UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
CIÊNCIA DA COMPUTAÇÃO

UM ESTUDO COMPARATIVO DE ALGORITMOS DE AGRUPAMENTO DE DADOS PARA DADOS DE DOCAGEM MOLECULAR

TRABALHO DE GRADUAÇÃO

Otávio Machado

Santa Maria, RS, Brasil

2014

UM ESTUDO COMPARATIVO DE ALGORITMOS DE AGRUPAMENTO DE DADOS PARA DADOS DE DOCAGEM MOLECULAR

Otávio Machado

Trabalho de Graduação apresentado ao curso de Ciência da Computação da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para a obtenção do grau de

Bacharel em Ciência da Computação

Orientador: Prof.ª Dr.ª Ana T. Winck

Trabalho de Graduação N.379
Santa Maria, RS, Brasil

2014

Universidade Federal de Santa Maria
Centro de Tecnologia
Ciência da Computação

A Comissão Examinadora, abaixo assinada,
aprova o Projeto de Trabalho de Graduação

UM ESTUDO COMPARATIVO DE ALGORITMOS DE AGRUPAMENTO DE DADOS PARA DADOS DE DOCAGEM MOLECULAR

elaborado por
Otávio Machado

como requisito parcial para obtenção do grau de
Bacharel em Ciência da Computação

COMISSÃO EXAMINADORA:

Ana T. Winck, Dr.ª
(Presidente/Orientador)

Giovani R. Librelotto, Dr. (UFSM)

Sérgio L. S. Mergen, Dr. (UFSM)

Santa Maria, 1 de Dezembro de 2014.

AGRADECIMENTOS

Primeiramente, gostaria de generalizar meus agradecimentos a todos que apareceram ou participaram da minha vida nos últimos cinco anos da graduação. Vocês, de uma forma ou outra, me ensinaram algo e contribuíram de forma positiva na minha vida acadêmica, profissional ou pessoal.

Obrigado aos meus pais, que me deram a mão e me mostraram o mundo com uma visão que muito aprecio até hoje. À minha mãe, Cleonice, que me ensina todos os dias, a cada palavra, escrita ou falada, o amor incondicional que é o de mãe, e ao meu pai, Ricardo, que me ensinou, não só em palavras, mas em atitudes, que o nosso vínculo, enquanto família, é capaz de vencer absolutamente tudo. Algum dia quero ser capaz de transmitir um pouco de todo o sentimento, sabedoria e união que vocês me passaram. Obrigado à minha irmã, Cláudia, que sempre me acolheu de braços e coração abertos. Tu sempre fostes um exemplo de pessoa, aos meus olhos por ser exatamente quem tu és, e tu me ensinaste muito, mesmo às vezes sem querer - e às vezes querendo, Dom Manuel, o Venturoso. - Obrigado aos meus irmãos, Ezequiel e Tiago, que não só fazem meus dias mais felizes, mesmo que à distância, mas também me fazem crer na consciência das próximas gerações, e a todos os meus familiares, que sempre me apoiaram em todos os aspectos.

Quero também agradecer em especial a alguns amigos, que não só contribuíram com a minha vida, mas que com certeza fazem de mim uma pessoa muito melhor. A ti, Mariane, porque és, para mim, sempre Naíra. À Ana Schwendler, que desde o momento que nos reconhecemos, mesmo enquanto colegas de profissão, sabia que tinha mais uma irmã na vida. Ao meu namorado - e futuro marido - Dayvson, que me ensinou os significados de amor e sonho, e fez dos últimos anos só mais um do qual eu não preciso acordar. Também deixo meu obrigado a todos os amigos que não podem deixar de serem mencionados: Júnior, Larissa, Marcelo, Guilherme, Ivo, dentre tantos outros.

Gostaria de mencionar o Governo Federal, que investiu no ensino público e me proporcionou um estudo de quatro anos em uma universidade de alta qualidade, e ao programa Ciência Sem Fronteiras e o CPNq, que me proporcionaram a maior e melhor experiência acadêmica, cultural e social que eu podia ter imaginado, que foram meus dez meses em Portugal.

Obrigado aos meus amigos que conheci nesses dez meses, e que fizeram desse tempo mais um sonho em vida: Marília, minha eterna companheira de quarto, Caio e Guilherme, que

juntos formamos uma segunda família durante essa vivência, Johan, Thyago, Renato e Orlando, que são amizades que levo pelo resto da minha vida, mesmo que agora distantes.

À professora e orientadora Ana Winck, que realizou o meu desejo de pesquisar e desenvolver voltando-me à Bioinformática, por todas as reuniões, decisões, repetições sobre aquilo que não entrava na minha cabeça e toda a ajuda que eu precisei, os meus mais sinceros agradecimentos. À coordenadora Andrea, e ao secretário de Ciência da Computação, Marcelo, por sempre estarem dispostos a ajudar, principalmente no meu último semestre de curso, o meu muito obrigado. Obrigado também aos professores Giovani e Sérgio, que aceitaram fazer parte da banca avaliadora deste trabalho.

Por fim, não posso deixar de agradecer aos colegas Alberto, Thiago e Tháygoro, que me ensinaram que passar vinte e quatro horas corridas simulando deadlocks pode ser divertido, pelo companheirismo e por todas as risadas.

“A melhor maneira de prever o futuro é inventá-lo.”

— Alan Kay

RESUMO

Trabalho de Graduação
Ciência da Computação
Universidade Federal de Santa Maria

UM ESTUDO COMPARATIVO DE ALGORITMOS DE AGRUPAMENTO DE DADOS PARA DADOS DE DOCAGEM MOLECULAR
AUTOR: OTÁVIO MACHADO
ORIENTADOR: ANA T. WINCK
Local da Defesa e Data: Santa Maria, 1 de Dezembro de 2014.

Em Bioinformática, tratamos com uma grande quantidade de dados do contexto da biologia, buscando obter, a partir deles, informações valiosas. Desenho Racional de Fármacos é uma área da bioinformática, a qual é desenvolvido a partir da interação entre duas moléculas, chamadas Receptor e Ligante. Através de experimentos de simulação por Docagem Molecular, busca-se o melhor posicionamento dessas duas moléculas buscando a melhor ligação possível entre as duas. A qualidade desta ligação é medida pelo FEB - Energia livre de ligação. Através dessas simulações, os melhores ligantes para um determinado receptor são testados in vitro e uma nova droga pode surgir. Uma estratégia para melhor obter esses resultados é considerar tanto o receptor como o ligante como estruturas flexíveis. Entretanto, ao considerar essa flexibilidade, o custo computacional para experimentos de docagem torna-se muito alto. Nesse sentido, o algoritmo 3D-Tri busca encontrar conformações promissoras para futuros experimentos de docagem, a partir de uma estratégia de árvore de decisão, baseada em agrupamento de dados. O objetivo do presente trabalho é tratar esse grande volume de dados, e contribuir para a evolução do algoritmo 3D-Tri, através da análise de algoritmos de agrupamento de dados para uma possível redução no tempo de execução de futuros experimentos de docagem molecular.

Palavras-chave: Bioinformática. Desenho Racional de Fármacos. Agrupamento de Dados.

ABSTRACT

Undergraduate Final Work
Federal University of Santa Maria

A COMPARATIVE STUDY OF DATA CLUSTERING ALGORITHMS WITH DOCKING DATA
AUTHOR: OTÁVIO MACHADO
ADVISOR: ANA T. WINCK
Defense Place and Date: Santa Maria, October 1, 2014.

In Bioinformatics, we deal with a big amount of biology-related data, trying to obtain, from them, valuable information. Rational Drug Design is one field of research in bioinformatics, which is developed from the interaction between two molecules, called Receptor and Ligand. Through molecular docking simulations, we try to find the best positioning of these two molecules, searching for the best possible binding between them. The binding quality is measured by the FEB - Free Energy of Binding. Through these simulations, the best ligands to a receptor are tested in vitro and a new drug may be created. A strategy to get these results regards considering both the receptor and the ligand as flexible structures. However, the computational cost increases when considering such a flexibility. Hence, the 3D-Tri algorithm tries to find promising conformations to further molecular dockings experiments from a decision-tree strategy based on clustering analysis. The objective of this work is to treat this large amount of data, as well as contribute to evolve the 3D-Tri algorithm, with the analysis of data clustering algorithms to a possible time reduction in the execution of future docking experiments.

Keywords: Bioinformatics, Rational Drug Design, Data clustering.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Bioinformática é uma área interdisciplinar, a qual utiliza de técnicas computacionais para o tratamento de dados biológicos. Uma das áreas de estudo da bioinformática é o Desenho Racional de Fármacos (KUNTZ, 1992), o qual investiga, fundamentalmente, a interação entre receptores (macromoléculas) e ligantes (pequenas moléculas candidatas a fármaco). É na docagem molecular que se investiga o melhor encaixe do ligante em um receptor, de modo que um ligante deve interagir com um receptor para que as funções do receptor sejam inibidas (LYBRAND, 1995). Um dos grandes desafios está em realizar experimentos de Docagem Molecular, considerando-se o receptor flexível, o qual pode exercer diferentes formas. Para simular essa variação, são realizadas simulações de Dinâmica Molecular (LIN et al., 2002), utilizando-se cada conformação da dinâmica em um experimento de docagem molecular.

