CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

DESENVOLVIMENTO DE MATRIZES DE DISTÂNCIAS PARA REPRESENTAR INTERAÇÕES DE PROTEÍNAS COMBINADAS COM ALGORITMOS DE AGRUPAMENTO

OTAVIANO MARTINS MONTEIRO

Orientador: Thiago de Souza Rodrigues
Centro Federal de Educação Tecnológica de Minas Gerais

Coorientador: Sandro Renato Dias
Centro Federal de Educação Tecnológica de Minas Gerais

BELO HORIZONTE
AGOSTO DE 2023

OTAVIANO MARTINS MONTEIRO

DESENVOLVIMENTO DE MATRIZES DE DISTÂNCIAS PARA REPRESENTAR INTERAÇÕES DE PROTEÍNAS COMBINADAS COM ALGORITMOS DE AGRUPAMENTO

Tese apresentada ao Programa de Pós-graduação em Modelagem Matemática e Computacional do Centro Federal de Educação Tecnológica de Minas Gerais, como requisito parcial para a obtenção do título de Doutor em Modelagem Matemática e Computacional.

Área de concentração: Modelagem Matemática e Computacional

Linha de pesquisa: Métodos Matemáticos Aplicados

Orientador: Prof. Dr. Thiago de Souza Rodrigues
Centro Federal de Educação Tecnológica de Minas Gerais

Coorientador: Prof. Dr. Sandro Renato Dias
Centro Federal de Educação Tecnológica de Minas Gerais

CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS

MODELAGEM MATEMÁTICA E COMPUTACIONAL

BELO HORIZONTE

AGOSTO DE 2023

Monteiro, Otaviano Martins.
M775d Desenvolvimento de matrizes de distâncias para representar interações de proteínas combinadas com algoritmos de agrupamento / Otaviano Martins Monteiro. – 2023.
122 f. : il.

Tese apresentada ao Programa de Pós-Graduação em Modelagem Matemática e Computacional.
Orientador: Thiago de Souza Rodrigues.
Coorientador: Sandro Renato Dias.
Inclui referências.
Tese (doutorado) – Centro Federal de Educação Tecnológica de Minas Gerais.

1. Análise por agupamento. 2. Matrizes de distâncias. 3. Proteínas I. Rodrigues, Thiago de Souza. II. Dias, Sandro Renato. III. Centro Federal de Educação Tecnológica de Minas Gerais. IV. Título.

CDD: 519.53

Elaboração da ficha catalográfica pelo Bibliotecário Jeferson N. Auto de Oliveira, CRB-6 2825 / BIB-NG / CEFET-MG

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
COORDENAÇÃO DO PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

ATA DA 60ª sessão pública DE APRESENTAÇÃO E DEFESA DE TESE DE DOUTORADO, REQUISITO PARA A OBTENÇÃO DO TÍTULO DE DOUTOR EM MODELAGEM MATEMÁTICA E COMPUTACIONAL. Em 24 de agosto de 2023, através de videoconferência, Campus II, CEFET-MG, reuniu-se, às 14 horas, a Banca Examinadora, designada para este fim, pelo Colegiado do Programa de Pós-Graduação Stricto Sensu em Modelagem Matemática e Computacional, constituída por: Dr. Thiago de Souza Rodrigues (Orientador), Dr. Sandro Renato Dias (Coorientador), Dr. Rogério Martins Gomes, Dr. André Rodrigues da Cruz, Dr. Gustavo Campos Menezes e Dr. Alisson Marques da Silva, para examinar o trabalho do doutorando **Otaviano Martins Monteiro,** sob o título *"DESENVOLVIMENTO DE MATRIZES DE DISTÂNCIAS PARA REPRESENTAR INTERAÇÕES DE PROTEÍNAS COMBINADAS COM ALGORITMOS DE CLUSTERING".* O Prof. Dr. Thiago de Souza Rodrigues, como Presidente da Banca Examinadora, declarou aberta a sessão, passando a palavra ao candidato, para que este expusesse sua Tese de doutorado. Terminada a exposição, o Presidente passou a palavra aos membros da Banca Examinadora, que iniciaram a arguição, na seguinte ordem: Dr. Rogério Martins Gomes, Dr. André Rodrigues da Cruz, Dr. Gustavo Campos Menezes, Dr. Alisson Marques da Silva, Dr. Sandro Renato Dias e Dr. Thiago de Souza Rodrigues. Terminada a arguição, reuniu-se a Banca Examinadora, isoladamente, para deliberação. Terminada a reunião, o Presidente deu conhecimento ao candidato de que sua Tese foi aprovada, devendo a redação final da Tese de Doutorado incorporar as contribuições da Banca Examinadora. A Banca definiu ainda que a entrega dos exemplares com a redação final deverá ser feita em, no máximo, 90 (noventa) dias. Nada mais havendo a tratar, o Presidente declarou encerrada a sessão, cujas atividades são registradas nesta Ata, lavrada, a qual assina, juntamente com os demais membros da Banca Examinadora. Belo Horizonte, 24 de agosto de 2023.

Prof. Dr. Thiago de Souza Rodrigues (Orientador)
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Sandro Renato Dias (Coorientador)
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Rogério Martins Gomes
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. André Rodrigues da Cruz
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Gustavo Campos Menezes
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Alisson Marques da Silva
Centro Federal de Educação Tecnológica de Minas Gerais

Emitido em 28/08/2023

ATA DE DEFESA DE TESE Nº 54/2023 - DECOM (11.56.03)

(Nº do Protocolo: NÃO PROTOCOLADO)

(Assinado digitalmente em 29/08/2023 14:04)
ALISSON MARQUES DA SILVA
PROFESSOR ENS BASICO TECN TECNOLOGICO
CTINFDV (11.50.29)
Matrícula: ###529#8

(Assinado digitalmente em 28/08/2023 11:55)
ANDRE RODRIGUES DA CRUZ
PROFESSOR ENS BASICO TECN TECNOLOGICO
DECOM (11.56.03)
Matrícula: ###837#6

(Assinado digitalmente em 28/08/2023 11:30)
GUSTAVO CAMPOS MENEZES
PROFESSOR ENS BASICO TECN TECNOLOGICO
DCCN (11.58)
Matrícula: ###504#1

(Assinado digitalmente em 28/08/2023 11:44)
ROGERIO MARTINS GOMES
PROFESSOR ENS BASICO TECN TECNOLOGICO
DECOM (11.56.03)
Matrícula: ###66#4

(Assinado digitalmente em 28/08/2023 11:26)
SANDRO RENATO DIAS
PROFESSOR ENS BASICO TECN TECNOLOGICO
DDE (11.48)
Matrícula: ###444#8

(Assinado digitalmente em 28/08/2023 10:56)
THIAGO DE SOUZA RODRIGUES
PROFESSOR DO MAGISTERIO SUPERIOR
DECOM (11.56.03)
Matrícula: ###518#3

Agradecimentos

Primeiramente agradeço a Deus pelo dom da vida, pela sabedoria e pelo conforto nos momentos difíceis.

Agradeço aos meus orientadores Thiago de Souza Rodrigues e Sandro Renato Dias pelas orientações e ensinamentos que recebi.

Ao Rogério Martins Gomes pelas contribuições na reta final desta tese de doutorado.

Ao Higor Coimbra Amorim, que eu tive a oportunidade de coorientá-lo no TCC. Ajudá-lo nesta etapa também contribuiu para o meu aprendizado.

A todos os alunos e professores do MMC por todo o apoio que recebi.

A todos os meus amigos e familiares, em especial a minha esposa Ana Flávia Gonçalves Rocha Monteiro, minha filha Mariana Rocha Monteiro, meu irmão Cristiano Martins Monteiro, minha mãe Creuzoni Martins Monteiro, meu pai Geraldo Monteiro da Silva *(in memoriam)*, meu padrasto Adelmo Aparecido Gomes, meu sogro Geraldo Lúcio Rocha e minha sogra Cimira Zoé Gonçalves Rocha.

Ao Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) pelo suporte oferecido e pela bolsa de estudos utilizada no início desta pesquisa no mestrado (4 meses de bolsa de estudos) e nos dois primeiros anos do doutorado.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo financiamento desta pesquisa, que incluiu 20 meses de bolsa de estudos recebidas no mestrado, além de 30 meses de bolsa de estudos adquiridas no doutorado.

Resumo

As proteínas são macromoléculas formadas por aminoácidos e estão presentes em todos os seres vivos. Várias proteínas tiveram suas estruturas tridimensionais resolvidas experimentalmente e foram armazenadas através de arquivos de texto em bancos de dados biológicos como o *Protein Data Bank* (PDB). Essas informações proteicas podem ser utilizadas por *softwares*, como o LSQKAB, que verificam similaridades tridimensionais de proteínas através de sobreposições entre os átomos das estruturas comparadas. No entanto, a realização de sobreposições atômicas requer um alinhamento preciso entre os átomos de duas estruturas por meio de movimentos de rotação e translação. Esse procedimento é computacionalmente intensivo, sendo classificado como NP-Completo. Portanto, a realização de múltiplas sobreposições atômicas, algo frequente em *softwares* que propõem mutações em proteínas, acarreta em um elevado custo computacional. Assim sendo, o propósito deste estudo consiste em elaborar abordagens fundamentadas em matrizes de distâncias, combinadas com algoritmos agrupamento (*clustering*) com o intuito de criar conjuntos de interações de proteínas que compartilham conformações tridimensionais semelhantes. O objetivo principal é alcançar soluções de alta precisão e desempenho notável, com o propósito de minimizar a necessidade de realizar sobreposições atômicas. Com o intuito de cumprir esses objetivos, foram desenvolvidas matrizes de distâncias baseadas em diferentes abordagens. A Matriz de Ângulos (MA) foi desenvolvida a partir dos ângulos dos átomos. A Matriz de Distâncias Completa Mista (MDCM) foi desenvolvida através da fusão de diferentes técnicas. A Matriz de Distâncias Reduzida cujos Centroides são Carbonos Alfa (MDRCCA), a Matriz de Distâncias Reduzida a partir de um Ponto entre os Carbonos Alfa (MDRPCA), além da Matriz de Pontos Médios (MPM) foram desenvolvidas a partir da importância dos átomos de carbonos alfa (CA). A concepção da MPM também foi influenciada pela importância das distâncias entre todos os átomos na estrutura, uma vez que essas distâncias são cruciais para o enovelamento da mesma. Essas estratégias foram integradas a algoritmos de agrupamento e os resultados subsequentes foram comparados com o método de busca da ferramenta RID, por ser uma ferramenta especialista em trabalhar com interações de proteínas, além da *atomic Cutoff Scanning Matrix* (aCSM) por ser uma das versões da *Cutoff Scanning Matrix* (CSM), que é considerada o estado da arte na geração de assinaturas em grafos proteicos, e com a Matriz de Distâncias Completa (MDC), que apresentou resultados superiores ao método de busca da RID e aCSM, nos primeiros trabalhos deste projeto. Os resultados foram satisfatórios, principalmente os alcançados pela MPM, que superou as demais técnicas na maioria dos experimentos.

Palavras-chave: Interações proteicas. Matriz de distâncias. *Clustering*. Matriz de Pontos Médios (MPM). Matrizes de Distâncias Reduzidas. Matriz de Distâncias Completa Mista.

Abstract

Proteins are macromolecules formed by amino acids and are present in all living beings. Several proteins had their three-dimensional structures experimentally resolved and were stored through text files in biological databases such as the Protein Data Bank (PDB). These proteins information can be used by softwares, such as LSQKAB, which verify three-dimensional similarities of proteins through superimpositions between the atoms of compared structures. However, to perform atomic superpositions requires a precise alignment between the atoms of two structures through rotation and translation movements. This procedure is computationally intensive, being classified as NP-Complete. In this way, carrying out multiple atomic overlaps, something frequent in software that propose mutations in proteins, entails a high computational cost. Therefore, the purpose of this study is to develop approaches based on distance matrices, combined with clustering algorithms in order to create groups of protein interactions that share similar three-dimensional conformations. The main objective is to achieve solutions of high precision and remarkable performance, with the purpose of minimizing the need to carry out atomic superimpositions. In order to fulfill these objectives, distance matrices based on different approaches were developed. The Angle Matrix (MA) was developed from the angles of atoms. The Mixed Complete Distance Matrix (MDCM) was developed by merging different techniques. The Reduced Distance Matrix whose Centroids are Alpha Carbons (MDRCCA), the Reduced Distance Matrix from a Point Between Alpha Carbons (MDRPCA) and the Midpoint Matrix (MPM) were developed from the importance of alpha carbons atoms (CA). The design of MPM was also influenced by the importance of the distances between all atoms in the structure, because these distances are crucial for its folding. These strategies were integrated to clustering algorithms and the subsequent results were compared with the RID search method, because it is a specialist tool in working with protein interactions, atomic Cutoff Scanning Matrix (aCSM), as it is one of the versions of the Cutoff Scanning Matrix (CSM), which is considered the state of the art in the generation of signatures in protein graphs, and with the Complete Distance Matrix (MDC), which presented superior results to RID search method and aCSM, in the first works of this project. The results were satisfactory, mainly those achieved by MPM, which outperformed the other techniques in most experiments.

Keywords: Protein Interactions. Distance matrix. clustering. Midpoints Matrix (MPM). Reduced Distance Matrices. Mixed Complete Distance Matrix (MDCM).

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Lista de Abreviaturas e Siglas

aCSM	*atomic Cutoff Scanning Matrix*
ANOVA	Análise de Variância
CSM	*Cutoff Scanning Matrix*
GB	*Gigabyte*
MA	Matriz de Ângulos
MDC	Matriz de Distâncias Completa
MDCM	Matriz de Distância Completa Mista
MDRCCA	Matriz de Distâncias Reduzida cujos Centroides são Carbonos Alfa
MDRPCA	Matriz de Distâncias Reduzida construída à partir de um Ponto entre os Carbonos Alfas
MPM	Matriz de Pontos Médios
Proteus	*Protein Engineering Supporter*
RID	*Residue Interaction Database*
RMSD	*Root-Mean-Square Deviation*
SSV	*Structural Signature Variation*
TB	*Terabyte*

Lista de Símbolos

Å	Angstrom
ϵ	Epsilon
$\sum$	Somatório
$\forall$	Para todo
ψ	*Psi*
ϕ	*Phi*

Sumário

Capítulo 1

Introdução

As proteínas são macromoléculas biológicas formadas por um conjunto de aminoácidos distintos, ligados em sequências lineares. Essas macromoléculas estão presentes em todos os seres vivos e desempenham diversas funções como catálise, regulação e função imune (NELSON; COX, 2014). Devido à abundância e importância das proteínas, o estudo dessas macromoléculas pode beneficiar diversos segmentos. Dentre as áreas beneficiadas estão a saúde, através do desenvolvimento de fármacos e vacinas, o meio-ambiente, no aprimoramento de enzimas para a degradação de contaminantes, além do segmento industrial, com o aprimoramento de enzimas digestivas que são utilizadas em vários processos (DIAS, 2012).

Uma das formas de beneficiar diversos segmentos através das proteínas consiste em propor mutações entre um ou mais aminoácidos dessas estruturas. O objetivo das propostas de mutações pode ser de aumentar ou diminuir a estabilidade, bem como a flexibilidade da proteína, podendo envolver a manutenção da função proteica (DIAS, 2012). De acordo com Hunter (1993), dentre as forças que estabilizam as proteínas estão as pontes dissulfeto, que são ligações covalentes formadas pelos átomos de enxofre das cisteínas, que após oxidarem, tornam-se cistinas. Essas ligações são responsáveis por manter a estabilidade conformacional (estrutura tridimensional) de uma proteína, ligando partes distantes de uma cadeia polipeptídica ou de diferentes cadeias (DIAS, 2012). Gromiha e Selvaraj (2004) e Pace et al. (2011) complementam que as pontes dissulfeto, juntamente com pontes de hidrogênio, pares iônicos e interações de *van der Waals*, definem a forma e mantém a estrutura da proteína estável.

Diversas estruturas de proteínas, bem como pontes dissulfeto e ligações de hidrogênio, foram estudadas por vários pesquisadores e armazenadas em banco de dados biológicos, como o *Protein Data Bank* (PDB). Este banco de dados armazena por meio de arquivos de texto, informações sobre as estruturas tridimensionais das proteínas, como distâncias atômicas e interações (BERMAN et al., 2000). A partir das informações armazenadas nos

arquivos do PDB pode-se comparar as conformações tridimensionais das proteínas através de sobreposições atômicas realizadas após o alinhamento estrutural.

Dentre os *softwares* que realizam os procedimentos de alinhamento estrutural e sobreposições atômicas, com informações proteicas no formato do PDB, está o LSQKAB (WINN et al., 2011). Esse *software* é baseado no algoritmo de Kabsch (1976) e aplica várias transformações de coordenadas, visando o alinhamento estrutural, através do ajuste de um subconjunto de átomos em relação ao subconjunto correspondente da outra estrutura comparada. As transformações de coordenadas realizadas pelo LSQKAB possibilitam o cálculo da matriz de rotação ideal entre duas estruturas de proteínas, possibilitando a realização de sobreposições atômicas, minimizando o *Root-Mean-Square Deviation* (RMSD). Este *software* também retorna os resultados das sobreposições atômicas através de uma lista com todas as diferenças de distâncias dos pares de átomos das duas estruturas comparadas (deslocamento), em arquivos de texto chamados de deltas (CCP4, 2011).

As sobreposições atômicas são fundamentais para comparar o enovelamento das estruturas proteicas, além de ser uma das etapas utilizadas por *softwares* que realizam proposições de mutações de proteínas. No entanto, conforme Crescenzi et al. (1998), Oliveira (2016) e Melo-Minardi (2021), o problema do alinhamento estrutural para realizar a sobreposição atômica não é simples de ser resolvido, assim pertencendo à classe NP-Completo. A maior dificuldade ocorre pelas proteínas serem frequentemente resolvidas em coordenadas tridimensionais diferentes, tornando necessário a realização de vários movimentos de rotação e translação entre as duas estruturas comparadas, para encontrar o melhor ajuste entre ambas (MELO-MINARDI, 2021). Deste modo, de acordo com Dias (2012), utilizar o LSQKAB para trabalhar com todos os registros de uma base de dados é uma tarefa demorada computacionalmente, devido aos diversos cálculos que essa ferramenta realiza.

Dias (2012) propôs uma alternativa para comparar diversas estruturas de proteínas em um tempo viável, através da ferramenta *Residue Interaction Database* (RID), trabalhando juntamente com o LSQKAB. A estratégia consistiu em escolher um registro para ser a referência. Em seguida, o LSQKAB foi utilizado para realizar sobreposições atômicas contra os demais registros desta mesma base de dados, assim gerando um arquivo de deltas para cada comparação. Dias (2012) ainda criou um valor de *score* calculando a distância Euclidiana a partir de cada par de arquivo delta gerado, para otimizar a busca de pares similares. Os resultados obtidos foram bons, mas não ótimos, pois arquivos com valores de distâncias atômicas diferentes podem obter o mesmo valor de *score*.

Existem *softwares* que trabalham com bases de dados de arquivos no formato do PDB a partir de outras estratégias. Dentre estes exemplos estão a *Cutoff Scanning Matrix* (CSM) e suas versões, como a *atomic Cutoff Scanning Matrix* (aCSM). Estes algoritmos são considerados o estado da arte na classificação de estruturas proteicas em larga escala. De

acordo com Pires et al. (2011), a CSM gera assinaturas para grafos biológicos, com base no padrão de distâncias dos átomos de uma estrutura proteica. As assinaturas geradas são utilizadas em tarefas de classificação. A aCSM é uma versão atômica da CSM e suporta assinatura de *pockets* baseada em grafos (PIRES et al., 2013).

A partir de uma das etapas da CSM, foi desenvolvida uma metodologia baseada em matriz de distâncias completa (MDC), visando o aperfeiçoamento do método de busca da ferramenta RID. O *software* desenvolvido possibilita que interações de proteínas no formato do PDB sejam representadas através de uma matriz de distâncias, construída pelo cálculo da distância Euclidiana entre todos os átomos da interação, e posteriormente sejam agrupadas por técnicas de *clustering*. A precisão da MDC foi superior à aCSM e à RID nos primeiros cenários comparados, além de ser mais rápida (MONTEIRO, 2017) (MONTEIRO; DIAS; RODRIGUES, 2020).

Embora a MDC tenha gerado resultados interessantes, há a convicção de que é viável elaborar novas abordagens para representar interações de proteínas, visando a obtenção de acurácias ainda maiores e a redução dos tempos de processamento. É igualmente válido explorar a combinação dessas inovações com diversos algoritmos de agrupamento, além de avaliá-las em relação a novos conjuntos de dados. É importante mencionar que nas etapas iniciais, o *k-means* foi o único algoritmo de agrupamento empregado, e apenas uma base de dados de pontes dissulfeto foi utilizada.

Portanto, o escopo do presente estudo engloba a formulação de novas metodologias ancoradas em matrizes de distâncias, com o objetivo de promover aprimoramentos tanto na acurácia quanto no desempenho computacional. Além disso, pretende-se avaliar essas abordagens em conjunto com diferentes técnicas de agrupamento, abrangendo variadas bases de dados. As abordagens mais promissoras desenvolvidas nesse estudo podem ser posteriormente integradas à ferramenta RID, com o propósito de aprimorar seu método de busca.

As estratégias para construir matrizes de distâncias neste estudo foram concebidas a partir de abordagens distintas. Através da análise dos ângulos dos átomos nas interações, surgiu a Matriz de Ângulos (MA). Utilizando a tática de mesclar horizontalmente diferentes técnicas de representação de proteínas, originou-se a Matriz de Distâncias Completa Mista (MDCM). Em consideração à importância dos átomos de carbono alfa (CA) – situados no núcleo das cadeias principais e fundamentais para o dobramento proteico, dado que as cadeias laterais derivam deles –, desenvolveram-se a Matriz de Distâncias Reduzida cujos Centroides são Carbono Alfa (MDRCCA), a Matriz de Distâncias Reduzida a partir de um Ponto entre os Carbonos Alfa (MDRPCA) e, além disso, a Matriz de Pontos Médios (MPM). A funcionalidade desta última matriz de distâncias (MPM) foi inspirada em dois elementos: a relevância dos átomos de CA e a influência das distâncias entre todos os átomos no

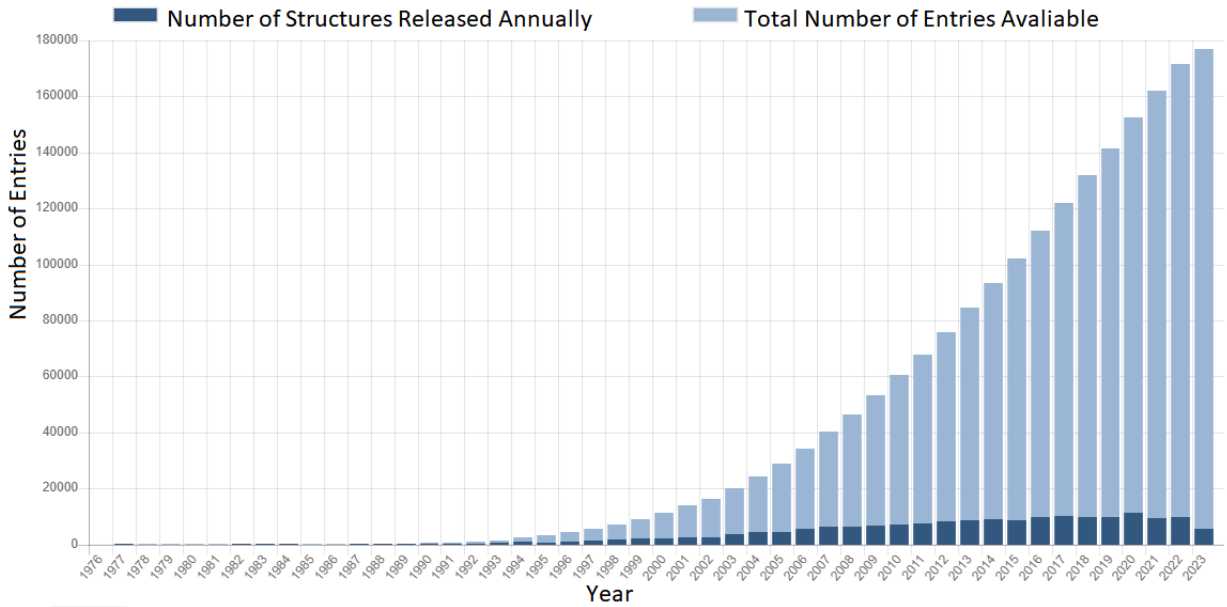
processo de dobramento proteico (similar ao conceito empregado na MDC).

Essas abordagens foram então combinadas com três algoritmos de agrupamento que possuem diferenças consideráveis em seus funcionamentos: *k-means clustering*, *hierarchical clustering* e OPTICS (PEDREGOSA et al., 2011). Os algoritmos de agrupamento utilizados foram obtidos pela biblioteca *scikit-learn* da linguagem de programação Python. Os experimentos foram executados utilizando três bases de dados distintas: a primeira contendo informações sobre pontes dissulfeto, a segunda focada em ligações de hidrogênio e a terceira resultante da união dessas duas bases de dados. Os resultados obtidos foram satisfatórios, com destaque notável para a MPM, que superou todas as outras técnicas em diversos cenários de testes. Esses experimentos consistiram em realizar sobreposições atômicas entre as interações que ficaram no mesmo agrupamento e avaliar o RMSD e o maior deslocamento de cada sobreposição realizada.

1.1 Justificativa

A cada ano que passa, os bancos de dados biológicos (como o PDB), estão com mais estruturas proteicas armazenadas. Dentre essas estruturas estão as que foram obtidas através de experimentos de cristalografia de raios-X. Conforme RSCB (2023), a cristalografia de raios-x é um método pra determinar a localização de cada átomo em relação aos outros átomos na molécula. A Figura 1 ilustra o crescimento anual do armazenamento de estruturas obtidas por cristalografia de raios-X no PDB.

Figura 1 – Estatísticas do PDB: a figura publicada por RSCB (2023) ilustra o crescimento de estruturas de experimentos de cristalografia de raios-X lançados por ano



Fonte: Adaptada de RSCB (2023)

De acordo com a Figura 1, a ilustração de cor azul escuro no gráfico indica a quantidade

de estruturas de experimentos de cristalografia de raio-X inseridas no ano. A ilustração na cor azul claro indica a soma acumulada de estruturas armazenadas no decorrer dos anos. Deste modo, nota-se que a quantidade de armazenamento anual tem oscilações, mas com uma tendência de aumento, principalmente nos últimos anos. É importante ressaltar que os valores do ano de 2023 estão menores do que os dos últimos anos por ser o ano atual, no qual ainda acontecerão novas inserções de estruturas de proteínas (esse gráfico foi gerado em 01 de agosto de 2023). Esse aumento de inserções de proteínas é benéfico para a área de bioinformática, inclusive para *softwares* que realizam proposições de mutações proteicas, pois a cada dia que passa pode-se realizar novas propostas de mutações, a partir das novas estruturas cadastradas. No entanto, esse constante aumento de dados biológicos contribui para a necessidade de desenvolver novas estratégias que possibilitem trabalhar com esses dados de um modo eficiente e rápido.

Os dados armazenados no PDB desempenham um papel fundamental em diversas aplicações, como a identificação de mutações específicas, a comparação tridimensional de estruturas proteicas, entre outras tarefas. Para realizar essas operações, é comum recorrer ao alinhamento estrutural para realizar sobreposições atômicas. No entanto, esses procedimentos dependem da realização de movimentos de translação e rotação dos átomos de uma proteína (ou de um subconjunto, como uma interação) em relação a outra estrutura que está sendo comparada. Como as estruturas proteicas são resolvidas em coordenadas tridimensionais distintas, esse desafio se enquadra na categoria de problemas NP-Completo, tornando-se uma tarefa computacionalmente custosa (CRESCENZI et al., 1998), (OLIVEIRA, 2016) e (MELO-MINARDI, 2021).

A complexidade de processamento se amplifica especialmente quando lidamos com conjuntos de dados extensos e quando a busca por grupos de estruturas com conformações tridimensionais semelhantes também é necessária. Deste modo, é necessário desenvolver metodologias que ofereçam representações precisas e eficientes das coordenadas atômicas nas interações proteicas. O objetivo é minimizar a dependência de múltiplos alinhamentos estruturais e sobreposições atômicas. Diante deste cenário, é crucial explorar estratégias fundamentadas em matrizes de distâncias para expressar as interações entre proteínas. A utilização desse tipo de estrutura de dados confere a vantagem de manter as interações direcionadas no espaço tridimensional, além de facilitar a organização dos dados em grupos coesos. Isso representa uma alternativa viável para a comparação simultânea de diversas interações proteicas, de forma mais ágil, uma vez que dispensa a necessidade de alinhamentos estruturais e sobreposições atômicas.

1.2 Motivação

Softwares de proposição de mutações sítio dirigidas podem ajudar especialistas da área a analisar e propor mutações entre estas estruturas, podendo trazer melhorias para estas macromoléculas. O desenvolvimento de técnicas que possam aprimorar a precisão destas ferramentas, bem como a performance computacional são importantes.

De acordo com Dias (2012) e Barroso et al. (2020), o estudo das proteínas, bem como suas melhorias à partir de mutações, podem beneficiar diversos segmentos. Dentre os exemplos estão a saúde, no desenvolvimento de fármacos e vacinas. Além da melhora das enzimas, usadas nos processos digestivos para a produção de novos alimentos e biocombustíveis.

A primeira base de dados utilizada neste trabalho é de pontes dissulfeto. Existem estudos que provam que a estabilidade de uma proteína pode aumentar através da introdução de uma nova ponte dissulfeto em uma região distinta desta mesma proteína (PANTOLIANO et al., 1989) (PANTOLIANO et al., 1988). De acordo com Almog et al. (2002), através do estudo de duas variantes da enzima BPN', foi constatado que a estabilidade desta proteína aumentou 1.000 vezes com a inserção de 10 pontos de mutações, incluindo a adição de uma nova ponte dissulfeto. Existem ainda estudos, como em Sakaguchi et al. (2008), no qual foi constatado que as proteínas perdiam estabilidade na medida em que as suas pontes dissulfeto eram removidas.

A segunda base de dados utilizada neste trabalho foi formada por ligações de hidrogênio. Conforme Hubbard e Haider (2010) e Dias (2012) as pontes de hidrogênio são importantes para o enovelamento das proteínas, sendo responsáveis pela manutenção das estruturas secundárias. De acordo com Nelson e Cox (2014), Bowie (2011) e Pace et al. (2014) as pontes de hidrogênio também contribuem muito para a estabilidade das proteínas.

1.3 Objetivos

O objetivo geral deste trabalho é desenvolver matrizes de distâncias que possibilitem a representação tridimensional de interações de proteínas. Os objetivos específicos são:

1. Desenvolver matrizes de distâncias a partir de diferentes estratégias: relação dos carbonos alfa com os demais átomos, ângulos e mesclagem de matrizes de distâncias.
2. Combinar as técnicas desenvolvidas neste trabalho com algoritmos de *clustering* para validar as metodologias propostas, através de sobreposições atômicas entre as interações que ficarem no mesmo agrupamento;
3. Realizar testes em diferentes bases de dados;
4. Comparar as metodologias desenvolvidas neste trabalho com outras técnicas, bem como realizar análises estatísticas para determinar qual metodologia de representação

de interações proteicas e qual algoritmo de *clustering* mostraram-se mais adequados para o problema em questão.

5. Realizar análise de performance computacional.

1.4 Organização do Trabalho

O trabalho está organizado da seguinte maneira: o Capítulo 2 contém os trabalhos relacionados com esta tese de doutorado. O Capítulo 3 refere-se à fundamentação teórica, explicando os principais itens para o bom entendimento deste trabalho, como as proteínas, pontes dissulfeto, ligações de hidrogênio e alguns conceitos computacionais como matrizes de distâncias e *clustering*. O Capítulo 4 exibe a metodologia de pesquisa, os *hardwares* e *softwares* utilizados, as bases de dados utilizadas para os experimentos, a estrutura dos registros da base de dados, entre outras informações. O capítulo 5 refere-se ao desenvolvimento deste trabalho, detalhando as matrizes de distâncias desenvolvidas. O Capítulo 6 contém os resultados do trabalho, bem como a análise dos mesmos. O Capítulo 7 mostra a conclusão e os trabalhos futuros.

Capítulo 2

Trabalhos Relacionados

A predição de estruturas tridimensionais de proteínas é uma das tarefas mais desafiadoras da área de bioinformática. De acordo com Dias (2007) e Laimer et al. (2015) os métodos de predição de proteínas podem ser divididos principalmente em duas categorias: baseados em sequência e na estrutura. Este Capítulo está dividido do seguinte modo: a seção 2.1 refere-se aos trabalhos baseados na sequência e a seção 2.2 é sobre os *softwares* baseados na estrutura. Ressalta-se que como o presente trabalho é focado na estrutura, essa seção será mais detalhada.

2.1 *Softwares* de Predição de Proteínas Baseados na Sequência

Conforme Carpentier e Chomilier (2019), dentro da categoria de *softwares* baseados em sequência, existe uma divisão entre ferramentas que utilizam apenas informações sequenciais e *softwares* que combinam informações estruturais e sequenciais visando o aperfeiçoamento do alinhamento sequencial.

De acordo com Dias (2007) os *softwares* que baseiam-se apenas em informações sequenciais buscam por resíduos de aminoácidos que repetem-se em duas ou mais sequências de proteínas, visando a identificação de regiões similares que possam indicar relações funcionais ou evolucionárias entre estas sequências. Como exemplo de *softwares* desta categoria pode-se citar iPTREE-STAB, desenvolvido em Huang, Gromiha e Ho (2007) que utiliza árvore de decisão acoplada com algoritmo de reforço adaptativo, além de árvore de classificação e regressão. O objetivo do iPTREE-STAB consiste em predizer mudanças de estabilidade de proteínas após substituições de aminoácidos. MuStab é outro *software* focado apenas nas informações de sequências, sendo desenvolvido em Teng, Srivastava e Wang (2010) e utiliza aprendizado de máquina. O MuStab também possibilita a predição de mudanças de estabilidade proteica através da substituição de aminoácidos. Outros

exemplos de *softwares* baseados apenas em informações de sequências proteicas são as versões do Kalign (1, 2 e 3), implementadas respectivamente em Lassmann e Sonnhammer (2005), Lassmann, Frings e Sonnhammer (2009) e Lassmann (2020). Estes *softwares* possibilitam o múltiplo alinhamento de sequências de proteínas.

Dentre os *softwares* que combinam informações estruturais e sequenciais das proteínas pode-se citar o PROMALS3D. Esse *software* constrói alinhamentos de sequências múltiplas consistentes, reunindo informações e utilizando-se de validações de níveis sequenciais e estruturais sobre as proteínas de entrada (PEI; KIM; GRISHIN, 2008). Outro *software* pertencente a esta categoria é o *Formatt*, que possibilita o alinhamento múltiplo de estruturas, utilizando também informações de sequência para aprimorar os resultados (DANIELS; NADIMPALLI; COWEN, 2012). Chakrabarty e Parekh (2022) é um trabalho recente que combina análise sequencial e estrutural para estudar a família de proteínas anquirina. Chakrabarty e Parekh (2022) classificaram as proteínas anquirinas humanas em *clusters* com base na similaridade sequencial. Além de realizar a análise estrutural baseada em redes de proteínas, mostrando a associação de resíduos conservados com resíduos topologicamente importantes identificados por medidas de centralidade de rede. Essa análise ajuda a entender o papel desses resíduos na estabilidade estrutural e na função dessa família de proteínas.

2.2 *Softwares* de Predição de Proteínas Baseados na Estrutura

Os *softwares* de predição de proteínas baseados na estrutura frequentemente realizam comparações, alinhamentos e sobreposições atômicas em estruturas tridimensionais de proteínas. As tarefas realizadas por estes *softwares* são mais complexas do que os procedimentos realizados por *softwares* baseados nas sequências das proteínas (DIAS, 2007). O presente trabalho encontra-se nesta categoria por trabalhar com as coordenadas tridimensionais das interações proteicas, podendo futuramente ser incorporado com algum *software* que realiza proposições de mutações com base na estrutura. Deste modo, escolheu-se dar maior ênfase aos trabalhos relacionados desta classe.

Hazes e Dijkstra (1988) desenvolveram o SSBOND, um *software* que auxilia na seleção de locais em uma proteína alvo de estrutura tridimensional conhecida, nos quais pode-se inserir uma nova ponte dissulfeto. A proposta para introdução das pontes dissulfeto envolve a modificação de pares de resíduos de aminoácidos para cisteínas. Essa abordagem busca aprimorar a termoestabilidade das estruturas mutadas, com a finalidade de expandir sua viabilidade para aplicações tanto comerciais quanto médicas. O algoritmo SSBOND tem como objetivo identificar e categorizar possíveis pares de resíduos de aminoácidos, levando

em consideração não apenas as distâncias entre os átomos de carbono beta, mas também os ângulos diédricos correspondentes. Dessa maneira, o algoritmo matematicamente calcula posições que atendam às especificações das distâncias entre o carbono alfa e o carbono beta, bem como entre o carbono beta e o potencial aminoácido a ser mutado. Além disso, o algoritmo verifica a conformidade do ângulo de ligação do carbono beta e a consistência da distância entre o aminoácido potencialmente mutado do resíduo 1 e o carbono beta no resíduo 2 em relação aos valores de referência para as pontes dissulfeto.

Outro trabalho importante na predição estrutural de proteínas é o de Sowdhamini et al. (1989), no qual foi desenvolvido um procedimento de modelagem computacional para avaliar critérios de inserções de pontes dissulfeto através de mutações sítio-dirigidas, visando o aumento da estabilização da proteína. Os autores definiram valores de distâncias entre átomos de uma mesma proteína para identificar uma ponte dissulfeto. De acordo com Sowdhamini et al. (1989) a distância máxima entre dois átomos de enxofre gama deve ser 2,04 Å, a distância entre os átomos de carbono alfa deve ser inferior a 6,5 Å, as distâncias entre os átomos de carbono beta deve ser de até 4,5 Å, além da distância entre o átomo de carbono beta e enxofre gama, que deve ser de no máximo 1,87 Å. Esses valores e critérios listados ainda são frequentemente utilizados para identificar pontes dissulfeto em arquivos de texto no formato do PDB, bem como realizar propostas de inserções de novas interações de pontes dissulfeto.

Um trabalho mais recente no qual realizou-se propostas de inserções de pontes dissulfeto, objetivando melhorias em macromoléculas foi proposto por Bashirova et al. (2019). Conforme observado por Bashirova et al. (2019) e Zanchetta (2013), as celulases são enzimas responsáveis pela degradação da celulose e são utilizadas em várias aplicações industriais, como na fabricação de algodão e papel. As celulases são divididas em três principais grupos conforme as suas atividades. Dentre estes grupos, pode-se citar as endoglucanases, que atuam na região interna da fibra de celulose, liberando compostos menores, formados por poucas unidades de glicose chamados oligossacarídeos (açúcares pequenos).

Bashirova et al. (2019) afirmam que as aplicações práticas das endoglucanases são frequentemente limitadas devido à perda de atividade enzimática em temperaturas industrialmente exigidas. Pequenos aumentos de temperatura resultam em mudanças na estrutura e no enovelamento da proteína, afetando sua função. Portanto, a estabilização térmica da enzima é um requisito essencial para seu uso eficiente em biotecnologia. Deste modo, Bashirova et al. (2019) realizaram uma abordagem de engenharia de ponte dissulfeto para melhorar a termoestabilidade da enzima EGLII. Sendo assim, a partir de simulações computadorizadas foram escolhidas variantes benéficas da EGLII. Os pares de resíduos identificados foram submetidos a mutações sítio dirigidas para inserir novas pontes dissulfeto. Os resultados indicaram que estas mutações aumentaram a termoestabilidade.

Além das pontes dissulfeto, as ligações de hidrogênio também são interações proteicas importantes. Dentre os principais trabalhos que caracterizam as pontes de hidrogênio pode-se citar o de Jeffrey (1997), no qual é afirmado que uma ponte de hidrogênio caracteriza-se como forte se o doador e o aceptor estiverem a uma distância entre 2.2 Å e 2.5 Å. Se a distância entre ambos for superior a 2.5 Å e até 3.2 Å, ela pode ser considerada como moderada. Com a distância acima de 3.2 Å, até 4.0 Å ela pode ser considerada como fraca. Apesar de considerar que a distância entre o aceptor e o doador superior a 3.2 Å torna a ponte de hidrogênio fraca, Jeffrey (1997) ressalta que ainda é possível haver uma ponte de hidrogênio com força significativa quando o doador e o aceptor estão a uma distância de 3.5 Å.

O desenvolvimento de *softwares* que realizam proposições de mutações com base na estrutura, além de levar em consideração as distâncias que caracterizam as interações, devem validar também se as conformações tridimensionais do alvo (a ser mutado) e os candidatos a mutação são similares. De acordo com Kufareva e Abagyan (2012) e Campelo (2017), o *Root-Mean-Square Derivation* (RMSD) é a forma mais comum de comparar as similaridades de estruturas tridimensionais de proteínas. Dentre os principais trabalhos que envolvem o RMSD está o de Kabsch (1976), no qual foi desenvolvido o cálculo da matriz de rotação ideal, que pode ser utilizado entre duas estruturas de proteínas, para minimizar o RMSD entre ambas. Baseado em Kabsch (1976), o LSQKAB, desenvolvido por Winn et al. (2011) possibilita a comparação de similaridades tridimensionais entre estruturas proteicas com base nas coordenadas atômicas x, y e z. Através desta ferramenta é possível realizar sobreposições atômicas de um único átomo ou de um conjunto específico de átomos contidos em arquivos texto, no formato do PDB. Este *software* ainda possibilita obter o resultado detalhado da sobreposição atômica em um arquivo de texto, chamado de delta, contendo as diferenças de distâncias para cada par de átomo comparado (CCP4, 2011).

O LSQKAB e os demais *softwares* desenvolvidos a partir do algoritmo de Kabsch são amplamente utilizados na bioinformática. No entanto, o alinhamento estrutural feito pelo LSQKAB para obter a matriz de rotação ideal, bem como o cálculo da sobreposição atômica feita após esse alinhamento, não são simples de serem resolvidos computacionalmente, pertencendo à classe de problemas NP-Completo (CRESCENZI et al., 1998), (OLIVEIRA, 2016) e (MELO-MINARDI, 2021). Deste modo, é comum que *softwares* que tenham que realizar várias sobreposições atômicas, como os que propõem mutações em proteínas, utilizem estratégias para reduzir a quantidade de sobreposições atômicas necessárias, visando o ganho de performance computacional. Um exemplo é a ferramenta *Residue Interaction Database* (RID).

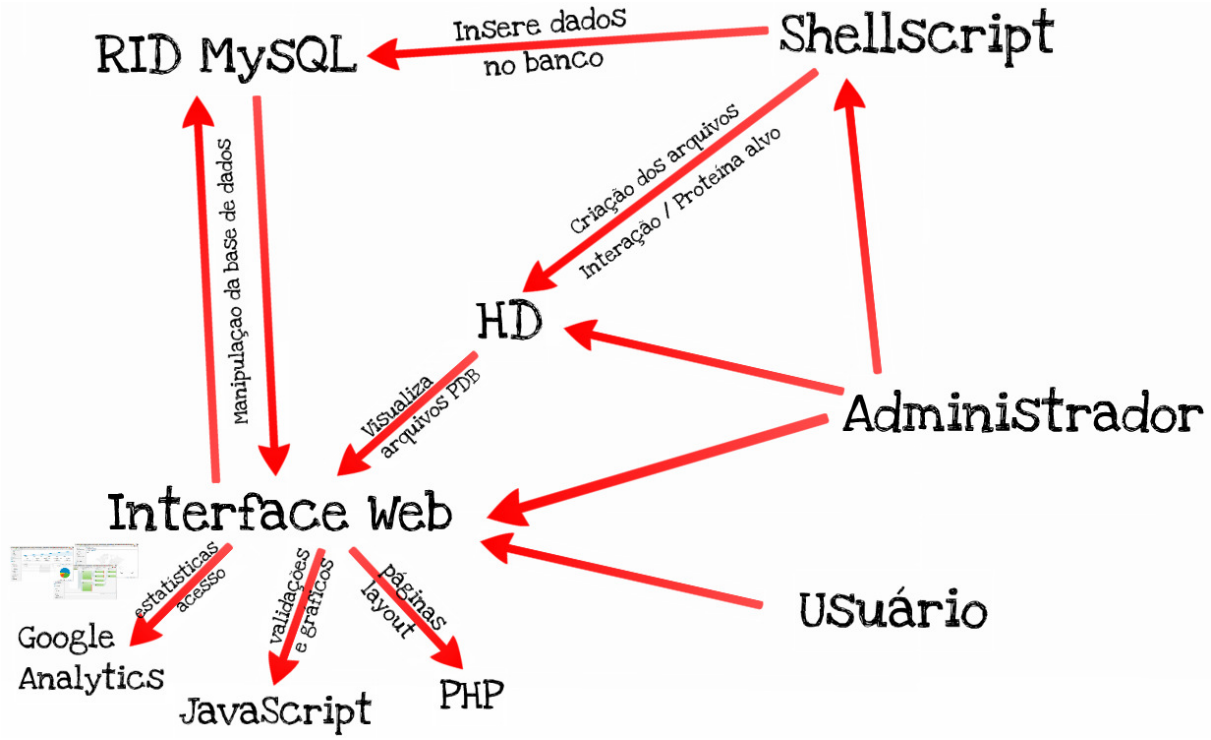
A RID consiste em um algoritmo e uma nova base de dados que propõem mutações entre pares de resíduos de proteínas com estruturas tridimensionais conhecidas, visando aumen-

tar a estabilidade da proteína, focando na conformação da cadeia principal. As mutações sugeridas mantêm o enovelamento da proteína alvo, para que ela continue com a sua função. Dias (2012) desenvolveu uma estratégia para trabalhar com o LSQKAB de um modo hábil na busca de pares interagentes da ferramenta. O LSQKAB foi utilizado na ferramenta RID para sobrepor todos os arquivos de interações da base de dados contra um arquivo de referência, gerando um registro de delta com as diferenças de distâncias atômicas para cada comparação (DIAS, 2012). Dias (2012) ainda otimizou a busca de pares interagentes da ferramenta RID, calculando um valor de *score* para cada par de arquivos deltas gerados. Este cálculo foi baseado na distância euclidiana a partir das diferenças de distâncias atômicas dentre os arquivos deltas. Este procedimento ocorreu para diminuir o custo da busca de pares similares que o LSQKAB teria para comparar todos os arquivos da base de dados, diminuindo a necessidade de sobreposições de $O(nInteracoesBanco * nParesCandidatos)$ para $O(nInteracoesBanco + nParesCandidatos)$, no qual $nInteracoesBanco$ refere-se ao número de interações da base de dados e $nParesCandidatos$ refere-se ao número de pares candidatos a mutação.

Um exemplo sobre o funcionamento da estratégia da RID consiste na seguinte situação: se o banco de dados tem 30.000 interações de proteínas e a proteína alvo contém 200 pares, também no formato de arquivo texto, a solução exata seria comparar cada um desses 200 pares com cada um das 30.000 interações armazenadas no banco de dados. Assim, seriam realizadas sobreposições atômicas em 6.000.000 (6 milhões) de possíveis combinações de arquivos de textos com informações de proteínas (30.000 x 200). Essa quantidade ainda poderia ser dobrada para 12.000.000 (12 milhões), se for considerado que durante o alinhamento estrutural, bem como nas sobreposições atômicas, uma estrutura é mantida como "fixa" e a outra é ajustada à essa primeira estrutura. Ao inverter as estruturas "fixa" e "móvel", os resultados podem ficar ligeiramente diferentes. Deste modo, a solução aproximada realizada pela RID consiste em determinar uma interação para ser a referência e realizar a sobreposição atômica de todas as 30.000 interações do banco de dados contra essa referência, gerando 30.000 arquivos de deltas (1 x 30.000). Do mesmo modo, realizar as sobreposições atômicas entre essa mesma referência e os 200 pares da proteína alvo, gerando 200 arquivos de deltas (1 x 200). Posteriormente a ferramenta RID ainda calcula um *score* para pontuar essas estruturas proteicas, com base nos valores obtidos nos arquivos de deltas. De um modo prático, a ferramenta RID tem o seu funcionamento com base no diagrama ilustrado pela Figura 2.

De acordo com a Figura 2, o administrador tem acesso aos códigos-fonte escritos em *Shellscript*, inserções diretas no HD e acesso à interface web. O usuário comum tem acesso apenas à interface web, que possibilita a manipulação da base de dados através de inserções de proteínas no formato do PDB, além de permitir ao usuário a visualização de validações, gráficos e estatísticas sobre as proteínas.

Figura 2 – Diagrama da ferramenta RID



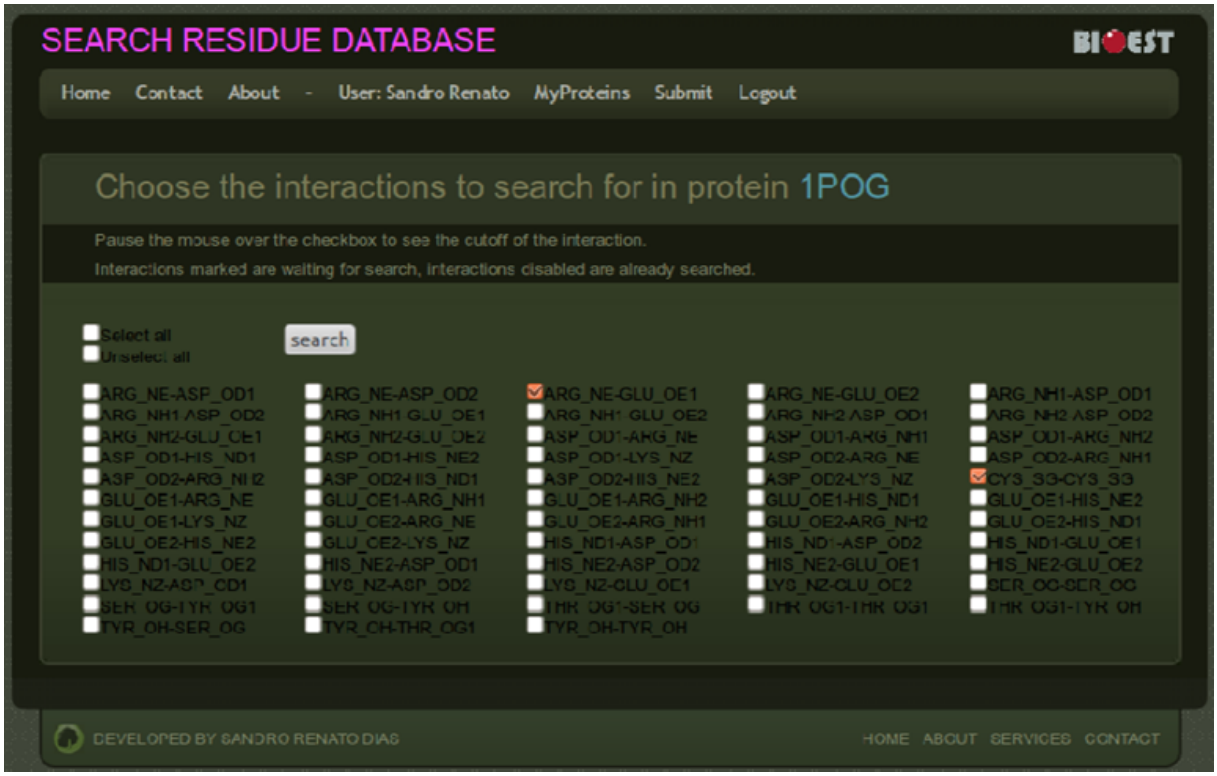
Fonte: Dias (2012)

Conforme Dias (2012), o usuário pode escolher como o sistema carregará as informações da estrutura proteica que ele deseja analisar. Além da opção de importar um arquivo no formato do PDB, o usuário pode informar o código do PDB da estrutura ou selecionar algum registro já importado anteriormente. Após essa escolha, o *software* apresenta a relação das interações existentes no banco de dados da RID, que podem ser utilizadas na estrutura submetida pelo usuário. A Figura 3 ilustra a escolha do usuário por trabalhar com a proteína 1POG.

