completo de distâncias - todas as distâncias entre quaisquer pares de átomos são conhecidas. Esse problema pode ser resolvido em tempo polinomial. Dong e Wu apresentaram um algoritmo em tempo polinomial que resolve este problema através de sistemas lineares (DONG; WU, 2002).

- Conjunto arbitrário de distâncias - são conhecidas somente as distâncias entre alguns átomos da molécula. Esse problema 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).

O Problema de Geometria de Distâncias é normalmente definido na sua forma contínua, entretanto Lavor et al. propuseram um modelo discreto do problema para o conjunto arbitrário de distâncias (LAVOR et al., 2011b), mostrado na Seção seguinte.

3.1.1.1 Versão Discreta do Problema

O Problema Discreto de Geometria de Distâncias Moleculares (do inglês *Discretizable Molecular Distance Geometry Problem - DMDGP*) é definido como dado um grafo simples ponderado não direcionado $G = (V, E, d)$ e uma ordem dada aos vértices $v_1, ..., v_n$ de V chamada *backbone ordering* que satisfaz os seguintes requisitos (LAVOR et al., 2011a):

1. E contém todas as 4-cliques de vértices consecutivos:
   $\forall i \in \{4, ..., n\} \quad \forall j, k \in \{i-3, ...i\} \quad (\{j, k\} \in E)$

2. A inequação triangular estrita $\forall i \in \{4, ..., n\} \quad d_{i-3,i-1} < d_{i-3,i-2} + d_{i-2,i-1}$ é verdadeira,

existe uma imersão válida $x : V \to \mathbb{R}^3$ tal que $||x_u - x_v|| = d_{uv}$ para todo $\{u, v\} \in E$?

As distâncias $d_{i-1,i}$ são chamadas de *comprimento de ligação* para $i \in \{2, ..., n\}$ e os ângulos $\theta_{i-2,i}$ entre os segmentos ligando os átomos v_{i-2}, v_{i-1} e v_{i-1}, v_i são chamados de *ângulos de ligação* para $i \in \{3, ..., n\}$.

O primeiro requisito requer que os comprimentos e ângulos de ligação entre átomos separados por três ligações consecutivas sejam conhecidas.

O segundo requisito força os ângulos de ligação a não serem múltiplos de π. E significa que a imersão do átomo v, denotada por x_v, está na interseção das três esferas S_3 , S_2 e S_1 centralizadas em x_{v-3}, x_{v-2} e x_{v-1} de raios $d_{v-3,v}$, $d_{v-2,v}$ e $d_{v-1,v}$ como pode ser visto na Figura 3.1.

Segundo Lavor (LAVOR et al., 2011a) a interseção entre as três esferas pode ser:

- um ponto - que tem probabilidade 0;

- dois pontos;

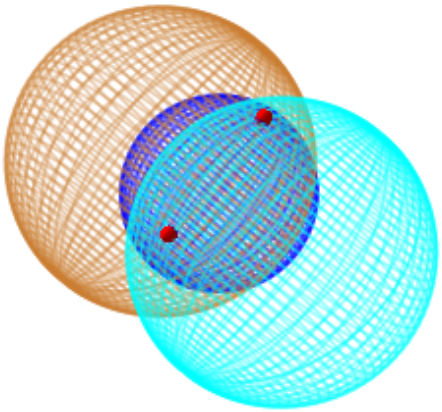


Figura 3.1: Interseção de três esferas.

- um círculo - impossível neste caso por causa da restrição da inequação.

A discretização exige que todos os 4-cliques de vértices consecutivos sejam subgrafos de G, cada 3-subcliques $k_3^v = \{v-3, v-2, v-1\}$ é usada para testar a desigualdade de triângulos. Se as duas suposições forem verdadeiras então cada átomo terá somente duas possibilidades de posição, como mostrado na Figura 3.2.

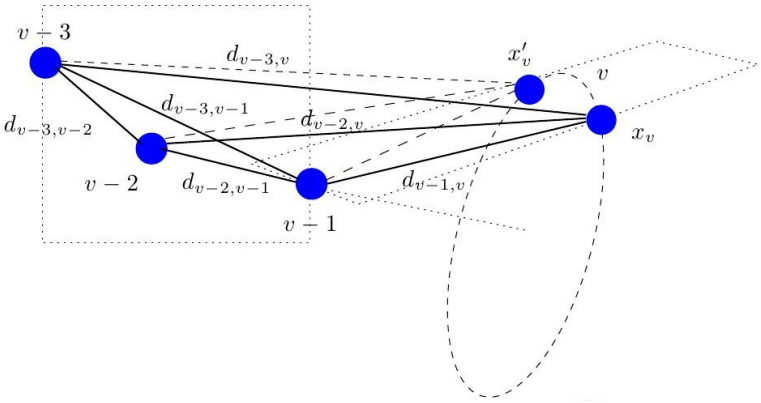


Figura 3.2: Possibilidade de posição de v (SILVA; LAVOR; OCHIAND, 2008).

Dada uma instância do MDGP para determinar se é um problema discreto basta encontrar uma ordem dos vértices que satisfaça os dois requisitos.

3.1.2 Estratégias de Resolução

Existem vários algoritmos para resolver o MDGP, a grande maioria deles resolve a versão contínua clássica do problema, como o *Geometric Build-up*. Entretanto para a versão discreta do problema 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.

A maioria dos métodos usa uma das duas seguintes técnicas matemáticas para resolver o problema:

- Resolução de sistemas (não)lineares formados a partir das equações das distâncias Euclideanas interatômicas - usa distâncias interatômicas para calcular a interseção de esferas resolvidas através de sistemas lineares.

- Resolução do sistema de matrizes formados a partir dos dados de coordenadas internas da molécula - utiliza ângulos de ligação, torção (diedro) para calcular a interseção de esferas, através de técnicas de multiplicação de matrizes.

No próximo capítulo serão apresentados os resultados parciais obtidos com o desenvolvimento da pesquisa, envolvendo a implementação de alguns algoritmos da literatura e principalmente a adaptação de um método de resolução do MDGP se baseando nas duas formas apresentadas acima para o cálculo da interseção de quatro esferas.

Existem vários trabalhos envolvendo o Problema de Geometria de Distâncias Moleculares, nessa Seção eles estão divididos em duas classificações: os trabalhos que resolvem o problema de conjunto completo de distâncias onde todas as distâncias entre todos os pares de átomos são conhecidas e aqueles para o conjunto arbitrário de distâncias onde somente um pequeno conjunto de distâncias é fornecida.

3.1.2.1 Métodos para o Conjunto Completo de Distâncias

O principal trabalho para essa classe especial do problema é o *A linear-time algorithm for solving the molecular distance geometry problem with exact inter-atomic distances* desenvolvido por Dong e Wu, publicado no journal *Global of Optimization* em 2002 (DONG; WU, 2002).

Nesse trabalho, os autores propuseram um algoritmo de tempo polinomial que usa a técnica de distâncias interatômicas, resolvendo o problema através de decomposição de uma matriz de distâncias formada pelas distâncias conhecidas .

Se todas as distâncias são conhecidas, é possível organiza-las em uma matriz $d = [d_{i,j}]$ onde $d_{i,j}$ corresponde a distância entre os átomos i e j. O problema é encontrar o conjunto de coordenadas $x_1, x_2, ...x_n$ para os quais as distâncias $d_{i,j}$ entre os pontos i e j sejam

mantidas. Sendo x_i definido por $x_i = (u_i, v_i, w_i)^T$, onde u_i, v_i e w_i são as três coordenadas do átomo i.

O algoritmo é baseado na relação das distâncias euclidianas entre os pontos e suas coordenadas. No espaço tridimensional, se as distâncias entre quatro átomos que não pertencem ao mesmo plano são conhecidas é possível determinar as coordenadas dos demais átomos a partir das distâncias entre o átomo desconhecido e os quatro átomos fixos. Como mostrado na Figura 3.3.

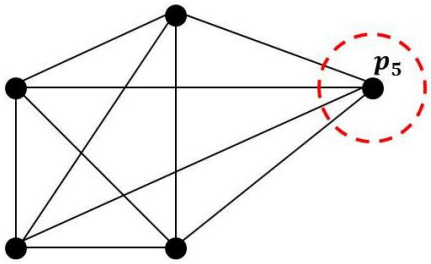


Figura 3.3: Quinto ponto determinado a partir das distâncias para os outros quatro.

O Algoritmo 1 é divido em duas etapas:

- Primeira etapa - os quatro primeiros átomos são fixados determinando suas coordenadas;
- Segunda Etapa - os demais átomos são encontrados a partir dos quatro.

Algoritmo 1: Algoritmo Dong Wu

Fixar átomo x_1, x_2, x_3 e x_4

para $i = 5$ *até* n **faça**
| Encontrar o átomo i usando os quatro átomos fixos

fim para

3.1.2.1.1 Primeira Etapa

O primeiro átomo é colocado na origem $u_1 = 0$, $v_1 = 0$ e $w_1 = 0$. O segundo átomo é fixado em um dos eixos, por exemplo no primeiro eixo colocando $u_2 = d_{1,2}$, $v_2 = 0$ e $w_2 = 0$.

O terceiro átomo é escolhido dentre os restantes que não estejam na mesma linha determinada pelos dois primeiros átomos e colocado no plano formado pelos eixos, por exemplo entre o primeiro e segundo eixo ficando assim com $w_3 = 0$, como mostrado na

Figura 3.4. As duas demais coordenadas são encontradas pelas distâncias do átomo para os dois primeiros:

$$u_3^2 + v_3^2 = d_{3,1}^2 \tag{3.3}$$

$$(u_3 - u_2)^2 + v_3^2 = d_{3,2}^2 \tag{3.4}$$

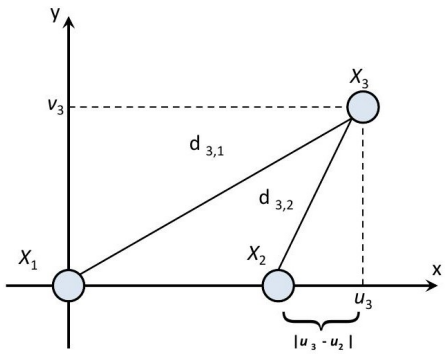


Figura 3.4: Adição do terceiro átomo.

Isolando v_3^2 em (3.3): $v_3^2 = d_{3,1}^2 - u_3^2$

Substituindo em (3.4):

$$\begin{aligned}
(u_3 - u_2)^2 + d_{3,1}^2 - u_3^2 &= d_{3,2}^2 \\
\cancel{u_3^2} - 2u_3u_2 + u_2^2 + d_{3,1}^2 - \cancel{u_3^2} &= d_{3,2}^2 \\
2u_3u_2 &= u_2^2 + d_{3,1}^2 - d_{3,2}^2 \\
u_3 &= \frac{d_{3,1}^2 - d_{3,2}^2 + u_2^2}{2u_2}
\end{aligned}$$

E então as coordenadas podem ser encontradas pelas equações:

$$\begin{aligned}
u_3 &= \frac{(d_{3,1}^2 - d_{3,2}^2)}{2u_2} + \frac{u_2}{2} \\
v_3 &= \pm(d_{3,1}^2 - u_3^2)^{\frac{1}{2}}
\end{aligned}$$

O valor de v_3 pode ser positivo ou negativo sem que isso afete a estrutura final da molécula, então o valor positivo é escolhido.

O quarto átomo é escolhido dentre os restantes desde que não possa ser colocado no mesmo plano formado pelos três primeiros átomos. As coordenadas do átomo podem ser encontradas com as equações de distâncias euclidianas entre o átomo e os três primeiros fixos.

$$u_4^2 + v_4^2 + w_4^2 = d_{4,1}^2 \tag{3.5}$$
$$(u_4 - u_2)^2 + v_4^2 + w_4^2 = d_{4,2}^2 \tag{3.6}$$
$$(u_4 - u_3)^2 + (v_4 - v_3)^2 + w_4^2 = d_{4,3}^2 \tag{3.7}$$

Isolando w_4^2 de (3.5): $w_4^2 = d_{4,1}^2 - u_4^2 - v_4^2$.
Substituindo em (3.6):

$$\begin{aligned}
(u_4 - u_2)^2 + \cancel{y_4^2} + d_{4,1}^2 - u_4^2 - \cancel{y_4^2} &= d_{4,2}^2 \\
\cancel{u_4^2} - 2u_4u_2 + u_2^2 + d_{4,1}^2 - \cancel{u_4^2} &= d_{4,2}^2 \\
2u_4u_2 &= d_{4,1}^2 - d_{4,2}^2 + u_2^2 \\
u_4 &= \frac{d_{4,1}^2 - d_{4,2}^2 + u_2^2}{2u_2}
\end{aligned}$$

Partindo de (3.7) para isolar v_4 e substituindo w_4^2 de (3.6), $w_4^2 = d_{4,2}^2 - (u_4 - u_2)^2 - v_4^2$:

$$\begin{aligned}
(u_4 - u_3)^2 + (v_4 - v_3)^2 + d_{4,2}^2 - (u_4 - u_2)^2 - v_4^2 &= d_{4,3}^2 \\
(u_4 - u_3)^2 + \cancel{y_4^2} - 2v_4v_3 + v_3^2 + d_{4,2}^2 - (u_4 - u_2)^2 - \cancel{y_4^2} &= d_{4,3}^2 \\
2v_4v_3 = (u_4 - u_3)^2 + v_3^2 + d_{4,2}^2 - (u_4 - u_2)^2 - d_{4,3}^2 \\
v_4 = \frac{d_{4,2}^2 - d_{4,3}^2 - (u_4 - u_2)^2 + (u_4 - u_3)^2 + v_3^2}{2v_3}
\end{aligned}$$

E então as coordenadas do quarto átomo podem ser encontradas pelas equações:

$$\begin{aligned}
u_4 &= \frac{d_{4,1}^2 - d_{4,2}^2}{2u_2} + \frac{u_2}{2} \\
v_4 &= \frac{d_{4,2}^2 - d_{4,3}^2 - (u_4 - u_2)^2 + (u_4 - u_3)^2}{2v_3} + \frac{v_3}{2} \\
w_4 &= \pm(d_{4,1}^2 - u_4^2 - v_4^2)^{\frac{1}{2}}
\end{aligned}$$

O valor de w_4 pode ser positivo ou negativo correspondendo a duas estruturas simétricas espelhadas. O valor positivo é escolhido, mas para obter a segunda estrutura basta trocar os sinais de todos os w_i para $i >= 4$.

Os átomos restantes podem ser encontrados apartir das coordenadas dos quatro primeiros através das distâncias entre eles e os quatro fixos, como mostrado na segunda etapa.

3.1.2.1.2 Segunda Etapa

As distâncias entre os átomos podem ser escritas formalmente como:

$$||x_i - x_j|| = d_{i,j} \quad para \quad i, j = 1, 2, ...n$$

Ou equivalente

$$\begin{aligned} ||x_i||^2 &= d_{i,0}^2 \\ ||x_i - x_j||^2 &= d_{i,j}^2 \end{aligned} \tag{3.8}$$

A equação (3.8) é equivalente a:

$$d_{i,j}^2 = ||x_i||^2 - 2x_i^T x_j + ||x_j||^2$$

Considera-se as coordenadas dos quatro átomo conhecidas como:

$$\begin{aligned} x_1 &= (u_1, v_1, w_1)^T \qquad x_2 = (u_2, v_2, w_2)^T \\ x_3 &= (u_3, v_3, w_3)^T \qquad x_4 = (u_4, v_4, w_4)^T \end{aligned}$$

A partir das posições conhecidas é possível determinar a coordenada $x_i = (u_i, v_i, w_i)^T$ do átomo i. Como as distâncias entre todos os átomos são conhecidas, pode-se definir as distâncias entre o átomo i e j sendo $j = 1, 2, 3, 4$ como:

$$||x_i - x_1|| = d_{i,1} \qquad ||x_i - x_2|| = d_{i,2}$$
$$||x_i - x_3|| = d_{i,3} \qquad ||x_i - x_4|| = d_{i,4}$$

equivalente a:

$$||x_i - x_1||^2 = ||x_i||^2 - 2x_i^T x_1 + ||x_1||^2 = d_{i,1}^2$$
$$||x_i - x_2||^2 = ||x_i||^2 - 2x_i^T x_2 + ||x_2||^2 = d_{i,2}^2$$
$$||x_i - x_3||^2 = ||x_i||^2 - 2x_i^T x_3 + ||x_3||^2 = d_{i,3}^2$$
$$||x_i - x_4||^2 = ||x_i||^2 - 2x_i^T x_4 + ||x_4||^2 = d_{i,4}^2$$

considerando $x_i = (u_i, v_i, w_i)^T$ então:

$$||x_i||^2 - 2u_iu_1 - 2v_iv_1 - 2w_iw_1 + ||x_1||^2 = d_{i,1}^2 \tag{3.9}$$
$$||x_i||^2 - 2u_iu_2 - 2v_iv_2 - 2w_iw_2 + ||x_2||^2 = d_{i,2}^2 \tag{3.10}$$
$$||x_i||^2 - 2u_iu_3 - 2v_iv_3 - 2w_iw_3 + ||x_3||^2 = d_{i,3}^2 \tag{3.11}$$
$$||x_i||^2 - 2u_iu_4 - 2v_iv_4 - 2w_iw_4 + ||x_4||^2 = d_{i,4}^2 \tag{3.12}$$

Subtraindo a equação (3.9) da equação (3.10):

$$\cancel{||x_i||^2} - 2u_iu_2 - 2v_iv_2 - 2w_iw_2 + ||x_2||^2 - \cancel{||x_i||^2} +$$
$$+2u_iu_1 + 2v_iv_1 + 2w_iw_1 - ||x_1||^2 = d_{i,2}^2 - d_{i,1}^2$$

$$2u_i(u_1 - u_2) + 2v_i(v_1 - v_2) + 2w_i(w_1 - w_2) =$$
$$= (||x_1||^2 - ||x_2||^2) - (d_{i,1}^2 - d_{i,2}^2)$$

Fazendo a mesma subtração de (3.9) nas equações (3.11) e (3.12):

$$2u_i(u_1-u_3)+2v_i(v_1-v_3)+2w_i(w_1-w_3)= \\ =(||x_1||^2-||x_3||^2)-(d_{i,1}^2-d_{i,3}^2)$$

$$2u_i(u_1-u_4)+2v_i(v_1-v_4)+2w_i(w_1-w_4)= \\ =(||x_1||^2-||x_4||^2)-(d_{i,1}^2-d_{i,4}^2)$$

Em forma de matriz as equações se reduzem a $Ax_i = b_i$, onde:

$$A = 2\begin{bmatrix} u_1-u_2 & v_1-v_2 & w_1-w_2 \\ u_1-u_3 & v_1-v_3 & w_1-w_3 \\ u_1-u_4 & v_1-v_4 & w_1-w_4 \end{bmatrix}$$

$$b_i = \begin{bmatrix} (||x_1||^2-||x_2||^2)-(d_{i,1}^2-d_{i,2}^2) \\ (||x_1||^2-||x_3||^2)-(d_{i,1}^2-d_{i,3}^2) \\ (||x_1||^2-||x_4||^2)-(d_{i,1}^2-d_{i,4}^2) \end{bmatrix}$$

e

$$x_i = \begin{bmatrix} u_i \\ v_i \\ w_i \end{bmatrix}$$

Com a resolução desse sistema linear é possível encontrar a coordenada do átomo i.

3.1.2.2 Métodos para o Conjunto Incompleto de Distâncias

Para o conjunto arbitrário de distâncias os trabalhos foram divididos em dois tipos, os voltados para a versão contínua e aqueles para versão discreta do problema.

3.1.2.2.1 Versão Contínua - *Geometric Build Up*

O trabalho *A Geometric Build-Up Algorithm for Solving the Molecular Distance Ge-*

ometry Problem with Sparse Distance Data de Qunfeng Dong e Zhijun Wu publicado no journal *Journal of Global Optimization* em 2003 (DONG; WU, 2003) mostra que quando o problema fornece apenas um conjunto incompleto de distâncias, o conjunto base fixo não pode ser usado já que algumas distâncias não são conhecidas. Mas é possível usar a mesma idéia de determinar a coordenada de um átomo atráves de quatro já imersos no plano, quando as distâncias desses átomos para o átomo i forem conhecidas.

Ao invés de usar uma base fixa, para cada átomo não imerso procura-se um conjunto base de quatro átomos já fixados dos quais são conhecidas as distâncias entre qualquer um dos quatro átomos e o não fixo. Se esses quatro átomos são encontrados, as coordenadas para o quinto átomo podem ser calculadas através do sistema de equações mostrado na Sessão 3.1.2.1.2, o GBU mostrado no Algoritmo 2 continua até que todos os átomos estejam fixados.

Algoritmo 2: Algoritmo GBU

Fixar os quatro átomos iniciais x_1, x_2, x_3 e x_4

repita

  para $i \in$ *conjunto de átomos não fixados* **faça**

    Procurar o conjunto base de quatro átomos fixados para o átomo i

    Fixar o átomo i usando a base encontrada

  fim para

  se *nenhum átomo for fixado no* ***para*** *anterior* **então**

    Parar **repita**, estrutura parcialmente encontrada

  fim se

até *todos os átomos estejam fixados*;

3.1.2.3 Método para a Versão Discreta - Algoritmo *Branch and Prune* (BP)

O trabalho The discretizable molecular distance geometry problem de Carlile Lavor, Leo Liberti, Nelson Maculan e Antonio Mucherino publicado no journal *Computacional Optimization and Aplication* em 2011 apresenta o Branch-and-Prune que é um dos algoritmos que resolvem o MDGP quando somente é conhecido um conjunto incompleto de distâncias (LAVOR et al., 2011a).

O algoritmo atribui uma ordem aos átomos da molécula, e para cada átomo $v \in V$ que será imerso no $\mathbb{R}^3$ são feitas duas suposições para resolução:

- São conhecidas imersões válidas para todos os átomos que precedem v;

- As arestas $\{v-3,v\}$, $\{v-2,v\}$ e $\{v-1,v\} \in E$. São conhecidas as distâncias $d_{v-3,v-1}$, $d_{v-3,v-2}$ e $d_{v-2,v-1}$ e a inequação triangular estrira $d_{v-3,v-1} < d_{v-3,v-2} + d_{v-2,v-1}$ é verdadeira.

A segunda suposição significa que a imersão do átomo v, denotada por x_v, está na interseção de três esferas como pode ser visto na Figura 3.1. O algoritmo Branch-and-Prune proposto por Lavor et al. (LAVOR et al., 2011b), mostrado no Algoritmo 3 possui cinco argumentos de entrada:

- grafo G;
- vértice v que será imerso em $\mathbb{R}^3$;
- subconjunto $U \subseteq N(v)$ sendo $N(v) =$ conjunto de adjacente de v com $|U| = 3$;
- uma imersão válida x' para o subgrafo $G[U]$;
- um conjunto X de imersões válidas de G já encontradas.

Algoritmo 3: BranchAndPrune(G, v, U, x', X)

P=interseção das 3 esferas $S^2(x'_u, d_{u,v})$ para $u \in U$
para todo $p \in P$ **faça**
  $x = (x'_u, p)$
  se *x é uma imersão válida* **então**
    se *x é o último átomo* **então**
      Acrescenta x a X
    senão
      $U' = (U - min(U)) \cup \{v\}$ BranchAndPrune($G, v+1, U', x, X$)
    fim se
  fim se
fim para todo

A recursão começa com uma chamada BranchAndPrune($G, 4, \{1,2,3\}, y, \varnothing$), onde y é uma imersão válida para os átomos $\{1,2,3\}$. O algoritmo constrói uma árvore binária na qual cada nível v representa possíveis posições espaciais p para o vértice v. No fim da execução o conjunto X contém todas as imersões válidas de G estendendo x'.

A poda pode ser feita pelo método *Direct Distance Feasibility (DFF)* que considera as distâncias de v para os vértices do subconjunto $\overline{U} = \{u \in V \mid u < v \wedge u \in N(v) \wedge u \notin U\}$, ou seja, vértices que tem distâncias conhecidas para v e não foram utilizados para

determinar sua posição. Se $||x_u - x_v|| \neq d_{u,v}$ então x_v não é uma imersão válida e toda subárvore abaixo dele pode ser podada.

A Figura 3.5 mostra a árvore binária resultante da execução do algoritmo Branch-and-Prune. Após fixar os três primeiros átomos de base no plano cartesiano, para cada átomo da proteína o BP gera duas possibilidades de posições matematicamente válidas usando a intersecção de três esferas. Para testar se a nova ramificação gerada é válida, o algoritmo procura no conjunto de entrada por uma distância conhecida entre o vértice atual e um já imerso que não fez parte da determinação de sua posição, então um teste de poda é feito usando o método DDF, se o método invalida a incorporação, então toda a subárvore é podada.

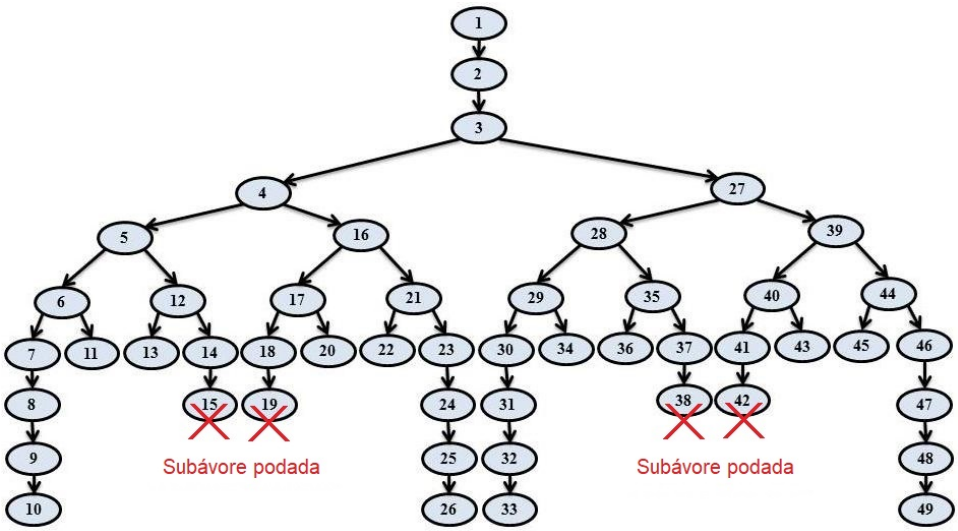


Figura 3.5: Árvore binária resultante do branch-and-prune.

Considerando que em uma molécula com n átomos onde cada átomo i tem coordenada $x_i \in \mathbb{R}^3$, existe uma ligação covalente entre o átomo i e o $i+1$ de comprimento igual a distância euclidiana entre x_i e x_{i+1}. De acordo com Liberti (LIBERTI; LAVOR; MACULAN, 2007) "O ângulo de ligação $\theta_i \in [0, \pi]$ é o ângulo entre os seguimentos ligando os átomos $i-2, i-1$ e $i-1, i$ (para todo $i = 3, ..., n$). O ângulo de torção $\omega_i \in [0, 2\pi]$ é o ângulo entre as normais dos planos definidos pelos átomos $i-3, i-2, i-1$ e $i-2, i-1, i$ (para todo $i = 4, ..., n$)".

Os ângulos de ligação $\theta_3, \theta_4, ..., \theta_n$ são calculados através da lei dos cossenos utilizando as distâncias conhecidas entre os átomos. Os três primeiros átomos são fixados utilizando as distâncias e os ângulos de ligação. A partir do quarto átomo são utilizados também os ângulos de torção calculados pela fórmula (SILVA; LAVOR; OCHIAND,

2008):

$$\cos w_{i-3,i} = \frac{d_{i-3,i-2}^2 + d_{i-2,i}^2 - 2d_{i-3,i-2}d_{i-2,i}\cos\theta_{i-2,i}\cos\theta_{i-1,i+1} - d_{i-3,i}^2}{2d_{i-3,i-2}d_{i-2,i}\sin\theta_{i-2,i}\sin\theta_{i-1,i+1}}$$

Tendo as distâncias entre os átomos, os ângulos de ligação e ângulos de torção, as coordenadas cartesianas $x_i = (u_i, v_i, w_i)$ para cada átomo na molécula podem ser obtidos pelas seguintes fórmulas (SILVA; LAVOR; OCHIAND, 2008),(LIBERTI; LAVOR; MACULAN, 2007):

$$\begin{bmatrix} u_i \\ v_i \\ w_i \\ 1 \end{bmatrix} = B_1 B_2 \ldots B_i \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \forall i = 1, \ldots n$$

onde

$$B_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, B_2 = \begin{bmatrix} -1 & 0 & 0 & -d_{1,2} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

,

$$B_3 = \begin{bmatrix} -\cos\theta_{1,3} & -\sin\theta_{1,3} & 0 & -d_{2,3}\cos\theta_{1,3} \\ \sin\theta_{1,3} & -\cos\theta_{1,3} & 0 & d_{2,3}\cos\theta_{1,3} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

,

$$B_i = \begin{bmatrix} -\cos\theta_{i-2,i} & -\sin\theta_{i-2,i} & 0 & -d_{i-1,i}\cos\theta_{i-2,i} \\ \sin\theta_{i-2,i}\cos\omega_{i-3,i} & -\cos\theta_{i-2,i}\cos\omega_{i-3,i} & -\sin\omega_{i-3,i} & d_{i-1,i}\sin\theta_{i-2,i}\cos\omega_{i-3,i} \\ \sin\theta_{i-2,i}\sin\omega_{i-3,i} & -\cos\theta_{i-2,i}\sin\omega_{i-3,i} & \cos\omega_{i-3,i} & d_{i-1,i}\sin\theta_{i-2,i}\sin\omega_{i-3,i} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

3.1.3 Trabalhos Relacionados

Lavor (LAVOR, 2018) apresentada uma nova ordem nos átomos da proteína, baseada nas informações químicas das proteínas obtidas pelos experimentos de RMN, o problema é então formulado como um problema de busca combinatorial.

A principal contribuição do artigo é combinar informações da geometria de proteínas (hipótese de geometria rígida, plano peptídico e quiralidade) e experimentos de RMN para modelar o problema do cálculo de proteínas 3D usando dados de RMN. Uma nova ordem chamama ordem HC é proposta, mostrada na Figura 3.6, onde $i = 2, ..., p-1, H^{1'}$ é o segundo hidrogênio ligado a N^1 e $O^{p'}$ é o segundo oxigênio ligado a Cp.

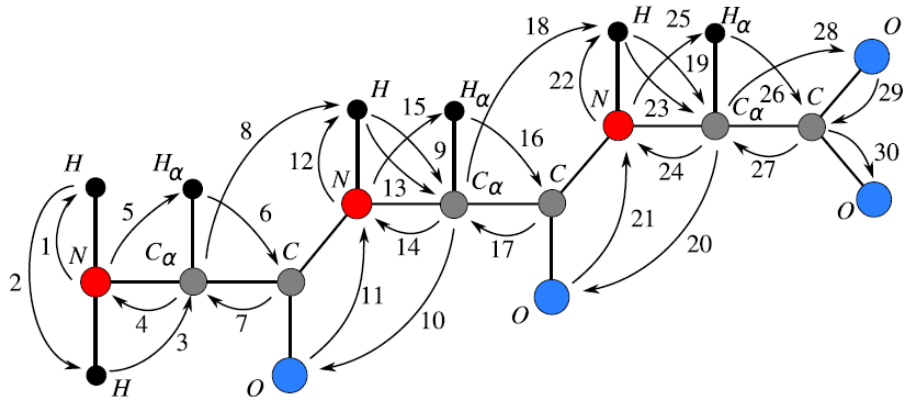


Figura 3.6: Ordem HC (LAVOR, 2018).

Alencar et al. (ALENCAR; LAVOR; LIBERTI, 2018) apresentam as propriedades teóricas de um algoritmo para encontrar uma realização de uma matriz de distância euclidiana (completa) $n \times n$ na menor dimensão de incorporação possível. São explicadas as propriedades teóricas de um novo algoritmo que determina se uma dada matriz simétrica oca, com diagonal zero, com elementos não negativos é um EDM, então o algoritmo cal-

cula a dimensão de incorporação da matriz, juntamente com uma incorporação real. Este artigo trata apenas do caso de distâncias exatas.