Dados provenientes de Docagem molecular e de dinâmica molecular são de grande escala, que podem conter ruído e estão relacionados com as coordenadas geométricas da disposição dos átomos de cada molécula. Por tratar-se de um grande volume de dados, tanto em termos de conformações do receptor como em termos da quantidade de lgantes disponiveis em bancos de dados (o Zinc (IRWIN; SHOICHET, 2005), atualmente, possui mais de 35 milhões de compostos disponóveis), o teste de todos os compostos com todas as conformações torna-se computacionalmente custoso. Nesse sentido, uma estratégia para reduzir o tempo de execução está em dentificar as melhores conformações de dinâmica molecular para futuros experimentos de docagem molecular.

A forma como essas conformações promissoras são identifcadas torna-se um desafio, de modo que algoritmos de agrupamento de dados podem se tornar úteis. Trabalhos iniciais já foram realizados (MACHADO et al., 2011) (WINCK, 2012) para incorpor algoritmos de agrupamento de dados como forma de indentificar tais conformações. Buscando evoluir os resultados dos trabalhos supracitados, o presente trabalho busca melhor explorar o trabalho desenvolvido em (WINCK, 2012), no que se refere à aplicação de agrupamento de dados sobre dados de dinâmica molecular e docagem molecular. Para isso, é preciso definir quais algoritmos de agrupamento de dados lidam de forma razoável com este conjunto de dados. Neste sentido, este trabalho apresenta de um estudo de algoritmos de agrupamento de dados, com uma análise de sua eficácia quando utilizados para agrupar dados tridimensionais de átomos das diferentes conformações que uma proteína pode apresentar.

O objetivo geral deste trabalho é analisar algoritmos de agrupamentos de dados de lógicas diferenciadas, de forma a encontrar quais tipos de algoritmos são eficientes quando utilizados em conjunto com os dados tridimensionais provenientes de resultados de Docagem Molecular. Além disso, proporcionar um framework que permita ao usuário definir qual tipo de agrupamento ele gostaria de aplicar ao conjunto de dados. Para tanto, os seguintes passos são levados em considerção:

- Analisar algoritmos de agrupamento de dados de acordo com o resultado de sua aplicação em um conjunto de dados específico;

- Definir para quais tipos de dados cada tipo de algoritmo de agrupamentos de dados é capaz de trabalhar melhor;

- Descobrir qual método melhor se adapta ao problema de dados de docagem molecular.

2 REVISÃO BIBLIOGRÁFICA

Este capítulo apresenta uma revisão acerca dos temas e métodos abordados neste trabalho. Pela parte da bioinformática, é descrito sobre Desenho Racional de Fármacos, apresentando as suas etapas e estratégias de desenvolvimento. Pelo ponto de vista computacional, é realizada uma revisão a respeito dos diferentes algoritmos de agrupamento de dados.

2.1 Desenho Racional de Fármacos

Inicialmente, bioinformática era definida como uma área interdisciplinar envolvendo biologia, ciência da computação, matemática e estatística para analisar diferentes tipos de dados biológicos (Mount, 2004). Atualmente, com o desenvolvimento da computação e o advento da era Genômica, Luscombe et. al (LUSCOMBE; GREENBAUM; GERSTEIN, 2001) definem a área como biologia em termos de moléculas e a aplicação da computação para entender e organizar a informação associada com esses dados biológicos em larga escala.

Apesar de ser relativamente nova, a Bioinformática é muito ampla e pode ser dividida em diversas sub-áreas. Lesk (LESK, 2002) aponta algumas principais frentes, como genomas, proteomas, alinhamento de árvores filogenéticas, biologia de sistemas, estruturas de proteínas e descoberta de fármacos - sendo, a última, o foco deste trabalho.

O Desenho Racional de Fármacos, ou RDD - Rational Drug Design - é a área da Bioinformática cujo objetivo fundamental é explorar a interação entre receptores e ligantes com o objetivo de encontrar algum ligante que possa se tornar um possível inibidor para alguma proteína associada a uma determinada doença (LYBRAND, 1995). Ligantes são definidos como moléculas que se ligam a outras moléculas biológicas, chamadas receptores, para realizar ou inibir funções específicas (BALAKIN, 2009). O processo de RDD é composto de quatro etapas (KUNTZ, 1992):

1. A primeira etapa consiste em isolar um alvo específico, chamado receptor, (proteínas, receptores de membrana, DNA, RNA, etc.). A partir da análise computacional da estrutura tridimensional (3D) dessa proteína armazenada em um banco de dados estrutural como o Protein Data Bank (PDB) (H.M. BERMAN et al., 2000), é possível apontar prováveis regiões de ligação, por exemplo, regiões onde uma pequena molécula, chamada ligante, pode se ligar a esse receptor;

2. Baseado nas prováveis regiões de ligação identificadas na etapa anterior é selecionado um conjunto de prováveis candidatos a ligantes que podem se ligar a essa região no receptor. As diferentes conformações que um dado ligante pode assumir dentro do sítio de ligação de uma determinada proteína podem ser simuladas por software de docagem molecular como AutoDock3.0.5 (MORRIS et al., 1998);

3. Os ligantes que teoricamente obtiveram melhores resultados nas simulações são experimentalmente sintetizados e testados;

4. Baseado nos resultados experimentais, o medicamento é gerado ou o processo retorna à etapa 1 com pequenas modificações no ligante.

   O restante deste trabalho está organizado conforme segue. O Capítulo 2 apresenta uma revisão bibliográfica acerca dos assuntos relacionados ao desenho racional de fármacos e de algoritmos de agrupamento de dados. No Capítulo é apresentada uma motivação e metodologia desenvolvida. O Capítulo 4 mostra os resultados obtidos, seguido da conclusão e referências bibliográficas.

A ligação entre receptores e ligantes pode ser avaliada através de um método chamado Docagem Molecular, cujo propósito é avaliar a qualidade das possibilidades de ligação destas duas moléculas. Um ligante se unirá a um receptor para exercer uma função fisiológica vinculada à ligação dessa com outras moléculas, e essas ligações determinam se as funções serão estimuladas ou inibidas. É somente em locais específicos, chamados sítios de ativação, onde essas ligações ocorrem, pois a força dessa ligação não depende somente da forma como as duas moléculas se ligam, mas também da energia favorável à essa ligação - chamada de FEB, *free energy of binding*. Estas interações, que ocorrem a nível atômico, são medidas através da quantidade de energia despendida. Ou seja, quanto mais negativa, mais forte é a ligação realizada e, portanto, mais efetivo o resultado da função do ligante.

Para que as possíveis ligações sejam avaliadas, no entanto, é necessário levar em conta muitos fatores, como a mobilidade do ligante e do receptor, o efeito do ambiente no receptor (proteína), a distribuição de carga no ligante, e outras interações dos mesmos com a água, que complicam muito a descrição desse processo. Existem muitos algoritmos que buscam simular estas condições mas, em grande parte, apenas a flexibilidade do ligante é considerada, tornando o receptor rígido. Uma das soluções apresentadas na literatura é executar uma série de experimentos de docagem molecular, sendo cada um desses experimentos uma conformação do

receptor gerada por uma simulação por Dinâmica Molecular. Uma simulação por DM é uma das técnicas computacionais mais versáteis e amplamente utilizadas para o estudo de macromoléculas biológicas (GUNSTEREN; BERENDSEN, 1990). Com simulações pela DM é possível estudar o efeito explícito de ligantes na estrutura e estabilidade das proteínas, os diferentes parâmetros termodinâmicos envolvidos, incluindo energias de interação e entropias.

Essa foi a estratégia relatada em WINCK et al. (2010) para os trabalhos sendo desenvolvidos. Porém, gerar uma simulação deste tipo traz dois grandes problemas: o tempo necessário para executar estes experimentos e o grande volume de dados gerados. Tratar esse grande volume de dados para uma possível redução no tempo de execução de futuros experimentos de docagem molecular é o objetivo principal deste trabalho. Através de um estudo mais aprofundado dos tipos de algoritmos de Agrupamento de dados, poderemos entender como dados tão complexos se comportam, e quais abordagens trazem uma maior eficiência na descoberta das conformações mais importantes de serem analisadas.

2.2 Algoritmos de Agrupamento de Dados

A técnica de agrupamento de dados, forma não-supervisionada de aprendizado de máquina, busca organizar um conjunto de objetos em grupos, de acordo com a sua semelhança. Assim, os métodos enquadrados nesta classificação buscam maximizar a similaridade de objetos de um mesmo grupo, e minimizar a similaridade entre objetos de grupos distintos (HAN; KAMBER, 2006). Técnicas de agrupamento de dados geralmente não são aplicáveis eficientemente a qualquer tipo de dados. Assim, percebe-se a importância do estudo dos algoritmos e da análise de dados e contexto específicos de forma a escolher um algoritmo que melhor se enquadre na resolução de cada problema.

2.2.1 k-means

O k-means (HARTIGAN; WONG, 1979) é uma técnica simples de agrupamento de dados particional - ou seja, que divide os dados em grupos onde cada elemento destes dados está contido em exatamente um único grupo -, utilizada para particionar uma população n-dimensional em k conjuntos, baseados em amostras. Este processo é capaz de gerar grupos razoavelmente eficientes em relação à similaridade dos elementos dentro de cada conjunto. É um método facilmente programável, e computacionalmente econômico, capaz de processar um

grande conjunto de dados (TAN; STEINBACH; KUMAR, 2005).

Seu desenvolvimento foi baseado na expectativa de ser capaz de particionar uma população de forma ótima, mas, em geral, este procedimento não será capaz de encontrar a optimalidade - excetuando casos especiais. Segundo o autor (TAN; STEINBACH; KUMAR, 2005), no entanto, ainda não há um método geral de agrupamento de dados que seja capaz de sempre chegar a uma partição ótima de um conjunto de dados.