Através da Figura 3 pode-se observar as possíveis interações do banco de dados da ferramenta RID que podem ser selecionadas pelo usuário para serem utilizadas na proteína submetida (PDB 1POG). Após selecionar as interações desejadas, o sistema ilustra através de pontos no gráfico, os pares do banco que têm conformações muito próximas da proteína alvo. A Figura 4 ilustra o cenário em que o usuário escolheu procurar por apenas pontes dissulfeto.

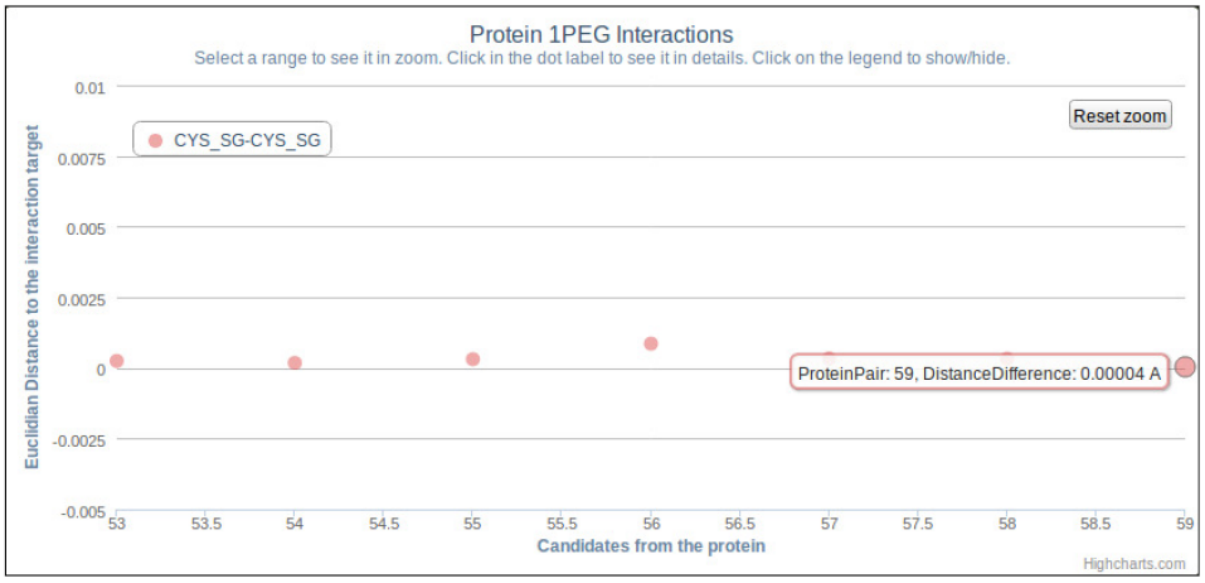
Através da Figura 4 pode-se clicar em um dos possíveis pares para visualizar a sua identificação (59) e o valor do *score* (0,00004 Å) calculado entre a interação da proteína alvo e a interação do banco de dados. O objetivo do uso do *score* é utilizá-lo como um identificador/assinatura. Quanto menor o valor do *score*, maior será a similaridade da conformação da cadeia principal do par de aminoácidos da proteína alvo (submetida pelo usuário) com a conformação da cadeia principal do par de aminoácidos interagentes contido

Figura 3 – Escolha das interações a serem pesquisadas pela ferramenta RID



Fonte: Dias (2012)

Figura 4 – Possíveis pares da estrutura submetida

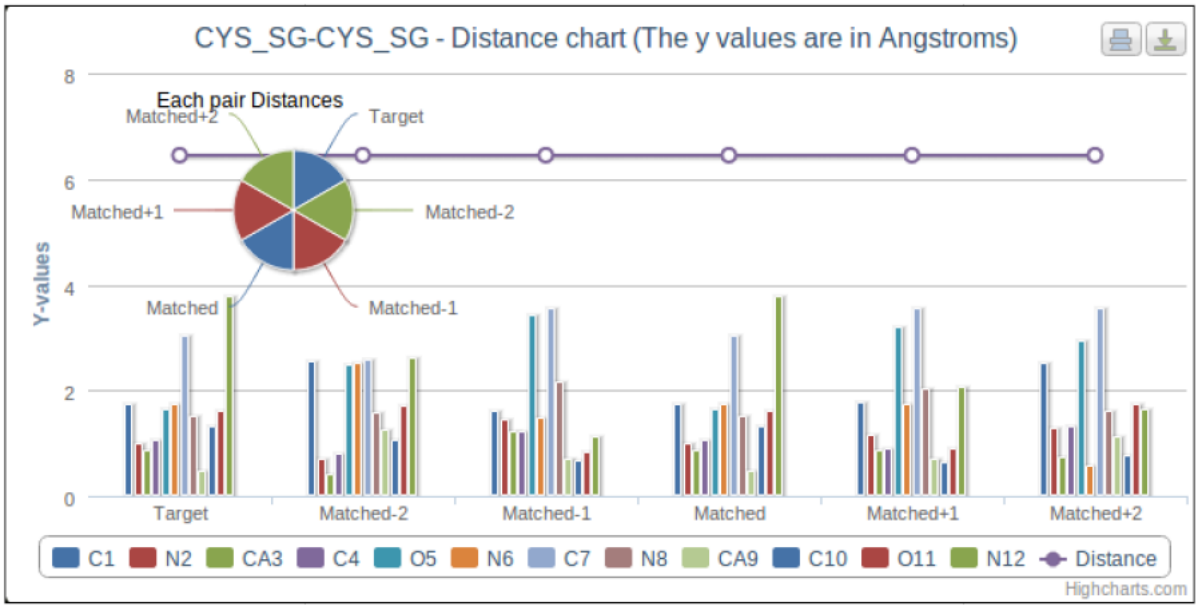


Fonte: Dias (2012)

no banco de dados e sugerido pelo sistema.

A Figura 5 ilustra com mais detalhes o possível par, indicando o par da proteína como *Target* e o par da interação do banco como *Matched*. Os *Matcheds* combinados com 1, 2, -1 e -2 são dois elementos posteriores e anteriores ao encontrado como melhor interação.

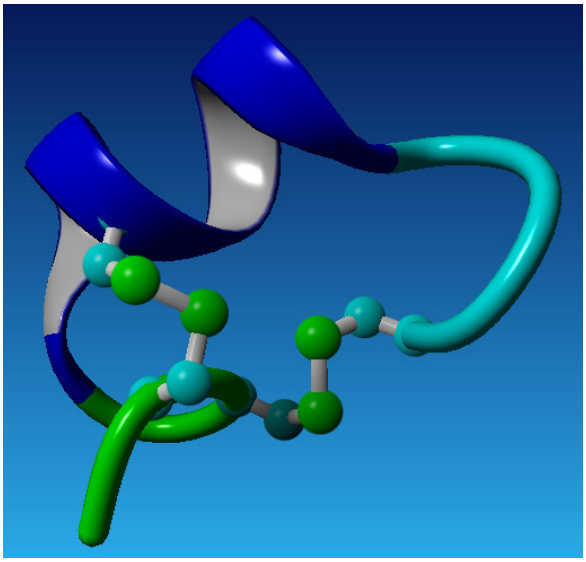
Figura 5 – Análises do par selecionado



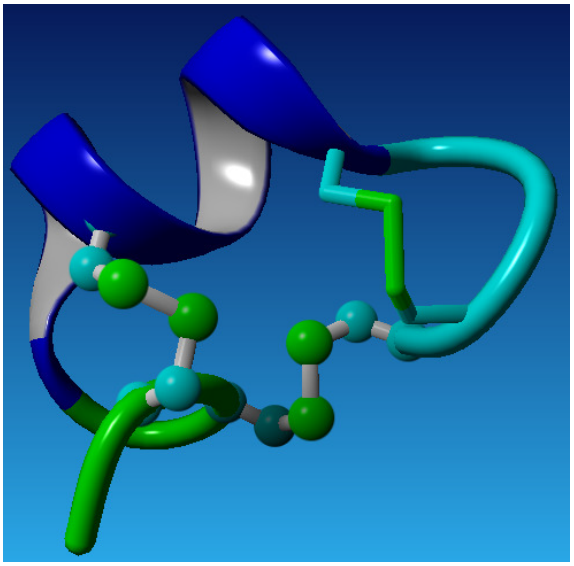
Fonte: Dias (2012)

Figura 6 – Estrutura da proteína PDB 1PEN antes e após a inclusão de uma ponte dissulfeto proposta pela RID

(a) 1PEN antes da proposta de inclusão



(b) 1PEN após a inclusão da ponte dissulfeto



Fonte: Dias (2012)

Através da Figura 5 pode-se observar três modalidades diferentes para a análise entre a proteína alvo (*target*) e as interações retornadas. O gráfico de pizza indica as porcentagens dos *scores* calculados, para compará-los no formato de fatias. A curva plotada acima do gráfico de pizza indica os valores em Angstrons dos *scores* encontrados. O gráfico de barras apresenta para cada barra a distância daquele átomo ao respectivo átomo do arquivo utilizado como referência. Um dos resultados da RID consistiu em uma proposição de mutação na proteína PDB 1PEN, ilustrada pela Figura 6.

A Figura 6 ilustra uma proposta de mutação feita pela RID, a partir de uma proteína resolvida em Hu et al. (1996) e armazenada no PDB com o título 1PEN. De acordo com Dias (2012) a 1PEN é um polipeptídeo pequeno, que possui uma cadeia de 109 átomos, totalizando 16 resíduos em uma única cadeia. Através da Figura 6 (a) pode-se notar que a 1PEN já possuía duas pontes dissulfeto, ilustradas no formato de *ball and stick*. A Figura 6 (b) ilustra a proposta de inclusão de uma nova ponte dissulfeto, representada no formato de *stick*. A ponte dissulfeto em questão foi obtida a partir da proteína PDB 1GAI, através de uma busca feita pela RID em sua base de dados, composta por interações extraídas do PDB. A proposta de inserção ocorreu por esta ponte dissulfeto ter conformação tridimensional similar à cadeia principal da 1PEN. Esta proposta de inserção pode aumentar a estabilidade da 1PEN (DIAS, 2012). De um modo geral, os resultados da RID foram bons, mas podem ser aperfeiçoados.

O *Protein Engineering Supporter* (Proteus) é outro *software* que propõem pares de mutações em uma estrutura tridimensional alvo. Conforme Barroso et al. (2020), as sugestões de mutações deste *software* são baseadas em contatos observados em outras estruturas conhecidas do PDB. O Proteus parte do pressuposto de que se a estrutura alvo tiver um par de resíduos de aminoácidos não interagentes e este par for trocado por um par interagente, o resultado pode aumentar a estabilidade da estrutura alvo. Barroso et al. (2020) ainda afirma que essa troca pode ocorrer somente se a conformação da cadeia principal dos resíduos envolvidos no contato for conservada. Além de não esperarem nenhum impedimento estérico entre as mutações propostas e os demais átomos.

De acordo com Barroso et al. (2020), para determinar se uma mutação pode ser sugerida, o Proteus realiza o alinhamento estrutural entre pares de tríades de uma proteína alvo com os pares de tríades de uma proteína do banco de dados. Para não ter que realizar comparações de RMSD entre 175.257 estruturas proteicas (toda a base de dados), o que demandaria alto custo computacional, o Proteus utilizou outras estratégias para reduzir a quantidade de comparações, descartando pares de tríades com baixas possibilidades de apresentarem estruturas similares. Deste modo, Barroso et al. (2020) introduziram um filtro antes da etapa de alinhamento estrutural, usando o *Structural Signature Variation* (SSV), apresentado em Mariano et al. (2019).

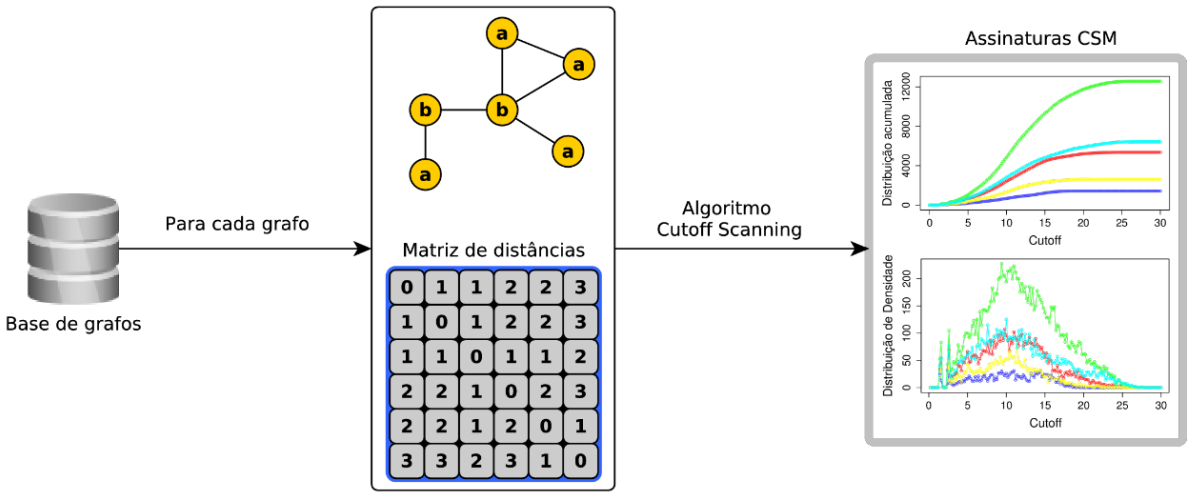
Barroso et al. (2020) afirma que para calcular a similaridade, o SSV utilizou a distância Euclidiana entre assinaturas de pares de tríades alvo reduzidas e toda assinatura reduzida do banco de dados do Proteus. De acordo com Barroso et al. (2020), apesar do SSV apresentar um método de comparação rápido com um grande número de verdadeiros positivos, foi observado também uma grande quantidade de falsos positivos. No entanto, Barroso et al. (2020) afirma que utilizar o SSV como filtro foi importante por reduzir o número de possíveis alinhamentos estruturais de 700 milhões pra 2 milhões, economizando um

considerável custo computacional e tornando viável o uso deste sistema em larga escala, na versão web.

O SSV utiliza a *atomic Cutoff Scanning Matrix* (aCSM), desenvolvida por Pires et al. (2013), para construir assinaturas estruturais. A aCSM é focada em tarefas de predição de ligantes de proteínas em larga escala, possibilitando também a geração do vetor de atributos considerando todos os átomos da estrutura (PIRES et al., 2013). No funcionamento da aCSM, a estrutura de uma proteína é convertida em um grafo, onde os átomos são os vértices e as arestas são definidas por um conjunto de distâncias de cortes entre os átomos (BARROSO et al., 2020). A aCSM é uma das versões da *Cutoff Scanning Matrix* (CSM). Além dessa versão, pode-se citar outras variações deste *software*, tais como: mCSM que foi apresentada em Pires, Ascher e Blundell (2014), pkCSM desenvolvida em Pires, Blundell e Ascher (2015), CSM-lig apresentada em Pires e Ascher (2016), mmCSM-PPI que foi construída em Rodrigues, Pires e Ascher (2021), pdCSM-cancer apresentada em Al-Jarf et al. (2021), entre outras. Cada uma dessas alternativas refere-se a uma expansão do algoritmo CSM, sendo adaptado para atuar em cenários mais específicos.

Conforme Campelo (2017), a CSM é considerada o estado arte na classificação de estruturas proteicas. Esse *software* foi desenvolvido por Pires et al. (2011) e obteve resultados satisfatórios em diversos cenários como na anotação automática proteica, predição de ligantes e atividade de pequenas moléculas (PIRES et al., 2011). De acordo com Pires et al. (2011), essa ferramenta trabalha com proteínas no formato do PDB. O seu funcionamento consiste em gerar vetores de atributos, que representam padrões de distâncias entre os nós de um grafo. Cada resíduo é representado por um centroide. Os vetores de atributos obtidos podem ser utilizados em tarefas de classificação. O fluxograma da CSM pode ser observado através da Figura 7.

Figura 7 – Fluxograma da CSM



Fonte: Pires (2012)

De acordo com a Figura 7, a primeira etapa desse algoritmo consiste em calcular a distância euclidiana entre todos os pares dos centróides que representam os resíduos. No trabalho original, Pires et al. (2011) realizaram experimentos definindo os carbonos alfa e beta como centroides. Os valores obtidos através destes cálculos foram inseridos em uma matriz de distâncias. Pires et al. (2011) definiram um intervalo de distâncias *(cutoffs)* a ser considerado (de 0,0 até 30,0 Å) e um passo (0,2 Å). As distâncias geradas, contidas nesses intervalos foram avaliadas e a frequência de pares de resíduos dentro do intervalo definido foram computadas. Deste modo, foi gerado para cada registro avaliado, um vetor com 151 valores. Unidos, estes vetores compõem a matriz CSM. Cada linha desta matriz representa uma proteína e cada coluna indica a frequência de pares de resíduos a uma distância menor ou igual a cada passo do valor de corte. A matriz CSM construída pode ser utilizada em tarefas de classificação. Como etapa final, foi utilizado o *Singular Value Decomposition* (SVD), para reduzir a dimensionalidade e os ruídos dos dados, assim diminuindo o tempo gasto com processamento. Nos trabalhos de Lima (2011) e Lima (2015), o CSM também foi utilizado juntamente com diferentes algoritmos de *clustering*: *k-means clustering*, *k-medoids* e *espectral clustering* para agrupar famílias de proteínas.

A partir da versatilidade da CSM e de suas versões, mais especificamente da aCSM, Monteiro (2017) desenvolveu uma metodologia baseada em Matriz de Distâncias Completa (MDC), para representar interações de proteínas, que pode ser combinada com técnicas de *clustering*. A MDC realiza o cálculo da distância Euclidiana entre todos os átomos da interação e armazena os resultados de cada interação em uma linha da matriz MDC. Essa matriz pode ser agrupada por técnicas de *clustering*. O objetivo da MDC é possibilitar o agrupamento dessas interações conforme as diferenças de distâncias atômicas, para obter grupos de interações com conformações tridimensionais similares. Os experimentos em Monteiro (2017) e Monteiro, Dias e Rodrigues (2020) foram realizados com arquivos de interações de pontes dissulfeto, extraídas do PDB pela ferramenta RID. Nesses experimentos, os resultados da MDC superaram as precisões do método de busca da RID e da aCSM, indicando que a MDC apresentou resultados iniciais promissores para trabalhar com interações de proteínas.

A técnica de agrupameto utilizada nos primeiros experimentos da MDC foi o *k-means clustering*. Conforme Karim et al. (2021) além desse algoritmo de agrupamento, existem diferentes técnicas que já foram testadas em problemas de bioinformática, tais como: *Gaussian Mixed Model* (GMM), *hierarchical clustering* e *OPTICS* . Essas técnicas apresentam diferentes vantagens e limitações. Amorim (2020) testou a combinação da MDC com a técnica de agrupamento GMM, para avaliar se essa combinação superaria os resultados da combinação entre MDC e *k-means clustering*. No entanto, conforme Amorim (2020) o uso do *k-means* apresentou uma melhor acurácia na maioria dos cenários testados, além de uma melhor performance computacional. Apesar disso, ainda torna-se viável testar outros algoritmos de

agrupamento para o problema em questão. Duas possibilidades são o *hierarchical clustering* por possuir o seu funcionamento baseado em hierarquias e o OPTICS por trabalhar com base na densidade das amostras (estratégias diferentes do *k-means clustering* e do GMM).

Além de testar outras técncias de agrupamento no problema em questão, pode-se desenvolver novas formas de representar as interações proteicas através de matrizes de distâncas. Existem diversos trabalhos que destacam a obtenção de bons resultados ao focar na relação dos átomos de carbono alfa (CA) com os demais átomos, por esses átomos estarem no centro da cadeia principal, serem importantes para o enovelamento das proteínas, além da cadeia lateral originar-se a partir desses átomos. Dentre os trabalhos que destacam a importância desses átomos pode-se citar Zhou, Yu e Anh (2007), no qual são utilizadas as coordenadas tridimensionais desses átomos para distinguir e agrupar proteínas de diferentes classes estruturais, com base na análise de quantificação de recorrência dos átomos de carbono alfa. Os autores ainda afirmam que muitas conformações das cadeias de aminoácidos são possíveis devido à rotações da cadeia em torno dos átomos de carbono alfa. Essas variações conformacionais são responsáveis pelas diferenças nas estruturas tridimensionais das proteínas. Pires et al. (2011) também ressalta os resultados satisfatórios que foram obtidos ao considerar os átomos de carbono alfa como centroides, para o cálculo de distância Euclidiana entre resíduos de aminoácidos intra-cadeia, visando a assinatura de grafos biológicos. De acordo com os autores, os experimentos com os átomos de carbono alfa obtiveram resultados melhores do que ao utilizar os átomos de carbono beta como centroides. Campelo (2017) inseriu pontos fictícios entre os átomos de carbono alfa para obter uma linearidade geométrica artificial da cadeia principal. Essa linearidade obtida foi coerente com a original. Ou seja, os pontos artificiais inseridos entre os átomos de CA reproduziram com boa precisão o posicionamento dos átomos nitrogênio (N) e carbono (C), sem a influência dos ângulos diedrais (*phi* e *psi*). Com estes resultados, os autores sugerem que não são as posições do nitrogênio e carbono os fatores determinantes no incremento da precisão da classificação, mas o fortalecimento do caráter geométrico linear da cadeia principal, obtida pela inserção dos pontos artificiais entre os átomos de CA.

Apesar de Campelo (2017) ter conseguido uma linearidade geométrica artificial da cadeia principal sem a influência dos ângulos *phi* e *psi*, alguns trabalhos recentes avaliaram os efeitos desses ângulos diedrais associados a mudanças conformacionais na proteína *spike*, gerando informações importantes na luta contra a pandemia de covid-19 (RAY; LE; ANDRICIOAEI, 2021) (GUPTA; CHAKRABARTI, 2022) (BASTIDAS, 2023). Dentre esses trabalhos, Bastidas (2023) foi o que mais focou nos ângulos diedrais, analisando o comportamento desses ângulos da proteína *Spike* ao longo do tempo, por meio de dinâmica molecular para identificar as suas oscilações. Dentre os resultados obtidos está a observação de que para aminoácidos menos abundantes na proteína *Spike*, existem certas frequências em que os ângulos diedrais nunca oscilam, em contraste com os aminoácidos relativamente mais

abundantes. Isso sugere que os componentes de frequência das oscilações dos ângulos diedrais também podem ser uma função da posição na estrutura primária, pois em quanto mais posições um aminoácido é encontrado, mais frequências ele pode amostrar.

O estudo da proteína *Spike* também conta com trabalhos que focaram nos átomos de carbono alfa, para analisar essa proteína ao longo do tempo. Um desses trabalhos consiste em Wei et al. (2023), no qual desenvolveu-se uma sequência de mapas de contato através dos átomos de carbono alfa da proteína *Spike*, de modo a representar a trajetória de evolução dessa proteína, para examinar a rota de regulação alostérica do SARSCoV-2. Esse trabalho também utilizou dinâmica molecular e aprendizado de máquina.

Diante desses trabalhos relacionados, entende-se que é possível desenvolver novas matrizes de distâncias para representar interações de proteínas, através de diferentes abordagens, que podem futuramente ser incorporadas à ferramenta RID. Acredita-se que focar nos átomos de carbono alfa pode ser uma estratégia eficaz diante da sua importância para a proteína. Também torna-se viável desenvolver matrizes de distâncias que levem em consideração os ângulos das proteínas, além de construir estruturas de dados que combinem diferentes estratégias de representações dessas macromoléculas, tanto por mesclagem de matrizes, tanto por formas mais robustas de realizar os cálculos para construí-las. Diante dos resultados interessantes da MDC, torna-se viável incluir novos algoritmos de clustering para agrupar as matrizes construídas através de novas abordagens, além do *k-means clustering*. Dentre as possíveis estratégias estão os algoritmos baseados na densidade (como o OPTICS) e algoritmos baseados na hierarquia (como o *hierarchical clustering*).

Capítulo 3

Fundamentação Teórica

Neste capítulo são descritos os principais temas para o bom entendimento deste trabalho. As primeiras seções referem-se aos conhecimentos necessários de biologia. Deste modo, a Seção 3.1 é sobre as proteínas, a 3.2 refere-se às pontes dissulfeto, a 3.3 refere-se às ligações de hidrogênio, a Seção 3.4 é referente aos ângulos *phi* e *psi* e a Seção 3.5 explica o alinhamento estrutural e o processo de sobreposição atômica. As últimas seções deste capítulo estão relacionadas aos conceitos de computação. A Seção 3.6 é referente à matriz de distâncias e a Seção 3.7 refere-se às técnicas de agrupamentos utilizadas neste trabalho: *k-Means clustering*, OPTICS e *hierarchical clustering*.

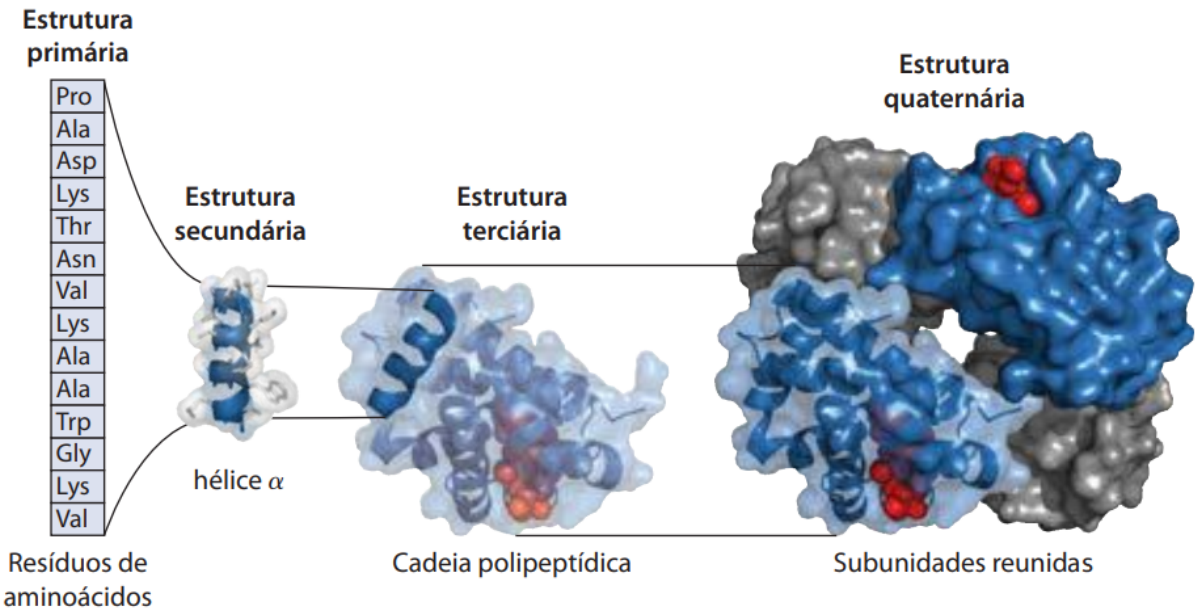
3.1 Proteínas

De acordo com Moraes et al. (2013) e Nelson e Cox (2014), as proteínas são as macromoléculas biológicas mais abundantes e estão presentes em todos os seres vivos. Essas macromoléculas podem realizar diversas funções biológicas, podendo ser classificadas conforme suas próprias funcionalidades, i.e., proteínas estruturais, transportadoras, reguladoras e enzimas.

A estruturação dos tecidos é feita pelas proteínas estruturais, que fornecem proteção e resistência às estruturas biológicas. Um exemplo é o colágeno, que está presente em cartilagens e tendões. As proteínas transportadoras atuam no transporte eficiente de muitas moléculas. Estão presentes nas membranas da célula, no espaço intracelular e também no plasma sanguíneo. Um exemplo é a hemoglobina dos eritrócitos, que é responsável pelo transporte de oxigênio dos alvéolos pulmonares para os tecidos. As proteínas reguladoras têm participação na regulação da atividade celular ou fisiológica. Como exemplo, estão muitos hormônios proteicos, que são sintetizados por glândulas endócrinas. Ao serem liberados na corrente sanguínea, podem estimular ou inibir vias metabólicas regulando, desta forma, processos celulares, como a insulina, por exemplo. O grupo de proteínas que exibe atividades catalíticas são chamados de enzimas (MORAES et al., 2013).

As proteínas também podem ser classificadas em quatro diferentes níveis de organização estrutural. Sendo eles primário, secundário, terciário e quartenário (MORAES et al., 2013) (NELSON; COX, 2014). Esses níveis estruturais são ilustrados na Figura 8 .

Figura 8 – Níveis de organização estrutural das proteínas.



Fonte: Nelson e Cox (2014)

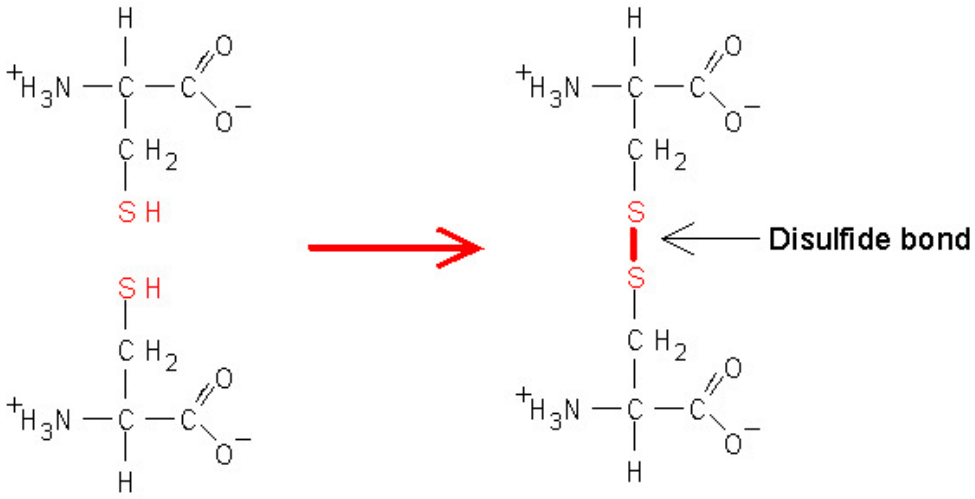
Conforme Nelson e Cox (2014) e Moraes et al. (2013), a estrutura primária refere-se à sequência de aminoácidos unidos entre si por ligações peptídicas. A estrutura secundária é dada pela forma em que os aminoácidos se interagem, formando um arranjo regular de resíduos de aminoácidos em um segmento de uma cadeia polipeptídica, no qual cada resíduo e seus vizinhos estão espacialmente relacionados do mesmo modo. Com essas interações, é criada uma conformação espacial de polipeptídeo. A estrutura terciária consiste na conformação tridimensional completa de uma cadeia polipeptídica. A conformação nativa é mantida pelas múltiplas interações não covalentes (como pontes dissulfeto) que se formam entre os resíduos ou grupamentos prostéticos da proteína. A estrutura quaternária é referente a localização espacial de múltiplas cadeias polipeptídicas, associadas de uma maneira estável.

3.2 Pontes Dissulfeto

Conforme Hunter (1993), as pontes dissulfeto são ligações covalentes, formadas pelos átomos de enxofre das cisteínas, que após oxidarem, tornam-se cistinas. Estas ligações são responsáveis por manter a estabilidade conformacional (estrutura tridimensional) de uma

proteína, ligando partes distantes de uma cadeia polipeptídica ou de diferentes cadeias. A Figura 9 ilustra uma ponte dissulfeto.

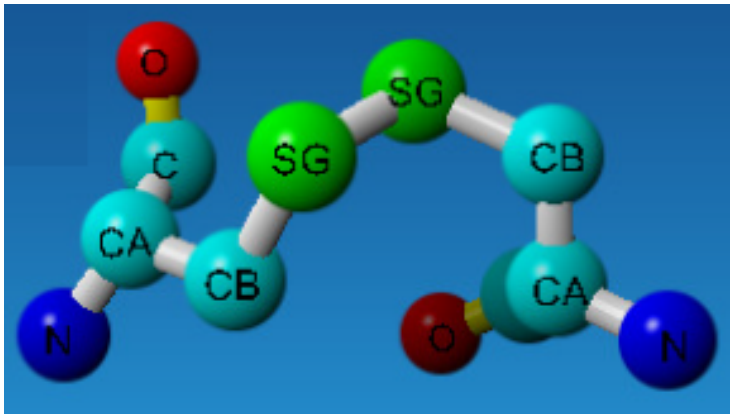
Figura 9 – Ponte dissulfeto formada pela interação de duas cisteínas



Fonte: Yu (2013)

As pontes dissulfeto são ligações covalentes muito fortes (YU, 2013). A construção de uma ponte dissulfeto, conforme ilustrado pela Figura 9, ocorre através da ligação de duas cisteínas. O aminoácido cisteína possui grupos tiol -SH em sua cadeia lateral. Dois resíduos de cisteína podem interagir resultando em uma ligação -S-S-. Após essa união, estas moléculas são chamadas cistinas. A Figura 10 ilustra tridimensionalmente uma ponte dissulfeto.

Figura 10 – Visualização tridimensional de uma ponte dissulfeto



Fonte: Dias (2012)

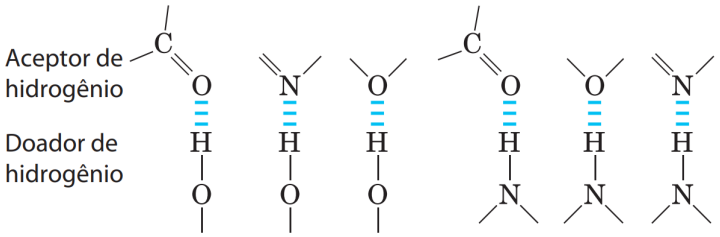
Conforme a Figura 10, pode-se observar a construção de uma ponte dissulfeto através da ligação de duas cisteínas pelos átomos de Enxofre Gama (SG), na cor verde. A ponte dissulfeto estabiliza a cadeia principal de cada lado interagente. A cadeia principal de cada

lado é composta pelos átomos de Nitrogênio (N), Carbono Alfa (CA), Carbono (C) e Oxigênio (O).

3.3 Ligações de Hidrogênio

De acordo com Hubbard e Haider (2010) e Dias (2012), as pontes de hidrogênio fornecem a maioria das interações direcionais que sustentam o enovelamento e a estrutura das proteínas, sendo também responsáveis pela manutenção das estruturas secundárias como α-hélices e folhas β. Isto deve-se pela interação de hidrogênio ocorrer devido ao compartilhamento de um átomo de hidrogênio entre um doador de próton e um aceptor de próton, podendo ser interpretada como um estágio intermediário na transferência de um próton de um ácido AH para uma base B, ocorrendo entre moléculas polares. Através da Figura 11 pode-se observar como as pontes de hidrogênio são formadas.

Figura 11 – Formações das pontes de hidrogênio



Fonte: Nelson e Cox (2014)

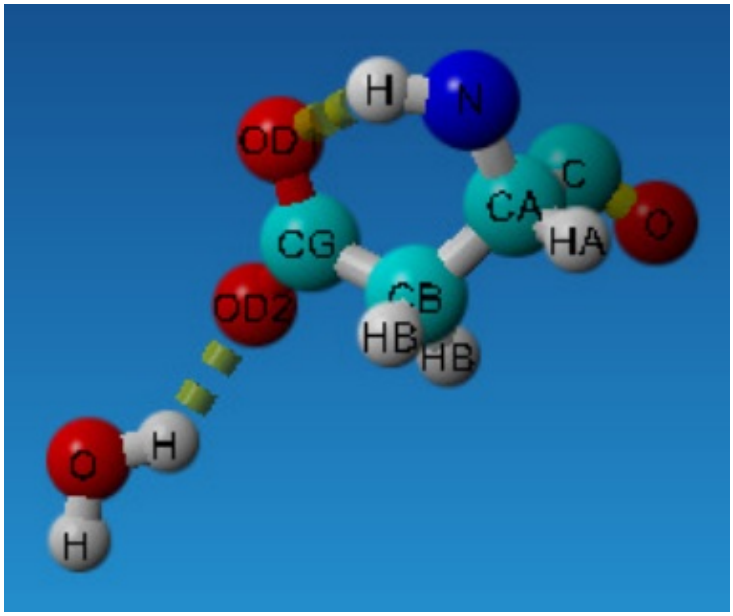
De acordo com a Figura 11, a ponte de hidrogênio é formada entre um átomo eletronegativo (hidrogênio aceptor, que frequentemente é um oxigênio ou um nitrogênio que contém um único par de elétrons), e um átomo de hidrogênio ligado covalentemente a outro átomo eletronegativo (o doador de hidrogênio) (NELSON; COX, 2014). A Figura 12 ilustra duas pontes de hidrogênio no formato de *ball* e *stick*.

Conforme Dias (2012) a Figura 12 ilustra duas ligações de hidrogênio. A primeira pode ser observada entre o átomo oxigênio delta 2 (OD2), que foi isolado do restante da cadeia e o hidrogênio (H) da molécula de água próxima do resíduo. A outra ponte de hidrogênio localiza-se entre os átomos oxigênio delta 1 (OD1) do resíduo e o H ligado ao nitrogênio (N) da cadeia principal do resíduo.

3.4 Ângulos *Phi* e *Psi*

De acordo com Nelson e Cox (2014) e Al-Karadaghi (2022) os ângulos dos dobramentos da estrutura terciária das proteínas são essenciais para a determinação da conformação nativa das proteínas. Dentre estes ângulos, pode-se citar os ângulos torsionais *phi* e *psi*. O

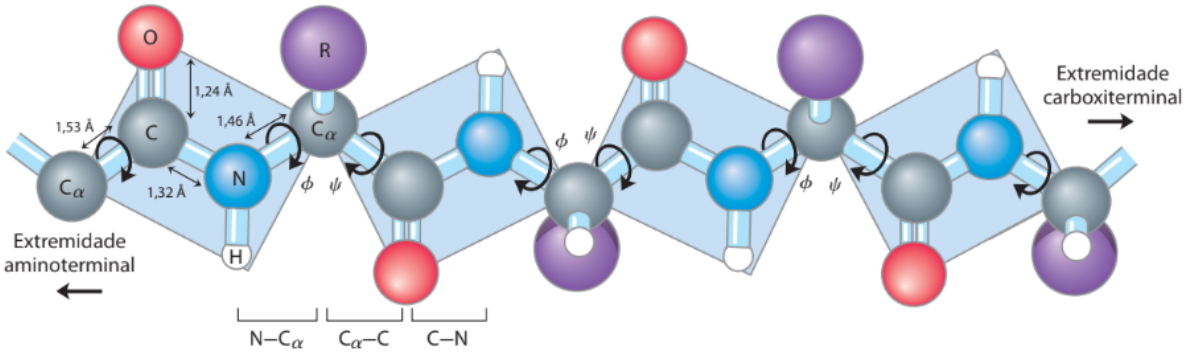
Figura 12 – Visualização tridimensional de ligações de hidrogênio



Fonte: Dias (2012)

ângulo *phi* é calculado entre o carbono alfa e o nitrogênio. O ângulo *psi* é calculado entre o carbono alfa e o carbono. A Figura 13 ilustra estes ângulos ao longo da cadeia principal.

Figura 13 – Cadeia principal com os ângulos *phi* e *psi*



Fonte: Nelson e Cox (2014)

A Figura 13 ilustra os carbonos na cor cinza (incluindo os carbonos alfas) e os átomos de nitrogênio são ilustrados na cor azul claro. Nesta figura pode-se observar os ângulos *psi*, representados por ψ, calculados entre os carbonos alfa e os carbonos. Também é possível notar os ângulos *phi*, representados por ϕ, calculados entre os carbonos alfa e os nitrogênios.

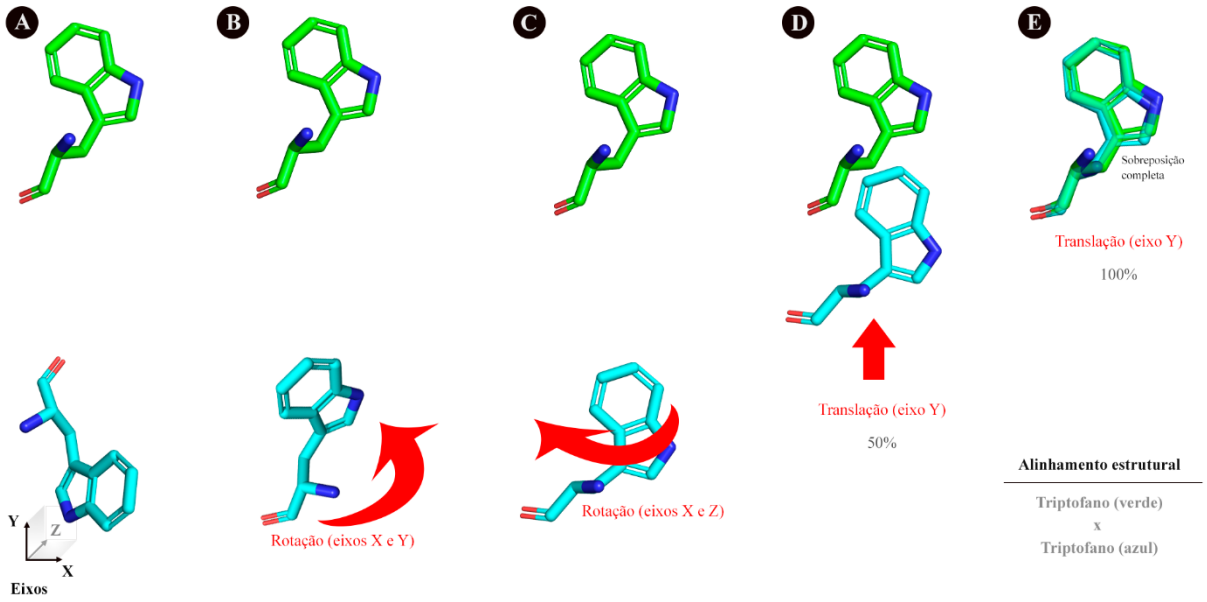
3.5 Alinhamento Estrutural e Sobreposições Atômicas

A estrutura tridimensional de uma proteína está diretamente ligada à sua atividade biológica (ANTCZAK et al., 2016). Diversas pesquisas sugerem que a evolução tende a conservar a estrutura tridimensional das proteínas, bem como de outras macromoléculas. Portanto, similaridades estruturais entre duas proteínas podem indicar relações evolutivas e funções comuns (KOLODNY; LINIAL, 2004) (SANTOS; MARTINS; MARIANO, 2021). Deste modo, conhecer e comparar as estruturas das proteínas é uma tarefa importante para a área de bioinformática. De acordo com Kolodny e Linial (2004) o alinhamento estrutural tem o objetivo de comparar duas estruturas tridimensionais de moléculas, como as proteínas, sobrepondo uma estrutura em relação a outra e estabelecendo equivalências entre as estruturas comparadas, com base na forma e na conformação tridimensional.

De acordo com Santos, Martins e Mariano (2021) o alinhamento estrutural contém 3 fundamentos. O primeiro refere-se a considerar o espaço tridimensional. As proteínas são compostas por aminoácidos, que por sua vez são formados por átomos. Cada átomo é representado por coordenadas x, y e z. O segundo fundamento trata-se de que apenas uma estrutura é alterada. Em um alinhamento par-a-par, apenas as coordenadas de uma estrutura devem ser movimentadas. As coordenadas atômicas da outra estrutura devem permanecer fixas, sendo utilizadas como referências para a realização do alinhamento. No caso de alinhar mais de duas estruturas, uma delas será a estrutura-referência e as demais serão alinhadas em relação a ela. O terceiro fundamento trata-se das operações de rotação e translação. No processo de alinhamento estrutural pode-se realizar apenas estas duas operações. Como os aminoácidos são interligados por ligações covalentes, ao aplicar o movimento de rotação em um dos átomos, as coordenadas dos demais átomos interligados serão afetadas. A Figura 14 ilustra o processo de alinhamento estrutural entre dois aminoácidos triptofano.

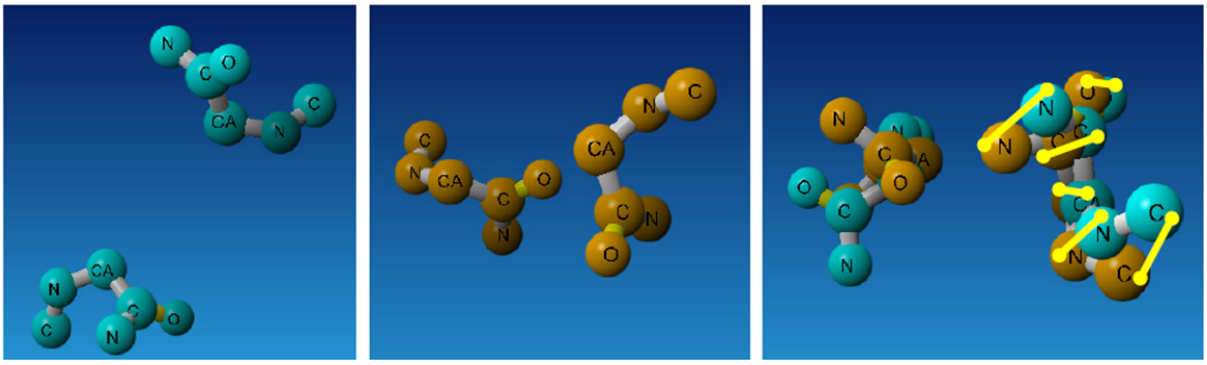
A Figura 14 ilustra os dois aminoácidos triptofano no formato de *sticks* com as cores verde e azul ciano. A parte (A) da Figura 14 ilustra os aminoácidos em diferentes sistemas de coordenadas, com orientações e posições distintas. A parte (B) da Figura 14 ilustra o movimento de rotação do triptofano azul, com o objetivo de aproximar-se da posição final do triptofano verde. A divisão (C) da Figura 14 indica uma nova rotação do triptofano azul visando uma maior aproximação da posição da estrutura verde. A divisão (D) da Figura 14 ilustra o movimento de translação realizado pela estrutura azul. Por fim, a parte (E) da Figura 14 ilustra o último movimento de translação do triptofano azul, no qual esta estrutura sobrepõem atomicamente o triptofano verde. Como resultado final deste alinhamento estrutural obtêm-se uma sobreposição atômica no qual cada átomo da estrutura azul está sobreposto sobre o seu átomo correspondente do triptofano verde. A Figura 15 ilustra mais um exemplo de sobreposição atômica.

Figura 14 – Sobreposição atômica entre dois aminoácidos triptofano



Fonte: Santos, Martins e Mariano (2021)

Figura 15 – Sobreposição atômica entre duas interações de proteínas



Fonte: Dias (2012)

A Figura 15 ilustra um segundo exemplo de sobreposição atômica obtida após o alinhamento estrutural. Nesta ilustração, na parte esquerda, são exibidos 12 átomos na cor azul claro no formato de *balls* e *sticks*. No centro da Figura 15 pode-se notar uma outra estrutura, também no formato de *balls* e *sticks* e com 12 átomos, mas na cor laranja. A ilustração à direita da Figura 15 refere-se ao resultado final da sobreposição atômica entre as duas estruturas comparadas. Após realizar a sobreposição atômica é possível medir o quanto as estruturas comparadas são similares tridimensionalmente. Dentre as formas mais comuns de fazer esta análise estão a validação quanto ao maior deslocamento e o cálculo do RMSD.

O maior deslocamento consiste em avaliar a maior diferença de distância entre os pares

de átomos sobrepostos. As linhas de amarelo indicadas na direita da Figura 15 ilustram este procedimento. A avaliação do maior deslocamento pode ser obtida através do *software* LSQKAB. Este *software* realiza o alinhamento estrutural e a sobreposição atômica entre estruturas proteicas, possibilitando o retorno do maior deslocamento através da informação *Maximum Displacement*, ou através da geração de um arquivo de delta contendo as informações das diferenças de distâncias atômicas entre cada par de átomos sobrepostos, no qual a maior destas diferenças de distâncias refere-se ao maior deslocamento. A Figura 16 mostra a estrutura do arquivo delta "1ejg_CYS-3-A_CYS-40-A_mc6.pdb-1cbn_CYS-3-A_CYS-40-A_mc6.pdb.deltas".

Figura 16 – Exemplo de um arquivo delta

```
0.015       2C   A              2C   A
0.017       3N   A              3N   A
0.033       3CA  A              3CA  A
0.031       3C   A              3C   A
0.032       3O   A              3O   A
0.023       4N   A              4N   A
0.031      39C   A             39C   A
0.013      40N   A             40N   A
0.029      40CA  A             40CA  A
0.031      40C   A             40C   A
0.047      40O   A             40O   A
0.026      41N   A             41N   A
```

Fonte: Monteiro, Dias e Rodrigues (2020)

De acordo com Dias (2012) e Monteiro, Dias e Rodrigues (2020), o nome do arquivo delta é dado pela união dos nomes dos registros que o formou, separados por "-". Neste exemplo, os arquivos formadores foram o "1ejg_CYS-3-A_CYS-40-A_mc6.pdb" (arquivo utilizado como referência) e o "1cbn_CYS-3-A_CYS-40-A_mc6.pdb". A primeira coluna refere-se ao valor das diferenças das distâncias atômicas entre estes arquivos (informação utilizada para avaliar o maior deslocamento). As próximas duas colunas referem-se às informações da primeira interação proteica e as duas últimas à segunda interação. Estas informações são: número do resíduo da proteína, identificação do átomo e cadeia. Ao avaliar o maior deslocamento desta comparação, pode-se notar o valor 0,047 Å.

A forma mais comum de comparar as similaridades tridimensionais entre estruturas proteicas é através do RMSD, cuja fórmula pode ser visualizada através da equação 1.

$$RMSD(v,w) = \sqrt{1/n \sum_{i=1}^{n}((v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz})^2)} \quad (1)$$

De acordo com a equação 1, v e w são dois arquivos de interações, n é o total de átomos de cada registro, v_{ix} e w_{ix} referem-se aos valores das coordenadas x das duas estruturas comparadas. Do mesmo modo, v_{iy} e w_{iy} referem-se aos valores das coordenadas y das duas estruturas. Por fim, v_{iz} e w_{iz} representam os valores da coordenada z. O RMSD retorna um único valor para definir o quanto duas estruturas são similares. Conforme Tsai (2003) duas estruturas proteicas que tenham o RMSD de até 0,5 Å possuem uma forte similaridade estrutural. Este valor é utilizado em outros trabalhos, como em Barroso et al. (2020).

