UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

EDUARDO SPIELER DE OLIVEIRA

Um Algoritmo Genético de Chaves Aleatórias Viciadas para o problema de Atracamento Molecular

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Orientador: Prof. Dr. Márcio Dorn

Porto Alegre
2016

CIP — CATALOGAÇÃO NA PUBLICAÇÃO

Spieler de Oliveira, Eduardo

Um Algoritmo Genético de Chaves Aleatórias Viciadas para o problema de Atracamento Molecular / Eduardo Spieler de Oliveira. – Porto Alegre: PPGC da UFRGS, 2016.

98 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2016. Orientador: Márcio Dorn.

1. Atracamento Molecular. 2. Otimização. 3. Algoritmo Genético. 4. Docking. I. Dorn, Márcio. II. Título.

AGRADECIMENTOS

Agradeço aos meus pais e ao meu orientador.

Um Algoritmo Genético de Chaves Aleatórias Viciadas para o problema de Atracamento Molecular

RESUMO

O Atracamento Molecular é uma importante ferramenta utilizada no descobrimento de novos fármacos. O atracamento com ligante flexível é um processo computacionalmente custoso devido ao número alto de graus de liberdade do ligante e da rugosidade do espaço de busca conformacional representando a afinidade entre o receptor e uma molécula ligante. O problema é definido como a busca pela solução de menor energia de ligação proteína-ligante. Considerando uma função suficientemente acurada, a solução ótima coincide com a melhor orientação e afinidade entre as moléculas. Assim, o método de busca e a função de energia são partes fundamentais para a resolução do problema. Muitos desafios são enfrentados para a resolução do problema, o tratamento da flexibilidade, algoritmo de amostragem, a exploração do espaço de busca, o cálculo da energia livre entre os átomos, são alguns dos focos estudados. Esta dissertação apresenta uma técnica baseada em um Algoritmo Genético de Chaves Aleatórias Viciadas, incluindo a discretização do espaço de busca e métodos de agrupamento para a multimodalidade do problema de atracamento molecular. A metodologia desenvolvida explora o espaço de busca gerando soluções diversificadas. O método proposto foi testado em uma seleção de complexos proteína-ligante e foi comparado com softwares existentes: AutodockVina e Dockthor. Os resultados foram estatisticamente analisados em termos estruturais. O método se mostrou eficiente quando comparado com outras ferramentas e uma alternativa para o problema de Atracamento Molecular.

Palavras-chave: Atracamento Molecular. Otimização. Algoritmo Genético. Docking.

A Biased Random Key Genetic Algorithm for the Molecular Docking problem

ABSTRACT

Molecular Docking is a valuable tool for drug discovery. Receptor and flexible Ligand docking is a very computationally expensive process due to a large number of degrees of freedom of the ligand and the roughness of the molecular binding search space. A Molecular Docking simulation starts with a receptor and ligand unbounded structures and the algorithm tests hundreds of thousands of ligands conformations and orientations to find the best receptor-ligand binding affinity by assigning and optimizing an energy function. Despite the advances in the conception of methods and computational strategies for search the best protein-ligand binding affinity, the development of new strategies, the adaptation, and investigation of new approaches and the combination of existing and state-of-the-art computational methods and techniques to the Molecular Docking problem are clearly needed. We developed a *Biased Random-Key Genetic Algorithm* as a sampling strategy to search the protein-ligand conformational space. The proposed method has been tested on a selection of protein-ligand complexes and compared with existing tools AutodockVina and Dockthor. Compared with other traditional docking software, the proposed method has the best average Root-Mean-Square Deviation. Structural results were statistically analyzed. The proposed method proved to be efficient and a good alternative to the molecular docking problem.

Keywords: Molecular Docking. Optimization. Genetic Algorithm.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

DRFBE	Desenho Racional de Fármacos Baseado em Estrutura
BRKGA	*Biased Random-Key Genetic Algorithms*
AM	Atracamento Molecular
VS	*Virtual Screening*
CAPRI	*Critical Assessment of Prediction of Interactions*
EBI	*European Bioinformatics Institute*
PDB	*Protein Data Bank*
AG	Algoritmo Genético
AGL	Algoritmo Genético Lamarckiano
SSGA	*Steady State Genetic Algorithm*
EG	Evolução Diferencial
AM	Algoritmos Meméticos
SA	*Simulated Annealing*
OEP	Otimização por Enxame de Partículas
RMSD	*Root-mean-square deviation*
GPU	*Graphics Processing Unit*
ILS	*Iterated Local Search algorithm*
DMRTS	*Dynamic Modified Restricted Tournament Selection*
RMN	Ressonância Magnética Nuclear

SUMÁRIO

1 INTRODUÇÃO

Luscombe et al. (2001) descreve os principais objetivos da Bioinformática como sendo: a organização dos dados de uma maneira que permita que pesquisadores tenham um fácil acesso a estes dados e possam submeter novas entradas produzidas; o desenvolvimento de ferramentas e recursos que auxiliem os pesquisadores na análise destes dados; e ainda, o uso destas ferramentas computacionais para analisar dados e interpretar os resultados. A Bioinformática Estrutural pode ser apresentada como a área de estudo da estrutura de moléculas, tais como: DNA, RNA, proteínas, entre outros compostos, e tem foco na representação, armazenamento, recuperação, análises e exibição de informações estruturais de macromoléculas biológicas (ALTMAN; DUGAN, 2003).

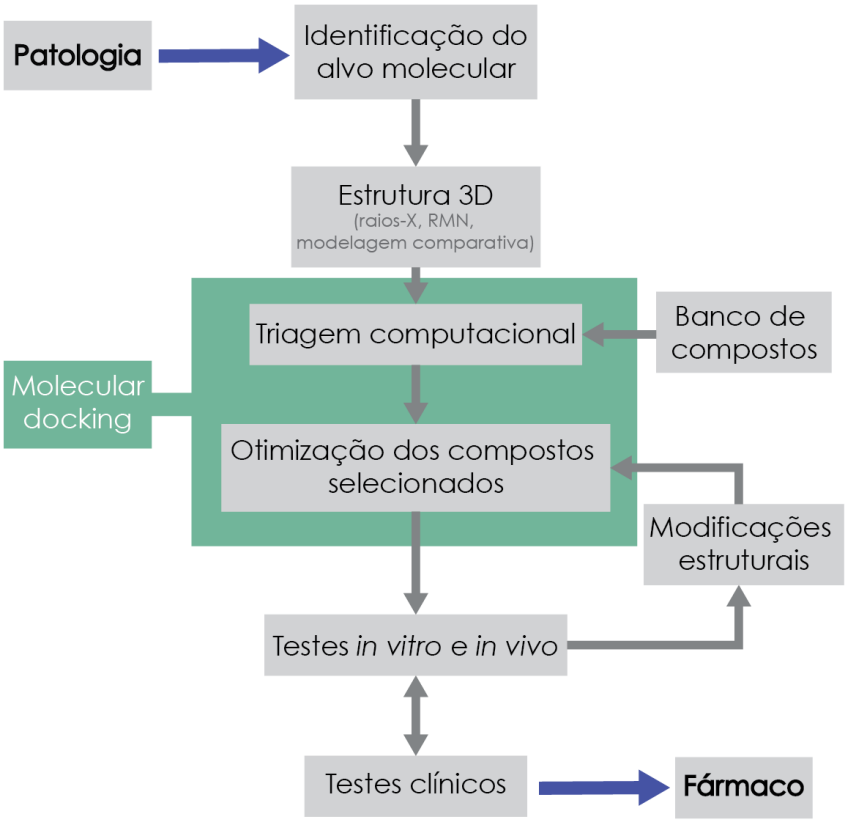
Um dos principais desafios da Bioinformática Estrutural é conhecido como o problema de Atracamento Molecular. O problema consiste em encontrar a orientação entre uma molécula ligante e uma molécula receptora que apresente a menor energia de interação (CAMACHO et al., 2014). Ferramentas de Atracamento Molecular objetivam a busca de um modelo que descreva a interação entre duas estruturas moleculares. O grau de dificuldade do problema está associado as estruturas moleculares, considerando o grande número de ângulos internos e mínimos locais no espaço (KUNTZ, 1992). O desenvolvimento de métodos e estratégias computacionais para o atracamento guiam o uso dessa técnica como uma ferramenta para o descobrimento de novos compostos químicos (fármacos) (BROOIJMANS; KUNTZ, 2003a). As principais estruturas utilizadas como receptor são proteínas. Proteínas ou polipeptídios são polímeros formados por 20 diferentes tipos de resíduos de aminoácidos que são ligados através de uma ligação peptídica (LESK, 2005). Cada proteína é definida por sua sequência única de resíduos de aminoácidos que em condições fisiológicas se enovelam em uma forma específica conhecida como estado nativo (ANFINSEN, 1973). São estruturas fundamentais para o organismo, suas funções variam desde construção de novos tecidos do corpo humano, transporte de substâncias, atuação no sistema de defesa do organismo, catalização de reações químicas, regulação de hormônios, entre outros.

Ferramentas de Atracamento proteína-ligante são atualmente importantes metodologias para o descobrimento de novos fármacos (SOUSA et al., 2013). Conhecer a forma 3D da proteína e do ligante implica na inferência de sua função. Todos esses fatores, aliados a necessidade de uma metodologia mais sistemática que objetive o estudo dos mecanismos envolvidos no processo de reconhecimento molecular, impulsionaram o

surgimento do paradigma conhecido como Desenho Racional de Fármacos Baseado em Estrutura (DRFBE *Structure-based Rational Drug Design*) (KUNTZ, 1992). O DRFBE é definido como o estudo de estruturas moleculares tridimensionais da molécula receptora para o desenho de compostos protótipos (moléculas ligantes candidatas a fármaco) tomando como base as informações estruturais e as interações envolvidas no processo de reconhecimento molecular receptor-ligante (MAGALHAES, 2006).

No DRFBE são utilizadas diferentes metodologias de Atracamento Molecular proteína-ligante, tanto para descoberta de novas substâncias bioativas, como para o refinamento e otimização de compostos bioativos previamente identificados. A Figura 1.1 apresenta as etapas dessa metodologia.

Figura 1.1: Etapas do desenho racional de fármacos baseado em estrutura; em verde são ilustradas as etapas onde aplicam-se ferramentas de Atracamento Molecular



Primeiramente é realizada a escolha adequada do alvo terapêutico, proteína ou enzima, relacionada à patologia, cuja função deve ser bloqueada ou ativada. É necessário, então, obter a estrutura molecular tridimensional do bioreceptor. Estruturas moleculares podem ser obtidas através de técnicas experimentais como a difração por raios-X e Ressonância Magnética Nuclear (RMN), ou por métodos computacionais como a modelagem

comparativa (CHOTHIA; LESK, 1992). Atualmente existem grandes bancos de estruturas moleculares de acesso público, como o *Protein Data Bank* (PDB) (BERMAN et al., 2000), que é um dos mais importantes banco de dados de estruturas 3-D de proteínas, com aproximadamente 120 mil[1] dados, e diversos bancos de estruturas moleculares de ligantes, como o Cambridge Structural Database (CSD), com aproximadamente 700 mil [2] estruturas e o ZINC (IRWIN; SHOICHET, 2005), com 35 milhões [3] de compostos.

Após obter a estrutura tridimensional da molécula, grandes bancos de estruturas são testados contra o alvo molecular, utilizando metodologias como a Triagem Virtual, a fim de identificar compostos biologicamente ativos candidatos a novos fármacos. Assim que um composto promissor é encontrado, são realizadas modificações na molécula, visando aumentar a resposta biológica desejada, e, também, para especificar a molécula para um determinado alvo molecular e para que se ajuste às características farmacocinéticas: absorção, distribuição, metabolismo e eliminação (NOLTING et al., 1996). Durante essa fase são utilizados metodologias mais acuradas de Atracamento Molecular, visando a identificação da conformação de ligações entre as moléculas selecionadas e a otimização dos compostos em termos de conformações químicas. Por fim, são realizados testes *in vitro* (em laboratório) e *in vivo* (em seres vivos) para que características como toxidez sejam analisadas (ROGERO et al., 2003).

O Atracamento Molecular pode ser descrito como um problema de otimização. Nesse problema busca-se prever a orientação de ambas as moléculas quando ligadas quimicamente de forma que esse complexo seja energicamente estável (GODOY et al., 2015). Devido ao grande número de orientações e conformações que ambos os complexos podem assumir, o uso de métodos determinísticos de otimização levaria um elevado tempo de execução, tornando inviável o uso dessas técnicas em um baixo tempo. Assim, para encontrar a conformação ótima entre o receptor e o ligante, com a menor energia, são utilizadas heurísticas e métodos computacionais inspirados na natureza, como, por exemplo, algoritmos evolutivos. Muitas metodologias e algoritmos foram propostos ao longo dos anos para tentar solucionar o problema de atracamento molecular proteína-ligante. Souza et al. (2013) apresenta diversos *softwares*, metodologias e parametrizações que foram desenvolvidos nos últimos 10 anos de pesquisa na área. Um dos maiores desafios computacionais é lidar com a alta flexibilidade dos complexos, incluindo os graus de liberdade dos átomos da estrutura. Essas diferentes abordagens são divididas, tradici-

[1] Em outubro de 2016 PDB http://www.rcsb.org
[2] Em outubro de 2016 CSD http://www.ccdc.cam.ac.uk/
[3] Em outubro de 2016 http://zinc.docking.org/

onalmente, em (KUNTZ, 1992): métodos de receptor-ligante rígidos, métodos de ligante flexível, e métodos de receptor e ligante flexíveis.

No método de receptor-ligante rígidos são considerados apenas os graus de liberdade translacionais e rotacionais da molécula ligante e do receptor. Atualmente, a maioria das ferramentas de Atracamento incluem a flexibilidade dos ângulos internos do ligante considerando, além dos graus de liberdade translacionais e rotacionais, os graus de liberdade conformacionais (MAGALHAES, 2006). Em ambas abordagens o receptor é considerado rígido, a estrutura da proteína é fixa na posição da estrutura determinada experimentalmente. Diversos estudos vem sendo realizados para a inclusão de flexibilidade do receptor (MACHADO et al., 2011; TEODORO; KAVRAKI, 2003; COZZINI et al., 2008; HUANG; ZOU, 2007; WONG, 2008; ALONSO; BLIZNYUK; GREADY, 2006; CHANDRIKA; SUBRAMANIAN; SHARMA, 2009; PANG; KOZIKOWSKI, 1994), porém, mesmo com a abordagem de receptor rígido, o docking de estruturas de ligantes grandes e altamente flexíveis é um grande desafio para esses algoritmos.

Há duas partes fundamentais no desenvolvimento de ferramentas de Atracamento Molecular: o método de busca, que deve considerar todas as possíveis soluções, graus de liberdade, e a função de energia, para avaliação a interação dos compostos. O algoritmo de busca objetiva percorrer o espaço de busca em um detalhamento suficiente a fim de encontrar o mínimo global da função de energia. O atracamento rígido considera o espaço de busca com diferentes posições de translação e rotação do ligante. O atracamento flexível adiciona os graus de liberdade internos da molécula, ou seja, são considerados os ângulos diedrais internos do ligante. Esse modelo conformacional permite uma simulação mais realista do Atracamento tal qual acontece na natureza.

A função de avaliação deve ser suficientemente realista para fornecer resultados compatíveis com o complexo determinado experimentalmente (BROOIJMANS; KUNTZ, 2003a). A função que representa as interações moleculares envolvidas no reconhecimento molecular proteína-ligante incluem: ligações de hidrogênio, interações de *van der Waals*, interações iônicas, interações hidrofóbicas, interações do tipo cátion-π, interações envolvendo anéis aromáticos do tipo π-π e empilhamento-T e coordenadas com íons metálicos (VERLI, 2014). A escolha de uma função de avaliação de energia que represente o sistema e as interações moleculares é de grande importância para o algoritmo de busca.

Problemas de Atracamento molecular enfrentam diversos desafios, (SOUSA et al., 2013) enumera três questões críticas para o Atracamento proteína-ligante: o tratamento da flexibilidade da proteína, a presença de estruturas moleculares de água e seus efeitos, e a

entropia da ligação química. A amostragem do ligante, a flexibilidade da proteína e a função de energia, referenciados em (HUANG; ZOU, 2010), são questões importantes para a resolução do problema. A amostragem se refere a geração de orientações e conformações próximas do sítio de ligação. A avaliação dessa orientação/conformação utilizando uma função de aptidão é fundamental para o algoritmo. Ademais, a velocidade em que o cálculo de energia é realizado e o custo computacional, relacionado com a quantidade de recursos necessários para um determinado algoritmo resolver o problema, envolvido no processo são aspectos importantes. A complexidade da função de avaliação infere no custo computacional do algoritmo, assim, a análise da relação entre o cálculo de energia e custo computacional é um desafio no problema de Atracamento Molecular.

Esse trabalho desenvolve um algoritmo que utiliza técnicas de amostragem para o cálculo de orientações e conformações do ligante em um espaço de busca discretizado seguindo uma metodogolia. A discretização do espaço de busca é utilizada também como parâmetro de similaridade entre as soluções, permitindo, assim, o uso de técnicas de agrupamento. Essa etapa visa criar soluções em todo o espaço de busca, que inclui o sítio de ligação da molécula receptora. A partir desse agrupamento, o algoritmo realiza uma competição local e global das soluções afim de diversificá-las. Para o cálculo de energia é utilizada uma função de campo de força semi-empirica. Definidas a preparação dos dados, discretização, competição das soluções, função de aptidão, um Algoritmo Genético de Chaves Aleatórias Viciadas foi aplicado como algoritmo de busca pelo mínimo global e, consequentemente, melhor conformação proteína-ligante.

1.1 Motivação

Com os avanços nos estudos no campo de biologia estrutural, há um aumento significativo do conhecimento de novas estruturas de proteínas, e, paralelamente, do número de estruturas de ligantes (ZHANG et al., 2012). Dessa forma, é necessário gerenciar eficientemente esses dados e desenvolver algoritmos que tornem o processo de descoberta de novos fármacos mais rápido e eficiente. Embora diversas ferramentas de Atracamento Molecular já tenham sido desenvolvidas, o problema ainda carece de uma ferramenta que o resolva de forma generalizada e acurada.

Estudos mostram que a abordagem utilizando algoritmos evolutivos, modelados para problemas de Bioinformática, podem gerar resultados superiores aos encontrados com algoritmos determinísticos. (SOUSA et al., 2013). O uso de algoritmos evolutivos

permite uma melhor exploração do espaço de busca, a utilização de funções complexas (não-diferenciáveis, multimodais e sujeita a restrições) e de rápida convergência. Dessa forma, o uso de uma meta-heurística adaptada ao problema pode gerar novos resultados em prol da resolução do mesmo. Algoritmos Genéticos são meta-heurísticas utilizados em estudos prévios de Atracamento Molecular e se mostram uma técnica promissora. A diversidade de parâmetros e variações, torna a técnica capaz de melhorar o processo de busca, além da possibilidade de serem aliados a outras técnicas.

Conseguir determinar a conformação proteína-ligante com acurácia e com baixo custo, pode gerar benefícios em vários campos de pesquisa como Medicina, Bioinformática e para a indústria farmacêutica (TRAMONTANO; LESK, 2006). O Atracamento Molecular é uma área carente de técnicas computacionais robustas e eficazes, na qual diferentes métodos podem ser aplicados, e cujos avanços científicos seriam significativos.

1.2 Objetivos gerais

O objetivo geral dessa pesquisa é desenvolver um algoritmo para o problema de Atracamento Molecular. Considerando o aumento de estruturas 3-D de proteínas determinadas experimentalmente ao longo dos últimos anos [4], o número de moléculas ligantes disponíveis em bancos de dados como o CSD, ZINC, e o longo e custoso processo de testes para o atracamento dessas estruturas por vias experimentais, a proposta de uma nova abordagem para a resolução desse problema vem de encontro com o auxílio ao descobrimento de novos fármacos.

A escolha das estruturas utilizadas, da representação desses dados, de uma função de energia que descreve as interações moleculares e dos parâmetros do algoritmo são exemplos de variáveis que influenciam a resolução do problema. Assim, desenvolver uma meta-heurística orientada ao problema, assim como propôr e testar diferentes representações de dados, exploração do espaço de busca e diferentes técnicas computacionais são os principais objetivos do trabalho. Para isso, são considerados os desafios da área e testadas diferentes implementações, em busca de um algoritmo que possibilite a predição da orientação da estrutura proteína-ligante.

Nessa nova abordagem para predição da orientação proteína-ligante serão realizadas operações sobre os dados, que podem ser definidas em quatro etapas:

[4] Endereço onde é apresentado o crescimento do número de estruturas 3-D de proteínas disponíveis no PDB: http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=molType-protein

- Etapa 1: preparação das estruturas. Este é um processo padrão em qualquer estudo de Atracamento Molecular. Nessa etapa é definido um conjunto de teste e cada estrutura é realizada sua preparação: verificando-se a necessidade de adição ou remoção de de átomos ou resíduos (no caso do receptor), verificação do posicionamento da cadeia lateral, da necessidade de protonação, entre outras modificação estruturais dos compostos.
- Etapa 2: atracamento molecular das estruturas. Nessa etapa o algoritmo é aplicado sobre as moléculas, testando-se diferentes parâmetros, em um determinado número de avaliações de energia.
- Etapa 3: análise do resultados. A partir da definição de uma parametrização ideal para um conjunto de teste, o algoritmo é executado por um determinado número de vezes para toda a seleção de complexos.
- Etapa 4: avaliação dos resultados. Por fim, serão utilizados métodos estatísticos de avaliação, análise do RMSD (*Root-mean-square deviation*), que avalia a distância interatômica de moléculas, possibilitando, dessa forma, avaliar a similaridade dos resultados com as estruturas experimentais, além da avaliação dos valores de energia e convergência do algoritmo.

1.3 Estrutura

No Capítulo 2 é apresentada a fundamentação biológica do trabalho, são explanadas as definições das moléculas receptoras e ligantes e suas interações. São apresentados os tipos de funções de cálculo de energia livre e suas definições. Os principais bancos de dados onde foram adquiridos os arquivos de representação dos complexos para esse trabalho são apresentados. A função biológica relacionada a estrutura é discutida nesse Capítulo. O posicionamento de ferramentas de Atracamento Molecular entre técnicas de Triagem Virtual é explicado e ilustrado. Por fim, é apresentado o Atracamento rígido e flexível.

No Capítulo 3 são apresentados as técnicas e algoritmo empregados para a resolução do problema de Atracamento Molecular. A base computacional parte da representação dos dados biológicos e das categorias de métodos aplicadas no problema. O experimento CAPRI é apresentado de forma a guiar as ultimas técnicas e metodologias que vem sendo aplicadas, que são posteriormente discutidas. É dado, ainda, o enfoque

em duas ferramentas, cujos resultados servem de comparação com os obtidos nesse trabalho. Por fim, são apresentados os principais desafios na área de Atracamento Molecular atualmente.

No Capítulo 4 é apresentado o método proposto, assim, são definidos a representação de dados escolhida e é descrita a função de energia, a metodologia utilizada para a discretização do espaço de busca. O algoritmo é, então, descrito detalhadamente, assim como as técnicas de agrupamento, reinicialização, competições local e global. Finalmente, são ilustrados toda a metodologia unindo todas as técnicas em forma de diagramas.

No Capítulo 5 são descritos os métodos de avaliação, os dados selecionados e os experimentos realizados. São apresentadas os diferentes conjuntos de organização dos dados e uma descrição sobre os mesmos. Os experimentos incluem a parametrização do algoritmo, os testes de Atracamento rígido e flexível e a comparação com outras ferramentas. Por fim, são realizadas as conclusões da Dissertação e são discutidos futuros trabalhos.

2 FUNDAMENTAÇÃO BIOLÓGICA

O principal objetivo do Atracamento Molecular é otimizar a interação entre a molécula receptora e uma molécula ligante. A interação entre duas moléculas é avaliada por meio de uma função de energia, essa função realiza cálculos a partir da ligação química e interação desses dois compostos. Assim, é importante entender cada uma dessas estruturas e as interações proteína-ligante, o cálculo da energia livre entre moléculas, a função desses complexos, que são relacionados a sua forma estruturas, e os conceitos de Atracamento rígido e Atracamento flexível. Além disso, nesse Capítulo são apresentados os principais banco de dados utilizados para aquisição dessas estruturas e o papel de ferramentas de Atracamento Molecular dentro do processo de Triagem Virtual de fármacos.

2.1 Estruturas Moleculares: Ligantes e Receptores

Um fármaco é uma estrutura molecular capaz de gerar ou bloquear um reação biológica no organismo (BARREIRO; FRAGA, 2014). Essa micromolécula, no Atracamento Molecular definida como a estrutura ligante, interage com outra molécula receptora (proteínas ou enzimas) gerando uma resposta biológica.

Proteínas são polímeros sintetizados pelas células a partir de aminoáciodos (VERLI, 2014), essas biomoléculas estáveis são capazes de adotar diversos arranjos tridimensionais. A função de muitas proteínas está relacionada com a ligação com outras moléculas, por serem moléculas dinâmicas sua função quase invariavelmente depende da interação com outros moléculas. Essas subtas mudanças em sua conformação refletem a vibração molecular e os movimentos de resíduos de aminoácidos. As características transientes de proteínas e ligantes é crítico para a vida, permitindo para a um organismo responder rapidamente a mudança em um ambiente ou circunstâncias metabólicas (LEHNINGER; NELSON; COX, 2004).

Um ligante pode ser qualquer outra molécula, inclusive outra proteína. Fármacos são as moléculas ligantes, estruturas fundamentais para o Atracamento Molecular. A molécula ligante se conecta em uma região da proteína chamada sítio de ligação, uma área complementar ao ligante em tamanho, forma, cargas e características hidrofóbicas ou hidrofílicas. Essa interação é de tanta especifidade que a proteína é capaz de discriminar entre milhares de moléculas e seus ambientes realizando a ligação química com apenas uma ou poucas estruturas de ligantes. Essa seletividade é de vital importância para manter

o alto grau de ordem em um sistema vivo.

Uma das estruturas estudadas nesse trabalho é a HIV-protease, enzima atacada pelo vírus HIV, cuja replicação no organismo causa a Síndrome da Imonodeficiência Adquirida (AIDS). Os fármacos desenvolvidos conseguem se encaixar no sítio de ligação dessa molécula receptora, fazendo com que essa enzima seja bloqueada e consequentemente o vírus fique impedido de sua reprodução. A Figura 2.1 apresenta a estrutura tridimensional do complexo é representado pela código PDB 1AJV, cuja molécula receptora é a estrutura HIV-protease sem a molécula ligante, já a Figura 2.2 apresenta a molécula ligante (em vermelho) ligada ao mesmo receptor.

Figura 2.1: Molécula receptora HIV-protease (representada em surface e cartoon), em destaque a o sítio ativo da molécula

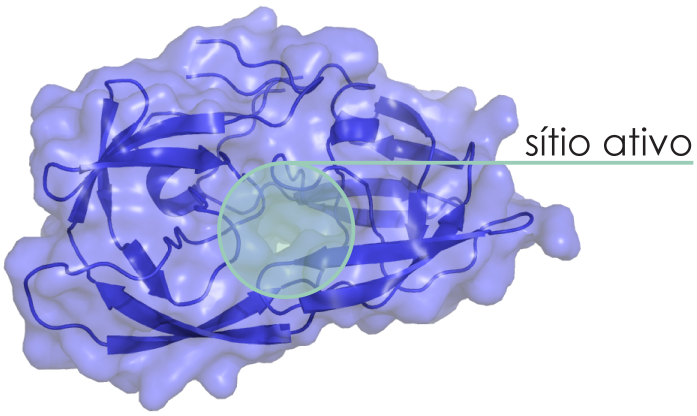
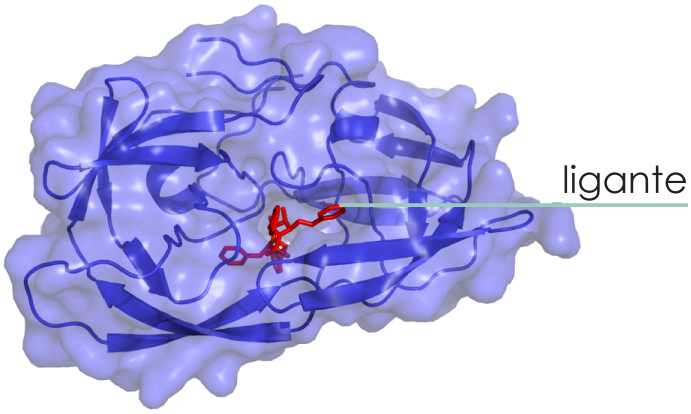


Figura 2.2: Molécula receptora HIV-protease (representada em surface e cartoon) com a molécula ligante (em vermelho)



O problema de Atracamento Molecular é definido, portanto, como a busca pela a melhor conformação entre essas duas moléculas. Considerando o sítio de ligação da molécula receptora, essa posição ideal possibilita que a molécula receptora ative a função do fármaco ou a iniba. As características físico-químicas são responsáveis pela afinidade e especificidade do ligante e do receptor. Já as características estruturais determinam o arranjo espacial das moléculas, cujas variações nessas estruturas são translações, mudanças na orientação e rotações das ligações covalentes.

2.2 Interações Ligante-Receptor

A formação de complexos, cujas ligações, estáveis ou transientes, de duas os mais moléculas, promove comunicações intra e intermoluculares (EISENSTEIN; KATZIR, 2004). Essas interações são fundamentais para quase todos os processos em um organismo vivo (DUNN, 2007). As interações de um fármaco com o seu sítio de ação no sistema biológico ocorrem durante a chamada fase farmacodinâmica e são determinadas pela resultante entre forças intermoleculares atrativas e repulsivas, isto é, interações hidrofóbicas, eletrostáticas e estéricas (BARREIRO; FRAGA, 2014).