O funcionamento do processo do algoritmo de k-means consiste na escolha de k grupos, que representa a quantidade de partições que serão feitas na população de dados. Cada um destes k conjuntos é, inicialmente, representado por um único ponto, geralmente aleatório (TAN; STEINBACH; KUMAR, 2005). A partir destes pontos, chamados de centróides, cada ponto da população de dados é, então, atribuído ao grupo cujo centróide está mais próximo. Quando todos os pontos da população foram atribuídos, é necessário ajustar o centróide de cada grupo, baseado nos pontos que lhe foram atribuídos. O algoritmo básico do k-means é formalmente descrito no Algoritmo 1.

Tabela 2.1 – Algoritmo básico k-means.

Algoritmo 1. Algoritmo Básico k-means
1: Selecionar K pontos como os centróides iniciais
2: repetir
3: Formar k clusters ao atribuir cada ponto para o centróide mais próximo a ele
4: Recomputar o centróide de cada cluster
5: até que os centróides não tenham suas posições mudadas

O primeiro passo é ilustrado na Figura 2.1(a), onde pontos são selecionados como os centróides iniciais. No segundo passo, os pontos são atribuídos aos centróides, e os centróides são novamente atualizados. Nos passos 2, 3 e 4, que são mostrados nas Figuras 2.1(b), (c) e (d), respectivamente, dois dos centróides se movem para os dois menores grupos de pontos na parte de baixo das figuras. O algoritmo termina na Figura 2.1 (d), porque os centróides não têm suas posições mudadas, e parte-se do princípio que os centróides identificaram o agrupamento natural dos pontos.

A principal convenção a respeito da definição de centróides iniciais é de inicializá-los aleatoriamente. Mas esta convenção pode resultar na produção de grupos diferentes de uma execução para outra do algoritmo. Ainda, há a possibilidade de que os grupos resultantes da execução sejam grupos muito ruins, ou que sejam feitas muitas iterações do algoritmo até que

ele convirja em seus grupos. Assim, a escolha dos centróides iniciais é a escolha chave na execução deste algoritmo (TAN; STEINBACH; KUMAR, 2005).

Existem técnicas que podem aumentar as chances de bons agrupamentos de dados resultantes da execução do algoritmo, como:

- Múltipla execução do k-means e posterior escolha dos centróides iniciais iguais à execução que gerou um menor erro quadrático entre os elementos de cada grupo.
- Execução prévia de um algoritmo hierárquico, para que se escolha, na iteração onde houver k grupos, seus centróides como os centróides iniciais para o k-means.
- Escolher o primeiro centróide inicial aleatoriamente, ou definí-lo como o centróide de todos os pontos do conjunto. A partir daí, definir os próximos centróides como o ponto mais longínquo de todos os centróides já definidos previamente.

Cada uma destas soluções possui características boas e ruins, como: a primeira exige uma capacidade de processamento maior dos dados, já que é exigida a múltipla execução do algoritmo para a definição dos centróides. A segunda, por sua vez, exige que o tamanho do conjunto de dados não seja muito grande - na medida de algumas centenas para alguns milhares de dados -. Enquanto a terceira pode selecionar pontos de ruído como centróides iniciais, resultando em uma execução errônea do k-means.

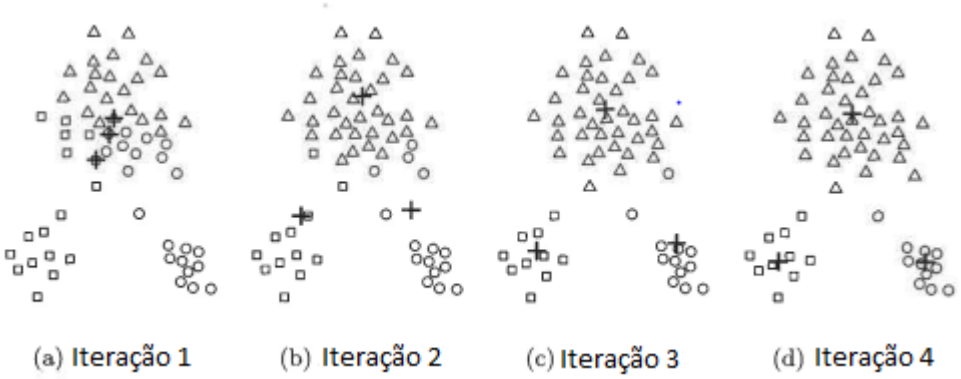

Figura 2.1 – Utilização do algoritmo de k-means para encontrar três clusters nos dados amostrados. (TAN; STEINBACH; KUMAR, 2005)

O algoritmo k-means não é somente utilizado para o agrupamento de dados, mas também, segundo (MACQUEEN, 1975), para Classificação de relevância, aproximação de uma distribuição geral, teste de independência entre muitas variáveis, árvores de classificação baseadas em distância e procedimento de melhoria em dois passos.

Há, no entanto, alguns problemas envolvidos na utilização do processo de k-means no agrupamento de dados, como a possibilidade de obtenção de grupos vazios durante o processo de atribuição e a presença de outliers, que são dados errôneos ou ruído. O processo também não é capaz de lidar com grupos que não sejam centralizáveis, ou grupos de diferentes tamanhos e densidades, como mostrado nas Figuras 2.2, 2.3 e 2.4, embora, segundo (TAN; STEINBACH; KUMAR, 2005) ele seja capaz de encontrar subgrupos puros se o número k de grupos for grande o suficiente. De uma forma geral, o k-means é restrito para problemas onde os dados se separam em grupos que sejam centralizáveis. Para estes problemas há derivações do algoritmo k-means, como o Bisecting k-means, k-medoid, k-means, dentre outros, cujas definições vão além do escopo deste trabalho.

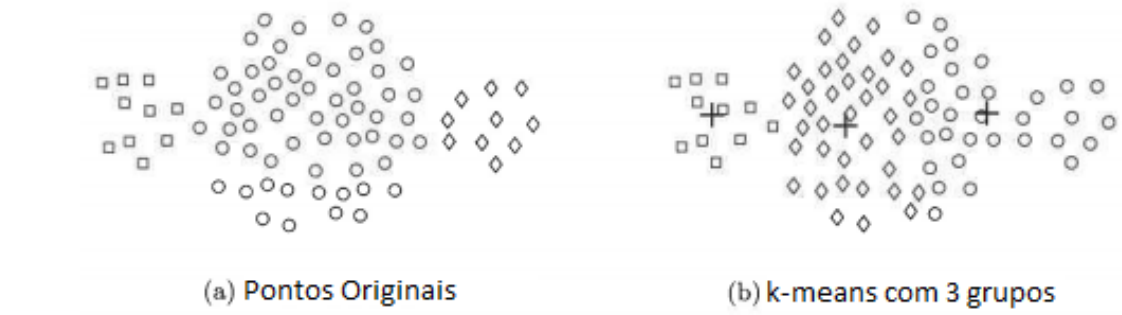

Figura 2.2 – k-means com grupos de tamanhos diferentes. (TAN; STEINBACH; KUMAR, 2005)

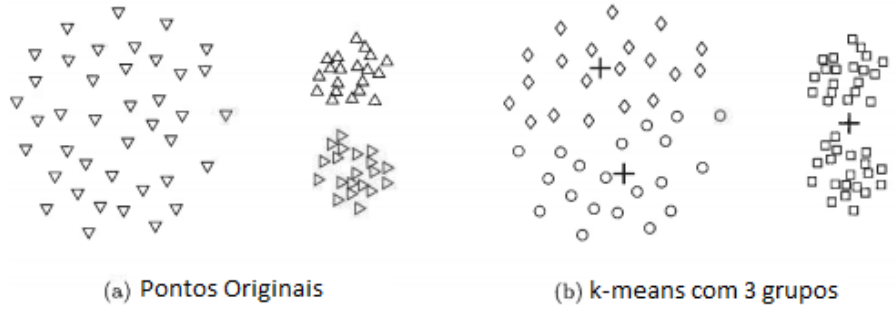

Figura 2.3 – k-means com grupos de densidades diferentes. (TAN; STEINBACH; KUMAR, 2005)

2.2.2 Algoritmo de Ward

O método de Ward é um algoritmo de agrupamento de dados hierárquico e, por definição (WARD, 1963), é um método que gera, para os dados existentes em um conjunto, s níveis de

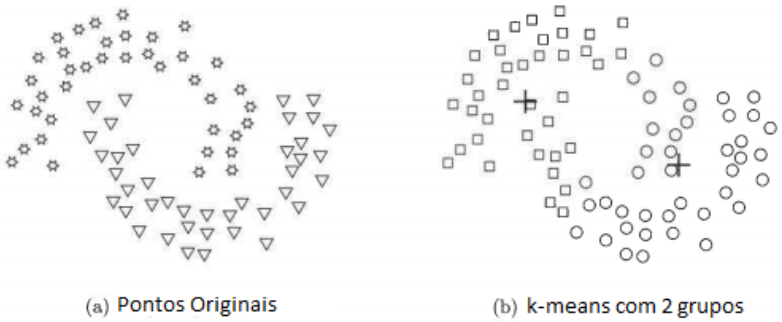

Figura 2.4 – k-means com grupos não-esféricos/não-globulares. (TAN; STEINBACH; KUMAR, 2005)

agrupamento, onde o nível mais baixo é definido por k grupos, cada um contendo apenas um dado do conjunto, e o nível mais alto com apenas um grupo, contendo todos os dados, de forma que permite ao usuário uma maior gama de possibilidades durante a decisão de qual camada ou nível será mais benéfica na aplicação do método a um problema específico

Os algoritmos hierárquicos são visualmente descritos através de dendrogramas ou diagrama de aninhamentos, como mostrado na Figura 2.5, e se diferem nas formas de tomada de decisão - passo 3, no Algoritmo 2 - de quais dados serão agrupados.