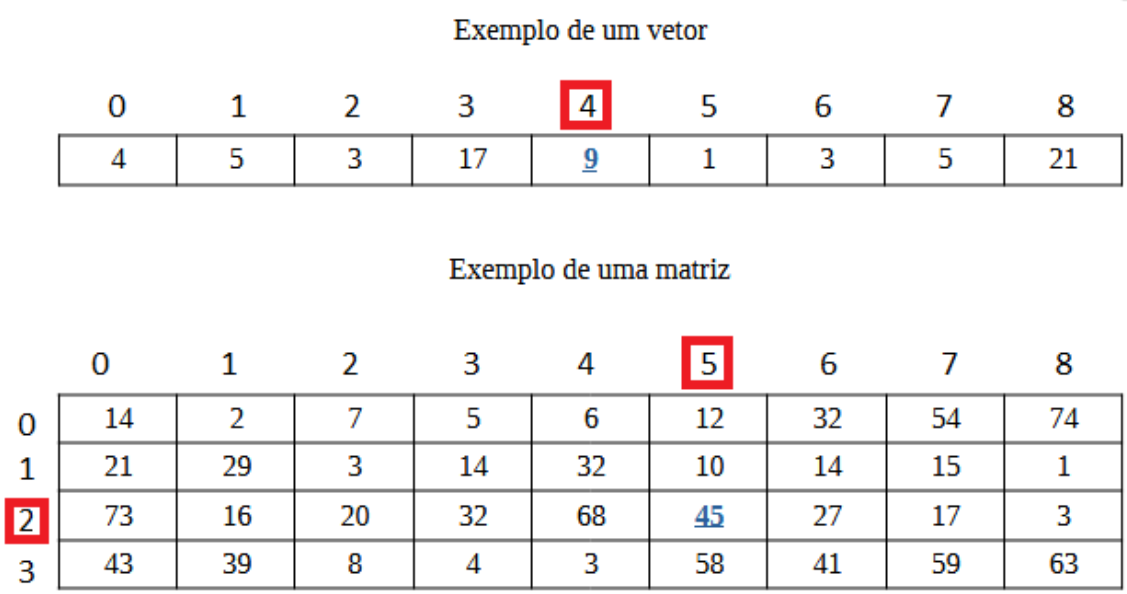
A avaliação através do maior deslocamento é mais rigorosa do que a avaliação pelo RMSD. Por exemplo, ao definir a maior tolerância como 0,5 Å, obrigatoriamente as 12 diferenças de distâncias entre todos os átomos das duas interações proteicas comparadas devem ter até 0,5 Å. No entanto, se definir a mesma tolerância quanto ao RMSD é possível que um ou mais pares de átomos comparados tenham valores superiores à 0,5 Å, desde que os demais pares de átomos compensem estando mais próximos.

3.6 Matriz de Distâncias

A matriz é um tipo de dado que permite a representação de variáveis de um mesmo tipo, armazenadas de forma contínua na memória. Cada valor da matriz pode ser acessado individualmente através de seu índice. A matriz pode ser unidimensional (também chamada de vetor ou *array*), ou pode ter mais dimensões (PINHO, 2020). A Figura 17 ilustra um vetor e uma matriz.

Através da Figura 17 pode-se observar na parte superior um vetor de 9 posições (começando pelo índice 0). Cada uma destas posições pode armazenar um valor, ou em determinadas implementações, um índice pode conter uma outra estrutura de dados, como um novo vetor, por exemplo. No entanto, nesse exemplo, ao acessar o índice de número 4 do vetor (destacado na cor vermelha), será obtido o valor 9 (na cor azul e sublinhado). A Figura 17 ainda ilustra na parte inferior, uma matriz de 4 linhas e 9 colunas. As matrizes possibilitam armazenar dados de uma forma organizada. Neste exemplo, ao acessar o índice de linha 2, juntamente com o índice de coluna 5, será obtido o valor 45. A estrutura de dados da matriz pode ser utilizada para armazenar as distâncias entre pontos, com cada posição representando a distância entre dois pontos. Deste modo, é possível armazenar todas as distâncias de um conjunto de dados (NEI; KUMAR, 2000). O uso da matriz de distâncias também pode ser aplicado no estudo de proteínas, pode-se utilizar cada linha da matriz

Figura 17 – Exemplo de um vetor e de uma matriz



Fonte: Elaborada pelo autor (2023)

para representar uma interação de proteína e cada coluna da mesma matriz pode indicar as informações de distâncias atômicas sobre a respectiva interação proteica.

3.7 Técnicas de *Clustering* Utilizadas Neste Trabalho

De acordo com Xu e Wunsch (2008), a definição de *cluster* não é universalmente aceita. Entretanto, Xu e Wunsch (2008) listam algumas descrições que podem ser consideradas como válidas. Dentre estas definições, pode-se citar Everitt (1979), no qual afirma-se que um *cluster* pode ser entendido como um conjunto de entidades que são similares entre si e as entidades contidas nos outros *clusters* são distintas do primeiro *cluster* mencionado.

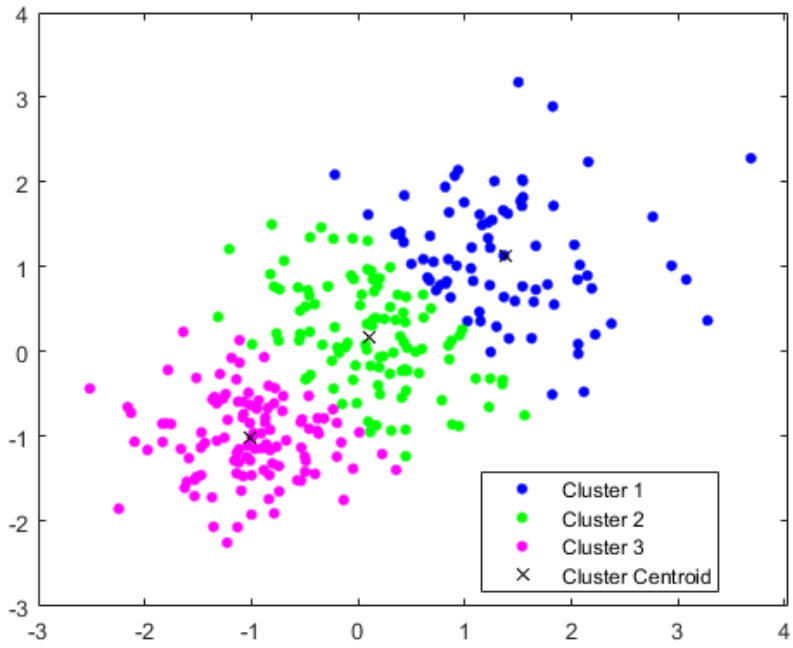
Existem diferentes algoritmos de *clustering*, que possibilitam agrupar os dados com base em diferentes maneiras, tais como: similaridade, densidade e distribuição estatística. Estas técnicas são frequentemente utilizadas para agrupar dados do mundo real, como estruturas tridimensionais de proteínas (KARIM et al., 2021). Neste trabalho foram utilizadas as técnicas *k-means clustering*, OPTICS e *hierarchical clustering*.

3.7.1 *K- Means Clustering*

De acordo com MathWorks (2006b), o *k-means clustering* é uma técnica de agrupamento que particiona os dados em “k” grupos distintos, baseados na distância para o centroide

de cada *cluster*. O centroide é o ponto para o qual a soma das distâncias de todos os elementos de um grupo é a menor possível. A Figura 18 ilustra os *clusters* formados pelo *k-means* e seus respectivos centroides.

Figura 18 – Ilustração do *k-means clustering*



Fonte: MathWorks (2006b)

Através da Figura 18 pode-se observar a formação de três *clusters*. O centroide de cada grupo é indicado por "x". O *k-means clustering* minimiza a cada iteração a soma das distâncias de cada item em relação ao centroide do seu grupo, movendo objetos entre *clusters*, até que a soma não possa ser mais diminuída. Por padrão, o *k-means clustering* utiliza a distância Euclidiana como métrica para este cálculo. Conforme Ali (2021) as etapas deste algoritmo são as seguintes:

1. Definir os centroides iniciais;
2. Calcular a distância Euclidiana de cada ponto de dados para os centroides selecionados na etapa 1;
3. Formar *clusters* atribuindo cada registro ao agrupamento do centroide que ele tenha a menor distância;
4. Obter a média de cada *cluster* formado. Os pontos médios serão os novos centroides.

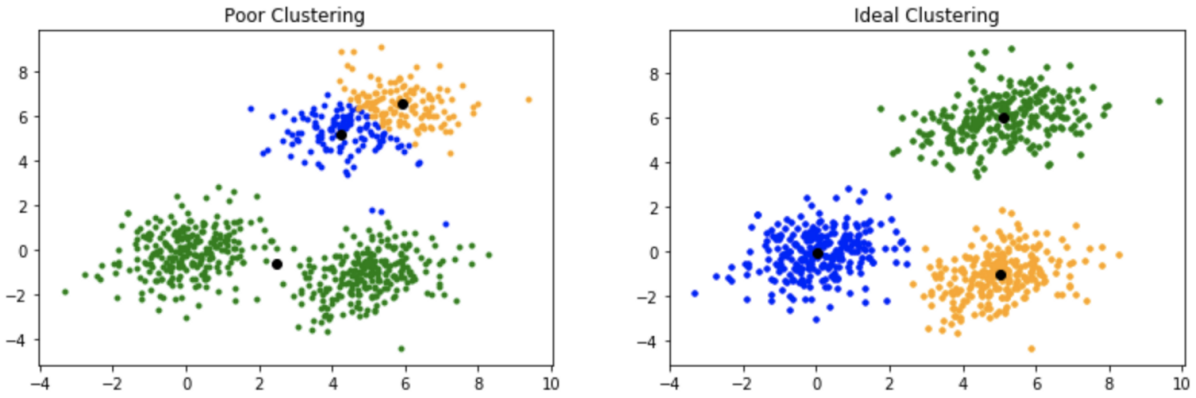
De acordo com Pedregosa et al. (2011) e Ali (2021) a escolha do algoritmo de inicialização do *k-means clustering* é uma etapa importante, que pode impactar na formação dos agrupamentos. Neste trabalho utilizou-se a inicialização do *k-means clustering* com a entrada *k-means++*, implementada por Pedregosa et al. (2011) através da biblioteca *Scikit-*

learn na linguagem de programação Python. O algoritmo de inicialização "*k-means++*" utiliza uma heurística com amostras baseadas em probabilidades empíricas das contribuições dos pontos para encontrar "sementes" de centroides (PEDREGOSA et al., 2011). Essa estratégia diminui o tempo de processamento e melhora o resultado final na maioria dos casos se comparado com a forma de inicialização original do *k-means*, por possibilitar que os centroides sejam inicializados mais distantes uns dos outros. Se os pontos definidos como centroides iniciais estiverem muito próximos, há uma grande possibilidade deles encontrarem um *cluster* em sua região local e o agrupamento real ser misturado com outro *cluster* (PEDREGOSA et al., 2011) e (ALI, 2021). A Figura 19 ilustra a diferença na formação final dos agrupamentos com base na forma de inicialização.

Figura 19 – Diferença da formação final dos *clusters* a partir do algoritmo de inicialização

(a) Formação ruim de agrupamentos

(b) Formação ideal de agrupamentos (inicialização através do *K-Means ++*)



Fonte: Ali (2021)

Através da Figura 19 (a) pode-se notar que quando os centroides iniciais ficam muito próximos, os registros que seriam de um mesmo agrupamento são inseridos em grupos distintos e ainda, registros que seriam de agrupamentos diferentes são inseridos no mesmo grupo. Ou seja, se a definição dos centroides iniciais for inadequada, com centroides muito próximos, o algoritmo *k-means clustering* pode não conseguir identificar a formação ideal durante o processamento. A Figura 19 (b) ilustra a formação ideal dos *clusters* após a escolha dos centroides obtida pela inicialização com o *k-means++*, que contribui para a boa qualidade dos agrupamentos. As etapas do algoritmo de inicialização *k-means++* são descritas a seguir conforme Pedregosa et al. (2011) e Ali (2021):

1. O primeiro centroide é escolhido aleatoriamente;
2. Calcular a distância Euclidiana entre o centroide e todos os outros pontos da base de dados. O ponto mais longe será o próximo centroide;
3. Criar agrupamentos em torno desses centroides, associando cada ponto ao centroide mais próximo.
4. O ponto que estiver mais distante de seu centroide será o próximo centroide;

5. O último passo consiste em repetir os passos 3 e 4 até que todos os centroides sejam escolhidos.

3.7.1.1 Complexidade do *K-Means* Clustering

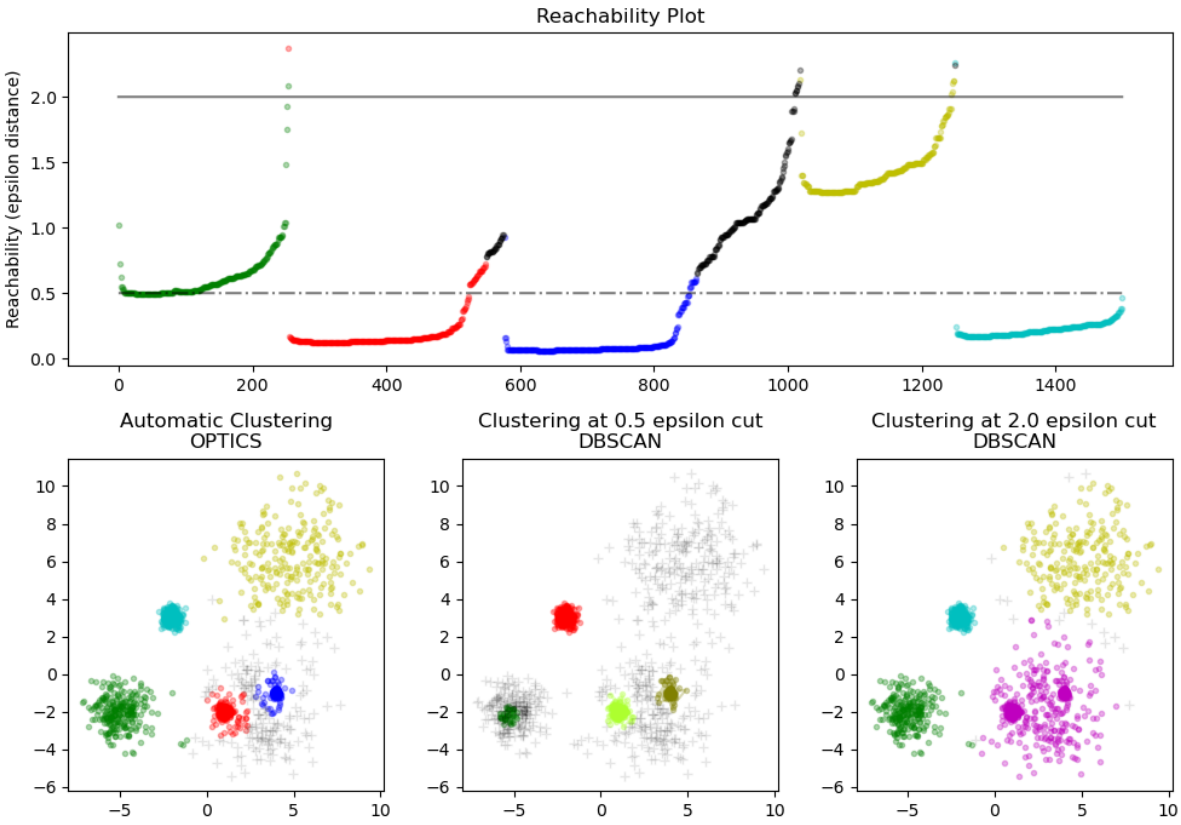
De acordo com Pedregosa et al. (2011) a complexidade do algoritmo de inicialização *k-means ++* é $O(log\ k)$. Sobre o processamento do *K-Means*, a complexidade média é dada por $O(k\ n\ T)$, em que k refere-se ao número de grupos, n é o número de amostras e T consiste no número de iterações. Conforme Arthur e Vassilvitskii (2006) a complexidade do pior caso é dada por $O(n^{(k+2/p)})$, em que p é o número de *features*, ou seja, a quantidade de dimensões (colunas) que estão sendo agrupadas. A complexidade de espaço consumido por memória RAM é de $O(n(p+c))$, onde c é o número de centroides (JIN; HAN, 2010).

3.7.2 *Ordering Points To Identify the Clustering Structure* (OPTICS)

De acordo com Ankerst et al. (1999) o *Ordering Points To Identify the Clustering Structure* (OPTICS) é um algoritmo de *clustering* que não produz o agrupamento explícito de um conjunto de dados. O OPTICS cria uma ordenação aumentada dos dados, representando a estrutura agrupada com base na densidade. A ideia básica do OPTICS é similar ao *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN). Entretanto, o OPTICS aborda uma das principais fraquezas do DBSCAN: detectar *clusters* significativos em dados de densidade variável. Para essa finalidade, a base de dados é ordenada linearmente, com os pontos espacialmente mais próximos tornando-se vizinhos na ordenação. Além disso, armazena-se uma distância especial (ϵ) para considerar que dois pontos pertençam ao mesmo grupo. A Figura 20 ilustra a diferença entre o agrupamento feito pelo OPTICS e o DBSCAN.

Através da Figura 20 pode-se observar como o OPTICS trabalha com agrupamentos de diferentes densidades. Na parte superior da Figura 20 é possível notar o gráfico de alcançabilidade, indicado pela distância ϵ, a qual refere-se à distância máxima entre dois pontos. Nessa primeira parte do gráfico ainda é possível observar que os dados a serem agrupados são de diferentes densidades, no qual cada cor indica uma densidade. Na parte inferior dessa mesma figura é possível observar no canto esquerdo o resultado do agrupamento com o OPTICS. Cada grupo foi formado respeitando a densidade definida na parte superior da figura. Os pontos escuros na parte superior da Figura 20 tornaram-se *outliers* no agrupamento (pontos de cores mais claras e mais distantes dos centros dos agrupamentos). No centro da parte inferior da Figura 20 pode-se observar o agrupamento através do DBSCAN com o parâmetro ϵ no valor de 0,5. Como resultado nota-se uma grande quantidade de *outliers*, além da não identificação do *cluster* na parte superior direita. A parte inferior à direita da Figura 20 ilustra os resultados com o DBSCAN ao definir o valor de ϵ com 2,0. Nesse caso, a quantidade de *outliers* foi pequena. No entanto, com esse

Figura 20 – Diferença de *clustering* entre OPTICS e DBSCAN

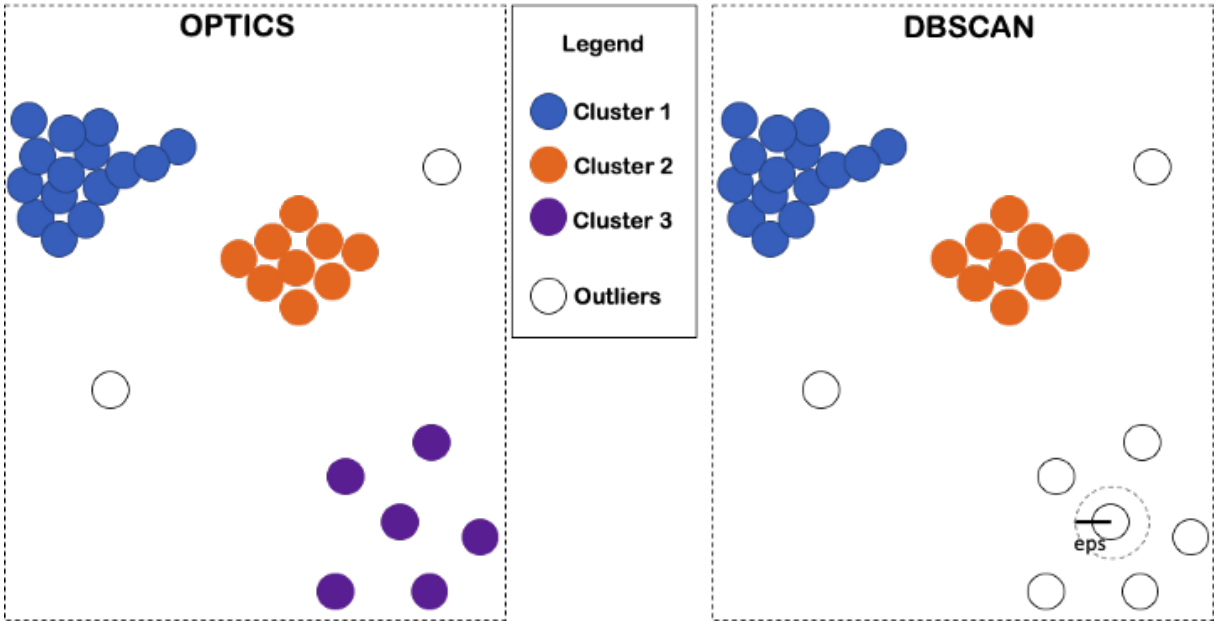


Fonte: Pedregosa et al. (2011)

grande valor de ϵ não foi possível identificar todos os 5 *clusters*, apenas 4 agrupamentos foram identificados. Os grupos ilustrados pelo OPTICS na cor vermelho e azul tornaram-se um único grupo indicado pelo DBSCAN na cor roxo. Através da Figura 21 pode-se observar um outro exemplo da diferença de agrupamento entre o OPTICS e o DBSCAN.

Através da Figura 21 é possível notar como o OPTICS trabalhou melhor as diferentes densidades, formando 3 *clusters* e apenas 2 registros ficaram distantes dos agrupamentos, tornando-se *outliers*. Com o DBSCAN apenas 2 *clusters* foram formados e todos os registros que seriam do agrupamento da cor roxo tornaram-se *outliers*, pois foi definido um pequeno valor de ϵ. Por outro lado, se o valor de ϵ fosse maior, o agrupamento na cor roxo seria identificado, mas possivelmente os *clusters* azul e laranja seriam unidos em um único agrupamento (YUFENG, 2022). Deste modo, com base nos dois exemplos citados, é possível afirmar que os grupos formados pelo OPTICS foram mais coerentes com a realidade ao trabalhar com bases de dados de densidades variadas, sendo possível abstrair com maior precisão relações intrísecas dos dados.

Figura 21 – Vantagem do OPTICS sobre o DBSCAN



Fonte: Yufeng (2022)

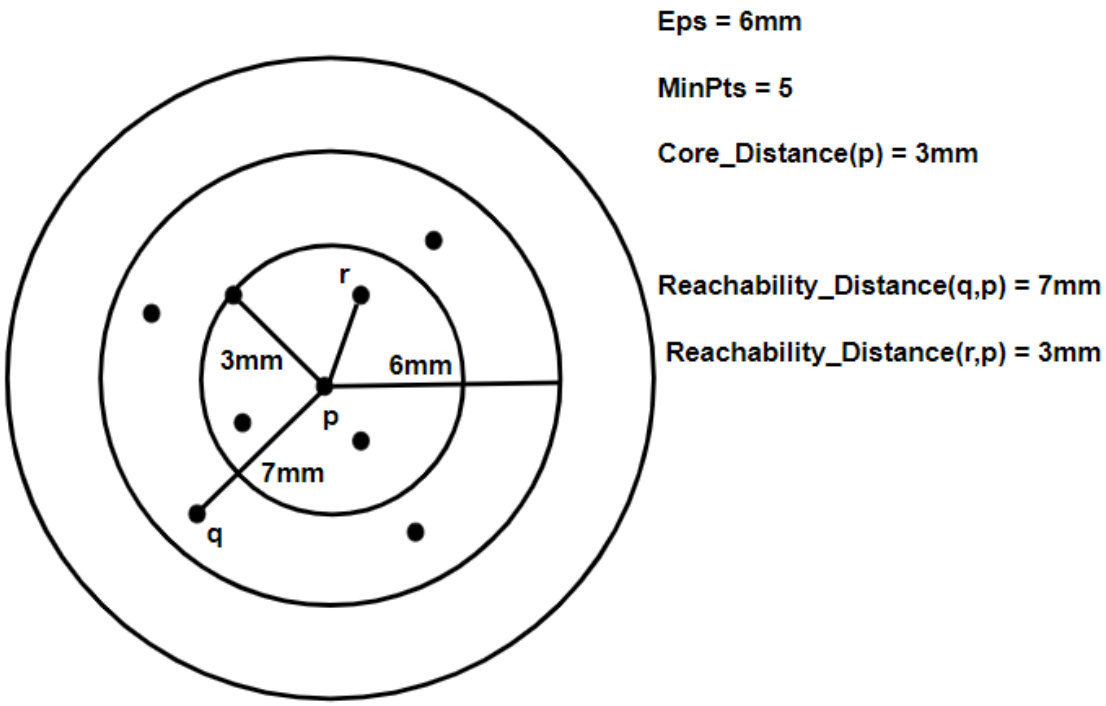
3.7.2.1 Funcionamento do Algoritmo OPTICS

Conforme Gupta (2019), Sinclair (2019) e Pereira (2010), as principais definições relacionadas ao OPTICS são descritas a seguir:

- *Epsilon* (ϵ): consiste na distância máxima entre duas amostras para serem consideradas da mesma vizinhança. No OPTICS é comum definir esse valor como “infinito” para possibilitar a formação de agrupamentos em diferentes escalas.
- Distância de núcelo: É o valor mínimo do raio necessário para classificar um determinado ponto como um ponto do núcleo. Se o ponto dado não for um ponto do núcleo, então o valor será indefinido.
- MinPts: número mínimo de pontos necessários para formar um *cluster*. Um ponto *p* é um ponto de núcleo se no mínimo MinPts pontos (contando o ponto *p*) forem encontrados em sua vizinhança dentro do valor definido por ϵ.
- Distância de acessibilidade: Consiste na distância entre *p* e outro objeto. Se esse outro objeto pertencer ao núcelo, a distância de acessibilidade será o mesmo valor da distância de núcleo. A Figura 22 ilustra os termos mencionados.

Conforme a Figura 22 a circunferência mais interna refere-se ao núcleo. O valor "Eps = 6mm" indica a distância máxima para dois registros estarem no mesmo agrupamento. Pode-se observar que essa distância ultrapassa o núcelo, alcançando pontos na região de borda (que também faz parte do *cluster*). A Figura 22 ainda possibilita verificar que o parâmetro “MinPts = 5” foi atendido, pois o agrupamento tem mais de 5 registros. Por fim,

Figura 22 – Ilustração das principais definições do OPTICS



Fonte: Gupta (2019)

pode-se notar que a distância de alcançabilidade entre pontos dentro do núcleo será o próprio valor do núcleo (distância entre *r* e *p*) e a distância de um elemento do núcleo para um registro de fora (distância de *p* para *q*) será calculada através da métrica de distância definida (distância Euclidiana, por exemplo). A partir dessas definições pode-se explicar os passos para a construção do algoritmo OPTICS.

De acordo com Sinclair (2019) o primeiro passo do algoritmo OPTICS consiste em calcular as distâncias do núcleo em todos os pontos de dados no conjunto. Em seguida, todo o conjunto de dados é percorrido e as distâncias de acessibilidade são atualizadas. Cada ponto é processando apenas uma vez. Ao processar um ponto de dados, a sua posição na ordenação e sua acessibilidade são definidas. O próximo ponto de dados a ser processado será o de menor distância de acessibilidade em relação ao ponto atual. Deste modo, o algoritmo mantem os *clusters* próximos uns dos outros na ordenação de saída. O próximo passo consiste em extrair os agrupamentos a partir das distâncias de acessibilidade.

3.7.2.2 Complexidade do OPTICS

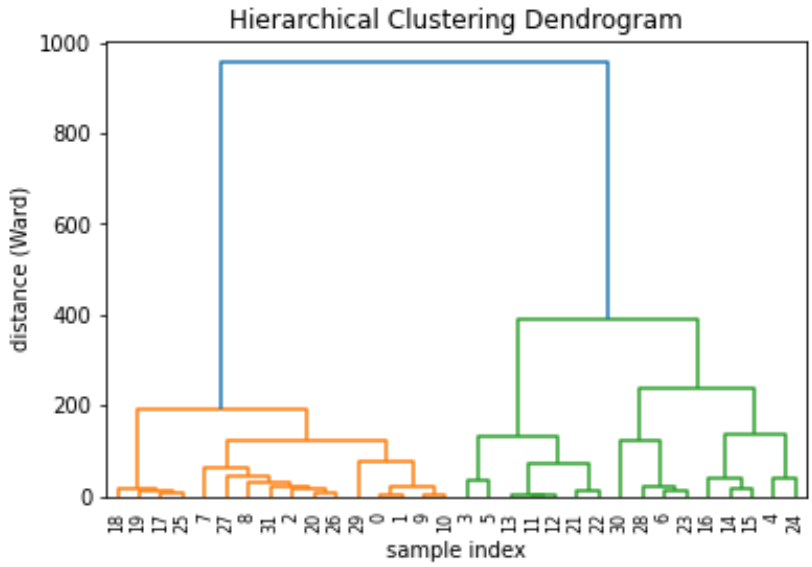
Conforme Ankerst et al. (1999) o tempo geral de execução do OPTICS é de $O(n\ log\ n)$. No entanto, um alto valor de ϵ permite a formação de grupos com diferentes densidades, mas influencia no tempo de execução. Se o valor de ϵ for maior do que a maior distância da base

de dados, a complexidade será quadrática (ANKERST et al., 1999), (KAMIL; AL-MAMORY, 2021).

3.7.3 *Hierarchical Clustering*

O *hierarchical clustering* é um método de análise de agrupamentos que busca a construção de uma hierarquia de *clusters* de uma forma aninhada, mesclando ou dividindo os grupos sucessivamente. Essa hierarquia de agrupamentos é representada como uma árvore (ou dendrograma). A raiz da árvore é o único *cluster* que reúne todas as amostras, as folhas são os grupos com apenas uma amostra (N.RAJALINGAM; K.RANJINI, 2011) e (PEDREGOSA et al., 2011). A Figura 23 ilustra um dendrograma obtido através do *hierarchical clustering*.

Figura 23 – Exemplo de um dendrograma construído através do *hierarchical clustering*



Fonte: Holtz (2018)

Através da Figura 23 é possível observar no eixo "x" a identificação de cada registro. No eixo "y" pode-se notar as distâncias através do método de ligação Ward. Deste modo pode-se avaliar a distância de um agrupamento para outro em diferentes níveis. As estratégias para o *hierarchical clustering* dividem-se principalmente em duas categorias: aglomerativas e divisivas. A primeira categoria consiste em uma abordagem *bottom-up* (de baixo para cima). Cada observação começa no seu *cluster* e os pares de agrupamentos são mesclados ao subir na hierarquia. O dendrograma da Figura 23 é um exemplo desse grupo. A segunda categoria trata-se de uma abordagem “de cima para baixo”. Todas as observações começam em um *cluster* e as divisões são executadas recursivamente à medida que se desce na hierarquia (N.RAJALINGAM; K.RANJINI, 2011) e (PEDREGOSA et al., 2011).

3.7.3.1 Critérios de Ligação *Bottom-up* no *Hierarchical Clustering*

No *hierarchical clustering* cada critério de ligação determina a métrica utilizada na estratégia de unir os agrupamentos. Na abordagem *Bottom-up* pode-se destacar 4 critérios de ligações: Ward, *complete*, *average* e *single*. Conforme Pedregosa et al. (2011) e SciPy (2022) esses critérios de ligações são usados para computar a distância $d(s,t)$ entre dois agrupamentos s e t. O algoritmo começa com uma floresta de *clusters* que ainda não foram usados na hierarquia que está sendo formada. Quando dois *clusters* dessa floresta são combinados em um simples agrupamento u, então s e t são removidos da floresta e u é adicionado à floresta. Quando apenas um agrupamento permanece na floresta, o algoritmo para e esse *cluster* torna-se a raiz. A cada iteração do algoritmo responsável por fazer a ligação, a matriz de distância $d[i,j]$ é atualizada para refletir a distância do *cluster* recém-formado u com os demais *clusters* na floresta. Essa matriz corresponde à distância entre os *clusters* i e j (PEDREGOSA et al., 2011) e (SCIPY, 2022). A seguir são mostradas 4 diferentes formas de calcular os critérios de ligações na abordagem *bottom-up* citadas anteriormente. Para essa finalidade deve-se considerar u e v como *clusters*, sendo que o agrupamento u é a combinação de s e t. Sendo v um *cluster* remanescente na floresta, que não seja u, os seguintes métodos calculam a distância entre o mais novo grupo formado u e v:

- *Ward*: de acordo com Pedregosa et al. (2011) e SciPy (2022) o critério de ligação Ward minimiza a soma das diferenças quadradas dentro de todos os *clusters*. Trata-se de uma abordagem de minimização de variância voltada para o agrupamento hierárquico, que mescla recursivamente um par de *clusters* se eles aumentarem de forma mínima a variação interna do agrupamento. Uma nova entrada $d(u,v)$ é calculada através de:

$$d(u,v) = \sqrt{\frac{|v|+|s|}{T}d(v,s)^2 + \frac{|v|+|t|}{T}d(v,t)^2 - \frac{|u|}{T}d(s,t)^2} \tag{2}$$

  no qual: u é o mais novo *cluster* unido, consistindo dos agrupamentos s e t. A variável v é um *cluster* não utilizado na floresta $T = |v| + |s| + |t|$. A cardinalidade deste algoritmo é dada por $|*|$. Este método também é conhecido como algoritmo incremental.
- *Complete*: Conforme Pedregosa et al. (2011) e SciPy (2022) o critério de ligação *complete* minimiza a distância máxima entre observações de pares de *clusters*. Uma nova entrada $d(u,v)$ é calculada por:

$$d(u,v) = max(dist(u[i], v[j])) \tag{3}$$

  Para todos os pontos i no *cluster* u e todos os pontos j no agrupamento v. Este método também é conhecido como algoritmo do ponto mais distante ou algoritmo Voor Hees.

- *Average*: De acordo com Pedregosa et al. (2011) e SciPy (2022) o critério de ligação *average* minimiza a média de distâncias entre todas observações de pares de *clusters*. Uma nova entrada $d(u, v)$ é calculada com:

$$d(u, v) = \sum_{ij} \frac{d(u[i], v[j])}{(|u| * |v|)} \tag{4}$$

  Para todos os pontos i e j, onde u e $|v|$ são as cardinalidades dos *clusters* u e v, respectivamente. Este método também é conhecido como algoritmo UPGMA.
- *Single*: conforme Pedregosa et al. (2011) e SciPy (2022) o método de ligação *single* minimiza a distância entre as observações mais próximas de pares de *clusters*. Uma nova entrada $d(u, v)$ é calculada do seguinte modo:

$$d(u, v) = min(dist(u[i], v[j])) \tag{5}$$

  Para todos os pontos i no *cluster* u e todos os pontos j no agrupamento v. Este método também é conhecido como algoritmo do ponto mais próximo.

A Figura 24 ilustra as diferenças de agrupamentos desses 4 critérios.

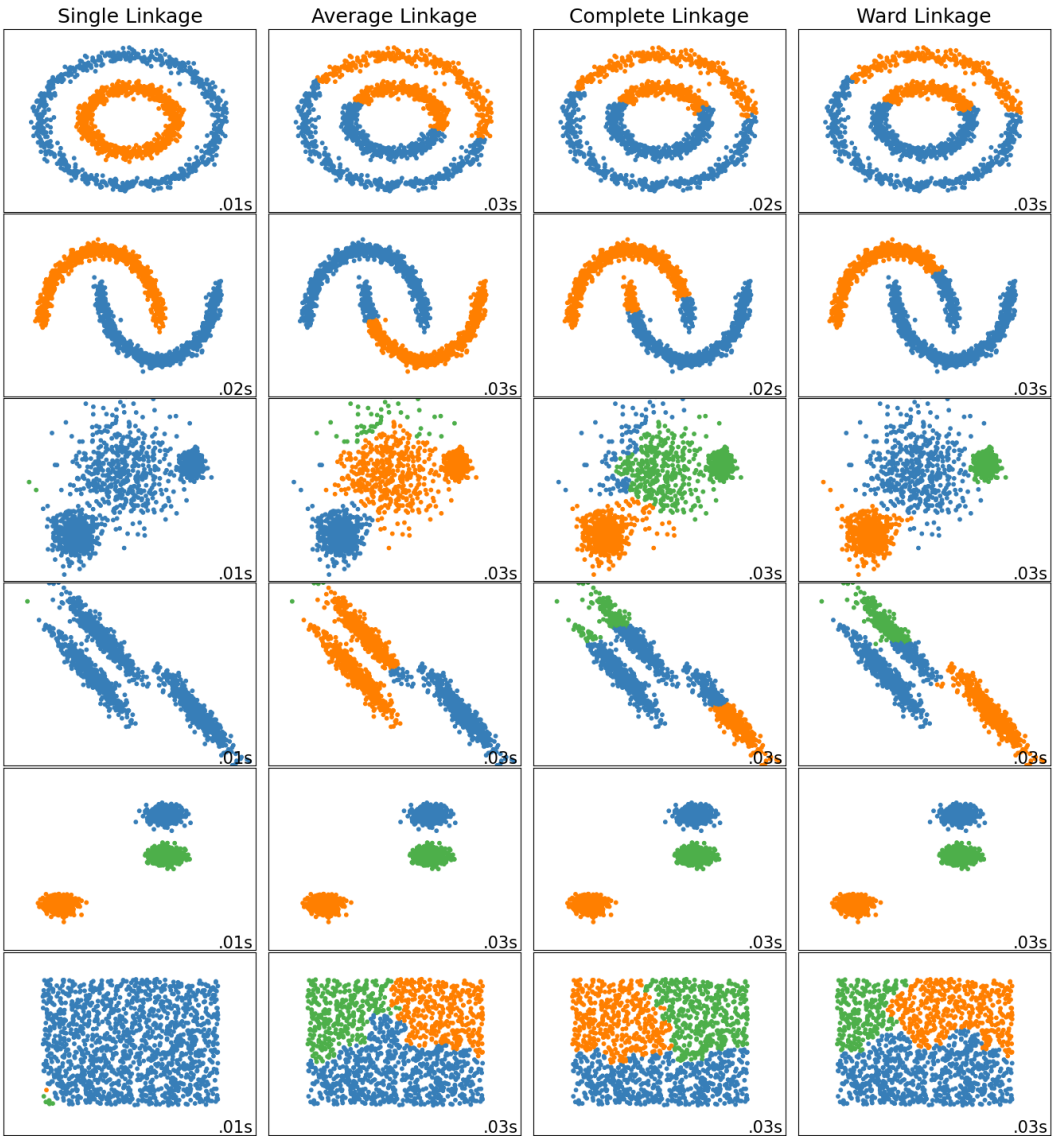
Conforme Pedregosa et al. (2011) as principais observações que podem ser realizadas sobre as diferenças dos critérios são:

- O método de ligação *single* é mais rápido e funciona bem em dados que não são globulares. No entanto, este critério de ligação não funciona bem ao trabalhar em bases de dados que possuem ruídos;
- Os critérios de ligações *average* e *complete* funcionam bem em *clusters* globulares sem ruídos;
- O Ward é o método de ligação mais efetivo ao trabalhar com dados que tenham ruídos.

3.7.3.2 Complexidade do *Hierarchical Clustering*

Conforme Roy e Chakrabarti (2017) a complexidade de tempo do *hierarchical clustering* é dada por $O(kn^2)$, em que o número de elementos a serem agrupados é dado por n e o número de *clusters* é dado por k. De acordo com Patlolla (2018) o espaço de memória RAM requerido pelo *hierarchical clustering* é muito alto devido a necessidade de armazenamento da matriz de similaridades. Deste modo, o espaço de memória RAM requerido é $O(n^2)$.

Figura 24 – Diferentes tipos de ligações



Fonte: Pedregosa et al. (2011)

Capítulo 4

Desenvolvimento de Metodologias Baseadas em Matrizes de Distâncias para Representar Interações de Proteínas Visando a Comparação de Similaridades Atômicas

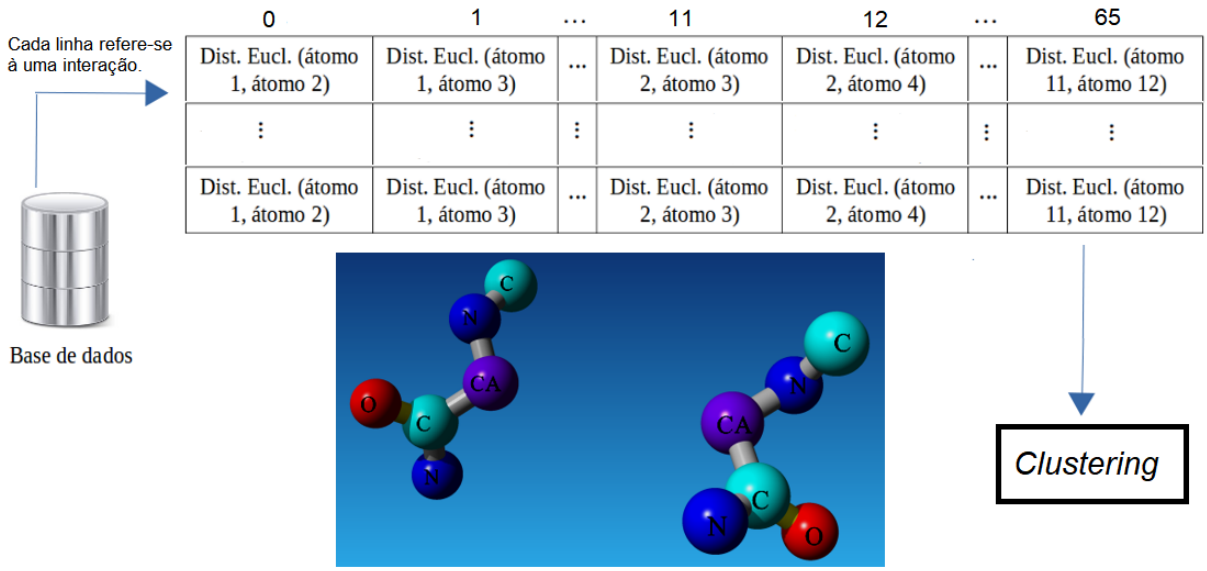
Este Capítulo relata o desenvolvimento de metodologias baseadas em matrizes de distâncias para verificar as similaridades de proteínas. A Seção 4.1 refere-se ao desenvolvimento da Matriz de Distâncias Completa (MDC). A Seção 4.2 é sobre a Matriz de Distâncias Reduzida cujos Centroides são Carbonos Alfa (MDRCCA). A Seção 4.3 refere-se à Matriz de Distâncias Reduzida a partir de um Ponto entre os Carbonos Alfas (MDRPCA). A Seção 4.4 refere-se à Matriz de Ângulos (MA). A Seção 4.5 é sobre a Matriz de Distâncias Mista (MDCM). A Seção 4.6 é referente à Matriz de Pontos Médios (MPM).

4.1 Matriz de Distâncias Completa (MDC)

De acordo com Monteiro (2017) a MDC é uma metodologia baseada em matriz de distâncias que pode trabalhar juntamente com técnicas de *clustering* para possibilitar o agrupamento de interações de proteínas no formato do PDB, conforme as diferenças de distâncias atômicas. A Figura 25 ilustra o fluxograma desta metodologia.

De acordo com a Figura 25, o fluxograma da MDC tem como primeira etapa calcular a distância Euclidiana para as coordenadas x, y e z entre todos os átomos de uma mesma interação proteica. Os cálculos realizados para cada interação de proteína são armazenados em uma linha da matriz de distâncias e cada coluna refere-se à um valor de distância

Figura 25 – Fluxograma da MDC



Fonte: Adaptado de Monteiro (2017) e Monteiro, Dias e Rodrigues (2020)

Euclidiana calculado. Como cada interação da base de dados contém 12 átomos, foram obtidos 66 valores para representar cada registro. A Figura 25 contém ainda a ilustração tridimensional de uma interação proteica para facilitar a compreensão de como a matriz MDC é construída. A primeira coluna da matriz armazena o cálculo da distância Euclidiana do primeiro átomo de carbono do lado esquerdo (átomo de carbono que está na parte superior da Figura 25) com o átomo de nitrogênio que faz ligação com este carbono. A segunda coluna da matriz refere-se ao cálculo da distância Euclidiana entre o primeiro átomo de carbono (átomo utilizado no cálculo anterior) e o carbono alfa do lado esquerdo. Deste modo, de maneira sucessiva realiza-se o cálculo da distância Euclidiana entre todos os átomos desta interação, até que a última coluna seja a distância Euclidiana do átomo de nitrogênio com o átomo de oxigênio ambos do lado direito da Figura 25. O cálculo da distância Euclidiana entre os átomos de um mesmo arquivo interagente, bem como a sua representação em matriz de distâncias, têm como objetivo estabelecer a orientação espacial das estruturas tridimensionais. Obter a orientação espacial dos átomos de cada interação mostra-se importante para a metodologia obter resultados coerentes com algoritmos que realizam a sobreposição atômica. Por fim, utiliza-se um algoritmo de *clustering* para agrupar a matriz de distâncias construída, com o objetivo de formar *clusters* que tenham interações com conformações tridimensionais similares (MONTEIRO, 2017). O Algoritmo 1 ilustra o cálculo da distância Euclidiana entre os átomos de cada interação proteica para construir a matriz de distâncias da MDC.

De acordo com o Algoritmo 1, para construir a matriz MDC é necessário receber como entrada a matriz "MtInterac" . Essa estrutura de dados contém as coordenadas atômicas x,

Algoritmo 1: Algoritmo para construir a matriz MDC

```
Entrada: MtInterac[ ][ ];
Saida: MatrizMDC[ ][ ];
Função ConstroiMatrizMDC
indiceColuna ← 0;
para cada registro ∈ MtInterac[ ][ ] faça
    para i ← 0 até 11 faça
        para j ← i + 1 até 11 faça
            MatrizMDC[registro.id][indiceColuna] ←
              EuclDist(MtInterac[registro.id][i[0, 1, 2]], MtInterac[registro.id][j[0, 1, 2]]);
            indiceColuna + +;
        fim
        indiceColuna ← 0;
    fim
fim
retorna MatrizMDC[ ][ ];
```

y e z de cada interação proteica extraída do PDB. Cada linha da "MtInterac" refere-se à uma interação e cada coluna trata-se de um átomo desta interação, na qual cada índice desta matriz de entrada armazena um vetor de três posições referentes aos valores das coordenadas x, y e z de cada átomo. O algoritmo para construir a matriz MDC necessita de 3 comandos de repetições aninhados, indicados por **for**. O primeiro **for** é responsável por garantir a realização deste procedimento para todas as interações. Os dois **for** internos realizam o cálculo da distância Euclidiana entre todos os átomos de uma mesma interação e os valores são armazenados na variável "MatrizMDC" na qual cada linha refere-se à uma interação e cada coluna refere-se a um valor de distância Euclidiana. No final do Algoritmo 1 pode-se observar que "MatrizMDC" é retornada, podendo ser utilizada pelo *K-Means Clustering*.

Conforme Monteiro (2017) e Monteiro, Dias e Rodrigues (2020), os primeiros experimentos da MDC juntamente com o *K-Means Clustering* foram realizados com 16.383 arquivos de interações de pontes dissulfeto, extraídas do PDB pela ferramenta RID. Esta metodologia identificou pequenas diferenças de distâncias atômicas entre os registros, classificando-os em *clusters* distintos, conforme ilustrado pela Figura 26.

Conforme a Figura 26, as posições centrais de ambos os *clusters* (31,32 e 33), tiveram maiores diferenças de valores, quando comparadas com os registros do outro *cluster*. Os valores contidos nas demais posições foram mais similares.

Após a formação dos grupos Monteiro (2017) e Monteiro, Dias e Rodrigues (2020) ainda realizaram sobreposições atômicas através do LSQKAB, entre todos os registros que ficaram no mesmo *cluster*. Os resultados foram comparados com os algoritmos aCSM e RID. Os resultados obtidos foram satisfatórios, apresentando uma precisão superior aos algoritmos

Figura 26 – Exemplo de alguns *clusters* formados pela MDC

Alguns registros contidos no *cluster* 1

Id	0	1	…	31	32	33	…	64	65
9.243	1,336	2,455	…	1,333	6,821	5,562	…	1,333	2,255
9.244	1,337	2,456	…	1,334	6,852	5,589	…	1,334	2,258
9.248	1,337	2,460	…	1,335	6,880	5,613	…	1,334	2,261

Alguns registros contidos no *cluster* 251

Id	0	1	…	31	32	33	…	64	65
561	1,340	2,455	…	1,323	4,954	4,580	…	1,442	2,305
557	1,334	2,438	…	1,327	4,965	4,779	…	1,334	2,257
8.515	1,328	2,442	…	1,327	4,925	4,833	…	1,331	2,262

Fonte: Adaptado de Monteiro, Dias e Rodrigues (2020)

comparados. Também foi constatado um tempo menor para realizar os experimentos.

4.1.1 Complexidade de Tempo da MDC

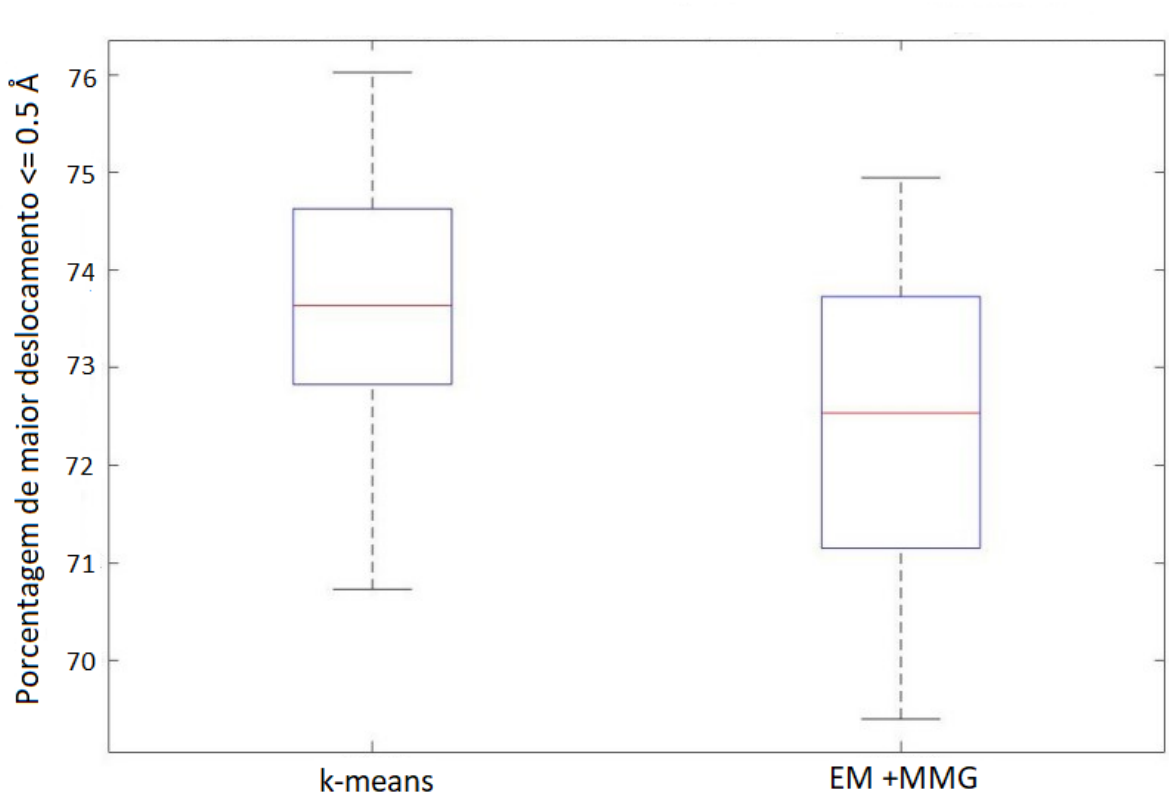
A complexidade de tempo necessária para representar uma interação proteica em uma linha da matriz MDC é de $O(b((n^2 - n)/2))$, em que b é o tamanho da base de dados e n é a quantidade de átomos da interação proteica. Todas as bases de dados avaliadas nesse trabalho contém 12 átomos. A complexidade quadrática refere-se a duas estruturas de repetições necessárias para calcular a distância Euclidiana entre todos os pares de átomos. No entanto, a complexidade tem uma subtração por n, pois não é necessário calcular a distância de um átomo em relação a si mesmo. A divisão por 2 ocorre por somente ser necessário calcular a distância Euclidiana entre o mesmo par de átomo apenas uma vez durante a execução das estruturas de repetições aninhadas.