O processo de ligação de uma proteína e um ligante acontece junto com um mudança conformacional na proteína que faz com que do sítio de ligação seja complementar ao ligante. Esse processo permite uma ligação mais justa das duas moléculas. Trabalhos como (GABB; JACKSON; STERNBERG, 1997) utilizam esse conceito de complementaridade de formas para prever a atracamento de proteínas. Junto com a mudança conformacional da proteína, a interação de cargas pontuais dispersas com o campo de *Coulomb* e eletroestática são elementos considerados no estudo. Dessa forma, é possível perceber que um grande número interações intra-e intermoleculares estão envolvidas no reconhecimento molecular de receptores e ligantes.

As principais interações entre os complexos biomoleculares, postulado por (PAULING; DELBRUCK, 1940), são as interações de *van der Waals*, caracterizadas pela atração de molélucas apolares que apresentem dipolo induzido, interações eletroestáticas, cujas forças resultam em uma atração ou repulsão entre as cargas e dependem de uma constante dielétrica do meio e da distância intermolecular das cargas, e ligações de hidrogênio, ligação química em que apenas dois elétrons são compartilhado por três átomos. Essas interações são importantes para a estabilidade do complexo biomolecular (BENITE; MACHADO; BARREIRO, 2007).

Outro fator importante nas interações receptor-ligante são os solventes, como moléculas de água, que podem modificar características estruturais dos sítios de ligação (Pauling; Delbruck, 1940). A maioria das proteínas passam pelo processo de enovelamento e funcionam em ambiente aquoso. Dados estruturais e termodinâmicos indicam que água em um complexo proteína-ligante pode contribuir para a ligação química (LADBURY, 1996). A entropia do sistema também é alterada quando o solvente é adicionado, já que as superfícies apolares liberam e desorganizam as moléculas de água. Esse aumento da entropia do solvente com o ocultamento das superfícies apolares é conhecido como efeito hidrofóbico (BALDWIN, 2014). As mudanças na entropia do sistema alteram a estabilidade do complexo, como a perda da entropia rotacional e translacional, e variações na entropia vibracional e conformacional da biomolécula. A água participa ainda nas interações na parte interna da proteína. Algumas proteínas podem ser parcialmente ou completamente preenchidas por água em seus canais. Em resumo, moléculas de água próximas da estrutura da proteína fazem parte da estrutura da proteína, já que elas determinam a conformação das cadeias laterais expostas, estabilizam o fim das estruturas secundárias, e ainda ocupam posições em sítios ativos onde influenciam ligações e, às vezes, catalizações (RICHARDSON, 1981).

Cofatores como coenzimas e grupos prostéticos são substâncias orgânicas (coenzimas) ou inorgânicas necessárias para o funcionamento de enzimas. Muitas proteínas, em particular enzimas, conseguem apenas realizar sua função bioquímica se conectadas, primeiramente, a uma molécula diferente (KEPPEL, 1991). As principais coenzimas são vitaminas, que podem em alguns casos estar fortemente ligadas a proteina, como ions de metais como zinco e cobre. Coenzimas como NADH e ATM são vitais para o metabolismo celular.

Ligações com metais conseguem estabilizar uma estrutura tridimensional de uma proteína. Como estruturas fundamentais de muitas sistemas biológicos, metais, por vezes, neutralizam cargas negativas que em outra situação iria repelir-se. Metais são ainda usados como fator catalizador de atividades em enzimas. Algumas proteínas contém agrupamentos que incluem diversos átomos de metais ou íons. Agrupamentos de metais orgânicos podem ainda fazer parte da proteína e em enzimas específicas.

2.3 Cálculo da energia livre

O cálculo da energia livre de uma estrutura necessita de métodos computacionais robustos(FRENKEL; SMIT, 2002). Determinar de maneira acurada e com baixo custo computacional a energia do complexo receptor-ligante é um campo de estudo no campo descobrimento de fármacos. Determinar a afinidade de ligação de compostos conseguiria prever quais compostos seriam mais propensos a serem sintetizados. A necessidade de uma avaliação rápida, por vezes, leva ao uso de funções que aproximam a avaliação do complexo. Algumas técnicas utilizam funções simplificadas em fase de execução para por fim inserir elementos na funções para uma melhor avaliação. Diferentes funções de energia vem sido utilizadas por programas de Atracamento, as principais podem ser divididas em três classes principais (KITCHEN; FURR J. R., 2004): funções baseadas em campo de força, funções empíricas e semi-empíricas, funções baseadas em conhecimento.

2.3.1 Funções baseadas em campo de força

Um campo de força pode ser descrito como um campo vetorial que descreve as forças agindo sobre uma partícula em várias posições no espaço. Funções baseadas em campo de força quantificam a soma de duas energias, a energia de interação receptor-ligante e a energia interna do ligante. A maioria das funções de campo de força consideram somente uma conformação da proteína, o que causa a omissão do cálculo da energia interna da proteína, simplificando a avaliação de energia. Campos de força amplamente utilizados são: GROMOS (GUNSTEREN, 1987), AMBER (WEINER et al., 1984), CHARMM (CORNELL; CIEPLAK, 1995; BROOKS, 1983) e MMFF94 (HALGREN, 1996a; HALGREN, 1996b).

As interções entre ligante e receptor são frequentemente descritas utilizando parâmetros de energia de *van der Waals* e eletrostática. O termo de energia potencial de *van der Walls* é dado pela energia potencial de Lennard-Jones. Termos eletrostáticos são inferidos pela formulação de *Coulomb* com uma função de que avalia a distância entre cargas e suas contribuições carga a carga. A forma funcional de energia interna do ligante é geralmente bastante similiar com a interação receptor-ligante, incluindo também termos de *van der Walls* e eletrostática.

A energia potencial eletrostática é representada como um par de interação de *Cou-*

lomb entre as moléculas, descritas na Eq. 2.1:

$$E_{coul}(r) = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \frac{q_i q_j}{4\pi\epsilon r_{ij}} \tag{2.1}$$

onde N é o número de átomos na molécula A e B, respectivamente, e q representa a carga de cada átomo.

A energia potencial de *van der Walls* para o tratamento de interações de não-ligados é geralmente modelado como a função de Lennard-Jones, como descrito na Eq 2.2.

$$E_{vdW}(r) = \sum_{j=1}^{N_A} \sum_{i=1}^{N_B} 4\pi[(\frac{\sigma_{ij}}{r_{ij}})^{12}) - (\frac{\sigma_{ij}}{r_{ij}})^{6})] \tag{2.2}$$

onde σ é o poço de potencial e r é a distância (finita) na qual o potencial inter-partícula é zero.

Funções baseadas em campo de força tem grandes limitações, pois são originalmente formuladas para a modelagem de contribuições entálpicas para a estrutura e energias, não incluindo, dessa forma, solvatação e termos de entropia (KITCHEN; FURR J. R., 2004). Essas funções exigem também a introdução de delimitadores de distâncias para o tratamento de interações de complexos não-ligados, o que é feito de maneira empírica.

2.3.2 Funções empíricas ou semi-empíricas

Funções empíricas são inferidas a partir de dados experimentais. Essas funções analisam energias de ligação e/ou conformação, como a soma de diversas funções parametrizadas, primeiramente proposto por (BÖHM, 1992). A formulação de funções de avaliação empíricas é baseada na ideia de que a energia de ligação podem ser aproximadas pela soma de termos individuais não relacionados. Os coeficientes de vários termos são obtidos pela análise de regressão usando energias de ligação experimentalmente determinadas, e ainda, informações estruturais da cristalografia por raios-X.

A formulação de funções empíricas ou semi-empíricas é frequentemente mais simples do que funções de avaliação de campo de força, embora muitos dos termos de contribuições individuais tenham partes iguais aos termos mecânicos do campo de força molecular. A vantagem desse tipo de função é que os termos são, geralmente, simples de avaliar. A desvantagem desses métodos é a necessidade de utilizar dados experimen-

tais para desenvolver a regressão e adaptação, o que acaba rendendo diferentes fatores de pesos para vários termos (SCHNEIDER; BöHM, 2002). Como consequência, termos de diferentes parametrizações não são facilmente recombinados para uma nova função de energia.

Ferramentas de atracamento molecular como o GOLD (JONES; WILLETT; GLEN, 1995) utilizam os termos de avaliação de ligações de hidrogênio baseado em valores empíricos para a força de ligação entre diferentes átomos de hidrogênio. A função (MORRIS et al., 1998) utiliza termos baseados em campo de força e termos semi-empíricos, embora os termos sejam baseados em campo de força, seus pesos são multiplicados por termos obtidos experimentalmente. A função semi-empírica do software AutoDock (MORRIS et al., 2009) faz o re-escalonamento dos coeficientes nos termos da função de energia da mecânica molecular, além de incluir dois novos termos. Esses termos incluem o efeito da solvatação na interação receptor-ligante e conseguem fazer uma estimativa da perda de graus de conformação do ligante quando ocorre a ligação com o receptor. Programas como o LUDI (BÖHM, 1994) e FlexX (RAREY et al., 1996) também implementam funções empíricas, incluindo termos de ligação de hidrogênio, ponte salina, efeito hidrofóbico e entropia.

Em funções empíricas a formulação pode ser bem variada, como termos para as interações de não-ligados. Podem também incluir contribuições não-entálpicas, chamados de termos rotor. Esses termos aproximam as penalidades de entropia da ligação elevando o peso do somatório do número de ângulos diedrais nos ligantes. Todavia, termos utilizados atualmente para a aproximação da entropia ou energia de solvatação incorporem descrições incompletas desses efeitos em ligações proteína-ligante (SCHNEIDER; BöHM, 2002).

2.3.3 Funções baseadas em conhecimento

Funções baseadas em conhecimento são formuladas a partir da reprodução de resultados experimentais de estruturas. Para inferir essas funções, complexos receptor-ligante são modelados utilizando relações simples de potenciais de átomo par a par e um número de interações de átomos é definida dependendo do ambiente molecular. Assim, como ocorre em métodos empíricos, funções baseadas em conhecimento tentam implicitamente capturar os efeitos da ligação que são difíceis de modelar explicitamente (WANG; LU; WANG, 2003). Inferindo, assim, funções acuradas e de baixo custo computacional.

Entre os fatores que são incluídos nessas funções estão os potenciais de força média (PMF)(MUEGGE, 2000; MUEGGE, 2001; MUEGGE; MARTIN, 1999) para a avaliação da energa livre. Drugscore (GOHLKE; HENDLICH; KLEBE, 2000) inclui também a correções de acessibilidade do solvente para avaliar a interação proteína-ligante. SMoG (DEWITTE; SHAKHNOVICH, 1996) é outra ferramenta que utiliza a mesma classe em diversos termos de sua função de avaliação de energia. A maior vantagem na utilização desse tipo de função é a simplicidade e baixo custo computacional, permitindo, dessa forma, uma análise em uma grande base de dados para a Triagem Virtual. Sua desvantagem é o fato de que a derivação dessas funções são basicamente informações experimentais de moléculas limitados, dessa forma, pelo número de complexos proteínas-ligantes utilizados para sua composição (ZHANG et al., 2005).

2.4 Bancos de dados

Para a representação computacional de estruturas biológicas são utilizadas diversas ténicas. Inicialmente a primeira técnica aplicada é a cristalografia de raio-X. A critologrfia de raios-X é um técnica sofisticada que, simplificadamente, consiste em fazer passar um raio-X através de um cristal da substância estudada. Com a difusão do feixe padrões de intensidade podem ser interpretados as distribuições de átomos dos cristais, por difração, extraindo-se, assim, informações sobre a estrutura molecular do complexo. Bancos de dado como o *Protein Data Bank* (PDB; http://www.rcsb.org/pdb/),(BERMAN et al., 2000), e ZINC (http://zinc.docking.org/),(IRWIN; SHOICHET, 2005), utilizam, entre outras técnicas, a cristalografia de raios-X e disponibilizam essas moléculas para estudos científicos.

O banco de dados PDB é mais difundido repositório de dados estruturais de biomoléculas. O PDB foi criado em 1971 pelo *Brookhaven National Laboratories (BNL)* como um repositório de estruturas cristalográficas de macromoléculas. Além da representação computacional da molécula, o banco de dados ainda provê informações sobre o método de aquisição, resolução, entre outros, informações essas de grande importância para a análise dessas estruturas. Outros centros de aquisições de estruturas atuam juntamente com o PDB para aquisição de estruturas como o European *Bioinformatics Institute* (EBI). Os dados passam por um processo de validação assegurando a qualidade do modelo atômico depositado.

O banco de dados ZINC é a mais vasta coleção de compostos químicos, comerci-

almente disponíveis e preparados para Triagem Virtual, melhor explanado na Seção 2.6. O repositório é uma ferramenta de pesquisa que disponibiliza compostos químicos para alvos biológicos, incluindo fármacos comerciais. Aliados a outros 20 bancos de dados, o foco do repositório é em compostos para o Atracamento Molecular. Além das estruturas biológicas o ZINC também disponibiliza informações sobre a estrutura, como resolução, flexibilidade, entre outras informações químicas.

Para a obtenção do formato computacional das estruturas utilizadas nesse trabalho foram utilizados os bancos de dados PDB e Zinc. A partir da obtenção dos mesmos, as estruturas foram preparadas, visando formatar os arquivos em uma representação do algoritmo e juntamente com uma análise biológica das estruturas.

2.5 Função biológica relacionada a estrutura

Um dos principais desafios a partir da modelagem de estruturas biológicas é a inferência de sua função. Proteínas que possuem uma origem evolutiva em comum compartilham também uma estrutura similar, são chamadas de proteínas homólogas. Entretanto, em alguns casos, proteínas que não possuem uma origem evolutiva similar podem compartilhar a mesma topologia. Assim, a relação estrutura receptora e função biológica é uma tarefa difícil de ser inferida.

As principais dificuldades nesse campo de estudo podem ser resumidas nos fatos de que: proteínas homólogas podem ser originadas por duplicação de genes e evolução subsequentemente e, assim, adquirirem uma diferente função; alguns enovelamentos são adotados por proteínas para o desenvolvimento de diversas funções; proteínas podem ter um enovelamento novo ainda não observado (TRAMONTANO, 2006).

A partir da análise estrutural da proteína é possível perceber algumas características relacionadas à sua função. Por exemplo, quais resíduos estão expostos ao solvente e quais se encontram no centro da proteína, definido-os, respectivamente, como hidrofílicos ou hidrofóbicos. A forma quaternária da proteína pode também ser inferida pela estrutura, já que a forma observada por meio da cristalografia é geralmente a topologia biologicamente ativa da molécula.

Outra característica que pode ser inferida pela estrutura da proteína é a presença de motivos locais, cujas funções podem ser identificadas pela estrutura. Mesmo que motivos locais funcionais não possam ser detectados, ainda é possível analisar fendas na superfícies da proteína e identificar a presença de certos aminoácidos, cadeias laterais que

estejam envolvidos na atividade catalizadora.

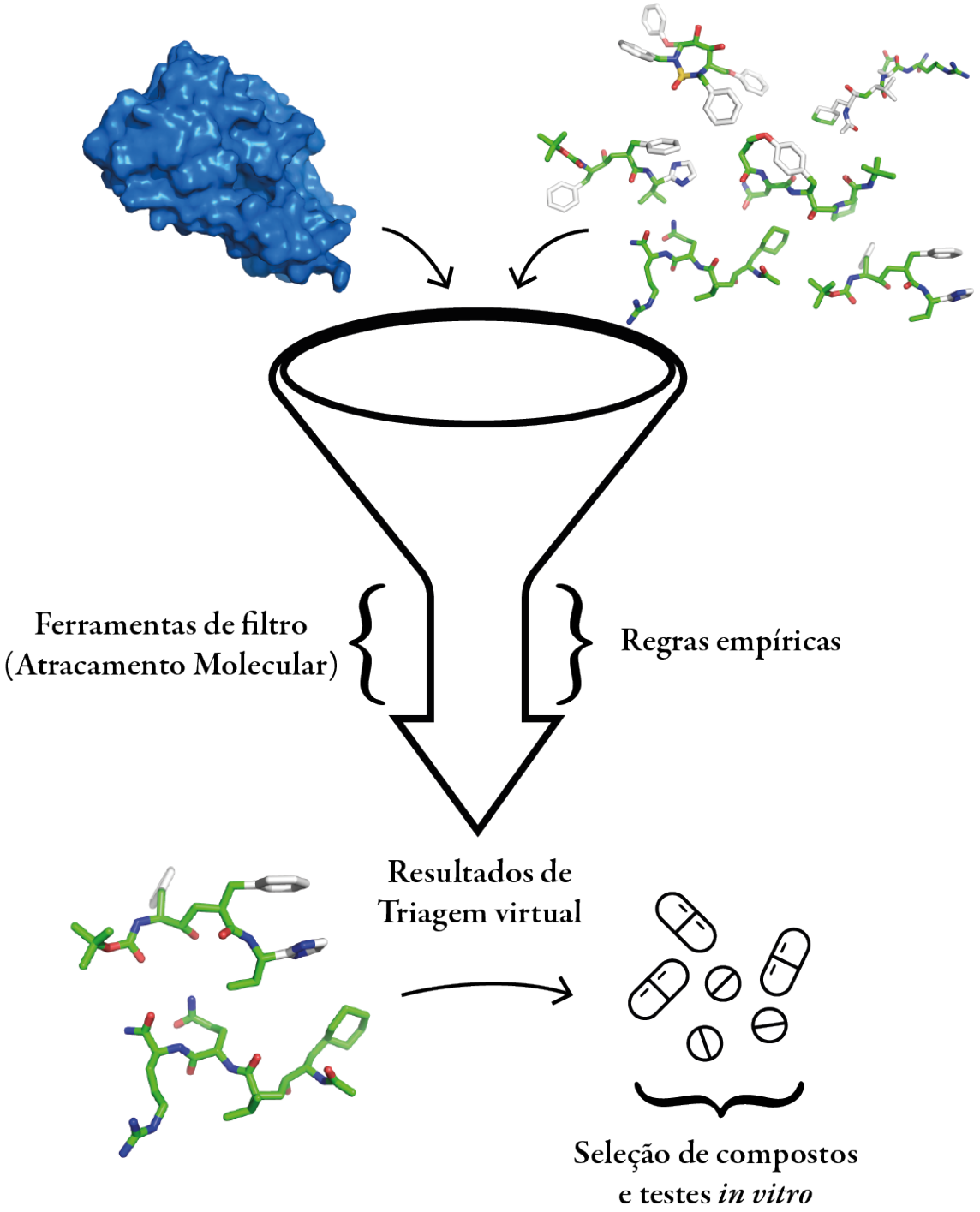
Portanto, a função da proteína está associada a sua forma estrutural, o que torna de suma importância a identificação da conformação proteína-ligante através do Atracamento Molecular. Além da estrutura final com as duas moléculas acopladas, a posição de cada átomo infere características ao complexo, a presença de íons, moléculas de água e mesmo a posição de certos amino-ácidos revelam características químicas relacionadas à função da proteína.

2.6 Triagem Virtual

Triagem Virtual (*Virtual Screening*) é uma técnica computacional utilizada no descobrimento de novos fármacos. Busca-se identificar estruturas mais propensas para a ligação em uma molécula alvo, normalmente uma proteína ou uma enzima. Pode ser considerado com um filtro que reduz a quantidade de compostos químicos presentes em bancos de dados que serão futuramente testados. A seleção de estruturas em bases de dados químicas é uma metodologia bem estabelecida para encontrar novos candidatos a fármacos, considerando uma estrutura tridimensional alvo conhecida (WALTERS; STAHL; MURCKO, 1998). Com o aumento de alvos farmacêuticos preditos, métodos de Triagem Virtual sem dúvidas terão papel fundamental na área farmacogenômica para encontrar os primeiros compostos alvos, especialmente em compostos em que não há informação sobre potenciais ligantes (BISSANTZ; FOLKERS; ROGNAN, 2000).

Os métodos atuais de Triagem Virtual passam primeiramente por ferramentas de Atracamento Molecular, onde é possível prospectar a afinidade de ligação de duas moléculas, e, então, por uma avaliação de energia. Assim, a ferramenta de Atracamento Molecular é utilizada em uma base de dados de compostos com o foco de eliminar estruturas não desejadas. A busca de ligantes através de métodos computacionais que consideram a estrutura 3D de um alvo terapêutico é chamada triagem baseada na estrutura alvo-molecular (SCHNEIDER; BöHM, 2002). O nível de sofisticação de ferramentas de Triagem Virtual e sua dependência de contexto cresce com o conhecimento disponível de uma droga em particular e com o padrão de interação receptor-ligante (KELLENBERGER et al., 2004). A Figura 2.3 ilustra o processo de Triagem Virtual e o Atracamento Molecular nesse processo.

Figura 2.3: Ilustração do processo de Triagem Virtual: primeiramente é seleciona uma molécula receptora e uma série de candidatos a fármacos, ou moléculas ligantes, através de regras empíricas e ferramentas de Atracamento Molecular são selecionados moléculas para a desenho de fármacos e posteriores testes *in vitro*

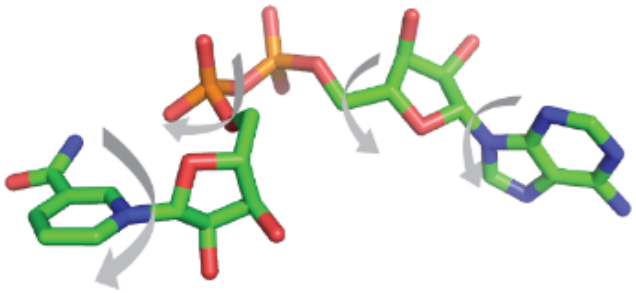


2.7 Atracamento Rígido e Flexível

Um dos fatores a ser considerado em uma técnica de Atracamento Molecular é a flexibilidade das moléculas receptora e ligante. Durante o processo de Atracamento ocorrem mudanças conformacionais dessas moléculas, já que a topologia dessas estruturas se molda para que a ligação química ocorra da forma mais estável possível. Adicionar a flexibilidade é uma característica importante já que simula de maneira mais realística o complexo molecular, porém aumenta em muita a complexidade do problema, tornando-o em alguns casos inviável devido ao número de graus de liberdade de certas moléculas.

O atracamento rígido considera apenas as variações translacionais e rotacionais da estrutura. A complexidade do problema diminui já que se considera as rotações do ligante e da proteína como corpos rígidos. No problema de Atracamento flexível são considerados os graus de liberdade dos ângulos diedrais, como ilustrado na Figura 2.4. Atualmente, a maioria dos programas de atracamento incluem além da liberdade rotacional e translacional, as rotações diedrais que modificam a conformação da estrutura.

Figura 2.4: Graus de liberdade dos ângulos diedrais de um ligante



Em ambos as propostas de inclusão ou não de flexibilidade ao ligante a estrutura da proteína é mantida fixa na posição determinada experimentalmente. Todavia, essa rigidez não condiz com a realidade biológica, já que a molécula passa por mudanças conformacionais assim como o ligante. Adicionar a flexibilidade ao receptor aumenta em muito a complexidade do problema, levando em conta o número de átomos presentes na estrutura proteica. A flexibilidade na proteína é fundamental para entender a formas em que fármacos exercem seus efeitos biológicos, suas posição no sítio de ligação, suas orientação, cinética de ligação, metabolismo e transporte (TEAGUE, 2003). Alguns estudos conseguem incluir a flexibilidade em alto grau na proteína (VERDONK et al., 2003; JONES; WILLETT, 1995; CLAUSSEN; BUNING CM.; LENGAUER, 2001), porém a inclusão da flexibilidade é um grande desafio para ferramentas de Atracamento.

Outra metodologia é inclusão da flexibilidade apenas no sítio de ligação receptor-ligante ou em partes da molécula receptora. Embora essa metodologia não represente as interação biológicas com tanta exatidão, chega-se a uma aproximação do modelo biológico suficientemente acurada. Alguns estudos utilizam essa metodologia (TROTT; OLSON, 2010; WEI et al., 2004; FISCHER et al., 2014), e mostram resultados próximos as estruturas experimentais. Métodos de inclusão parcial no receptor também incluem bibliotecas de rotâmeros que são valores preferenciais dos ângulos das cadeias laterais de resíduos de aminoácidos. Nesse modelo o algoritmo deve fazer uma busca exaustiva sobre todas as conformações preferencias de cada aminoácido. Essa abordagem pode ser vista nos trabalhos de (LEACH, 1994) e (JACKSON; GABB; STERNBERG, 1998).

2.8 Conclusão

O alto custo e tempo necessário para a criação de um fármaco estimulam o desenvolvimento de técnicas computacionais que acelerem esse processo. Ferramentas de Atracamento Molecular e Triagem Virtual já são aplicadas no processo de desenho e descoberta de novos candidatos a fármacos. Na pesquisa científica diversas técnicas vem sendo desenvolvidas aliando-se ao desenvolvimento computacional e algoritmos, porém muitos desafios são enfrentados considerando a alta complexidade do problema.

3 TÉCNICAS E ALGORITMOS DE ATRACAMENTO MOLECULAR

Técnicas de Atracamento Molecular são caracterizadas por abordagens geométricas ou de energia. Métodos que exploram a geometria das moléculas analisam alinhamentos factíveis entre receptores e ligantes conhecidos, examinam suas ligações químicas e avaliam seus efeitos estereoquímicos (KUNTZ et al., 1982). O alto número de ângulos internos tornam esses modelos mais simplificados e, assim, menos acurados quando comparados aos métodos de avaliação de energia. Na abordagem de avaliação de energia é realizado o cálculo de energia livre dos complexos, são testadas diferentes conformações e a busca pelo menor valor do potencial energético das estruturas.

Métodos de Atracamento Molecular baseados em energia utilizam diferentes representação das estruturas tridimensionais e técnicas de otimização. A forma computacional que a estrutura é representada passa pela modelagem dos compostos e retratam as interações físico-químicas dos complexos. A busca pelo mínimo global da função de energia equivalente ao teste de diferentes conformações receptor-ligante e a avaliação de suas interações atômicas. Se esta função é suficientemente acurada, a conformação nativa da estrutura coincide com o mínimo global de energia (COMBS et al., 2013). Diversas técnicas de otimização foram desenvolvidas e aplicadas ao problema, no entanto, muitos desafios ainda são encontrados (HUANG; ZOU, 2010).

3.1 Representação de estruturas moleculares

Computacionalmente proteínas e moléculas ligantes podem ser representados de três formas: por superfície, por grade ou por átomos (HALPERIN et al., 2002). A representação por superfície é utilizada principalmente em ferramentas de Atracamento Molecular proteína-proteína. Estruturas de proteínas permitem o estudo de suas características de superfície baseados em sua contribuição atômica. Esses métodos tentam alinhar como pontos da superfície minimizando ângulos entres as superfícies de moléculas opostas (ANDREI et al., 2012). O uso de grades de energia potencial foi primeiramente proposto por (GOODFORD, 1985), utilizando essa representação para a minimização da energia. A ideia básica pressupõe o armazenamento de informações sobre as contribuições energéticas pontos de uma grade os quais são lidos durante a avaliação de energia do ligante. Normalmente esses pontos armazenam dois tipos de potenciais: eletrostático e *van der Waals* (SCHNEIDER; BöHM, 2002). A representação por átomos é utilizada

em conjunto com uma função de energia potencial durante o processo de avaliação de aptidão. Considerando a quantidade de átomos presentes no complexo receptor-ligante com número de interação de pares de átomos, essa representação pode se mostrar computacionalmente custosa.

A representação da estrutura ligante, estrutura menor que o receptor, é geralmente definida por átomos, definindo coordenadas cartesianas para cada átomo e suas ligações químicas. Cada ligação covalente possui um ângulo diedral associado. Os ângulos diedrais definem a conformação da estrutura, dessa forma, permitir a variação dos mesmos significa levar em consideração a flexibilidade do ligante (SIMONSEN et al., 2013). Perturbações randômicas translacionais, rotacionais da estrutura completa e dos ângulos internos são realizadas dentro de um sítio de ligação a fim de encontrar a posição e conformação da estrutura que apresente a menor energia dentro do sistema.

A estrutura tridimensional é representada por coordenadas x, y e z de cada átomo. Assim, a representação das variações translacionais das estruturas são representadas utilizando o equacionamento 3.1.

$$(x, y, z) \rightarrow (x + \Delta x, y + \Delta y, z + \Delta z) \tag{3.1}$$

Às coordenadas de cada átomo são adicionados variáveis randômicas Δx, Δy e Δz. Essa operação translada a molécula ligante como uma estrutura rígida. Variações translacionais podem ser realizadas em apenas algumas coordenadas, diversificando a aleatoriedade da operação de translação. Variar apenas uma coordenada de translação significa transladar em apenas um eixo a estrutura biológica. Valores randômicos de translação devem respeitar o espaço de busca, que deve incluir o sítio de ligação da proteína.