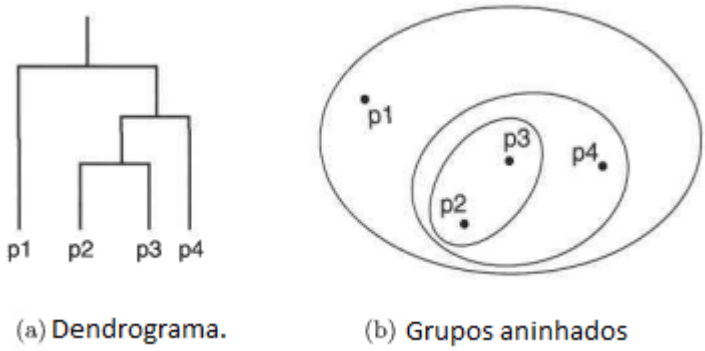

Figura 2.5 – Agrupamento de dados hierárquico de quatro pontos mostrado como um dendrograma e diagrama de aninhamento. (TAN; STEINBACH; KUMAR, 2005)

Especificamente para o método de Ward, segundo a interpretação de (TAN; STEINBACH; KUMAR, 2005) a respeito da proposição original de Joe Ward, a proximidade entre dois grupos é definida como o aumento no erro quadrático que resulta da junção de dois grupos.

O cálculo do aumento no erro quadrático na junção de dois grupos é calculado como disposto na Equação 2.1:

Tabela 2.2 – Algoritmo hierárquico básico.

Algoritmo 2. Algoritmo Hierárquico Básico
1: Computar a matriz de proximidade, se for necessário.
2: repetir
3: Juntar os dois grupos mais similares, ou próximos.
4: Atualizar a matriz de proximidade para refletir a proximidade entre o novo grupo e os grupos restantes.
5: até que só reste um grupo.

$$\Delta(\mathrm{A}+\mathrm{B}) = \sum_{i \in \mathrm{A} \bigcup \mathrm{B}} || \overrightarrow{x}_i - \overrightarrow{m}_{\mathrm{A} \bigcup \mathrm{B}} ||^2 - \sum_{i \in \mathrm{A}} || \overrightarrow{x}_i - \overrightarrow{m}_{\mathrm{A}} ||^2 - \sum_{i \in \mathrm{B}} || \overrightarrow{x}_i - \overrightarrow{m}_{\mathrm{B}} ||^2 = \frac{n_A n_B}{n_A + n_b} || \overrightarrow{m}_A - \overrightarrow{m}_B ||^2 \quad (2.1)$$

Onde:

x_j é o elemento j do conjunto sendo analisado, m_j é o centro do grupo j e n_j, o número de pontos existente neste grupo. Δ é o que se chama o custo de combinação entre os grupos A e B.

As técnicas de agrupamento de dados hierárquicas são normalmente utilizadas quando o problema a ser resolvido exige a criação de uma hierarquia. Há estudos que sugerem que estes algoritmos podem produzir grupos de melhor qualidade, mas que esta classe de algoritmos de agrupamento de dados são computacionalmente custosos tanto em tempo quanto em espaço. (TAN; STEINBACH; KUMAR, 2005).

2.2.3 DBSCAN

O DBScan - *Density Based Spatial clustering*, ou agrupamento espacial baseado em densidade - é um algoritmo desenvolvido com o intuito de classificar dados em grandes bases de dados espaciais, cumprindo com três requerimentos:

1. Um mínimo de conhecimento do domínio, para que os parâmetros de entrada sejam definidos, porque os parâmetros ótimos geralmente não são conhecidos quando lida-se com uma grande base de dados;

2. Descobrimento de grupos de formatos variados, porque, em bases de dados espaciais, o formato dos grupos pode ser esférico, esticado, linear, alongado, etc;

3. Boa eficiência para grandes bases de dados, ou seja, bases de dados com mais do que alguns poucos mil dados. (ESTER et al., 1996).

Ainda segundo ESTER et al. (1996), nenhum dos algoritmos mais conhecidos de agrupamento de dados era capaz de cumprir com os três requerimentos citados.

Este método, por ser baseado em Densidade, utiliza uma abordagem baseada em centro, ou seja, dado um raio, descobre-se qual o número de itens próximos ao item que está sendo analisado, e classifica cada um dos pontos do conjunto de dados em uma de três categorias:

- Pontos de núcleo: dentro de um dos conjuntos, este dado está localizado no meio do grupo. Um dado é de núcleo se, através da abordagem baseada em centro, ele possuir um mínimo MinPts de pontos dentro de seu raio - chamado, pelos autores, de Eps -. O ponto A, na figura 2.6, ilustra um ponto de núcleo.

- Ponto de Borda: Um dado é considerado ponto de borda se, em seu raio, houver menos de MinPts pontos, e algum destes pontos for um Ponto de Núcleo. Ele ainda é considerado um elemento do grupo. O ponto B, na figura 2.6, ilustra um ponto de borda.

- Ruído: Se dentro do raio de um dado houver menos que MinPts pontos e nenhum destes pontos for um ponto de núcleo, o dado é considerado ruído, e não será agrupado. O ponto c na figura 2.6 ilustra um ponto de ruído.

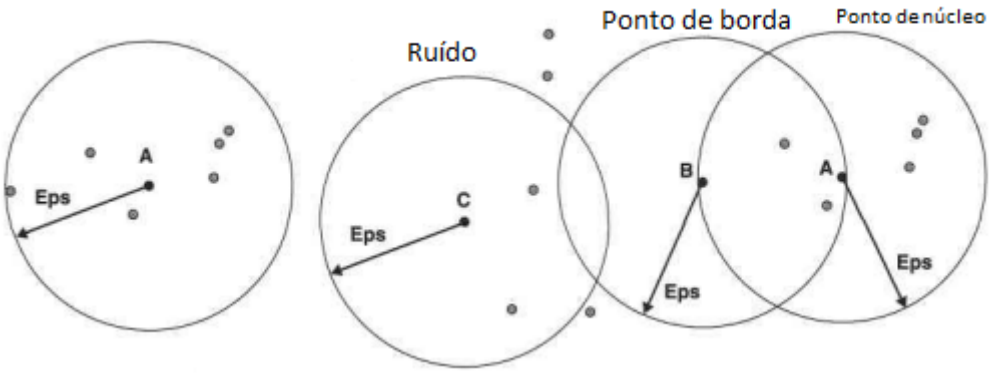

Figura 2.6 – Ilustração de três diferentes formas de classificação dos pontos em um grupo de dados. (TAN; STEINBACH; KUMAR, 2005)

Para a definição do MinPts, e do Eps, existem diversas abordagens. A mais comum é a verificação da distância de cada ponto até o seu k-ésimo vizinho, chamada de k-dist. Para

pontos que estão no mesmo grupo, k-dist será baixo desde que k seja menor que o tamanho deste grupo. Então, através dessa premissa, é possível descobrir qual o k que resulta na menor distância k-dist e esta distância será uma boa opção de Eps. Já um bom MinPts seria o próprio valor de k.

O algoritmo 3 é o método simplificado, desenvolvido com a lógica mostrada acima.

Tabela 2.3 – Algoritmo de agrupamento de dados por densidade básico.

Algoritmo 3. Algoritmo de agrupamento de dados por densidade básico
1: Classificar todos os pontos como pontos de núcleo, de borda ou ruído;
2: Eliminar pontos de ruído;
3: Marcar pontos de núcleo que estão dentro do raio um do outro;
4: Faça, de cada conjunto de pontos de núcleo interconectados, um grupo
5: Associe cada ponto de borda ao grupo de um dos pontos de núcleo que este ponto de borda está interligado.

Esta técnica, embora seja resistente à ruídos e capaz de encontrar grupos de formatos diferenciados, ainda tem alguns problemas para encontrar grupos que possuírem formatos muito diferenciados uns dos outros. Também foram notados problemas em agrupamentos de dados de alta dimensionalidade, principalmente por sua abordagem ser por densidade (TAN; STEINBACH; KUMAR, 2005).

2.2.4 Meanshift

O Mean Shift é um método não supervisionado e não paramétrico para estimar o gradiente de uma função de probabilidade, tendo dados amostrados (CHENG, 1995). Ele é utilizado não apenas para agrupamento de dados, mas também para reconhecimento de padrões e filtragem de dados. Esta técnica funciona como mostrado no algoritmo 4.

Tabela 2.4 – Algoritmo MeanShift básico.

Algoritmo 4. Algoritmo MeanShift básico
1: Fixar uma janela centrada em cada ponto do conjunto de dados;
2: O centro de massa (média dos pontos) é calculado para cada janela;
3: Com base na equação 2.2, o deslocamento para a média é calculado;
4: O centro da janela é deslocado para a média calculada no passo anterior;
5: O processo é repetido até que ocorra a convergência.

$$M_h(\overrightarrow{x}) = \frac{1}{n_{\overrightarrow{x}}} \sum_{\overrightarrow{x}_i \in S_h(\overrightarrow{x})} \overrightarrow{x}_i - \overrightarrow{x} \tag{2.2}$$

A equação 2.2 calcula o deslocamento do centro da janela. xi representa a média dos pontos dentro da hiperesfera S de raio h, centrada em x, contendo nx pontos.