4.1.2 Comparação dos Resultados da MDC ao Substituir o *K-Means Clustering* pelo *Gaussian Mixed Model*

A primeira técnica de agrupamento utilizada na MDC foi o *k-means clustering* (MONTEIRO, 2017) (MONTEIRO; DIAS; RODRIGUES, 2020). Através de Amorim (2020) foram realizados testes que consistiram em comparar a precisão e a performance computacional da MDC ao substituir o *k-means clustering* pelo o Modelo de Mistura de Gaussianas (GMM) ajustado pelo algoritmo Expectativa-Maximização. Amorim (2020) realizou experimentos que consistiram em representar 16.383 interações de pontes dissulfeto através da MDC e realizar agrupamentos com essas duas técnicas de *clustering*. A Figura 27 ilustra uma das comparações realizadas, a qual avalia a porcentagem total de sobreposições atômicas

que tiveram até 0,5 Å de maior deslocamento, dentre as interações que ficaram no mesmo agrupamento.

Figura 27 – Comparação da precisão da MDC com o *k-means clustering* e com o *Gaussiana Mixed Model*



Fonte: Adaptado de Amorim (2020)

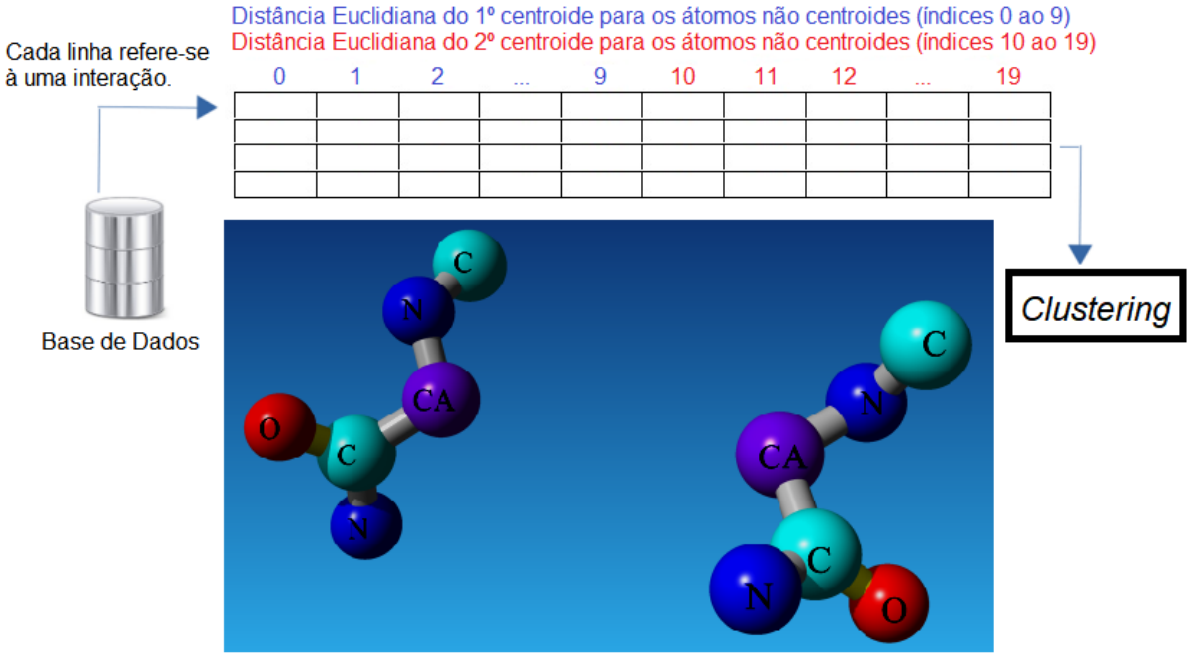
Conforme a Figura 27, pode-se notar através dos *boxplots* que a precisão da MDC combinada com o *K-Means clustering* foi superior do que os resultados da combinação entre MDC e GMM. Além dos experimentos de precisão, Amorim (2020) avaliou o tempo necessário para realizar o agrupamento com essas diferentes técnicas de *clustering*. Como resultado constatou-se que para o cenário avaliado o *K-Means clustering* é aproximadamente 3 vezes mais rápido do que o GMM.

4.2 Matriz de Distâncias Reduzida Cujos Centroides são os Carbonos Alfas (MDRCCA)

A partir da MDC, foi desenvolvida uma nova metodologia, a Matriz de Distâncias Reduzida cujos Centroides são os Carbonos Alfas (MDRCCA). O funcionamento desta metodologia consiste em definir os átomos de carbono alfa para serem os centroides e construir a matriz de distâncias a partir do cálculo da distância Euclidiana entre as coordenadas x, y e z destes

centroides em relação às coordenadas dos demais átomos. Os átomos de carbono alfa foram escolhidos para serem os centroides por estarem no centro da cadeia principal e também pelo fato da cadeia lateral ser originada à partir destes átomos. O fluxograma da MDRCCA é ilustrado pela figura Figura 28.

Figura 28 – Fluxograma da MDRCCA



Fonte: Adaptado de Monteiro, Dias e Rodrigues (2021)

De acordo com a Figura 28, cada interação de proteína foi representada através de uma linha da matriz MDRCCA. Como cada interação da base de dados utilizada possui 12 átomos, dos quais 2 são de carbonos alfa (considerados cetroides), cada linha foi construída com 20 colunas (índices 0 ao 19). Na construção da MDRCCA as 10 primeiras posições (do índice 0 ao 9) referem-se ao cálculo da distância Euclidiana para as coordenadas x, y e z do primeiro átomo de carbono alfa (primeiro centroide) para as mesmas coordenadas dos outros 10 átomos que não são centroides. As outras 10 posições referem-se ao cálculo da distância Euclidiana do outro átomo de carbono alfa (segundo centroide), em relação aos outros átomos que não são centroides (do índice 10 ao 19). Do mesmo modo que realizado em Monteiro, Dias e Rodrigues (2020), o cálculo da distância Euclidiana e a representação em matriz, têm o objetivo de estabelecer a orientação espacial dos átomos contidos no arquivo de interação (MONTEIRO; DIAS; RODRIGUES, 2021). Posteriormente, a matriz MDRCCA pode ser agrupada por técnicas de *clustering*. O Algoritmo 2 mostra a construção da MDRCCA.

De acordo com o Algoritmo 2 a função "ConstroiMatrizMDRCCA" necessita do mesmo

Algoritmo 2: Algoritmo para construir a matriz MDRCCA

```
Entrada: MtInterac[ ][ ];
Saida: MatrizMDRCCA[ ][ ];
Function ConstroiMatrizMDRCCA
para cada registro ∈ MtInterac[ ][ ] faça
    indiceMDRCCA ← 0;
    //Calculos do primeiro centroide (índice 2)
    para i ← 0 até 11 faça
        MatrizMDRCCA[registro.id][indiceMDRCCA] ←
         EuclDist(MtInterac[registro.id][2[0, 1, 2]], MtInterac[registro.id][i[0, 1, 2]])
         indiceMDRCCA + +
        se i == 2||i == 8 então
            continue
        fim
    fim
    //Calculos do segundo centroide (índice 8)
    para i ← 0 até 11 faça
        MatrizMDRCCA[registro.id][indiceMDRCCA] ←
         EuclDist(MtInterac[registro.id][8[0, 1, 2]], MtInterac[registro.id][i[0, 1, 2]])
         indiceMDRCCA + +
        se i == 2||i == 8 então
            continue
        fim
    fim
fim
retorna MatrizMDRCCA[ ][ ]
```

parâmetro de entrada da MDC, a matriz com as interações, representada pela variável "MtInterac". O algoritmo tem um comando de repetição **for** que é responsável por garantir que todas as interações armazenadas na variável de entrada "MtInterac" participarão do cálculo de distância Euclidiana. Dentro desta estrutura de repetição pode-se observar outros dois comandos **for**, responsáveis por calcular a distância Euclidiana dos átomos de carbonos alfa em relação aos demais átomos e armazenar os valores na "MatrizMDRCCA". É possível observar que o primeiro **for** interno à estrutura de repetição principal possibilita o calculo da distância Euclidiana do primeiro centroide que está localizado na terceira coluna da matriz de entrada (índice 2), com os demais átomos que não são centroides. Por este motivo foi inserido o comando **if** para evitar o cálculo da distância Euclidiana entre os átomos das colunas de índice 2 e 8 (átomos de carbono alfa). A segunda estrutura de repetição interna possibilita o cálculo da distância Euclidiana do segundo centroide com os demais átomos que não são centroides. Ao fim do processamento, é retornada a matriz MDRCCA que será utilizada no procedimento de *clustering*. A Figura 29 ilustra alguns *clusters* que foram formados em um experimento da MDRCCA ao trabalhar com uma base de dados de pontes dissulfeto. Mais especificamente, foram ilustrados alguns registros dos *clusters* 1 e 409.

Figura 29 – Exemplo de alguns *clusters* formados pela MDRCCA

Alguns registros contidos no *cluster* 1

Id	0	1	...	9	10	11	...	18	19
9.243	2,455	1,470	…	5,191	7,485	6,203	…	2,398	2,435
9.244	2,457	1,468	…	5,160	7,515	6,220	…	2,398	2,439
15.298	2,422	1,462	…	5,190	7,605	6,310	…	2,405	2,450
15.299	2,429	1,460	…	5,340	7,630	6,365	…	2,404	2,435

Alguns registros contidos no *cluster* 409

Id	0	1	...	9	10	11	...	18	19
15.055	2,386	1,441	…	4,730	4,191	4,346	…	2,427	2,415
15.056	2,377	1,439	…	4,778	4,130	4,317	…	2,412	2,401
15.876	2,433	1,459	…	4,876	4,171	4,410	…	2,392	2,426
15.877	2,434	1,455	…	4,850	4,082	4,325	…	2,392	2,425

Fonte: Adaptado de Monteiro, Dias e Rodrigues (2021)

Conforme a Figura 29, os valores das posições iniciais (índices 0 e 1) e os valores das posições finais (índices 18 e 19) foram ligeiramente diferentes entre os registros de um *cluster* para as pontes dissulfeto do outro grupo. Por outro lado, as posições centrais (9, 10 e 11) foram muito diferentes para registros de *clusters* distintos. As diferenças entre as 20 posições de cada registro contribuíram para o algoritmo de *clustering* formar os grupos.

4.2.1 Complexidade de Tempo da MDRCCA

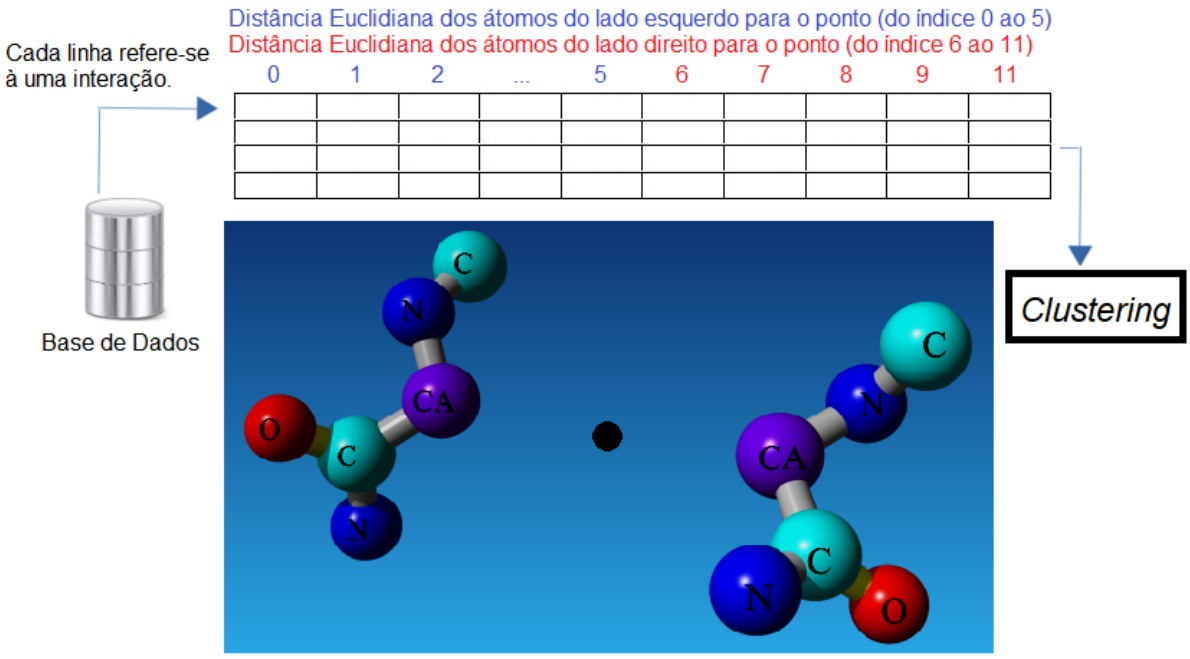
A complexidade de tempo necessária para representar uma interação proteica em uma linha da matriz MDRCCA é de $O(b(nc - cc))$, em que b refere-se ao tamanho da base de dados e n é a quantidade de átomos contida na interação da proteína e c refere-se a quantidade de átomos definidos como centroides. Para construir a matriz MDRCCA calcula-se diretamente a distância Euclidiana dos átomos definidos como centroides em relação aos demais átomos. Deste modo, não é necessário utilizar estruturas de repetições aninhadas, evitando que ele tenha a complexidade quadrática.

4.3 Matriz de Distâncias Reduzida através de um Ponto Entre os Carbonos Alfa (MDRPCA)

A Matriz de Distâncias Reduzida através de um Ponto Entre os Carbonos Alfa (MDRPCA) tem como fundamento o cálculo de um ponto entre os átomos de carbono alfa. A partir deste ponto fictício, calcula-se a distância Euclidiana entre todos os átomos contidos no arquivo de interação em relação a este ponto. Deste modo, são necessárias apenas 12 colunas da matriz para armazenar os cálculos de distância Euclidiana para cada arquivo

de interação. O fluxograma da MDRPCA é ilustrado pela Figura 30 (MONTEIRO; DIAS; RODRIGUES, 2021).

Figura 30 – Fluxograma da MDRPCA



Fonte: Adaptado de Monteiro, Dias e Rodrigues (2021)

Conforme a Figura 30, para cada arquivo de interação contido na base de dados, calcula-se primeiro o ponto entre os dois carbonos alfas deste arquivo. Em seguida, calcula-se a distância Euclidiana de todos os átomos do arquivo interagente para este ponto. As distâncias do ponto para os átomos contidos no lado esquerdo da interação são armazenadas nas 6 primeiras colunas da matriz. Os cálculos referentes aos átomos do lado direito da interação em relação ao ponto, são armazenados nas 6 últimas posições (MONTEIRO; DIAS; RODRIGUES, 2021). O Algoritmo 3 detalha o processo de construção da matriz MDRPCA.

De acordo com o Algoritmo 3 pode-se construir a matriz MDRPCA através da mesma variável de entrada utilizada nas construções das matrizes MDC e MDRCCA. Na construção da MDRPCA pode-se observar que dentro do **for** principal, para cada interação calcula-se primeiro o ponto entre os átomos de carbono alfa, indicados pelos índices 2 e 8. Em seguida, calcula-se a distância Euclidiana entre todos os átomos e o ponto calculado anteriormente, no qual cada valor é armazenado em uma coluna da matriz MDRPCA. Ao fim do processamento, esta matriz é utilizada em *clustering*. A Figura 31 ilustra alguns *clusters* formados pela MDRPCA juntamente com o *k-means clustering*.

De acordo com a Figura 31, ao comparar alguns registros dos *clusters* 1 e 55, pode-se

Algoritmo 3: Algoritmo para construir a matriz MDRPCA

Entrada: $MtInterac[\,][\,];$
Saida: $MatrizMDRPCA[\,][\,];$
Function ConstroiMatrizMDRPCA
para *cada* $registro \in MtInterac[]$ **faça**
  $pontoEntreCAs[0,1,2] \leftarrow$
   $CaulcPonto(MtInterac[registro.id][2[0,1,2]], MtInterac[registro.id][8[0,1,2]]);$
  para $i \leftarrow 0$ *até 11* **faça**
    $MatrizMDRPCA[registro.id][i] \leftarrow$
     $EuclDist(MtInterac[registro.id][i[0,1,2]], pontoEntreCAs[0,1,2])$
  fim
fim
retorna $MatrizMDRPCA[\,][\,];$

Figura 31 – Exemplos de alguns *clusters* formados pela MDRPCA

Alguns registros contidos no *cluster* 1

Id	0	1	...	4	5	6	...	10	11
6.135	4,441	3,619	…	4,705	5,600	4,242	…	4,702	4,594
6.136	4,447	3,630	…	4,694	5,633	4,236	…	4,733	4,595
11.290	4,428	3,797	…	4,397	5,428	5,595	…	4,634	4,835
11.291	4,412	3,789	…	4,411	5,422	4,577	…	4,641	4,838

Alguns registros contidos no *cluster* 55

Id	0	1	...	4	5	6	...	10	11
11.071	2,966	2,190	…	4,750	5,027	4,034	…	4,355	4,385
11.075	3,027	2,222	…	4,728	5,018	4,043	…	4,320	4,424
11.077	3,073	2,299	…	4,739	5,001	4,093	…	4,281	4,453
1.393	3,147	2,135	…	4,586	5,140	4,311	…	4,074	4,980

Fonte: Adaptado de Monteiro, Dias e Rodrigues (2021)

observar uma maior diferença nos valores das duas primeiras posições de cada grupo. As posições centrais e finais apresentaram ligeiras diferenças dos arquivos de interações de um grupo para o outro.

4.3.1 Complexidade de Tempo da MDRPCA

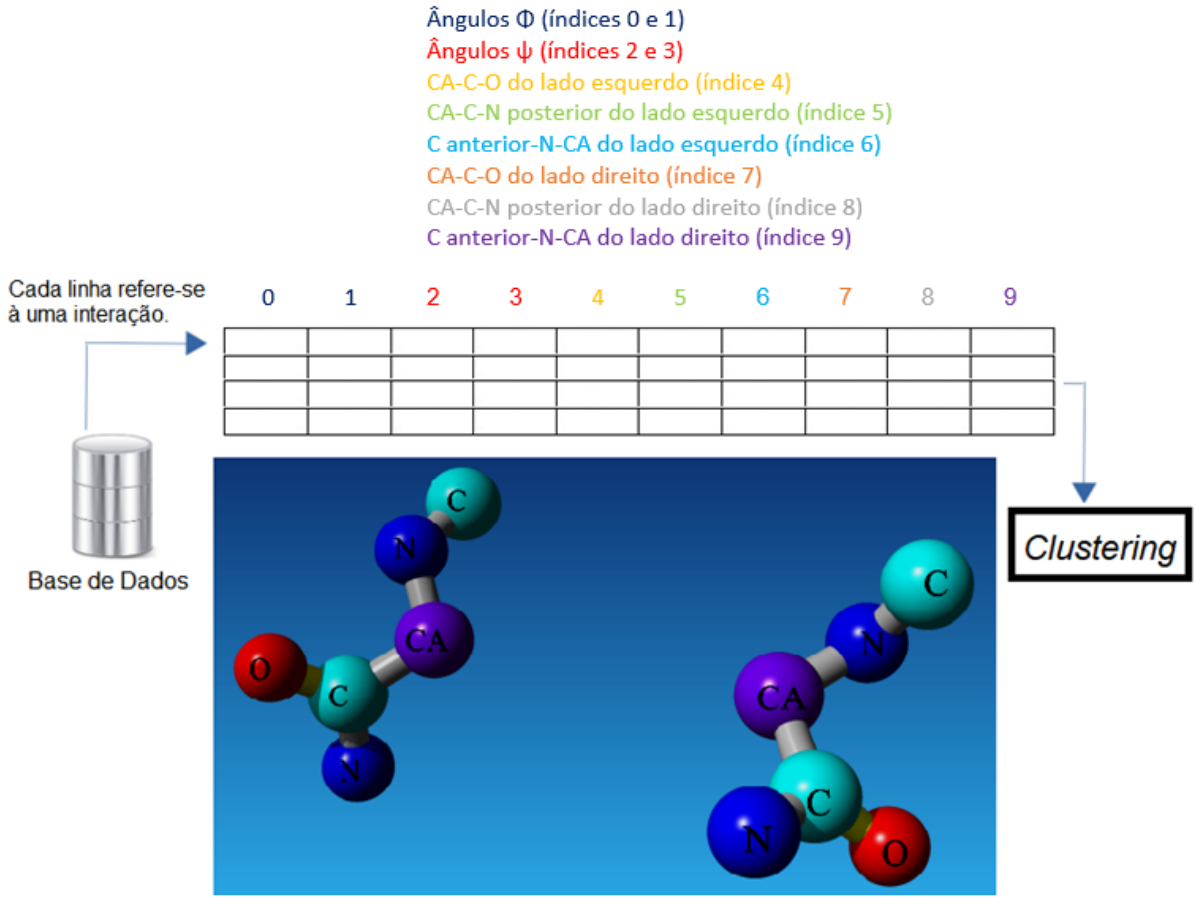
O primeiro passo para representar uma interação proteica em uma linha da matriz MDRPCA é calcular o ponto entre os carbonos alfas. Esse cálculo é feito de um modo direto, com a complexidade de $O(1)$ para cada interação, sendo $O(b)$ ao considerar a base de dados completa, em que b é o tamanho da base de dados. O próximo passo consiste em calcular a distância Euclidiana entre esse ponto calculado e todos os átomos da interação. Deste modo, a complexidade é de $O(bn)$, em que n é a quantidade de átomos. A MDRPCA é a forma mais rápida de representar uma interação proteica dentre as metodologias apresentadas

neste trabalho.

4.4 Matriz de Ângulos (MA)

A Matriz de Ângulos (MA) consiste em uma técnica de representar interações de proteínas utilizando somente valores dos ângulos, na medida de radianos, de cada interação proteica. Trata-se de uma matriz em que cada coluna refere-se a um determinado ângulo e cada linha indica um arquivo de interação. Através da MA cada interação é representada por 10 valores, sendo 4 referentes aos ângulos *phi* e *psi*. Os outros 6 valores referem-se aos seguintes ângulos, calculados para os lados esquerdo e direito da interação: CA-C-O, CA-C-N (posterior) e C (anterior)-N-CA. Os cálculos desses ângulos foram feitos pelas funções *get_phi_psi_list* e *calc_angle*, fornecidas pela biblioteca Biopython (BIOPYTHON, 2023). Deste modo, a matriz de ângulos pode ser ilustrada pela Figura 32.

Figura 32 – Matriz de Ângulos



Fonte: Elaborada pelo autor (2023)

Como pode ser observado pela Figura 32, as duas primeiras posições da MA referem-se aos ângulos ϕ, os dois próximos índices são referentes aos ângulos ψ. As próximas 3 posições

referem-se respectivamente aos ângulos CA-C-O, CA-C-N (posterior) e C (anterior)-N-CA do lado esquerdo. As últimas três posições referem-se a esses ângulos, mas para os átomos do lado direito. O algoritmo 4 ilustra a construção da matriz MA.

Algoritmo 4: Algoritmo para construir a matriz MA

Entrada: $MtInterac[\,][\,]$;
Saída: $MatrizMA[\,]$;
Function ConstroiMatrizMA
para *cada* $registro \in MtInterac[\,][\,]$ **faça**
$MA[registro.id][0,1] \leftarrow CaulculaAnguloPi(MtInterac[registro.id][:])$;
$MA[idRegistro][2,3] \leftarrow CaulculaAnguloPsi(MtInterac[registro.id][:])$;
$MA[registro.id][4] \leftarrow CalcAngle(MtInterac[registro.id][2[0,1,2]], MtInterac[registro.id][3[0,1,2]], MtInterac[registro.id][5[0,1,2]])$;
$MA[registro.id][5] \leftarrow CalcAngle(MtInterac[registro.id][2[0,1,2]], MtInterac[registro.id][3[0,1,2]], MtInterac[registro.id][6[0,1,2]])$;
$MA[registro.id][6] \leftarrow CalcAngle(MtInterac[registro.id][0[0,1,2]], MtInterac[registro.id][1[0,1,2]], MtInterac[registro.id][2[0,1,2]])$;
$MA[registro.id][7] \leftarrow CalcAngle(MtInterac[registro.id][8[0,1,2]], MtInterac[registro.id][6[0,1,2]], MtInterac[registro.id][10[0,1,2]])$;
$MA[registro.id][8] \leftarrow CalcAngle(MtInterac[registro.id][9[0,1,2]], MtInterac[registro.id][11[0,1,2]], MtInterac[registro.id][6[0,1,2]])$;
$MA[registro.id][9] \leftarrow CalcAngle(MtInterac[registro.id][6[0,1,2]], MtInterac[registro.id][7[0,1,2]], MtInterac[registro.id][8[0,1,2]])$;
fim
retorna $MatrizMA[\,][\,]$;

Conforme o algoritmo da construção da MA pode-se notar que é necessário apenas uma estrutura de repetição. Também é possível observar que os ângulos são calculados através de acesso direto às respectivas posições da matriz de entrada. A Figura 33 ilustra como a MA difere registros de dois agrupamentos distintos.

Conforme a Figura 33 pode-se observar que ao longo das 10 colunas da MA, os registros dos *clusters* 177 e 153 tiveram maiores diferenças nas posições 1, 2 e 3. Para as demais posições, as interações obtiveram valores mais próximos.

4.5 Matriz de Distâncias Completa Mista (MDCM)

A Matriz de Distâncias Completa Mista (MDCM) consiste na concatenação horizontal das metodologias MDC, MDRPCA e MA. Esta metodologia foi desenvolvida para verificar se a união das informações obtidas por essas técnicas resultaria em agrupamentos com acurácias satisfatórias. A MDC obtém informações da distância Euclidiana entre todos os átomos da interação proteica. A MDRPCA tem as informações da distância Euclidiana de todos os átomos da interação em relação a um ponto entre os carbonos alfas. Os ângulos são importantes para determinar a conformação nativa da proteína. Ou seja, torna-se viável

Figura 33 – Exemplo de alguns *clusters* formados pela MA

Alguns registros do *cluster* 177

Id	0	1	2	3	4	5	6	7	8	9
757	-1,70	0,376	-1,51	-0,11	2,114	2,018	2,127	2,107	0,220	2,097
758	-1,69	0,3682	-1,54	-0,06	2,140	2,011	2,142	2,100	0,2119	2,135
5.852	-2,08	0,409	-1,49	0,207	2,11	2,06	2,114	2,086	0,414	2,14
5.853	-2,06	0,433	-1,45	0,1455	2,120	2,031	2,120	2,096	0,420	2,114

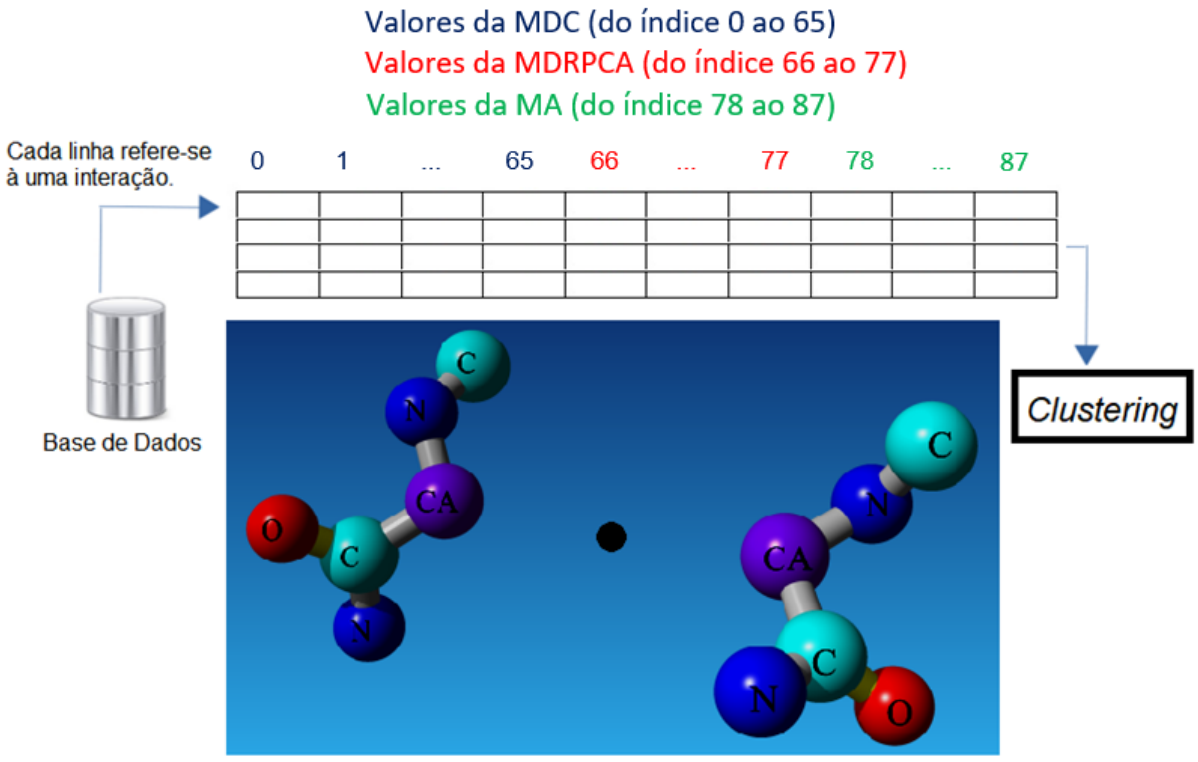
Alguns registros do *cluster* 153

Id	0	1	2	3	4	5	6	7	8	9
261	-2,01	-0,41	-1,31	2,964	2,093	2,042	2,121	2,118	0,1932	2,144
263	-2,05	-0,44	-1,34	2,975	2,098	2,037	2,115	2,115	0,184	2,152
1.785	-1,13	-0,29	-1,39	2,70	2,10	2,09	2,124	2,120	0,360	2,104
4.917	-1,12	-2,89	-1,39	2,711	2,078	2,080	2,110	2,141	0,367	2,123

Fonte: Elaborada pelo autor (2023)

verificar se a combinação de todas essas informações contribuirá positivamente para a qualidade dos agrupamentos ou se o excesso de informações serão ruídos que prejudicarão na formação dos *clusters*. A Figura 34 ilustra a construção da MDCM.

Figura 34 – Matriz MDCM



Fonte: Elaborada pelo autor (2023)

Conforme a Figura 34, a MDCM é composta por 88 valores para representar cada interação. As primeiras 66 posições referem-se à MDC. Os próximos 12 valores foram obtidos pela MDRPCA. Por fim, os 10 últimos valores referem-se aos valores da MA. A Figura 35 ilustra como a MDCM difere as interações para inseri-las em dois agrupamentos distintos.

Figura 35 – Exemplo de alguns *clusters* da MDCM com o k-means clustering

Alguns registros do *cluster* 5

Id	0	...	3	...	65	66	67	68	69	...	86	87
6.207	1,324	...	3,296	...	2,227	4,998	3,843	3,037	3,066	...	0,247	2,146
6.208	1,332	...	3,254	...	2,245	5,016	3,868	3,023	3,073	...	0,264	2,133
9.339	1,327	...	3,296	...	2,259	4,972	3,784	2,969	2,972	...	0,261	2,16
9.340	1,33	...	3,254	...	2,278	4,947	3,772	2,910	3,003	...	0,282	2,155

Alguns registros do *cluster* 254

Id	0	...	3	...	65	66	67	68	69	...	86	87
1.376	1,342	...	4,662	...	2,277	4,698	3,895	2,658	3,364	...	0,424	2,071
2.053	1,316	...	4,467	...	2,237	4,407	3,654	2,427	3,099	...	0,458	2,136
5.763	1,326	...	4,609	...	2,244	4,661	3,916	2,657	3,373	...	0,426	2,153
5.852	1,336	...	4,630	...	2,20	4,750	3,970	2,728	3,403	...	0,415	2,139

Fonte: Elaborada pelo autor (2023)

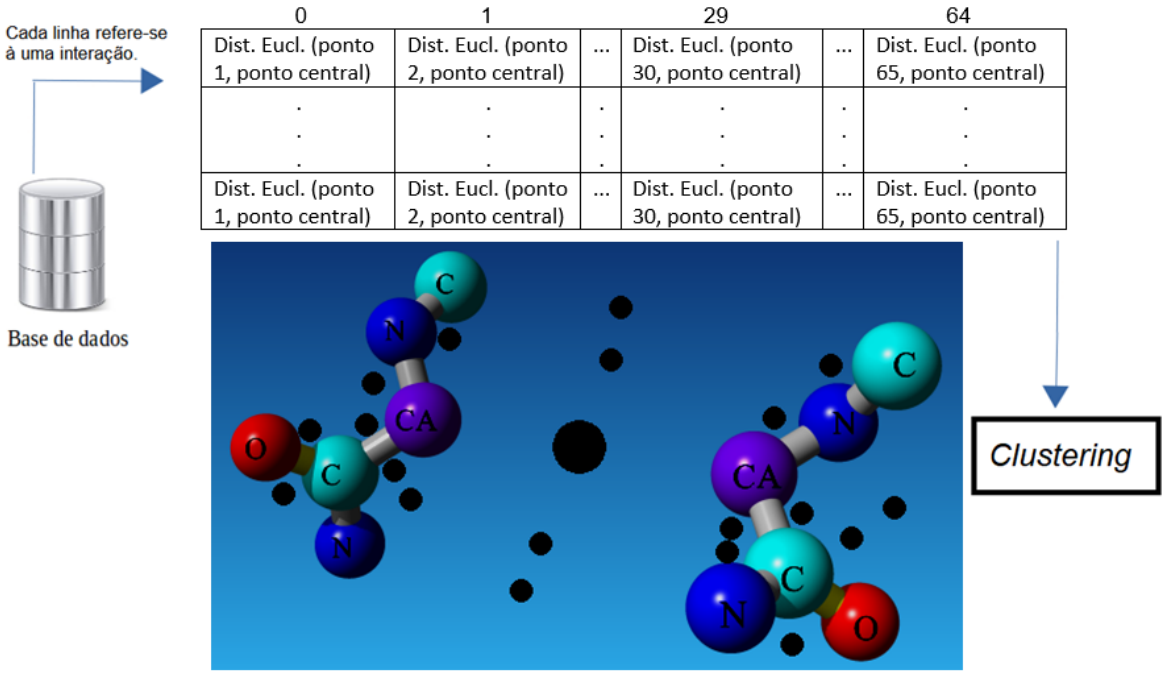
Através da Figura 35 pode-se observar que ao longo das 88 posições da MDCM, os registros dos agrupamentos 5 e 254 obtiveram diferenças em algumas posições. Dentre as posições ilustradas, pode-se notar maiores discrepâncias nos índices 3, 66, 68 e 86. As demais colunas ilustradas obtiveram valores parecidos de um *cluster* para o outro.

4.6 Matriz de Pontos Médios (MPM)

A Matriz de Pontos Médios (MPM) consiste em combinar as principais estratégias das construções da MDC e da MDRPCA. No primeiro passo desta metodologia calcula-se um ponto entre os dois carbonos alfas (como feito na MDRPCA). O segundo passo consiste em calcular um ponto médio entre cada possível dupla de átomos existente na interação. Em seguida, calcula-se a distância Euclidiana entre o ponto central aos carbonos alfas e cada ponto obtido no passo 2. Deste modo obtêm-se uma matriz de 65 colunas e *n* linhas, em que *n* é a quantidade de interações. É importante ressaltar que a MPM tem uma coluna a menos do que a MDC devido ao ponto entre os carbonos alfas ser o ponto central. Assim não existe a necessidade de calculá-lo novamente no passo 2 e consequentemente no passo 3. A Figura 36 ilustra a construção da MPM.

Por meio da Figura 36 pode-se notar com maior destaque o ponto central, na cor preta, entre os carbonos alfas. Os outros pontos referem-se aos pontos médios entre os demais

Figura 36 – Matriz MPM



Fonte: Elaborada pelo autor (2023)

átomos da interação. É importante ressaltar que para uma melhor visualização da Figura 36 não foram ilustrados todos os 65 pontos, escolheu-se ilustrar apenas alguns desses pontos. A Figura 36 ainda ilustra que a primeira posição (índice 0) será o resultado da distância Euclidiana do ponto médio entre o átomo de carbono anterior à cadeia principal do lado esquerdo e o nitrogênio da cadeia principal também do lado esquerdo, em relação ao ponto central. Deste modo, a segunda coluna será o valor da distância Euclidiana do ponto entre o átomo de carbono do lado esquerdo e o carbono alfa também do lado esquerdo, com o ponto central. E assim, sucessivamente até a coluna 65 ser o resultado da distância Euclidiana do ponto entre o oxigênio do lado direito e o nitrogênio posterior, também do lado direito, com o ponto central. O Algoritmo 5 detalha a construção da MPM.

De acordo com o algoritmo 5, para cada interação, primeiro calcula-se o ponto entre os átomos de carbono alfa. Em seguida, através de dois comandos de repetições **for**, calcula-se um ponto entre cada possível par de átomos. Ainda dentro dessas estruturas de repetições, é realizado o cálculo da distância Euclidiana entre o ponto central aos carbonos alfas e o ponto entre cada par de átomos. O valor obtido é inserido na matriz MPM. A Figura 37 ilustra dois agrupamentos obtidos através da MPM.

De acordo com a Figura 37 pode-se observar que a maioria das posições ilustradas de algumas interações dos *clusters* 10 e 163 foram diferentes. Dentre as posições ilustradas na Figura 37 é possível observar maiores diferenças nos índices 0, 31, 34, 63 e 64.

Algoritmo 5: Algoritmo para construir a matriz MPM

Entrada: $MtInterac[\,][\,]$;
Saida: $MatrizMPM[\,][\,]$;
Function ConstroiMatrizMPM
$indiceColuna \leftarrow 0$;
para *cada* $registro \in MtInterac[\,][\,]$ **faça**
  $pontoEntreCAs[0,1,2] \leftarrow$
   $CaulcPonto(MtInterac[idRegistro][2[0,1,2]], MtInterac[idRegistro][8[0,1,2]])$;
  para $i \leftarrow 0$ *até 11* **faça**
    para $j \leftarrow 0$ *até 11* **faça**
      $if((i! = 2\&\&i! = 8)||(j! = 2\&\&j! = 8))\ pontoAtomos[0,1,2] \leftarrow$
       $CaulcPonto(MtInterac[idRegistro][i[0,1,2]], MtInterac[idRegistro][j[0,1,2]])$;
      $MatrizMPM[idRegistro][indiceColuna] \leftarrow$
       $EuclDist(pontoEntreCAs[0,1,2], pontoAtomos[0,1,2])$;
      $indiceColuna++$;
    fim
    $indiceColuna \leftarrow 0$;
  fim
fim
retorna $MatrizMPM[\,][\,]$;

Figura 37 – Exemplo de alguns *clusters* da MPM com o k-means clustering

Alguns registros do *cluster* 10

Id	0	1	...	31	32	33	34	...	63	64
387	4,204	3,616	...	1,425	1,255	0,748	1,112	...	4,522	4,746
556	4,118	3,552	...	1,316	1,201	0,762	1,068	...	4,922	4,724
1.016	4,317	3,807	...	1,238	1,126	0,759	1,063	...	4,579	4,841
2.297	4,229	3,629	...	1,417	1,243	0,752	0,982	...	4,428	4,640

Alguns registros do *cluster* 163

Id	0	1	...	31	32	33	34	...	63	64
445	4,055	3,376	...	1,800	1,159	0,770	0,901	...	3,169	3,636
1.342	4,064	3,500	...	1,632	1,379	0,761	0,486	...	3,510	3,703
1.747	4,098	3,409	...	1,783	1,139	0,763	0,935	...	3,126	3,585
3.958	3,553	2,898	...	1,724	1,252	0,762	0,551	...	3,640	4,114

Fonte: Elaborada pelo autor (2023)

4.6.1 Complexidade de Tempo da MPM

O primeiro passo para representar uma interação proteica em uma linha da matriz MPM é calcular o ponto entre os carbonos alfas, como feito na MDRPCA. Esse cálculo é feito de um modo direto, com a complexidade de $O(1)$. O próximo passo consiste em calcular um ponto entre cada possível par de átomos e em seguida calcular a distância Euclidiana de cada ponto obtido em relação ao ponto central. A complexidade é de $O((n^2 - n/2 - 1) * 2)$, que

engloba o processamento necessário para calcular os pontos médios, bem como calcular as distâncais Euclidianas. Essa complexidade é parecida com a MDC, mas com algumas diferenças. A primeira consiste na subtração por 1, pois como o ponto central é o ponto entre os carbonos alfas, não é necessário calculá-lo novamente durante o processamento das estruturas de repetições e inserir este valor na matriz MPM, que seria 0. Para atender a essa finalidade, foi inserido no código-fonte uma estrutura de seleção **if** para validar se no mínimo um dos átomos do par não é de carbono alfa, para possibilitar o cálculo do ponto médio. No entanto, o acréscimo deste **if** resultou em dobrar a quantidade de comparações, assim justificando a multiplicação por 2 na complexidade deste algoritmo. Ou seja, essa validação aumentou o tempo para construir a MPM. Entretanto, o tempo para realizar os agrupamentos serão menores, devido à redução do tamanho da matriz em uma coluna.

Capítulo 5

Metodologia

Neste Capítulo são abordados todos os passos que foram seguidos para realizar os experimentos. A Seção 5.1 apresenta os *hardwares* e *softwares* utilizados. A Seção 5.2 contém informações sobre as bases de dados utilizadas. A Seção 5.3 mostra como foi definida a quantidade ideal de *clusters* através do método *Silhouette*. A Seção 5.4 detalha os experimentos de precisão realizados. A Seção 5.5 aborda a repetição de experimentos e a análise estatística.

5.1 Hardwares e *Softwares* Utilizados

Os experimentos da base de dados 1, que contém 16.383 arquivos de interações, foram desenvolvidos em um notebook Samsung Expert X23, Intel (R) Core (TM) i5, com 8 GB RAM e sistema operacional Ubuntu 20.10 Para os experimentos com a base de dados 2, que contém 129.705 ligações de hidrogênio, e para os experimentos em que as duas bases de dados foram unidas, formando a base de dados 3, que contém 146.088 interações proteicas, foram utilizados diferentes computadores disponibilizados em nuvens através do *Google Colaboratory* e *Google Cloud*.

De acordo com Google (2022a) e David (2022) o *Google Colaboratory* é um serviço de nuvem com as opções gratuita ou paga, hospedado pelo próprio *Google*. Esse serviço permite ao usuário escrever e executar códigos na linguagem de programação Python, através do navegador, não havendo a necessidade de fazer alguma instalação prévia (SANTOS, 2022) (GOOGLE, 2022a). Para os experimentos deste trabalho, utilizou-se a versão gratuita do *Google Collaboratory*, que fornece os seguintes recursos computacionais: 12 GB de memória RAM, 2 vCPUs e GPU K80, com a arquitetura Kepler (DAVID, 2022). Esses recursos de *hardware* foram utilizados para trabalhar com as bases de dados 2 e 3, possibilitando o agrupamento das técnicas de representações de proteínas com o OPTICS. No entanto, esses recursos computacionais não foram suficientes para realizar os experimentos com as demais técnicas de *clustering* avaliadas nesse trabalho: *K-Means* e

hierarchical clustering. Deste modo, utilizou-se também serviços do *Google Cloud*.

Conforme Google (2022b), Google (2022c) e Google (2022d), o *Google Cloud* fornece diversos serviços de computação em nuvem voltados para empresas e pesquisas computacionais, tais como: computação de alto desempenho, mineração de dados, banco de dados, entre outros. Trata-se de um serviço pago, mas ao realizar o cadastro nessa plataforma é possível receber créditos que variam de $300 (trezentos dólares) até $5,000 (cinco mil dólares) para utilizá-los no site. Através do *Google Cloud* utilizou-se os seguintes recursos de *hardware*: máquina e2-custom-16-19456, plataforma *Intel Broadwell*, arquitetura x86/64 e memória RAM de 16 GB. Esses recursos foram suficientes para o uso do *k-means clustering* com as bases de dados maiores. No entanto, esses *hardwares* ainda não foram suficientes para suportar os experimentos com o *hierarchical clustering*, por insuficiência de memória RAM. Estima-se que seria necessário mais de 4 TB de memória RAM para agrupar a base de dados 2 através do *hierarchical clustering*.

As implementações, bem como os agrupamentos foram realizados através do *software Visual Studio Code*, com a linguagem de programação Python. O programa de computador LSQKAB na versão 6.5 foi utilizado para comparar o RMSD e o maior deslocamento entre as estruturas que ficaram no mesmo *cluster*. Foram elaborados *scripts* em Python, bem como códigos para a criação de gráficos no Matlab r2020a, além de *scripts* em *Shell Script*, que foram executados por meio do prompt.

5.2 Bases de Dados utilizadas

A base de dados 1 contém 16.383 interações de pontes dissulfeto, extraídas do PDB pela ferramenta RID, no formato de arquivos de texto. Trata-se da mesma base de dados utilizada em Monteiro (2017), Monteiro, Dias e Rodrigues (2020), Amorim (2020), Monteiro, Dias e Rodrigues (2021) e Monteiro, Dias e Rodrigues (2021). A base de dados 2 também foi extraída do PDB pela ferramenta RID. No entanto, trata-se de uma base de dados consideravelmente maior, com 129.705 ligações de hidrogênio. A base de dados 3 consistiu na união das bases de dados 1 e 2, totalizando 146.088 interações. Cada interação contida nessas diferentes bases de dados possuem o mesmo formato e a mesma quantidade de átomos.

De acordo com Dias (2012) e Monteiro, Dias e Rodrigues (2021), a ferramenta RID obteve para cada lado interagente de cada ponte dissulfeto, bem como de cada ligação de hidrogênio, a cadeia principal, um átomo do resíduo anterior à cadeia principal e um átomo do resíduo posterior à cadeia principal. Alguns átomos foram ignorados pela RID e não foram inseridos nos arquivos de texto. Dentre os átomos excluídos estão: carbonos beta, enxofre gama e os átomos da cadeia lateral. Estes átomos foram ignorados porque o foco dessa

base de dados é a manutenção da cadeia principal, para conservar a função da proteína em um evento de mutação. Deste modo, estes átomos ignorados não são necessários para este trabalho. A Figura 38 mostra a estrutura dos arquivos que compõem a base de dados. Como exemplo, foi ilustrado o arquivo “1ejg_CYS-3-A_CYS-40-A_mc6.pdb”, pertencente à base de dados 1 e consequentemente presente na base de dados 3.

Figura 38 – Exemplo de um arquivo de interação

```
CRYST1   40.824   18.498   22.371  90.00  90.47  90.00 P 1 21 1      2
SCALE1      0.024495  0.000000  0.000201        0.00000
SCALE2      0.000000  0.054060  0.000000        0.00000
SCALE3      0.000000  0.000000  0.044702        0.00000
ATOM    1  C   THR A   2      14.172  10.757   7.221  1.00  2.52        C
ATOM    2  N   CYS A   3      13.438  11.192   8.264  1.00  2.26        N
ATOM    3  CA  CYS A   3      13.607  10.675   9.600  1.00  2.06        C
ATOM    4  C   CYS A   3      12.210  10.413  10.164  1.00  1.94        C
ATOM    5  O   CYS A   3      11.324  11.246   9.996  1.00  2.78        O
ATOM    6  N   CYS A   4      12.015   9.277  10.850  1.00  1.74        N
ATOM    7  C   THR A  39      20.535  13.026  11.330  1.00  4.95        C
ATOM    8  N   CYS A  40      19.748  11.982  11.477  1.00  4.08        N
ATOM    9  CA  CYS A  40      18.460  12.120  12.139  1.00  3.64        C
ATOM   10  C   CYS A  40      18.660  12.202  13.669  1.00  3.83        C
ATOM   11  O   CYS A  40      19.515  11.523  14.227  1.00  4.96        O
ATOM   12  N   PRO A  41      17.847  13.009  14.329  1.00  4.47        N
END
```

Fonte: Monteiro (2017)

A primeira linha do arquivo é referente a descrição da célula cristalográfica utilizada para extrair diversas informações da estrutura atômica, sendo identificada por CRYST1. As linhas 2, 3 e 4, indicados por SCALEn mostram os operadores para a transformação de coordenadas ortogonais em coordenadas da célula cristalográfica. Abaixo destas 4 primeiras linhas, é iniciada a seção de coordenadas. Esta seção é uma das mais importantes do arquivo. A partir de informações contidas em PDB (2010), foi possível construir a Tabela 1 para explicar detalhadamente os itens desta seção.

O primeiro conjunto de caracteres da seção de coordenadas refere-se ao nome do registro contido naquela linha. O segundo conjunto de caracteres indica o número de série do átomo. A terceira sequência mostra o nome do átomo, sendo: carbono (C), nitrogênio (N), carbono alfa (CA) e oxigênio (O). A quarta *string* refere-se à indicação de localização alternativa. As próximas quatro sequências de caracteres mostram respectivamente o nome do resíduo, a identificação da cadeia, o número de sequência do resíduo e o código de inserção do

Tabela 1 – Seção de Coordenadas de um arquivo de Interação

Colunas	Tipo de Dados	Campo	Definição
1 - 6	Nome do registro	ATOM	
7 - 11	Inteiro	serial	Número de série do átomo
13 - 16	String	name	Nome do átomo
17	Caracter	altLoc	Indicador de localização alternativa
18 - 20	Nome do resíduo	resName	Nome do resíduo
22	Caracter	chainId	Identificação da cadeia
23 - 26	Inteiro	resSeq	Número de sequência do resíduo
27	Caracter	iCode	Código da inserção de resíduos
31 - 38	Real	x	Coordenada ortogonal para x em angstrons
39 - 46	Real	y	Coordenada ortogonal para y em angstrons
47 - 54	Real	z	Coordenada ortogonal para z em angstrons
55 - 60	Real	occupancy	Probabilidade do átomo estar naquele local
61 - 66	Real	tempFactor	Fator de Temperatura
77 - 78	String	element	Símbolo do elemento alinhado à direita
79 - 80	String	charge	Carga

Fonte: Adaptado de PDB (2010)

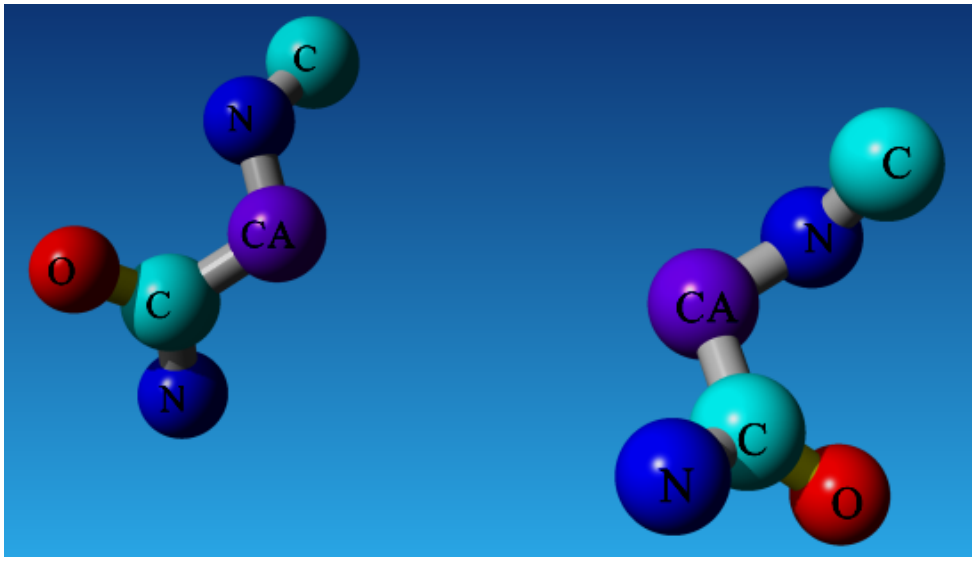
resíduo. As *strings* 9 , 10 e 11 indicam as coordenadas atômicas em Å para os eixos x, y e z. As próximas 4 sequências de caracteres referem-se à probabilidade do átomo estar naquela localização, fator de temperatura, elemento do símbolo alinhado à direita e a carga. A ponte dissulfeto “1ejg_CYS-3-A_CYS-40-A_mc6.pdb”, representada pela Figura 38 não contém informações para os campos de indicação de localização alternativa, código de inserção e carga. A Figura 39 ilustra tridimensionalmente a estrutura do arquivo de interação “1ejg_CYS-3-A_CYS-40-A_mc6.pdb”.