A fim de variar a rotação da molécula, as operações apresentadas na Equação 3.2 são realizadas em cada coordenada da biomolécula. Para realizar a operação são definidos quatro valores, três representam um vetor de referência e o quarto referente a um ângulo Θ. O vetor de referência é definido a partir de coordenadas de um átomo, no qual a estrutura é rotacionada, assim, o átomo de referência permanece fixo, enquanto a estrutura rota em torno desse referencial. A partir das coordenadas desse átomo é definido, primeiramente, um vetor unitário $u = (u_x, u_y, u_z)$. Esse vetor é geralmente escolhido no centro de massa da molécula, porém qualquer ponto da estrutura pode ser escolhido, considerando que sobre esse vetor é realizada a operação. Um quadrivetor $Q = (q_0, q_1, q_2, q_3)$ define as operações geométricas de rotação. Essas operações são arranjadas em uma matriz R conhecida como matriz de rotação, como na Equação 3.2. A matriz de rotação é então

multiplicada por todos os pontos da moléculas, o resultado final são novas coordenadas com todos os átomos rotacionados Θ radianos.

$$\begin{aligned} q_0 &= cos(\frac{\Theta}{2}) \\ q_1 &= u_x(1 - q_0 q_0)^{\frac{1}{2}} \\ q_2 &= u_y(1 - q_0 q_0)^{\frac{1}{2}} \\ q_3 &= u_z(1 - q_0 q_0)^{\frac{1}{2}} \end{aligned} \tag{3.2}$$

$$R = \begin{bmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(q_1 q_2 - q_0 q_3) & 2(q_1 q_3 + q_0 q_2) \\ 2(q_1 q_2 - q_0 q_3) & q_0^2 - q_1^2 + q_2^2 - q_3^2 & 2(q_2 q_3 - q_0 q_1) \\ 2(q_1 q_3 - q_0 q_2) & 2(q_2 q_3 - q_0 q_1) & q_0^2 - q_1^2 - q_2^2 + q_3^2 \end{bmatrix}$$

Utilizando esse equacionamento, Ref. (MAGALHAES, 2006), são minimizadas as operações trigonométricas, sendo necessário fornecer coordenadas referentes ao vetor de referência e ao ângulo de rotação. A rotação dos ângulos diedrais é realizada pelo mesmo equacionamento, entretanto a definição do vetor unitário é feita a partir da ligação covalente. Os pontos que formam vetor são as coordenadas dos dois átomos da ligação química.

No Atracamento rígido são realizadas somente operações de translação e rotação na estrutura, não são consideradas, assim, operação de rotação dos ângulos diedrais. No problema de Atracamento flexível são considerados também os ângulos internos (ligações covalentes) da molécula ligante, ou seja, há uma mudança conformacional do complexo molecular.

Além de considerar os ângulos internos de rotação da molécula ligante algumas abordagens incluem a flexibilidade parcial ou total da estrutura receptora (COZZINI et al., 2008). Métodos de Dinâmica Molecular, Monte Carlo e Algoritmos evolutivos são utilizados, combinando-os, por vezes, com Bibliotecas de Rotâmeros (CHANDRIKA; SUBRAMANIAN; SHARMA, 2009) ou Grades de Energia (EISENSTEIN; KATZIR, 2004). Incluir a flexibilidade parcial ou total do receptor aumenta a complexidade do problema, assim, é necessário o uso de métodos de busca mais robustos.

3.2 Categorias de métodos de busca

Para determinar a melhor conformação receptor-ligante algoritmos de busca são aplicados a fim de encontrar o mínimo global da função de avaliação de energia do complexo. Algoritmos de busca podem ser classificados em três grupos de acordo com a metodologia aplicada para explorar a flexibilidade do ligante: busca sistemática, determinística e estocástica (BROOIJMANS; KUNTZ, 2003b).

Algoritmos de busca sistemática consideram todos os graus de liberdade molecular através de um conjunto de valores explorados de forma combinacional. Nesses algoritmos a molécula ligante é dividida em fragmentos rígidos e flexíveis que são incorporados ao sítio de ligação conectando partes da molécula e somatizando-as até obter a estrutura completa. Por essa razão são conhecidos como algoritmos de construção incremental ou baseados em fragmentos. Primeiramente, um fragmento núcleo é adicionado ao sítio de ligação, em seguida, para cada novo fragmento, é realizada uma busca conformacional considerando um conjunto de valores relativos aos graus de liberdade, ângulos diedrais, do ligante.

Em métodos determinísticos o estado atual do sistema determina as modificações a serem feitas guiando para o próximo estado. O resultado final é altamente dependente do estado inicial da estrutura, pois, dada uma mesma configuração inicial do sistema e uma mesma parametrização, o estado final será o mesmo (GUEDES; MAGALHãES; DARDENNE, 2013). Algoritmos determinísticos são utilizados frequentemente quando existe uma relação clara entre as características de uma possível solução e sua utilidade para um dado problema (WEISE, 2009). Métodos clássicos de Atracamento Molecular utilizam essa técnica para otimização da energia, assim como em métodos de simulação por dinâmica molecular.

Métodos estocásticos aplicados ao problema de Atracamento Molecular modificam randomicamente os graus de liberdade da molécula (translacional, rotacional e conformacional) e, a cada passo, geram uma diversidade de soluções. Para o problema de Atracamento Moleculas, algoritmos evolutivos são métodos estocásticos utilizados para encontrar um mínimo global de energia da ligação proteína-ligante, como Algoritmos Genéticos (GA-*Genetic Algorithms*), Evolução Diferencial (DE- Differential Evolution), Algoritmos Meméticos (MA-*Memetic Algorithms*), Optimização por Enxame de Partículas (OEP), Arrefecimento Simulado (SA-*Simulated Annealing*), Algoritmo de Colônia de Formigas, entre outros. As primeiras aplicações de Algoritmos Genéticos (GA) para o

problema de Atracamento Molecular foram desenvolvidos em (JUDSON et al., 1995). O princípio básico de algoritmos evolutivos é baseado em implementações de mutação, recombinação, seleção, e avaliações de aptidão em um conjunto de soluções para um determinado problema. Cada uma dessas operações são realizadas a fim de aumentar a diversidade de soluções e prever uma convergência prematura.

Algoritmos evolutivos oferecem diversas vantagens sobre métodos de otimização lineares, considerando que eles lidam com um variedade de soluções em um espaço de busca. Estes métodos requerem apenas o valor da função objetivo e conseguem lidar com problemas de otimização multi-modais(problemas que possuam muitos mínimos locais e a busca possa ficar estagnada em uma solução sub-ótima), descontinuidades no espaço de busca, valores da função objetivo com ruído ou problemas de mudanças dinâmicas (DEVI; SIVA; COUMAR, 2015). As soluções podem ter diferentes estruturas de dados, representações mistas e são utilizadas problemas onde a um espaço de busca complexo é definido (WEISE, 2009). A estrutura básica de um algoritmo evolutivo é apresentada no Algoritmo 1.

```
1  t ← 0 ;
2  inicializaP(t) ;
3  avaliaP(t) ;
4  while not termina(P(t)) do
5  |  P'(t) ← seleciona_I(P(t)) ;
6  |  P''(t) ← recombina_I(P'(t)) ;
7  |  P'''(t) ← muta_I(P''(t)) ;
8  |  Avalia(P''(t)) ;
9  |  P(t + 1) ← seleciona_II(P'''(t) ⋃ P(t)) ;
10 |  t ← t + 1 ;
11 end
12 Retorna melhor individuo X;
```

Algoritmo 1: Pseudocódigo simplificado de um algoritmo evolutivo

Algoritmos genéticos tem se destacado na solução de busca e otimização devido a uma série de vantagens em relação a outros métodos. São métodos estocásticos inspirados na genética e no processo de evolução natural. Uma das vantagens de um algoritmo genético é a simplificação que eles permitem na formulação e solução de problemas de otimização. Outras vantagens são a utilização de regras de transição probabilisticas, a possibilidade de utilizar funções não diferenciáveis, além de não requerer informações adicionais (como derivadas) sobre a função a otimizar. O uso de uma população de indivíduos, onde cada indivíduo representa uma possível solução para o problema, é uma

vantagem em relação a métodos determinísticos que convergem para uma única solução. GA's podem ainda ser hibridizados com outras técnicas, pois são algoritmos altamente adaptáveis.

A escolha do método de Atracamento Molecular é feita primeiramente analisando a complexidade do problema, representação dos dados, espaço de busca, entre outros aspectos matemáticos. São analisados, então, os métodos já desenvolvidos, muitos dos quais são amplamente divulgados através do CAPRI (*Critical Assessment of Prediction of Interactions*).

3.3 CAPRI

O CAPRI é um evento comunitário anual organizado pelo *European Bioinformatics Institute* (EBI) onde são submetidos trabalhos relacionados ao problema de Atracamento Molecular. Os experimentos do CAPRI apresentam importantes contribuições para o desenvolvimento de métodos de predição de interações proteína-proteína e ligante-proteína (VREVEN et al., 2013). Nesse experimento dados de uma estrutura são fornecidos aos participantes que realizam os testes em suas ferramentas. As coordenadas do complexo são divulgados após os testes, dando a oportunidade para grupos de pesquisa prever a conformação entre as moléculas. Os resultados são, então, divulgados e são feitas avaliações sobre os resultados de cada ferramenta. O evento oferece a oportunidade de participantes de testar seus métodos de predição em moléculas em que não foram publicados suas estruturas experimentais (JANIN, 2010).

No última edição do evento, em 2015, 11 complexos foram submetidos a ferramentas de 67 grupos de pesquisa na tentativa de predição da conformação dessas moléculas. Complexos com inibidores de enzimas, os quais apresentam poucas mudanças conformacionais, conseguiram atingir ótimos resultados, com valores abaixo de $1,5$ Å. Casos que envolvem grandes mudanças conformacionais não atingiram bons resultados ao predizer o modelo correto. Técnicas desenvolvidas focam no uso de meta-heurísticas, algoritmos evolutivos, diferentes funções de energia, métodos de busca baseados em Busca Conhecimento em Base de Bados. A análise de diversos métodos é feita na seção 3.4.

Analisar os resultados do CAPRI é uma maneira de conhecer os métodos utilizados atualmente e os resultados obtidos por diversos grupos de pesquisa. Além de métodos, são testadas funções de aptidão e técnicas computacionais que auxiliam na escolha de metodologias para a resolução problema de Atracamento Molecular.

3.4 Metaheurísticas utilizadas em Atracamento Molecular

Muitas ferramentas e métodos de Atracamento Molecular foram desenvolvidas ao longo de anos de pesquisa na área. O problema, no entanto, carece de uma solução que o resolva de maneira eficiente, considerando todas as variáveis biológicas e as limitações computacionais. A principal métrica de análise de soluções é em termos de Desvio Médio Quadrático das posições dos átomos (RMSD-*Root-mean-square deviation of atomic positions*) medida em angstroms (Å).

A maioria das técnicas faz, primeiramente, o *re-docking de estruturas*, método que utiliza complexos já atracados para testar a habilidade da ferramenta em predizer a conformação receptor-ligante. O Atracamento rígido é utilizado nessa etapa de forma a simplificar o problema e garantir que o algoritmo consiga resolvê-lo de maneira eficaz. Muitas vezes a flexibilidade da molécula ligante é aplicada gradualmente, adicionam-se a graus de liberdade gradualmente.

Durante os últimas duas décadas, diferentes meta-heurísticas tem sido aplicadas como métodos de busca para solucionar o problema de Atracamento Molecular (LAMEIJER et al., 2005). O trabalho de (SOUSA et al., 2013) revisa uma década de ferramentas desenvolvidas, onde diferentes estratégias foram aplicadas em aproximadamente 50 *softwares*. (CAMACHO et al., 2014) faz um estudo comparativo entre as meta-heurísticas mais utilizadas: Algoritmos Genéticos(GA) (LóPEZ-CAMACHO et al., 2013; MAGALHAES; BARBOSA; DARDENNE, 2004; JONES et al., 1997), Evolução Diferencial (DE) (KUKKONEN; LAMPINEN, 2005) e Otimização por Enxame de Partículas (PSO - *Particle Swarm Optimization*) (NEBRO et al., 2009; JANSON; MERKLE; MIDDENDORF, 2008). O tamanho da população, taxa de mutação, taxa de cruzamento, número de gerações, entre outros são parâmetros a ser considerados em algoritmos evolutivos.

No trabalho de Morris, (MORRIS et al., 2009), foram estudadas 180 estruturas baseados na proteína HIV-protease. A ferramenta desenvolvida, Autodock4, é uma das principais referências em *softwares* de Atracamento Molecular. A técnica desenvolvida baseia-se em um Algoritmo Genético Lamarckiano (AGL) e utiliza uma função semi-empírica de energia. Inicialmente foram utilizados 7 complexos com graus de liberdade sendo aos poucos adicionados. Os resultados dessa etapa chegaram a valores abaixo de $1,14$ Å, em 10 execuções para cada estrutura, o valor médio chegou a 0.88 Å, e desvio padrão de $0,25$ Å. A segunda etapa utilizou uma seleção de 170 estruturas, onde 100 delas chegaram a valores de RMSD abaixo de $3,5$ Å. O trabalho se destaca pelo uso de uma

grade de energia potencial em que é adicionado a flexibilidade parcial da proteína e uma função de energia própria.

No trabalho de Camacho, (CAMACHO et al., 2014), são comparadas 3 meta-heurísticas: Otimização por Enxame de Particulas (OEP), Evolução Diferencial (ED) e Algoritmos Genéticos (GA). No trabalho foi desenvolvido um *framework* que incopora a avaliação de energia da ferramenta Autodock 4.2. Foram testadas 83 estruturas, que incluem ligantes de diferentes tamanhos e flexibilidades. Os resultados chegaram em valores acima de 10 Å, onde cada estrutura passou por 30 execuções e 1500000 avaliações de energia. Nos experimentos o algoritmo de evolução diferencial chegou em uma convergência mais tardia em relação às outras técnicas, porém com melhores resultados. O GA demonstrou uma convergência rápida, porém suas soluções estagnam após 250000 avaliações de energia. O trabalho foca na comparação de algoritmos evolutivos aplicados ao problema de AM e sua avaliação de convergência.

No trabalho de Pippel, (MEIER et al., 2010), foi desenvolvido o *framework* ParaDocks que implementa um algoritmo de OEP e Otimização por Colônia de Formigas. A ferramenta opera paralelamente com uma Unidade de Processamento Gráfico (GPU-Graphics Processing Unit) e uma Unidade Central de Processamento (CPU-Central Processing Unit) fazendo a predição com ligantes flexíveis. O trabalho objetiva o uso de diferentes funções de energia em uma seleção de 13 estruturas. Em 73% das instâncias testadas os resultados chegaram em RMSD's abaixo de $2,0$ Å. A função de energia PMF04 (MUEGGE, 2006) se mostrou mais acurada entre as funções testadas GOLD (VERDONK et al., 2003), BLEEP (MITCHELL et al., 1999a; MITCHELL et al., 1999b) e DRUGSCORE (GOHLKE; HENDLICH; KLEBE, 2000). Os resultados mostram a eficácia de meta-heurísticas para o problema de AM e uma comparação entre funções de aptidão.

Técnicas de otimização utilizando meta-heurísticas variam em diversos aspectos. A parametrização de cada algoritmo leva a soluções distintas. Além do método de otimização, diversas variáveis, como o algoritmo de amostragem, representação de dados, representam grande influência nas solução alcançadas. Nesse trabalho os resultados são comparados com as ferramentas Autodock Vina (TROTT; OLSON, 2010) e Dockthor (MAGALHÃES et al., 2014), assim, nas Seções 3.4.1 e 3.4.2 são analisados esses dois trabalhos.

3.4.1 Autodock Vina

AudoDock Vina[1] é uma ferramenta de Atracamento Molecular e Triagem Virtual que oferece acurados resultados e função de energia para predições de conformações proteína-ligante. O *software* permite o uso de multi processadores, uma rápida avaliação de energia, e sua alta performance o torna uma das ferramentas mais citadas na área. No desenvolvimento do método de otimização foram testados uma variedade de abordagens incluindo GA, PSO e SA, até ser definido como algoritmo de busca a técnica de Busca Local Iterada (ILS-*iterated local search algorithm*).

O algoritmo utiliza uma sucessão de passos que incluem mutação e otimização local. A quantidade de passos é definido de forma adaptável em cada execução de acordo com a complexidade do problema. Diversas execuções são realizadas a partir de conformações randômicas das moléculas. A ferramenta utiliza *multithreading*, dessa forma, é possível utilizar o paralelismo de *hardware* com memória compartilhada. O algoritmo de otimização mantém uma seleção de mínimos locais relevantes encontrados e combina-os em execuções distintas utilizando-os em um processo de refinamento e agrupamento.

A ferramenta utiliza o formato de arquivo PDBQT de modo a fazer-se compatível com outras versões do *software*, além de compatível com ferramentas auxiliares, como AutoDock Tools (MORRIS et al., 2009), para a preparção de arquivos, escolha do espaço de busca e visualização dos resultados. Outras etapas de preparação previamente necessárias são dispensáveis já que o Autodock Vina calcula seu próprio *grid map* (MORRIS et al., 2009) e faz o agrupamentos e ranqueamento de resultados. Outras configurações previamente definidas é o número máximo de átomos de cada estrutura, o número de ângulos de rotação e o tamanho máximo do *grid map*, entre outros parâmetros que são fixados em tempo de compilação e se adaptam de acordo com a entrada.

Os testes para a validação da ferramenta incluíram 190 complexos proteína-ligante. Nos experimentos o receptor foi tratado como rígido e o ligante flexível com um número de ângulos de rotação variando de 0 a 32. Além da preparação das estruturas é necessário definir um arquivo de configuração definindo o tamanho do espaço de busca, selecionado como 15 Å, e o ponto central dessa área, definido manualmente em cada estrutura. Nas coordenadas do ponto central são adicionados 5 Å, a fim de garantir que o espaço de busca não está centralizado na estrutura experimental, o que tornaria a busca tendenciosa. Os resultados mostram valores de RMSD menores de 2 Å, em 78% das moléculas testadas.

[1]Disponível em: http://vina.scripps.edu/

3.4.2 Dockthor

A ferramenta Dockthor [2] implementa um Algoritmo Genético de multi-soluções, chamado de Torneio de Seleção Restrito Dinamicamente Modificado (DMRTS - *Dynamic Modified Restricted Tournament Selection*). O método utiliza um critério de inserção baseado na similaridade e um torneio dinâmico para preservar boas soluções e aumentar a diversidade na população do Algoritmo Genético.

O algoritmo *Steady State Genetic Algorithm* (SSGA) desenvolvido não faz a separação por gerações da população, já que cada recombinação criada é imediatamente testado para inserção na população. A população de possíveis candidatos evolui à medida que são aplicados os operadores genéticos. O critério de parada foi definido como um número máximo de avaliações de energia. A representação dos dados é feita por um vetor referente às translações, rotações e conformações da estrutura tridimensional dentro do sítio de ligação, onde cada possível solução é gerada randomicamente.

A partir da seleção de indivíduos e geração de uma recombinação, o método de DMRTS define a inserção dos mesmos na população. A exploração da diversidade de solução é fundamental para o SSGA prosposto, assim, indivíduos novos substituem soluções similares para aumentar a capacidade de busca do algoritmo. O critério de similaridade adotado é a distância euclidiana entre as soluções. Cada solução é ranqueada a fim de selecionar as melhores e piores soluções de cada agrupamento. Esse ranqueamento define a substituição ou não do vetor solução.

Os testes realizados pela ferramenta Dockthor incluíram o *re-docking* e *cross-docking* de 5 ligantes baseados na proteína HIV-protease, variando seus ângulos internos de 12 até 20. O método foi testado também em uma diversidade de 34 complexos proteína-ligantes de 18 famílias de proteínas. A comparação de performance foi realizada contra outras ferramentas de Atracamento Molecular: GOLD, Autodock Vina, GLIDE.

Considerando um limiar de $2,5$ Å, a ferramenta foi bem sucedida em $91,2\%$ dos testes, enquando GOLD e Autodock Vina atingiram $82,4\%$ e GLIDE $97,0\%$ de sucesso nas mesmas estruturas testadas. A ferramenta conseguiu ainda em 82.4% das estruturas resultados abaixo de $2,0$ Å. Os resultados indicam que o método realiza um amostragem eficaz do espaço de busca conformacional, produzindo uma boa diversidade de soluções. Os resultados indicam que o método realiza um amostragem eficaz do espaço de busca conformacional, produzindo uma boa diversidade de soluções.

[2] Disponível em: http://dockthor.lncc.br/

3.5 Desafios em Atracamento Molecular

O problema de Atracamento Molecular enfrenta diversos desafios tanto no âmbito de sua complexidade matemática como na capacidade de representação das interações físico-químicas dos complexos moleculares. Em (SOUSA et al., 2013) são enumerados três dos principais desafios para a predição da conformação proteína-ligante: o tratamento da flexibilidade da proteína; a presença de moléculas de água estruturais e seus efeitos no Atracamento; e a entropia de ligação. Em (HUANG; ZOU, 2010) são discutidos os desafios de amostragem do ligante e funções de energia acuradas para o problema. Em (MURRAY, 2007) são citados a representação das estruturas ligante e receptora, o papel de moléculas de água nas interações químicas e métodos de busca e sua velocidade.

A representação tridimensional da estrutura ligante deve levar em consideração os estados tautoméricos da molécula, já que ligantes mudam sua conformação ao fazer a ligação com proteínas. Muitas ferramentas consideram os ângulos torcionais de flexibilidade permitindo a rotação de ligações covalentes. Muitos programas realizam o Atracamento com o ligante rígido e umas série de conformações preestabelecidas. Outros aspectos geométricos durante testes preliminares de predição a fim de reduzir o número de graus de liberdade para o algoritmo de busca.

A representação tridimensional da proteína também enfrenta os desafios de possíveis estados tautoméricos da molécula, em particular para algumas estruturas são desafios os alternativos estados de protonação. Uma possível solução é a definição manual de possíveis estados baseado na análise do sítio de ligação e de conformações conhecidas da molécula ligante, embora essa solução possa não funcionar quando os estados de protonação variam com ligantes distintos. Outro importante aspecto é a flexibilidade do receptor, já que muitas ferramentas consideram-na como uma molécula rígida. Uma solução adotada por algumas metodologia é considerar certas conformações, porém essa técnica não considera o fato de que a conformação proteíca é modificada pela influência da molécula ligante. Outras metodologias incluem a flexibilidade parcial, calculando as contribuições energéticas da molécula no sítio de ligação.

A amostragem do ligante é um elemento básico do AM, onde dada um molécula alvo, o algoritmo de amostragem deve gerar orientações e conformações dentre do sítio de ligação. O sítio de ligação pode ser experimentalmente determinado ou manualmente definido. Existem três abordagens utilizadas para solucionar o problema de amostragem (HUANG; ZOU, 2010): combinação de formas, busca sistemática e algoritmos estocásti-

cos, explanados na seção 3.2.

Moléculas de água mediam as interações entre as estruturas receptoras e ligante. Em problemas de Atracamento Molecular, por vezes, água é considera como parte da proteína. O desafio, nesse caso, é determinar se para um certo ligante o potencial de mediação da partícula de água deve ser incluída no processo de Atracamento ou se deve ser eliminado. Três diferentes opções são adotadas por ferramentas de AM: omitir moléculas de água; permitir aquelas de contribuição energética relevante para o sistema; incluir todas as moléculas, considerando-as no modelo de avaliação de energia.

Encontrar uma função de energia acurada é um dos principais desafios no AM, já que muitas das avaliações de aptidão são inadequadas. Funções desenvolvidas e presentes em várias ferramentas conseguem reproduzir as ligações experimentais de 70%-80% dos complexos. Entretanto, quando adicionados a flexibilidade de proteínas e ligantes, suas topologias, e valência geométrica das moléculas, muitas vezes esses funções fazem uma avaliação longe do ideal. Em compensação, uma função rigorosa de energia seria muito custosa computacionalmente, considerando a análise de diversos modos de ligação. Assim, funções de energia assumem simplificações para mediar uma avaliação suficientemente acurada e um baixo custo computacional.

Efeitos de entropia tem uma grande contribuição para o cálculo de energia, que incluem a redução de graus de liberdade rotacionais e translacionais do ligante, mudanças na forma da proteína e do ligante, e no arranjo de camadas de água sobre os solutos. A entropia é, todavia, ignorada em muitas funções a fim de simplificar o cálculo de energia. A eliminação desse termo deve-se ao alto custo computacional, especialmente em ferramentas de Atracamento proteína-ligante, em que a eficiência computacional é um fator importante. Embora existam algumas tentativas de inclusão de entropia em funções de energia, a formulação incorporando esses termos ainda é um grande desafio.

Todos os desafios apresentados, além de considerados para o desenvolvimento de ferramentas de AM, devem ser ponderados com as variáveis de custo e capacidade computacional. Maximizar a acurácia de métodos de predição enquanto se minimiza o custo computacional requerido é um grande desafio para a área. Dessa forma, o método proposto nessa dissertação considera os desafios presentes no campo de pesquisa e propõe uma metodologia que busca aliar um bom desempenho computacional e representação biológica.

3.6 Conclusão

Diversas técnicas vem sendo aplicadas para o problemas de Atracamento Molecular, porém o problema ainda carece de solução generalizada. O uso de algoritmos de otimização e de funções de energia acuradas são chaves principais para a resolução do problema. Assim, desenvolver um método de Atracamento requer a investigação de metodologias já desenvolvidas, suas vantagens e buscar técnicas ainda não exploradas no campo de pesquisa. Desde a representação dos dados, categorias de métodos e os desafios encontrados foram analisados e ponderados a fim de desenvolver uma metodologia que contribua para a resolução do problema. Duas das ferramentas abordadas desenvolvem técnicas que alcançaram bons resultados e, então, seus resultados são utilizados a fim de comparação com a ferramenta desenvolvida.

4 MÉTODO PROPOSTO

O método proposto objetiva analisar as variáveis biológicas e desafios em Atracamento Molecular para criar uma abordagem utilizando uma meta-heurística para o problema de otimização, um algoritmo de amostragem que gere conformações proteína-ligante e um modelo de exploração do espaço de busca que gere soluções diversificadas para o problema. Primeiramente, é realizada a preparação das estruturas selecionadas para os testes, onde os dados são analisados e é definida uma representação para as soluções. A segunda etapa trata da definição de uma função de energia, aliando uma análise eficiente das ligações químicas com um baixo custo computacional, sendo definida uma função energia para avaliação da qualidade das soluções geradas. Em seguida, é analisado o problema de otimização, nesse ponto é explorada a metodologia desenvolvida: a discretização do espaço de busca, Algoritmo Genético de Chaves Aleatórias Viciadas com agrupamento e competição entre soluções. Essa abordagem traz vantagens ao gerar uma diversidade de soluções pois explora um campo de busca discretizado, fazendo com que o algoritmo de amostragem gere conformações em todo a área de busca na proteína, aliado a aplicação de algoritmo evolutivo robusto para o problema de Atracamento Molecular.

4.1 Preparação e representação das estruturas moleculares

Para cada arquivo adquirido no banco de dados PDB foi realizada a preparação das estruturas a fim de obter uma representação mais acurada do problema biológico de forma computacionalmente factível, além de definições sobre os graus de liberdade e efeitos dos solventes. A preparação passa pelas etapas de: remoção e adição de átomos, inclusão de cargas e conversão do arquivo. A partir dessa preparação é feita a representação da estrutura para o algoritmo de otimização.

A ferramenta Pymol (Schrödinger, LLC, 2015) foi utilizada para a visualização das moléculas, remoção e adição de átomos. Das estruturas cristalográficas foram removidos moléculas pequenas, como solventes, íons sem interações com o complexo, água, entre outros. Essa remoção foi definida para simplificação do cálculo de energia, devido a complexidade da inclusão desses elementos na formulação. Dessa forma, considera-se que a proteína encontra-se no vácuo.

Para a segunda parte da preparação a ferramenta AutodockTools (ADT) (MORRIS et al., 2009) foi utilizada para gerar o arquivo com as coordenadas de cada átomo. Os

arquivos foram convertidos para o formato PDBQT. Nesse formato, o arquivo representa o ligante adicionado valores de cargas de cada átomo, informações sobre as ligações químicas presentes na molécula e sobre ângulos diedrais ativos para rotação. Nessa etapa foi definido um limite de 10 ângulos torcionais para a molécula ligante, esse valor foi definido a fim de determinar uma complexidade máxima para o problema. Algumas moléculas possuem muitos ângulos de rotação internos, entretanto todos os complexos escolhidos para realização dos testes possuem, no mínimo, 10 ângulos diedrais. Foram adicionados ao ligante átomos de cargas parciais e hidrogênios utilizando a ferramenta Open Babel (O'BOYLE et al., 2011). No formato PDBQT são unidas as moléculas de carbono com as moléculas de hidrogênio, representando uma molécula de carga equivalente, cujas coordenadas do átomo de carbono são mantidos. A molécula receptora representa os átomos de hidrogênio e carbono da mesma forma.

A última parte da preparação é a definição de um arquivo de configuração. São definidos nesse arquivo o ponto central do espaço de busca. Além das coordenadas x, y e z do ponto central é definida a largura, altura e comprimento da caixa que limita o espaço de busca. Para isso, para cada estrutura é testado um cubo virtual utilizando a ferramenta ADT, esse espaço deve conter o sítio de ligação e espaço suficiente para o deslocamento do ligante. Foi definido para a estrutura *HIV-protease* um cubo de 11 Å, pois foi considerado suficientemente grande para englobar quase toda a proteína e incluir seu sítio ativo.