Segundo CHENG (1995), se o tamanho da janela for muito grande, todos os pontos convergirão para uma única posição. Se for muito pequeno, e ela cobrir apenas o próprio ponto no qual ela está centralizada, existirá um grupo para cada ponto. Assim, se o tamanho da janela estiver entre os dois extremos, grupos diferentes, ou máximos locais serão encontrados, e estes serão os centros de cada grupo. A figura 2.7 (a) mostra um conjunto aleatório de pontos, enquanto as figuras 2.7(b, c, d, e, f) ilustram os diferentes resultados de agrupamento, quando diferentes tamanhos de janela são utilizados. Uma das principais características do MeanShift é

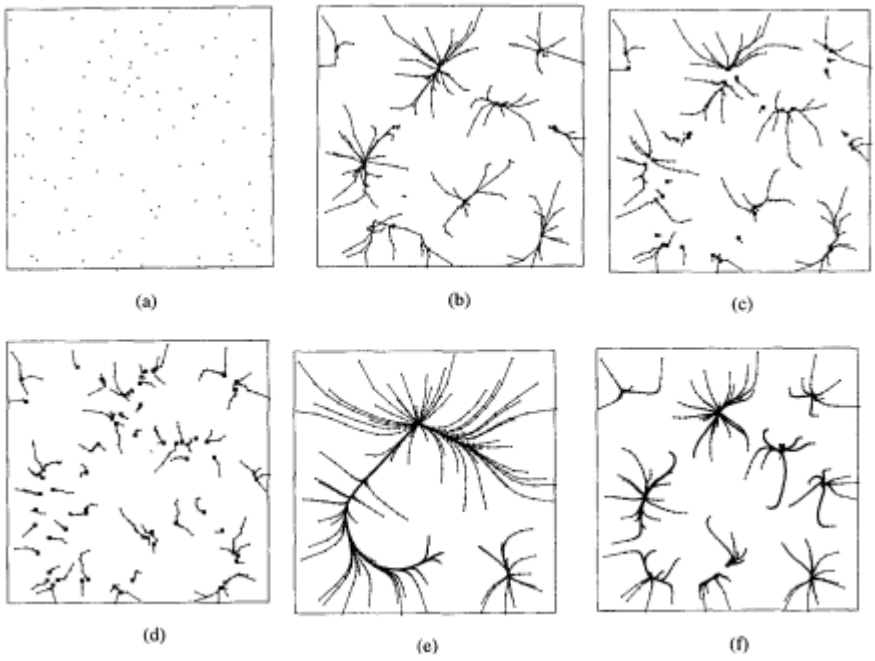

Figura 2.7 – Forma de agrupamento dos pontos de acordo com o tamanho da janela. (CHENG, 1995)

ser um algoritmo não-paramétrico. Porém, ele ainda precisa que se defina o tamanho da janela que será utilizada. Ele pode não funcionar bem com dados de alta dimensionalidade.

3 MOTIVAÇÃO E METODOLOGIA

Sabe-se que é em experimentos *in-silico* de docagem molecular se investiga e avalia o melhor encaixe e conformação de um ligante em uma cavidade do receptor, cujo resultado desse encaixe pode ser avaliado a partir de um valor de FEB liberado. Diferentes trabalhos foram desenvolvidos com o objetivo de minerar dados provenientes de docagem molecular e de dinâmica moleculr para tentar encontrar aquelas conformações de um receptor que sejam o mais promissoras para um dado ligante.

Apesar dos esforços de mineração de dados realizados sejam relevantes, Em WINCK (2012) é apresentado um algoritmo de indução de árvore de regressão, baseado em agrupamento dados, para melhor distinguir essas conformações, avaliando a proteína em nível atômico. Isso é, o algoritmo denominado 3D-Tri considera propriedades tridimensionais do receptor para predizer o valor de FEB.

3.1 Dataset tridimensional

O formato dataset utilizado neste trabalho, que foi proposto por WINCK (2012), contém as informações tridimensionais dos átomos de um receptor, de modo que, para cada átomo do receptor é identificada suas coordenadas tridimensionais no formato x, y, z. Este dataset tridimensional é assim formado:

- Cada 3 colunas do dataset indica uma coordenada espacial do átomo do receptor.
- Por considerar as coordenadas espaciais, o dataset pode ter uma dimesão de até 3 vezes a quantidade de átomos do receptor utilizado.
- Cada linha do dataset diz respeito a uma conformação do receptor obtida a partir de uma simulação de dinâmica molecular.
- As células do dataset dizem respeito à posição especial de cada átomo para uma determinada conformação
- Os datasets são gerados individualmente por ligante utilizado em experimentos de docagem. Nesse sentido, o último atributo do dataset contém o valor de FEB obtido do experimento de docagem molecular com o ligante que determina o dataset e com cada conformação que compõe o dataset.

A Tabela 3.1 ilustra uma estruturação do dataset tridimensional, onde as colunas indicam o coordenada espacial seguido do número do átomo sendo representado, bem como o nome, número do resíduo de aminoácido e a sigla do átomo a qual a coordenada pertence.

Tabela 3.1 – Exemplo de um dataset tridimensional. Fonte: (WINCK, 2012)

$x1$ ALA1_N	$y1$ ALA1_N	$z1$ ALA1_N	...	$x4008$ LEU268_OXT	$y4008$ LEU268_OXT	$z4008$ LEU268_OXT	FEB
15,838	-20,060	8,807	...	-20,647	-17,858	-3,495	-11,22
15,665	-19,974	7,918	...	-20,600	-17,957	-3,176	-11,21
...	...	...	...	...	...	...	...
14,959	-14,885	-18,370	...	21,498	16,662	-11,545	-1,00

3.2 Algortimo 3D-Tri

O algoritmo 3D-Tri busca induzir uma árvore de regressão a qual considera cada nodo da árvore como um átomo, e suas arestas buscam indicar se esse átomo está dentro ou fora de um intervalo ideal para este átomo, sendo representado pela posição inicial e final de suas coordenadas: $[(x_i, x_f)(y_i, y_f)(z_i, z_f)]$, i indica a posição inicial e f a posição final. A Figura 3.1 como nodo do algoritmo 3D é particionado.

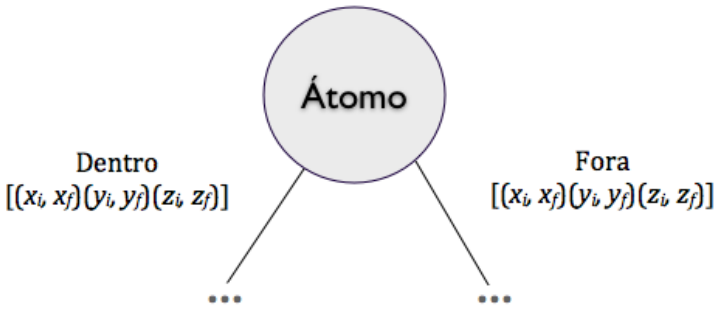

Figura 3.1 – Divisão de um nodo pelo pelo algoritmo proposto. Fonte: (WINCK, 2012)

O algoritmo 3D-Tri é composto de dois principais módulos: o primeiro busca encontrar o intervalo ideal para cada átomo enquanto o segundo refere-se à indução da árvore de regressão. Para a definição do intervalo ideal é proposta uma estratégia de agrupamento de dados.

O algoritmo de 3D-Tri propõe que o intervalo ideial para cada átomo seja identificado a partir de agrupamento de dados. Essa estratégia baseia-se, basicamente, na execução do algoritmo de K-means, com dois grupos (k=2) de modo que, para cada grupo gerado, é calculado o FEB médio dos objetos do grupo. A execução do algoritmo de agrupamento é efetuada para cada átomo do dataset tridimensional.

Ao selecionar o grupo com menor valor de FEB médio, gera-se uma estratégia de ranquiamento com a qual é possível identificar os menores e maiores valores para cada coordenada e, assim, determinar os seus limites iniciais e finais para a construção do intervalo $[(x_i, x_f)(y_i, y_f)(z_i, z_f)]$.

A proposta original do algoritmo 3D-Tri executa algoritmo de K-means com dois grupos para cada átomo. Entretanto, a escolha do algoritmo de K-means, bem como a escolha do número de grupos implementada, não é fundamentada como sendo a melhor escolha para o conjunto de dados tridimensionais. Este trabalho busca realizar uma avaliação de diferentes algoritmos de agrupamento para os dados tridimensionais e, assim, identificar aquele que melhor possa contribuir para uma futura geração do melhor intervalo atômico.

3.3 Metodologia

O objetivo deste trabalho se dá na criação de um framework de algoritmos básicos de agrupamento de dados, contendo métodos de lógica diferenciada na forma de resolução deste problema.

O trabalho de WINCK (2012) define o grupo de melhores conformações - isto é, imagens estáticas da interação entre receptor e ligante - através do algoritmo de k-means, com o parâmetro k definido em 2. Este trabalho também se propõe ao, na criação deste framework, realizar uma análise de forma a buscar, dentre os métodos selecionados, o método que melhor se destacasse na sua aplicação com os dados de conformação obtidos. Esse entendimento tem sua importância, pois busca-se a possibilidade de, em trabalhos futuros, desenvolver um algoritmo específico para este tipo de dados. A escolha de qual forma de agrupamento funciona melhor com este tipo de dados se torna, então, uma base diretiva.

Para isso, buscou-se primeiramente entender os tipos básicos de agrupamentos de dados e seus principais algoritmos. A aplicação em contexto geral dos mesmos foi levada em conta na escolha dos métodos, de forma que algoritmos que não funcionam comumente bem para dados semelhantes não tinham prioridade de desenvolvimento no framework proposto. Ao fim, os algoritmos de k-means - método particional -, o algoritmo de Ward - método hierárquico - e os algoritmos DBSCAN e MeanShift - métodos baseados em densidade - foram selecionados para desenvolvimento.