A Figura 39 ilustra os dois lados interagentes que formam o arquivo de interação. O carbono mais à esquerda (C) é um átomo do resíduo anterior à cadeia principal do lado esquerdo. Os próximos átomos deste lado formam esta cadeia principal. São eles: nitrogênio (N), carbono alfa (CA), carbono (C) e oxigênio (O). O último nitrogênio (N) do lado esquerdo é um átomo do resíduo posterior à cadeia principal. O lado direito do arquivo de interação também contém 6 átomos, que seguem o mesmo padrão do lado esquerdo.

5.3 Tratamento da Base de Dados

A maioria das interações utilizadas neste trabalho tem a ordem dos átomos como descrito na Figura 38, sendo esta ordem: C-N-CA-C-O-N-C-N-CA-C-O-N. No entanto, observou-se que algumas interações estavam com os átomos em ordens trocadas, o que poderia interferir no funcionamento e nos resultados das metodologias baseadas em matriz de distâncias. A Figura 40 ilustra uma interação que está com os átomos em ordens trocadas.

Figura 39 – Forma tridimensional da interação “1ejg_CYS-3-A_CYS-40-A_mc6.pdb”



Fonte: Elaborada pelo autor (2023)

Figura 40 – Arquivo de interação com a ordem dos átomos trocada

```
CRYST1   52.710   52.710   43.410  90.00  90.00  90.00 P 43 21 2     8
SCALE1      0.018972  0.000000  0.000000        0.00000
SCALE2      0.000000  0.018972  0.000000        0.00000
SCALE3      0.000000  0.000000  0.023036        0.00000
ATOM      1  C   ALA A  13      49.393   7.654  32.078  1.00 24.24           C
ATOM      2  N   CYS A  14      49.401   6.880  31.000  1.00 23.23           N
ATOM      3  CA  CYS A  14      50.225   5.686  31.035  1.00 22.21           C
ATOM      4  C   CYS A  14      51.695   6.025  30.831  1.00 24.52           C
ATOM      5  O   CYS A  14      52.046   7.139  30.455  1.00 26.12           O
ATOM      6  N   ARG A  15      52.585   5.072  31.125  1.00 26.54           N
ATOM      7  C   GLY A  37      48.896   3.966  25.756  1.00 21.91           C
ATOM      8  N   CYS A  38      48.161   4.850  26.426  1.00 21.43           N
ATOM      9  C   CYS A  38      45.844   5.803  26.382  1.00 21.81           C
ATOM     10  O   CYS A  38      46.289   6.931  26.593  1.00 21.61           O
ATOM     11  CA ACYS A  38      46.747   4.608  26.708  0.38 22.68           C
ATOM     12  N   ARG A  39      44.619   5.569  25.942  1.00 22.11           N
END
```

Fonte: Elaborada pelo autor (2023)

A Figura 40 ilustra a interação "1qlq_CYS-14-A_CYS-38-A_mc6.pdb". Pode-se observar através do destaque em vermelho que os átomos desta interação estão na ordem C-N-CA-C-O-N-C-N-C-O-CA-N. Ou seja, os 8 primeiros átomos e o 12º estão na ordem correta. No entanto, o 9º, o 10º e o 11º átomos estão em posições diferentes do padrão dos demais arquivos interagentes. O fato do segundo átomo de carbono alfa desta interação ser o 11º ao invés de ser o 9º átomo, como nos demais arquivos, poderia induzir ao erro na MDRPCA e na MPM ao calcular o ponto entre os carbonos alfa, caso não fosse validada as

reais posições dos átomos. Essa discrepância na disposição dos átomos também poderia potencialmente levar a equívocos na determinação dos centróides no contexto da MDRCCA. Além disso, a matriz MDC apresentaria cálculos de distâncias Euclidianas em sequências invertidas, o que poderia comprometer a integridade na formação dos agrupamentos.

Para resolver este problema desenvolveu-se um *script* em Python que faz a leitura dos arquivos interagentes e salva todas as informações de coordenadas atômicas em um novo arquivo de texto, padronizando os átomos para a sequência mais comum dos arquivos da base de dados: C-N-CA-C-O-N-C-N-CA-C-O-N. Essa padronização obtida por esse pré-processamento de dados contribuiu para facilitar as validações necessárias durante o desenvolvimento das metodologias baseadas em matrizes de distâncias.

5.4 Quantidade Ideal de *Clusters*

Conforme Rousseeuw (1987) o *silhouette* é um método de interpretação e validação da consistência de *clusters* em uma base de dados. De acordo com Pedregosa et al. (2011) o coeficiente *silhouette* é calculado através da distância média intra-cluster (a) e a distância média do cluster mais próximo (b) para cada amostra, i.e a distância entre uma amostra e o *cluster* mais próximo que a amostra não faça parte. A equação 6 ilustra o cálculo do coeficiente para uma amostra:

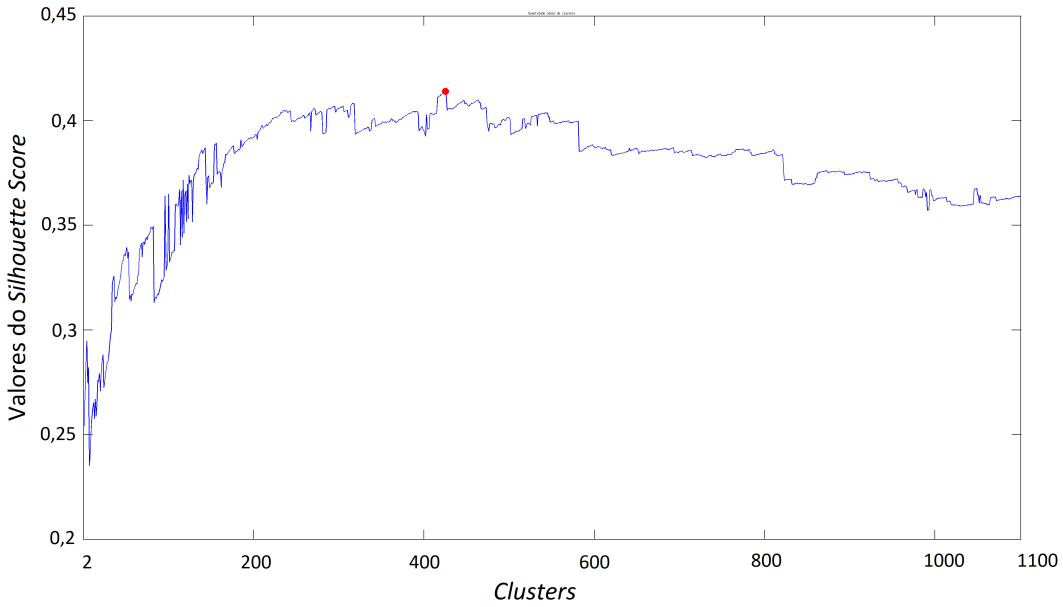
$$silhouette = \frac{(b-a)}{max(a,b)} \tag{6}$$

Através da equação 6 pode-se observar que quanto mais próxima uma amostra estiver dos demais elementos do seu grupo (o valor *a* for pequeno) e mais distante das amostras contidas em *clusters* vizinhos (o valor *b* for grande), maior será o coeficiente *silhouette*. De acordo com Pedregosa et al. (2011) o *silhouette* é um algoritmo de maximização, no qual é retornado um valor entre -1 e 1. Quando um valor negativo é retornado, normalmente significa que a amostra está em um *cluster* errado e que possivelmente ela se encaixaria melhor em um agrupamento vizinho. Se o valor retornado for próximo de 0 significa que o registro está no limite de decisão entre dois grupos. Quando o valor retornado é próximo de 1 significa que a amostra encaixou-se adequadamente no seu *cluster* atual, ficando longe dos agrupamentos vizinhos.

Calcular a média dos valores de *silhouette* para todas amostras e repetir esse cálculo considerando diferentes quantidades de agrupamentos é uma maneira de determinar a quantidade mais adequada de *clusters* para a base de dados (ROUSSEEUW, 1987) e (PEDREGOSA et al., 2011). Deste modo, realizou-se a validação da quantidade ideal de agrupamentos para os experimentos com o *k-means clustering* e o *hierarchical clustering*.

Para a base de dados 1, calculou-se a média do *silhouette score* dentre todas as amostras para o intervalo de grupos de 2 até 1.100 *clusters*. Para o *k-means*, esse procedimento foi repetido para 31 sementes diferentes. A Figura 41 ilustra a média do *silhouette score* para cada quantidade de agrupamentos, conforme sementes do *k-means clustering*.

Figura 41 – Calculo da média do *silhouette score* para uma execução da MPM combinada com o *k-means clustering*



Fonte: Elaborada pelo autor (2023)

De acordo com a Figura 41, os valores obtidos pelas médias das amostras no cálculo do *silhouette score*, tiveram oscilações com tendência de alta até a quantidade de 425 grupos (maior valor encontrado, em destaque no gráfico). Também é possível notar que ao passar de 425 *clusters* as oscilações continuaram, mas dessa vez com tendência de queda. Para as bases de dados 2 e 3, devido a limitações de recursos computacionais não foi possível validar o *silhouette score* para todas as possíveis quantidades de agrupamentos, pois como essas bases de dados são maiores, a quantidade de grupos a serem formados também tornou-se maior e consequentemente o tempo de execução aumentou. Deste modo, realizou-se três processamentos, o primeiro foi mais superficial e o segundo e terceiro foram mais refinados. O primeiro processamento consistiu calcular o *silhouette* com o passo de 5.000 agrupamentos até notar uma evidente tendência de queda. Por exemplo, primeiro calculou-se o *silhouette* para 2 grupos, depois 5.000 agrupamentos, em seguida 10.000, depois 15.000 e assim sucessivamente até 30.000 *clusters*. Deste modo, foi possível obter os dois maiores valores "superficiais" do *silhouette score*. O segundo processamento consistiu em calcular o *silhouette* com o passo de 500 entre as duas quantidades de grupos encontradas no procedimento anterior, acrescentando no intervalo de busca 2.500 antes

do primeiro valor e 2.500 após o segundo valor, para o caso de existir um maior valor de *silhouette score* um pouco antes ou um pouco depois do passo de 5.000. Por exemplo, se os maiores valores encontrados forem o *score* de 20.000 e 25.000, então o valor de *silhouette* seria calculado para o intervalo de 17.500 até 27.500, com o passo de 500. Por fim, realizou-se um novo processamento, mas agora com o passo de 50 entre os dois maiores valores de *silhouette* encontrados no segundo processamento, sendo também acrescido 250 antes do primeiro valor e 250 depois do segundo valor. Ou seja, se os dois maiores valores encontrados forem referentes a 21.000 e 22.000, o intervalo de busca seria entre 20.750 e 22.250, com o passo de 50. É importante ressaltar que essa estratégia utilizada para as bases de dados 2 e 3 não garante que o maior valor de *silhouette score* será encontrado, mas é provável que seja encontrado um valor próximo do maior.

As quantidades ideais de *clusters* utilizadas pelo OPTICS foram obtidas automaticamente pelo próprio algoritmo de *clustering*. Definiu-se a quantidade mínima de duas interações por grupo e a maior distância possível entre duas amostras de um mesmo agrupamento como infinito. Deste modo, o algoritmo do OPTICS formou os agrupamentos automaticamente.

5.5 Realização dos Experimentos de Precisão

Os experimentos tiveram como objetivo avaliar diversas metodologias, incluindo MPM, MDCM, MA, MDC, MDRCCA, MDRPCA, o método de busca RID e aCSM, em relação à formação de agrupamentos utilizando diferentes técnicas de *clustering*. Em certos casos, neste estudo, algumas abordagens de representação de proteínas não foram utilizadas em conjunto com todas as técnicas de *clustering*, por duas razões possíveis. A primeira razão está relacionada ao fato de que a combinação entre a representação das proteínas e a técnica de agrupamento não ter resultado em um número de grupos ideal, de acordo com o critério do método *silhouette*. Em algumas dessas combinações, apenas dois grupos foram considerados ideais, o que não se adequa ao escopo do trabalho, uma vez que isso certamente resultaria em uma precisão muito baixa. A segunda razão é que algumas dessas combinações já foram testadas em estudos anteriores e não alcançaram níveis satisfatórios de acurácia.

A acurácia dos grupos formados foi verificada através de sobreposições atômicas feitas com o LSQKAB. Estas sobreposições ocorreram entre todos os registros que estiveram no mesmo grupo. Cada registro foi comparado duas vezes com todas as outras pontes dissulfeto de seu *cluster*. Na primeira comparação, um destes registros permaneceu "fixo" e o outro foi alinhado ao mesmo. Na segunda comparação, ocorreu o oposto entre estes arquivos. Nestas comparações verificou-se o RMSD e o maior deslocamento entre os 12 pares de átomos de cada estrutura comparada.

A validação do RMSD foi obtida através da informação *RMS Displacement*, retornada pelo LSQKAB. A validação do maior deslocamento pode ser obtida através da informação *Maximum Displacement xyz*, também retornada pelo LSQKAB. Também é possível obter o maior deslocamento ao avaliar a maior distância do arquivo de deltas gerado pela comparação entre as duas estruturas. No entanto, visando reduzir o tempo de processamento, o maior deslocamento foi avaliado através da informação *Maximum Displacement xyz*. Apesar dessas estratégias, as validações do RMSD e do maior deslocamento demandaram de muito tempo de processamento computacional, pela necessidade de ter que realizar o alinhamento estrutural para possibilitar sobreposições atômicas em diversos arquivos de interações de proteínas, um procedimento que se enquadra na classe NP-Completo. Sendo assim, essas validações foram algumas das tarefas mais complicadas e demoradas deste trabalho.

A partir da obtenção dos valores de RMSD e maior deslocamento, foram realizadas análises estatísticas através da Anova e do teste de Tukey, considerando a tolerância de 0,5 Å para avaliar os valores obtidos. A escolha pela Anova e pelo teste de Tukey ocorreu por possibilitarem comparar simultaneamente os resultados das técnicas avaliadas.Também foi verificado que os valores obtidos pelas técnicas atendem as premissas de normalidade, independência e homocedasticidade que são exigidas pela Anova (CAMPELO, 2015).

5.6 Repetição de Experimentos e Análise Estatística

Os algoritmos de *clustering* OPTICS e *hierarchical clustering* são determinísticos. Ou seja, ao definir os parâmetros, a mesma saída será obtida, não existindo uma "semente" para modificar o resultado de uma execução para a outra. No entanto, durante o processo de construção de agrupamentos com o *k-means clustering*, os centroides definidos inicialmente com base em um valor de semente, interferem no resultado final. Deste modo, cada experimento com o *k-means* foi repetido por 31 vezes, para obter uma quantidade considerável de resultados, que possibilite avaliar as médias das execuções e a realização de análises estatísticas. Conforme Montgomery (2019), dada uma distribuição normal, ao repetir o experimento por mais de 30 vezes obtêm-se uma quantidade suficiente de resultados para fazer uma análise estatística adequada.

As análises estatísticas foram feitas através da Análise de Variância (ANOVA) para 1 fator e também através do teste de Tukey. A ANOVA consiste em um procedimento que compara a distribuição de 3 ou mais grupos, podendo ser utilizada quando os resíduos cumprem as premissas de normalidade, independência e homocedasticidade (CAMPELO, 2015). Com os resultados da ANOVA, pode-se utilizar o teste de Tukey para verificar se ocorreram diferenças significativas entre os resultados das técnicas comparadas (MATHWORKS, 2005).

Capítulo 6

Análise e Discussão dos Resultados

Este Capítulo é dividido em três seções. A Seção 6.1 refere-se aos experimentos realizados com a primeira base de dados, de 16.383 interações de pontes dissulfeto. A Seção 6.2 é referente aos experimentos com a segunda base de dados, de 129.705 interações de pontes de hidrogênio. A Seção 6.3 indica os resultados dos experimentos com a terceira base de dados, que consiste na união das bases de dados 2 e 3, formando uma base de dados com 146.088 interações de proteínas.

6.1 Resultados Obtidos com a Base de Dados 1

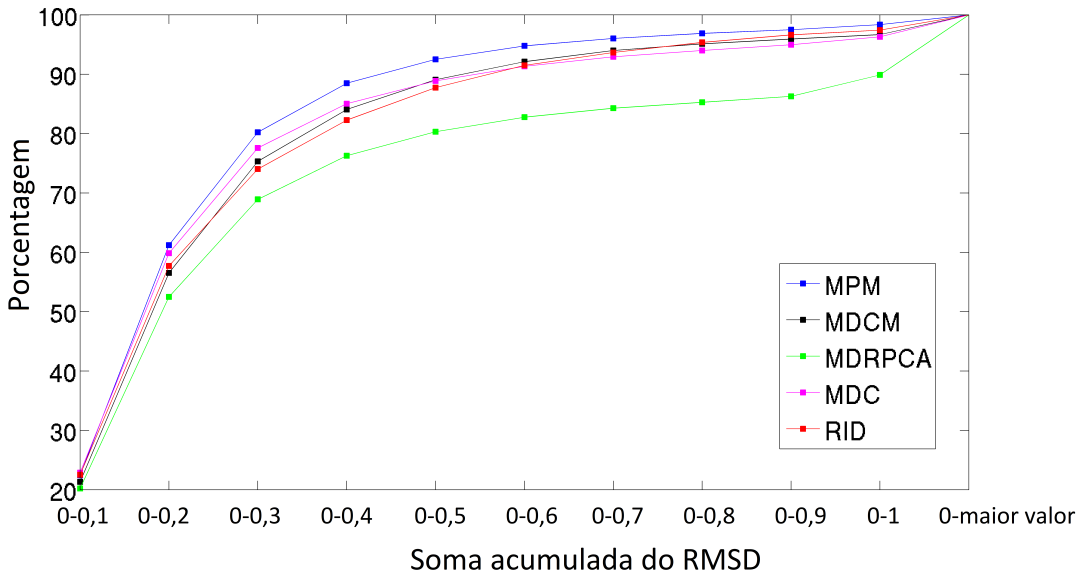
Para a base de dados 1 foram realizados experimentos para as diferentes técnicas de representações de proteínas, combinadas com três diferentes técnicas de agrupamentos: *k-means clustering*, OPTICS e *hierarchical clustering*. Os resultados obtidos para cada técnica de agrupamentos são descritos respectivamente nas Subseções 6.1.1, 6.1.2 e 6.1.3.

6.1.1 *K-Means Clustering*

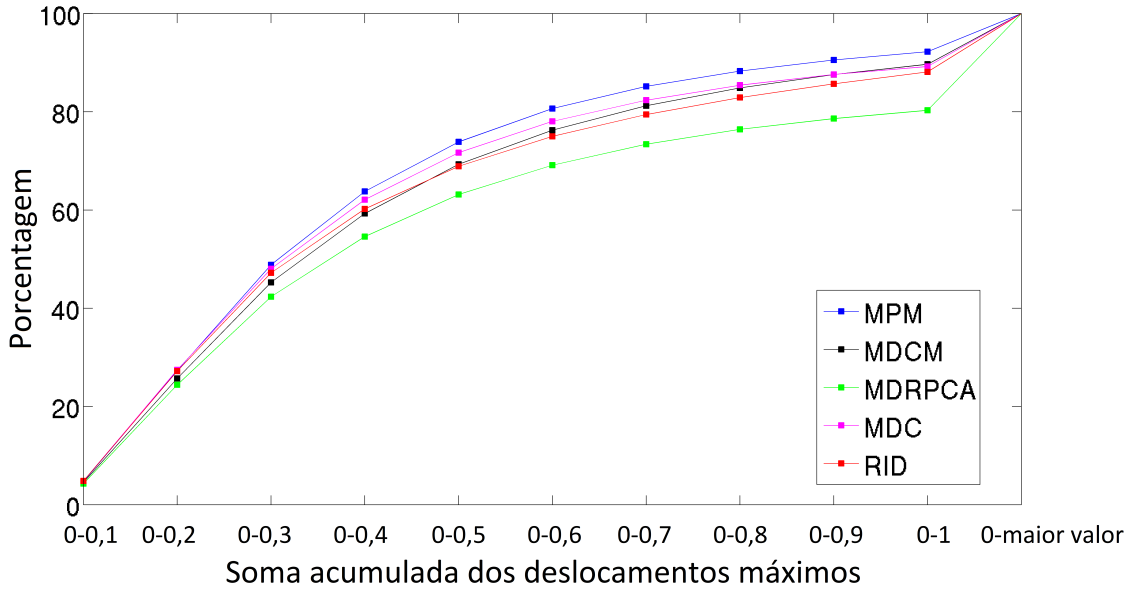
Através do *k-means clustering* foram realizadas 31 execuções de experimentos para cada técnica de representação de proteínas, que atendessem aos dois critérios informados na seção de metodologia. Deste modo, foram definidas 5 técnicas para a realização dos experimentos: MPM, MDCM, MDRPCA, MDC e o método de busca da RID. A Figura 42 (a) ilustra a soma acumulada do valor percentual médio quanto ao RMSD para as técnicas avaliadas. A Figura 42 (b) ilustra a soma acumulada do valor percentual médio quanto ao maior deslocamento. Pode-se observar que para essas análises, quanto maior for o valor da soma acumulada nas faixas iniciais (mais à esquerda), indicará um melhor resultado. Desse modo é indicado que uma respectiva técnica conseguiu obter um maior percentual de sobreposições com baixos valores de RMSD e maior deslocamento, indicando que foi possível construir grupos de interações proteicas com conformações tridimensionais similares.

Figura 42 – Soma acumulada dos resultados com a base de dados 1, através do *k-means clustering*.

(a) Resultados conforme o RMSD



(b) Resultados conforme o maior deslocamento



Fonte: Elaborada pelo autor (2023)

De acordo com a Figura 42 (a) todas as 5 técnicas obtiveram valores percentuais próximos ao considerar o RMSD de até 0,1 Å. No entanto, com a soma acumulada de até 0,2 Å nota-se que a MPM obtém um maior percentual nesta faixa do que as demais técnicas, seguida pela MDC. Na próxima faixa, da soma acumulada de até 0,3 Å, percebe-se que a superioridade da MPM aumenta em relação às demais técnicas, continuando com o melhor resultado dentre todas as técnicas até o fim deste gráfico. Através da Figura 42 (a) também é possível notar que na soma acumulada até a faixa de 0,5 Å (valor que indica uma forte similaridade estrutural) a precisão da MPM é consideravelmente superior do que as demais técnicas. Nesta faixa também nota-se que a MDC e a MDCM obtiveram acurácias próximas, mas com uma leve vantagem da MDCM. O método de busca da RID obteve uma precisão um pouco abaixo dessas metodologias. A MDRPCA obteve resultados inferiores do que as demais técnicas avaliadas em todos os valores de somas acumuladas.

Conforme a Figura 42 (b) na avaliação do maior deslocamento pode-se notar que nas duas primeira faixas de somas acumuladas (0,1 Å e 0,2 Å) os valores das cinco técnicas avaliadas ficaram próximos. A partir da soma acumulada de 0,3 Å a MPM começa a apresentar resultados melhores do que as demais metodologias, mantendo a superioridade até o fim deste gráfico. De um modo geral, a MDC ficou em segundo lugar nessa avaliação. O método de busca da RID e a MDCM apresentaram resultados próximos até a soma acumulada de 0,5 Å. Do mesmo modo que na avaliação pelo RMSD, a MDRPCA apresentou resultados inferiores em todas as faixas de somas acumuladas avaliadas.

As Figuras 42 (a) e (b) indicam que nas avaliações de RMSD e maior deslocamento, a MPM apresentou os melhores resultados. Na avaliação do RMSD a MDCM foi a metodologia na qual os resultados mais se aproximaram da MPM. No entanto, na outra comparação, a MDC foi a metodologia que mais se aproximou da MPM. A estratégia de busca da RID ficou um pouco abaixo dessas metodologias. Todas essas técnicas superaram a MDRPCA nesses cenários avaliados. Ao avaliar as Figuras 42 (a) e (b) de modo conjunto, ainda pode-se perceber o quanto a avaliação pelo maior deslocamento é mais rigorosa se comparada com a avaliação pelo RMSD. Enquanto todas metodologias obtiveram mais de 20% de sobreposições com até 0,1 Å de RMSD, apenas aproximadamente 5% tiveram o maior deslocamento com até 0,1 Å. Este mesmo comportamento pode ser observado durante todos os valores de somas acumuladas, inclusive para os resultados superiores a 1,0 Å. A Tabela 2 detalha os principais resultados obtidos através dos agrupamentos realizados.

A Tabela 2 está ordenada pelas técnicas que apresentaram os melhores resultados, somando o percentual de RMSD e maior deslocamento. Pode ser notado por essa tabela que a MPM apresentou os melhores resultados para a avaliação do RMSD e também do maior deslocamento, considerando a tolerância de 0,5 Å. A terceira, a quarta e a quinta coluna da Tabela 2 indicam respectivamente a menor, a média e a maior quantidade ideal

Tabela 2 – Resultados do *k-means clustering* para a base de dados 1

Técnica	RMSD <= 0,5Å	Mai. D. <= 0,5Å	Mín. *cl.*	Med. *cl.*	Max.*Cl.*	*cl.* 1 reg.
MPM	92,52%	73,88%	271	366,94	523	0,32
MDC	88,83%	71,68%	233	350,58	461	0,10
MDCM	89,08%	69,32%	172	388,16	631	1,68
RID	87,76%	68,61%	224	379,55	611	0,55
MDRPCA	80,32%	63,16%	191	300,13	487	0,13

Fonte: Elaborada pelo autor (2023)

de agrupamentos conforme o método *silhouette*, para cada técnica. Pode-se observar que todas as 5 técnicas obtiveram variações consideráveis da quantidade ideal de grupos. Essa variação interferiu diretamente nos *boxplots*. As Figuras 43 (a) e (b) referem-se aos *boxplots* com as porcentagens de aproveitamentos totais das 31 execuções de cada técnica de representação de proteínas, considerando até 0,5 Å de RMSD e maior deslocamento.

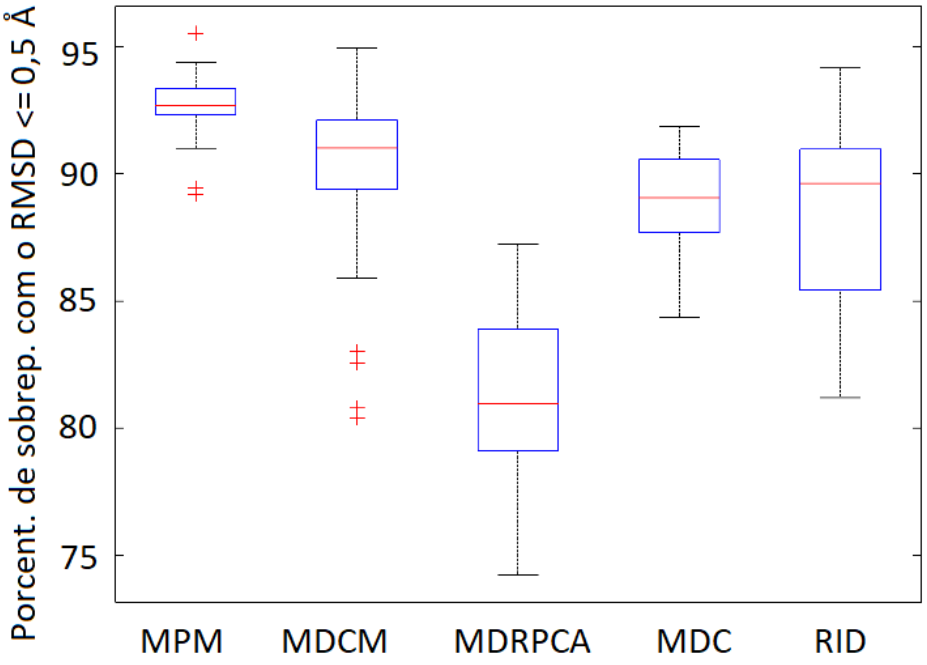
De acordo com a Figura 43 (a), pode-se notar o mesmo comportamento de superioridade dos resultados da MPM em relação às demais técnicas, que também foram observados pelo gráfico de somas acumuladas e pela Tabela 2. A Figura 43 (a) também indica que a MDCM, MDC e o método de busca da RID superaram a MDRPCA. Conforme a Figura 43 (b), ao avaliar o maior deslocamento as cinco metodologias obtiveram resultados mais próximos, com uma leve vantagem da MPM.

Ao avaliar os *boxplots* da Figura 43 (a) e (b), juntamente com a Tabela 2 pode-se notar uma relação direta entre a quantidade ideal de *clusters* indicada pelo *silhouette* para cada semente e o resultado obtido. Ou seja, como o foco deste trabalho consiste em realizar sobreposições atômicas entre interações que ficaram em um mesmo agrupamento, se a quantidade total de grupos for alta, a tendência é que os resultados sejam melhores. Por exemplo, a MPM obteve 2 *outliers* nos dois *boxplots*, sendo um indicando um valor bem acima de seus resultados (melhor resultado geral) e o outro *outlier* foi bem abaixo da sua média. As sementes que originaram as execuções que terminaram com esses resultados foram as mesmas que indicaram a maior e a menor quantidade ideal de agrupamentos da MPM, sendo 523 e *271 clusters* respectivamente. Com base nessa análise, pode-se observar através da Tabela 2 que a MDCM teve uma semente que indicou 631 agrupamentos como quantidade ideal e a estratégia de busca da RID obteve uma semente que resultou em 611 grupos. Mesmo com essas quantidades de *clusters*, a melhor execução individual foi obtida pela MPM, que também obteve o melhor aproveitamento geral, sem ter a maior quantidade média de *clusters*.

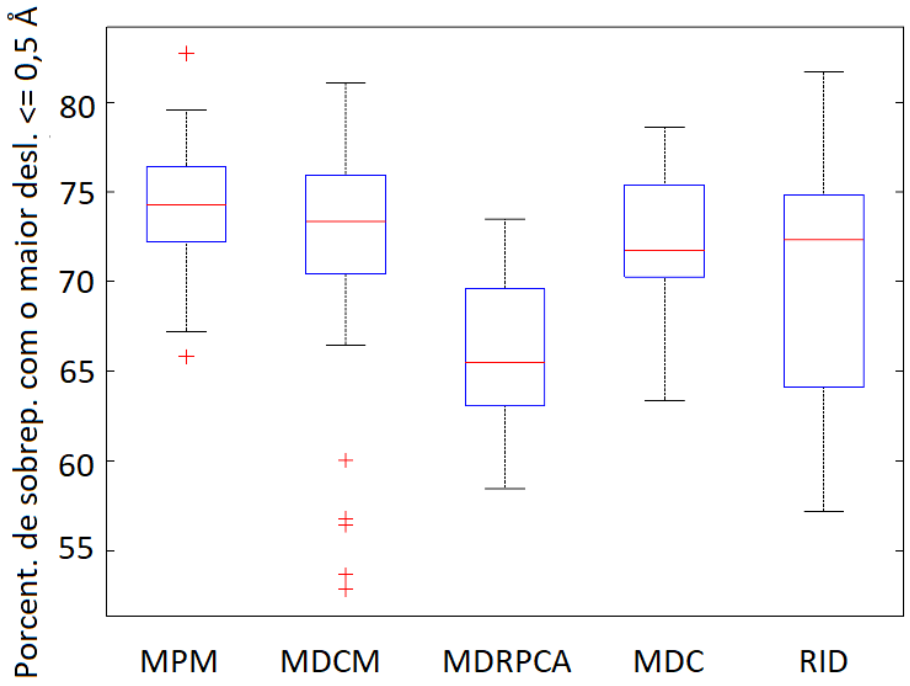
Outra avaliação que pode ser feita através destes *boxplots*, juntamente com a Tabela 2, trata-se das grandes variações de acurácias obtidas pelas técnicas MDCM, MDRPCA e

Figura 43 – *Boxplots* com as porcentagens de aproveitamentos totais da base de dados 1, utilizando o *k-means clustering*

(a) *Boxplots* considerando RMSD de até 0.5 Å



(b) *Boxplots* considerando o maior deslocamento de até 0.5 Å



Fonte: Elaborada pelo autor (2023)

RID. Essas variações ocorreram principalmente pelas diferentes quantidades ideais de agrupamentos obtidas por cada semente testada. Por exemplo, a MDCM obteve uma semente que adequou-se melhor com 172 agrupamentos. Por outro lado, essa metodologia também obteve uma outra semente que resultou em 631 grupos. A quantidade média de agrupamentos foi 388. Essas diferentes quantidades de grupos contribuíram para os 4 *outliers* formados, além aumentar a variação dos resultados dos *boxplot*. O método de busca da RID não apresentou *outliers*, mas as quantidades ideais de agrupamentos variaram de 224 até 611, resultando em uma grande amplitude dos resultados contidos no *boxplot*, principalmente na avaliação do maior deslocamento.

Os *boxplots* mostram de uma forma gráfica a variação dos dados observados. No entanto, é necessário realizar uma análise estatística para confirmar quais técnicas obtiveram vantagens significativas em relação às outras. Deste modo, foi verificado se as amostras obtidas cumpriam as premissas para a realização da Análise de Variância (ANOVA), sendo essas premissas: normalidade, independência e homocedasticidade. A Figura 44 ilustra a verificação das premissas da ANOVA.

As Figuras 44 (a) e (b) ilustram a validação de normalidade. Cada resíduo (pontos azul) são comparados com os seus respectivos valores esperados (linha vermelha). Pequenas variações como nestas figuras são toleráveis.

A premissa de independência visa garantir que uma observação não influencia a outra, de modo que os resultados não sejam previsíveis. As Figuras 44 (c) e (d) indicam que esta premissa também foi cumprida por não existir um padrão definido entre os valores dos resíduos, não sendo possível prever o próximo valor.

O cumprimento da premissa de homocedasticidade é ilustrado pelas Figuras 44 (e) e (f). Nesta premissa, os dados devem estar igualmente distribuídos, com os resíduos obtendo valores acima e abaixo de 0. Pequenas variações são toleráveis.

Após verificar que todas as premissas foram cumpridas, foi possível realizar o experimento da ANOVA para 1 fator. As tabelas da ANOVA para os resultados de RMSD e maior deslocamento são ilustradas respectivamente pelas Tabelas 3 e 4.

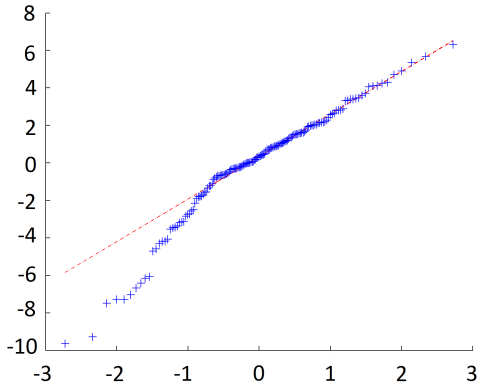
Tabela 3 – Tabela ANOVA para os experimentos da primeira base de dados, considerando o RMSD de até 0,5 Å

Fonte de var.	Soma dos quad.	Graus de lib.	Quad. méd.	F-est.	Prob>F
Colunas	2391,46	4	597,84	68,31	8,04181e-33
Erro	1312,75	150	8,752		
Total	3704,2	154			

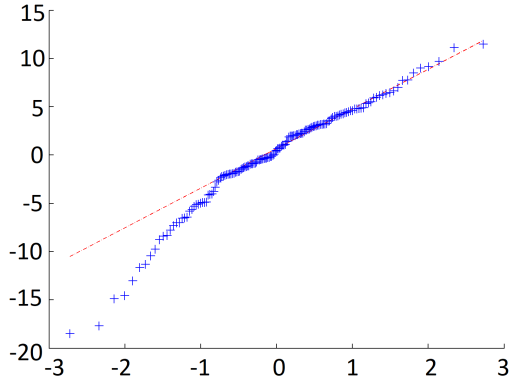
Fonte: Elaborada pelo autor (2023)

Figura 44 – Premissas para utilizar a ANOVA considerando a porcentagem total de sobreposições com até 0,5 Å de RMSD e maior deslocamento, para a base de dados 1

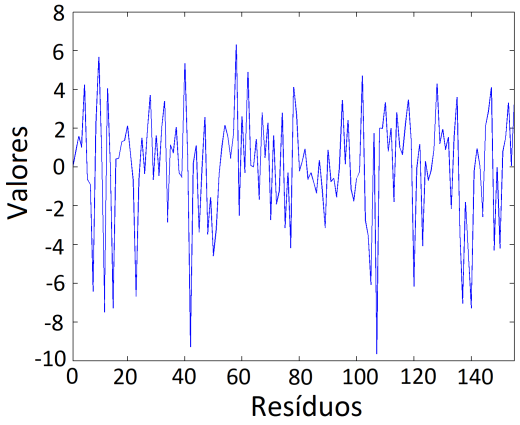
(a) Normalidade dos experimentos do RMSD



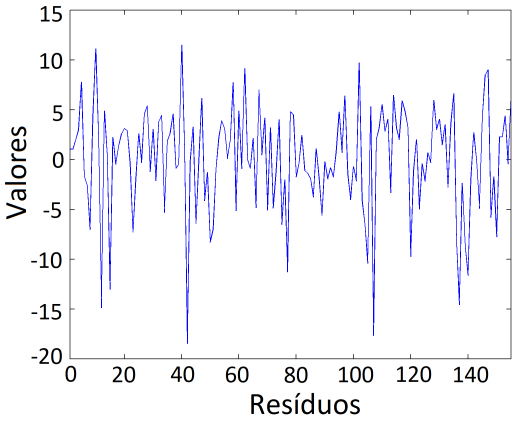
(b) Normalidade dos experimentos do maior deslocamento



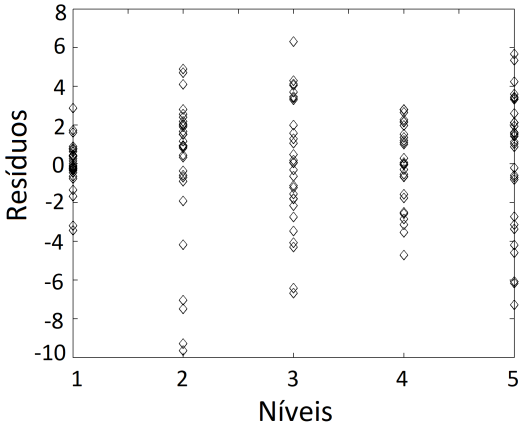
(c) Independência dos experimentos do RMSD



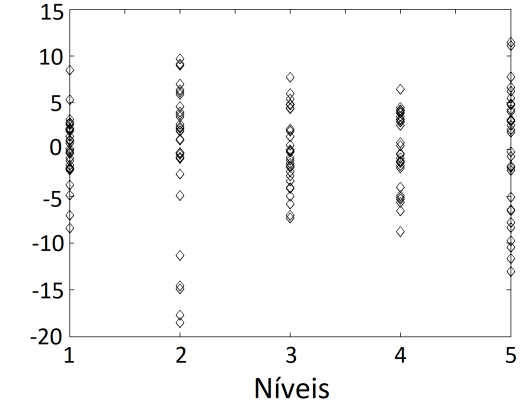
(d) Independência dos experimentos do maior deslocamento



(e) Homocedasticidade dos experimentos do RMSD



(f) Homocedasticidade dos experimentos do maior deslocamento



Fonte: Elaborada pelo autor (2023)

Tabela 4 – Tabela ANOVA para os experimentos da primeira base de dados, considerando o maior deslocamento de até 0,5 Å

Fonte de var.	Soma dos quad.	Graus de lib.	Quad. méd.	F-est.	Prob>F
Colunas	1227,24	4	306,809	10,49	1,60828e-07
Erro	4387,11	150	29,247		
Total	5614,35	154			

Fonte: Elaborada pelo autor (2023)

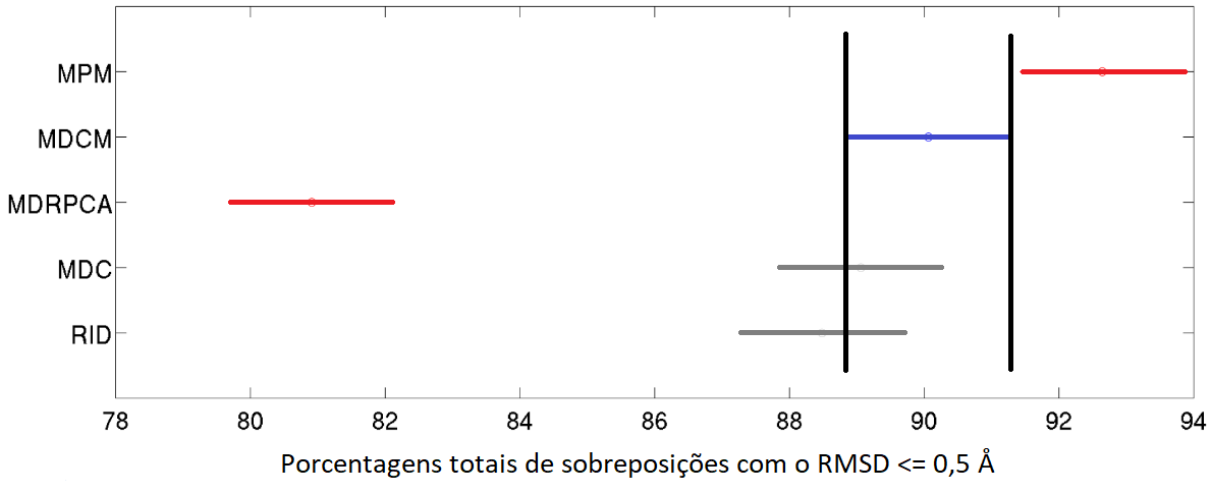
Conforme MathWorks (2006a) a ANOVA para 1 fator testa a hipótese nula de que todas as médias do grupo são iguais entre si, contra a hipótese alternativa de que pelo menos uma média é diferente das outras. Um valor p menor que o nível de significância definido (0,01) indica que pelo menos uma das médias da amostra é significativamente diferente das demais. De acordo com as Tabelas 3 e 4, o valor p obtido para o teste F (sexta coluna) indicou que ao menos uma das técnicas avaliadas possui a precisão diferente das demais. A partir dos resultados da ANOVA pode-se realizar o teste de Tukey para verificar com maior precisão quais técnicas obtiveram resultados significativamente diferentes das outras técnicas. O nível de confiança definido foi de 99%. Os resultados do teste de Tukey são ilustrados pela Figura 45.

Conforme a Figura 45 (a), ao avaliar a porcentagem total de sobreposições que obtiveram o RMSD com até 0,5 Å, pode-se notar através das linhas verticais próximas aos resultados da MDC, que os resultados da MPM ficaram na frente dessas linhas verticais, indicando que a MPM obteve resultados significativamente superiores em relação às demais técnicas, com o nível de confiança de 99%. Ao comparar as técnicas MDCM, MDC e a estratégia de busca da RID, pode-se constatar através das linhas verticais, que os resultados dessas três técnicas foram próximos, não ocorrendo diferenças significativas para o nível de confiança definido. A MDRPCA obteve resultados estatisticamente inferiores em relação a todas as técnicas comparadas.

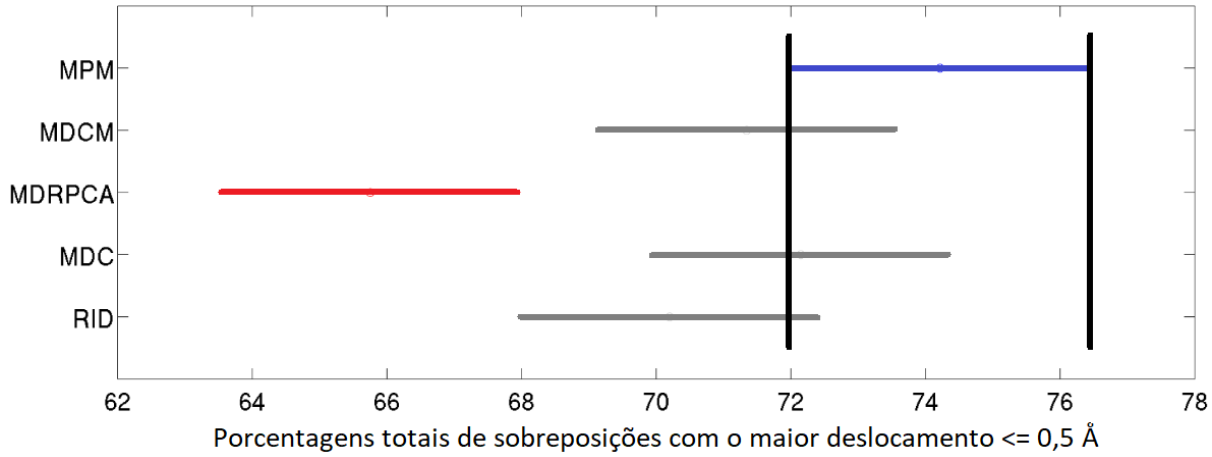
A Figura 45 (b) refere-se à avaliação estatística quanto ao percentual de sobreposições realizadas, que resultaram no maior deslocamento de até 0,5 Å. Neste experimento pode-se notar que a MPM obteve uma leve vantagem em relação à MDCM, MDC e RID. No entanto, nesse experimento não ocorreu diferença significativa entre essas técnicas. As únicas diferenças significativas para o nível de confiança definido foram da MPM, MDCM e MDC que superaram a MDRPCA. O método de busca da RID apresentou uma vantagem em relação à MDRPCA, mas sem indicar uma diferença significativa. Nesse experimento também pode-se constatar que a RID obteve um empate técnico com todas as outras técnicas comparadas.

Figura 45 – Teste de Tukey com as porcentagens de aproveitamentos para a primeira base de dados

(a) Teste de Tukey considerando RMSD de até 0,5 Å



(b) Teste de Tukey considerando o maior deslocamento de até 0,5 Å



Fonte: Elaborada pelo autor (2023)

6.1.1.1 Análise da Performance Computacional

A Tabela 5 ilustra o tempo médio de execução para construir os dados a serem agrupados, o tempo para realizar os agrupamentos e o tempo total. Os dados de tempo nesta tabela estão no formato de minutos : segundos. Os valores foram arredondados para o segundo mais próximo. Apesar da etapa de alinhamento estrutural para realizar sobreposições atômicas ser muito demorada, ela não foi considerada nas análises de tempo deste trabalho, pois ela foi empregada com o intuito de validar os resultados, assim não fazendo parte diretamente das composições das metodologias apresentadas e avaliadas neste trabalho.

Conforme a Tabela 5, a MDRPCA foi a metodologia mais rápida em todas as etapas. Os principais motivos foram: poder trabalhar diretamente com os arquivos de interações

Tabela 5 – Tempo para realização dos experimentos com o *k-means clustering* para a base de dados 1

Técnica	Tempo const. dados	Tempo agrup.	Tempo total
MDRPCA	00:01	00:22	00:23
MDC	00:09	00:34	00:43
MPM	00:19	00:31	00:50
RID	02:05	00:28	02:33
MDCM	02:43	00:41	03:24

Fonte: Elaborada pelo autor (2023)

(ao contrário da RID), não ser necessário calcular ângulos (como na MDCM) e não ser necessário realizar vários cálculos de distâncias Euclidiana como na (MPM, MDC e MDCM). É importante ressaltar que nessa tabela não foram contabilizados o tempo gasto para encontrar a interação de referência utilizada pela ferramenta RID. Também não foram contabilizados os tempos gastos para fazer as validações de sobreposições atômicas com o LSQKAB (procedimento feito por todas as técnicas para medir a acurácia).

6.1.2 OPTICS

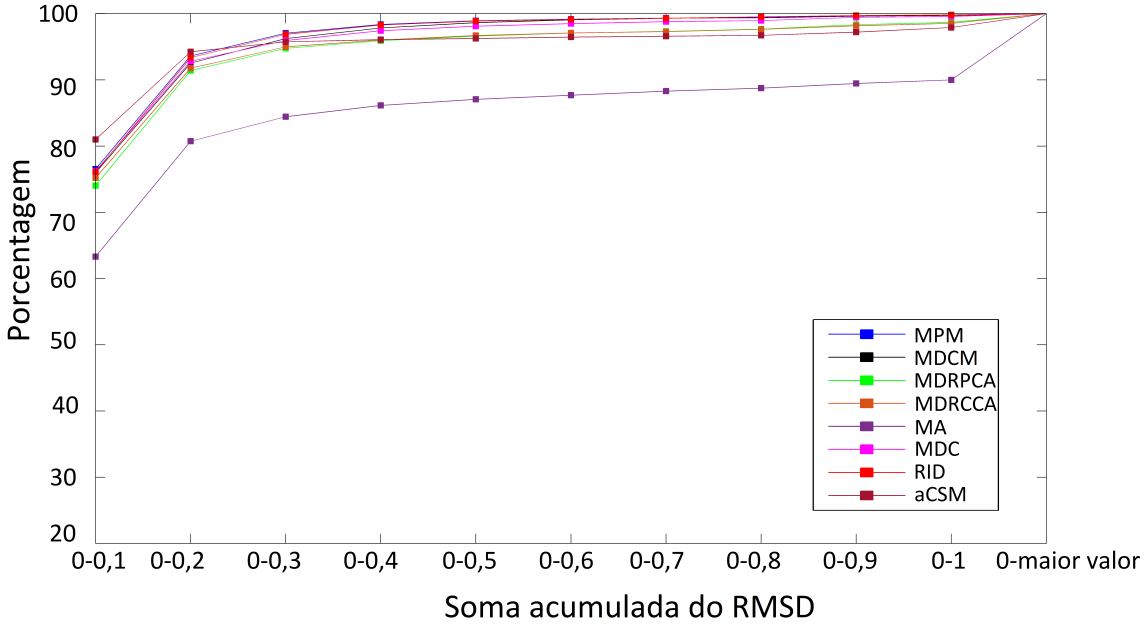
Ao contrário do *k-means clustering*, o OPTICS é um algoritmo determinístico, que não tem o resultado final influenciado por sementes. Deste modo, foi realizado apenas uma execução de experimento para cada técnica que atendesse a quantidade mínima de *clusters* indicada na seção de metodologia. A outra condição, de filtrar técnicas a partir de resultados de trabalhos anteriores, não foi aplicada a esses experimentos, pois o presente trabalho é o primeiro a combinar essas técnicas de representações de proteínas e o OPTICS. Assim, todas as 8 técnicas de representações de proteínas estudadas neste trabalho foram avaliadas juntamente com o OPTICS.