Mucherino (MUCHERINO, 2018) apresentada um algoritmo adaptado do Branch an Prune (algoritmo de ramificação e poda) para o espaço unidimensional euclidiano e a as distâncias em intervalos, que é chamado BP1. O algoritmo BP1 usa a poda de retorno, que permite refinar todos os intervalos de posição do vértice, para que eles contenham apenas posições viáveis. O BP1 pode ser aplicado a todas as instâncias da dimensão 1 para as quais, na ordem dada do vértice, existe pelo menos um vértice de referência para cada vértice do gráfico. No artigo é explicado que para dimensões maiores que 1 ainda é um problema em aberto conceber métodos e algoritmos para uma exploração completa do conjunto de soluções.

3.2 Modelagem Molecular de Moléculas Complexas

Moléculas complexas são compostas de carbono ligado a outros elementos, especialmente oxigênio, hidrogênio e nitrogênio. São moléculas semelhantes a polímeros, como as proteínas, devido a imensa variedade possível desses polímeros, eles são de complexos. (KOBAYASHI, 2011)

O termo Modelagem Molecular se refere aos métodos teóricos e técnicas computacionais para modelar ou mimetizar o comportamento das moléculas. A seguir é mostrado algumas definições e metódos de modelagem de moléculas de proteína.

3.2.1 Bioquímica das Proteínas

As Proteínas são compostos orgânicos de alto peso molecular, construídas a partir de aminoácidos arranjados em várias sequências específicas e ligados entre si através de ligações peptídicas . São consideradas as macromoléculas mais importantes das células e chegam a constituir quase 50% de suas massas.

Os aminoácidos são compostos que carregam em suas moléculas, como o nome sugere, um grupo amino (básico) e um grupo carboxílico (ácido). Seus componentes principais são carbono, hidrogênio, oxigênio, nitrogênio e enxofre, algumas podem conter também fósforo, ferro, zinco e cobre.

Análises de vários tipos de proteínas mostram que todas as proteínas são formadas a

partir de um conjunto padrão de vinte aminoácidos, chamados de $\alpha-$aminoácidos (VOET; VOET, 2006). Os $\alpha-$aminoácidos são formados por um carbono central (C_α) ligado a quatro grupos: grupo amina (NH_2), grupo carboxila ($COOH$), hidrogênio e cadeia lateral que é específica para cada aminoácido.

A Figura 3.7 mostra a estrutura geral dos $\alpha-$aminoácidos (VOET; VOET, 2006), onde são apresentadas as diferentes regiões que constituem um aminoácido. Os $\alpha-$aminoácidos unem-se através de **ligação peptídica**. Quando a proteína está sendo sintetizada o radical carboxila de um aminoácido perde um grupamento $-OH$ (hidroxila) e o radical amina de outro aminoácido perde um átomo de hidrogênio $-H$. Os aminoácidos então se unem e o OH se liga ao H, gerando como produto uma molécula de água. A ligação peptídica ocorre entre o carbono (C) de um aminoácido e o nitrogênio (N) do aminoácido próximo, classificada como ligação covalente ($C-N$) (VOET; VOET, 2006).

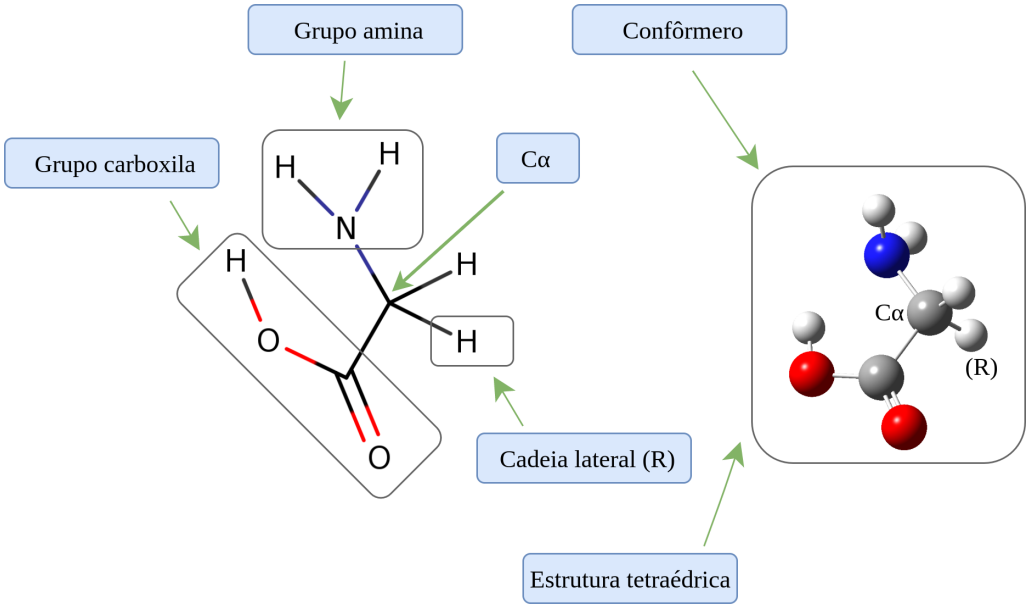


Figura 3.7: Estrutura do aminoácido Glicina (Gly), o menos complexo dentre todos os aminoácidos, contendo apenas um único hidrogênio como cadeia lateral.

Os aminoácidos são substâncias anfotéricas, ou seja, podem se comportar como ácido ou base liberando nesta ordem H ou OH em uma reação. Se o aminoácido estiver dissolvido em um meio ácido ele se torna carregado positivamente e se o meio for alcalino ele se torna carregado negativamente. Se a reação for entre dois aminoácidos o grupo amina de um libera um H se ligando ao grupo carboxila do outro que libera um OH formando uma peptídeo a mais de H_2O.

3.2.1.1 Classificação dos aminoácidos

Os aminoácidos se dividem em três grupos de classificação de acordo com a polaridade da cadeia lateral: aminoácidos com cadeias laterais apolares, aminoácidos com cadeias laterais polares sem carga e aminoácidos com cadeias laterais polares com carga (positiva ou negativa).

Aminoácidos com cadeias laterais apolares também chamados de hidrofóbicos são menos solúveis em água. O grupo inclui a glicina, cinco aminoácidos com cadeias laterais hidrocarbonadas alifática hidrocarbonadas (alanina, valina, leucina, isoleucina e prolina), dois com anéis aromáticos (fenilalanina e triptófano) e um com enxofre (metionina). Os aminoácidos apolares são mostrados na Figura 3.8.

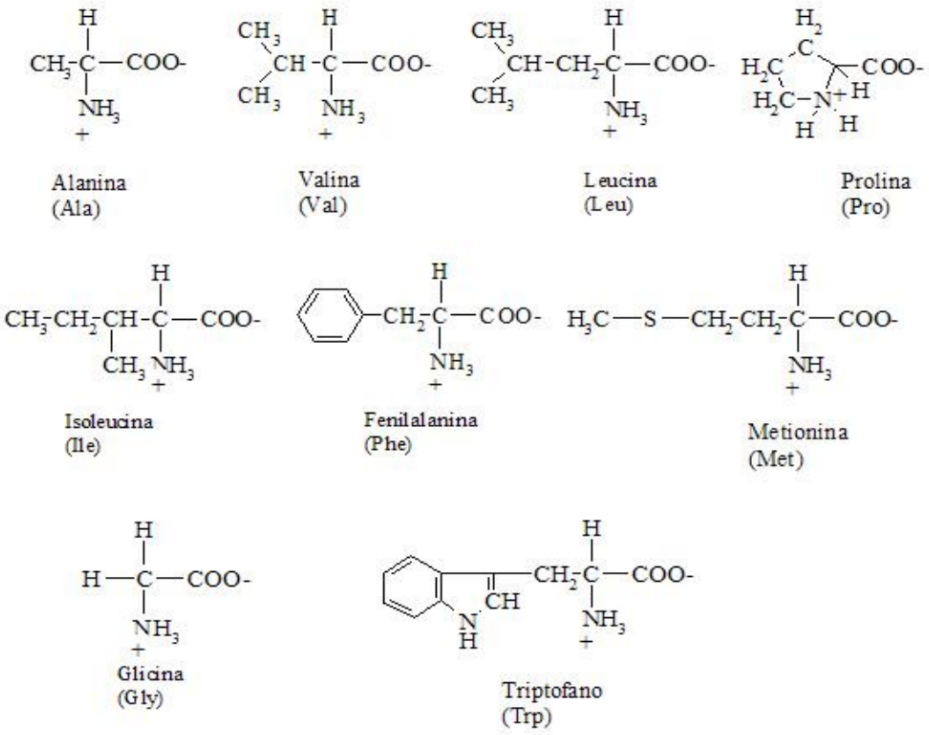


Figura 3.8: Aminoácidos com cadeias laterais apolares.

Aminoácidos com cadeias laterais polares sem carga contém grupos funcionais não carregados que podem formar ligações de hidrogênio com a água. O grupo inclui aminoácidos hidroxilados (serina, treonina e tirosina), amídicos (asparagina e glutamina) e sulfídricos (cisteína). Os aminoácidos polares sem carga são mostrados na Figura 3.9.

Aminoácidos com cadeias laterais polares com carga podem ser carregados positiva ou negativamente. Os aminoácidos nos quais a cadeia lateral apresenta carga positiva

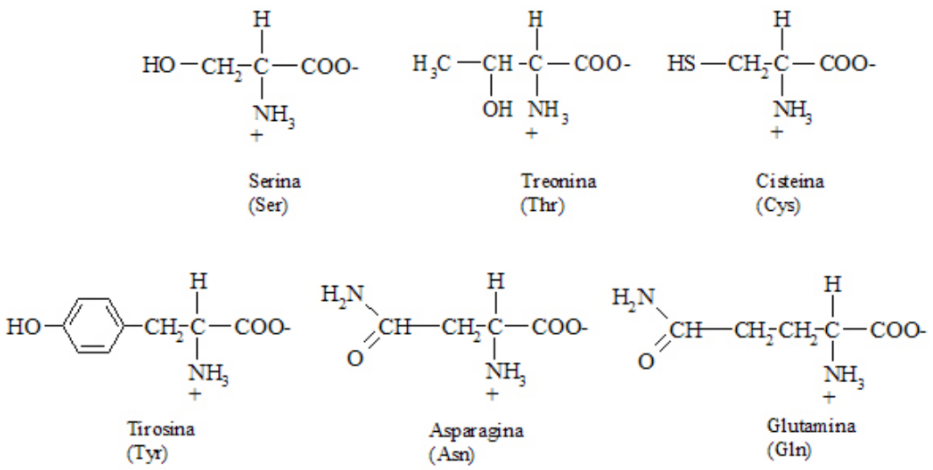


Figura 3.9: Aminoácidos com cadeias laterais polares sem carga.

têm todos seis carbonos em sua composição, são eles: lisina, arginina e histidina. Os aminoácidos nos quais a cadeia lateral apresenta carga negativa são os ácidos aspártico e glutâmico, ambos com um grupo carboxílico a mais que é inteiramente ionizado. Os aminoácidos polares carregados positivamente são mostrados na Figura 3.10 e os carregados negativamente são mostrados na Figura 3.11.

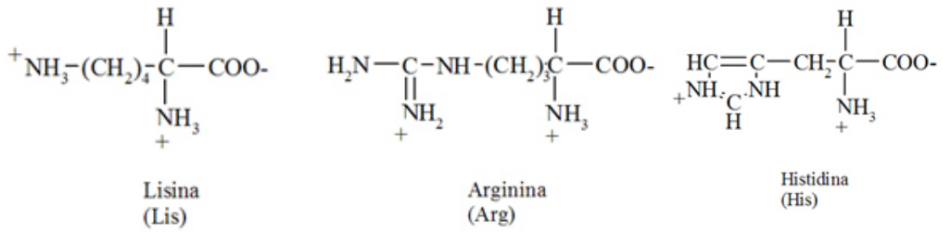


Figura 3.10: Aminoácidos com cadeias laterais polares com carga positiva.

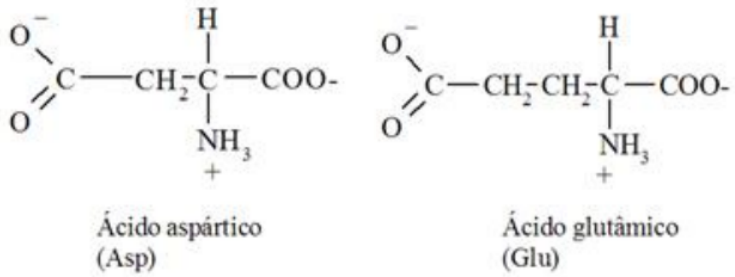


Figura 3.11: Aminoácidos com cadeias laterais polares com carga negativa.

3.2.1.2 Ligações Peptídicas

O $\alpha-$aminoácidos unem-se através de ligação peptídica, a ligação ocorre entre o grupo carboxila de um deles e o grupo amino do outro formando um peptídio e liberando uma molécula de água, como pode ser visto na Figura 3.12.

Quando a proteína está sendo sintetizada o radical carboxila de um aminoácido perde um grupamento $-OH$ (hidroxila) e o radical amina de outro aminoácido perde um átomo de hidrogênio $-H$. Os aminoácidos então se unem e o OH se liga ao H, produzindo uma molécula de água. A ligação peptídica ocorre entre o carbono (C) de um aminoácido e o nitrogênio (N) do aminoácido vizinho, classificada como ligação covalente ($C-N$).

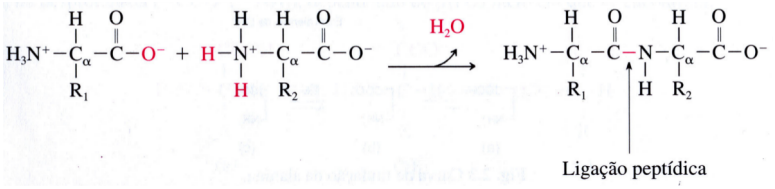


Figura 3.12: Ligação peptídica.

As moléculas que se formam a partir de dois aminoácidos são chamados de dipeptídeo, três formam um tripeptídeo, quatro produzem um tetrapeptídeo, alguns poucos (3-10) são chamados de oligopepídeos (do grego oligo, pouco) e aquelas formados por muitos resíduos de aminoácidos são denominados de polipeptídios (do grego poli, muitos). Uma proteína pertence à categoria dos polipeptídios, já que são constituídas por um número expressivo de aminoácidos.

As proteínas variam de comprimento entre 40 e 33.000 resíduos de aminoácidos e como a massa média de um resíduo é de 110 D(dalton), a sua massa molecular pode variar de 4 a 3600 kD.

Pode existir um número gigantesco de moléculas protéicas, uma vez que para cada resíduo existe um conjunto base de vinte aminoácidos disponíveis para utilização. Ou seja, para um dipeptídeo há vinte diferentes escolhas para o primeiro resíduo e mais vinte para o segundo, gerando um total de $20^2 = 400$ possíveis dipeptídeos.

Seguindo o mesmo raciocínio, para um tripeptídeo existe mais vinte possibilidades para o terceiro resíduo, resultando em $20^3 = 8000$ possíveis tripeptídeos. Considerando que uma proteína pequena tenha em média 100 aminoácidos, há $20^{100} = 1,27x10^{130}$ possibilidades de arranjo, mas a natureza criou apenas uma pequena fração dessa possibilidade.

3.3 Os Desafios da Predição Estrutural

Sabe-se que a conformação espacial da estrutura é decisivo na função desempenhada pelas proteínas e está diretamente ligada ao seu enovelamento tridimensional, constituindo-se no principal fator que rege as interações biológicas (LEVY; WOLYNES; ONUCHIC, 2004). O enovelamento de proteínas é crítico em muitas doenças, incluindo diabetes tipo II, assim como doenças neurodegenerativas como Alzheimer e Parkinson. Apesar de grandes avanços neste campo, ainda não é possível predizer de forma precisa as estruturas tridimensionais das proteínas (DILL; MACCALLUM, 2012).

Assim, as pesquisas neste campo possuem valor imprescindível na descoberta de novos fármacos, no entendimento de processos metabólicos e no mecanismo de doenças variadas. No ano de 1968, Cyrus Levinthal percebeu que embora exista um grande número de conformações possíveis a uma proteína, seu dobramento para a estrutura mais estável ocorre com extrema precisão e de forma muito rápida. Ainda hoje, as regras pelas quais a natureza realiza a busca conformacional permanece incerta (DILL; MACCALLUM, 2012).

A predição teórica da estrutura proteica teve grandes avanços pelo grupo de pesquisa de John Moult em 1994 conforme a primeira versão do CASP (Critical Assessment of protein Structure Prediction). Atualmente os algoritmos de predição estrutural baseiam-se na suposição de que sequências semelhantes levam a estruturas semelhantes. Sendo assim, as estruturas podem ser preditas usando modelagem comparativa a partir de estruturas já conhecidas. Esse processo é conhecido como busca por homologia. Entretanto, quando não há estruturas cristalográficas disponíveis com uma sequência com alto de grau de homologia, predizer com precisão a estrutura torna-se muito difícil ou quase impossível. Predições com essas características denominam-se de *ab-initio* ou *primeiros princípios* (HARDIN; POGORELOV; LUTHEY-SCHULTENCHOW, 2002).

Nesse contexto, a predição estrutural tornou-se um problema tanto computacional quanto físico-químico. O que parece impossível à primeira vista, não parece impossível aos processos biológicos que encontram a solução ao enovelamento como um problema de otimização global constituído por uma série de problemas menores de otimização local.

Inexplicavelmente proteínas, por processos metabólicos, podem sofrer conversão rapidamente para complexos estados nativos de menor energia com enovelamento altamente

especificado, às vezes em microssegundos. O enovelamento específico de uma proteína, ainda que complexo, representa a menor entropia possível dentro de um propósito funcional, e por isso a estrutura nativa obtida é termodinamicamente estável.

Em outras palavras, qualquer tentativa de predição *ab initio* de estruturas proteicas deve levar em conta a descrição tridimensional do enovelamento como resultado de interações diversas como ligações de hidrogênio, iônicas, atrações de van der Waals e principalmente contatos hidrofóbicos (DILL et al., 2008). Não é tarefa fácil, uma vez que esta descrição necessitará descrever as distância interatômicas, os ângulos interatômicos e os ângulos diedrais ou de torção entre planos de átomos. Como exemplo do grau de dificuldade tomemos o caso dos ângulos em uma sequência de aminoácidos.

3.3.1 Métodos Utilizados na Determinação Estrutural de Proteínas

A conformação espacial da estrutura proteica devido ao enovelamento tridimensional da sequência de aminoácidos é fundamental na função desempenhada pela proteína no maquinário biológico (LAVOR et al., 2011a). Erros de sequenciamento e na descrição de distâncias e ângulos interatômicos resultam em uma estrutura irreal da proteína.

Métodos que descrevam rigorosamente a conformação espacial de proteínas têm utilizados técnicas experimentais e teóricas como é mostrado na Figura 3.13. Experimentalmente, a estrutura tridimensional pode ser obtida por dois métodos: cristalografia de Raios-X e Ressonância Magnética Nuclear (RMN).

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, em geral na presença de moléculas de água, 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 que não podem ser cristalizada. As técnicas de obtenção de estruturas são variadas e dependem fortemente de auxílio computacional para construção ou inferência de um modelo viável (PURSLOW et al., 2020; GUERRY; HERRMANN, 2011). Em geral as coordenadas atômicas entre conjuntos de pares de átomos são modeladas a partir do sinal de acoplamento de ressonância magnética dos núcleos atômicos, em geral de átomos de hidrogênio (CAVALLI et al., 2007). O sinal RMN fornece apenas um pequeno subconjunto de distâncias entre os átomos de uma molécula, pois só obtêm distâncias entre pares de átomos que estejam

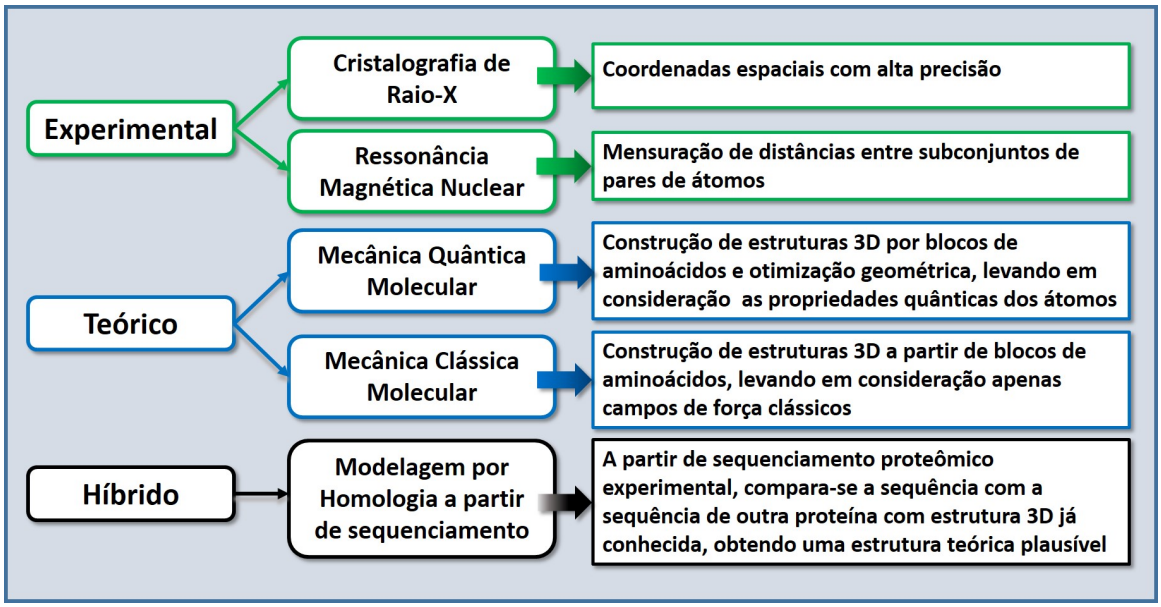


Figura 3.13: Classificação geral dos métodos de obtenção de estrutura terciária de uma proteína.

próximos na faixa de 5 a 6 (FIDALGO et al., 2012).

Os métodos teóricos de construção de estruturas proteicas são muitos variados, de forma geral há duas abordagens básicas, a mecânica clássica e a mecânica quântica. Na clássica apenas as interações de campos de força clássicos, como interações de Coulomb, forças de Van der Waals, entre outras, são aplicados nos átomos que compõe a estrutura.

Na mecânica quântica, o sistema é construído levando em consideração as propriedades quânticas da matéria. A modelagem clássica é computacionalmente barata em termos de processamento, na quântica os custos são proibitivos quando se trata de proteínas de grande massa molecular (o que é comum a grande maioria das proteínas). Entre estes dois métodos há uma profusão de métodos híbridos que mesclam dados experimentais e abordagens teóricas, sejam clássicas, quânticas, mistas (CHOW; ZHANG, 2008; ELSTNER; FRAUENHEIM; SUHAI, 2003) ou de complexidade computacional (SILVA; LAVOR; OCHIAND, 2008).

Em qualquer método usado para construir e validar a estrutura de uma proteína o problema será basicamente o mesmo: 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 (*Molecular Distance Geometry Problem* - MDGP) e formulado tradicionalmente como um problema de otimização contínua (SILVA; LAVOR; OCHIAND, 2008).

3.3.1.1 Cristalografia de Raio X

A cristalografia de raio X é uma ferramenta para se obter a informação estrutural da molécula, desenvolvida em 1912 por William Henry Bragg e William Lawrence Bragga partir de um trabalho mais desenvolvido por Máximo von Laue. Von Laue descobriu que brilhando raios X através de um cristal do sulfato de cobre em uma placa fotográfica, os pontos da difracção que se relacionaram à estrutura cristalina da amostra foram produzidos.

A cristalografia de raios-X só pode ser usada para examinar cristais sólidos com um arranjo regular de átomos. Pode-se estudar minerais, por exemplo, e muitos outros compostos, tais como sal ou açúcar. Também se pode estudar o gelo, mas só até que ele derreta.

Usando a radiação eletromagnética, a cristalografia consegue determinar a estrutura molecular e atômica de um cristal. Os raios X difratam em sentidos específicos quando passam pela estrutura do cristal, com a análise das intensidades e dos ângulos destes feixes, a posição dos elétrons dentro da estrutura cristalina podem ser determinados gerando assim a estrutura tridimensional da molécula. Como mostrado na Figura 3.14

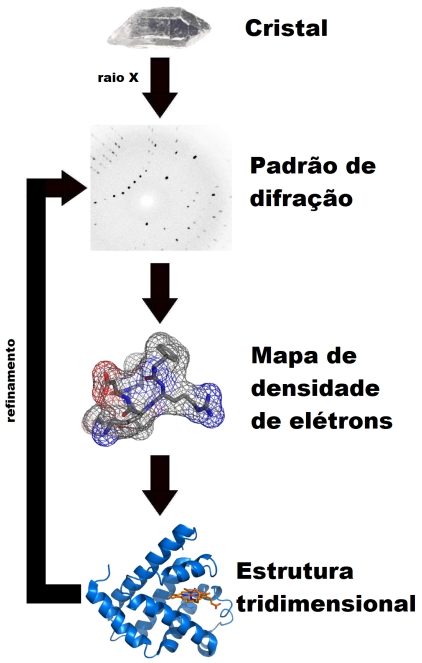


Figura 3.14: Metodologia da Cristalografia de raio X.

3.3.1.2 Ressonância Magnética Nuclear - RMN

O fenômeno da Ressonância Magnética Nuclear (RMN) foi detectado pela primera vez em 1946, inicialmente utilizada somente para estudar pequenas moléculas, mas com o avanço tecnológico começou a ser possível trabalhar com estrutura de proteínas e outras macromoléculas de nível atômico (BOTTON, 2007).

O processo de obtenção da estrutura por RMN utiliza técnicas de RMN pulsada, onde após ser escolhida a largura de pulso de RF adequada, todos os núcleos de um mesmo isótopo da amostra líquida (átomos de um mesmo elemento, mesmo número de prótons e neutons diferentes) são excitados simultaneamente e o sinal observado imediatamente após o pulso contém as frequências de todos os núcleos da amostra, chamado de *free-induction decay* ou FID (MUNTE, 2001).

Esses FIDs são então detectados e armazeados em intervalos de tempo regulares durante um período, gerando assim uma representação digital do FID. Esses valores são então processados usando a transformada de Fourier, gerando então espectros no domínio das frequências, além de informações de deslocamento químico e acoplamento escalar, conforme mostrado na Figura 3.15 Esses dados podem então ser processados, analisados e assinalados para a obtenção de informações de distâncias da estrutura tridimensional. A metologia do experimento de RMN é mostrado na Figura 3.16.

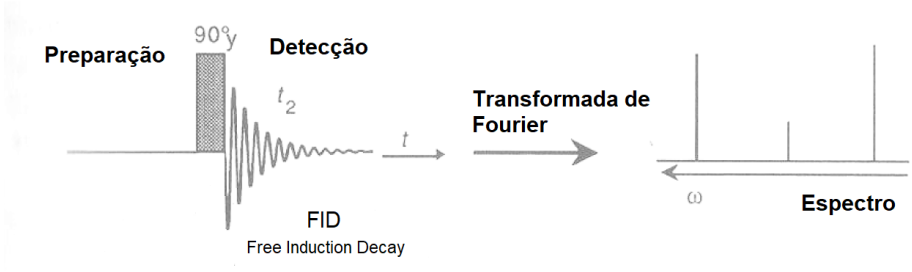


Figura 3.15: Experimento de RMN.

3.3.1.3 Assinalamento de Proteínas

A forma mais simples de assinalamento de proteína envolve ^{15}N,^{13}C marcado e a utilização dos espectros CBCANNH e CBCA(CO)NH.

As proteínas grandes fornecem espectros de RMN ruins, uma vez que eles ´caem mais lentamente´. Por isso, os espectros CBCANNH e CBCA(CO)NNH de proteínas maiores (mais de 150 resíduos), muitas vezes não são de qualidade suficiente para ser capaz de

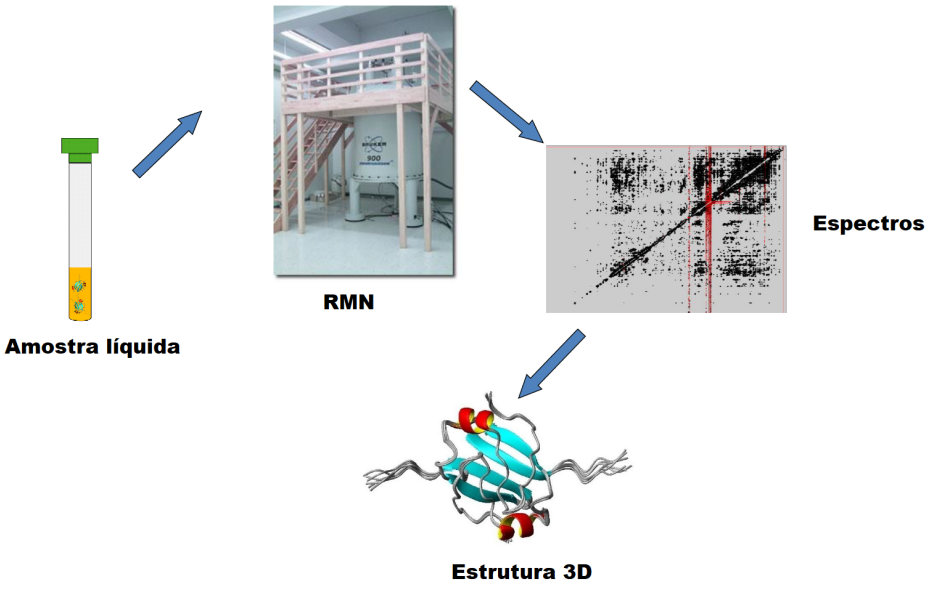


Figura 3.16: Metodologia da RMN.

realizar um trabalho completo de assinalamento. Neste caso, uma boa opção é a utilização dos espectros HNCA, HN(CO)CA, HNCO e HN(CA)CO.

Para grandes proteínas (geralmente mais de 250 resíduos) pode tornar-se necessário deuterar a proteína. Neste caso variantes do CBCANNH e CBCA(CO)NH são usados que são menos sensíveis, mas, devido às melhores propriedades de relaxamento da proteína o espectro pode melhorar.

Em princípio, é bom usar campos mais altos (750 MHz ou mais) com as proteínas maiores, uma vez que é possível utilizar técnicas Trosy e a maior resolução ajuda quando o grau de sobreposição é grande. No entanto, é interessante notar que o CBCA(CO)NNH ou HN(CO)CA são susceptíveis de diminuir em qualidade devido ao efeito de aumento da anisotropia carbonila deslocamento químico (*carbonyl chemical shift anisotropy*). Assim, mesmo para as proteínas de 200 ou mais aminoácidos é melhor gravar conjuntos de dados 3D de 600 MHz.

Experiências em 3D são geralmente baseadas em experimentos 2D e por isso a maneira mais fácil de pensar em um 3D é de um 2D que é então estendido em uma terceira dimensão.

Por exemplo, um HNCO se baseia numa HSQC 2D (Figura 3.17 a), de modo que os eixos x e y são 1H e 15N, respectivamente. o espectro é então estendido para uma terceira dimensão, que é uma dimensão 13C. Assim, os picos HSQC agora não apenas se encontram em um plano, mas eles vão se levantado para a terceira dimensão de acordo

com o valor 13C ppm do grupo CO precedendo o grupo NH (Figura 3.17 b e c).

Agora é possível olhar para o espectro 3D a partir de vários ângulos diferentes e cada vez visualizar um plano diferente. A dimensão 1H é geralmente deixada na dimensão x e na maioria dos casos, a dimensão de 13C é vista ao longo do eixo y, deixando a 15N para formar o plano z. Então é possível olhar um espectro 1H-13C (2D) em diferentes lugares ao longo da dimensão 15N (Figura 3.17 d e e)

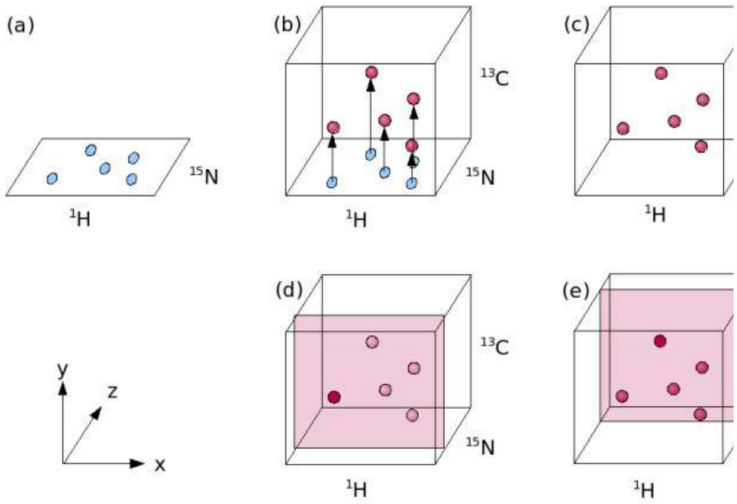


Figura 3.17: Espectro HNCO HSQC.