Após a escolha dos algoritmos, deu-se início aos seus desenvolvimentos. Buscou-se a melhor linguagem para que a programação e posteriores teste e análise fossem realizados. Por

lidar apenas com dados, estatísticas e cálculos matemáticos, sem necessidade de interfaces ou estruturas de dados mais complexas, decidiu-se por utilizar uma linguagem de script. Por questão de facilidade de escrita e leitura, estabilidade e quantia de usuários e bibliotecas disponíveis, a linguagem Python foi escolhida.

Definida a linguagem, buscou-se bibliotecas relacionadas a agrupamento de dados, para que os algoritmos pudessem ser executados com menos possibilidades de erros e maior generalidade na definição de parâmetros. A escolha do Scikit-Learn (PEDREGOSA et al., 2011) pareceu apropriada por se tratar de um conjunto de métodos extensivamente utilizados, simples, bem documentada e, também, estável.

Com os algoritmos desenvolvidos e os dados obtidos através de docagem molecular, foram realizadas repetidas execuções dos algoritmos com diferenciação em seus parâmetros de execução.

A análise se baseou no fato de que cada conformação possui um valor de FEB - energia livre de ligação, que define a força da interação entre as proteínas simuladas e, portanto, a força de atuação do ligante no receptor -. Este dado nos ajuda a definir a qualidade de cada conformação, pois quanto menor seu valor, melhor e mais importante a conformação se torna na análise da qualidade do efeito da droga testada no receptor. A média de FEB das conformações em cada um dos grupos foi realizada, e o grupo com o menor valor médio do dado é considerado o melhor grupo resultante do agrupamento.

Para que se fosse definido um algoritmo ou tipo de algoritmo de agrupamento de dados que melhor se qualificasse na aplicação desse formato de dados, procurou-se o grupo que, além de possuir o melhor valor de FEB médio, tivesse, em comparação com os grupos restantes, a maior diferença. Isso é, aquele que tivesse a maior diferença dos valores médio de FEB intergrupos. Isso se justifica pois o grupo que tiver não apenas um FEB médio maior, mas também mais distante dos demais, tende a ser mais relevante.

4 RESULTADOS

Neste trabalho, foi realizado um framework de algoritmos básicos de agrupamento de dados, contendo, respectivamente, os métodos de k-means, Ward, DBSCAN e MeanShift. Estes algoritmos foram desenvolvidos e sua eficiência foi analisada quando aplicados em dados reais, que representam a interação entre um receptor e um ligante. A figura 5.1 mostra o formato dos dados que foram utilizados.

SS	281	281	281	283	283	283	284	284	284	285	285	285
1	-1306	-1.57	7377	-732	-2715	6635	-985	-3.63	7169	-1407	-2775	5234
2	-1411	-1.38	7218	-909	-2521	6397	-1087	-3435	6964	-1.66	-2683	5.06
3	-1725	-1577	7549	-1171	-2689	6778	-1385	-3611	7318	-1846	-2789	5.38
4	-1547	-1324	7578	-1066	-2529	6977	-1284	-3363	7644	-1814	-2703	5632
5	-1774	-1591	7621	-1257	-2702	6782	-1.42	-3651	7295	-1912	-2763	5353
6	-1.48	-1581	7274	-1049	-2716	6451	-1325	-3607	7014	-1742	-2716	5114
7	-2042	-1885	7643	-1374	-2916	6845	-1525	-3841	7403	-1994	-2987	5376
8	-1903	-1712	7551	-1244	-2797	6757	-1439	-3737	7274	-1958	-2957	5392
9	-1473	-1943	7445	-993	-3041	6.66	-1282	-3986	7.12	-1642	-3106	5252
10	-1474	-1274	7523	-1018	-2353	6647	-1278	-3301	7118	-1663	-2381	5152

Figura 4.1 – Formato dos dados que foram utilizados

Na primeira linha da figura 4.1, temos a nomenclatura das informações que serão utilizadas a seguir: a primeira coluna conta com o número da conformação, ou seja, cada linha é uma conformação da interação entre o receptor e o ligante. A cada três colunas, então, temos o mesmo número sendo repetido. Este é o número identificador de um átomo das proteínas, em que a primeira coluna possui o valor no eixo x, a segunda coluna, o valor no eixo y, e a terceira, o valor no eixo z.

Os algoritmos citados foram selecionados por serem usados para conjuntos de dados diferenciados. Cada um possui uma lógica de agrupamento diferente, e dessa forma é possível analisar quais os tipos de agrupamento funcionam melhor com o tipo de dados que obtemos.

Ao analisar a eficácia destes algoritmos, uma técnica foi definida: cada algoritmo deve ser aplicado ao conjunto de dados com diferentes configurações de seus parâmetros iniciais, conforme segue:

- Para o algoritmo de k-means, entre as execuções mudou-se a quantidade de grupos formados.
- Para o algoritmo de Ward, o critério de parada - que, também, é o número de grupos formados.
- Para o DBSCAN, o EPS e o número mínimo de pontos.

- Para o MeanShift, não houve alterações. O algoritmo é capaz de calcular automaticamente qual o melhor raio para otimização dos grupos.

Cada execução do algoritmo resulta na quantidade de grupos criados e no FEB médio de cada grupo. O FEB é o valor que define a qualidade das ligações entre o ligante e o receptor para cada conformação e, portanto, é o parâmetro central da nossa análise de eficiência. Quanto menor o valor de FEB, melhor é a interação entre as duas proteínas e, portanto, maior ou melhor serão os efeitos desejado daquela interação. A figura 4.2 mostra um exemplo do formato de saída da execução dos algoritmos

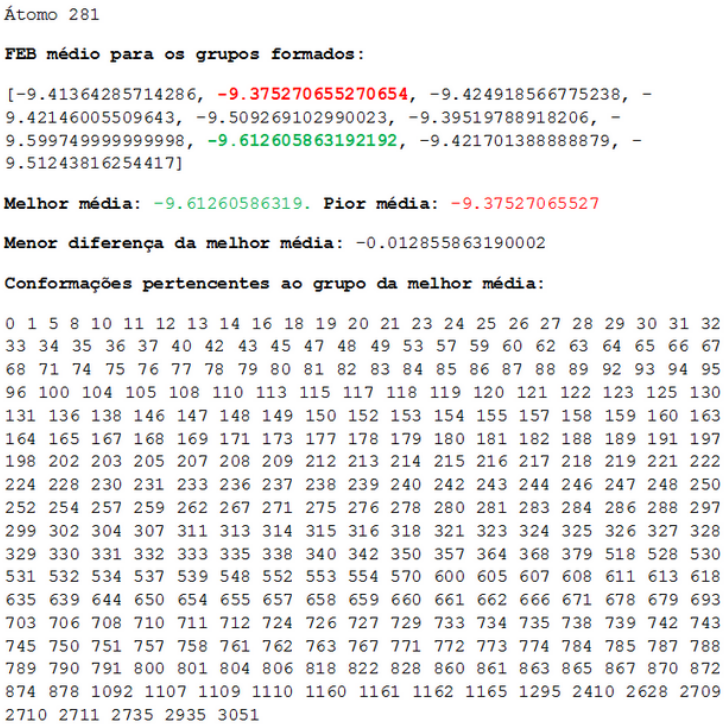

Figura 4.2 – Ilustração do formato de saída da execução dos algoritmos desenvolvidos.

Os algoritmos foram desenvolvidos em Python, versão 2.7, utilizando uma biblioteca de aprendizado de máquina chamada Scikit-learn.

4.1 Resultados de execução: k-means

A princípio, buscava-se mudar não só o número de grupos formados pelo algoritmo, mas também a inicialização de seus centróides. Como é conhecido, a inicialização aleatória dos centróides pode resultar na formação de grupos totalmente diferentes a cada execução do algoritmo, o que traz uma aleatoriedade não desejada.

Mas ao executar os primeiros testes, com qualquer quantia de grupos formados, percebeu-se que, embora o valor de FEB médio mudasse - e, portanto, os elementos contidos em cada grupo também mudassem -, ao verificar o grupo contendo o melhor valor médio de FEB, podia-se perceber que a grande maioria de seus membros era constante. Isto pode ser verificado através da tabela 4.1, que mostra execuções do algoritmo para cinco e dez grupos, respectivamente. Execuções com mais de dez ou menos de cinco grupos resultaram em grupos contendo os melhores valores médios de FEB quando comparados com execuções para outros valores para k. f

Como resultado da análise, percebeu-se que o k-means consegue, através de múltiplas execuções e posterior análise estatística dos elementos do melhor grupo encontrado, convergir em um bom grupo de conformações. A figura 4.3 ilustra em três dimensões o resultado do agrupamento do algoritmo de k-means também para o átomo 281.