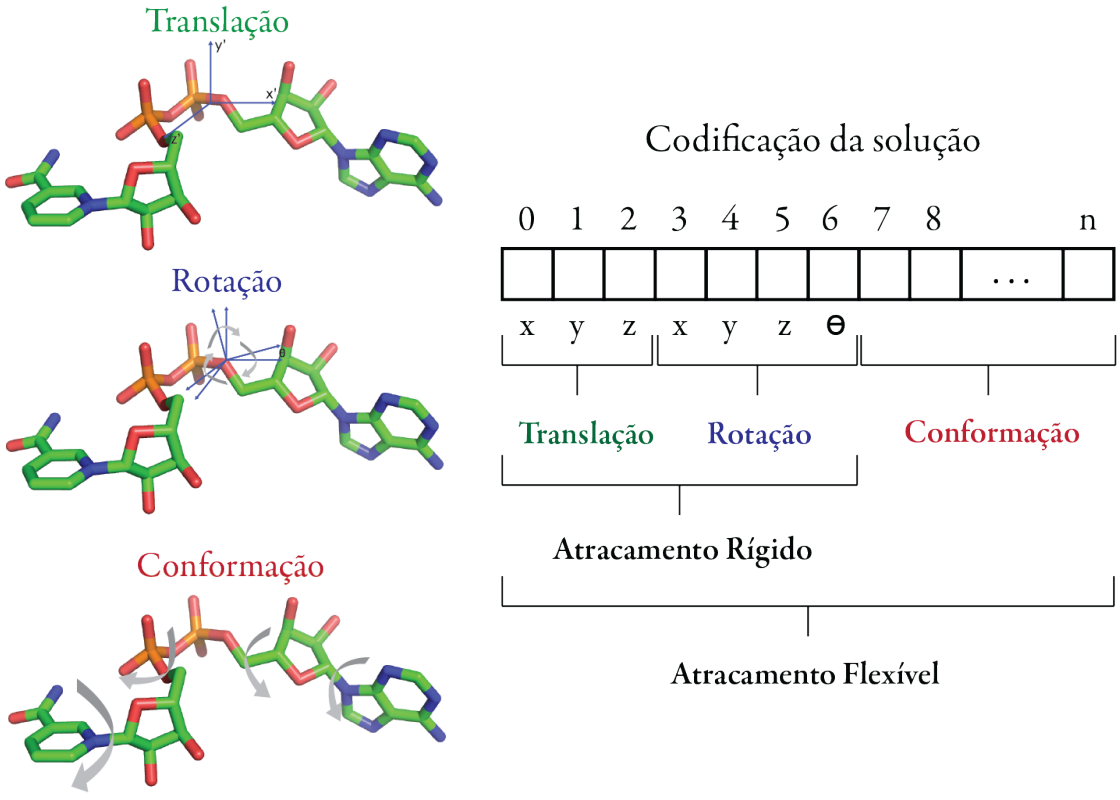
A Figura 4.1 apresenta a representação da solução adotada para o problema. A codificação é representada com um vetor de três coordenadas x,y e z referentes a translação da molécula, quatro parâmetros rotacionais, x, y, z e θ referentes as coordenadas do vetor de rotação e um ângulo de rotação global da estrutura ligante, e os ângulos diedrais referentes aos graus de liberdade da molécula ligante. A operação de rotação é realizada nas ligações covalentes, porém o vetor é fixo na direção da ligação química, assim é preciso inferir apenas um ângulo de rotação na codificação da solução. A Figura 4.1 mostra, ainda, a representação dos ângulos diedrais de uma molécula ligante e por fim o representação do processo de Atracamento Molecular proteína-ligante.

O algoritmo evolutivo desenvolvido utiliza o conceito de população, ou conjunto de soluções. Assim, cada vetor-solução é chamado de indivíduo, que contém posições, genes, referentes as translações, rotações e conformações do ligante. A quantidade de posições desse vetor n varia de acordo com o número de ângulos diedrais do ligante. Cada indivíduo representa uma solução dentro do espaço de busca. Assim, cada posição

recebe um valor randômico dentro dos limites do campo de busca, essa solução é avaliada de acordo com uma função de aptidão.

Em testes de Atracamento Molecular são realizados, primeiramente, testes de Atracamento rígido, como uma etapa prévia ao Atracamento flexível. Essa etapa visa simplificar o problema e testar a capacidade do algoritmo de otimização. O Atracamento rígido considera apenas os graus de liberdade translacionais e rotacionais do ligantes, assim, são gerados valores randômicos apenas para os 7 primeiros genes, como ilustra a Figura 4.1. O Atracamento flexível adiciona ao vetor solução as posições restantes para gerar cada solução. Todo deslocamento é realizado a partir do preenchimento das posições do vetor com valores randômicos e é realizado a partir de um ponto referência definido como o centro do campo de busca, selecionado para cada estrutura no arquivo de configuração.

Figura 4.1: Codificação da solução: x,y e z são valores de translação de toda a estrutura a partir de uma molécula de referência, os próximos 4 valores x, y e z e θ representam o vetor unitário e o ângulo de rotação para a molécula, os seguintes valores representam os ângulos diedrais referentes a rotação das ligações covalentes da estrutura ligante.



4.2 Função de energia utilizada

A metodologia de Atracamento Molecular proposta nesse trabalho utiliza uma abordagem de avaliação baseada em energia. Considerar todos os graus de liberdade do complexo proteína-ligante e todas as variáveis químico-físicas teriam um custo computacional elevado (PEARLMAN; CHARIFSON, 2001). Assim, são realizadas simplificações e aproximações de modelos de energia, para chegar em um tempo de execução viável. Como parte do processo de desenvolvimento de fármacos, métodos computacionais de Atracamento objetivam automatizar etapas, cujas soluções podem posteriormente ser refinadas.

Para avaliar a qualidade da solução o algoritmo utiliza uma função de aptidão, que mede a energia de interação do complexo proteína-ligante. Nesse trabalho a energia é calculada utilizando a função de energia do Autodock Vina (TROTT; OLSON, 2010). A função calcula o energia total de ligação do complexo. As Equações 4.1, 4.2 e 4.3 descrevem o cálculo de energia utilizado.

$$\Delta G = (V^{L-L}_{bonded} - V^{L-L}_{unbonded}) + (V^{R-R}_{bonded} - V^{R-R}_{unbonded}) + (V^{R-L}_{bonded} - V^{R-L}_{unbonded} + \Delta G_{conf}) \quad (4.1)$$

$$\begin{aligned} V = W_{vdw} \sum_{i,j} (\frac{A_{ij}}{r^{12}_{ij}} - \frac{B_{ij}}{r^{6}_{ij}}) + W_{hbond} \sum_{i,j} E(t) (\frac{C_{ij}}{r^{12}_{ij}} - \frac{D_{ij}}{r^{10}_{ij}}) + \\ W_{elec} \sum_{i,j} \frac{q_i qj}{\epsilon(r_{ij}) r_{ij}} + W_{sol} \sum_{i,j} (S_i V_j + S_j V_i) e^{\frac{-r^2_{ij}}{2\sigma^2}} \end{aligned} \quad (4.2)$$

$$\Delta G_{conf} = W_{conf} N_{tors} \quad (4.3)$$

A Equação 4.1 representa a função de energia livre que utiliza a transição dos estados de ligados e não-ligados para calcular a energia de interação entre o ligante (L) e o receptor (R). O campo de força nessa equação realiza seis avaliações de potencial e considera o termo conformacional de entropia. Basicamente o cálculo total de energia considera diferentes estados de cada molécula e de suas interações, adicionando-se, por fim, o termo de entropia do sistema que ocorre no último termo, de interação entre os compostos químicos.

Na Equação 4.2 os pesos W_{vdw}, W_{hbond},W_{conf}, W_{elec} e W_{sol} representam ligações de hidrogênio, forças torcionais, interações eletrostáticas e dessolvatação, respectivamente. A variável r_{ij} corresponde a distância entre cada átomo do complexo. Os

parâmetros de *Lennard-Jonnes* para os potenciais máximos entre dois átomos são representados por A_{ij}, B_{ij} , C_{ij} e D_{ij}. No segundo termo da Equação, o termo $E(t)$ representa o direcionalidade dependente do ângulo na ligação de hidrogênios. No terceiro termo são calculados as variáveis eletrostáticas de *Coulomb*, e finalmente, o quarto termo é calculado a partir do volume (V) de átomos ao redor do complexo, dados por um peso S.

A Equação 4.3 apresenta o cálculo do termo de entropia da Equação 4.1. O termo N_{tors} é referente a variação da entropia de acordo com as mudanças conformacionais das moléculas receptora e ligante. O termo W_{conf} é o peso dado para o essa equação e, por consequência, para sua influência no cálculo total da energia.

Em cada termo das Equações 4.1, 4.2 e 4.1 há valores de pesos associados. Essas variáveis determinam a influência daquele termo para a equação. O termo de *Lennard Jonnes* r^{-12}, por exemplo, por vezes eleva em muito o valor de cálculo do potencial nas interações do ligante e do receptor, fazendo com conformações muito próximas da experimental tenham valores de energia altos. Assim, é possível ponderar a influência desse termo no cálculo.

Para a utilização da função proposta, os arquivos devem estar em formado compatível com a Autodock Vina, assim, a etapa de preparação das estruturas é obrigatória e deve preceder o cálculo de energia. Os arquivos representando as moléculas devem estar no formato PBDQT, com as coordenadas de cada átomo, informações sobre as ligações químicas e valores de carga.

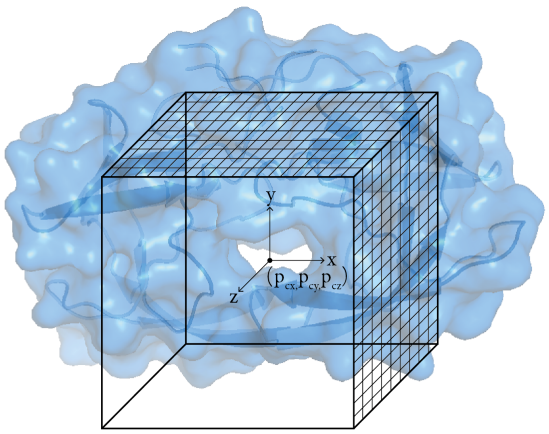
O objetivo de otimização problema de Atracamento Molecular é encontrar o mínimo global da função de energia. Para todas as soluções geradas pelo algoritmo, é realizado o cálculo de energia. Considerando o número de soluções geradas pelo algoritmo, essa tarefa é uma das mais custosas computacionalmente. Assim, são aplicadas metologias para simplificar o cálculo, tanto matematicamente ao considerar os termos mais relevantes para o cálculo de energia livre, como através de métodos computacionais. Como é o caso da utilização de grades de energia, (ZHANG et al., 2008; MENG; SHOICHET; KUNTZ, 1992; LUTY et al., 1995), e exploração do espaço de busca (MAGALHAES, 2006) . Ao considerar a dificuldade do problema o uso de meta-heurísticas é altamente recomendável (BLUM et al., 2011). A meta-heurística desenvolvida, detalhada na Seção 4.6, explora o espaço conformacional com um grande número de combinações e uma convergência rápida.

4.3 Proposta de descrição do espaço de busca

Nesse trabalho é adotada uma metodologia de inclusão de flexibilidade do receptor apenas no sítio de ligação. Para essa abordagem é necessário avaliar um tratamento as interações atômicas de longo alcance. Considerando N como o número de átomo nas complexo molecular, o custo computacional para o cálculo de energia cresce N^2, deve-se levar em consideração que esse complexo conta com milhões de átomos. Considerar todos os átomos para o cálculo de energia seria computacionalmente inviável. Dessa forma, a metodologia de Atracamento Molecular baseada em grade (LUTY et al., 1995) todos os potenciais relacionados ao átomo receptor são pré-calculados e armazenados em cada ponto de uma grade/malha tridimensional construída para englobar o sítio-ativo do complexo receptor (MENG; SHOICHET; KUNTZ, 1992).

Nessa metodologia cada ponto da grade calcula previamente a influência eletrostática e dos potenciais de *Lennard Jonnes*, considerando a molécula receptora como rígida. Dessa forma, não são negligenciados os átomos mais distantes do sítio de ligação, ou seja, as ligações de longo alcance, já que essas interações são fundamentais para o processo de atracamento. Os valores de cada ponto são determinados por uma interpolação dos potenciais calculados. Na Figura 4.2 é ilustrado o espaço de busca delimitado por uma grade de energia. A metodologia aplicada nesse trabalho utiliza para o cálculo da grade apenas os

Figura 4.2: Espaço de busca delimitado por uma grade de energia. Nessa representação são calculadas as contribuições energéticas da molécula receptora em cada ponto da grade e é adicionado a flexibilidade parcial da molécula receptora



átomos da molécula receptora. Assim, o cálculo agrega todas as interações moleculares da proteína, incluindo as de longo alcance, e consegue reduzir o custo computacional, já

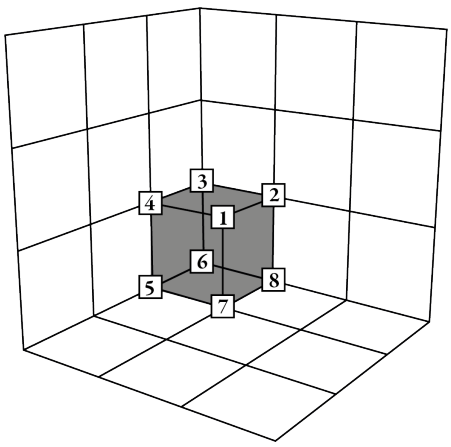
que não é necessário incluir todas as moléculas da proteína no cálculo de energia durante a execução do algoritmo.

A interpolação realizada em cada ponto da grade considera os termos de longo alcance através da formação de células. Cada célula da grade possui oito pontos localizados nos vértices de um cubo. A discretização da célula é feita por meio da Equação 4.4.

$$\begin{aligned} P_1 &= [x(i), y(j), z(k)] \\ P_2 &= [x(i+\Delta t), y(j), z(k)] \\ P_3 &= [x(i), y(j+\Delta t), z(k)] \\ P_4 &= [x(i), y(j), z(k+\Delta t)] \\ P_5 &= [x(i+\Delta t), y(j+\Delta t), z(k)] \\ P_6 &= [x(i+\Delta t), y(j), z(k+\Delta t)] \\ P_7 &= [x(i), y(j+\Delta t), z(k+\Delta t)] \\ P_8 &= [x(i+\Delta t), y(j+\Delta t), z(k+\Delta t)] \end{aligned} \tag{4.4}$$

A Equação 4.4 mostra a discretização que é feita de maneira uniforme, em cada um dos oito pontos P_N é adicionado uma translação Δt. Essa translação parte de um ponto referencial, a origem do espaço de busca, e permite uma indexação automática de cada célula da grade. A Figura 4.3 ilustra a indexação dos oito pontos da célula para o cálculo da energia potencial.

Figura 4.3: Pontos indexados em uma célula da grade de energia (MAGALHAES; BARBOSA; DARDENNE, 2004)



A energia potencial calculada em cada célula considera os termos não-ligados, que geram um alto custo computacional, como rígidos. O potencial devido a presença

desses átomos é avaliada em qualquer ponto p segundo a função de energia apresentada na Seção 4.2. Para os outros átomos da molécula receptora, os potenciais Φ são calculados de acordo com as Equações 4.5, 4.6 e 4.7.

$$\Phi^A(p) = \sum_j \frac{(A_{ij})^{\frac{1}{2}}}{r_{jp}^{12}} \tag{4.5}$$

$$\Phi^B(p) = \sum_j \frac{(B_{ij})^{\frac{1}{2}}}{r_{jp}^{6}} \tag{4.6}$$

$$\begin{aligned} \Phi^{ES}(p) &= \sum_j \frac{(q_j)}{\epsilon r_{jp}} \\ &= \sum_j \frac{(q_j)}{r_{jp}^2} \end{aligned} \tag{4.7}$$

O quarto termo de dessolvatação é calculado pelas Equações 4.8 e 4.9.

$$\Phi^{DES,EXP}(p) = \sum_j f_i \ \ exp(\frac{-r_{jp}^2}{2\sigma^2}) \tag{4.8}$$

$$\Phi^{DES,MAS}(p) = \sum_j S_j \ \ exp(\frac{-r_{jp}^2}{2\sigma^2}) \tag{4.9}$$

Na Equação 4.5 considera-se j como cada átomos rígido, e cada i como átomo móvel. O parâmetro de solvatação S_i e o volume f_i, é localizado no ponto da grade p, então a dessolvatação explicita do átomo i, por átomos rígidos, é calculada pela Equação 4.8. A solvatação por átomos rígidos pelo átomo i é igual a $f_i\Phi^{DES,MAS}(p)$.

A grade corresponde ao potencial das ligações de hidrogênio e é construída utilizando o mesmo equacionamento da função de energia. Por fim, para calcular o potencial de qualquer ponto no espaço é utilizado a interpolação de oito pontos, cuja Equação 4.10 representa a interação total de átomos rígidos e móveis. Um campo de força pode ser calculado pelo gradiente negativo da grade de potencial energético. O valor de força associado a cada átomo é estimado pela interpolação dos valores das derivadas negativas.

$$\begin{aligned} E = \sum_i (A_{ij})^{\frac{1}{2}}\Phi^A(x_i) + (B_{ij})\Phi^B(x_i) + q_i\Phi^{ES}(x_i) + \\ S_i\Phi^{DES,EXP}(x_i) + f_i\Phi^{DES,MAS}(p) \end{aligned} \tag{4.10}$$

A célula é construída para englobar os átomos do receptor, que pode conter parte ou toda a molécula. A grade eletrostática é formada por um volume cúbico e deve ser posicionado por uma determinada coordenada. Assim, é possível definir uma caixa posicionada no espaço de busca que inclui o sítio de ligaçao da proteína. O mapeamento da grade de energia é calculado pela ferramenta Autdock Vina. O arquivo de configuração permite a definição do tamanho da área de busca, delimitando área com um valor em angstroms (Å), e do ponto central, que será posicionada a caixa.

O espaço de busca definido pela grade de energia representa a figura de um cubo ou uma caixa. O tamanho definido deve englobar o sítio de ligação e deve ter espaço suficiente para permitir as translações, rotações e conformações da molécula ligante. O centro do campo de busca é posicionado em um ponto randômico da proteína, a partir desse ponto são adicionados em uma direção aleatória o valor de 5 Å, garantindo que o centro não coincida com o ponto exato de ligação da molécula ligante, e assim, tornar a busca não tendenciosa. A partir do tamanho e centro desse espaço definido por um arquivo de configuração é definida a discretização da área de busca.

O volume da caixa é definido a partir de três variáveis: Δx, Δy e Δz, correspondentes a altura, largura e comprimento, respectivamente. As medidas do tamanho do cubo são medidas em Angstrons (Å) e ao multiplicarem-se formam o volume total de uma área discretizada. O ponto central é definido pelas variáveis: p_{cx}, p_{cy} e p_{cz}, ou seja, coordenadas x, y e z. A partir dessas variáveis foram calculados volumes correspondentes a cubos menores dentro do espaço total de busca. A Equação 4.11 representa o cálculo do volume V_c para um o cubo central do campo de busca V_{c1}.

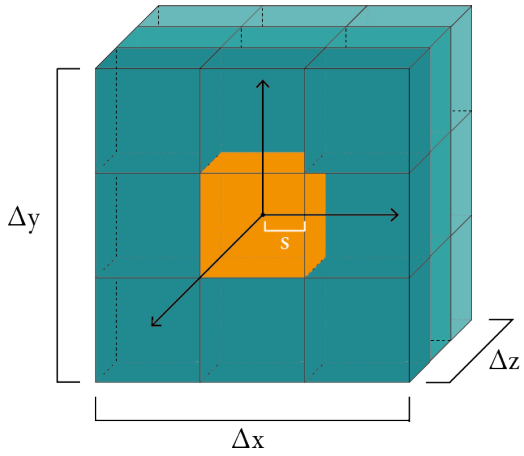
$$\begin{aligned} V_{c1} = [|p_{cx} + \frac{\Delta x}{s}| + |p_{cx} - \frac{\Delta x}{s}|] \times \\ [|p_{cy} + \frac{\Delta y}{s}| + |p_{cy} - \frac{\Delta y}{s}|] \times \\ [|p_{cz} + \frac{\Delta z}{s}| + |p_{cz} - \frac{\Delta z}{s}|] \end{aligned} \tag{4.11}$$

A Equação 4.11 representa o cálculo de volume do cubo central da área de busca, em destaque (laranja) na Figura 4.4. A variável s representa o número de regiões em que o espaço é dividido, ou seja, a variável s influencia o número de cubos formados. A partir da análise das estruturas e definição do tamanho da caixa, totalmente dependente dos tamanhos do ligante e sítio de ligação da proteína, é definido o valor da variável s. Na

Equação 4.11 é apresentada a formulação de um cubo, para o o cálculo de V_{cn} cubos são considerados vetores criados nas extensões dos valores de Δx, Δy e Δz.

A Figura 4.4 ilustra o espaço de busca discretizado com valor de $s = 6$. Segundo Equação 4.11 os valores de Δx, Δy e Δz são divididos gerando três cubos por aresta. Cada aresta é dividida pelo valor de s e cada espaço segue a formula de cálculo dos cubos. Considerando altura, largura e comprimento, calculados pelo módulo do ponto central mais um deslocamento nos três eixos, o número total de mini-cubos é $\beta = 27$. A discretização do espaço de busca é utilizada como critério de similaridade para a formação de agrupamentos. A divisão de cada espaço define uma translação em relação ao ponto central, assim, o número de mini-cubos coincide com o número de agrupamentos. A criação de um agrupamentos de soluções é descrita na Seção 4.5.

Figura 4.4: Discretização do espaço de busca. Em laranja é destacado a região central do cubo criado a partir da Equação 4.11



A partir da discretização desse espaço, soluções são geradas ocupando todas as regiões do campo de busca. O algoritmo, então, gera soluções randômicas, utilizando a representação proposta na Seção 4.1. Os três primeiros posições do vetor que representa a molécula ligante representam os graus de liberdade de translação, estes valores são limitados pelo espaço de busca, já que a molécula deve permanecer por completo dentro da área de busca. Além de explorar essa área, a discretização é utilizada como critério de similaridade para o agrupamento dessas soluções.

4.4 Algoritmo Genético de Chaves Aleatórias Viciadas

Problemas de otimização global buscam o mínimo ou o máximo de uma função em um domínio discreto ou contínuo. A função de aptidão aplicada, Seção 4.2, para a avaliação da qualidade de ligação entre as moléculas receptoras e ligante é caracterizada por ser: multi-modal, contínua, diferenciável e sujeita a restrições do espaço de busca devido a limitação por um espaço factível biologicamente. Assim, a solução $x^* \in S \subset \mathbb{R}^n$, onde S é uma região definida como espaço de busca, explicada na Seção 4.3. O resultado ótimo é formulado como: $f(x^*) \leq f(x) | \forall x \in S$, onde a função objetivo é definida $f : S \longrightarrow \mathbb{R}$.

A função $f(x)$ para o problema de otimização e suas restrições são apresentadas na Equações 4.12 e 4.13, respectivamente:

$$min \;\; f(x), x = x_1, x_2, ..., x_n \tag{4.12}$$

Sujeito a:

$$\begin{aligned} g_i(x) &\subseteq [\frac{-\Delta x}{2}, \frac{\Delta x}{2}] \\ h_i(x) &\subseteq [\frac{-\Delta y}{2}, \frac{\Delta y}{2}] \\ l_i(x) &\subseteq [\frac{-\Delta z}{2}, \frac{\Delta z}{2}] \end{aligned} \tag{4.13}$$

As restrições $g_i(x)$, $h_i(x)$ e $l_i(x)$ são referentes a espaço de busca delimitado por um tamanho que varia de $\frac{-\Delta x}{2}$ a $\frac{\Delta x}{2}$ no plano x e as mesma variações para os planos y e z, respectivamente.

A partir da definição da função a ser otimizada e suas restrições, e considerando a função de energia proposta na Seção 4.2, foi desenvolvido um Algoritmo Genético de Chaves Aleatórias Viciadas (BRKGA - *Biased Random-Key Genetic Algorithms*), algoritmo primeiramente proposto por, Bean (BEAN, 1994). Esses algoritmos, uma variação de GA's, tem sido aplicados para diversos problemas de otimização (NORONHA; RESENDE; RIBEIRO, 2011; RESENDE, 2012; GOULART et al., 2011; PRASETYO; FAUZA G.; LEE, 2015). Eles utilizam o conceito de chaves aleatórias para codificar a solução do problema de otimização, representando as possíveis soluções e, como em GA's, a meta-heurística trabalha com diversos indivíduos em uma população.

Em um Algoritmo Genético indivíduos, chamados de cromossomos, são soluções

geradas de forma aleatória para formar uma populção. O algoritmo utiliza o conceito de sobrevivência do mais adequado, onde são eliminadas soluções em que a função de aptidão retorna os piores resultados. Cada iteração do método é chamada de geração, onde são aplicadas operações de mutação e cruzamento. As iterações ocorrem em laço até que um critério de parada seja satisfeito.

Algoritmos Genéticos com Chaves Aleatórias (RKGA) representam suas soluções como vetores em um determinado intervalo. Um algoritmo decodificador organiza os vetores e a avalia a aptidão de cada solução. A estratégia de decoficicação da solução é dependente do problema (TANGPATTANAKUL; JOZEFOWIEZ; LOPEZ, 2013). A população de vetores-solução evolui em iterações chamadas de gerações. A população inicial é definida por um número p de vetores. Cada componente da solução, posição do vetor, é chamado de gene, cada gene é gerado de forma independente.

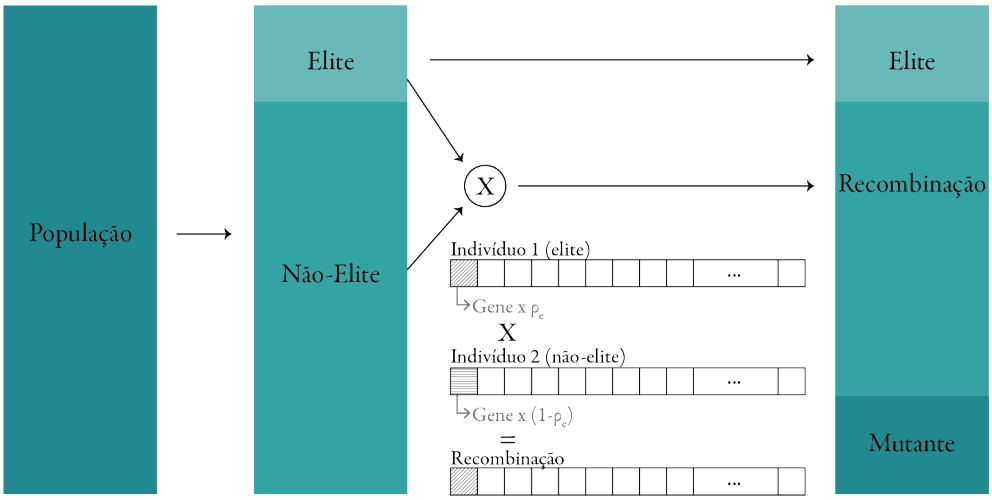
Após a avaliação de energia de cada elemento da população e organização das mesmas pelo algoritmo determinístico decodificador, a população é particionada em dois grupos: grupo de elite (p_e), os quais obtiveram as melhores avaliações de energia em cada agrupamento do espaço de busca, e o grupo de não-elite ($p - p_e$ ou $p_{\overline{e}}$).

Em cada geração (k) são formados novos indivíduos através das operações de cruzamento e mutação de modo a evoluir a população. Primeiramente, todos os indivíduos do grupo de elite (p_e) são copiado para a nova população da geração $k + 1$ sem qualquer modificação. Então, é realizada a operação de mutação, onde são gerados um número de vetores de chaves aleatórias p_m, ou vetores-solução, da mesma forma como é inicializada a população. Com isso, novos indivíduos randômicos são permanentemente introduzidos na população, mantendo a diversidade da mesma. Após a introdução de p_e e p_m é a realizada a operação de cruzamento nos em $p - p_e - p_m$ indivíduos necessários para completar a população p. Essa operação gera uma recombinação resultante da mescla das posições dos vetores-solução de dois indivíduos.

A Figura 4.5 ilustra a dinâmica de evolução em um BRKGA. Na esquerda é apresentada a população inicial dividida em elite, indivíduos de menor valor de energia em cada agrupamento, e não-elite, restante da população. A população de elite é copiada sem alterações para a população da geração $k + 1$ (na direita). A recombinação é o resultado do cruzamento entre um indíviduo da elite e um indivíduo do resto da população. Os indivíduos mutantes completam a população que é reformulada em cada geração. Por fim, o algoritmo decodificador organiza os novos vetores-solução de acordo com a avaliação da função de aptidão, já que os novos indivíduos podem migrar para a população de elite,

assim como indivíduos da elite podem mudar de grupo. As soluções são dinamicamente realocadas na população, e os piores indivíduos são descartados em cada geração.

Figura 4.5: Dinâmica da evolução em um BRKGA: divisão da população, operações de cópia do grupo de elite, cruzamento gerando a recombinação de soluções e operação de mutação



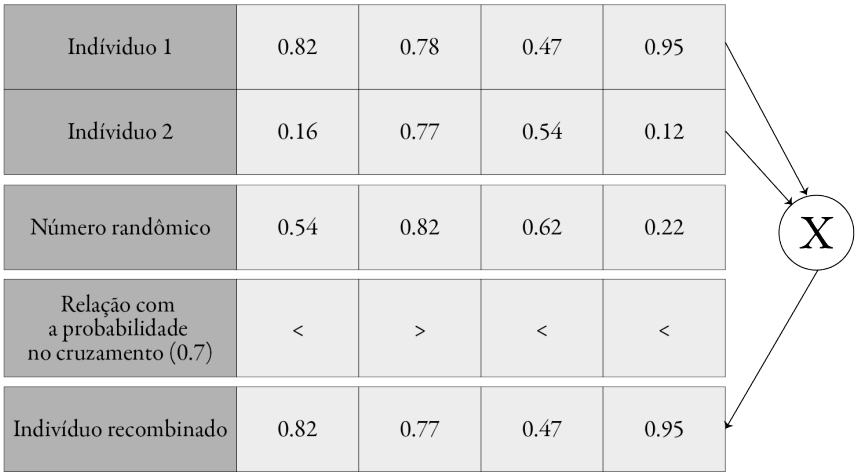
A operação de cruzamento é caracterizada pela seleção de genes através de critérios probabilísticos. O primeiro individuo é selecionado do grupo p_e da elite, o segundo é necessariamente um indivíduo do restante da população, ou $p_{\overline{e}}$. A probabilidade de um gene de um indivíduo do grupo de elite ser selecionado é maior que de um gene da população de não-elite. O fator que contribui para essa seleção é o cruzamento parametrizado uniforme (em inglês, *Parametrized Uniform Crossover*) incorporado no BRKGA.