Os experimentos com o OPTICS para a base de dados 1 foram realizados com 3 diferentes métricas de distâncias: *Minkowski*, Euclidiana e Euclidiana quadrática. Para essas três métricas, as extrações da quantidade ideal de agrupamentos foi feita automaticamente pelo algoritmo do OPTICS. A Figura 46 (a) ilustra a soma acumulada do valor percentual médio quanto ao maior deslocamento.

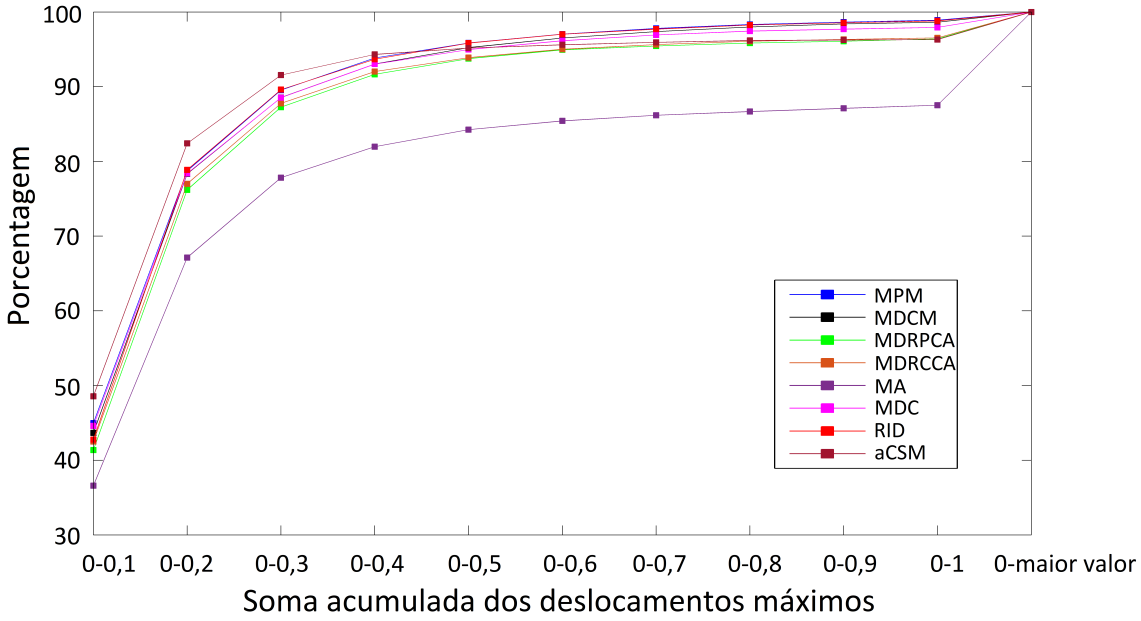
Conforme a Figura 46 (a) todas as técnicas de representações de proteínas obtiveram resultados satisfatórios ao combiná-las com o OPTICS, com maiores destaques para as metodologias MPM, MDCM, MDC, método de busca da RID e MDRPCA, que obtiveram os melhores resultados. Essas 5 técnicas obtiveram mais de 90% de sobreposições realizadas com até 0,2 Å. Ao considerar as sobreposições com até 0,5 Å de RMSD, as metodologias MPM, MDCM, MDC e RID apresentam acurácias superiores a 98%. Para essa faixa, a

Figura 46 – Soma acumulada com os resultados das sobreposições da base de dados 1, através do OPTICS, com a métrica de distância Minkowski

(a) Resultados conforme o RMSD



(b) Resultados conforme o maior deslocamento



Fonte: Elaborada pelo autor (2023)

MDRPCA, MDRCCA e aCSM também superaram 90% de precisão. Nesta análise, a Matriz de Ângulos apresenta um precisão próxima a 80%, ficando abaixo das demais técnicas avaliadas.

Através da Figura 46 (b) pode-se notar que na avaliação do maior deslocamento 7 técnicas de representações de proteínas obtiveram os resultados parecidos. Apenas a MA obteve resultados abaixo das demais. Nessa avaliação pode-se destacar principalmente a aCSM por apresentar os maiores valores das somas acumuladas de 0 até 0,4 Å. No entanto, ao considerar a soma acumulada até 0,5 Å, a MPM supera todas metodologias avaliadas. É importante ressaltar que embora o OPTICS tenha apresentado resultados melhores do que o *k-means clustering*, dois fatores contribuíram para esses resultados. O primeiro fator consiste no grande número de interações que foram descartadas pelo OPTICS durante o agrupamento. O segundo fator refere-se à grande quantidade de grupos determinada pelo OPTICS, que foi muito superior do que as quantidades determinadas pelo método *silhouette* para o *k-means*. A Tabela 6 indica as principais análises deste experimento.

Tabela 6 – Resultados do OPTICS com a métrica *Minkowski* para a base de dados 1

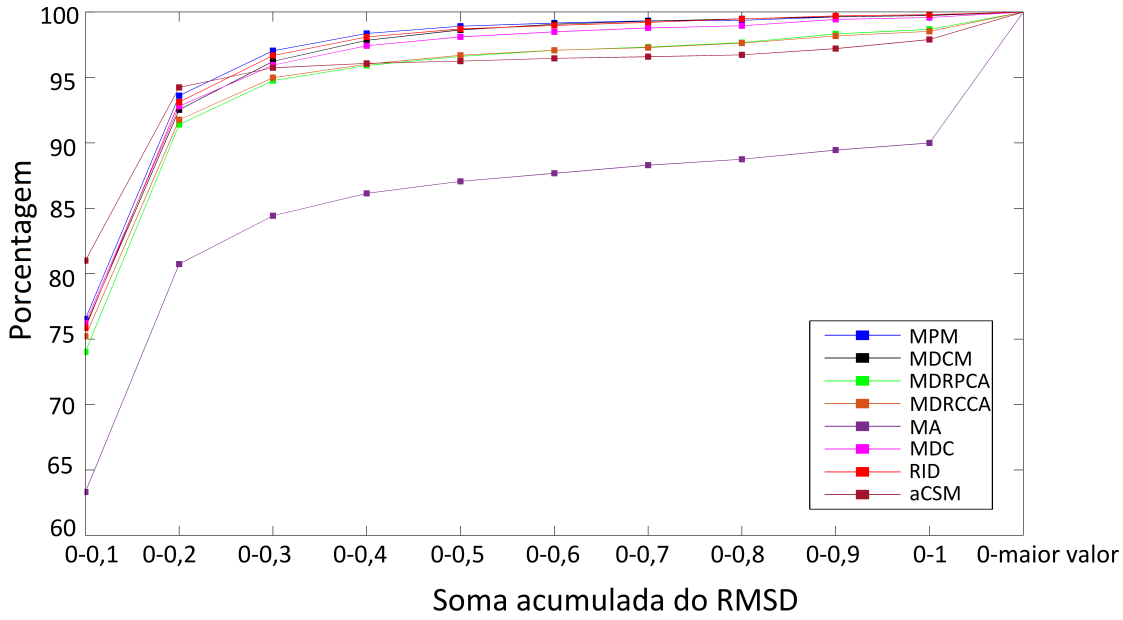
Técnica	RMSD <= 0,5Å	Maior D. <= 0,5Å	*Clusters*	Total Descart.	% Descart.
MPM	98,92%	95,85%	3.471	6.476	39,53%
RID	98,88%	95,83%	3.544	6.287	38,38%
MDCM	98,63%	95,24%	3.493	6.511	39,74%
MDC	98,10%	94,97%	3.478	6.560	40,04%
aCSM	96,25%	95,17%	2.832	8.641	52,74%
MDRCCA	96,70%	93,89%	3.505	6.615	40,38%
MDRPCA	96,61%	93,76%	3.551	6.486	39,59%
MAR	87,06%	84,25%	3.338	7.082	43,23%

Fonte: Elaborada pelo autor (2023)

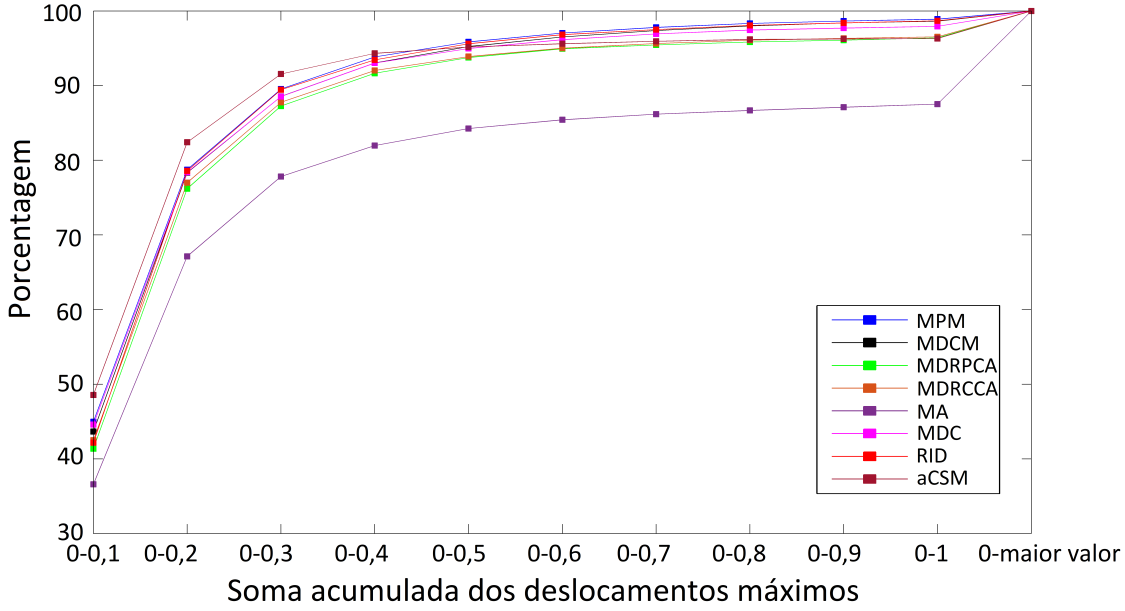
A Tabela 6 está ordenada pelas melhores precisões, considerando a soma dos percentuais de RMSD e maior deslocamento de até 0,5 Å (duas primeiras colunas). Deste modo, pode ser observado que a MPM foi a melhor, seguida pela RID, MDCM e MDC. A terceira coluna indica a quantidade de *clusters* definida automaticamente pelo OPTICS. Pode-se constatar que para todas as técnicas de representações de proteínas, a quantidade média de grupos foi quase 10 vezes maior do que a quantidade definida pelo método *silhouette* com o *k-means*. As duas últimas colunas referem-se à quantidade de interações descartadas pelo OPTICS durante o agrupamento. Para todas as técnicas, foram excluídas no mínimo 6.200 interações, chegando até a quantidade de 7.082 interações no caso da aCSM, que representou 52,74% da base de dados. A Figura 47 ilustra os resultados do OPTICS com a métrica de distância Euclidiana.

Figura 47 – Soma acumulada com os resultados das sobreposições da base de dados 1, através do OPTICS, com a métrica de distância Euclideana

(a) Resultados conforme o RMSD



(b) Resultados conforme o maior deslocamento



Fonte: Elaborada pelo autor (2023)

Através da Figura 47(a) pode-se observar um comportamento semelhante ao da validação do RMSD com a distância *Minkowski*. A MPM manteve-se com resultados superiores em relação às demais metodologias, mas com muita proximidade em relação aos resultados da MDCM, MDC e RID, que novamente obtiveram acurácias superiores a 98% ao avaliar o RMSD de até 0,5Å. A MDRPCA ficou um pouco abaixo dessas metodologias. Mas neste cenário ela superou novamente a MDRCCA, aCSM e MA.

A Figura 47 (b) indica que ao avaliar o maior deslocamento, novamente a aCSM obtêm os melhores resultados das somas acumuladas de até 0,4 Å. Sendo novamente superada pela MPM somente a partir do valor de soma acumulada de 0,5 Å. De um modo geral, novamente, 7 metodologias tiveram resultados muito próximos. Somente a MA ficou com resultados inferiores. A Tabela 7 indica as precisões obtidas nestes experimentos e informações sobre as interações descartadas.

Tabela 7 – Resultados do OPTICS com a métrica de distância Euclideana para a base de dados 1

Técnica	RMSD <= 0,5Å	Maior D. <= 0,5Å	*Clusters*	Total Descart.	% Descart.
MPM	98,92%	95,85%	3.471	6.476	39,53%
RID	98,71%	95,61%	3.738	5.669	34,60%
MDCM	98,63%	95,24%	3.493	6.511	39,74%
aCSM	96,25%	95,17%	2.832	8.641	52,74%
MDRCCA	96,69%	93,89%	3.506	6.615	40,38%
MDRPCA	96,61%	93,76%	3.551	6.486	39,59%
MDC	95,10%	94,97%	3.478	6.560	40,04%
MAR	87,06%	84,24%	3.338	7.082	43,23%

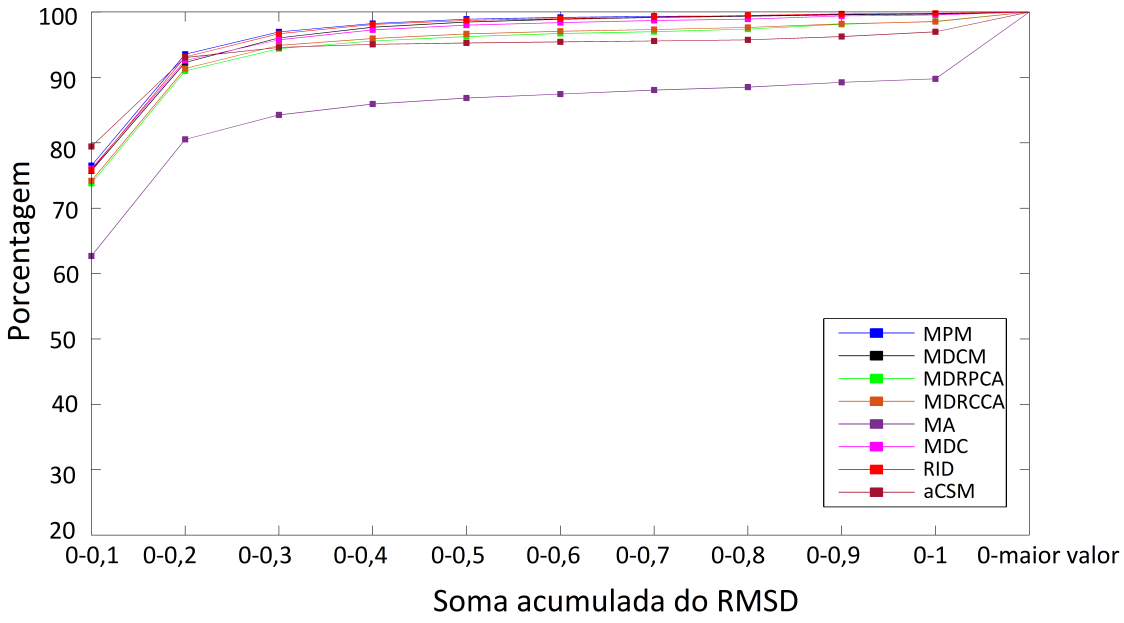
Fonte: Elaborada pelo autor (2023)

Através da Tabela 7 pode-se verificar que as metodologias MPM, estratégia de busca da RID e a MDCM obtiveram resultados muito próximos, mas com uma leve superioridade da MPM. No entanto, todas as metodologias tiveram grandes quantidades de agrupamentos e muitos descartes de interações durante o *clustering*. Também pode-se destacar que a alteração da métrica de distâncias de *Minkowski* para Euclideana contribuiu para uma leve redução de descartes ao agrupar a representação feita pela RID, de 6.287 interações descartadas para 5.669, assim evitando o descarte de 618 interações da base de dados. No entanto, para as demais técnicas comparadas, não ocorreram reduções consideráveis de descartes. A Figura 48 ilustra os resultados com a métrica de distância Euclidiana quadrática.

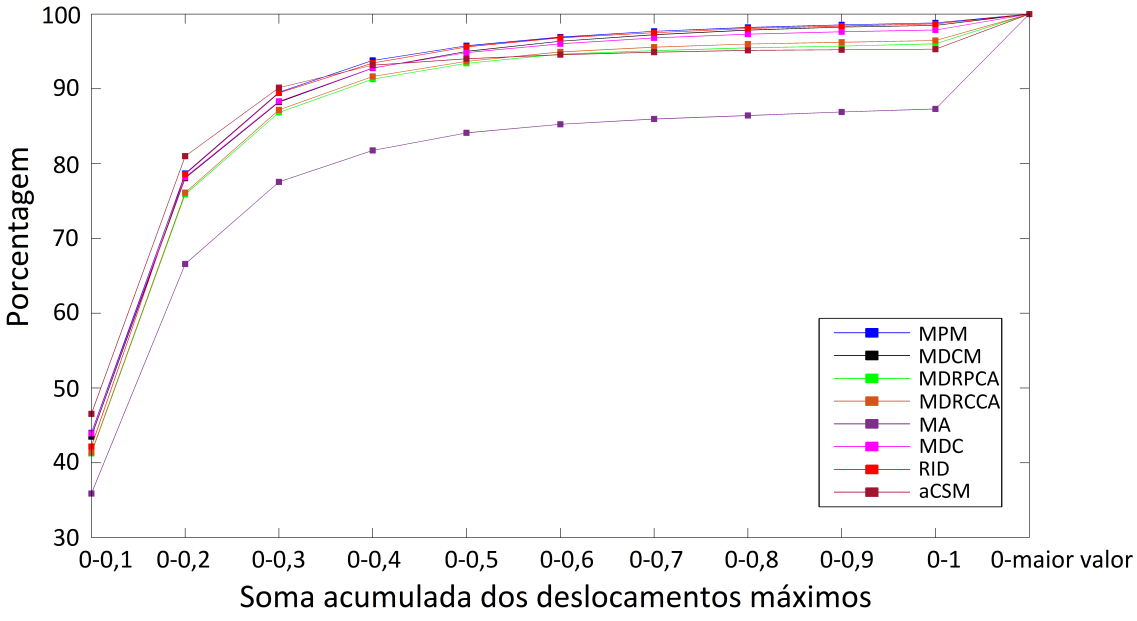
Conforme a Figura 48 (a) e (b) os comportamentos das avaliações do RMSD e maior deslocamento obtidos com a métrica de distância Eucliana quadrática foram próximos dos resultados obtidos pelas outras duas métricas de distâncias testadas anteriormente. Entretanto, na Tabela 8 pode-se notar uma pequena redução da quantidade de interações

Figura 48 – Soma acumulada com os resultados das sobreposições da base de dados 1, através do OPTICS, com a métrica de distância Euclideana quadrática

(a) Resultados conforme o RMSD



(b) Resultados conforme o maior deslocamento



Fonte: Elaborada pelo autor (2023)

descartadas durante o agrupamento, que ocorreu para todas as técnicas avaliadas.

Tabela 8 – Resultados do OPTICS com a métrica de distância Euclidiana quadrática para a base de dados 1

Técnica	RMSD <= 0,5Å	Maior D. <= 0,5Å	*Clusters*	Total Descart.	% Descart.
MPM	98,91%	95,76%	3.691	5.803	35,42%
RID	98,71%	95,61%	3.739	5.669	34,60%
MDCM	98,46%	95,01%	3.676	5.921	34,14%
MDC	97,99%	94,84%	3.666	5.924	36,16%
MDRCCA	96,65%	93,68 %	3.709	5.920	36,14%
aCSM	95,26%	94,02 %	3.250	7.217	44,05%
MDRPCA	96,25%	93,42 %	3.751	5.780	35,28%
MAR	86,85%	84,12%	3.550	6.422	39,20%

Fonte: Elaborada pelo autor (2023)

De acordo com a Tabela 8, a quantidade de interações descartadas durante o agrupamento diminuiu com a métrica de distância Euclidiana quadrática. Essa redução de registros excluídos ocorreu para todas as 8 metodologias avaliadas com o OPTICS. A técnica que apresentou a maior redução de descartes foi a aCSM, que com as outras métricas de distâncias descartou 8.641 interações, nesse último experimento descartou 7.217 registros, uma redução de 1.424 exclusões, quase 9% da base de dados. Deste modo, a métrica de distância Euclidiana quadrática apresentou uma melhor solução nos experimentos com o OPTICS, pois as acurácias de todas as técnicas de representações de proteínas continuaram altas, mesmo agrupando mais registros do que foi possibilitado pelas outras métricas de distâncias. Entretanto, é importante ressaltar que de um modo geral ainda ocorreram muitos descartes com o OPTICS.

6.1.2.1 Análise da Performance Computacional

A Tabela 9 apresenta o tempo total de execução para cada teste realizado com o método OPTICS. Diferentemente dos experimentos de agrupamento *k-means*, nesta tabela, o tempo requerido para construir a matriz que será agrupada não é listado separadamente, pois esses valores permanecem consistentes. Contudo, é importante mencionar que as três metodologias testadas com o método OPTICS não foram incluídas nos testes de agrupamento *k-means*. Portanto, os tempos necessários para construir as respectivas matrizes dessas metodologias são detalhados a seguir: a MDRCCA gasta 2 segundos para realizar este procedimento, a MA gasta 2 minutos e 43 segundos para construir a matriz de ângulos e a aCSM demora 3 minutos e 13 segundos para construir a matriz aCSM.

A Tabela 9 está ordenada conforme as metodologias mais rápidas neste experimento. Através desta tabela pode-se verificar que a MDRCCA foi a metodologia mais rápida, seguida pela MDRPCA. Nos resultados obtidos através do OPTICS, constata-se que essas duas

Tabela 9 – Tempo para realizar os experimentos com o OPTICS, para a base de dados 1

Técnica	Mét. Dist. Euclidiana.	Mét. Didt. Euclid. Quadrática	Mét. Dist. Minkowski
MDRCCA	00:50	00:55	02:29
MDRPCA	00:49	00:59	02:32
MPM	01:13	01:21	02:46
MDC	01:27	01:49	04:48
RID	02:34	02:42	03:10
MA	03:10	03:09	03:42
MDCM	04:05	04:32	08:00
aCSM	05:28	06:46	10:34

Fonte: Elaborada pelo autor (2023)

matrizes de distâncias reduzidas conseguiram cumprir seus objetivos, ou seja, o de formar agrupamentos com boas acurácias e de um modo rápido. A MPM foi a terceira metodologia mais rápida, seguida respectivamente por: MDC, RID, MA, MDCM e aCSM. Ao avaliar o tempo gasto computacionalmente conforme a métrica de distância utilizada no OPTICS, é possível observar que na maioria das vezes, utilizar a métrica de distância Euclidiana resultou no processamento mais rápido. A métrica de distância Euclidiana quadrática foi a segunda mais rápida, mas é importante levar em consideração que ela descartou uma quantidade menor de interações de proteínas, fazendo com que ela trabalhasse com mais registros durante o agrupamento. A métrica de distância *Minkowski* foi a mais demorada para todas as técnicas de representações de proteínas em que ela foi testada.

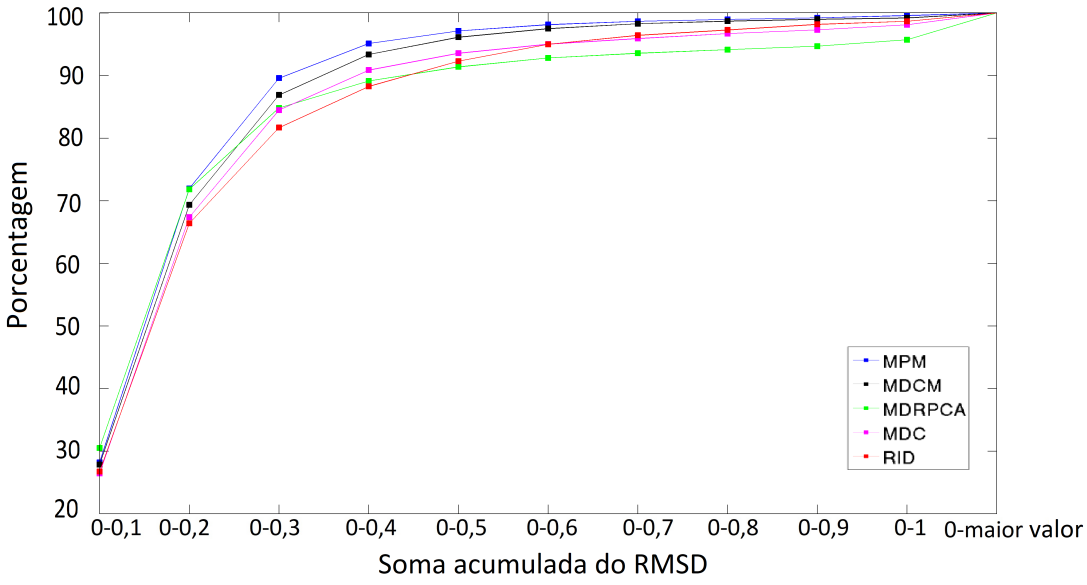
6.1.3 *Hierarchical Clustering*

Do mesmo modo que o OPTICS, o *hierarchical clustering* é um algoritmo determinístico, que não é impactado por sementes. Deste modo, realizou-se apenas uma execução de experimento combinando cada técnica de representação de proteínas com um diferente tipo de ligação no *hierarchical clustering*, desde que a quantidade mínima de grupos fosse atendida na verificação com o *silhouette*. A validação de resultados obtidos em trabalhos anteriores também não se aplica aos experimentos com o *hierarchical clustering*, pois o presente trabalho é o primeiro a combinar as técnicas de representações de proteínas avaliadas neste trabalho com este algoritmo de agrupamentos. A Figura 49 (a) e (b) indicam os gráficos de soma acumulada para a validação do RMSD e maior deslocamento, com o método de ligação Ward. As técnicas de representações de proteínas que participaram deste experimento foram: MPM, MDCM, MDRPCA, MDC e RID.

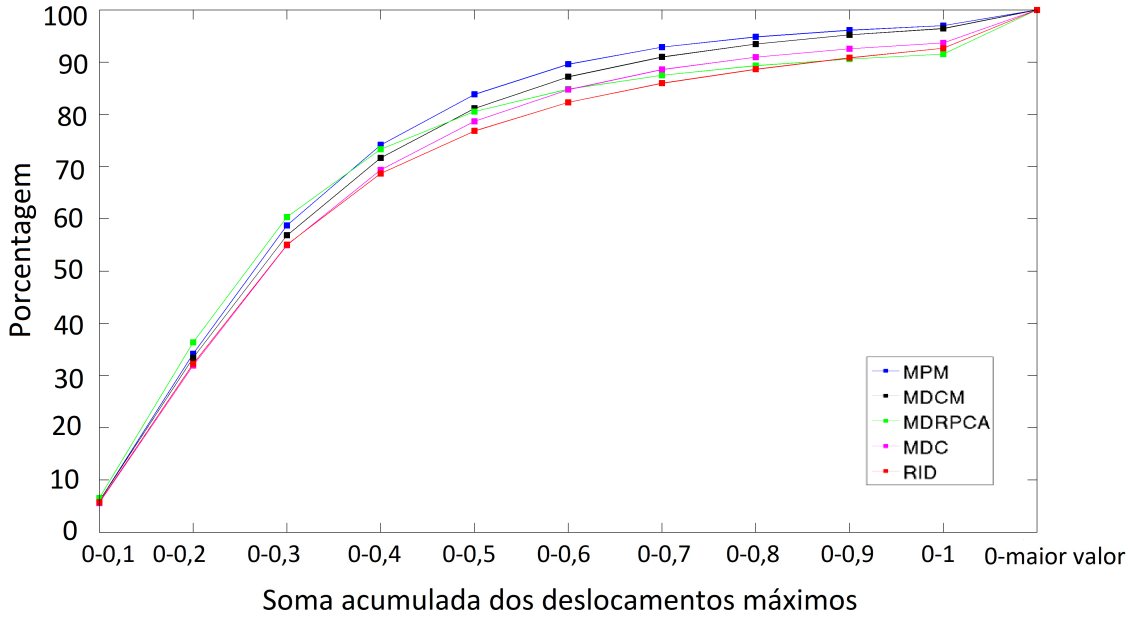
Conforme a Figura 49 (a) pode-se observar que na avaliação do RMSD com até 0,1 Å, a MDRPCA tem resultados superiores do que as demais metodologias. No entanto, a partir da soma acumulada de 0,1 Å, a MPM torna-se a técnica de maior precisão, até o final deste

Figura 49 – Soma acumulada com os resultados das sobreposições da base de dados 1, através do *Hierarchical clustering*, com o método de ligação Ward

(a) Resultados conforme o RMSD



(b) Resultados conforme o maior deslocamento



Fonte: Elaborada pelo autor (2023)

gráfico. Na faixa de soma acumulada de até 0,5 Å , as melhores acurácias alcançadas foram respectivamente pelas técnicas: MPM, MDCM, MDC, método de busca da RID e MDRPCA.

De acordo com a Figura 49 (b), na avaliação do maior deslocamento, as cinco metodologias começam com os valores muito próximos, mas com uma superioridade da MDRPCA até a soma acumulada de 0,3 Å. No entanto, a partir da soma acumulada de até 0,4 Å, a MPM tornou-se a metodologia de maior precisão, mantendo os melhores resultados até o fim do gráfico. Na faixa de soma acumulada de até 0,5 Å as melhores acurácias alcançadas foram respectivamente pelas técnicas: MPM, MDCM, MDRPCA, MDC e RID. A Tabela 10 indica as principais informações desse experimento.

Tabela 10 – Resultados do *hierarchical clustering* com o método de ligação Ward para a base de dados 1

Técnica	RMSD <= 0,5Å	Maior D. <= 0,5Å	*Clusters*	Quant. *Cl.* 1 inter.
MPM	97,12%	83,82%	583	0
MDCM	96,11%	81,15%	583	4
MDC	93,55%	78,68 %	475	0
MDRPCA	91,38%	80,54%	539	0
RID	92,30%	76,84 %	476	0

Fonte: Elaborada pelo autor (2023)

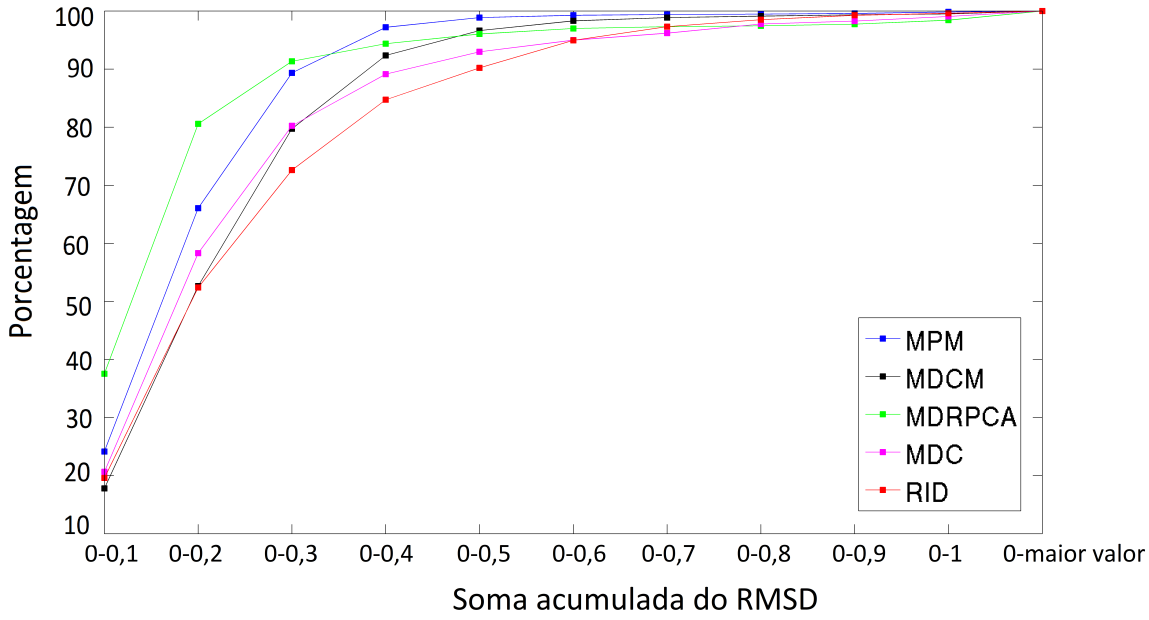
A Tabela 10 contém os principais resultados, ordenados pelas maiores acurácias, considerando os valores de RMSD e maior deslocamento de até 0,5 Å. Nessa tabela pode-se notar que o método de ligação Ward obteve uma melhor combinação com a técnica de representação de agrupamentos MPM. Nessa avaliação obteve-se 97,12% de acurácia ao considerar o RMSD de até 0,5 Å e 83,82 % ao considerar o maior deslocamento de até 0,5 Å. Também é possível constatar que todas as outras 4 técnicas de representações de proteínas avaliadas obtiveram resultados satisfatórios ao combiná-las com o método de ligação Ward, podendo ainda destacar que poucos *clusters* tiveram apenas 1 registro (somente 4 ocorrências na MDCM).

Para o método de ligação *complete*, as mesmas técnicas de representação de proteínas participaram do experimento. A Figura 50 (a) e (b) ilustram as avaliações do RMSD e do maior deslocamento respectivamente.

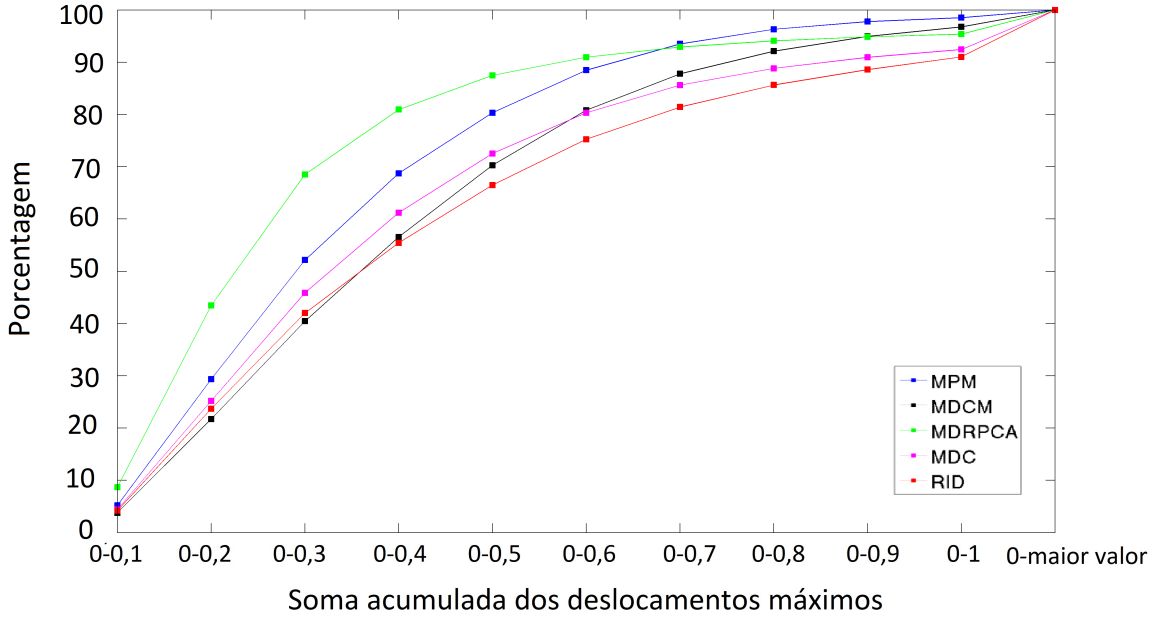
Conforme a Figura 50 (a) pode-se observar que até a soma acumulada de 0,3 Å, a MDRPCA apresenta superioridade nos resultados em relação às demais metodologias. No entanto, a partir da soma acumulada de 0,4 Å, a MPM começa a apresentar os melhores resultados deste experimento, mantendo-os até o fim do gráfico. Ao considerar a soma acumulada de até 0,5 Å a ordem de melhores resultados foram: MPM, MDCM, MDRPCA, MDC e RID. Conforme a Figura 50 (b), a MDRPCA apresentou os melhores resultados durante todas as

Figura 50 – Soma acumulada com os resultados das sobreposições da base de dados 1, através do *Hierarchical clustering*, com o método de ligação *complete*

(a) Resultados conforme o RMSD



(b) Resultados conforme o maior deslocamento



Fonte: Elaborada pelo autor (2023)

faixas de somas acumuladas avaliadas, principalmente nas faixas de 0 até 0,4 Å. A MPM obteve o segundo melhor resultado nesse experimento, seguida pela MDC, MDCM e RID. A Tabela 11 mostra as principais informações deste experimento.

Tabela 11 – Resultados do *hierarchical clustering*, com o método de ligação *complete*, para a base de dados 1

Técnica	RMSD <= 0,5Å	Maior D. <= 0,5Å	*Clusters*	Quant. *cl.* 1 inter.
MDRPCA	96,03%	87,48%	1.588	216
MPM	98,87 %	80,34%	1.166	168
MDCM	96,63%	70,28%	905	108
MDC	92,98%	72,53%	1.606	122
RID	90,20%	66,50%	853	87

Fonte: Elaborada pelo autor (2023)

Conforme a Tabela 11, a MDRPCA foi a técnica de representação de proteínas que obteve uma melhor combinação com o método de ligação *complete*. Embora a MPM tenha obtido um melhor resultado na avaliação do RMSD de até 0,5 Å, a MDRPCA apresentou a melhor acurácia ao considerar a avaliação do RMSD e do maior deslocamento. No entanto, ao comparar os resultados do método de ligação *complete* com o Ward (Tabela 10) pode-se notar que através do método *complete* cada técnica de representação de proteínas formou mais grupos e a quantidade de *clusters* que tiveram apenas 1 registro aumentou consideravelmente para todas as técnicas.

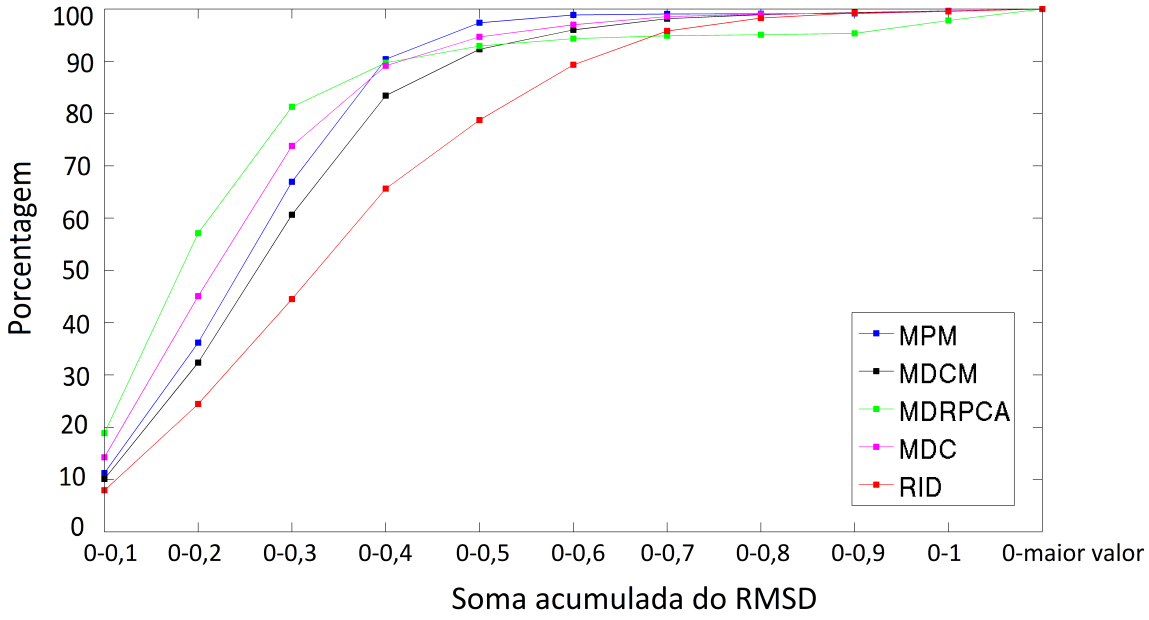
As Figuras 51 (a) e (b) ilustram os resultados das avaliações de RMSD e maior deslocamento obtidos com o método de ligação *average*. Participaram deste experimento: MPM, MDCM, MDRPCA, MDC, RID.

Conforme a Figura 51 (a), pode-se notar que na avaliação do RMSD, a MDRPCA apresentou os melhores resultados até a soma acumulada de 0,3 Å. A partir da próxima faixa de soma acumulada de porcentagens (0,4 Å) a MPM passou a obter os melhores resultados. Ao considerar os valores de RMSD de até 0,5 Å, as melhores metodologias foram respectivamente: MPM, MDC, MDRPCA, MDCM e RID.

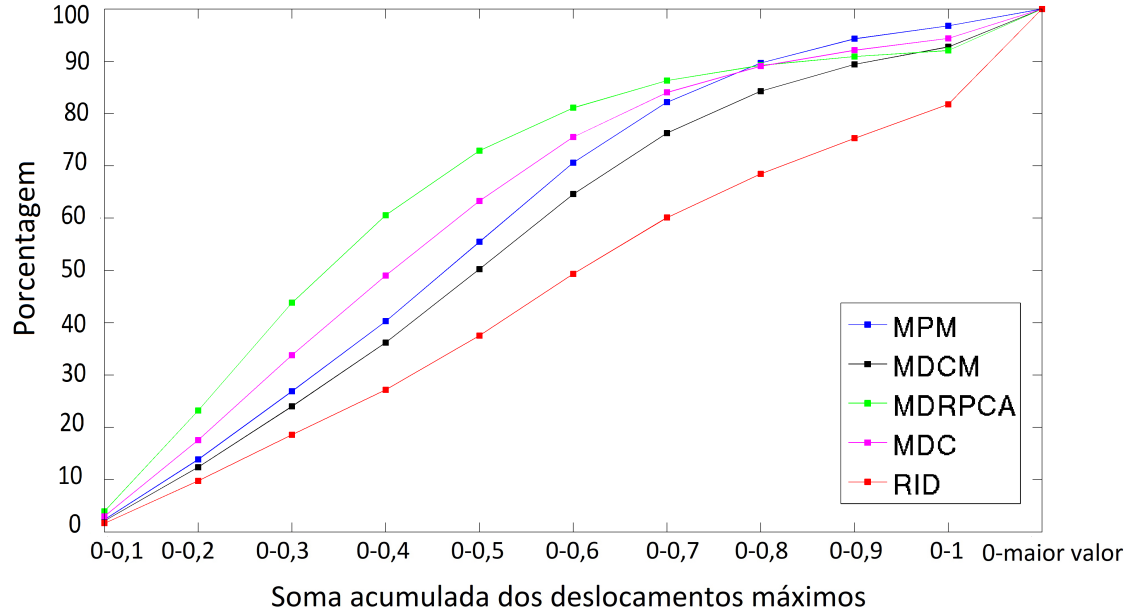
Através da Figura 51 (b), pode-se observar a superioridade dos resultados da MDRPCA nas principais faixas de somas acumuladas. Nesta avaliação, a MDC ficou em segundo lugar, seguida pela MPM, MDCM e RID. Ao avaliar juntamente as Figuras 51 (a) e 51 (b) pode-se notar que a combinação entre a MDRPCA e o *hierarchical clustering* com o método de ligação *average* obteve uma boa acurácia na formação de grupos de interações de proteínas com baixos valores de RMSD e maior deslocamento. A Tabela 12 ilustra as principais informações deste experimento.

Figura 51 – Soma acumulada com os resultados das sobreposições da base de dados 1, através do *Hierarchical clustering*, com o método de ligação *average*

(a) Resultados conforme o RMSD



(b) Resultados conforme o maior deslocamento



Fonte: Elaborada pelo autor (2023)

Tabela 12 – Resultados do *hierarchical clustering*, com o método de ligação *average*, para a base de dados 1

Técnica	RMSD <= 0,5Å	Maior D. <= 0,5Å	*Clusters*	Quant. *cl.* 1 inter.
MDRPCA	92,93%	72,90%	1.425	293
MDC	94,66%	63,29%	1.380	289
MPM	97,37%	55,46%	1.341	290
MDCM	92,31%	50,24%	1.067	219
RID	78,75%	37,54%	816	149

Fonte: Elaborada pelo autor (2023)

Conforme a Tabela 12, a MDRPCA apresentou a melhor combinação de resultados considerando o RMSD e o maior deslocamento de até 0,5 Å. Embora todas as técncias com exceção do método de busca da RID tenham obtido mais de 92% acurácia ao considerar o RMSD de até 0,5 Å, é importante ressaltar que de um modo geral, os resultados obtidos na avaliação do maior deslocamento foram ruins para todas as técnicas de representações de proteínas. Nesse experimento também é possível destacar que cada técnica obteve grandes quantidades de agrupamentos definidos como ideal conforme o método *silhouette*, bem como tiveram muitos *clusters* com apenas 1 registro, o que não é interessante para o problema em questão, pois esses registros não irão ser sobrepostos atomicamente com nenhuma outra interação.

As Figuras 52 (a) e 52 (b) ilustram os resultados para o método de ligação *single*. Para estes experimentos foram testadas apenas duas técnicas de representações de proteínas, a MPM e a RID. Conforme o método *silhouette*, todas as outras metodologias não apresentaram quantidades de grupos compatíveis com o problema em questão.

Conforme a Figura 52 (a) e (b), tanto para a avaliação do RMSD e quanto para o maior deslocamento, a MPM apresentou resultados consideravelmente supreriores do que os resultados apresentados pela RID. A Tabela 13 ilustra as principais informações deste experimento.

Tabela 13 – Resultados do *hierarchical clustering* com o método de ligação *single* para a base de dados 1

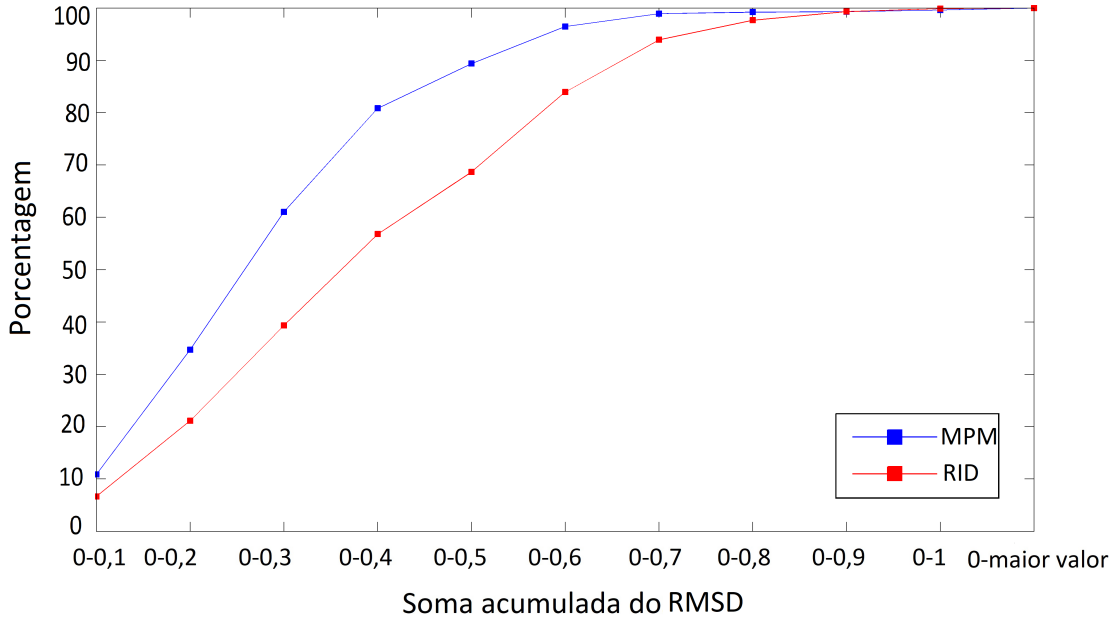
Técnica	RMSD <= 0,5Å	Maior D. <= 0,5Å	*Clusters*	Quant. *cl.* 1 inter.
MPM	89,37%	51,68%	3.078	1.041
RID	68,66%	32,86%	2.838	969

Fonte: Elaborada pelo autor (2023)

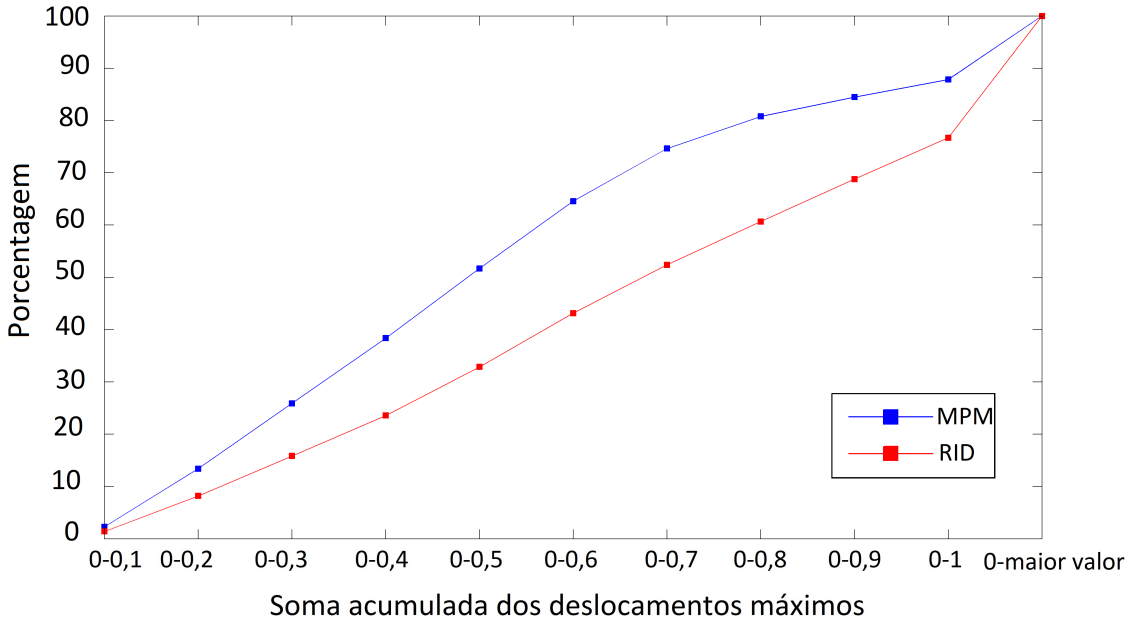
Conforme a Tabela 13, os resultados da MPM foram muito superiores do que os da RID. No entanto, essas acurácias ficaram abaixo dos outros métodos de ligações do *hierarchical*

Figura 52 – Soma acumulada com os resultados das sobreposições da base de dados 1, através do *Hierarchical clustering*, com o método de ligação *single*

(a) Resultados conforme o RMSD



(b) Resultados conforme o maior deslocamento



Fonte: Elaborada pelo autor (2023)

clustering que foram testados (Ward, *complete* e *average*). Além disso, o método *single* formou muitos agrupamentos com apenas 1 registro, o que não é interessante para o problema em questão.