A maioria dos espectros de tripla ressonância utilizados para assinalamento do backbone tem um 1H, 15N e 13C em cada dimensão. Vários outros tipos de espectros, principalmente os 3D NOESY e HCCH-TOCSY/COSY tem duas dimensões 1H e uma dimensão 15N ou 13C. Neste caso, as duas dimensões 1H são vistas em x e y e o 15N ou 13C é deixado no plano z.

3.3.1.3.1 Assinalamento do backbone com tripla ressonância

Assinalamento do backbone com tripla ressonância é baseada nos espectros CBCANNH e CBCA(CO)NHH. A ideia é que o CBCANNH correlaciona cada grupo NH, com os deslocamentos químicos C_α e C_β do seu próprio resíduo (fortemente) e do resíduo anterior (fracamente). O CBCA(CO)NNH apenas correlaciona o grupo NH aos deslocamentos químicos C_α e C_β do resíduo anterior. As Figuras 3.18, 3.19 e 3.20 mostram como o espectro pode ser usado para ligar um grupo NH para o próximo em uma cadeia longa.

Os deslocamentos químicos de C_α e C_β adotam valores característicos de cada tipo de aminoácido. Alguns destes, como alanina, serina, treonina e glicina são fáceis de

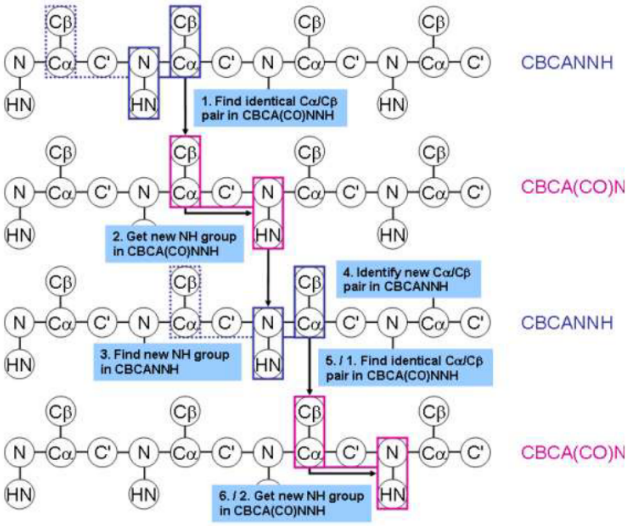


Figura 3.18: Assinalamento com CBCA(CO)NNH.

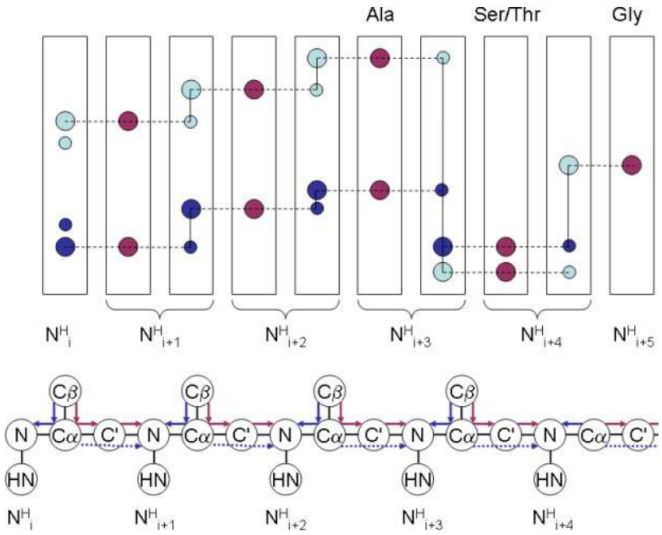


Figura 3.19: Assinalamento com CBCA(CO)NNH.

identificar como os seus deslocamentos químicos C_β são muito diferentes das dos outros aminoácidos (e no caso de glicina não há C_β). Valina, isoleucina e prolina também se destacam pelo fato de que eles têm deslocamentos químicos C_α mais baixos do que os normais.

Uma vez que uma cadeia de grupos NH, com os seus correspondentes deslocamentos químicos C_α e C_β foi construído, a identificação de alguns dos tipos de aminoácidos torna possível combinar esta cadeia com a sequência. Por exemplo, se foi encontrada a sequência xxxSxxAx e esta sequência aparece apenas uma vez na proteína em estudada, então a atribuição específica de sequência pode ser realizada.

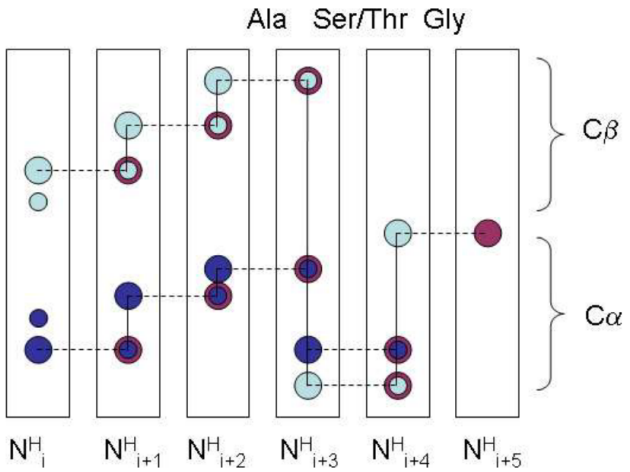


Figura 3.20: Assinalamento com CBCA(CO)NNH.

Em alguns casos se a sua proteína é grande (por exemplo mais de 200 resíduos), a qualidade dos espectros CBCANNH e CBCA(CO)NNH não é muito boa. As ressonâncias de C_β podem não ser visíveis acima do nível de um ruído. Neste caso, é possível utilizar os deslocamentos químicos C_α e C', em vez de os deslocamentos químicos C_α e C_β,para andar de um resíduo para o próximo.

Os espectros HNCA e HN(CO)CA fornecem a mesma informação que os espectros CBCANNH e CBCA(CO)NNH, só que sem as ressonâncias C_β. Para complementar é possível utilizar os experimentos HNCO e HN(CA)CO. Estes ligam cada grupo NH(i), com o C'(i-1) (HNCO), com ou C'(i) e C' (i-1)(HN(CA)CO). Os resíduos são agora ligados como mostrado na Figura 3.21.

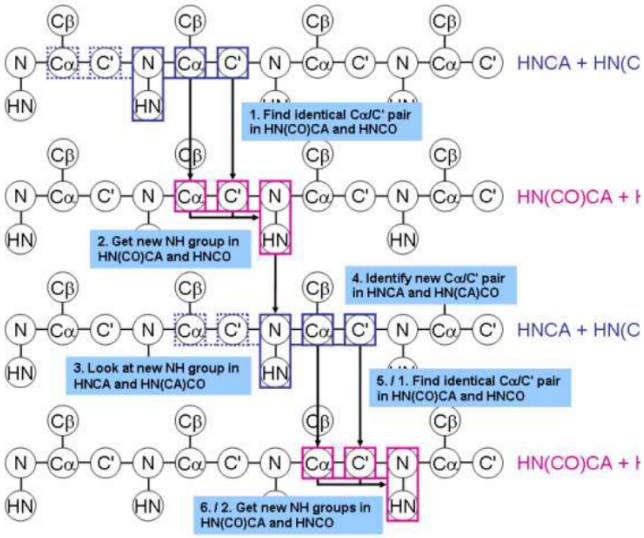


Figura 3.21: Assinalamento com HNCA e HNCOCA.

A vantagem de utilizar o espectro de HNCA e HNCO como base é que eles são mais sensíveis que o CBCANNH e, assim, a qualidade dos espectros é melhor. A desvantagem é que os deslocamentos quimicos de C_α e C' fornecem menos informações sobre o tipo de aminoácido e são menos dispersos.

3.3.1.3.2 Assinalamento Cadeia Lateral

Vários métodos e espectros estão disponíveis para o assinalamento da cadeia lateral, dependendo do tamanho da proteína e da quantidade de tempo disponível espectrômetro.

Um método fácil é começar com um conjunto de espectros HBHA(CO)ONH, HCC(CO)NNH e CC(CO)NNH. Estes experimentos fornecem os deslocamentos químicos do hidrogênio e do carbono da cadeia lateral do resíduo anterior a cada grupo NH.

Para cadeias laterais mais longas nem todos os picos podem ser necessariamente visíveis, de modo que isso pode não ser suficiente. Em alguns casos, poderá ser difícil distinguir entre um H_β e um H_δ. Além disso, a ligação de hidrogênio que está ligado ao carbono também não é fornecida. Isto é relevante no caso de Valinas onde há dois grupos metilo: pode ser possível identificar os deslocamentos químicos de ambos carbonos metilo e hidrogênio metilo, mas não vai ser conhecida quem é ligado a quem, como pode ser visualizado na Figura 3.22.

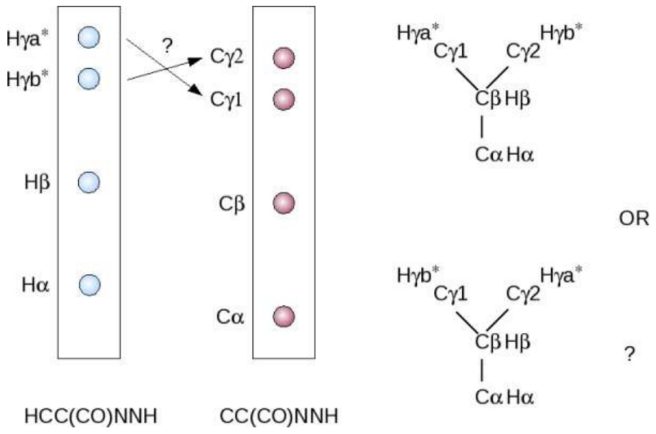


Figura 3.22: Dificuldade em distinguir entre um H_β e um H_δ.

O espectro mais útil para atribuição de cadeia lateral é o HCCH-TOCSY juntamente com o espectro HCCH-COSY. O HCCH-TOCSY irá para cada posição de carbono mostrar em uma dimensão o deslocamento químico do hidrogênio que está ligado ao carbono e na outra dimensão os hidrogênios que pertencem a cadeia lateral. Há uma enorme

quantidade de informações contidas neste espectro, para grandes proteínas pode tornar-se bastante cheia.

O HCCH-COSY é parecido, mas que em vez de ver todos os hidrogênios pertencentes a uma determinada cadeia lateral na terceira dimensão, só é possível ver aqueles que estão ligados a átomos de carbono vizinhos. A Figura 3.23 mostra as faixas que são mostradas para um resíduo Valina nos espectros HCCH-TOCSY e HCCH-COSY.

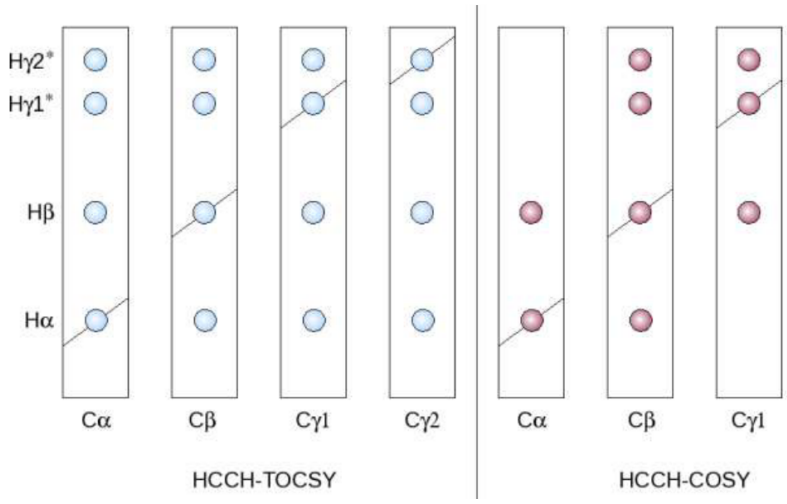


Figura 3.23: Assinalamento cadeia lateral HCCH TOCSY COSY.

O princípio geral por trás do uso do espectro HCCH-TOCSY é a seguinte: Usando os deslocamentos químicos de C_α e C_β conhecidos a partir da assinalamento de backbone. Partindo desses pontos no espectro, é possível obter imediatamente os deslocamentos químicos de $H_\alpha e\ H_\beta$ encontrando faixas em cada deslocamento de carbono que têm picos nos mesmos valores ppm de hidrogênio. É possível ver mais picos para os átomos H_γ e H_δ (se presente nesse tipo de aminoácido particular). Ao navegar por estes deslocamento de hidrogênio é possível identificar os deslocamentos dos carbonos que estão ligados aos hidrogênios como mostrado na Figura 3.24.

3.4 Validação Bioquímica - Gráfico de Ramachandran

O gráfico de Ramachandran descreve os ângulos de torção $\phi - \psi$ do backbone da proteína. Desta forma, fornece uma visão geral da conformação de uma proteína. Os ângulos $\phi - \psi$ se agrupam em regiões distintas no gráfico de Ramachandran, onde cada região corresponde a uma estrutura secundária específica.

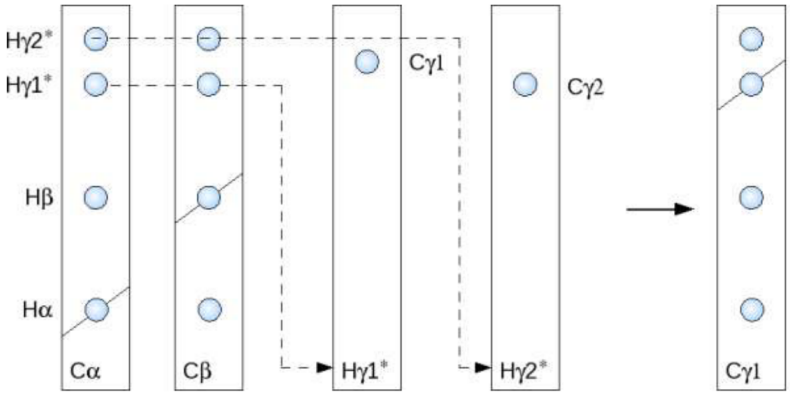


Figura 3.24: Assinalamento com HCCH TOCSY COSY.

Os ângulos de torção ϕ (phi) e ψ (psi) determinam a conformação do backbone da proteína, 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).

Em princípio, os ângulos diédricos ϕ e ψ nos aminoácidos podem ter qualquer valor entre $+180$ e -180, mas 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; COX, 2017). Um modelo de gráfico de Ramachandran onde há linhas de contorno destacadas em verde para as áreas permitidas aos ângulos diedrais, e em branco as áreas que causam colisão é mostrado na Figura 3.25.

3.5 Ferramentas de Reconstrução Estrutural

Determinar a estrutura de compostos orgânicos é um problema que praticamente surgiu junto com a química orgânica, porém os processos de elucidação eram baseados em experimentos muito simples e com grandes margens de erro. Entretanto as análises instrumentais evoluíram consideravelmente após a primeira metade do século XX. Estes avanços são devidos à consolidação do métodos espectrométricos, sendo estes a ressonância magnética nuclear, a espectrometria no ultravioleta, no infravermelho e de massas.

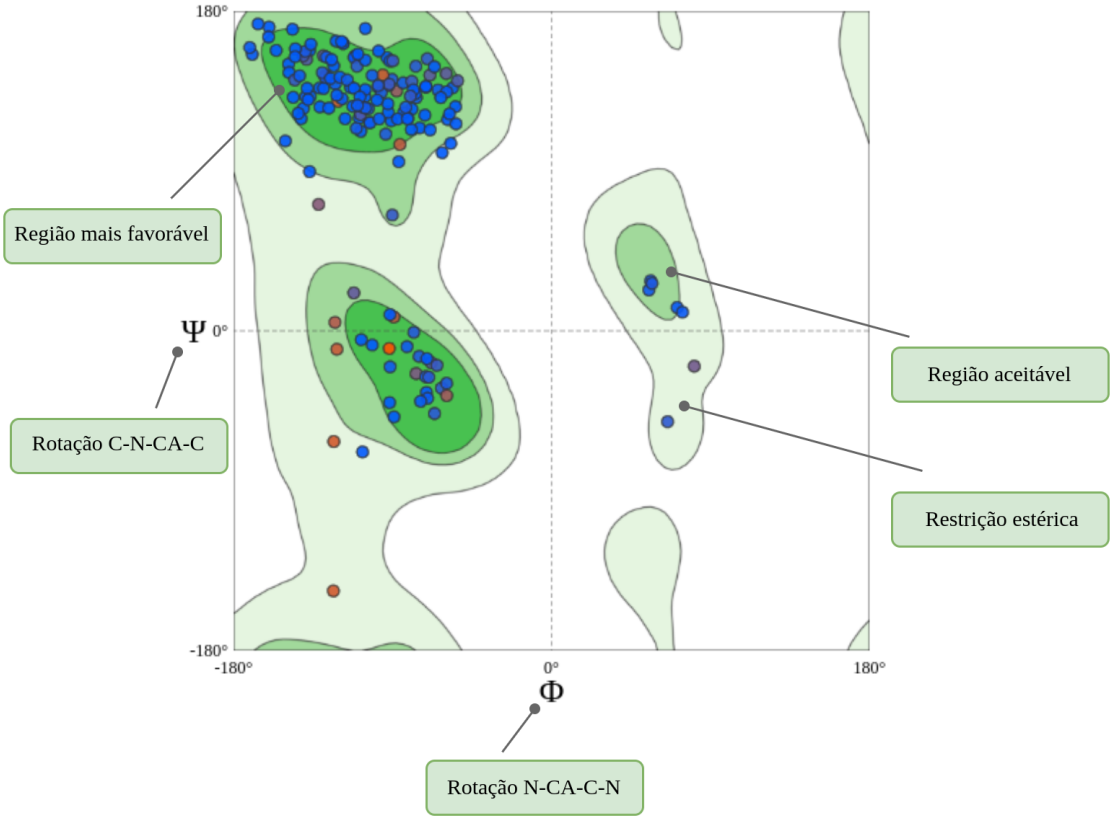


Figura 3.25: Regiões do gráfico de Ramachandran tendo como exemplo a validação para a estrutura ORF8 (PDB ID: 7JX6) obtida com o módulo “Structure Assessment” na plataforma MolProbity (WILLIAMS et al., 2018).

Com a crescente geração de dados espectrométricos, as coletâneas destas informações já não eram eficientes dentro do ambiente de pesquisa, tanto pelo volume, tanto pela dificuldade de achar os dados de um composto específico. Contudo, ferramentas computacionais se tornaram uma opção, dado o desenvolvimento expressivo da informática. O conjunto de técnicas que visam utilizar a computação para automatizar a determinação estrutural é chamado de "elucidação estrutural auxiliada por computador"do inglês CASE("Computer-Assisted Structure Elucidation").

Pioneiro na determinação estrutural, o programa DENDRAL compara os dados do espectro de entrada com pequenos fragmentos de estruturas de poucos átomos com seus respectivos deslocamentos químicos, gerando uma lista de fragmentos compatíveis com o espectro, deixando o usuário escolher as estruturas desejadas para poder então combiná-las em um grupo de possíveis estruturas completas.

Vale destacar o SISTEMAT, uma ferramenta pioneira Brasileira e um dos primeiros sistemas que utilizam IA desenvolvido para desempenhar funções além da elucidação estrutural. O SISTEMAT é um sistema modular com vários programas que vão de banco

de dados espectrométrico até o SISCONT que trabalha com dados de Ressonância Magnética Nuclear e procura fragmentos compatíveis para o usuário propor as estruturas.

Os sistemas de predição estrutural utilizam diversas estratégias algorítmicas na etapa de geração das estruturas, dentre elas IAs heurísticas como o DENDRAL,SISTEMAT, DARC/EPIOS, ACD/StrucEluc, I-Tasser HOUDINI e SESAMI; Algoritmos genéticos como GENIUS e SENECA, que também emprega o método de Faulon, além de Redes neurais como o PSIPRED.

Referente a validação das estruturas geradas destacam-se programas especialistas como ChemMNR que utiliza regras de adição; IA de Rede Neural e Procura em banco de dados aplicando código HOSE.

Capítulo 4

Estudo de Caso sobre SARS-CoV-2

Em dezembro de 2019, um novo coronavírus causou um surto de doença pulmonar na cidade de Wuhan, capital da província de Hu-bei, na China, e desde então se espalhou pelo mundo.

O vírus foi chamado de SARS-CoV-2, devido o genoma do RNA ser cerca de 82% idêntico ao coronavírus SARS (SARS-CoV). A doença causada pela SARS-CoV-2 é chamada de COVID-19. Em 11 de março, a Organização Mundial da Saúde (OMS) declarou o surto uma pandemia. Em 15 de março, existem mais de 170.000 casos em todo o mundo, com uma taxa de mortalidade de casos de cerca de 3,7% (ZHANG et al., 2020).

Segundo o Ministério da saúde os primeiros coronavírus humanos foram isolados pela primeira vez em 1937. Mas somente em 1965 que o vírus foi chamado coronavírus devido seu formato ser parecido com uma coroa. Muitas pessoas se infectam com os coronavírus comuns ao longo da vida, sendo as crianças pequenas mais propensas a se infectarem com o tipo mais comum do vírus. Os coronavírus mais comuns que infectam humanos são o alpha coronavírus $229E$ e $NL634$ e beta coronavírus $OC43, HKU1$ (MS, 2020).

Em desembro de 2019 os primeiros casos do novo SARS-CoV-2 foram detectados na cidade de Wuhan, na China e em pouco tempo a doença se espalhou pela China e pelo mundo. No Brasil já foram mapeados mais de 427 genomas do vírus, a partir do sequenciamento foi possível determinar como se deu a evolução e a disseminação de três grandes linhagens do novo coronavírus em território brasileiro.

Por exemplo, 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, com os primeiros infectados identificados vindo dos Estados Unidos e Europa. En-

quanto aqui no Brasil, o novo coronavírus teria sofrido duas mutações sequenciais, as quais formaram a cepa mais atuante no país, que consiste na sub-linhagem brasileira B.1.1 (B1.1.BR). Esta é considerada a cepa responsável pela maior parte da transmissão comunitária do vírus no país, pois foi a única encontrada em 18 infectados que não tinham feito viagem internacional recentemente ao contágio, o que propiciou tal conclusão.

Além desta cepa brasileira dominante, B.1.1, há cepas diferentes com pequenas variações, com outras cinco em destaque, sendo: A.2, B.1, B.2.1, B.2.2 e B.6. Isso corrobora com a afirmação de que a mutação do vírus SARS-CoV-2 está sendo relativamente lenta, de forma que ainda não se detectou uma linhagem diferente que poderia resultar em formas diferentes (mais graves ou não) da doença.

Nesse contexto, no estado do amazonas, um grupo de pesquisadores da Fundação Oswaldo Cruz (Fiocruz-Amazonas) identificou e obteve a primeira sequência do genoma do vírus SARS-CoV-2 de um paciente infectado na região Norte do Brasil. A análise da sequência de nucleotídeos (dentre um total de 29 789) revelou um total de nove mutações em comparação com o primeiro caso em Wuhan, China. Mutações não-sinônimas no receptor nsp4 com a variante Phe3071Tyr foram então observadas. Assim como alterações em OFR3a com a mutação Gly196Val. As mutações também surgiram no ORF8 com a variante Leu84Ser (NASCIMENTO et al., 2020).

Nas próximas seções são mostrados definições de vírus, a origem do SARS-CoV-2 e suas variações.

4.1 Aspectos Bioquímicos do Vírus

Vírus são bioestruturas moleculares muito menores que células e que ainda é incerto a classificação como organismos vivos. Ainda que haja discussão acerca de seu estado vital, o fato é que os vírus não possuem hialoplasma e ribossomos, e não apresentam o maquinário metabólico e bioquímico mínimo necessário à produção de sua própria energia metabólica, bem como reprodução autônoma. Nesse sentido, os vírus não podem captar nutrientes, utilizar energia ou realizar qualquer atividade biossintética.

Vale esclarecer que proteína biossintética refere-se a toda proteína que é sintetizada em uma estrutura celular, e nesse sentido os vírus são incapazes de produzir rotas biossintéticas de proteínas em sua estrutura (GRAMMBITTER et al., 2019). Sua dependência

de estruturas celulares complexas para sua multiplicação e metabolismo impede que seja considerado como um ser vivo, não sendo capaz de crescer em tamanho e de se dividir. Fora de uma célula hospedeira o vírus é basicamente uma partícula proteica inerte, um vírion (termo usado para se referir a uma única partícula viral que não se encontra dentro de uma célula hospedeira) com informação codificada capaz de infectar células vivas.

Em termos biológicos estruturais os vírus são relativamente simples quando comparados a uma célula qualquer e são formados basicamente por uma cápsula proteica envolvendo o material genético que pode ser DNA, RNA ou ambos no caso de um citomegalovírus.

As proteínas que formam um vírus são chamadas de proteínas virais, que por definição é qualquer componente proteico que faz parte de um vírus e desempenha uma função específica em sua constituição. Sua classificação depende da função e pode ser do tipo: estrutural, não estrutural, regulatória e acessória.

Os vírus em geral são estruturas muito diminutas (a maioria com 20-300 nm de diâmetro) podendo ser vistos unicamente utilizando-se um microscópio eletrônico, não sendo possível visualizá-los com microscopia óptica por conta do comprimento de onda da luz.

Uma cepa viral é reconhecida quando possui propriedades fenotípicas únicas. Quando uma cepa apresenta antigenicidade única e estável denomina-se então de sorotipo. Portanto, estes constituem linhagens estáveis que se diferenciam entre si ao longo do tempo. Nota-se que sorotipos de um vírus podem apenas ser neutralizados pelos seus anticorpos específicos e não por anticorpos correspondentes a outros serotipos, embora também seja possível a neutralização-cruzada (MAHY; REGENMORTE, 2010). Existe um grande debate se a dinâmica de mutação viral é regida por fatores estocásticos (neutros) ou determinísticos. Caso os processos estocásticos prevaleçam sobre a evolução viral, então haverá influência de interações epistáticas e uma aleatoriedade inerente à resistência a medicamentos e vacinas (HOLMES; HARVEY; MAY, 2009).

O SARS-CoV-2 pertence ao grupo dos coronavírus, da família Coronavidae da ordem Nidovirales. Os coronavírus recebem este nome graças à camada proteica semelhante a uma coroa que reveste o vírus (GONZALEZ et al., 2003). Infecções decorrentes desse grupo são comuns em animais e determinadas mutações foram capazes de contaminar humanos, algumas das variantes são apenas resfriados comuns, mas espécies como o SARS-CoV-2 podem causar problemas graves de saúde (DOLHNIKOFF et al., 2020).

A proteína Spike (S) reveste todos os coronavírus e é responsável por fazer a ligação do vírus com as células hospedeiras (LI, 2016). Esta proteína possui uma forma específica que dá origem ao nome da família, estruturas parecidas com espinhos, dispostas ao redor da parte externa do vírus, foram associadas com as pontas de uma coroa. A Figura 4.1 mostra uma elucidação da estrutura do coronavírus como uma coroa.

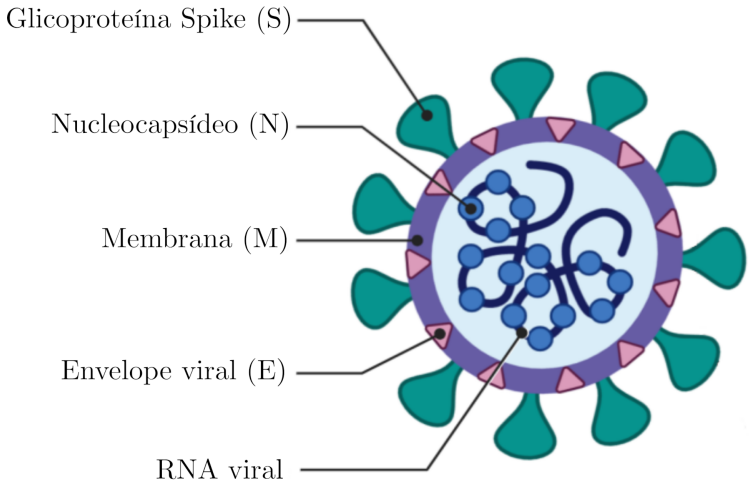


Figura 4.1: Estrutura geral do novo coronavírus.

4.2 Coronavírus Humano

Dentro da subfamília Orthocoronavidae existem 4 gêneros sendo estes: *Alphacoronavirus, Betacoronavirus, Gammacoronavirus* e *Deltacoronavirus*. Pertencentes a estes gêneros, são 7 classes de Coronavírus capazes de infectar humanos, todas de origem evolutiva animal (HELMY et al., 2020; HU et al., 2015; BRANDÃO, 2018; WOO et al., 2009). A Tabela 4.1 mostra uma classificação taxonômica e histórica.

4.2.1 SARS-CoV

O SARS-CoV foi o primeiro coronavírus a manifestar sintomas graves em humanos. A doença causada por infecção de SARS-CoV é chamada de síndrome respiratória aguda grave (SARS), causando febre, dores de cabeça, calafrios, dores musculares, garganta seca e diarreia (DROSTEN et al., 2003).

Os primeiros casos desta síndrome acorreram na china em 2002 que rapidamente se espalhou por 12 países da América, Europa e Ásia, contaminando mais de 8000 pessoas e ocasionando cerca de 800 óbitos.

Tabela 4.1: Origem e classificação dos 7 (sete) Coronavírus que afetam seres humanos. As imagens foram obtidas na plataforma BioRender.

Gênero	Espécie	Origem Evolutiva	Ano de descoberta	Doença Causada
Alphacoronavirus	HCoV-229E		1960	Resfriado
Betacoronavirus	SARS-CoV		2002	SARS
Betacoronavirus	HCoV-OC43		2004	Resfriado
Alphacoronavirus	HCoV-NL63		2004	Resfriado
Betacoronavirus	HCoV-HKU1		2005	Resfriado
Betacoronavirus	MERS-CoV		2012	MERS
Betacoronavirus	SARS-CoV-2	Desconhecido	2019	COVID-19

A glicoproteína Spike do coronavírus SARS (SARS-CoV) (mostrada na Figura 4.3) é responsável pela interação entre o vírus e seu receptor celular, a enzima conversora da angiotensina 2 (ACE2), intermediado pelo domínio de ligação ao receptor (RBD).

Os detalhes atômicos dessas estruturas contribuíram no desenvolvimento de tratamentos para a SARS, esclarecendo o processo de infecção e servindo de base para inibidores virais e variantes do RBD (LI et al., 2005; ZIEBUHR, 2005; GALLAGHER; BUCHMEIER, 2001; BOSCH et al., 2004).

4.2.2 MERS-CoV

O MERS-CoV foi o segundo coronavírus a causar sintomas graves em humanos. A doença causada por sua infecção é a síndrome respiratória do Oriente Médio (MERS), originária de um surto na Arábia Saudita em 2012 que se espalhou pelo Oriente médio e em alguns países da Europa e África.

Em outro surto de MERS na Coreia do Sul em 2015, cerca de 200 pessoas foram infectadas e 36 morreram, incidente associado a um viajante que havia retornado do oriente médio. Os sintomas mais comuns de MERS são febre, tosse, diarreia e falta de ar, podendo ocasionar sintomas mais graves, geralmente em pessoas que possuem comorbidades (GROOT et al., 2013; ZUMLA; HUI; PERLMAN, 2015; BADAWI; RYOO, 2016).