Seja ρ_e a probabilidade de um gene do indivíduo de elite ser selecionado e n o número de posições do vetor-solução, é definida uma probabilidade maior para esse gene em todas as posições. Por exemplo, caso $\rho > 0.7$ o indivíduo de elite tem 70% de chance de ter seu gene selecionado. Ou seja, para $i = 1, ..., n$, onde o iésimo compontente da variável de recombinação $c(i)$ tem a probabilidade ρ_e de selecionar um componte e_i do indivíduo da elite. A probabilidade de um gene da população de não-elite ($\overline{e}$) é igual a $1 - p_{\overline{e}}$.

A Figura 4.6 ilustra a operação de cruzamento com valores reais atribuídos aos vetores-solução, ou indivíduos. Os dois indivíduos possuem 4 posições, para o cada posição é gerado um número randômico entre 0 e 1, como se uma moeda fosse jogada. Essa moeda é, porém, viciada de acordo com um valor de probabilidade atribuído a ela, $\rho_e = 0.7$ no exemplo. Caso o número aleatório relacionado com o critério de probabilidade seja menor, é selecionado o gene do primeiro indivíduo, caso contrário é selecionado

o gene do segundo indivíduo. O resultado final é um indivíduo recombinado, com suas posições vetoriais mescladas entre as duas soluções.

Figura 4.6: Cruzamento parametrizado uniforme em um BRKGA



A Referência (GOULART et al., 2011) sugere a utilização dos valores da Tabela 4.1 para parametrização do algoritmo.

Tabela 4.1: Valores recomendados para os parâmetros do BRKGA

Parâmetros	Valor recomendado
p	$p = a.n$
p_e	$0,10p \leq p_e \leq 0,25p$
p_m	$0,10p \leq p_m \leq 0,30p$
ρ_e	$0,5p \leq \rho_e \leq 0,80p$

4.5 Agrupamento e competições global e local

O agrupamento é uma importante técnica na biologia computacional (KOZAKOV et al., 2005). Métodos de agrupamento estendem o uso de meta-heurísticas ao promover a formação de sub-populações de soluções similares. Esses métodos vêm sido desenvolvidos para reduzir os efeitos do desequilíbrio genético resultantes do operador de seleção de AG (SARENI; KRAHENBUHL, 1998). O desequilíbrio genético é uma mudança na frequência de variação genética ema população devido a amostragem aleatória de indivíduos. A geração de novas soluções, cruzando valores de outras soluções, método

conhecido como cruzamento, tem na variação desses valores um papel fundamental nas operações de algoritmos evolutivos. O uso de nichos mantém a diversidade da população e permite que o AG investigue diferentes áreas do espaço conformacional em paralelo. Além disso, evita um dos principais problema em AG, quando os processos evolutivo convergem rápido demais para uma solução que possa estar presa em um mínimo local (JASSADAPAKORN; CHONGSTITVATANA, 2011).

O método de agrupamento desenvolvido utiliza como critério de similaridade os valores de translação das soluções. Assim, segundo a representação apresentada na Seção 4.1, os três valores iniciais do vetor de solução são variações translacionais que se encaixam dentro do espaço de busca. Ao gerar uma nova solução, seja por inicialização da população, cruzamento ou mutação, é verificado, dentro do espaço discretizado, a qual grupo pertence o novo vetor. É importante ressaltar que a operação de cruzamento possibilita que soluções de dois grupos gerem uma solução que se encaixe em um terceiro grupo distinto, dessa forma, mantendo a diversidade da população. Apenas soluções de mesmo grupo, quando cruzadas, geram uma solução no mesmo grupo.

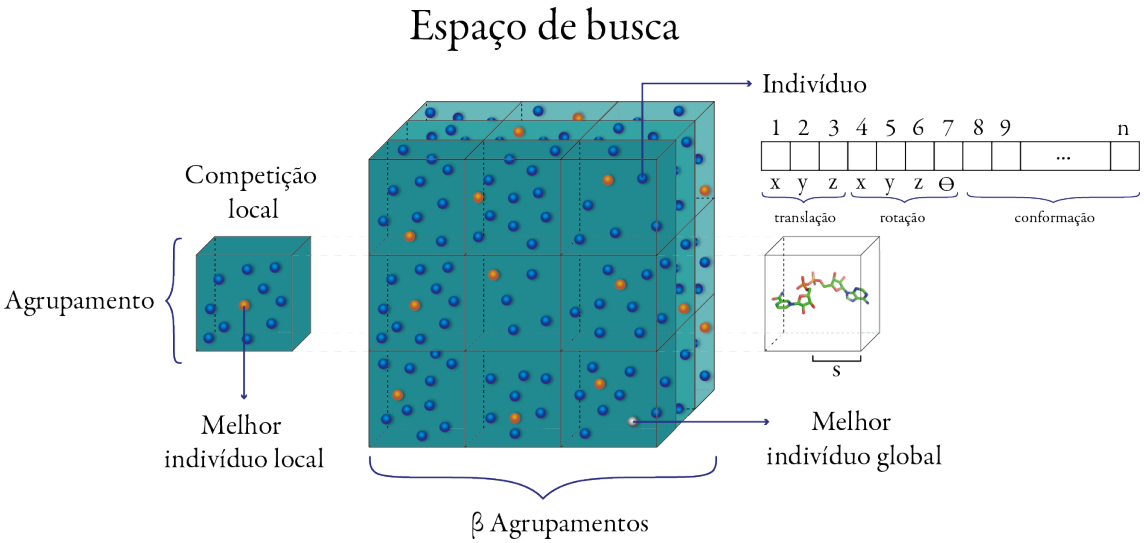
A partir do uso do método de agrupamento, foi possível implementar funções de competição global e local entre as soluções. Em cada cubo do espaço discretizado soluções são inicialmente agrupadas, tanto na inicialização da população, como na geração de soluções randômica na operação de mutação. Na operação de cruzamento, o vetor solução gerado entra na população se a avaliação de energia do mesmo foi menor que alguma solução presente no agrupamento. Caso seja, a conformação de pior avaliação de energia é retirada da população. Assim, em cada agrupamento é definido o melhor indivíduo, aquele com menor energia. Globalmente é feita a organização de todas as soluções, onde os melhores resultados de cada agrupamento estão agrupados na população de elite do BRKGA. Nesse contexto, as melhores soluções globais concorrem para encontrar a melhor solução global.

Um método de reinicialização da população foi implementado a partir da tentativa de introduzir novas soluções em cada agrupamento. Para cada solução é avaliado seu valor de energia, caso haja um valor maior entre os vetor no cubo, a pior solução é substituída. Porém, passados um número de avaliações de energia, verificou-se que novas soluções não são inseridas nos grupos, pois já atingiram-se valores baixos de energia em cada conformação. Nesse caso a população está estabilizada e pode, assim, estar presa em um mínimo global. A reinicialização da população foi desenvolvida de forma que, caso uma nova solução tente entrar na população e não consiga por um determinado número de

vezes, a população é reinicializada, gerando novas soluções randômicas, entretanto, são preservados os indivíduos da elite, e os melhores resultados de cada grupo.

A Figura 1.1 representa o espaço de busca discretizado, a representação das soluções, a melhor solução local e a melhor global. Cada solução é ilustrada por uma esfera, correspondente a um vetor de translação, rotação e conformação da molécula ligante. As esferas laranjas são as melhores soluções locais, de cada grupo, as azuis, outras possíveis soluções em competição local. A esfera branca é a melhor solução global, representa o resultado final do algoritmo, ou a melhor conformação proteína-ligante.

Figura 4.7: Discretização do espaço de busca, representação de cada solução, agrupamento de soluções, melhor solução local (laranja) e melhor solução global (branco)



4.6 Algoritmo BRKGA com agrupamento de soluções e competições global e local

O algoritmo proposto nesse trabalho une a discretização do espaço de busca, com o agrupamento de solução, competições global e local em uma Algoritmo Genético de Chaves Aleatórias Viciadas. O Pseudocódigo 2 mostra o algoritmo desenvolvido. A primeira parte do algoritmo descreve a busca pelos dados de configuração que definem o ponto central e tamanho do campo de busca. A partir dessas restrições é realizada a discretização do campo de busca. É importante ressaltar que o discretização é uma etapa prévia que independe da meta-heurística. As variáveis de entrada do algoritmo são o tamanho da população P, a porcentagem de indivíduos de elite P_e, de mutação P_m, número de posições do vetor-solução n, a probabilidade do cruzamento ρ_e e o número de

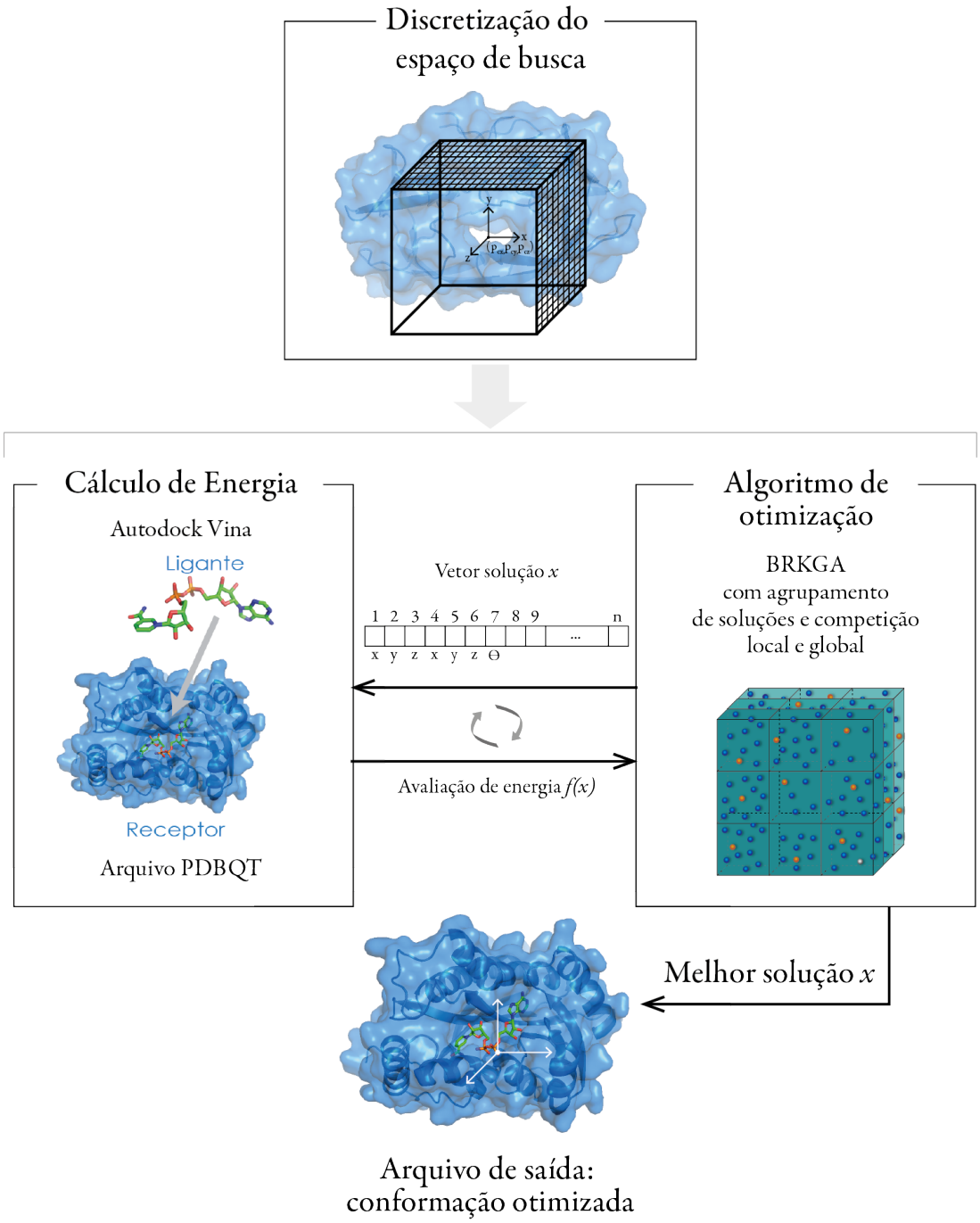
agrupamentos L. A variável c, se refere ao indivíduo recombinado.

```
    Data: |P|, |P_e|, |P_m|,|n|, |ρ_e|, L
 1  Busca dados do arquivo de configuração;
 2  Define ponto central ;
 3  Define tamanho do campo de busca ;
 4  Discretiza o espaço de busca de acordo com restrições;
 5  while não atingir número de avaliações de energia do
 6      P ← inicializa n vetores chaves aleatórias ;
 7      if restart < restart_criterion then
 8          Agrupa soluções em L agrupamentos ;
 9          Avalia a energia de cada solução em P ;
10          Divide P em p_e e p_ē ;
11          Inicializa a população da próxima geração: P^+ ← P_e ;
12          Gera o grupo de mutantes P_m ;
13          Agrupa P_m ;
14          Adiciona P_m: P^+ ← P^+ ⋃ P_m ;
15          foreach i ← 1 to |P| − |P_e| − |P_m| do
16              Seleciona um indivíduo a de P_e;
17              Seleciona um indivíduo b de P_ē;
18              foreach j ← 1 to n do
19                  Randomiza uma variável boleana B com a probabilidade ρ de
                    resultar Verdade;
20                  if B == True then
21                      c[j] ← a[j]
22                  end
23                  else
24                      c[j] ← b[j]
25                  end
26              end
27              if energy(recombinação c) < energy(pior solução em L(c)) then
28                  adiciona recombinação c à população P^+
29              end
30              else
31                  restart_criterion + +
32              end
33          end
34          Atualiza população P ← P_+;
35          Decodifica populaçãoP;
36          Encontra a melhor solução X^+ in P: X^+ ← argmin (f(x)|X ∈ P);
37      end
38  end
39  Retorna melhor solução X;
```

Algoritmo 2: Pseudocódigo do Algoritmo de Chaves Aleatórias Viciadas: discretização do campo de busca, reinicialização, competição global e local

O processo completo é ilustrado na Figura 4.8. A primeira etapa é a discretização do espaço de busca, em seguida, é executado o algoritmo de otimização agrupando soluções e iniciando a competição local e global, enquanto diferentes soluções são geradas a função de aptidão avalia os melhores resultados. Por fim, a saída do sistema é um arquivo PDBQT com as coordenadas do ligante para a melhor orientação ligante-receptor encontrada pelo algoritmo.

Figura 4.8: Diagrama do algoritmo BRKGA com agrupamento de soluções e competições global e local



4.7 Conclusão

A metodologia proposta utiliza uma técnica de discretização do espaço de busca, onde um modelo matemático de um cubo é explorado a fim de gerar soluções diversas dentro desse espaço. A partir dessa discretização é aplicado um Algoritmo Genético de Chaves Aleatórias Viciadas que tem a vantagem de representar as soluções através de um vetor, conseguindo, assim, representar a conformação da molécula ligante. O algoritmo também utiliza uma operação de cruzamento probabilístico, o que prioriza soluções de menor energia para o cruzamento. O algoritmo separa as soluções em grupos, o que permitiu a utilização de um método de agrupamento, unido à discretização. A formação de agrupamentos permitiu também o desenvolvimento de uma competição de soluções, método que mantém a diversidade e evita que o algoritmo fique preso em mínimos locais.

5 EXPERIMENTOS E RESULTADOS

Nesse capítulo são apresentados os experimentos realizados aplicando a metodologia desenvolvida nessa Dissertação. Na primeira seção são apresentados os métodos de avaliação, na segunda seção os complexos selecionados para os testes. Nas seções seguintes são apresentados os resultados dos experimentos para Atracamento Rígido, parametrização do algoritmo e Atracamento flexível. Os testes foram executados na máquina MS Azure Standard DS5 v2, com processadores de 16 núcleos Xeon E5-2673 v3 (Haswell), 2.4 GHz de clock. O código foi desenvolvido na linguagem de programação Python.

5.1 Métodos de avaliação

Durante o processo de fabricação de fármacos busca-se a conformação entre uma molécula receptora alvo e um fármaco, ou molécula ligante. As informações sobre essa conformação são desconhecidas, assim, são utilizadas informações disponíveis sobre a ligação entre complexos existentes. Algoritmos de Atracamento são avaliados de acordo com alguns critérios como:

- Sucesso: relativo a função de energia. No planejamento de fármacos o valor de energia da ligação entre os complexos é a única informação que indica a qualidade da conformação. Dessa forma, o valor mais baixo de energia encontrado pelo algoritmo deve ser o mais próximo possível com a estrutura obtida experimentalmente. Um algoritmo que encontre soluções de baixa energia mas que obtenha conformações muito distintas das experimentais não tem confiabilidade suficiente.

- Confiabilidade: capacidade de encontrar uma conformação de energia baixa em um determinado número de avaliações. A confiabilidade está relaciona com a capacidade do algoritmo de otimização de encontrar o mínimo global da função de energia, que deve ocorrer dentro do limite estipulado de avaliações de energia ou tempo de execução do algoritmo.

- Eficácia: relacionado a capacidade de encontrar conformações próximas em diferentes execuções do algoritmo. Além de chegar em valores de energia baixos, as conformações encontradas em diferentes execuções devem ser mais aproximadas possíveis, em termo de RMSD. O sucesso do algoritmo é alcançado quando a con-

formação de menor energia coincide com uma conformação factível biologicamente em diferentes execuções do algoritmo.

O algoritmo deve encontrar o mínimo global da função de energia, considerando diferentes execuções esse valor de energia deve ser o mais aproximado possível. A principal medida tanto para avaliação das estruturas testadas como para comparação com outros métodos, é o RMSD. Para duas estruturas a e b, de uma molécula idêntica, o RMSD é definido como na Equação 5.1, onde a_{ix}, b_{ix}, a_{iy}, b_{iy}, a_{iz} e b_{iz}, representam as coordenadas x, y e z dos átomos a_i e b_i, respectivamente.

$$RMSD_{ab} = \sqrt{\frac{1}{n}\sum_{i=1}^{j}((a_{ix} - b_{ix})^2 + (a_{iy} - b_{iy})^2) + (a_{iz} - b_{iz})^2)} \tag{5.1}$$

É importante ressaltar que os valores de RMSD são calculados apenas em casos em que a conformação proteína-ligante é conhecida, como um método de avaliação da qualidade da ligação dos complexos. Na predição de novos fármacos são utilizados somente os valores obtidos pela função de energia, assim como nas operações do algoritmo de Atracamento Molecular proposto.

5.2 Dados para os testes

Para a realização dos experimentos foram selecionadas 50 estruturas dos bancos de dados PDB e ZINC, descritos na Seção 2.4. As 50 estruturas foram organizadas em 4 conjuntos de teste, organizados segundo o tamanho de cada complexo. Os conjuntos foram divididos de acordo com ligantes de tamanho pequeno, médio e grande, nos conjunto 1,2 e 3, respectivamente. O quarto grupo representa moléculas com um receptor diferente. No artigo de (CAMACHO et al., 2014) 47 estruturas em comum com esse trabalho foram utilizadas, permitindo, assim, um estudo comparativo dos resultados obtidos.

O primeiro conjunto contém 10 estruturas baseadas no receptor *HIV-protease* proposto pelo artigo (MORRIS et al., 2009) para validação da função de energia, mesma função utilizada nesse trabalho. Esse primeiro conjunto foi selecionado a fim de executar testes relativos ao Atracamento rígido e posteriores testes de parametrização do algoritmo. As estruturas ligantes nesse conjunto possuem diferentes tamanhos e resolução. Seus nomes, códigos PDB e resolução são apresentados na Tabela 5.1.

Tabela 5.1: Conjunto de testes 1: Complexos proteína-ligante, código PDB e solução das estruturas

Molécula	**Código PDB**	**Resolução (Å)**
HIV-1 protease/AHA006	1AJV	$2,00$
HIV-1 protease/AHA001	1AJX	$2,00$
HIV-1 protease/Macrocyclic peptidomimetic inhibitor 8	1D4K	$1,85$
HIV-1 protease/AHA047	1G2K	$1,95$
HIV-1 protease/U75875	1HIV	$2,00$
HIV-1 protease/KNI-272	1HPX	$2,00$
HIV-1 protease/GR126045	1HTF	$2,20$
HIV-1 protease/Q8261	1HVH	$1,80$
HIV-1 protease/U100313	2UPJ	$3,00$

O segundo conjunto de dados contém complexos também baseados na molécula receptora HIV-protease porém o conjunto possui moléculas ligantes de maior tamanho e características heterogêneas, em relação aos átomos do complexo. O conjunto possui 20 moléculas, cujas informações de nome, código PDB e resolução são apresentadas na Tabela 5.2.

Tabela 5.2: Conjunto de testes 2: Complexos proteína-ligante, código PDB e solução das estruturas

Molécula	**Código PDB**	**Resolução (Å)**
HIV-1 protease/U-89360E	1A9M	$2,30$
HIV-1 protease/HYDROLASE INHIBITOR	1AAQ	$2,50$
HIV-1 protease/PEPTIDOMIMETIC INHIBITOR 4	1B6L	$1,75$
HIV-1 protease/PEPTIDOMIMETIC INHIBITOR 6	1B6M	$1,85$
HIV-1 protease/SB203386	1BDL	$2,80$
HIV-1 protease/SB203386	1BDR	$2,80$
HIV-1 protease/U89360E	1GNM	$2,30$
HIV-1 protease/U89360E	1GNO	$2,30$
HIV-1 protease/NOVEL GAMMA-TURN MIMETIC	1HBV	$2,30$
HIV-1 protease/hydroxyethylene-based inhibitors	1HEG	$2,20$
HIV-1 protease/CGP 53820	1HIH	$2,20$
HIV-1 protease/VX-478	1HPV	$1,90$
HIV-2 protease/ L-735,524	1HSG	$2,00$
HIV-1 protease/PENICILLIN-DERIVED	1HTE	$2,80$
JE-2147-HIV	1KZK	$1,90$
PROTEASE B/OVOMUCOID INHIBITOR	1SGB	$1,80$
HIV-1 protease/SB203386	1TCX	$2,30$
GCAA RNA TETRALOOP	1ZIH	$2,30$
Deuterated gammaE crystallin	1ZIR	$1,36$
HIV-1 protease/AMINIMIDE PEPTIDE ISOSTERE	3AID	$2,50$

O terceiro conjunto possui 17 estruturas, também baseadas no receptor *HIV-protease*, com ligantes de maior tamanho. Essas estruturas são apresentadas na Tabela 5.3.

Tabela 5.3: Conjunto de testes 3: Complexos proteína-ligante, código PDB e solução das estruturas

Molécula	**Código PDB**	**Resolução (Å)**
HIV-1 protease/PEPTIDOMIMETIC INHIBITOR 1	1B6J	$1,85$
HIV-1 protease/PEPTIDOMIMETIC INHIBITOR 7	1B6P	$2,00$
HIV-1 protease/Macrocyclic peptidomimetic inhibitor 8	1D4K	$1,85$
HIV-1 protease/MACROCYCLIC PEPTIDOMIMETIC	1D4L	$1,75$
HIV-1 protease/HYDROLASE INHIBITOR	1HEF	$2,20$
HIV-1 protease/KNI-272	1HXW	$1,80$
HIV-1 protease/multi-resistant mutant	1IZH	$1,90$
HIV-1 protease/(hydroxyethyl)amide isostere	1JLD	$2,50$
HIV-1 protease/MULTI-DRUG RESISTANT	1K6C	$2,20$
HIV-1 protease/MULTI-DRUG RESISTANT	1K6P	$2,25$
HIV-1 protease/MULTI-DRUG RESISTANT	1K6T	$2,20$
HIV-1 protease/MULTI-DRUG RESISTANT	1K6V	$2,00$
HIV-1 protease/CYCLIC PEPTIDOMIMETIC	1MTR	$1,75$
HIV-1 protease/Lopinavir	1MUI	$2,80$
HIV-1 protease/inhibitor complex	2BPX	$2,80$
HIV-1 protease/L-700,417	4PHV	$2,10$
HIV-1 protease/ACETYL-PEPSTATIN	5HVP	$2,00$

O conjuto 4 consiste em 3 estruturas, cujo receptor é a estrutura alvo da bactéria de tuberculose (*Mycrobacterium tuberculosis*), chamada de ENOYL-ACYL CARRIER PROTEIN (ACP) REDUCTASE, testadas com 3 diferentes ligantes, cujos código PDB são apresentados na Tabela 5.4.

Tabela 5.4: Conjunto de testes 4: Complexos proteína-ligante, código PDB/ZINC e solução das estruturas

Molécula	**Código PDB**	**Resolução (Å)**
ENOYL-ACYL CARRIER PROTEIN (ACP) REDUCTASE	1ENY/NAD	$2,20$
Triclosan	TCL	$2,00$
Ethionamide	ETH	$2,00$

Todas as estruturas passaram pela preparação descrita na Seção 4.1. Nessas estruturas foram selecionados 10 graus de liberdade de ângulos diedrais ativos, valor mínimo de rotações de todas as estruturas testadas. A proteína HIV-protease possui um sítio ativo em forma de túnel o qual engloba o ligante, ou molécula inibidora. Essa caraterística facilita a definição do centro do campo de busca.

5.3 Resultados de parametrização

O método proposto prevê uso de parâmetros como: tamanho da população e critério de reinicialização. Nos experimentos de parametrização foram utilizados o primeiro conjunto de dados em três seleções de parâmetros. O primeiro parâmetro (TP) se refere ao tamanho da população, ou o número de soluções gerada pelo algoritmo, a escolha desses valores foi feita de acordo com a recomendação (GOULART et al., 2011): $TP = 200$, $TP = 400$ e $TP = 800$ indivíduos. O segundo parâmetro testado (R) se refere ao critério de reinicialização do algoritmo, foram testados os valores de $R = 500$, $R = 1000$ e $R = 1500$.

A tabela 5.5 apresenta os resultados obtidos em 8 execuções do algoritmo para as 3 seleções de parâmetros. O número de avaliações de energia foi fixado em 100 mil avaliações, valor definido ao considerar a convergência do algoritmo que acontece no mínimo com o número de avaliações definido. Na primeira coluna são apresentadas as 10 estruturas, em seguida a variável TP (tamanho da população), variável R (valor de reinicialização), a melhor solução dentre as 8 execuções, os valores de energia e RMSD, a média de RMSD e energia das 8 execuções e, por fim, a variável σ (Desvio Padrão) dos valores médios. Toda nova solução gerada pelo algoritmo entra em competição com as soluções de seu agrupamento, caso sua função de aptidão resulte um valor menor do que os que se encontrem no agrupamento, ele entra na população, caso contrário é incrementada a variável R. Com o aumento do tamanho da população, maior é o número de avaliações de energia por geração, assim foram gradativamente aumentados os valores de R, por isso foram testados 3 valores de R equivalentes ao aumento do população.

Os resultados apresentados selecionaram a parametrização de $TP = 400$ e $R = 1000$ por obter os melhores resultados para as 10 estruturas testadas. Nos testes as estruturas com essa parametrização chegou a conformações de $RMSD < 1,0$, com excessão da estrutura 1HPX. Nessa estrutura as 3 parametrizações alcançaram resultados em que o atracamento ocorreu fora do sítio de ligação. Os resultados nas 9 estruturas mostraram a eficácia do algoritmo, que atingiu ótimos resultados para o primeiro conjunto de dados. De acordo com esses valores o algoritmo foi parametrizado para os testes com os outros conjuntos de dados.

Tabela 5.5: Resultados para as execuções de parametrização do algoritmo

PDB	**TP**	**R**	**Melhor Solução**		**Média**		σ
			Energia [a]	RMSD [b]	Energia	RMSD	
	200	500	-9,605	0,449	-10,522	0,582	0,114
1AJV	400	1000	**-10,603**	**0,426**	**-9,628**	**0,622**	**0,115**
	800	1500	-9,119	5,290	-10,349	7,878	1,976
	200	500	-11,015	0,317	-10,456	0,728	0,360
1AJX	400	1000	**-10,007**	**0,586**	**-9,031**	**0,847**	**0,233**
	800	1500	-9,119	5,000	-8,533	6,803	1,165
	200	500	-12,866	0,161	-12,408	0,139	0,024
1BV9	400	1000	**-12,919**	**0,113**	**-12,902**	**0,129**	**0,016**
	800	1500	-12,639	5,747	-12,031	8,118	2,233
	200	500	-12,010	0,403	-11,812	0,694	0,270
1D4K	400	1000	**-12,983**	**0,403**	**-11,558**	**0,680**	**0,201**
	800	1500	-12,920	6,443	-12,490	8,101	1,801
	200	500	-4,232	1,039	-5,454	9,532	6,929
1G2K	400	1000	**-5,605**	**0,764**	**-7,273**	**7,621**	**7,191**
	800	1500	-9,850	6,443	-9,133	8,101	1,801
	200	500	-11,127	0,443	-10,312	0,493	0,108
1HIV	400	1000	**-11,442**	**0,479**	**-11,354**	**0,635**	**0,117**
	800	1500	-11,847	7,790	-11,124	8,286	1,719
	200	500	-4,091	11,763	-4,912	13,999	1,571
1HPX	400	1000	**-4,419**	**11,713**	**-4,133**	**11,964**	**2,887**
	800	1500	-3,164	13,992	-3,844	17,882	2,612
	200	500	-7,912	0,873	-8,029	4,237	2,934
1HTF	400	1000	**-7,484**	**0,933**	**-7,028**	**1,800**	**1,372**
	800	1500	-7,692	4,598	-7,019	7,030	1,877
	200	500	-7,126	1,079	-7,031	4,334	3,246
1HVH	400	1000	**-7,816**	**0,760**	**-7,099**	**1,095**	**0,244**
	800	1500	-7,869	4,653	-8,028	6,527	1,481
	200	500	-10,217	1,194	-10,263	1,700	0,342
2UPJ	400	1000	**-12,011**	**1,125**	**-11,447**	**1,249**	**0,068**
	800	1500	-10,103	4,078	-9,231	6,266	1,700

[a](kcal/mol) e [b]Å(Angstroms).