6.1.3.1 Análise da Performance Computacional

A Tabela 14 ilustra o tempo total dos agrupamentos realizados através do *hierarchical clustering*, com os quatro métodos de ligações que foram testados. É importante ressaltar que para o método de ligação *single*, somente a MPM e o método de busca da RID conseguiram atender aos critérios estabelecidos na Seção de metodologia. E assim, infelizmente as técnicas MDRPCA, MDC e MDCM não foram combinadas com esse método de ligação.

Tabela 14 – Tempo para realização dos experimentos com o *hierarchical clustering* para a base de dados 1

Técnica	Ward	*Complete*	*average*	*single*
MDRPCA	00:12	00:11	01:19	-
MDC	00:31	00:32	00:40	-
MPM	00:44	00:40	00:41	00:30
RID	02:22	02:21	02:23	02:21
MDCM	03:03	03:04	03:14	-

Fonte: Elaborada pelo autor (2023)

Conforme a Tabela 14, a MDRPCA obteve o menor tempo total nas comparações feitas com os métodos de ligações Ward e *complete*. A MDC e a MPM foram as mais rápidas na combinação com o método de ligação *average*. A MPM também foi a mais rápida no experimento com o método *single*. Com exceção da MPM, todas as técnicas de representações de proteínas que foram testadas neste cenário tiveram uma maior demora com o método de ligação *average*.

6.2 Resultados Obtidos com a Base de Dados 2

Para a base de dados 2, que contém 129.705 interações de pontes de hidrogênio, foram realizados experimentos com o *k-means clustering* na subseção 6.2.1, e com o OPTICS na subseção 6.2.2. Para esta base de dados não foi possível realizar experimentos com o *hierarchical clustering* devido ao alto consumo de memória RAM deste algoritmo.

6.2.1 *K-Means Clustering*

Como a segunda base de dados é aproximadamente 8 vezes maior do que a primeira, e deste modo o processamento tornou-se mais demorado, optou-se por selecionar apenas as 3 técnicas que obtiveram as melhores acurácias nos experimentos com o *k-means*

na primeira base de dados, para avaliá-las melhor neste novo cenário. As técnicas de representações de proteínas selecionadas foram: MPM. MDCM e MDC. A Figura 53 (a) e (b) ilustram as somas acumuladas dos percentuais de sobreposições atômicas, conforme as distâncias de RMSD e maior deslocamento.

Conforme a Figura 53 (a) pode-se observar que a MDC e a MDCM começam com os melhores resultados até a soma acumulada de 0,2 Å de RMSD. Entretanto, a partir da soma acumulada de 0,3 Å, a MPM começa a apresentar os melhores resultados e os mantêm até o final deste gráfico. A Figura 53 (a) ainda indica que as médias das 3 metodologias comparadas foram superiores a 95% ao definir o RMSD de até 0,5 Å. Nesta avaliação, a média da MPM superou 97%. A Figura 53 (b) ilustra um comportamento um pouco parecido ao avaliar o maior deslocamento. No entanto, pode-se notar que os percentuais de acurácias das 3 metodologias foram levemente inferiores, devido ao maior rigor da avaliação do maior deslocamento. Também pode-se notar que a superioridade da MPM na soma acumulada de até 0,5 Å não foi tão destacada como na avaliação do RMSD. Esses comportamentos também foram constatados nos experimentos da primeira base de dados. A Tabela 15 contém as principais análises deste experimento.

Tabela 15 – Resultados do *k-means clustering* para a base de dados 2

Técnica	RMSD <= 0,5Å	Mai. D. <= 0,5Å	Mín. *cl.*	Med. *cl.*	Max.*Cl.*	*cl.* 1 reg.
MPM	97,03%	94,20%	19.900	21.436	23.050	2.152
MDCM	96,18%	93,87%	21.500	22.275	23.100	2.611
MDC	95,65%	93,48%	21.300	22.547	24.200	2.213

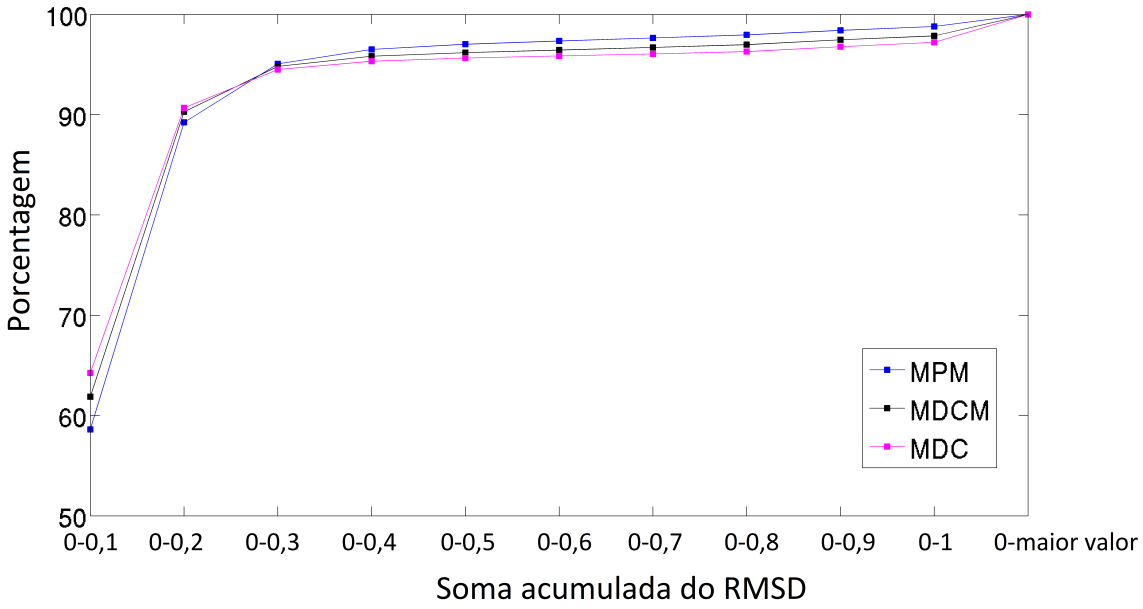
Fonte: Elaborada pelo autor (2023)

Conforme a Tabela 15, nota-se que a acurácia de todas as metodologias foram altas, principalmente a precisão da MPM, que foi superior tanto na avaliação do RMSD quanto na avaliação do maior deslocamento. Pode-se destacar também que os experimentos com essa base de dados obtiveram maiores acurácias do que os experimentos com a primeira base de dados. Um dos motivos deve-se ao método *silhouette* ter retornado quantidades maiores de grupos, como a quantidade ideal, para cada experimento. Mas é importante ressaltar que a MPM conseguiu os melhores resultados utilizando quantidades de *clusters* menores do que a MDC e a MDCM. O ponto negativo das técnicas de representações de proteínas trabalharem com uma maior quantidade de grupos deve-se à quantidade de interações de proteínas que ficaram sozinhas em seus respectivos *clusters*, o que não é interessante para o problema, pois essas interações não serão sobrepostas. A Figura 54 contém os *boxplots* considerando as acurácias totais de cada uma das 31 execuções de experimentos, considerando o RMSD e o maior deslocamento de até 0,5 Å.

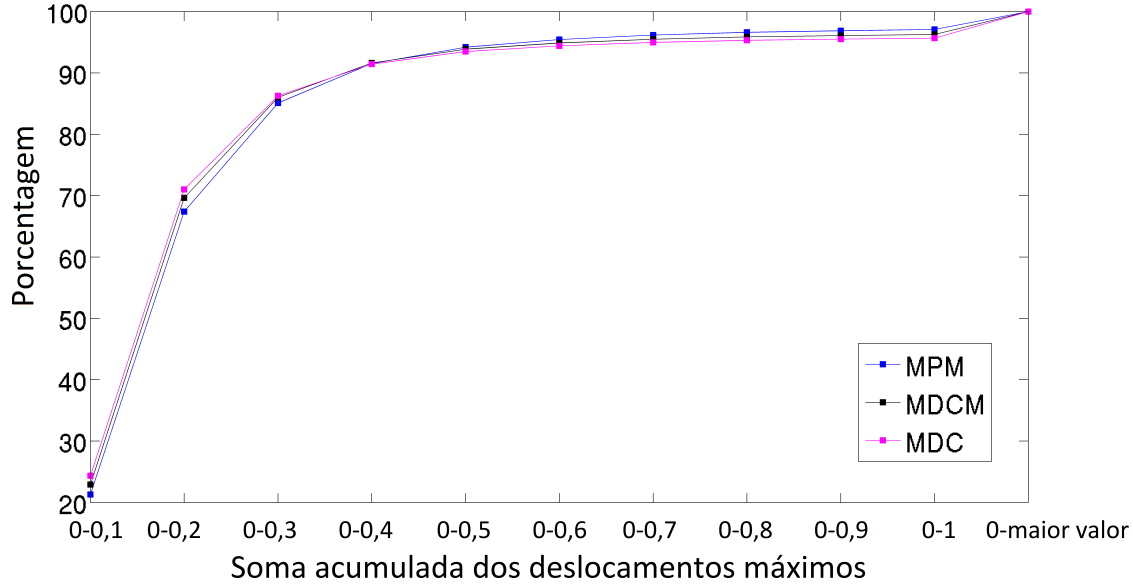
Conforme a Figura 54, os resultados da MPM foram melhores do que os resultados das

Figura 53 – Soma acumulada dos resultados com a base de dados 2, através do *k-means clustering*

(a) Resultados conforme o RMSD



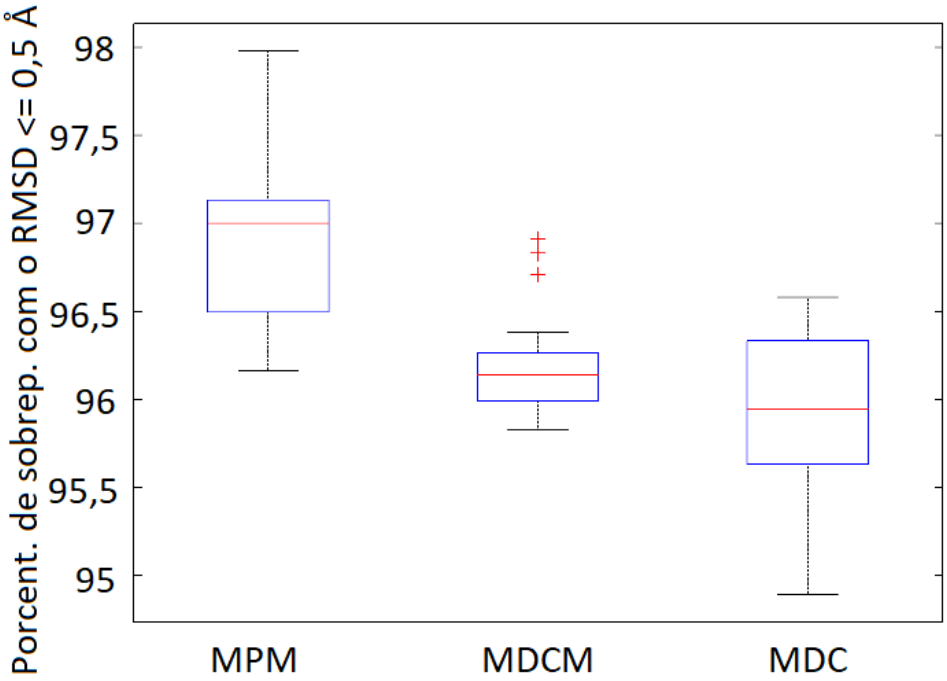
(b) Resultados conforme o maior deslocamento



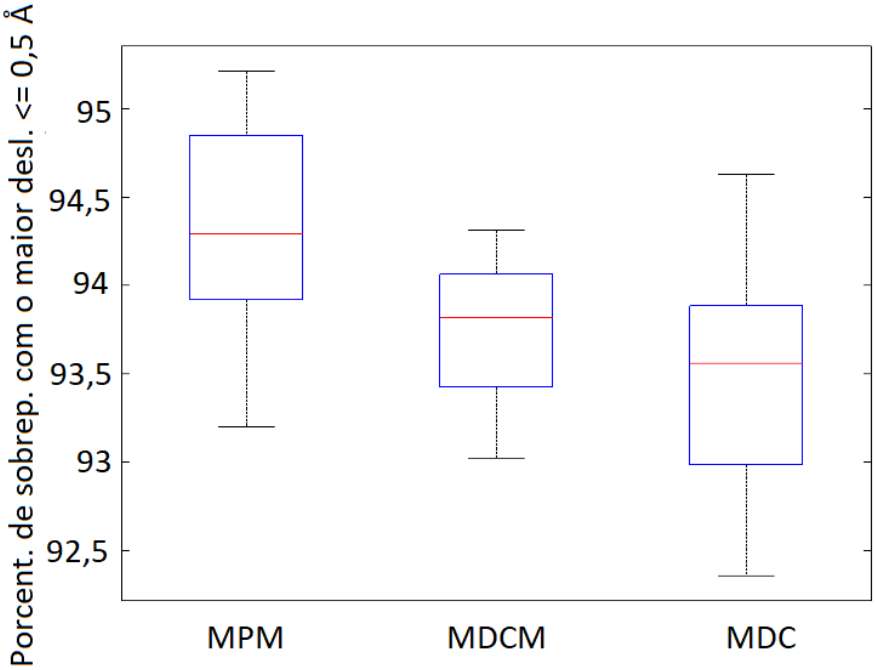
Fonte: Elaborada pelo autor (2023)

Figura 54 – *Boxplots* com as porcentagens de aproveitamentos totais da base de dados 2, utilizando o *k-means clustering*

(a) *Boxplots* considerando RMSD de até 0.5 Å



(b) *Boxplots* considerando o maior deslocamento de até 0.5 Å



Fonte: Elaborada pelo autor (2023)

demais metodologias comparadas. Ao combinar a análise deste *boxplot* com a Tabela 15, pode-se notar um comportamento similar ao da primeira base de dados. Ou seja, mesmo trabalhando com menos grupos, a MPM conseguiu ter as melhores acurácias deste experimento. Também constata-se que na avaliação do maior deslocamento, as precisões das metodologias ficaram mais próximas se comparadas com os resultados obtidos na avaliação do RMSD, em que a superioridade da MPM foi mais visível. Também foram realizados testes estatísticos através da tabela ANOVA e do teste de Tukey. A Figura 55 ilustra o cumprimento das premissas da ANOVA (normalidade, independência e homocedasticidade).

Conforme as Figuras 55 (a) e 55 (b) pode-se observar que a premissa de normalidade foi atendida tanto para a validação pelo RMSD quanto para a validação do maior deslocamento. Os valores da Figura 55 (b), referentes ao teste de normalidade dos resultados da avaliação do maior deslocamento ficaram muito coerentes com a linha tracejada na cor vermelha, indicando a normalidade dos dados. Na Figura 55 (a), embora alguns valores tenham ficado um pouco afastados da linha vermelha, pode-se entender que a premissa de normalidade dos dados também foi cumprida, pois a maior parte dos valores ficaram acima ou muito próximos desta linha tracejada e a ANOVA suporta estes pequenos desvios. Conforme as Figuras 55 (c) e 55 (d) é possível observar que a premissa de independência dos dados também foi atendida para ambas avaliações. Ou seja, os valores são imprevisíveis, de modo que a partir de um determinado valor obtido, não é possível determinar o próximo valor. As Figuras 55 (e) e 55 (f) ilustra que a premissa de homocedasticidade também foi atendida para ambos os casos, pois os resíduos ficaram bem distribuídos, sendo tanto maiores como menores do que 0.

Após verificar que todas as premissas foram atendidas, realizou-se o experimento da ANOVA para 1 fator. As tabelas da ANOVA para os resultados de RMSD e maior deslocamento são ilustradas respectivamente pelas Tabelas 16 e 17.

Tabela 16 – Tabela ANOVA para os experimentos da segunda base de dados, considerando o RMSD de até 0,5 Å

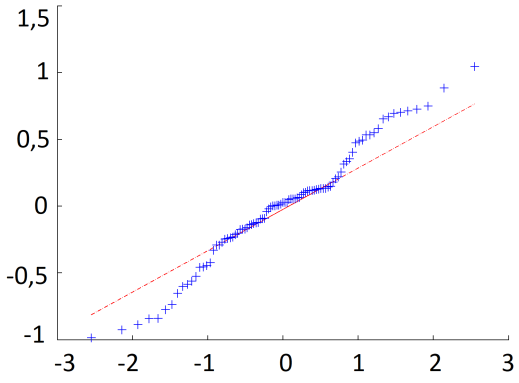
Fonte de var.	Soma dos quad.	Graus de lib.	Quad. méd.	F-est.	Prob>F
Colunas	18,5732	2	9,28658	49,92	2,59523e-15
Erro	16,7438	90	0,18604		
Total	35,317	92			

Fonte: Elaborada pelo autor (2023)

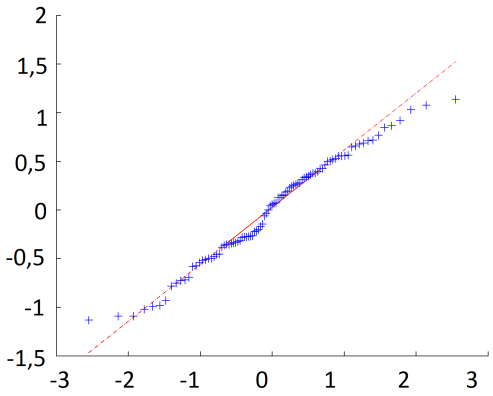
De acordo com as Tabelas 16 e 17, os valores p obtidos para o teste F (sexta coluna) de ambos os casos, foram menores do que 0,01, valor estabelecido para o nível de confiança de 99%. Deste modo, tanto para a avaliação do RMSD quanto para a avaliação do maior

Figura 55 – Premissas para utilizar a ANOVA considerando a porcentagem total de sobreposições com até 0,5 Å de RMSD e maior deslocamento, para a base de dados 2

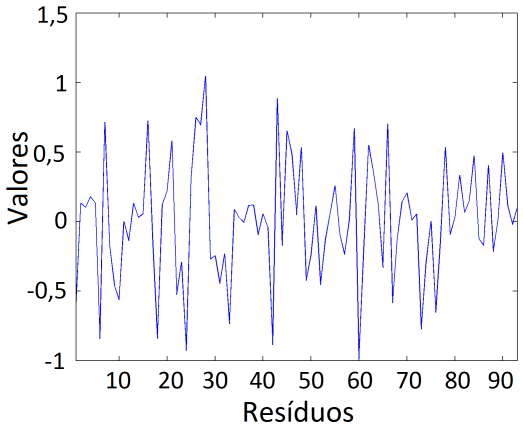
(a) Normalidade dos experimentos do RMSD



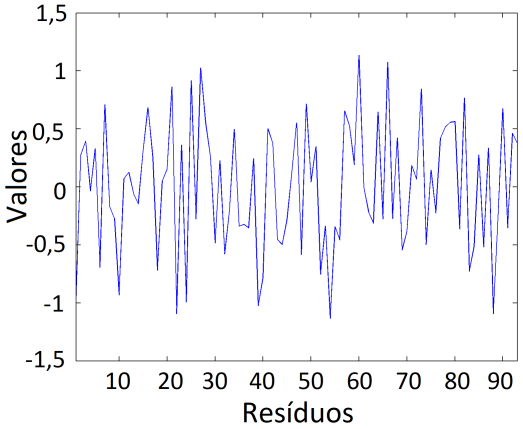
(b) Normalidade dos experimentos do maior deslocamento



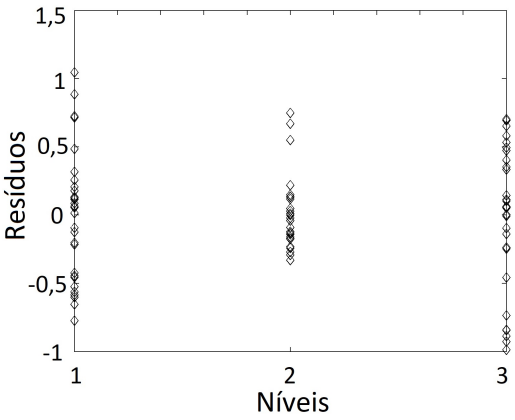
(c) Independência dos experimentos do RMSD



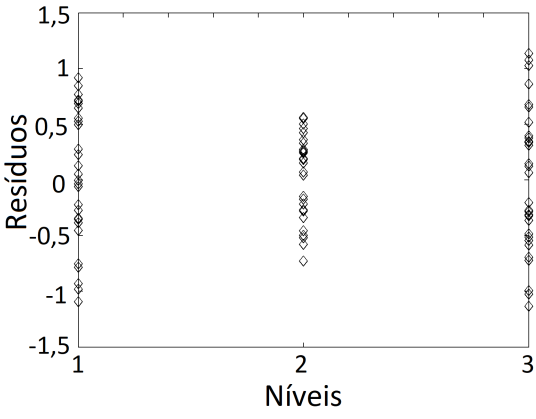
(d) Independência dos experimentos do maior deslocamento



(e) Homocedasticidade dos experimentos do RMSD



(f) Homocedasticidade dos experimentos do maior deslocamento



Fonte: Elaborada pelo autor (2023)

Tabela 17 – Tabela ANOVA para os experimentos da segunda base de dados, considerando o maior deslocamento de até 0,5 Å

Fonte de var.	Soma dos quad.	Graus de lib.	Quad. méd.	F-est.	Prob>F
Colunas	10,3613	2	5,18065	16,52	7,75399e-07
Erro	28,2276	90	0,31364		
Total	38,5889	92			

Fonte: Elaborada pelo autor (2023)

deslocamento, ao menos uma das técnicas avaliadas obteve resultados estatisticamente diferente das demais. A partir dos resultados da ANOVA pode-se realizar o teste de Tukey para verificar com maior precisão quais técnicas obtiveram resultados significativamente diferentes das outras técnicas, com o nível de confiança de 99%. Os resultados do teste de Tukey são ilustrados pela Figura 56.

Conforme a Figura 56 (a), ao avaliar a porcentagem total de sobreposições que obtiveram o RMSD com até 0,5 Å, pode-se notar através das linhas verticais próximas aos resultados da MDCM, que ocorreu um empate estatístico entre ela e a MDC. Através dessa mesma figura também é possível observar que os resultados da MPM superaram essas duas metodologias, com o nível de confinaça de 99%. A Figura 56 (b) indica um comportamento similar para a avaliação estatística quanto ao percentual de sobreposições realizadas, que resultaram no maior deslocamento de até 0,5 Å. Nesta figura pode-se notar que a MPM também ficou à frente das demais metodologias comparadas. Assim, também obtendo uma diferença estatística conforme o nível de confiança definido. Novamente, a MDCM e a MDC tiveram um empate dentro do nível de confiança definido.

6.2.1.1 Análise da Performance Computacional

A Tabela 18 armazena os tempos necessários para construir a base de dados e realizar os agrupamentos com o *k-means clustering* para a segunda base de dados. O computador utilizado para realizar estes experimentos foram obtidos através do *Google Cloud*, cujas configurações foram detalhadas na seção de metodologia.

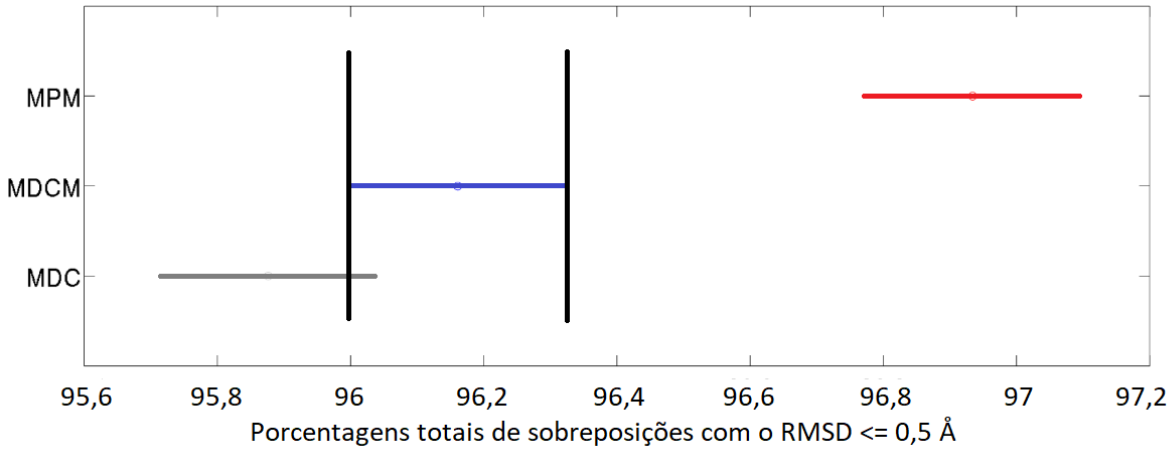
Tabela 18 – Tempo para realização dos experimentos com o *k-means clustering* para a base de dados 2

Técnica	Tempo const. dados	Tempo agrup.	Tempo total
MPM	02:11	43:31	45:42
MDC	01:15	47:43	48:58
MDCM	19:13	49:41	01:08:54

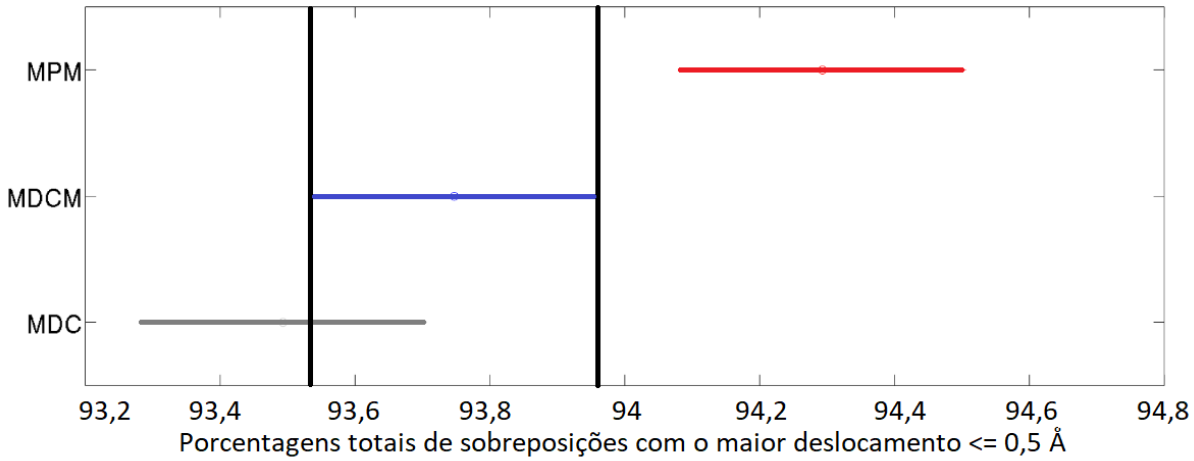
Fonte: Elaborada pelo autor (2023)

Figura 56 – Teste de Tukey com as porcentagens de aproveitamentos para a segunda base de dados

(a) Teste de Tukey considerando RMSD de até 0.5 Å



(b) Teste de Tukey considerando o maior deslocamento de até 0.5 Å



Fonte: Elaborada pelo autor (2023)

Conforme a Tabela 18, pode-se observar que a MDC gastou menos tempo para construir os dados a serem agrupados. Esse comportamento já era esperado pelos resultados obtidos nos experimentos da primeira base de dados, visto que a construção da matriz a ser agrupada é uma operação de tempo linear, influenciada pelo tamanho da base de dados. No entanto, a MPM foi a metodologia mais rápida na etapa de agrupamento e consequentemente foi a mais rápida no tempo total. Os dois fatores que mais contribuíram o performance da MPM durante o agrupamento foram: a quantidade de colunas a serem agrupadas, visto que dentre as metodologias testadas neste cenário ela é a que tem menos colunas. O outro fator refere-se à quantidade de grupos indicada como ideal pelo método *silhouette*, que foi menor do que as demais técnicas comparadas e assim, demandou de menos tempo.

6.2.2 OPTICS

Ao testar todas as técnicas de representações de proteínas com o OPTICS, para agrupar a segunda base de dados, verificou-se que todas as 8 técnicas atenderam ao critérios estabelecido para a realização dos experimentos. No entanto, todos os agrupamentos foram feitos apenas com a métrica de distância Euclidiana quadrática, devido ao fato desta métrica de distância ter possibilitado uma diminuição no número de interações descartadas pelo OPTICS ao trabalhar com a primeira base de dados. A Figura 57 ilustra as avaliações do RMSD e do maior deslocamento.

Conforme a Figura 57 pode-se observar que a aCSM apresentou o melhor resultado para a análise de até 0,1 Å, tanto para o RMSD, bem como para o maior deslocamento. No entanto, nas próximas faixas de somas acumuladas, outras metodologias conseguem obter resultados melhores do que a aCSM. Na faixa de soma acumulada de até 0,2 Å na avaliação do RMSD, algumas metodologias já superam a precisão da aCSM. Na avaliação do maior deslocamento, a aCSM passa a ser superada na faixa de soma acumulada de até 0,3 Å. Ao considerar a soma acumulada de até 0,5 Å, a MPM apresenta a maior precisão na análise do RMSD e a estratégia de busca da RID apresenta a maior acurácia na análise do maior deslocamento. No entanto, de um modo geral, tanto na avaliação do RMSD quanto no maior deslocamento, pode-se constatar que várias metodologias obtiveram resultados muito próximos. Apenas a MA apresentou resultados consideravelmente abaixo das demais técnicas comparadas. A Tabela 19 contém os principais resultados deste experimento.

Tabela 19 – Resultados do OPTICS com a métrica de distância Euclidiana quadrática para a base de dados 2

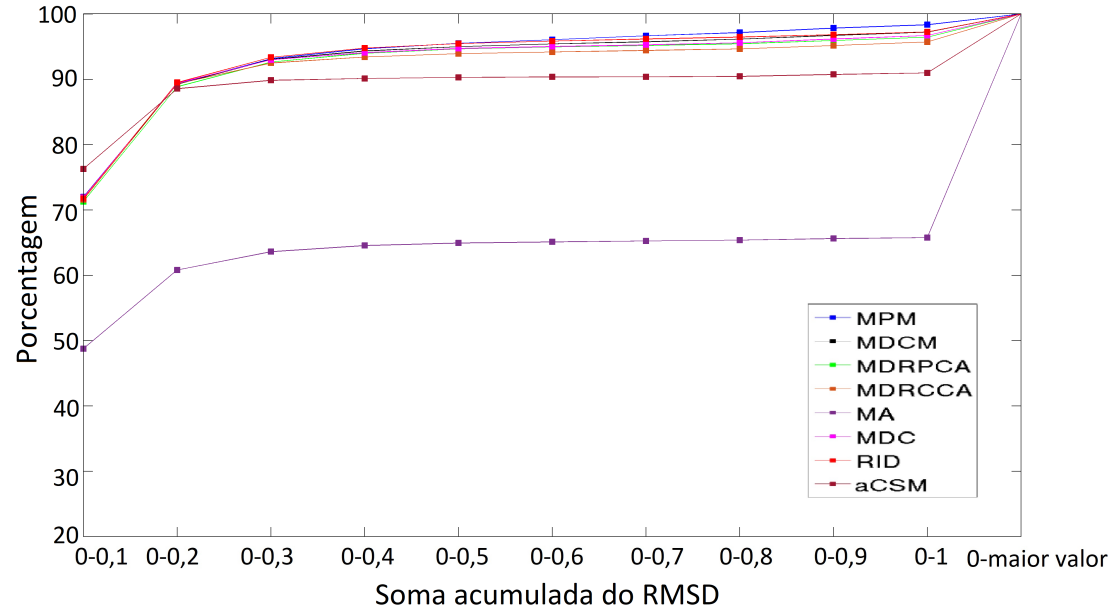
Técnica	RMSD <= 0,5Å	Maior D. <= 0,5Å	*Clusters*	Total Descart.	% Descart.
RID	95,47%	92,71%	35.516	29.450	22,71%
MPM	95,51%	92,58%	35.555	29.215	22,52%
MDCM	95,00%	92,33%	35.451	29.499	22,74%
MDC	94,70%	92,15%	35.595	28.962	22,33%
MDRPCA	94,64%	91,98 %	35.676	29.215	22,52%
MDRCCA	93,96%	91,52 %	36.040	27.825	21,45%
aCSM	90,28%	89,45 %	30.470	26.911	20,75%
MA	64,93%	64,13%	29.136	49.279	37,99%

Fonte: Elaborada pelo autor (2023)

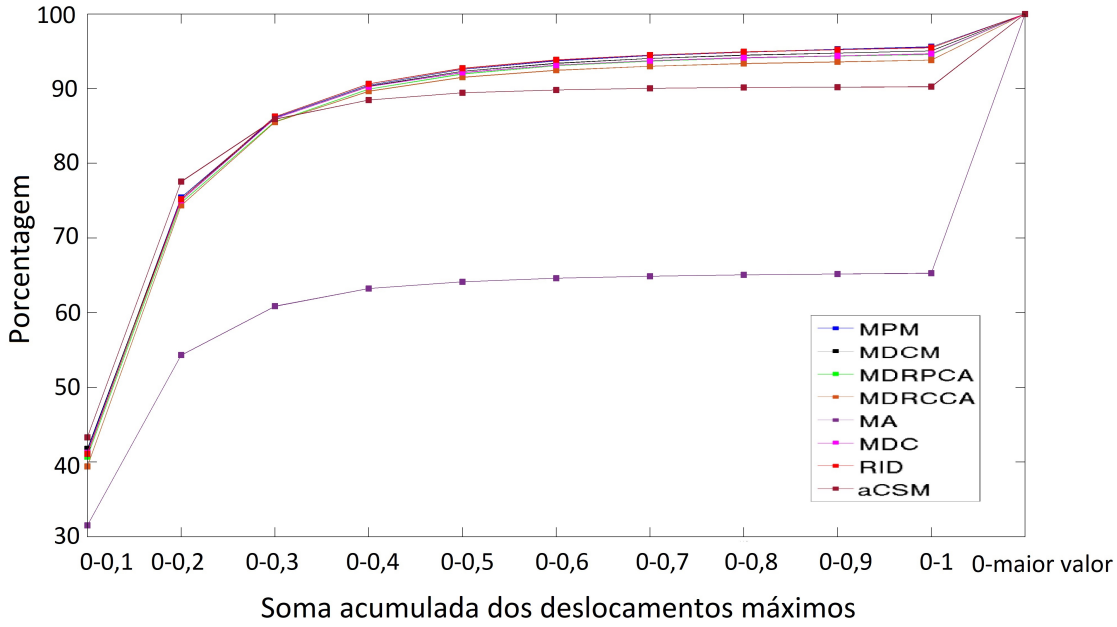
Conforme a Tabela 19, as acurácias de 6 metodologias ficaram com valores muitos próximos ao avaliar o RMSD e o maior deslocamento de até 0,5 Å. Essas metodologias foram: RID, MPM, MDCM, MDC, MDRPCA e MDRCCA. Apesar de ter começado com a melhor acurácia ao considerar o gráfico de soma acumulada de até 0,1 Å de RMSD e maior deslocamento, conforme visto através da Figura 57, a aCSM obteve o sétimo melhor

Figura 57 – Soma acumulada com os resultados das sobreposições da base de dados 2, através do OPTICS, com a métrica de distância Euclideana quadrática

(a) Resultados conforme o RMSD



(b) Resultados conforme o maior deslocamento



Fonte: Elaborada pelo autor (2023)

percentual de aproveitamento ao considerar o RMSD e o maior deslocamento de até 0,5 Å. A MA obteve resultados abaixo de todas as demais metodologias.

6.2.2.1 Análise da Performance Computacional

A Tabela 20 contém o tempo gasto para realizar os experimentos da base de dados 2 com o OPTICS. Ela está ordenada conforme a metodologia mais rápida, considerando o tempo total.

Tabela 20 – Tempo para realizar os experimentos com o OPTICS com a base de dados 2

Técnica	Tempo Total
MDRPCA	20:47
MDRCCA	23:19
RID	39:14
MA	44:21
MPM	01:35:23
MDC	01:56:45
MDCM	02:12:07
aCSM	05:08:19

Fonte: Elaborada pelo autor (2023)

Através da Tabela 20 pode-se observar que as matrizes de distâncias reduzidas (MDRPCA e MDRCCA) foram as metodologias mais rápidas. O método de busca da RID e a MA apresentaram demoras esperadas na etapa de transformar os dados para os formatos a serem agrupados, mas compensaram o tempo gasto na etapa de agrupamento. Em seguida, estão as técnicas mais robustas, que agrupam quantidades maiores de informações, por trabalharem com quantidades maiores de colunas (MPM, MDC MDCM e aCSM). É importante ressaltar que os parâmetros da aCSM foram alterados para a realização dos testes com esta base de dados, pois como as distâncias entre átomos das interações desta base são maiores do que as distâncias da primeira base de dados, foi necessário realizar essa alteração para obter os padrões de distâncias de um modo mais coerente. Assim, o intervalo de distância calculado foi entre 0 e 50 Å, com o passo de 0,01 Å, resultando em uma matriz de 500 colunas. O aumento do tamanho da matriz a ser agrupada, contribuiu para o aumento do tempo de processamento.

6.3 Resultados Obtidos com a Base de Dados 3

A base de dados 3 consiste na combinação das bases de dados 1 e 2, que totaliza 146.088 interações de pontes dissulfeto e de hidrogênio. Foram realizados experimentos com o *k-means clustering* na Subseção 6.3.1, e com o OPTICS na Subseção 6.3.2. Para esta

base de dados também não foi possível realizar experimentos com o *hierarchical clusering* devido ao alto consumo de memória RAM deste algoritmo.

6.3.1 *K-Means Clustering*

Para os experimentos do *k-means clustering* com a terceira base de dados, foram realizados testes somente com a MPM, MDCM e MDC. As Figuras 58 (a) e 58 (b) ilustram as somas acumuladas dos percentuais de sobreposições atômicas, conforme as distâncias de RMSD e maior deslocamento.

Conforme a Figura 58 (a) na avaliação do RMSD, a MDC obteve os melhores resultados até 0,1 Å. No entanto, a partir da soma acumulada de 0,2 Å, a MPM passou a obter a melhor acurácia e manteve os melhores resultados até o final deste gráfico. Também pode-se notar que a partir da soma acumulada de 0,2 Å, a MDCM também supera os resultados da MDC e apresenta a segunda melhor precisão até o final deste gráfico. Através da Figura 58 (b) observa-se que na avaliação do maior deslocamento, as três técncias obtiveram resultados muito parecidos. No entanto, constata-se que nas primeiras faixas de valores de maior deslocamento, a MDC apresentou uma pequena superioridade em relação as outras técnicas, mas a partir da soma acumulada de 0,4 Å, esta metodologia começa a ser superada pela MPM e pela MDCM. Ao chegar na soma acumulada de 0,5 Å, a MPM apresenta os melhores resultados, seguida pela MDCM e MDC respectivamente. A Tabela 21 ilustra os principais resultados desta análise.

Tabela 21 – Resultados do *k-means clustering* para a base de dados 3

Técnica	RMSD <= 0,5Å	Mai. D. <= 0,5Å	Mín. *cl.*	Med. *cl.*	Max.*Cl.*	*cl.* 1 reg.
MPM	97,17%	92,61%	19.000	21.366	22.800	1.614
MDCM	96,10%	92,06%	21.500	22.683	24.100	2.213
MDC	95,21%	91,84%	19.850	22.242	25.050	1.785

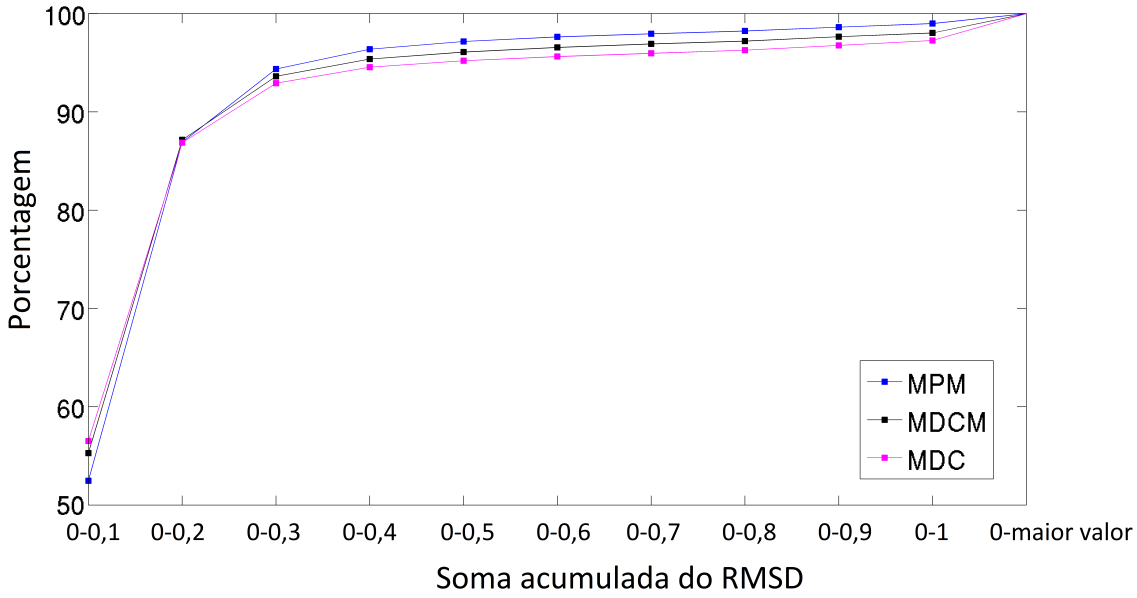
Fonte: Elaborada pelo autor (2023)

Conforme a Tabela 21, a MPM apresentou as melhores precisões nas análises de RMSD e maior deslocamento, além de utilizar menos *clusters* e ainda contribuiu para a formação de menos grupos com apenas uma interação. A MDCM apresentou acurácias superiores à MDC. No entanto, pode-se observar que a MDCM utilizou uma quantidade maior de grupos, o que facilita o trabalho do *k-means*, além de formar muitos grupos com apenas uma interação. A Figura 59 contém os *boxplots* considerando as acurácias totais de cada uma das 31 execuções de experimentos, considerando o RMSD e o maior deslocamento de até 0,5 Å.

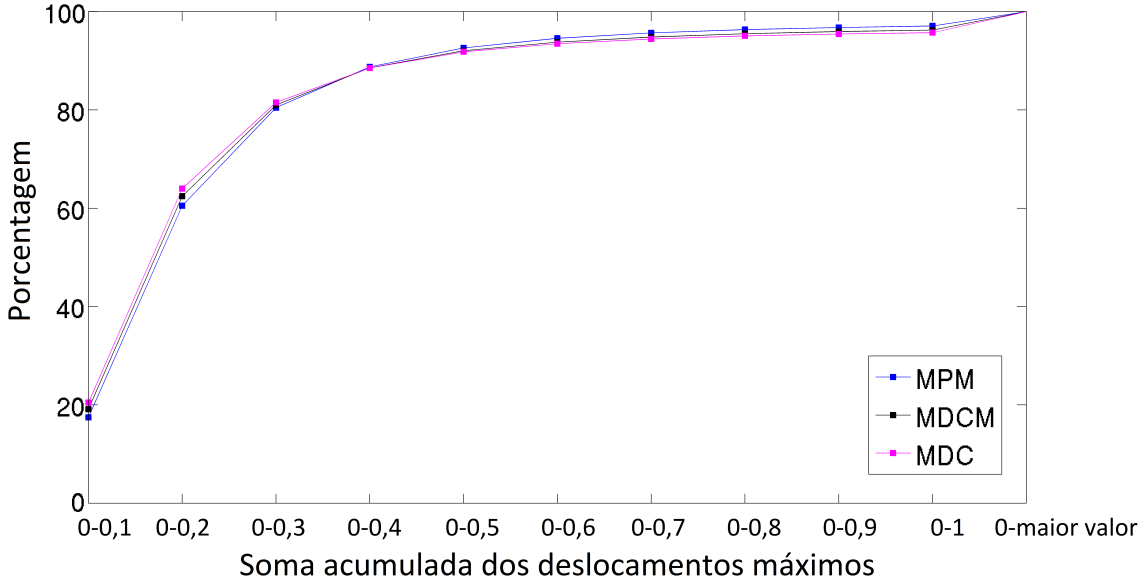
Conforme as Figuras 59 (a) e 59 (b), os resultados da MPM foram melhores do que os resultados das demais metodologias comparadas, principalmente na avaliação do RMSD.

Figura 58 – Soma acumulada dos resultados da base de dados 3, com o *k-means clustering*

(a) Resultados conforme o RMSD



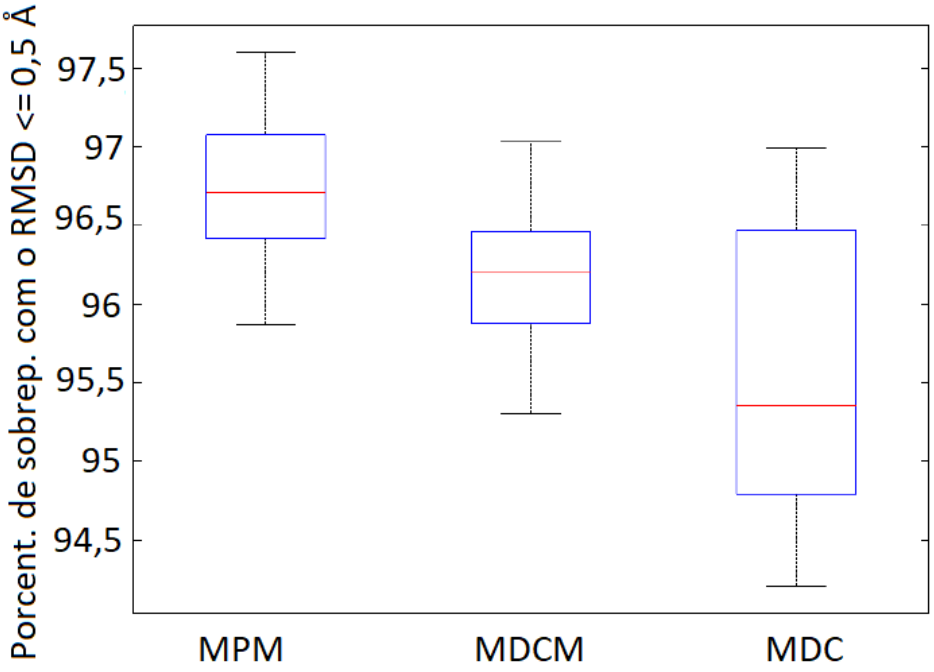
(b) Resultados conforme o maior deslocamento



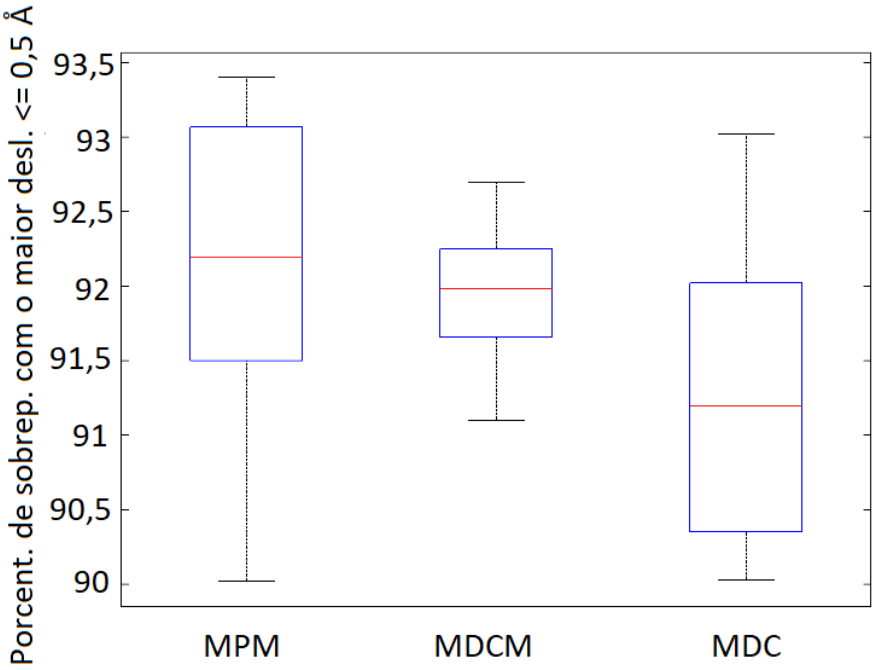
Fonte: Elaborada pelo autor (2023)

Figura 59 – *Boxplots* com as porcentagens de aproveitamentos totais da base de dados 3, utilizando o *k-means clustering*

(a) *Boxplots* considerando RMSD de até 0.5 Å



(b) *Boxplots* considerando o maior deslocamento de até 0.5 Å



Fonte: Elaborada pelo autor (2023)

Ao analisar esses *boxplots* juntamente com a Tabela 21 pode-se constatar que novamente, a MPM conseguiu obter as melhores execuções de experimentos, mesmo trabalhando com uma quantidade menor de grupos. Mas é importante ressaltar que a MDCM e a MDC também apresentaram resultados interessantes.

Em seguida realizou-se testes estatísticos através da tabela ANOVA e do teste de Tukey, como nas bases de dados anteriores. A Figura 60 ilustra o cumprimento de todas as premissas da ANOVA: normalidade, independência e homocedasticidade.

Conforme as Figuras 60 (a) e 60 (b) é possível constatar que a premissa de normalidade foi atendida para as análises de RMSD e também na avaliação do maior deslocamento. Os valores, representados por pontos na cor azul, ficaram bem alinhados com a reta na cor vermelha. Conforme as Figuras 60 (c) e 60 (d) é possível observar que a premissa de independência dos dados também foi atendida para ambas avaliações, indicando que os valores são imprevisíveis, não sendo possível determinar o próximo valor a partir do valor atual. As Figuras 60 (e) e 60 (f) ilustram que a premissa de homocedasticidade também foi atendida para ambas as avaliações, de modo que os resíduos ficaram bem distribuídos, sendo maiores, bem como menores do que 0.

Após verificar que todas as premissas foram atendidas, realizou-se o experimento da ANOVA para 1 fator. As tabelas da ANOVA para os resultados de RMSD e maior deslocamento são ilustradas respectivamente pelas Tabelas 22 e 23.