O vírus pode infectar humanos, primatas, porcos e morcegos (ZIELECKI et al., 2013). No corpo humano, o MERS-CoV tem com alvo células alveolares do tipo II e células epiteliais não ciliadas, sendo seu receptor celular a dipeptidil peptidase 4 (DDP4) (veja

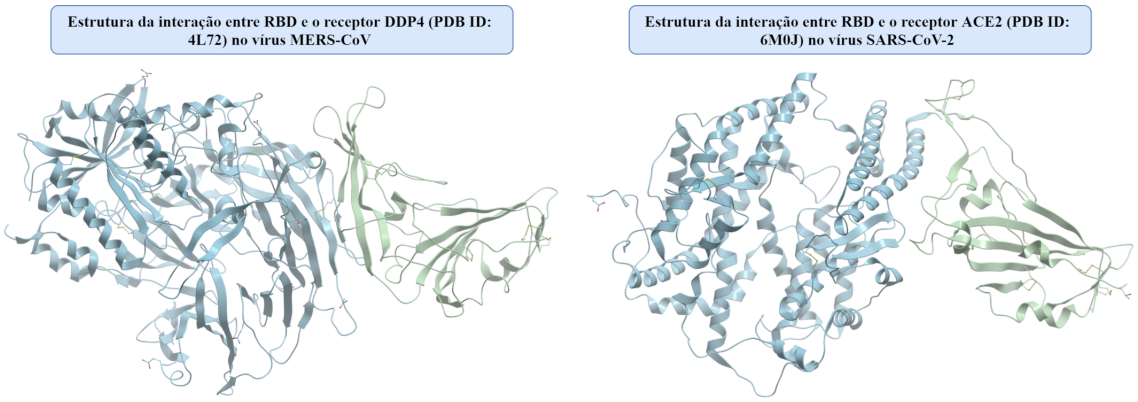


Figura 4.2: Comparação entre as estruturas cristalográficas do vírus MERS-CoV onde foi apresentado o complexo RBD com o receptor celular DDP4 (PDB ID: 4L72) (WANG et al., 2013), enquanto ao lado direito da figura encontra-se o complexo ACE2-RBD (PDB ID: 6M0J) (WANG et al., 2020) referente ao vírus SARS-CoV-2. A visualização foi por intermédio da versão gratuita do software ICM-MolBrowser 3.8.7.

Figura 4.2), diferente do SARS-CoV, que tem com alvo células células epiteliais ciliadas, que expressam ACE2 (HUI et al., 2019; RAJ et al., 2013).

4.2.3 SARS-CoV-2

O SARS-CoV-2 é o coronavírus mais notável, uma vez associado à atual pandemia. COVID-19 é a doença causada pela sua infecção, os sintomas mais comuns são tosse seca e febre, porém em casos mais graves os sintomas podem variar de perda do olfato até erupções cutâneas.

Cerca de 80% dos casos são assintomáticos ou leves e 20% são infecções graves e gravíssimas, podendo evoluir para sepse, falência dos órgãos, pneumonia grave com insuficiência respiratória e morte (GORBALENYA et al., 2020; CDC, 2020).

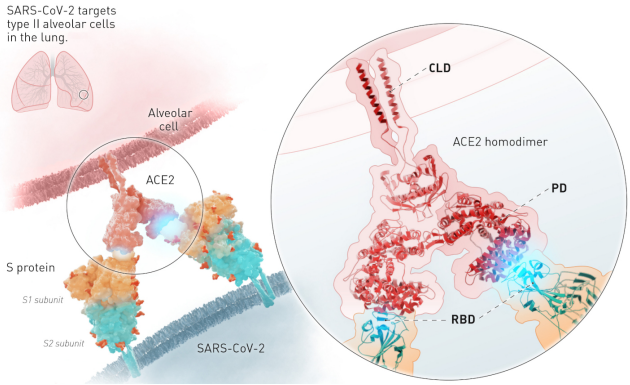


Figura 4.3: Mecanismo de reconhecimento molecular entre a glicoproteína Spike do vírus SARS-CoV-2 e os receptores ACE2 nos alvéolos pulmonares (COGNITION, 2020).

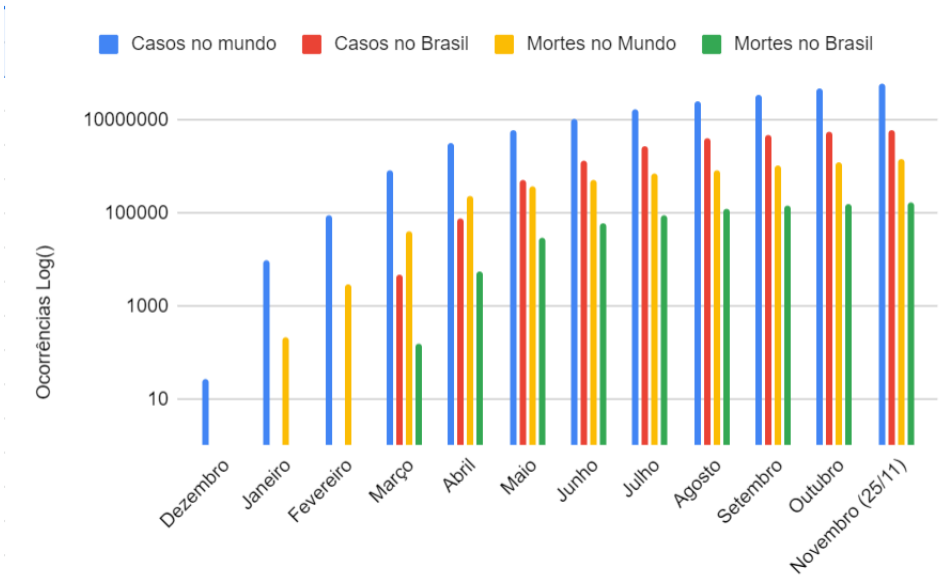


Figura 4.4: Gráfico da quantidade de ocorrências de infecção pela COVID-19 no Brasil e no mundo.

Os primeiros casos da COVID-19 foram descobertos em dezembro de 2019 na cidade de Wuhan, na China. O SARS-CoV-2 é transmitido pelo ar, através de gotículas expelidas pelo trato respiratório humano, tornando-o em um vírus de fácil contágio (WHO, 2020b). No dia 26 de fevereiro, o Brasil teve o primeiro caso confirmado e em 11 de março de 2020, a Organização Mundial da Saúde (OMS) decretou a pandemia da COVID-19.

Atualmente não há tratamento para infecção por SARS-CoV-2, os cuidados com pacientes mais graves consistem em tratar apenas os sintomas e complicações da doença. Assim como o SARS-CoV, o SARS-CoV-2 interage com a célula receptora por intermédio da glicoproteína Spike e a proteína ACE-2, no qual estudos dessas moléculas e os seus mecanismos de ativação estão contribuindo para o desenvolvimento de inibidores virais e vacinas, destacando a importância da análise molecular dessas estruturas (XU et al., 2020; PRIGENT et al., 2010).

4.3 Linhagens Emergentes do Vírus SARS-CoV-2

Apesar do número relativamente grande de linhagens emergentes, estima-se que a taxa de mutação ao vírus SARS-CoV-2 é de $1,1 \times 10^{-3}$ mutações/sítio a cada ano, que é portanto considerada moderada para um vírus. Possivelmente este resultado decorre da Nsp14 que possui um mecanismo central de correção/revisão das mutações que eventualmente ocorrem (DUCHENE et al., 2020).

O Reino Unido relatou uma nova variante em outubro de 2020 mas que se espalhou apenas na metade de dezembro. Inicialmente as preocupações surgiram no condado de Kent localizado ao sudeste da Inglaterra, que apresentou um inesperado aumento na transmissão do SARS-CoV-2, ainda que o país estivesse contendo a disseminação (KUPFERSCHMIDT, 2021).

Denomina-se VUI 202012/01 definida por inúmeras mutações não-sinônimas na glicoproteína Spike constituindo a nova linhagem B.1.1.7. As principais substituições de aminoácidos reportadas foram D614G, A222V, N439K, Y453F e N501Y enquanto as deleções foram nas posições 69-70 e 144-145 cujos resultados foram obtidos com mais de 24 746 amostras do vírus na região britânica.

Dentre as alterações de aminoácidos acredita-se que a mais relevante poderia ser N501Y pois influenciaria na interação com alguns anticorpos neutralizantes enquanto a variante A222V teria um efeito estimado entre 40% e 70% no aumento da transmissibilidade (WHO, 2020a). Inesperadamente também verificou-se que a deleção na posição 69/70 afetou a precisão de alguns diagnósticos de RT-PCR (WHO, 2020a).

Em estudos de laboratório, a variante E484K tornou o vírus menos suscetível ao plasma convalescente (KUPFERSCHMIDT, 2021). Resultados preliminares mostraram que as variantes britânicas não alteraram de forma significativa a imunogenicidade perante a vacina desenvolvida pela Pfizer e BioNtech (XUPING et al., 2021). Nota-se que mutações são esperadas para o SARS-CoV-2 sendo a ampla maioria apenas marcadores regionais, embora estudos mais profundos ainda são necessários (COG-UK, 2020).

Ainda há dúvidas se as vacinas precisarão de administração periódica para manter a imunogenicidade em consequência de mutações. Apesar de tudo, temos como exemplo os vírus do sarampo e da poliomielite que até hoje não sofreram mutações a ponto de anular a eficácia de vacinas (KUPFERSCHMIDT, 2021).

Recentemente descobriu-se no sequenciamento genético de quatro viajantes do estado do Amazonas, Brasil em sua ida ao Japão onde apresentaram uma nova linhagem B.1.1.28. Este clado é constituído pelo conjunto de 3 (três) mutações na glicoproteína Spike: K417N, E484K e N501Y. Possivelmente o clado evoluiu de uma linhagem viral que já circulava no estado do Amazonas desde abril de 2020 mas que surgiu provavelmente apenas ao final de 2020.

Estas mutações podem ser resultado de uma pressão evolutiva em consequência de

milhões de pessoas terem sido infectadas ao redor do mundo (NAVECA et al., 2021). Embora tenham havido um número relativamente grande de mudanças no genoma no SARS-CoV-2 acarretando em diferentes linhagens e clados, o vírus conserva a sequência inicial surgida na província de Wuhan (NAVECA et al., 2021).

A linhagem P.2 foi detectada em várias regiões do Brasil, incluindo a cidade de Manaus, lugar onde foi constatado que dois pacientes foram reinfectados com SARS-CoV-2 (RESENDE et al., 2021). Além disso, há evidências *in vitro* de que a presença da mutação E484K reduz a eficácia de anticorpos policlonais derivados de soro convalescente (GREANEY et al., 2021).

4.4 Vacinas Administradas Contra a COVID-19

A tecnologia de vacinas evoluiu significativamente nas últimas décadas, incluindo a abordagem com mRNA e de proteína recombinante. Mediante o auxílio da genômica, proteômica e imunologia molecular, tem-se um aumento impressionante no conhecimento dos vírus.

A variação antigênica em consequência do surgimento de novas cepas ainda representa um grande desafio na imunogenicidade. Um importante exemplo é o vírus influenza, cuja surgimento de inúmeras variantes a cada ano torna então necessário uma estratégia de vacinação anual (WEISSMAN et al., 2020).

O sequenciamento genético ao SARS-CoV-2 foi rapidamente disponibilizado por pesquisadores chineses, sendo o ponto de partida para a construção de vacinas. A proteína S na superfície viral é o alvo principal de todas as vacinas, pois interage com o receptor ACE2, e sendo assim anticorpos direcionados à Spike poderiam interferir diretamente nesta ligação, neutralizando o vírus.

Atualmente estima-se haver mais de 192 vacinas em diferentes fases de pesquisa e ensaios clínicos, embora apenas 15 já foram aprovadas pelas agências regulamentadoras ou estão na Fase III (BIOWORLD, 2020). A Figura 4.5 mostra um diagrama resumindo algumas vacinas já aprovadas em estudos clínicos de Fase III, todas as taxas de eficácia global utilizadas na figura correspondem a dados atualizados conforme o dia 05 de Fevereiro de 2020 (SOARES; HARTMAN, 2020).

A vacina baseada em mRNA, expressa o antígeno alvo no paciente cujo transporte é

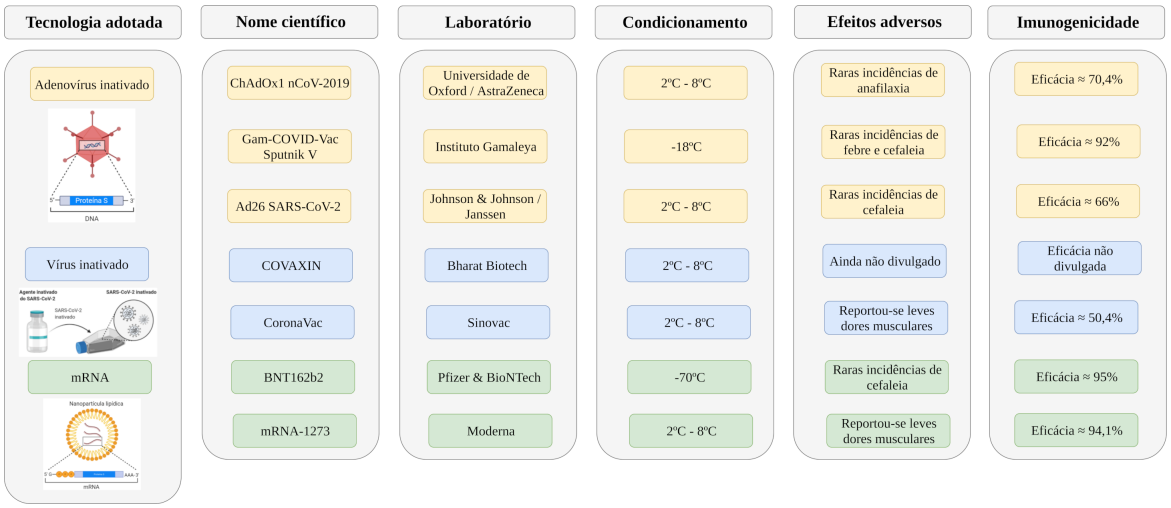


Figura 4.5: Diagrama resumindo algumas vacinas já aprovadas em estudos clínicos de Fase III. Os diagramas estruturais das vacinas foram obtidos na plataforma BioRender.

encapsulado em nanopartículas de lipídios, tendo sendo co-desenvolvidas pela Moderna e NIH. Existem também as vacinas baseadas em vetor viral tal como a desenvolvida pela Universidade de Oxford em parceria com a AstraZeneca. A indústria Johnson & Johnson vem desenvolvendo vacinas utilizando-se um vetor de adenovírus (AMANAT; KRAMMER, 2020). Dentre todas as vacinas apresentadas apenas a desenvolvida pelo laboratório Johnson & Johnson e Janssen pode ser administrada em uma única dose (AMANAT; KRAMMER, 2020).

O desenvolvimento clínico de vacinas começa com ensaios de fase I para avaliar a segurança. Posteriormente são seguidos por ensaios de fase II, onde é formulado as doses que apresentem eficácia e finalmente são realizados ensaios de fase III, nos quais a eficácia e a segurança de uma vacina devem ser demonstradas simultaneamente.

As etapas de vacinação são normalmente em 2 (duas) fases com um intervalo de 3-4 semanas após a primeira dose. Consequentemente a imunogenicidade possivelmente será alcançada apenas 1-2 semanas após a segunda dose (AMANAT; KRAMMER, 2020).

Mesmo para vacinas contra o vírus influenza da gripe, para as quais existem vasta infraestrutura de produção, a demanda em uma pandemia ultrapassa a capacidade de produção, sendo então um grande desafio a todos os governos.

Apesar de tudo as variantes no genoma viral, até agora, não ocasionaram em cepas aparentemente resistentes às vacinas em produção. Inesperadamente um estudo realizado por um grupo de cientistas descobriu que as cepas G poderiam até mesmo facilitar a interação entre vacina e as proteínas Spike, e consequentemente induzindo a produção de

mais anticorpos.

Ainda sim, é fundamental restringir sua disseminação para assim reduzir a probabilidade de mutação, em particular, na proteína de pico (WEISSMAN et al., 2020). Em um estudo recente, as linhagens surgidas no Reino Unido (B.1.1.7) e na África do Sul (B.1.351) não afetaram de forma crítica a imunogenicidade na neutralização viral induzida por duas doses da vacina mRNA BNT162b2 (Pfizer & BioNTech) (XUPING et al., 2021).

4.5 Trabalhos Relacionados

Dentre os pesquisadores que estão buscando criar o medicamento para tratar os doentes está um grupo da Universidade de Lübeck, na Alemanha, que publicou um artigo na Science. Nesse artigo é apresentada a estrutura 3D da principal protease do vírus SARS-CoV-2, usando um aparelho de raio-X conhecido como Bessy II. Eles mostram que essa proteína está envolvida diretamente na reprodução do coronavírus e que uma análise de sua arquitetura 3D permitiria o desenvolvimento de medicamentos que inibiriam a multiplicação do microrganismo.

Zhang (ZHANG et al., 2020) mostra o alvo principal dos medicamentos que é a principal protease (M^{pro}, também chamada de $3CL^{pro}$), devido ao seu papel essencial no processamento das poliproteínas que são traduzidas a partir do RNA viral, inibir a atividade dessa enzima bloquearia a replicação viral.

No artigo é mostrada a estrutura SARS-CoV-2 M^{pro} e o inibidor de $\alpha-$cetoamida, que foi derivado de um inibidor previamente projetado, mas com a ligação amida $P3-P2$ incorporada em um anel de piridona para aumentar a meia-vida do composto no plasma. Com base na estrutura, Figura 4.6 , desenvolveram um composto de chumbo como inibidor do SARS-CoV-2 M^{pro}, Figura 4.7 (ZHANG et al., 2020).

Hilgenfeld (HILGENFELD, 2014) faz uma revisão mostrando as principais contribuições que a cristalografia macromolecular fez nos últimos anos para exemplificar as estruturas e mecanismos das proteases essenciais dos coronavírus, sendo as principais a protease (M^{pro}) e a protease (PL^{pro}).

Nessa revisão é explicado o surgimento do coronavírus *SARS* em 2002-2003 e do coronavírus *MERS* 10 anos depois e as origens desses vírus. É mostrado também a estru-

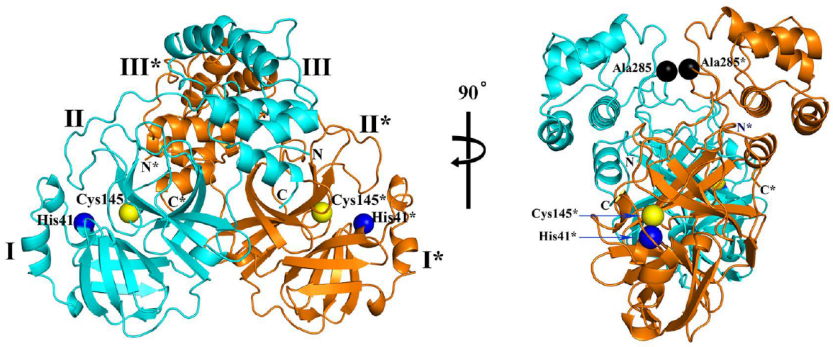


Figura 4.6: Estrutura tridimensional do SARS-CoV-2 M^{pro}, em duas visões diferentes (ZHANG et al., 2020).

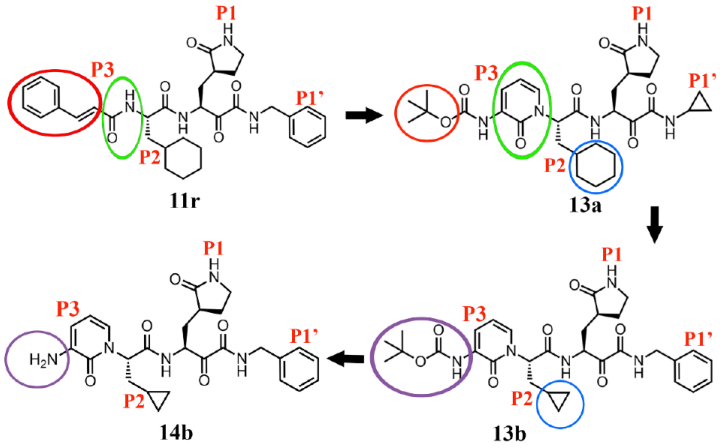


Figura 4.7: Estruturas químicas dos inibidores de α−cetoamida 11r, 13a, 13b e 14b. Os círculos coloridos destacam as modificações de uma etapa de desenvolvimento para a próxima(ZHANG et al., 2020).

tura cristalina do coronavírus SARS M^{pro} e sua dependência do pH , além dos esforços para projetar inibidores com base nessas estruturas. A estrutura cristalina do coronavírus SARS PL^{pro} e seu complexo também é discutido com a ubiquitina, assim como seu ortólogo do coronavírus *MERS*.

Capítulo 5

Resultados

Neste capítulo são apresentados os trabahos realizados durante as duas fases de pesquisa do mestrado. A primeira onde foram realizadas as implementações dos principais algoritmos para os tipos do Problema de Geometria de Distâncias Moleculares, feita validação e análise comparativa entre os algoritmos e uma visita técnica ao Centro Nacional de Ressonância Magnética Nuclear da UFRJ, onde foi possível analisar o método de obtenção de dados através da Ressonância Magnética Nuclear e realizar o assinalamento de uma proteína.

Ainda durante a *Fase I* foi proposta uma modificação do método *Branch-and-Prune* (BP) original para as duas formas de abordagens, considerando a interseção de quatro esferas. Os experimentos nesta fase foram realizados utilizando dados de proteínas já conhecidas, retiradas do banco *Protein Data Bank* (PDB) (BERMAN et al., 2000) e um fragmento dA1rC2rG3dC4T5rC6rG7dA8rG9 gerada pelo professor Kelson Mota (OLIVEIRA et al., 2021; COSTA et al., 2021) da UFAM com o programa Hyperchem.

Na *Fase II* surgiu a seguinte dúvida: todas as estruturas geradas pelos algoritmos utilizados na primeira fase são quimicamente válidas? A construção das estruturas pelos algoritmos garantia que toda a estrutura 3D gerada seria matematicamente válida, mas não garantia a validade química do resultado, então procuramos uma maneira de validar as estruturas, validando cada uma delas após sua geração, além disso também fizemos um estudo de caso com algumas variações do SARS-CoV-2 encontradas principalmente no Amazonas.

5.1 FASE I - Imersão e Manipulação do Problema

Os algoritmos para o problema com **conjunto arbitrário** podem ser utilizados na resolução do problema para **conjunto completo**, como todas as distâncias são fornecidas o algoritmo usaria somente uma parte das distância conhecidas. Mas o contrário não é válido já que o algoritmo para conjunto completo poderia tentar utilizar uma distância que não existe.

Para a classe especial do problema de geometria de distâncias que ocorre quando as distâncias entre todos os pares de átomos são conhecidas foi implementado o algoritmo de tempo polinomial proposto por Dong e Wu que usa a técnica de distâncias interatômicas, resolvendo o problema através de decomposição de uma matriz de distâncias formada pelas distâncias conhecidas (DONG; WU, 2002).

Para versão contínua do problema com conjunto incompleto de distâncias, o algoritmo *Geometric Build Up* (GBU) foi implementado utilizando o método de resolução através de sistemas de equações com distâncias interatômicas.

Para versão discreta do problema com conjunto arbitrário de distâncias foram implementados os algoritmos *Branch and Prune* original utilizando o método de coordenadas internas e uma adaptação do BP foi proposta com interseção de quatro esferas usando o método de distâncias interatômicas.

5.1.1 Algoritmo de Geração das Instâncias de Teste

Para realização dos testes de geração das estruturas foi necessário fazer um pré-processamento de dados, utilizando os arquivos do *Protein Data Bank* como entrada de um algoritmo que calcula as distâncias entre os átomos. O algoritmo de pré-procesamento gera como saída arquivos contendo os átomos da molécula, qual aminoácido pertencem e um intervalo de distância entre eles.

Para cada proteína testada foram gerados quatro tipos de instâncias de teste, a partir do arquivo PDB referente a essa proteína. Foram geradas instâncias para os dois tipos de MDGP, ou seja, algumas foram geradas simulando um conjunto completo de distâncias entre os átomos e outras simulando um conjunto arbitrário de distância, sendo que esse conjunto arbitrário foi conseguido utilizando somente as distâncias menores que 6 Å, simulando assim os dados da Ressonância Magnética Nuclear.

Além de diferenciar pelo tipo de conjunto de distâncias, foi gerado também algumas utilizando somente os átomos do backbone e algumas utilizando backbone e cadeia lateral conforme é mostrado na Figura 5.1.

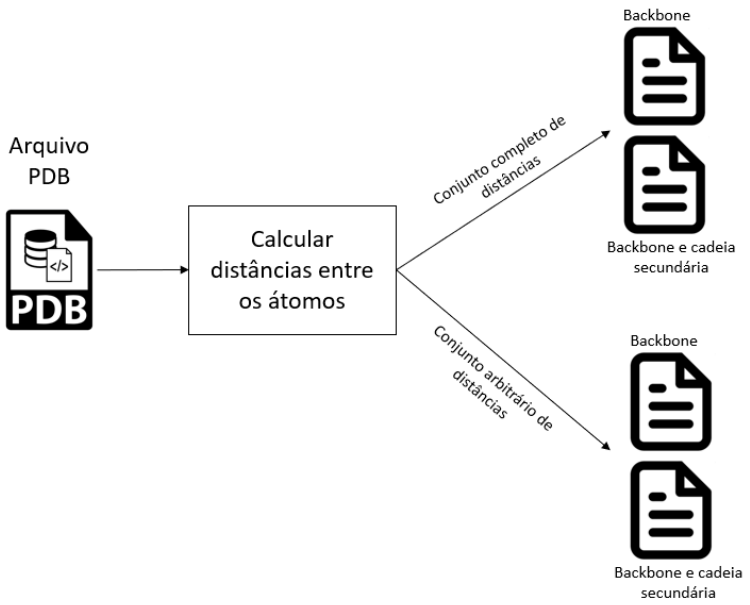


Figura 5.1: Geração das instâncias de teste.

Os quatro arquivos de teste gerados para cada proteína receberam o nome da referência da proteína no banco PDB seguido pelo tipo de arquivo gerado, *CC* para conjunto completo, *CP* para conjunto parcial e *ct*6 para as que foram simulação do RMN com corte de distâncias menores que 6 Å. Logo os arquivos seguiram a seguinte lógica e nomenclatura:

- Uma com o conjunto completo de distâncias (sem corte) e com toda a estrutura (backbone e secundária) - nomeada *refPDB_CC*.
- Uma com o conjunto de distâncias menores que 6Å (corte 6) e com toda a estrutura (backbone e secundária) - nomeada *refPDB_ct6_CC*.
- Uma com o conjunto completo de distâncias (sem corte) do backbone - nomeada *refPDB_CP*.
- Uma com o conjunto de distâncias menores que 6Å (corte 6) do backbone - nomeada *refPDB_ct6_CP*

5.1.2 Algoritmo proposto - BP com Quatro Esferas

O *Branch and Prune* original usa a interseção de três esferas para calcular a possível posição do átomo seguinte, gerando duas possibilidades. Como a interseção de quatro esferas é somente um ponto como mostrado na Figura 5.2, se for possível encontrar uma distância entre o átomo atual e outro diferente dos três anteriores é possível gerar uma quarta esfera e usar a interseção das quatro como a posição única do átomo.

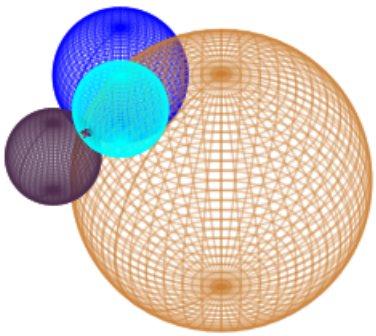


Figura 5.2: Interseção de quatro esferas.

O Algoritmo 4 segue todas as regras do *Branch and Prune*, começa fixando os quatro primeiros átomos como o BP original. Mas a partir do quinto átomo, além de utilizar os três átomos anteriores, procura um quarto átomo já fixado (que tenha distância conhecida para o átomo que se deseja fixar) para ser o centro da quarta esfera.

Quando a quarta esfera é encontrada, o sistema de equações da interseção de esferas mostrada na Seção 3.1.2.1.2 é usado para encontrar a coordenada do átomo. Sempre que for possível encontrar a quarta esfera a solução do problema não abre ramificações na árvore de busca, assim grande parte da estrutura é conseguida de forma linear como mostrado na Figura 5.3.

Algoritmo 4: Algoritmo BP com quatro esferas

Fixar os átomos x_1, x_2, x_3 e x_4 como o BP original

para $i = 5$ *até* n **faça**

  Procurar um átomo $x_v | v < i - 3$ que tenha distância conhecida para o átomo i

  Fixar o átomo i usando os átomos x_{i-1}, x_{i-2}, x_{i-3} e x_v

fim para

5.1.3 Variações de Podas Testadas no Algoritmo Branch and Pune

Durante os experimentos realizados com o algoritmo Branch and Prune, foram testadas algumas modificações no procedimento de poda da árvore de geração de posições dos átomos. A poda padrão do BP verifica se as distâncias entre a posição do átomo atual gerado e todos os anteriormente gerados que possuem informações de distância no arquivo de entrada são compatíveis, admitindo um intervalo de tolerância. Se qualquer uma delas não for válida a posição é invalidada e a árvore é podada.

Foram testamos duas formas diferentes de poda, na primeira maneira chamada de **Poda pela Base**, utilizamos somente os três átomos da base fixada no início do processo como validação de poda, se todas as distâncias conhecidas entre o átomo fixado e os da base forem compatíveis a posição é validada e a árvore continua a ser gerada, mas se um deles for incompatível, a posição é invalidade e o ramo é podado.

A segunda maneira pode ser mais caracterizada como condição de aceite do que poda, chamada de **Apto com Um**. Nesse experimento foram utilizadas todas as posições dos átomos anteriores para validar a continuação do ramo. Foi verificado se as distâncias entre a posição do átomo atual gerado e todos os anteriormente gerados que possuem informações de distância no arquivo de entrada são compatíveis, se uma delas fosse compatível a posição do átomo era aceita como válida e a ramificação da árvore continuava gerando as demais posições.

5.1.4 Comparação dos Algoritmos

Algoritmos para versões contínuas e discreta do MDGP foram estudados, implementados na linguagem de programação C e testados: o algoritmo *polinomial de Dong e Wu* para o conjunto completo de distâncias; o algoritmo *Geometric Build Up*; o método *Branch and Prune* original e uma adaptação do *Branch and Prune*, utilizando o cálculo da interseção de quatro esferas.

Para testar se os valores obtidos pela implementação dos algoritmos BP original e BP com quatro esferas estavam corretos, os resultados obtidos foram comparados com as saídas do software *MD-Jeep*, que é um *Branch and Prune* implementado por Mucherino et al. (MUCHERINO; LIBERTI; LAVOR, 2010), para as mesmas entradas.

O BP modificado foi comparado com o BP original (implementação feita durante a pesquisa) e com o Software MD-Jeep, utilizando as instâncias disponibilizadas pelos ide-

alizadores do algoritmo, que foram geradas a partir de arquivos do banco PDB retirando os átomos da cadeia lateral, calculando as distâncias entre os átomos e considerando somente as distâncias menores que 6 Å.

Para testar o algoritmo polinomial de Dong & Wu, foram criadas instâncias com o conjunto completo de distâncias entre pares de átomos. As proteínas utilizadas foram as mesmas dos demais algoritmos, mas na geração das instâncias a partir de arquivos do banco PDB foram considerados apenas os átomos do *backbone* (Carbono C, Nitrogênio N e Carbono alfa $C\alpha$) e calculadas todas as distâncias entre os átomos.

Os resultados dos testes para o algoritmo Polinomial de Dong & Wu, BP original e BP com quatro esferas utilizando instâncias com distâncias no máximo de 6Å são mostrado na Tabela 5.1.

Tabela 5.1: Tabela de comparação dos Algoritmos. Molécula - nome da proteína do PDB, |V| - número de átomos, |E| - quantidade de distâncias conhecidas, Enum - número de átomos fixados, Tsoma - soma total das distâncias entre todos os átomos fixados.