5.4 Resultados de Atracamento Rígido

Definida a parametrização do algoritmo, foram realizados os testes de Atracamento rígido. Nesse teste são gerados valores randômicos de translação e rotação da molécula ligante, as operações de conformação não são realizadas no algoritmo. Esse teste visa avaliar a capacidade do algoritmo de realizar bons resultados para uma simplificação do problema, além de validar a função de energia.

O conjunto de dados 1, Tabela 5.1, foi selecionado para esses testes. As 10 estruturas passaram por 8 execuções, cujo critério de parada foi 100 mil avaliações de energia, valor mínimo de avaliação que o algoritmo precisa para atingir a convergência. A Tabela 5.6 mostra os resultados obtidos, na primeira coluna o código PDB da estrutura, em seguida os valores de energia (kcal/mol) e RMSD (Å) da melhor solução encontrada nas 8 execuções e, por fim, a média de energia, desvio padrão da média de energia, RMSD médio, e desvio padrão da média de RMSD de todas as execuções.

Tabela 5.6: Resultados de Atracamento rígido para 10 estruturas em 30 execuções de 1 milhão avaliações de energia

PDB	**Melhor Solução**		**Média**			
	Energia [a]	RMSD [b]	Energia	σ	RMSD	σ
1AJV	-10,884	0,183	-10,866	0,012	0,249	0,0282
1AJX	-10,757	0,241	-12,862	0,132	0,145	0,050
1BV9	-12,942	0,071	-12,862	0,132	0,145	0,050
1D4K	-13,136	0,301	-13,078	0,050	0,412	0,060
1G2K	-11,208	0,120	-11,192	0,013	0,167	0,022
1HIV	-11,793	0,069	-11,724	0,128	0,149	0,087
1HPX	-2,726	17,41	-2,71	0,008	17,538	0,063
1HTF	-6,723	0,293	-6,291	0,034	0,345	0,067
1HVH	-5,912	0,545	-5,878	0,022	0,574	0,023
2UPJ	-9,308	0,644	-9,293	0,008	0,701	0,0231

[a](kcal/mol) e [b]Å(Angstroms).

Os resultados validam o algoritmo a estratégia de busca e função de avaliação, já que para 9 das 10 estruturas estudadas o algoritmo atingiu valores de RMSD abaixo de $1,0$ Å. A estrutura 1HPX não obteve resultados bons em nenhuma das parametrizações

devido a complexidade da estrutura em termos de tamanho. Para esse complexo foram realizados testes utilizando outras ferramentas, Seção 5.6, a fim de comparar os resultados obtidos pelo algoritmo. A melhor execução ocorreu na estrutura 1BV9, com RMSD de $0,071$ Å. Os resultados de Atracamento rígido validam a meta-heurística para a resolução problema, assim, foram executados os testes de Atracamento flexível descritos na Seção 5.5.

5.5 Resultados de Atracamento Flexível

Os testes de Atracamento flexíveis utilizaram um conjunto extenso de estruturas, no total 50 complexos proteína-ligante foram submetidas ao algoritmo de Atracamento Molecular. As estruturas foram divididas em 4 grupos, como apresentados nas Tabelas 5.1, 5.2, 5.3 e 5.4. O primeiro conjunto foi utilizado para parametrização, testes de Atracamento rígido e comparação com outras ferramentas, assim, os resultados apresentados nessa Seção englobam os conjuntos 2, 3 e 4. Para todos os testes foi mantida a configuração testada na primeira etapa de parametrização. Cada estrutura passou por 30 execuções do algoritmo, cujo critério de parada foi 1 milhão de avaliações de energia.

A Tabela 5.7 apresenta os resultados para 20 estruturas do segundo conjunto. Apresentando na primeira coluna o código PBD, na segunda a melhor solução (RMSD e energia), e os valores médios de energia e RMSD, e seus desvios padrão, para as 30 execuções.

Analisando os valores dos melhores resultados das 30 execuções é possível observar que 9 estruturas obtiveram um resultado menor que $2,0$ Å, com destaque (valor em negrito) para a estrutura 1B6L que obteve o valor de RMSD de $0,769$ Å. O pior resultado foi a estrutura 1KZK, cujo valor de RMSD no melhor resultado (em vermelho) foi de $9,043$ Å, e média de $10,521$. O baixo valor de desvio padrão mostra que o algoritmo obteve na maioria das execuções valores próximos de RMSD e energia. Na análise estrutural, pôde-se observar que o ligante foi posicionado em uma região fora do sítio de ligação, não encontrando o mínimo global da função de energia. O melhor resultado para as médias de enegia e RMSD foi encontrado nos testes com o complexo 1A9M, com RMSD médio de $1,204$ Å. Para essa estrutura o algoritmo performou uma conformação bastante semelhante a estrutura obtida experimentalmente.

Em um panorama geral, 6 estruturas obtiveram valores médios de RMSD abaixo de $2,0$ Å, $13,0$ abaixo de $3,0$ Å, ou seja, em 65% do conjunto de testes o algoritmo al-

cançou bons resultados. Em outras 4 estruturas, 20% da amostra, os resultados variaram entre $3 - 6$ Å, resultados satisfatórios. No restante, 3 estruturas, os resultados obtidos foram maiores de $6,0$ Å, resultados considerados ruins. O algoritmo manteve uma boa média para 13 estruturas, com baixos valores de desvio padrão. Considerando a dificuldade do conjunto de teste, com diversificadas estruturas em questão de tamanho e ângulos diedrais, os resultados alcançados são considerados bons.

Tabela 5.7: Resultados de Atracamento flexível para o conjunto de teste 2: 30 execuções e 1 milhão de avaliações de energia; comparação com os resultados obtidos pela ferramenta Autodock Vina (CAMACHO et al., 2014). Os valores destacados representam a melhor solução encontrada(em negrito) e a pior solução encontrada (em vermelho).

PDB	**Melhor Solução**		**Média**				**Autodock Vina**	
	Energia [a]	RMSD [b]	Energia	σ	RMSD	σ	Energia	RMSD
1A9M	-12,846	1,019	**-12,322**	**0,355**	**1,204**	**0,148**	-2,970	9,560
1AAQ	-11,295	2,204	-10,688	0,400	2,495	0,1935	-7,010	12,700
1B6L	**-14,017**	**0,769**	-12,851	0,484	1,666	0,430	-10,170	13,040
1B6M	-17,887	1,102	-16,015	0,980	1,913	0,464	-11,500	11,880
1BDL	-7,669	3,003	-7,420	0,194	3,885	0,335	-5,890	10,540
1BDR	-12,846	2,252	-11,945	0,588	2,663	0,322	-4,500	12,300
1GNM	-9,44	0,977	-9,374	0,116	1,223	0,242	-18,690	12,040
1GNO	-9,554	0,791	-9,252	0,197	1,263	0,339	-14,710	11,420
1HBV	-14,004	2,0144	-13,255	0,468	2,410	0,165	-4,52	12,150
1HEG	-9,655	6,276	-8,702	0,421	6,872	0,367	-5,89	10,580
1HIH	-10,623	1,7161	-10,217	0,231	3,135	1,804	-3,120	12,700
1HPV	-12,695	3,149	-11,787	0,421	4,482	0,478	-3,670	12,160
1HSG	-14,019	1,720	-13,057	0,533	4,703	1,625	-5,700	11,990
1HTE	-9,043	0,722	-8,812	0,186	1,325	0,183	-6,920	12,320
1KZK	33,822	9,534	42,283	13,248	10,521	0,343	-7,510	11,740
1SBG	-13,755	2,386	-13,058	0,3238	2,716	0,183	-4,720	11,200
1TCX	-12,4	2,222	-12,110	0,169	2,629	0,442	-3,910	12,250
1IZH	-14,013	1,138	-12,973	0,580	2,050	0,621	-3,150	11,740
1Z1R	-15,260	2,358	-14,182	0,792	3,259	0,96	-8,670	12,520
3AID	-12,319	2,252	-11,894	0,396	2,506	0,189	-5,370	11,840

[a](kcal/mol) e [b]Å(Angstroms).

Na Tabela 5.7 são apresentados os valores obtidos pela ferramenta Autodock Vina,

valores de RMSD e energia extraídos do Artigo (CAMACHO et al., 2014). Nesse artigo foram testadas 4 meta-heurísticas para o problema de otimização, e foi considerado o mesmo conjunto de teste. Os dados das duas últimas colunas são referentes ao melhor valor médio de resultados obtidos para cada estrutura, ou seja, das 4 meta-heurísticas selecionadas no estudos, o melhor valor alcançado foi selecionado para fins de comparação. É possível observar que os resultados alcançados mantém uma média de RMSD acima de 10 Å, com excessão da estrutura 1A9M, que alcançou um valor de $9,56$. Os resultados alcançados pelo método desenvolvido superam em todos as estruturas os valores alcançados pelo *software* Autodock Vina, utilizando o mesmo os mesmo parâmetros de sítio de ligação e número de avaliações de energia.

Os testes realizados no terceiro conjunto de dados, que conta com 15 estruturas, seguiram a mesma parametrização e número de avaliações de energia. Esse conjunto possui estruturas cuja resoluções variam em $1,75 - 2,8$ Å. Na Tabela 5.8, em negrito, estão as estruturas 1D4K que alcançou o melhor resultado em com os valores de RMSD de $0,747$ Å, além da estrutura 1JLD que obteve o melhor valores de média de RMSD com $1,318$ Å. Os valores de desvio padrão se mantiveram abaixo de $1,0$, com exceção da estrutura 1D4L que teve valores de RMSD distintos em suas execuções. O pior resultado (em vermelho) foi a estrutura 1B6J com RMSD médio de $18,026$ Å. Em uma análise geral dos resultados é possível observar que $93,3$ % das estruturas o RMSD permaneceu com valores abaixo de $3,0$ Å, resultados muito bons considerando a complexidade das estruturas testadas. Considerando que os dados utilizados para testes também são encontrado em (CAMACHO et al., 2014), a última coluna apresenta os resultados alcançados pela ferramenta Autodock Vina. Na média de 30 execuções com 1 milhão de avaliações de energia, em 14 das estruturas testadas a ferramenta proposta superou os resultados alcançados pelo Autodock Vina. Em todos os testes executados pela ferramenta os valores de RMSD foram superiores a 3 Å.

Os resultados de Atracamento flexível para o terceiro conjunto obtiveram em $66,6\%$ dos casos um valor de RMSD abaixo de $2,0$ Å. Os desvios padrão, tanto de energia como de RMSD se mantiveram baixos mostrando que o algoritmo realiza resoluções semelhantes nas 30 execuções. Entretanto, em alguns casos, como acontece no segundo grupo, o algoritmo encontra o mínimo local da função em uma conformação em que o ligante se posiciona fora do sítio de ligação, em uma região externa da proteína, o que gera um RMSD alto. Por isso, incorporou-se a comparação com outras ferramentas, onde pôde-se observar que ocorre a mesma situação, inclusive em outras estruturas em que

o método proposto alcança melhores resultados. A média de resultados de RMSD das estruturas é acima de $10,0$ Å.

Tabela 5.8: Resultados de Atracamento flexível para o conjunto 3; comparação com os resultados obtidos pela ferramenta Autodock Vina. Os valores destacados representam a melhor solução encontrada(em negrito) e a pior solução encontrada (em vermelho).

PDB	**Melhor Solução**		**Média**				**Autodock Vina**	
	Energia [a]	RMSD [b]	Energia	σ	RMSD	σ	Energia	RMSD
1B6J	-6,694	16,903	-6,329	0,243	18,026	0,642	-5,740	11,340
1B6P	-22,783	0,955	-20,884	1,294	2,495	0,476	-7,720	12,540
1D4K	**-15,411**	**0,747**	-14,449	0,883	1,784	1,3129	-11,280	11,910
1D4L	-15,763	0,963	-13,627	1,306	5,496	4,292	-13,280	11,120
1HEF	-8,351	2,633	-7,531	0,514	6,642	1,368	-4,330	7,630
1HXW	-14,143	1,927	-13,699	0,252	2,399	0,271	-4,140	11,400
1IZH	-12,698	1,255	-9,356	0,311	1,452	0,114	-3,150	11,740
1JLD	-14,378	1,059	**-12,191**	**0,22**	**1,318**	**0,214**	-5,110	12,140
1K6C	-18,096	1,454	-15,233	1,195	2,829	1,469	-8,640	12,670
1K6P	-18,096	1,454	-15,237	1,195	2,82	1,469	-8,850	12,600
1K6T	-18,102	2,227	-16,907	0,841	2,747	0,429	-10,24	11,940
1MTR	-17,905	1,328	-17,058	0,410	1,874	0,242	-10,240	12,140
1MUI	-15,027	2,804	-14,308	0,270	3,546	0,339	-4,870	11,120
2BPX	-14,381	2,029	-13,223	0,633	5,764	1,517	-6,610	12,640
5HVP	-11,960	1,368	-9,376	1,759	6,314	4,411	-9,010	11,800

[a](kcal/mol) e [b]Å(Angstroms).

Os últimos testes de Atracamento flexível englobam o conjunto 4, formado por 3 estruturas. A estrutura receptora difere dos outros conjunto de dados. A estrutura ligante é difere nos 3 casos, são elas: NAD, ETH e TCL, o complexo do receptor acoplado com o ligante NAD tem por referência o código PDB 1ENY. Os testes seguem a estrutura de 30 execuções e 1 milhão de avaliações de energia.

Os resultados mostram o melhor resultado na estrutura de ligante ETH com RMSD de $0,51$ Å. Nas outras estruturas o algoritmo não conseguiu sucesso ao prever a orientação das estruturas com valores de RMSD maiores que 3 Å. Os resultados mostram que para uma estrutura complexa como é o ligante NAD os resultados não conseguem prever uma conformação próxima da estrutura cristalográfica.

Tabela 5.9: Resultados de Atracamento flexível para o conjunto 4. Os valores destacados representam a melhor solução encontrada(em negrito) e a pior solução encontrada (em vermelho).

PDB	**Melhor Solução**		**Média**			
	Energia [a]	RMSD [b]	Energia	σ	RMSD	σ
1ENY/NAD	-4,521	10,779	-4,502	0,012	10,841	0,0324
ETH	**-8,601**	**0,510**	**-8,593**	**0,005**	**0,549**	**0,017**
TCL	-4,182	3,372	-4,168	0,014	7,344	1,159

[a](kcal/mol) e [b]Å(Angstroms).

5.6 Comparação com outras ferramentas

Os resultados de Atracamento flexível para o conjunto 1 são apresentados como uma comparação com as ferramentas Autodock Vina e Dockthor. A ferramenta Autodock Vina utiliza a mesma função de energia aplicada no algoritmo proposto, e a ferramenta Dockthor utiliza um Algoritmo Genético, por isso a comparação visa avaliar as três metodologias em relação aos resultados alcançados. Nos testes foram realizadas 30 execuções para cada uma das 10 estruturas e 1 milhão de avaliações de energia foi selecionado com o critério de parada para os algoritmos. Foram selecionados os mesmos 10 graus de liberdade para as estruturas ligantes, o mesmo centro do campo de busca e o mesmo número de pontos de grade para as três ferramentas. As configurações de parametrização dos algoritmos foram mantidas padrão, com valores que optimizam os resultados. A discretização das grades de energia foi selecionada como $0,25$ Å(valor padrão para o programa Dockthor), o que gera um número estimado de pontos de grade de 531441, para cada estrutura foi selecionado um ponto central que foi configurada para cada programa e o tamanho do espaço busca foi mantido em 11 Å.

Os resultados são apresentados na Tabela 5.10. Em 7 estruturas, $77,7$ % do total, o algoritmo BRKGA proposto obteve melhores resultados em comparações as outras ferramentas. Na estrutura 1AJX, a ferramenta Dockthor alcançou a melhor execução com o RMSD de $0,362$ Å, porém na média de execuções o algoritmo BRKGA obteve melhores resultados. Na estrutura 1G2K, embora a melhor execução tenha sido do algoritmo BRKGA, a ferramenta Autodock Vina obteve o melhor valor médio de $0,559$ Å. Nas outras 7 estruturas os valores de RMSD do algoritmo BRKGA superou os resultados nos testes envolve as outras duas ferramentas. A estrutura 1HPX que havia obtido resultados

não satisfatórios nos testes de Atracamento Rígido obteve na melhor execução o valor de RMSD de $8,204$ Å. O resultado, embora insatisfatório, superou os resultados obtidos pelas outras ferramentas. A ferramenta Dockthor não executou os testes para a estrutura 1HVH, por isso não foi possível realizar a comparação entre as três ferramentas.

Tabela 5.10: Resultados de comparação do algoritmo BRKGA com as ferramentas Autodock Vina e Dockthor: melhor solução (energia e RMSD), e média (energia, desvio padrão, RMSD e desvio padrão)

PDB	**Ferramentas**	**Melhor Solução**		**Média**			
		Energia [a]	RMSD [b]	Energia	σ	RMSD	σ
1AJV	BRKGA	**-9,605**	**0,299**	**-11,367**	**0,582**	**0,382**	**0,050**
	Vina	-11,800	0,853	-11,75	0,273	0,996	0,737
	Dockthor	21,331	0,620	22,341	1,348	0,793	0,186
1AJX	BRKGA	-11,545	0,416	**-11,409**	**0,079**	**0,542**	**0,151**
	Vina	-12,000	1,032	-11,926	0,044	1,097	0,098
	Dockthor	**36,629**	**0,3622**	38,308	6,448	0,680	0,092
1BV9	BRKGA	**-12,958**	**0,1004**	**-12,95**	**0,004**	**0,421**	**0,014**
	Vina	-8,400	4,807	-7,683	1,059	6,655	2,266
	Dockthor	88,247	23,413	88,421	0,157	24,829	0,692
1D4K	BRKGA	**-15,342**	**0,451**	**-14,636**	**0,344**	**0,653**	**0,122**
	Vina	-2,800	22,258	-2,566	0,250	22,857	0,359
	Dockthor	15523,430	19,375	15525,890	3,290	21,120	0,822
1G2K	BRKGA	**-11,851**	**0,258**	-11,607	0,072	0,564	0,276
	Vina	-12,9	0,504	**-12,77**	**0,496**	**0,559**	**0,027**
	Dockthor	0,895	0,574	1,291	0,770	0,794	0,108
1HIV	BRKGA	**-13,04**	**0,416**	**-12,981**	**0,051**	**0,472**	**0,0407**
	Vina	-2,100	21,390	-1,960	0,106	23,334	0,860
	Dockthor	99,266	19,003	106,890	3,150	21,015	1,533
1HPX	BRKGA	**-6,472**	**8,204**	**-5,842**	**0,319**	**15,285**	**2,663**
	Vina	-1,200	21,274	-1,083	0,059	22,052	0,441
	Dockthor	99,062	16,485	102,822	1,539	18,455	1,165
1HTF	BRKGA	**-10,528**	**0,588**	**15,445**	**1,384**	**-9,631**	**0,346**
	Vina	-10,700	8,1009	-10,383	0,381	8,331	0,269
	Dockthor	63,370	14,310	66,795	1,957	15,443	0,566
1HVH	BRKGA	**-9,702**	**0,913**	**-9,631**	**0,063**	**1,173**	**0,136**
	Vina	-11,600	4,868	-10,383	0,152	7,074	1,232
	Dockthor	-	-	-	-	-	-
2UPJ	BRKGA	**-12,054**	**1,125**	**-11,985**	**0,068**	**1,453**	**0,308**
	Vina	-11,000	1,678	-10,896	0,049	9,819	1,538
	Dockthor	286,341	0,867	295,107	4,498	7,626	3,011

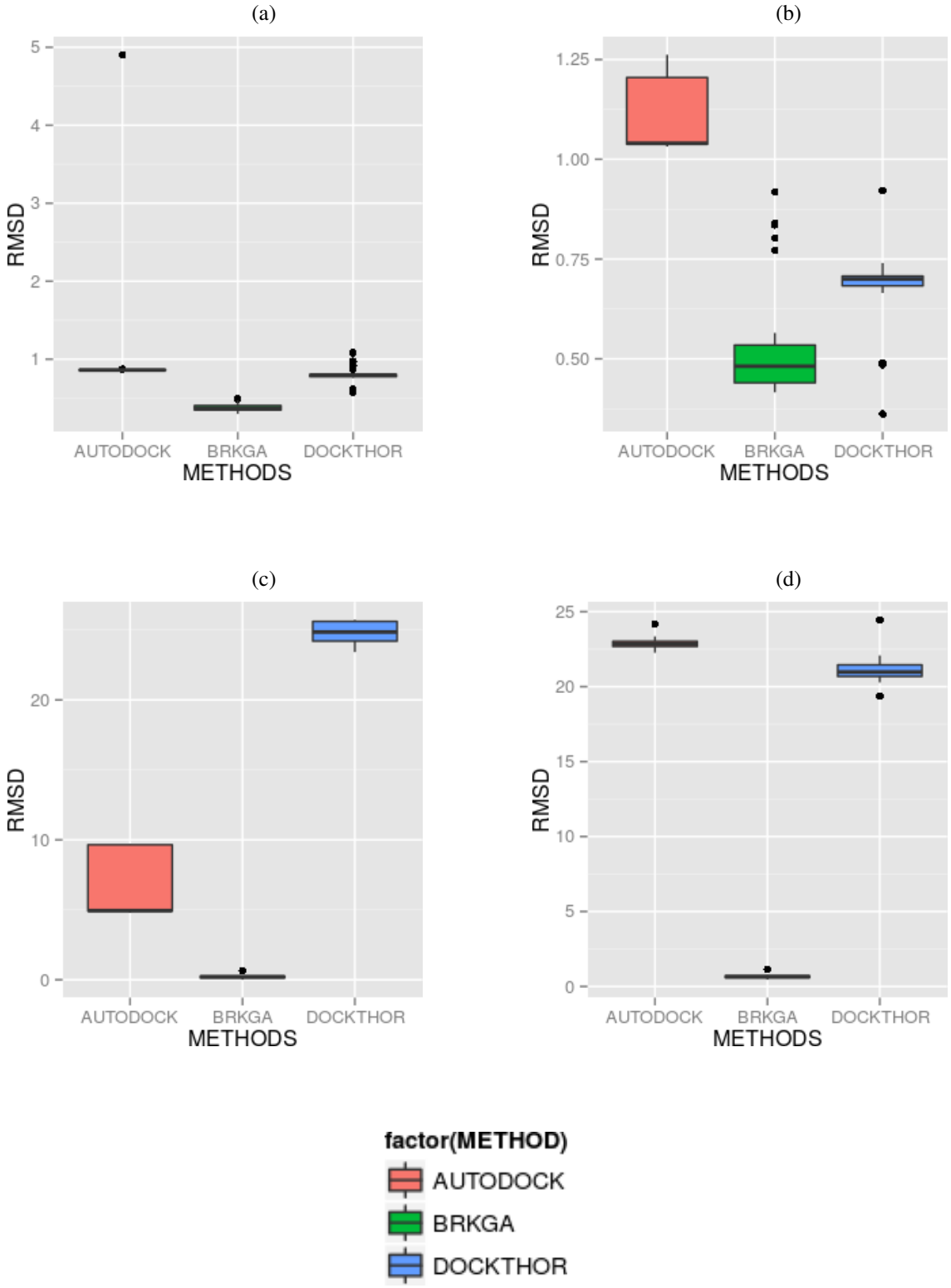
[a](kcal/mol) e [b]Å(Angstroms).

Para uma análise mais aprofundada de comparação entre as ferramentas foram gerados diagramas de caixa (boxplot) das 30 execuções. No eixo vertical do diagrama de caixa é representada a variável a ser analisada, no caso os valores de RMSD encontrados em cada execução. No eixo horizontal os fatores de interesse são as três ferramentas: Autodock Vina (vermelho), BRKGA (verde) e Dockthor (azul). O diagrama procura obter, localizar e analisar a variação de RMSD em cada execução independente. As seguintes informações são apresentadas: a mediana e os quartis (onde o quartil inferior contém 25% das menores medianas e o quartil superior contém 75% de todas as medidas); o símbolo central de cada gráfico representa a mediana, o segmento de reta vertical conecta o topo da caixa ao maior valor observado. São apresentados também valores atípicos, ou dados que são muito diferentes do conjunto, são casos de estudo para uma avaliação da validade dessa execução para o conjunto de soluções encontrado. Nessas situações, execuções resultaram em valores de RMSD muito distintos da maioria das soluções. O diagrama, em resumo, mostra onde estão localizados 50% dos valores mais prováveis, a mediana e os valores de RMSD extremos.

Na Figura 5.1 são apresentados os diagramas de caixa para as estruturas 1AJV, 1AJX, 1BV9 e 1D4K. Na estrutura 1AJV os três métodos obtive valores baixos de RMSD e mantiveram a distribuição baixa com poucos valores destoantes. Apenas uma das execuções da ferramenta Autodock Vina apresentou um valor de RMSD destoante da mediana. Na estrutura 1AJX o algoritmo BRKGA obteve os melhores resultados porém com uma variação da distribuição maior que a ferramenta Dockthor e ainda com algumas execuções com valores maiores de RMSD que a mediana da outra ferramenta, porém, com melhores valores em média. A estrutura 1D4K tem melhores resultados obtidos pelo algoritmo BRKGA, com alta variação no algoritmo Autodock Vina e altos valores de RMSD na ferramenta Dockthor. Resultados que se repetiram nos testes envolvendo a estrutura 1D4K, onde o algoritmo proposto BRKGA obteve melhores resultados.

Nessa análise é possível perceber que o método proposto obteve uma distribuição padrão em 3 das 4 estruturas, com baixos valores de desvio padrão. A estrutura 1AJX difere em seus resultados, pois em algumas execuções obteve RMSD's acima da mediana, aumentando o desvio padrão da distribuição. Entretanto, os resultados foram muito bons para as 3 ferramentas, com valores, em todas as execuções, abaixo de $2,0$ Å. A estrutura 1D4K obteve resultados muito bons somente no método proposto BRKGA, a média de RMSD para as outras ferramentas foi acima de $15,0$ Å.

Figura 5.1: Diagrama de caixa das estrtururas (a) 1AJV, (b) 1AJX, (c) 1BV9 e (d) 1D4K, comparando os valores de RMSD para três ferramentas: Autodock Vina, BRKGA e Dockthor



As Tabelas 5.11, 5.12, 5.13 e 5.14 mostram os testes de Tukey para as estruturas 1AJV, 1AJX, 1BV9 e 1D4K, respectivamente. Essa análise é proposta por que cada execução gera amostras independentes, isso significa que uma observação não é influenciada pela anterior; a variância dentro de cada grupo é igual àquela dentro dos grupos, assim cada tratamento contribui de forma igual para a soma dos quadrados; e os valores de RMSD seguem uma distribuição normal. O teste faz comparação das ferramentas de duas a duas, definindo a menor diferença significativa utilizando a amplitude da distribuição. As tabelas apresentam o centro de variância entre duas ferramentas, o limite inferior e superior, e por fim, o valor P (referente a qualidade da amostragem).