Tabela 22 – Tabela ANOVA para os experimentos da terceira base de dados, considerando o RMSD de até 0,5 Å

Fonte de var.	Soma dos quad.	Graus de lib.	Quad. méd.	F-est.	Prob>F
Colunas	22,1129	2	11,0565	27,19	5,79e-10
Erro	36,5952	90	0,4066		
Total	58,7082	92			

Fonte: Elaborada pelo autor (2023)

Tabela 23 – Tabela ANOVA para os experimentos da terceira base de dados, considerando o maior deslocamento de até 0,5 Å

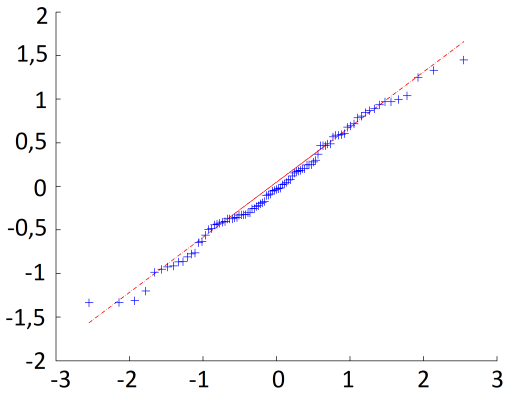
Fonte de var.	Soma dos quad.	Graus de lib.	Quad. méd.	F-est.	Prob>F
Colunas	9,8701	2	4,93503	6,67	0,002
Erro	66,5901	90	0,73989		
Total	76,4601	92			

Fonte: Elaborada pelo autor (2023)

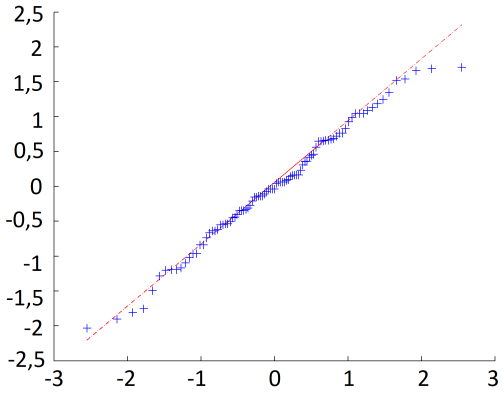
De acordo com as Tabelas 22 e 23, os valores p obtidos para o teste F (sexta coluna) de ambas avaliações, foram menores do que 0,01, valor estabelecido para o nível de confiança

Figura 60 – Premissas para utilizar a ANOVA considerando a porcentagem total de sobreposições com até 0,5 Å de RMSD e maior deslocamento, para a base de dados 3

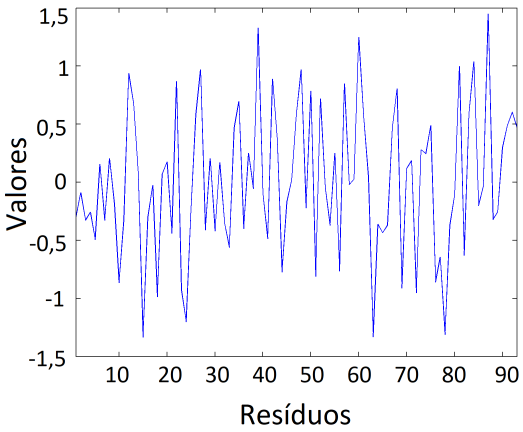
(a) Normalidade dos experimentos do RMSD



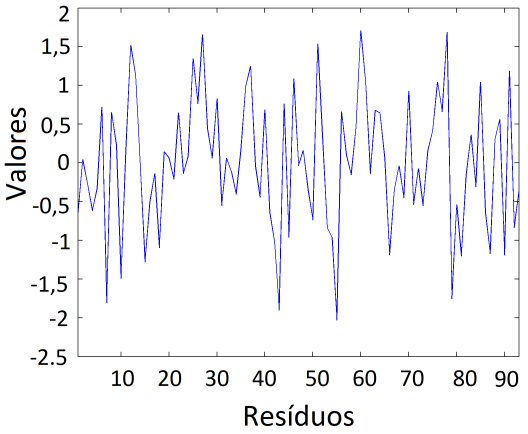
(b) Normalidade dos experimentos do maior deslocamento



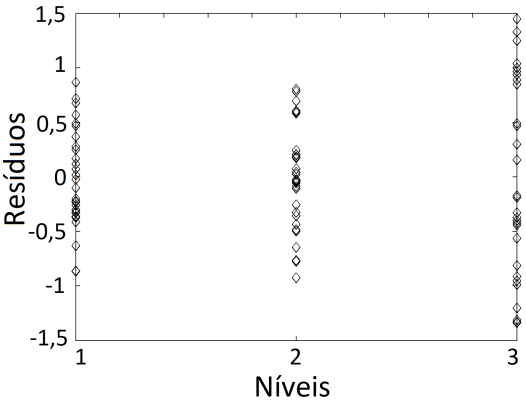
(c) Independência dos experimentos do RMSD



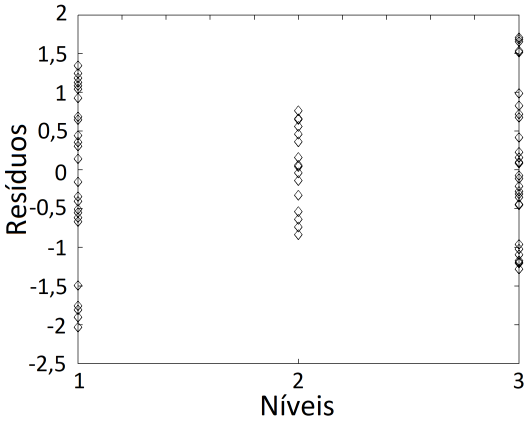
(d) Independência dos experimentos do maior deslocamento



(e) Homocedasticidade dos experimentos do RMSD



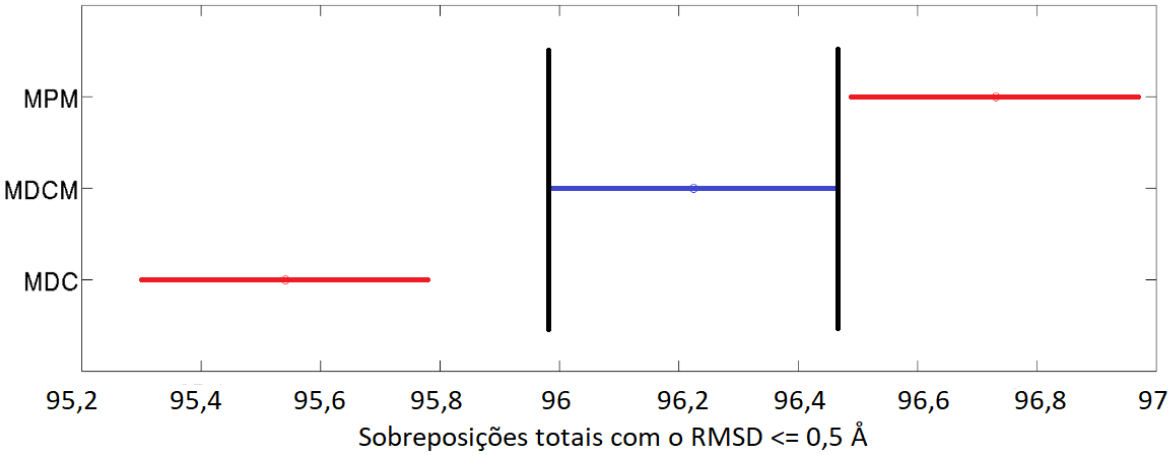
(f) Homocedasticidade dos experimentos do maior deslocamento



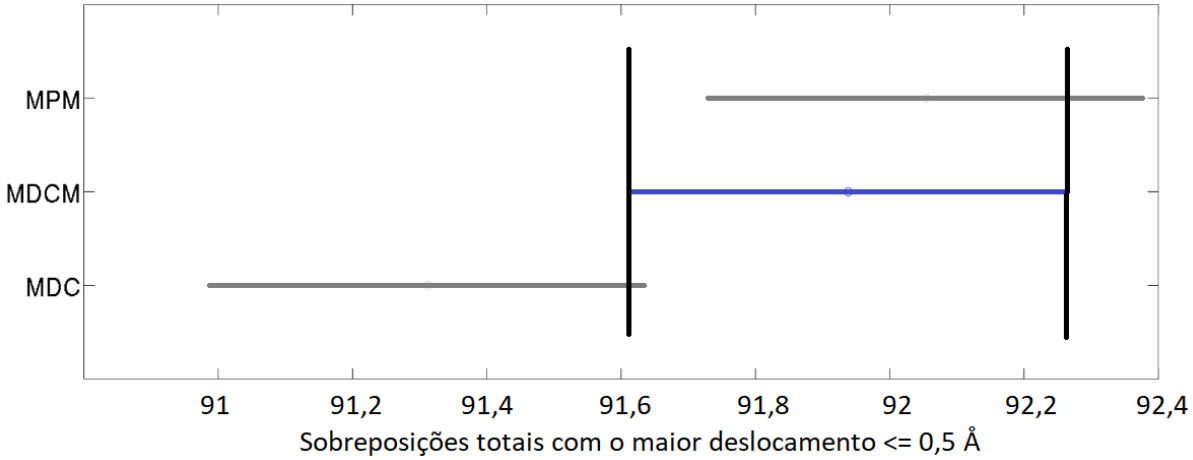
Fonte: Elaborada pelo autor (2023)

Figura 61 – Teste de Tukey com as porcentagens de aproveitamentos para a base de dados 3

(a) Teste de Tukey considerando RMSD de até 0,5 Å



(b) Teste de Tukey considerando o maior deslocamento de até 0,5 Å



Fonte: Elaborada pelo autor (2023)

de 99%. É possível observar que na avaliação do maior deslocamento o valor do teste F ficou mais próximo da tolerância do que o resultado na avaliação do RMSD, indicando que para os valores de maior deslocamento, as diferenças de acurácias entre as técnicas avaliadas, foram menores do que na avaliação do RMSD, seguindo os mesmos resultados obtidos através das outras análises. No entanto, tanto para a avaliação do RMSD quanto para a avaliação do maior deslocamento, ao menos uma das técnicas avaliadas possuiu resultados estatisticamente diferente das demais. A partir dos resultados da ANOVA pode-se realizar o teste de Tukey para verificar com maior precisão qual técnica (ou quais técnicas) obtiveram resultados significativamente diferentes das demais, com o nível de confiança de 99%. Os resultados do teste de Tukey são ilustrados pela Figura 61.

Conforme a Figura 61 (a), ao avaliar a porcentagem total de sobreposições que obtiveram o

RMSD com até 0,5 Å, pode-se constatar através das linhas verticais próximas aos resultados da MDCM, que a MPM superou estatisticamente, com o nível de confiança de 99%, as outras duas técnicas avaliadas. Através dessa figura também é possível observar que a MDCM superou estatisticamente a MDC. A Figura 61 (b) indica através das linhas verticais que ocorreu um empate estatístico entre a MPM e a MDCM, com uma leve vantagem da MPM. Também pode-se observar a ocorrência de um empate para o nível de confiança definido entre a MDC e a MDCM. A única diferença siginificativa na avaliação do maior deslocamento foi entre a MPM e a MDC, no qual a Matriz de Pontos Médios superou a MDC.

6.3.1.1 Análise da Performance Computacional

A Tabela 24 ilustra o tempo necessário para realizar os experimentos combinando as 3 metodologias de representações de proteínas juntamente com o *k-means clustering*.

Tabela 24 – Tempo para realização dos experimentos com o *k-means clustering* para a base de dados 3

Técnica	Tempo const. dados	Tempo agrup.	Tempo total
MPM	02:24	45:53	48:17
MDC	01:26	48:15	49:41
MDCM	21:01	01:01:47	01:22:48

Fonte: Elaborada pelo autor (2023)

Conforme pode ser observado pela Tabela 24, para a terceira base de dados, novamente a MPM foi a metodologia mais rápida. O seu desempenho é explicado principalmente por dois motivos: menos colunas de informações a serem agrupadas e quantidade ideal de grupos menor do que as demais técnicas comparadas.

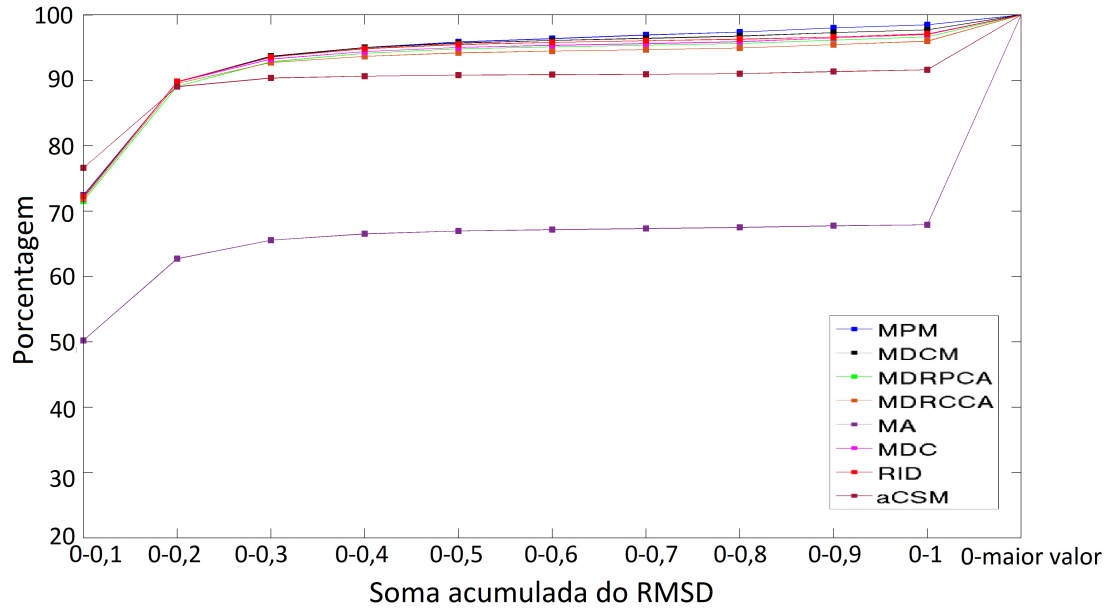
6.3.2 OPTICS

Do mesmo modo que nas bases de dados 1 e 2, todas as 8 técnicas de representações de proteínas participaram dos experimentos, por atenderem ao critérios estabelecido. A Figura 62 ilustra a soma acumulada dos valores de RMSD e maior deslocamento obtidos através das sobreposições realizadas.

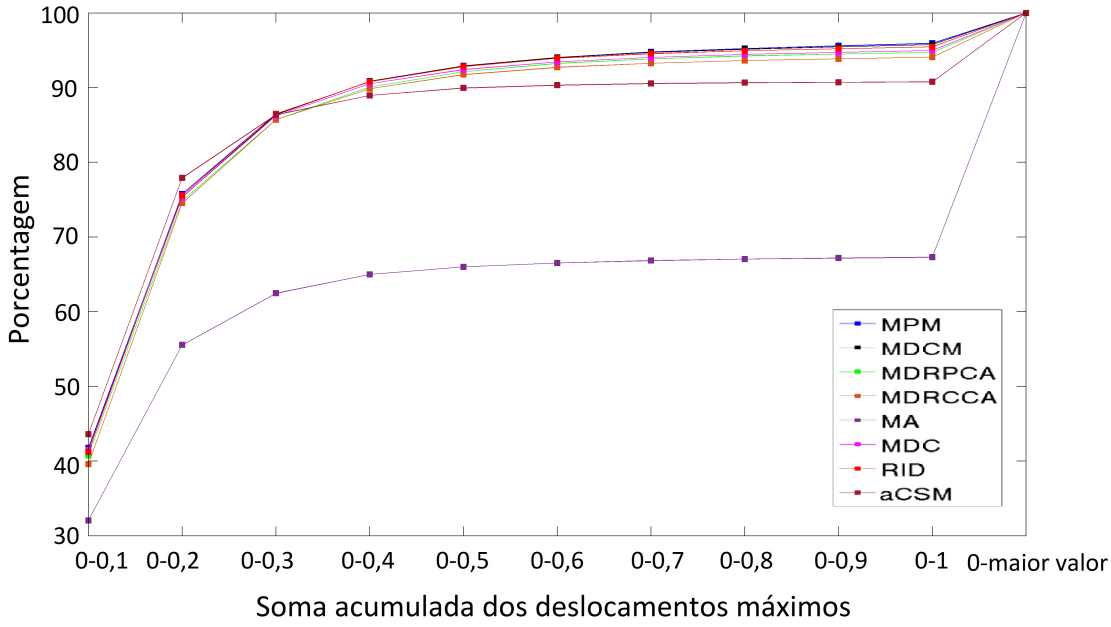
Conforme pode ser observado pela Figura 62, estas avaliações obtiveram comportamentos similares aos resultados da base de dados 2. A aCSM obteve os melhores resultados para a soma acumulada de até 0,1 Å. No entanto, nas próximas faixas de somas acumuladas, nota-se que outras metodologias ultrapassam a aCSM. Novamente, 6 técnicas continuaram com os melhores valores de acurácias, os quais ficaram bem próximos (MPM, MDCM, RID,

Figura 62 – Soma acumulada com os resultados das sobreposições da base de dados 3, através do OPTICS, com a métrica de distância Euclideana quadrática

(a) Resultados conforme o RMSD



(b) Resultados conforme o maior deslocamento



Fonte: Elaborada pelo autor (2023)

MDC, MDRPCA e MDRCCA). No entanto, a ordem das melhores precisões tiveram algumas alterações. A Tabela 25 ilustra os principais resultados obtidos nesse experimento.

Tabela 25 – Resultados do OPTICS com a métrica de distância Euclidiana quadrática para a base de dados 3

Técnica	RMSD <= 0,5Å	Maior D. <= 0,5Å	*Clusters*	Total Descart.	% Descart.
MDCM	95,70%	92,92%	37.705	40.006	27,38%
MPM	95,86%	92,91%	39.246	34.881	23,88%
RID	95,48%	92,83%	39.545	34.468	23,59%
MDC	95,03%	92,42%	39.257	34.901	23,89%
MDRPCA	94,81%	92,13 %	39.424	35.011	23,97%
MDRCCA	94,20%	91,74 %	39.746	33.759	23,11%
aCSM	90,80%	89,94 %	33.777	50.172	34,34%
MA	66,94%	66,01%	32.531	56.170	38,45%

Fonte: Elaborada pelo autor (2023)

Conforme a Tabela 25, a MDCM foi a metodologia que obteve a melhor precisão, somando a acurácia da avaliação do RMSD com a acurácia obtida na verificação do maior deslocamento de até 0,5 Å. No entanto, pode-se observar que essa metodologia foi a terceira que mais descartou interações durante o agrupamento. A MPM obteve a segunda melhor precisão somando os resultados do RMSD e do maior deslocamento. Mas se for considerar somente o resultado do RMSD de até 0,5 Å, ela apresentou a melhor acurácia, inclusive descartando uma quantidade menor de registros do que as demais técnicas. De um modo geral, nos vários experimentos realizados com o OPTICS, ao considerar o RMSD e o maior deslocamento de 0,5 Å, pode-se observar que sempre a MDCM, MPM, RID, MDC, MDRPCA e MDRCCA obtêm resultados muito próximos, ocorrendo pequenas variações de um experimento para outro. Muitas vezes, alguns fatores influenciam na determinação de qual dessas técnicas obterá o melhor resultado do cenário em questão. Neste caso, por exemplo, o total de interações descartadas pela MDCM influenciou para que ela superasse as demais metodologias. No entanto, é importante ressaltar que a quantidade de descartes da MDCM foi menor do que as realizadas pela MA e pela aCSM, que inclusive obtiveram precisões inferiores do que as outras 6 técnicas comparadas.

6.3.2.1 Análise da Performance Computacional

A Tabela 26 ilustra o tempo gasto para realizar os experimentos da base de dados 3 com o OPTICS.

Conforme a Tabela 26, foi possível observar que a ordem das metodologias mais rápidas foi a mesma da base de dados 2. No entanto, como esta base de dados é maior, o tempo das metodologias também foram maiores.

Tabela 26 – Tempo para realizar os experimentos com o OPTICS para a base de dados 3

Técnica	Tempo Total
MDRPCA	29:35
MDRCCA	32:26
RID	46:31
MA	52:18
MPM	01:45:33
MDC	02:06:17
MDCM	02:26:29
aCSM	06:41:37

Fonte: Elaborada pelo autor (2023)

Capítulo 7

Conclusão

Neste trabalho foram desenvolvidas diferentes técnicas de representações de interações de proteínas, que foram combinadas com diferentes técnicas de *clustering*. Dentre as metodologias de representações de proteínas, pode-se destacar os resultados alcançados pela Matriz de Pontos Médios (MPM). Essa matriz de distâncias apresentou uma boa combinação com os diferentes algoritmos de *clustering* testados, obtendo uma maior acurácia na maioria dos cenários testados, inclusive obtendo vantagem estatisticamente significativa com o nível de confiança de 99%. Os bons resultados da MPM devem-se ao fato da sua construção ter sido planejada levando em consideração os átomos de carbono alfa, que são os principais átomos das interações, e ainda por ter levado em consideração as distâncias entre todos os demais átomos, pois essas informações são importantes para determinar o enovelamento da proteína.

Sobre as demais técnicas de representações de proteínas, pode-se concluir que a Matriz de Distâncias Completa (MDC) e a Matriz de Distâncias Completa Mista (MDCM) também obtiveram resultados interessantes ao serem combinadas com as diferentes técnicas de agrupamentos. A Matriz de Distâncias Reduzida a parir de um Ponto entre os Carbonos Alfas (MDRPCA) apresentou uma boa combinação com o *hierarchical clustering*, principalmente com os métodos de ligação *complete* e *average*. Outra vantagem da MDRPCA deve-se ao seu desempenho computacional, sendo a técnica de representação de proteínas mais rápida na etapa de construção dos dados a serem agrupados. A Matriz de Distâncias Reduzida cujos Carbonos Alfas são Centroides (MDRCCA), *Residue Interaction Database* (RID) e *atomic Cutoff Scanning Matrix* (aCSM) obtiveram precisões satisfatórias ao serem combinadas com o OPTICS. No entanto a Matriz de Ângulos (MA) obteve acurácia inferior nessa combinação com o OPTICS.

Ao utilizar o OPTICS juntamente com todas as técnicas de representações de proteínas pode-se concluir que esse algoritmo de *clustering* contribuiu para a obtenção de resultados com acurácias satisfatórias, além de demandar de poucos recursos computacionais. Os

experimentos do OPTICS com as maiores bases de dados foram realizados em computadores disponibilizados pelo Google Colab. No entanto, também verificou-se que seria possível realizar estes experimentos com o OPTICS em um notebook de 8 GB de memória RAM e processador *intel core i5 7 th Gen*. O ponto negativo do uso do OPTICS deve-se à quantidade de descartes realizados durante o agrupamento. Através dos testes com a primeira base de dados verificou-se que a métrica de distância Euclidiana quadrática possibilitou para essa base de dados um percentual de descartes levemente menor, mas no entanto, ainda ocorreram muitas exclusões durante o *clustering*. Ao trabalhar com a segunda base de dados de 129.705 interações, tamanho quase 8 vezes maior do que a primeira, notou-se que o percentual de descartes diminuiu consideravelmente. Deste modo, trabalhou-se com a hipótese do percentual de interações descartadas diminuir na media em que a base de dados era aumentada. Entretanto, ao trabalhar com o OPTICS na terceira base de dados, de 146.088 interações, obteve-se mais registros descartados do que ao trabalhar com a segunda base de dados.

Outro algoritmo de agrupamento testado juntamente com as técnicas de representações de interações proteicas foi o *hierarchical clustering*. Nos testes para a primeira base de dados foram obtidos resultados interessantes principalmente ao utilizar o método de ligação Ward, no qual obteve-se acurácias satisfatórias para todas as técnicas de representações de proteínas, além de ocorrerem poucos casos de interações ficarem sozinhas no mesmo grupo. Os métodos de ligações *complete* e *average* também apresentaram resultados interessantes, principalmente ao trabalhar com a MDRPCA. No entanto, esses métodos de ligações formaram muitos grupos com apenas uma interação. O método de ligação do *hierarchical clustering* utilizado neste trabalho, que obteve menos compatibilidade com as técnicas de representações de proteínas foi o *single*. Este método de ligação foi incompatível com a maioria das técnicas de representações de proteínas, sendo compatível apenas com a MPM e a RID. Entretanto, as acurácias obtidas não foram satisfatóiras, além de ocorrerem muitos grupos com apenas uma interação. O principal ponto negativo do *hierarchical clustering* verificado neste trabalho foi o alto consumo de memória RAM, que impossibilitou testá-lo com as bases de dados maiores, algo que foi possivel de ser realizado com o OPTICS e com o *k-Means clustering*.

O *k-means clustering* apresentou boas acurácias ao ser utilizado em conjunto com as técnicas de representações de proteínas. As vantagens desse algoritmo devem-se principalmente ao seu consumo de memória RAM ser relativamente baixo, sendo possível trabalhar com as bases de dados maiores em um computador de 16 GB de memória RAM (não foi verificado se foi utilizado algum recurso SSD para complementar o processamento). A outra vantagem deve-se ao fato do *k-means clustering* construir poucos grupos com apenas 1 registro. A principal desvantagem do *k-means clustering* deve-se aos seus resultados serem influenciados por sementes. Ou seja, algumas sementes podem resultar em pequenas

quantidades de agrupamentos tidas como ideais e consequentemente interferir na qualidade do agrupamento. Entretanto, os pontos positivos deste algoritmo de agrupamento superam os pontos negativos e pode-se afirmar que dentre os algoritmos testados para essas bases de dados o *k-means clustering* mostrou-se o mais adequado.

7.1 Trabalhos Futuros

Como continuidade deste trabalho, pode-se realizar alguns trabalhos futuros, que são descritos nas próximas subseções.

7.1.1 Incorporar a MPM e o *K-Means Clustering* na RID

Esta possibilidade de trabalho futuro consiste em aperfeiçoar a ferramenta *Residue Interaction Database* (RID), substituindo o seu método de busca original pela MPM juntamente com o *k-means clustering* (ou outras técnicas utilizadas neste trabalho), visando o aperfeiçoamento da busca de pares interagentes. Também é uma possibilidade disponibilizar essa nova versão da RID em um servidor web, para que divesos usuários possam utilizá-la.

7.1.2 Reduzir Algumas Linhas da Matriz Antes de Utilizar o *Hierarchical Clustering*

O *hierarchical clustering* apresentou resultados interessantes ao trabalhar com a primeira base de dados. No entanto, o seu alto consumo de memória RAM impossibilitou o seu uso com as bases de dados maiores. Esta proposta de trabalho futuro consiste em fazer um pré-processamento na matriz a ser agrupada, de modo a definir antes de utilizar o *hierarchical clustering*, interações que muito provavelmente estarão no mesmo *cluster*. Deste modo, será possível selecionar somente uma dessas interações similares para o *hierarchical clustering*, com o objetivo de reduzir o tamanho da matriz a ser agrupada e consequentemente diminuir o consumo de memória RAM. Após o agrupamento, as demais interações poderão ser inseridas no mesmo *cluster* da interação que foi escolhida para representá-las.

Referências

AL-JARF, R. et al. pdcsm-cancer: using graph-based signatures to identify small molecules with anticancer properties. **Journal of Chemical Information and Modeling**, ACS Publications, v. 61, n. 7, p. 3314–3322, 2021. Citado na página 17.

ALI, M. **Implementing K-Means Clustering with K-Means++ Initialization in Python.** Geek Culture, 2021. Disponível em: <https://abrir.link/l1l2l.> Acesso em: 03 de novembro de 2022. Citado 2 vezes nas páginas 31 e 32.

ALMOG, O. et al. Structural basis of thermostability analysis of stabilizing mutations in subtilisin bpn. **Journal of Biological Chemistry**, ASBMB, v. 277, n. 30, p. 27553–27558, 2002. Citado na página 6.

AL-KARADAGHI, S. **Torsion Angles and the Ramachandran Plot**. Protein Structures, 2022. Disponível em: <https://proteinstructures.com/structure/ramachandran-plot/>. Acesso em: 01 de novembro de 2022. Citado na página 24.

AMORIM, H. C. **Busca de Similaridade de Estruturas usando o Modelo de Mistura de Gaussianas**. 2020. Citado 4 vezes nas páginas 18, 44, 45 e 59.

ANKERST, M. et al. Optics: Ordering points to identify the clustering structure. **ACM Sigmod record**, ACM New York, NY, USA, v. 28, n. 2, p. 49–60, 1999. Citado 3 vezes nas páginas 33, 36 e 37.

ANTCZAK, M. et al. Structural alignment of protein descriptors–a combinatorial model. **BMC bioinformatics**, Springer, v. 17, n. 1, p. 1–14, 2016. Citado na página 26.

ARTHUR, D.; VASSILVITSKII, S. How slow is the k-means method? In: **Proceedings of the twenty-second annual symposium on Computational geometry**. [S.l.: s.n.], 2006. p. 144–153. Citado na página 33.

BARROSO, J. R. M. et al. Proteus: An algorithm for proposing stabilizing mutation pairs based on interactions observed in known protein 3d structures. **BMC bioinformatics**, BioMed Central, v. 21, n. 1, p. 1–21, 2020. Citado 4 vezes nas páginas 6, 16, 17 e 29.

BASHIROVA, A. et al. Disulfide bond engineering of an endoglucanase from penicillium verruculosum to improve its thermostability. **International Journal of Molecular Sciences**, v. 20, n. 7, 2019. ISSN 1422-0067. Disponível em: <https://www.mdpi.com/1422-0067/20/7/1602>. Citado na página 10.

BASTIDAS, O. H. Time dependent dihedral angle oscillations of the spike protein of sars-cov-2 reveal favored frequencies of dihedral angle rotations. **bioRxiv**, Cold Spring Harbor Laboratory, p. 2023–01, 2023. Citado na página 19.

BERMAN, H. M. et al. The protein data bank. **Nucleic acids research**, Oxford University Press, v. 28, n. 1, p. 235–242, 2000. Citado na página 1.

BIOPYTHON. **Bio.PDB.vectors module**. 2023. Disponível em: <https://biopython.org/docs/dev/api/Bio.PDB.vectors.html>. Acesso em: 23 de julho de 2020. Citado na página 51.

BOWIE, J. U. Membrane protein folding: how important are hydrogen bonds? **Current Opinion in Structural Biology**, v. 21, n. 1, p. 42–49, 2011. ISSN 0959-440X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0959440X10001661>. Citado na página 6.

CAMPELO, F. **Design and Analysis of Experiments 06 Simple Comparisons**. 2015. Disponível em: <https://github.com/fcampelo/Design-and-Analysis-of-Experiments/blob/master/06-SimpleComparisons/Chapter06.pdf>. Acesso em: 10 de abril de 2021. Citado na página 66.

CAMPELO, J. A. F. G. **Um estudo sobre o uso da geometria poligonal da cadeia principal como uma assinatura estrutural mais conservada que o empacotamento de resíduos**. Tese (Doutorado) — Universidade Federal de Minas Gerais - UFMG, 2017. Disponível em: <https://repositorio.ufmg.br/handle/1843/BUOS-B5MKJ3>. Acesso em: 22 de fevereiro de 2021. Citado 3 vezes nas páginas 11, 17 e 19.

CARPENTIER, M.; CHOMILIER, J. Protein multiple alignments: sequence-based versus structure-based programs. **Bioinformatics**, Oxford University Press, v. 35, n. 20, p. 3970–3980, 2019. Citado na página 8.

CCP4. **LSQKAB (CCP4: Supported Program)**. 2011. Disponível em: <http://www.ccp4.ac.uk/html/lsqkab.html>. Acesso em: 23 de abril de 2020. Citado 2 vezes nas páginas 2 e 11.

CHAKRABARTY, B.; PAREKH, N. Sequence and structure-based analyses of human ankyrin repeats. **Molecules**, MDPI, v. 27, n. 2, p. 423, 2022. Citado na página 9.

CRESCENZI, P. et al. On the complexity of protein folding. In: **Proceedings of the thirtieth annual ACM symposium on Theory of computing**. [S.l.: s.n.], 1998. p. 597–603. Citado 3 vezes nas páginas 2, 5 e 11.

DANIELS, N. M.; NADIMPALLI, S.; COWEN, L. J. Formatt: Correcting protein multiple structural alignments by incorporating sequence alignment. **BMC bioinformatics**, BioMed Central, v. 13, n. 1, p. 1–8, 2012. Citado na página 9.

DAVID. **Alternative to Colab Pro: Comparing Google's Jupyter Notebooks to Gradient Notebooks (Updated!)**. Paperspace blog, 2022. Disponível em: <https://blog.paperspace.com/alternative-to-google-colab-pro/>. Acesso em: 04 de outubro de 2022. Citado na página 58.

DIAS, S. R. **Residue interaction database: proposição de mutações sítio dirigidas com base em interações observadas em proteínas de estrutura tridimensional conhecida**. Tese (Doutorado) — Universidade Federal de Minas Gerais - UFMG, 2012. Disponível em: <https://repositorio.ufmg.br/bitstream/1843/BUOS-9GQKBD/1/tese_sandrord_vfinal9_ok.pdf>. Acesso em: 14 de abril de 2021. Citado 14 vezes nas páginas 1, 2, 6, 12, 13, 14, 15, 16, 23, 24, 25, 27, 28 e 59.

DIAS, U. M. **Predição da Função das Proteínas Sem Alinhamentos Usando Máquinas de Vetor de Suporte**. Dissertação (Mestrado) — Universidade Federal de Alagoas, 2007. Disponível em: <http://www.repositorio.ufal.br/bitstream/riufal/808/1/Dissertacao_UlissesMartinsDias_2007.pdf>. Acesso em: 06 de fevereiro de 2022. Citado 2 vezes nas páginas 8 e 9.

EVERITT, B. S. Unresolved problems in cluster analysis. **Biometrics**, JSTOR, p. 169–181, 1979. Citado na página 30.

GOOGLE. **Damos-lhe as boas-vindas ao Colaboratory**. Google, 2022. Disponível em: <https://colab.research.google.com/>. Acesso em: 06 de outubro de 2022. Citado na página 58.

GOOGLE. **Google Cloud para pesquisadores**. 2022. Disponível em: <https://cloud.google.com/edu/researchers>. Acesso em: 06 de outubro de 2022. Citado na página 59.

GOOGLE. **Sonhe, crie e transforme com o Google Cloud**. 2022. Disponível em: <https://cloud.google.com/?hl=pt-br>. Acesso em: 06 de outubro de 2022. Citado na página 59.

GOOGLE. **Supere desafios reais dos negócios no Google Cloud**. 2022. Disponível em: <https://cloud.google.com/free?hl=pt-br>. Acesso em: 06 de outubro de 2022. Citado na página 59.

GROMIHA, M. M.; SELVARAJ, S. Inter-residue interactions in protein folding and stability. **Progress in biophysics and molecular biology**, Elsevier, v. 86, n. 2, p. 235–277, 2004. Citado na página 1.

GUPTA, A. **ML | OPTICS Clustering Explanation**. Geeks for Geeks, 2019. Disponível em: <https://www.geeksforgeeks.org/ml-optics-clustering-explanation/>. Acesso em: 11 de outubro de 2022. Citado 2 vezes nas páginas 35 e 36.

GUPTA, A. M.; CHAKRABARTI, J. Effect on the conformations of the spike protein of sars-cov-2 due to mutation. **Biotechnology and Applied Biochemistry**, Wiley Online Library, 2022. Citado na página 19.

HAZES, B.; DIJKSTRA, B. W. Model building of disulfide bonds in proteins with known three-dimensional structure. **Protein Engineering, Design and Selection**, Oxford University Press, v. 2, n. 2, p. 119–125, 1988. Citado na página 9.

HOLTZ, Y. **Basic Dendrogram**. the Python Graph Gallery, 2018. Disponível em: <https://www.python-graph-gallery.com/400-basic-dendrogram>. Acesso em: 03 de novembro de 2022. Citado na página 37.

HU, S.-H. et al. The 1.1 å crystal structure of the neuronal acetylcholine receptor antagonist, α-conotoxin pnia from conus pennaceus. **Structure**, Elsevier, v. 4, n. 4, p. 417–423, 1996. Citado na página 16.

HUANG, L.-T.; GROMIHA, M. M.; HO, S.-Y. iptree-stab: interpretable decision tree based method for predicting protein stability changes upon mutations. **Bioinformatics**, Oxford University Press, v. 23, n. 10, p. 1292–1293, 2007. Citado na página 8.

HUBBARD, R. E.; HAIDER, M. K. Hydrogen bonds in proteins: role and strength. **eLS**, John Wiley & Sons, Ltd, 2010. Citado 2 vezes nas páginas 6 e 24.

HUNTER, L. Molecular biology for computer scientists, artificial intelligence and molecular biology. **Google Scholar Google Scholar Digital Library Digital Library**, 1993. Citado 2 vezes nas páginas 1 e 22.

JEFFREY, G. A. **An introduction to hydrogen bonding**. [S.l.]: Oxford university press New York, 1997. v. 12. Citado na página 11.

JIN, X.; HAN, J. K-means clustering. In: ____. **Encyclopedia of Machine Learning**. Boston, MA: Springer US, 2010. p. 563–564. ISBN 978-0-387-30164-8. Disponível em: <https://doi.org/10.1007/978-0-387-30164-8_425>. Citado na página 33.

KABSCH, W. A solution for the best rotation to relate two sets of vectors. **Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography**, International Union of Crystallography, v. 32, n. 5, p. 922–923, 1976. Citado 2 vezes nas páginas 2 e 11.

KAMIL, I. S.; AL-MAMORY, S. O. Enhancement of optics'time complexity by using fuzzy clusters. **Materials Today: Proceedings**, Elsevier, 2021. Citado na página 37.

KARIM, M. R. et al. Deep learning-based clustering approaches for bioinformatics. **Briefings in Bioinformatics**, Oxford University Press, v. 22, n. 1, p. 393–415, 2021. Citado 2 vezes nas páginas 18 e 30.

KOLODNY, R.; LINIAL, N. Approximate protein structural alignment in polynomial time. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 101, n. 33, p. 12201–12206, 2004. Citado na página 26.

KUFAREVA, I.; ABAGYAN, R. Methods of protein structure comparison. **Homology Modeling: Methods and Protocols**, Springer, p. 231–257, 2012. Citado na página 11.

LAIMER, J. et al. Maestro-multi agent stability prediction upon point mutations. **BMC bioinformatics**, BioMed Central, v. 16, n. 1, p. 1–13, 2015. Citado na página 8.

LASSMANN, T. **Kalign 3: multiple sequence alignment of large datasets**. [S.l.]: Oxford University Press, 2020. Citado na página 9.

LASSMANN, T.; FRINGS, O.; SONNHAMMER, E. L. Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. **Nucleic acids research**, Oxford University Press, v. 37, n. 3, p. 858–865, 2009. Citado na página 9.

LASSMANN, T.; SONNHAMMER, E. L. Kalign–an accurate and fast multiple sequence alignment algorithm. **BMC bioinformatics**, BioMed Central, v. 6, n. 1, p. 1–9, 2005. Citado na página 9.

LIMA, E. B. d. **Utilizando Agrupamento Com Restrições e Agrupamento Espectral Para Integração de Dados de Enzimas**. Dissertação (Mestrado) — Universidade Federal de Minas Gerais - UFMG, 2011. Disponível em: <https://repositorio.ufmg.br/bitstream/1843/SLSS-8GQGQC/1/elisaboarilima.pdf>. Acesso em: 04 de março de 2023. Citado na página 18.

LIMA, E. B. d. **Detecção de Subfamílias Proteicas Isofuncionais Utilizando Integração de Dados e Agrupamento Espectral**. Tese (Doutorado) — Universidade Federal de Minas Gerais - UFMG, 2015. Disponível em: <https://repositorio.ufmg.br/bitstream/1843/BUOS-APTNCE/1/tese_elisa_boari_de_lima_2011695257.pdf>. Acesso em: 04 de março de 2023. Citado na página 18.

MARIANO, D. C. B. et al. A computational method to propose mutations in enzymes based on structural signature variation (ssv). **International Journal of Molecular Sciences**, v. 20, n. 2, 2019. ISSN 1422-0067. Disponível em: <https://www.mdpi.com/1422-0067/20/2/333>. Citado na página 16.

MATHWORKS. **multcompare**. 2005. Disponível em: <https://www.mathworks.com/help/stats/multcompare.html>. Acesso em: 23 de maio de 2021. Citado na página 66.

MATHWORKS. **anova1**. 2006. Disponível em: <https://www.mathworks.com/help/stats/anova1.html>. Acesso em: 15 de abril de 2021. Citado na página 74.

MATHWORKS. **K-Means Clustering**. 2006. Disponível em: <http://www.mathworks.com/help/stats/k-means-clustering.html>. Acesso em: 23 de abril de 2020. Citado 2 vezes nas páginas 30 e 31.

MELO-MINARDI, R. C. de. **T1:E1 Sobreposição estrutural de proteínas**. Online Bioinfo Bioinformática. Youtube, 2021. Disponível em: <https://www.youtube.com/watch?v=pDdtWT23Zso&t=5s>. Acesso em: 24 de fevereiro de 2023. Citado 3 vezes nas páginas 2, 5 e 11.

MONTEIRO, O. M. **Desenvolvimento de uma metodologia baseada em matriz de distâncias para a verificação de similaridades de proteínas**. Dissertação (Mestrado) — Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), 2017. Disponível em: <https://sig.cefetmg.br/sigaa/verArquivo?idArquivo=2014217&key=017616e954e6ef20026810838e63f61e>. Acesso em: 17 de junho de 2020. Citado 8 vezes nas páginas 3, 18, 41, 42, 43, 44, 59 e 60.

MONTEIRO, O. M.; DIAS, S. R.; RODRIGUES, T. d. S. Reduced distance matrix to verify the similarity between protein structures. **Brazilian Archives of Biology and Technology**, SciELO Brasil, v. 64, 2021. Citado 3 vezes nas páginas 46, 48 e 59.

MONTEIRO, O. M.; DIAS, S. R.; RODRIGUES, T. de S. Desenvolvimento de uma metodologia baseada em matriz de distancias para a verificacao de similaridades de proteinas. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**, v. 7, n. 1, 2020. Citado 8 vezes nas páginas 3, 18, 28, 42, 43, 44, 46 e 59.

MONTEIRO, O. M.; DIAS, S. R.; RODRIGUES, T. de S. Desenvolvimento da mdrpca para a verificar similaridades de proteinas. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**, 2021. Citado 3 vezes nas páginas 49, 50 e 59.

MONTGOMERY, D. **Design and Analysis of Experiments**. [S.l.]: John Wiley & Sons, 2019. Citado na página 66.

MORAES, C. et al. **Série em biologia molecular e celular: métodos experimentais no estudo de proteínas**. [S.l.]: Fundação Oswaldo Cruz–Fiocruz, Rio de Janeiro, 2013. Citado 2 vezes nas páginas 21 e 22.

NEI, M.; KUMAR, S. Phylogenetic inference: distance methods. **Molecular evolution and phylogenetics**, p. 89, 2000. Citado na página 29.

NELSON, D. L.; COX, M. M. **Princípios de Bioquímica de Lehninger**. [S.l.]: Artmed, 2014. Citado 6 vezes nas páginas 1, 6, 21, 22, 24 e 25.

N.RAJALINGAM; K.RANJINI. Article: Hierarchical clustering algorithm - a comparative study. **International Journal of Computer Applications**, v. 19, n. 3, p. 42–46, April 2011. Full text available. Citado na página 37.

OLIVEIRA, M. d. S. **Identificação de padrões conformacionais em estruturas experimentais e projeto de metaheurísticas para a predição da estrutura tridimensional de proteínas**. 2016. Citado 3 vezes nas páginas 2, 5 e 11.

PACE, C. N. et al. Contribution of hydrophobic interactions to protein stability. **Journal of molecular biology**, Elsevier, v. 408, n. 3, p. 514–528, 2011. Citado na página 1.

PACE, C. N. et al. Contribution of hydrogen bonds to protein stability. **Protein Science**, Wiley Online Library, v. 23, n. 5, p. 652–661, 2014. Citado na página 6.

PANTOLIANO, M. W. et al. The engineering of binding affinity at metal ion binding sites for the stabilization of proteins: subtilisin as a test case. **Biochemistry**, ACS Publications, v. 27, n. 22, p. 8311–8317, 1988. Citado na página 6.

PANTOLIANO, M. W. et al. Large increases in general stability for subtilisin bpn through incremental changes in the free energy of unfolding. **Biochemistry**, ACS Publications, v. 28, n. 18, p. 7205–7213, 1989. Citado na página 6.

PATLOLLA, C. R. **Understanding the concept of Hierarchical clustering Technique**. Towards Data Science, 2018. Disponível em: <https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec>. Acesso em: 07 de novembro de 2022. Citado na página 39.

PDB. **Coordinate Section**. 2010. Disponível em: <http://www.wwpdb.org/documentation/file-format-content/format33/sect9.html>. Acesso em: 15 de fevereiro de 2020. Citado 2 vezes nas páginas 60 e 61.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Citado 10 vezes nas páginas 4, 31, 32, 33, 34, 37, 38, 39, 40 e 63.

PEI, J.; KIM, B.-H.; GRISHIN, N. V. Promals3d: a tool for multiple protein sequence and structure alignments. **Nucleic acids research**, Oxford University Press, v. 36, n. 7, p. 2295–2300, 2008. Citado na página 9.

PEREIRA, C. M. M. **OPTICS: Ordering Points to Identify the Clustering Structure**. Universidade de São Paulo, 2010. Disponível em: <http://wiki.icmc.usp.br/images/1/1f/OPTICS_CassioPereira.pdf>. Acesso em: 13 de outubro de 2022. Citado na página 35.

PINHO, M. S. **Matriz e Vetores**. 2020. Disponível em: <http://www.inf.pucrs.br/~pinho/LaproI/Vetores/Vetores.htm>. Acesso em: 20 de abril de 2020. Citado na página 29.

PIRES, D. E.; ASCHER, D. B. Csm-lig: a web server for assessing and comparing protein–small molecule affinities. **Nucleic acids research**, Oxford University Press, v. 44, n. W1, p. W557–W561, 2016. Citado na página 17.

PIRES, D. E.; ASCHER, D. B.; BLUNDELL, T. L. mcsm: predicting the effects of mutations in proteins using graph-based signatures. **Bioinformatics**, Oxford University Press, v. 30, n. 3, p. 335–342, 2014. Citado na página 17.

PIRES, D. E.; BLUNDELL, T. L.; ASCHER, D. B. pkcsm: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. **Journal of medicinal chemistry**, ACS Publications, v. 58, n. 9, p. 4066–4072, 2015. Citado na página 17.

PIRES, D. E. et al. Cutoff scanning matrix (csm): structural classification and function prediction by protein inter-residue distance patterns. **BMC genomics**, BioMed Central, v. 12, n. 4, p. S12, 2011. Citado 4 vezes nas páginas 3, 17, 18 e 19.

PIRES, D. E. et al. acsm: noise-free graph-based signatures to large-scale receptor-based ligand prediction. **Bioinformatics**, Oxford University Press, v. 29, n. 7, p. 855–861, 2013. Citado 2 vezes nas páginas 3 e 17.

PIRES, D. E. V. **CSM: Uma Assinatura para Grafos Biológicos Baseada em Padrões de Distâncias**. Tese (Doutorado) — Universidade Federal de Minas Gerais - UFMG, 2012. Disponível em: <http://homepages.dcc.ufmg.br/~dpires/publications/thesis.pdf>. Acesso em: 29 de marco de 2021. Citado na página 17.

RAY, D.; LE, L.; ANDRICIOAEI, I. Distant residues modulate conformational opening in sars-cov-2 spike protein. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 118, n. 43, p. e2100943118, 2021. Citado na página 19.

RODRIGUES, C. H.; PIRES, D. E.; ASCHER, D. B. mmcsm-ppi: predicting the effects of multiple point mutations on protein–protein interactions. **Nucleic Acids Research**, Oxford University Press, v. 49, n. W1, p. W417–W424, 2021. Citado na página 17.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, Elsevier, v. 20, p. 53–65, 1987. Citado na página 63.

ROY, S.; CHAKRABARTI, A. Chapter 11 - a novel graph clustering algorithm based on discrete-time quantum random walk. In: BHATTACHARYYA, S.; MAULIK, U.; DUTTA, P. (Ed.). **Quantum Inspired Computational Intelligence**. Boston: Morgan Kaufmann, 2017. p. 361–389. ISBN 978-0-12-804409-4. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128044094000115>. Citado na página 39.

RSCB. **Estatísticas PDB: crescimento de estruturas de experimentos de cristalografia de raios-X lançados por ano**. 2023. Disponível em: <https://www.rcsb.org/stats/growth/growth-xray>. Acesso em: 01 de agosto de 2023. Citado 2 vezes nas páginas vii e 4.

SAKAGUCHI, M. et al. Role of disulphide bonds in a thermophilic serine protease aqualysin i from thermus aquaticus yt-1. **Journal of biochemistry**, Oxford University Press, v. 143, n. 5, p. 625–632, 2008. Citado na página 6.

SANTOS, T. G. **Google Colab: o que é, tutorial de como usar e criar códigos**. Alura, 2022. Disponível em: <https://www.alura.com.br/artigos/google-colab-o-que-e-e-como-usar>. Acesso em: 06 de outubro de 2022. Citado na página 58.

SANTOS, V. D.; MARTINS, P.; MARIANO, D. Alinhamentos estruturais: métodos de sobreposição de proteínas e outras moléculas. **BIOINFO - Revista Brasileira de Bioinformática e Biologia Computacional**, Editora Alfahelix, v. 1, n. 1, p. 1, 2021. Citado 2 vezes nas páginas 26 e 27.

SCIPY. **scipy.cluster.hierarchy.linkage**. The SciPy community, 2022. Disponível em: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html>. Acesso em: 04 de novembro de 2022. Citado 2 vezes nas páginas 38 e 39.

SINCLAIR, C. **Clustering Using OPTICS: A seemingly parameter-less algorithm**. Towards Data Science, 2019. Disponível em: <https://towardsdatascience.com/clustering-using-optics-cac1d10ed7a7>. Acesso em: 13 de outubro de 2022. Citado 2 vezes nas páginas 35 e 36.

SOWDHAMINI, R. et al. Stereochemical modeling of disulfide bridges. criteria for introduction into proteins by site-directed mutagenesis. **Protein Engineering, Design and Selection**, Oxford University Press, v. 3, n. 2, p. 95–103, 1989. Citado na página 10.

TENG, S.; SRIVASTAVA, A. K.; WANG, L. Sequence feature-based prediction of protein stability changes upon amino acid substitutions. **BMC genomics**, BioMed Central, v. 11, n. 2, p. 1–8, 2010. Citado na página 8.

TSAI, C. S. **An introduction to computational biochemistry**. [S.l.]: John Wiley & Sons, 2003. Citado na página 29.

WEI, Y. et al. Allosteric regulation of sars-cov-2 spike protein revealed by contrastive machine learning. 2023. Citado na página 20.

WINN, M. D. et al. Overview of the ccp4 suite and current developments. **Acta Crystallographica Section D: Biological Crystallography**, International Union of Crystallography, v. 67, n. 4, p. 235–242, 2011. Citado 2 vezes nas páginas 2 e 11.

XU, R.; WUNSCH, D. **Clustering**. [S.l.]: John Wiley & Sons, 2008. v. 10. Citado na página 30.

YU, X. **Application for Organic Nitrogen compounds - Role of Protein in Hair**. 2013. Disponível em: <http://chemysteryworld.blogspot.com/2013/09/application-for-organic-nitrogen_22.html>. Acesso em: 22 de setembro de 2019. Citado na página 23.

YUFENG. **Understanding OPTICS and Implementation with Python**. Towards Data Science, 2022. Disponível em: <https://towardsdatascience.com/understanding-optics-and-implementation-with-python-143572abdfb6>. Acesso em: 11 de outubro de 2022. Citado 2 vezes nas páginas 34 e 35.

ZANCHETTA, A. **Celulases e suas aplicações**. 2013. Disponível em: <http://www.rc.unesp.br/ib/ceis/mundoleveduras/2013/Celulases.pdf>. Acesso em: 09 de agosto de 2021. Citado na página 10.

ZHOU, Y.; YU, Z.-G.; ANH, V. Cluster protein structures using recurrence quantification analysis on coordinates of alpha-carbon atoms of proteins. **Physics Letters A**, Elsevier, v. 368, n. 3-4, p. 314–319, 2007. Citado na página 19.