Molécula	**|V|**	**|E|**	**D&W**		**BP original**		**BP4**	
			Enum	**Tsoma**	**Enum**	**Tsoma**	**Enum**	**Tsoma**
1a70	291	1628	291	659553.151	291	659553.151	27	3512.112
1bpm	1443	9303	1443	31436395.882	1443	31436395.882	27	3659.986
1crn	138	846	138	117779.167	138	117779.167	30	3321.266
1fs3	372	2209	372	1273285.784	372	1273285.784	30	2862.799
1hoe	222	1259	222	354460.508	222	354460.508	27	2977.760
1mbn	459	3200	459	2078741.920	459	2083523.246	10	186.396
1mqq	2032	13016	2032	69452237.209	2032	69452298.455	27	2640.510
1pht	249	1448	249	473924.361	249	473924.361	21	1737.185
1poa	354	2201	354	1137943.707	354	1137943.707	32	3160.257
1ppt	108	660	108	77191.496	108	77191.496	27	3275.104
1ptq	150	829	150	140790.146	150	140790.146	24	2367.833
1rgs	792	4936	792	8237455.610	792	8237429.209	11	256.620
fragmento	290	41905	290	574516.598	26	1362.497	14	292.687

Na Tabela 5.3 são mostradas as quantidades de átomos fixados e a soma das distâncias entre todos os átomos da molécula. Como o BP4 gera somente uma estrutura parcial, são mostradas também a soma do GBU e BP para a mesma quantidade de átomos geradas pelo BP4. Na Tabela 5.2 são mostrados as quantidades de átomos gerados pelo BP e BP4 e os átomos visitados pelo BP4 e GBU para encontrar a base.

O algoritmo BP original implementado gerou as mesmas coordenadas que o MD-Jeep para todas as instâncias de testes e foi em alguns casos mais rápido devido a estruturas de

Tabela 5.2: Tabela de comparação de número de átomos gerados pelo BP e BP4 e visitados pelo GBU e BP4 com conjunto de distâncias menores ou iguais a 10Å.

Molécula	**|V|**	**|E|**	**Átomos fixados**	**Átomos gerados**		**Átomos visitados**	
				BP	**BP4**	**GBU**	**BP4**
1a70	291	7565	43	63	43	561	444
1crn	138	3238	39	64	39	373	232
1fs3	372	9912	32	52	32	211	118
1hoe	222	6029	36	54	36	430	333
1mbn	459	11457	33	39	33	228	134
1pht	249	6181	37	54	37	507	408
1poa	354	9289	27	47	27	126	51
1ppt	108	2081	48	67	48	785	640
1ptq	150	3550	57	88	57	1226	1067

dados diferentes utilizadas.

Como no BP com quatro esferas a base de cálculo para os átomos varia a cada átomo e o método utilizado é o de sistemas de equações com distâncias euclidianas para a interseção de esferas, ocorre um acúmulo de erro a cada átomo calculado. Uma vez que um átomo utiliza a posição do átomo anterior que já tem erro, o acúmulo de erro acaba ultrapassando o erro permitido e a enumeração de posições para.

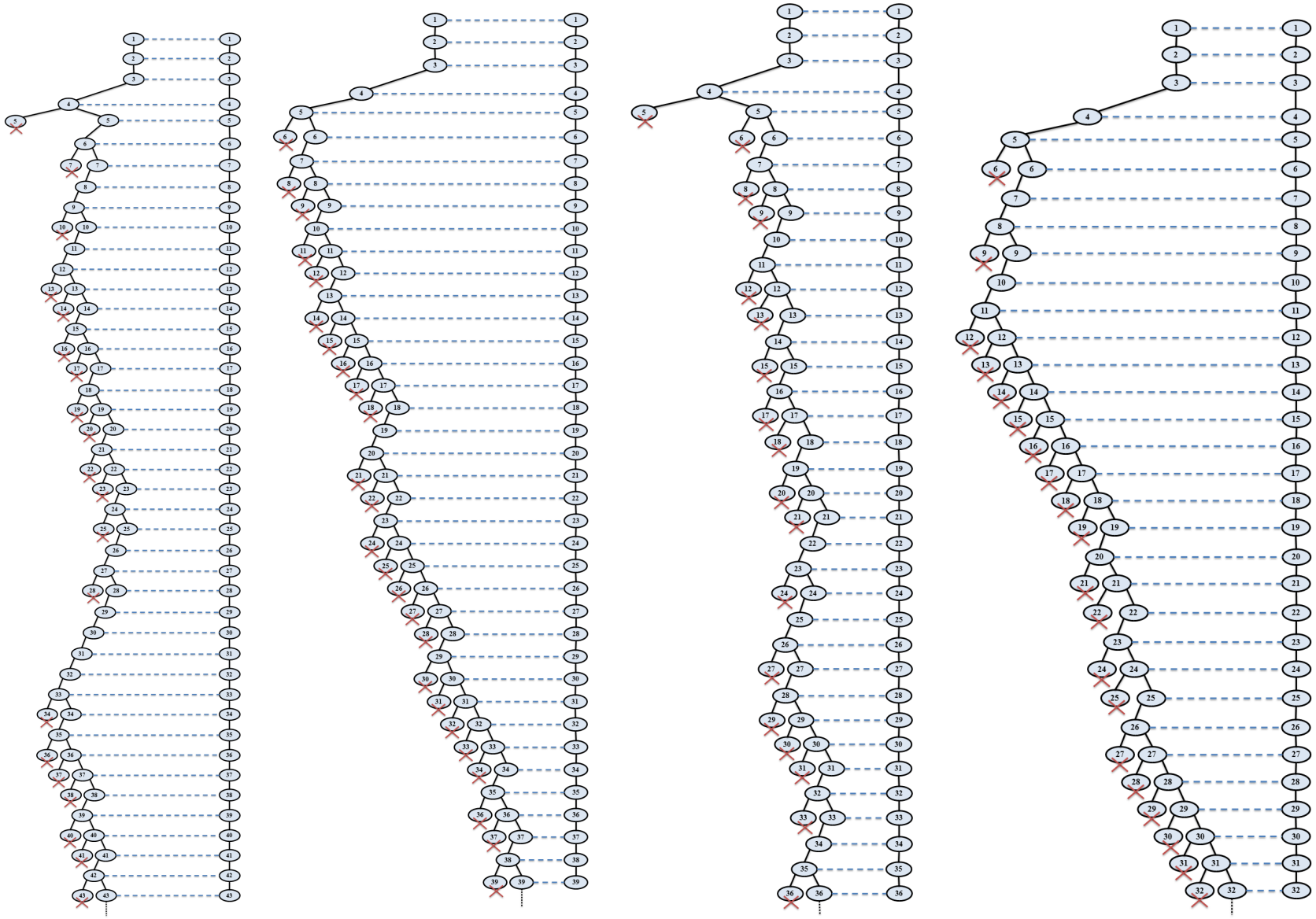
Logo o BP modificado implementado aplicando sistemas de distâncias euclidianas gera as mesmas coordenadas que o MD-Jeep para todas as intâncias, mas apenas a um determinado nível ou quantidade de átomos fixos, porque na resolução de sistemas lineares ocorre um acúmulo de erros na determinação da posição que impede a aplicação desta técnica para determinar a estrutura da molécula inteira. Somente uma estrutura parcial é obtida e esta versão está, portanto, limitada por este contexto.

Na Figura 5.3 são mostradas as árvores geradas pelo BP para algumas instâncias e os passos feitos pelo BP4 para calcular os mesmos átomos. O BP sempre gera duas possibilidades de posição para cada átomo, gerando assim uma árvore de possibilidades, entretanto no BP4 sempre que for possível encontrar a quarta esfera, a solução do problema não abre ramificações na árvore de busca, assim grande parte da estrutura é conseguida de forma linear.

Como o formato PDB é um padrão compatível com muitos programas de visualização de moléculas, todos os algoritmos geram como saída arquivos do tipo PDB contendo as

Tabela 5.3: Tabela de comparação entre os algoritmos GBU, BP e BP4 com conjunto de distâncias menores ou iguais a 10Å.

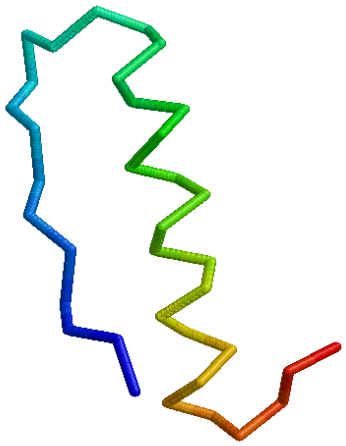
Molécula	**|V|**	**|E|**	**Enumeração completa**			**Enumeração parcial**			
			Enum	**GBU**	**BP**	**Enum**	**GBU**	**BP**	**BP4**
				Tsum	**Tsoma**		**Psoma**	**Psoma**	**Psoma**
1a70	291	7565	291	660002.280	659553.151	43	9977.118	9977.118	9977.031
1crn	138	3238	138	218754.097	117779.167	39	5899.918	5899.918	5899.940
1fs3	372	9912	372	1785242.399	1273285.784	32	3324.448	3324.448	3324.470
1hoe	222	6029	222	354460.520	354460.508	36	6505.256	6505.256	6505.244
1mbn	459	11457	459	2231055.949	2078741.915	33	3613.976	3613.976	3613.952
1pht	249	6181	249	484666.602	473924.361	37	8371.016	8371.016	8371.107
1poa	354	9289	354	1136892.154	1137943.707	27	1991.926	1991.926	1991.874
1ppt	108	2081	108	77191.496	77191.496	48	13437.097	13437.097	13437.104
1ptq	150	3550	150	140790.105	140790.146	57	18931.732	18931.731	18931.775



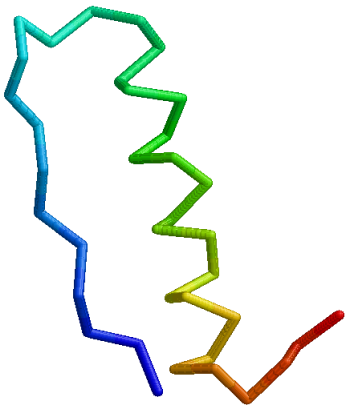
(a) Molecule 1A70 (b) Molecule 1CRN (c) Molecule 1hoe (d) Molecule 1fs3

Figura 5.3: Comparação das etapas de geração dos átomos nos algoritmos BP e BP4.

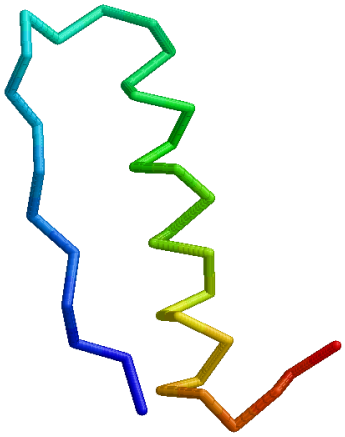
informações das coordenadas dos átomos. Esses arquivos de saída foram utilizados no programa RasMol disponível em http://www.rasmol.org/ para geração de representações das moléculas, assim a Figura 5.4 mostra uma comparação entre as estruturas geradas pelos algoritmos e a molécula original do banco PDB.



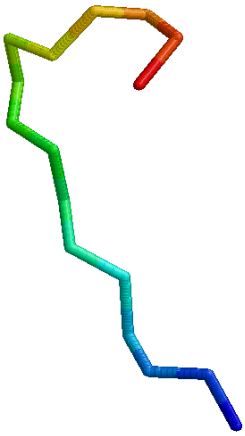
(a) Arquivo do banco PDB



(b) Saída BP original



(c) Saída algoritmo polinomial



(d) Saída algoritmo BP quatro esferas

Figura 5.4: Estruturas geradas pelos algoritmos comparados com a original do banco PDB para o hormônio pancreático contido na instância 1PPT.

5.1.5 Assinalamento da Proteína TRX1

O assinalamenteo da proteína TRx1, foi realizado com dados da RMN adquiridos durante visita técnico-científica ao Centro Nacional de Ressonância Magnética Nuclear (CNRMN), coordenado pelo pesquisador Fábio Almeida do Instituto de Bioquímica Médica (IBqM/UFRJ), na Universidade Federal do Rio de Janeiro (UFRJ).

As tioredoxinas são proteínas pequenas, estáveis ao calor, presentes em todos os organismos vivos. Eles estão envolvidos em um grande número de processos celulares. A Trx1 é uma das três isoformas diferentes da tioredoxina que compoem a levedura Sacharomyces cerevisiae (PINHEIRO et al., 2007).

5.1.5.1 Dados de Entrada do Assinalamento

Os espectros obtidos pela RMN do amostra da proteína TRx1 em solução foram assinalados de acordo os deslocamentos químicos dos aminoácidos da proteína. Para o assinalamento foram necessários:

- Sequência (ordem) dos 103 aminoácidos que formam a proteína Trx1, mostrada na Figura 5.5 : MVTQFKTASE FDSAIAQDKL VVVDFYATWC GPCKMIAPMI EKFSEQYPQA DFYKLDVDEL GDVAQKNEVS AMPTLLLFKN GKEVAKVVGA NPAAIKQAIA ANA

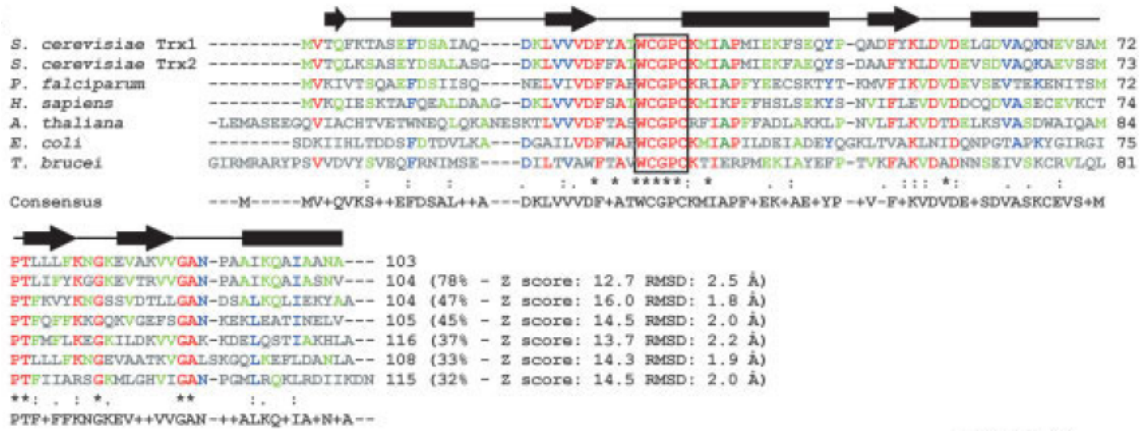


Figura 5.5: Alinhamento sequencial de Trx1 e proteínas estruturalmente homólogas de diferentes organismos (PINHEIRO et al., 2007).

- Tabela de deslocamento químico do (C_α) e (C_β) de cada um dos aminoácidos existentes. Mostrada no Apêndice A.

- Espectros gerados pela RMN - dados de saída da ressonância fornecidos pelo CNRMN.

- Software CCPN que é um programa gráfico para visualização do espectro, atribuição de dados e análise de RMN. Esse SW foi utilizado para visualizar os desloca-

mentos químicos fornecidos pelos espectros e a atribuição da ordem dos aminoácidos de acordo com seus deslocamentos de (C_α) e (C_β) (VRANKEN et al., 2005).

5.1.5.2 Resultado do Assinalamento

Os principais desafios e dificuldades de determinar as estruturas de proteínas usando RMN são: proteínas têm milhares de sinais, atribuir o sinal específico para cada átomo, milhares de interações entre átomos também precisam ser atribuídos e a necessidade de transformar a representação de espectros através de interações entre átomos em coordenadas espaciais (VRANKEN et al., 2005).

Durante a visita ao CNRMN, conseguimos identificar uma das maneiras de determinação da estrutura da proteína por RMN, o fluxo realizado seguiu os passos mostrados na Figura 5.6. As etapas de determinação da estrutura proteína por RMN são: Preparar a amostra; Fazer os experimentos de RMN; Fazer assinalamento do backbone e da cadeia lateral; Encontrar as restrições estruturais: distâncias, ângulos de torsão, restrição de orientação; Calcular e validar a estrutura.

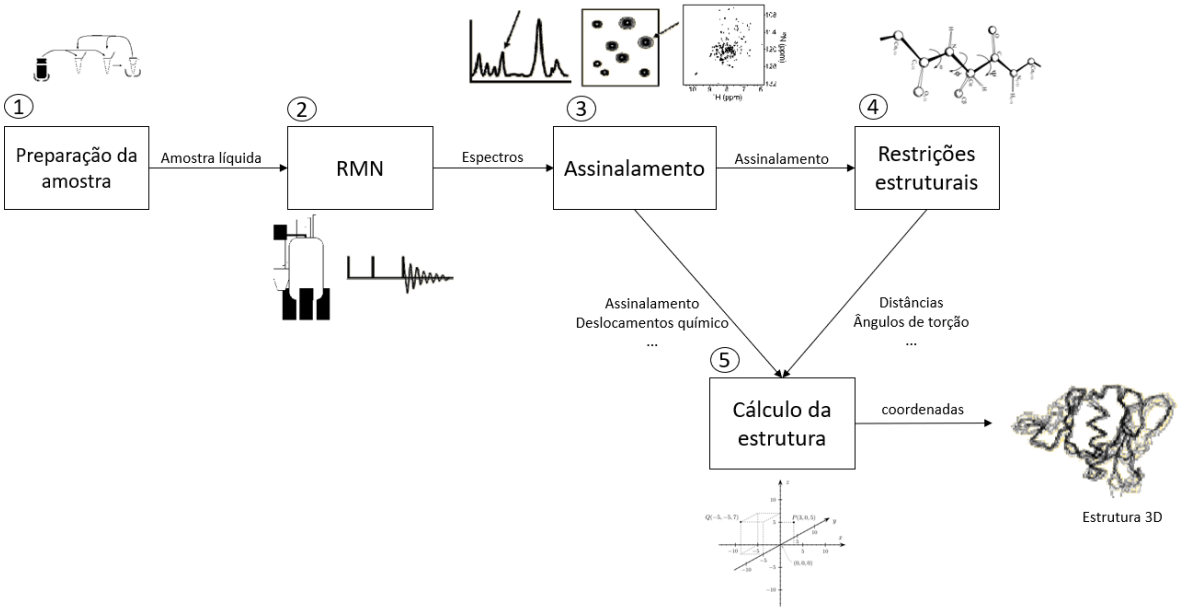


Figura 5.6: Fluxo para predição da estrutura 3D de uma proteína realizado no CNRMN

O fluxo de obtenção da estrutura tridimensional a partir da RMN começa na etapa 1 preparando a amostra em meio líquido para os experimentos de RMN, essa etapa foi realizada por técnicos do CNRMN. Essa etapa tem como saída a amostra líquida da proteína.

Na etapa 2 a amostra é inserida no equipamento de RMN e em intervalos de tempo

pulsos eletromagnéticos são feitos na amostra, após esses pulsos são detectadas e processadas as frequências de todos os núcleos da amostra para obtenção de várias informações que compõem os experimentos, como por exemplo os deslocamentos químicos. essa etapa tem como entrada a amostra líquida da proteína e gera como saída os espectros com deslocamentos químicos.

Na etapa 3 é feito o assinalamento do backbone e da cadeia lateral, que consiste em os espectros obtidos na etapa anterior serem analisados utilizando por exemplo os deslocamentos químicos dos carbonos (C_α) e (C_β) para ordenar os sinais de acordo com a sequência de aminoácidos da proteína. Tem como entrada os espectros para assinalamento e gera o assinalamento da proteína como saída.

Nessa fase com a orientação dos professores Fábio Almeira e Gisele Amorim foi realizado o assinalamento da proteína utilizando o programa CCPNmr Analyses (VRANKEN et al., 2005) e a tabela de deslocamento (Apêndice A) para identificar o tipo do aminoácido (Seção 3.2.1.1) e para identicar o aminoácido anterior ou posterior criando assim a cadeia de sequência de sinais (ordem). As Figuras do assinalamento completo são mostradas no Apêndice B.

Na etapa 4 outros cálculos são feitos a partir do assinalamento para obtenção de restrições estruturais da proteína como ângulos de torção e distâncias entre átomos. Tem como entrada assinalamento da proteína e gera como saída os ângulos de torsão. Essa fase foi feita com o programa TalosN (TALOS-N..., a), os dados do assinalamento foram enviados para o Talos N Server disponível em (TALOS-N..., b) e os arquivos de saída com os valores dos ângulos foram recebidos no email cadastrado no momento do upload.

Na etapa 5 os dados de assinalamento e as restrições conseguidas nas etapas anteriores são utilizados para calcular e validar a estrutura tridimensional da proteína. Essa fase foi realizada com o programa Aria (ARIA...,), que pode receber como entrada as restrições estrturais e tem como saída a estrura 3D da proteína.

5.2 FASE II - Validação Bioquímica e Estudo de caso

Durante a segunda fase do mestrado foram realizados trabalhos na tentativa de encontrar uma maneira de validar quimicamente as estruturas geradas pelos algoritmos e realizado o estudo de caso envolvendo as proteínas do SARS-CoV-2.

5.2.1 Metodologia Fase II

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 5.7 para o cálculo estrutural das variantes.

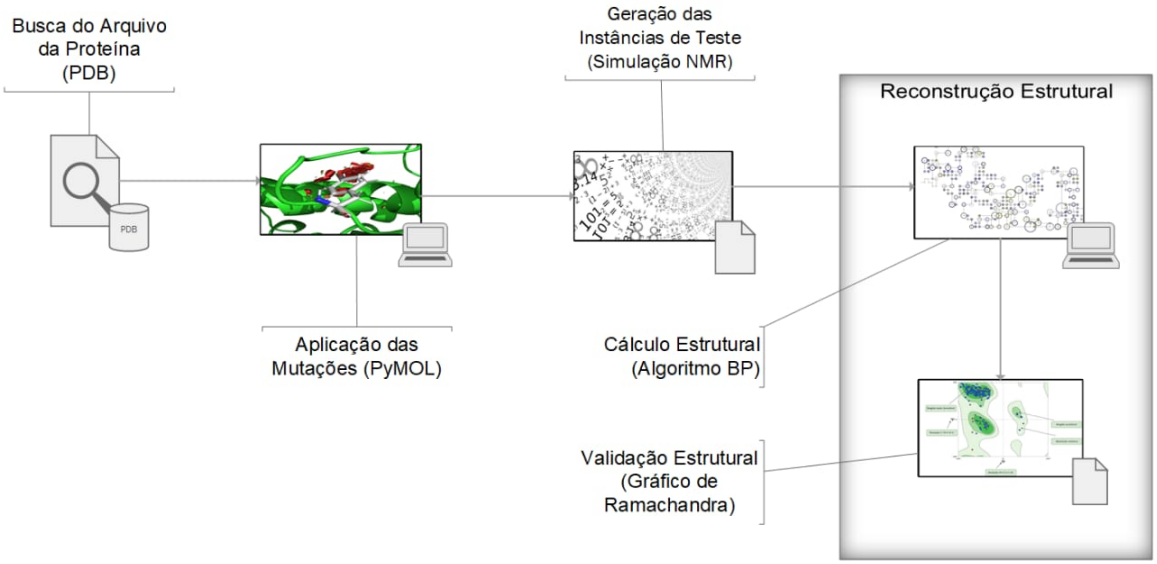


Figura 5.7: Fluxo do cálculo estrutural das variantes encontradas no Amazonas.

Primeiramente foi realizada uma busca do arquivo original sem mutação da proteína do vírus, no banco de dados de proteínas RCSB PDB (BERMAN et al., 2000). Foi 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 , 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 estru-

tura 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 3.4).

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.

5.2.2 Estudo de Caso - Proteínas do Vírus SARS-CoV-2

Nesta seção são mostradas algumas estruturas 3D geradas a partir de experimentos realizados com proteínas do novo coronavírus. Para a realização dos testes foram criadas instâncias de teste, a partir de arquivos do banco de dados RCSB PDB (BERMAN et al., 2000), que simularam os dados de RMN de acordo com as regras explicadas na Seção 5.1.1. As estruturas do SARS-CoV-2 escolhidas para os testes e outros estudos são mostradas na Tabela 5.4.

Tabela 5.4: Informações cristalográficas das estruturas estudadas nesta dissertação (NELSON et al., 2020; KERN et al., 2020; NELSON; HALL; FREMONT, 2020; HANKE et al., 2020; XU et al., 2021).

Proteína	Massa (kDa)	Átomos	Resíduos	Resolução	Tipo
6W37	7.54	525	66	2.9	ORF7a
6XDC	64.33	3.150	386	2.9	ORF3a
7JX6	24.00	1.822	192	1.61	ORF8
6YWK	95,51	6.935	861	2,2 Å	Nsp3
6WTT	105,02	7.430	906	2,15 Å	Protease principal
6XS6	417,78	18.294	2.357	3.70Å	
7BWJ	71,63	4.820	636	2,58Å	
6M0j	97,14	6.571	791	2,45Å	

A plataforma adotada para gerar os diagramas de Ramachandran utilizados na validação das saídas do algoritmo foi MolProbity (<https://swissmodel.expasy.org/assess>) (CHEN et al., 2010) com o módulo *Structure Assessment* implementado na plataforma SWISS-Model (WATERHOUSE et al., 2018) pela qual obtemos a porcentagem de aminoácidos pertencentes às regiões favoráveis e com restrições estéricas.

O valor RMSD da sobreposição entre as soluções preditas pelo algoritmo desenvolvido e a respectiva estrutura cristalográfica foi na plataforma PDBeFold (<https://www.ebi.ac.uk/msd-srv/ssm/>) (KRISSINEL; HENRICK, 2004). Sendo que a obtenção do arquivo .pdb já com as estruturas alinhadas foi mediante a ferramenta TM-Align (<https://zhanglab.ccmb.med.umich.edu/TM-align/>) (ZHANG; SKOLNICK, 2005).

5.2.2.1 Geração das Variantes Encontradas no Amazonas

Em consequência da indisponibilidade de estruturas cristalográficas com as variantes no estado do Amazonas, foi necessário simular a geração dessas variantes do vírus juntamente com um grupo de pesquisas de química da UFAM, liderado pelo professor Kelson Mota.

A mutação *in silico* (Figura 5.8) foi efetuada com a plataforma CHARMM-GUI <http://www.charmm-gui.org/?doc=input/pdbreader> com o módulo *PDB READER* e funcionalidade *Add mutation* partindo-se das estruturas de referência ORF3a (PDB ID: 6XDC) e ORF8 (PDB ID: 7JX6), gerando ao final o arquivo de topologia .pdb que possui compatibilidade com o campo de força CHARMM36. Nota-se que antes da submissão do arquivo .pdb ao algoritmo desenvolvido houve uma minimização estrutural prévia. O software utilizado para o alinhamento foi Schrödinger Maestro 2020-3 com o módulo *Superposition* entre todos os átomos.

Realizou-se uma minimização estrutural para que fosse possível remover conflitos estéricos na predição conformacional das proteínas ORF3a (PDB ID: 6XDC) e ORF8 (PDB ID: 7JX6) em decorrência das variantes no Amazonas. Todos os arquivos de entrada foram gerados com o auxílio da ferramenta QwikMD incluso no software VMD 1.4.3 (HUMPHREY; DALKE; SCHULTEN, 1996). Sendo assim mediante o algoritmo NAMD 2.14 (PHILLIPS et al., 2005) via simulated annealing foi realizado o aquecimento gradual da temperatura até atingir 300,0 K e resfriamento sob $1ps \cdot K^{-1}$. Considerou-se efeitos de solvatação de moléculas de água mediante o modelo TIP3P.

Figura 5.8: Etapas adotadas para minimização estrutural das estruturas ORF3a e ORF8 contendo as variantes no Amazonas até o passo final de alinhamento estrutural com a respectiva estrutura sem quaisquer mutações.

Houve a inserção de íons Na^+ e Cl^- sob a concentração fisiológica de 0,15 M para atingir a neutralidade do sistema. A minimização foi realizada sob 1000 passos e posterior simulação MD em aproximadamente 2 ns sob pressão de 1 atm e temperatura constante de 300,0 K com ensemble NPT. A pressão manteve-se constante com o pistão estocástico de Langevin e temperatura mediante o método de Nosé-Hoover. As interações eletrostáticas de longo-alcance foram estimadas com o algoritmo de Particle-Mesh-Ewald (PME) a uma distância de corte de 14,0 Å. O passo para cada iteração foi o tempo padrão de 2 fs para evitar instabilidades no sistema.

Utilizou-se a ferramenta I-Mutant 3.0 (CAPRIOTTI; FARISELLI; CASADIO, 2005) que implementa a abordagem de machine-learning SVM para assim predizer $\Delta\Delta G$ em consequência das variantes emergentes no complexo ACE2-RBD (PDB ID: 7DF4) onde foi adotada a condição fisiológica em $pH = 7,4$, temperatura de 25°C e algoritmo SVM3 de classificação ternária onde denota-se estabilidade ($\Delta\Delta G \geq +0,5 kcal \cdot mol^{-1}$), desestabilização ($\Delta\Delta G < -0,5 kcal \cdot mol^{-1}$) e neutralidade ($-0,5 \leq \Delta\Delta G \geq +0,5$).

5.2.2.2 Validação da Predição para ORF7

Foram geradas quatro instâncias de teste a partir do PDB referentes a essa proteína:

- Uma com o conjunto completo de distâncias (sem corte) e com toda a estrutura (backbone e secundária) - nomeada $6w37_CC$
- Uma com o conjunto de distâncias menores que 6Å (corte 6) e com toda a estrutura (backbone e secundária) - nomeada $6w37_ct6_CC$
- Uma com o conjunto completo de distâncias (sem corte) do backbone - nomeada $6w37_CP$
- Uma com o conjunto de distâncias menores que 6Å (corte 6) do backbone - nomeada $6w37_ct6_CP$

Para o cálculo da estrutura tridimensional da proteína 6W37 foi utilizado o algoritmo Brach-and-Prune. Foram calculadas e analisadas todas as possibilidades de estrutura da árvore de geração que o algoritmo constrói, todas as estruturas geradas respeitam as restrições matemáticas do problema, mas algumas delas podem ferir restrições físico-químicas

das proteínas então foi utilizado o gráfico de Ramachandran para validar quais as estruturas geradas são aceitáveis quimicamente.

O primeiro teste realizado foi para a instância $6w37_ct6_CP$ que gerou duas estruturas tridimensionais como soluções (estrutura X,Y,Z), que respeitam as restrições matemáticas de distância. Contudo, ao gerar os gráficos de Ramachandran para as duas foi observado que a primeira gerou um gráfico de Ramachandran que respeitava as regiões válidas, sendo assim uma estrutura aceitável. Entretanto a segunda solução foi totalmente invalidada porque gerou um gráfico de Ramachandran com praticamente todos os pontos em regiões erradas, como mostrado na Figura 5.9.

Os demais testes realizados com as instâncias da proteína $6w37$ seguiram o padrão do primeiro, geraram duas estruturas 3D como solução, contudo após validação uma delas mostrou-se inconsistente, como mostrado na Figura 5.9. Um ponto interessante que identificamos 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 predizer toda a estrutura testada, sendo uma de suas soluções válida de acordo com as restrições matemáticas e físico-químicas. Contudo, mais testes estão sendo feitos para validar essa constatação observada.

Na sobreposição estrutural entre a solução válida para a predição do ORF8 (PDB ID: 6W37) e a respectiva estrutura cristalográfica obteve-se um RMSD tendendo à zero. Consequentemente, o algoritmo foi capaz de reconstruir a proteína de forma muito consistente. Nota-se que optamos pelo resultado de alinhamento com maior probabilidade, isto é, o de menor RMSD. Em contrapartida, a solução com mais inconsistências (6w37_-CC_sol_1) no diagrama de Ramachandran foi justamente a que teve baixo alinhamento com um RMSD de 3,06 Å.

5.2.2.3 Predição Estrutural Induzida Pelas Variantes Amazonenses G196V e L84S

Após a validação de teste com a estrutura ORF7 e assim constatando a eficácia do algoritmo implementado, foi possível então predizer o novo enovelamento da proteína ORF3a com a variante G196V assim como a mutação L84S pertencente a ORF8 que acometeram o estado do Amazonas.