Tabela 5.11: Teste de Tukey: 1AJV

Ferramentas	**Centro**	**Limite Inferior**	**Limite Superior**	**valor-p**
BRKGA-AUTODOCK	-0,614	-0,880	-0,348	0,00
DOCKTHOR-AUTODOCK	-0,201	-0,467	0,063	0,172
DOCKTHOR-BRKGA	0,412	0,146	0,678	0,001

Tabela 5.12: Teste de Tukey: 1AJX

Ferramentas	**Centro**	**Limite Inferior**	**Limite Superior**	**valor-p**
BRKGA-AUTODOCK	-0,554	-0,626	-0,482	0,00
DOCKTHOR-AUTODOCK	-0,416	-0,488	-0,344	0,00
DOCKTHOR-BRKGA	0,138	0,065	0,210	4,85

Tabela 5.13: Teste de Tukey: 1BV9

Ferramentas	**Centro**	**Limite Inferior**	**Limite Superior**	**valor-p**
BRKGA-AUTODOCK	-6,44	-7,289	-5,601	0,00
DOCKTHOR-AUTODOCK	18,17	17,329	19,017	0,00
DOCKTHOR-BRKGA	24,619	23,774	25,463	0,00

Tabela 5.14: Teste de Tukey: 1D4K

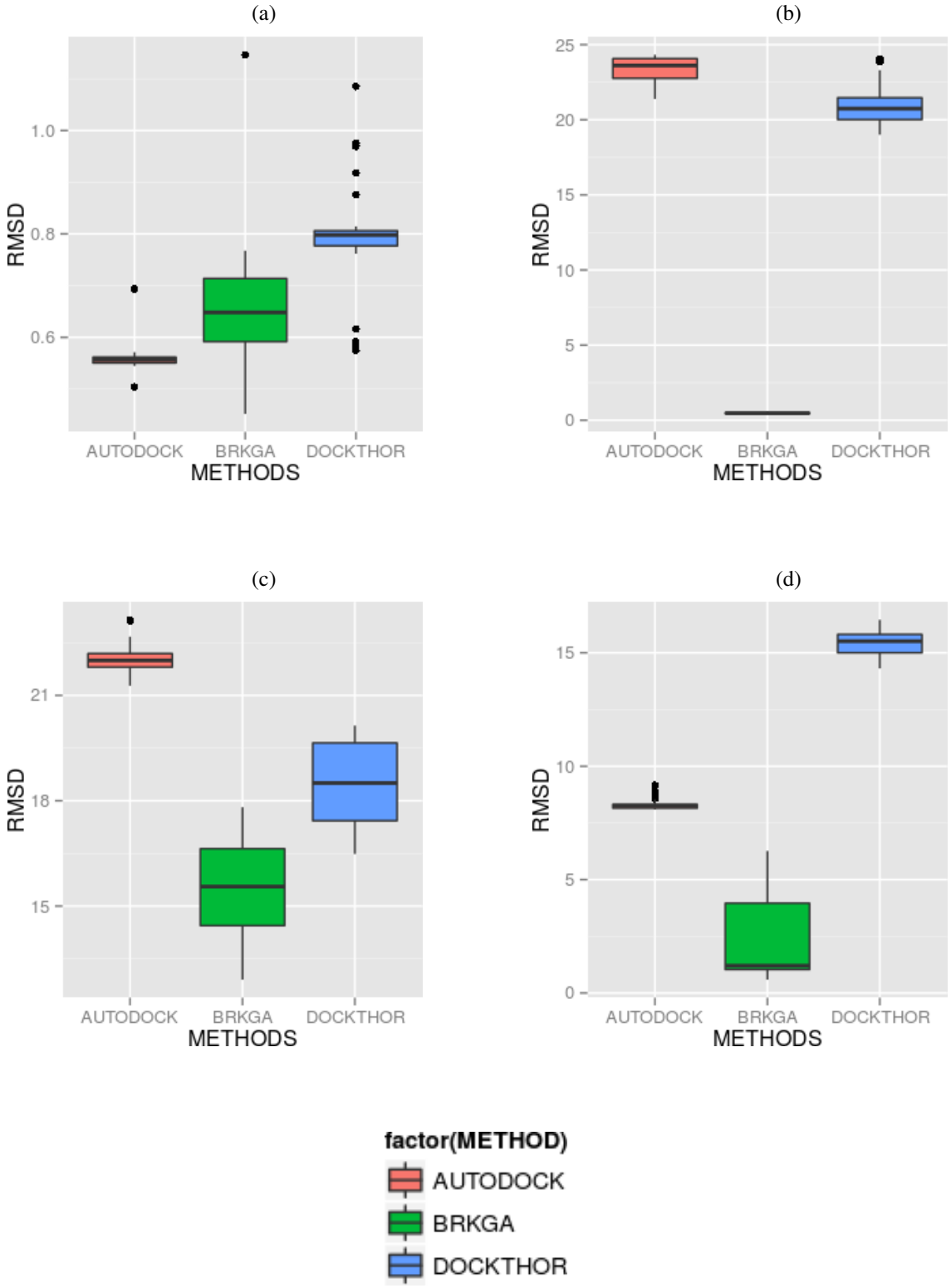
Ferramentas	**Centro**	**Limite Inferior**	**Limite Superior**	**valor-p**
BRKGA-AUTODOCK	-22,203	-22,525	-21,881	0,00
DOCKTHOR-AUTODOCK	-1,733	-2,055	-1,411	0,00
DOCKTHOR-BRKGA	20,470	20,148	20,148	0,00

A análise de variância busca validar o desempenho das ferramentas, assim, a Tabela 5.11 mostra que o desempenho entre as ferramentas AutoDock e Dockthor foi bem semelhante com valor-p (probabilidade de significância) de 10%, mostrando a similaridade dos resultados. O valor permaneceu em 0 % para a comparação com o BRKGA, mostrando que a amostragem é independente e validando o resultado superior apresentado no diagrama de caixa 5.1 (a). Na Tabela 5.12 a tabela mostra na comparação entre BRKGA-Dockthor um valor-p alto, devido aos valores discrepantes gerados pelo BRKGA, com RMSD's maiores que $0,75$ Å, ao mesmo que a ferramenta Dockthor alcançou em algumas execuções resultados melhores que o algoritmo BRKGA, assim as duas ferramentas se equivalem nos resultados, alcançando em diferentes execuções soluções semelhantes. Nas Tabelas 5.13 e 5.14 o P-valor se manteve em 0 %, validando o melhor desempenho do BRKGA apresentado no diagrama de caixas. A comparação das 3 ferramentas pelo teste de Tukey valida os diagramas apresentados.

Na Figura 5.2 são apresentados os diagramas de caixa para as estruturas 1G2K (a), 1HIV (b), 1HPX (c) e 1HTF (d). Na estrutura 1G2K (a), os resultados da ferramenta Autodock são em média melhores, com RMSD menores de $0,5$ Å. A variância do BRKGA nessa estrutura é maior, com dois resultados destoantes, já que uma execução obteve RMSD maior de $1,3$ Å, e outra que atingiu o melhor resultado de todas as execuções, $0,258$ como apresentado na Tabela 5.10. Para a estrutura 1HIV (b) os resultados do algoritmo BRKGA superam as outras duas ferramentas, assim como na estrutura 1HTF (d), apesar da maior variância nesse caso. Na estrutura 1HPX a ferramenta BRKGA possui amostras de menor RMSD e variância semelhante a ferramenta Dockthor.

Nessa análise é possível observar que o algoritmo BRKGA obteve melhores resultados em 3 das 4 estrututas. Apesar de uma variância maior nas estruturas 1HPX e 1HTF, a mediana manteve valores mais baixos que as outras estruturas. Nos testes realizados a ferramenta Dockthor obteve os piores resultados, médias acima de $20,0$ Å, para as estruturas 1HIV e 1HPX. Os resultados para a estrutura 1G2K para essa ferramenta destoam, atingindo valores altos de desvio padrão. Em relação a variância das soluções, a ferramenta Autodock Vina teve nas 4 estruturas melhores resultados, ou seja, nas 30 execuções os resultados obtidos pela ferramenta são bastante semelhantes. Com exceção da estrutura 1HIV em que o algoritmo BRKGA obteve valores baixos de RMSD e semelhantes dentre as execuções.

Figura 5.2: Diagrama de caixa das estrtururas 1G2K (a), 1HIV (b), 1HPX (c) e 1HTF (d), comparando os valores de RMSD para três ferramentas: Autodock Vina, BRKGA e Dockthor



O teste de Tukey para as estruturas 1G2K, 1HIV, 1HPX e 1HTF são apresentados nas Tabelas 5.15, 5.16, 5.17 e 5.18, respectivamente. Nas tabelas são apresentados as comparações entre as 3 ferramentas, mostrando o centro da variância e seus limites mínimos e máximos. Por fim, os valores-p, nas 4 tabelas são apresentados, onde os valores iguais a zero validam, dessa forma, os respectivos diagramas de caixa.

Tabela 5.15: Teste de Tukey: 1G2K

Ferramentas	**Centro**	**Limite Inferior**	**Limite Superior**	**valor-p**
BRKGA-AUTODOCK	0,094	0,035	0,153	0,00
DOCKTHOR-AUTODOCK	0,235	0,176	0,294	0,00
DOCKTHOR-BRKGA	0,140	0,081	0,199	0,00

Tabela 5.16: Teste de Tukey: 1HIV

Ferramentas	**Centro**	**Limite Inferior**	**Limite Superior**	**valor-p**
BRKGA-AUTODOCK	-22,861	-23,487	-22,236	0,00
DOCKTHOR-AUTODOCK	-2,318	-2,943	-1,693	0,00
DOCKTHOR-BRKGA	20,543	19,917	21,168	0,00

Tabela 5.17: Teste de Tukey: 1HPX

Ferramentas	**Centro**	**Limite Inferior**	**Limite Superior**	**valor-p**
BRKGA-AUTODOCK	-6,606	-7,268	-5,944	0,00
DOCKTHOR-AUTODOCK	-3,596	-4,258	-2,934	0,00
DOCKTHOR-BRKGA	3,009	2,347	3,671	0,00

Tabela 5.18: Teste de Tukey: 1HTF

Ferramentas	**Centro**	**Limite Inferior**	**Limite Superior**	**valor-p**
BRKGA-AUTODOCK	-6,150	-6,839	-5,461	0,00
DOCKTHOR-AUTODOCK	7,111	6,422	7,801	0,00
DOCKTHOR-BRKGA	13,262	12,572	13,951	0,00

A última estrutura a ser analisada estatisticamente é o complexo 2UPJ, cujo diagrama de caixa é apresentado na Figura 5.2. Os resultados de variância mostram o melhor desempenho da ferramenta BRKGA. As ferramentas Dockthor e Autodock obtiveram em algumas execuções valores de RMSD similares aos melhores resultados e em uma execução o valor mais baixo foi alcançado pelo Dockthor. Assim, na Tabela 5.19 são analisados

os centro e limites de variância das 3 ferramentas. O P-valor validam os valores apresentados no diagrama de caixa, garantindo os melhores valores do BRKGA para a amostragem apresentada.

Figura 5.3: Diagrama de caixa da estrturura 2UPJ, comparando os valores de RMSD para três ferramentas: Autodock Vina, BRKGA e Dockthor

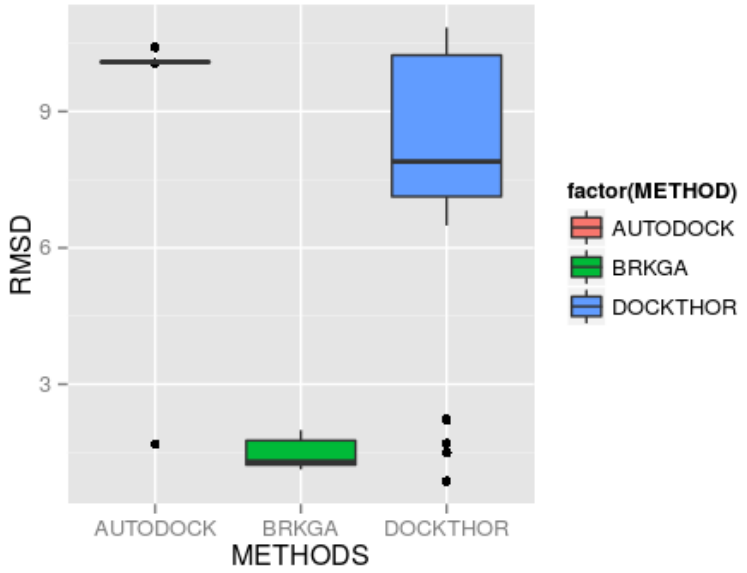


Tabela 5.19: Teste de Tukey: 2UPJ

Ferramentas	**Centro**	**Limite Inferior**	**Limite Superior**	**valor-p**
BRKGA-AUTODOCK	-8,365	-9,572	-7,159	0,00
DOCKTHOR-AUTODOCK	-2,193	-3,40	-0,986	0,00
DOCKTHOR-BRKGA	6,172	4,965	7,379	0,00

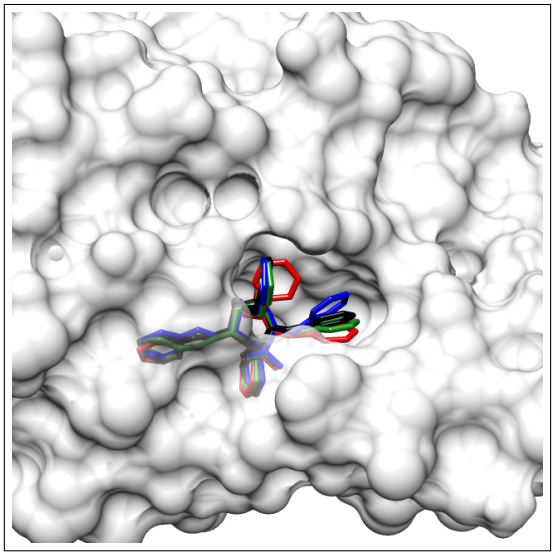
Nas Figuras 5.4 e 5.5 são apresentados os melhores resultados de cada execução para as três ferramentas testadas. As conformações de cada ligante podem ser observadas de acordo com as diferentes cores: preto, estrutura experimental, em verde, o resultado do algoritmo BRKGA, em vermelho, resultado do Autodock Vina, e em azul, o resultado da ferramenta Dockthor. É importante ressaltar que as ilustrações são referentes aos melhores resultados obtidos por cada ferramenta, assim, essa análise não reflete a média de execuções e distribuição das soluções, todavia a capacidade do algoritmo de atingir uma conformação ótima dentro do número de execuções proposto.

A Figura 5.4 (a) apresenta os resultados para a estrutura 1AJV, o melhor resultado de cada ferramenta é bastante aproximado, como apresentado no diagrama de caixa 5.1 (a), as três ferramentas obtiveram bons resultados nas 30 execuções. Para a estrutura

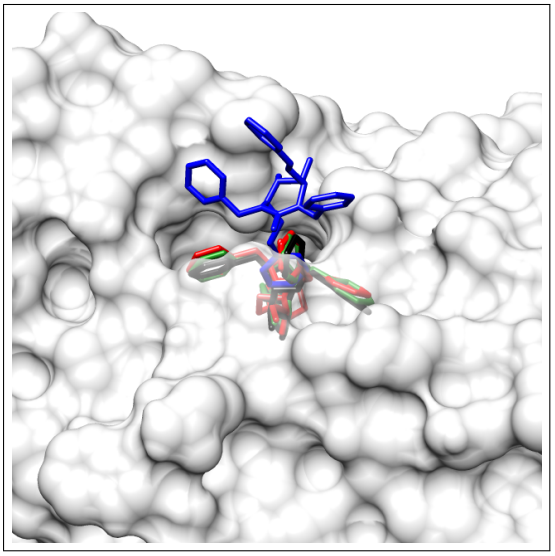
1AJX, Figura 5.4 (b), os resultados são similares para as melhores execuções. Todas as estruturas se posicionam no sítio de ligação da molécula receptora. Ao contrário, na estrutura 1BV9, Figura 5.4 (c), os resultados do algoritmo BRKGA foram bastante similares com a estrutura experimental, já os resultados das outras duas ferramentas, alcançaram uma conformação em que o complexo ligante se encontra foram do sítio de ligação. O mesmo ocorre na estrutura 1D4K, o algoritmo BRKGA obteve resultados melhores, encontrando o sítio de ligação e posicionando o ligante perto da conformação experimental.

Figura 5.4: Melhores resultados para os testes de comparação das estruturas 1AJV, 1AJX, 1BV9 e 1D4K. Em preto a estrutura experimental, em verde o melhor resultado do algoritmo BRKGA, em vermelho o resultado da ferramenta Autodock e em azul o resultado da ferramenta Dockthor.

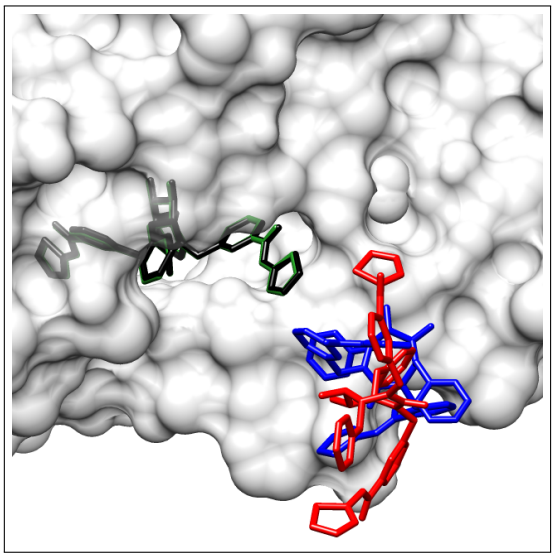
(a) 1AJV



(b) 1AJX



(c) 1BV9



(d) 1D4K

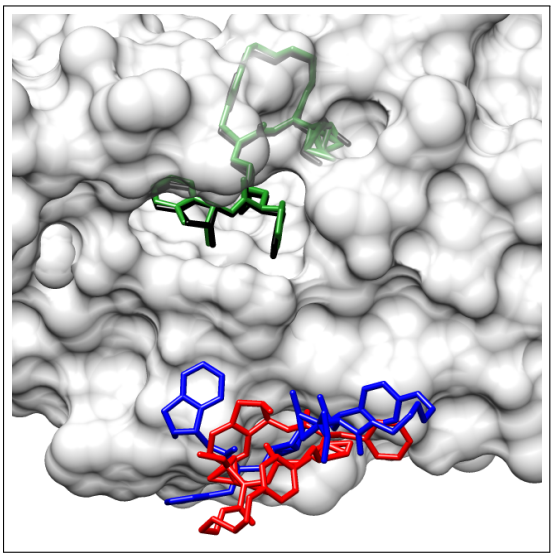
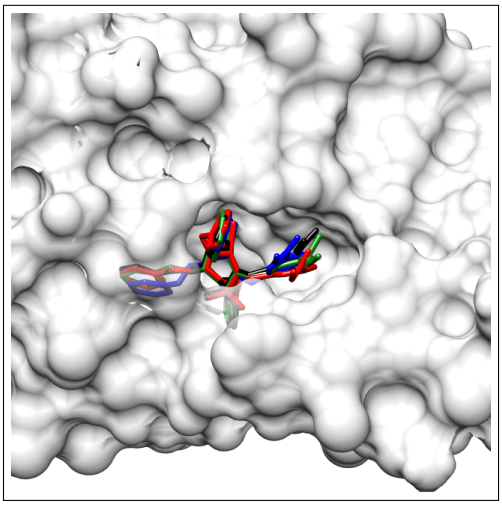
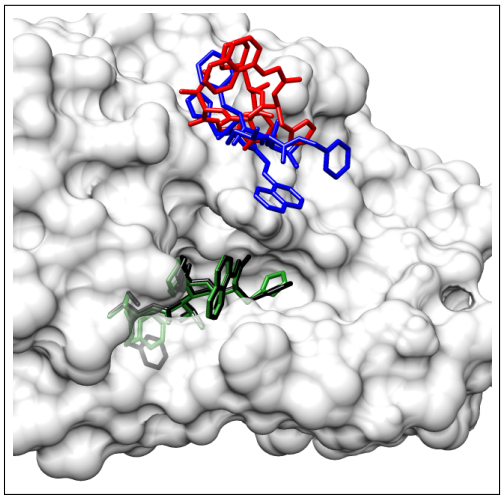


Figura 5.5: Melhores resultados para os testes de comparação das estruturas 1G2K, 1HIV, 1HPX, 1HTF e 2UPJ. Em preto a estrutura experimental, em verde a melhor resultado do algoritmo BRKGA, em vermelho o resultado da ferramenta Autodock e em azul o resultado da ferramenta Dockthor.

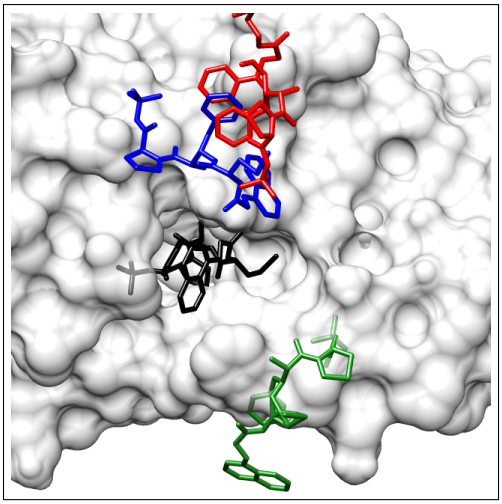
(a) 1G2K



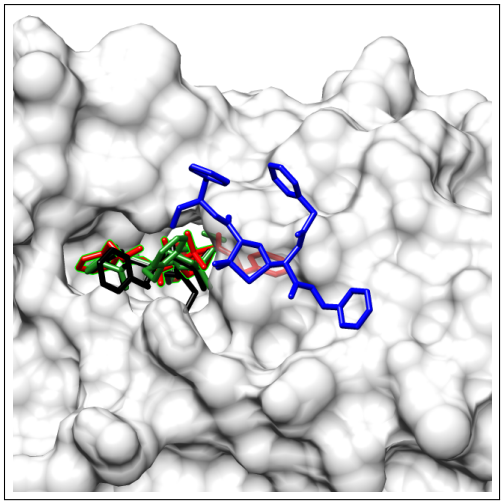
(b) 1HIV



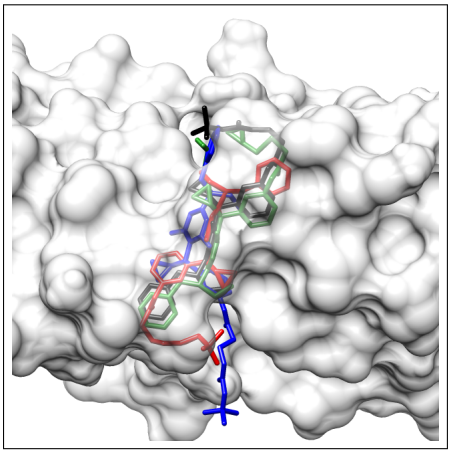
(c) 1HPX



(d) 1HTF



(e) 2UPJ



Na Figura 5.5 são apresentadas as estruturas 1G2K, 1HIV, 1HTX, 1HTF e 2UPJ. Na estrutura 1G2K, 5.5 (a), os três algoritmos obtiveram em suas melhores execuções conformações próximas dos dados experimentais. Já na estrutura 1HIV, 5.5 (b), o resultado do algoritmo BRKGA foram próximos do experimental, enquanto os resultados das outras duas ferramentas encontraram uma conformação fora do sítio de ligação da proteína. Para a estrutura 1HPX, 5.5 (c), as conformações encontradas foram, para as três ferramentas, fora do sítio de ligação. Devido a complexidade da estrutura, todos os valores de RMSD alcançados nas execuções foram acima de $10,0$ Å. Os resultados para as estruturas 1HTF e 2UPJ, 5.5 (d) e 5.5 (e), respectivamente, foram próximos do complexo cristalográfico.

5.7 Avaliação geral dos resultados

A avaliação geral dos resultados é analisada por meio dos infográficos apresentados nas Figuras 5.6, 5.7, 5.8 e 5.9. Essa representação mostra em porcentagem o sucesso obtido em cada conjunto de dados, representando como círculos cada solução encontrada. Foram selecionados 3 espaços com RMSD menores que $2,0$ Å, entre $2,0 - 3,0$ Å, e valores maiores que $3,0$ Å. Resultados de RMSD entre $1,5 - 2,0$ Åsão considerados bem sucedidos (HEVENER et al., 2009).

A Figura 5.6 apresenta os resultados do primeiro conjunto que contém 10 estruturas. Esse primeiro conjunto foi utilizado para a parametrização do algoritmo, testes de Atracamento rígido e comparação com as ferramentas Dockthor e Autodock Vina. Nos resultados de Atracamento flexível 80 % das estruturas alcançam valores de RMSD abaixo de $2,0$ Å. O melhor valor de RMSD foi de $0,1004$ Å, alcançado em uma execução da estrutura 1BV9, o pior valor, $8,204$ Å, para a estrutura 1HPX.

Na Figura 5.7 é apresentada a taxa de sucesso para o conjunto 2, as 20 estrutura são representadas por círculos, ditribuídos em 3 partes, valores com RMSD abaixo de $2,0$ Å, valores de RMSD entre $2 - 3$ Å, e valores de RMSD acima de $3,0$ Å, cujas porcentagens chegam a 30%, 35% e 35% respectivamente. O melhor resultado de RMSD de $0,769$ e pior de $9,534$ também são apresentados, valores alcançados pelas estruturas 1B6L e 1KZK, respectivamente.

A Figura 5.8 apresenta a taxa de sucesso para o terceiro conjunto, onde $66,6\%$ das estruturas atingiram RMSD abaixo de $2,0$ Å, cuja melhor estrutura, 1D4K, alcançou o valor de $0,747$ Å. Nesse conjunto, ainda, $26,6\%$ dos resultados obtiveram RMSD entre

Figura 5.6: CONJUNTO 1: Infográfico dos resultados de Atracamento flexível. Cada círculo representa uma solução, os círculos maiores representam a melhor (verde) e pior (vermelho) solução nesse conjunto de estruturas, a porcentagem é a taxa de sucesso dentro das faixas de valores de RMSD menores de 2 Å, e maiores de 3 Å

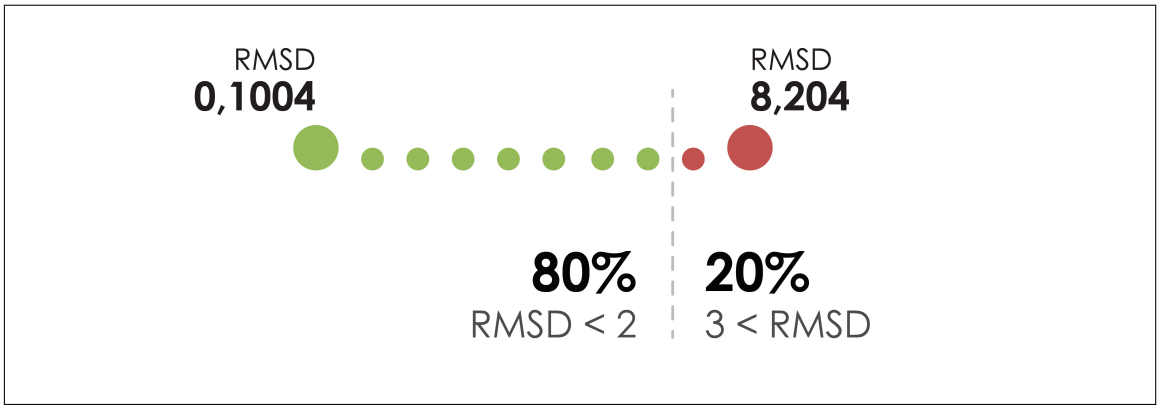
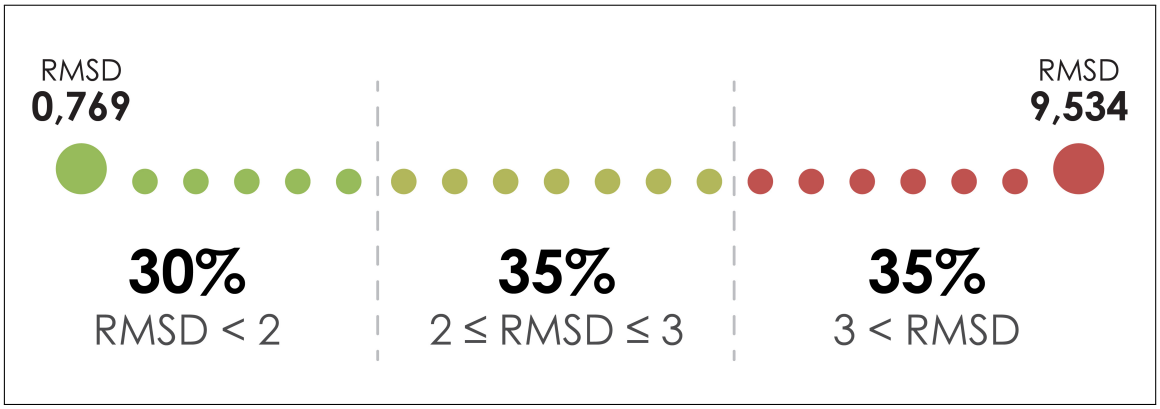


Figura 5.7: CONJUNTO 2: Infográfico dos resultados de Atracamento flexível. Cada círculo representa uma solução, os círculos maiores representam a melhor (verde) e pior (vermelho) solução nesse conjunto de estruturas, a porcentagem é a taxa de sucesso dentro das faixas de valores de RMSD menores de 2 Å, entre 2-3 Åe maiores de 3 Å



$2,0$ e $3,0$ Å. Apenas uma estrutura, 1B6J teve valores maiores de $3,0$ Å. .

A taxa de sucesso do último conjunto é apresentado na Figura 5.9. De 3 estruturas testadas, o algoritmo performou um bom resultado em somente uma delas, atingindo o valor de $0,510$ Å, no ligante ETH. Nas outras duas estruturas os valores médios de RMSD foram maiores de 7 Å. A complexidade dessas estruturas foram uma dificuldade pro algoritmo que obteve na melhor execução para o ligante NAD o valor de RMSD de $10,779$ Å.

A análise geral dos resultados mostra que em $62,2\%$ dos resultados o algoritmo obteve valores de RMSD abaixo de $2,0$ Å. Embora as taxa de acerto para o conjunto 4 tenha sido de apenas $33,3\%$, o algoritmo obteve bons resultados para o problema de predição da conformação proteína-ligante.

Figura 5.8: CONJUNTO 3: Infográfico dos resultados de Atracamento flexível. Cada círculo representa uma solução, os círculos maiores representam a melhor (verde) e pior (vermelho) solução nesse conjunto de estruturas, a porcentagem é a taxa de sucesso dentro das faixas de valores de RMSD menores de 2 Å, entre 2-3 Å, e maiores de 3 Å

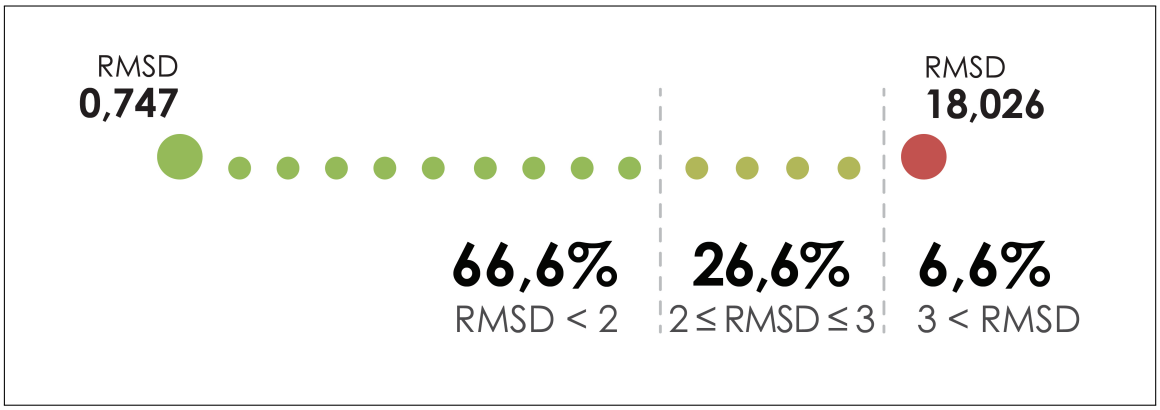
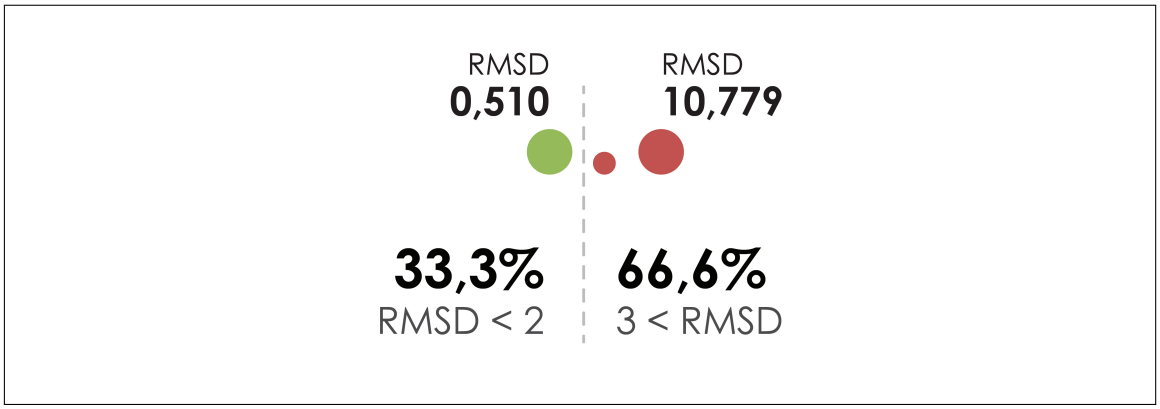


Figura 5.9: CONJUNTO 4: Infográfico dos resultados de Atracamento flexível. Cada círculo representa uma solução, os círculos maiores representam a melhor (verde) e pior (vermelho) solução nesse conjunto de estruturas, a porcentagem é a taxa de sucesso dentro das faixas de valores de RMSD menores de 2 Å, e maiores de 3 Å



6 CONCLUSÃO E TRABALHOS FUTUROS

Um Algoritmo de Chaves Aleatórias Viciadas foi desenvolvido como uma ferramenta de Atracamento Molecular que possa ser utilizada no processo de descobrimento de novos fármacos. Para desenvolver essa metodologia foram analisadas as dificuldades atuais, biológicas e computacionais. A ferramenta desenvolvida busca aliar os desafios biológicos analisando as interações físico-químicas envolvidas no processo de Atracamento com custo computacional e tempo factível para a resolução do problema.