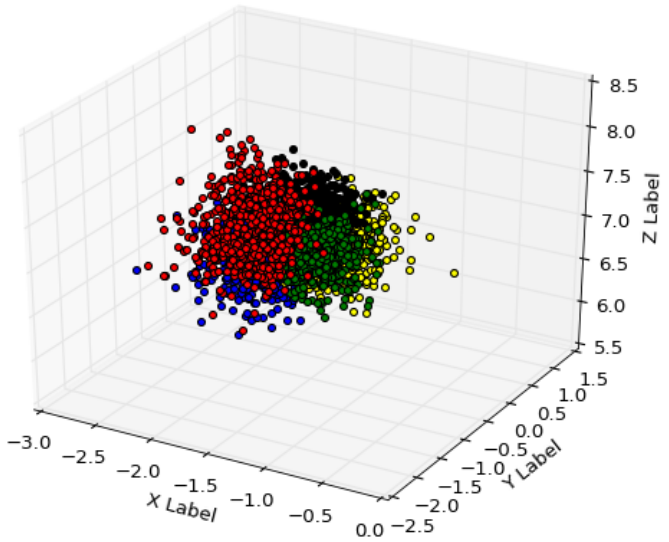

Figura 4.3 – Resultado de uma execução do algoritmo k-means com k = 5, onde o grupo com menor FEB médio é o colorido em vermelho.

4.2 Resultados de execução: algoritmo de Ward.

O algoritmo de Ward não se baseia em inicialização aleatória de centróides, e sim na junção de grupos cujo centróide está mais próximo e posterior ajuste do mesmo. Portanto, não apresenta aleatoriedade nos resultados para cada critério de parada selecionado. A tabela 4.2 mostra resultados das execuções do algoritmo de Ward. Para a formação de um número de grupos menor que cinco ou maior que vinte, o resultado de FEB médio não trazia resultados

satisfatórios.

Como resultado da análise, percebeu-se que, embora sua melhor execução - em relação à diferença do melhor FEB médio com outros - ainda não seja comparável com os resultados dos outros algoritmos, concluiu-se que algoritmos hierárquicos provavelmente não lidam bem com o formato de dados utilizado. Na figura 4.4, representamos a execução do algoritmo para que o critério de parada seja a formação de dez grupos.

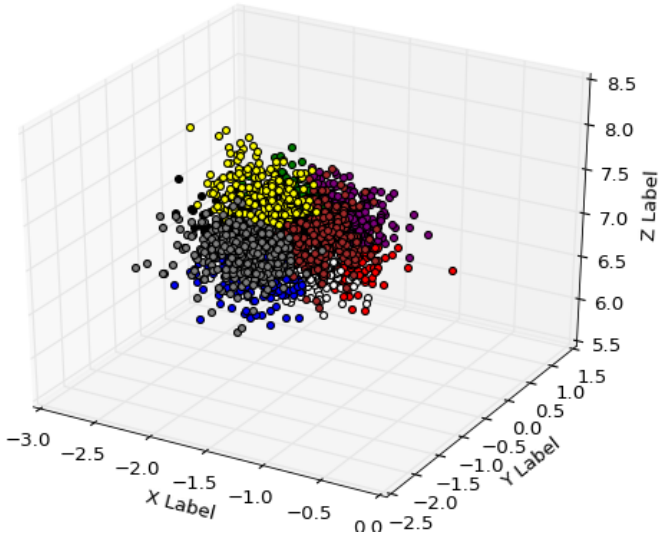

Figura 4.4 – Resultado de uma execução do algoritmo Ward com 10 grupos, onde o grupo com menor FEB médio é o colorido em cinza.

4.3 Resultados de execução: algoritmos DBSCAN e MeanShift.

A princípio, a análise dos algoritmos foi feita a partir dos resultados da sua execução para o primeiro átomo disponível, tendo como parâmetros modificados o eps, - o raio que parte de cada ponto formado pelos dados no conjunto de dados - e o número mínimo de pontos, que é o que delimita quais pontos serão classificados como pontos de núcleo ou de borda. A primeira execução do DBSCAN sobre este átomo trouxe resultados precários, principalmente porque, já que buscamos a área do espaço vetorial onde, para cada átomo, ele se encontra com mais frequência, produzindo um melhor resultado de FEB, algoritmos baseados em Densidade pareciam a melhor opção de tipo de agrupamento de dados a ser feita. Embora, para este determinado átomo, o algoritmo conseguisse separar grupos com valores médios de FEB muito altos, a distinção do melhor grupo para o segundo melhor valor era relativamente baixa - pouco

mais de 0.1 de diferença, nos melhores casos -, além de gerar grupos com pouquíssimas conformações. A tabela 4.3 mostra os resultados obtidos com diferentes configurações dos parâmetros citados.

A análise do MeanShift foi realizada de forma não-paramétrica, pois o algoritmo é capaz de calcular automaticamente qual o tamanho da janela que trará os melhores resultados de agrupamento. Por ser não paramétrico, o algoritmo calculou a janela ótima para cada um dos átomos do conjunto de dados, gerando, para cada átomo, um número diferente de grupos formados. Isso evidenciou uma característica importante destes dados: visto que cada conformação é uma imagem estática da interação e movimento das duas proteínas - receptor e ligante - através do tempo, certos átomos - os que geravam menos grupos, nos testes com o MeanShift -, provavelmente possuíam movimentação uniforme e constante, totalmente aleatória ou parcialmente aleatória, resultando na dificuldade de distinguir grupos dado um valor de janela muito baixo - seja definido como parâmetro no DBSCAN ou automaticamente, no MeanShift -, enquanto outros átomos - que geravam dois ou mais grupos no MeanShift - se movimentavam mantendo uma constância em determinadas posições do espaço. A tabela 4.4 mostra os resultados do MeanShift, e a figura 4.5 mostra o resultado do agrupamento para o átomo que obteve melhor resultado durante sua execução.

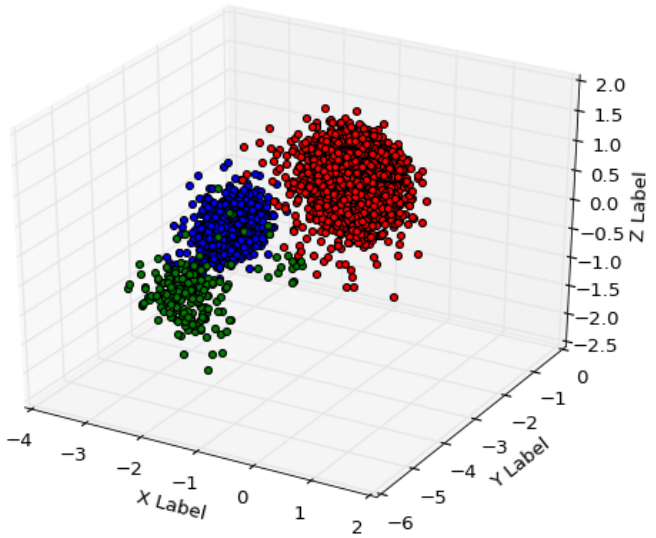

Figura 4.5 – Resultado de uma execução do algoritmo MeanShift para o átomo 2192. Ele conta com 3 grupos, onde o grupo com menor FEB médio está colorido em vermelho.

Descoberta esta propriedade, o algortimo DBSCAN foi mais uma vez utilizado, desta vez no átomo que obteve melhor resultado de agrupamento nos testes do MeanShift. Os valores de FEB médio dos grupos gerados, bem como os elementos de cada grupo podem ser vistos na

tabela 4.5.

Embora não tenha gerado o maior FEB médio de todas as execuções do algoritmo, os resultados do novo teste do DBSCAN geraram maior diferença entre o grupo com o melhor valor médio de FEB comparado à execução anterior, mas ainda menor que a gerada pelo MeanShift. Além disso, ao verificar a composição dos grupos, percebe-se que grande parte das conformações se encontram no grupo com o maior valor de FEB médio. A figura 4.6 mostra a composição dos grupos em forma de gráfico.

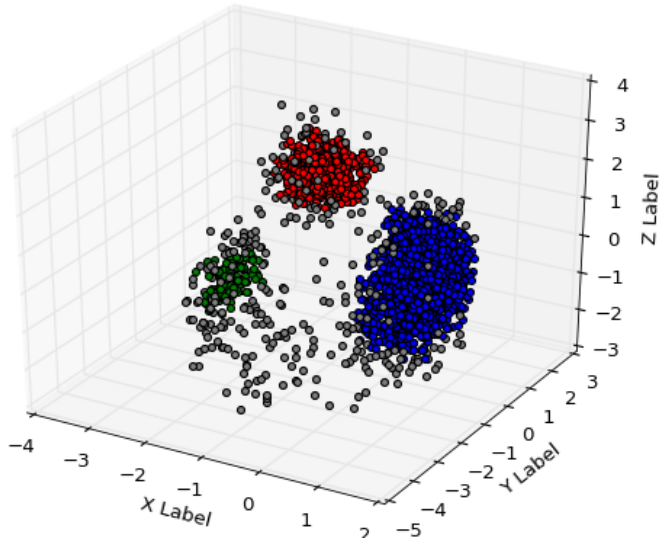

Figura 4.6 – Resultado de uma execução do algoritmo DBSCAN para o átomo 2191. Ele conta com 4 grupos, onde o grupo com menor FEB médio está colorido em azul, e os valores considerados ruído estão coloridos em cinza.

4.4 Considerações

Os experimentos realizados nos permitiram ter uma visão mais abrangente quanto ao funcionamento dos algoritmos sobre os dados tridimensionais. É importante destacar os resultados dos algoritmos por densidade, de onde identificamos que existem ätomos cujo movimento é mais livre e átimos cujo movimento se concentra em algumas áreas do espaço. Além disso foi possível, pelo algoritmo DBSCAN, identificar os grupos de átomo que tem um bom valor dFEB médio muito alto para aqueles de movimento mais livre, mas com uma diferença para os demais grupos baixa, com um número reduzido de conformações no melhor grupo.ara átomos de movimento limitado, é capaz de trazer resultados muito bons, com grupos de tamanho considerável. algoritmo MeanShift trabalha muito bem com átomos de movimento limitado, resultando na

maior distância de FEB médio do melhor grupo em relação aos demais, mas não é capaz de trabalhar com átomos com movimento livre - para estes átomos. Assim, descobrimos que os algoritmos de agrupamento de dados baseados em Densidade são os que trazem resultados mais satisfatórios quando aplicados a este tipo de dados.