Ao realizar o alinhamento estrutural na ORF3a, as diferenças foram relativamente pequenas com RMSD de 3,29 Å. Em relação à estrutura ORF8, houve um deslocamento

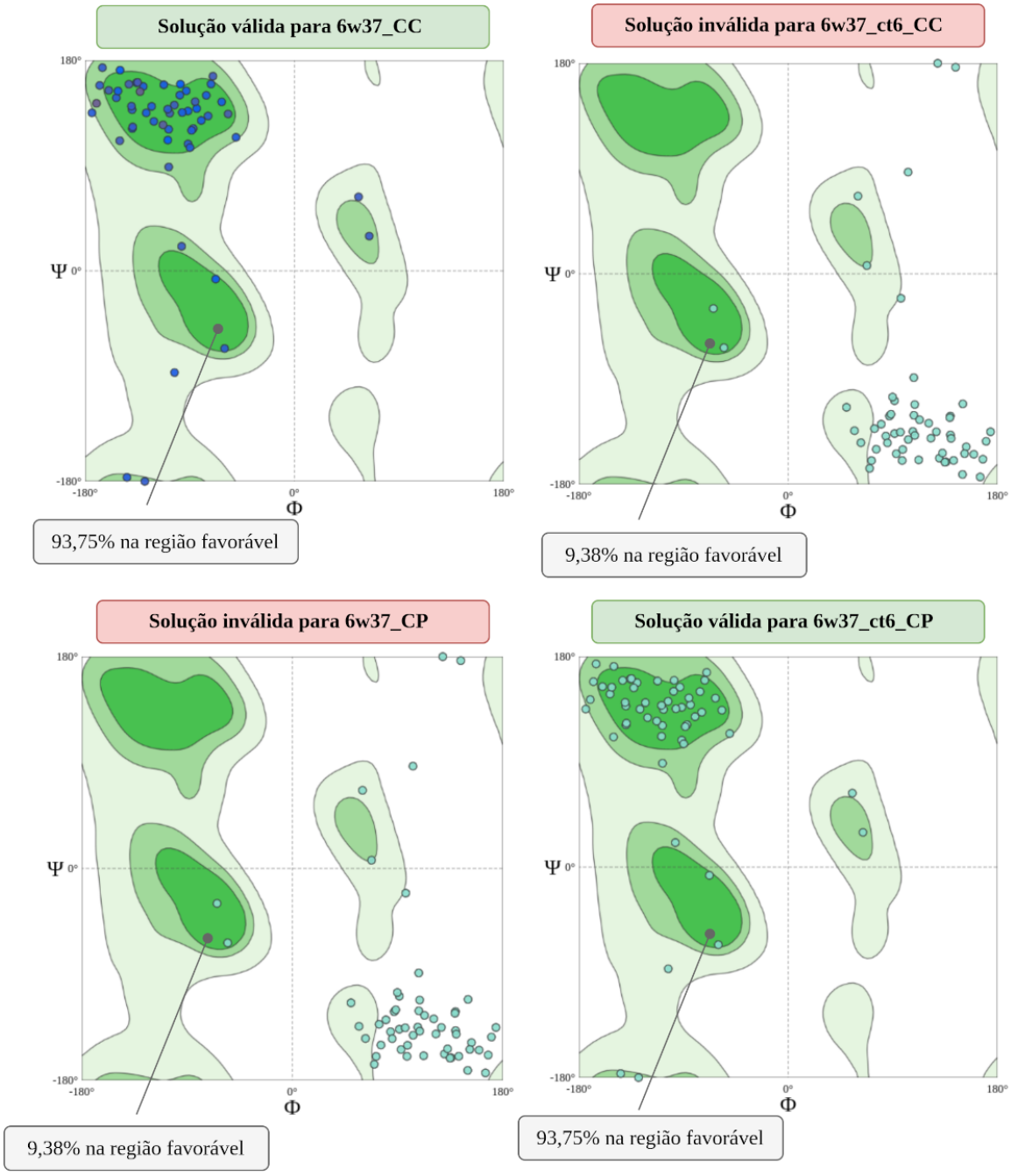


Figura 5.9: Diagramas de Ramachandran referente às soluções previstas pelo algoritmo para ORF7 (PDB ID: 6W37). Todas as imagens foram construídas com o auxílio da plataforma MolProbity.

com RMSD de 0,27 Å em comparação à variante. Embora as mudanças possam ter sido apenas em decorrência da natureza estocástica da dinâmica molecular, a mutação pode realmente ter induzido sutis alterações conformacionais constatadas após a minimização estrutural. A sobreposição das estruturas sem mutações e com as variantes é mostrada na Figura 5.10

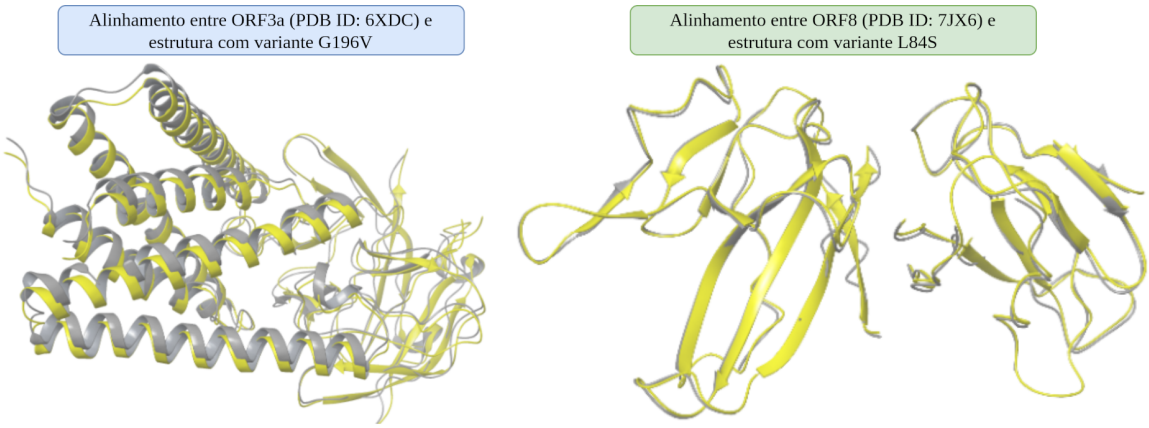


Figura 5.10: Sobreposição estrutural entre as estruturas cristalográficas sem quaisquer mutações da ORF3a (PDB ID: 6XDC) e ORF8 (PDB ID: 7JX6) e as respectivas variantes G196V e L84S.

A solução de maior probabilidade para estrutura ORF8 (PDB ID: 7JX6) acarretou em um total de 91,07% dos resíduos na região mais favorável, como apresentado na Figura 5.11. A estrutura cristalográfica apresentou exatamente o mesmo número de aminoácidos nesta região, embora alguns ângulos de torção na predição ainda não são condizentes com a cristalografia. O alinhamento estrutural propiciou um RMSD de 0 Å, indicando que o algoritmo pôde reconstruir de forma correta a estrutura de entrada.

Referente à estrutura ORF3a (PDB ID: 6XDC), a predição mais consistente apresentou 95,93% dos resíduos na região mais favorecida. E assim como nos testes anteriores, a estrutura obtida experimentalmente apresentou o mesmo percentual na região sem impedimentos estéricos no diagrama de Ramachandran e cujo RMSD de alinhamento foi 0 Å. Os gráficos de ramachandran feitos para a estruturas são apresentados na Figura 5.12

5.2.2.4 Estudo da Linhagem Amazonense B.1.1.28 de clado P.1/P.2

Todas as possibilidades de estrutura da árvore de geração que o algoritmo BP constrói foram calculadas e analisadas, todas as estruturas geradas respeitaram as restrições matemáticas do problema, mas algumas delas foram consideradas quimicamente inválidas

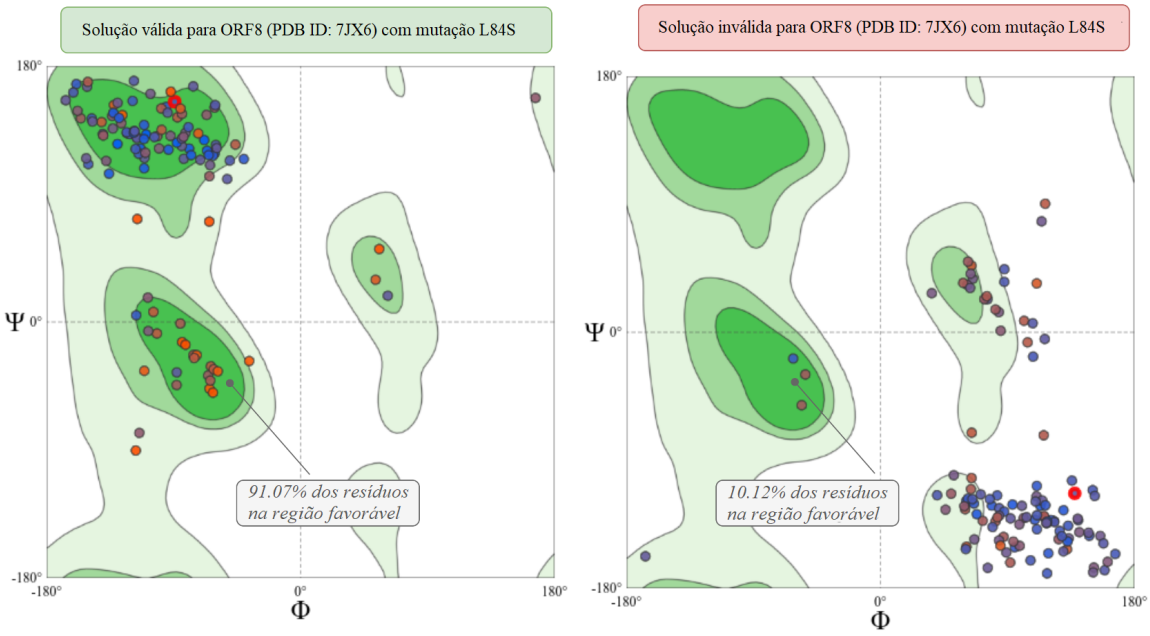


Figura 5.11: Diagramas de Ramachandran referente às soluções previstas pelo algoritmo para ORF8 (PDB ID: 7JX6). Todas as imagens foram construídas com o auxílio da plataforma MolProbity.

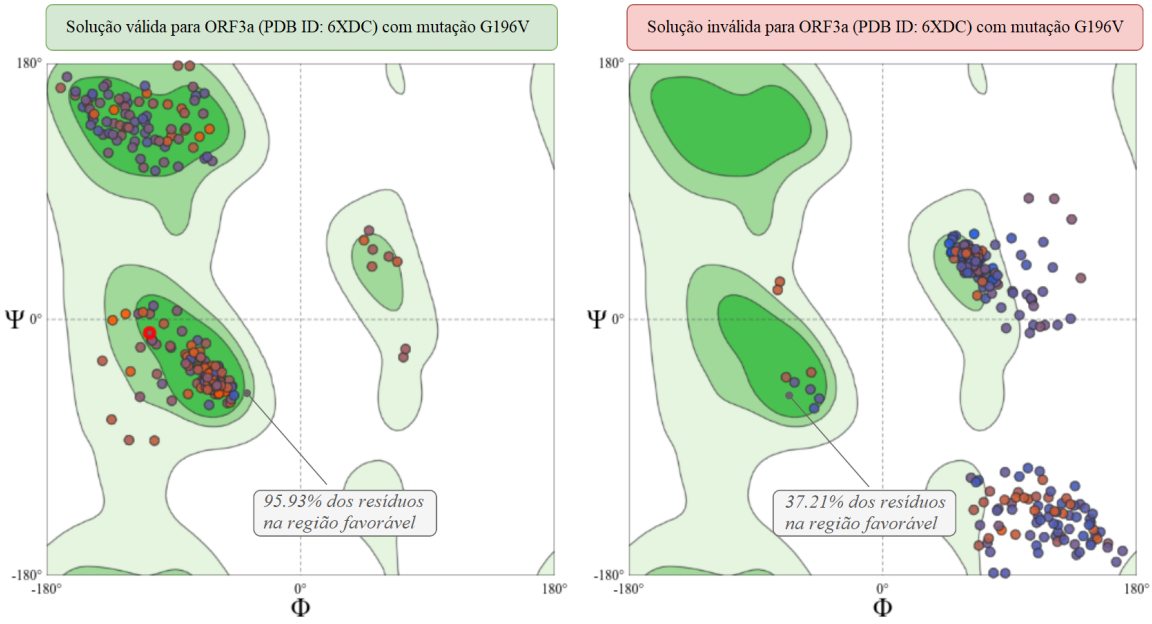


Figura 5.12: Diagramas de Ramachandran referente às soluções previstas pelo algoritmo para ORF3a (PDB ID: 6XDC). Todas as imagens foram construídas com o auxílio da plataforma MolProbity.

quando feriram as restrições físico-químicas das proteínas ao gerar o 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. Porém, quando os gráficos de Ramachandran foram gerados para as duas soluções, verificou-se que a primeira foi totalmente invalidada porque gerou um gráfico com quase todos os pontos em regiões erradas. O segundo porém, gerou um gráfico de Ramachandran que respeitou as regiões válidas, sendo uma estrutura aceitável.

A estrutura do complexo ACE2-RBD (PDB ID: 6M0J) contendo as mutações da cepa P.1 considerada quimicamente válida apresentou um total de $97,06\%$ dos aminoácidos da região sem impedimento estérico, conforme mostrado em Figura 5.13. A estrutura cristalográfica de referência apresentou $97,08\%$ de resíduos, e portanto uma diferença muito pequena em relação à reconstrução, sendo um importante indicativo da grande consistência do algoritmo implementado neste trabalho. Além disso, a solução mais consistente mostrou $2,92\%$ dos ângulos de torção sob condições marginais, enquanto $0,38\%$ sob condições de impedimento estérico total.

Os testes para o anticorpo-antígeno (PDB ID: 7BWJ) seguiu o padrão do primeiro, geraram duas estruturas 3D como solução, porém após a validação, uma delas se mostrou inconsistente. A estrutura reconstruída apresentou $93,97\%$ dos aminoácidos em uma região sem impedimento estérico, como mostrado em Figura 5.13. Assim, embora o algoritmo não seja capaz de remover os conflitos estéricos induzidos pela mutagênese, a reconstrução estrutural foi satisfatoriamente próxima da estrutura cristalográfica.

Na Figura 5.14 é possível observar que a solução mais consistente reconstruída pelo algoritmo mostrou um RMSD tendendo a 0Å para o complexo ACE2-RBD contendo P.1 quando alinhado com a estrutura cristalográfica do PDB ID: 6M0J . Por outro lado, a solução inválida gerou um RMSD maior de $7,54$Å.

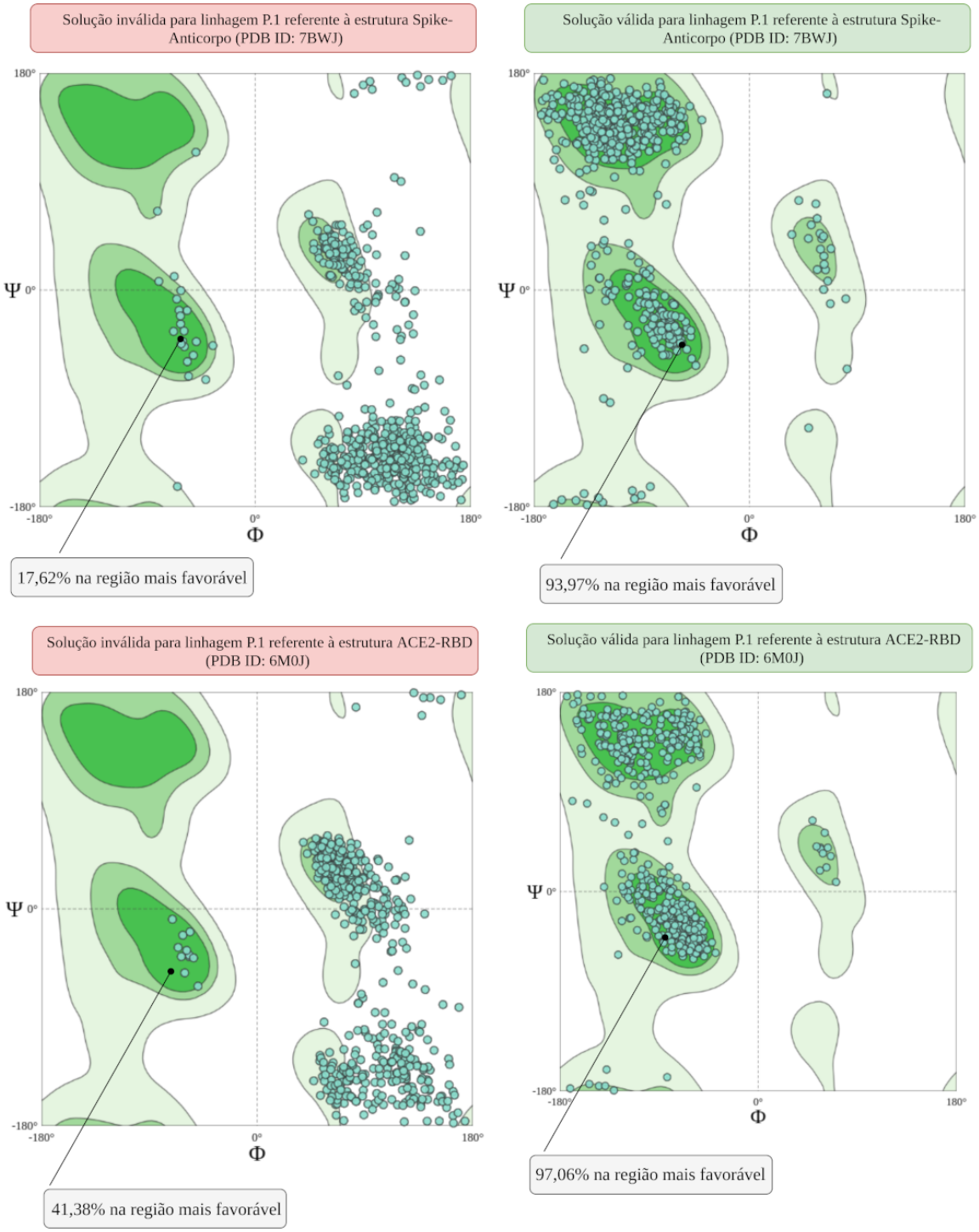


Figura 5.13: Diagramas de Ramachandran para as soluções obtidas no algoritmo BP para a cepa P.1. As estruturas de referência onde a mutagênese e a respectiva reconstrução foram aplicadas foram os complexos ACE2-RBD (PDB ID: 6M0J) e Spike-Anticorpo (PDB ID: 7BWJ). Todas as imagens foram construídas com o auxílio da plataforma MolProbity.

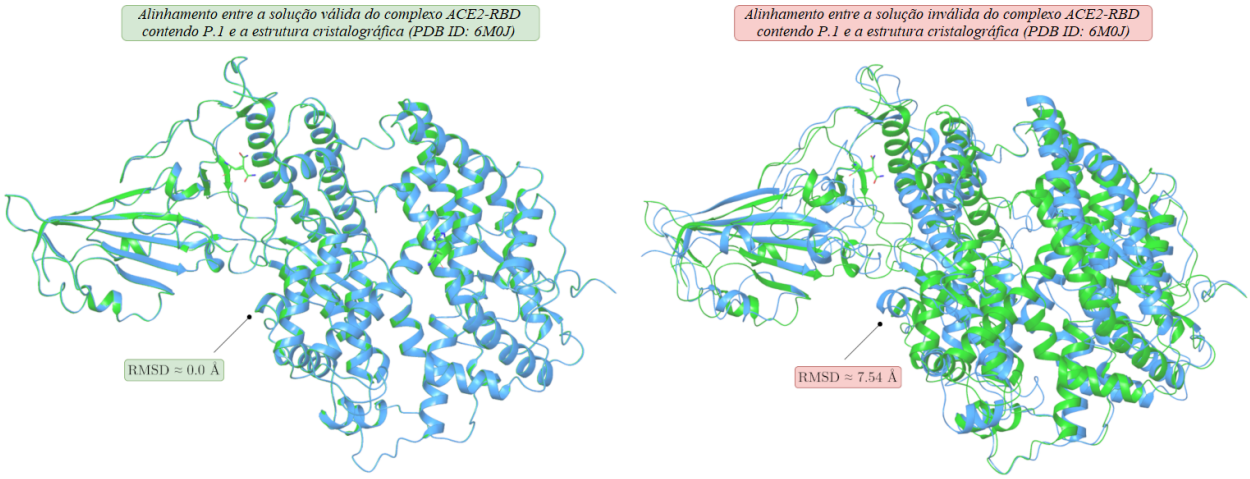


Figura 5.14: Alinhamento estrutural no software Schrödinger Maestro 2020-4 entre soluções válidas e inválidas contendo P.1 em comparação com a respectiva estrutura cristalográfica do complexo ACE2-RBD (PDB ID: 6M0J).

5.2.2.4.1 Análise Química da Linhagem Amazonense

A partir da análise dos valores médios das ligações RMSD, RMSF, SASA e Hidrogênio, conforme Tabela 5.5, em geral pode ser visto que mutações da variante P.1 estabilizaram a estrutura ACE2-RBD, embora em alguns resíduos centrais foi observado um aumento na flexibilidade estrutural de acordo com a análise RMSF. Um comparativo entre as mutações é mostrado na Figura 5.15.

Tabela 5.5: Comparação entre as variantes P.1 e P.2 em relação aos valores médios de alguns parâmetros resultantes da dinâmica molecular na faixa de $18ns$ para ACE2-RBD (PDB ID: 6M0J) que quantificam mudanças estruturais no Spike Região RBD na interação com ACE2.

Análise	**Sem mutação**	**Linhagem P.1**	**Linhagem P.2**
Média RMSD	(2.05 ± 0.32)Å	(1.76 ± 0.26)Å	(1.81 ± 0.34)Å
Média RMSF	(1.22 ± 0.51)Å	(1.06 ± 0.38)Å	(1.33 ± 0.76)Å
Ligações de hidrogênio	197 ± 12	202 ± 12	203 ± 11
Média SASA	$(378.55 \pm 3.64)nm^2$	$(375.04 \pm 3.16)nm^2$	$(373.35 \pm 3.32)nm^2$
Contatos nativos	0.9918 ± 0.0015	0.9895 ± 0.0016	0.9877 ± 0.0019
Média Rg	(31.37 ± 0.30)Å	(31.64 ± 0.18)Å	(31.48 ± 0.18)Å

É importante destacar que 3 (três) análises corroboram a hipótese de maior estabilidade da ACE2-RBD em função de P.2, entre elas: Valores médios mais baixos de RMSD e RMSF e maior formação de ligações de hidrogênio. Enquanto apenas 2 (duas) análises refletem a hipótese de maior instabilidade, sendo elas: Maior raio de giração (Rg) e menores contatos nativos.

A maior estabilização no complexo ACE2-RBD como resultado de certas mutações pode ser uma explicação de porque o vírus seguiu uma evolução convergente como já relatado em alguns estudos experimentais (LAM et al., 2020; BOBAY; ODONNELL; OCHMAN, 2020; KEMP et al., 2021; HODCROFT et al., 2021), embora as causas até então são desconhecidos.

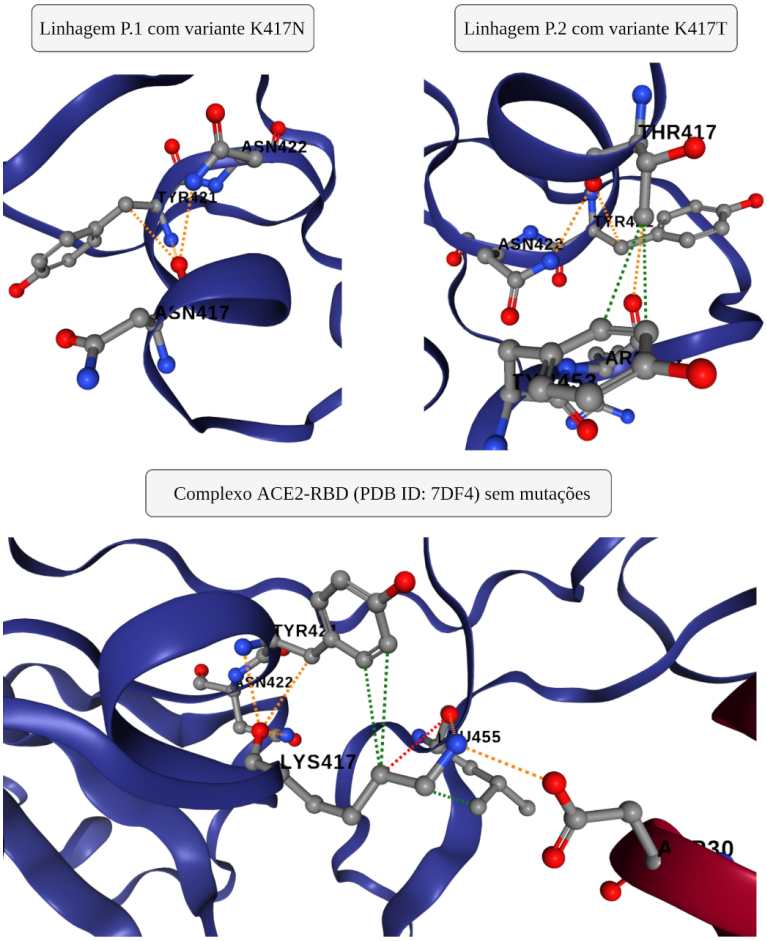


Figura 5.15: Comparativo entre as interações químicas formadas no complexo ACE2-RBD (PDB ID: 7DF4) em consequência das mutações K417N e K417T referentes às linhagens P.1 e P.2, respectivamente. Todos os diagramas foram gerados na plataforma DynaMut2 (RODRIGUES; PIRES; ASCHER, 2021). As linhas tracejadas em Verde representam os contatos hidrofóbicos, as linhas em Vermelho correspondem às ligações de Hidrogênio e por fim a cor Laranja refere-se às interações polares.

Através dessas simulações é possível verificar que existe uma tendência a uma maior

estabilidade estrutural nas mutações mais frequentes. Em outras palavras, mutações que foram repetidas em várias cepas, como E484K e N501Y, tendem a ter maior estabilidade termodinâmica. Os impactos das linhas P.1 e P.2 ficarão mais claros à medida que as simulações forem repetidas ou com o aumento do intervalo de tempo, o que geraria resultados mais conclusivos, convergência das flutuações e maior reprodutibilidade.

Efetuou-se uma análise de $\Delta\Delta G$ em consequência das variantes que compõem a linhagem P.1/P.2 com auxílio da ferramenta I-Mutant 3.0 (CAPRIOTTI; FARISELLI; CASADIO, 2005). Acredita-se que a variante N501Y surgida no Reino Unido a priori não afetaria de forma significativa a interação ACE2-RBD (PDB ID: 7DF4) e portanto a eficácia das atuais vacinas em consequência de uma desestabilização não crítica de $0,08$ $kcal \cdot mol^{-1}$.

Por outro lado, K417N surgida no Amazonas gerou uma inesperada desestabilização de -1,50 $kcal \cdot mol^{-1}$ no reconhecimento ACE2-RBD, o que embora possa diminuir a probabilidade de invasão celular poderia dificultar o reconhecimento da resposta imune induzida pelas vacinas. Além disso, a desestabilização crítica em decorrência de K417N poderia ter uma correlação com o aumento sem precedentes no mês de Janeiro-2021 em infecções e óbitos no estado do Amazonas por conta da COVID-19.

A linhagem de clado P.2 distingue-se pela mutação K417T que gera uma desestabilização menos significativa embora ainda crítica de $-1,35$ $kcal \cdot mol^{-1}$. Nota-se que mutações onde há desestabilização, tendem a sumir por pressão evolutiva, visto que causam impacto negativo na função protéica (ANCIEN et al., 2018). Em relação à linhagem P.2 a variante K417T mostrou uma maior estabilidade em relação à K417N tendo como resultado $-0,82$ $kcal \cdot mol^{-1}$.

No intuito de que obter maior convicção dos resultados de estabilidade do complexo ACE2-RBD, foi realizado a predição de $\Delta\Delta G$ mediante a ferramenta mCSM (PIRES; ASCHER; BLUNDELL, 2013) onde: N501Y ($-0,757 kcal \cdot mol^{-1}$); K417N ($-1,582 kcal \cdot mol^{-1}$); E484K ($-0,081 kcal \cdot mol^{-1}$). Novamente a mutação K417N da linhagem P.1 apresentou uma desestabilização crítica assim como ocorreu na ferramenta I-Mutant 3.0. Quando analisado a linhagem P.2 a mutação K417T gerou uma desestabilização de $-1,594$ $kcal \cdot mol^{-1}$ corroborando a hipótese de modificações no resíduo 417 na proteína Spike poder ser o maior responsável pelo aumento da transmissibilidade do vírus no Amazonas.

Por intermédio da ferramenta DynaMut2 (RODRIGUES; PIRES; ASCHER, 2021) pôde-se compreender o impacto das mutações nas ligações químicas. A mutação K417N pertencente à linhagem P.1 é considerada desestabilizante em consequência de haver a perda de 1 (uma) ligação de Hidrogênio, 3 (três) contatos hidrofóbicos além de 2 (duas) interações polares na interface ACE2-RBD. O mesmo padrão foi observado na linhagem P.2 com a variante K417T, embora com maior estabilidade em consequência de ter perdido menos ligações com o desaparecimento de 1 (uma) ligação de Hidrogênio, 1 (um) contato hidrofóbico e 1 (uma) interação polar.

Mediante o auxílio da plataforma PDBePISA, verificou-se que uma mutação preocupante na linhagem B.1.1.28 P.1/P.2 seria a mesma que acometeu o Reino Unido. Nisto a mutação N501Y acarretou na formação de uma ligação de hidrogênio entre os resíduos Tyr501-Lys353 a uma distância de 2,90 Å o que explicaria o $\Delta_i G \approx -6,5\ kcal \cdot mol^{-1}$ em comparação à estrutura sem quaisquer mutações onde $\Delta_i G \approx -5,9\ kcal \cdot mol^{-1}$. Isto talvez fosse a causa da maior transmissibilidade na linhagem britânica B.1.1.7, mas também indicaria a priori que a linhagem no Amazonas também seria caracterizada por um aumento de transmissibilidade.

5.2.2.5 Testes realizados com outras variantes do vírus para validar a metologia

Os demais testes realizados com as instâncias de teste seguiram o padrão dos anteriores. Cada teste gerou como solução duas ou mais estruturas 3D matematicamente válidas, porém, após validação com o gráfico de Ramachandran, uma delas sempre foi inconsistente, sendo quimicamente inválida. Um resumo dos resultados é mostrado na Tabela 5.6.

Os gráficos de Ramachandran apresentados na Figura 5.16 mostram a comparação entre a estrutura cristalográfica referente ao arquivo PDB ID: 6WTT e as duas estruturas reconstruídas no teste. A estrutura considerada válida obteve um percentual de resíduos na região favorável igual a da estrutura cristalográfica de 94,78%, enquando a solução inválida apresentou uma porcentagem muito baixa de 26,33%, o que invalidou a estrutura.