A parametrização do algoritmo foi definida testando diferentes valores até que uma configuração ótima do algoritmo fosse alcançada. Estratégias distintas em relação a alguns tópicos foram testadas: a variação do tamanho da população inicial do algoritmo e métodos de reinicialização. A adição de uma grade de energia para o cálculo de energia e flexibilização parcial do receptor foi uma técnica proposta junto com a função de energia escolhida para avaliação da qualidade de previsão da orientação receptor-ligante. Além disso, foi proposto para o algoritmo a discretização do espaço de busca e a utilização do mesmo como critério de similaridade para a formação de agrupamentos, além de um algoritmo de competição local e global das soluções. Todas as técnicas foram gradualmente adicionadas ao algoritmo, seguindo metodologias desenvolvidas na área de Atracamento Molecular e analisando as dificuldades encontradas no algoritmo. As vantagens dessas técnicas englobam a melhor diversificação das soluções dentro do espaço de busca, o uso de técnicas de agrupamento, nas quais diferentes parâmetros puderam ser testados, e técnicas de competição que resultaram em melhorares soluções.

Para testar esse algoritmo foram selecionados 50 estruturas, em sua maioria baseadas na molécula HIV-protease. Dentro desse conjunto de estruturas os testes contaram também com complexos baseados na proteína receptora alvo da macro-molécula de tuberculose, como um conjunto que fosse distinto dos testes já aplicados. Os conjuntos foram separados de acordo com o tamanho da molécula ligante e das diferentes moléculas receptoras. O primeiro conjunto foi organizado com 10 estruturas, com resolução abaixo de $2,0$ Å, o segundo conjunto com 21 estruturas, com resolução abaixo de $3,0$ Å, o terceiro conjunto com 17 estruturas, de tamanhos e resoluções variadas, e o último conjunto com 3 estruturas e diferente molécula receptora. Para todos os conjuntos de dados foram executados exaustivos testes, com 30 execuções para cada estrutura e 1 milhão de avaliações de energia como critério de parada do algoritmo, esses valores foram escolhidos de acordo com testes executados na literatura.

Inicialmente o algoritmo foi testado em uma simplificação do problema, conhecido como Atracamento rígido. Nesses testes são eliminadas as conformações dos ângulos internos diedrais da molécula ligante, assim, pôde-se avaliar a acurácia do algoritmo para posterior aumento da dificuldade do problema ao incluir a conformação do ligante. Os resultados obtidos mostraram que o algoritmo consegue valores baixos de RMSD, em 90% abaixo de $2,0$ Å, condizendo seus resultados com as estruturas obtidas experimentalmente. Esses resultados motivaram a utilização da técnica para testes em que se adiciona a conformação do ligante.

Os resultados de Atracamento flexível obtiveram 80% de sucesso para o primeiro conjunto com valores de RMSD abaixo de $2,0$ Å, 75% dos resultados do segundo conjunto tiveram valores abaixo de $3,0$ Å, e $33,3\%$ dos resultados do conjunto 4 obtiveram RMSD abaixo de $2,0$ Å. Os resultados mostram que o algoritmo consegue prever uma orientação ligante-receptor condizendo com a estrutura cristalográfica em $62,6\%$ dos casos.

A inclusão dos métodos foi gradual, onde se buscou primeiramente uma taxa de sucesso de mais de 80% para o primeiro conjunto de dados. Primeiramente o algoritmo BRGKA foi testado com diferentes populações, logo foi adicionada a discretização do campo de busca, depois o agrupamento de soluções e por fim a competição entre soluções. Cada etapa obteve melhores resultados até que se alcançou a taxa de sucesso de 90% para o primeiro conjunto e uma parametrização ótima para o algoritmo.

Um estudo comparativo de três ferramentas foi realizado, comparando a ferramenta desenvolvida com dois *softwares*: Autodock Vina e Dockthor, em prol de comparar a metodologia e função de energia empregadas. O algoritmo limitou os graus conformacionais da estrutura ligante em 10 ângulos, número de ângulos diedrais mínimo encontrado em todas as moléculas testadas, esse limite foi aplicado para os testes nas três ferramentas. Os testes buscaram utilizar a parametrização ótima de cada ferramenta, de forma a atingir os melhores resultados em cada execução. Os resultados de comparação mostram que a ferramenta desenvolvida supera os resultados, do conjunto de 10 estruturas testadas, em 90% dos casos.

O estudo comprovou que a implementação de uma técnica que alia a discretização do espaço de busca com um algoritmo BRKGA, com técnicas de agrupamento e competições, além e a utilização da função de energia da ferramenta Autodock Vina, é um ferramenta eficaz para o problema de Atracamento Molecular flexível. O algoritmo proposto aproxima o problema real ao utilizar uma função de energia que considera os

efeitos entrópicos e solvatação ao adicionar a flexibilidade parcial da molécula receptora; o ligante é considerado flexível, considerando certos graus de liberdade da molécula.

Os resultados obtidos guiam para estudos de aperfeiçoamento do algoritmo desenvolvido, embora a ferramenta tenha obtido bons resultados para as estruturas cujo complexo receptor é a proteína HIV-protease, os resultados utilizando outra biomolécula receptora não foram satisfatórios, com sucesso em apenas 33% dos testes. Assim, é necessário que o algoritmo seja validado em um conjunto maior e mais diverso, com diferentes moléculas receptoras e ligantes.

Uma proposta de aperfeiçoamento é a utilização metodologia de otimização de multi-objetivo, nesse caso haveriam mais de uma função a ser otimizada. Uma opção seria separar a função de energia em intra e extra moleculares, ou ainda, calcular separadamente os termos ligados e termos não-ligados. Nessa abordagem o algoritmo convergiria para uma solução única de conformação que obtivesse o mínimo em ambas as funções.

O algoritmo pode melhorar ainda os efeitos relacionados aos solventes, já que as estruturas passaram, na fase de preparação, por uma remoção de átomos e termos não-ligados, considerando-os no vácuo, o que biologicamente não acontece. A adição de maiores graus de liberdade para o ligante é um desenvolvimento necessário para o algoritmo, já que alguns dos ligantes testados possuem mais do que os 10 ângulos diedrais aos quais foram limitados pelo método. A adição de flexibilidade do receptor é uma proposta mais avançada, pois aumentaria a complexidade e custo computacional, porém trataria do problema de uma forma biologicamente mais acurada.

Outra proposta para trabalhos futuros são testes para descoberta e planejamento de fármacos, onde não se conheça a estrutura cristalográfica. Esses testes serviriam para aplicação da ferramenta como auxílio em Triagem Virtual e para o Atracamento de complexos protótipos e verificariam a capacidade de prever a orientação ligante-receptor para novos fármacos.

REFERÊNCIAS

ALONSO, H.; BLIZNYUK, A. A.; GREADY, J. E. Combining docking and molecular dynamic simulations in drug design. **Medicinal research reviews**, v. 26, n. 5, p. 531–568, 2006.

ALTMAN, R. B.; DUGAN, J. M. Defining bioinformatics and structural bioinformatics. In: **Struct, Bioinf.** [S.l.: s.n.], 2003. v. 44, chp. 1, p. 1–14.

ANDREI, R. M. et al. Intuitive representation of surface properties of biomolecules using bioblender. **BMC bioinformatics**, v. 13, n. 4, p. 1, 2012.

ANFINSEN, C. B. Principles that govern the folding of protein chains. **Science**, v. 181, n. 4096, p. 223–230, 1973.

BALDWIN, R. L. Dynamic hydration shell restores kauzmann's 1959 explanation of how the hydrophobic factor drives protein folding. **Nat. Acad. Sci.**, v. 111, n. 36, p. 13052–13056, 2014.

BARREIRO, E. J.; FRAGA, C. A. M. **Química Medicinal: As bases moleculares da ação dos fármacos**. [S.l.: s.n.], 2014.

BEAN, J. C. Genetic algorithms and random keys for sequencing and optimization. **ORSA J. Comp.**, v. 6, n. 2, p. 154–160, 1994.

BENITE, A. M. C.; MACHADO, S. d. P.; BARREIRO, E. J. Uma visão da química bioinorgânica medicinal. **Química Nova**, v. 30, p. 2062–2067, 00 2007.

BERMAN, H. M. et al. The protein data bank. **Nucl. Acids Res.**, Oxford University Press, v. 28, n. 1, p. 235–242, 2000.

BISSANTZ, C.; FOLKERS, G.; ROGNAN, D. Protein-based virtual screening of chemical databases. 1. evaluation of different docking/scoring combinations. **J. Med. Chem.**, v. 43, n. 25, p. 4759–4767, 2000.

BLUM, C. et al. Hybrid metaheuristics in combinatorial optimization: A survey. **Appl. Soft Comput.**, v. 11, n. 6, p. 4135 – 4151, 2011.

BÖHM, H. J. Ludi: rule-based automatic design of new substituents for enzyme inhibitor leads. **J. of Comp-Aid. Mol. Des.**, v. 6, n. 6, p. 593–606, 1992.

BÖHM, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. **J. Comput.-Aided Mol. Des.**, v. 8, n. 3, p. 243–256, 1994.

BROOIJMANS, N.; KUNTZ, I. D. Molecular recognition and docking algorithms. **Annual Review of Biophysics and Biomolecular Structure**, v. 32, n. 1, p. 335–373, 2003.

BROOIJMANS, N.; KUNTZ, I. D. Molecular recognition and docking algorithms. **Byophys. Biomol. Struct.**, n. 32, p. 335–373, 2003.

BROOKS, B. R. Charmm: A program for macromolecular energy, minimization and dynamics calculations. **J. Comput. Chem.**, v. 4, n. 2, p. 187–217, 1983.

CAMACHO, E. L. et al. Solving molecular flexible docking problems with metaheuristics: A comparative study. **Appl. Soft Comput.**, v. 28, n. 28, p. 379–393, 2014.

CHANDRIKA, B. R.; SUBRAMANIAN, J.; SHARMA, S. D. Managing protein flexibility in docking and its applications. **Drug discovery today**, v. 14, n. 7, p. 394–400, 2009.

CHOTHIA, C.; LESK, A. M. The relation between the divergence of sequence and structure in proteins. **The EMBO Journal**, v. 5, n. 5, p. 823–826, 1992.

CLAUSSEN, H.; BUNING CM., R. M.; LENGAUER, T. Flexe: efficient molecular docking considering protein structure variations1. **J. Mol. Bio.**, v. 308, n. 2, p. 377 – 395, 2001.

COMBS, S. A. et al. Small-molecule ligand docking into comparative models with rosetta. **Nat. Prot.**, v. 8, n. 7, p. 1277–1299, 2013.

CORNELL, W. D.; CIEPLAK, P. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. **J. Amer. Chem. Soc.**, v. 117, n. 19, p. 5179–5197, 1995.

COZZINI, P. et al. Target flexibility: An emerging consideration in drug discovery and design†. **J. Med. Chem.**, v. 51, n. 20, p. 6237–6255, 2008.

DEVI, R. V.; SIVA, S. S.; COUMAR, M. S. Evolutionary algorithms for de novo drug design – a survey. **Appl. Soft Comput.**, v. 27, p. 543 – 552, 2015.

DEWITTE, R. S.; SHAKHNOVICH, E. I. Smog: de novo design method based on simples, fast, and accurate free energy estimates. 1. methodology and supporting evidence. **J. A. Chem. Soc.**, v. 118, n. 47, p. 11733–11744, 1996.

DUNN, M. F. Protein-ligand interactions: General description. **Eng. Life Sci.**, v. 2, n. 3, p. 22–101, 2007.

EISENSTEIN, M.; KATZIR, E. K. On proteins, grids, correlations, and docking. **Comp. Rendus Bio.**, v. 327, n. 5, p. 409 – 420, 2004.

FISCHER, M. et al. Incorporation of protein flexibility and conformational energy penalties in docking screens to improve ligand discovery. **Nat. Chem.**, v. 6, n. 7, p. 575–583, 2014.

FRENKEL, D.; SMIT, B. Chapter 7 - free energy calculations. In: FRENKEL, D.; ; SMIT, B. (Ed.). **Understanding Molecular Simulation (Second Edition)**. Second edition. San Diego: Academic Press, 2002. p. 167 – 200.

GABB, H. A.; JACKSON, R. M.; STERNBERG, M. J. Modelling protein docking using shape complementarity, electrostatics and biochemical information. **J. of Mol. Bio.**, v. 272, n. 1, p. 106 – 120, 1997.

GODOY, M. J. G. et al. Solving molecular docking problems with multi-objective metaheuristics. **Molecules**, v. 20, n. 20, p. 10154–10155, 2015.

GOHLKE, H.; HENDLICH, M.; KLEBE, G. Knowledge-based scoring function to predict protein–ligand interactions. **J. Mol. Biol**, v. 2000, p. 337–356, 2000.

GOODFORD, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. **J. of Med. Chem.**, ACS Publications, v. 28, n. 7, p. 849–857, 1985.

GOULART, N. et al. Biased random-key genetic algorithm for fiber installation in optical network optimization. In: **2011 IEEE Congress of Evolutionary Computation (CEC)**. [S.l.: s.n.], 2011. p. 2267–2271.

GUEDES, I. A.; MAGALHãES, C. S. d.; DARDENNE, L. E. Receptor-ligand molecular docking. **International Union for Pure and Applied Biophysics (IUPAB) and Springer-Verlag Berlin Heidelberg 2013**, n. 6, p. 75–87, 2013.

GUNSTEREN, B. W. F. v. **Groningen Molecular Simulation (GROMOS) Library Manual**. [S.l.], 1987.

HALGREN, T. A. Merck molecular force field. i. basis, form, scope, parameterization, and performance of mmff94. **J. Comput. Chem.**, v. 17, n. 5-6, p. 490–519, 1996.

HALGREN, T. A. Merck molecular force field. ii. mmff94 van der waals and electrostatic parameters for intermolecular interactions. **J. Comput. Chem.**, v. 17, n. 5-6, p. 520–552, 1996.

HALPERIN, I. et al. Principles of docking: An overview of search algorithms and a guide to scoring functions. **Proteins: Struct., Funct., Bioinf.**, v. 47, n. 4, p. 409–443, 2002.

HEVENER, K. E. et al. Validation of molecular docking programs for virtual screening against dihydropteroate synthase. **J. Chem. Inf. Mol.**, v. 49, n. 2, p. 444–460, 2009.

HUANG, S.; ZOU, X. Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. **Proteins: Struct., Funct., Bioinf.**, v. 66, n. 2, p. 399–421, 2007.

HUANG, S.; ZOU, X. Advances and challenges in protein-ligand docking. **Int. J. Mol. Sci.**, v. 11, n. 8, p. 3016, 2010.

IRWIN, J. J.; SHOICHET, B. K. Zinc - a free database of commercially available compounds for virtual screening. **J. Chem. Inf. Model.**, v. 45, n. 1, p. 177–182, 2005.

JACKSON, R. M.; GABB, H. A.; STERNBERG, M. J. Rapid refinement of protein interfaces incorporating solvation: application to the docking problem1. **J. Mol. Bio.**, v. 276, n. 1, p. 265–285, 1998.

JANIN, J. Protein-protein docking tested in blind predictions: the capri experiment. **Mol. BioSys.**, v. 6, n. 12, p. 2351–2362, 2010.

JANSON, S.; MERKLE, D.; MIDDENDORF, M. Molecular docking with multi-objective particle swarm optimization. **Appl. Soft Comput.**, v. 8, n. 1, p. 666–675, 2008.

JASSADAPAKORN, C.; CHONGSTITVATANA, P. Self-adaptation mechanism to control the diversity of the population in genetic algorithm. **arXiv preprint arXiv:1109.0085**, 2011.

JONES, G.; WILLETT, P. Docking small-molecule ligands into active sites. **C. Opin. Biotech.**, v. 6, n. 6, p. 652–656, 1995.

JONES, G.; WILLETT, P.; GLEN, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. **J. Mol. Bio.**, v. 245, n. 1, p. 43–53, 1995.

JONES, G. et al. Development and validation of a genetic algorithm for flexible docking1. **J. Mol. Bio.**, v. 267, n. 3, p. 727 – 748, 1997.

JUDSON, R. S. et al. Docking flexible molecules: A case study of three proteins. **J. Comput. Chem.**, v. 16, n. 11, p. 1405–1419, 1995.

KELLENBERGER, E. et al. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. **Proteins: Struct., Funct., Bioinf.**, v. 57, n. 2, p. 225–242, 2004.

KEPPEL, G. **Design and analysis: A researcher's handbook**. [S.l.]: Prentice-Hall, Inc, 1991.

KITCHEN, D. B.; FURR J. R., B. J. Docking and scoring in virtual screening for drug discovery: methods and applications. **Nat Rev Drug Discov**, v. 3, n. 2, p. 935 – 949, 2004.

KOZAKOV, D. et al. Optimal clustering for detecting near-native conformations in protein docking. **Biophysical journal**, v. 89, n. 2, p. 867–875, 2005.

KUKKONEN, S.; LAMPINEN, J. Gde3: The third evolution step of generalized differential evolution. In: **IEEE Congress on Evolutionary Computation (CEC'2005)**. [S.l.: s.n.], 2005. p. 443–450.

KUNTZ, D. Struc.-based strat. drug design dis. **Science**, v. 257, n. 257, p. 1078–1082, 1992.

KUNTZ, I. D. et al. A geometric approach to macromolecule-ligand interactions. **J. Mol. Bio.**, v. 161, n. 2, p. 269 – 288, 1982.

LADBURY, J. E. Just add water! the effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. **Chemistry and Biology**, v. 3, n. 12, p. 973 – 980, 1996.

LAMEIJER, E. W. et al. Evolutionary algorithms in drug design. **Nat. Comp.**, n. 4, p. 177–243, 2005.

LEACH, A. R. Ligand docking to proteins with discrete side-chain flexibility. **J. Mol. Bio.**, v. 235, n. 1, p. 345–356, 1994.

LEHNINGER, A.; NELSON, D. L.; COX, M. M. **Principles of Biochemistry**. 4. ed. New York, NY, USA: W.H. Freeman, 2004.

LESK, A. M. **Introduction to Bioinformatics**. 2. ed. [S.l.]: Oxford University Press, 2005.

LUSCOMBE, N. M.; GREENBAUM, D.; GERSTEIN, M. What is Bioinformatics? A proposed definition and overview of the field. **Methods Inf. Med.**, New Haven, CT, USA., v. 40, n. 4, p. 346–358, 2001.

LUTY, B. A. et al. A molecular mechanics/grid method for evaluation of ligand-receptor interactions. **J. Comput. Chem.**, v. 16, n. 4, p. 454–464, 1995.

LóPEZ-CAMACHO, E. et al. jmetalcpp: optimizing molecular docking problems with a c++ metaheuristic framework. **Bioinformatics**, 2013.

MACHADO, K. S. et al. Fredows: a method to automate molecular docking simulations with explicit receptor flexibility and snapshots selection. **BMC genomics**, v. 12, n. 4, p. 1, 2011.

MAGALHAES, C. S. D. **Algoritmos Geneticos para o Problema de Docking Proteina-Ligante**. Thesis (Doutorado) — Laboratorio Nacional de Computacao Cientifica, Petropolis, RJ, Brasil, 2006.

MAGALHAES, C. S. d.; BARBOSA, H. J.; DARDENNE, L. E. A genetic algorithm for the ligand-protein docking problem. **Gen. Mol. Bio.**, v. 27, p. 605 – 610, 00 2004.

MAGALHÃES, C. S. de et al. A dynamic niching genetic algorithm strategy for docking highly flexible ligands. **Inf. Sci.**, v. 289, p. 206–224, 2014.

MEIER, R. et al. Paradocks: A framework for molecular docking with population-based metaheuristics. **J. Chem. Inf. Model.**, v. 50, n. 5, p. 879–889, 2010.

MENG, E. C.; SHOICHET, B. K.; KUNTZ, I. D. Automated docking with grid-based energy evaluation. **J. Comp. Chem.**, v. 13, n. 4, p. 505–524, 1992.

MITCHELL, J. B. O. et al. Bleep—potential of mean force describing protein–ligand interactions: I. generating potential. **J. Comput. Chem.**, v. 20, n. 11, p. 1165–1176, 1999.

MITCHELL, J. B. O. et al. Bleep—potential of mean force describing protein–ligand interactions: Ii. calculation of binding energies and comparison with experimental data. **J. Comp. Chem.**, v. 20, n. 11, p. 1177–1185, 1999.

MORRIS, G. M. et al. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. **J. Comp. Chem.**, v. 19, n. 14, p. 1639–1662, 1998.

MORRIS, G. M. et al. Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. **J. Comput. Chem.**, v. 30, n. 16, p. 2785–2791, 2009.

MUEGGE, I. A knowledge-based scoring function for protein-ligand interactions: Probing the reference state. **Persc. Drug Disc. and Des.**, v. 20, n. 1, p. 99–114, 2000.

MUEGGE, I. Effect of ligand volume correction on pmf scoring. **J. Comput. Chem.**, v. 22, n. 4, p. 418–425, 2001.

MUEGGE, I. Pmf scoring revisited. **J. Med. Chem.**, v. 49, n. 20, p. 5895–5902, 2006.

MUEGGE, I.; MARTIN, Y. C. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. **J. Med. Chem.**, v. 42, n. 5, p. 791–804, 1999.

MURRAY, W. Illustration of current challenges in molecular docking. **Struct.-Based Drug Discovery**, v. 5, p. 201, 2007.

NEBRO, A. et al. Smpso: A new pso-based metaheuristic for multi-objective optimization. In: **2009 IEEE Symposium on Computational Intelligence in Multicriteria Decision-Making**. [S.l.: s.n.], 2009. p. 66–73.

NOLTING, A. et al. Pharmacokinetic-pharmacodynamic modeling of the antibiotic effect of piperacillin in vitro. **Pharma.l Res.**, v. 13, n. 1, p. 91–96, 1996.

NORONHA, T. F.; RESENDE, M. G.; RIBEIRO, C. C. A biased random-key genetic algorithm for routing and wavelength assignment. **J. Glob. Optim.**, v. 50, n. 3, p. 503–518, 2011.

O'BOYLE, N. M. et al. Open babel: An open chemical toolbox. **J. Cheminf.**, v. 3, n. 1, p. 1–14, 2011.

PANG, Y. P.; KOZIKOWSKI, A. P. Prediction of the binding sites of huperzine a in acetylcholinesterase by docking studies. **J. Comp.-aid. molecular design**, v. 8, n. 6, p. 669–681, 1994.

PAULING, L.; DELBRUCK, M. The nature of the intermolecular forces operative in biological processes. **Science**, American Association for the Advancement of Science, v. 92, n. 2378, p. 77–79, 1940.

Pauling, L.; Delbruck, M. The nature of the intermolecular forces operative in biological processes. **Science**, v. 92, p. 77–79, 1940.

PEARLMAN, D. A.; CHARIFSON, P. S. Are free energy calculations useful in practice? a comparison with rapid scoring functions for the p38 map kinase protein system. **J. Med. Chem.**, v. 44, n. 21, p. 3417–3423, 2001.

PRASETYO, H.; FAUZA G., A. Y.; LEE, S. H. Survey on applications of biased-random key genetic algorithms for solving optimization problems. In: **Industrial Engineering and Engineering Management (IEEM)**. [S.l.: s.n.], 2015. p. 863–870.

RAREY, M. et al. A fast flexible docking method using an incremental construction algorithm. **J. Mol. Chem.**, Elsevier, v. 261, n. 3, p. 470–489, 1996.

RESENDE, M. G. C. Biased random-key genetic algorithms with applications in telecommunications. **J. Span. Soc. of Stat. Oper. Res.**, v. 20, n. 1, p. 130–153, 2012.

RICHARDSON, J. S. Advances in protein chemistry. In: **The Anatomy and Taxonomy of Protein Structure**. [S.l.]: Academic Press, 1981. v. 34, p. 167 – 339.

ROGERO, S. O. et al. Teste in vitro de citotoxicidade: estudo comparativo entre duas metodologias. **Materials Research**, v. 6, p. 317–320, 2003.

SARENI, B.; KRAHENBUHL, L. Fitness sharing and niching methods revisited. **Trans. Evol. Comp**, v. 2, n. 3, p. 97–106, 1998.

SCHNEIDER, G.; BöHM, H. J. Virtual screening and fast automated docking methods. **Drug Discovery Today**, v. 7, n. 1, p. 64 – 70, 2002.

Schrödinger, LLC. The pymol molecular graphics system, version 1.8, schrödinger, llc. 2015.

SIMONSEN, M. et al. Gpu-accelerated high-accuracy molecular docking using guided differential evolution. **Nat. Comp. Ser.**, p. 349–368, 2013.

SOUSA, S. et al. Protein-ligand docking in the new millennium a retrospective of 10 years in the field. **Curr. Med. Chem.**, v. 20, n. 18, p. 2296–2314, 2013.

TANGPATTANAKUL, P.; JOZEFOWIEZ, N.; LOPEZ, P. Biased random key genetic algorithm with hybrid decoding for multi-objective optimization. In: **FedCSIS, 2013 Federated Conference on**. [S.l.: s.n.], 2013. p. 393–400.

TEAGUE, S. J. Implications of protein flexibility for drug discovery. **Nature R. Drug Disc.**, v. 2, n. 7, p. 527–541, 2003.

TEODORO, M. L.; KAVRAKI, L. E. Conformational flexibility models for the receptor in structure based drug design. **Current pharmaceutical design**, Bentham Science Publishers, v. 9, n. 20, p. 1635–1648, 2003.

TRAMONTANO, A. Protein structure prediction: Concepts and applications. **PROTEOMICS**, v. 6, n. 19, p. 5364–5364, 2006.

TRAMONTANO, A.; LESK, A. M. **Protein structure prediction: concepts and applications**. 1. ed. Weinheim, Germany: [s.n.], 2006.

TROTT, O.; OLSON, A. J. Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. **J. Comput. Chem.**, v. 31, n. 2, p. 455–461, 2010.

VERDONK, M. L. et al. Improved protein-ligand docking using gold. **Proteins: Struct., Funct., Bioinf.**, v. 52, n. 4, p. 609–623, 2003.

VERLI, H. Níveis de informação biológica. In: **Bioinformática: da Biologia à Flexibilidade Moleculares**. [S.l.: s.n.], 2014. chp. 2, p. 14–37.

VREVEN, T. et al. Performance of zdock in capri rounds 20–26. **Proteins: Struct., Funct., Bioinf.**, v. 81, n. 12, p. 2175–2182, 2013.

WALTERS, W. P.; STAHL, M. T.; MURCKO, M. A. Virtual screening—an overview. **Drug Discovery Today**, v. 3, n. 4, p. 160–178, 1998.

WANG, R.; LU, Y.; WANG, S. Comparative evaluation of 11 scoring functions for molecular docking. **J. Med. Chem.**, v. 46, n. 12, p. 2287–2303, 2003.

WEI, B. Q. et al. Testing a flexible-receptor docking algorithm in a model binding site. **J. Mol. Bio.**, v. 337, n. 5, p. 1161 – 1182, 2004.

WEINER, S. J. et al. A new force field for molecular mechanical simulation of nucleic acids and proteins. **J. A. Chem. Soc.**, v. 106, n. 3, p. 765–784, 1984.

WEISE, T. Global optimization algorithms-theory and application. **Self-Published,**, Citeseer, p. 25–26, 2009.

WONG, C. F. Flexible ligand-flexible protein docking in protein kinase systems. **J. Bioch. Biophys. Acta - Prot. Proteo.**, v. 1784, n. 1, p. 244–251, 2008.

ZHANG, C. et al. A knowledge-based energy function for protein-ligand, protein-protein, and protein-dna complexes. **J. Med. Chem.**, v. 48, n. 7, p. 2325–2335, 2005.

ZHANG, S. et al. Dovis: an implementation for high-throughput virtual screening using autodock. **Bmc Bioinformatics**, v. 9, n. 1, p. 1, 2008.

ZHANG, Y. et al. Using ensemble methods to deal with imbalanced data in predicting protein-protein interactions. **Comput. Biol. Chem.**, v. 36, p. 36–41, 2012.