Tabela 4.1 – Exemplo de execução do algortimo k-means para k = 10 e k = 5, do átomo 281.

k	Código do átomo	Menor média de FEB	Maior média de FEB	Menor diferença de FEB médio	Número de conformações no grupo
10	281	-9.61260586319	-9.37527065527	-0.012855863190002	309
10	281	-9.62243589744	-9.37257142857	-0.01708706023	313
10	281	-9.61996453901	-9.36577342048	-0.0313701615	283
10	281	-9.59024193548	-9.59024193548	-0.00230880478	248
10	281	-9.67786259542	-9.37247787611	-0.11193039203	132
5	281	-9.5987537092	-9.39943538269	-0.04727946408	338
5	281	-9.61014354067	-9.3965323993	-0.13050498645	628
5	281	-9.59968609865	-9.37936378467	-0.11848454826	670
5	281	-9.58853608247	-9.58853608247	-0.00433813025	486
5	281	-9.59885496183	-9.37078212291	-0.12156238541	656

Tabela 4.2 – Execução do algortimo de Ward para parada com 10, 5 e 20 grupos, do átomo 281.

Critério de parada	Átomo	Menor FEB médio	Maior FEB médio	Menor diferença de FEB médio	Número de Conformações no grupo
10 grupos	281	-9.59829015544	-9.38603053435	-0.02596872686	387
5 grupos	281	-9.58875409836	-9.4109223301	-0.10961036747	611
20 grupos	281	-9.66309090909	-9.36435582822	-0.04451948051	56

Tabela 4.3 – Exemplo de execução do algortimo DBSCAN com modificações nos parâmetros eps e minsamples, para o átomo 281.

Parâmetros	Átomo	Menor FEB médio	Maior FEB médio	Menor diferença de FEB médio	Número de conformações no grupo
eps=0.3, minsamples=5	281	-9.806	-9.15714285714	-0.076	5
eps=0.4, minsamples=5	281	-9.77666666667	-9.45689751661	-0.13166666667	3
eps=0.35, minsamples=5	281	-9.76571428571	-9.45205893052	-0.06904761904	7
eps=0.35, minsamples=10	281	-9.721	-9.365	-0.12655555555	10
eps=0.32, minsamples=20	281	-9.61324324324	-9.39750237417	-0.16051597051	37

Tabela 4.4 – Exemplo de execução do algortimo MeanShift, mostrando os átomos que o algoritmo criou três ou mais grupos

Átomo	Número de grupos criados	Menor FEB médio	Maior FEB médio	Menor diferença de FEB médio
293	3	-9.66370833333	-9.43116840114	-0.17813786353
2191	4	-9.51630155211	-9.30892491468	-0.17117334698
2192	3	-9.51259627867	-9.30914383562	-0.20345244305
2412	3	-9.692	-9.42347361809	-0.19067292225
2904	3	-9.57457052797	-9.37754843019	-0.15952548292
2911	3	-9.49958139535	-9.42139079334	-0.03573294071

Tabela 4.5 – Execução do algoritmo DBSCAN para o átomo 2191, considerado pelo MeanShift o átomo que agrupava seus movimentos em quatro áreas.

Parâmetros	Átomo	Menor FEB médio	Maior FEB médio	Menor diferença de FEB médio	Número de conformações no grupo
eps=0.3, min$_samples = 5$	2191	-9.51659854678	-9.30608534323	-0.02659854678	2203
eps=0.4, min$_samples = 5$	2191	-9.51539724811	-9.304	-0.04289724811	2254
eps=0.35 min$_samples = 10$	2191	-9.51473181818	-9.30723048327	-0.18871486902	2236
eps=0.35 min$_samples = 20$	2191	-9.51543018336	-9.30599182004	-0.17989172182	2128

5 CONCLUSÃO

Já era esperado que não seria encontrado um algoritmo básico capaz de encontrar o grupo ótimo de conformações. Através das análises, no entanto, descobrimos que os resultados do algoritmo de k-means foram melhores que o esperado, enquanto o algoritmo de Ward, de técnica hierárquica, teve a avaliação menos satisfatória. Embora o DBSCAN, executado sozinho, consiga gerar grupos com valores médios de FEB extremamente altos, com maior consistência de bons resultados, mesmo em átomos cujo movimento não é limitado mas com um pequeno número de conformações, e o MeanShift, executado sozinho, consiga gerar grupos com a melhor das menores distâncias de FEB médio, a junção da execução do algoritmo MeanShift para descobrir quais átomos possuem uma constância em sua movimentação, com posterior execução do MeanShift foi capaz de eleger um grande grupo de conformações úteis, as separando das demais. A tabela 5.1 mostra as conclusões que foram tomadas de acordo com a análise das seções 4 e 5.

Ainda assim, através dos testes foi descoberta uma propriedade interessante da movimentação dos átomos do conjunto de dados: alguns se mantém, em grande parte, em determinadas áreas do espaço, enquanto outros têm movimentação constante ou aleatória. Os que se mantém em determinadas áreas do espaço são melhor agrupados pelo algoritmo MeanShift, enquanto os demais parecem ter melhor resultado com a aplicação do DBSCAN. Ademais, essa propriedade abre espaço para uma discussão extremamente interessante: Seriam essas áreas do espaço frequentemente habitadas pelos átomos descobertos na execução do MeanShift prováveis sítios ativos de ligação?

Tabela 5.1 – Resumo dos resultados da análise dos algoritmos de acordo com sua eficácia na aplicação dos dados apresentados

Algoritmo	Média de FEB	Maior consistência	Distâncias de FEB médio	Conformações no grupo
k-means				
Ward				
DBSCAN	X	X		
MeanShift			X	
DBSCAN em átomos importantes				X

REFERÊNCIAS

BALAKIN, K. Pharmaceutical data mining: approaches and applications for drug discovery. **New York: John Wiley & Sons**, [S.l.], 2009.

CHENG, Y. Mean shift, mode seeking, and clustering. **IEEE Trans. Pattern Anal. Mach. Intell. 17 (8): 790–799**, [S.l.], 1995.

ESTER, M. et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. **Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)**, [S.l.], 1996.

GUNSTEREN, W. van; BERENDSEN, H. Computer simulation of molecular dynamics: methodology, applications, and perspectives in chemistry. **Angewandte Chemie International Edition in English, vol. 29, pp. 992–1023.**, [S.l.], 1990.

MORGANKAUFAMN (Ed.). **Data Mining**: concepts and techniques. 2nd.ed. [S.l.: s.n.], 2006.

HARTIGAN, J. A.; WONG, M. A. A K-means clustering algorithm. **Applied Statistics, vol.28, pp.100–108.**, [S.l.], 1979.

H.M. BERMAN, H. et al. PDB - Protein Data Bank. **Nucleic Acids Research, vol. 28, 2000, pp. 235-242**, [S.l.], 2000.

IRWIN, J.; SHOICHET, B. ZINC – a free database of commercially available compounds for virtual screening. **Journal of Chemical Information and Modeling, 45, 1, 177–182**, [S.l.], 2005.

KUNTZ, I. D. Structure-based strategies for drug design and discovery. **Science, vol. 257, pp. 1078-1082**, [S.l.], 1992.

LESK, A. Introduction to Bioinformatics. **New York: Oxford University Press, pp. 320**, [S.l.], 2002.

LIN, J.-H. et al. Computational drug design accommodating receptor flexibility: the relaxed complex scheme. **Journal of American Chemical Society. 124:5632-5633**, [S.l.], 2002.

LUSCOMBE, N.; GREENBAUM, D.; GERSTEIN, M. What is bioinformatics? a proposed definition and overview of the field. **Methods Information in Medicine, vol. 40, 2001, pp. 346–358.**, [S.l.], 2001.

LYBRAND, T. P. Ligand-protein docking and rational drug design. **Structural Biology, vol. 5, 1995, pp. 224–228**, [S.l.], 1995.

MACHADO, K. et al. Mining flexible-receptor docking data. **WIREs Data Mining and Knowledge Discovery, vol. 1, 2011, pp. 532-541**, [S.l.], 2011.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. **5-th Berkeley Symposium on Mathematical Statistics and Probability**, [S.l.], 1975.

MORRIS, G. M. et al. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. **Journal of Computational Chemistry, vol. 19, 1998, pp. 1639–1662**, [S.l.], 1998.

PEDREGOSA, F. et al. Scikit-learn: machine learning in Python. **Journal of Machine Learning Research**, [S.l.], v.12, p.2825–2830, 2011.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. Introduction to Data Mining. **Boston: Addison Wesley, 2nd Edition, 769.**, [S.l.], 2005.

WARD, J. H. J. Hierarchical Grouping to Optimize an Objective Function. **Journal of the American Statistical Association, vol. 58, n.301, pp.236-244.**, [S.l.], 1963.

WINCK, A. **3D-Tri**: um algoritmo de indução de Árvore de regressão para propriedades tridimensionais - um estudo sobre dados de docagem molecular considerando a flexibilidade do receptor. 2012. Tese (Doutorado em Ciência da Computação) — Pontífica Universidade Católica do Rio Grande do Sul.

WINCK, A. et al. Supporting intermolecular interaction analyses of flexible-receptor docking simulations. **IADIS International conference of Applied computing, 2010, pp. 183-190.**, [S.l.], 2010.