Para a estrutura da variante PDB ID: 6XS6, a estrutura cristalográfica apresentou uma porcentagem de 93,88% dos resíduos na região favorável. A solução de maior probabilidade reconstruída pelo BP para esta estrutura acarretou um percentual um pouco menor, um total de 93,83% dos resíduos na região favorável, como apresentado na Figura 5.17. A estrutura invalidada pelo gráfico de Ramachandran gerou apenas 24,32% dos resíduos

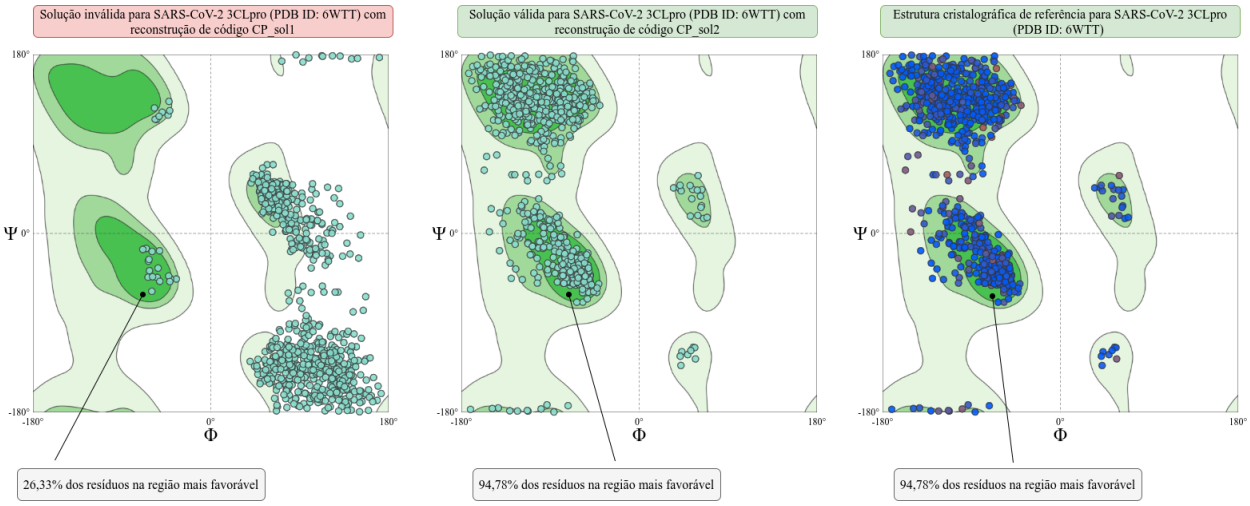


Figura 5.16: Diagramas de Ramachandran referente às soluções previstas pelo algoritmo para (PDB ID: 6WTT). Todas as imagens foram construídas com o auxílio da plataforma MolProbity.

na região correta.

Os teste para o arquivo de referencia PDB ID: 6YWK são mostrados na Figura 5.18, a solução considerada quimicamente inválida apresentou um percentual de 39,37% na região válida, enquanto a quimicamente validada pelo gráfico de Ramachandran obteve 99,65% dos resíduos na região favorável. O percentual da estrutura válida foi igual ao da estrutura cristalográfica, o que mais uma vez mostrou a eficiência do algortmo Branch-and-Prune em obter pelo menos uma solução totalmente válida e muito próxima da real.

Tabela 5.6: Resumos dos testes realizados para validar a metodologia utilizada. As porcentagens de resíduos nas regiões favoráveis do gráfico de Ramachandran são apresentadas para ambas as soluções validadas e invalidadas pelos testes.

Proteína		**% dos Resíduos na Região Favorável**	
ID PDB	**Região**	**Solução Válida**	**Solução Inválida**
6W37	ORF7a	93.75%	9.38%
6YWK	Nsp3	99.65%	39.37%
6WTT	Main protease	94.78%	26.33%
6XS6	Spike-ACE2	93.83%	24.32%
6XDC	ORF3a	95.93%	37.21%
7JX6	ORF8	91.07%	10.12%
6M0J	ACE2-RBD	97.06%	41.38%
7BWJ	Antigen-antibody	93.97%	17.62%

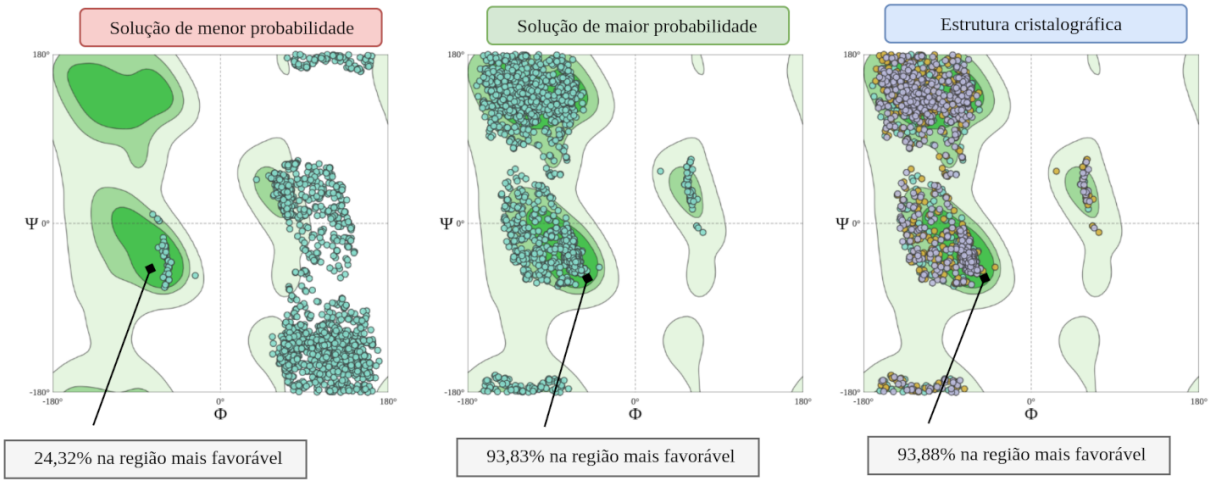


Figura 5.17: Diagramas de Ramachandran referente às soluções previstas pelo algoritmo para (PDB ID: 6XS6). Todas as imagens foram construídas com o auxílio da plataforma MolProbity.

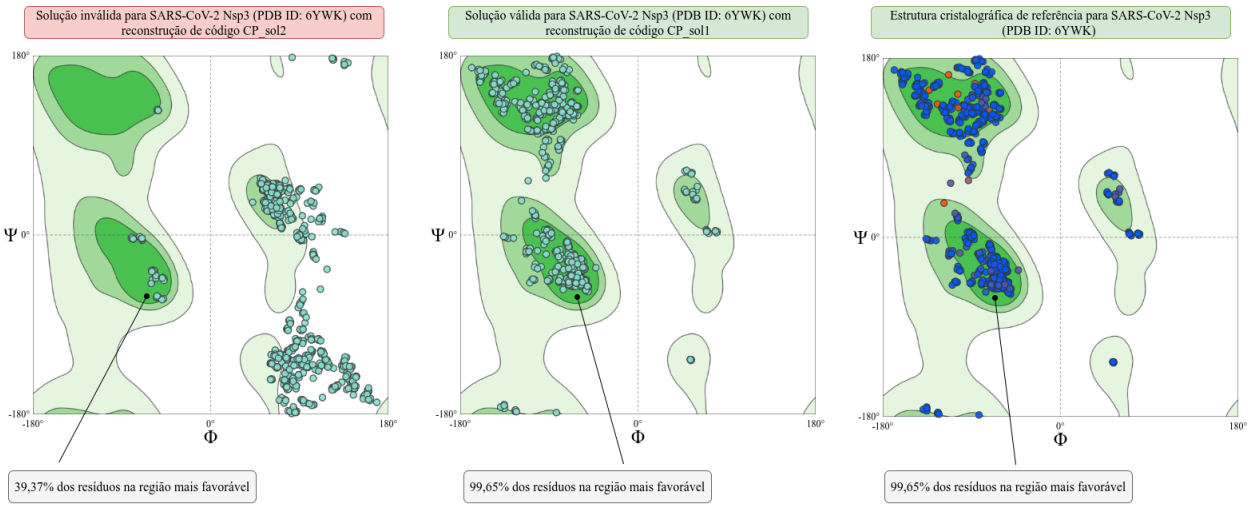


Figura 5.18: Diagramas de Ramachandran referente às soluções previstas pelo algoritmo para (PDB ID: 6YWK). Todas as imagens foram construídas com o auxílio da plataforma MolProbity.

Capítulo 6

Considerações Finais

Este trabalho abordou o Problema de Geometria de Distâncias Moleculares (*Molecular Distance Geometry Problem* - MDGP), no qual se tenta resolver a seguinte questão: dado um grafo $G = (V, E)$, ponderado e não-direcionado, existe uma imersão válida de G em $\mathbb{R}^3$?

A resposta a este questionamento tem sido fonte de intensas pesquisas, e neste trabalho foram apresentados as variações do problema, as principais abordagens teóricas e os principais métodos existentes. Foram mostrados dois tipos de abordagem matemática para a resolução do MDGP: solução de sistemas lineares de equações interatômicas de distância euclidiana, e a resolução de sistemas de coordenadas internas utilizando técnicas de multiplicação de matrizes.

Foram apresentados, também, algoritmos que resolvem os dois tipos clássicos de MDGP, um para o conjunto completo de distâncias e os algoritmos *Geometric Build Up* (GBU), *Branch-and-Prune* (BP) original e *Branch-and-Prune* com quatro esferas para o caso onde se tem previamente apenas um subconjunto arbitrário de distâncias.

Experimentos computacionais foram feitos utilizando instâncias do *Protein Data Bank* (PDB) adaptados para os tipos do MDGP, com conjunto de distâncias completas e incompletas simulando resultados de experimentos de Ressonância Magnética Nuclear.

Na *Fase I* da pesquisa foram testados os algoritmos implementados e para o algoritmo BP original implementado foi feita uma comparação entre suas saídas e os resultados obtidos pelo software *MD-Jeep*, que é um *Branch and Prune* implementado pelos idealizadores da versão discreta do MDGP (MUCHERINO; LIBERTI; LAVOR, 2010). A comparação mostrou que a implementação gerou as mesmas coordenadas do *MD-Jeep*

para todas as instâncias de testes e foi em alguns casos mais rápido, o que validou a implementação feita. O BP com quatro esferas implementado aplicando sistemas de equações, gerou as mesmas coordenadas que o MD-Jeep até uma quantidade de átomos, porque na resolução de sistemas lineares ocorre um acúmulo de erros que impossibilitou a determinação da estrutura completa da molécula.

No ano de 2020 se espalhou pelo mundo um novo vírus chamado coronavírus (COVID-19), que é uma família de vírus que causam infecções respiratórias semelhante à gripe e tem sintomas como tosse, febre e, em casos mais graves, pneumonia. Esse processo infeccioso se alastrando pelo mundo mostrou o quão importante é o trabalho de análise estrutural das proteínas, pois com a estrutura tridimensional é possível estudar sua reprodução, que é uma característa intimamente relacionada à sua estrutura. Então, conhecendo sua estrutura é possível identificar medicamentos que ataquem pontos estratégicos mais vulneravéis do vírus. Ou seja, esse cenário de pandemia e busca pela cura evidenciou a necessidade da rápida e simples geração da estrutura 3D para o desenvolvimento de medicamentos que inibem a reprodução do SARS-CoV-2 (GALILEU, 2020).

Na *Fase II* da pesquisa enquanto os teste das proteínas do SARS-CoV-2 eram feitos e analisados, foi possível identificar a necessidade de validar o processo, verificar se todos os resultados de estruturas tridimensionais matematicamente válidas geradas eram possíveis de existir quimicamente, ou seja, se também eram quimicamente válidas.

A metodologia desenvolvida na *Fase II* desta dissertação mostrou-se eficaz na reconstrução de proteínas das variantes do vírus SARS-CoV-2, a partir de dados mutados das proteínas do vírus. Utilizando o algoritmo BP e validações de Ramachandran, foi possível constatar a grande consistência das reconstruções estruturais baseadas na emulação do sinal de RMN. Assim, percebe-se cada vez mais que o desafio de reconstruir proteínas tem sido resolvido com o grande auxílio da informática, pois os princípios físico-químicos que regem o dobramento ainda não são totalmente compreendidos.

O algoritmo Branch-and-Prune implementado e usado para testar as instâncias do vírus mostrou-se muito preciso em reconstruir proteínas a partir de seus dados cristalográficos, atuando como uma engenharia reversa para estudarmos as variantes do vírus SARS-CoV-2 que acometeram o estado do Amazonas. Um ponto interessante identificado ao testar as instâncias com o backbone e a cadeia lateral foi que embora o BP tenha sido projetado para encontrar a estrutura do backbone, foi possível reconstruir toda a

estrutura testada, uma de suas soluções sendo válida de acordo com a matemática e restrições físico-químicas. No entanto, mais testes estão sendo realizados para validar essa observação.

Dentre as perspectivas futuras, pretende-se automatizar todos os processos da metologia e também implementar uma heurística que receba como entrada a sequência de aminoácidos e que, apesar da alta complexidade computacional, possa prever 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 da proteína.

Como atividades adicionais da pesquisa foram desenvolvidos e apresentados alguns trabalhos em conferências e workshops, inicialmente na *FASE I* com os algortimos e implementaçãoes e posteriormente também alguns na *FASE II* com o estudo de caso sobre o SARS-CoV-2.

Em 2012 ocorreu a participação no Congresso Latino-Iberoamericano Invertigación Operativa/Simpósio Brasileiro de Pesquisa Operacional (CLAIO/SBPO) , com apresentação de trabalho envolvendo algoritmos para a determinação do número lista cromático de uma coloração com restrições (COELHO et al., 2012), cujas técnicas algorítmicas puderam ser aproveitadas para o entendimento e implementação dos métodos para o MDGP.

Ocorreu a participação no *XXXIII* Congresso da Sociedade Brasileira de Computação/*VII* e-Science workshop(CSBC 2013/e-Science) de 23 a 26 de Julho de 2013 em Maceió-AL com o trabalho *Uma proposta de adaptação do algoritmo Branch-and-Prune usando a interseção de quatro esferas para o Problema de Geometria de Distâncias Moleculares* desenvolvido juntamento com Rosiane de Freitas Rodrigues e Kelson Mota (SOUZA; FREITAS; MOTA, 2013).

Foi desenvolvido o trabalho *Constraint programming algorithms to determine the position of atoms in a molecule* e apresentado no The International Conference on Operations Research (OR 2013 Conference) de 3 a 6 de Setembro de 2013 pela profa. Dra. Rosiane de Freitas (SOUZA; FREITAS, 2013c).

Em setembro de 2013 ocorreu a participação na Escuela Latino-Iberoamericana de Verano en Investigación Operativa (XVII ELAVIO 2013) em Valencia (Espanha) de 8 a 12 de Setembro de 2013 com o trabalho Algoritmos baseados em programação por restrição para o Problema de Geometria de Distâncias Moleculares (SOUZA; FREITAS,

2013a).

Também em setembro de 2013 o trabalho Enumeration algorithms for determining the 3D structure of proteins foi apresentado no XXXIV Congreso nacional de Estadística e Investigación Operativa (SEIO 2013) em Castellón (Espanha) de 11 a 13 de Setembro de 2013 (SOUZA; FREITAS, 2013d).

No XLV Simpósio Brasileiro de Pesquisa Operacional (XLV SBPO) em Natal de 16 a 19 de Setembro de 2013 foi apresentado o trabalho Análise do problema de geometria de distâncias moleculares baseada em coordenadas internas e sistemas de distâncias interatômicas para interseção de esferas (SOUZA; FREITAS, 2013b).

Em outubro de 2013 o trabalho Sphere Intersection Algorithms for Molecular Distance Geometry Problem feito juntamente com o prof. Dr. Mario Salvatierra foi apresentado na Conferencia Latinoamericana en Informática (CLEI 2013) na Venezuela de 7 a 11 de Outubro de 2013 (SOUZA; FREITAS; SALVATIERRA, 2013).

Em Julho de 2021 ocorreu a participação no *XV* Brazilian E-Science Workshop (BreSci) do Congresso da Sociedade Brasileira de Computação com o trabalho *Estratégia Algorítmica para a Reconstrução e Validação da Estrutura Molecular de Variantes do SARS-CoV-2* (SOUZA et al., 2021a).

O trabalho *Improvement of SARS-CoV-2 macromolecule conformation by algorithmic structural prediction* foi aceito na Conferencia Latinoamericana de Informática (CLEI) de 2021, apresentado em Outubro de 2021 e escolhido como um dos três melhores artigos da conferência (SOUZA et al., 2021d). Após o evento, o trabalho foi selecionado para ter uma versão extendida publicada no CLEI Electronic Journal (CLEIej).

Além dos eventos de computação a pesquisa também foi apresentada em eventos de química, com os trabalhos *Comparative analysis among different structural minimization methods based on Branch-and-Prune structural reconstruction algorithm* (SOUZA et al., 2021b) e *Theoretical simulations for the Delta variant of SARS-CoV-2 for understanding the evolution of structural stability* (SOUZA et al., 2021c) que foram apresentadados no XXI Simpósio Brasileiro de Quimica Teórica em Novembro de 2021.

Em dezembro de 2021 foi publicada uma matéria intitulada *3D molecular reconstruction and functional analysis of SARS-CoV-2 Variants of Concern* sobre o estudo de caso desenvolvido nesta dissertação, na revista *IFORS Newsletter: OR and Development Section, International Federation of Operational Research Societies* (FREITAS et al., 2021).

Referências

ALENCAR, J.; LAVOR, C.; LIBERTI, L. Realizing euclidean distance matrices by sphere intersection. *Discrete Applied Mathematics*, 2018. Citado na página 30.

AMANAT, F.; KRAMMER, F. SARS-CoV-2 Vaccines: Status Report. *Immunity*, v. 52, n. 4, p. 583 – 589, 2020. ISSN 1074-7613. Citado na página 59.

ANCIEN, F. et al. Prediction and interpretation of deleterious coding variants in terms of protein structural stability. *Scientific Reports*, v. 8, n. 1, 2018. ISSN 2045-2322. Citado na página 88.

ARIA - Ambiguous Restraints for Iterative Assignment. <http://aria.pasteur.fr/>. Citado na página 74.

BADAWI, A.; RYOO, S. G. Prevalence of comorbidities in the middle east respiratory syndrome coronavirus (mers-cov): a systematic review and meta-analysis. *International Journal of Infectious Diseases*, Elsevier, v. 49, p. 129–133, 2016. Citado na página 54.

BERG, J. M. et al. *Biochemistry*. New York: [s.n.], 2012. Citado na página 47.

BERMAN, H. M. et al. *The Protein Data Bank*. 2000. 235-242 p. Disponível em: <https://doi.org/10.1093/nar/28.1.235>. Citado 3 vezes nas páginas 62, 75 e 76.

BIOWORLD. *Biopharma products in development for COVID-19*. 2020. <https://www.bioworld.com/COVID19products>. Acesso em: 31 de Dezembro de 2020. Citado na página 58.

BOBAY, L.-M.; ODONNELL, A. C.; OCHMAN, H. Recombination events are concentrated in the spike protein region of betacoronaviruses. *PLOS Genetics*, Public Library of Science, v. 16, n. 12, p. 1–14, 12 2020. Citado na página 87.

BOSCH, B. J. et al. Severe acute respiratory syndrome coronavirus (sars-cov) infection inhibition using spike protein heptad repeat-derived peptides. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 101, n. 22, p. 8455–8460, 2004. Citado na página 54.

BOTTON, L. M. P. G. de. *Estudos Estruturais de Proteínas de Xanthomonas axonopodis pv citri por Ressonância Magnética Nuclear*. Tese (Doutorado) — Universidade de São Paulo, São Paulo, september 2007. Citado na página 40.

BRANDÃO, P. E. Could human coronavirus oc43 have co-evolved with early humans? *Genetics and molecular biology*, SciELO Brasil, v. 41, n. 3, p. 692–698, 2018. Citado na página 53.

CAPRIOTTI, E.; FARISELLI, P.; CASADIO, R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Research*, v. 33, n. 2, p. 306–310, 07 2005. ISSN 0305-1048. Citado 2 vezes nas páginas 79 e 88.

CAVALLI, A. et al. Protein structure determination from nmr chemical shifts. *Proceedings of the National Academy of Sciences*, v. 104, n. 23, p. 9615–9620, 2007. Citado na página 37.

CDC. *Symptoms of Coronavirus Disease 2019 (COVID-19)*. 2020. <https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html>. Acesso em: 05 ago. 2020. Citado na página 55.

CHEN, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D*, v. 66, n. 1, p. 12–21, 2010. Citado na página 77.

CHOW, K.-C.; ZHANG, C.-T. Prediction of protein structural classes. *Critical Reviews in Biochemistry and Molecular Biology*, v. 30, n. 4, p. 275–349, 2008. Citado na página 38.

CNS. *Conselho Nacional de Saúde*. 2020. <https://conselho.saude.gov.br/recomendacoes-cns/1112-recomendac-a-o-n-022-de-09-de-abril-de-2020>. Citado na página 4.

COELHO, F. et al. Algoritmos branch-and-bound e dsatur para o número cromático de uma coloração com restrições. Congresso Latino-Iberoamericano Invertigación Operativa/Simpósio Brasileiro de Pesquisa Operacional (CLAIO/SBPO), 2012. Citado na página 94.

COG-UK. *COG-UK update on SARS-CoV-2 Spike mutations of special interest Report 1*. 2020. <https://www.cogconsortium.uk/wp-content/uploads/2020/12/Report-1_COG-UK_19-December-2020_SARS-CoV-2-Mutations.pdf>. Acessado em: 20 de Dezembro de 2020. Citado na página 57.

COGNITION, S. *COVID-19 & SARS-CoV-2*. 2020. Acessado em 26 de Novembro, 2020. Disponível em: <https://cognitionstudio.com/covid-19/>. Citado 2 vezes nas páginas xi e 55.

CORMEN, T. H. et al. *Algoritmos - Teoria e Prática*. 2th. ed. [S.l.]: Editora Campus, 2002. Citado 2 vezes nas páginas 8 e 12.

COSTA, R. A. et al. New insights into structural, electronic, reactivity, spectroscopic and pharmacological properties of bergenin: Experimental, dft calculations, md and docking simulations. *Journal of Molecular Liquids*, v. 330, 2021. Citado na página 62.

DILL, K. A.; MACCALLUM, J. L. The protein-folding problem, 50 years on. *Science*, American Association for the Advancement of Science, v. 338, n. 6110, p. 1042–1046, 2012. ISSN 0036-8075. Citado na página 36.

DILL, K. A. et al. The protein folding problem. *Annual Review of Biophysics*, v. 37, n. 1, p. 289–316, 2008. Citado na página 37.

DOLHNIKOFF, M. et al. Pathological evidence of pulmonary thrombotic phenomena in severe covid-19. *Journal of Thrombosis and Haemostasis*, Wiley Online Library, 2020. Citado na página 52.

DONG, Q.; WU, Z. A linear-time algorithm for solving the molecular distance geometry problem with exact inter-atomic distances. *Journal of Global Optimization*, v. 22, p. 365–375, 2002. Citado 4 vezes nas páginas 2, 16, 19 e 63.

DONG, Q.; WU, Z. A geometric build-up algorithm for solving the molecular distance geometry problem with sparse distance data. *Journal of Global Optimization*, v. 26, p. 321–333, 2003. Citado na página 26.

DROSTEN, C. et al. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *New England journal of medicine*, Mass Medical Soc, v. 348, n. 20, p. 1967–1976, 2003. Citado na página 53.

DUCHENE, S. et al. Temporal signal and the phylodynamic threshold of sars-cov-2. *Virus Evolution*, v. 6, n. 2, 08 2020. ISSN 2057-1577. Citado na página 56.

ELSTNER, M.; FRAUENHEIM, T.; SUHAI, S. An approximate DFT method for QM/MM simulations of biological structures and processes. *Journal of Molecular Structure: THEOCHEM*, v. 632, n. 1, p. 29–41, 2003. Citado na página 38.

FIDALGO, F. et al. Uma formulação numérica para resolução de problemas de geometria de distâncias moleculares. *Simpósio Brasileiro de Pesquisa Operacional*, 2012. Citado 3 vezes nas páginas 2, 16 e 38.

FMUSP. *Genoma do SARS-CoV-2 do primeiro caso de COVID-19 da América Latina sequenciado em 48 horas no Instituto Adolfo Lutz*. 2020. Https://www.fm.usp.br/fmusp/noticias/-genoma-do-sars-cov-2-do-primeiro-caso-de-covid-19-da-america-latina-sequenciado-em-48-horas-no-instituto-adolfo-lutz. Citado na página 4.

FREITAS, R. de. *Times Assíncronos para a Resolução de Problemas de Otimização Combinatória com Múltiplas Funções Objetivo*. Dissertação (Mestrado) — IC - Universidade Estadual de Campinas, UNICAMP, 1996. Citado na página 10.

FREITAS, R. de. *Caracterizações e algoritmos para problemas clássicos de escalonamento*. Tese (Doutorado) — COPPE - Programa de Engenharia de Sistemas e Computação - Universidade Federal do Rio de Janeiro, Rio de Janeiro, may 2009. Citado na página 12.

FREITAS, R. de et al. 3d molecular reconstruction and functional analysis of sars-cov-2 variants of concern. IFORS Newsletter: OR and Development Section, International Federation of Operational Research Societies. Edição Online, Dezembro 2021. Citado na página 95.

GALILEU. *Cientistas recriam em 3D proteína responsável pela multiplicação do coronavírus*. 2020. <https://revistagalileu.globo.com/Ciencia/Saude/noticia/2020/03/cientistas-recriam-em-3d-proteina-responsavel-pela-multiplicacao-do-coronavirus.html>. Citado na página 93.

GALLAGHER, T. M.; BUCHMEIER, M. J. Coronavirus spike proteins in viral entry and pathogenesis. *Virology*, Academic Press, v. 279, n. 2, p. 371–374, 2001. Citado na página 54.

GAREY, M. R.; JOHNSON, D. S. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. [S.l.: s.n.], 1979. Citado na página 11.

GOLDBARG, M. C.; PACCA, H.; LUNA, L. *Otimização Combinatória e Programação Linear*. 2th. ed. [S.l.]: Editora Campus, 2005. Citado 2 vezes nas páginas 11 e 12.

GONZALEZ, J. et al. A comparative sequence analysis to revise the current taxonomy of the family coronaviridae. *Archives of virology*, Springer, v. 148, n. 11, p. 2207–2235, 2003. Citado na página 52.

GORBALENYA, A. E. et al. Severe acute respiratory syndrome-related coronavirus: The species and its viruses–a statement of the coronavirus study group. BioRxiv, 2020. Citado na página 55.

GRAMMBITTER, G. L. C. et al. An uncommon type ii pks catalyzes biosynthesis of aryl polyene pigments. *Journal of the American Chemical Society*, v. 141, n. 42, p. 16615–16623, 2019. Citado na página 51.

GREANEY, A. J. et al. Comprehensive mapping of mutations to the sars-cov-2 receptor-binding domain that affect recognition by polyclonal human serum antibodies. *bioRxiv*, Cold Spring Harbor Laboratory, 2021. Citado na página 58.

GROOT, R. J. de et al. Commentary: Middle east respiratory syndrome coronavirus (mers-cov): announcement of the coronavirus study group. *Journal of virology*, Am Soc Microbiol, v. 87, n. 14, p. 7790–7792, 2013. Citado na página 54.

GUERRY, P.; HERRMANN, T. Advances in automated nmr protein structure determination. *Quarterly Reviews of Biophysics*, v. 44, n. 3, p. 257?309, 2011. Citado na página 37.

HANKE, L. et al. An alpaca nanobody neutralizes sars-cov-2 by blocking receptor interaction. *Nature Communications*, Nature, v. 11, n. 1, 2020. Citado 2 vezes nas páginas xiv e 76.

HARDIN, C.; POGORELOV, T. V.; LUTHEY-SCHULTENCHOW, Z. Ab initio protein structure prediction. *Current Opinion in Structural Biology*, v. 12, n. 2, p. 176–181, 2002. Citado na página 36.

HELMY, Y. A. et al. The covid-19 pandemic: a comprehensive review of taxonomy, genetics, epidemiology, diagnosis, treatment, and control. *Journal of Clinical Medicine*, Multidisciplinary Digital Publishing Institute, v. 9, n. 4, p. 1225, 2020. Citado na página 53.

HILGENFELD, R. From sars to mers: crystallographic studies on coronaviral proteases enable antiviral drug design. *The FEBS Journal*, 2014. Citado na página 60.

HODCROFT, E. B. et al. Emergence in late 2020 of multiple lineages of sars-cov-2 spike protein variants affecting amino acid position 677. *medRxiv*, Cold Spring Harbor Laboratory Press, 2021. Citado na página 87.

HOLMES, E. C.; HARVEY, P. H.; MAY, R. M. *The evolution and emergence of RNA viruses*. New York: Oxford University Press, 2009. ISBN 978?0?19?921112?8. Citado na página 52.

HU, B. et al. Bat origin of human coronaviruses. *Virology journal*, BioMed Central, v. 12, n. 1, p. 1–10, 2015. Citado na página 53.

HUI, P. et al. Tropism of the novel human betacoronavirus lineage c virus in human ex vivo and in vitro cultures, as an assessment of its potential transmissibility and pathogenesis in humans. In: *Health Research Symposium 2019*. Hong-Kong: [s.n.], 2019. Citado na página 55.

HUMPHREY, W.; DALKE, A.; SCHULTEN, K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, v. 14, n. 1, p. 33 – 38, 1996. ISSN 0263-7855. Citado na página 77.

KEMP, S. et al. Recurrent emergence and transmission of a sars-cov-2 spike deletion h69/v70. *bioRxiv*, Cold Spring Harbor Laboratory, 2021. Citado na página 87.

KERN, D. M. et al. Cryo-em structure of the sars-cov-2 3a ion channel in lipid nanodiscs. *bioRxiv*, Cold Spring Harbor Laboratory, 2020. Citado 2 vezes nas páginas xiv e 76.

KOBAYASHI, K. Complex organic molecules. In: GARGAUD, M. et al. (Ed.). *Encyclopedia of Astrobiology*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 352–352. ISBN 978-3-642-11274-4. Disponível em: <https://doi.org/10.1007/978-3-642-11274-4_337>. Citado na página 31.

KRISSINEL, E.; HENRICK, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D*, v. 60, n. 12, p. 2256–2268, 2004. Disponível em: <https://doi.org/10.1107/S0907444904026460>. Citado na página 77.

KUPFERSCHMIDT, K. Fast-spreading U.K. virus variant raises alarms. *Science*, American Association for the Advancement of Science, v. 371, n. 6524, p. 9–10, 2021. ISSN 0036-8075. Citado na página 57.

LAM, T. T.-Y. et al. Identifying sars-cov-2-related coronaviruses in malayan pangolins. *Nature*, v. 583, n. 7815, p. 282–285, 2020. ISSN 1476-4687. Citado na página 87.

LAVOR, C. et al. The discretizable molecular distance geometry problem. *Comput Optim Appl*, 2011. Citado 4 vezes nas páginas 2, 17, 26 e 37.

LAVOR, C. et al. Recent advances on the discretizable molecular distance geometry problem. *European Journal of Operational Research*, 2011. Citado 2 vezes nas páginas 17 e 27.

LAVOR, e. a. C. Minimal nmr distance information for rigidity of protein graphs. *Discrete Applied Mathematics*, 2018. Citado 2 vezes nas páginas x e 30.

LEVY, Y.; WOLYNES, P. G.; ONUCHIC, J. N. Protein topology determines binding mechanism. *Proceedings of the National Academy of Sciences*, National Academy of Sciences, v. 101, n. 2, p. 511–516, 2004. Disponível em: <https://doi.org/10.1073/pnas.2534828100>. Citado na página 36.

LI, F. Structure, function, and evolution of coronavirus spike proteins. *Annual review of virology*, Annual Reviews, v. 3, p. 237–261, 2016. Citado na página 53.

LI, F. et al. Structure of SARS Coronavirus Spike Receptor-Binding Domain Complexed with Receptor. *Science*, American Association for the Advancement of Science, v. 309, n. 5742, p. 1864–1868, 2005. ISSN 0036-8075. Citado na página 54.

LIBERTI, L.; LAVOR, C.; MACULAN, N. A branch-and-prune algorithm for the molecular distance geometry problem. *International Transactions in Operational Research*, 2007. Citado 2 vezes nas páginas 28 e 29.

LIBERTI, L. et al. Polynomial cases of the discretizable molecular distance geometry problem. 2011. Citado na página 17.

MACULAN, N. et al. The molecular distance geometry problem. *Elavio*, 2010. Citado na página 17.

MAHY, B. W. J.; REGENMORTE, M. H. V. V. Encyclopedia of general virology. In: ELSEVIER (Ed.). *Encyclopedia of General Virology*. Spain: Academic Press Elsevier, 2010. p. 23. ISBN 978-0-12-375146-1. Citado na página 52.

MOTA, K. Sobre potencialidades do estudo em distance geometry (dg). comunicação pessoal. 2013. Citado 2 vezes nas páginas 2 e 3.

MS. *Ministério da Saúde*. 2020. <https://coronavirus.saude.gov.br/>. Citado na página 50.

MUCHERINO, A. On the exact solution of the distance geometry with interval distances in dimension 1. *Springer International Publishing*, 2018. Citado na página 31.

MUCHERINO, A.; LIBERTI, L.; LAVOR, C. Md-jeep: An implementation of a branch and prune algorithm for distance geometry problems. In: FUKUDA, K. et al. (Ed.). *Mathematical Software – ICMS 2010*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. p. 186–197. ISBN 978-3-642-15582-6. Citado 2 vezes nas páginas 66 e 92.

MUNTE, C. E. *Ressonância Magnética Nuclear na Determinação de Estruturas de Proteínas: Aplicação à Mutante HiS15ALA de HPR de Staphylococcus Aureus, e ao Peptídeo-C da Proinsulina Humana*. Tese (Doutorado) — Instituto de Dpisica de São Carlos da Universidade de São Paulo, São Paulo, 2001. Citado na página 40.

NASCIMENTO, V. A. do et al. Genomic and phylogenetic characterisation of an imported case of SARS-CoV-2 in Amazonas State, Brazil. *Memórias do Instituto Oswaldo Cruz*, v. 115, 2020. ISSN 0074-0276. Citado na página 51.

NAVECA, F. et al. Phylogenetic relationship of SARS-CoV-2 sequences from Amazonas with emerging Brazilian variants harboring mutations E484K and N501Y in the Spike protein. Virological, 2021. Citado na página 58.

NELSON, C.; HALL, P.; FREMONT, D. Crystal structure of the sars-cov-2 orf8 protein. Protein Data Bank, 2020. Disponível em: <https://www.rcsb.org/structure/7jx6>. Citado 2 vezes nas páginas xiv e 76.

NELSON, C. et al. Structure of the SARS-CoV-2 ORF7a encoded accessory protein. Protein Data Bank, 2020. Disponível em: <https://www.rcsb.org/structure/6w37>. Citado 2 vezes nas páginas xiv e 76.

NELSON, D. L.; COX, M. M. *Lehninger Principles of Biochemistry*. 7. ed. New York: Macmillan Higher Education, 2017. ISBN 978-1-4641-2611-6. Citado na página 47.

NEWMAN, J. A review of techniques for maximizing diffraction from a protein crystal *in stilla*. *Acta Crystallographica Section D*, v. 62, n. 1, p. 27–31, 2006. Citado na página 37.

OLIVEIRA, A. A. d. et al. Larvicidal, adulticidal and repellent activities against aedes aegypti l. of two commonly used spices, origanum vulgare l. and thymus vulgaris l. *South African Journal of Botany*, v. 140, p. 17–24, 2021. Citado na página 62.

OLIVEIRA, S. R. de. *Orlistate, uma molécula estável*. 2007. <http://sro0.wordpress.com/2007/06/03/orlistate-uma-molecula-estavel/>. Citado na página 1.

OPAS. *Organização Pan-Americana de Saúde*. 2020. Citado na página 4.

PHILLIPS, J. C. et al. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, v. 26, n. 16, p. 1781–1802, 2005. Citado na página 77.

PINHEIRO, A. S. et al. Nmr solution structure of the reduced form of thioredoxin 1 from sacharomyces cerevisiae. *Proteins: Structure, Function, and Bioinformatics*, 2007. Citado 2 vezes nas páginas xii e 72.

PIRES, D. E. V.; ASCHER, D. B.; BLUNDELL, T. L. mcsm: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, v. 30, n. 3, p. 335–342, 11 2013. ISSN 1367-4803. Citado na página 88.

PRIGENT, H. et al. S peech effects of a speaking valve versus external peep in tracheostomized ventilator-dependent neuromuscular patients. *Intensive care medicine*, Springer, v. 36, n. 10, p. 1681–1687, 2010. Citado na página 56.

PROTEíNAS. 2011. <http://biologiadopreuni.blogspot.com.br/2011/03/proteinas.html>. Citado na página 1.

PURSLOW, J. A. et al. NMR Methods for Structural Characterization of Protein-Protein Complexes. *Frontiers in Molecular Biosciences*, v. 7, p. 9, 2020. Citado na página 37.

RAJ, V. S. et al. Dipeptidyl peptidase 4 is a functional receptor for the emerging human coronavirus-emc. *Nature*, Nature Publishing Group, v. 495, n. 7440, p. 251–254, 2013. Citado na página 55.

RESENDE, P. C. et al. Spike E484K mutation in the first SARS-CoV-2 reinfection case confirmed in Brazil, 2020. Virological, 2021. Disponível em: <https://virological.org/t/spike-e484k-mutation-in-the-first-sars-cov-2-reinfection-case-confirmed-in-brazil-2020/584>. Citado na página 58.

RODRIGUES, C. H.; PIRES, D. E.; ASCHER, D. B. Dynamut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. *Protein Science*, v. 30, n. 1, p. 60–69, 2021. Citado 3 vezes nas páginas xiii, 87 e 89.

Schrödinger, LLC. The PyMOL molecular graphics system, version 2.3.0. 2015. Citado na página 75.

SILVA, W.; LAVOR, C.; OCHIAND, L. S. Cálculo de estruturas de proteínas. *Simpósio Brasileiro de Pesquisa Operacional*, 2008. Citado 6 vezes nas páginas x, 2, 15, 18, 29 e 38.

SIT, A. *Solving distance geometry problems for protein structure determination*. Tese (Doutorado) — Iowa State University, 2010. Citado na página 16.

SOARES, C.; HARTMAN, T. *Tracking the vaccine race*. 2020. <https://graphics.reuters.com/HEALTH-CORONAVIRUS/VACCINE-TRACKER/xegpbqnlovq/>. Acesso em: 23 de Dezembro de 2020. Citado na página 58.

SOUZA, C. de et al. Estratégia algorítmica para a reconstrução e validação da estrutura molecular de variantes do sars-cov-2. SBC, Porto Alegre, RS, Brasil, p. 65–72, 2021. ISSN 2763-8774. Disponível em: <https://sol.sbc.org.br/index.php/bresci/article/view/15790>. Citado na página 95.

SOUZA, C. de et al. Comparative analysis among different structural minimization methods based on branch-and-prune structural reconstruction algorithm. Anais do XXI Simpósio Brasileiro de Química Teórica, 2021. Citado na página 95.

SOUZA, C. de et al. Theoretical simulations for the delta variant of sars-cov-2 for understanding the evolution of structural stability. Anais do XXI Simpósio Brasileiro de Química Teórica, 2021. Citado na página 95.

SOUZA, C. de et al. Improvement of sars-cov-2 macromolecule conformation by algorithmic structural prediction. 2021 XLVII Latin American Computing Conference (CLEI), p. 1–9, 2021. Citado na página 95.

SOUZA, C. de; FREITAS, R. de. Algoritmos baseados em programação por restrição para o problema de geometria de distâncias moleculares. XVII Escuela Latino-Iberoamericana de Verano en Investigación Operativa (XVII ELAVIO 2013), 2013. Citado na página 95.

SOUZA, C. de; FREITAS, R. de. Análise do problema de geometria de distâncias moleculares. 2013 XLV Simpósio Brasileiro de Pesquisa Operacional (XLV SBPO), 2013. Citado na página 95.

SOUZA, C. de; FREITAS, R. de. Constraint programming algorithms to determine the position of atoms in a molecule. The International Conference on Operations Research (OR 2013 Conference), 2013. Citado na página 94.

SOUZA, C. de; FREITAS, R. de. Enumeration algorithms for determining the 3d structure of proteins. SEIO 2013 XXXIV Congreso nacional de estadística e investigación operativa. VII Jornadas de Estadística Pública. Libro de abstracts, 2013. Citado na página 95.

SOUZA, C. de; FREITAS, R. de; MOTA, K. Uma proposta de adaptação do algoritmo branch-and-prune usando a interseção de quatro esferas para o problema de geometria de distâncias moleculares. XXXIII Congresso da Sociedade Brasileira de Computação/*VII* e-Science workshop (CSBC/e-Science), 2013. Citado na página 94.

SOUZA, C. de; FREITAS, R. de; SALVATIERRA, M. Sphere intersection algorithms for molecular distance geometry problem. 2013 XXXIX Latin American Computing Conference (CLEI), p. 1–4, 2013. Citado na página 95.

SZWARCFITER, J. L. *Grafos e Algoritmos Computacionais*. [S.l.]: Elsevier, 1986. Citado na página 9.

TALOS-N: Prediction of Protein Backbone and Sidechain Torsion Angles from NMR Chemical Shifts. <https://spin.niddk.nih.gov/bax/software/TALOS-N/>. Citado na página 74.

TALOS-N: Prediction of Protein Backbone and Sidechain Torsion Angles from NMR Chemical Shifts. <https://spin.niddk.nih.gov/bax/nmrserver/talosn/>. Citado na página 74.

VOET, D.; VOET, J. G. *Bioquímica*. [S.l.: s.n.], 2006. Citado na página 32.

VRANKEN, W. F. et al. The ccpn data model for nmr spectroscopy: Development of a software pipeline. *Proteins: Structure, Function, and Bioinformatics*, v. 59, n. 4, p. 687–696, 2005. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.20449>. Citado 2 vezes nas páginas 73 e 74.

WANG, N. et al. Structure of mers-cov spike receptor-binding domain complexed with human receptor dpp4. *Cell research*, Nature Publishing Group, v. 23, n. 8, p. 986–993, 2013. Citado 2 vezes nas páginas xi e 55.

WANG, X. et al. Crystal structure of 2019-nCoV spike receptor-binding domain bound with ACE2. *Nature*, v. 581, n. 7807, p. 215–220, 2020. Disponível em: <https://doi.org/10.1038/s41586-020-2180-5>. Citado 2 vezes nas páginas xi e 55.

WATERHOUSE, A. et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research*, v. 46, n. 1, p. 296–303, 2018. ISSN 0305-1048. Citado 2 vezes nas páginas 76 e 77.

WEISSMAN, D. et al. D614G Spike Mutation Increases SARS CoV-2 Susceptibility to Neutralization. *medRxiv*, Cold Spring Harbor Laboratory Press, 2020. Citado 2 vezes nas páginas 58 e 60.

WHO. *SARS-CoV-2 Variant ? United Kingdom of Great Britain and Northern Ireland*. 2020. <https://www.who.int/csr/don/21-december-2020-sars-cov2-variant-united-kingdom/en/>. Acesso em: 02 de Janeiro de 2020. Citado na página 57.

WHO. *World Health Organization*. 2020. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses>. Acesso em: 05 ago. 2020. Citado na página 56.

WILLIAMS, C. J. et al. Structure of mers-cov spike receptor-binding domain complexed with human receptor dpp4. *Protein science*, Wiley, v. 27, n. 1, p. 293–315, 2018. Citado 2 vezes nas páginas xi e 48.

WOO, P. C. et al. Coronavirus diversity, phylogeny and interspecies jumping. *Experimental Biology and Medicine*, SAGE Publications Sage UK: London, England, v. 234, n. 10, p. 1117–1127, 2009. Citado na página 53.

XU, C. et al. Conformational dynamics of SARS-CoV-2 trimeric spike glycoprotein in complex with receptor ACE2 revealed by cryo-EM. *Science Advances*, American Association for the Advancement of Science, v. 7, n. 1, 2021. Citado 2 vezes nas páginas xiv e 76.

XU, H. et al. High expression of ace2 receptor of 2019-ncov on the epithelial cells of oral mucosa. *International journal of oral science*, Nature Publishing Group, v. 12, n. 1, p. 1–5, 2020. Citado na página 56.

XUPING, X. et al. Neutralization of sars-cov-2 spike 69/70 deletion, e484k and n501y variants by bnt162b2 vaccine-elicited sera. *Nature Medicine*, 11 2021. Citado 2 vezes nas páginas 57 e 60.

ZHANG, L. et al. Crystal structure of sars-cov-2 main protease provides a basis for design of improved alfa-ketoamide inhibitors. *Science*, 2020. Citado 4 vezes nas páginas xi, 50, 60 e 61.

ZHANG, Y.; SKOLNICK, J. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic Acids Research*, v. 33, n. 7, p. 2302–2309, 2005. ISSN 0305-1048. Citado na página 77.

ZIEBUHR, J. The coronavirus replicase. In: ENJUANES, L. (Ed.). *Coronavirus Replication and Reverse Genetics*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005. p. 57–94. ISBN 978-3-540-26765-2. Disponível em: <https://doi.org/10.1007/3-540-26765-4_3>. Citado na página 54.

ZIELECKI, F. et al. Human cell tropism and innate immune system interactions of human respiratory coronavirus emc compared to those of severe acute respiratory syndrome coronavirus. *Journal of virology*, Am Soc Microbiol, v. 87, n. 9, p. 5300–5304, 2013. Citado na página 54.

ZUMLA, A.; HUI, D. S.; PERLMAN, S. Middle east respiratory syndrome. *The Lancet*, Elsevier, v. 386, n. 9997, p. 995–1007, 2015. Citado na página 54.

Apêndice A

Tabela de deslocamento

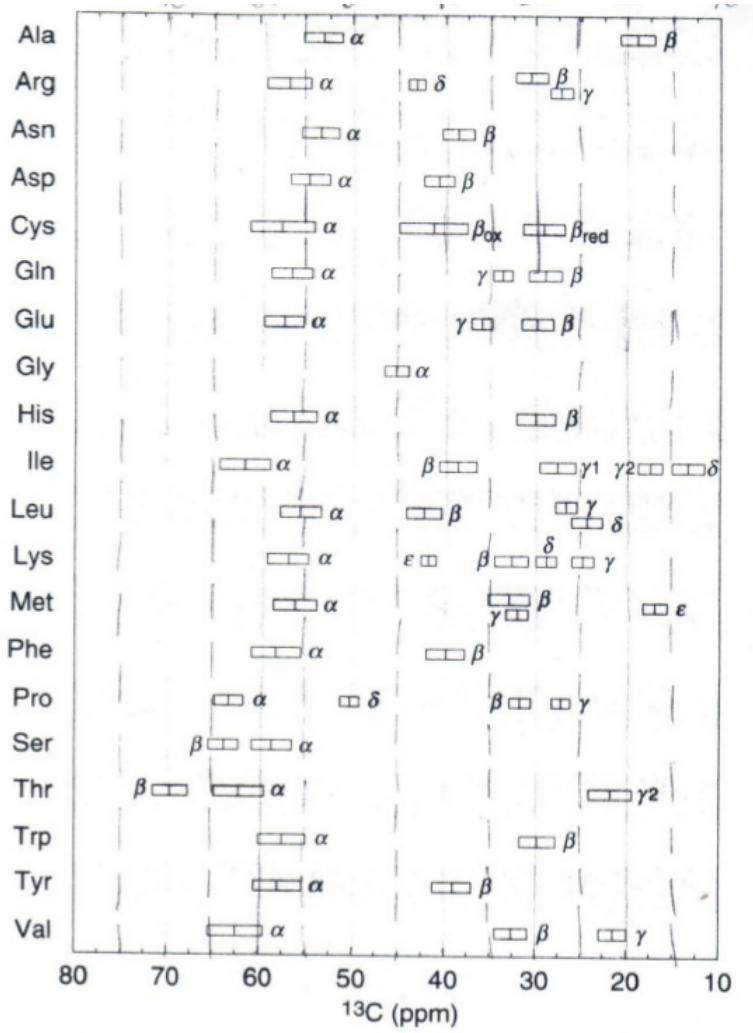


Figura A.1: Tabela de Deslocamento - fornecida no CNRMN.

Apêndice B

Assinalamento CCPNmr Analysis

B.1 Assinalamento do backbone com tripla ressonância

Assinalamento do backbone com tripla ressonância é baseada nos espectros CBCANNH e CBCA(CO)NHH. A ideia é que o CBCANNH correlaciona cada grupo NH, com os deslocamentos químicos C_α e C_β do seu próprio resíduo (fortemente) e do resíduo anterior (fracamente). O CBCA(CO)NNH apenas correlaciona o grupo NH aos deslocamentos químicos C_α e C_β do resíduo anterior. O espectro pode ser usado para ligar um grupo NH para o próximo em uma cadeia longa. Como mostrado nas Figura B.1, B.2 e B.3.

B.2 Assinalamento Cadeia Lateral

Um método fácil é começar com um conjunto de espectros HBHA(CO)ONH, HCC(CO)NNH e CC(CO)NNH. Estes experimentos fornecem os deslocamentos químicos do hidrogênio e do carbono da cadeia lateral do resíduo anterior a cada grupo NH. Como mostrado nas Figuras B.4, B.5 e B.6.

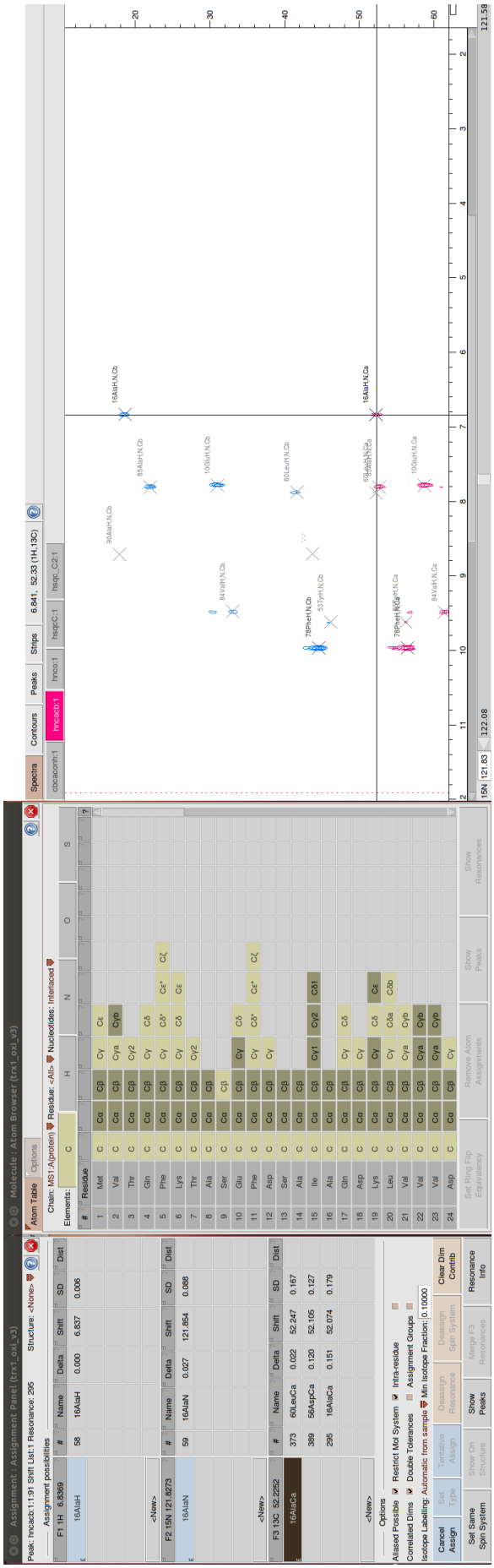


Figura B.1: Assinalamento C_α e C_β.

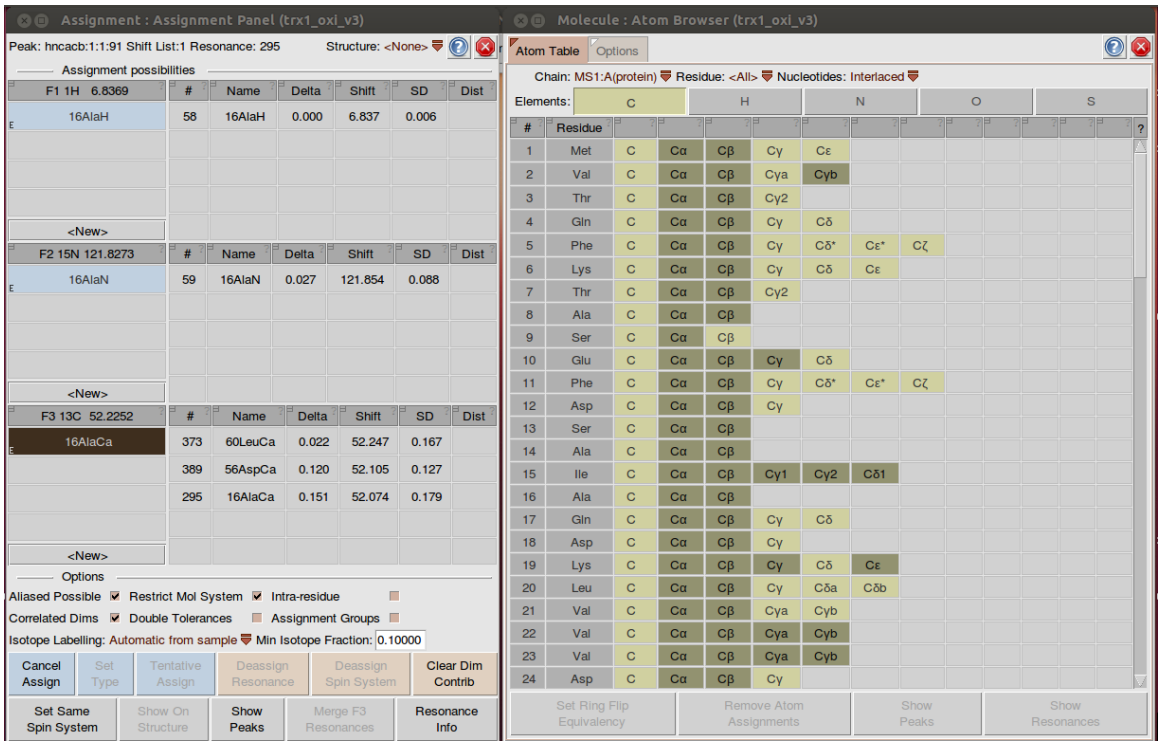


Figura B.2: Assinalamento C_α e C_β.

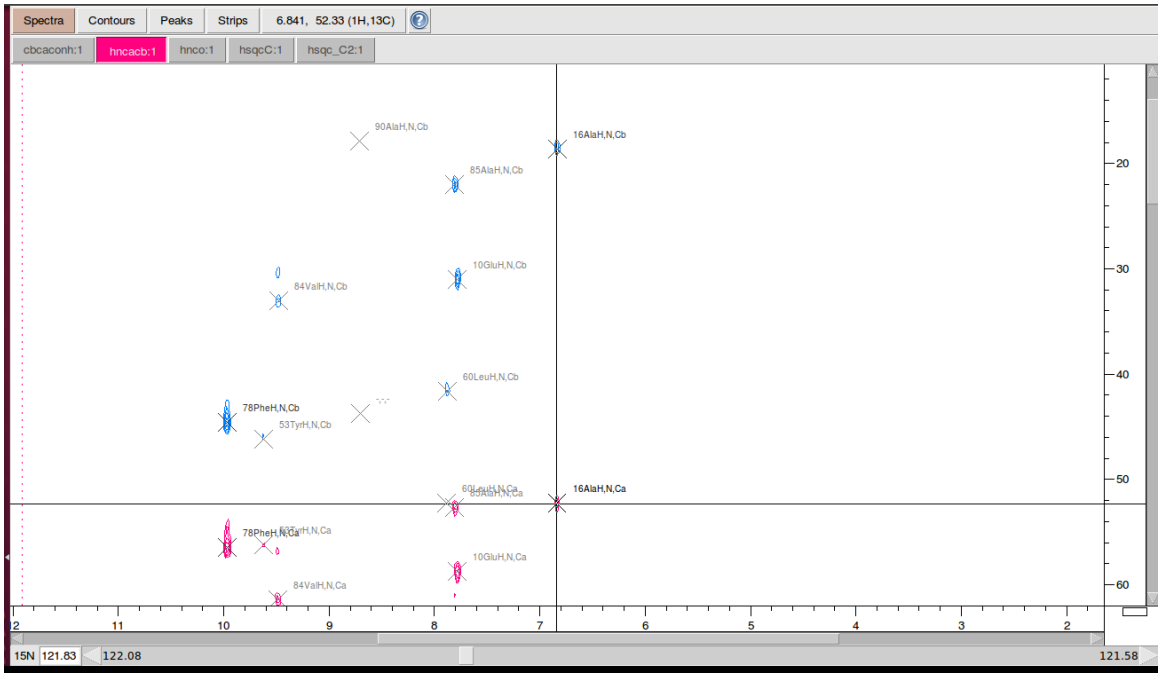


Figura B.3: Assinalamento C_α e C_β.

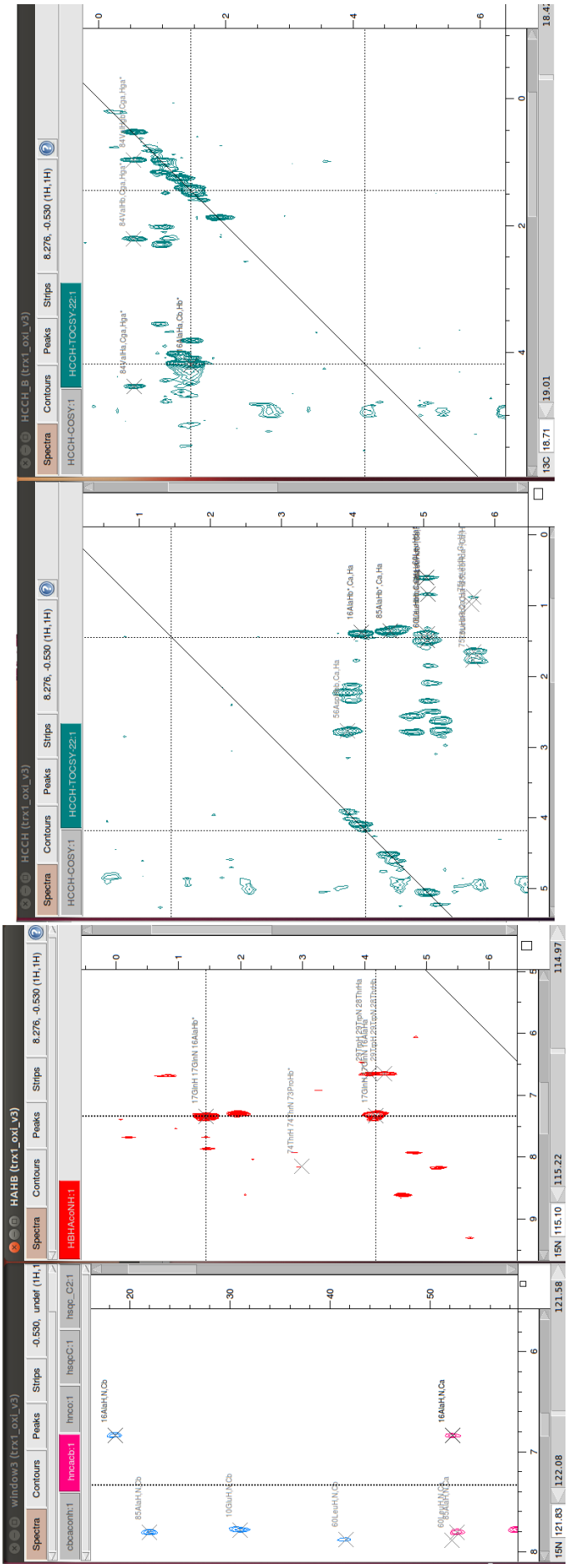


Figura B.4: Assinalamento da cadeia lateral.

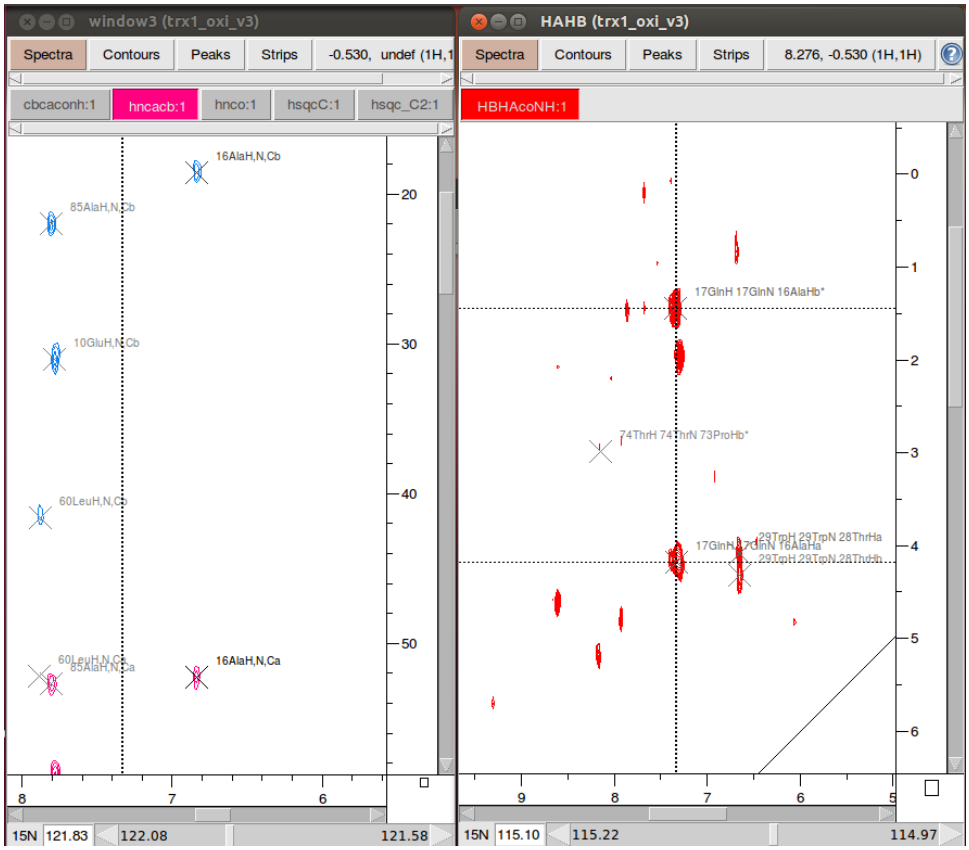


Figura B.5: Assinalamento da cadeia lateral.

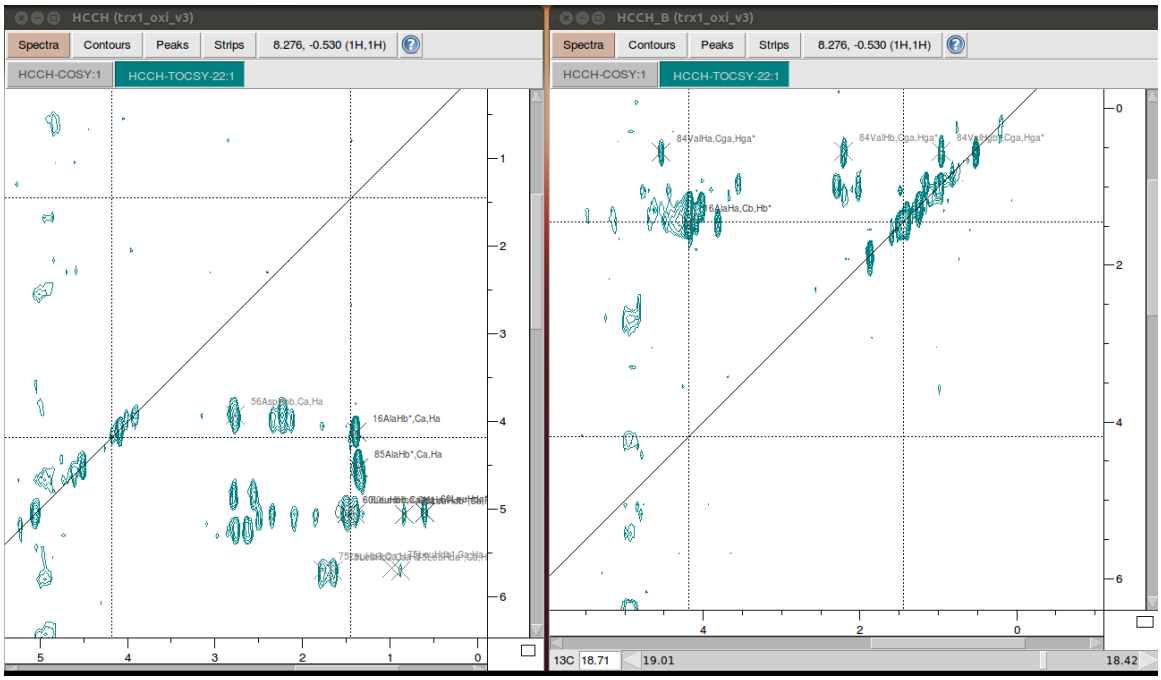


Figura B.6: Assinalamento da cadeia lateral.